

Chapter 2: Introduction to Time Series Analysis

Jan Skaloud* & Stéphane Guerrier†

*École Polytechnique Fédérale de Lausanne; †University of Geneva

This document was prepared with the help of Dr. Roberto Molinari, Dr. Mucyo Karemera, Gaetan Bakalli, Lionel Voirol,
Haotian Xu & Yuming Zhang

Material available online: XXXXXXXXXXXX



EPFL - January 2020

Introduction

Definition 2.1 (Time Series or TS).

A TS is a **stochastic process**, (i.e. a sequence of Random Variables (RV)), defined on a common probability space denoted as $(X_t)_{t=1,\dots,T}$ (i.e. X_1, X_2, \dots, X_T). Note that the time t is not continuous and belongs to discrete index sets. Therefore, we implicitly assume that:

- t is not random e.g. the time at which each observation is measured is known, and
- the time between two consecutive observations is constant.

Remark 1 (descriptive analysis).

In the classical time series theory, it is often useful to gain insight about a process by performing a descriptive analysis. While this approach may not be appropriate with inertial sensors, we shall briefly review it in the next few slides.

Introduction

Definition 2.2 (Descriptive Analysis).

Most time series analysis starts with displaying the data as a line plot on a graph. Time on the x-axis and variable on y-axis. Such graphs are often useful to assess various properties of the data at hand.

Time Series Graph/Plot:

- When recording values of the same variable over an extended period of time, it is difficult to discern any trend or pattern by simply looking at the values.
- However, when these data points are displayed on a plot (time on x-axis and X_t on y-axis), some features jump out.
- TS graph make trends easy to spot.
- These graphs more useful for small or moderate size data.

Descriptive Analysis

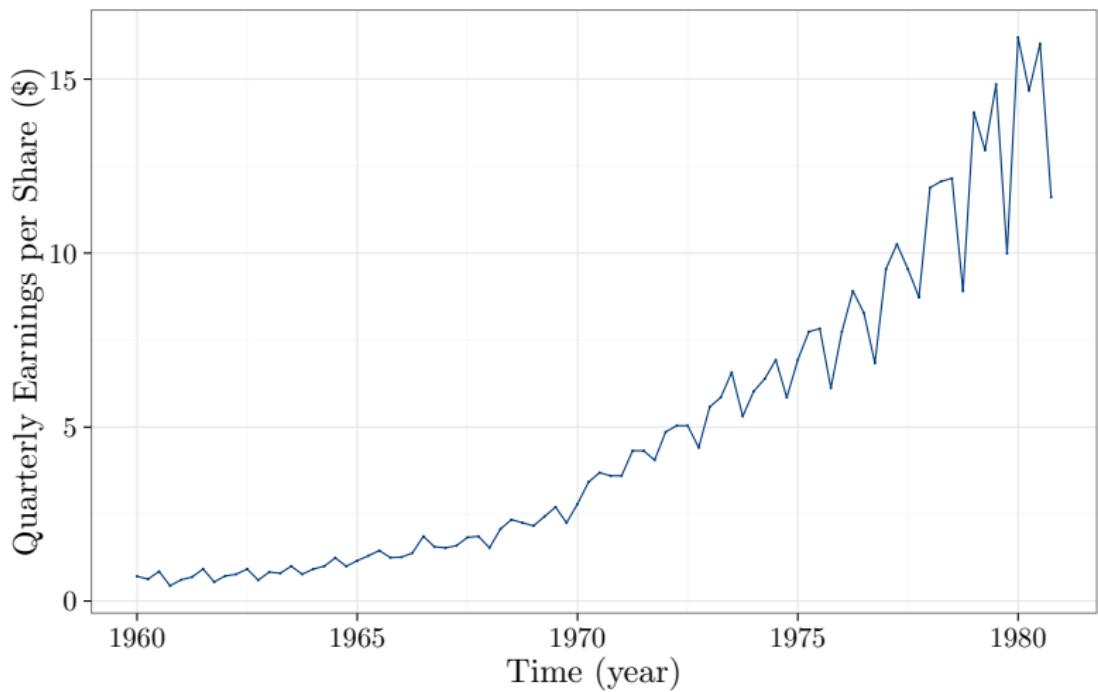
Question:

What do we want to check for in a time series data/graph?

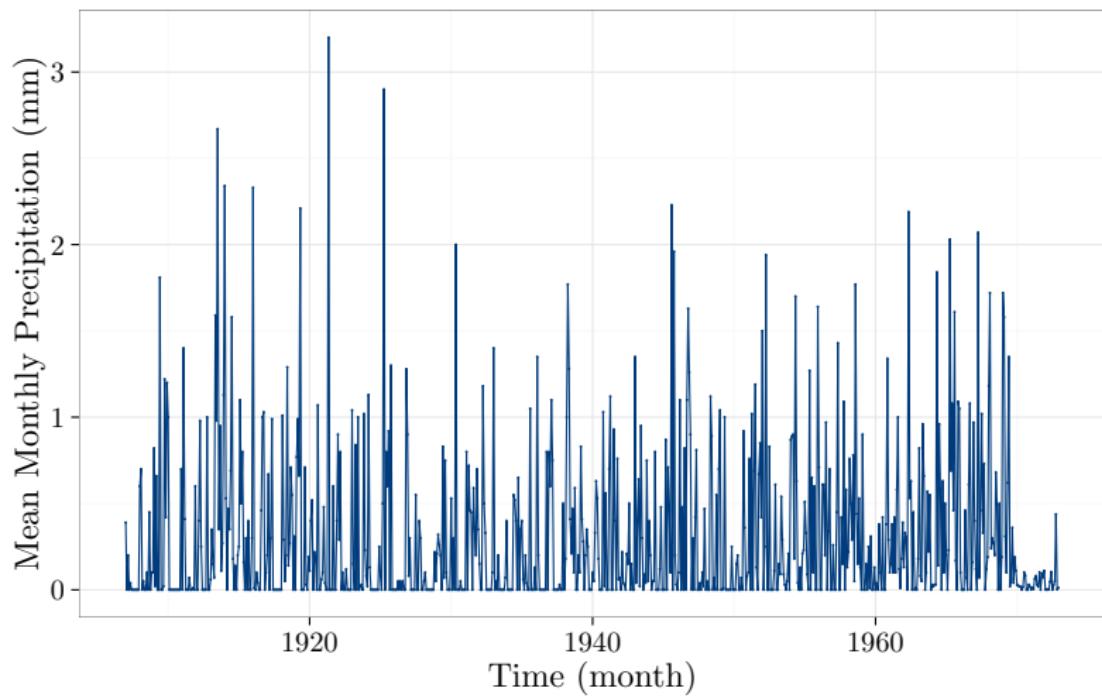
A possible answer:

- Trends:
 - Seasonal (e.g. business cycles)
 - Non-seasonal (e.g. impact of economic indicators on stock returns)
 - “Local” (e.g. vibrations observed before, during and after an earthquake)
- Changes in the **statistical properties**:
 - Mean (e.g. economic crisis)
 - Variance (e.g. earnings)
 - States (e.g bear/bull in finance)
- Model deviations (e.g. outliers)

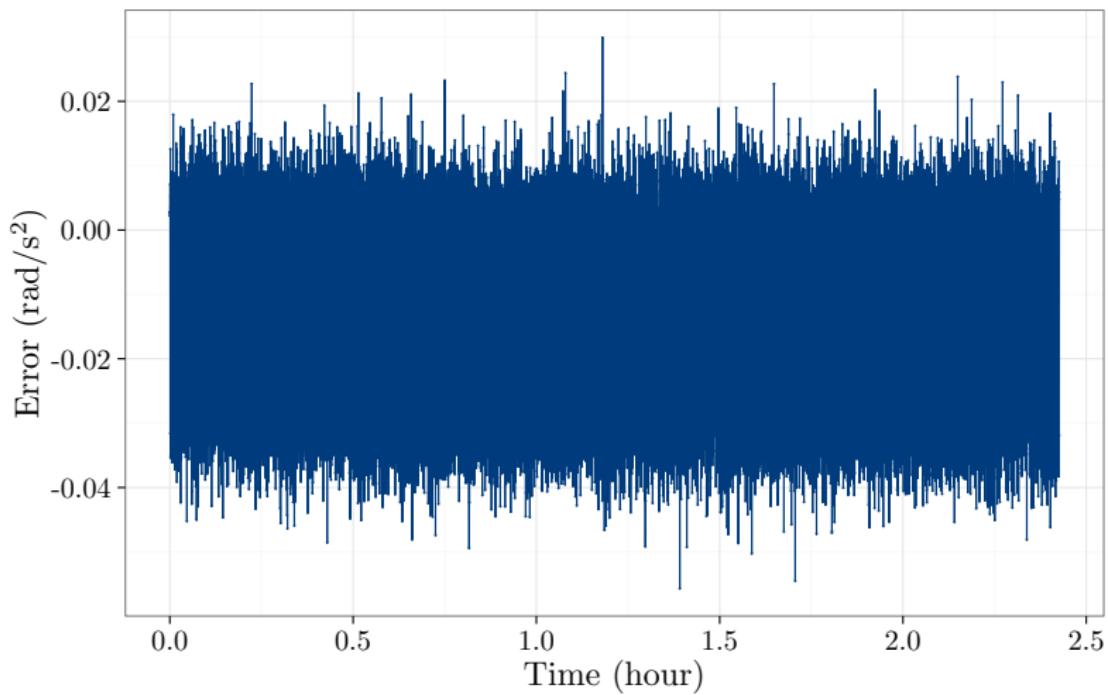
Example: Johnson and Johnson Quarterly Earnings



Example: Monthly Precipitation Data



Example: Inertial Sensor Data



Stochastic Processes Considered in this course

Definition 2.3 (Gaussian White Noise).

Gaussian White Noise (WN) with parameter $\sigma^2 \in \mathbb{R}^+$. This process is defined as

$$X_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

where “iid” stands for “independent and identically distributed”.

Definition 2.4 (Quantization Noise).

Quantization Noise (QN) with parameter $Q^2 \in \mathbb{R}^+$. This process has a PSD of the form

$$S_X(f) = 4Q^2 \sin^2\left(\frac{\pi f}{\Delta t}\right) \Delta t, \quad f < \frac{\Delta t}{2}.$$

Definition 2.5 (Drift).

Drift (DR) with parameter $\omega \in \Omega$ where Ω is either \mathbb{R}^+ or \mathbb{R}^- . This process is defined as $X_t = \omega t$.

Stochastic Processes Considered in this course

Definition 2.6 (Random walk).

Random walk (RW) with parameter $\gamma^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = X_{t-1} + \epsilon_t \text{ where } \epsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, \gamma^2) \text{ and } X_0 = 0.$$

Definition 2.7 (Auto-Regressive).

Auto-Regressive Process of Order 1 (AR1) process with parameter $\phi \in (-1, +1)$ and $v^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \stackrel{iid}{\sim} \mathcal{N}(0, v^2).$$

Stochastic Processes Considered in this course

Definition 2.8 (Gauss Markov).

Gauss Markov Process of Order 1 (GM) process with parameter $\beta \in \mathbb{R}$ and $\sigma_G^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = \exp(-\beta \Delta t) X_{t-1} + Z_t, \quad Z_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_G^2(1 - \exp(-2\beta \Delta t)))$$

where Δt denotes the time between X_t and X_{t-1} .

Remark 2 (GM and AR1).

A GM process is a one-to-one reparametrization of an AR1 process. In this course, we shall only discuss AR1 processes but all results remain valid for GM processes.

Stochastic Processes Considered in this course

Definition 2.9 (Composite stochastic processes).

A composite stochastic process is a sum of latent processes. In this course, we will always assume that these latent processes are independent.

Example:

The composite stochastic process: “2*AR1 + WN” is given:

$$Y_t = \phi_1 Y_{t-1} + Z_t, \quad Z_t \stackrel{iid}{\sim} \mathcal{N}(0, v_1^2)$$

$$W_t = \phi_2 W_{t-1} + U_t, \quad U_t \stackrel{iid}{\sim} \mathcal{N}(0, v_2^2)$$

$$Q_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$X_t = Y_t + W_t + Q_t,$$

where **only** (X_t) is observed.

Main purpose of TS analysis: Forecasting

- **Forecasting** is one of the main purpose of time series analysis. The question can be described as: if $(X_t)_{t=1,\dots,T}$ is an identically distributed sequence but is *not independent*, what is the “best” predictor for X_{T+h} for $h > 0$ (i.e. an estimator of $\mathbb{E}[X_{T+h}|X_T, \dots]$)?
- *A simple answer is that it depends on the “dependence” between X_1, \dots, X_T !*
- How could we measure this dependence?
- A first step is to extend the notation of covariance and correlation to time dependent sequences. We will refer to these notions as **autocovariance** and **autocorrelation**.
- The notion of autocovariance is an important one in time series analysis as it is closely linked to the concept of stationarity. Informally speaking, the latter creates a framework in which averages are “meaningful” (we will come back to this).

Review of Independence and Dependence

Definition 2.10 (Independence of two events).

Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition 2.11 (Independence of two random variables).

Two random variables X and Y with Cumulative Distribution Functions (CDF) $F_X(x)$ and $F_Y(y)$ (respectively) are **independent** if and only if their joint CDF $F_{X,Y}(x,y)$ is such that $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

Definition 2.12 (iid sequence).

The sequence X_1, X_2, \dots, X_T is said to be **iid** if and only if

$$\mathbb{P}(X_i < x) = \mathbb{P}(X_j < x) \quad \forall x \in \mathbb{R}, \forall i, j \in \{1, \dots, T\}, \text{ and}$$

$$\mathbb{P}(X_1 < x_1, X_2 < x_2, \dots, X_T < x_T) = \mathbb{P}(X_1 < x_1) \dots \mathbb{P}(X_T < x_T),$$

for any $T \geq 2$ and $x_1, \dots, x_T \in \mathbb{R}$.

Measuring (linear) dependence

Dependence between T RV is difficult to measure at one shot! So we consider just two random variables, X_t and X_{t+h} . Then, one common (linear) measure of dependence is the covariance between X_t and X_{t+h} , which is defined below.

Definition 2.13 (AutoCovariance).

The covariance between X_t and X_{t+h} , defined as the *AutoCovariance* or simply ACV, is denoted using the function $\gamma_X(t, t + h)$, i.e.

$$\gamma_X(t, t + h) \equiv \text{Cov}(X_t, X_{t+h}) = \mathbb{E}(X_t X_{t+h}) - \mathbb{E}(X_t)\mathbb{E}(X_{t+h}),$$

where

$$\mathbb{E}(X_t) = \int_{-\infty}^{\infty} x f(x) dx$$

$$\mathbb{E}(X_t, X_{t+h}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2) dx_1 dx_2,$$

where $f(x_1, x_2)$ denotes the joint density of X_t and X_{t+h} .

Measuring (linear) dependence

Remark 3 (Scale dependence).

Just as any covariance, the $\gamma_X(t, t + h)$ is “scale dependent” and therefore $\gamma_X(t, t + h) \in \mathbb{R}$.

- If $|\gamma_X(t, t + h)|$ is “close” to 0, then they are “less dependent”.
- If $|\gamma_X(t, t + h)|$ is “far” from 0, X_t and X_{t+h} are “more dependent”.

Remark 4 (ACV and independence).

$\gamma_X(t, t + h) = 0$ does not imply X_t and X_{t+h} are independent. However, if X_t and X_{t+h} are joint normally distributed then $\gamma_X(t, t + h) = 0$ implies that X_t and X_{t+h} are independent.

Measuring (linear) dependence

A measure of dependence related to the ACV is the autocorrelation. This is arguably the most commonly used metric in time series analysis.

Definition 2.14 (Autocorrelation).

The correlation between X_t and X_{t+h} is defined as the *autocorrelation* or simply ACF and is denoted using the function $\rho_X(t, t + h)$, i.e.

$$\rho_X(t, t + h) = \text{corr}(X_t, X_{t+h}) = \frac{\text{cov}(X_t, X_{t+h})}{\sqrt{\text{var}(X_t)} \sqrt{\text{var}(X_{t+h})}}$$

Remark 5 (Scale invariance).

Just as any correlation, $\rho_X(t, t + h)$ is scale free. Moreover, if $\rho_X(t, t + h)$ is “close” to ± 1 then this implies that there is “strong” (linear) dependence between X_t and X_{t+h} .

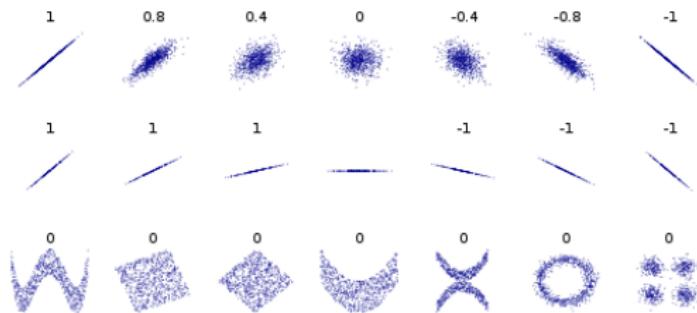
Measuring (linear) dependence

Remark 6 (Notation).

The notation $\gamma_X(t, t + h)$ and $\rho_X(t, t + h)$ is often simplified to $\gamma(t, t + h)$ and $\rho(t, t + h)$ when not ambiguous (i.e. only one time series is considered).

Remark 7 (Linear dependence and real dependence).

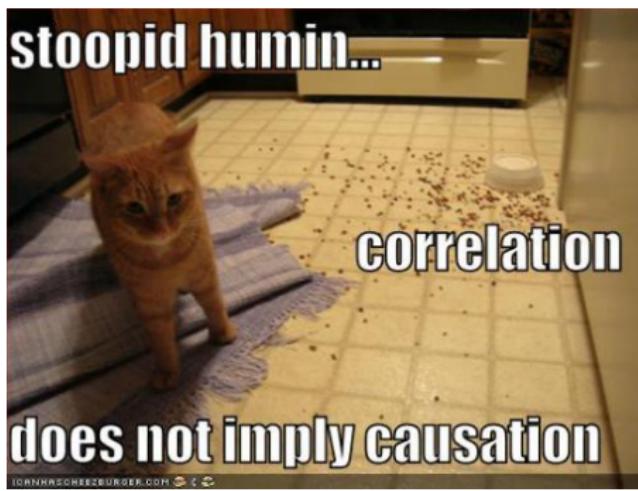
Covariance and correlation measure linear dependence. They are less helpful to measure monotonic dependence and they are much less helpful to measure nonlinear dependence. Nonlinear measures of dependence exist but we will not discuss this subject in this class. Here is an example:



Correlation == Causation?

Remark 8 (Causation).

Correlation *does NOT* imply causation. For example, if $\rho(t, t + h) \neq 0$ it does not imply that $X_t \rightarrow X_{t+h}$ is causal. Actually, real causation doesn't exist in Statistics but there exist approximated metric to measure this concept such as Granger causality (see Granger 1969). This idea is clearly illustrated in the image below.



Estimation of in the context of time series

Motivation:

Consider the simple (but strange!) model:

$$X_t \sim \mathcal{N}(0, Y_t^2) \text{ where } Y_t \text{ is unobserved and such that } Y_t \stackrel{iid}{\sim} \mathcal{N}(0, 1).$$

In this case, it is clear that the estimation of $\text{var}(X_t)$ is difficult (in fact X_t^2 is your best guess!) since only X_t is useful to estimate $\text{var}(X_t)$. This process is an example of a **non-stationary process** (we will see why in the next slides). On the other hand, if we consider the **stationarity process** such as:

$$X_t = \theta W_{t-1} + W_t \text{ where } W_t \stackrel{iid}{\sim} \mathcal{N}(0, 1).$$

Then, one can guess that a natural estimator of $\text{var}(X_t)$ is simply $\hat{\sigma}^2 = \frac{1}{T} \sum_{i=1}^T X_i^2$, because our hope is that averages are “**meaningful**” for such processes. In the next slides, we will formalize this idea through the concepts of stationarity.

Strong and Weak Stationarity

There exist two forms of stationarity, which are defined below:

Definition 2.15 (Strong Stationarity).

The joint probability distribution of $(X_t)_{t \in \mathbb{N}}$ is invariant under a shift in time, i.e.

$$\mathbb{P}(X_t \leq x_1, \dots, X_{t+k} \leq x_k) = \mathbb{P}(X_{t+h} \leq x_1, \dots, X_{t+h+k} \leq x_k)$$

for any time shift h and any x_1, x_2, \dots, x_k belong to the domain of X_t, \dots, X_{t+k} and $X_{t+h}, \dots, X_{t+h+k}$.

Definition 2.16 (Weak Stationarity).

The mean and autocovariance of the stochastic process are finite and invariant under a shift in time, i.e.

$$\mathbb{E}[X_t] = \mu < \infty,$$

$$\mathbb{E}[X_t^2] = \mu_2 < \infty,$$

$$\text{cov}(X_t, X_{t+h}) = \text{cov}(X_{t+k}, X_{t+h+k}) = \gamma(h).$$

Strong and Weak Stationarity

Why does stationarity matter?

The stationarity of X_t is important because it provides a framework in which averaging makes sense. Unless properties like mean and covariance are either fixed or “evolve” in a known manner, the concept of averaging is essentially meaningless.

Remark 9 (Implication on the ACV and ACF).

If a process is weakly stationary or strongly stationary and $\text{cov}(X_t, X_{t+h})$ exist for all $h \in \mathbb{Z}$. Then we have the ACV and ACF only depends on the lag between observations, i.e.

$$\begin{aligned}\gamma(t, t + h) &= \text{cov}(X_t, X_{t+h}) = \text{cov}(X_{t+k}, X_{t+h+k}) = \gamma(t + k, t + h + k) = \gamma(h), \\ \rho(t, t + h) &= \text{corr}(X_t, X_{t+h}) = \text{corr}(X_{t+k}, X_{t+h+k}) = \rho(t + k, t + h + k) = \rho(h).\end{aligned}$$

Strong and Weak Stationarity

Remark 10 (Properties of the ACV and ACF).

Remark 9 implies that the ACV and ACF have the following properties:

- $\gamma(0) = \text{var}[X_t] \geq 0$ and $\rho(0) = 1$.
- $\gamma(h) = \gamma(-h)$ and $\rho(h) = \rho(-h)$ (therefore they are both even functions).
- $|\gamma(h)| \leq \gamma(0)$ and $|\rho(h)| \leq 1$ for all $h \in \mathbb{Z}$.

The first two properties are direct for the properties of the covariance and correlation (i.e. $\text{cov}(X, X) = \text{var}(X)$ and $\text{cov}(X, Y) = \text{cov}(Y, X)$). However, the third property is less obvious and is a consequence of the Cauchy-Schwarz inequality, i.e.

$$\mathbb{E}^2[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \quad (1)$$

Using (1), we have

$$\begin{aligned}\gamma(h)^2 &= (\text{cov}(X_t, X_{t+h}))^2 = (\mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)])^2 \\ &\leq \mathbb{E}[(X_t - \mu)^2]\mathbb{E}[(X_{t+h} - \mu)^2] = \gamma(0)^2,\end{aligned}$$

which verifies the last properties.

Remarks: Strong and Weak Stationarity

Remark 11.

Neither type of stationarity implies the other one. This is being illustrated in the two examples presented in Appendix A [► Go to Appendix A](#). Note however that if X_t is Normal (Gaussian) with $\sigma^2 = \text{var}(X_t) < \infty$, then weak stationarity implies strong stationarity.

Remark 12.

From the definition of (weak) stationarity, it is easy to see that a WN or a QN processes is stationary while a RW process is not. The stationarity of an AR1 (or GM) is less obvious. In fact, an AR1 is stationary if $|\phi| < 1$. The derivation of this property is given in Appendix B [► Go to Appendix B](#).

Linear Processes

Definition 2.17 (Linear Processes).

A stochastic process (X_t) is said to be a linear process if it can be expressed as a linear combination of an iid sequence (which here is Gaussian for convenience), i.e.:

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

where $W_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Remark 13 (Properties of linear processes).

All linear processes are stationary and such that:

$$\mathbb{E}[X_t] = \mu,$$

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{h+j}.$$

Linear Processes

Remark 14 (Convergence of linear processes).

The latter condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ is required to ensure that the series has a limit and is related to the absolutely summable covariance structure (defined below).

Definition 2.18 (Absolutely summable covariance structure).

A process (X_t) is said to have an absolutely summable covariance structure if

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty.$$

Remark 15 (All linear process have an absolutely summable covariance structure).

Interestingly, the condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ is actually stronger than $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$. Indeed, we have that

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| \leq 2\gamma(0) \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right)^2 < \infty.$$

A Fundamental Representation

A Fundamental Representation

Autocovariances and autocorrelations also turn out to be very useful tools as they are one of the *fundamental representations of time series*.

If we consider a zero mean normally distributed process, it is clear that its joint distribution is fully characterized by the autocovariances $\mathbb{E}[X_t X_{t+h}]$ since the joint probability density only depends on these covariances.

Once we know the autocovariances we know everything there is to know about the process and therefore: **if two processes have the same autocovariance function, then they are the same process.**

Another Fundamental Representation

Fundamental Representation: the Power Spectral Density

For the same processes considered in the previous slide, another fundamental representation of a time series is given by the Power Spectral Density (PSD) which can be defined as

$$S_X(f) = \int_{-\infty}^{\infty} \gamma_X(h) e^{-ifh} dh,$$

where f is a frequency. Hence, the PSD is a Fourier transform of the autocovariance function which describes the variance of a time series over frequencies (with respect to lags h).

Given that the definition of the PSD, as for the autocovariance function, once we know the PSD we know everything there is to know about the process and therefore: **if two processes have the same PSD, then they are the same process.**

Estimation Problems with Dependent Data

Estimation in the context of time series is not as straightforward as in the iid case. In order to “warm-up”, let us start with the easiest case: the empirical mean.

Let (X_t) be a stationary time series, therefore we have that $\mu_t = \mathbb{E}[X_t] = \mu$ and the value of μ can be estimated by the sample mean, i.e.

$$\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t.$$

Using the properties of stationary process we have that:

$$\text{var}(\bar{X}) = \frac{1}{T^2} \text{cov}\left(\sum_{t=1}^T X_t, \sum_{s=1}^T X_s\right) = \frac{\gamma(0)}{T} \sum_{h=-T}^T \left(1 - \frac{|h|}{T}\right) \rho(h). \quad (2)$$

The derivation of (2) is instructive and is given in Appendix C [▶ Go to Appendix C](#). Moreover, some simulation-based and analytical methods to estimate $\text{var}(\bar{X})$ are discussed in Appendix D [▶ Go to Appendix D](#).

Estimation of $\gamma(h)$ and $\rho(h)$

We define here the “classical” estimator of $\gamma(h)$ and $\rho(h)$ as the sample autocovariance and autocorrelation functions. In the following section, we shall study the properties of these estimators:

Definition 2.19 (Sample autocovariance function).

