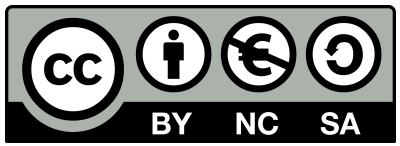
Chapter 2: Introduction to Time Series Analysis

Jan Skaloud* & Stéphane Guerrier†

*École Polytechnique Fédérale de Lausanne; †University of Geneva

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

Material available online: https://gmwm.netlify.com



EPFL - Winter 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 $(X_t)_{t=1,...,T}$ (i.e. X_1, X_2, ..., X_T). Note that the time t is not continuous but belongs to the discrete index set. Therefore, we implicitly assume that:

- t is not random (e.g. the time at which each observation is measured is known)
- 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 briefly review it in the next slides.

Introduction

Definition 2.2 (Descriptive Analysis).

A typical time series analysis starts displaying the data as a line plot on a graph. Time is on the x-axis and the variable on the 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 are more useful from small to moderate size data.

Descriptive Analysis

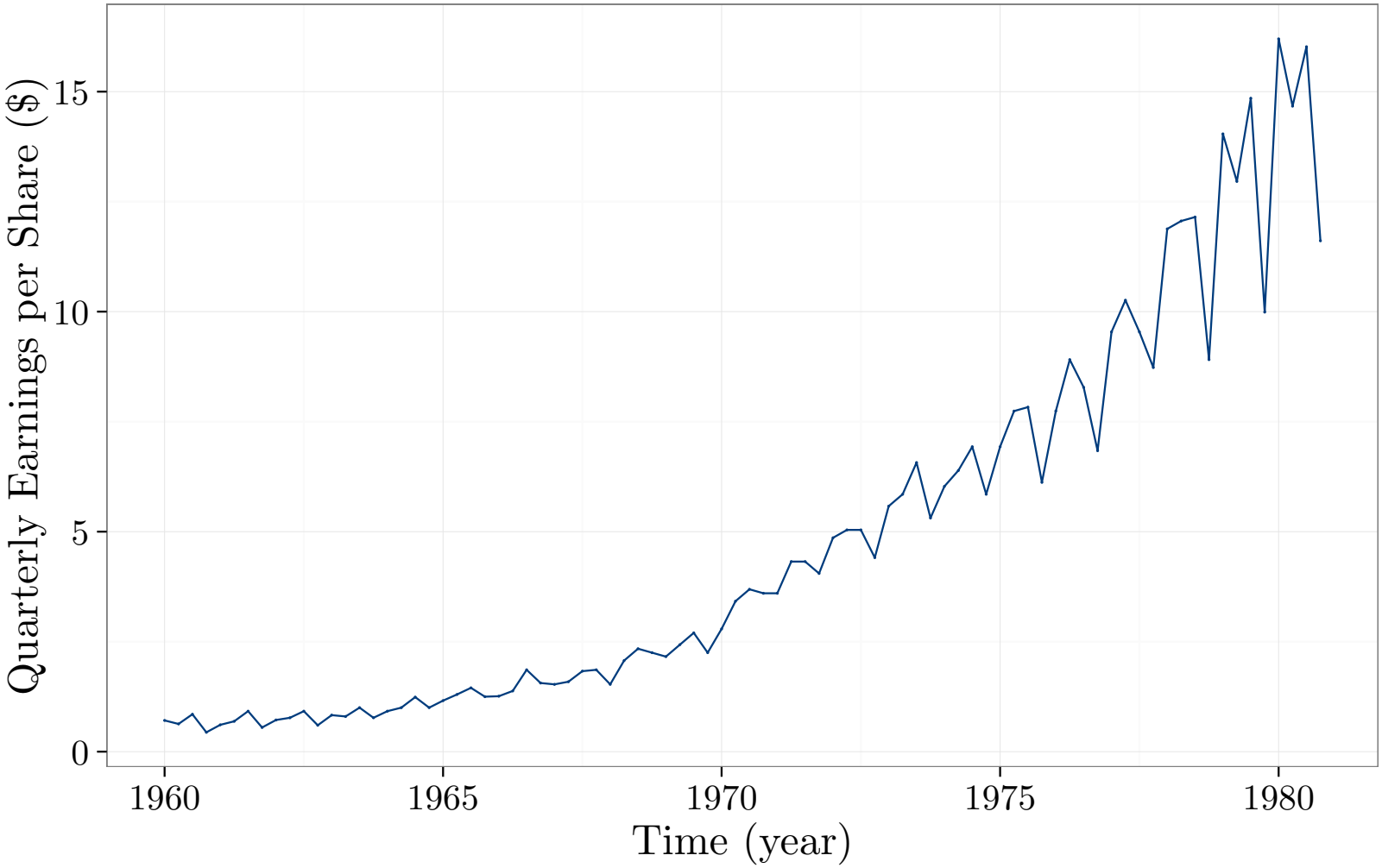
Question:

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

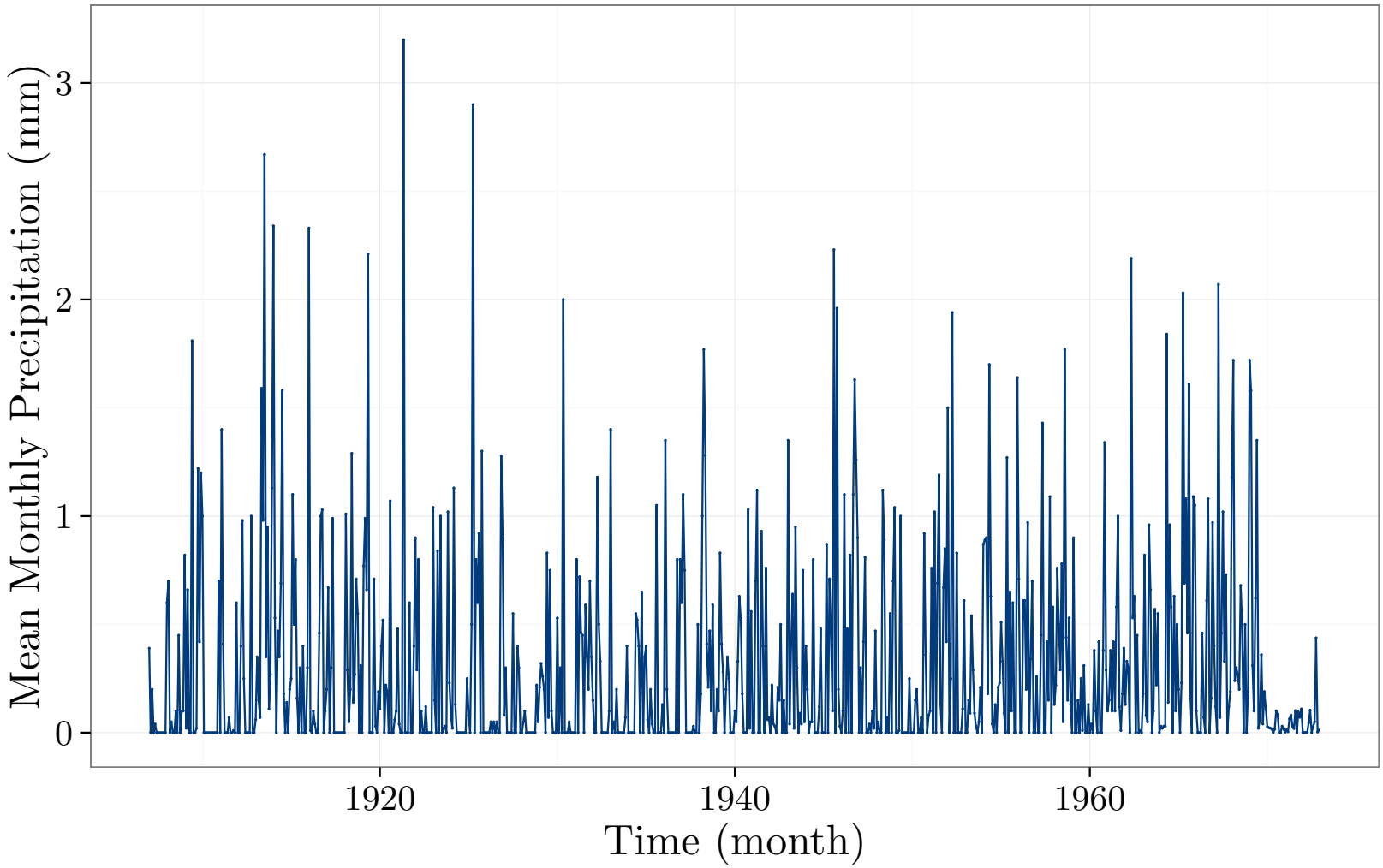
A possible answer:

- 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)

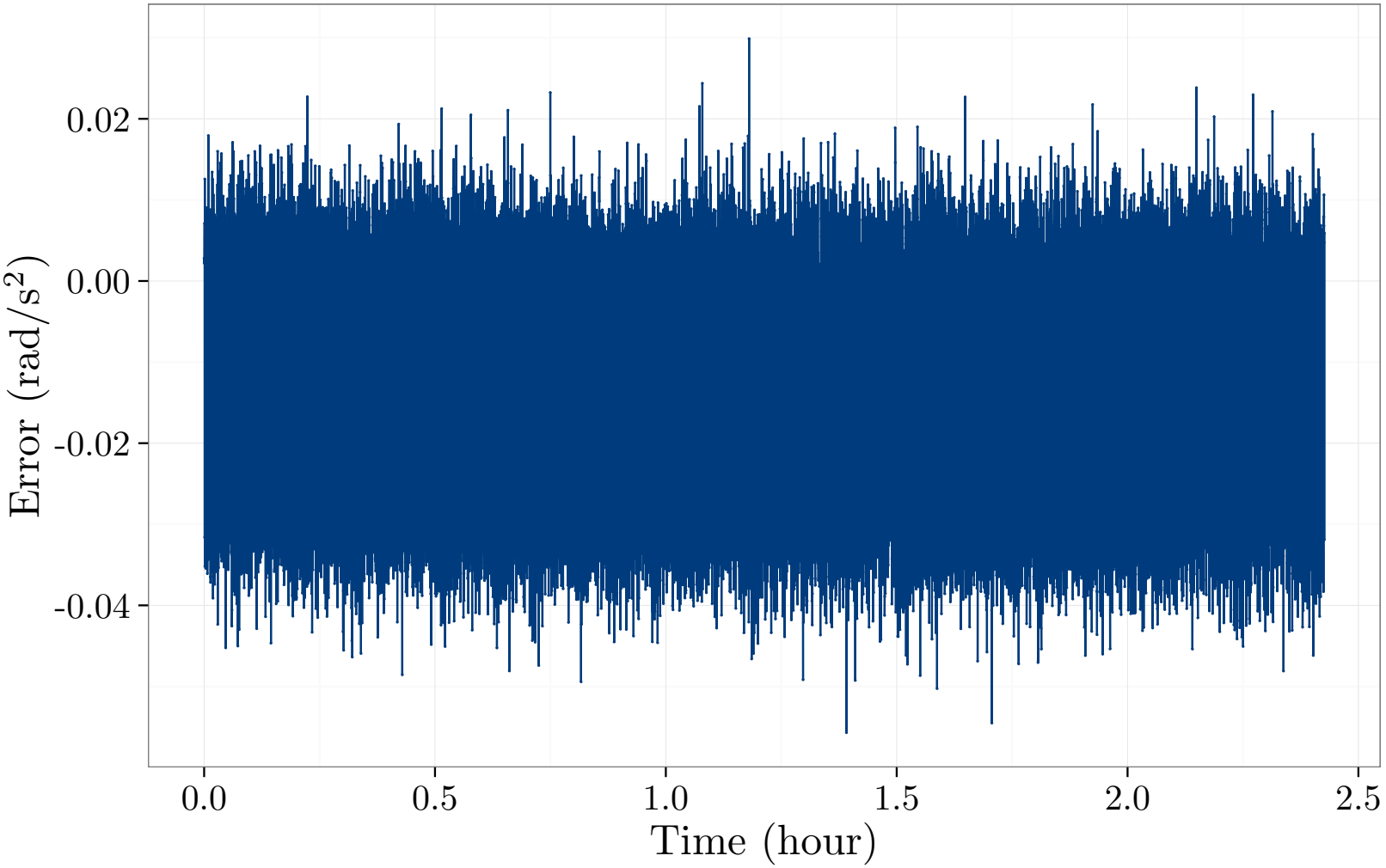
Example: Johnson and Johnson Quarterly Earnings



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:

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

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

Definition 2.6 (Random walk).

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

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

Definition 2.7 (Auto-Regressive).

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

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

Stochastic Processes Considered

Definition 2.8 (Gauss Markov).

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

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

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

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 by:

$$
\begin{aligned}
Y_t &= \phi_1 Y_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, v_1^2) \\
W_t &= \phi_2 W_{t-1} + U_t, \quad U_t \overset{iid}{\sim} \mathcal{N}(0, v_2^2) \\
Q_t &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\
X_t &= Y_t + W_t + Q_t,
\end{aligned}
$$

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

Main purpose of TS analysis: Forecasting

- **Forecasting** is one of the main purposes 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 related 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 $X_1, X_2, \ldots, X_T$ is said to be **iid** if and only if

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

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

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

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.

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

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 jointly 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}.

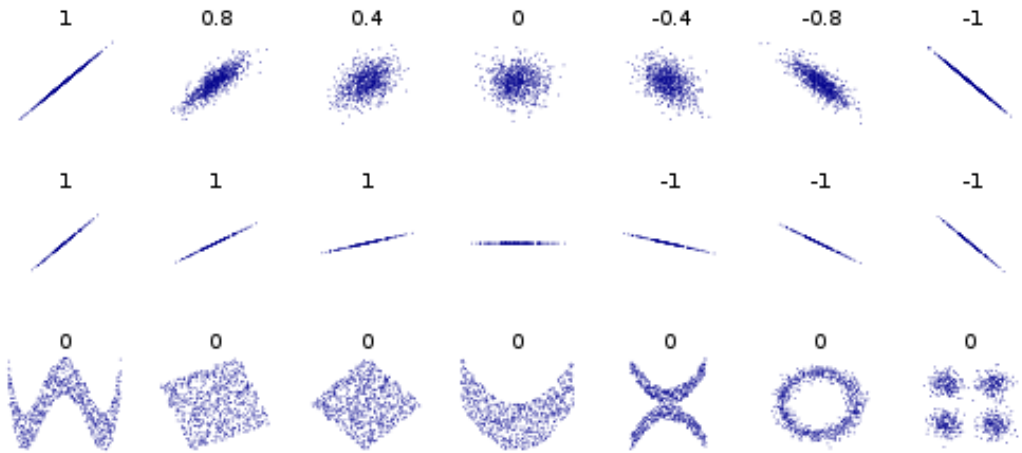
Measuring (linear) dependence

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).

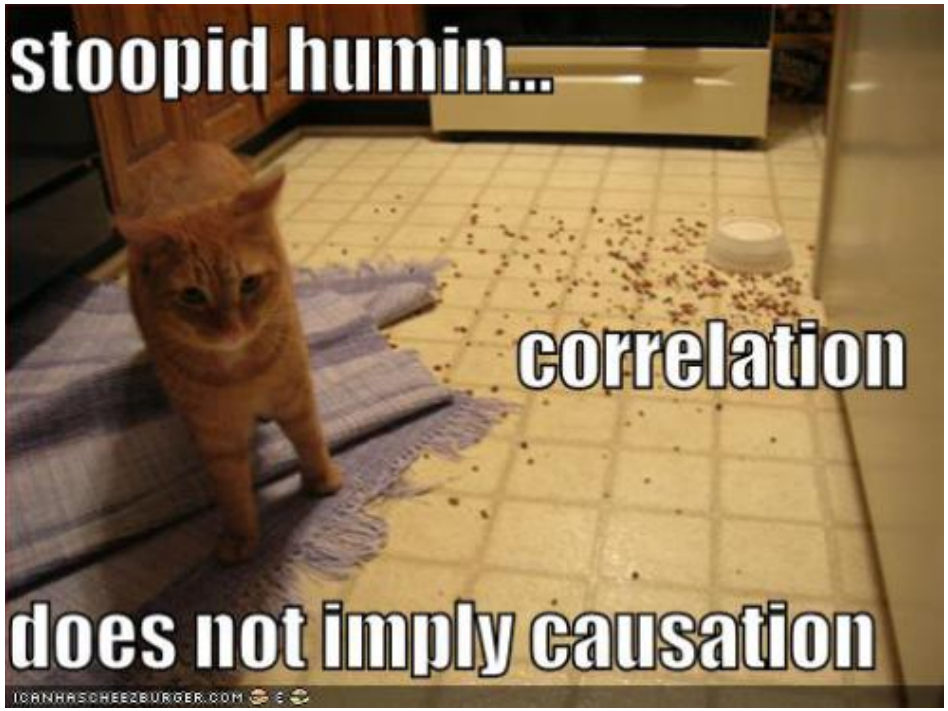
Covariance and correlation measure linear dependence. They are less helpful to measure monotonic dependence and they are much less helpful to measure nonlinear dependence. Nonlinear measures of dependence exist but we will not discuss this subject in this class. Here is an example:



Correlation == Causation?

Remark 8 (Causation).

Correlation *does NOT* imply causation. For example, $\rho(t, t+h) \neq 0$ does not imply that $X_t \to X_{t+h}$ 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 in the context of time series

Motivation:

Consider the simple (but strange!) model:

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

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 a **stationary process** such as:

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

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:

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.

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

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

Definition 2.16 (Weak Stationarity).

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

$$\begin{aligned}
\mathbb{E}[X_t] &= \mu < \infty, \\
\mathbb{E}\left[X_t^2\right] &= \mu_2 < \infty, \\
\operatorname{cov}(X_t, X_{t+h}) &= \operatorname{cov}(X_{t+k}, X_{t+h+k}) = \gamma(h).
\end{aligned}$$

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. Indeed the concept of averaging is meaningless unless properties like mean and covariance are either fixed or "evolve" in a known manner.

Remark 9 (Implication on the ACV and ACF).

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

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

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 derive directly from 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.

$$\mathbb{E}^2[XY] \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]. \tag{2.1}$$

Using (2.1), we have

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

which verifies the last properties.

Remarks: Strong and Weak Stationarity

Remark 11.

Neither type of stationarity implies the other one. This is 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 process 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 $W_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ 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 processes 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. 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 the lags h).

Given the definition of the PSD, as for the autocovariance function case, we can say that once we know the PSD we know everything there is to know about the process. 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:

$$\text{var}\left(\bar{X}\right) = \frac{1}{T^2}\,\text{cov}\left(\sum_{t=1}^{T} X_t, \sum_{s=1}^{T} X_s\right) = \frac{\gamma(0)}{T}\sum_{h=-T}^{T}\left(1 - \frac{|h|}{T}\right)\rho(h). \qquad (2.2)$$

The derivation of (2.2) is instructive and is given in Appendix C ▸ Go to Appendix C. Moreover, some simulation-based and analytical methods to estimate var $\left(\bar{X}\right)$ 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

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

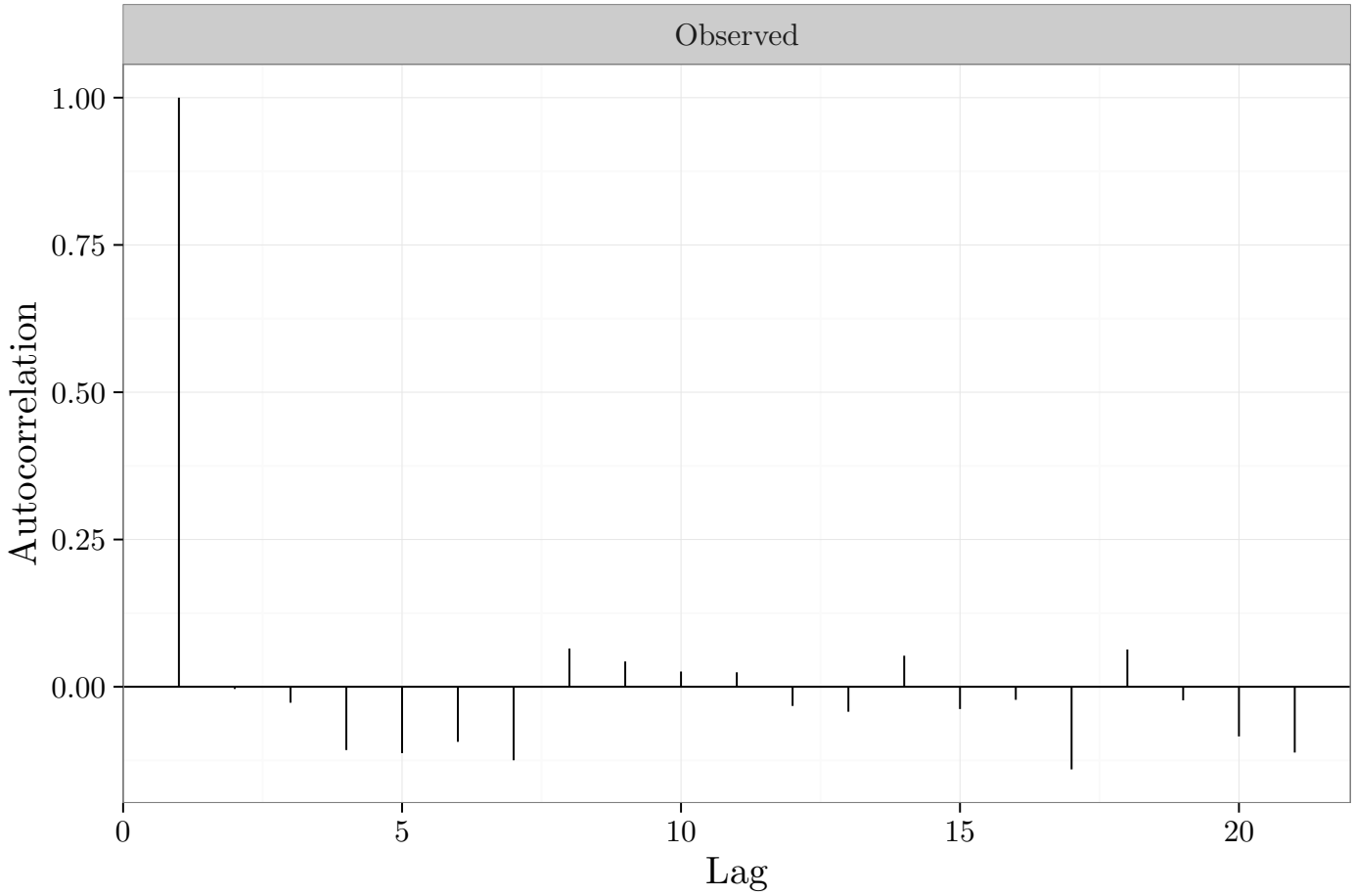
with $\hat{\gamma}(h) = \hat{\gamma}(-h)$ for $h = 0, 1, ..., k$, where k is a fixed integer.

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 $h = 0, 1, ..., k$, 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:

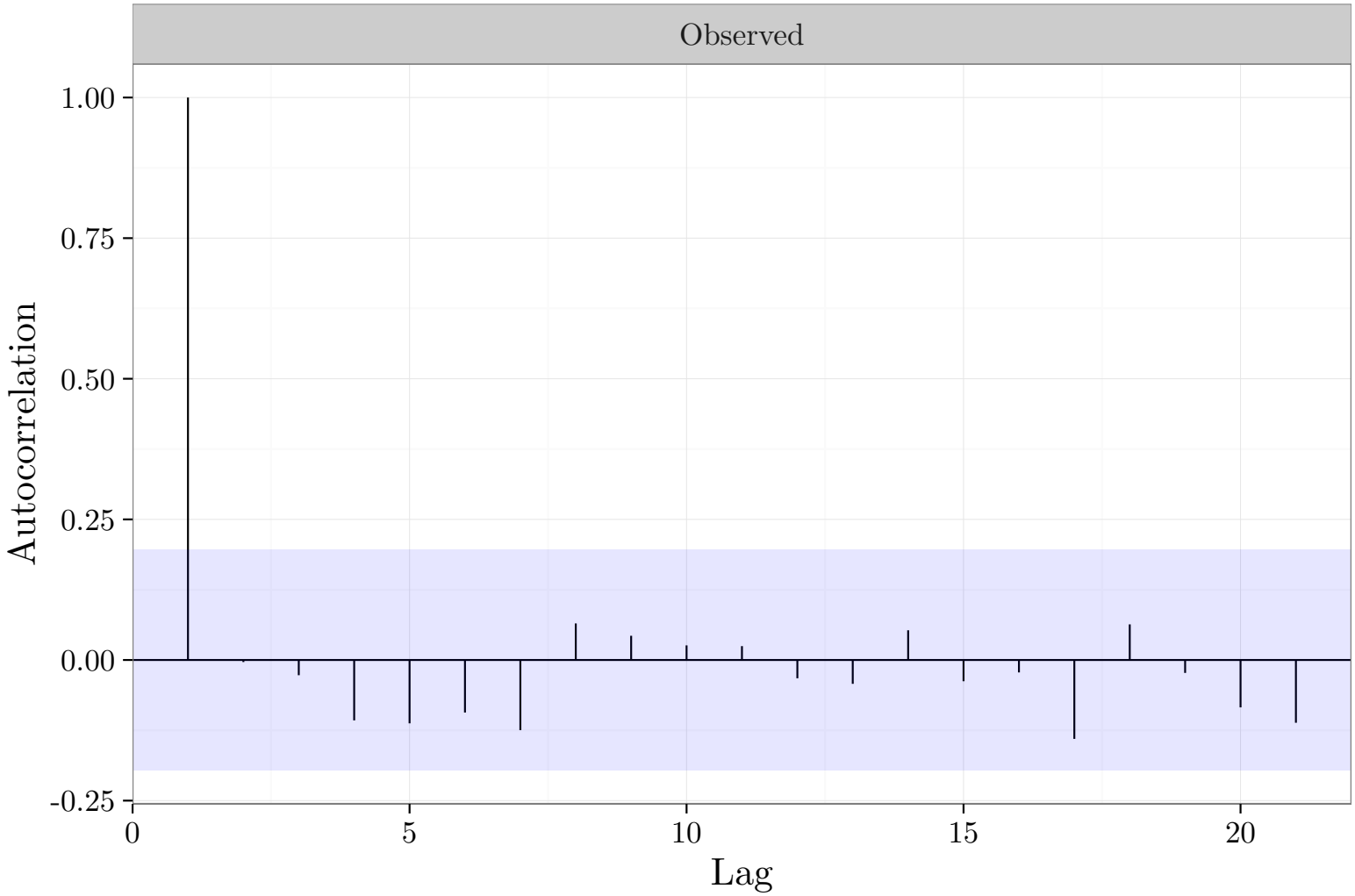
If (X_t) is a white noise then $\hat{\rho}(h)$ should be equal to 0 if $h \neq 0$. In practice, this is not the case because of the estimation error of $\hat{\rho}(h)$. The next result give us a way to assess whether the data comes from a completely random series or whether correlations are statistically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

Property:

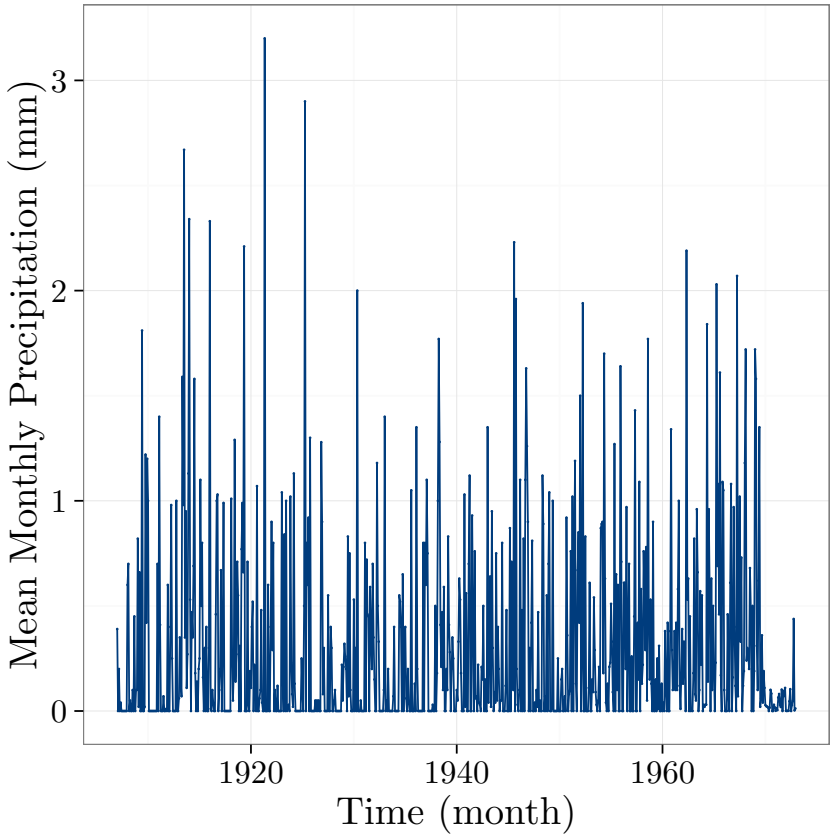
Under some technical conditions, if (X_t) is white noise and $h = 1, ..., H$ where H is fixed but arbitrary we have that:

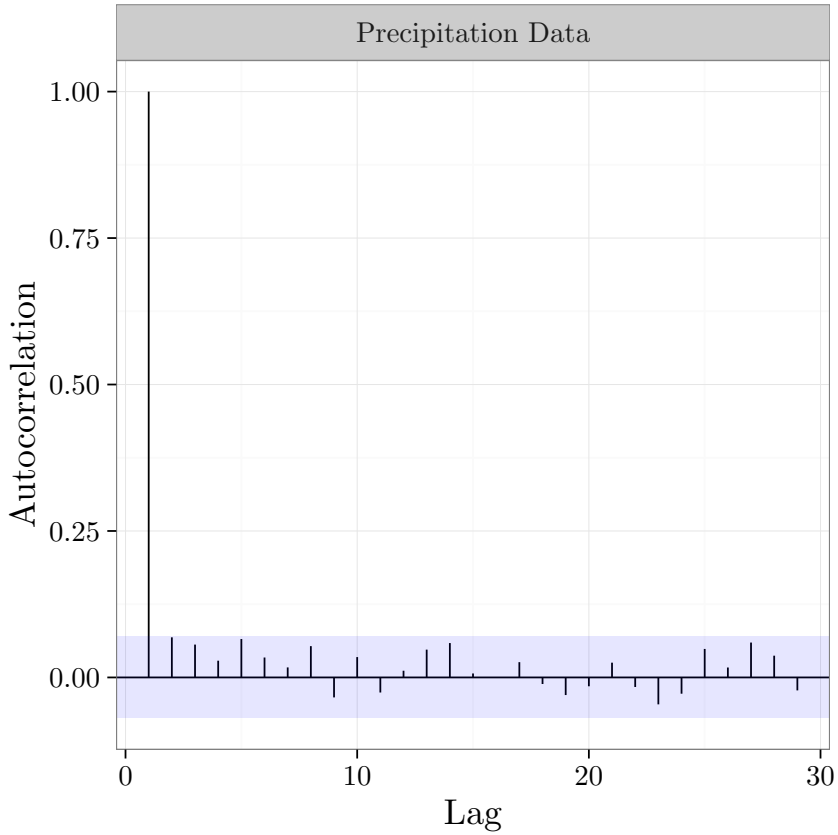
$$\sqrt{T}\left(\hat{\rho}(h) - \rho(h)\right) \xrightarrow{D} \mathcal{N}(0, 1).$$

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 later.

"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:

$$\begin{aligned} \mathrm{H}_0 : &\quad \rho(1) = \ldots = \rho(h) \\ \mathrm{H}_1 : &\quad \mathrm{H}_0 \text{ is false} \end{aligned}$$

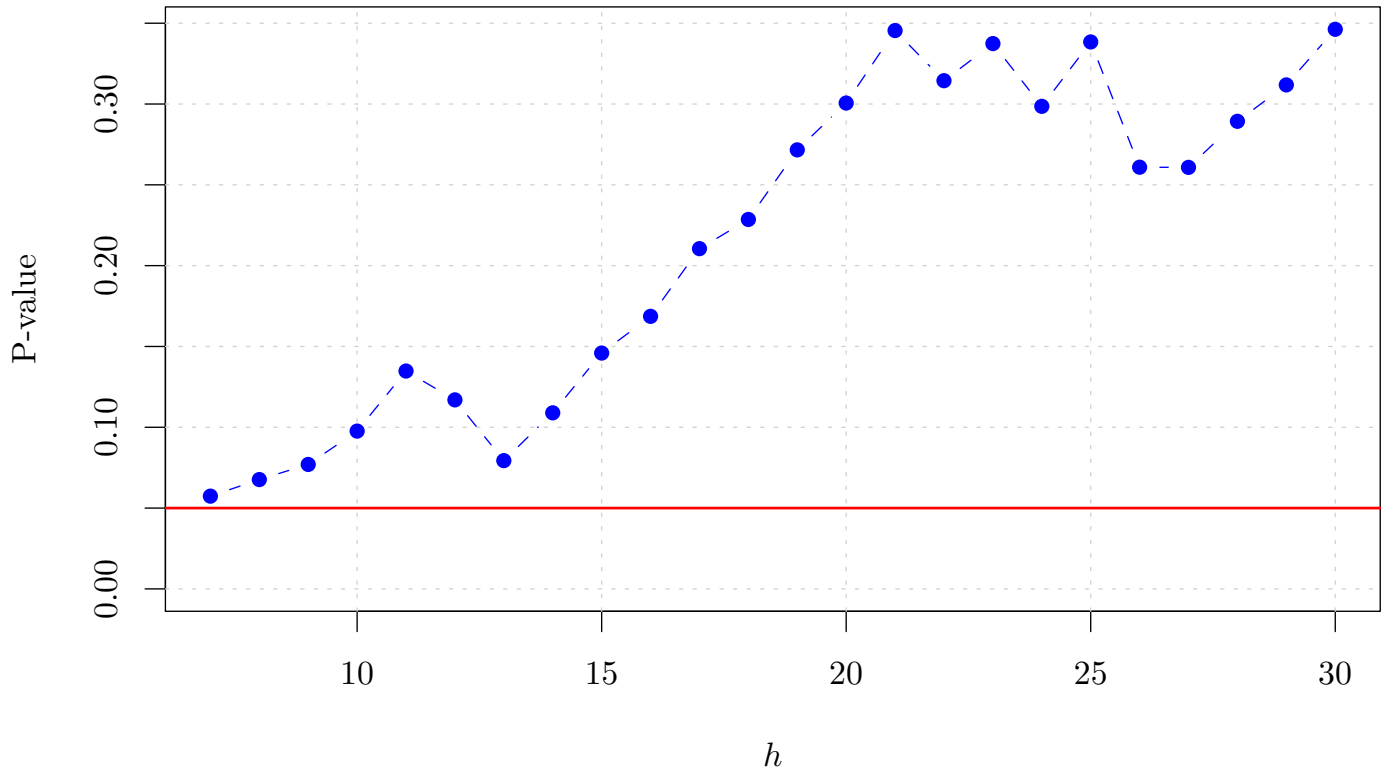
and is based on the test statistic

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

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 estimated to autocorrelation function.

An example:

Consider the following two processes:

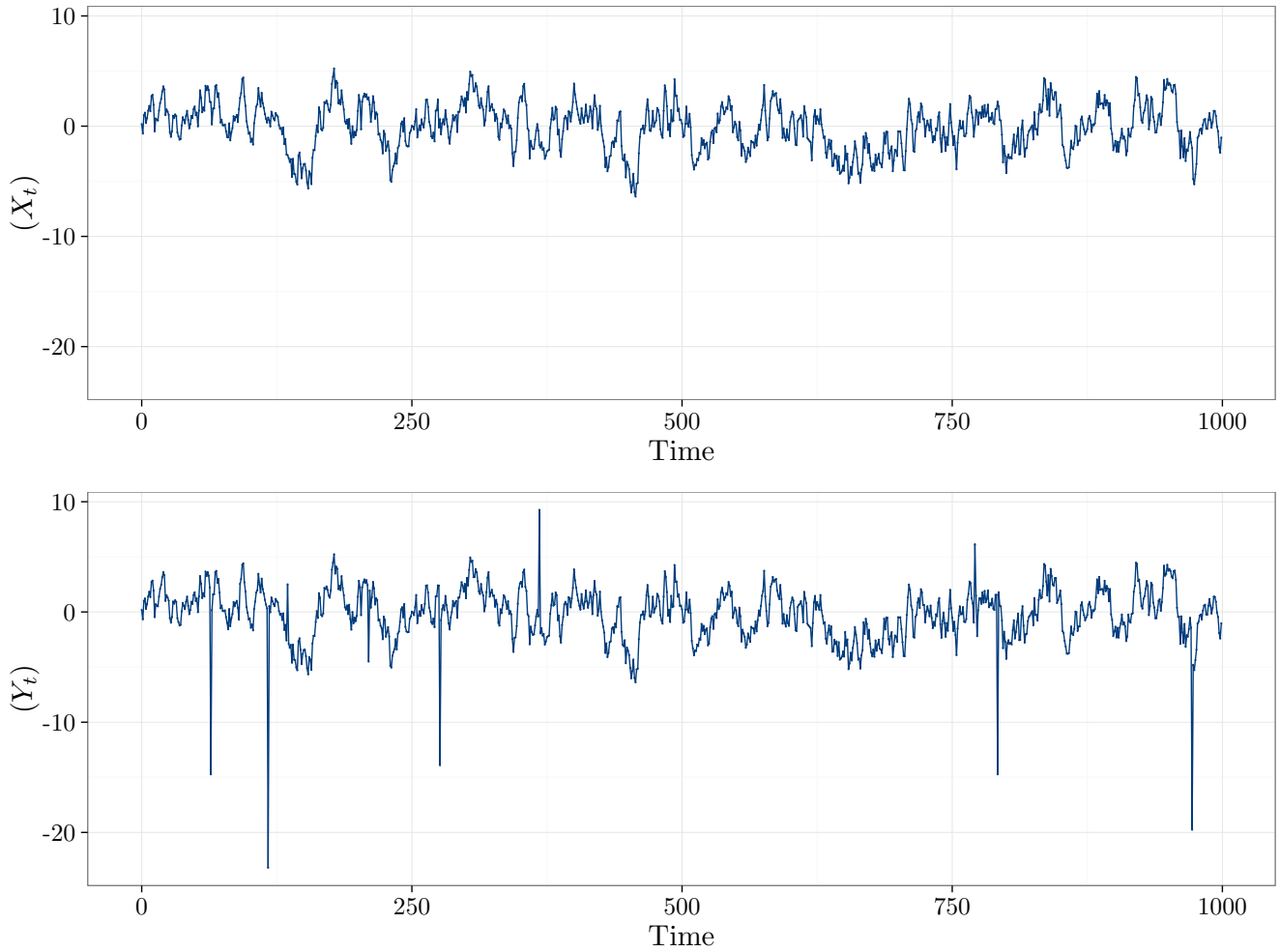
$$X_t = \phi X_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, \sigma_W^2)$$

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

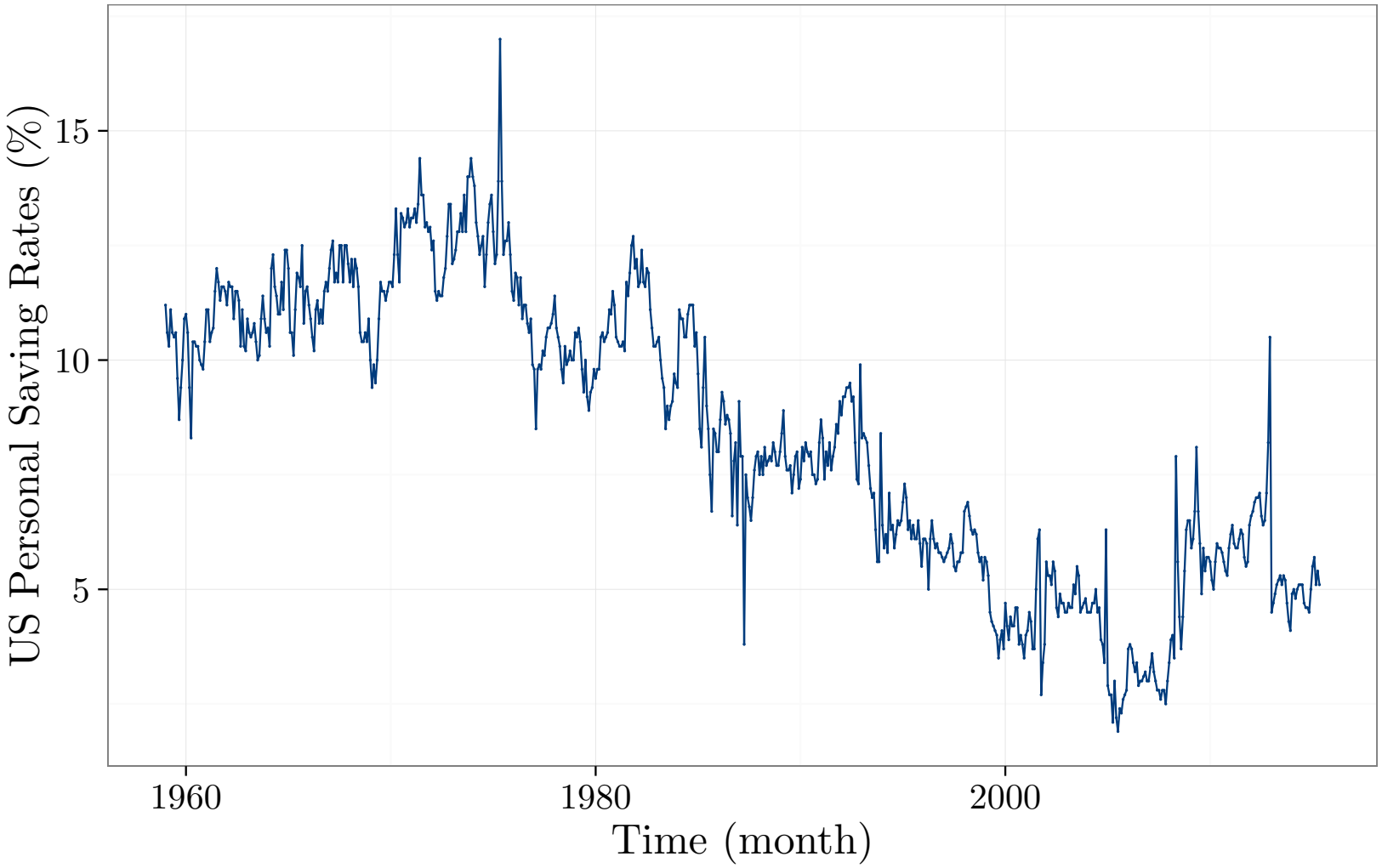
when ϵ is "small" and $\sigma_u^2 \gg \sigma_w^2$, the process (Y_t) can be interpreted as a "contaminated" version of (X_t).

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):

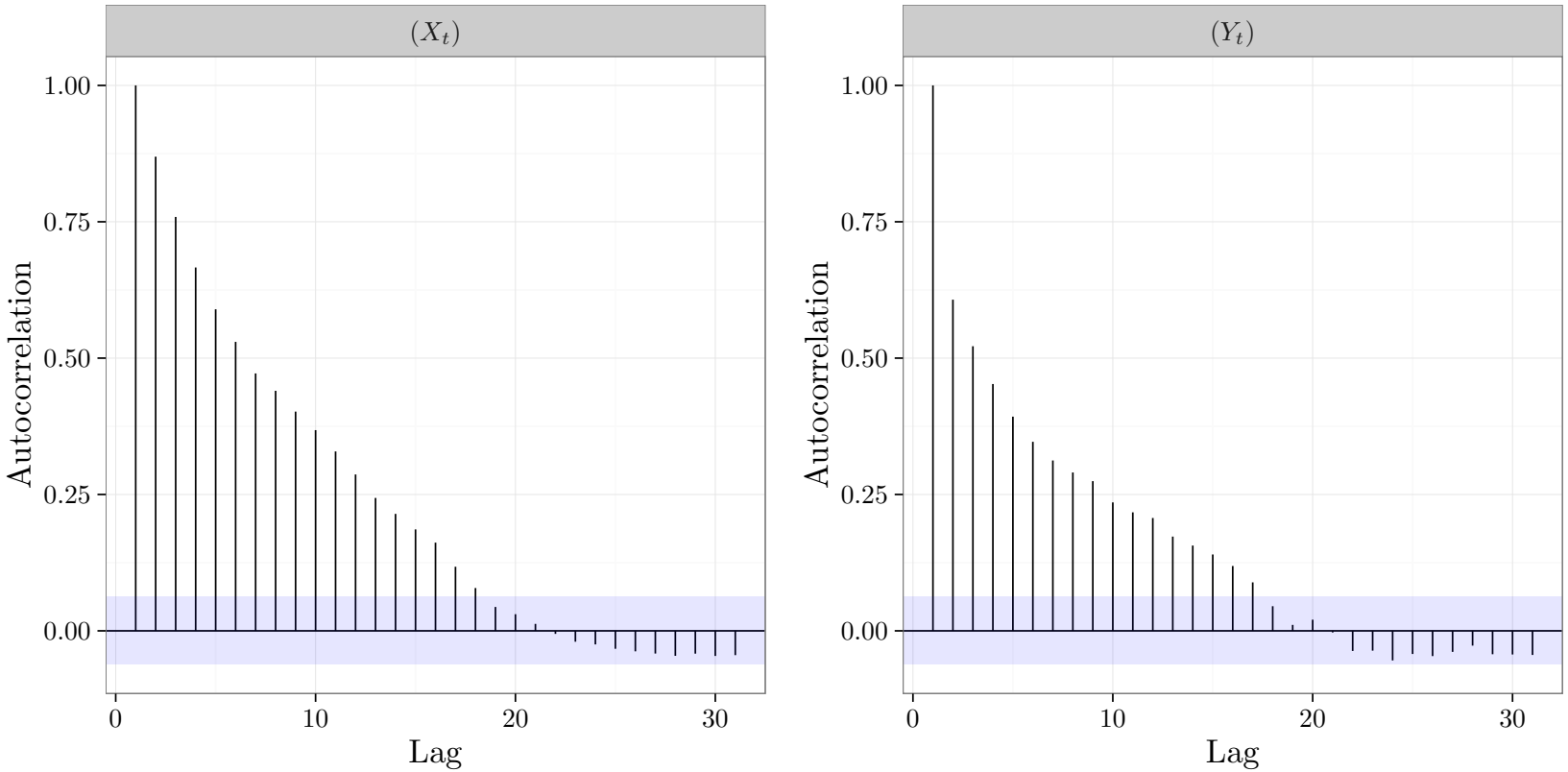


Remember: Personal US Saving Rates



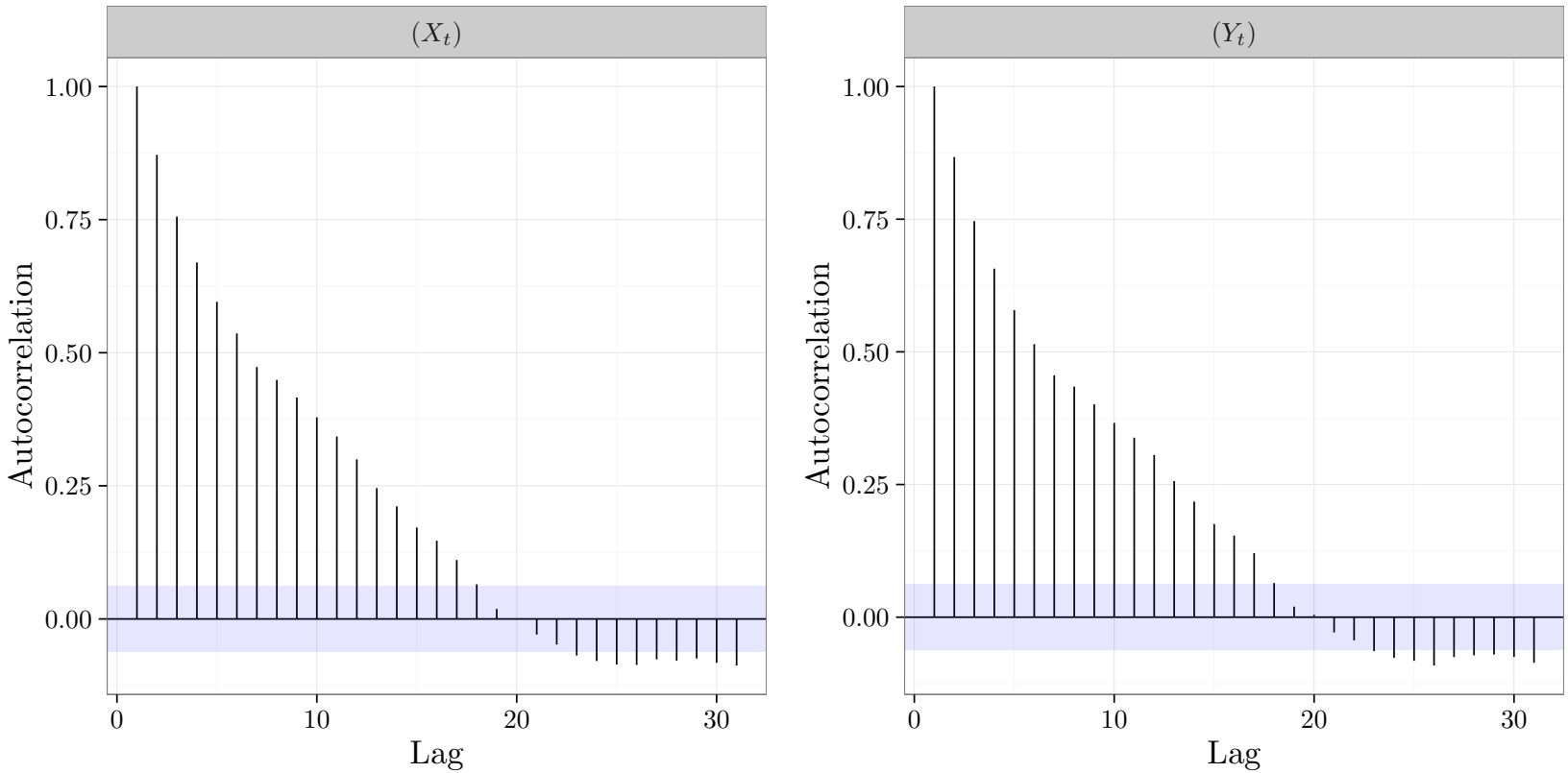
Robust Statistics - A Motivating Example

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



A robust estimator of the autocorrelation

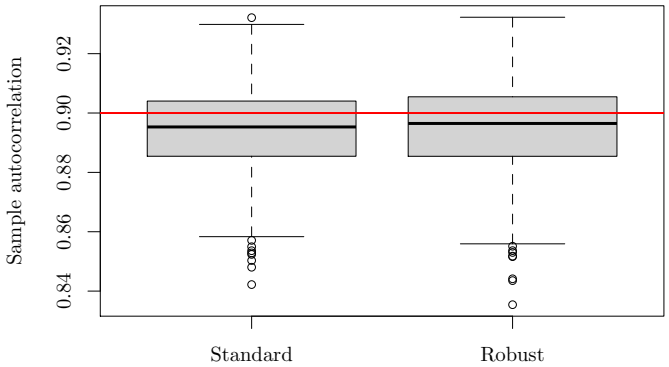
Several robust estimators of the autocorrelation have been introduced in the literature (see e.g. Ma and Genton, 2000). Such estimators are less efficient but tend to be less affected by extreme observations. Let us compare the robust ACFs of the two processes (using the R function `robacf`).