The sample autocovariance function is defined as

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (X_t - \bar{X})(X_{t+h} - \bar{X})$$

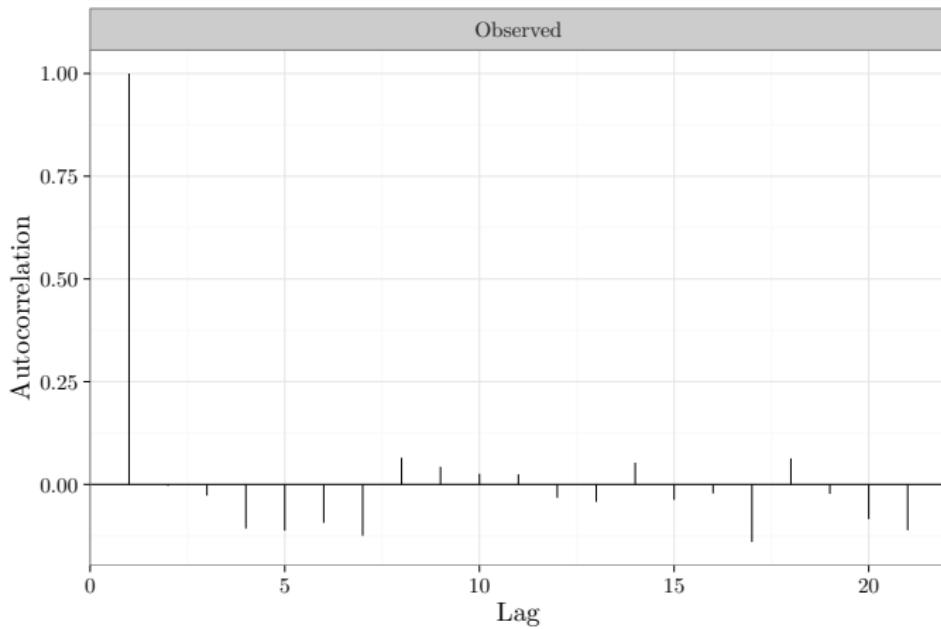
with $\hat{\gamma}(h) = \hat{\gamma}(-h)$ for $h = 0, 1, \dots, k$, where k is a fixed integer.

Definition 2.20 (Sample autocorrelation function).

The sample autocorrelation function is defined as $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$, with $\hat{\rho}(h) = \hat{\rho}(-h)$ for $h = 0, 1, \dots, k$, where k is a fixed integer.

An Example: a White Noise Process

Consider the estimated ACF of a simulated gaussian white noise process (i.e. $W_t \sim \mathcal{N}(0, 1)$) of length $T = 100$.



Estimation of $\gamma(h)$ and $\rho(h)$

Remark:

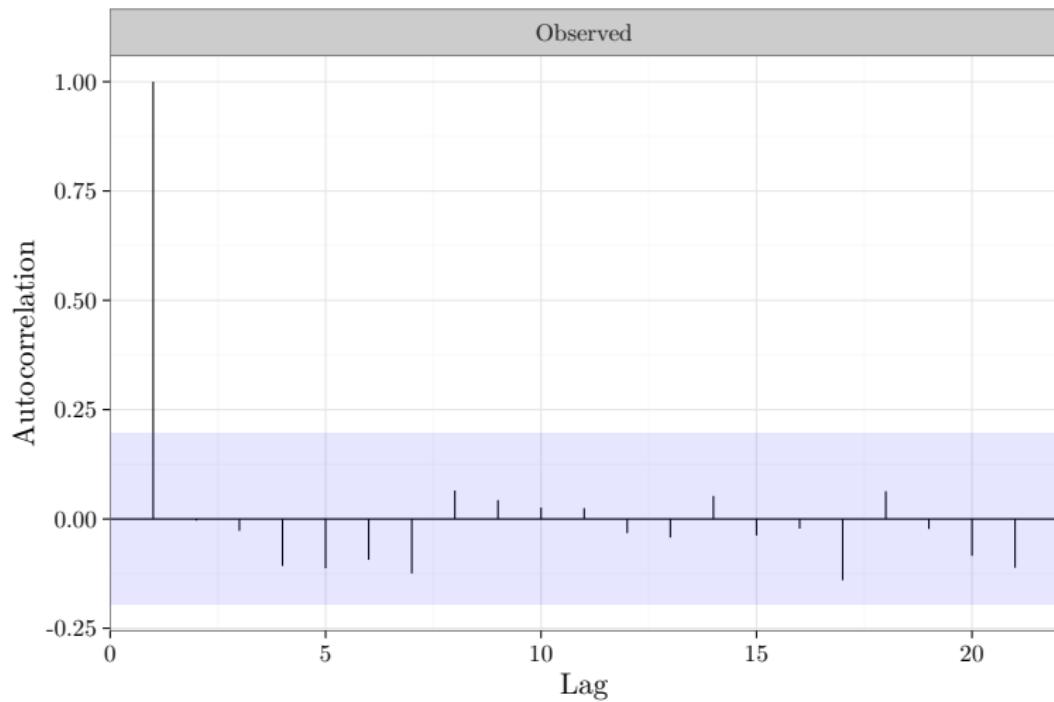
If (X_t) is a white noise then $\hat{\rho}(h)$ should be equal to 0 if $h \neq 0$. In practice, this is of course not the case due to the estimation error of $\hat{\rho}(h)$. The next result gives us a way to assess whether the data comes from a completely random series or whether correlations are statistically significant at some lags.

Property:

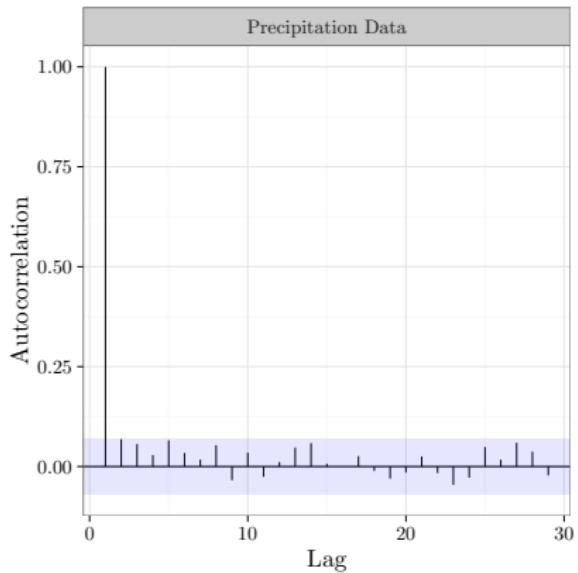
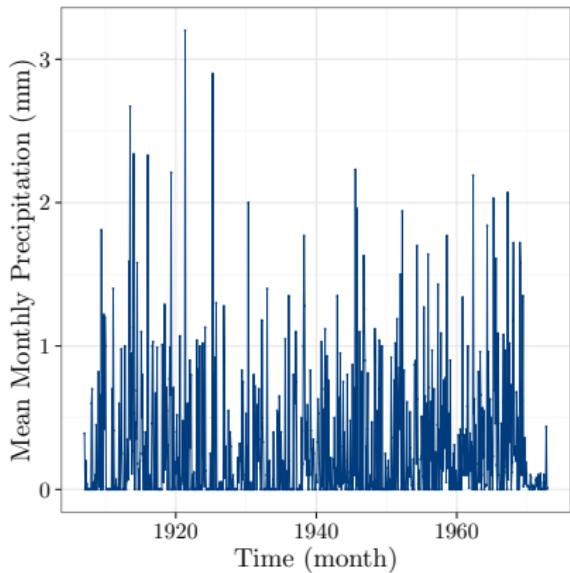
Under some technical conditions, if (X_t) is white noise and $h = 1, \dots, H$ where H is fixed but arbitrary we have that

$$\sqrt{T}(\hat{\rho}(h) - \rho(h)) \xrightarrow{D} \mathcal{N}(0, 1).$$

An Example: a White Noise Process - cont.



Example: ACF of Precipitation Data



Remark:

The “ACF” plot suggests an absence of linear dependence in this dataset. We will revisit this example latter.

"Whiteness" Testing

A first step in time series analysis is often to test if the time series at hand is an uncorrelated sequence. For this purpose, it seems natural to construct a test statistics based on

$$Q_{BP} = \sum_{k=1}^h \left(\sqrt{T} \hat{\rho}(k) \right)^2,$$

since such statistic should be "large" if the process is autocorrelated and asymptotically χ^2 distributed if the process is a white noise (remember that $\sqrt{T} \hat{\rho}(k) \sim \mathcal{N}(0, 1)$ asymptotically). In fact, this is known as the **Box-Pierce test**. Other tests are generally preferred as they provide better finite sample results.

"Whiteness" Testing

Ljung-Box Test

One of the most commonly used "whiteness" test is the Ljung-Box test. This approach aims to test the "overall" randomness based on a number of lags rather than testing the randomness at each distinct lag. It is therefore a *portmanteau test*. The Ljung-Box test is defined as:

$$\begin{aligned} H_0 &: \rho(1) = \dots = \rho(h) \\ H_1 &: H_0 \text{ is false} \end{aligned}$$

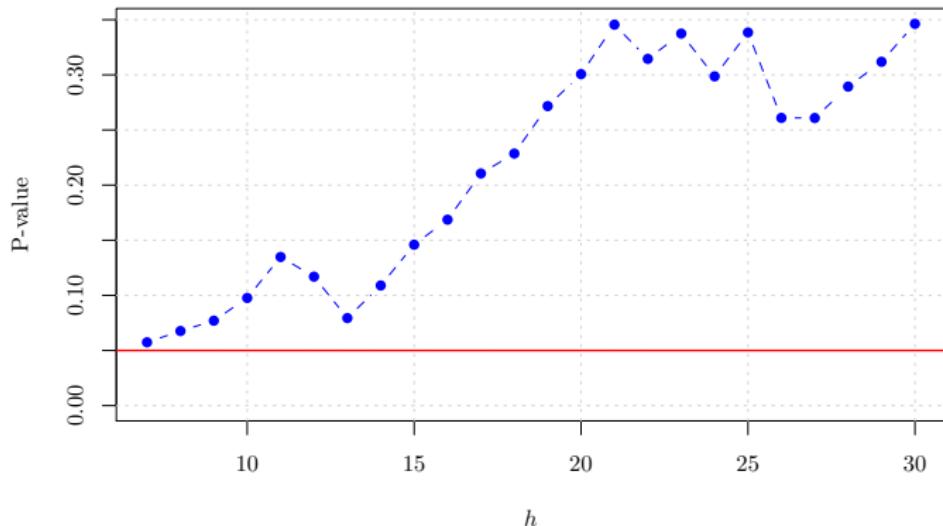
and is based on the test statistic

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}(k)^2}{n-k}.$$

It is easy to see that this statistic is equivalent to Q_{BP} and is therefore asymptotically χ^2 distributed with h degrees of freedom under H_0 .

Example: ACF of Precipitation Data - cont.

The choice of h is of course important and somewhat arbitrary. Instead of picking one specific value a graph with different values of h is often preferred. For example for the precipitation data we obtain the figure below. Based on this result we cannot reject H_0 .



Robust Statistics - A Motivating Example

As we have already seen, outliers are commonly observed in real time series data. It is then natural to wonder what is the impact of these “extreme” observations on the classical estimators such as the one we used to estimate the autocorrelation function.

An example:

Consider the following two processes:

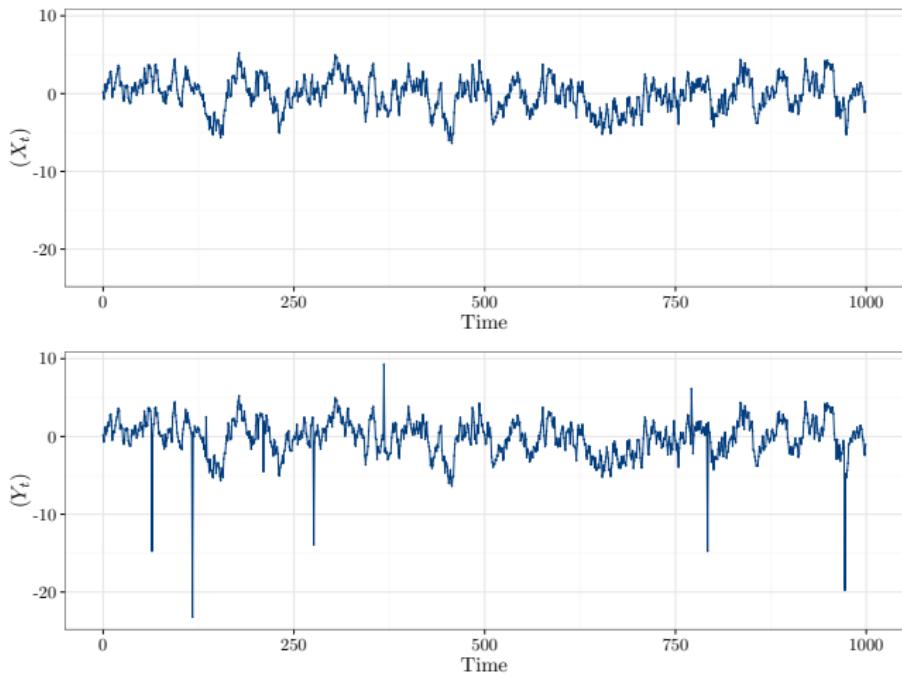
$$X_t = \phi X_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, \sigma_W^2)$$

$$Y_t = \begin{cases} X_t & \text{with probability } 1 - \epsilon \\ U_t & \text{with probability } \epsilon \end{cases}, \quad U_t \sim \mathcal{N}(0, \sigma_u^2)$$

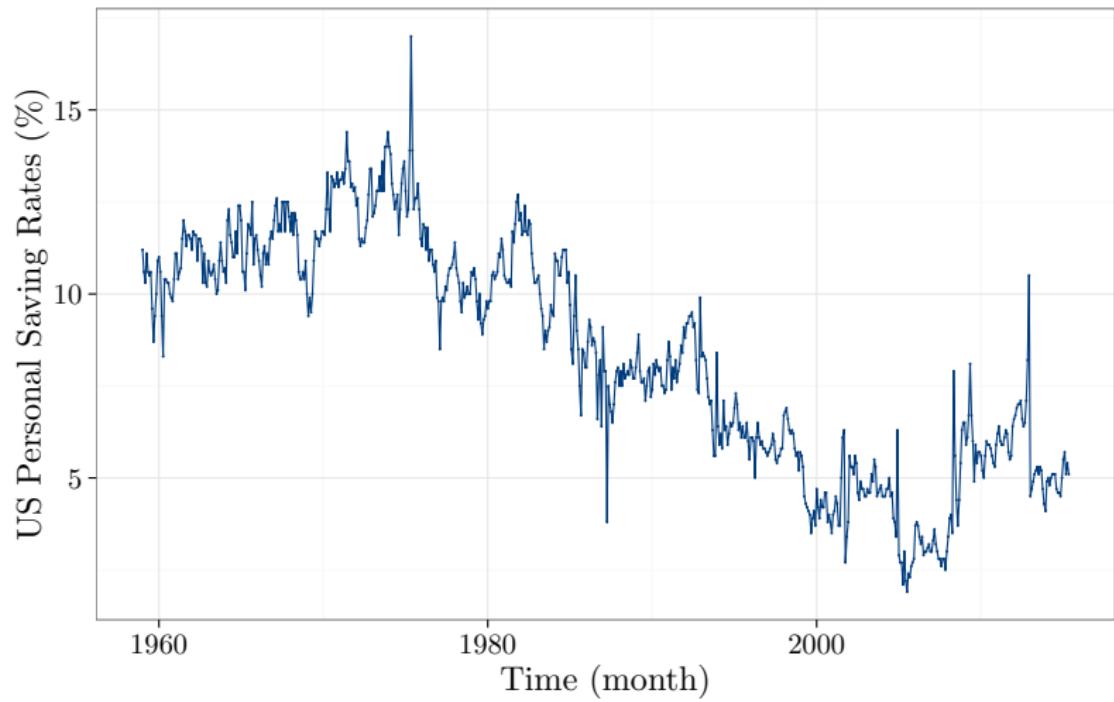
when ϵ is “small” and $\sigma_u^2 \gg \sigma_w^2$, the process (Y_t) can be interpreted as a “contaminated” version of (X_t) .

Robust Statistics - A Motivating Example

Let $\phi = 0.9$, $\sigma_u^2 = 1$, $\epsilon = 0.01$, σ_u^2 and $T = 10^3$ One realization of (X_t) and (Y_t) :

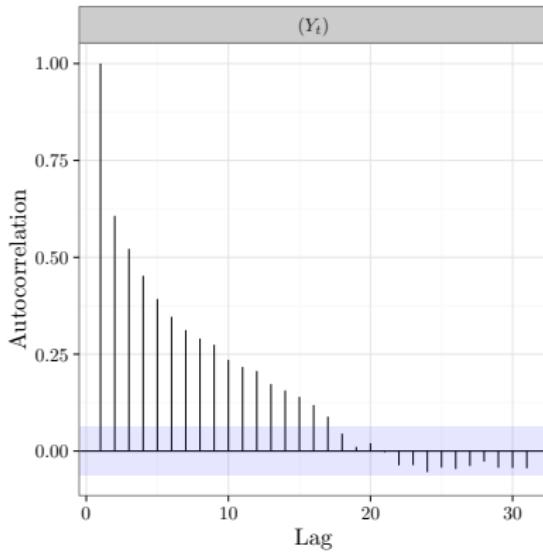
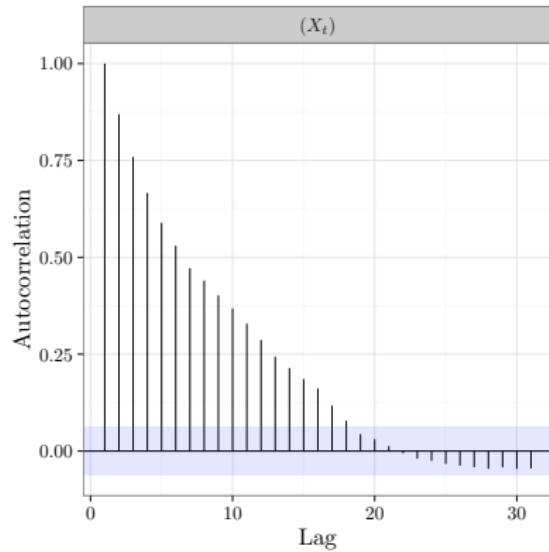


Remember: Personal US Saving Rates



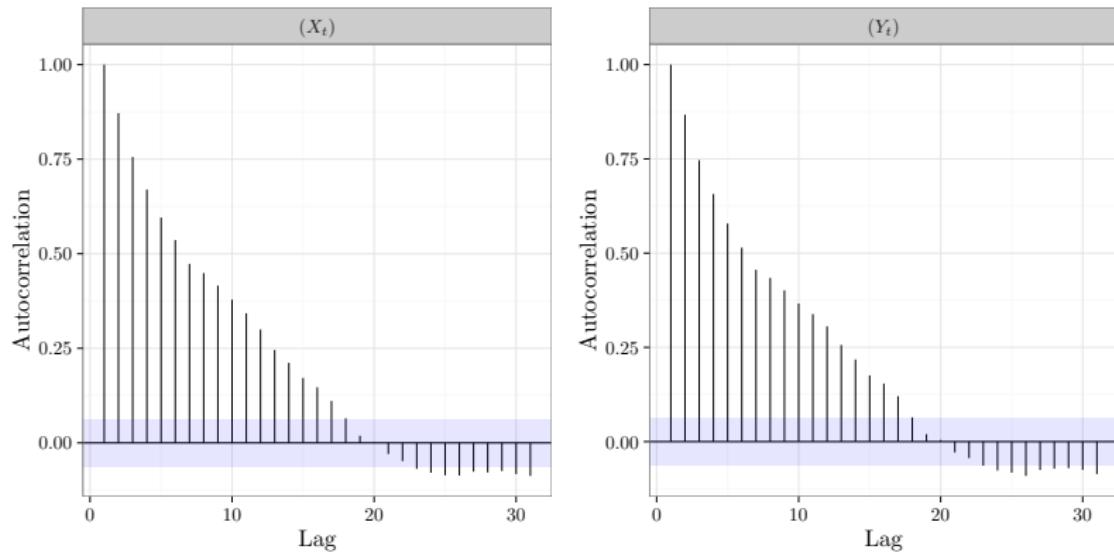
Robust Statistics - A Motivating Example

Consider the ACF of both processes. The extreme observations seem to “shrink” the estimated autocorrelations.



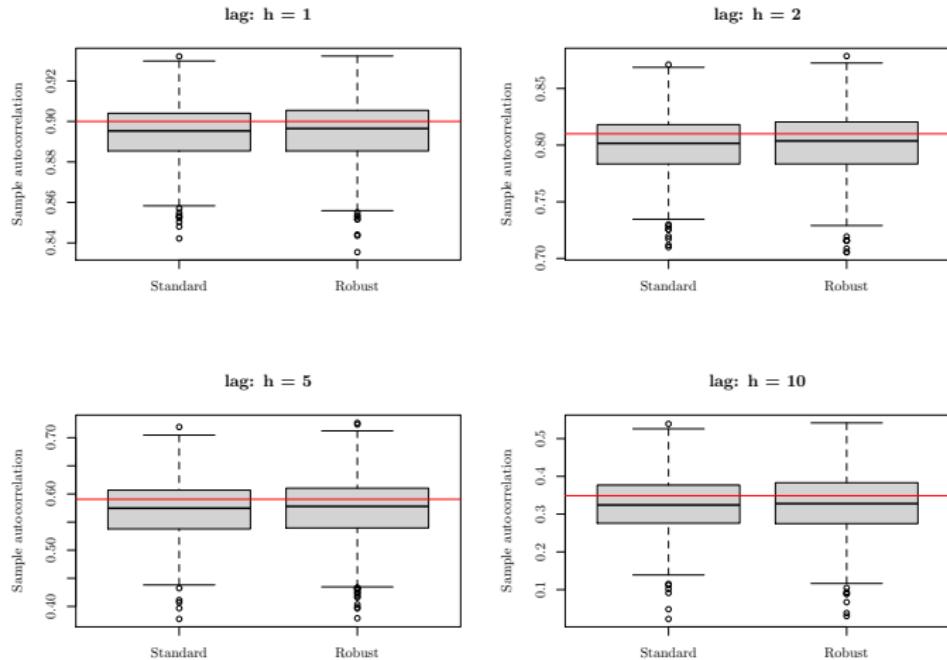
A robust estimator of the autocorrelation

Several robust estimators of the autocorrelation have been introduced in the literature (see e.g. Ma and Genton, 2000). Such estimators are less efficient but tends to be less affected by extreme observations. Let us compare the robust ACFs of the two processes (using the R function `robacf`).



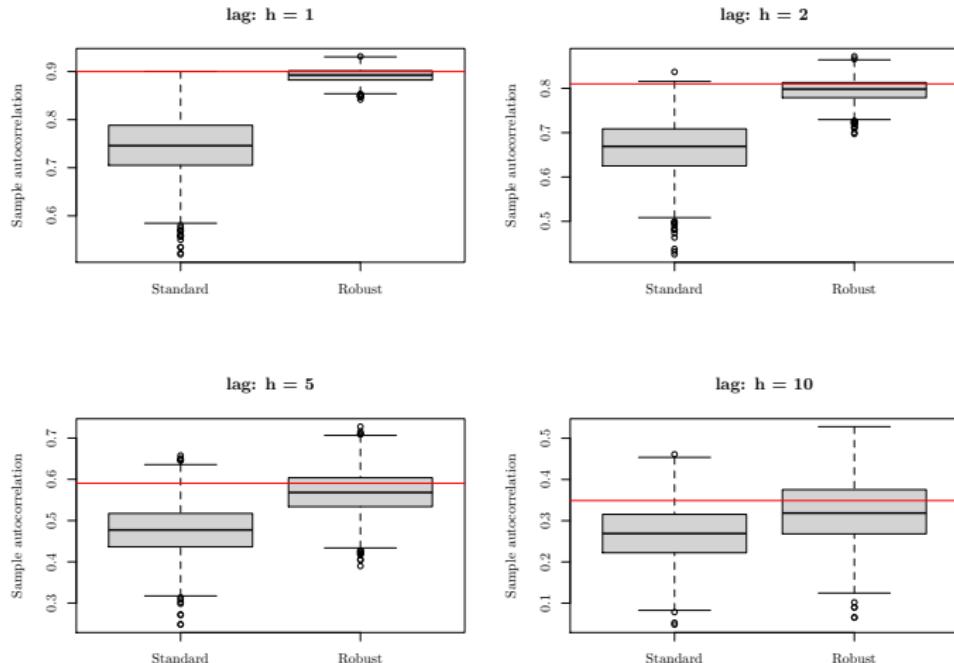
A small simulation study

Setting: $\phi = 0.9$, $\sigma_u^2 = 1$, $\epsilon = 0$, σ_u^2 , $T = 10^3$ and $B = 10^3$.



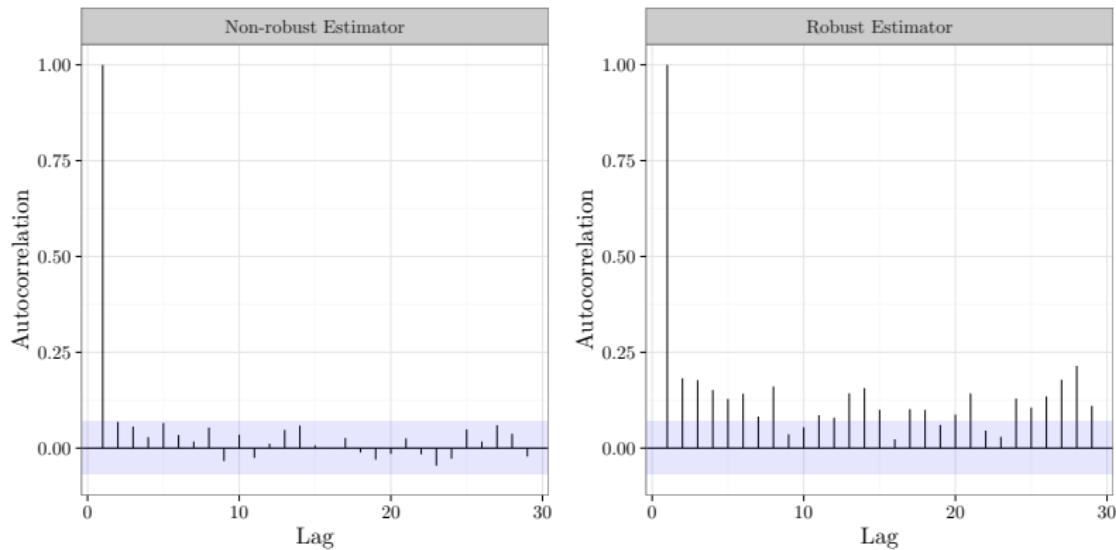
A small simulation study - cont.

Setting: $\phi = 0.9$, $\sigma_u^2 = 1$, $\epsilon = 0.01$, σ_u^2 , $T = 10^3$ and $B = 10^3$.



Example: ACF of Precipitation Data - cont.

The classical and robust ACFs seem significantly different. The robust estimator tends to indicate that the process is not uncorrelated.



ARMA Models - Modeling Paradigm

Modeling objective

A common measure used to assess many statistical models is their ability to reduce the input data to random noise. For example, we often say that a regression model “fits well” if its residuals ideally resemble iid random noise. We often settle for uncorrelated processes with data.

Prediction rationale:

If a model reduces the data to iid noise, then the model captures all of the relevant structure, at least in the sense that we obtain the decomposition

$$X_t = \mathbb{E}[X_t | X_{t-1}, \dots, X_0] + w_t = \hat{X}_t + W_t.$$

Motivation for ARMA models

- If $(X_t) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then the autocovariance is a *fundamental representation* of the process (i.e. it contains all the information about the process).
- The class of ARMA models is able to approximate a wide range of autocovariance structures with parsimonious (parametric) models in the sense that $\hat{\Sigma} \approx \Sigma(\hat{\theta})$.
- ARMA models typically depends on only a few parameters and these models are reasonably easy to estimate.
- The initial goal of time series modeling with ARMA models amounts to finding a model which can reduce X_t to iid noise.

Partial Autocorrelation Function

- We already discussed how the autocorrelation can be used to quantify the linear dependence in a time series.
- A related quantity is the Partial Autocorrelation Function or PACF, which is particularly useful to identify the order of autoregressive models or AR models (which we will discuss before ARMA models).
- This quantity has very similar properties compared to the ACF, therefore, a discussion on how it can be estimated as well as its asymptotic is omitted.

Definition

The partial autocorrelation function, ϕ_{hh} , can be defined for Gaussian processes as the “partial” correlation between the random variables X_{t+h} and X_t conditioning upon the intervening variables,

$$\phi_{h,h} = \text{Corr}(X_{t+h}, X_t | X_{t+1}, \dots, X_{t+h-1})$$

Autoregressive Models

Definition

An *autoregressive model of order p*, abbreviated AR(p) is of the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$$

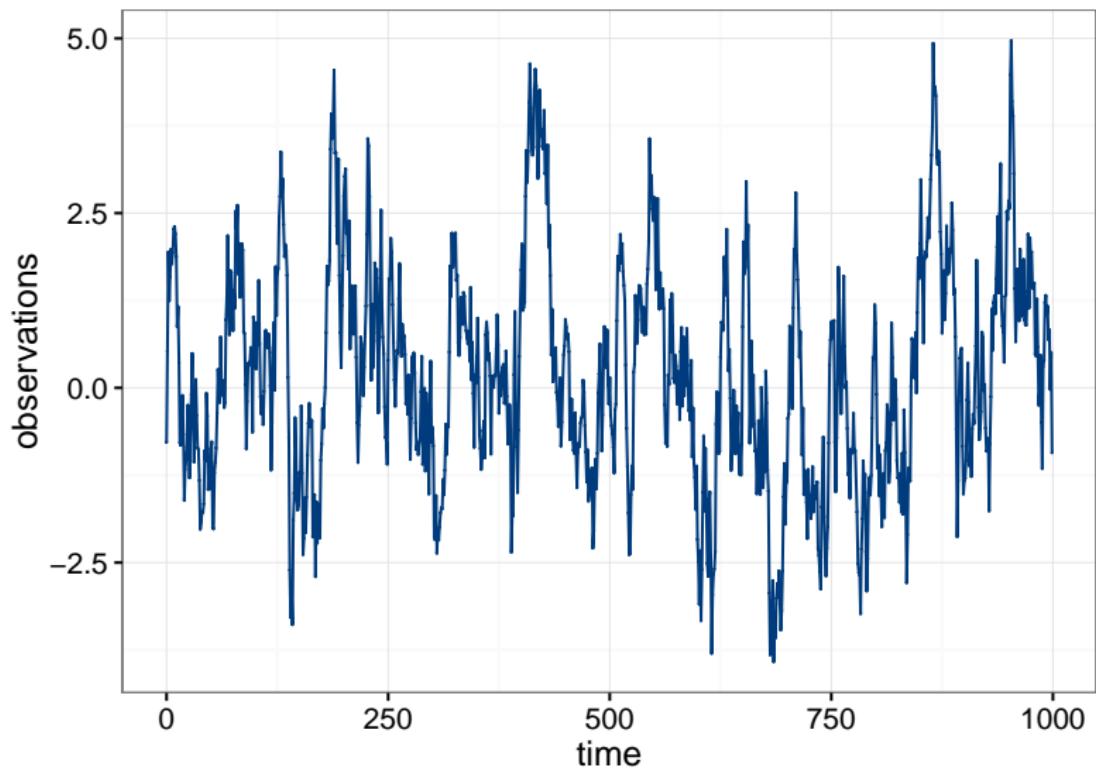
where X_t is stationary, ϕ_1, \dots, ϕ_p are constant ($\phi_p \neq 0$) and $W_t \sim \mathcal{N}(0, \sigma_w^2)$.

Remark:

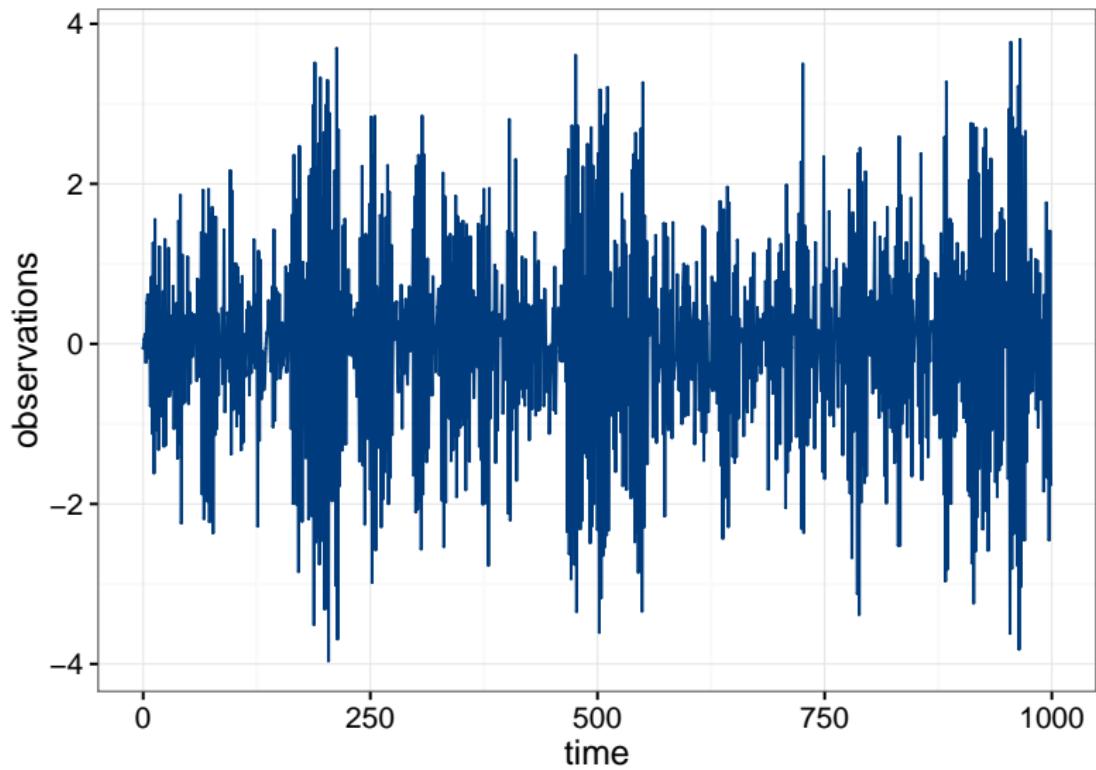
We assume in the above definition that the expected value of X_t is zero. If this is not the case, we can replace X_t by $X_t - \mu$ (where $\mu = \mathbb{E}[X_t]$) and obtain:

$$X_t = \alpha + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t.$$

Simulated Process: AR(1) process ($\phi = 0.9$)



Simulated Process: AR(1) process ($\phi = -0.9$)



Properties of an AR(1)

As we already discussed, the mean of an AR(1) is:

$$\mathbb{E}[X_t] = \sum_{j=0}^{\infty} \phi^j \mathbb{E}(W_{t-j}) = 0,$$

its autocovariance function is given by:

$$\gamma(h) = \text{Cov}(X_{t-h}, X_t) = \frac{\sigma_w^2 \phi^{|h|}}{1 - \phi^2},$$

and, therefore, its autocorrelation is:

$$\rho(h) = \phi^{|h|}.$$

Theoretical PACF of AR(1)

Theoretical PACF of an AR(1)

Using the definition of the PACF, we have

$$\phi_{1,1} = \text{corr}(X_t, X_{t+1}) = \frac{\text{cov}(X_t, X_{t+1})}{\text{var}(X_t)} = \frac{\phi\gamma(0)}{\gamma(0)} = \phi.$$

Next, we consider $\phi_{2,2}$, since

$$\begin{aligned}\text{cov}(X_t, X_{t+2}|X_{t+1}) &= \text{cov}(X_t, \phi X_{t+1} + W_{t+2}|X_{t+1}) \\ &= \text{cov}(X_t, \phi X_{t+1}|X_{t+1}) + \text{cov}(X_t, W_{t+2}|X_{t+1}) = 0\end{aligned}$$

we have that $\phi_{2,2} = 0$. In general, it is easy to verify that $\phi_{h,h} = 0$, $h > 1$.

Theoretical ACF and PACF of AR(p) models

- In general, the ACF of an AR(p) is a complicated function of the parameters but $\rho(h)$ tends to dampen, in a sinusoidal fashion, exponentially fast to zero as h increases.
- The PACF of an AR(p) is non-zero for the first p lags and zero for $h > p$.
- Example: consider the following models:

$$X_t = 0.95X_{t-1} + W_t$$

$$X_t = 1.5X_{t-1} - 0.75X_{t-2} + W_t$$

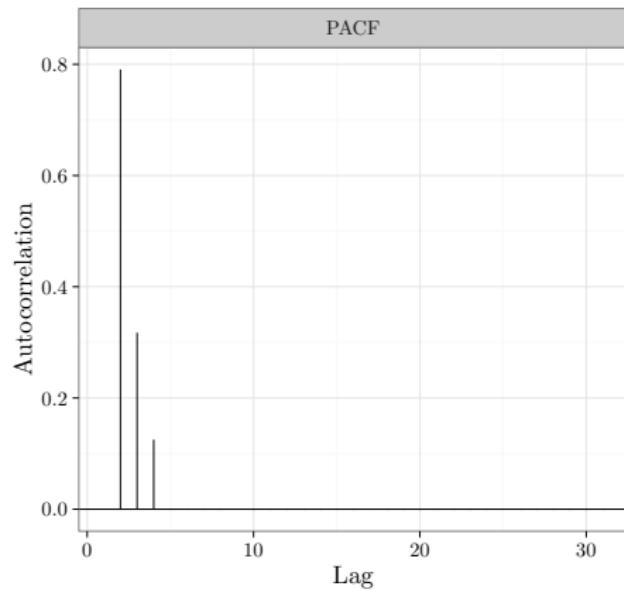
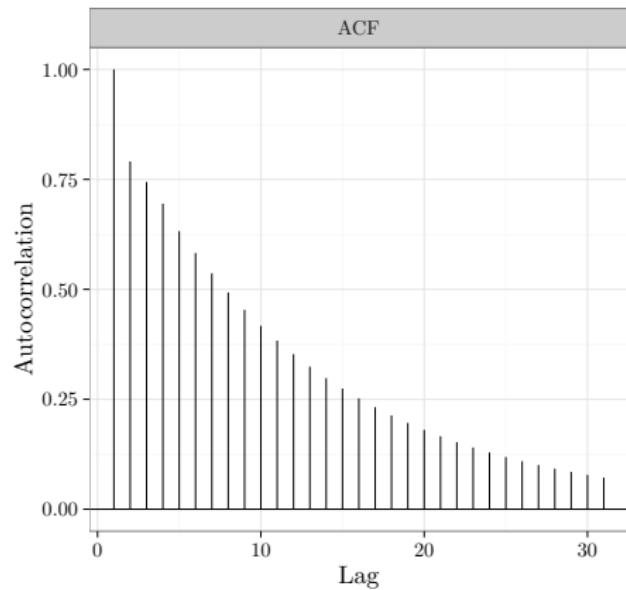
$$X_t = 0.5X_{t-1} + 0.25X_{t-2} + 0.125X_{t-3} + W_t$$

$$X_t = 0.5X_{t-1} + 0.75X_{t-2} - 0.8X_{t-3} + W_t$$

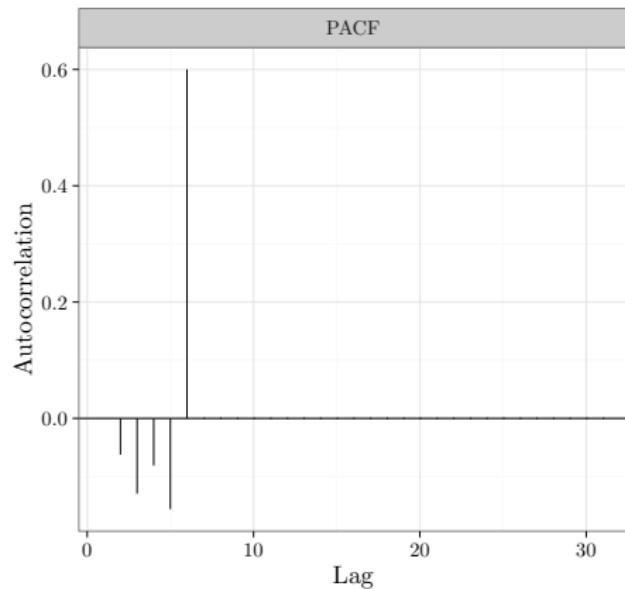
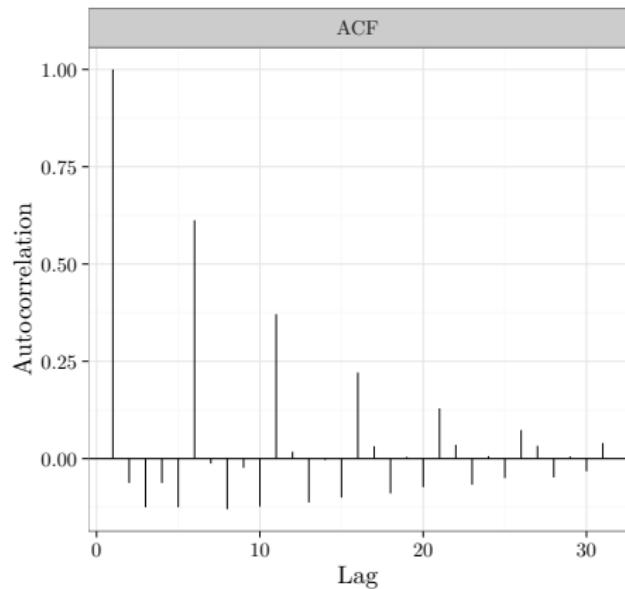
$$X_t = -0.1X_{t-2} - 0.1X_{t-4} + 0.6X_{t-5} + W_t$$

Can we recognize each model given their theoretical ACF/PACF? Let's try...

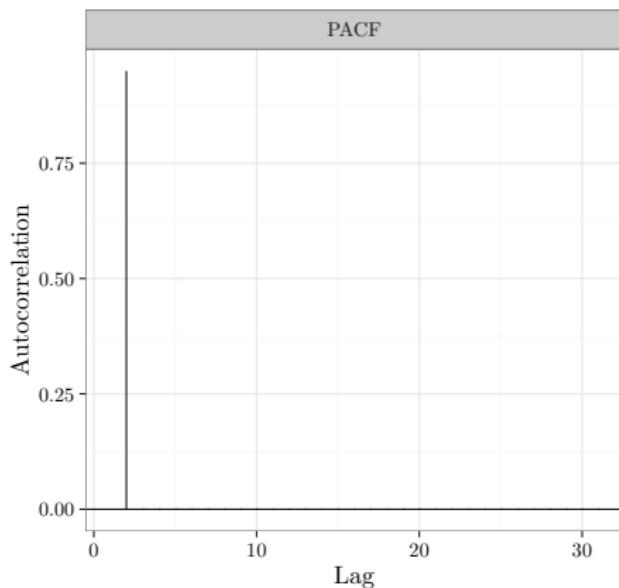
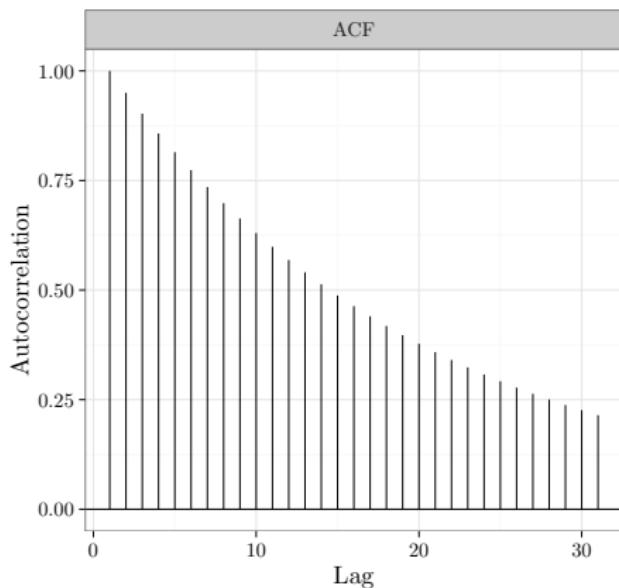
Theoretical ACF and PACF of AR(p) models



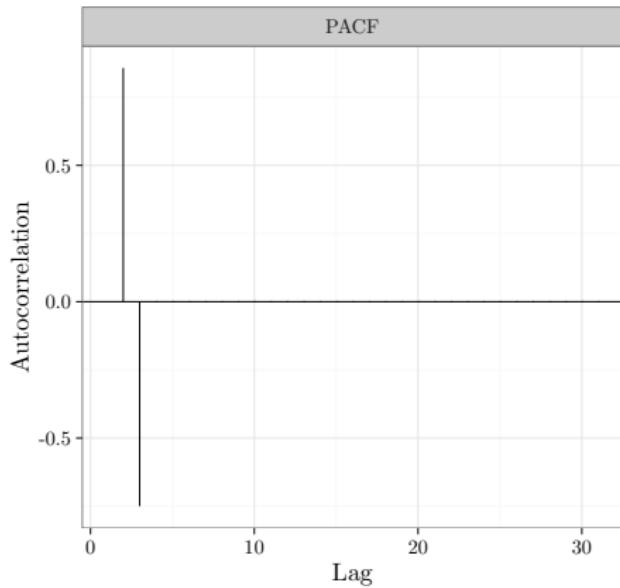
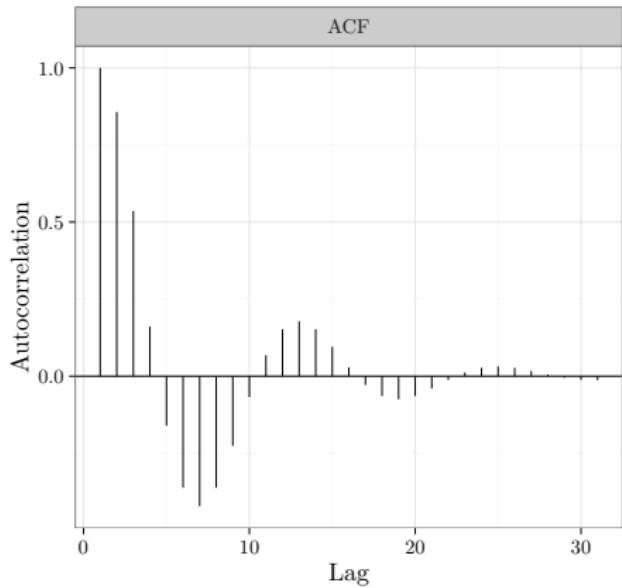
Theoretical ACF and PACF of AR(p) models



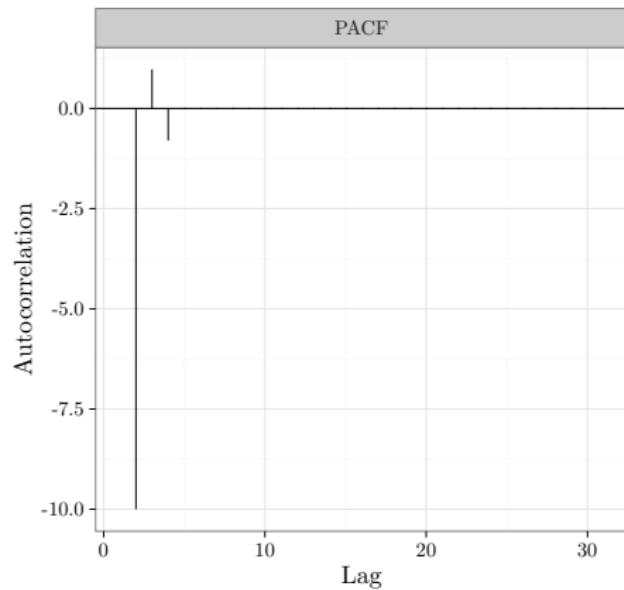
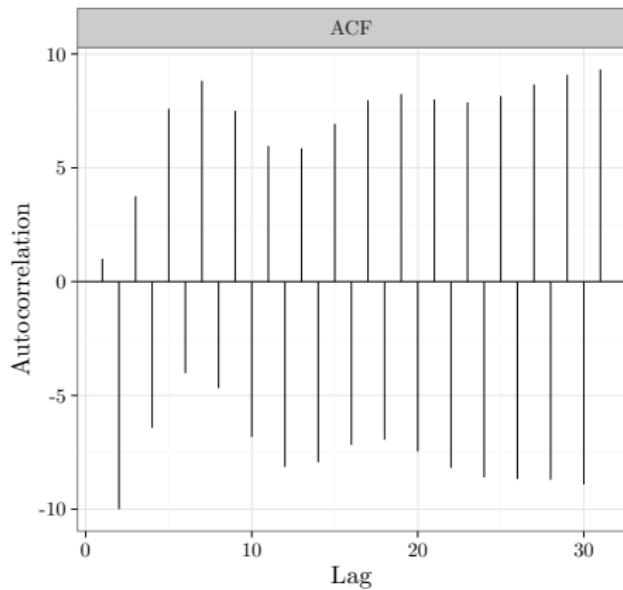
Theoretical ACF and PACF of AR(p) models



Theoretical ACF and PACF of AR(p) models



Theoretical ACF and PACF of AR(p) models



Causality of AR(p) Models

- Causality is an important property of AR(p) models.
- If a model is causal, it means (informally speaking) that this model uses only past observations (i.e. X_t, X_{t-1}, \dots) to predict the future (i.e. X_{t+1}, X_{t+2}, \dots).
- Of course, models that are not causal are rather useless to make predictions with because they need the future to predict the future!
- Let's consider a simple example with our favorite process an AR(1).

Example: Causality of an AR(1)

If $|\phi| < 1$, we already discussed that an AR(1) can be written as:

$$X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j},$$

therefore it is causal as X_t is a function of past observations of (W_t) .
If $|\phi| > 1$, something interesting happens and we have

$$X_{t-1} = \frac{1}{\phi} X_t - \frac{1}{\phi} W_t,$$

and therefore

$$X_t = - \sum_{j=1}^{\infty} \phi^{-j} W_{t+j},$$

so X_t depends on **future** observations!

Causality of AR(p) Models

To formally define causality, we first introduce the following two notational definitions.

Definition

We define the *backshift operator* by

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

Definition

The *autoregressive operator* is defined as

$$\phi(B) = 1 - \sum_{i=1}^p \phi_i B^i.$$

so AR(p) can be expressed as

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

Causality of AR(p) Models

A formal definition

An AR(p) is *causal*, iff X_t can be expressed by

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} = \psi(B) W_t, \quad t = 0, \pm 1, \pm 2, \dots,$$

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

Remarks:

- The verification of the causality of an AR(p) model is not straightforward and will not be presented here. This concept is closely related to the stationarity of AR(p) models.
- There exists several (numerical) methods that can be used to “check” if a model is stationary. These methods are implemented in most statistical software and will only estimate causal models.

Forecasting with AR(p) models

- When we forecast future values (i.e. X_{T+h} , $h > 0$), we have to estimate $\mathbb{E}[X_{T+h}|X_T, X_{T-1}, \dots, X_1]$. This quantity intuitively makes sense and relies on valid mathematical arguments (not presented here).
- Consider an AR(1), then we have

$$\mathbb{E}[X_{T+h}|X_T, X_{T-1}, \dots, X_1] = \mathbb{E}[\phi^h X_T | X_T, X_{T-1}, \dots, X_1] = \phi^h X_T.$$

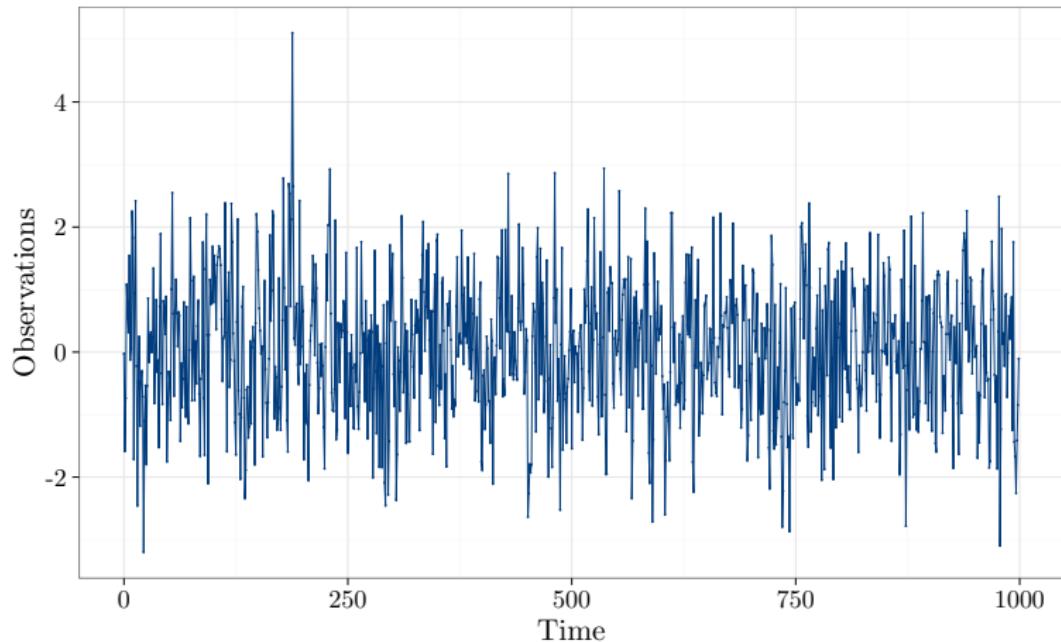
A natural estimator of the above quantity is therefore,

$$\widehat{\mathbb{E}}[X_{T+h}|X_T, X_{T-1}, \dots, X_1] = \widehat{\phi}^h X_T.$$

where $\widehat{\phi}$ is a suitable estimator (e.g. OLS, MLE) for ϕ .

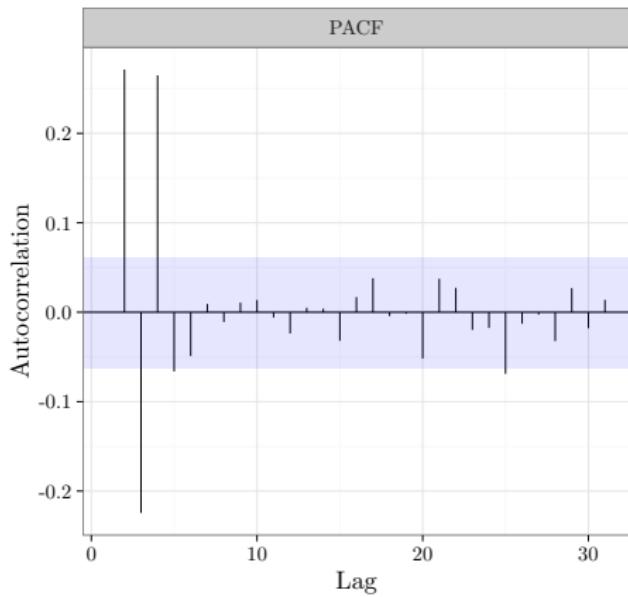
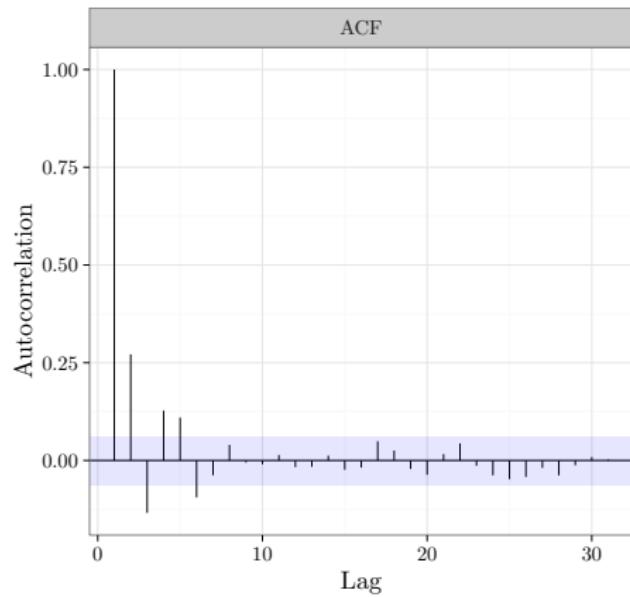
- Similar results can be obtained for an AR(p).
- This also creates a framework to compute residuals (in-sample). For example, in the case of an AR(1), we have: $r_t = X_t - \widehat{\phi} X_{t-1}$.