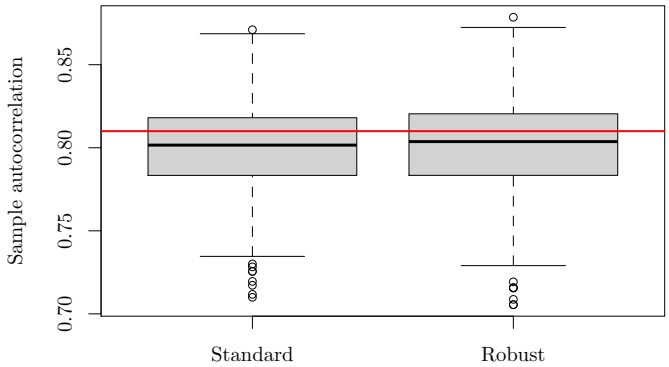
A small simulation study

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

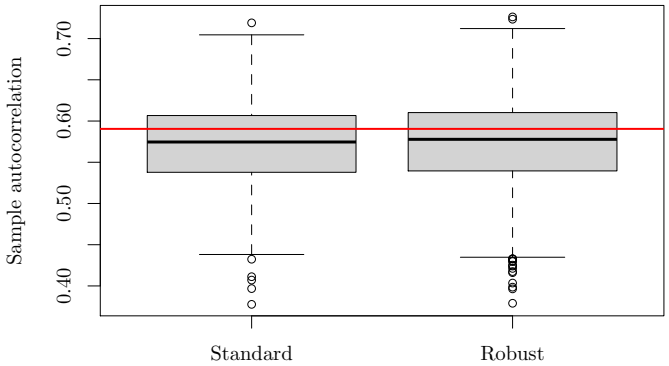
lag: h = 1



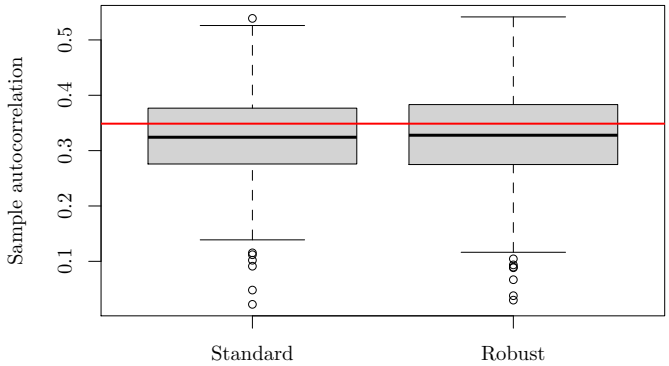
lag: h = 2



lag: h = 5

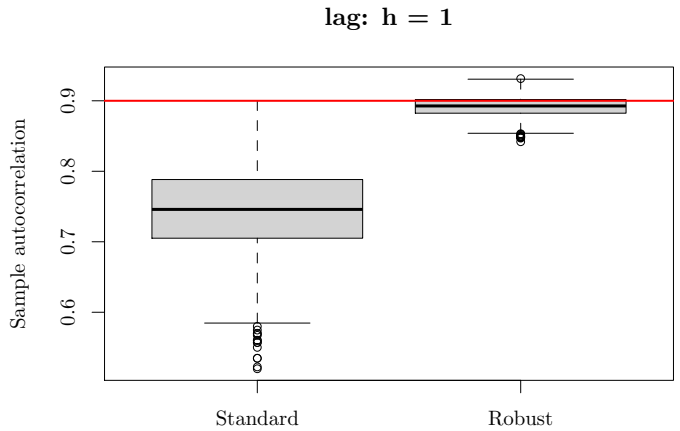


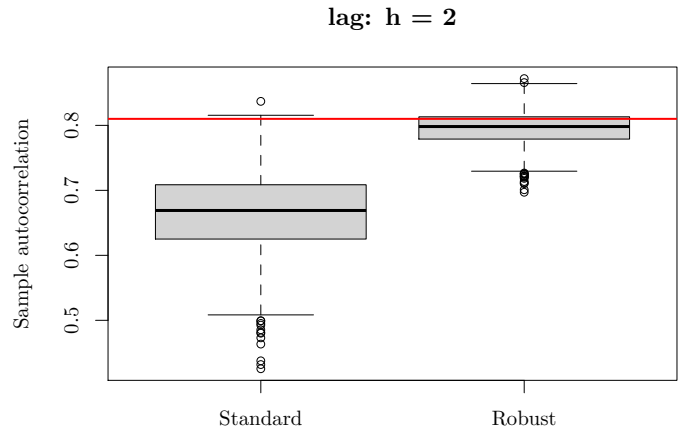
lag: h = 10

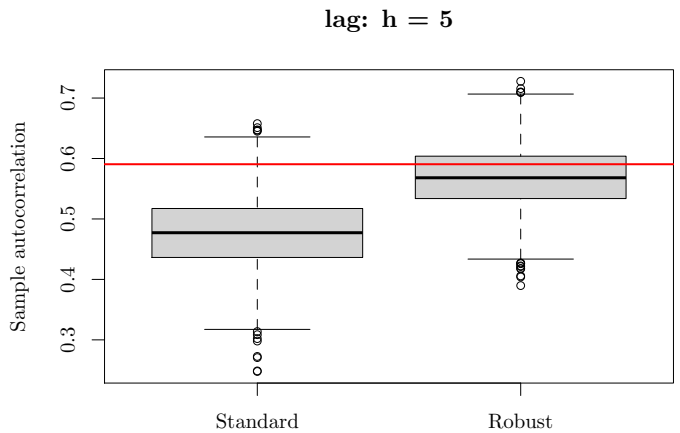


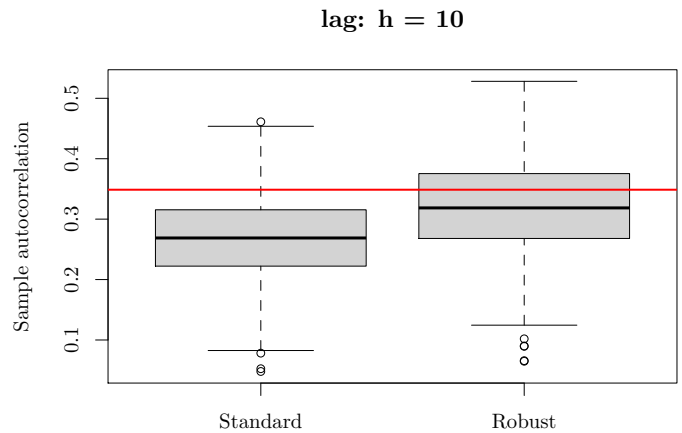
A small simulation study - cont.

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



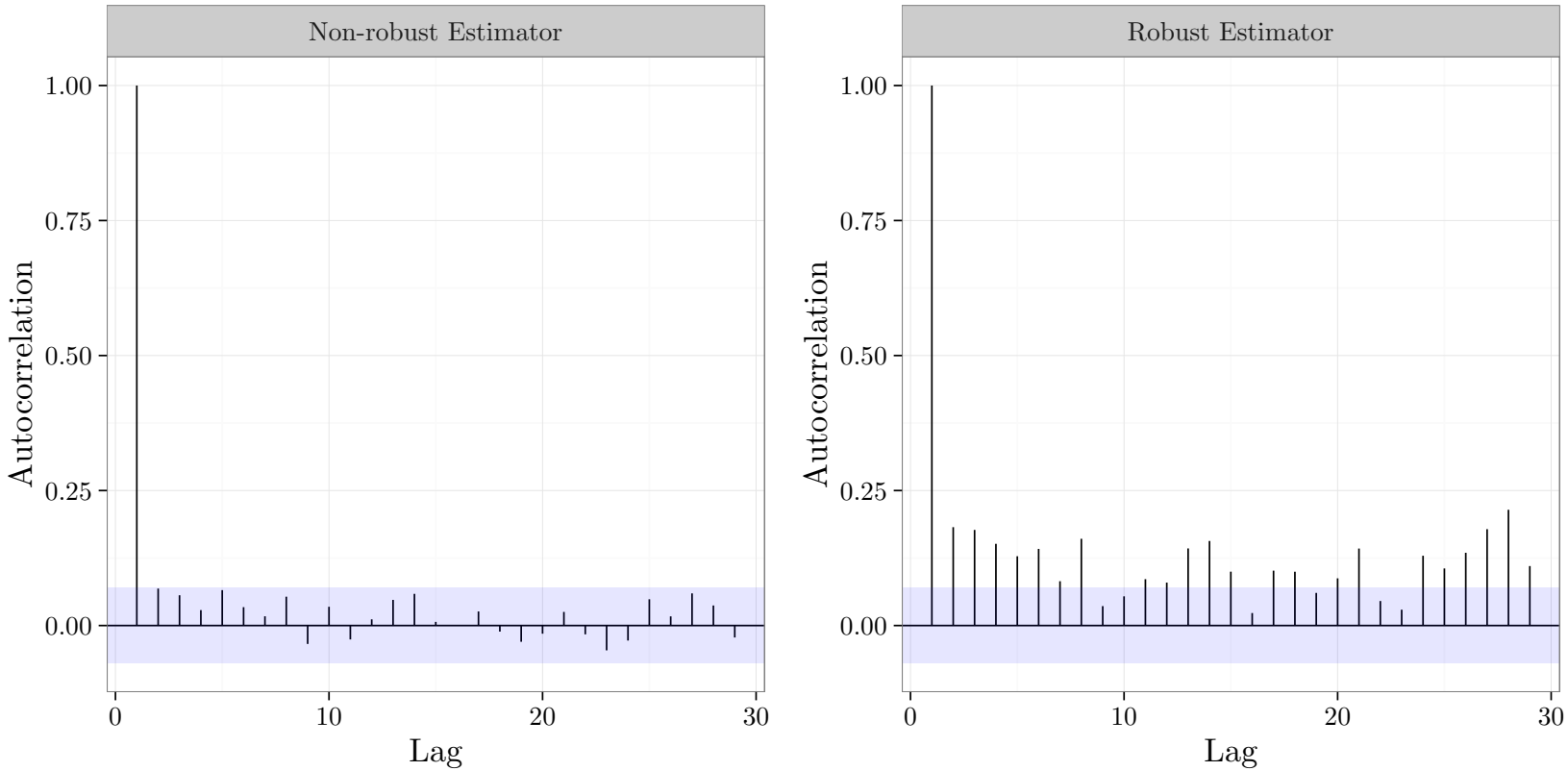






Example: ACF of Precipitation Data - cont.

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 and we obtain the decomposition:

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

Motivation for ARMA models

- If $(X_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, 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 $\hat{\mathbf{\Sigma}} \approx \mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$.
- ARMA models typically depends only on 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. It 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 properties, is omitted.

Definition

The partial autocorrelation function $(\phi_{h,h})$, 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:

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

Autoregressive Models

Definition

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

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

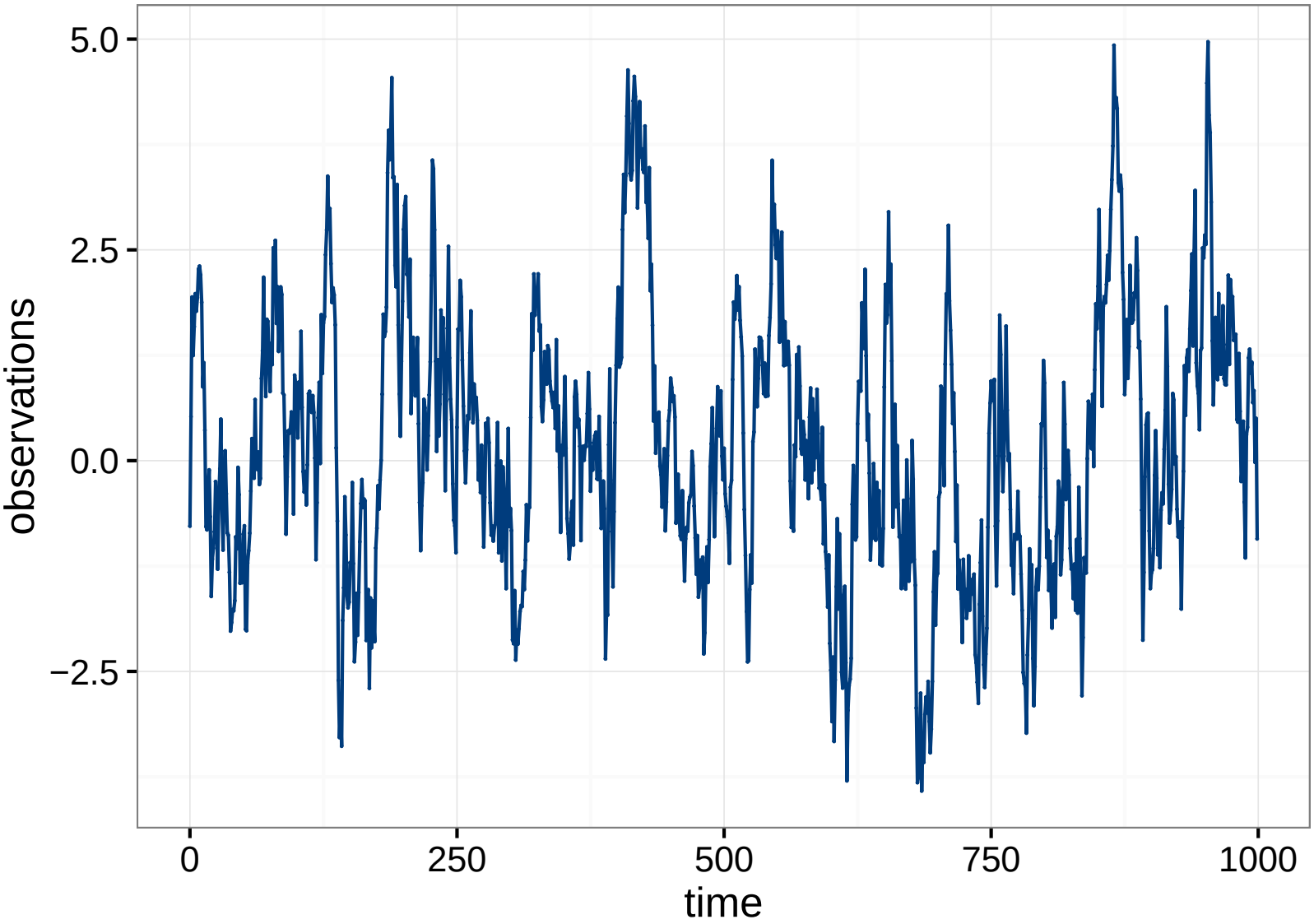
where X_t is stationary, $\phi_1, ..., \phi_p$ are constant ($\phi_p \neq 0$) and $W_t \sim \mathcal{N}(0, \sigma_w^2)$.

Remark:

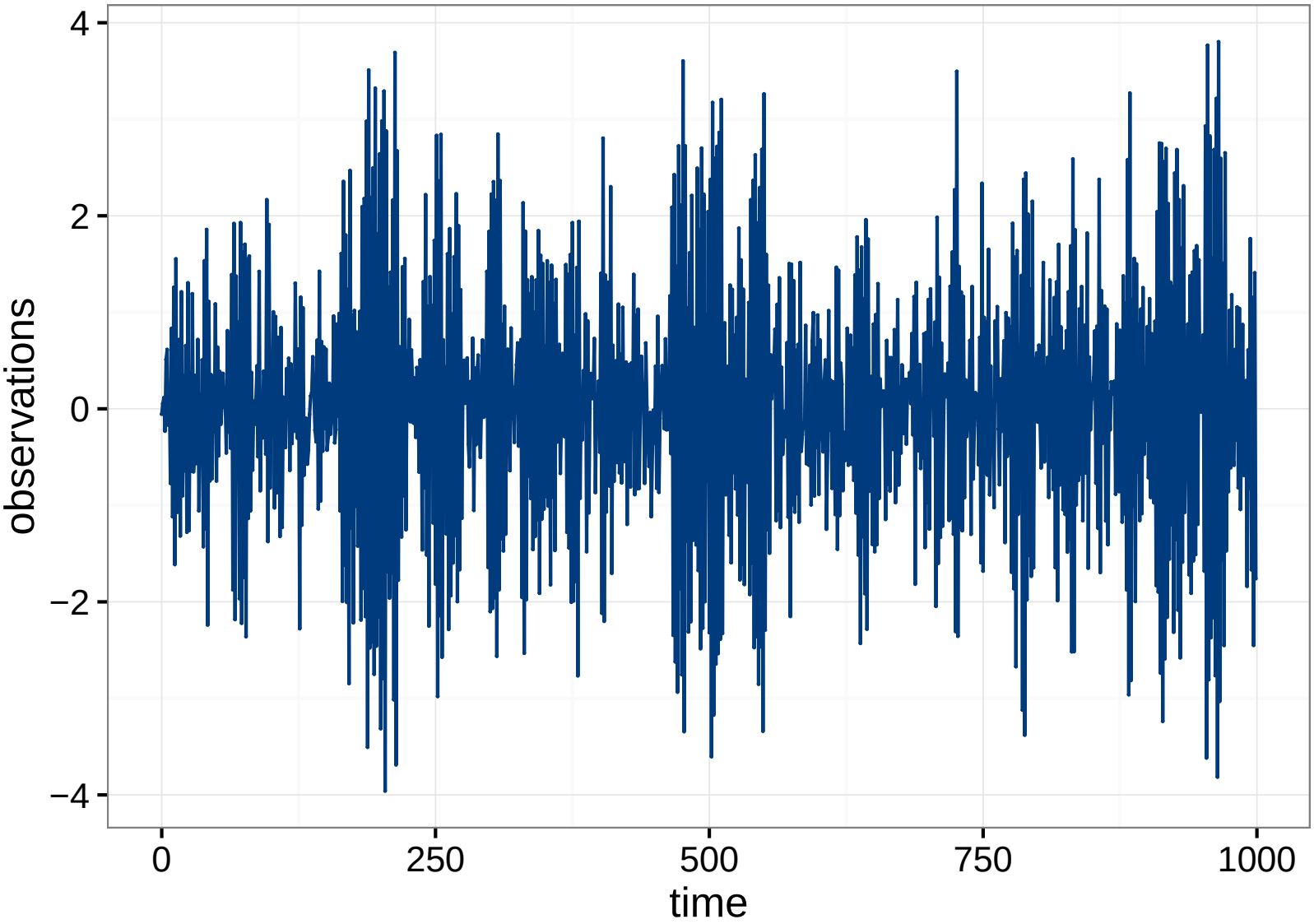
We assume, in the above definition, that the expected value of X_t is zero. If this is not the case, we can replace X_t by $X_t - \mu$ (where $\mu = \mathbb{E}[X_t]$) and obtain:

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

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:

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

and, therefore, its autocorrelation is:

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

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 $\phi_{h,h} = 0, \quad h > 1$.

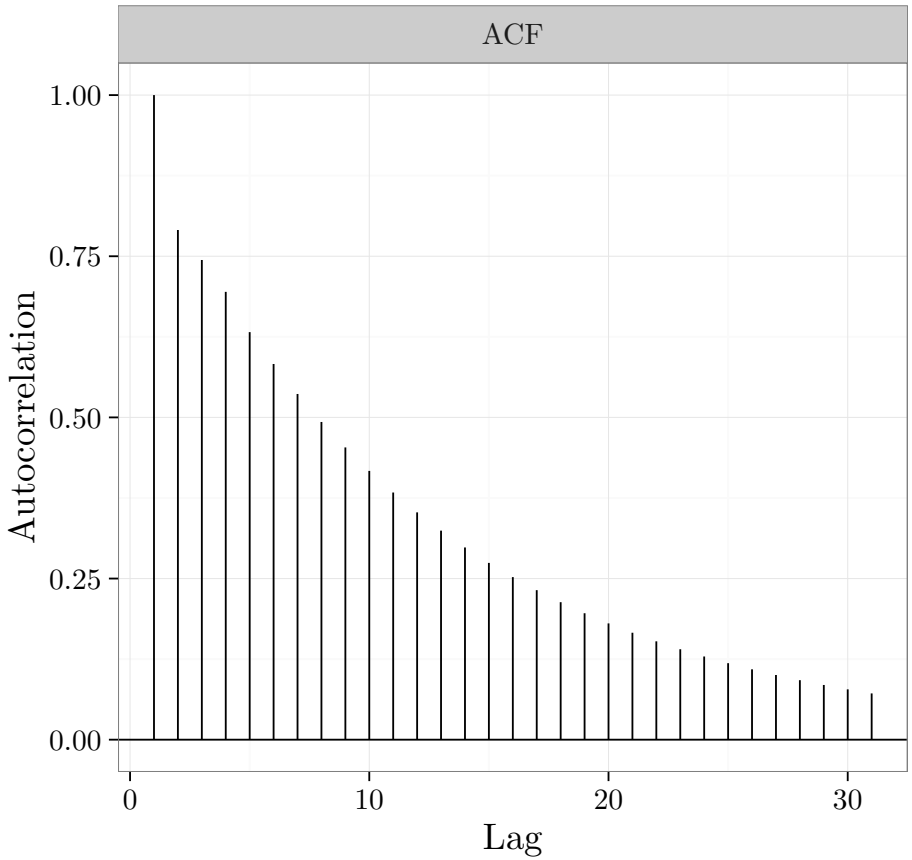
Theoretical ACF and PACF of AR(p) models

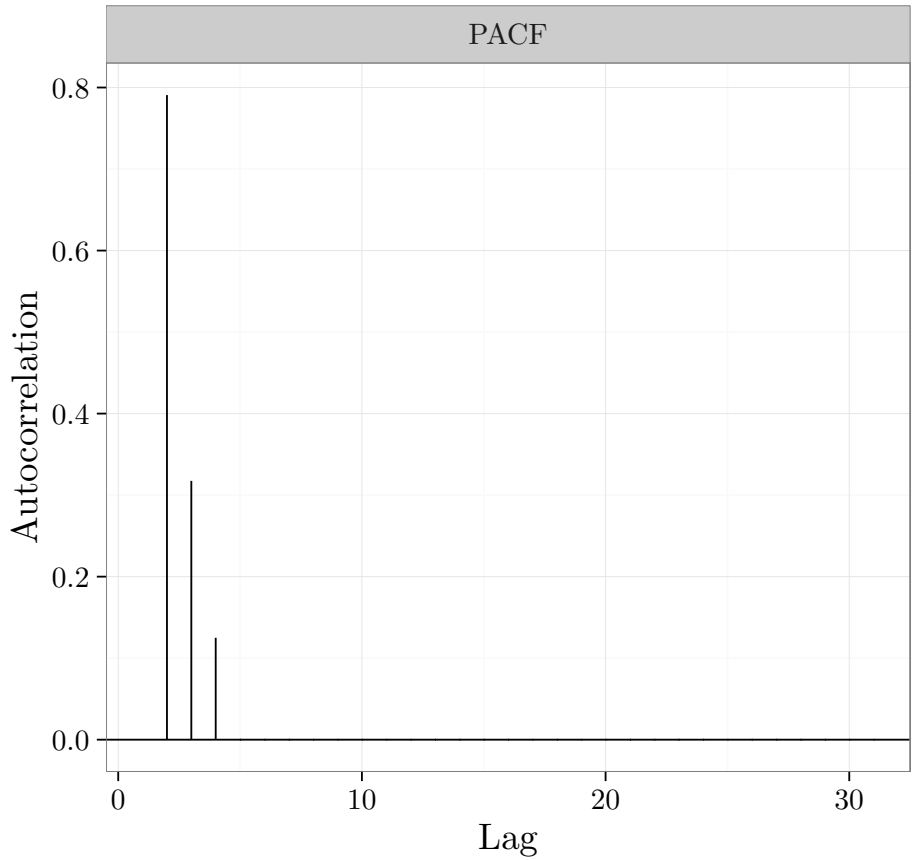
- In general, the ACF of an AR(p) is a complicated function of the parameters but $\rho(h)$ tends to dampen, in a sinusoidal fashion, exponentially fast to zero as h increases.
- The PACF of an AR(p) is non-zero for the first p lags and zero for $h > p$.
- As an Example, consider the following models:

$$
\begin{aligned}
X_t &= 0.95X_{t-1} + W_t \\
X_t &= 1.5X_{t-1} - 0.75X_{t-2} + W_t \\
X_t &= 0.5X_{t-1} + 0.25X_{t-2} + 0.125X_{t-3} + W_t \\
X_t &= 0.5X_{t-1} + 0.75X_{t-2} - 0.8X_{t-3} + W_t \\
X_t &= -0.1X_{t-2} - 0.1X_{t-4} + 0.6X_{t-5} + W_t
\end{aligned}
$$

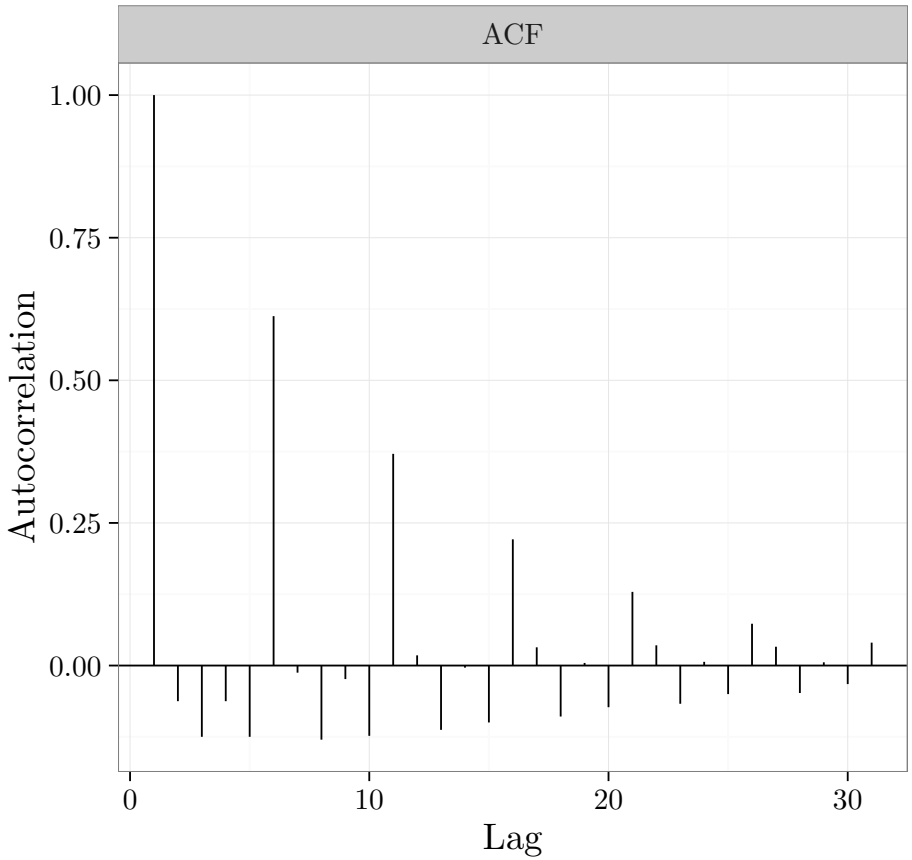
Can we recognize each model given their theoretical ACF/PACF? Let's try...

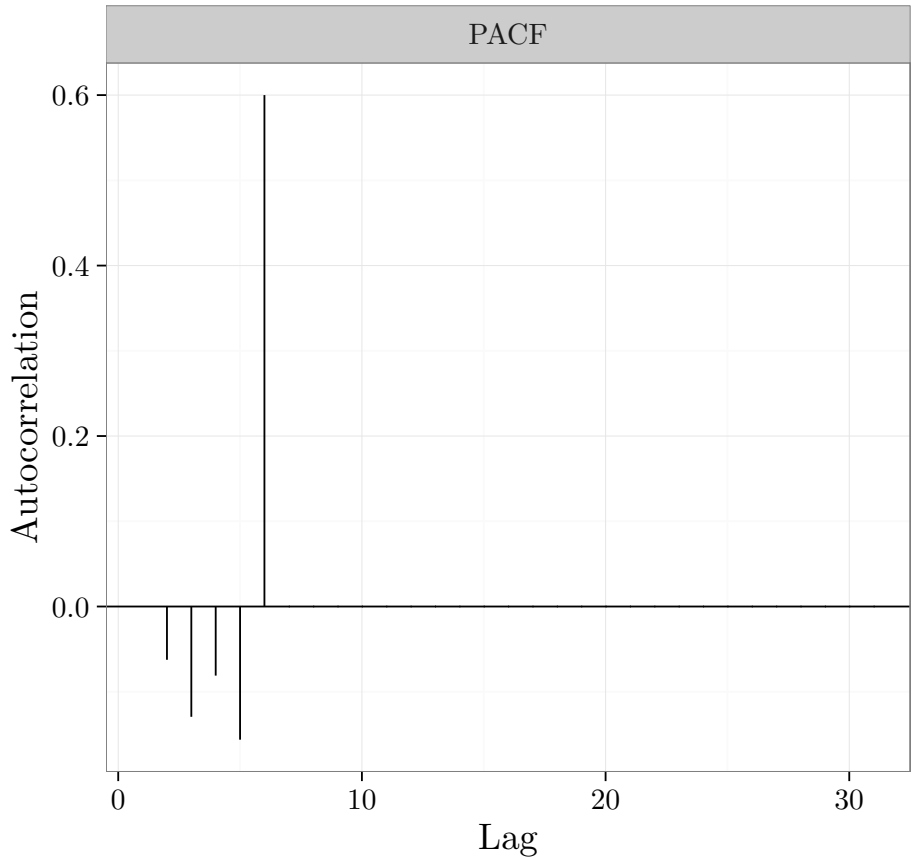
Theoretical ACF and PACF of AR(p) models



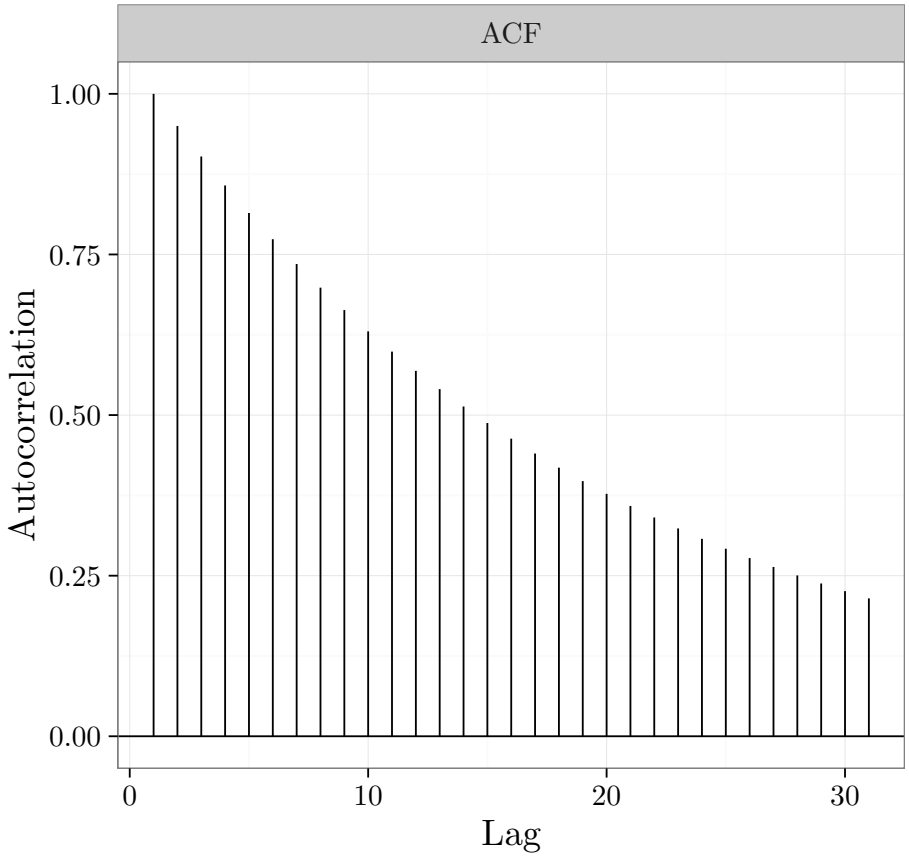


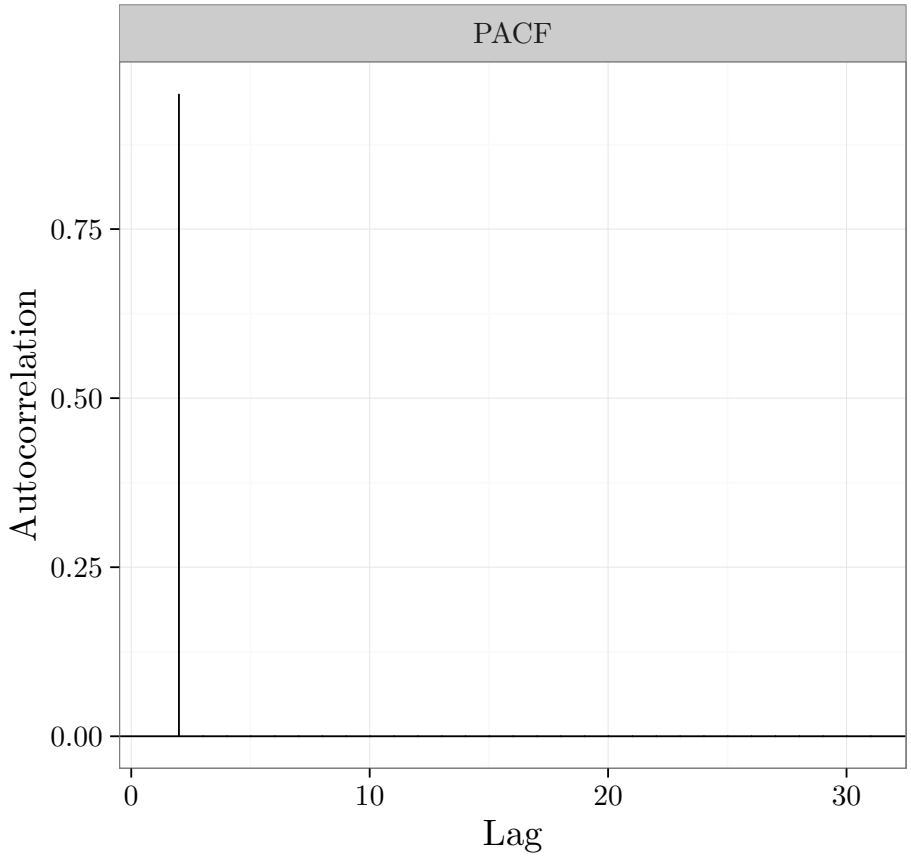
Theoretical ACF and PACF of AR(p) models



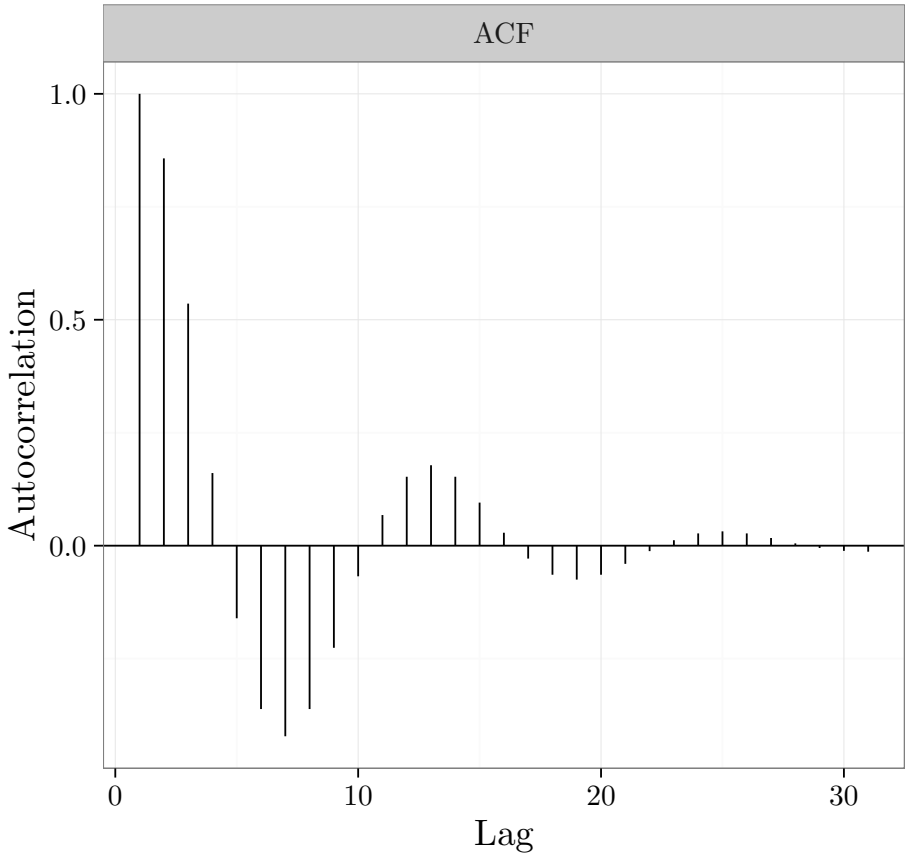


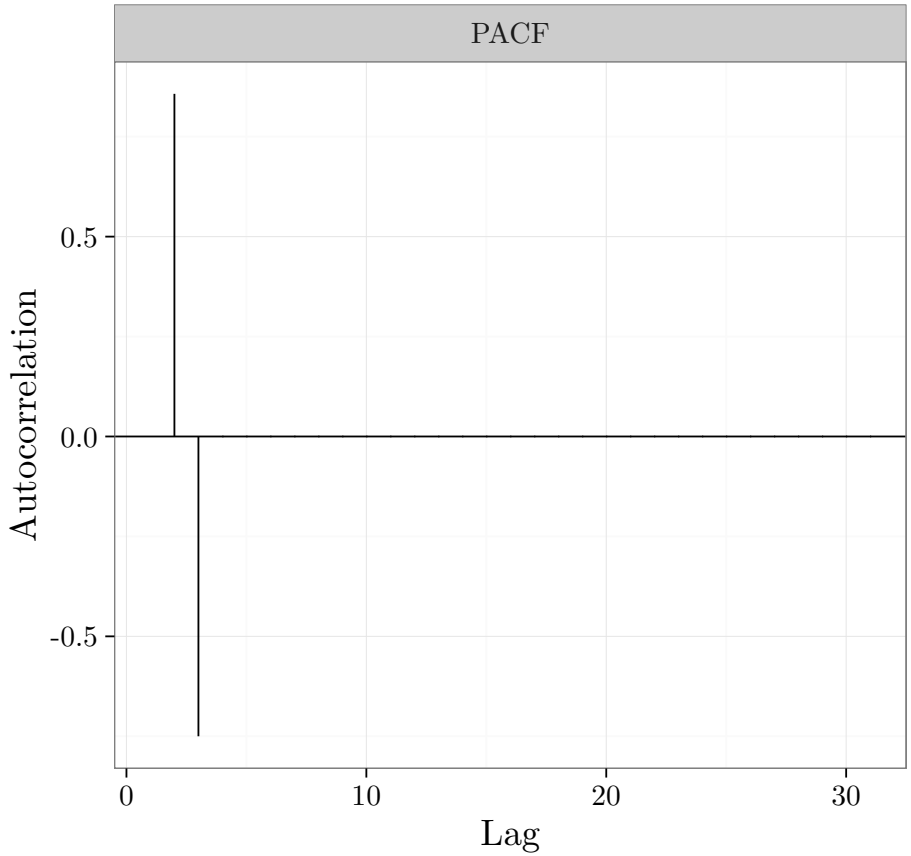
Theoretical ACF and PACF of AR(p) models



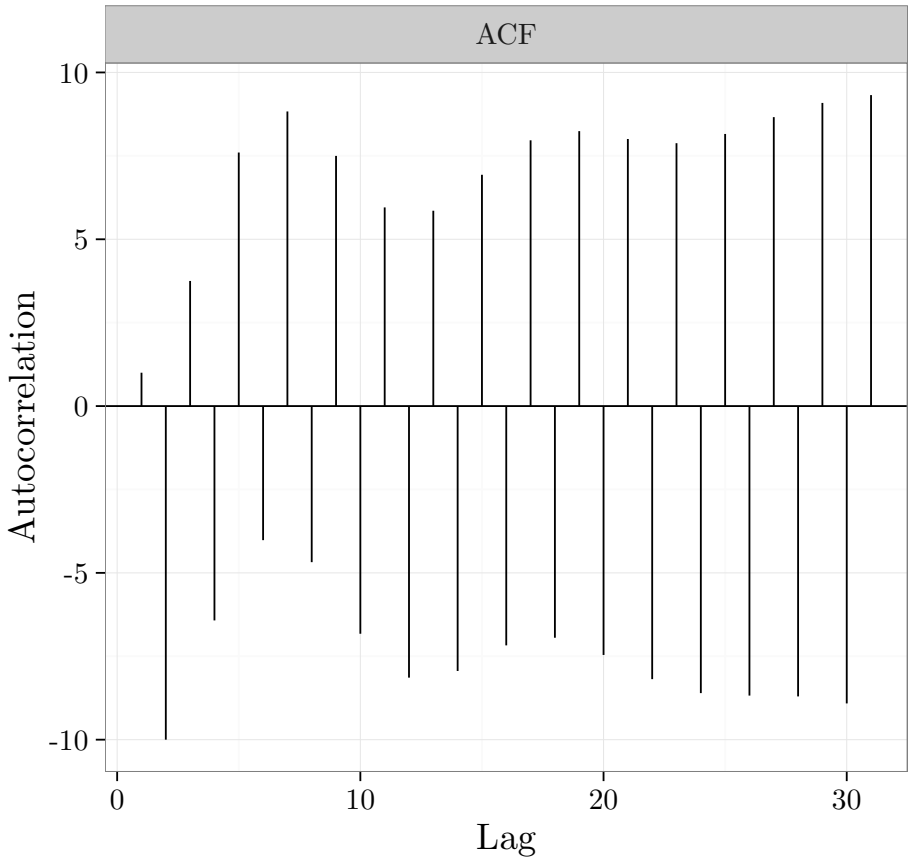


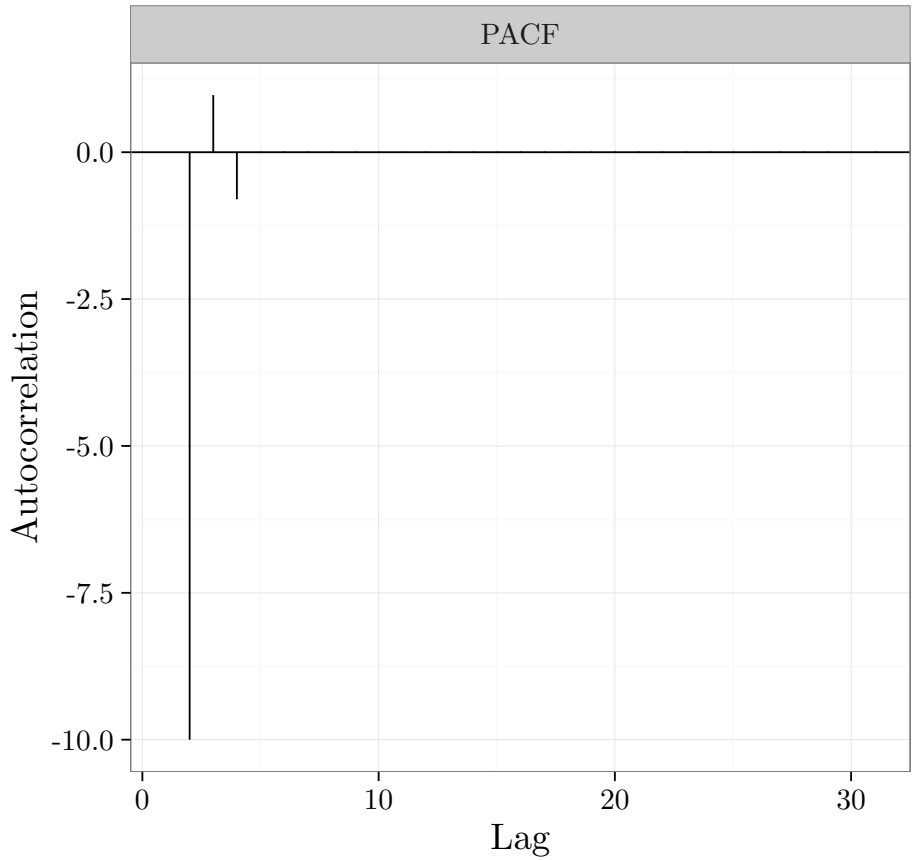
Theoretical ACF and PACF of AR(p) models





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, $X_{t-1}, ...$) to predict the future (i.e. X_{t+1}, $X_{t+2}, ...$).
- Of course, models that are not causal are rather useless to make predictions: 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 definitions of notations.