A Simulated Example: Order Identification



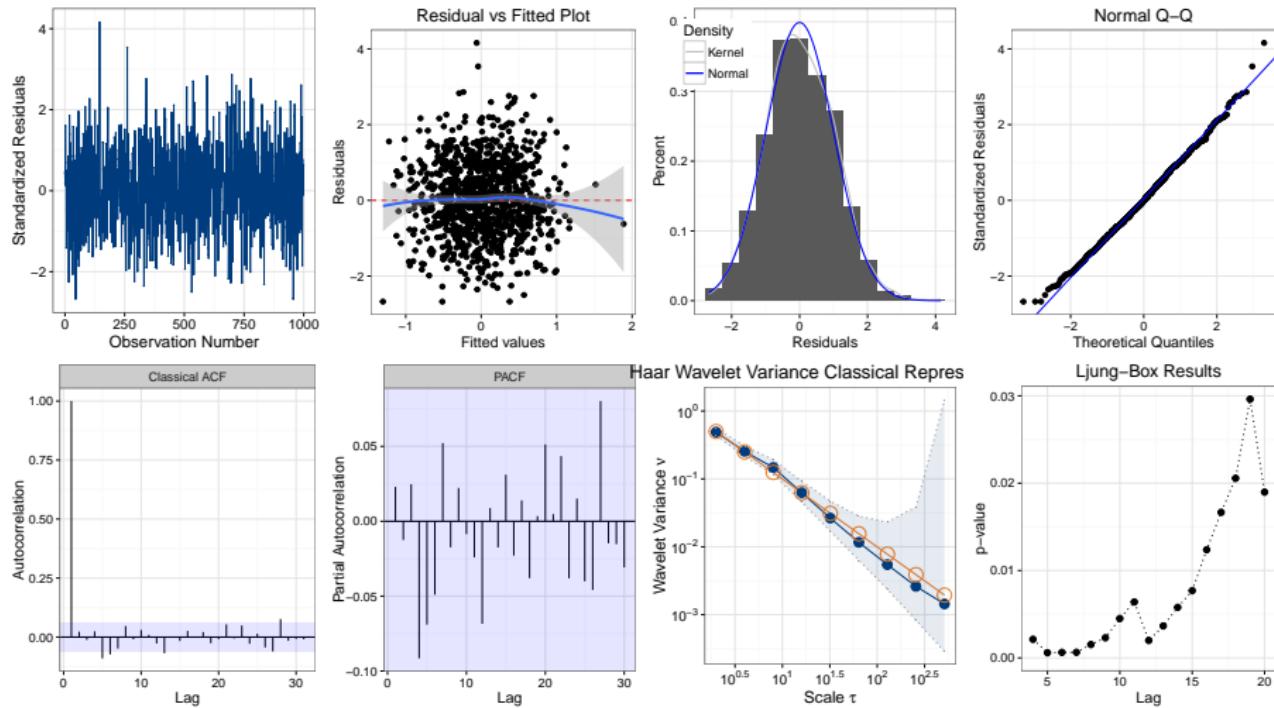
Let's try to find an appropriate AR model for this time series

A Simulated Example: ACF/PACF Graphs

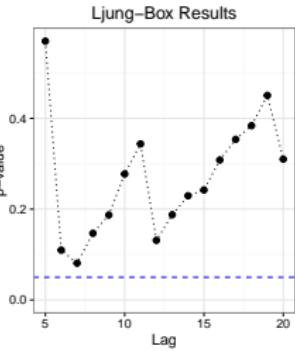
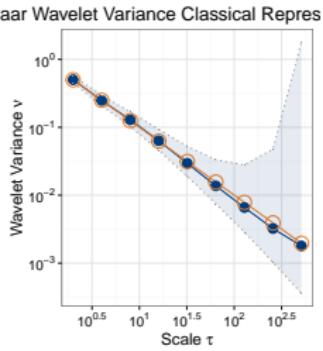
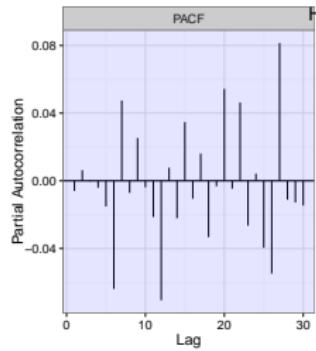
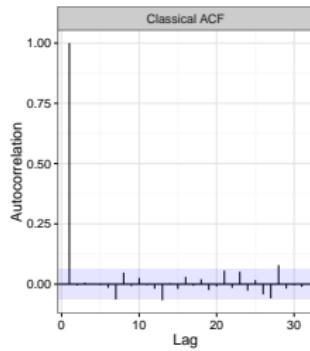
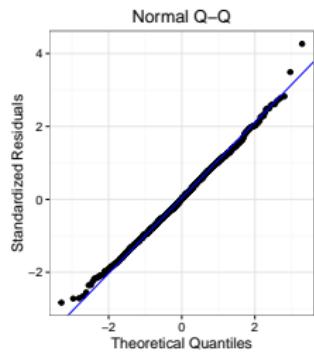
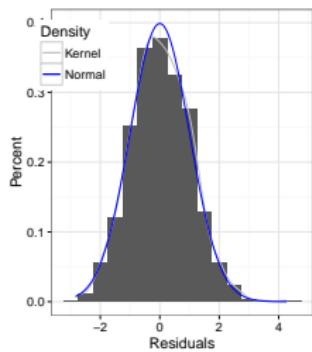
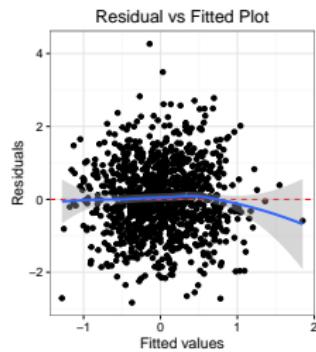
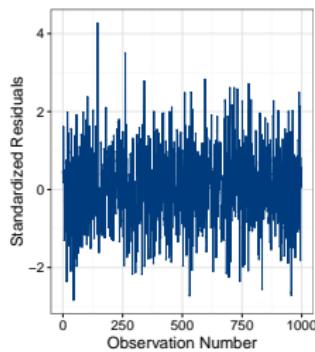


From this graph an AR(3) appears reasonable.

Diagnostic for AR(3)



Diagnostic for AR(4)



Model Selection

The approach we used in the previous slides is rather arbitrary and a method based on model selection criteria is often preferred. These criteria are estimators of the “prediction error” of a model. The most popular criteria used for model selection with time series models are presented below.

Definitions

Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn information criterion (HQ) are defined as:

$$\text{AIC} = -2 \log L + 2k$$

$$\text{BIC} = -2 \log L + \log(n)k,$$

$$\text{HQ} = -2 \log L + 2 \log(\log(n))k.$$

where L denotes the likelihood function at the estimated parameters and k the number of parameters in the candidate model.

Why are these criteria meaningful?

Consider the AIC:

The AIC is based on a divergence that informally speaking measures “how far” is the density of the estimated model compared to the “true” density. This divergence is called the Kullback-Leibler information which in this context can be defined for two densities of the same family as

$$KL = \frac{1}{n} E \left[E_0 \left[\log \left(\frac{f(y_0|\theta_0)}{f(y_0|\hat{\theta})} \right) \right] \right],$$

the expectations $E [\cdot]$ and $E_0 [\cdot]$ denote the expectation with respect to the densities of y and y_0 (conditionally on X). Informally speaking, this divergence measures how far $f(y_0|\theta_0)$ is from $f(y_0|\hat{\theta})$, where in the latter $\hat{\theta}$ is estimated on y , a sample independent from y_0 . The AIC is an unbiased estimator of KL .

Simulation Study

Consider two models:

$$X_t = 0.6X_{t-1} + 0.3X_{t-2} + W_t$$

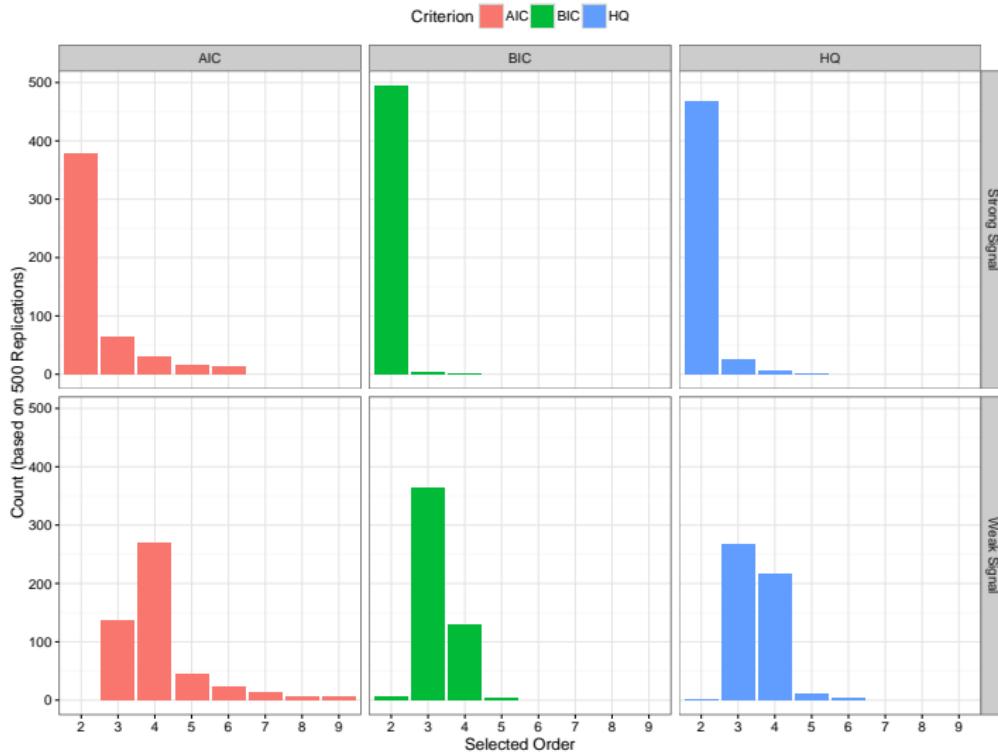
$$X_t = 0.5X_{t-1} + 0.25X_{t-2} + 0.125X_{t-3} + 0.0625X_{t-4} + W_t$$

The first one corresponds to a “strong” signal while the second one is called a “weak” signal. Based on the theoretical (i.e. asymptotic) properties of model selection criteria, we expect the BIC to perform very well for the first model, while the AIC should give “good” results on the second model.

Setting:

Sample size $T = 100$, Bootstrap replications $B = 500$.

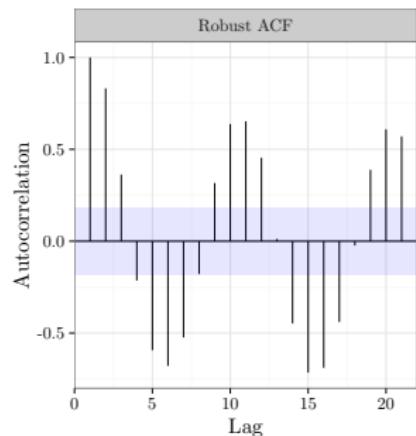
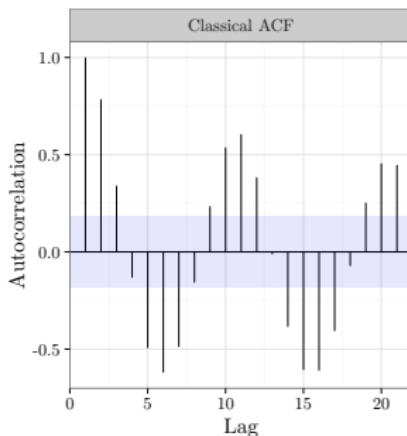
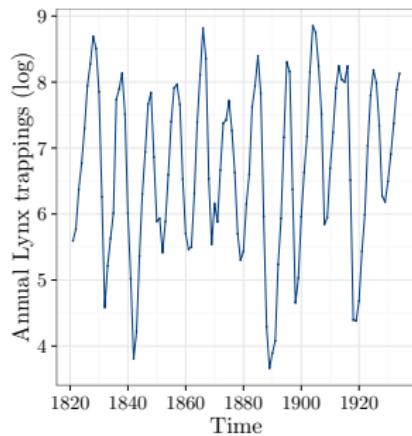
Simulation Study



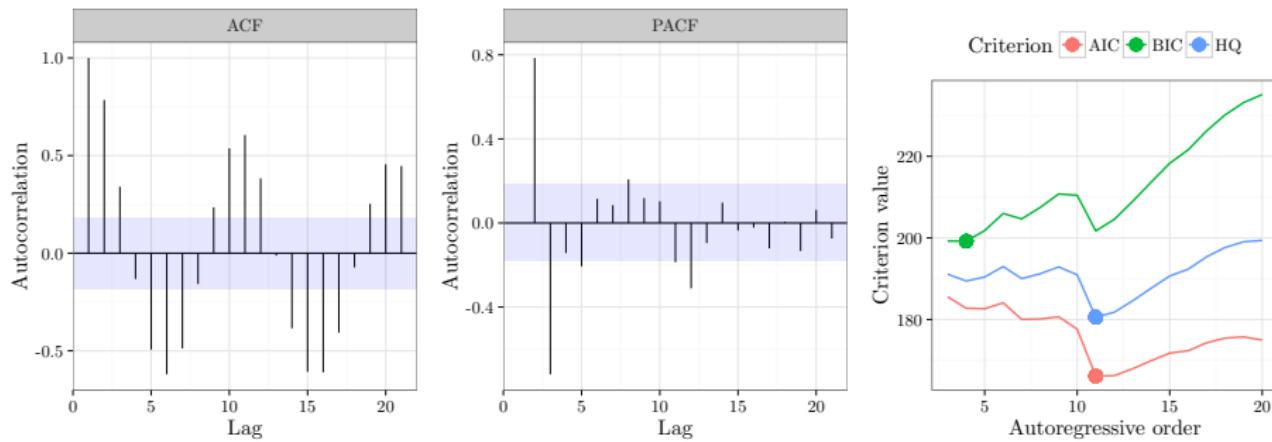
Application to Real Data: Lynx Dataset

Background:

We consider the annual numbers of lynx trappings for 1821-1934 in Canada.

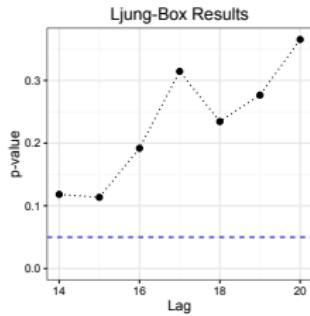
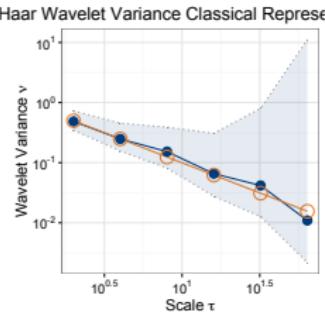
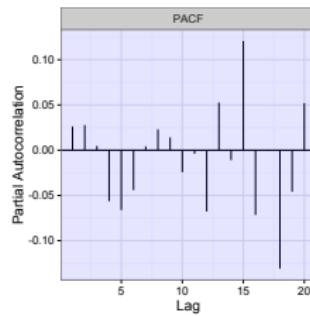
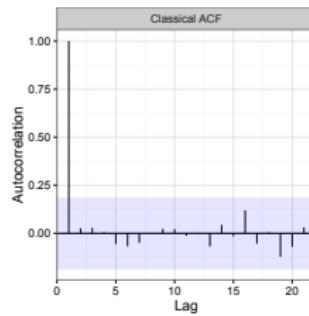
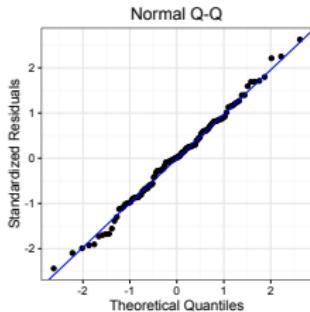
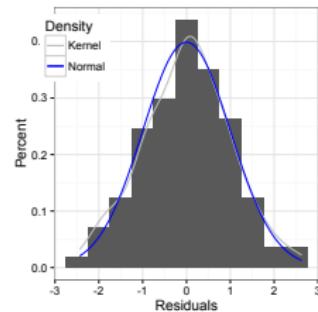
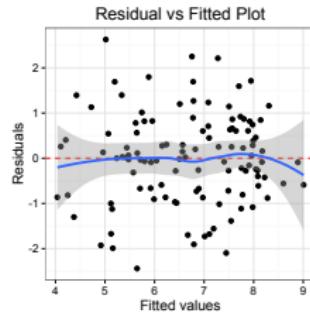
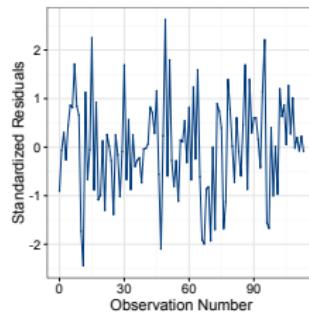


Application to Real Data: Lynx Dataset

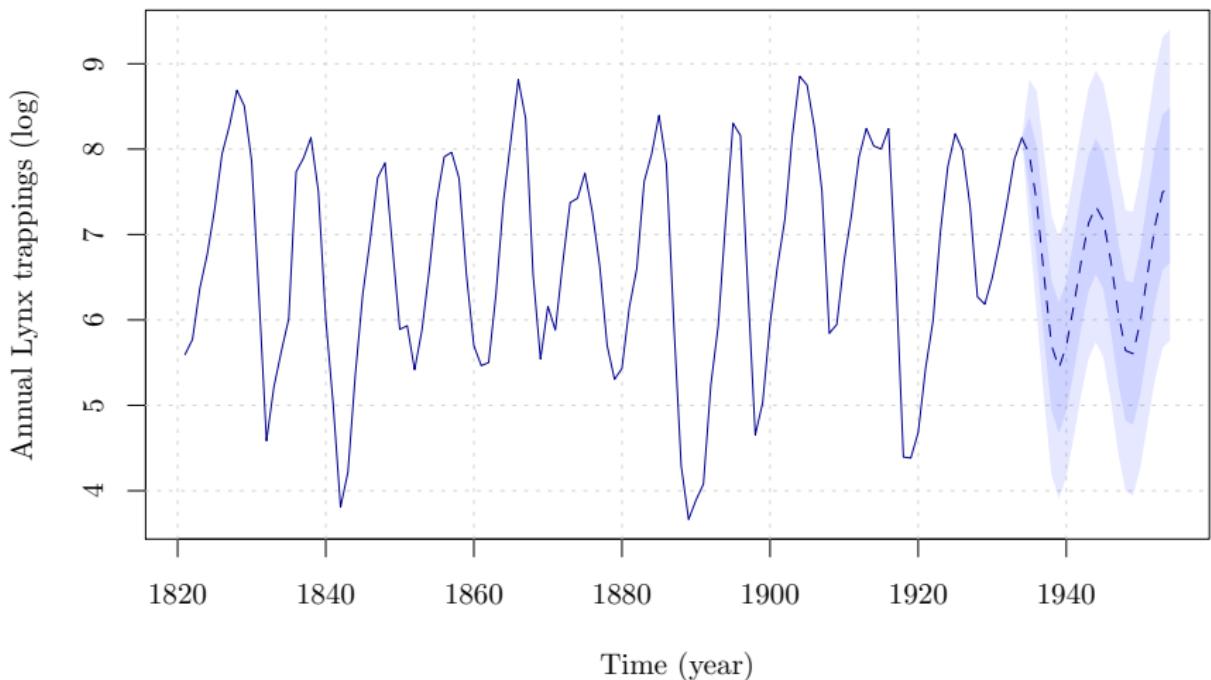


According to both AIC and HQ criterion an AR(11) seems appropriate.

Application to Real Data: Lynx Dataset



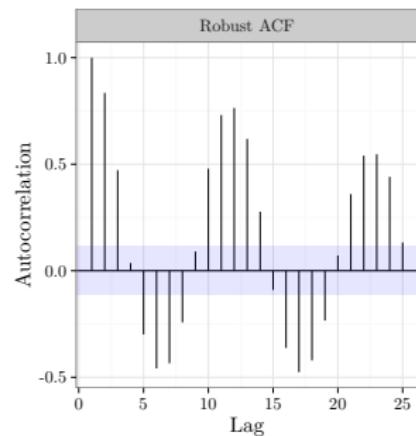
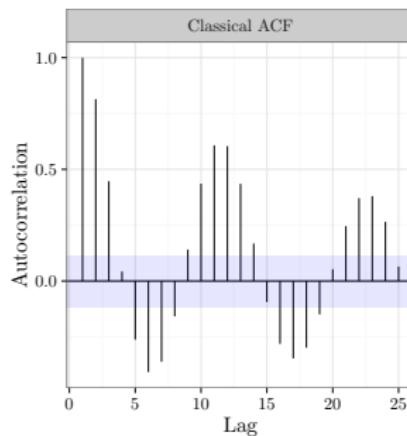
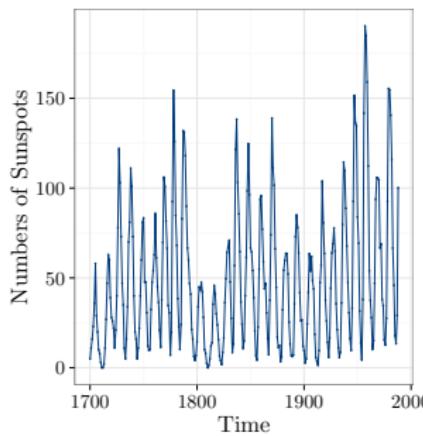
Application to Real Data: Lynx Dataset



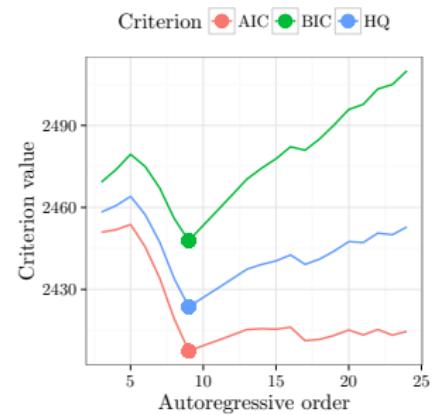
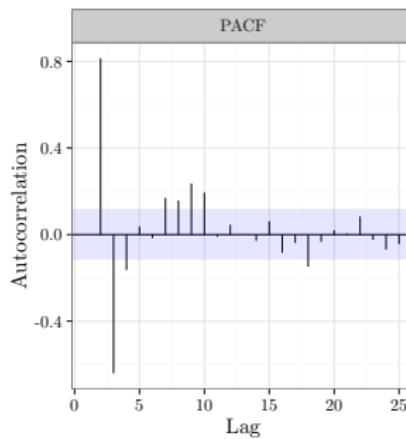
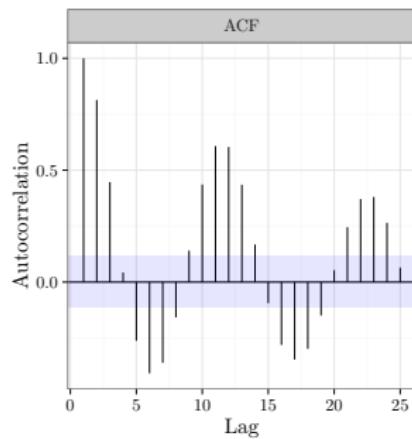
Application to Real Data: Sunspot Dataset

Background:

We consider the yearly numbers of sunspots from 1700 to 1988.

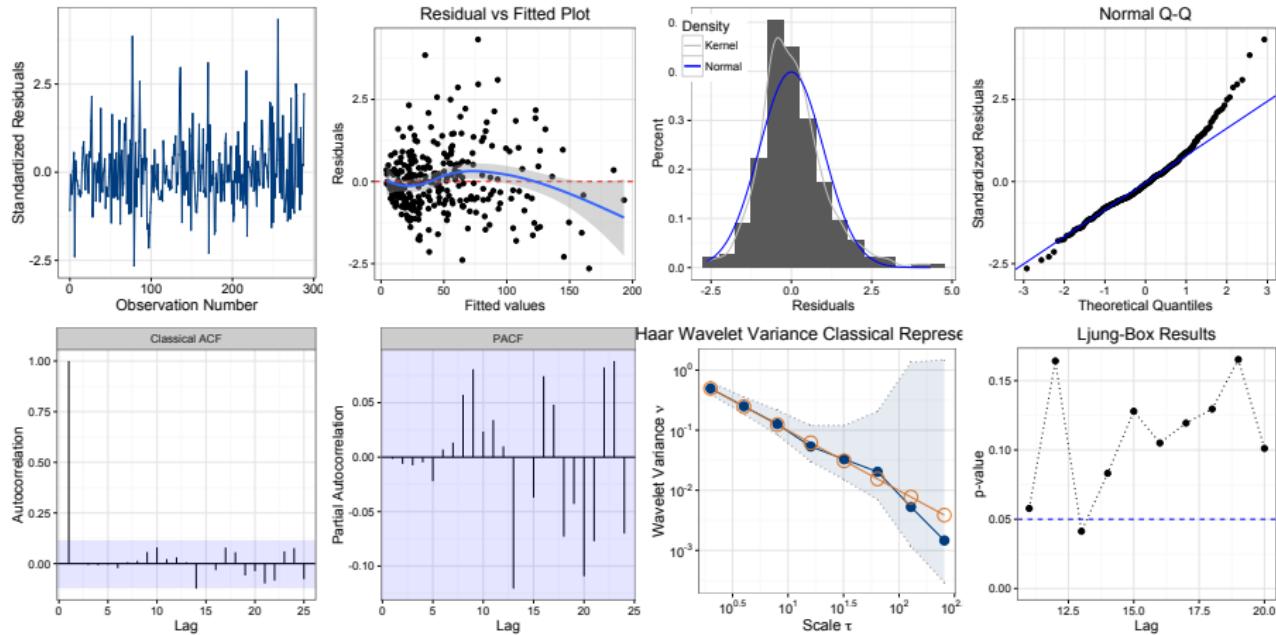


Application to Real Data: Sunspot Dataset

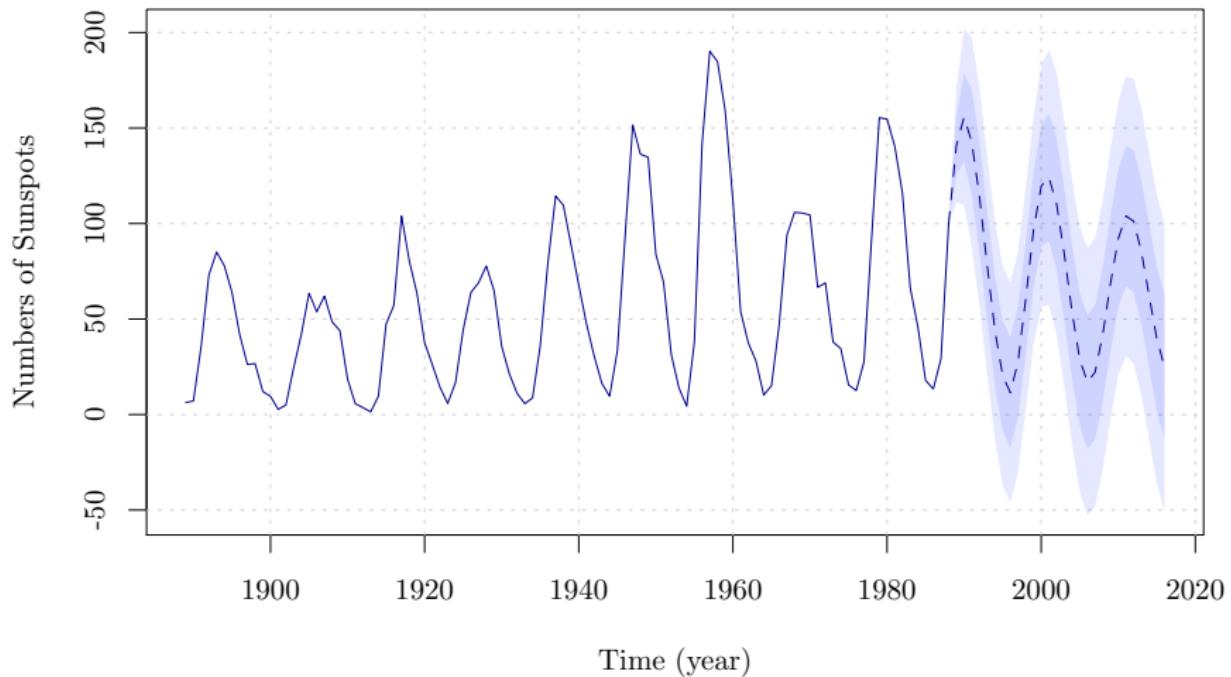


All criteria indicate that an AR(9) is appropriate.

Application to Real Data: Sunspot Dataset



Application to Real Data: Sunspots Dataset



Measuring Forecast Accuracy

The methods we used for model selection in the previous two examples are based on **strong parametric assumptions**. If your true model is indeed an AR(p), than they are (in the some sense) “optimal”. However, if this is not the case cross-validation like techniques are more appropriate.

A possible method:

- ① Split your time series of length T into two sub-series. The first one (i.e. training set) goes from 1 to n (where n is for example $0.8T$) and the second one (i.e. testing set) goes from $n + 1$ to T .
- ② Estimate the model you wish to evaluate on the training set and forecast next observation (i.e. X_{n+1}). Compute the difference between your forecast and the actual value of X_{n+1} .
- ③ Add observation X_{n+1} to the training set and let $n = n + 1$. Go to Step 2 until $n = T$.
- ④ Compute a suitable “score” to asses the quality of your model based on the empirical “prediction errors” vector.

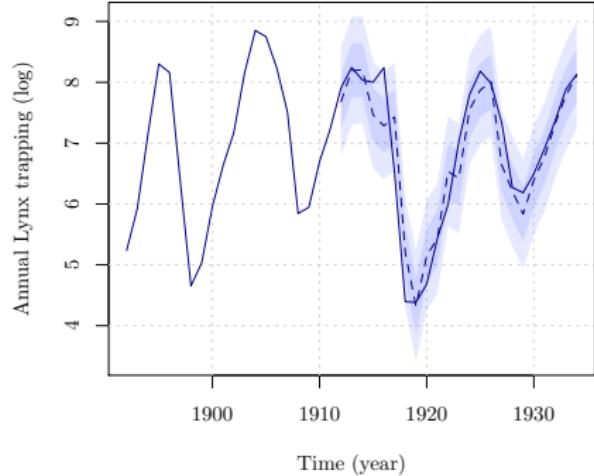
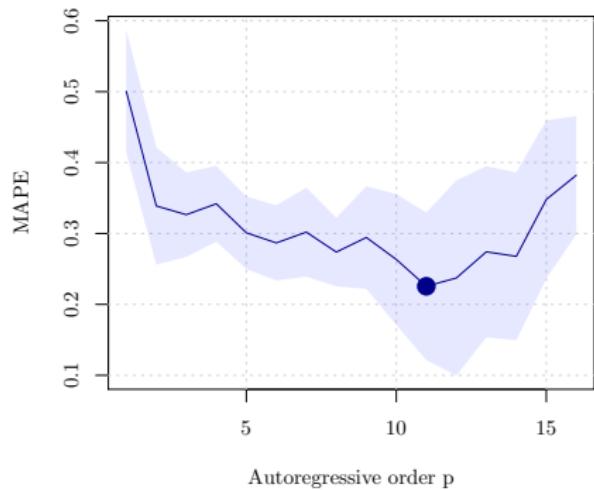
Measuring Forecast Accuracy

- Using the vector of differences between your forecast and the actual values of the time series, we can compute different metrics.
- Let $\hat{X}_h^{(1:j)}$ denote your prediction for the observation X_{j+1} based on a model whose parameters were estimated on the sample X_1, \dots, X_j , then the Median Absolute Prediction Error (or MAPE) is defined as

$$\text{MAPE} = \text{median} \left(\left(|\hat{X}_1^{(1:j)} - X_{j+1}| \right)_{j=1, \dots, n-1} \right).$$

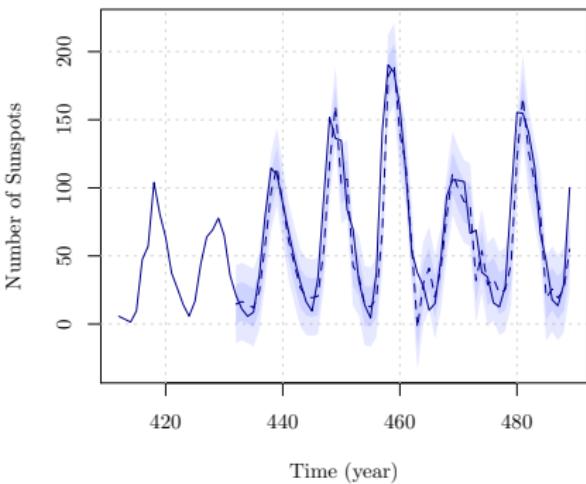
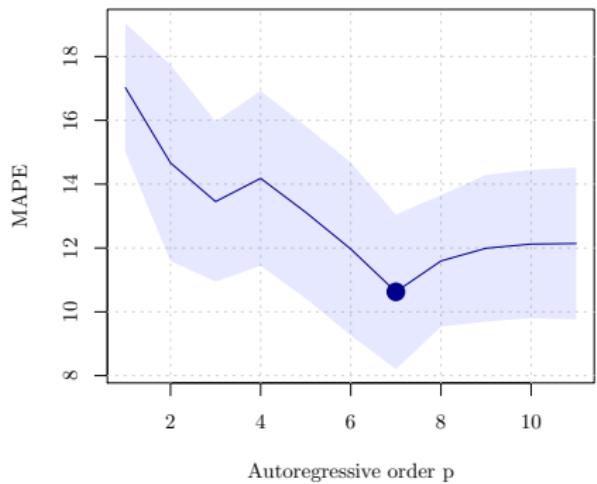
- Many other prediction errors can be derived such as the root mean squared error, mean absolute percentage error, mean absolute scaled error, and so on.
- The MAPE (and other similar metrics) can be used to assess the generalization of a model and to select a model.
- Let's see some examples...

MAPE for Lynx Dataset



This confirms that an AR(12) is a “good” choice.

MAPE for Sunspots Dataset



This suggests that an AR(7) might be sufficient instead of the AR(9) suggested by the AIC, BIC and HQ.

Moving Average Models

Definition

The *moving average model* of order q , or MA(q) model is defined as

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\theta_1, \dots, \theta_q$ are constant ($\max_{i=1, \dots, q} |\theta_i| > 0$) and $W_t \sim \mathcal{N}(0, \sigma^2)$.

Definition

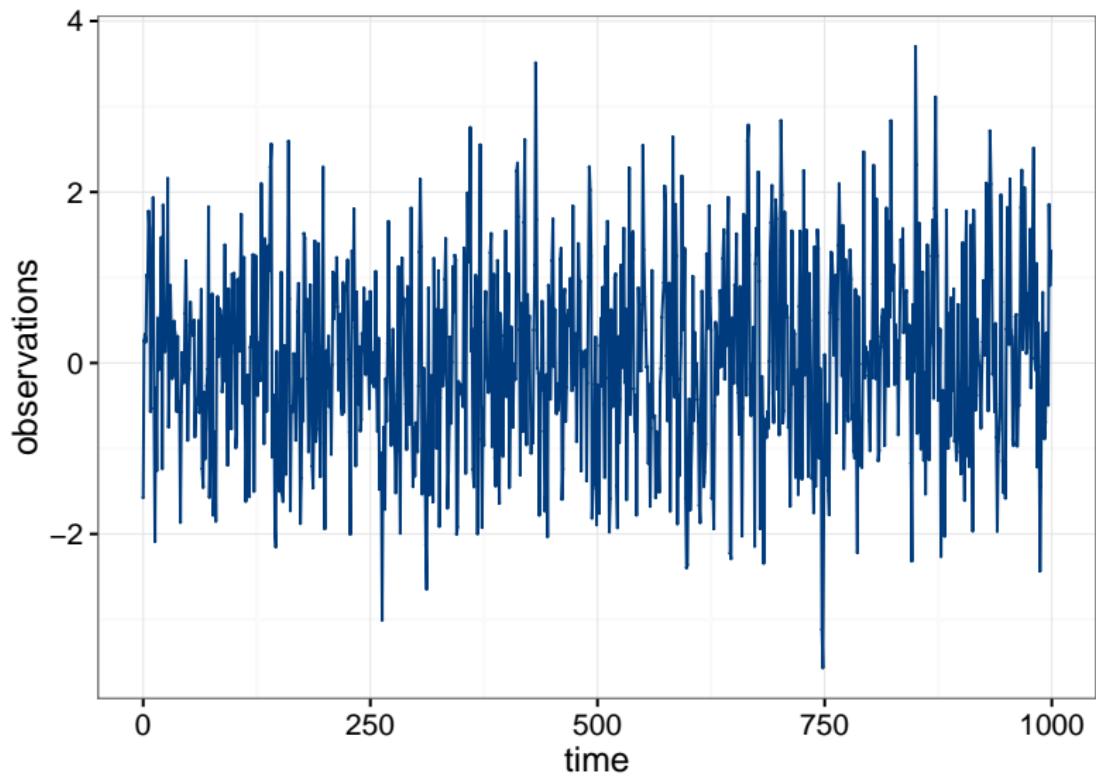
The *moving average operator* is defined

$$\theta(B) = 1 + \sum_{i=1}^q \theta_i B^i,$$

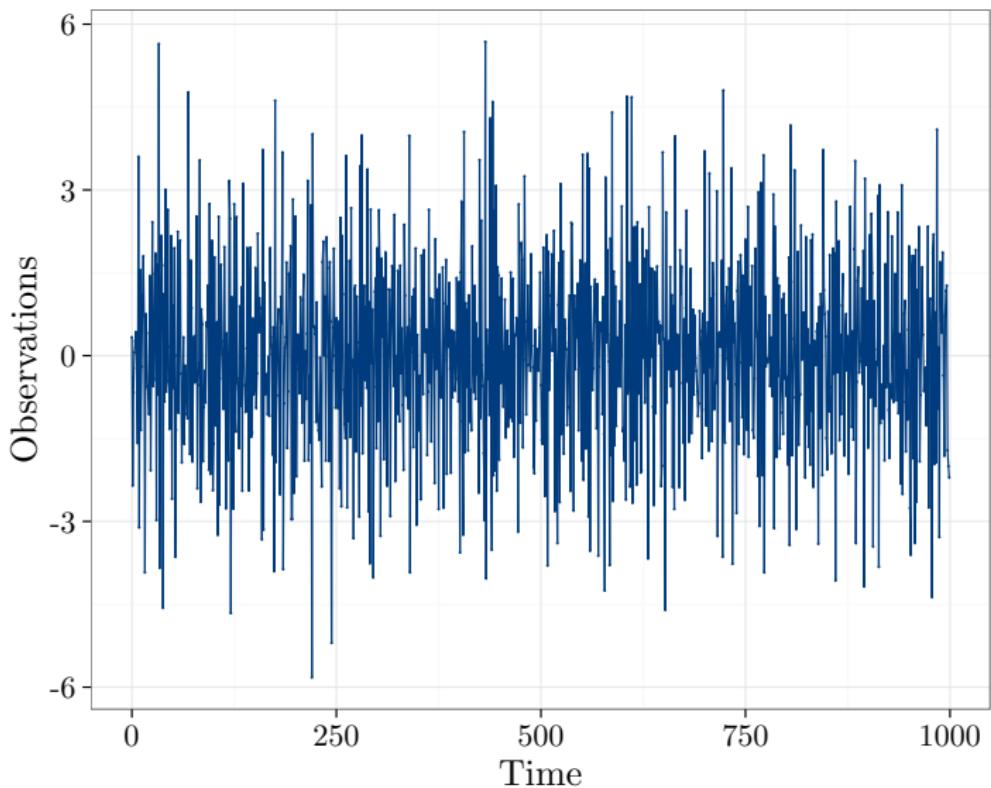
so an MA(q) can be expressed as

$$X_t = \theta(B) W_t.$$

Simulation of a MA(1) process ($\theta = 0.4$)



Simulation of a MA(1) process ($\theta = 0.65$)



Properties of an MA(1)

Definition

Consider an MA(1) process:

$$X_t = \theta W_{t-1} + W_t.$$

Then, its expected value is

$$\mathbb{E}[X_t] = \theta\mathbb{E}[W_{t-1}] + \mathbb{E}[W_t] = 0.$$

Its autocovariance function is given by:

$$\gamma(h) = \text{cov}(X_t, X_{t-h}) = \begin{cases} (1 + \theta^2)\sigma^2 & \text{if } h = 0, \\ \theta\sigma^2 & \text{if } |h| = 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

Properties of an MA(1)

Definition

If we divide the autocovariance by $\gamma(0)$, we obtain:

$$\rho(h) = \text{corr}(X_t, X_{t-h}) = \begin{cases} 1 & \text{if } h = 0, \\ \frac{\theta}{1+\theta^2} & \text{if } |h| = 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

Finally, the partial autocorrelation function is given by:

$$\phi_{k,k} = -\frac{(-\theta)^k (1 - \theta^2)}{(1 - \theta^{2(k+1)})}, \quad k \geq 1.$$

Non-uniqueness of MA Models

MA models are not unique:

Consider two processes:

$$\begin{aligned} X_t &= 2W_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, 1), \\ Y_t &= 0.5U_{t-1} + U_t, \quad U_t \sim \mathcal{N}(0, 4). \end{aligned}$$

For an MA(1) (i.e. $X_t = \theta W_{t-1} + W_t$), we have

$$\gamma(h) = \begin{cases} (\theta^2 + 1)\sigma^2 & \text{if } h = 0 \\ \theta\sigma^2 & \text{if } |h| = 1 \\ 0 & \text{if } |h| > 1 \end{cases}$$

Then, $\gamma_x(0) = 5$, $\gamma_x(1) = 2$, $\gamma_y(0) = 5$ and $\gamma_y(1) = 2$, so (X_t) and (Y_t) are the same process (since autocovariance corresponds to a fundamental representation with Gaussian processes)!

Invertibility of MA models

Definition

An MA(q) is *invertible* if it can be written as

$$\pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = W_t, \quad t = 0, \pm 1, \pm 2, \dots,$$

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

- By considering only invertible MA processes we ensure that they are **unique**.
- Invertibility for MA(q) plays a less much important role than causality for AR(p).
- **Example:** An MA(1) is invertible if $|\theta| < 1$.
- Similarly to causality “checks”, there exist (numerical) methods that can be used to assess this. These methods are implemented in statistical software and only invertible MA processes will be estimated.

Forecasting with MA processes

Forecasting is more difficult with $MA(q)$ processes. Indeed, if (W_t) was observed together with (X_t) , we would have simply consider

$$\mathbb{E}[X_{T+1} | X_T, \dots, X_1, W_T, \dots W_1] = \theta W_T,$$

and use $\hat{\theta}W_T$ as an estimator of the above quantity. Unfortunately, the process (W_T) is unobserved and several methods can be used. The most common ones are the innovation algorithm and the Kalman filter. We will not discuss the details on these techniques but the techniques are implemented in most statistical software.

Theoretical ACF and PACF of MA(q) models

- In general, the PACF of an MA(q) is a complicated function of the parameters but $\phi(h)$ tends to dampen, in a sinusoidal fashion, exponentially fast to zero as h increases.
- The ACF of an MA(q) is non-zero for the first q lags and zero for $h > q$.
- Example: consider the following models:

$$X_t = -0.5W_{t-1} + W_t$$

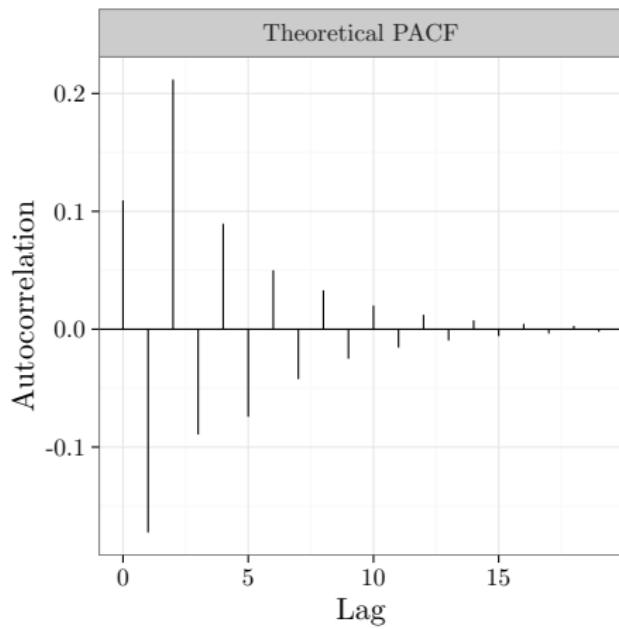
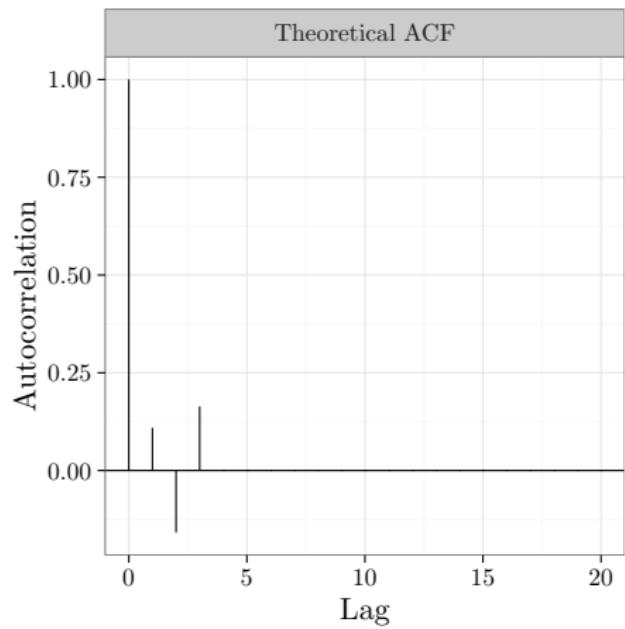
$$X_t = -0.9W_{t-1} - 0.5W_{t-2} + W_t$$

$$X_t = 0.7W_{t-1} - 0.5W_{t-2} + 0.3W_{t-3} + W_t$$

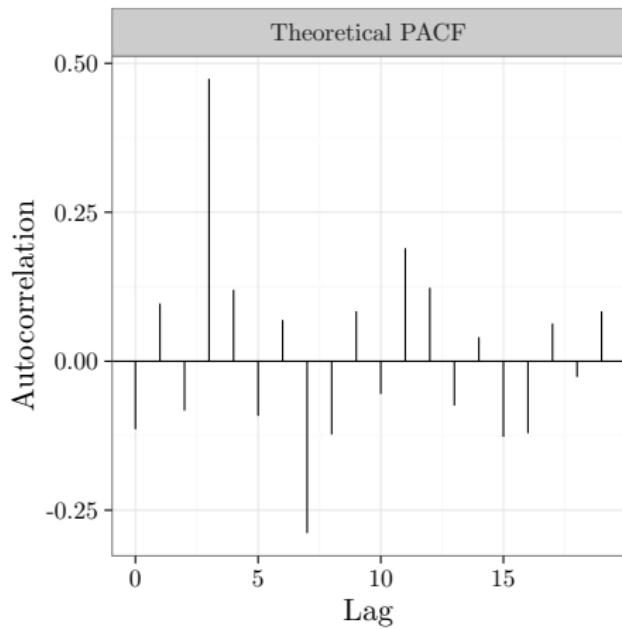
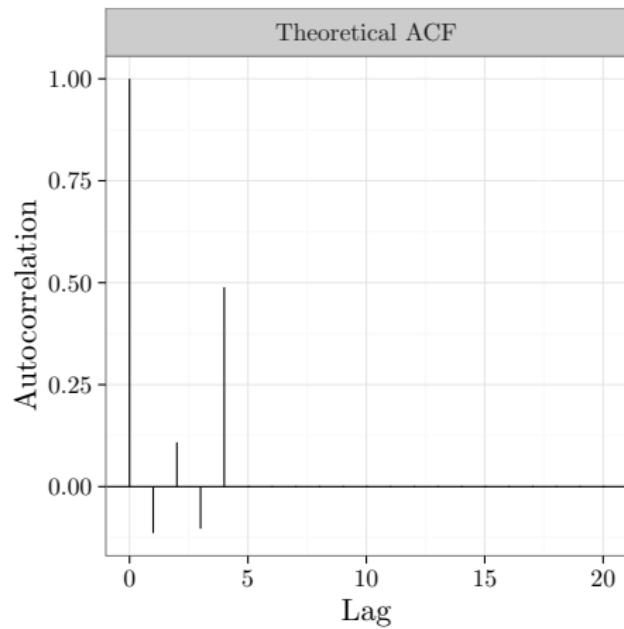
$$X_t = -0.1W_{t-1} + 0.1W_{t-2} - 0.1W_{t-3} + 0.9W_{t-4} + W_t$$

Can we recognize each model given their theoretical ACF/PACF?
Let's try...

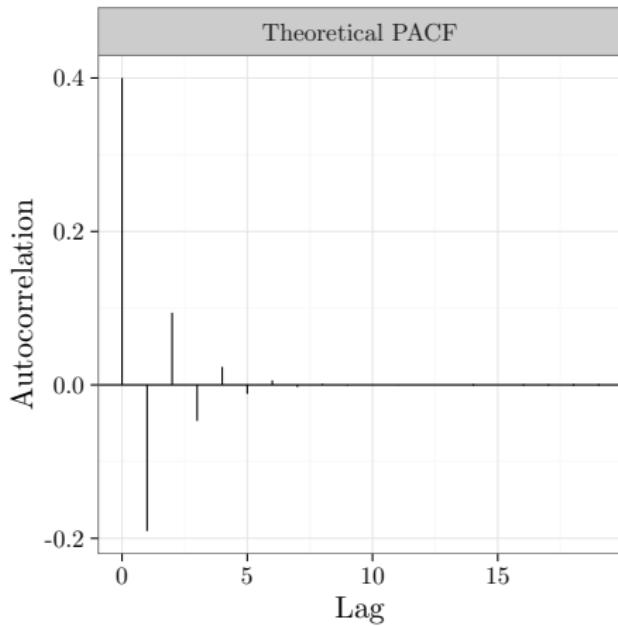
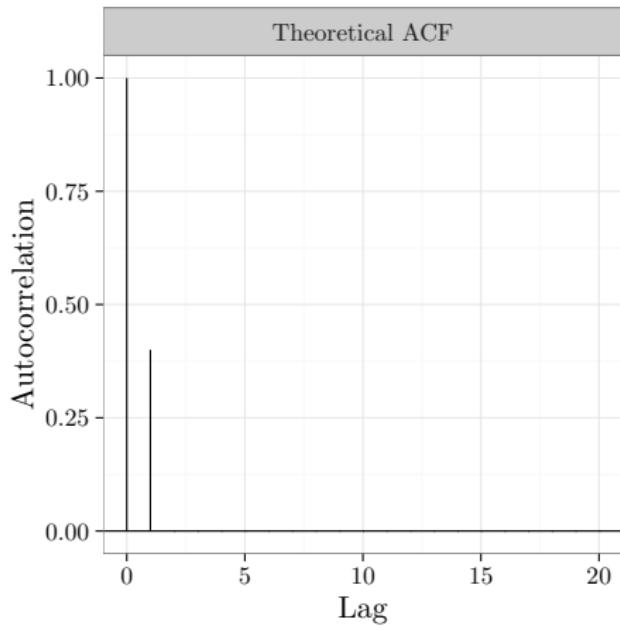
Theoretical ACF and PACF example



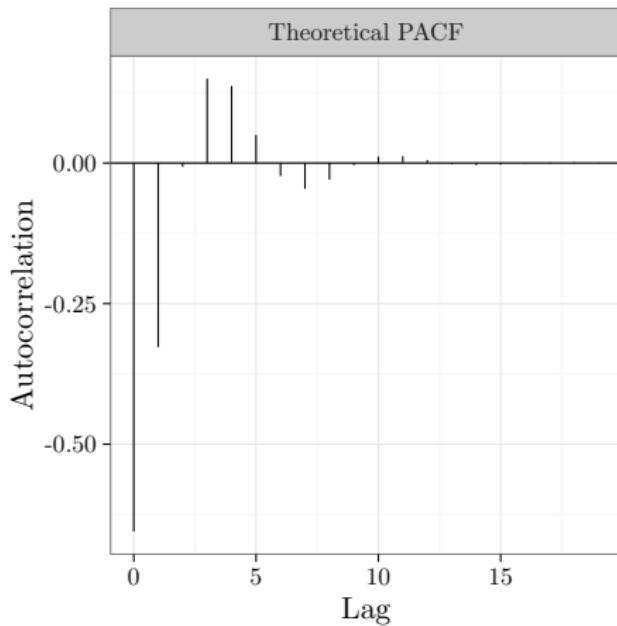
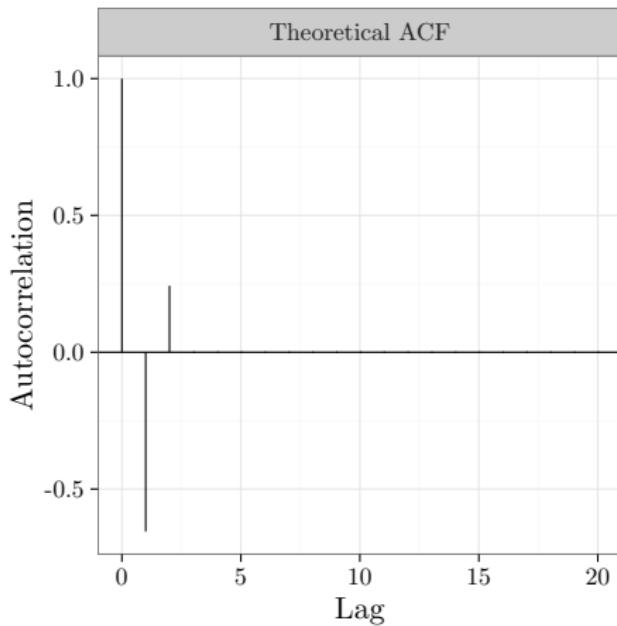
Theoretical ACF and PACF example