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:

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

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

Remarks:

- It is not straightforward to verify the causality of an AR(p) model and for this reason it will not be presented here. The 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 method 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 $\mathbb{E}\left[X_{T+h}|X_T, X_{T-1}, ..., X_1\right]$. This quantity intuitively makes sense and relies on valid mathematical arguments (not presented here).
- Consider an AR(1), then we have:

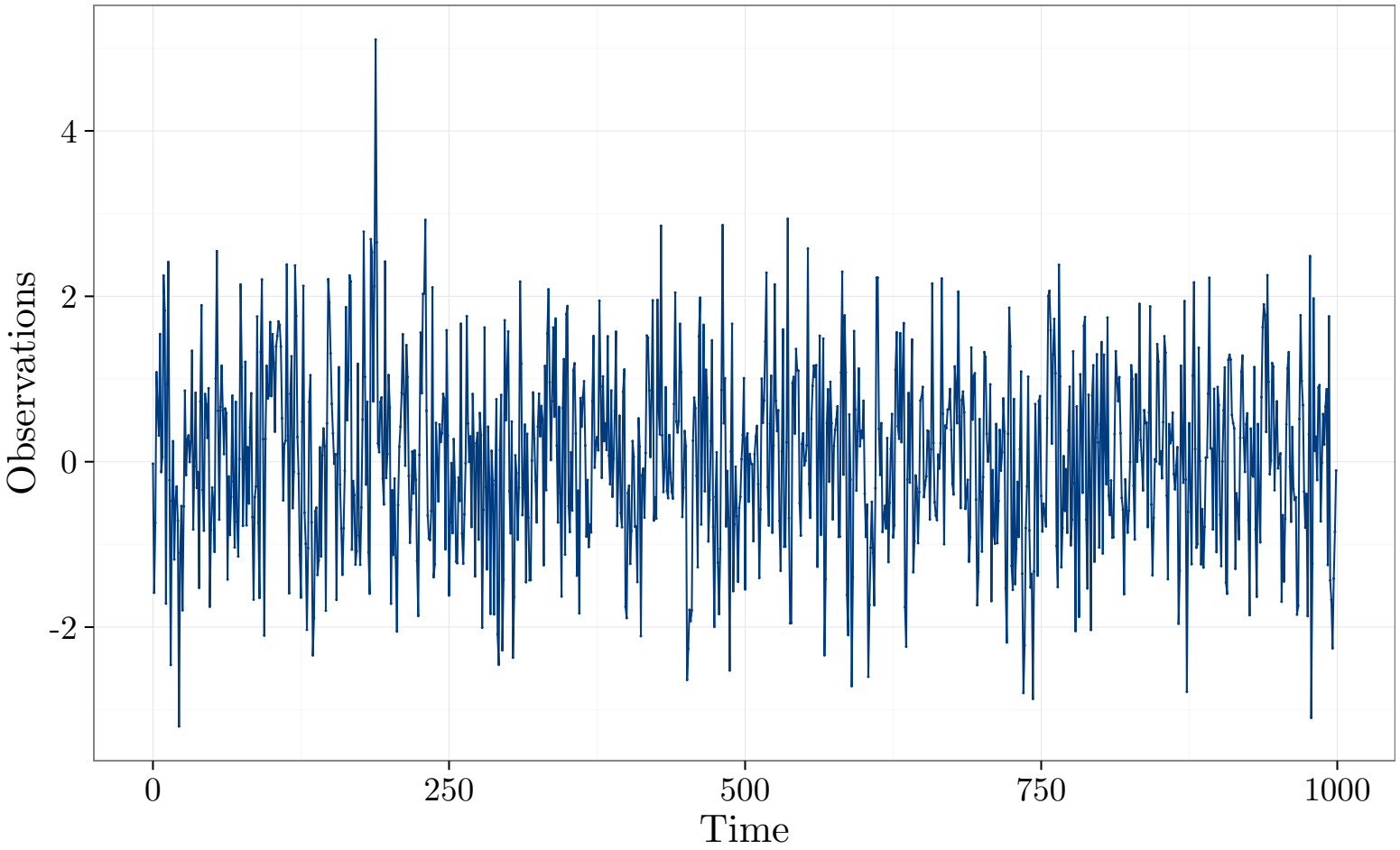
$$\mathbb{E}\left[X_{T+h}|X_T, X_{T-1}, ..., X_1\right] = \mathbb{E}\left[\phi^h X_T|X_T, X_{T-1}, ..., X_1\right] = \phi^h X_T.$$

  A natural estimator of the above quantity is therefore,

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

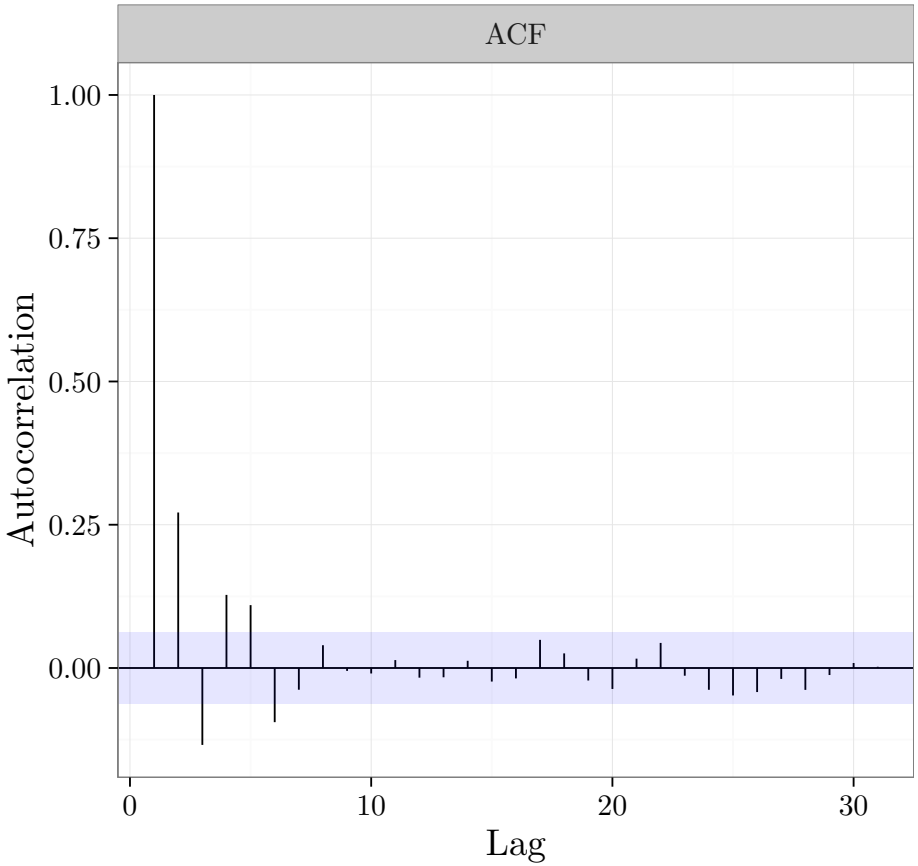
  where $\hat{\phi}$ is a suitable estimator (e.g. OLS, MLE) for ϕ.
- Similar results can be obtained for an AR(p).
- This also creates a framework to compute residuals (in-sample). For example, in the case of an AR(1), we have: $r_t = X_t - \hat{\phi} X_{t-1}$.

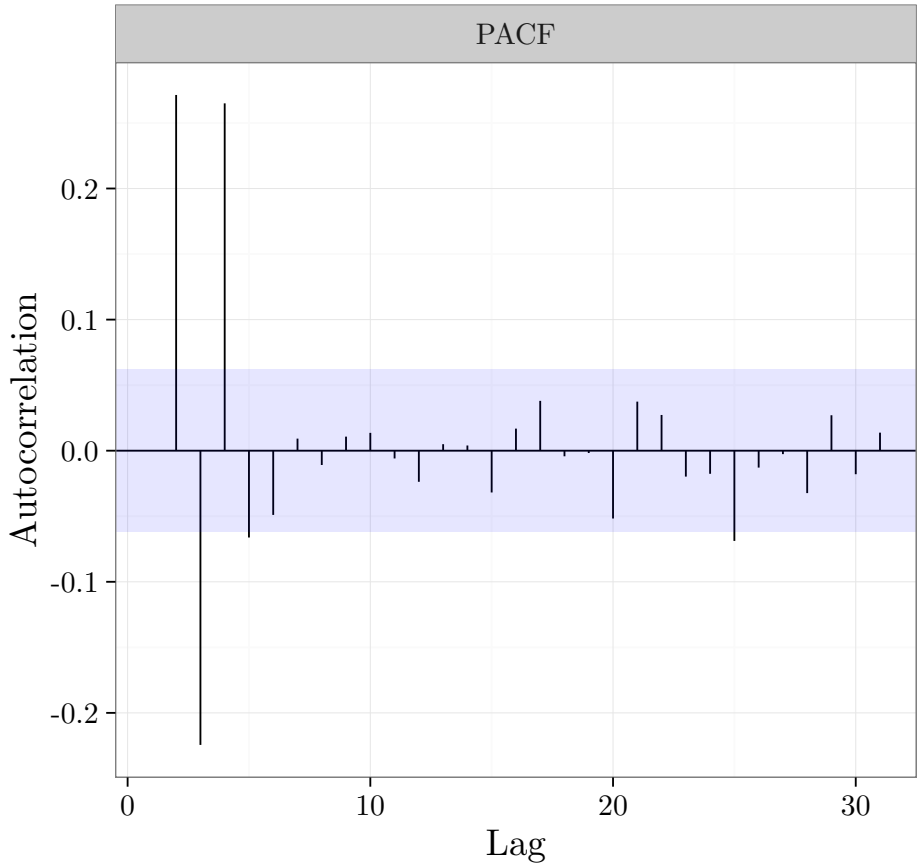
A Simulated Example: Order Identification



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

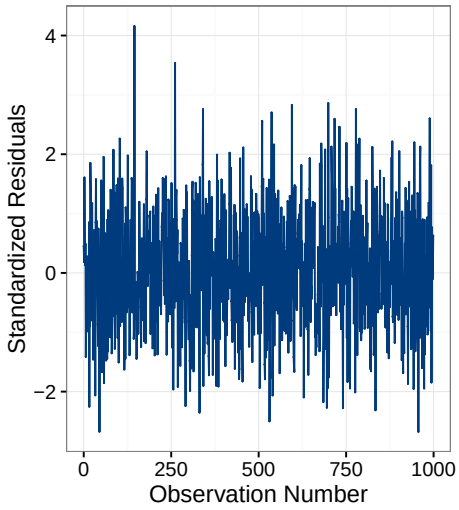
A Simulated Example: ACF/PACF Graphs

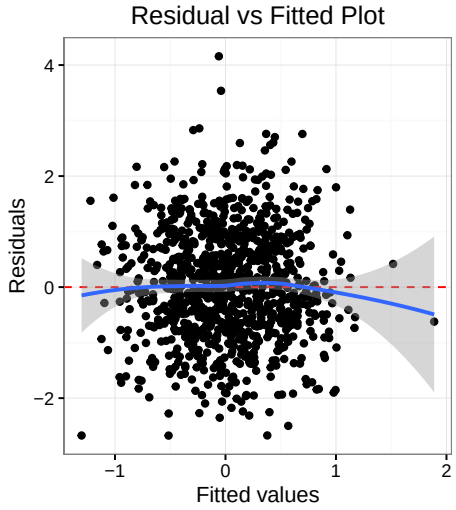


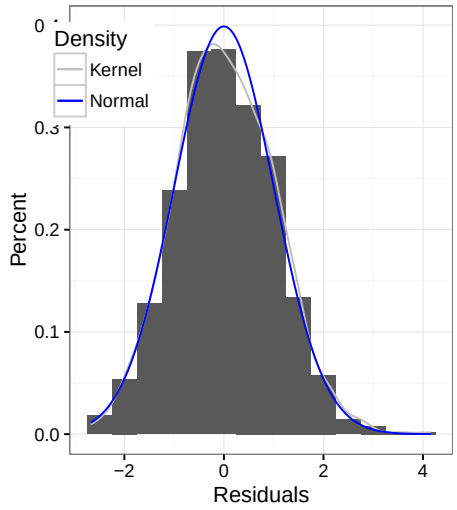


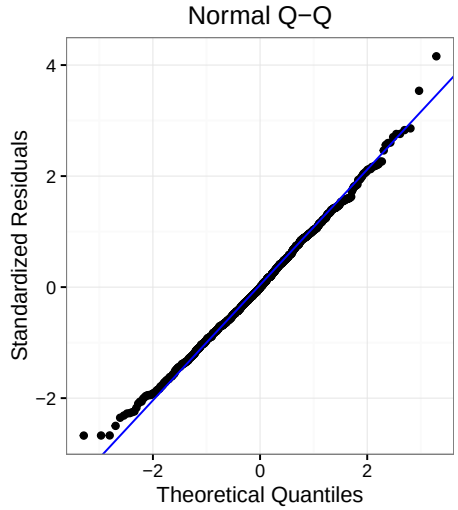
From this graph an AR(3) appears reasonable.

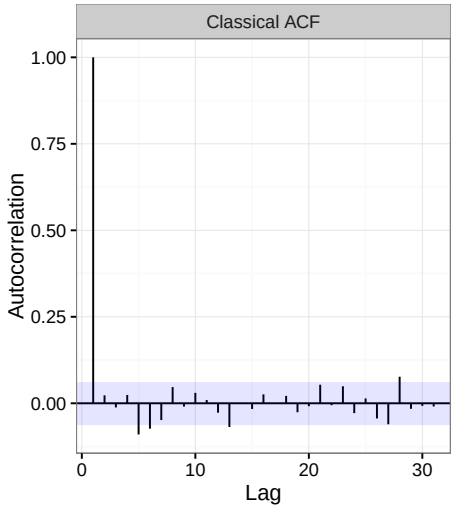
Diagnostic for AR(3)

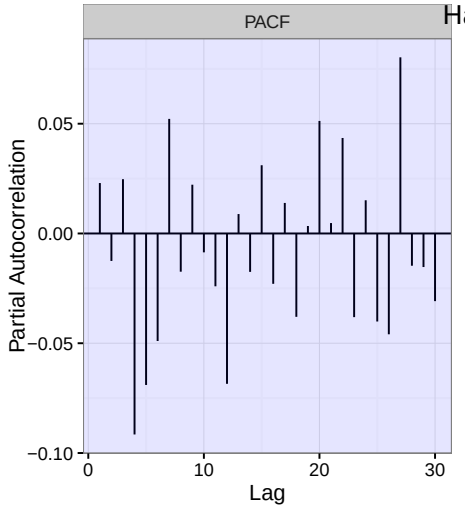


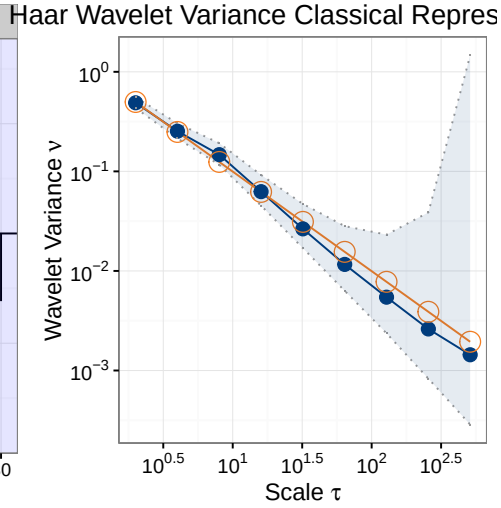


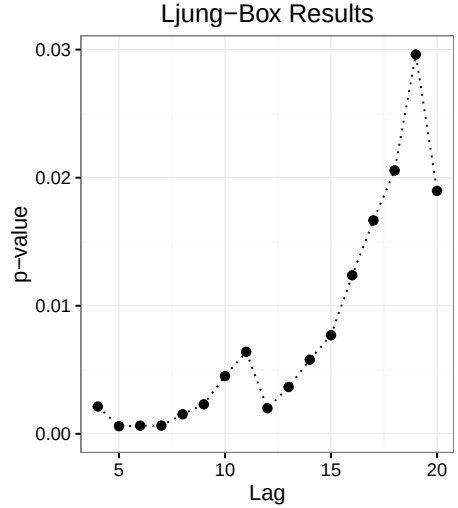




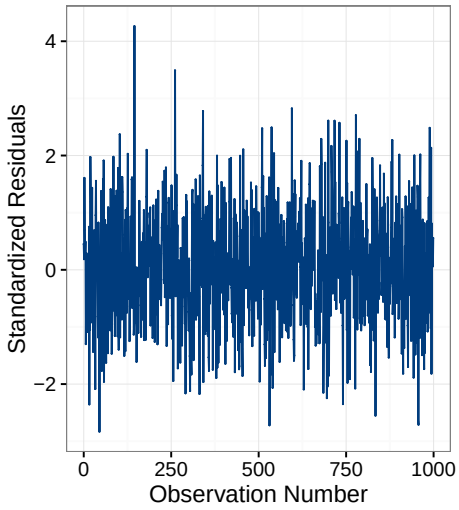


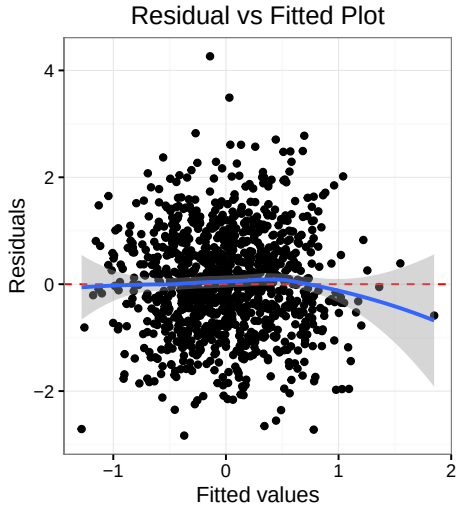


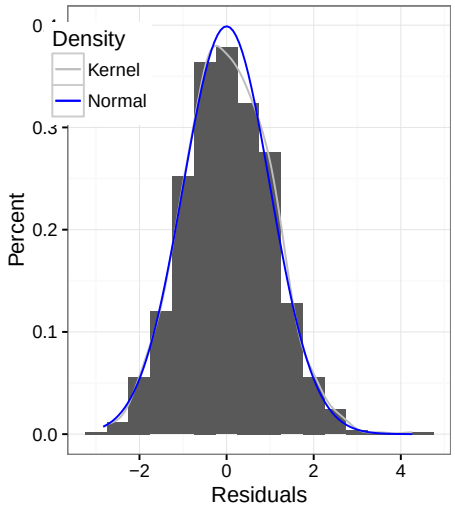


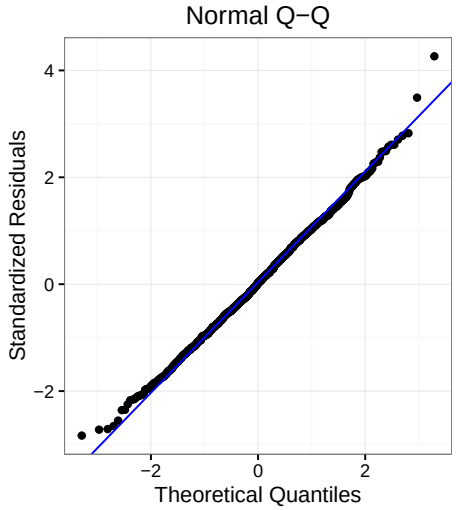


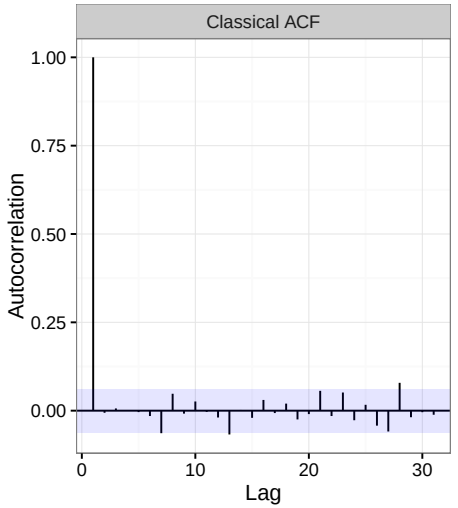
Diagnostic for AR(4)

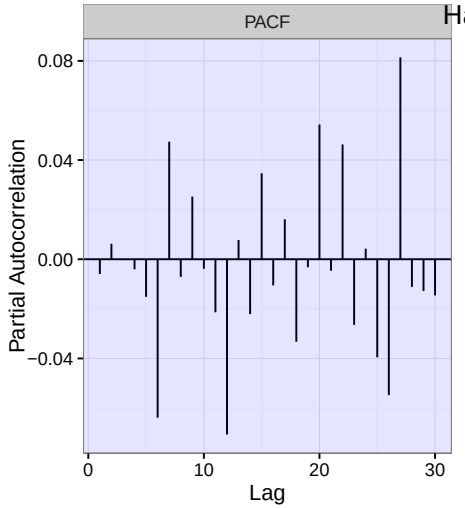


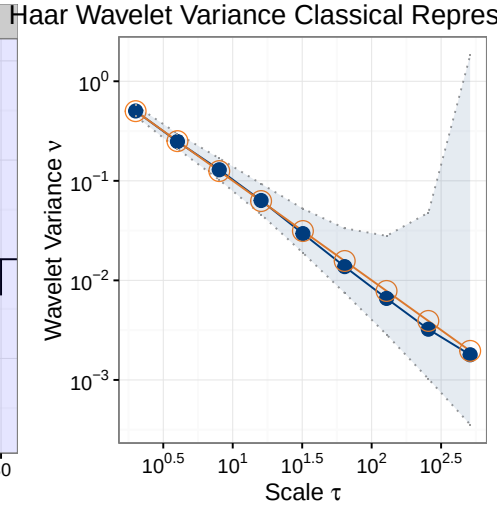


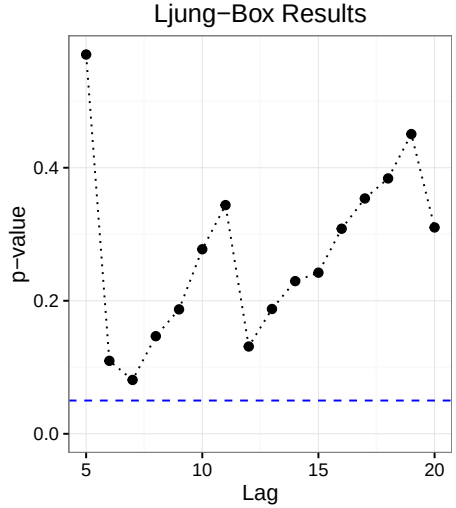












Model Selection

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:

$$\begin{aligned} \text{AIC} &= -2\log L + 2k \\ \text{BIC} &= -2\log L + \log(n)k, \\ \text{HQ} &= -2\log L + 2\log(\log(n))k. \end{aligned}$$

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:

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

the expectations $E[\cdot]$ and $E_0[\cdot]$ denote an 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:

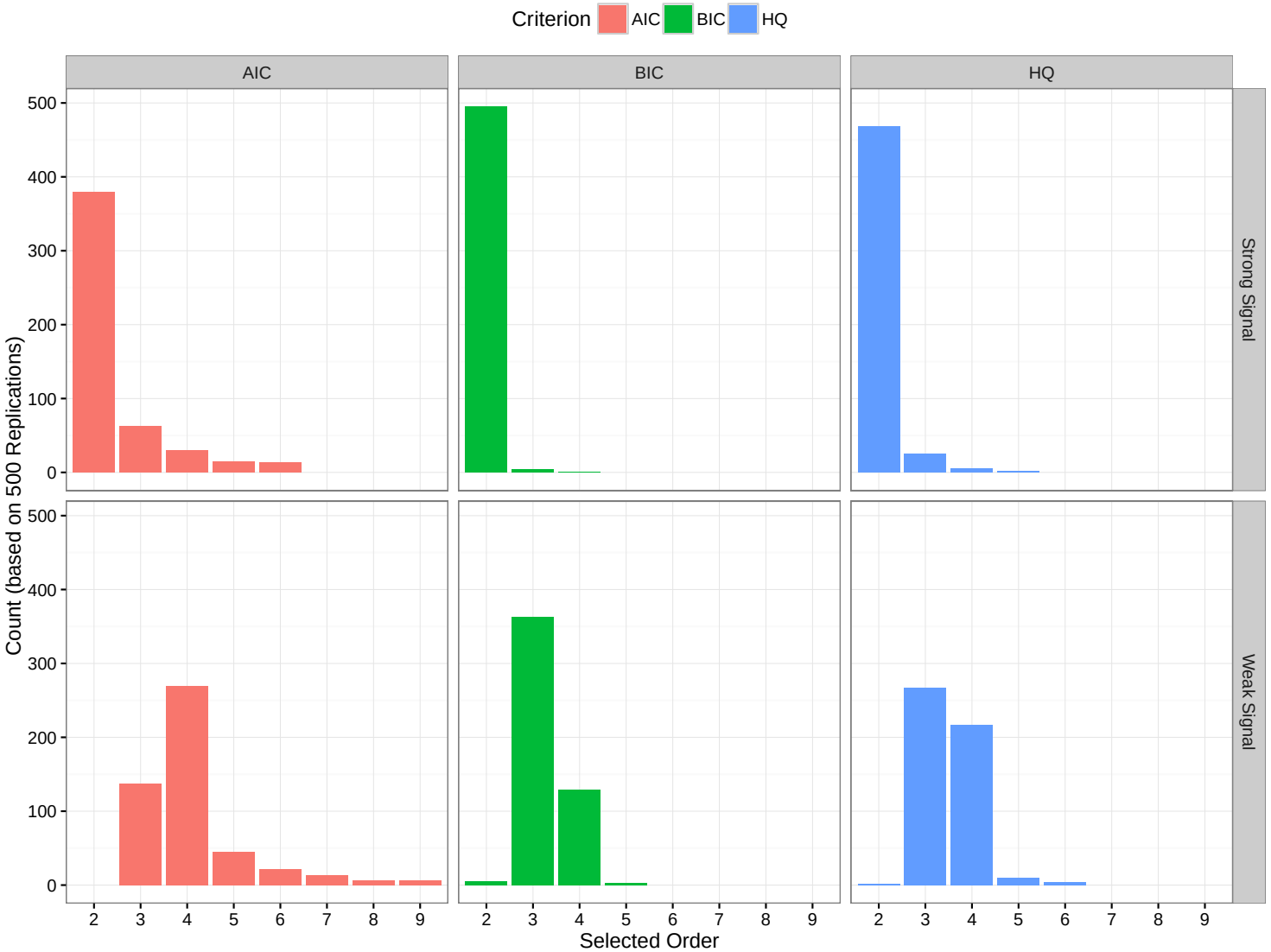
$$X_t = 0.6X_{t-1} + 0.3X_{t-2} + W_t$$
$$X_t = 0.5X_{t-1} + 0.25X_{t-2} + 0.125X_{t-3} + 0.0625X_{t-4} + W_t$$

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 $B = 500$.

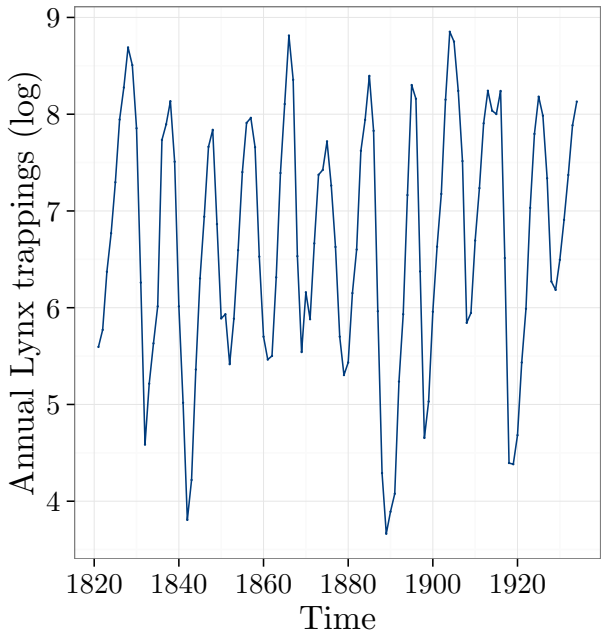
Simulation Study

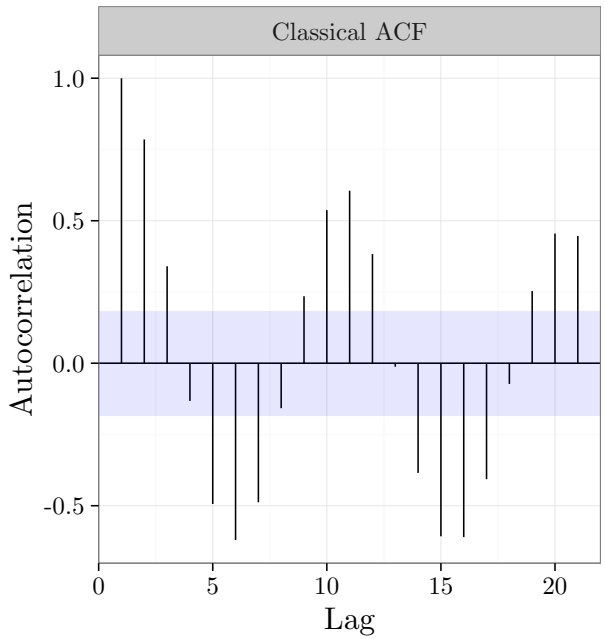


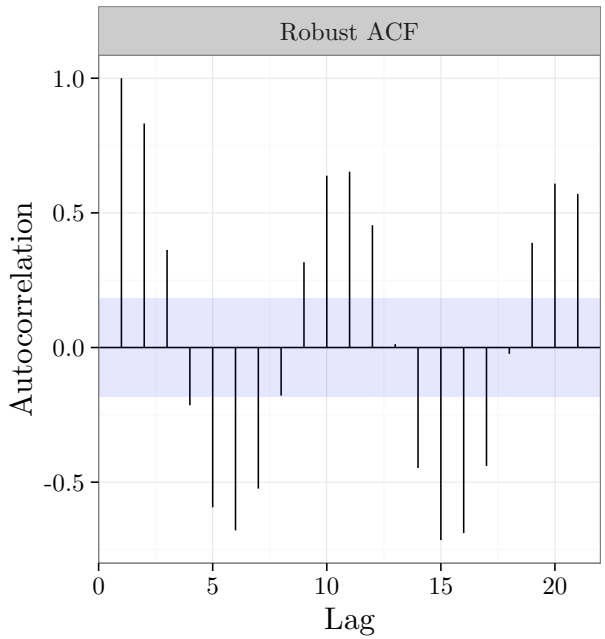
Application to Real Data: Lynx Dataset

Background:

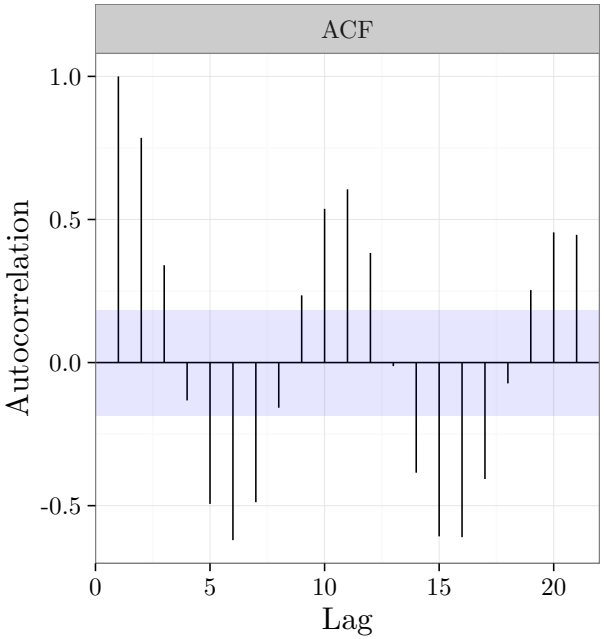
We consider the annual numbers of lynx trappings for 1821-1934 in Canada.

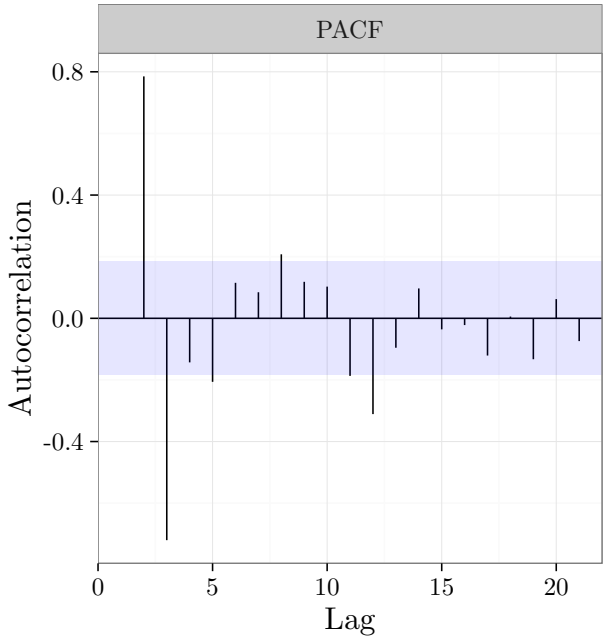


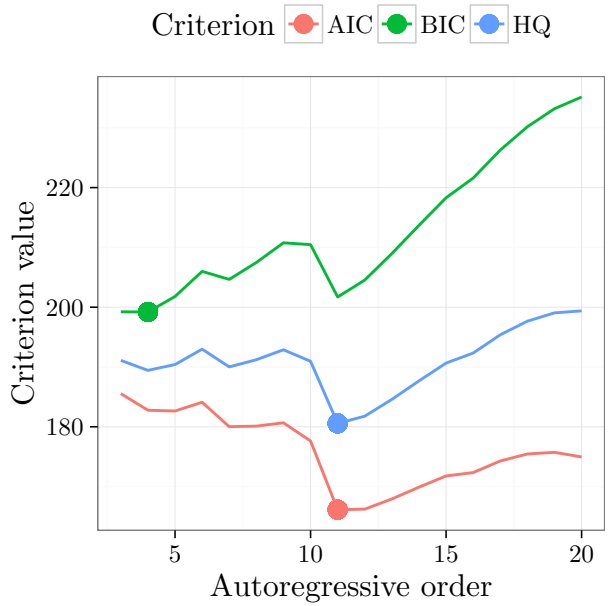




Application to Real Data: Lynx Dataset

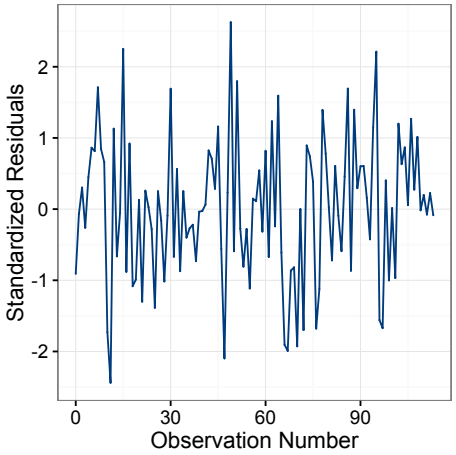


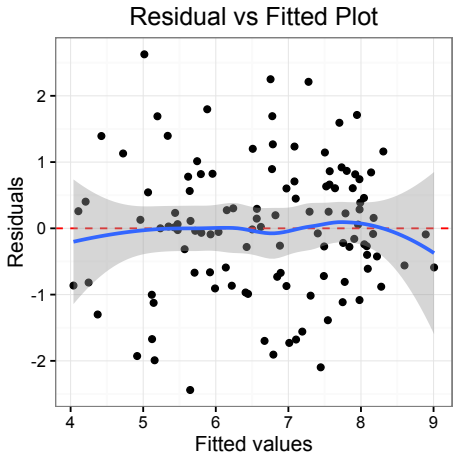


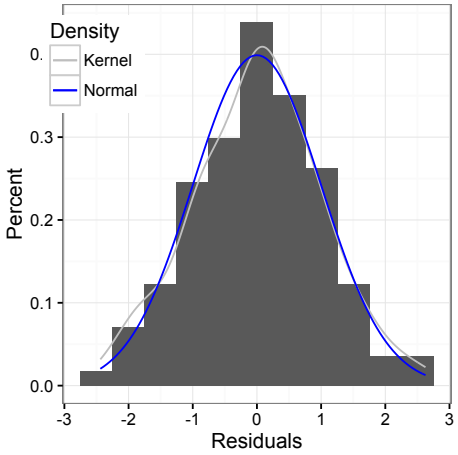


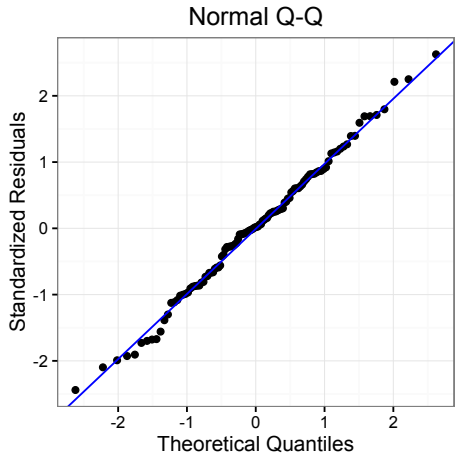
According to both AIC and HQ criterion an AR(11) seems appropriate.

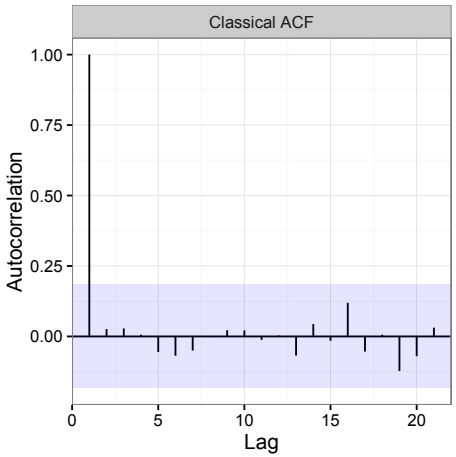
Application to Real Data: Lynx Dataset

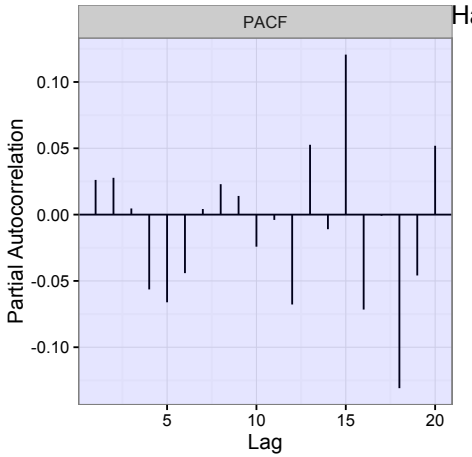


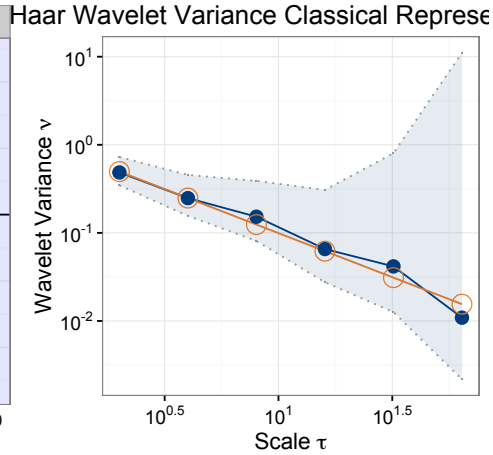


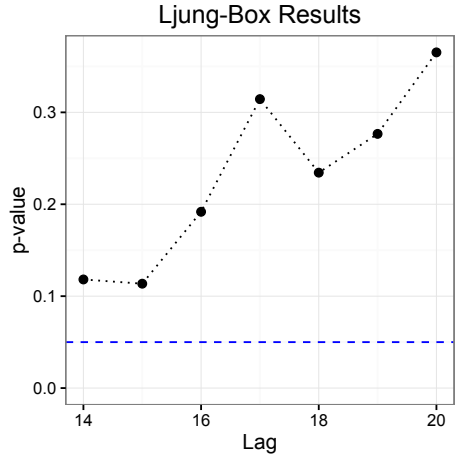




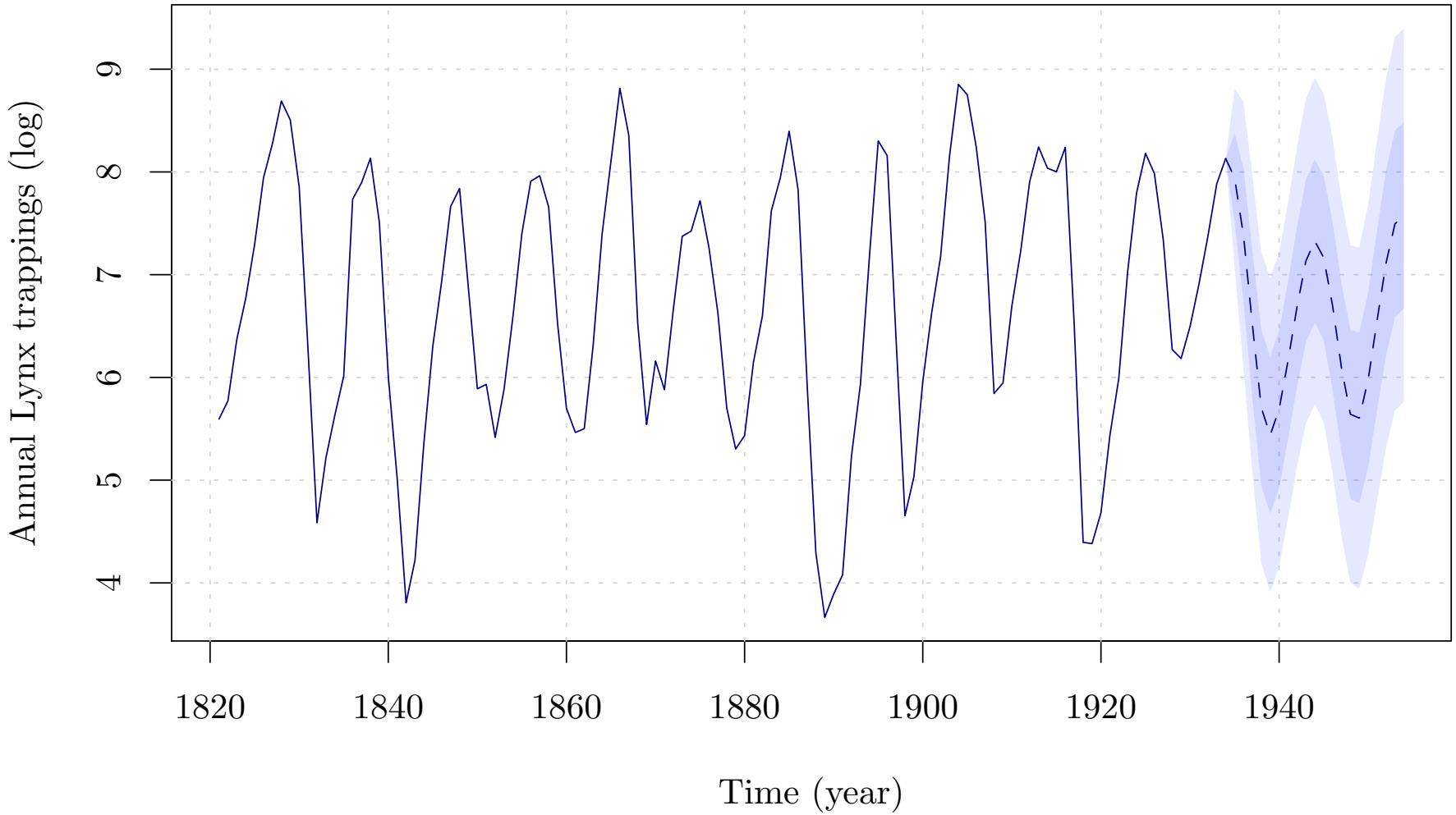








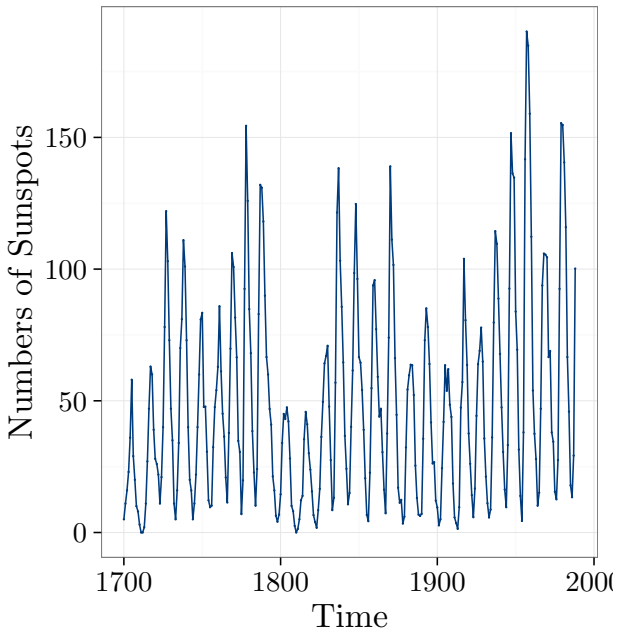
Application to Real Data: Lynx Dataset

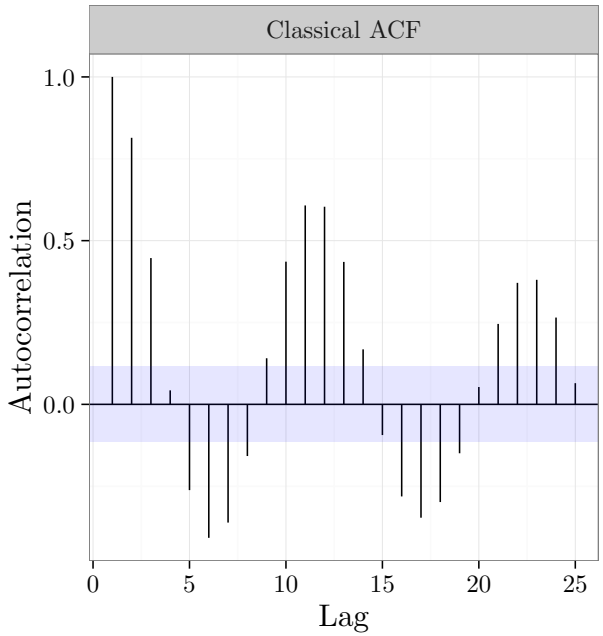


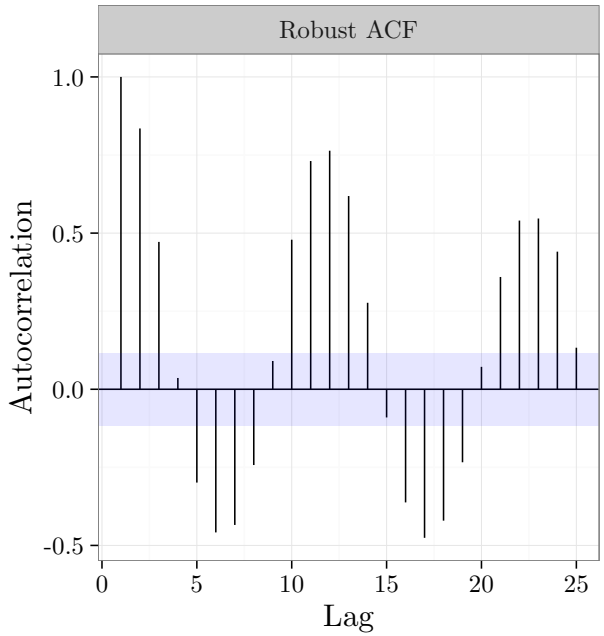
Application to Real Data: Sunspot Dataset

Background:

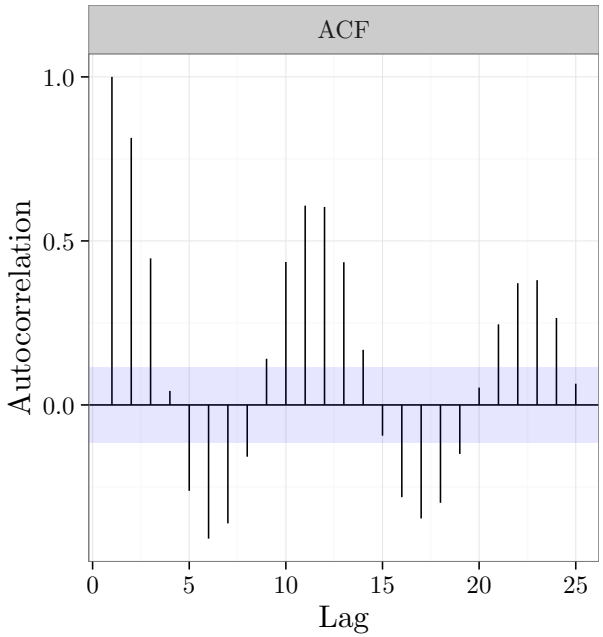
We consider the yearly numbers of sunspots from 1700 to 1988.