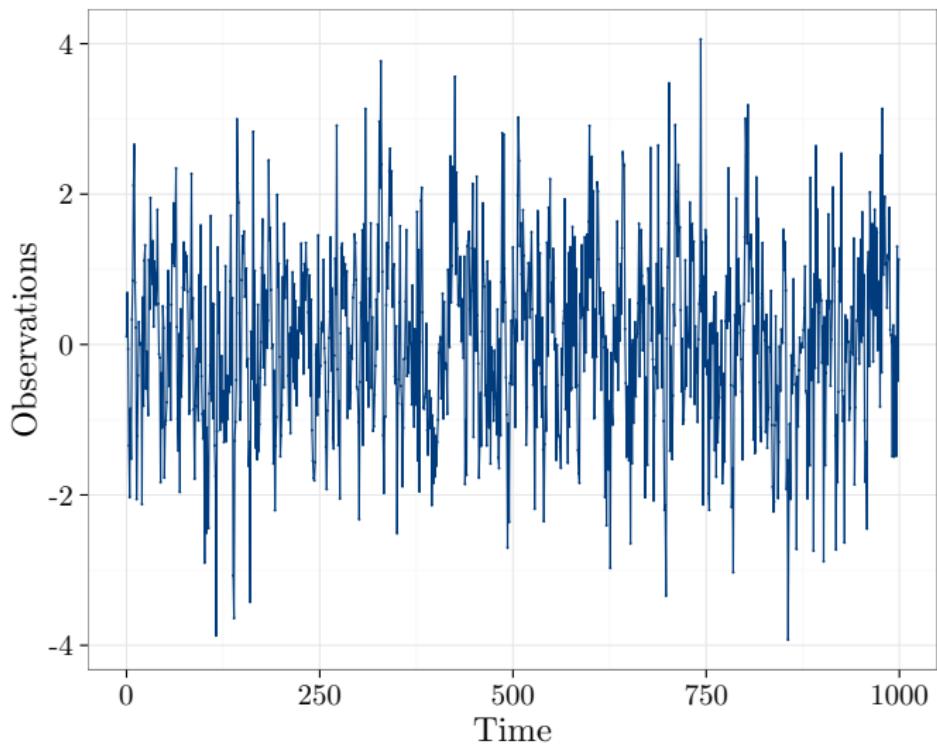
Theoretical ACF and PACF example



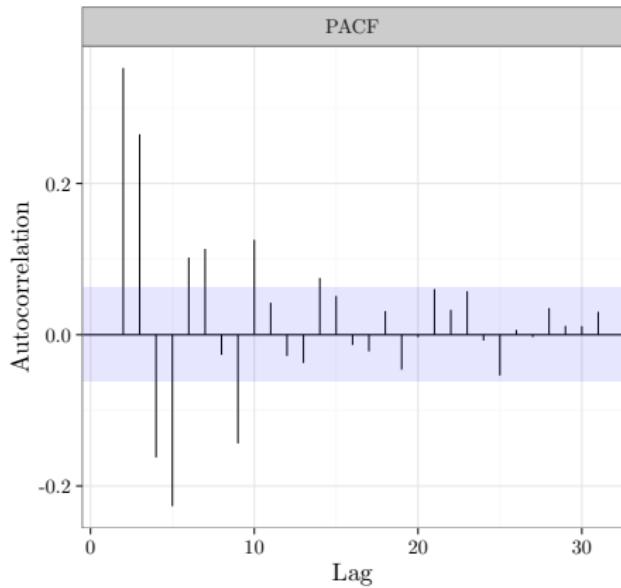
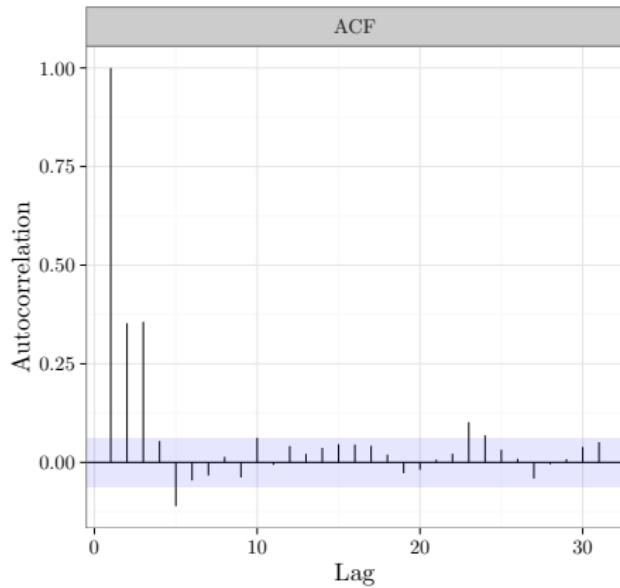
Theoretical ACF and PACF example



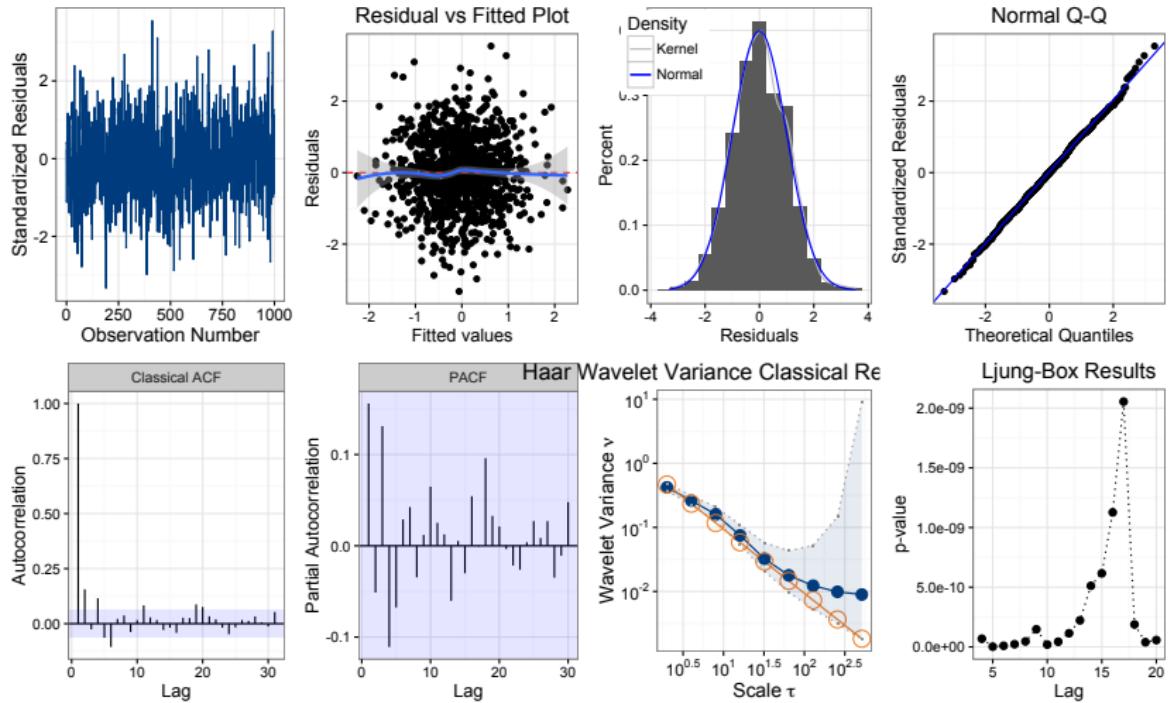
A Simulated Example: Order Identification



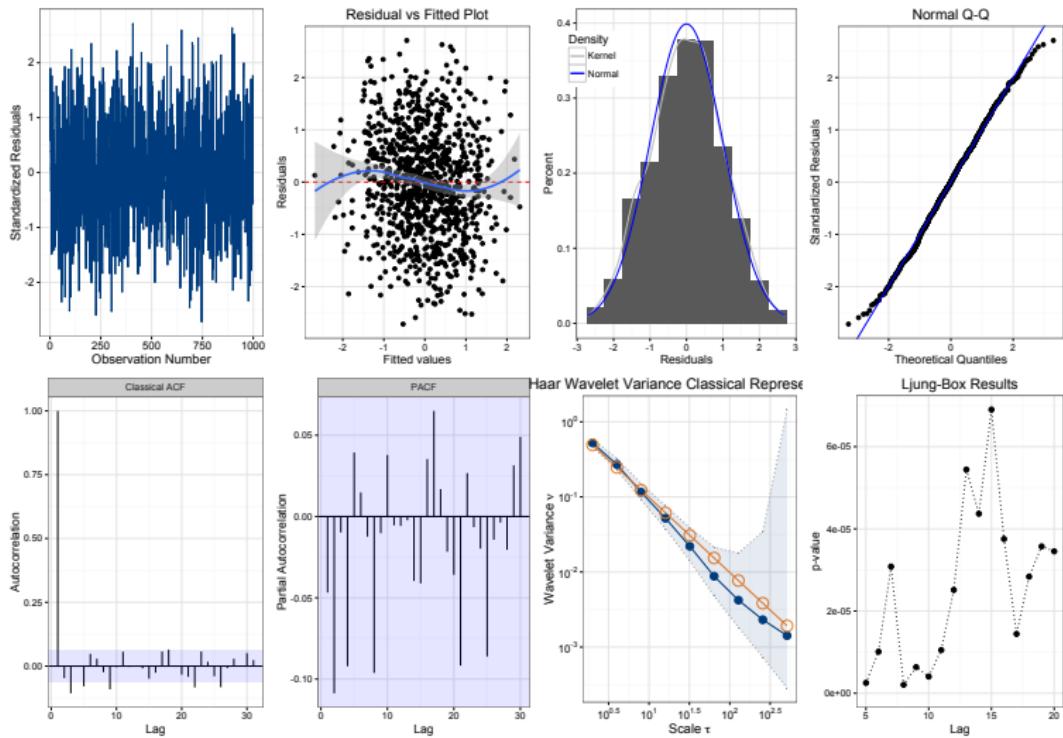
A Simulated Example: ACF/PACF Graphs



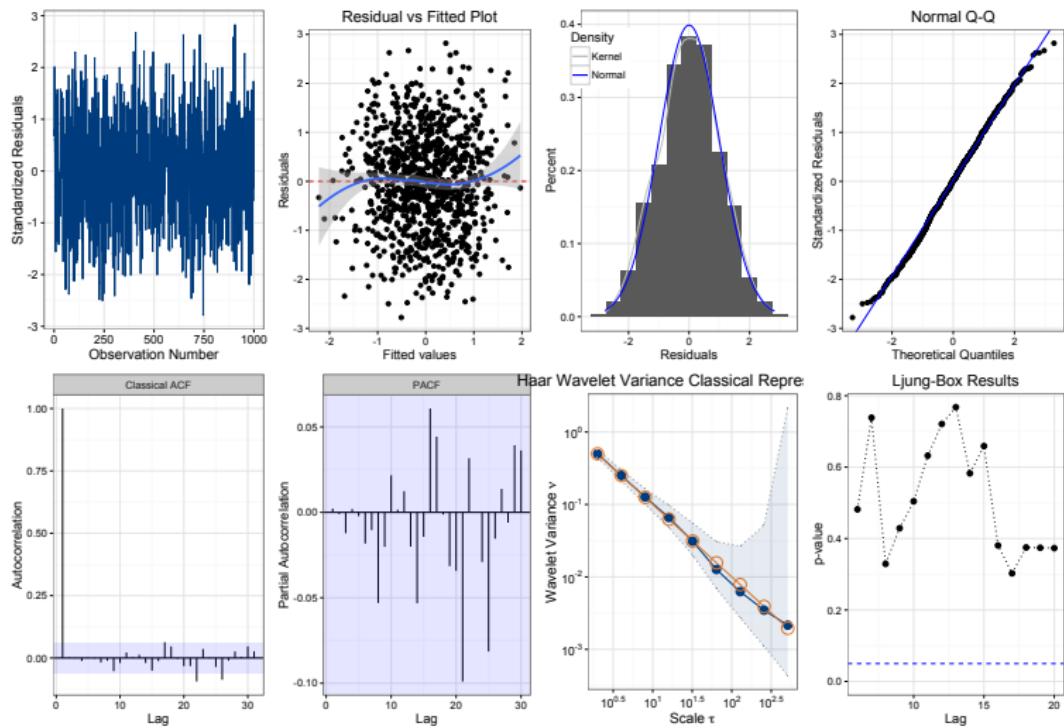
Diagnostic for MA(2)



Diagnostic for MA(3)



Diagnostic for MA(4)



ARMA models

Definition:

A process (X_t) is an ARMA(p, q) process if (X_t) (or $(X_t - \mathbb{E}[X_t])$) satisfies the linear difference equation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $W_t \sim \mathcal{N}(0, \sigma_w^2)$. An ARMA(p, q) can be written in concise form as:

$$\phi(B) X_t = \theta(B) W_t,$$

where $\phi(z)$ and $\theta(z)$ are AR and MA polynomials:

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

ARMA models

Remarks:

- ① Assume $\mathbb{E}(X_t) = \mu = 0$. Otherwise you must substitute $X_t - \mu$ for X_t .
- ② If $p = 0$, then $\text{ARMA}(p = 0, q) = \text{MA}(q)$ process.
- ③ If $q = 0$, then $\text{ARMA}(p, q = 0) = \text{AR}(p)$ process.
- ④ We mainly work with ARMA models that are causal and invertible.
 - An ARMA model is causal and invertible if its AR part is causal and its MA part is invertible.
 - $\text{AR}(p)$ models are always *invertible* and $\text{MA}(q)$ models are always *causal*.
- ⑤ Forecasting with ARMA process relies on the same techniques used with MA process.

Parameter Redundancy of ARMA Models

ARMA can have “redundant” parameters:

Consider the following example:

$$X_t = W_t$$

$$X_t - 0.9X_{t-1} = W_t - 0.9W_{t-1}$$

$$X_t = 0.9X_{t-1} + W_t - 0.9W_{t-1}.$$

Therefore, $X_t = W_t$ can be written as $X_t = 0.9X_{t-1} + W_t - 0.9W_{t-1}$, which looks a lot like an ARMA(1,1).

Parameter Redundancy of ARMA Models

Identifying Redundant Parameters

To assess parameter redundancy of ARMA models, it is useful to express the models in operator form, in the case of the previous example:

$$X_t = 0.9X_{t-1} + W_t - 0.9W_t \iff (1 - 0.9B)X_t = (1 - 0.9B)W_t.$$

It clearly appears that $(1 - 0.9B)$ can be simplified in the above equation yielding to our original model $X_t = W_t$. **In general, if a model has autoregressive and moving average operators that share a common root then the model has redundant parameters.**

Parameter Redundancy of ARMA Models

Reducing the Parameter Redundancy of an ARMA

Consider the following example:

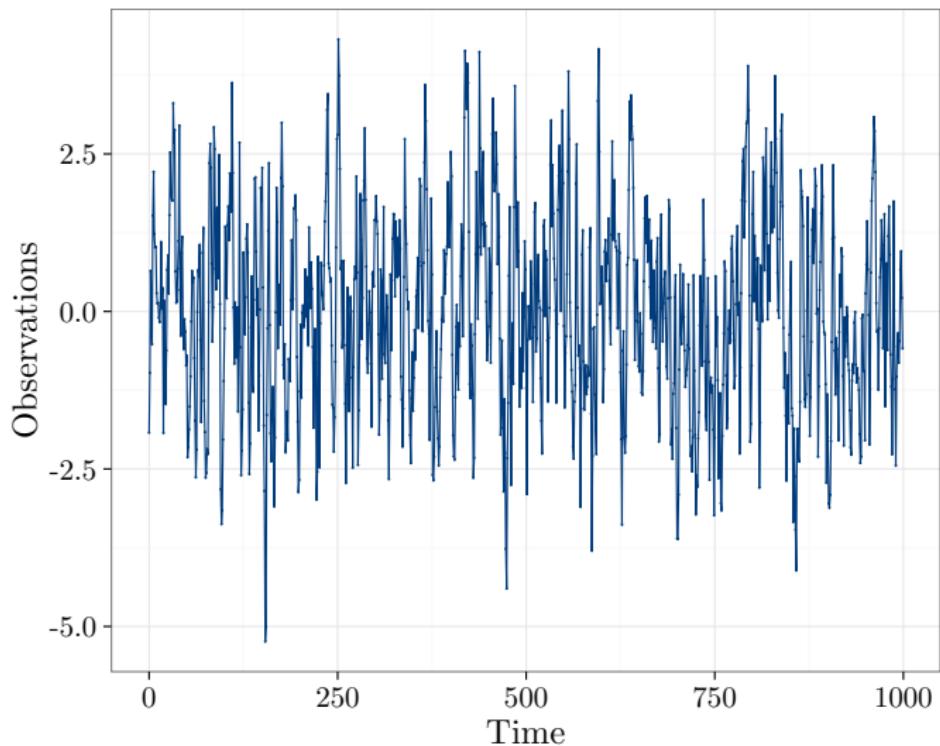
$$X_t = 0.3X_{t-1} + 0.1X_{t-2} + W_t + W_{t-1} + 0.16W_{t-2},$$

which is an ARMA(2,2). By rearranging the terms, we obtain

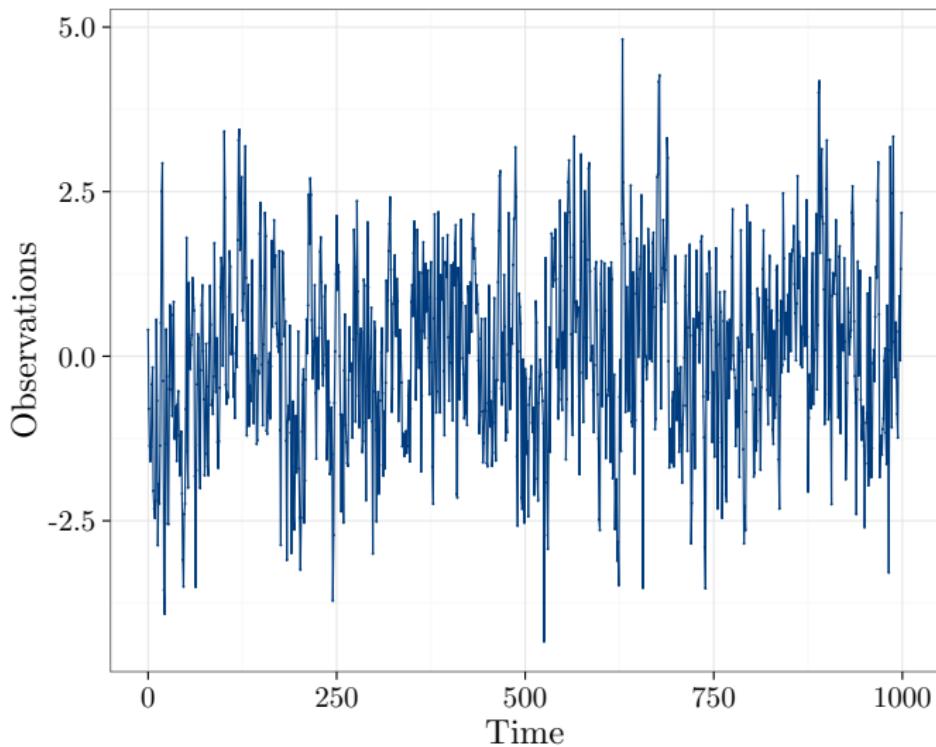
$$\begin{aligned} X_t &= 0.3X_{t-1} + 0.1X_{t-2} + W_t + W_{t-1} + 0.16W_{t-2} \\ X_t - 0.3X_{t-1} - 0.1X_{t-2} &= W_t + W_{t-1} + 0.16W_{t-2} \\ (1 - 0.3B - 0.1B^2)X_t &= (1 + B + 0.16B^2)W_t \\ (1 + 0.2B)(1 - 0.5B)X_t &= (1 + 0.2B)(1 + 0.8B)W_t \\ (1 - 0.5B)X_t &= (1 + 0.8B)W_t \\ X_t &= 0.5X_{t-1} + W_t - 0.8W_{t-1}. \end{aligned}$$

Therefore, our initial model is in fact an ARMA(1,1). Note that this model is causal (as $|\phi| < 1$) and invertible (as $|\theta| < 1$).

Simulation of a ARMA(1,1) ($\phi = 0.5, \theta = 0.5$)



Simulation of a ARMA(1,2) ($\phi = 0.9$, $\theta_1 = 0.1$, $\theta_2 = -0.8$)



Example of non-stationary process: Random Walk

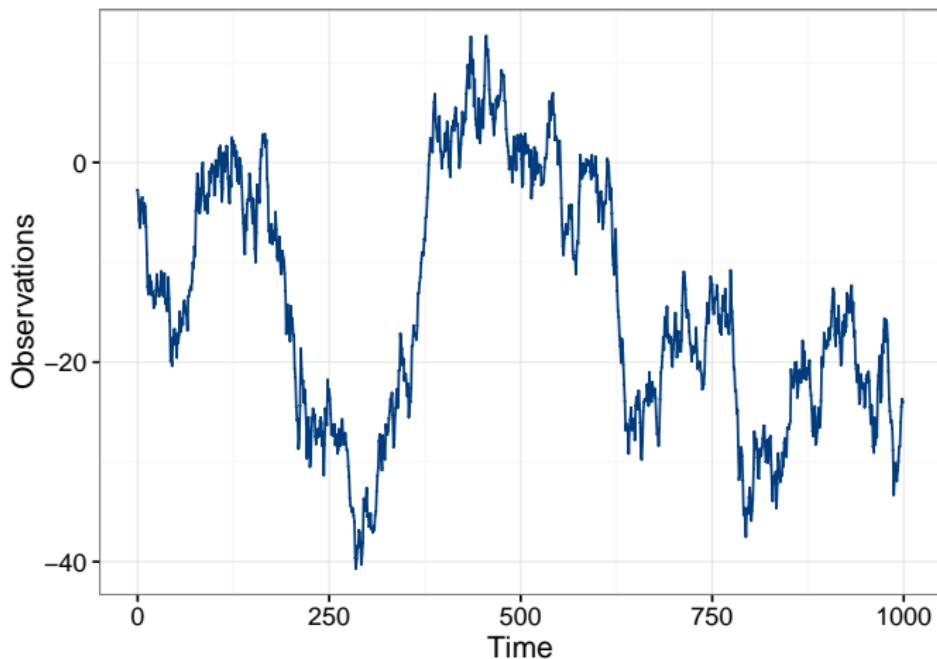


Figure: Simulation of a Random walk: $X_t = X_{t-1} + W_t$

Theoretical ACF/PACF for an ARMA(p, q) model

- ① Derivation the ACF/PACF of ARMA(p, q) is generally difficult.

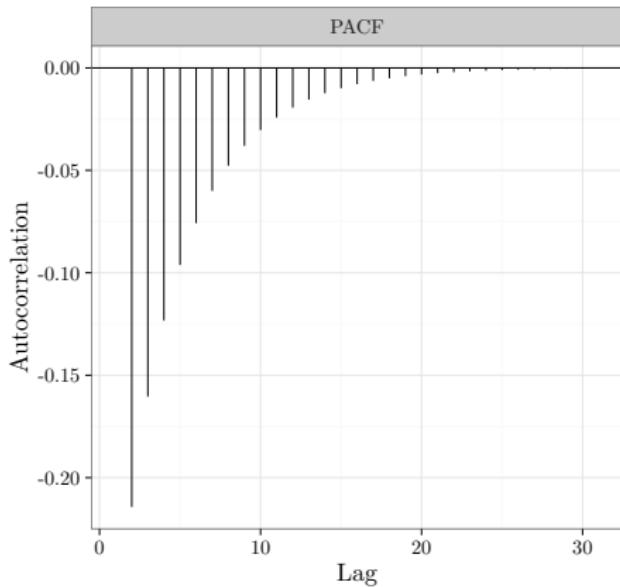
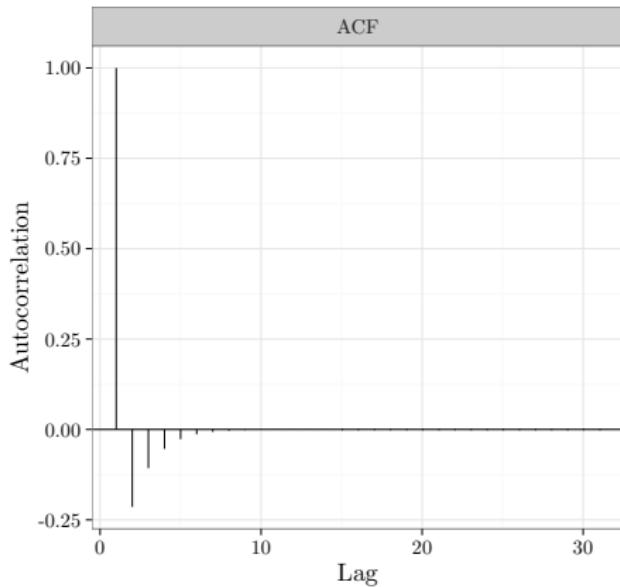
Example: Given an ARMA(1,1), we have $\rho(h) = \phi^{h-1} \rho(1)$ where

$$\rho(1) = \frac{(1 + \theta\phi)(\theta + \phi)}{1 + 2\phi\theta + \theta^2}.$$

- ② It is difficult to identify ARMA process from their ACF/PACF. Indeed, we have:
-

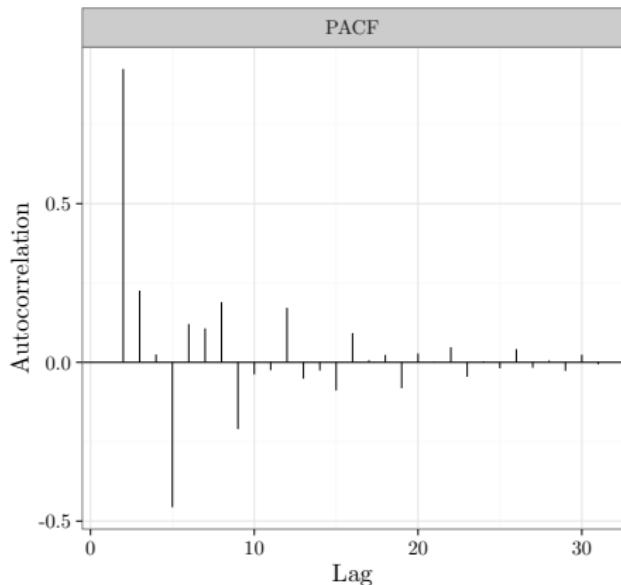
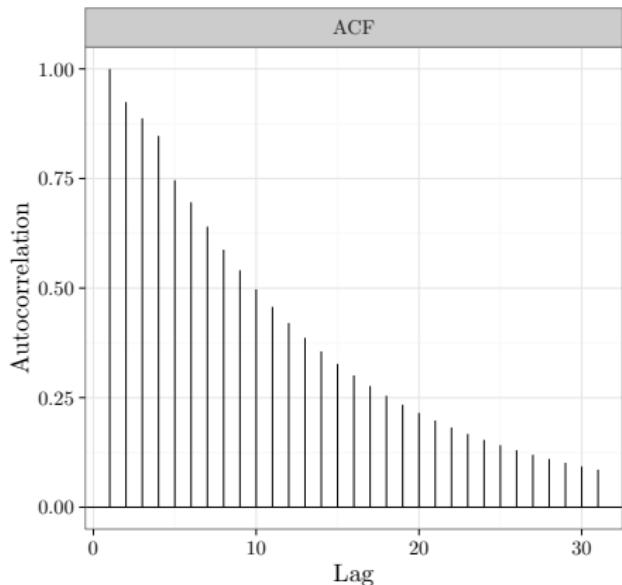
	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Order Identification of ARMA Models



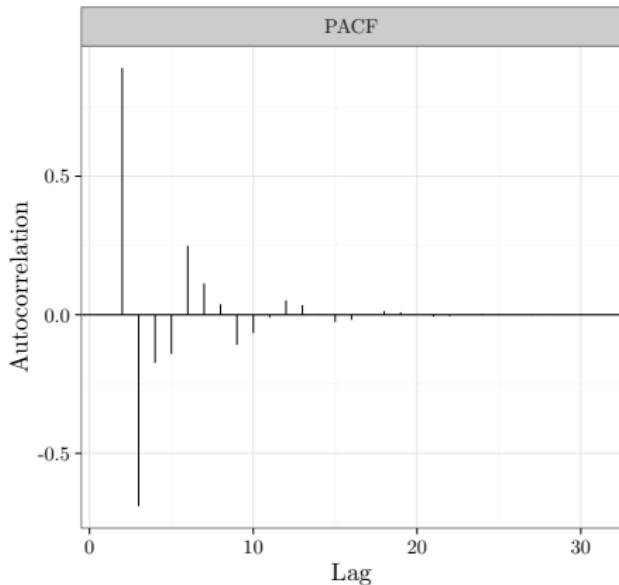
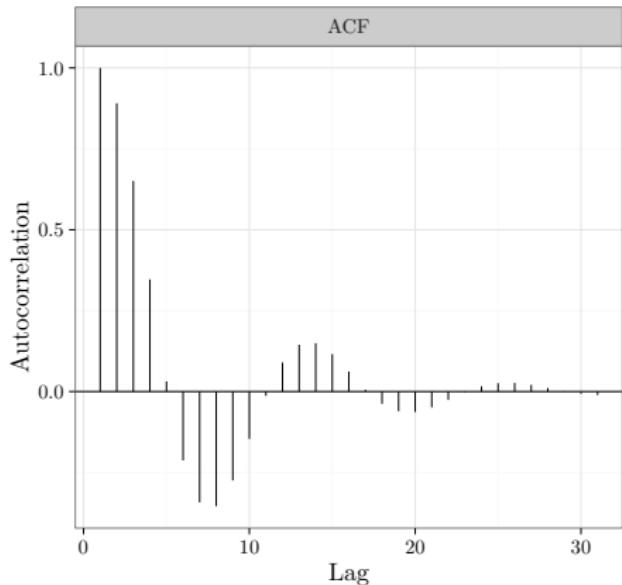
This an ARMA(1,1) process

Order Identification of ARMA Models



This an ARMA(3,4) process

Order Identification of ARMA Models



This an ARMA(2,3) process

Order Identification of ARMA Models

- ACF and PACF of ARMA models are difficult to interpret.
- It is generally easier to consider a list of candidate models and selected the “best” model in this list using a model selection criterion or an estimator of the prediction error (e.g. MAPE).
- Check diagnostic plot to assess if the model is reasonable for this time series.