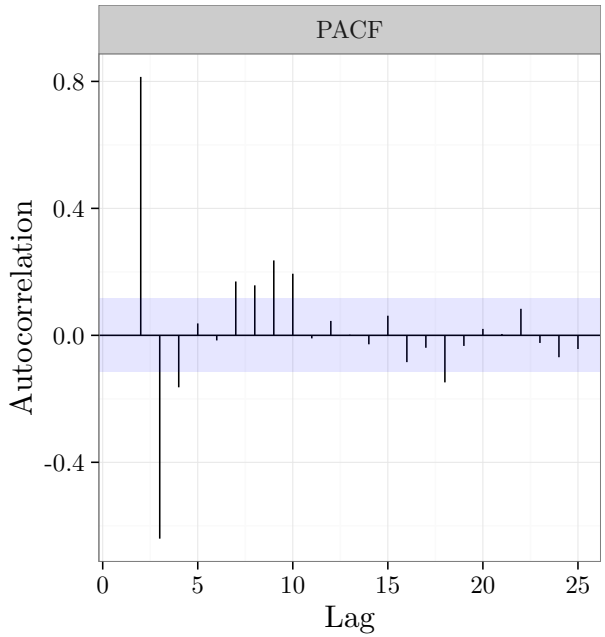


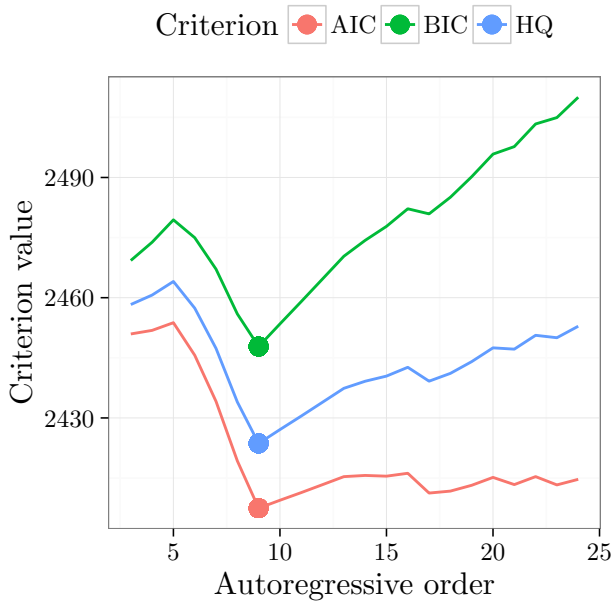




Application to Real Data: Sunspot Dataset

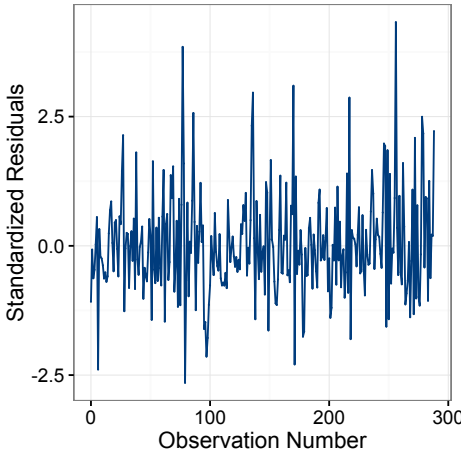


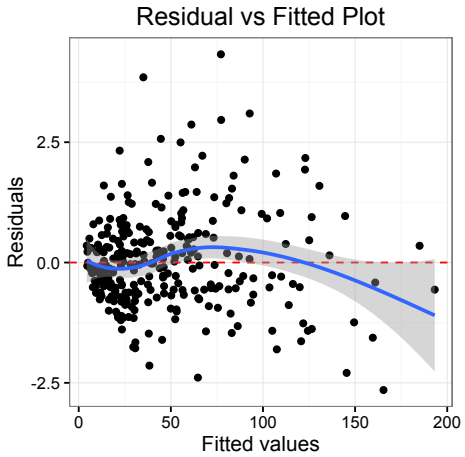


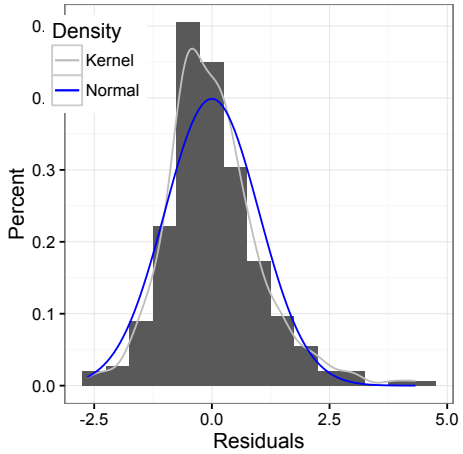


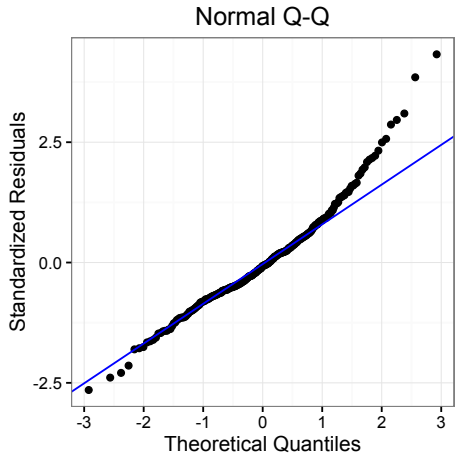
All criteria indicate that an AR(9) is appropriate.

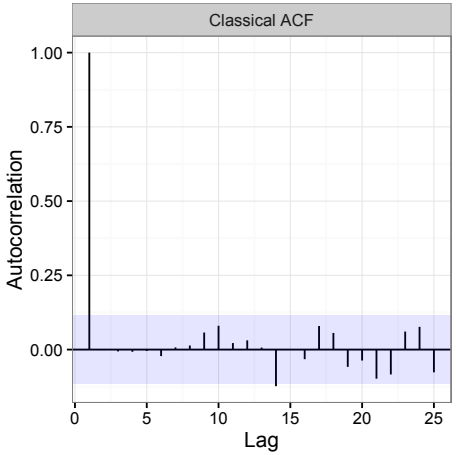
Application to Real Data: Sunspot Dataset

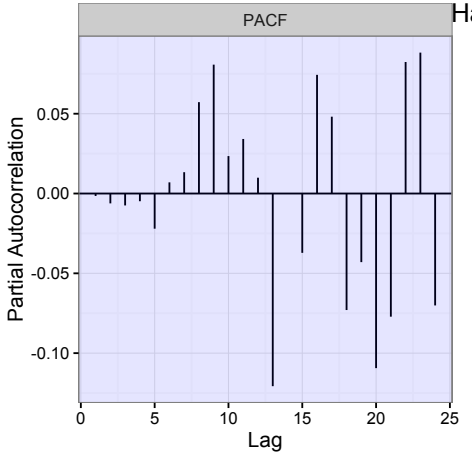


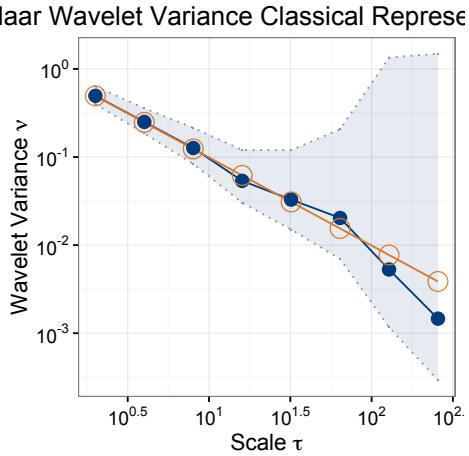


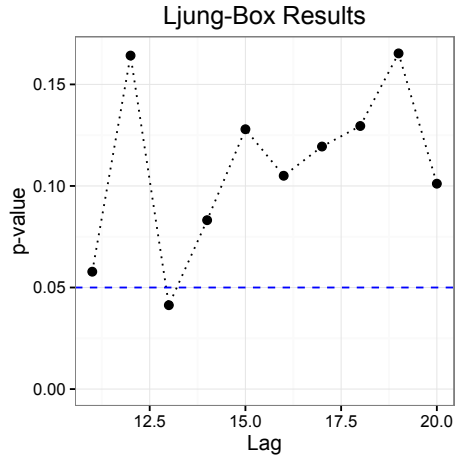




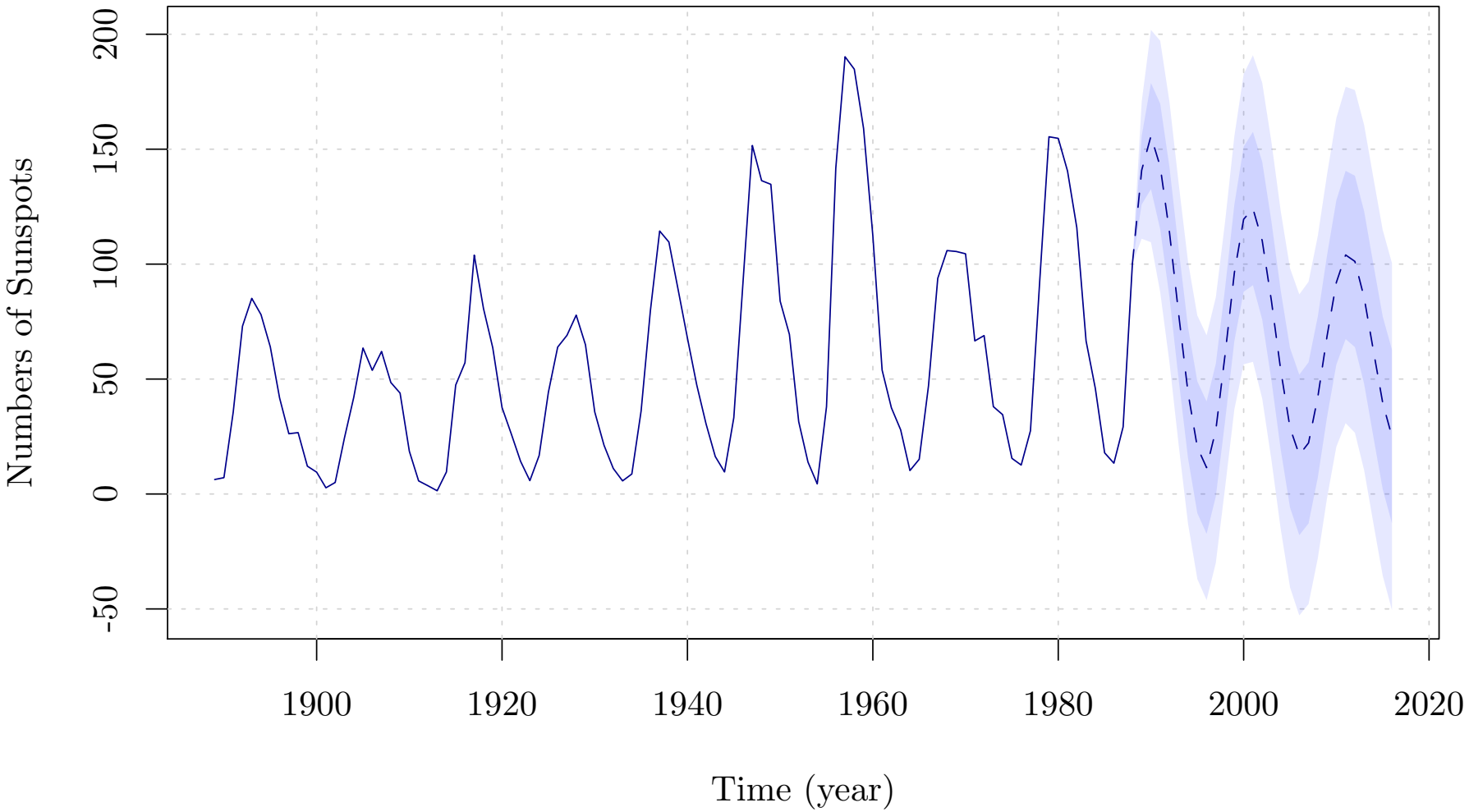








Application to Real Data: Sunspots Dataset



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.

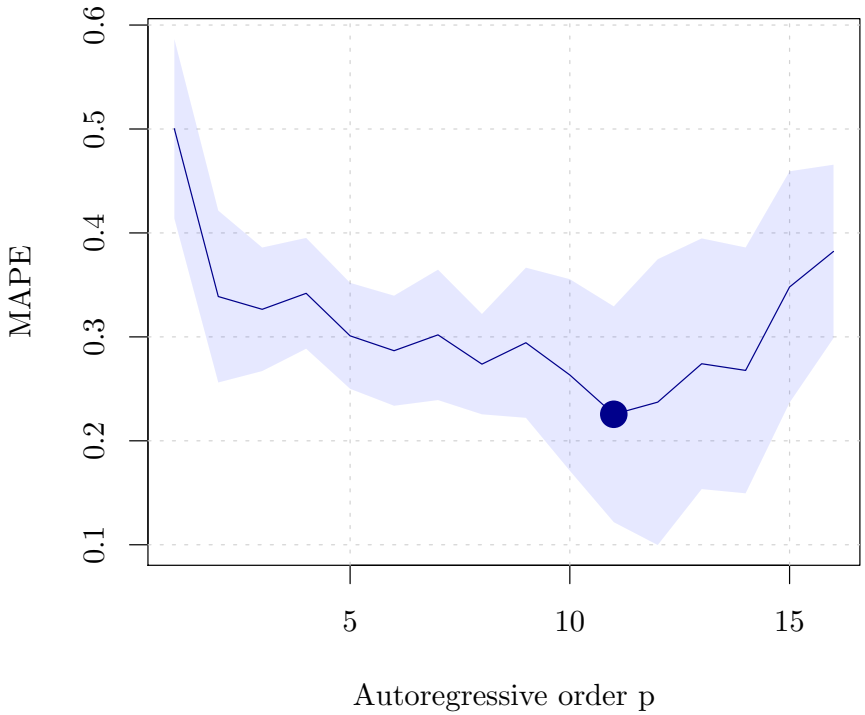
Measuring Forecast Accuracy

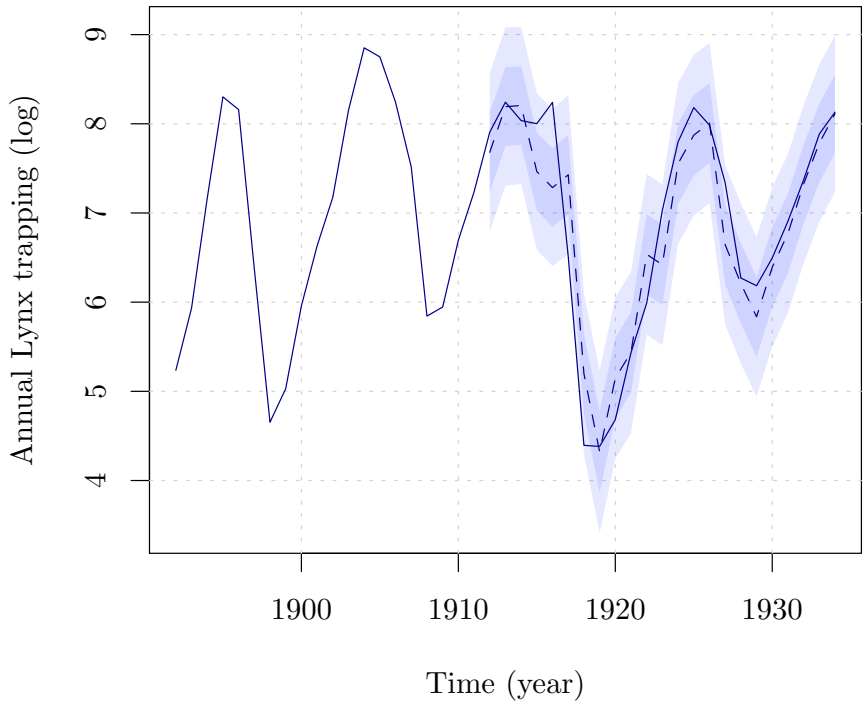
- Using the vector of differences between your forecast and the actual values of the time series, we can compute different metrics.
- Let $\hat{X}_h^{(1:j)}$ and obtain your prediction for the observation X_{j+h} based on a model whose parameters were estimated on the sample $X_1, ..., X_j$, then the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

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

- 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...

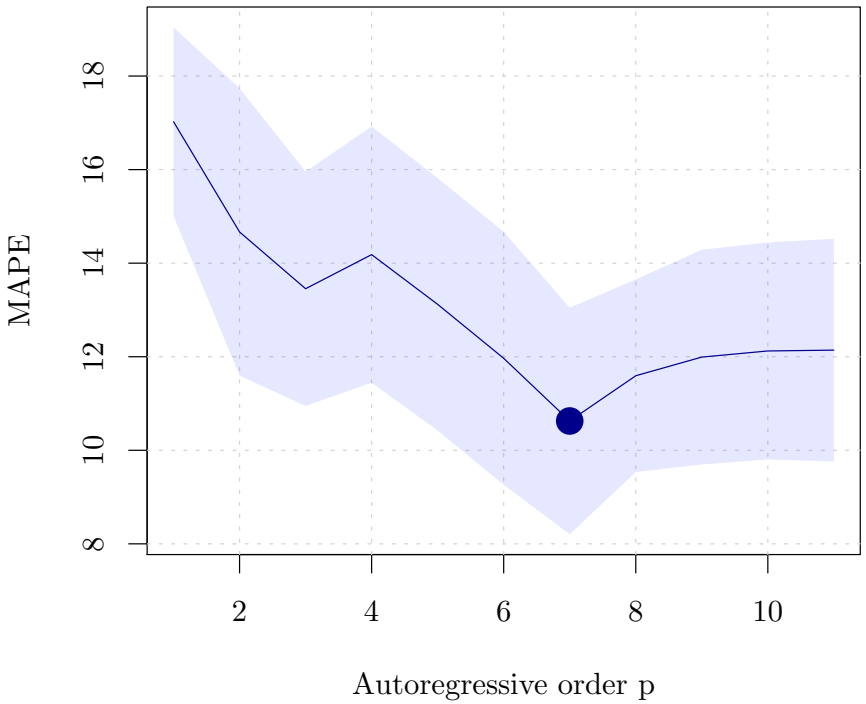
MAPE for Lynx Dataset

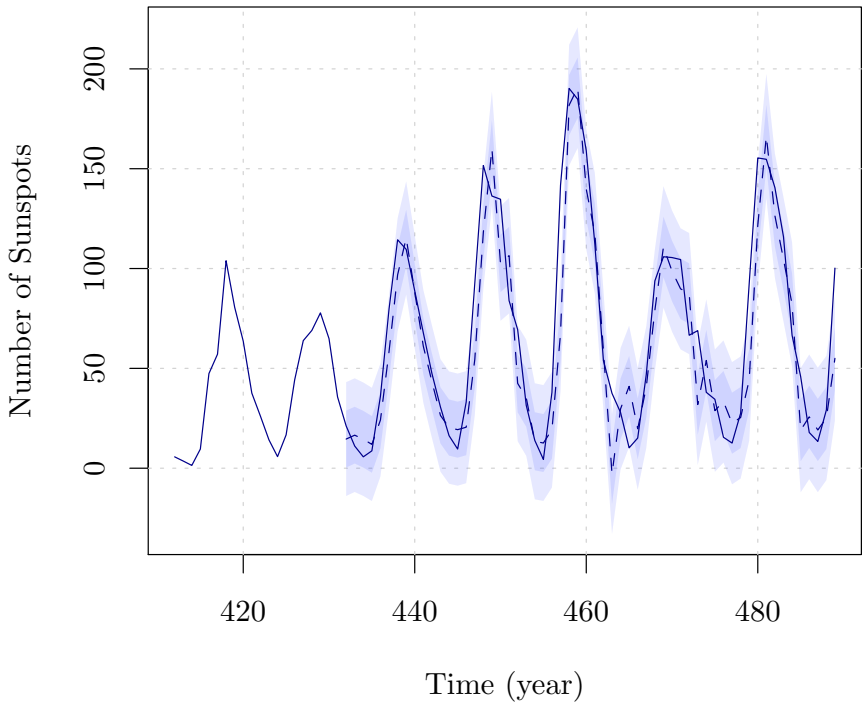




This confirms that an AR(12) is a "good" choice.

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

Definition

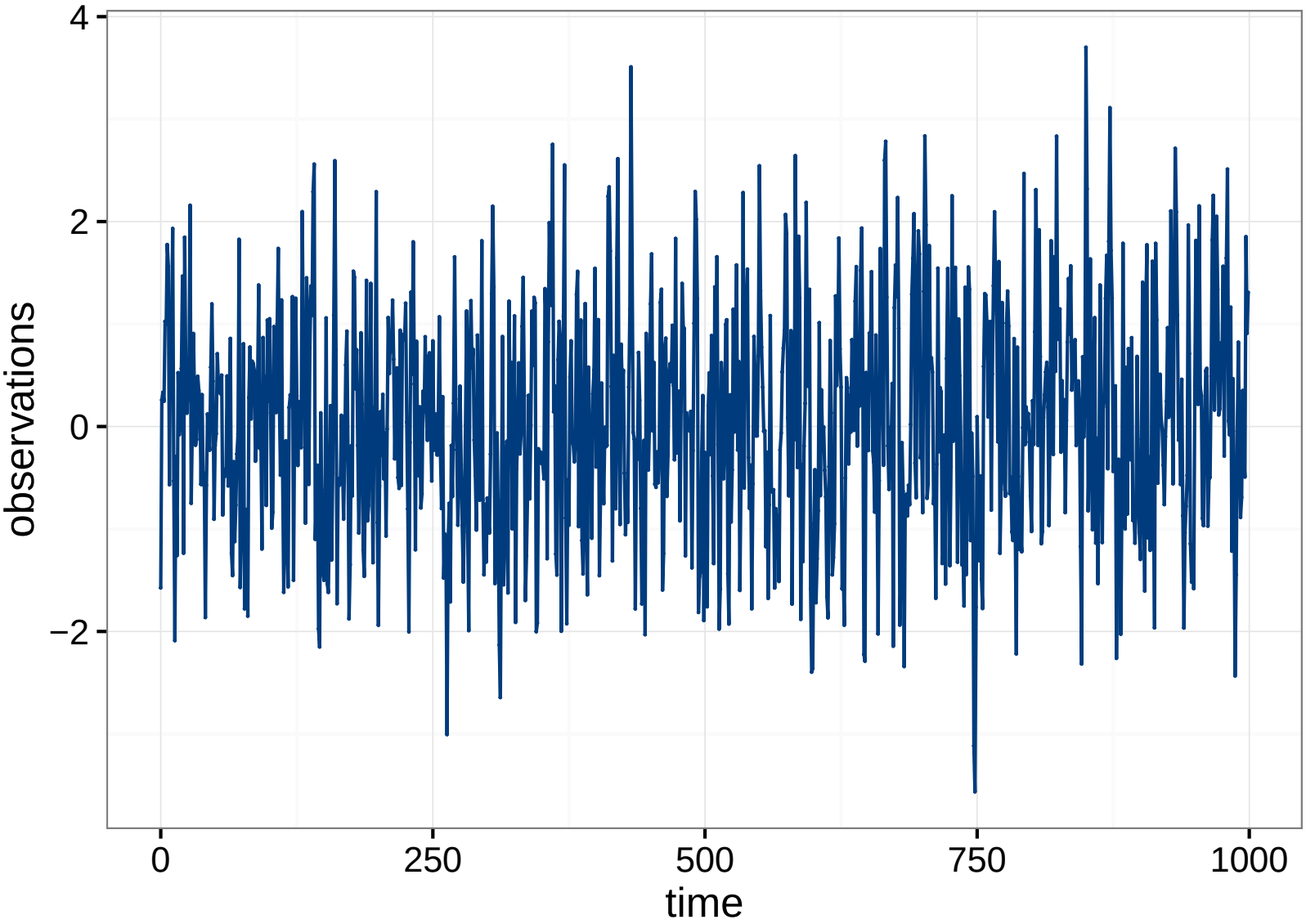
The *moving average operator* is defined

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

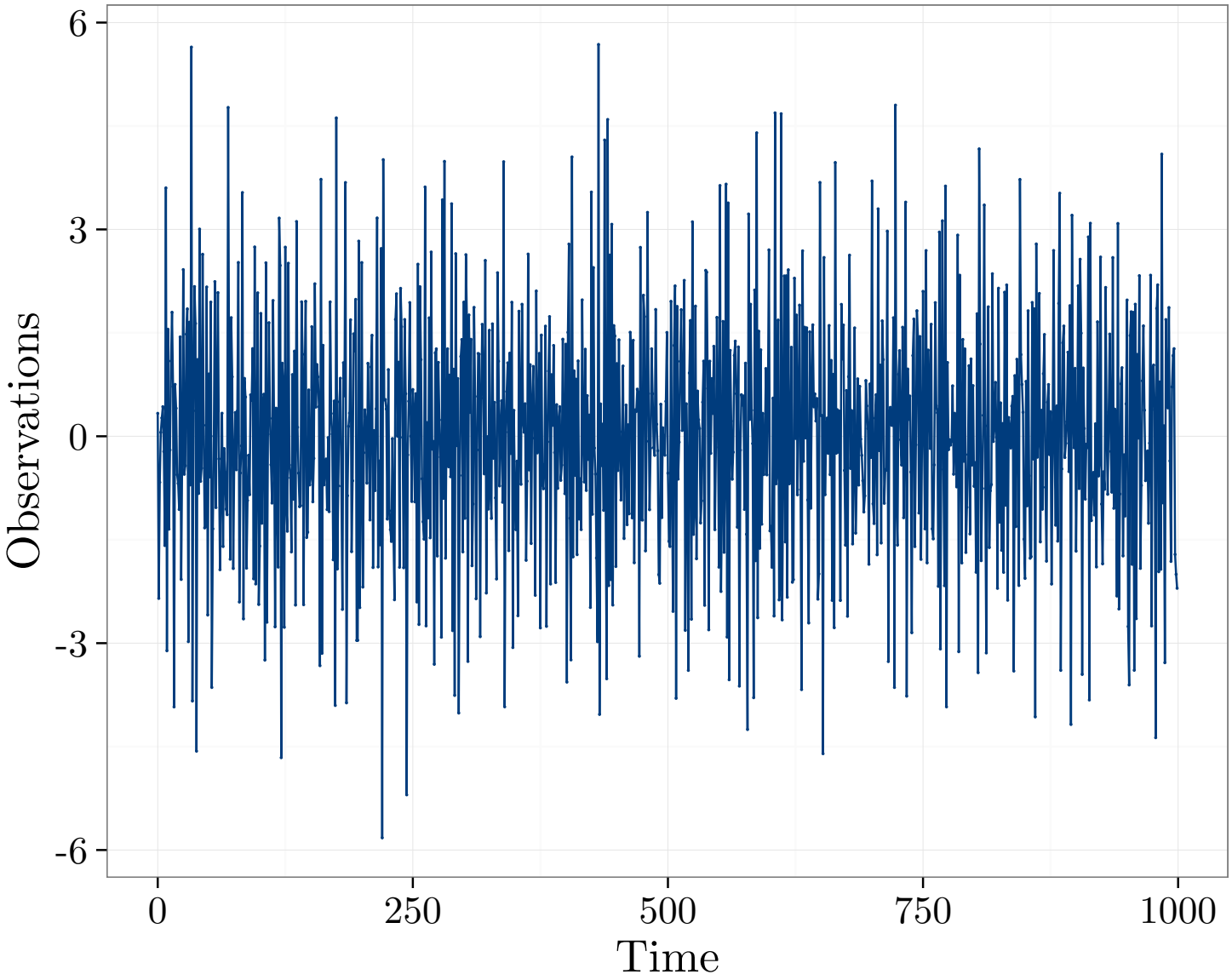
so an MA(q) can be expressed as

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

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:

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

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:

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

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

- By considering only invertible MA processes we ensure that they are **unique**.
- Invertibility for MA(q) plays a less 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 considered:

$$\mathbb{E}\left[X_{T+1} \mid X_T, ..., X_1, W_T, ...W_1\right] = \theta W_T,$$

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 of these techniques but the techniques are implemented in most statistical software.

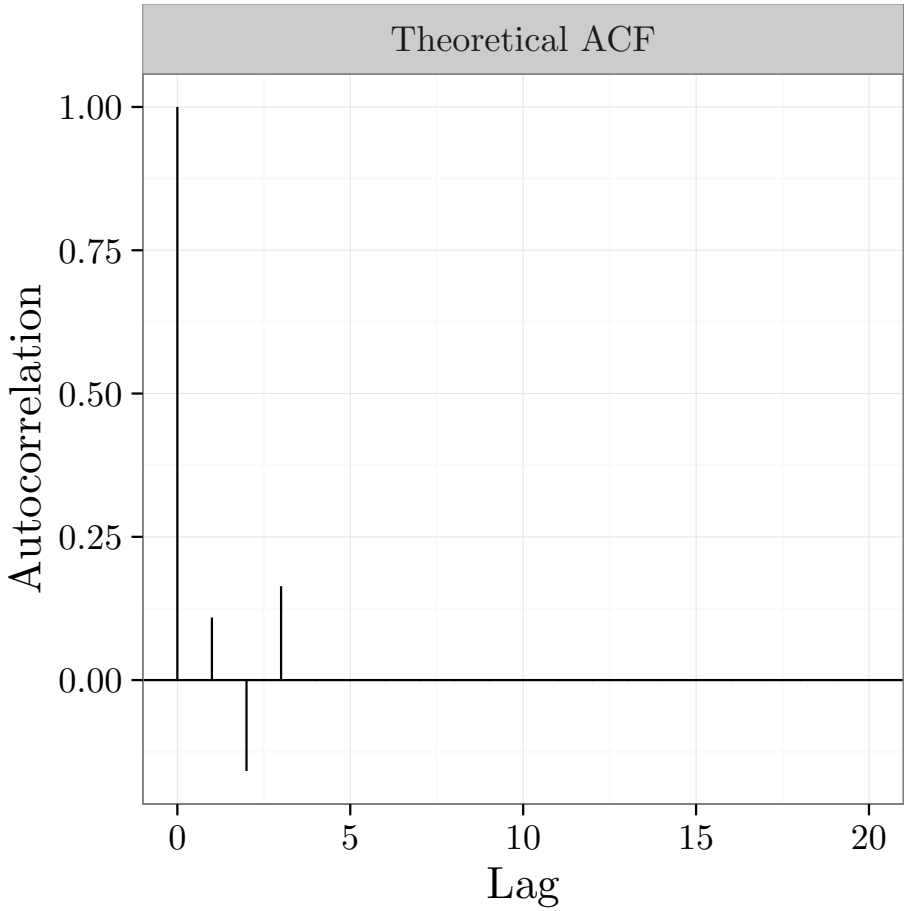
Theoretical ACF and PACF of MA(q) models

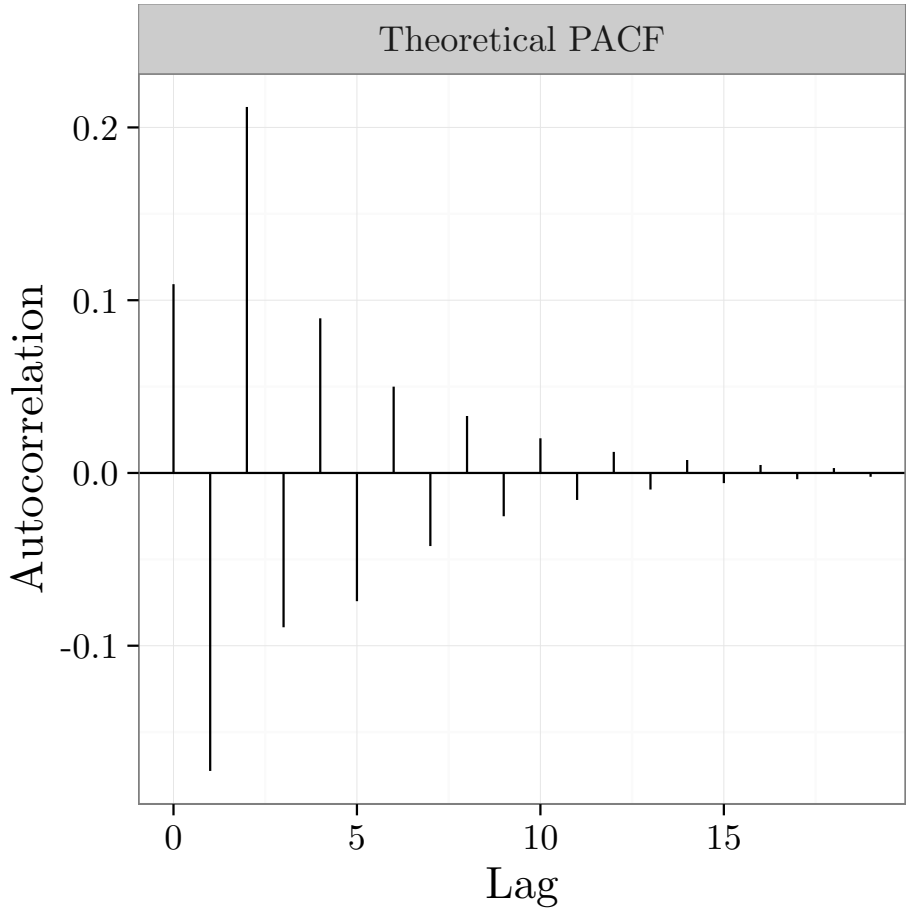
- In general, the PACF of an MA(q) is a complicated function of the parameters but $\phi(h)$ tends to dampen, in a sinusoidal fashion, exponentially fast to zero as h increases.
- The ACF of an MA(q) is non-zero for the first q lags and zero for $h > q$.
- Example: consider the following models:

$$X_t = -0.5W_{t-1} + W_t$$
$$X_t = -0.9W_{t-1} - 0.5W_{t-2} + W_t$$
$$X_t = 0.7W_{t-1} - 0.5W_{t-2} + 0.3W_{t-3} + W_t$$
$$X_t = -0.1W_{t-1} + 0.1W_{t-2} - 0.1W_{t-3} + 0.9W_{t-4} + W_t$$

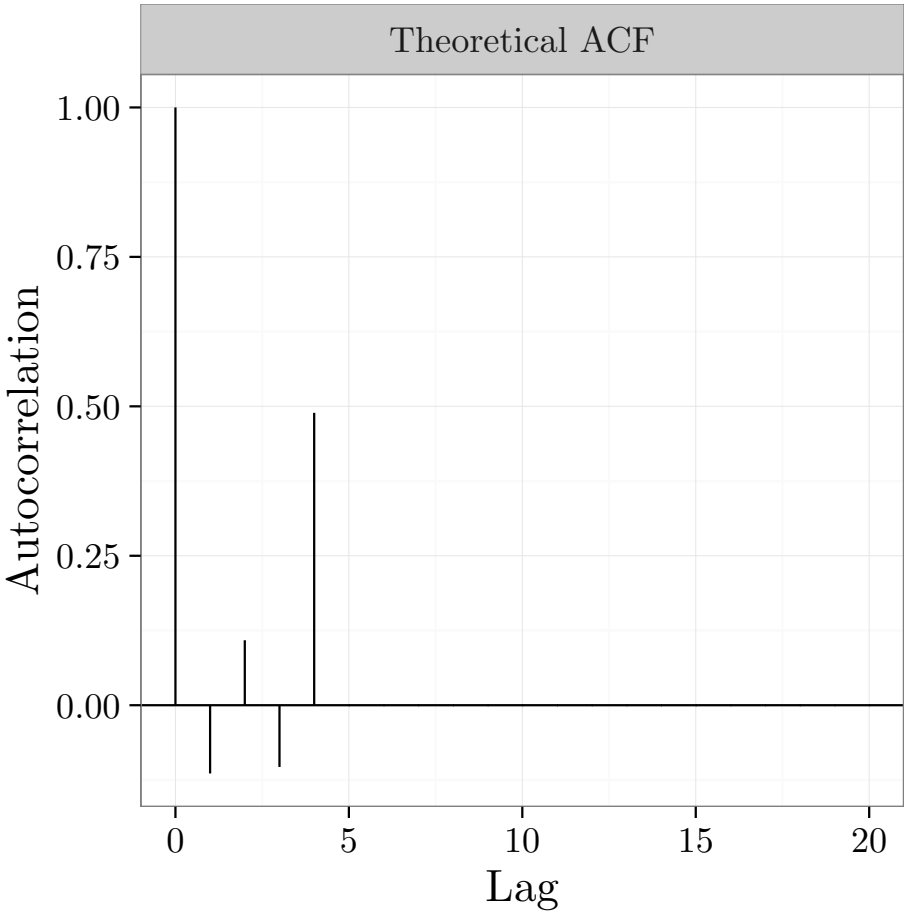
Can we recognize each model given their theoretical ACF/PACF? Let's try...