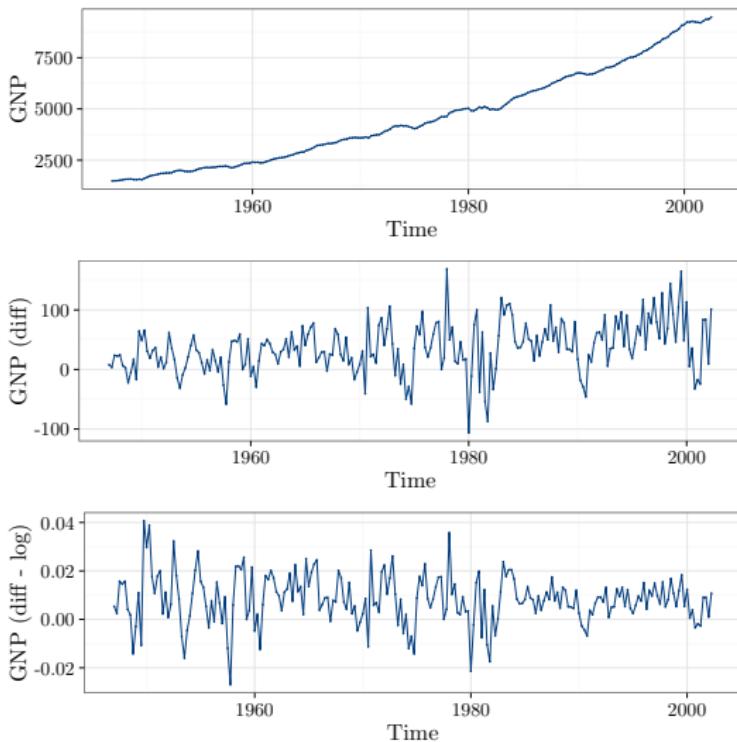
ARIMA Models

- Any time series data exhibits some non-stationary features but their first difference (i.e. $\nabla X_t = X_t - X_{t-1}$) can often be approximated by a stationarity.
- A simple example is: $X_t = X_{t-1} + W_t$ where $W_t \sim \mathcal{N}(0, \sigma^2)$ is non-stationary to ∇X_t is a white noise (which is stationary).
- Integrated ARMA (or **ARIMA**) models extend the class of ARMA models to include differencing. An ARIMA(p, d, q) is defined as

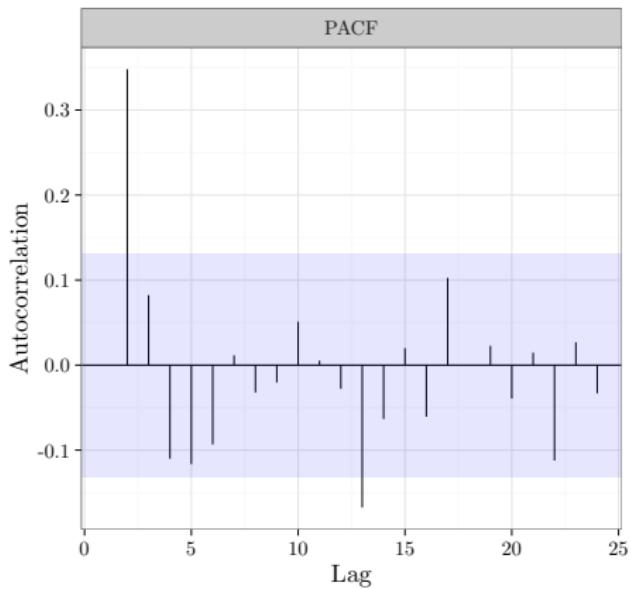
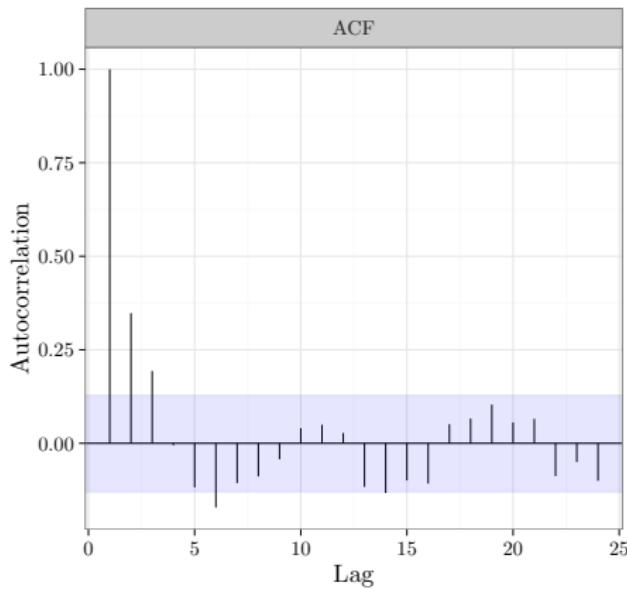
$$\phi(B)\nabla^d X_t = \theta(B)W_t$$

where $\phi(B)$ denotes the autoregressive operator, $\theta(B)$ the moving average operator, and $\nabla^d X_t = (1 - B)^d X_t$ represents the differencing operator.

Example: US GNP

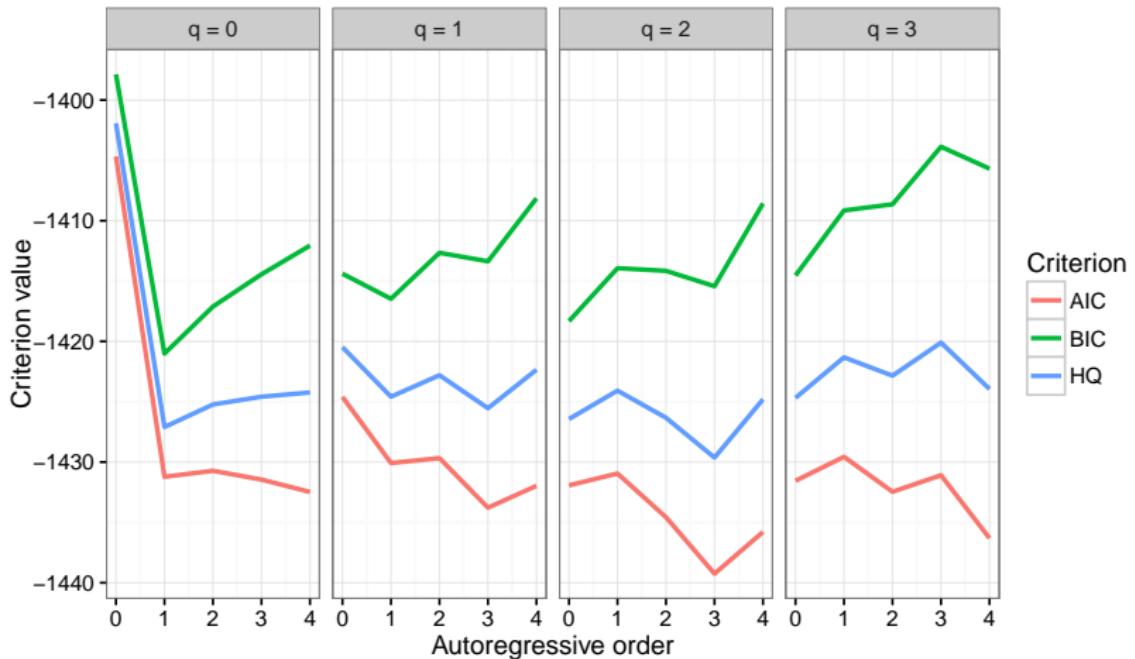


Example: US GNP

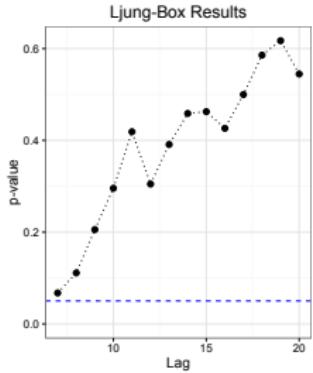
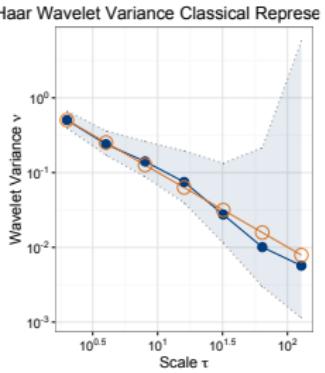
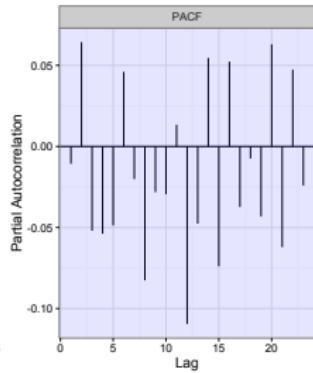
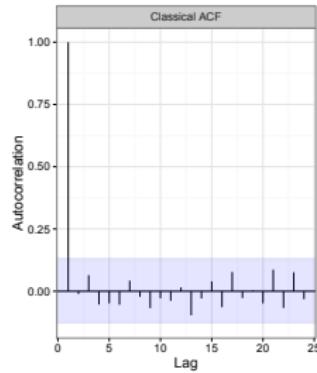
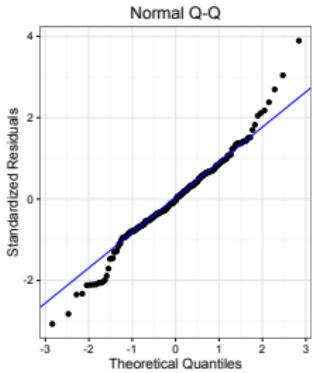
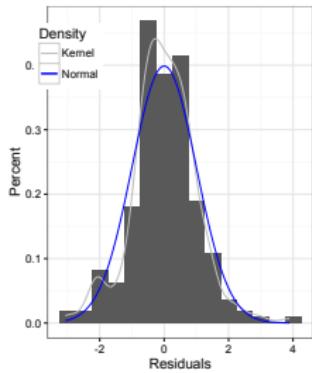
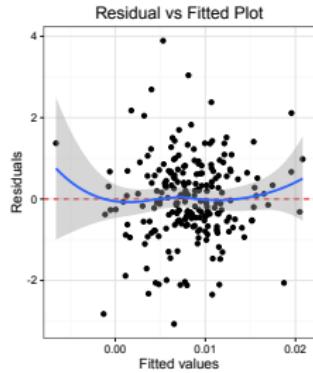
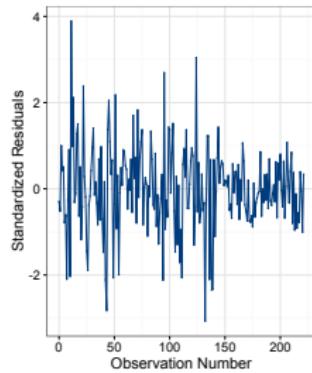


An AR(1) or MA(2) seems like potential candidates.

Example: US GNP



Example: US GNP



SARIMA Models

- In real time series data the dependence on the past tends to occur most strongly at **multiples of some underlying lag s** . For example, with quarterly economics data, there is often a strong dependence between the same quarter of different year (i.e. $s = 4$).
- Consider the following models as examples we could use for quarterly economics data:

$$\mathcal{M}_1 : X_t = \Phi X_{t-4} + W_t$$

$$\mathcal{M}_2 : X_t = \Phi X_{t-4} + W_t + \theta W_{t-1}.$$

SARIMA Models

The ACF/PACF of models \mathcal{M}_1 and \mathcal{M}_2 are given by:

$$\rho_1(h) = \begin{cases} 1 & \text{if } h = 0 \\ \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4 \\ 0 & \text{if } h \notin \mathbb{Z}_4 \cup \{0\} \end{cases}$$

$$\rho_2(h) = \begin{cases} 1 & \text{if } h = 0 \\ \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4 \\ \frac{\theta}{1+\theta^2} \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4^* \\ 0 & \text{if } h \notin \mathbb{Z}_4 \cup \mathbb{Z}_4^* \cup \{0\} \end{cases}$$

where $\mathbb{Z}_4 = \{4k : k \in \mathbb{Z} \setminus \{0\}\}$ and

$\mathbb{Z}_4^* = \{4k + 1 : k \in \mathbb{Z}\} \cup \{4k - 1 : k \in \mathbb{Z}\}$.

Let's derive these formulas...

ACF for \mathcal{M}_1

Assuming $|\Phi| < 1$, $k \in \mathbb{N}_+$ and $m : m/4 \notin \mathbb{N}_+$ we have

$$\begin{aligned}\text{var}(X_t) &= \text{var}(\Phi X_{t-4} + W_t) = \text{var}(W_t + \Phi W_{t-4} + \Phi^2 W_{t-8} + \dots) \\ &= \text{var}\left(\sum_{i=0}^{\infty} \Phi^i W_{t-4i}\right) = \sigma^2 \sum_{i=0}^{\infty} (\Phi^2)^i = \frac{\sigma^2}{1 - \Phi^2},\end{aligned}$$

$$\begin{aligned}\text{cov}(X_t, X_{t+4k}) &= \text{cov}(X_t, \Phi X_{t+4(k-1)} + W_{t+4k}) = \Phi \text{cov}(X_t, X_{t+4(k-1)}) \\ &= \Phi^k \text{var}(X_t) = \frac{\Phi^K \sigma^2}{1 - \Phi^2},\end{aligned}$$

$$\begin{aligned}\text{cov}(X_t, X_{t+m}) &= \text{cov}\left(\sum_{i=0}^{\infty} \Phi^i W_{t-4i}, \sum_{j=0}^{\infty} \Phi^j W_{t-4j+m}\right) \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \Phi^{i+j} \text{cov}(W_{t-4i}, W_{t-4j+m}) = 0.\end{aligned}$$

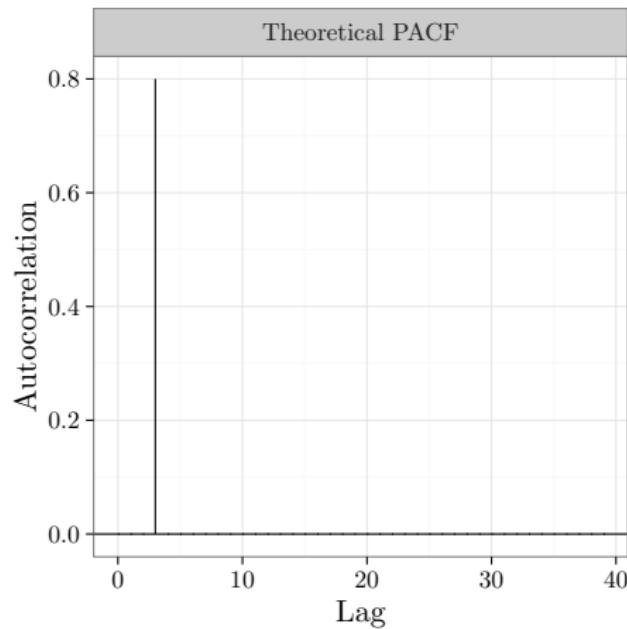
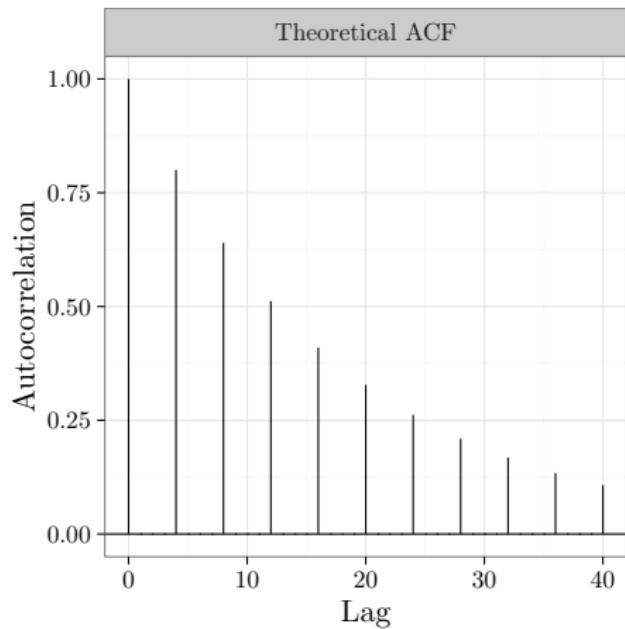
ACF for \mathcal{M}_1

Thus, the only nonzero correlation, aside from lag zero, is simply

$$\text{corr}(X_t, X_{t+4k}) = \Phi^k.$$

Since $\rho(h)$ is symmetric and $\rho(0) = 1$ we obtain

$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4 \\ 0 & \text{if } h \notin \mathbb{Z}_4 \cup \{0\}. \end{cases}$$

ACF/PACF for \mathcal{M}_1 with $\Phi = 0.8$ 

ACF for \mathcal{M}_2

For the second model, we have

$$\text{var}(X_t) = \text{var}(\Phi X_{t-4} + W_t + \theta W_{t-1}) = \Phi^2 \text{var}(X_{t-4}) + \sigma^2 (1 + \theta^2).$$

Since $\text{var}(X_t) = \text{var}(X_{t-4})$, we obtain

$$\text{var}(X_t) = \frac{\sigma^2 (1 + \theta^2)}{1 - \Phi^2}.$$

Next, we consider the following

$$X_t = \Phi X_{t-4} + W_t + \theta W_{t-1}$$

$$X_t X_{t-1} = \Phi X_{t-4} X_{t-1} + W_t X_{t-1} + \theta W_{t-1} X_{t-1}$$

$$\mathbb{E}[X_t X_{t-1}] = \Phi \mathbb{E}[X_{t-4} X_{t-1}] + \mathbb{E}[W_t X_{t-1}] + \mathbb{E}[\theta W_{t-1} X_{t-1}]$$

$$\gamma(1) = \Phi \gamma(3) + \theta \sigma^2.$$

ACF for \mathcal{M}_2

Moreover, we have

$$\begin{aligned}\gamma(4k) &= \text{cov}(X_t, X_{t+4k}) = \text{cov}(X_t, \Phi X_{t+4(k-1)} + W_{t+4k} + \theta W_{t+4k-1}) \\ &= \Phi \text{cov}(X_t, X_{t+4(k-1)}) = \Phi\gamma(4(k-1)).\end{aligned}$$

This implies that

$$\gamma(1) = \Phi\gamma(3) + \theta\sigma^2 = \Phi^2\gamma(-1) + \theta\sigma^2 = \Phi^2\gamma(1) + \theta\sigma^2$$

and therefore,

$$\gamma(1) = \frac{\theta\sigma^2}{1 - \Phi^2}.$$

Finally,

$$\begin{aligned}\gamma(4k+1) &= \Phi^k\gamma(1) = \Phi^k \frac{\theta\sigma^2}{1 - \Phi^2} \\ \gamma(4k-1) &= \Phi^k\gamma(-1) = \Phi^k\gamma(1) = \gamma(4k+1).\end{aligned}$$

ACF for \mathcal{M}_2

To sum up, we showed:

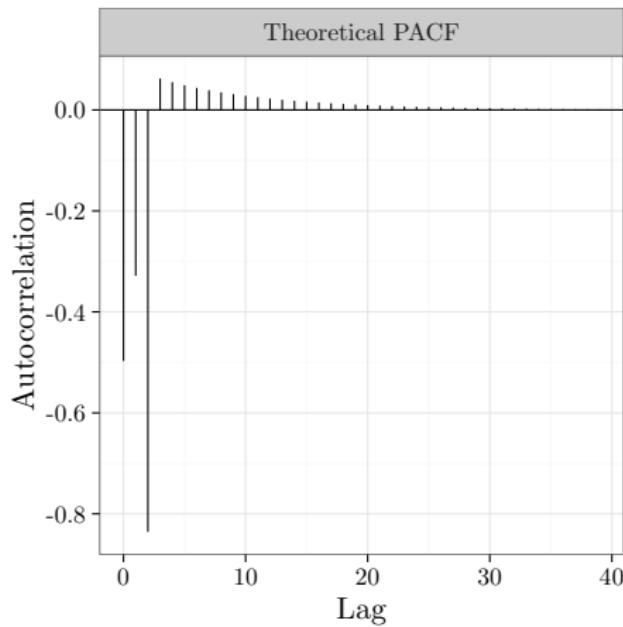
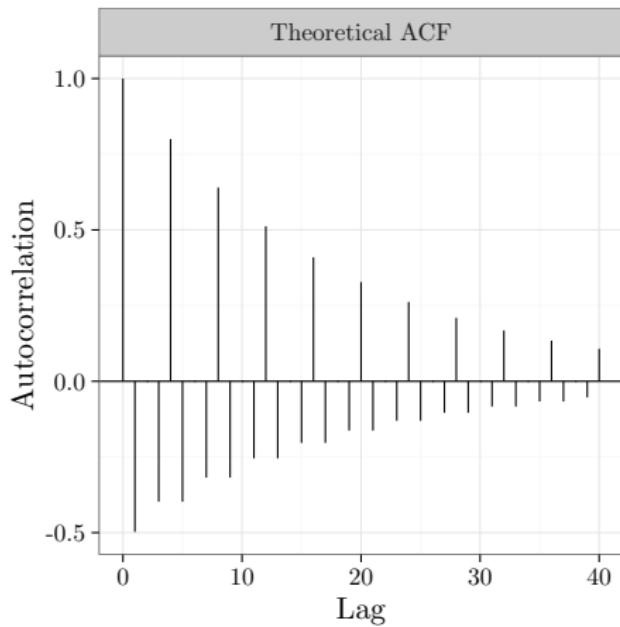
$$\gamma(0) = \frac{\sigma^2 (1 + \theta^2)}{1 - \Phi^2}$$

$$\gamma(4k) = \Phi^k \gamma(0)$$

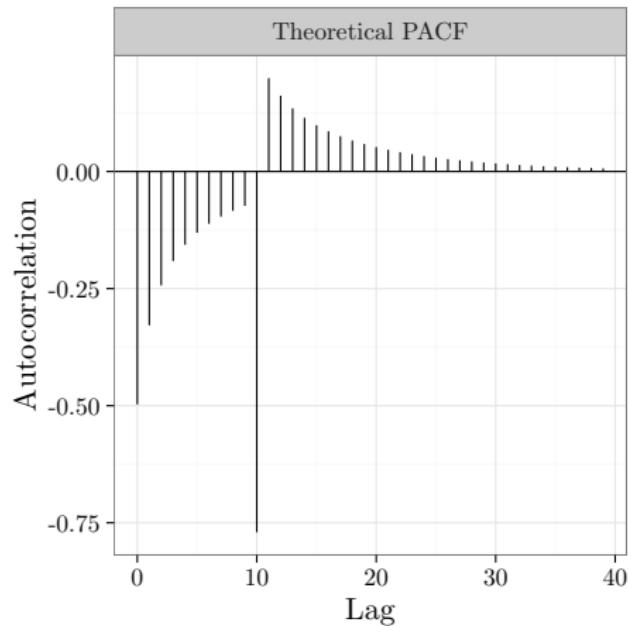
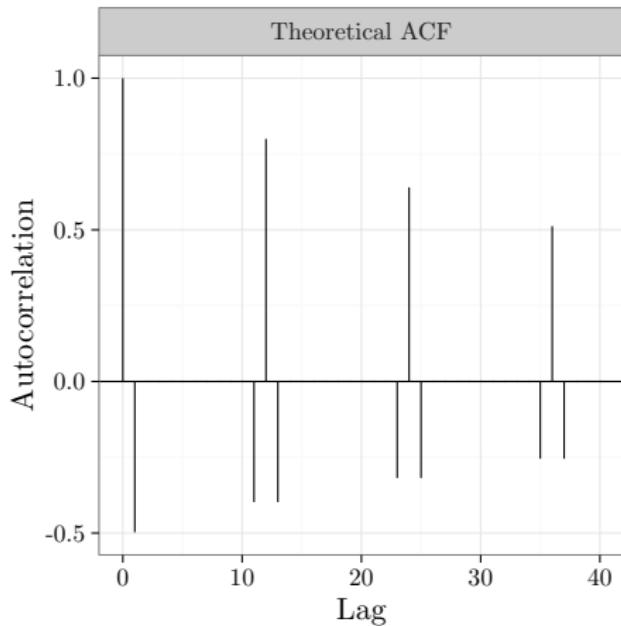
$$\gamma(4k+1) = \gamma(4k-1) = \Phi^k \frac{\theta \sigma^2}{1 - \Phi^2}.$$

By combining these results, we verify that:

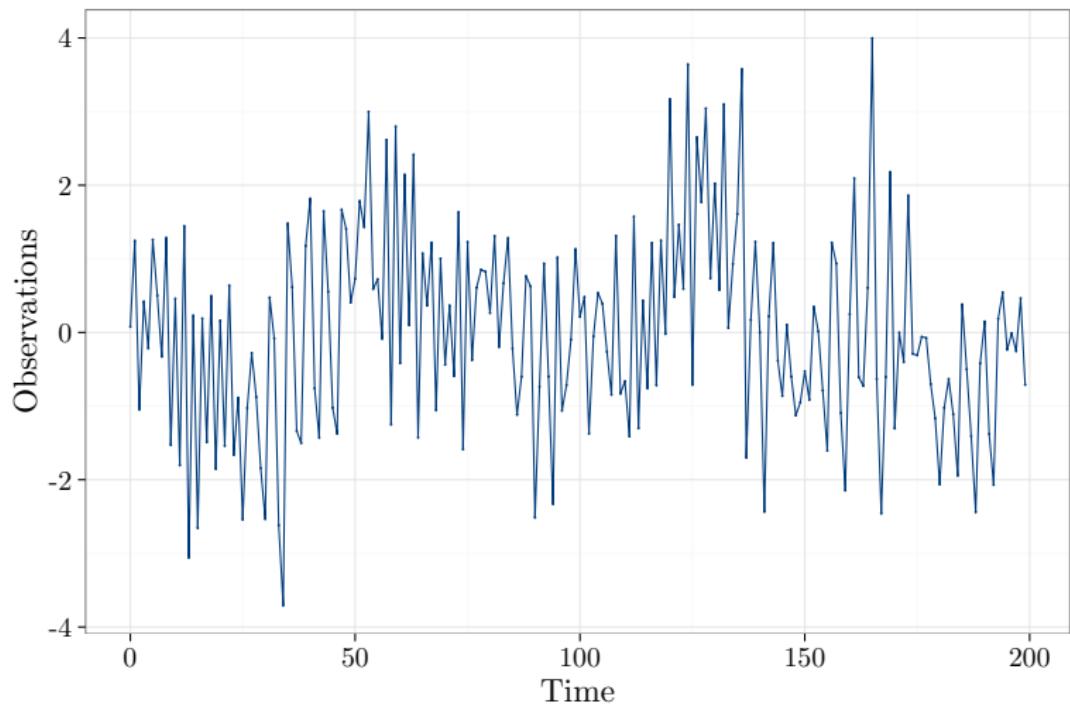
$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4 \\ \frac{\theta}{1+\theta^2} \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4^* \\ 0 & \text{if } h \notin \mathbb{Z}_4 \cup \mathbb{Z}_4^* \cup \{0\}. \end{cases}$$

ACF/PACF for \mathcal{M}_2 with $\Phi = 0.8$ and $\theta = -0.9$ 

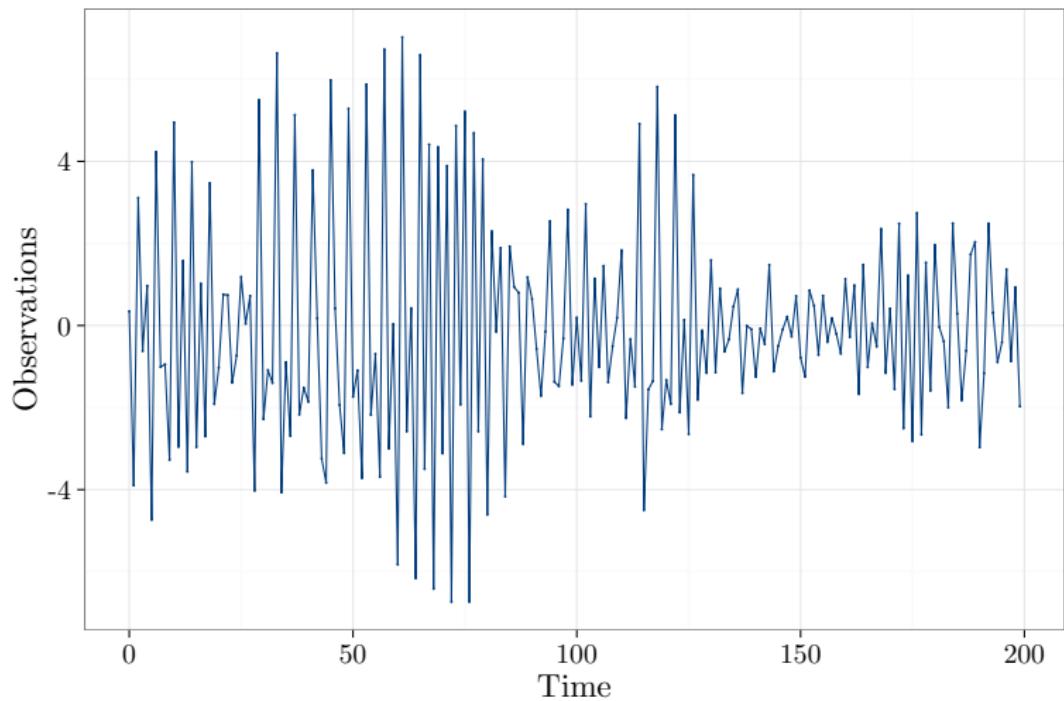
$$\text{ACF/PACF for } X_t = 0.8X_{t-12} - 0.9W_{t-1} + W_t$$



Simulated process: \mathcal{M}_1 with $\Phi = 0.8$



Simulated process: \mathcal{M}_1 with $\Phi = 0.8$ and $\theta = -0.9$



SARIMA Models

Definition:

A *Seasonal Autoregressive Integrated Moving Average model of order $(p, q, d) \times (P, Q, D)_s$* is in the form of:

$$\Phi(B)\phi(B)\nabla_s^D\nabla^d X_t = \delta + \Theta(B^s)\theta(B)W_t$$

with w_t the usual Gaussian white noise process and where the ordinary autoregressive and moving average process are represented by the polynomials $\phi(B)$ and $\theta(B)$ of order p and q , the seasonal part component by $\Phi(B)$ and $\Theta(B^s)$ of order P and Q and the seasonal difference component by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.