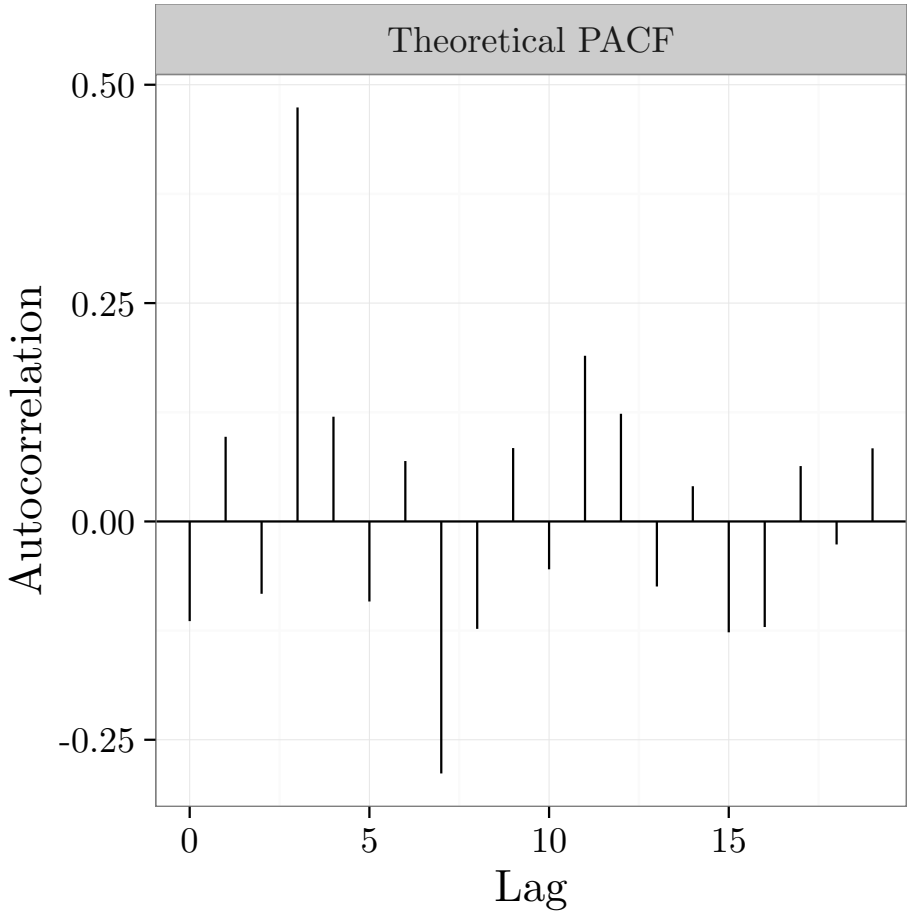
Theoretical ACF and PACF example



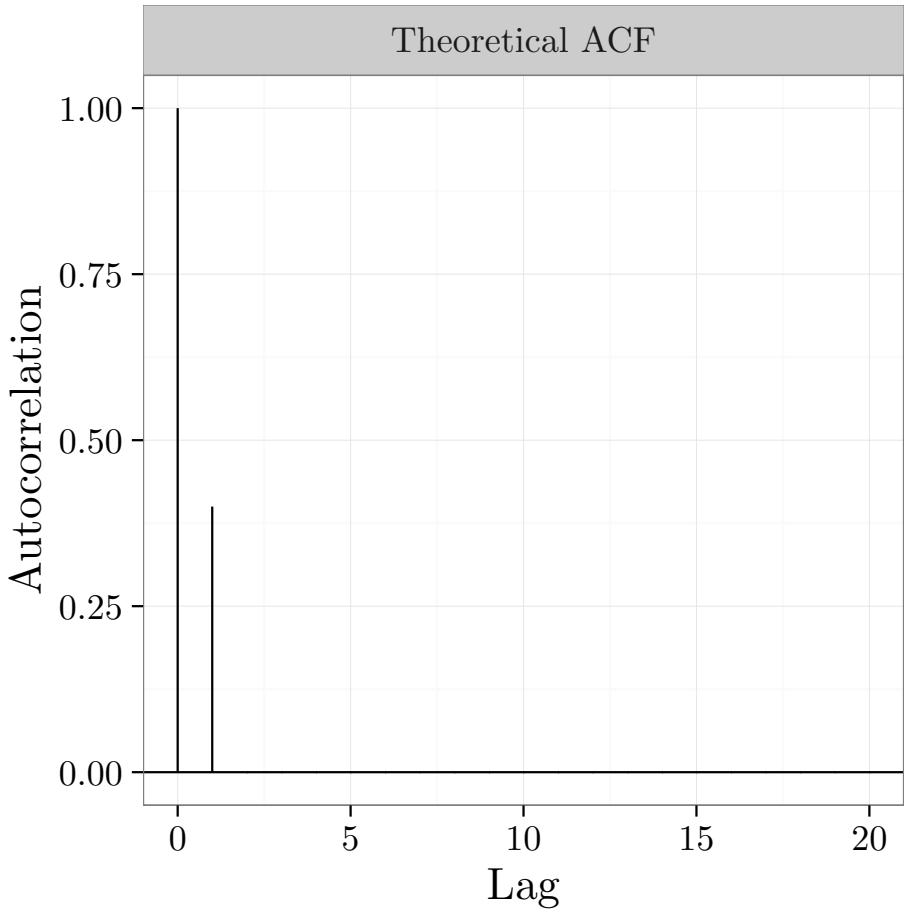


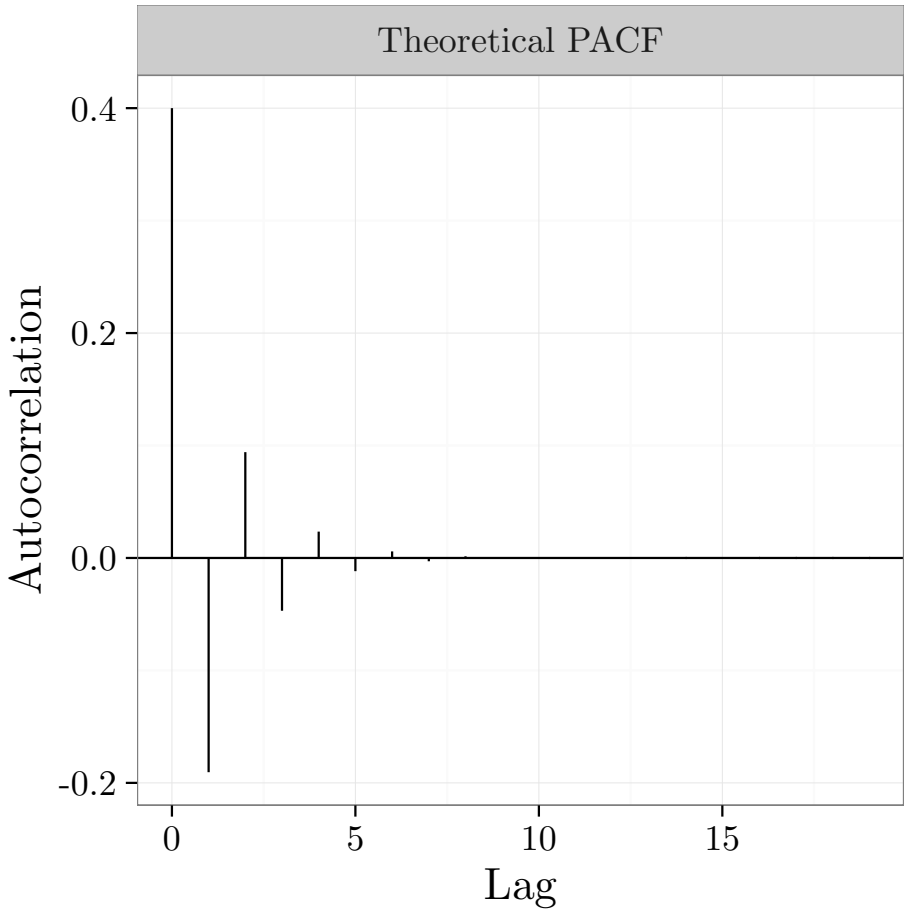
Theoretical ACF and PACF example



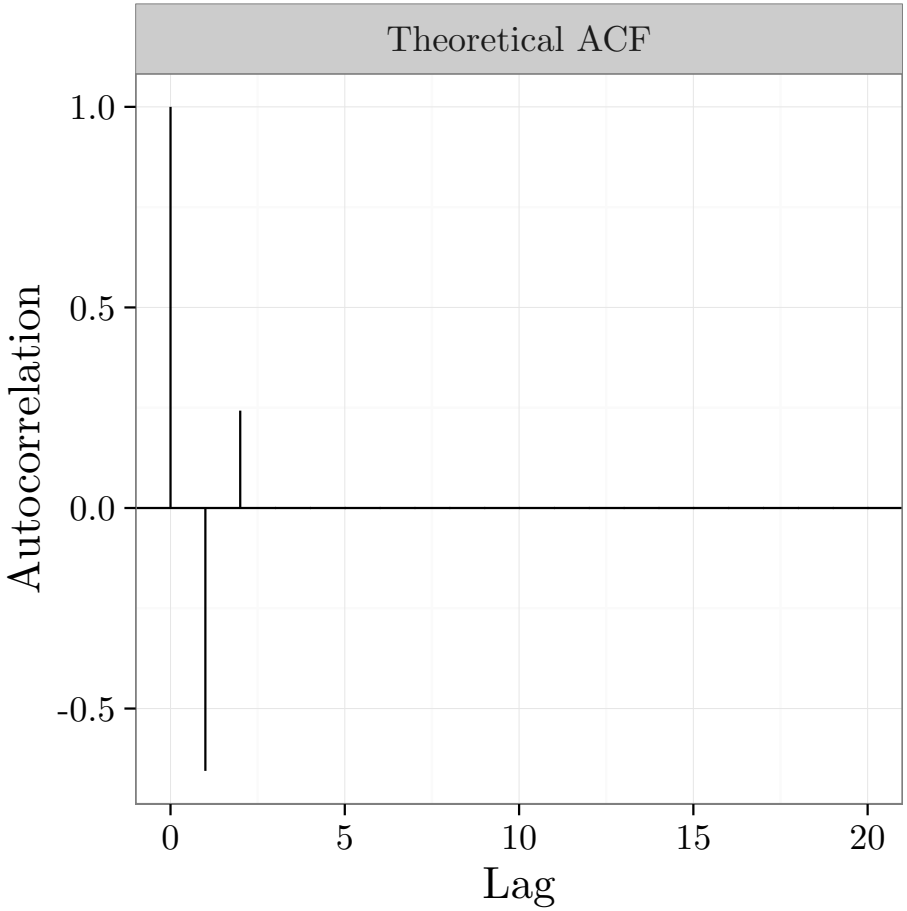


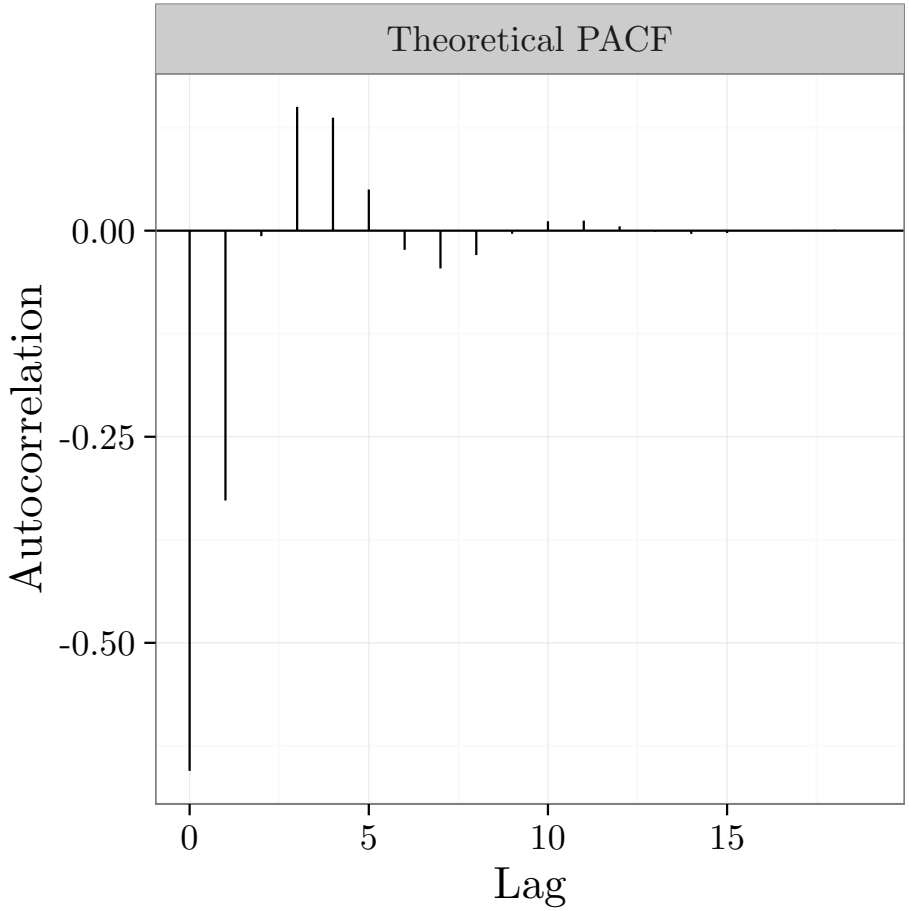
Theoretical ACF and PACF example



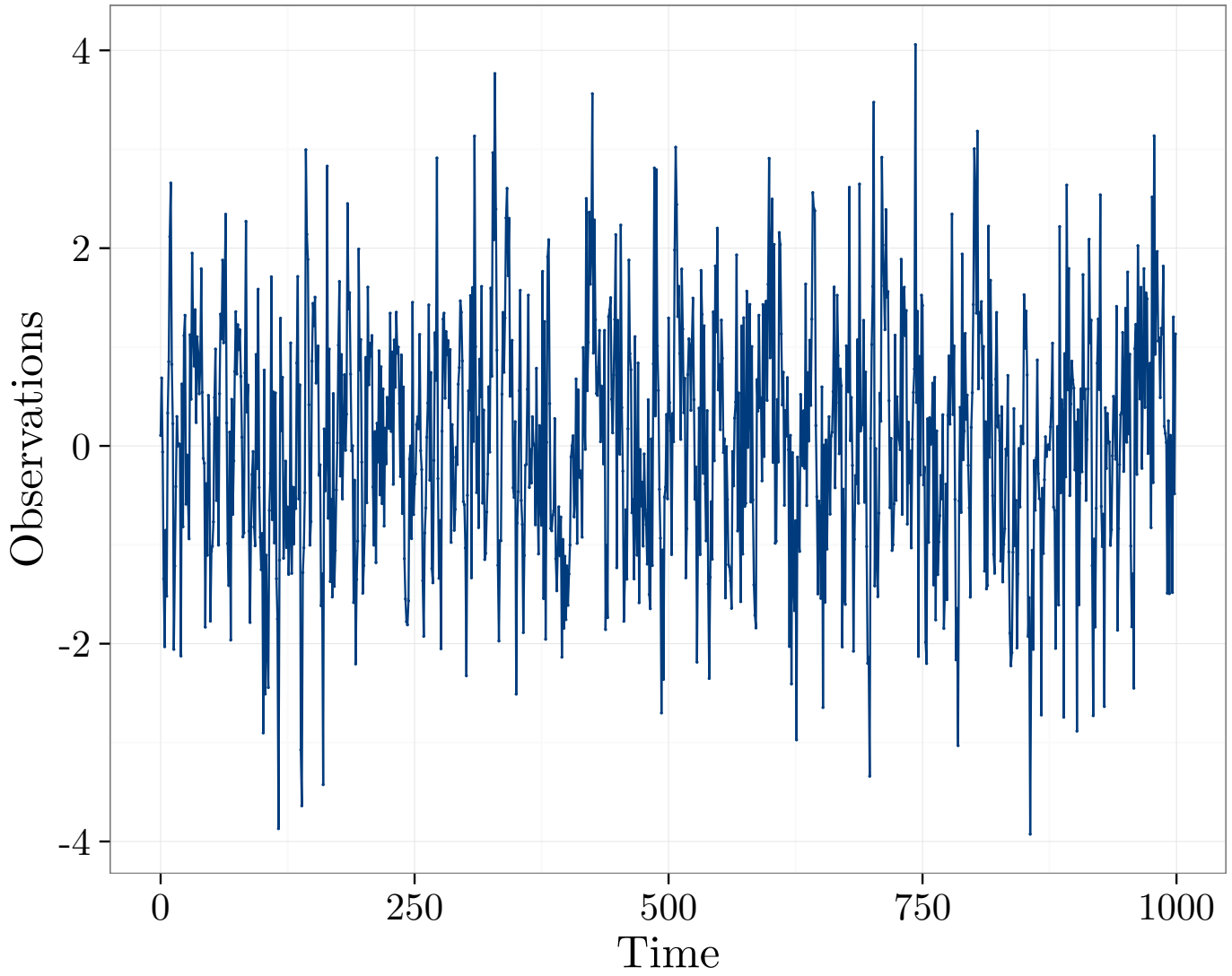


Theoretical ACF and PACF example

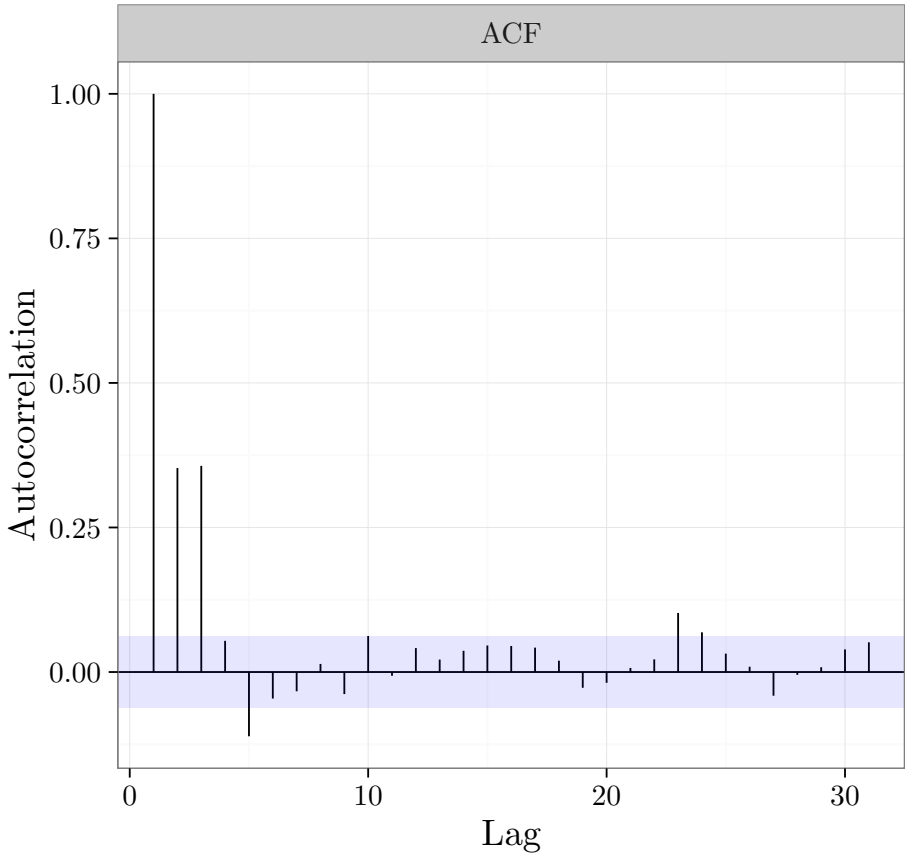


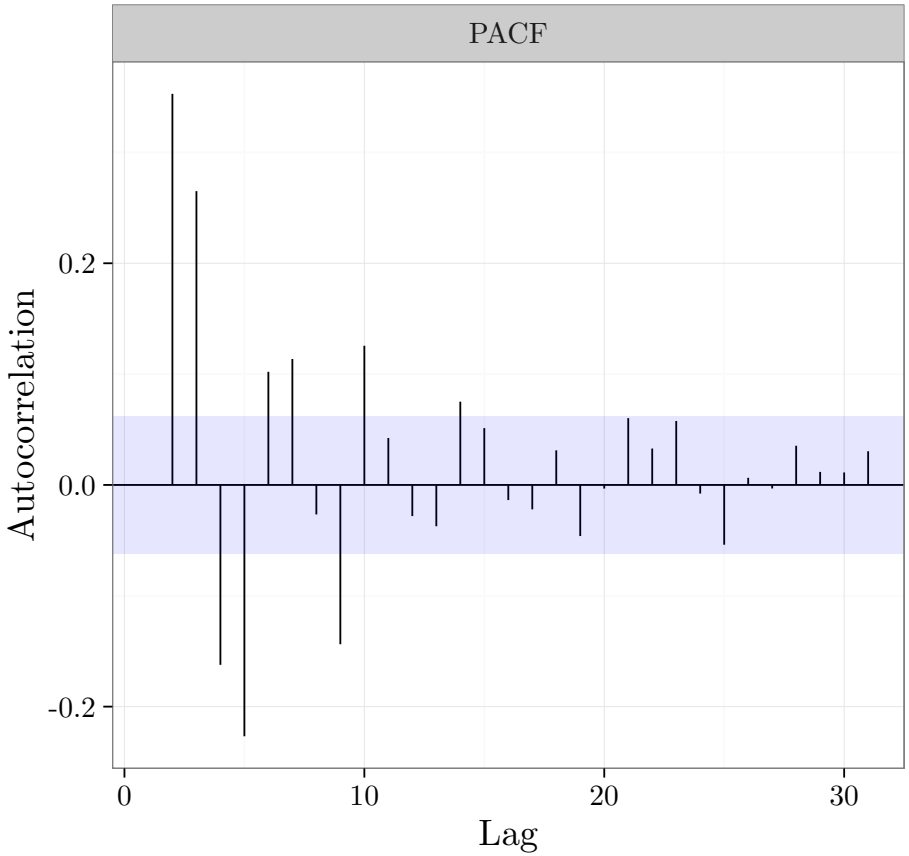


A Simulated Example: Order Identification

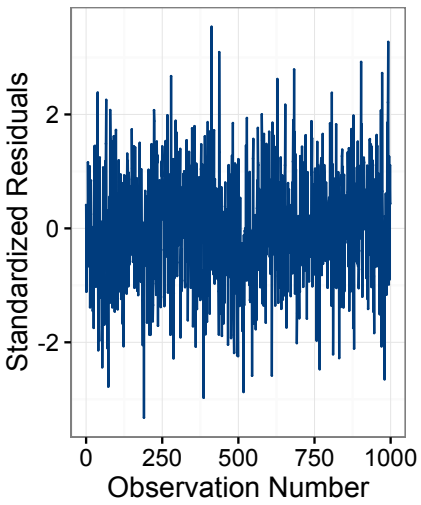


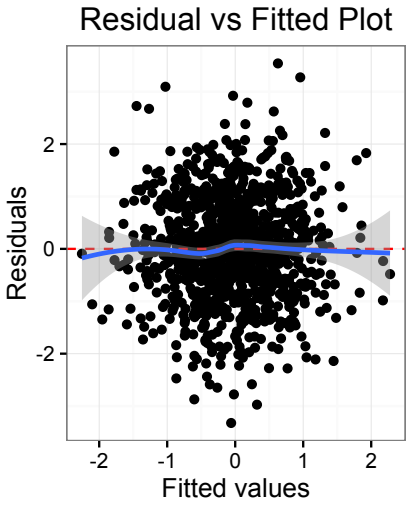
A Simulated Example: ACF/PACF Graphs

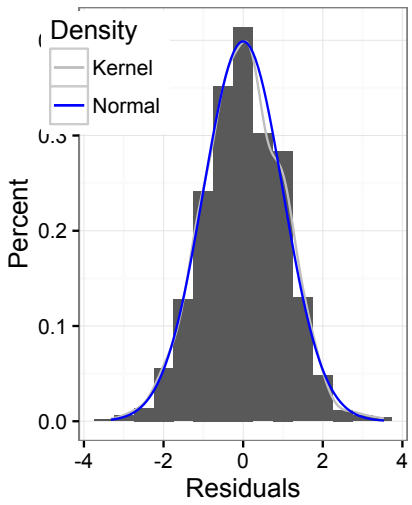


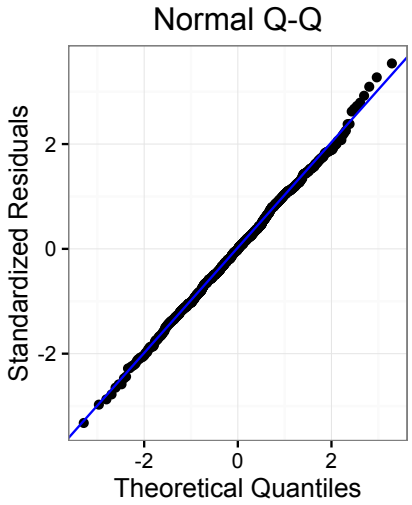


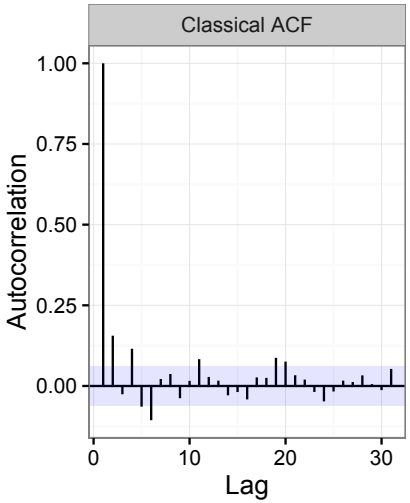
Diagnostic for MA(2)

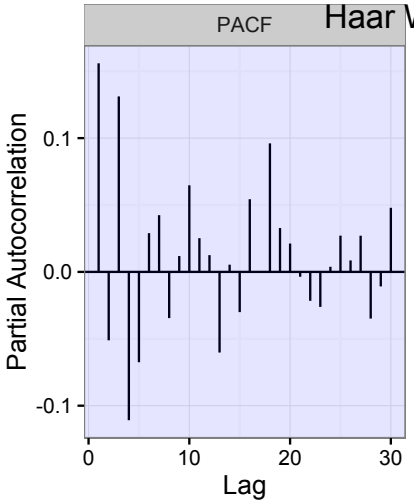


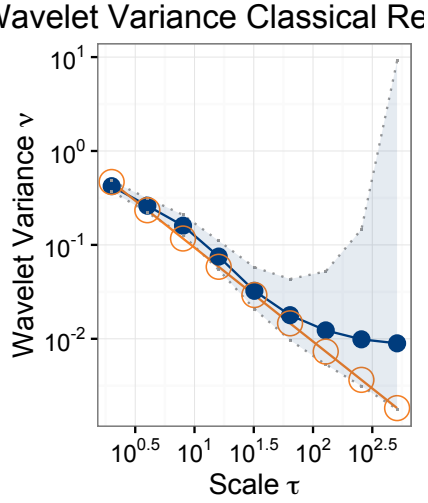


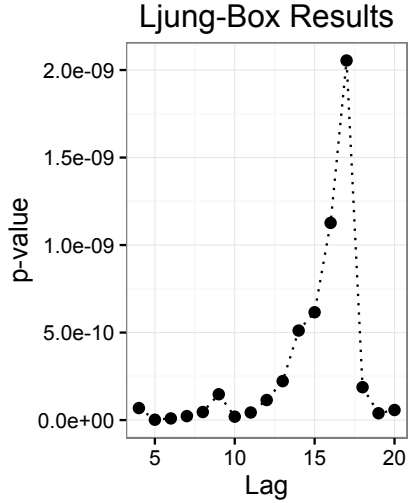




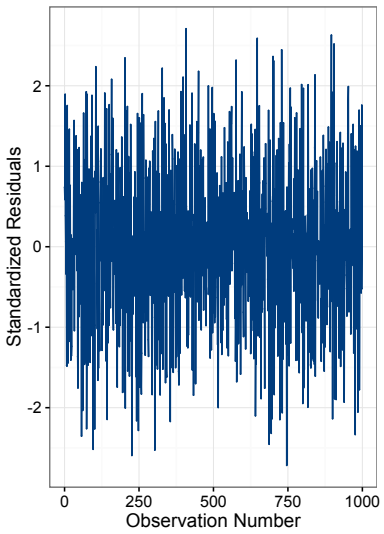


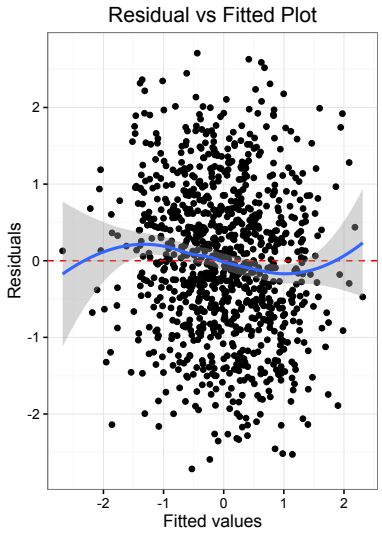


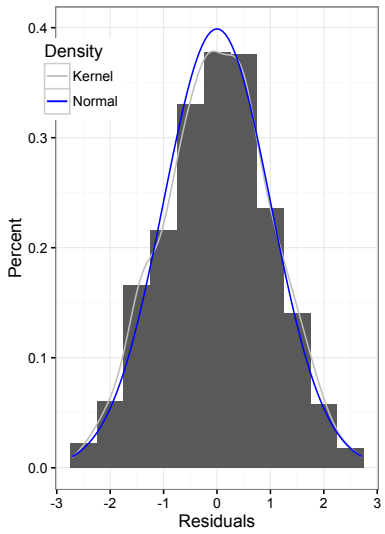


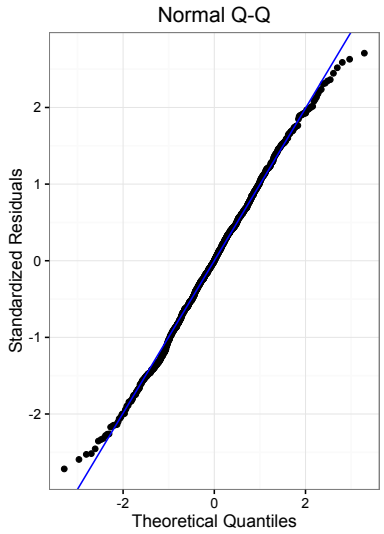


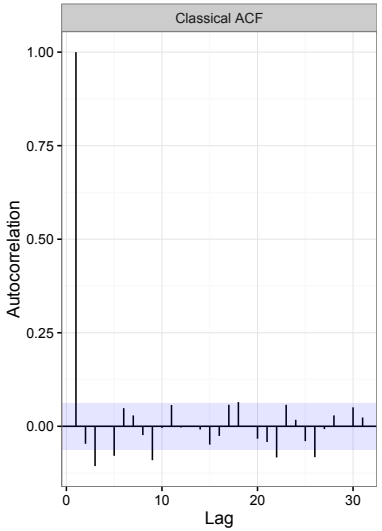
Diagnostic for MA(3)

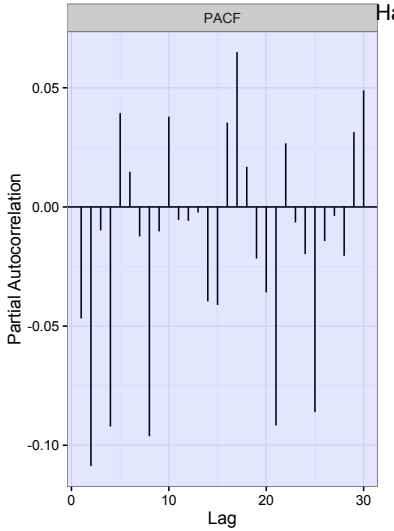


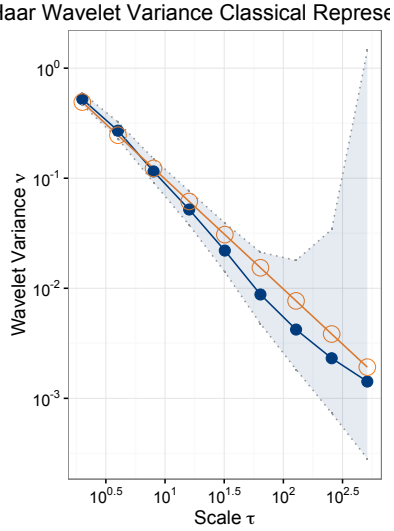


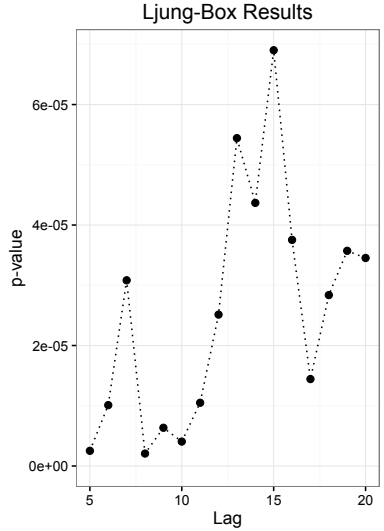




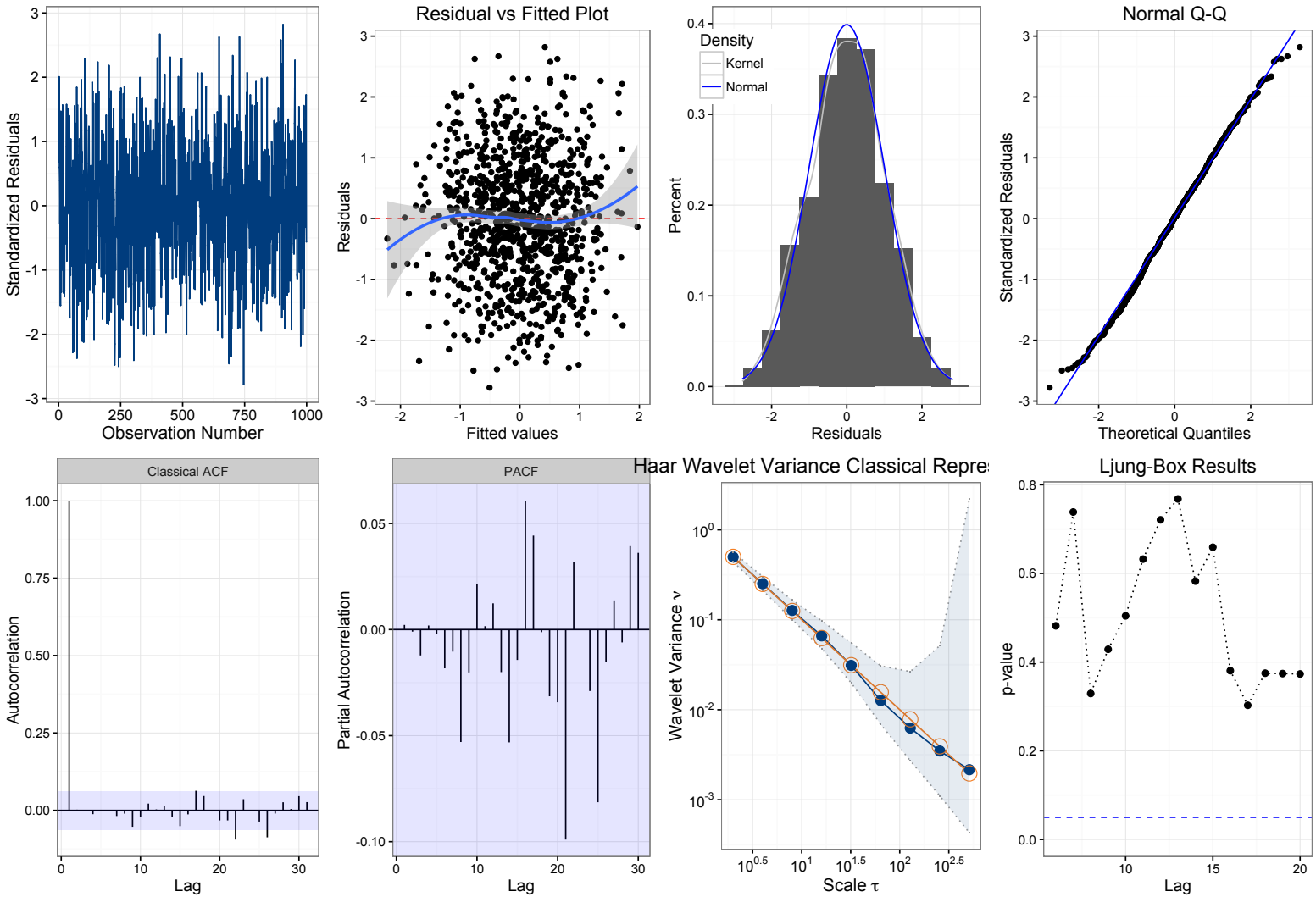








Diagnostic for MA(4)



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:

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

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

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 ARMA$(p = 0, q)$ = MA(q) process.
3. If $q = 0$, then ARMA$(p, q = 0)$ = 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.
   - AR(p) models are always *invertible* and 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:

$$\begin{aligned} X_t &= W_t \\ X_t - 0.9X_{t-1} &= W_t - 0.9W_{t-1} \\ X_t &= 0.9X_{t-1} + W_t - 0.9W_{t-1}. \end{aligned}$$

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 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:

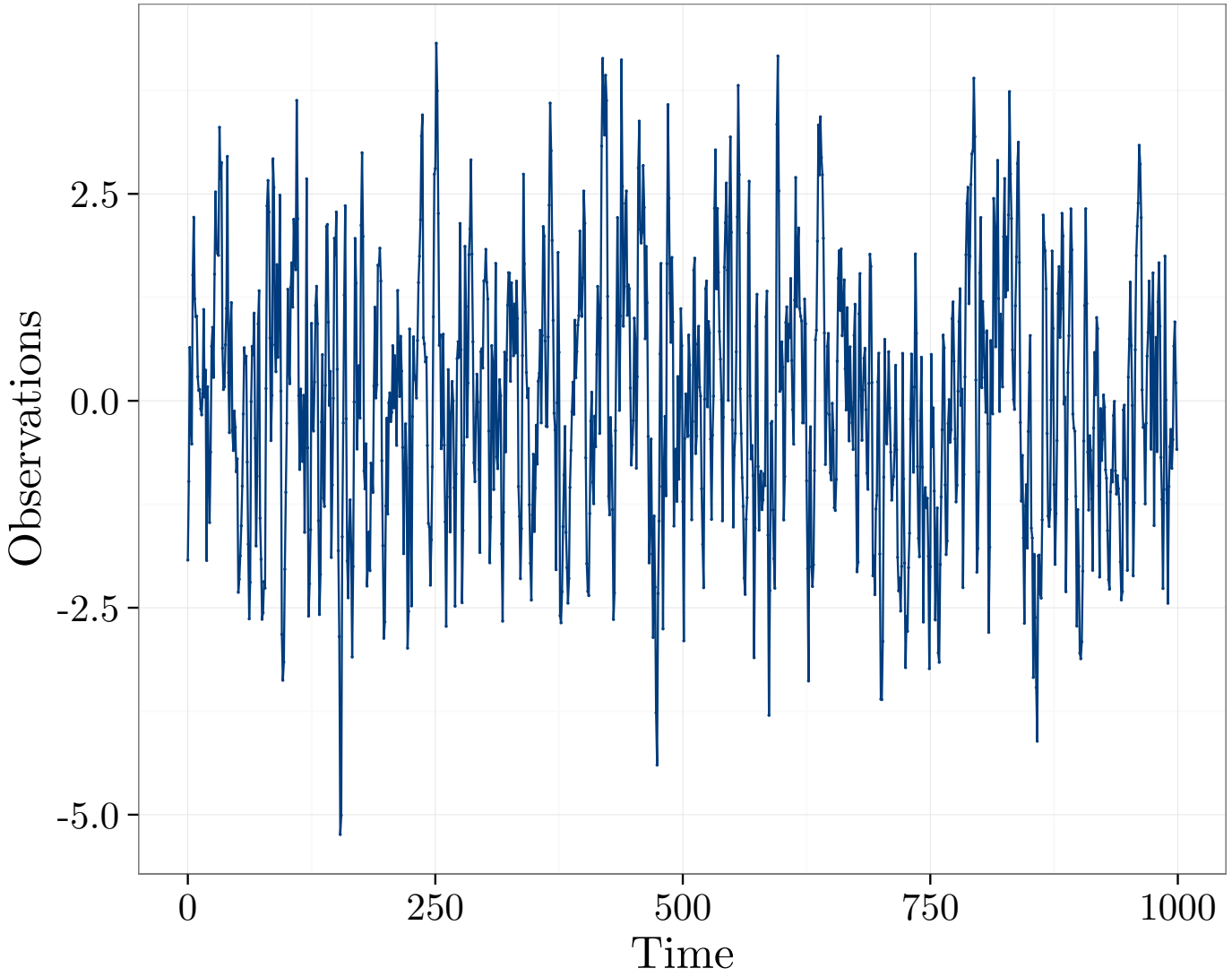
$$X_t = 0.3X_{t-1} + 0.1X_{t-2} + W_t + W_{t-1} + 0.16W_{t-2},$$

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

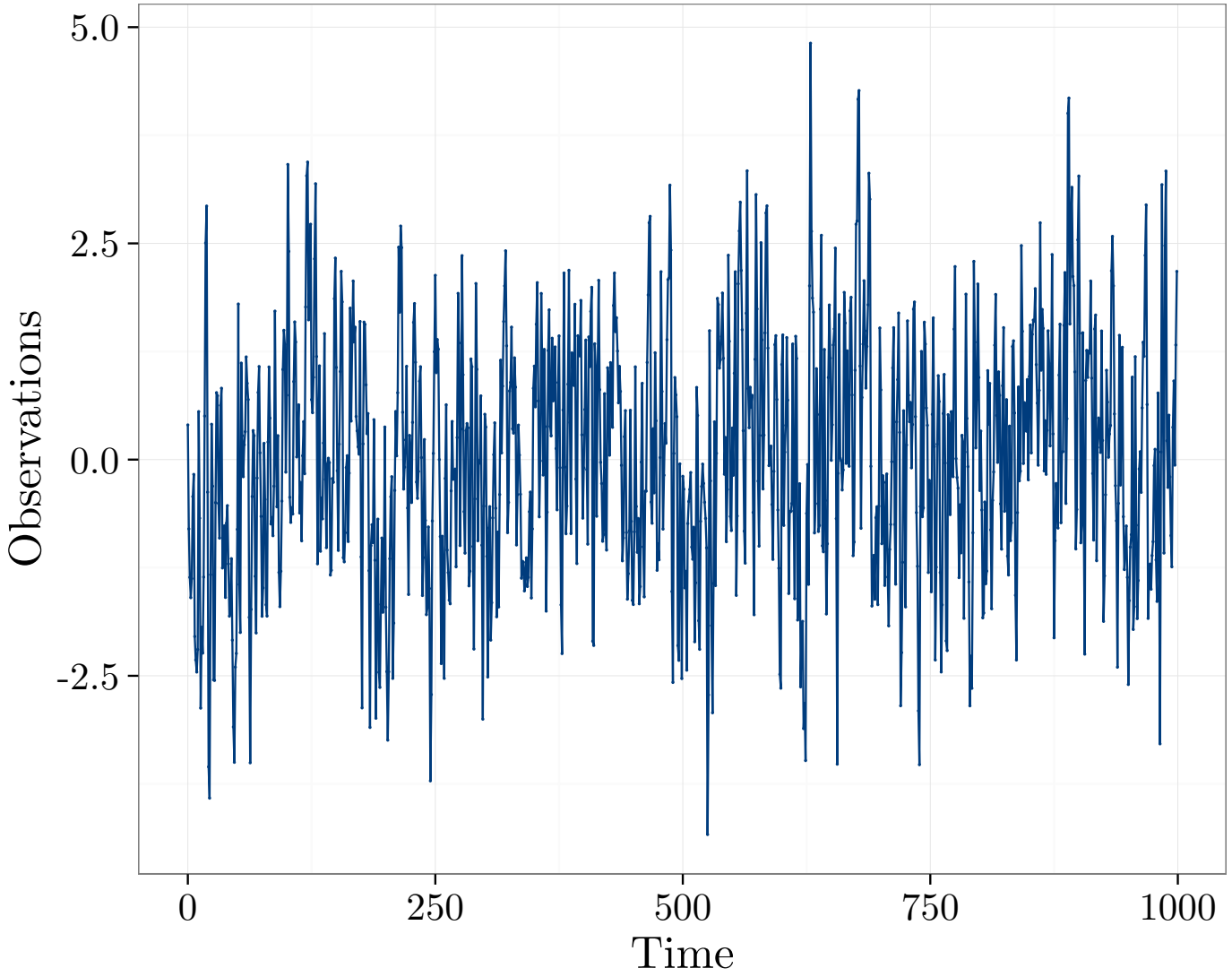
$$\begin{aligned}
X_t &= 0.3X_{t-1} + 0.1X_{t-2} + W_t + W_{t-1} + 0.16W_{t-2} \\
X_t - 0.3X_{t-1} - 0.1X_{t-2} &= W_t + W_{t-1} + 0.16W_{t-2} \\
(1 - 0.3B - 0.1B^2)X_t &= (1 + B + 0.16B^2)W_t \\
(1 + 0.2B)(1 - 0.5B)X_t &= (1 + 0.2B)(1 + 0.8B)W_t \\
(1 - 0.5B)X_t &= (1 + 0.8B)W_t \\
X_t &= 0.5X_{t-1}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

Therefore, our initial model was 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

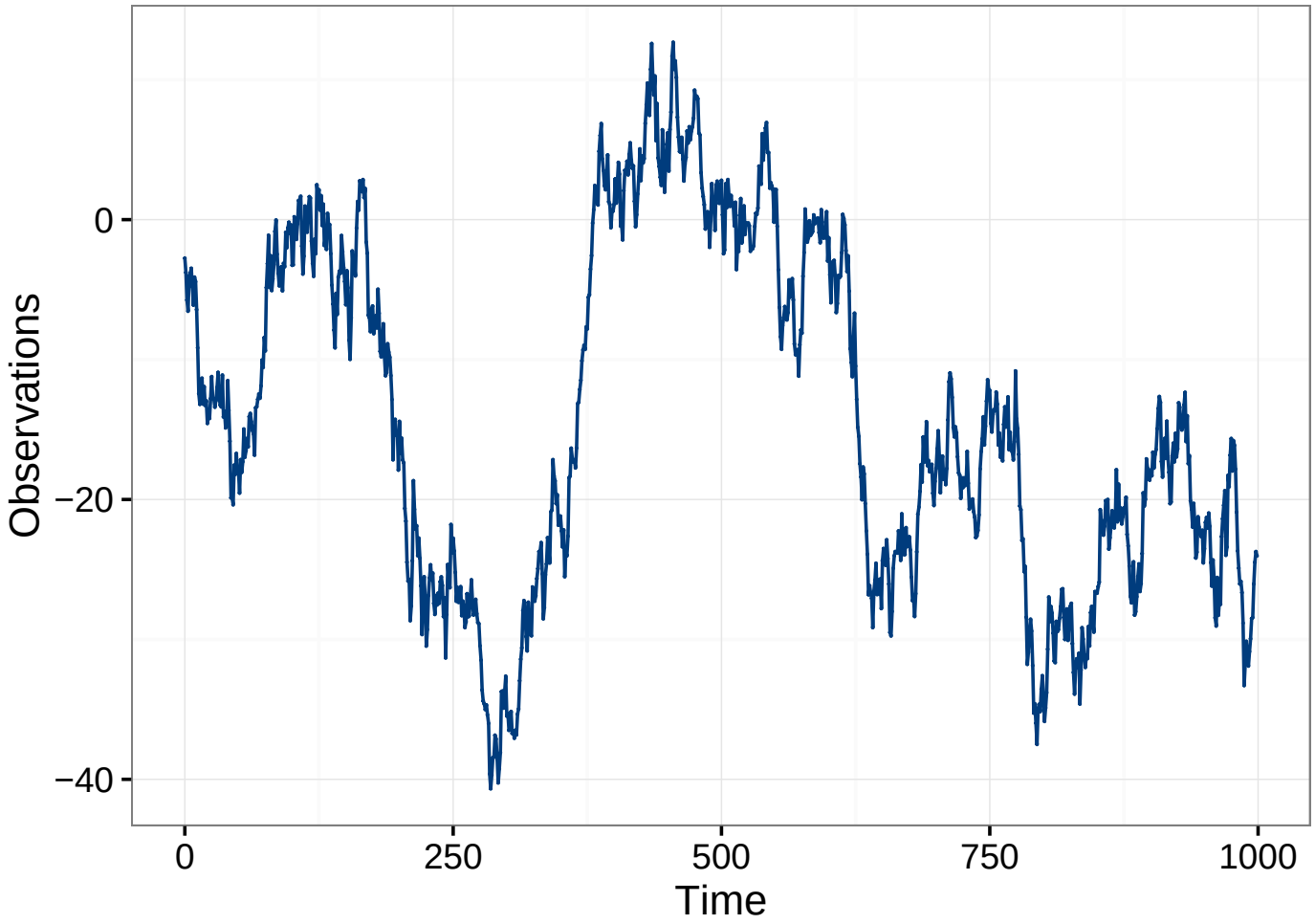


Figure: Simulation of a Random walk: $X_t = X_{t-1} + W_t$

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

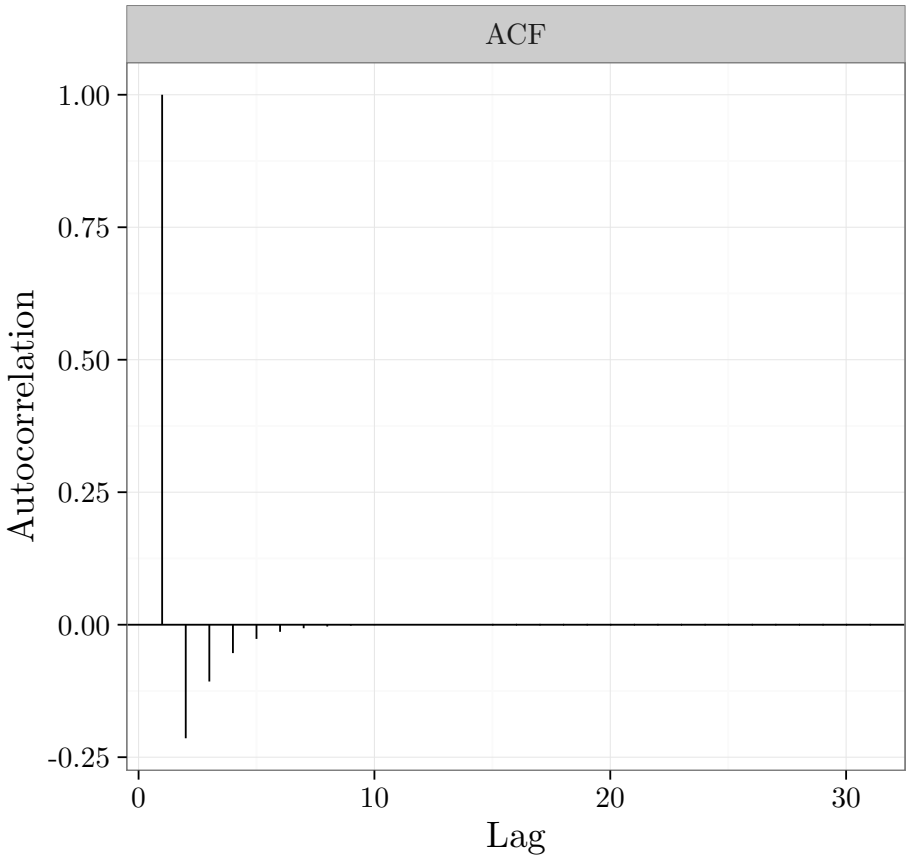
1. Derivation the ACF/PACF of ARMA(p,q) is generally difficult.
   Example: Given an ARMA(1,1), we have $\rho(h) = \phi^{h-1}\rho(1)$ where

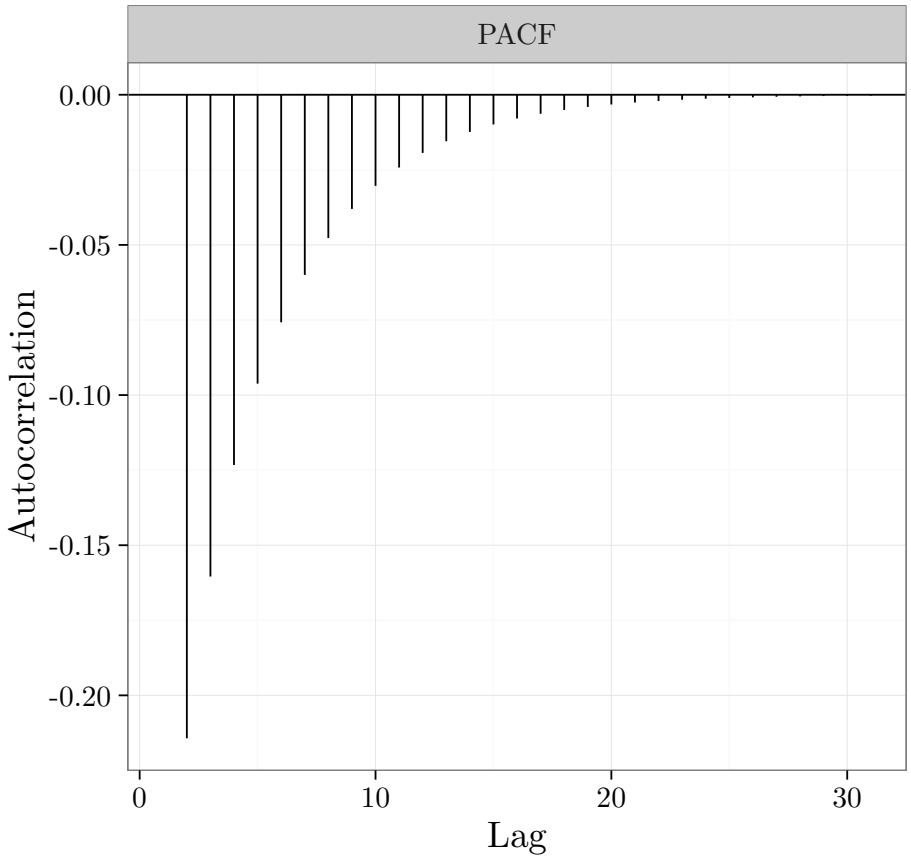
$$\rho(1) = \frac{(1+\theta\phi)(\theta+\phi)}{1+2\phi\theta+\theta^2}.$$

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

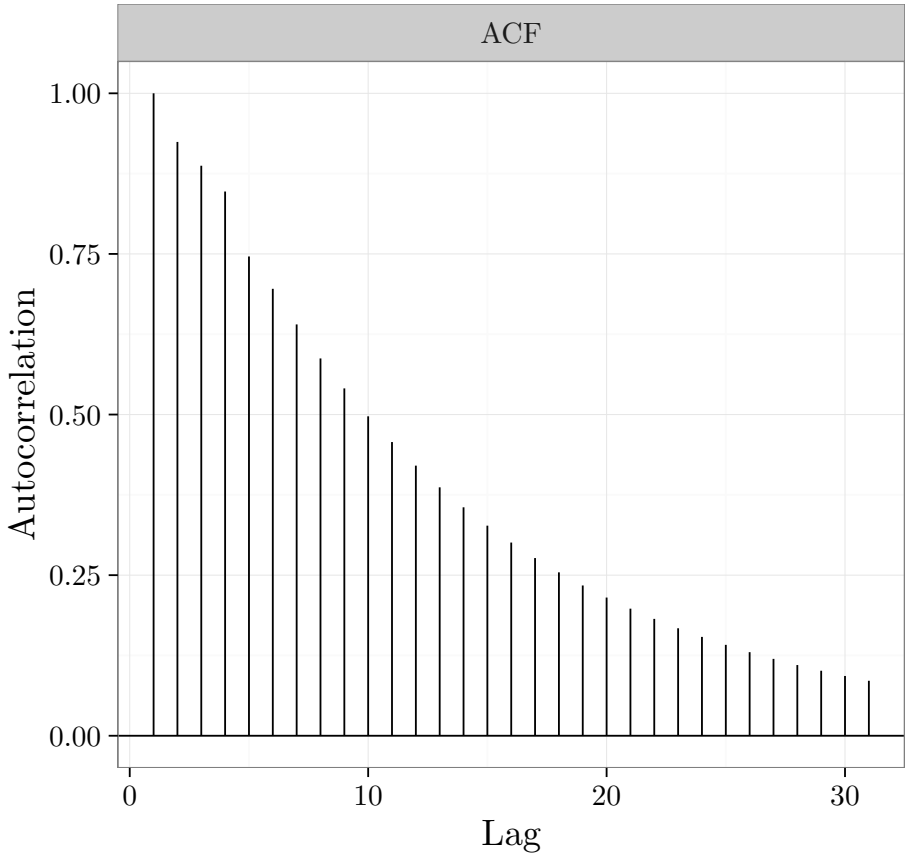
Order Identification of ARMA Models

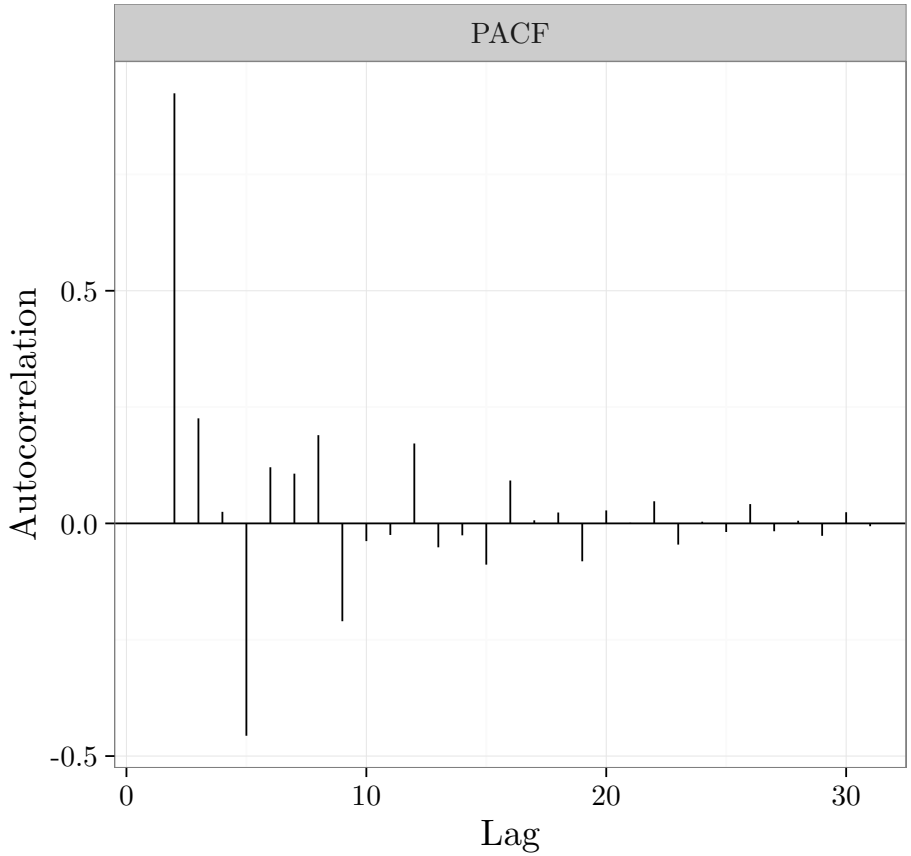




This an ARMA(1,1) process

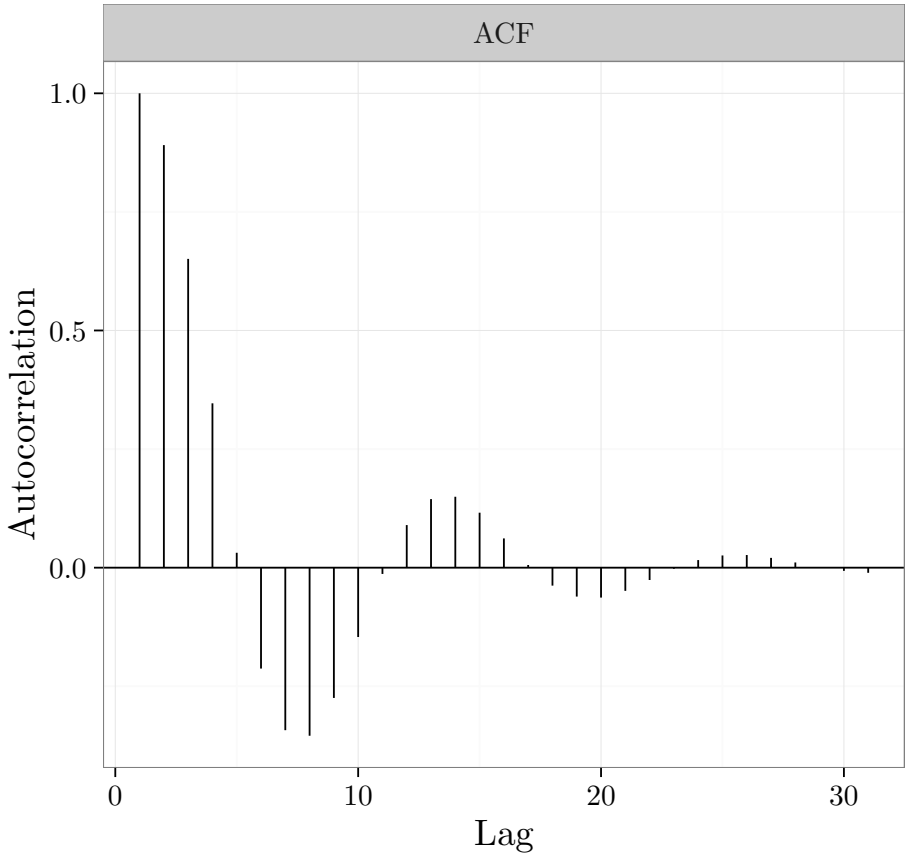
Order Identification of ARMA Models

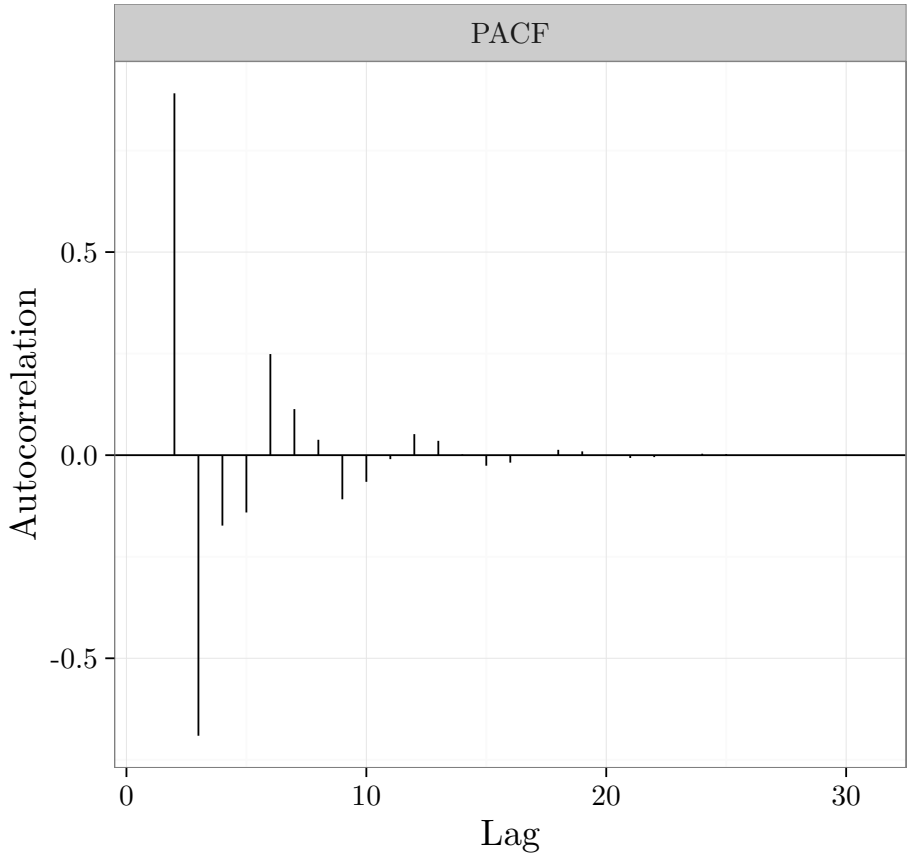




This an ARMA(3,4) process

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 plots to assess if the model is reasonable for this time series.

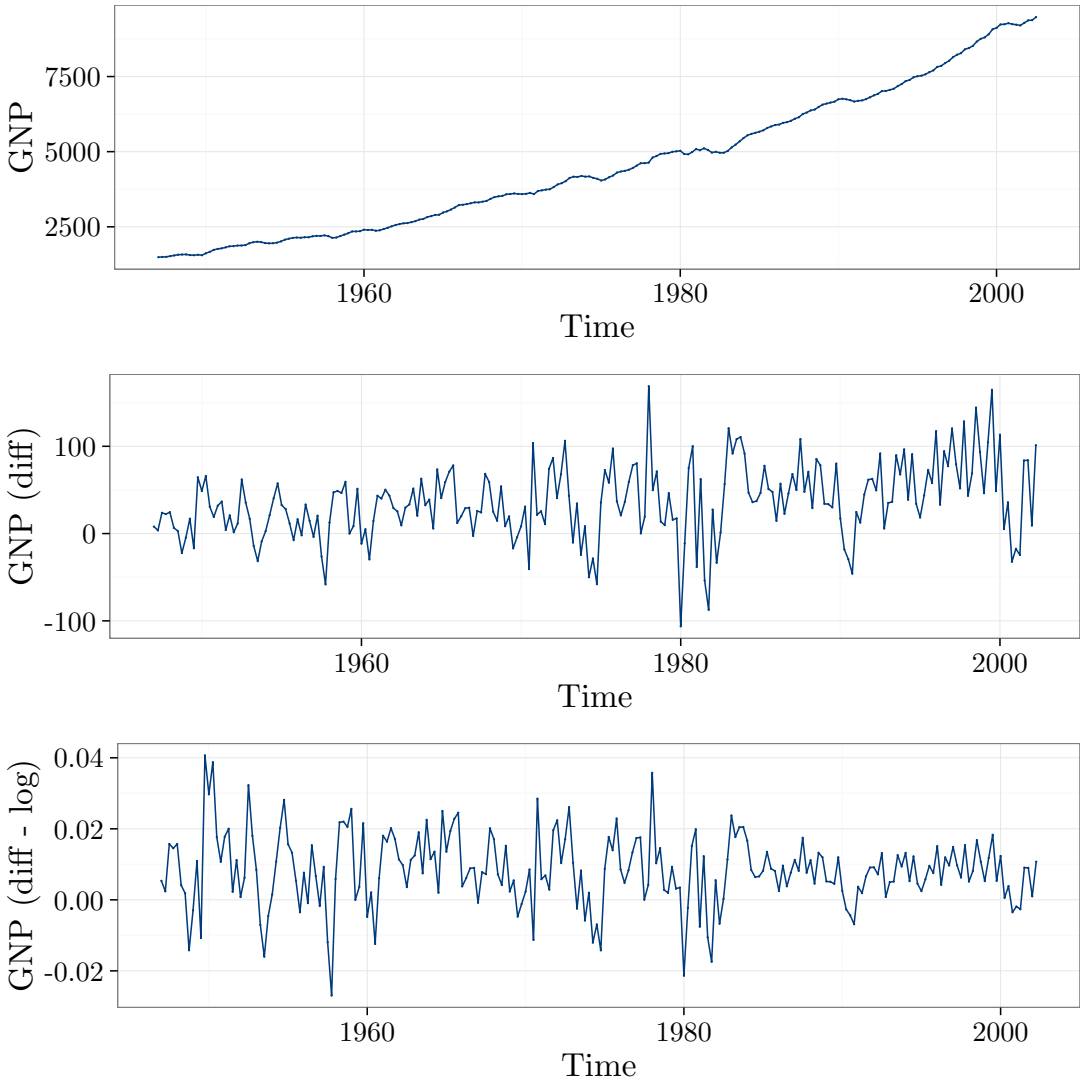
ARIMA Models

- 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 stationary process.
- 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

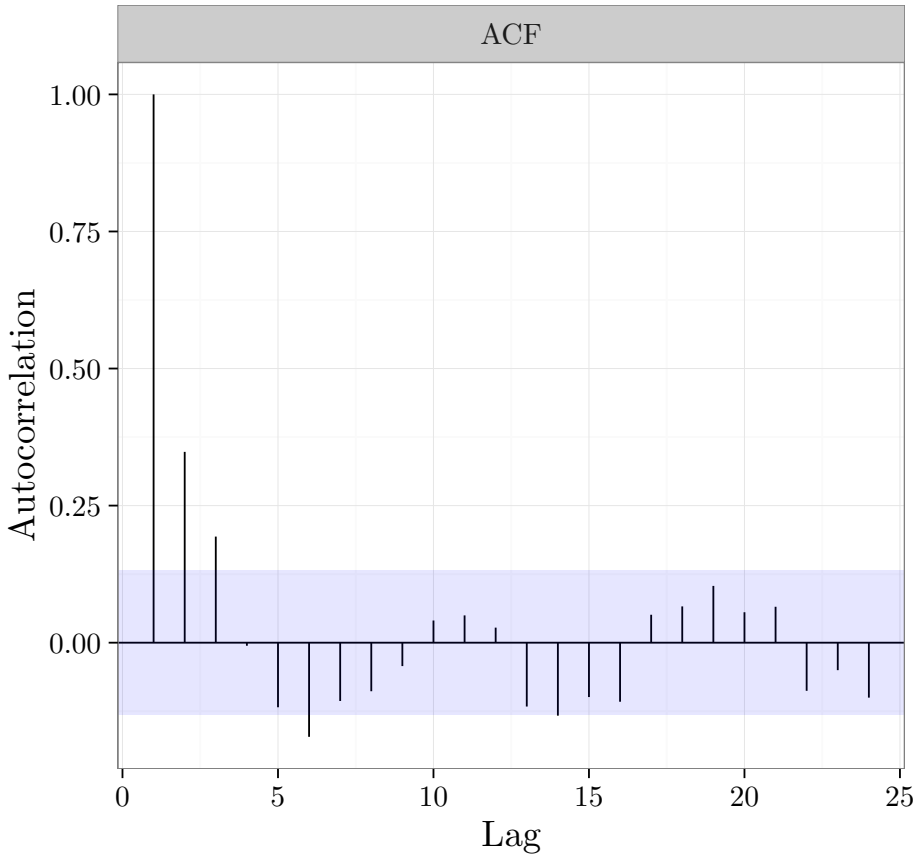
$$\phi(B)\nabla^d X_t = \theta(B)W_t$$

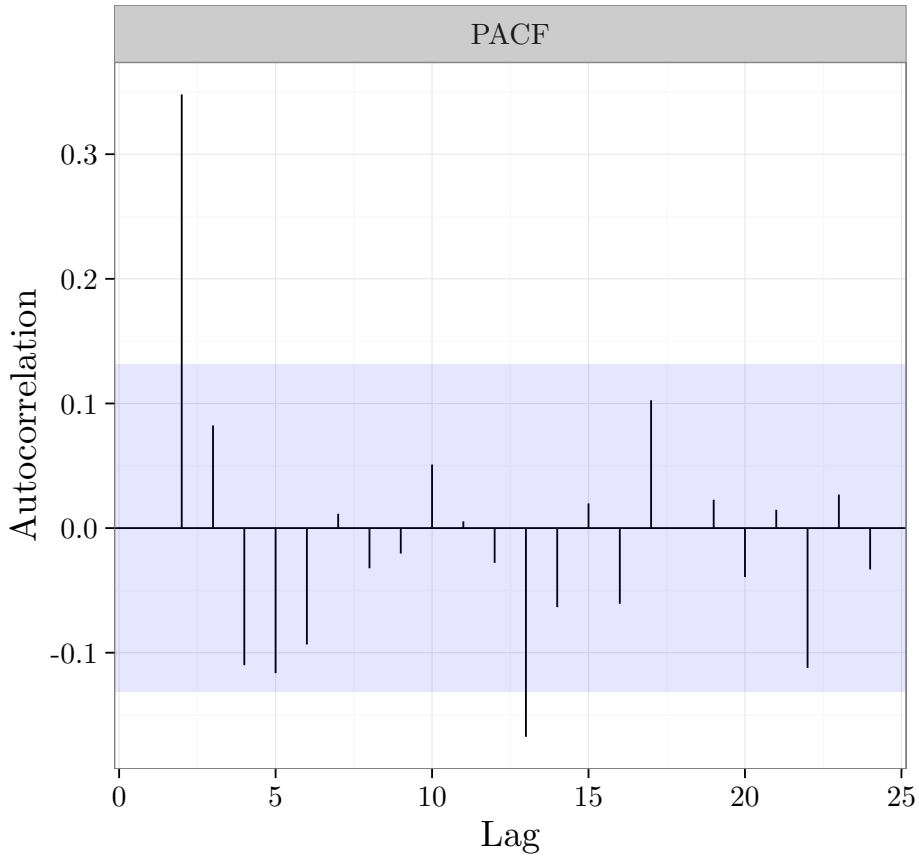
where $\phi(B)$ denotes the autoregressive operator, $\theta(B)$ the moving average operator, and $\nabla^d X_t = (1 - B)^d X_t$ represents the differencing operator.

Example: US GNP



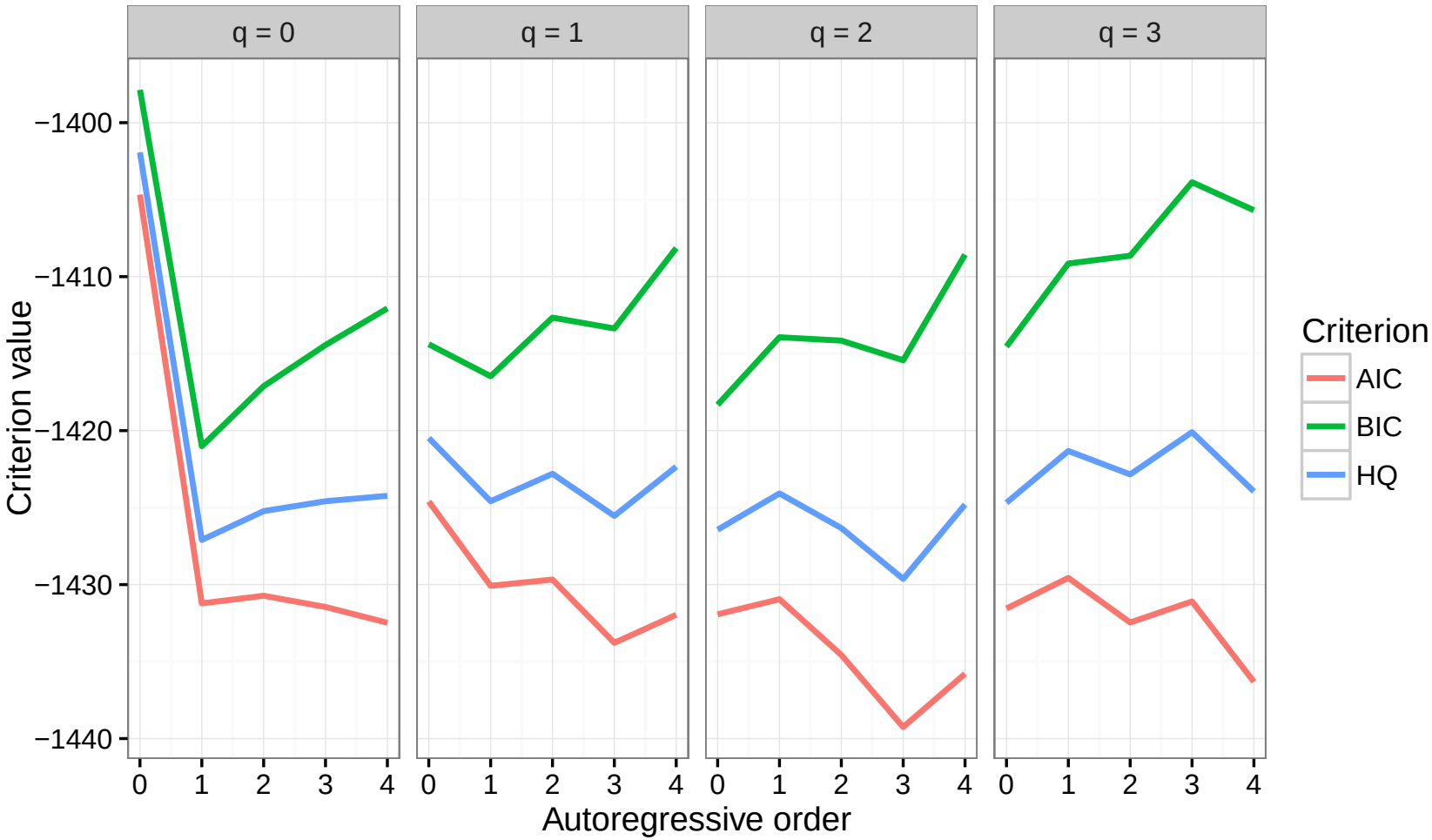
Example: US GNP



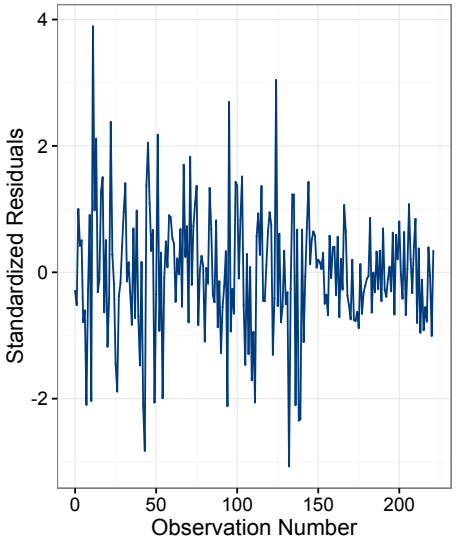


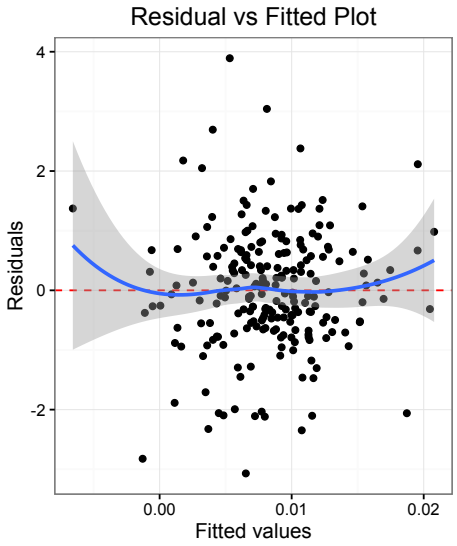
An AR(1) or MA(2) seem like potential candidates.

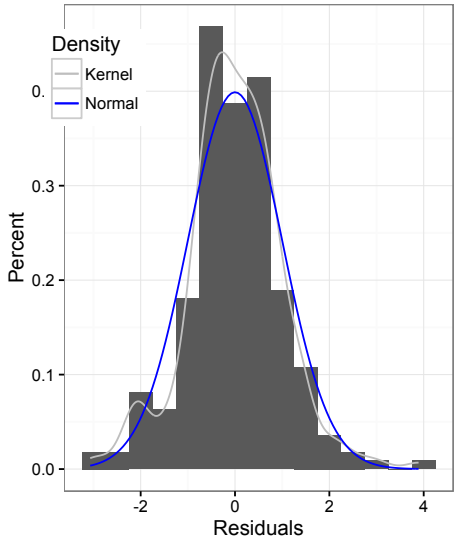
Example: US GNP

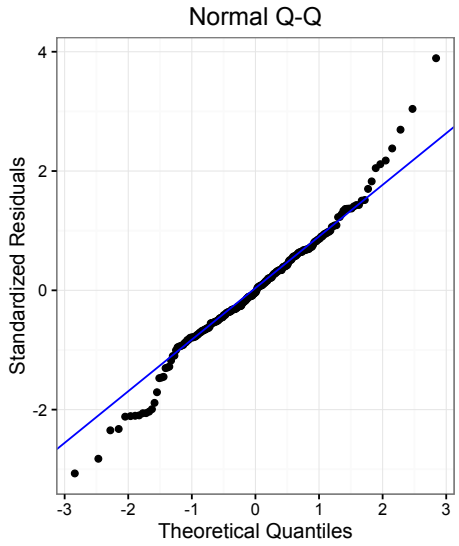


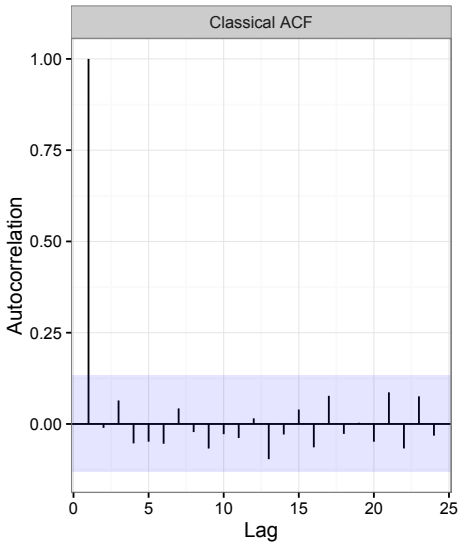
Example: US GNP

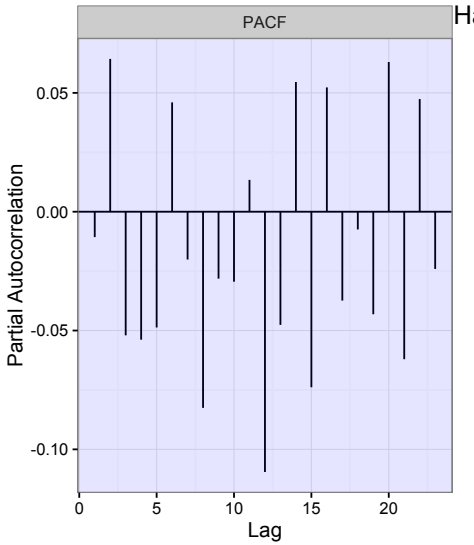


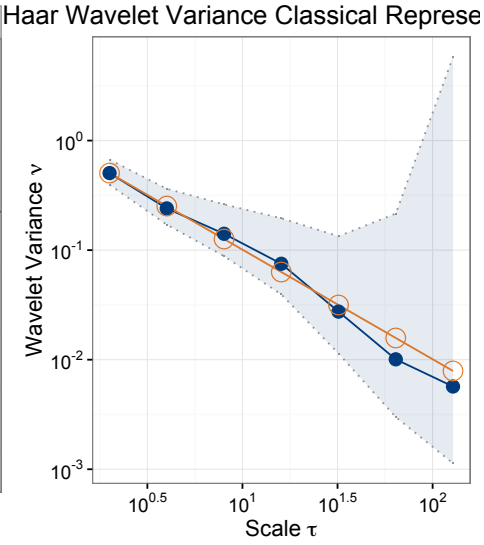


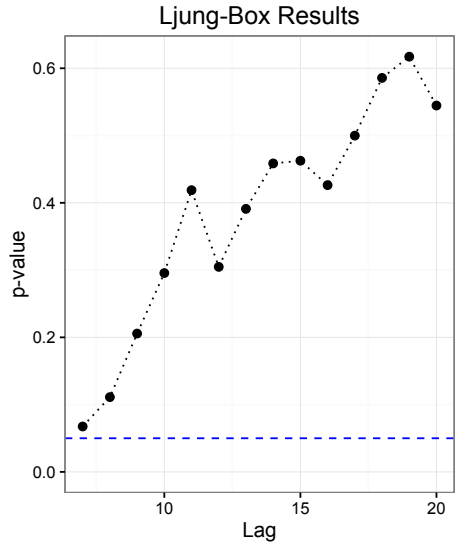












SARIMA Models

- 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:

$$\mathcal{M}_1: \quad X_t = \Phi X_{t-4} + W_t$$
$$\mathcal{M}_2: \quad X_t = \Phi X_{t-4} + W_t + \theta W_{t-1}.$$

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

$$\begin{aligned}
\operatorname{var}(X_t) &= \operatorname{var}(\Phi X_{t-4} + W_t) = \operatorname{var}\left(W_t + \Phi W_{t-4} + \Phi^2 W_{t-8} + \ldots\right) \\
&= \operatorname{var}\left(\sum_{i=0}^{\infty} \Phi^i W_{t-4i}\right) = \sigma^2 \sum_{i=0}^{\infty} \left(\Phi^2\right)^i = \frac{\sigma^2}{1-\Phi^2}, \\
\operatorname{cov}(X_t, X_{t+4k}) &= \operatorname{cov}\left(X_t, \Phi X_{t+4(k-1)} + W_{t+4k}\right) = \Phi \operatorname{cov}\left(X_t, X_{t+4(k-1)}\right) \\
&= \Phi^k \operatorname{var}(X_t) = \frac{\Phi^K \sigma^2}{1-\Phi^2}, \\
\operatorname{cov}(X_t, X_{t+m}) &= \operatorname{cov}\left(\sum_{i=0}^{\infty} \Phi^i W_{t-4i}, \sum_{j=0}^{\infty} \Phi^j W_{t-4j+m}\right) \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \Phi^{i+j} \operatorname{cov}(W_{t-4i}, W_{t-4j+m}) = 0.
\end{aligned}$$

ACF for $\mathcal{M}_1$

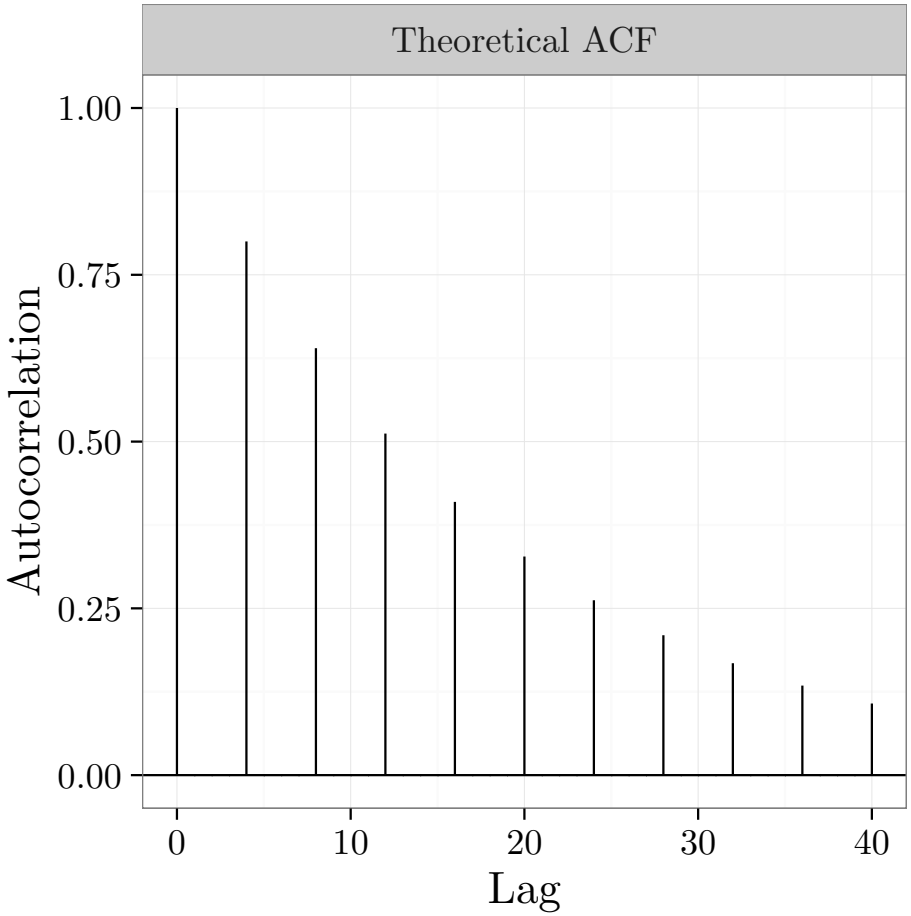
Thus, the only nonzero correlation, aside from lag zero, is simply

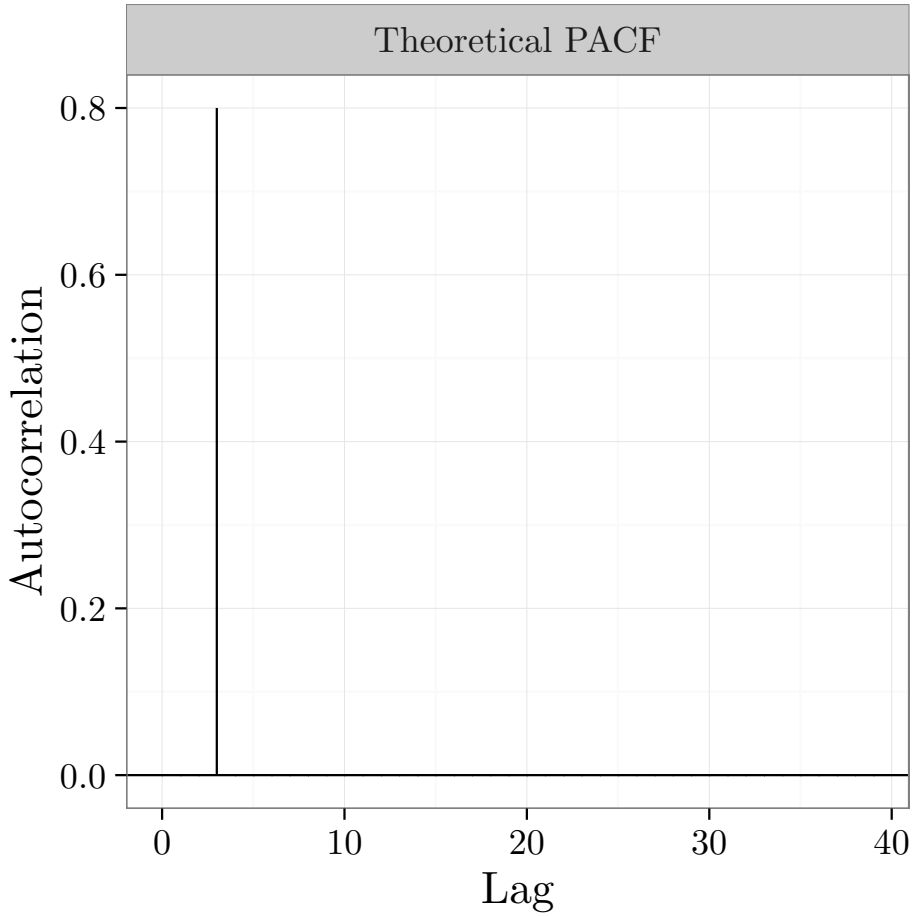
$$\text{corr}\,(X_t, X_{t+4k}) = \Phi^k.$$

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

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

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

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

Next, we consider the following:

$$\begin{aligned}
X_t &= \Phi X_{t-4} + W_t + \theta W_{t-1} \\
X_t X_{t-1} &= \Phi X_{t-4} X_{t-1} + W_t X_{t-1} + \theta W_{t-1} X_{t-1} \\
\mathbb{E}\left[X_t X_{t-1}\right] &= \Phi \mathbb{E}\left[X_{t-4} X_{t-1}\right] + \mathbb{E}\left[W_t X_{t-1}\right] + \mathbb{E}\left[\theta W_{t-1} X_{t-1}\right] \\
\gamma(1) &= \Phi\gamma(3) + \theta\sigma^2.
\end{aligned}$$

ACF for $\mathcal{M}_2$

Moreover, we have:

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

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 $\mathcal{M}_2$

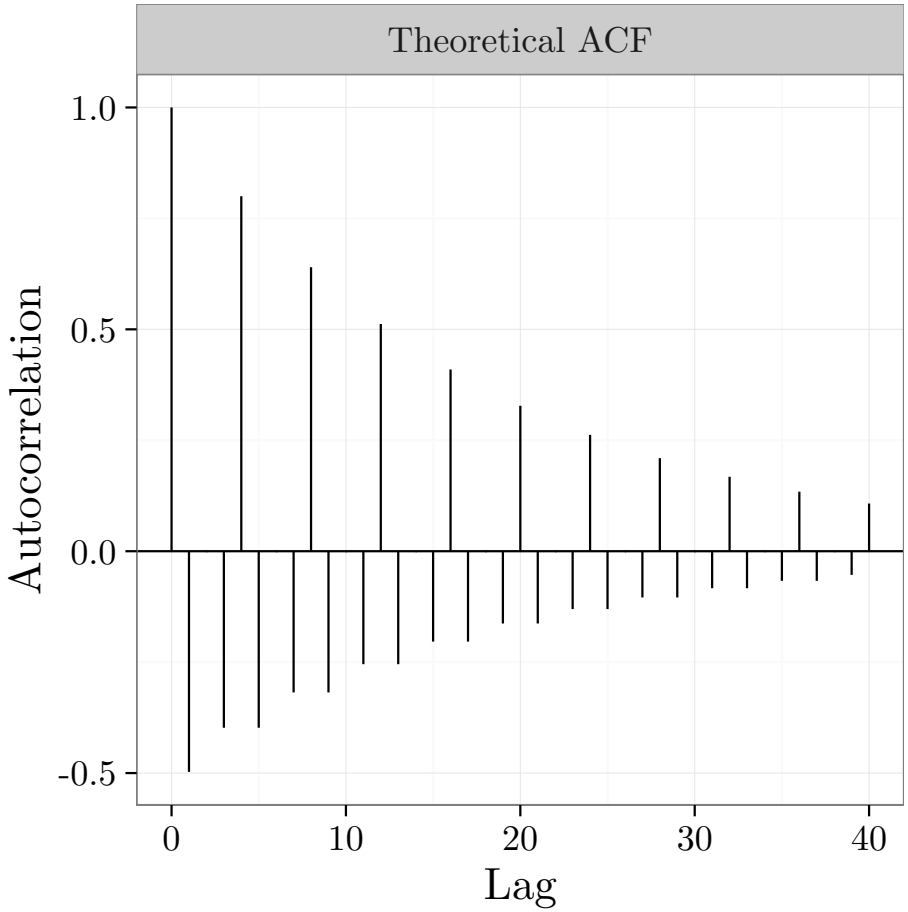
To sum up, we showed:

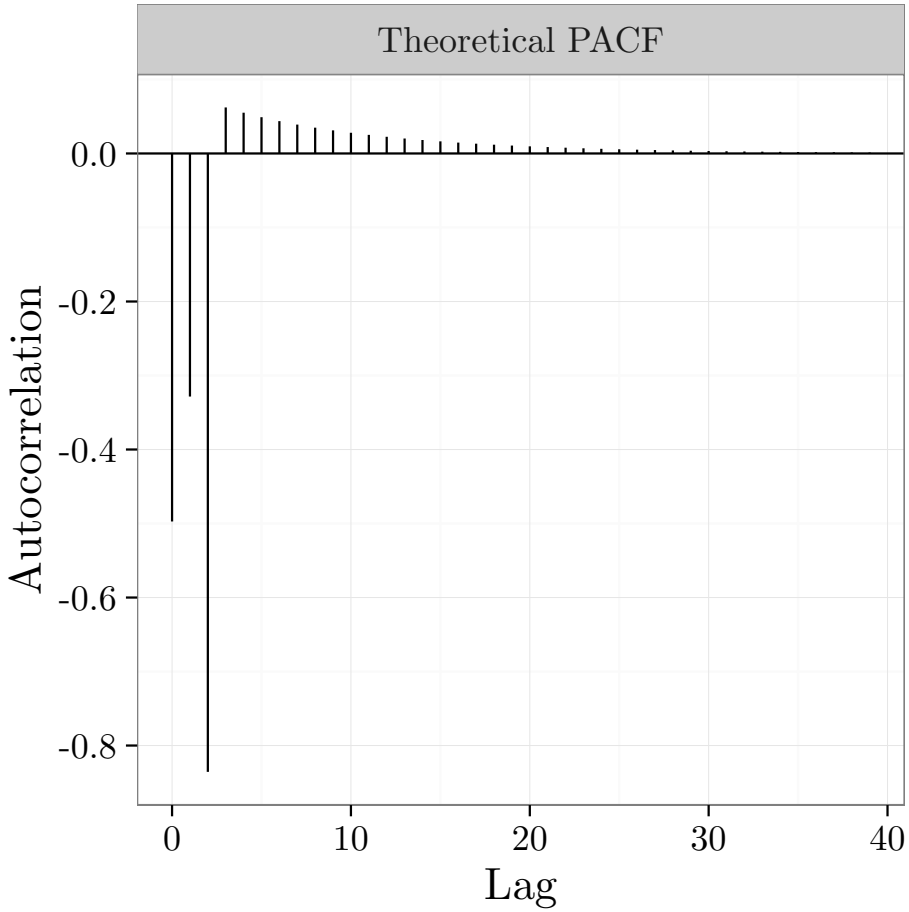
$$\gamma(0) = \frac{\sigma^2\left(1+\theta^2\right)}{1-\Phi^2}$$
$$\gamma(4k) = \Phi^k\gamma(0)$$
$$\gamma(4k+1) = \gamma(4k-1) = \Phi^k\frac{\theta\sigma^2}{1-\Phi^2}.$$

By combining these results, we verify that:

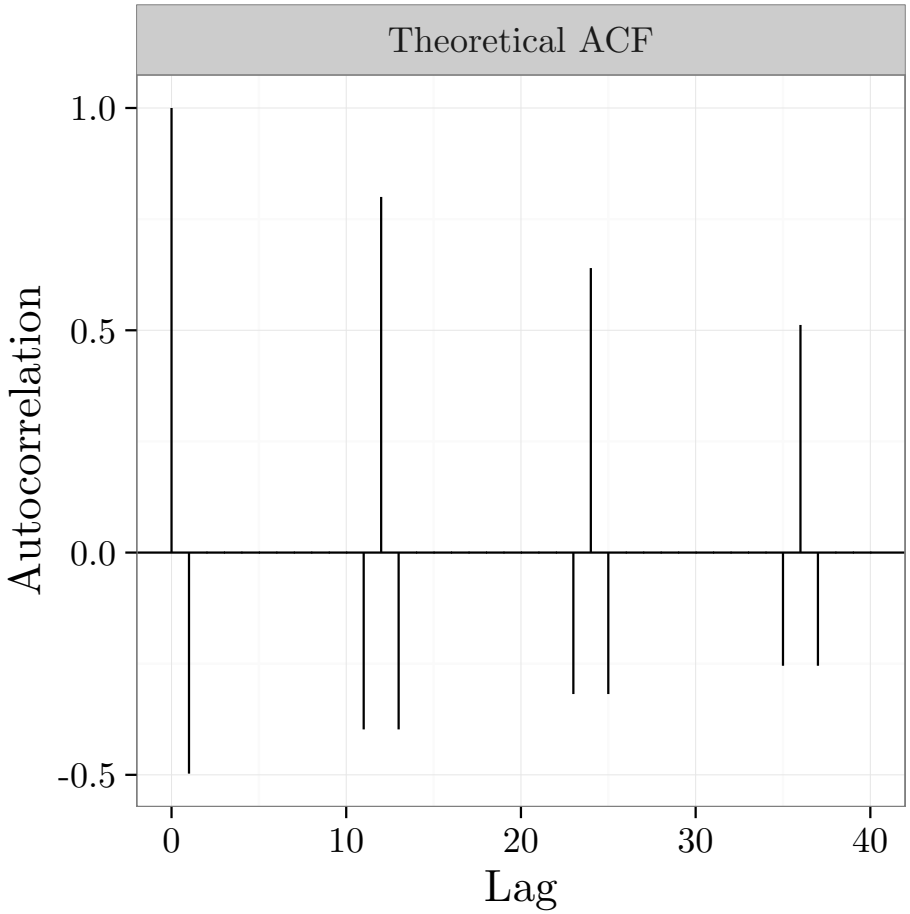
$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4 \\ \frac{\theta}{1+\theta^2}\Phi^{|h/4|} & \text{if } h \in \mathbb{Z}_4^* \\ 0 & \text{if } h \notin \mathbb{Z}_4 \cup \mathbb{Z}_4^* \cup \{0\}. \end{cases}$$

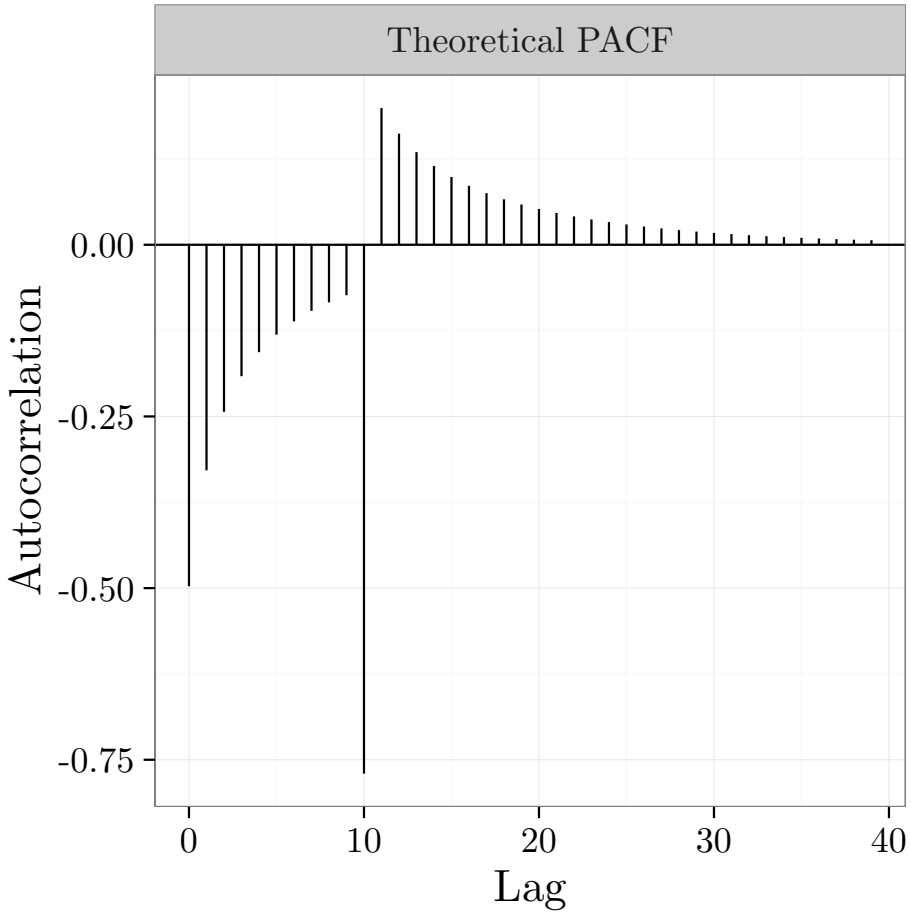
ACF/PACF for $\mathcal{M}_2$ with $\Phi = 0.8$ and $\theta = -0.9$



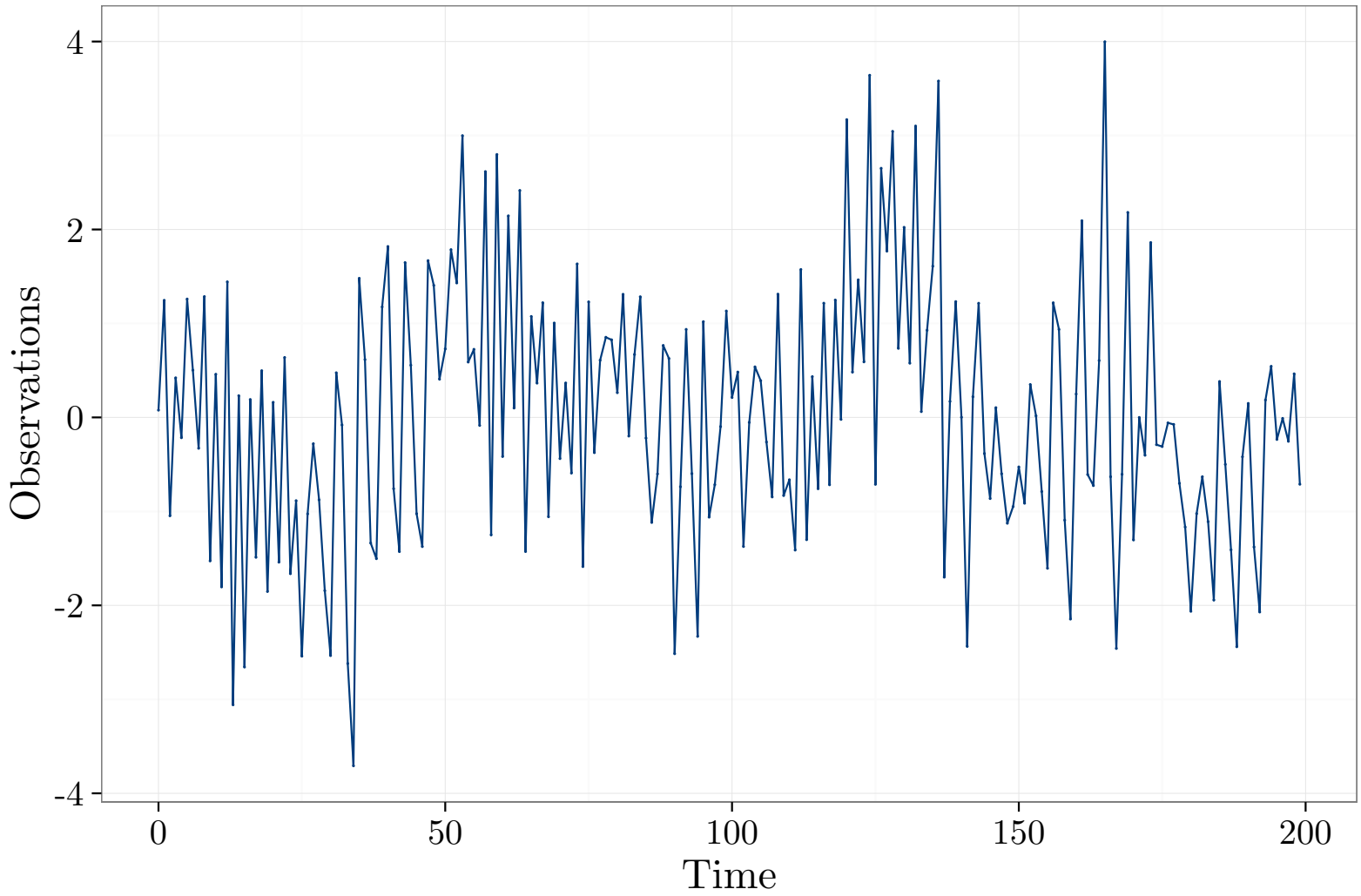


ACF/PACF for $X_t = 0.8X_{t-12} - 0.9W_{t-1} + W_t$

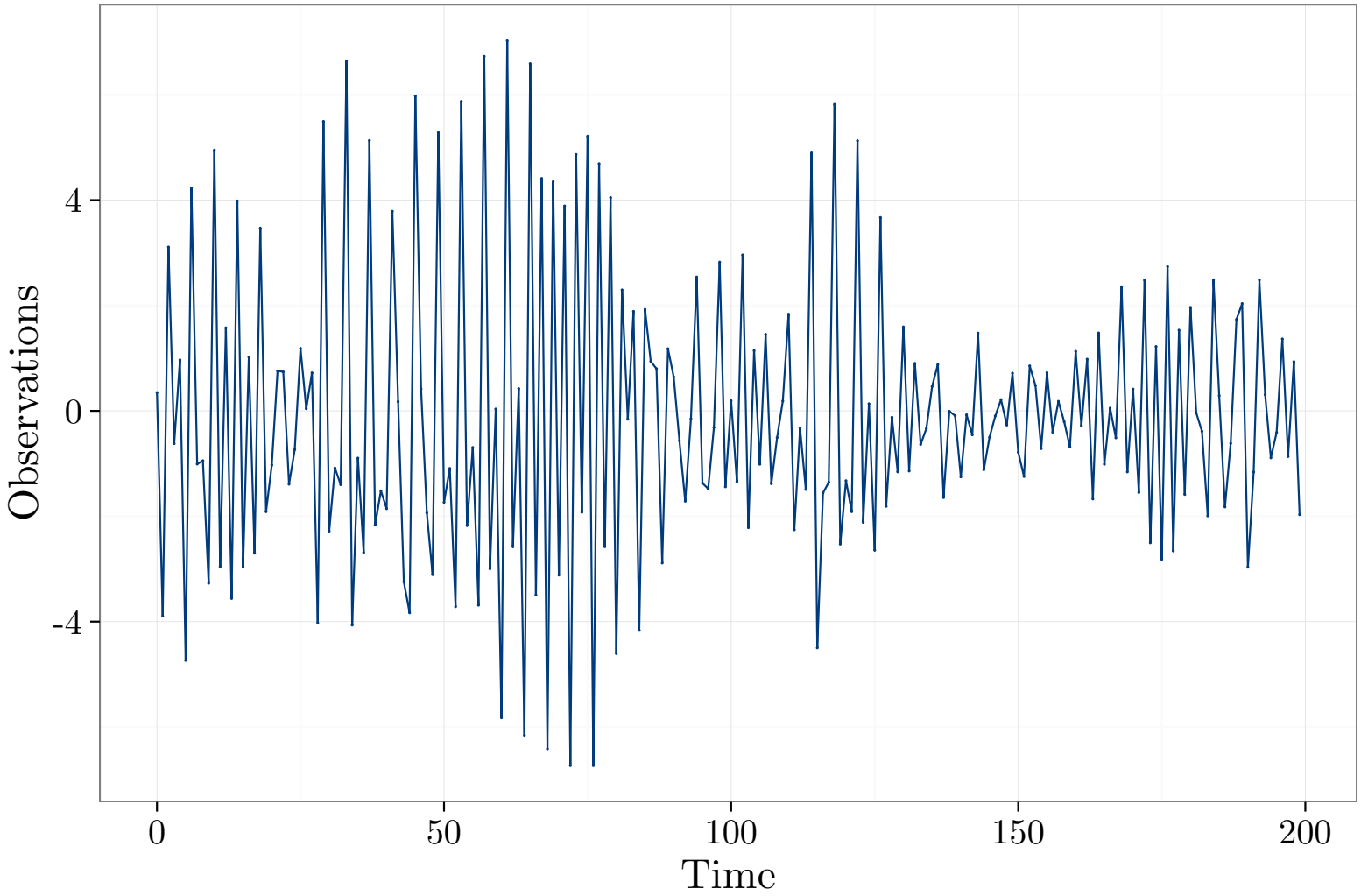




Simulated process: $\mathcal{M}_1$ with $\Phi = 0.8$



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* $(p, q, d)x(P, Q, D)_s$ is in the form of:

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

where w_t is the usual Gaussian white noise process. Besides, the ordinary autoregressive and moving average processes 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:

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

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

are the seasonal autoregressive and moving average operators.