Definition:

The operators:

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps}$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_Q B^{Qs},$$

are the seasonal autoregressive and moving average operators.

Some Examples

SARIMA Models

$$\mathcal{M}_1 : X_t = \Phi X_{t-4} + W_t$$

$$\mathcal{M}_2 : X_t = \Phi X_{t-4} + W_t + \theta W_{t-1}$$

$$\mathcal{M}_3 : X_t = \Phi_1 X_{t-12} + \Phi_2 X_{t-24} + W_t + \theta W_{t-1} + \Theta W_{t-12}$$

$$\mathcal{M}_4 : X_t = W_t + \theta W_{t-1} + \Theta_1 W_{t-6} + \Theta_2 W_{t-12}$$

$$\mathcal{M}_5 : X_t = \Phi X_{t-4} + W_t + \Theta W_{t-6}$$

Identified Models

Model 1 is a SARIMA(0,0,0)×(1,0,0)₄

Model 2 is a SARIMA(0,0,1)×(1,0,0)₄

Model 3 is a SARIMA(0,0,1)×(2,0,1)₁₂

Model 4 is a SARIMA(0,0,1)×(0,0,2)₆

Model 5 is a SARIMA(4,0,0)×(0,0,1)₆ or a SARIMA(0,0,6)×(1,0,0)₄

Identifying SARIMA coefficients

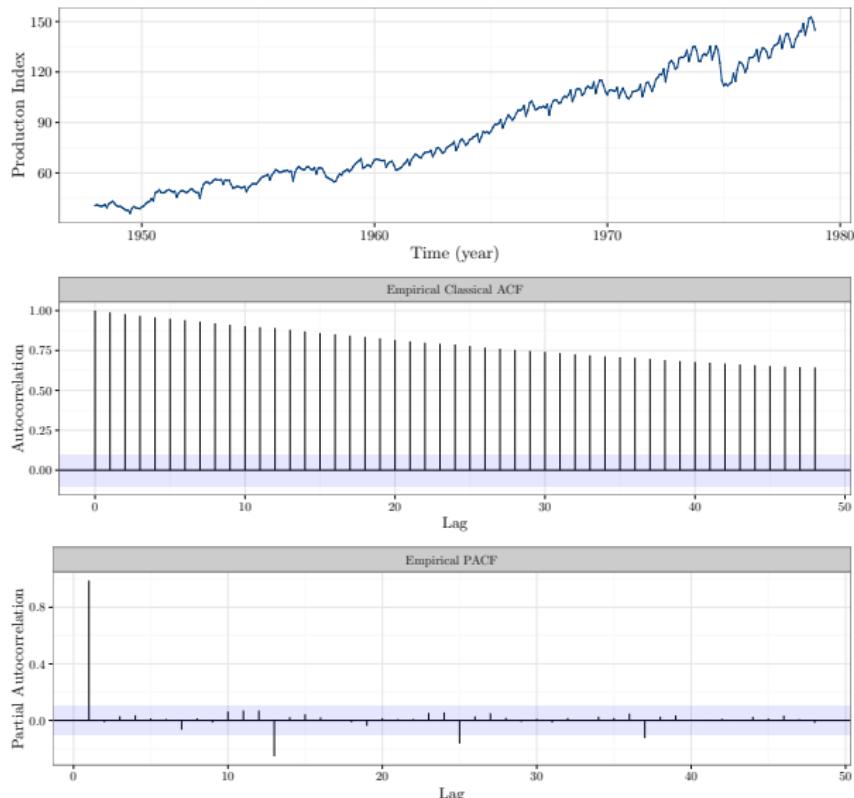
Identifying the order of a SARIMA model is a daunting task and there is “good” way to do it. Here is a possible method:

- Apply difference operators until a roughly stationary series is produced.
- Plot ACF/PACF of the “differenced” series, then:
 - Identify the seasonal period s ,
 - Identify P and Q using the following table (“seasonal lags” only):

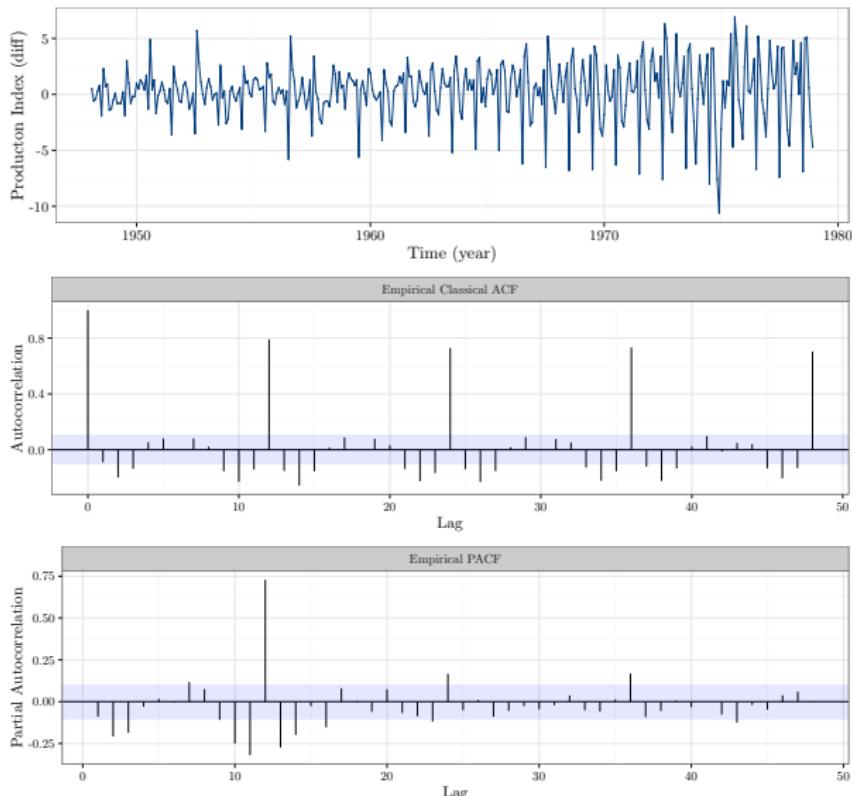
SAR(P)	SMA(Q)	SARMA(P,Q)
ACF Tails off	Cuts off after lag Q	Tails off
PACF Cuts off after lag P	Tails off	Tails off

- Identify p and q looking only at “nonseasonal” lags.
- If multiple models were identified in the previous step use AIC or BIC to select the “best” model.
- Evaluate whether the selected model is satisfactory using diagnostic plots.

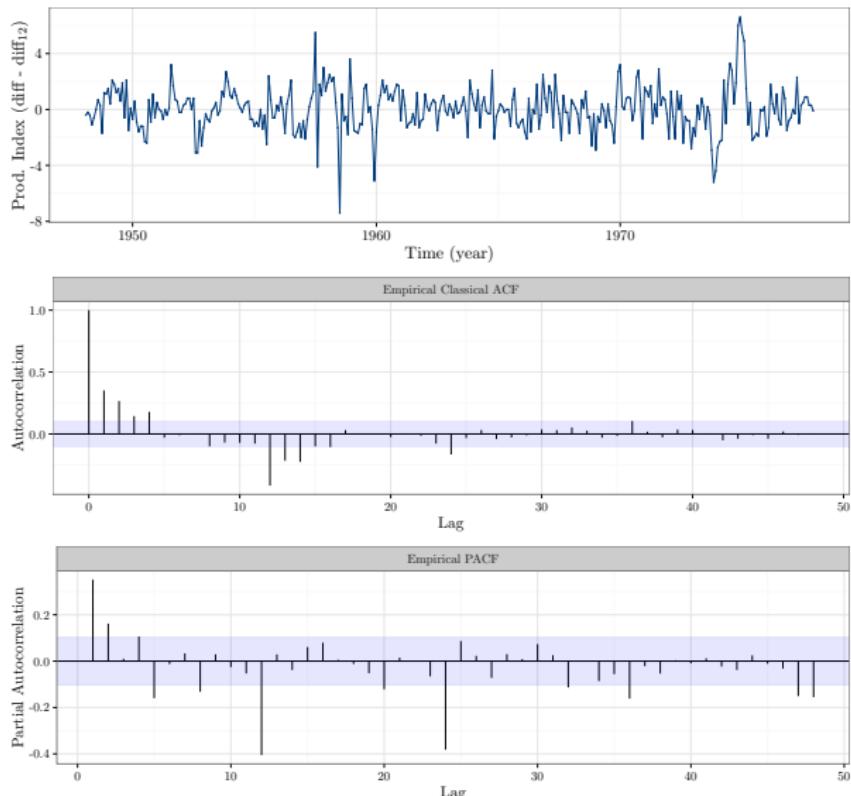
Example: Monthly Production Index Dataset



Example: Monthly Production Index Dataset



Example: Monthly Production Index Dataset



Example: Monthly Production Index Dataset

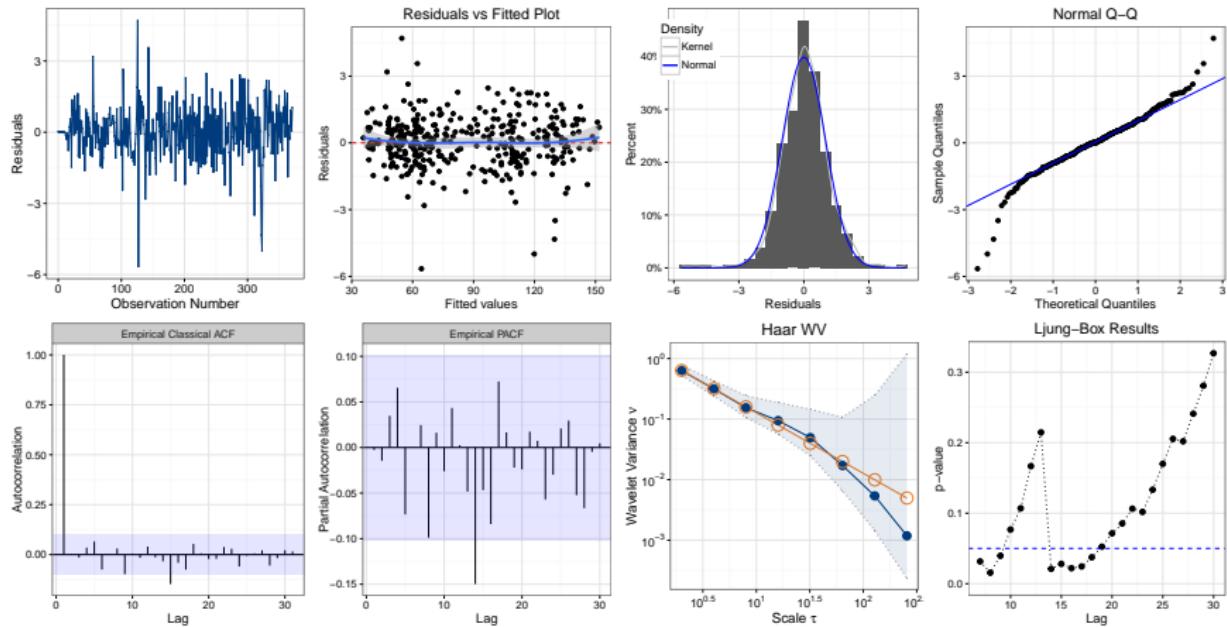
Using the approach described earlier:

- It seems that $d = 1$, $D = 1$ are appropriate.
- From ACF/PACF of the differenced series, we propose four models:
 - $s = 12$
 - $P = 0$ and Q is between 1 and 4.
 - $p = 2$ and $q = 0$
- Select “best” model: let $\mathcal{M}_i : \text{SARIMA}(2,1,0) \times (0,1,i)_{12}$,
 $i = 1, 2, 3, 4$. Then, we have

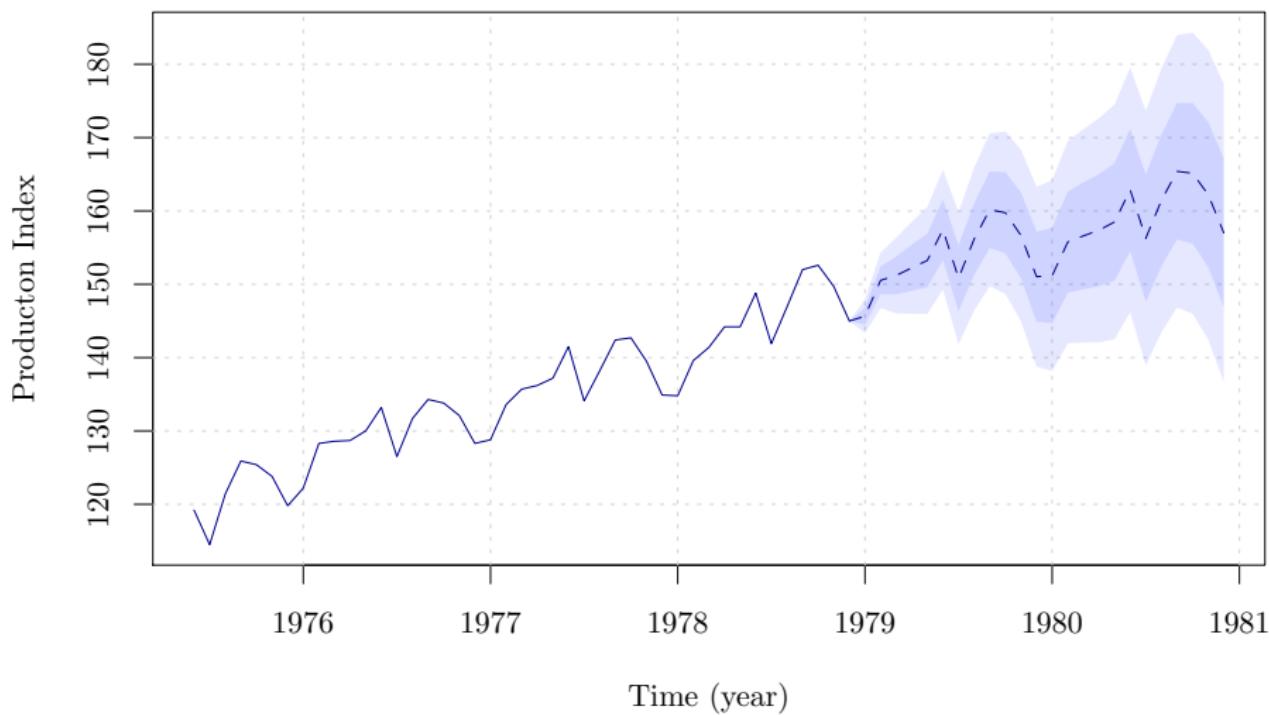
	\mathcal{M}_1	\mathcal{M}_2	\mathcal{M}_3	\mathcal{M}_4
AIC	1162.334	1163.712	1139.965	1141.898

- Diagnostic plot for \mathcal{M}_3 .

Example: Monthly Production Index Dataset



Example: Monthly Production Index Dataset



Appendix A: Examples on Strong and Weak Stationarity

Example 1 (Random walk): Consider a Gaussian random walk process X_t , as defined in 2.6 where initial value $X_0 = 0$. Since the model is “fixed in time” this process is clearly strongly stationary but not weakly stationary since

$$\text{var}(X_t) = \text{var}\left(\sum_{i=1}^t Z_i\right) = \sum_{i=1}^t \gamma^2 = t\gamma^2$$

and therefore $\mathbb{E}[X_t^2]$ does not exist.

Example 2 (mixtures): Let $X_t \stackrel{iid}{\sim} \exp(1)$ (i.e. exponential distribution with $\lambda = 1$) and $Y_t \stackrel{iid}{\sim} \mathcal{N}(1, 1)$. Then, let

$$Z_t = \begin{cases} X_t & \text{if } t \in \{2k | k \in \mathbb{N}\} \\ Y_t & \text{if } t \in \{2k + 1 | k \in \mathbb{N}\}. \end{cases}$$

Then, Z_t is weakly stationary but not strongly stationary.

[► Return to Remark 11](#)

Appendix B: Example: Weak Stationarity of an AR1

Consider an AR1 process (see Definition 2.7), defined as:

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \stackrel{iid}{\sim} \mathcal{N}(0, \nu^2),$$

with $|\phi| < 1$ and $\nu^2 < \infty$. Then, we have

$$X_t = \phi X_{t-1} + Z_t = \phi [\phi X_{t-2} + Z_{t-1}] + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t$$

$$\vdots$$

$$= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j}.$$

By taking the limit in k (which is perfectly valid as we assume $t \in \mathbb{Z}$) and assuming $|\phi| < 1$, we obtain

$$X_t = \lim_{k \rightarrow \infty} X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}, \tag{3}$$

since $|\phi| < 1$. This shows that this process is linear (see Definition 2.17).

Appendix B: Example: Weak Stationarity of an AR1

Therefore, we have

$$\mathbb{E}[X_t] = \sum_{j=0}^{\infty} \phi^j \mathbb{E}[Z_{t-j}] = 0$$

$$\text{var}(X_t) = \text{var}\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j}\right) = \sum_{j=0}^{\infty} \phi^{2j} \text{var}(Z_{t-j}) = \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} = \frac{\nu^2}{1 - \phi^2}.$$

Moreover, we obtain (assuming for notational simplicity that $|h| > 1$)

$$\text{cov}(X_t, X_{t+h}) = \phi \text{cov}(X_t, X_{t+h-1}) = \phi^2 \text{cov}(X_t, X_{t+h-2}) = \phi^h \text{cov}(X_t, X_t).$$

When $h \in \mathbb{Z}$ we obtain

$$\text{cov}(X_t, X_{t+h}) = \phi^{|h|} \text{cov}(X_t, X_t) = \phi^{|h|} \frac{\nu^2}{1 - \phi^2},$$

thus verifying the weak stationarity of the process.

[► Return to Remark 12](#)

Appendix C: Derivation of Equation (2)

The derivation of (2) allows us to introduce a common technique used in time series analysis. Let $\mathbf{1}$ denotes a unit vector of dimension T and $\mathbf{X} = [X_1, \dots, X_T]^T$, then we notice that \bar{X} can be expressed as follows

$$\bar{X} = \frac{1}{T} \mathbf{1}^T \mathbf{X}.$$

Moreover, we remember that if $\mathbf{Y} \in \mathbb{R}^k$ and a random variable and $\mathbf{A} \in \mathbb{R}^{h \times k}$ a fixed matrix we have

$$\text{var}(\mathbf{AY}) = \mathbf{A} \text{var}(\mathbf{Y}) \mathbf{A}^T.$$

Therefore, we have

$$\begin{aligned} \text{var}(\bar{X}) &= \frac{1}{T^2} \text{var}(\mathbf{1}^T \mathbf{X}) = \frac{1}{T^2} \mathbf{1}^T \text{var}(\mathbf{X}) \mathbf{1} \\ &= \frac{1}{T^2} \mathbf{1}^T \begin{bmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & & \gamma(T-2) \\ \vdots & & \ddots & \vdots \\ \gamma(T-1) & \dots & \dots & \gamma(0) \end{bmatrix} \mathbf{1}. \end{aligned}$$

Appendix C: Derivation of Equation (2)

By looking at the matrix $\text{var}(\mathbf{X})$ one can notice that it contains T times the term $\gamma(0)$, $2(T - 1)\gamma(1)$, $2(T - 2)\gamma(2)$ and so on. Therefore, we have

$$\text{var}(\bar{X}) = \frac{1}{T^2} (T\gamma(0) + 2(T - 1)\gamma(1) + 2(T - 2)\gamma(2) + \dots + \gamma(T - 1)).$$

Since $\gamma(h)$ is symmetric we also have

$$\text{var}(\bar{X}) = \frac{1}{T^2} \sum_{i=T}^T (T - |i|) \gamma(i) = \frac{\gamma(0)}{T} \sum_{i=T}^T \left(1 - \frac{|i|}{T}\right) \rho(i).$$

It is worth noting that in the iid case we have that $\text{var}(\bar{X}) = \frac{1}{T} \text{var}(X_1)$. This result can naturally also be obtained using (2) since $\text{var}(X_1) = \gamma(0)$ and $\sum_{i=T}^T \left(1 - \frac{|i|}{T}\right) \rho(i) = 1$.

► [Return to Equation \(2\)](#)

Appendix D: How to compute $\text{var}(\bar{X})$ in practice?

As in the previous example, let us consider a stationary AR1 process, i.e.

$$X_t = \phi X_{t-1} + Z_t, \quad \text{where } |\phi| < 1 \quad \text{and} \quad Z_t \stackrel{iid}{\sim} \mathcal{N}(0, \nu^2)$$

We already showed in Appendix B that $\gamma(h) = \phi^h \sigma^2 (1 - \phi^2)^{-1}$, therefore, we obtain (after some computations):

$$\text{var}(\bar{X}) = \frac{\nu^2 (T - 2\phi - T\phi^2 + 2\phi^{T+1})}{T^2 (1 - \phi^2) (1 - \phi)^2}. \quad (4)$$

Unfortunately, deriving such an exact formula is often difficult when considering more complex models. However, asymptotic approximations are often employed to simplify the calculation. For example, in our case we have

$$\lim_{T \rightarrow \infty} T \text{var}(\bar{X}) = \frac{\nu^2}{(1 - \phi)^2},$$

providing the following approximate formula:

$$\text{var}(\bar{X}) \approx \frac{\nu^2}{T (1 - \phi)^2}.$$

Appendix D: How to compute $\text{var}(\bar{X})$ in practice?

Alternatively, simulation methods can also be employed. For example, one could compute $\text{var}(\bar{X})$ as follows:

Step 1: Simulate under the assumed model, i.e. $X_t^* \sim F_{\theta_0}$, where F_{θ_0} denotes the true model (in this case an AR1 process).

Step 2: Compute \bar{X}^* (i.e. average based on (X_t^*)).

Step 3: Repeat Steps 1 and 2 B times.

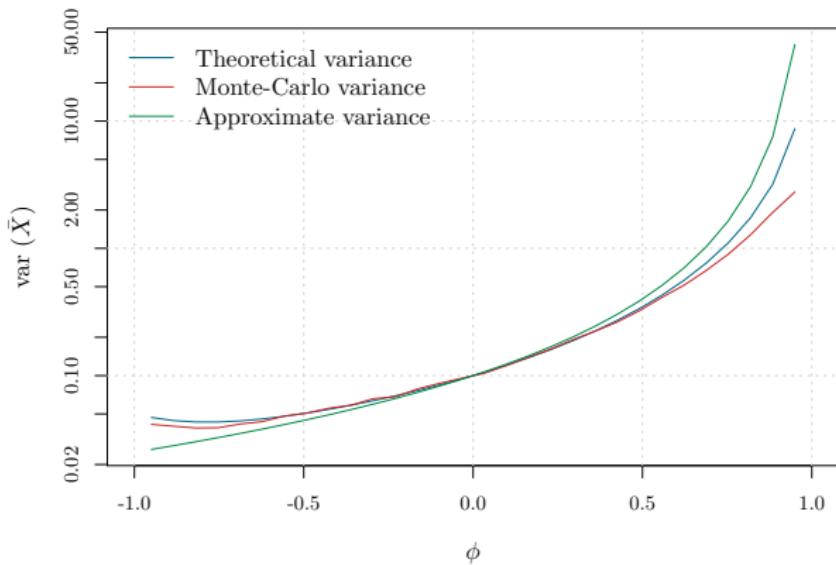
Step 4: Compute the empirical variance \bar{X}^* (based on B independent replications).

The above procedure is known as Monte-carlo method (it is actually a Monte-carlo integral) and is closely related to the concept of parametric bootstrap (see Efron and Tibshirani 1994) which is a very popular tool in statistics.

Appendix D: How to compute $\text{var}(\bar{X})$ in practice?

A numerical experiment

We consider $T = 10$, $B = 5000$ and grid of values for ϕ from -0.95 to 0.95.



► Return to Equation (2)