Some Examples

SARIMA Models

$$\mathcal{M}_1: \quad X_t = \Phi X_{t-4} + W_t$$
$$\mathcal{M}_2: \quad X_t = \Phi X_{t-4} + W_t + \theta W_{t-1}$$
$$\mathcal{M}_3: \quad X_t = \Phi_1 X_{t-12} + \Phi_2 X_{t-24} + W_t + \theta W_{t-1} + \Theta W_{t-12}$$
$$\mathcal{M}_4: \quad X_t = W_t + \theta W_{t-1} + \Theta_1 W_{t-6} + \Theta_2 W_{t-12}$$
$$\mathcal{M}_5: \quad X_t = \Phi X_{t-4} + W_t + \Theta W_{t-6}$$

Identified Models

Model 1 is a SARIMA$(0,0,0)\times(1,0,0)_4$
Model 2 is a SARIMA$(0,0,1)\times(1,0,0)_4$
Model 3 is a SARIMA$(0,0,1)\times(2,0,1)_{12}$
Model 4 is a SARIMA$(0,0,1)\times(0,0,2)_6$
Model 5 is a SARIMA$(4,0,0)\times(0,0,1)_6$ or a SARIMA$(0,0,6)\times(1,0,0)_4$

Identifying SARIMA coefficients

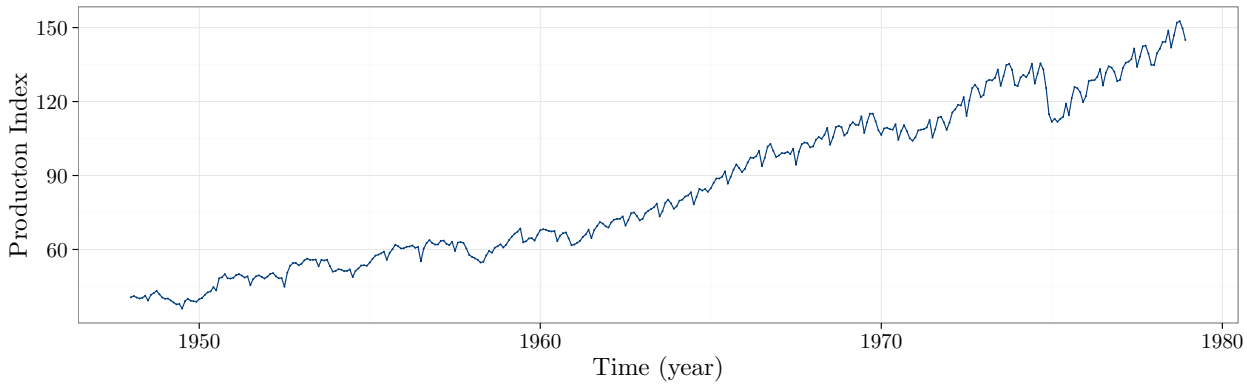
Identifying the order of a SARIMA model is a daunting task and there is a "good" way to do it. Here is a possible method:

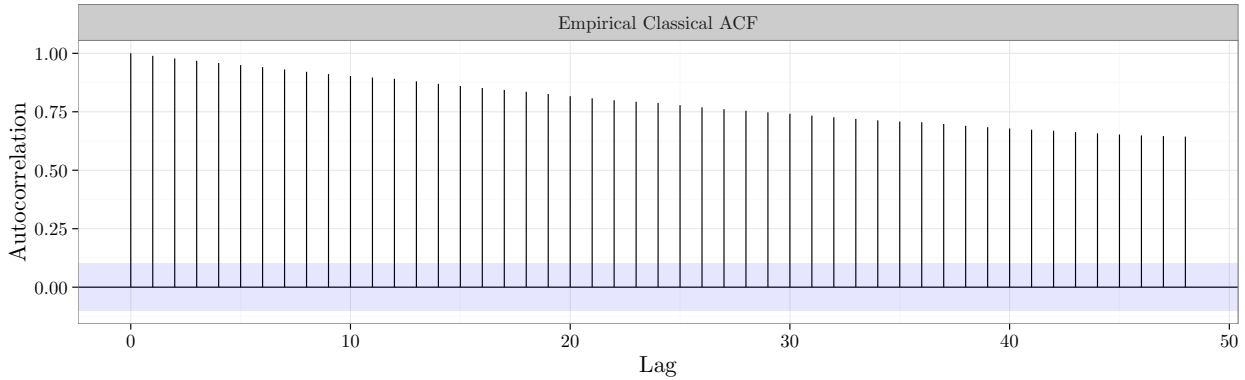
- Apply difference operators until a roughly stationary series is produced.
- Plot ACF/PACF of the "differenced" series, then:
  - Identify the seasonal period s,
  - Identify P and Q using the following table ("seasonal lags" only):

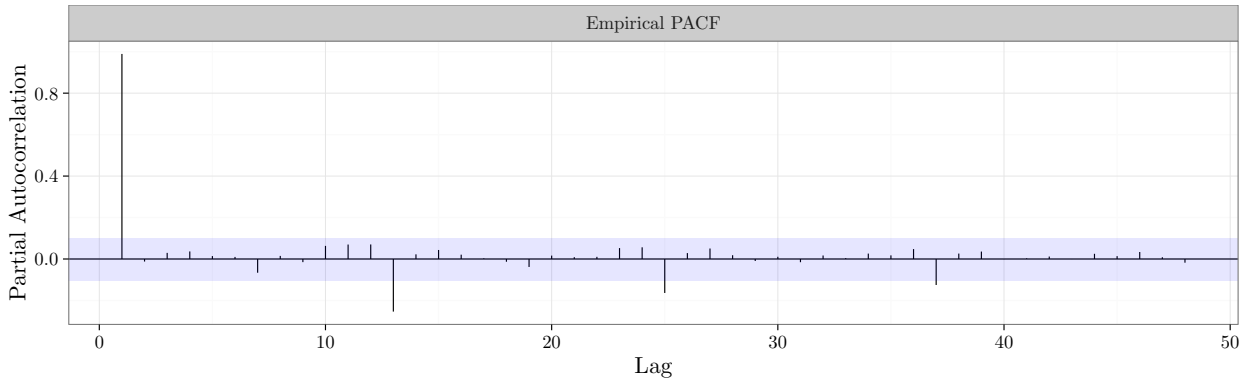
	SAR(P)	SMA(Q)	SARMA(P,Q)
ACF	Tails off	Cuts off after lag Q	Tails off
PACF	Cuts off after lag P	Tails off	Tails off

  - 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.

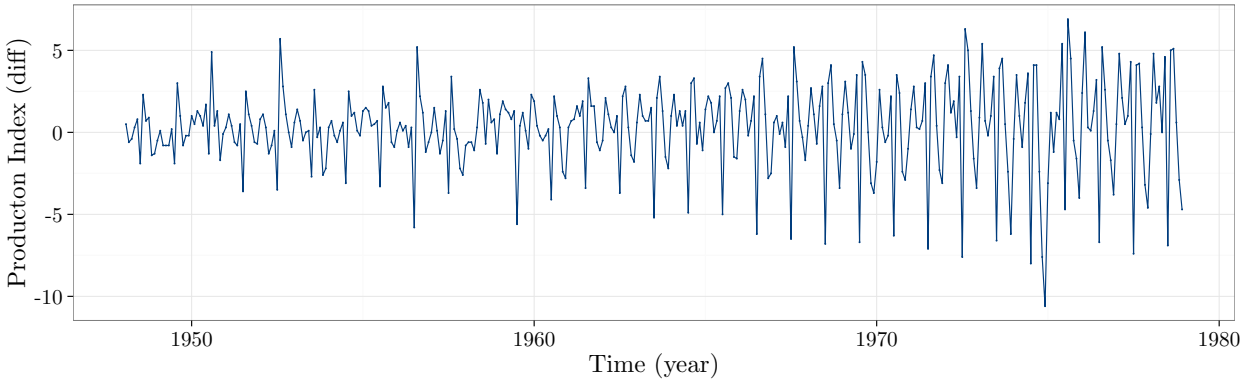
Example: Monthly Production Index Dataset

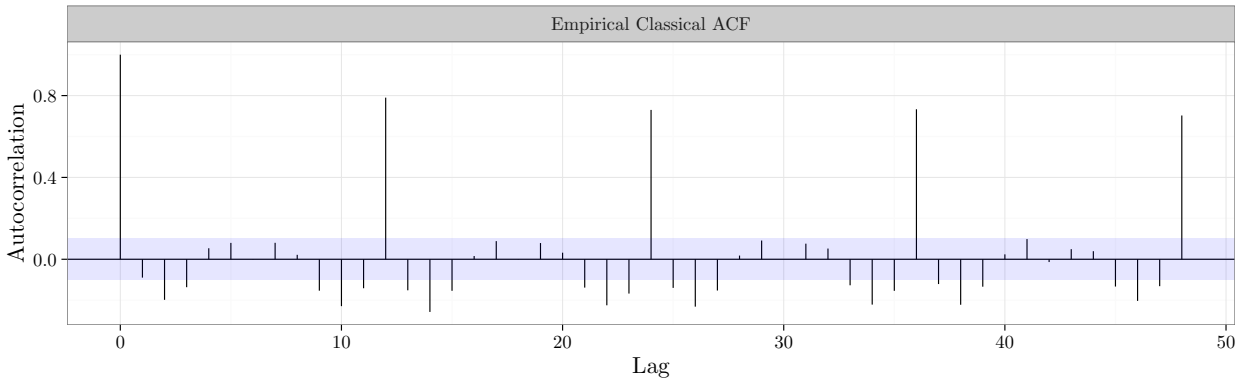


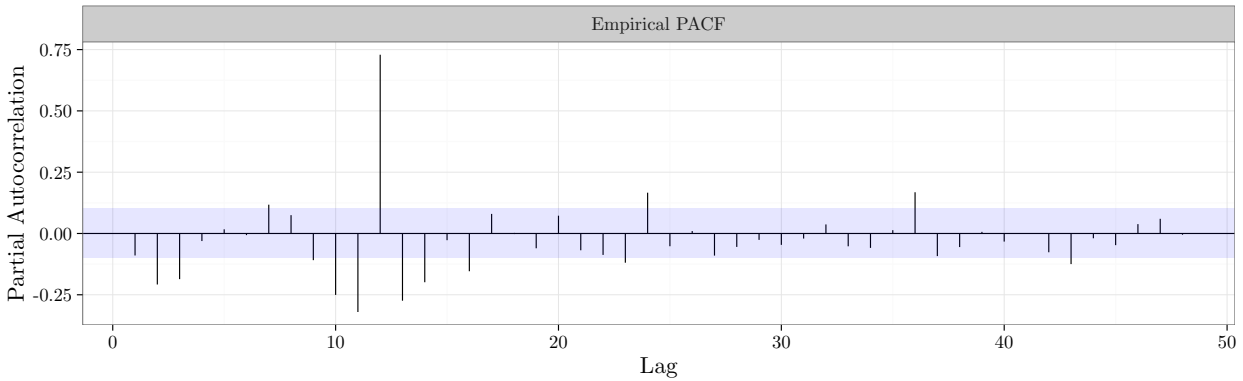




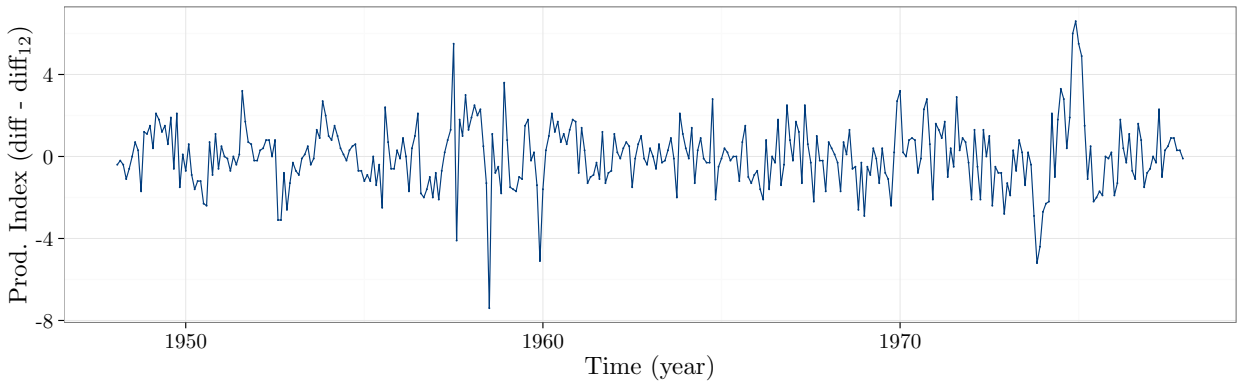
Example: Monthly Production Index Dataset

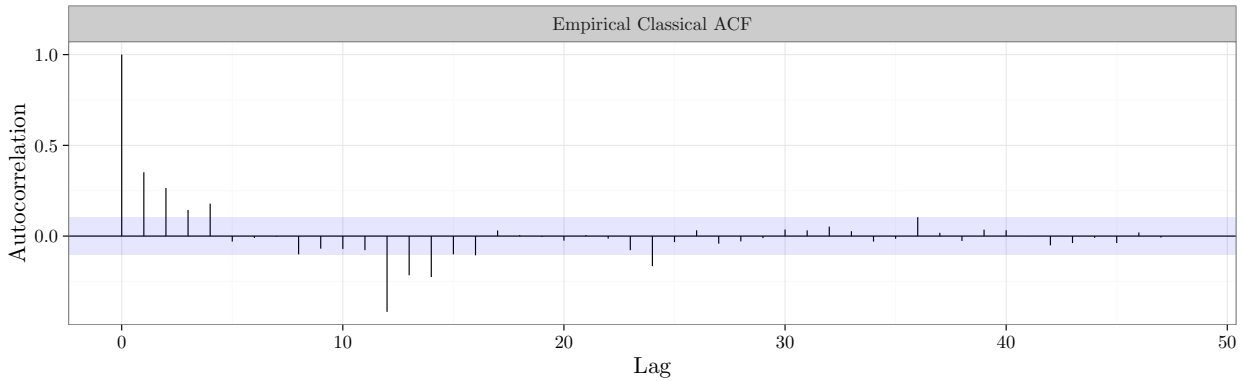


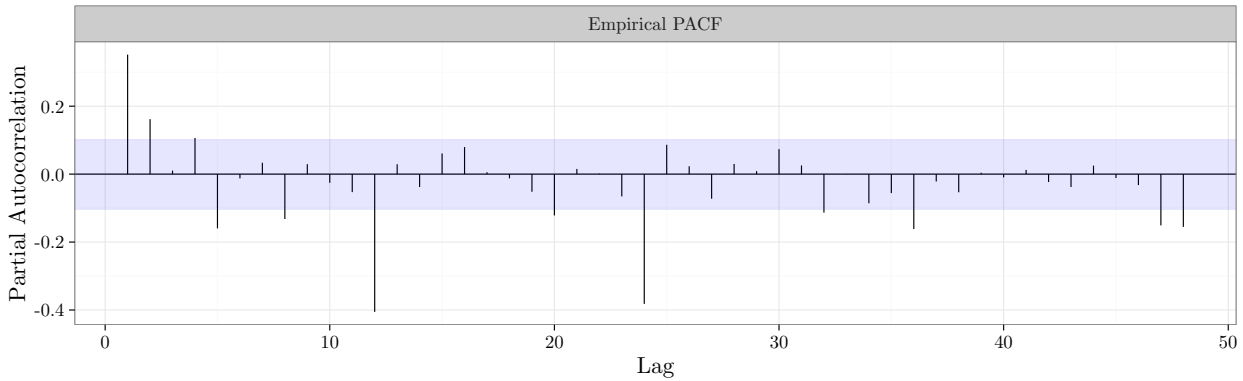




Example: Monthly Production Index Dataset







Example: Monthly Production Index Dataset

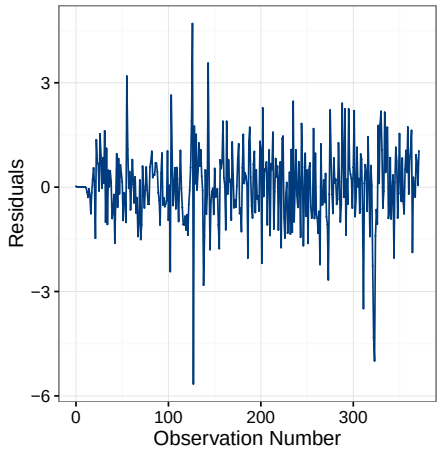
Using the approach described earlier:

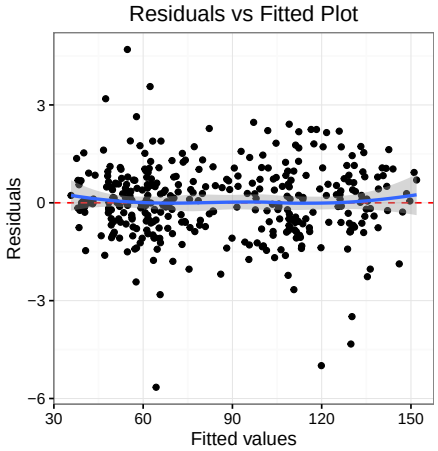
- It seems that $d = 1$, $D = 1$ are appropriate.
- From ACF/PACF of the differenced series, we propose four models:
  - $s = 12$
  - $P = 0$ and Q between 1 and 4.
  - $p = 2$ and $q = 0$
- Select "best" model: let $\mathcal{M}_i$: SARIMA(2,1,0)$\times$(0,1,i)$_{12}$, $i = 1, 2, 3, 4$. Then, we have

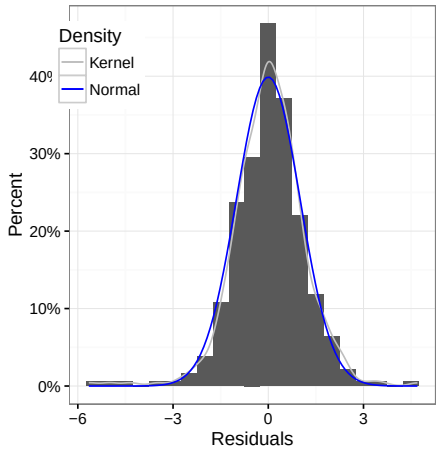
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
AIC	1162.334	1163.712	**1139.965**	1141.898

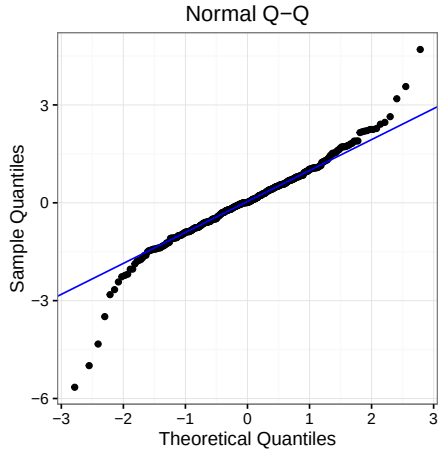
- Diagnostic plots for $\mathcal{M}_3$.

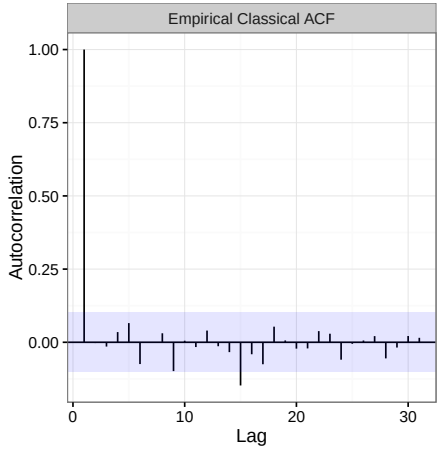
Example: Monthly Production Index Dataset

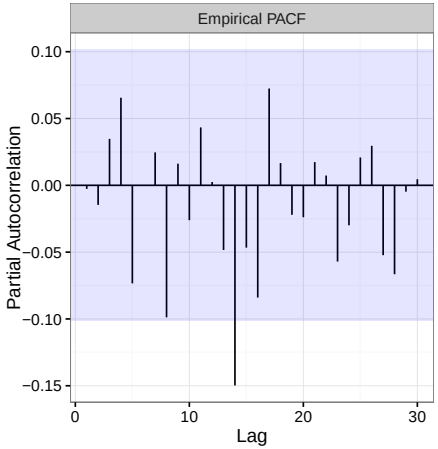


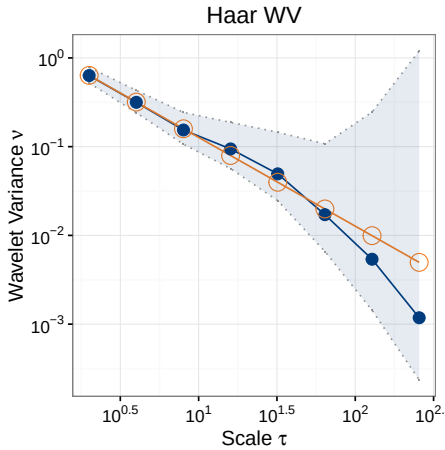


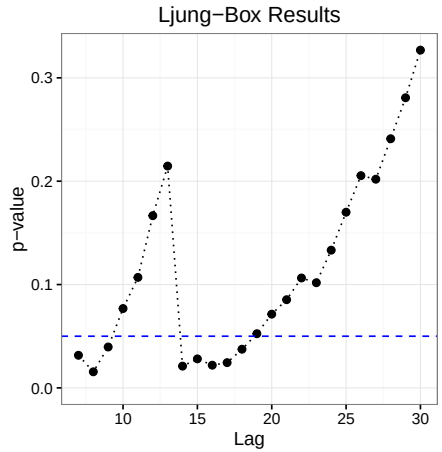




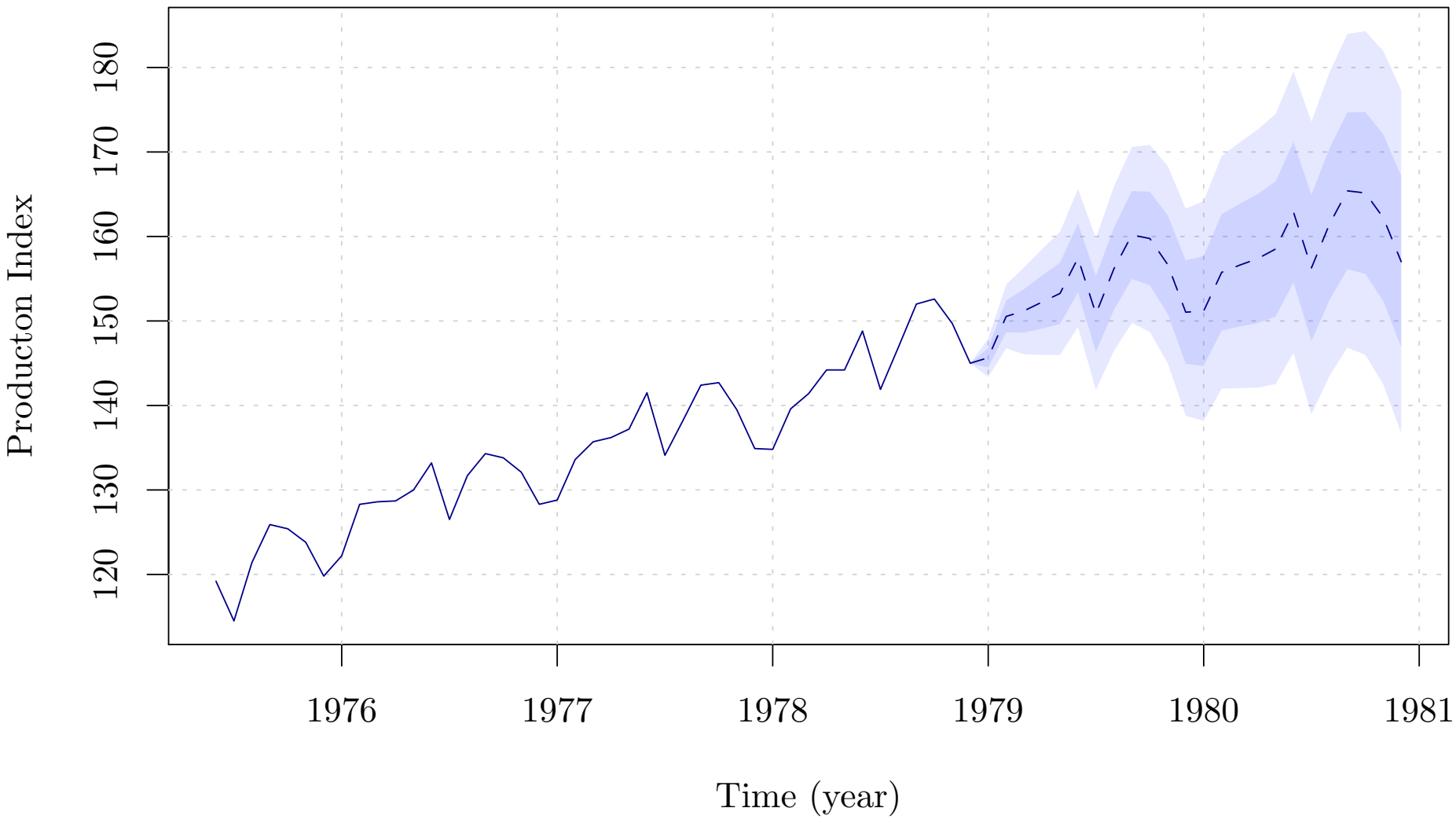








Example: Monthly Production Index Dataset



Statistical Estimators

In this section, we shall present an introduction to the study of the properties of statistical estimators. We will mainly focus on asymptotic properties and start by consider a general class of estimators.

Definition 2.21 (Extremum Estimator).

Many estimators have a **common structure**, which is often useful to study their asymptotic properties. One structure or framework is the class of estimators that maximize some objective function, known as extremum estimators, which can be defined as follows:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \ \hat{Q}_n(\boldsymbol{\theta}), \tag{2.3}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ denote, respectively, the parameter vector of interest and its set of possible values.

Remark:

The vast majority of statistical estimators can be represented as extremum estimators. This is for example the case of least squares, maximum likelihood or (generalized) method of moments estimators.

Example: Least Squares Estimator

Consider the linear model $\boldsymbol{y} = \mathbf{X}\beta_0 + \boldsymbol{\varepsilon}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a full-rank constant matrix, $[\boldsymbol{\varepsilon}]_i \overset{iid}{\sim} (0, \sigma_\varepsilon^2)$ and $\beta \in \mathcal{B} \subseteq \mathbb{R}^p$. Let $\hat{\beta}$ denote the Least Squares Estimator (LSE) of β_0, i.e.

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}.$$

This estimator is an extremum estimator since it can expressed as:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\text{argmax}} \; -||\boldsymbol{y} - \mathbf{X}\beta||_2^2,$$

similarly to our definition given in (2.3). ●

Example: Maximum Likelihood Estimator

Let $Z_1, \ldots, Z_n$ be an iid sample with pdf $f(z|\boldsymbol{\theta}_0)$. The Maximum Likelihood Estimator (MLE) is given by:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \ \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta}\right) \right]. \tag{2.4}$$

Therefore, the MLE can be seen as an example of extremum estimator with:

$$\hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta}\right) \right].$$

●

Remark: In (2.4) we are actually using a *normalized* log-likelihood instead of the actual log-likelihood. This has (in the vast majority of cases) no impact on the estimator but the normalized form is more convenient to use when we let $n \to \infty$.

Example: Generalized Method of Moment

Consider the same iid sample as in the previous example and suppose that there is a vector $\mathbf{g}(z|\boldsymbol{\theta})$ (i.e. a "moment function") such that $\mathbb{E}[\mathbf{g}(z|\boldsymbol{\theta}_0)] = 0$. Then, a possible estimator of $\boldsymbol{\theta}_0$ is:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}} \; - \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(z_i|\boldsymbol{\theta})\right]^T \widehat{\mathbf{W}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(z_i|\boldsymbol{\theta})\right], \tag{2.5}$$

where $\widehat{\mathbf{W}}$ is a positive definite matrix of appropriate dimension. Such estimators are called Generalized Method of Moments (GMM) estimators. They belong to the class of extremum estimators. ●

Remark: Instead of (2.5) we will often consider an alternative (but equivalent) definition of such estimator, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}} \; -||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}},$$

where $||\mathbf{x}||^2_{\mathbf{A}} = \mathbf{x}^T\mathbf{A}\mathbf{x}$, and where $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ denote, respectively, the empirical and model based moments.

Example: A simple GMM Estimator

Let $Z_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$ and $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)^T$. Suppose we want to estimate $\boldsymbol{\theta}_0$ by matching the first three empirical moments with their theoretical counterparts. In this case, a reasonable moment function or condition defining a GMM estimator is given by:

$$\mathbf{g}(Z|\boldsymbol{\theta}) = \begin{bmatrix} Z - \mu \\ Z^2 - \left(\mu^2 + \sigma^2\right) \\ Z^3 - \left(\mu^3 + 3\mu\sigma^2\right) \end{bmatrix}.$$

However, it can be noticed that $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(Z_i|\boldsymbol{\theta}) = \hat{\gamma} - \gamma(\boldsymbol{\theta})$, where $\hat{\gamma}$ and $\gamma(\boldsymbol{\theta})$ denote, respectively, the empirical and model-based moments, i.e.

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} Z_i \\ Z_i^2 \\ Z_i^3 \end{bmatrix}, \quad \gamma(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \\ \mu^3 + 3\mu\sigma^2 \end{bmatrix}.$$

Then, we can write our GMM estimator of $\boldsymbol{\theta}_0$ as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmin}} \; ||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}} \; -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}}. \tag{2.6}$$

Consistency of Statistical Estimators

In the next slides we will discuss the conditions for the consistency of extremum estimator, which is often denoted as $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$. We start by defining consistency.

Definition: Consistency

The estimator $\hat{\boldsymbol{\theta}}$ is said to be consistent if it converges in probability to $\boldsymbol{\theta}_0$, i.e.

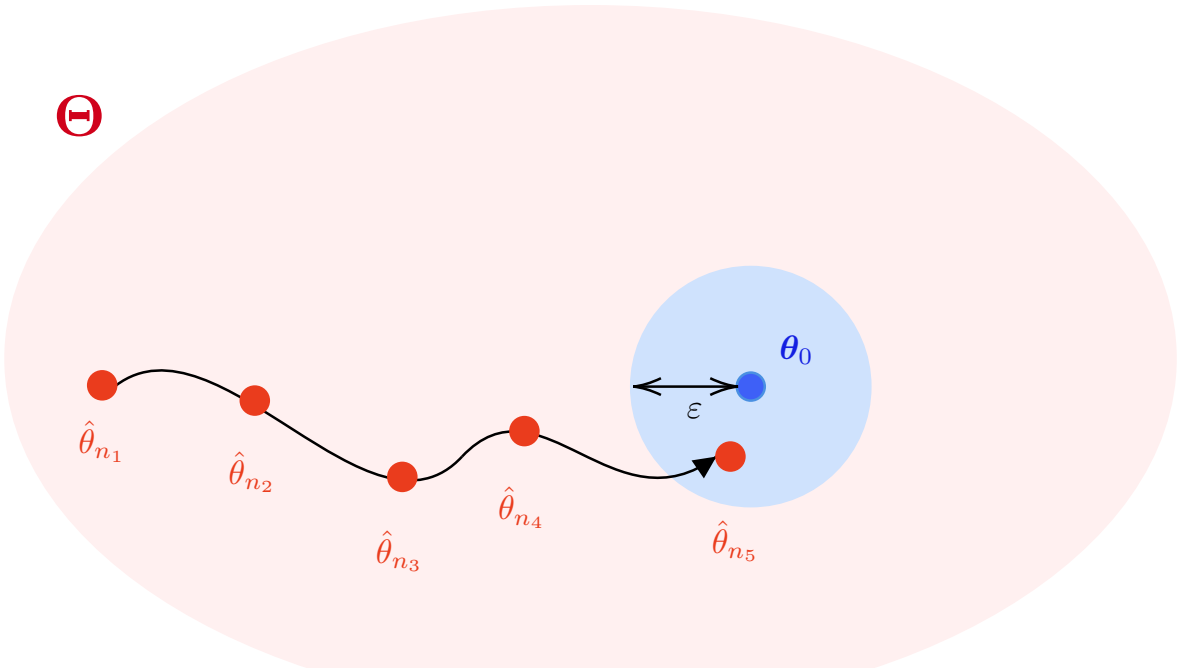
$$\lim_{n \to \infty} \Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon \right) = 0,$$

for all $\varepsilon > 0$.

Consistency - Interpretation

Interpretation:

In Layman's term consistency simply means that if n is "large enough" $\hat{\boldsymbol{\theta}}$ will be **arbitrarily close to θ_0** (i.e. inside of an hypersphere of radius ε centered at $\boldsymbol{\theta}_0$). This also means that the procedure (i.e. our estimator) based on unlimited data will be able to identify the underlying truth (i.e. $\boldsymbol{\theta}_0$).



Related Theorems

Consistency is often proven using the following two important results.

Theorem 2.22 (Weak Law of Large Numbers).

Suppose X_i are iid random variables with finite mean μ (i.e. $\mathbb{E}[X_i] = \mu$) and finite variance. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$.

Theorem 2.23 (Continuous Mapping Theorem).

Suppose $Y_n \xrightarrow{\mathcal{P}} \mu$, then $g(Y_n) \xrightarrow{\mathcal{P}} g(\mu)$ if $g(\cdot)$ is a continuous function.

A simple example on the consistency is presented in Appendix E

▸ Go to Appendix E.

Consistency of Extremum Estimators

When considering real-life problems, the approach based on Theorem 2.22 presented in Appendix E is in general not flexible enough and we generally rely on the results such as the one presented below.

Theorem 2.24 (Consistency of Extremum Estimators).

If there is a function $Q_0(\theta)$ such that:

(C.1) *$Q_0(\theta)$ is uniquely maximized in θ_0,*

(C.2) *Θ is compact*[a]*,*

(C.3) *$Q_0(\theta)$ is continuous in θ,*

(C.4) *$\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$*[b]*.*

then we have $\hat{\theta} \xrightarrow{\mathcal{P}} \theta$.

---

[a] Compact means that Θ is both closed (i.e. containing all its limit points) and bounded (i.e. all its points are within some fixed distance of each other).

[b] $\hat{Q}_n(\theta)$ is said to converges uniformly in probability to $Q_0(\theta)$ if $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{\mathcal{P}} 0$

Remarks: Consistency of Extremum Estimators

This theorem is an important result and provides a general approach to prove the consistency of a large class of estimators. A few remarks on the conditions of this result:

- Condition (C.1) is substantive and there are well-known examples where it fails. We will discuss further on how this assumption can (in some cases) be verified in practice.
- Condition (C.2) is also substantive as it requires that there exist some known bounds on the parameters. In practice, this assumption is often neglected although it is in most cases unrealistic to assume it.
- Conditions (C.3) and (C.4) are often referred to as "standard regularity conditions". They are typically satisfied. The verification of these conditions will be discussed further in this document.

Theorem 2.24: Sketch of the proof

The basic idea of the proof is the following. Under Condition (C.1) we have:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; Q_0(\boldsymbol{\theta}).$$

Condition (C.4) implies that:

$$\hat{Q}_n(\boldsymbol{\theta}) \xmapsto{\mathcal{P}} Q_0(\boldsymbol{\theta}),$$

therefore, it seems logical that:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; \hat{Q}_n(\boldsymbol{\theta}) \xmapsto{\mathcal{P}} \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; Q_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0.$$

Unfortunately, without additional conditions this simple proof is not correct. The main reason is that additional conditions (i.e. uniform convergence as well as Conditions (C.2) and (C.3)) are needed to ensure the validity of the above convergence given in orange. A formal proof of this result is given in Appendix F and is of course good to know! ▸ Go to Theorem Appendix F

Verification of Condition (C.1)

In general the verification of Condition (C.1) is difficult and is often assumed in the statistical literature. We present here two results, namely Lemma 2.25 from Newey and McFadden 1994 and Theorem 2 from Komunjer 2012, which allow the verification of this condition for GMM-type estimators (on which we shall focus here). A discussion on the verification of this condition for other estimators can for example be found in Chapter 7 of Baltagi 2008. The proof of Lemma 2.25 is given in Appendix G ▸ Go to Appendix G.

Lemma 2.25 (Identification GMM).

If $\mathbf{W} > 0$ *(i.e. positive definite) (where* $\mathbf{W}$ *is such that* $\widehat{\mathbf{W}} \xrightarrow{\mathcal{P}} \mathbf{W}$*),*
$\mathbf{g}_0(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{g}(z|\boldsymbol{\theta})]$, $\mathbf{g}_0(\boldsymbol{\theta}_0) = \mathbf{0}$ *and* $\mathbf{g}_0(\boldsymbol{\theta}) \neq \mathbf{0}$ *if* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *then*
$Q_0(\boldsymbol{\theta}) = -\mathbf{g}_0(\boldsymbol{\theta})^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta})$ *has a unique maximum at* $\boldsymbol{\theta}_0$.

Remark:

If we write a GMM estimator as in (2.6) then the condition $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$, if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, can be replaced by $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\theta}_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Verification of Condition (C.1)

Therefore, Lemma 2.25 shows that a GMM estimator verify Condition (C.1) in the case where $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbf{0}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (or alternatively $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\theta}_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$). The following theorem (which is a "simplified" version of Theorem 2 of Komunjer 2012) provides us with a way to verify this new condition.

Theorem 2.26 (Homeomorphism).

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Let $\mathbf{g}^*(\boldsymbol{\theta})$ denote a subset of p elements of $\mathbf{g}_0(\boldsymbol{\theta}) \in \mathbb{R}^q$, $q \geq p$ such that:

- *$\mathbf{g}^*(\boldsymbol{\theta})$ is in $\mathcal{C}^2$ (i.e. $\mathbf{g}^*(\boldsymbol{\theta})$ can be differentiated twice).*
- *For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $J(\boldsymbol{\theta})$ is nonnegative (or alternatively nonpositive), where $J(\boldsymbol{\theta}) := \det\left(\frac{\partial}{\partial \boldsymbol{\theta}^\intercal} \mathbf{g}^*(\boldsymbol{\theta})\right)$.*
- *$||\mathbf{g}^*(\boldsymbol{\theta})|| \to \infty$ whenever $||\boldsymbol{\theta}|| \to \infty$.*
- *For every $s \in \mathbb{R}^p$, the equation $\mathbf{g}^*(\boldsymbol{\theta}) = s$ has countably many (possibly zero) solutions in $\boldsymbol{\Theta}$.*

Then, $\mathbf{g}^*(\boldsymbol{\theta})$ is a Homeomorphism (i.e. $\mathbf{g}^*(\boldsymbol{\theta})$ is continuous and one-to-one).

Discussion on Theorem 2.26

Remarks:

- A direct consequence of Lemma 2.25 and Theorem 2.26 is that any GMM estimator with $\mathbf{W} > 0$, satisfying the conditions of Theorem 2.26, satisfies Condition (C.1) of Theorem 2.24.
- In addition, if one can show that $\mathbf{g}^c(\boldsymbol{\theta})$ is in $\mathcal{C}$ (where $\mathbf{g}^c(\boldsymbol{\theta})$ denotes the element of $\mathbf{g}_0(\boldsymbol{\theta})$ that are not in $\mathbf{g}^*(\boldsymbol{\theta})$) then Condition (C.3) of Theorem 2.24 is also verified.
- Note that when considering a GMM estimator of the form used in (2.6), one can simply verify the conditions of Theorem 2.24 with $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\theta})$.
- In Theorem 2 in Komunjer 2012 it is actually assumed that $\boldsymbol{\theta} \in \mathbb{R}^p$ while we assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We used this simplification to avoid an overly technical treatment of this topic. In fact, we assume here that there exists a one-to-one function $h(\cdot)$ such that $h : \mathbb{R}^p \mapsto \boldsymbol{\Theta}$. This condition is typically verified in practice.

Example: Proving Condition (C.1)

We revisit now the example presented in ▸ Example GMM. Let us say that $\widehat{\mathbf{W}}$ is such that $\widehat{\mathbf{W}} \xrightarrow{\mathcal{P}} \mathbf{W} > 0$. Then, we showed that:

$$\boldsymbol{\gamma}(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \\ \mu^3 + 3\mu\sigma^2 \end{bmatrix}.$$

Therefore, we define:

$$\mathbf{g}^*(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix} \quad \text{and} \quad g^c(\boldsymbol{\theta}) = \left[\mu^3 + 3\mu\sigma^2\right].$$

Since the elements of $\mathbf{g}^*(\boldsymbol{\theta})$ are polynomial in $\boldsymbol{\theta}$, the condition $\mathbf{g}^*(\boldsymbol{\theta}) \in \mathcal{C}^2$ is trivially satisfied. Next, we define:

$$\mathbf{A}(\boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}^T} \boldsymbol{\gamma}(\boldsymbol{\theta}) \in \mathbb{R}^{3\times 2} \quad \text{and} \quad \widetilde{\mathbf{A}}(\boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}^T} \mathbf{g}^*(\boldsymbol{\theta}) \in \mathbb{R}^{2\times 2}.$$

Example: Proving Condition (C.1)

Then, the determinant of $\widetilde{\mathbf{A}}(\boldsymbol{\theta})$ is equal to 1 as illustrated in the diagram below:

$$\mathbf{A}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\gamma}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & 0 \\ 2\mu & 1 \\ 3\left(\mu^2+\sigma^2\right) & 3\mu \end{bmatrix}, \qquad \det\left(\widetilde{\mathbf{A}}(\boldsymbol{\theta})\right) = 1.$$

Finally, the last two conditions of 2.26 are trivially satisfied since $||\mathbf{g}^*(\boldsymbol{\theta})||$ can only diverge if $||\boldsymbol{\theta}||$ diverges and since $\mathbf{g}^*(\boldsymbol{\theta}) = s$ has (one) countably solutions in $\boldsymbol{\Theta}$.

Then, it follows from Lemma 2.25 and Theorem 2.26 that the function $Q_0(\boldsymbol{\theta}) = -||\boldsymbol{\gamma}(\boldsymbol{\theta}_0) - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\mathbf{W}}$ is uniquely maximized in $\boldsymbol{\theta}_0$. ●

Verification of Condition (C.4)

In general, establishing the uniform convergence of $\hat{Q}_n(\boldsymbol{\theta})$ to $Q_0(\boldsymbol{\theta})$, i.e.

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0, \tag{2.7}$$

is not easy. A common strategy is the following:

- Show that $\hat{Q}_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} Q_0(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (i.e. pointwise convergence).
- Show that $\hat{Q}_n(\boldsymbol{\theta})$ is almost surely Lipschitz continuous, i.e.

$$\sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}_1) - \hat{Q}_n(\boldsymbol{\theta}_2)| \leq H||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||,$$

  where H is a random variable which is almost surely bounded, i.e. there exists a constant c such that $|H| < c$ almost surely.

Verification of Condition (C.4)

Then, the following theorem (which is a slightly adapted version of the Arzela-Ascoli Theorem) allows the verification of Condition (C.4).

Theorem 2.27 (modified Arzela-Ascoli).

Suppose that Θ is compact, for every $\theta \in \Theta$ we have $\hat{Q}_n(\theta) \xrightarrow{\mathcal{P}} Q_0(\theta)$ and $\hat{Q}_n(\theta)$ is almost surely Lipschitz continuous. Then, we have $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$.

Remark:

In this course, we will not discuss how to prove that $\hat{Q}_n(\theta)$ is almost surely Lipschitz continuous and we assume it for simplicity (more discussion on this topic can for example be found in Newey and McFadden 1994). Nevertheless, it is worth mentioning that this condition is almost always satisfied in practice and is therefore "reasonable" to assume.

In Appendix J we provide a practical example on how to prove (C.4) ▸ Go to Appendix J.

Law of Large Numbers for Dependent Processes

In general showing that $\hat{Q}_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} Q_0(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is done using one version of the law of large numbers (see e.g. Appendix J).

However, when X_i which are not iid random variables, Theorem 2.22 cannot be applied. The following theorem (taken from Proposition 7.5 of Hamilton 1994) generalizes this result for (weak) stationary processes with absolutely summable covariance structure (see Definition 2.18). A simple version of this proof based on Chebychev's inequality is given in Appendix H ▸ Go to Appendix H.

Theorem 2.28 (Weak Law of Large Numbers for Dependent Processes).

Suppose (X_t) is a (weak) stationary process with absolutely summable autocovariance structure, then

$$\bar{X}_T \xrightarrow{\mathcal{P}} \mathbb{E}[X_t].$$

Example: Consistency of the Sample Mean

Consider a stationary AR1 process and suppose we wish to study whether its sample mean converges in probability to its expected value. This time we consider consider a non-zero mean AR1, i.e.

$$(X_t - \mu) = \phi (X_{t-1} - \mu) + Z_t,$$

where $\mu = \mathbb{E}[X_t]$, $Z_t \overset{iid}{\sim} \mathcal{N}(0, \nu^2)$ and $\nu^2 < \infty$. This process can also be written as a linear process (see Definition 2.17):

$$X_i - \mu = \sum_{k=0}^{\infty} \phi^k Z_{i-k},$$

Since the process is stationary for $|\phi| < 1$ (see Appendix B), we have:

$$\gamma_h = \frac{\nu^2 \phi^{|h|}}{1 - \phi^2},$$

for $h \in \mathbb{Z}$.

Example: Consistency of the Sample Mean

Then, we have that:

$$\sum_{h=-\infty}^{\infty} |\gamma_k| = \sum_{h=-\infty}^{\infty} \frac{\nu^2 |\phi|^{|h|}}{1-\phi^2} < 2 \lim_{n\to\infty} \frac{\nu^2(1-|\phi|^{n+1})}{(1-\phi^2)(1-|\phi|)} = \frac{2\nu^2}{(1-\phi^2)(1-|\phi|)} < \infty,$$

implying that the process has an absolutely summable covariance structure (see Definition 2.18). Therefore applying Theorem 2.28, we can verify that:

$$\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{\mathcal{P}} \mathbb{E}[X_t] = \mu.$$

●

Consistency of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$

In Definitions 2.19 and 2.20 we defined the sample autocovariance and autocorrelation functions. In the following corollary we show that these estimators are both consistent. The proof of these results can be found in most time series textbooks but is also given in Appendix I to illustrate the use of Theorem 2.28 ▸ Go to Appendix I.

Corollary 2.29 (Consistency of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$).

Let (X_t) be such that:

- *(X_t) is weakly stationary,*
- *(X_t^2) has an absolutely summable covariance structure,*

then for all $|h| < \infty$ we have

$$\hat{\gamma}(h) \xrightarrow{\mathcal{P}} \gamma(h),$$

$$\hat{\rho}(h) \xrightarrow{\mathcal{P}} \rho(h).$$

Asymptotic Normality

The Central Limit Theorem (CLT) takes one step further than the law of large numbers. It identifies the limiting distribution of the (properly scaled) sum of random variables as a normal distribution, which allows us to do statistical inference (confidence intervals and hypothesis testing). The scale will tell us how fast this approximation converges to the normal distribution. These results are generally based on CLT with two most significant theorems used to prove asymptotic normality.

Theorem 2.30 (CLT for iid sequences).

Suppose X_i *are iid random variables with* $\mathbb{E}[X_i] = \mu$ *and* $\text{var}(X_i) = \sigma^2 < \infty$. *Let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, *then* $\sqrt{n}\left(\bar{X}_n - \mu\right) \overset{\mathcal{D}}{\longmapsto} \mathcal{N}(0, \sigma^2)$.

The above result can be extended to finite dimensional multivariate iid sequences by the Cramér-Wold device as follows.

Theorem 2.31 (CLT for iid multivariate sequences).

Suppose $\boldsymbol{X}_i$ *are iid random variables with* $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu} \in \mathbb{R}^d$ *and* $\text{cov}(\boldsymbol{X}_i) = \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. *Then, we have* $\sqrt{n}\left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \overset{\mathcal{D}}{\longmapsto} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Asymptotic Normality of Extremum Estimators

General results on the asymptotic normality of extremum estimators can be found in Newey and McFadden 1994. In this course, we shall restrict our attention to a simple example of GMM estimators ▸ Go to example. We will see that the asymptotic normality of extremum estimators is implied by combining the following results and techniques:

- the consistency of $\hat{\boldsymbol{\theta}}$,
- the Central Limit Theorem (see Theorem 2.31),
- Slutsky's Theorem (which allows to "mix" convergence in probability and distribution) and
- Taylor expansions.

Asymptotic Normality of Extremum Estimators

In the simple example we considered on GMM estimators, we have used Theorem 2.31 with:

$$\sqrt{n}\left(\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{bmatrix} Z_i - \mu_0 \\ Z_i^2 - \mu_0^2 - \sigma_0^2 \\ Z_i^3 - \mu_0^3 - 3\mu_0\sigma_0^2 \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \qquad (2.8)$$

where $\boldsymbol{\Sigma} = \text{cov}\left(\begin{bmatrix} Z_i & Z_i^2 & Z_i^3 \end{bmatrix}^T\right)$. In this example, the extremum estimator we consider is defined using the objective function:

$$Q_n(\boldsymbol{\theta}) = -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}}.$$

Central Limit Theorem

Then we can relate the asymptotic normality of the GMM estimator with the asymptotic normality of the moment conditions (i.e. Equation (2.8)) we have just shown. Since $\hat{\boldsymbol{\theta}}$ maximizes $Q_n(\boldsymbol{\theta})$, it satifies the following first order condition, i.e.

$$\left.\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}_3,$$

which is equivalent to:

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \left((\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta}))\right)^T \widehat{\mathbf{W}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right)\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}_3. \tag{2.9}$$

Using a Taylor expansion for $\boldsymbol{\gamma}(\hat{\boldsymbol{\theta}})$ around the true $\boldsymbol{\theta}_0$, we obtain:

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta}_0) + \left.\frac{\partial \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{2.10}$$

Under certain regularity conditions, we have (using Theorem 2.23) that:

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right)\right|_{\boldsymbol{\theta}} = -\left.\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\gamma}(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \xrightarrow{p} -\left.\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\gamma}(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Central Limit Theorem

After plugging in Equation (2.10) into Equation (2.9), and under certain regularity conditions and using Slutsky's Theorem, we obtain:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{D}^T \boldsymbol{\Sigma} \mathbf{D}\right), \tag{2.11}$$

where

$$\mathbf{D} = \left[||\frac{\partial}{\partial \boldsymbol{\theta}}\left(\gamma(\boldsymbol{\theta})\right)||\mathbf{w}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]^{-1}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\gamma(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)^T \mathbf{W}.$$

The derivation of (2.11) is due to the fact that:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) = -\overbrace{\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}}\left(\hat{\gamma}-\gamma(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^T \widehat{\mathbf{W}}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\left(\hat{\gamma}-\gamma(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)\right]^{-1}}^{\xrightarrow{p}\left[||\frac{\partial}{\partial \boldsymbol{\theta}}(\hat{\gamma}-\gamma(\boldsymbol{\theta}))||\mathbf{w}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]^{-1}}$$

$$\underbrace{\left(\frac{\partial}{\partial \boldsymbol{\theta}}\left(\hat{\gamma}-\gamma(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^T \widehat{\mathbf{W}}}_{\xrightarrow{p}\left(\frac{\partial}{\partial \boldsymbol{\theta}}(\hat{\gamma}-\gamma(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)^T \mathbf{W}} \sqrt{n}\left(\hat{\gamma}-\gamma(\boldsymbol{\theta}_0)\right) + o_p(1).$$

Central Limit Theorem and α-Mixing

For dependent processes, the validity of CLT requires the process to be "mixing" or "asymptotic independent". Suppose two events G and H are independent, then

$$|\Pr(G \cap H) - \Pr(G)\Pr(H)| = 0.$$

Based on this idea, many dependence measures have been developed. The most used is α-mixing coefficient, which is one of these dependence measures that can be easily verified for certain stochastic processes.

Definition: Mixing Coefficients

For a stochastic process $\{X_i\}_{i\in\mathbb{Z}}$, define the strong- or α-mixing coefficients as

$$\alpha(t_1, t_2) = sup\{|\Pr(A \cap B) - \Pr(A)\Pr(B)| : \ A \in \mathcal{F}_{-\infty}^{t_1}, B \in \mathcal{F}_{t_2}^{\infty}\},$$

where $\mathcal{F}_{-\infty}^{t_1} = \sigma(X_{-\infty}, \ldots, X_{t_1})$ and $\mathcal{F}_{t_2}^{\infty} = \sigma(X_{t_2}, \ldots, X_{\infty})$ are σ-algebras generated by the corresponding random variables.

If the process is stationary, then $\alpha(t_1, t_2) = \alpha(t_2, t_1) = \alpha(|t_1 - t_2|) := \alpha(\tau)$. If $\alpha(\tau) \to 0$ as $\tau \to \infty$, then the process is strong-mixing or α-mixing.

Central Limit Theorem and α-Mixing

Theorem 2.32 (Central Limit Theorem for α-Mixing Process).

Let (X_t) be a strictly stationary process with $\mathbb{E}[X_t] = 0$. $S_n := \sum_{t=1}^{n} X_t$ is the partial sum process with $\sigma_n^2 := \text{var}(S_n)$. Suppose (X_t) is α-mixing, and that for $\delta > 0$

$$\mathbb{E}\left[|X_t|^{2+\delta}\right] \leq \infty, \text{ and } \sum_{n=0}^{\infty} \alpha(n)^{\delta/2+\delta} \leq \infty.$$

Then

$$\lim_{n\to\infty} \frac{\sigma_n^2}{n} = \mathbb{E}\left[|X_t|^2\right] + 2\sum_{k=1}^{\infty} \mathbb{E}\left[X_1 X_k\right] := \sigma^2.$$

If $\sigma^2 > 0$, (X_t) obeys both the central limit theorem with variance σ^2 and the functional central limit theorem.

Remark 16 (Implication of α-mixing).

α-mixing $\Rightarrow$ the covariance structure is absolutely summable (see Definition 2.18).

Implication on the estimation of $\gamma(h)$ and $\rho(h)$

Using the results previously presented, let us quickly revisit the estimation and the inference of $\gamma(h)$ and $\rho(h)$.

Remark 17 (Significance of $\hat{\rho}(h)$).

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 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.

Theorem 2.33 (Distribution of $\hat{\rho}(h)$ in iid case).

If (X_t) is white noise (with finite variance) and $h = 1, ..., H$ where H is fixed but arbitrary we have that:

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

Implication on the estimation of $\gamma(h)$ and $\rho(h)$

Remark 18 (Confidence intervals for $\hat{\rho}(h)$).

Theorem 2.33 implies that an approximate confidence interval for $\hat{\rho}(h)$ (in the iid case) is given by:

$$\text{CI}(\rho(h), \alpha) = \hat{\rho}(h) \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{T}}$$

for $0 < h < k < \infty$ and where $z_{1-\frac{\alpha}{2}} := \mathbf{\Phi}^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the $\left(1 - \frac{\alpha}{2}\right)$ quantile of a standard normal distribution. Typically, for $\alpha = 0.05$ one would consider the following confidence interval:

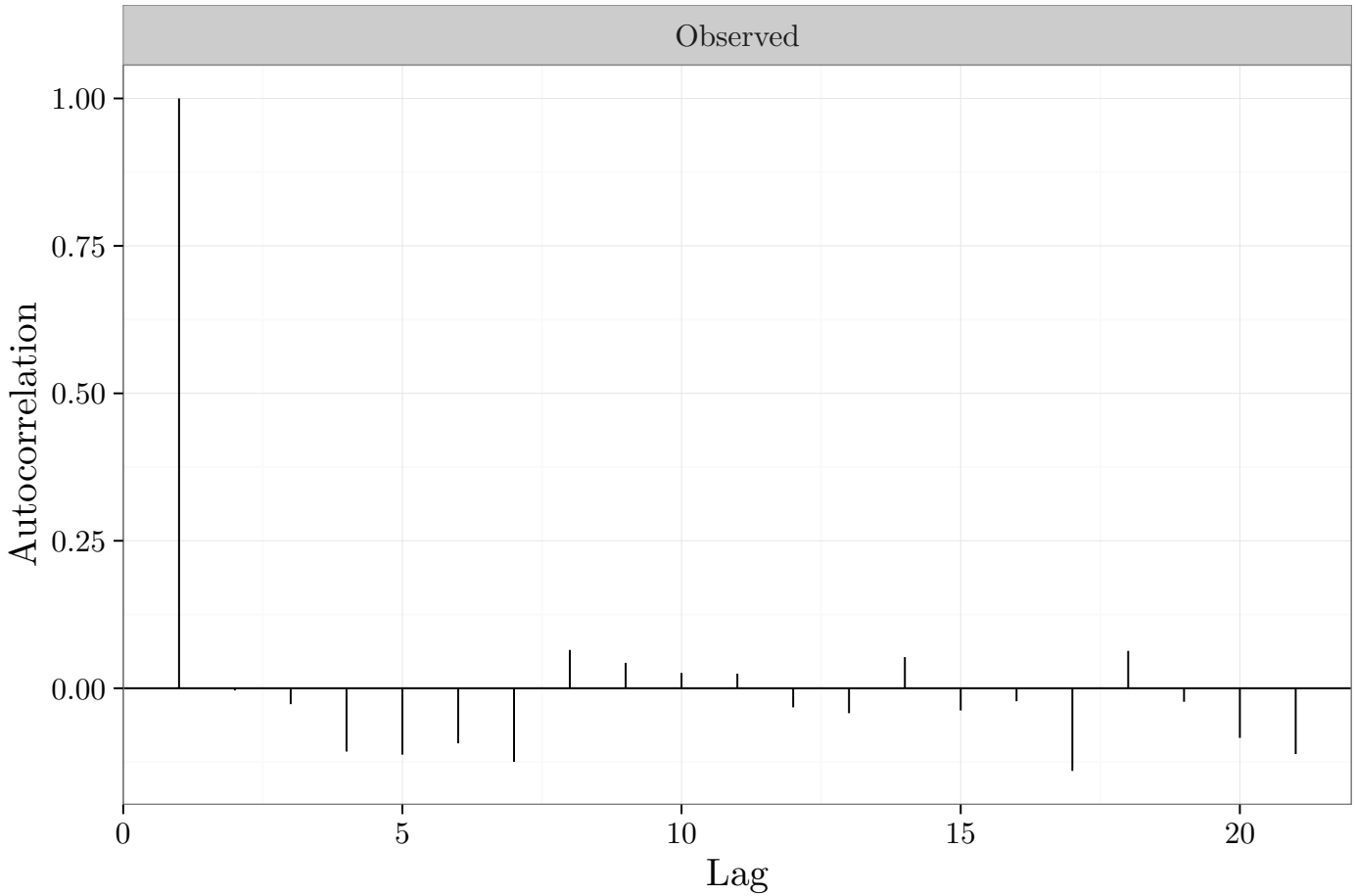
$$\text{CI}(\rho(h), 0.05) = \hat{\rho}(h) \pm \frac{2}{\sqrt{T}}$$

Remark 19 (Proof of Theorem 2.33).

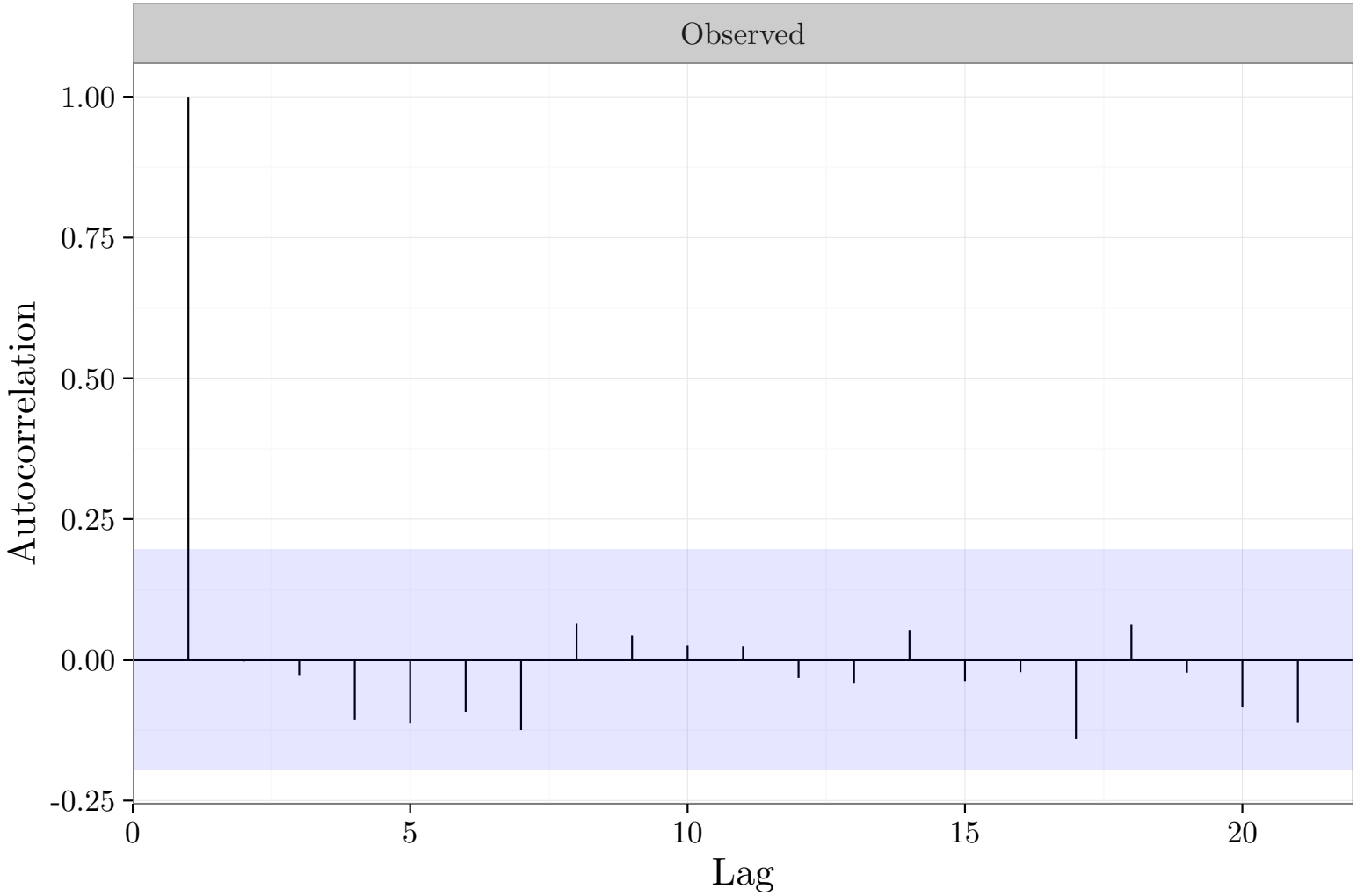
The proof of Theorem 2.33 is straightforward from the CLT and Delta method. It is therefore omitted from this document but can for example be found in Hamilton 1994.

Example: Sample Autocorrelation Function

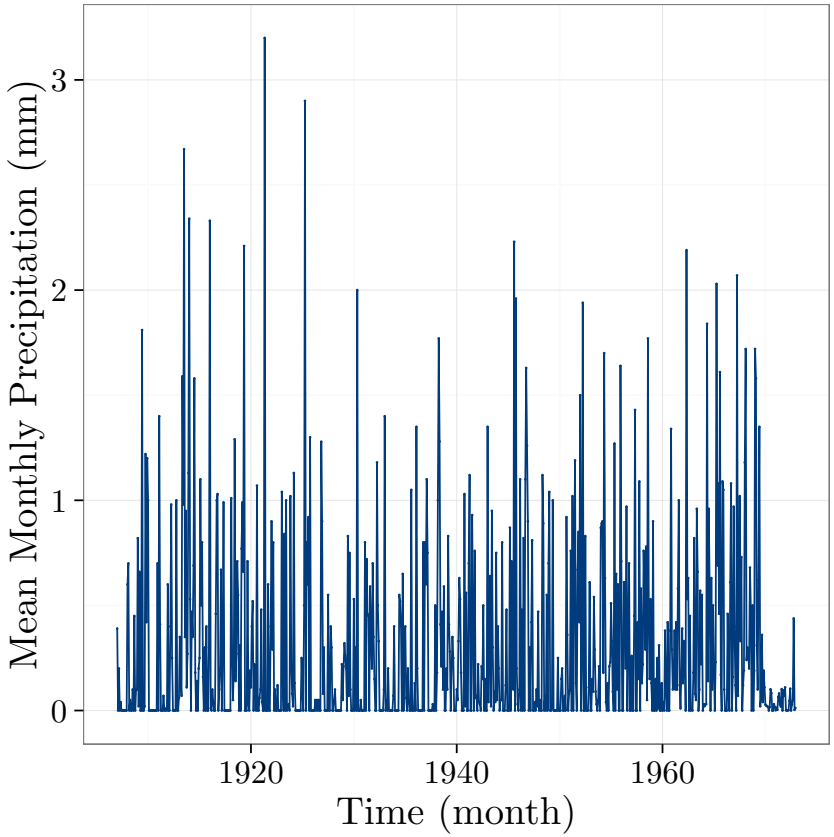
Consider the estimated ACF of a simulated WN process of length $T = 100$.

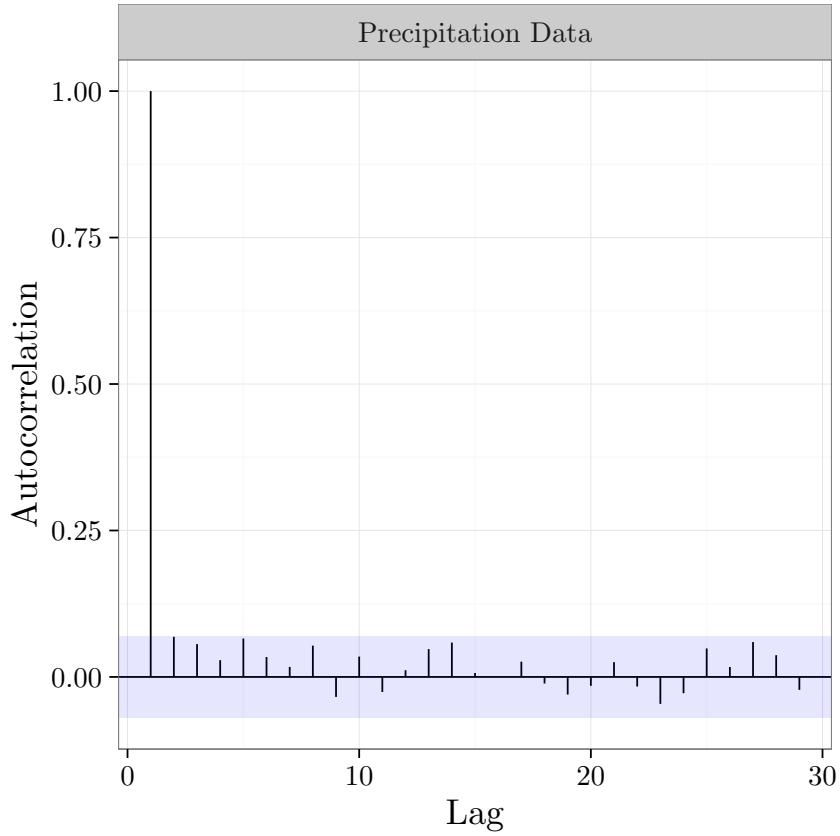


An Example: a White Noise Process



Example: ACF of Precipitation Data





Remark:

The "ACF" plot suggests an absence of linear dependence in this dataset.

References I

Baltagi, Badi H (2008). *A Companion to Theoretical Econometrics*. John Wiley & Sons.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods". In: *Econometrica: Journal of the Econometric Society*, pp. 424–438.

Hamilton, J. D. (1994). *Time Series Analysis*. Vol. 2. Princeton university press Princeton.

Komunjer, I. (2012). "Global Identification in Nonlinear Models with Moment Restrictions". In: *Econometric Theory* 28.4, p. 719.

Newey, W. K. and D. McFadden (1994). "Large Sample Estimation and Hypothesis Testing, V in Handbook of Econometrics". In: vol. 4. Elsevier, Amsterdam.

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:

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

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

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

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

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

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

$$\vdots$$

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

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

$$X_t = \lim_{k \to \infty} X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}, \tag{2.12}$$

since $|\phi| < 1$. This shows that the 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 (2.2)

The derivation of (2.2) allows us to introduce a common technique used in time series analysis. Let $\mathbf{1}$ denotes a unit vector of dimension T and $\mathbf{X} = [X_1, ..., X_T]^T$, 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 $\mathbf{Y} \in \mathbb{R}^k$ is a random variable and $\mathbf{A} \in \mathbb{R}^{h\times k}$ a fixed matrix we have:

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

Therefore, we have:

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

Appendix C: Derivation of Equation (2.2)

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:

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

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

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

It is worth noting that in the iid case we have that $\text{var}\left(\bar{X}\right) = \frac{1}{T}\,\text{var}(X_1)$. This result can naturally be obtained using (2.2) since $\text{var}(X_1) = \gamma(0)$ and $\sum_{i=T}^{T}\left(1-\frac{|i|}{T}\right)\rho(i) = 1$.

▸ Return to Equation (2.2)

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

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

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

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

$$\text{var}\left(\bar{X}\right) = \frac{\nu^2\left(T - 2\phi - T\phi^2 + 2\phi^{T+1}\right)}{T^2\left(1-\phi^2\right)\left(1-\phi\right)^2}. \tag{2.13}$$

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

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

providing the following approximate formula:

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

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

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

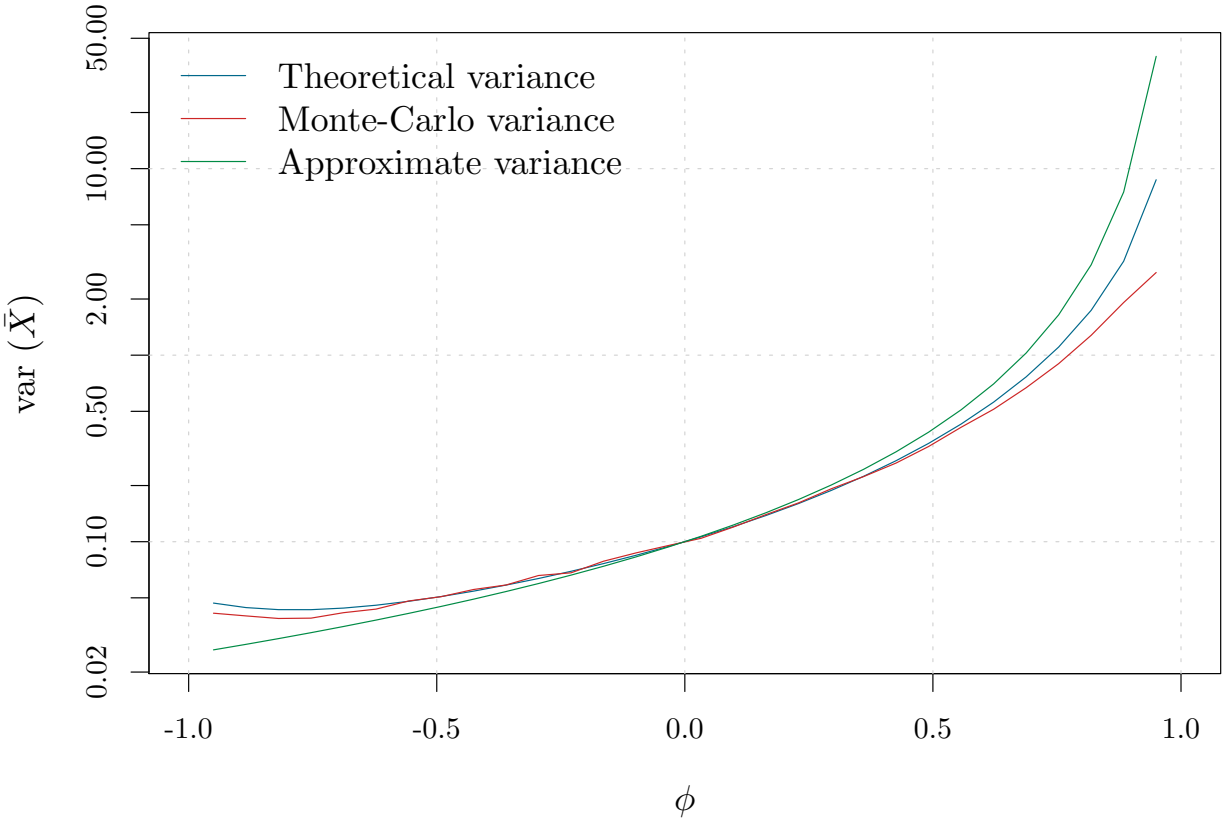
- Step 1: Simulate under the assumed model, i.e. $X_t^* \sim F_{\theta_0}$, where F_{θ_0} denotes the true model (in this case an AR1 process).
- Step 2: Compute $\bar{X}^*$ (i.e. average based on (X_t^*)).
- Step 3: Repeat Steps 1 and 2 B times.
- Step 4: Compute the empirical variance $\bar{X}^*$ (based 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 var $\left(\bar{X}\right)$ in practice?

A numerical experiment

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



Appendix E: Consistency - a simple example

Let $X_i \overset{iid}{\sim} \mathcal{E}(\lambda_0),\ \lambda_0 \in \mathbb{R}^+,\ i = 1, ..., n$. We wish to show that the MLE for λ_0 is consistent. Then, we have that the density of X is given by (assuming $X \geq 0$):

$$f(x|\lambda) = \lambda \exp(-\lambda x).$$

Therefore, the normalized log-likelihood function is given by

$$\mathcal{L}(\lambda|X_1, ..., X_n) = \log(\lambda) - \lambda \bar{X}_n,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. By taking the first derivative we obtain:

$$\frac{\partial}{\partial \lambda}\mathcal{L}(\lambda|X_1, ..., X_n) = \frac{1}{\lambda} - \bar{X}_n,$$

which implies that MLE is such that $\frac{\partial}{\partial \lambda}\mathcal{L}(\hat{\lambda}|X_1, ..., X_n) = 0$. Then:

$$\frac{1}{\hat{\lambda}} - \bar{X}_n = 0 \implies \hat{\lambda} = \bar{X}_n.$$

Appendix E: Consistency - a simple example

Finally, we verify that

$$\frac{\partial^2}{\partial \lambda^2} \mathcal{L}(\lambda | X_1, ..., X_n) = -\frac{1}{\lambda^2} < 0,$$

implying that $\hat{\lambda}$ is the maxima of $\mathcal{L}(\lambda | X_1, ..., X_n)$. Therefore, the MLE is a function of the sample mean $\bar{X}_n$. In this case, the consistency of $\hat{\lambda}$ is implied by Theorems 2.22 and 2.23. Indeed, it follows from the Weak Law of Large Numbers (i.e. Theorem 2.22) that $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$, where μ is given by:

$$\mu = \mathbb{E}[X_i] = \int_0^\infty x \lambda_0 \exp(-\lambda_0 x)\, dx = \frac{1}{\lambda_0}.$$

Since the function $f(x) = 1/x$ is continuous in $\mathbb{R}^+$, by the Continuous Mapping Theorem (i.e. Theorem 2.23) we know that $\hat{\lambda} \xrightarrow{\mathcal{P}} \lambda_0$, which concludes our example. ●

▸ Return to Theorem 2.22

Appendix F: Proof of Theorem 2.24

Let $\mathcal{G}$ be the ε-ball centered at $\boldsymbol{\theta}_0$ i.e. $\mathcal{G} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||_2 < \varepsilon\}$ for some $\varepsilon > 0$. Then $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$ is equivalent to:

$$\lim_{n\to\infty} \Pr(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon) = \lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) = 0.$$

We define $\gamma = Q(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} Q(\boldsymbol{\theta})$ which is strictly positive by 2.24. Then we have that $\hat{\boldsymbol{\theta}} \notin \mathcal{G}$ implies:

$$Q(\hat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}_0) - \gamma$$

and therefore:

$$\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) \leq \lim_{n\to\infty} \Pr(\mathcal{A})$$

where $\mathcal{A} = \left\{ Q(\hat{\boldsymbol{\theta}}) \leq Q(\boldsymbol{\theta}_0) - \gamma \right\}$.

Appendix F: Proof of Theorem 2.24

Next, we define the following events:

$$\mathcal{B} = \left\{ \left| Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}}) \right| > \gamma/3 \right\}$$
$$\mathcal{C} = \{ |Q_n(\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0)| > \gamma/3 \}$$
$$\mathcal{D} = \left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \gamma/3 \right\}$$

and we have that:

$$\begin{aligned} \Pr(\mathcal{A}) &\leq \Pr(\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C})) = \Pr\left(\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C})^c\right) + \Pr(\mathcal{B} \cup \mathcal{C}) \\ &= \Pr(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c) + \Pr(\mathcal{B} \cup \mathcal{C}). \end{aligned}$$

Appendix F: Proof of Theorem 2.24

It is easy to verify that $\Pr(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c) = \Pr(\varnothing)$ because if $\mathcal{A}$, $\mathcal{B}^c$ and $\mathcal{C}^c$ occur simultaneously we have that:

$$\begin{aligned} Q_n(\hat{\boldsymbol{\theta}}) &= Q(\hat{\boldsymbol{\theta}}) + Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}}) \leq Q(\hat{\boldsymbol{\theta}}) + |Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}})| \\ &\leq Q(\hat{\boldsymbol{\theta}}) + \gamma/3 \leq Q(\boldsymbol{\theta}_0) - 2\gamma/3 = Q(\boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_0) + Q_n(\boldsymbol{\theta}_0) - 2\gamma/3 \\ &\leq |Q(\boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_0)| + Q_n(\boldsymbol{\theta}_0) - 2\gamma/3 \leq Q_n(\boldsymbol{\theta}_0) - \gamma/3 < Q_n(\boldsymbol{\theta}_0) \end{aligned}$$

which contradicts (2.3). Moreover, $\Pr(\mathcal{B} \cup \mathcal{C})$ can be bounded as follow:

$$\begin{aligned} \Pr(\mathcal{B} \cup \mathcal{C}) &= \Pr\left(\sup_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \gamma/3 \right) \\ &\leq \Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \gamma/3 \right) = \Pr(\mathcal{D}). \end{aligned}$$

Appendix F: Proof of Theorem 2.24

Finally, we have that:

$$\begin{aligned}\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) \leq \lim_{n\to\infty} \Pr(\mathcal{A}) \leq \lim_{n\to\infty} \Pr\left(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c\right) + \Pr\left(\mathcal{B} \cup \mathcal{C}\right) \\ = \Pr(\varnothing) + \lim_{n\to\infty} \Pr\left(\mathcal{B} \cup \mathcal{C}\right) \leq \lim_{n\to\infty} \Pr(\mathcal{D}) = 0,\end{aligned}$$

which concludes the proof. ■

▸ Return to Sketch of the proof

Appendix G: Proof of Lemma 2.25

Since $\mathbf{W} > 0$ is nonsingular and there exists a unique $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $\mathbf{g}_0(\boldsymbol{\theta}_0) = \mathbf{0}$, for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have:

$$Q_0(\boldsymbol{\theta}) = -\mathbf{g}_0(\boldsymbol{\theta})^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta}) \leq -||\mathbf{g}_0(\boldsymbol{\theta}_0)^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta}_0)|| = Q_0(\boldsymbol{\theta}_0).$$

Therefore, $Q_0(\boldsymbol{\theta})$ has a unique maximum (0) at $\boldsymbol{\theta}_0$. ■

▸ Return to Lemma 2.25

Appendix H: Proof of Theorem 2.28

Since $\bar{X}_n - \mu$ has mean zero, its variance is

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = 1/n \sum_{k=-\infty}^{\infty} \zeta_k \leq 1/n \sum_{k=-\infty}^{\infty} |\zeta_k| \leq C/n.$$

By Chebychev's inequality, for all $\varepsilon > 0$

$$\Pr\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\varepsilon^2} \leq \frac{C}{\varepsilon^2 n},$$

so that

$$\lim_{n \to \infty} \Pr\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \to 0,$$

which concludes the proof. ■

▸ Return to Theorem 2.28

Appendix I: Proof of Corollary 2.29

The proof of the result follows directly from Theorem 2.28. (X_t^2) has an absolutely summable covariance structure which implies that both (X_t) and $(X_t X_{t-h})$, for all $h \in \mathbb{Z}$, have absolutely summable covariance structures. Under the conditions that (X_t) is weakly stationary and has absolutely summable covariance structure, by Theorem 2.29, we have $\bar{X}_T \xrightarrow{\mathcal{P}} \mu$. Let:

$$\tilde{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^{T} (X_t - \mu)(X_{t-h} - \mu).$$

Since $((X_t - \mu)(X_{t-h} - \mu))$ is stationary and has absolutely summable covariance structure, by Theorem 2.29 again, we have $\tilde{\gamma}(h) \xrightarrow{\mathcal{P}} \gamma(h)$. We can also show that $\sqrt{T}(\tilde{\gamma}(h) - \hat{\gamma}(h)) = o_p(1)$. Therefore, we obtain:

$$\hat{\gamma}(h) \xrightarrow{\mathcal{P}} \gamma(h).$$

Appendix I: Proof of Corollary 2.29

Similarly, we can show that:

$$(\hat{\gamma}(0), \hat{\gamma}(h))^T \xmapsto{\mathcal{P}} (\gamma(0), \gamma(h))^T .$$

Then by Theorem 2.23, we have:

$$\hat{\rho}(h) \xmapsto{\mathcal{P}} \rho(h),$$

which concludes the proof. ■

▸ Return to Corollary 2.29

Appendix J: Example - Proving Condition (C.4)

We return to the example presented in ▸ Example GMM, we define the following quantities:

$$\hat{Q}_n(\boldsymbol{\theta}) = -||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}}.$$

where we assumed that $\hat{\mathbf{W}} \xrightarrow{\mathcal{P}} \mathbf{W}$.

Since we have $Z_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$, we let

$$\mathbf{X}_i \equiv \begin{bmatrix} Z_i \\ Z_i^2 \\ Z_i^3 \end{bmatrix},$$

and therefore by the (weak) Law of Large Numbers (see Theorem 2.22), we have:

$$\hat{\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \xrightarrow{\mathcal{P}} \mathbb{E}[\mathbf{X}_i] = \boldsymbol{\gamma}(\boldsymbol{\theta}_0).$$

Next, we define $Q_0(\boldsymbol{\theta}) = -||\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\mathbf{W}}$ which is the same as $\hat{Q}_n(\boldsymbol{\theta})$ when replacing all the random quantities by the limiting term.

Appendix J: Example - Proving Condition (C.4)

We are now interested in showing that the difference $\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})$ converges in probability to 0. For ease of notation, we will define $\boldsymbol{\gamma}(\boldsymbol{\theta}_0)$ as γ_0:

$$\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}) \leq |\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| = \Big| \underbrace{||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}} - ||\gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\mathbf{W}}}_{\mathbf{A}} \Big| .$$

Developing the above expression $\mathbf{A}$ inside the absolute value, and defining without loss of generality that $\mathbf{W}^* = \widehat{\mathbf{W}} - \mathbf{W}$ where $\mathbf{W}$ is a symmetric matrix, we get:

$$\begin{aligned}
\mathbf{A} =& \hat{\boldsymbol{\gamma}}^T \widehat{\mathbf{W}} \hat{\boldsymbol{\gamma}} + \boldsymbol{\gamma}(\boldsymbol{\theta})^T \widehat{\mathbf{W}} \boldsymbol{\gamma}(\boldsymbol{\theta}) - 2\hat{\boldsymbol{\gamma}}^T \widehat{\mathbf{W}} \boldsymbol{\gamma}(\boldsymbol{\theta}) - \gamma_0^T \mathbf{W} \gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})^T \mathbf{W} \boldsymbol{\gamma}(\boldsymbol{\theta}) + 2\gamma_0^T \mathbf{W} \boldsymbol{\gamma}(\boldsymbol{\theta}) \\
=& \|\hat{\boldsymbol{\gamma}} - \gamma_0\|^2_{\widehat{\mathbf{W}}} - \gamma_0^T \widehat{\mathbf{W}} \gamma_0 + 2\hat{\boldsymbol{\gamma}}^T \widehat{\mathbf{W}} \gamma_0 - 2\hat{\boldsymbol{\gamma}}^T \widehat{\mathbf{W}} \boldsymbol{\gamma}(\boldsymbol{\theta}) + \boldsymbol{\gamma}(\boldsymbol{\theta})^T \widehat{\mathbf{W}} \boldsymbol{\gamma}(\boldsymbol{\theta}) \\
& - \gamma_0^T \mathbf{W} \gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})^T \mathbf{W} \boldsymbol{\gamma}(\boldsymbol{\theta}) + 2\gamma_0^T \mathbf{W} \boldsymbol{\gamma}(\boldsymbol{\theta}) \\
=& \|\hat{\boldsymbol{\gamma}} - \gamma_0\|^2_{\widehat{\mathbf{W}}} + \|\gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})\|^2_{\mathbf{W}^*} + 2\boldsymbol{\gamma}(\boldsymbol{\theta}) \widehat{\mathbf{W}} \gamma_0 - 2\gamma_0^T \widehat{\mathbf{W}} \gamma_0 \\
& - 2\hat{\boldsymbol{\gamma}}^T \widehat{\mathbf{W}} \boldsymbol{\gamma}(\boldsymbol{\theta}) + 2\gamma_0^T \mathbf{W} \boldsymbol{\gamma}(\boldsymbol{\theta}) \\
=& \|\hat{\boldsymbol{\gamma}} - \gamma_0\|^2_{\widehat{\mathbf{W}}} + \|\gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})\|^2_{\mathbf{W}^*} + 2\left(\gamma_0 - \boldsymbol{\gamma}(\boldsymbol{\theta})\right)^T \widehat{\mathbf{W}} \left(\hat{\boldsymbol{\gamma}} - \gamma_0\right).
\end{aligned}$$

Appendix J: Example - Proving Condition (C.4)

Therefore using the triangular inequality, we can write the following:

$$|\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \leq \underbrace{\left|\|\gamma(\boldsymbol{\theta}_0) - \gamma(\boldsymbol{\theta})\|^2_{\mathbf{W}^*}\right|}_{\equiv a_1} + \underbrace{\left|\|\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\|^2_{\widehat{\mathbf{W}}}\right|}_{\equiv a_2}$$
$$+ \underbrace{\left|2\left(\gamma(\boldsymbol{\theta}_0) - \gamma(\boldsymbol{\theta})\right)^T \widehat{\mathbf{W}}\left(\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\right)\right|}_{\equiv a_3}.$$

Before going on, suppose that $\boldsymbol{x}$ is a vector and $\mathbf{W}$ the above defined matrix. We have that $\boldsymbol{x}^T\mathbf{W}\boldsymbol{x} \leq \lambda_1\|\boldsymbol{x}\|^2$, where λ_1 is then largest eigenvalue of $\mathbf{W}$. Furthermore, lets also define the Frobenius norm of a matrix as $\|\mathbf{W}\| = (\sum_i^N \sum_j^J w_{i,j}^2)^{\frac{1}{2}}$ and $\|\mathbf{W}\|^2 = \sigma_{max} \leq \|\mathbf{W}\|$, where σ_{max} is the largest singular value of $\|\mathbf{W}\|$.

Example: Proving Condition (C.4)

Using those properties, we can investigate the terms in the above equation. Considering a_1 , we have that:

$$\begin{aligned} a_1 &\leq \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2 \lambda_1 \leq \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2 \|\mathbf{W}^*\| \\ &= \sum_{i=1}^{3} (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2 \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3} (\hat{w}_{i,j} - w_{i,j})^2}. \end{aligned}$$

By Conditions (C.1) to (C.3), $\sum_{i=1}^{3} (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2 \leq 3 \max_i (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2$ which is bounded.
By the previously mentioned on $\widehat{\mathbf{W}}$, we can also see that:

$$\sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3} (\hat{w}_{i,j} - w_{i,j})^2} \leq 3 \max_i \sqrt{(\hat{w}_{i,j} - w_{i,j})^2} \xrightarrow{\mathcal{P}} 0.$$

Appendix J: Example - Proving Condition (C.4)

Investigating the convergence of $\mathbf{a2} \leq \|\hat{\gamma} - \gamma_0\|^2 \|\widehat{\mathbf{W}}\|$, by the same conditions (C.1) to (C.3), and that the sample moments will converge to the population ones, we have that:

$$a_2 \xrightarrow{\mathcal{P}} 0.$$

Finally, considering the previous results on a_1 and a_2, we can see that the last term a_3 tends also to 0, therefore, we can say that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$|\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0.$$

●

▸ Return to the discussion on (C.4)