Applied Time Series Analysis
SS 2014 – Week 03

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, March 3, 2014

Applied Time Series Analysis
SS 2014 – Week 03

Where are we?

For most of the rest of this course, we will deal with (weakly) stationary time series. They have the following properties:

- $E[X_t] = \mu$
- $Var(X_t) = \sigma^2$
- $Cov(X_t, X_{t+h}) = \gamma_h$

If a time series is non-stationary, we know how to decompose into deterministic and stationary, random part.

Our forthcoming goals are:

- understanding the dependency in a stationary series
- modeling this dependency and generate forecasts

Applied Time Series Analysis
SS 2014 – Week 03

Autocorrelation

The aim of this section is to estimate, explore and understand the dependency structure within a stationary time series.

Def: **Autocorrelation**

$$Cor(X_{t+k}, X_t) = \frac{Cov(X_{t+k}, X_t)}{\sqrt{Var(X_{t+k}) \cdot Var(X_t)}} = \rho(k)$$

Autocorrelation is a dimensionless measure for the strength of the linear association between the random variables X_{t+k} and X_t.

There are 2 estimators, i.e. the lagged sample and the plug-in.
→ *see slides & blackboard for a sketch of the two approaches…*

Practical Interpretation of Autocorrelation

We e.g. assume $\rho(k) = 0.7$

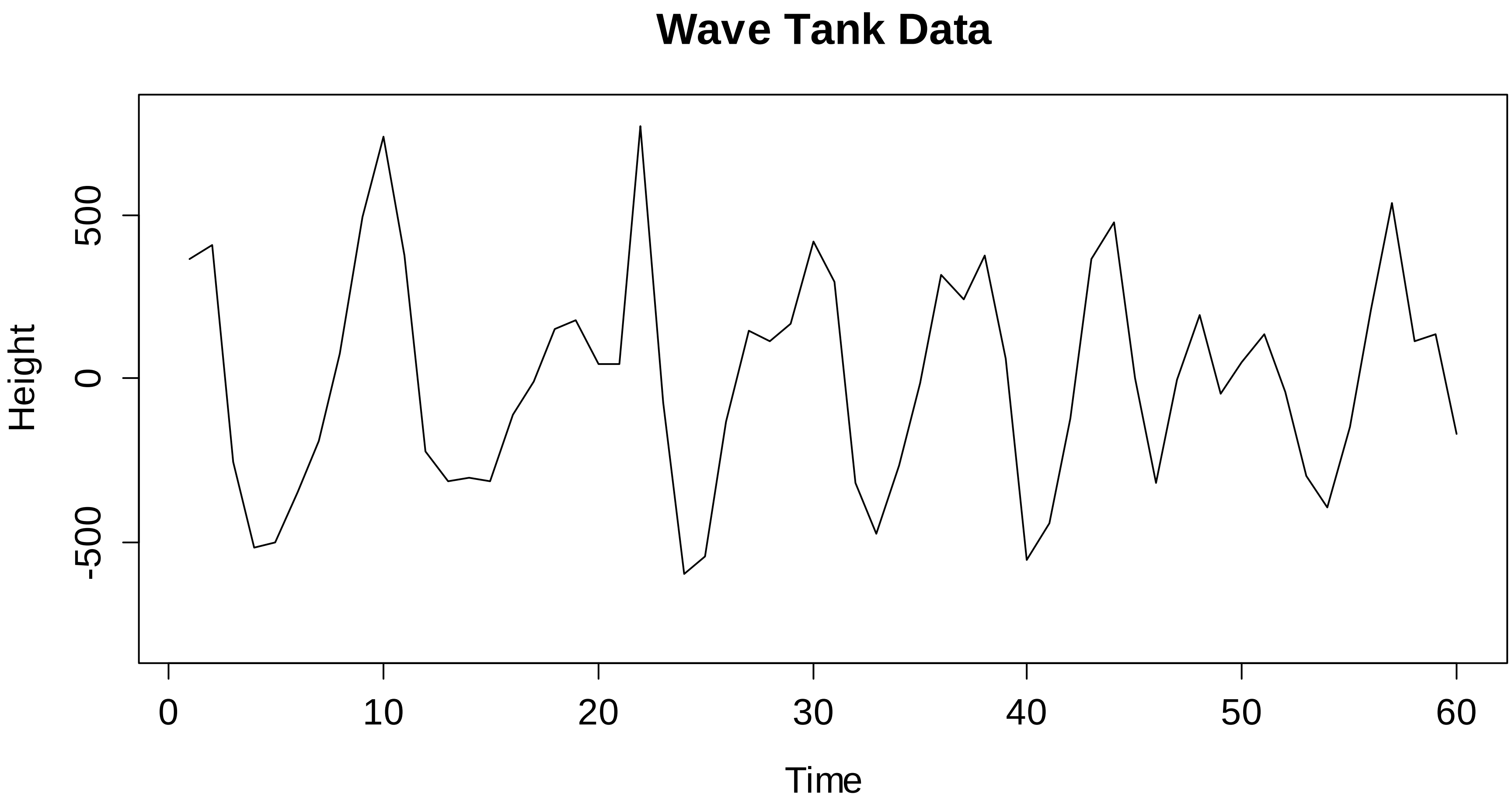
→ The square of the autocorrelation, i.e. $\rho(k)^2 = 0.49$, is the percentage of variability explained by the linear association between X_t and its predecessor X_{t-1}.

→ Thus, in our example, X_{t-1} accounts for roughly 49% of the variability observed in random variable X_t. Only roughly because the world is not linear.

→ From this we can also conclude that any $\rho(k) < 0.4$ is not a strong association, i.e. has a small effect on the next observation only.

Applied Time Series Analysis
SS 2014 – Week 03

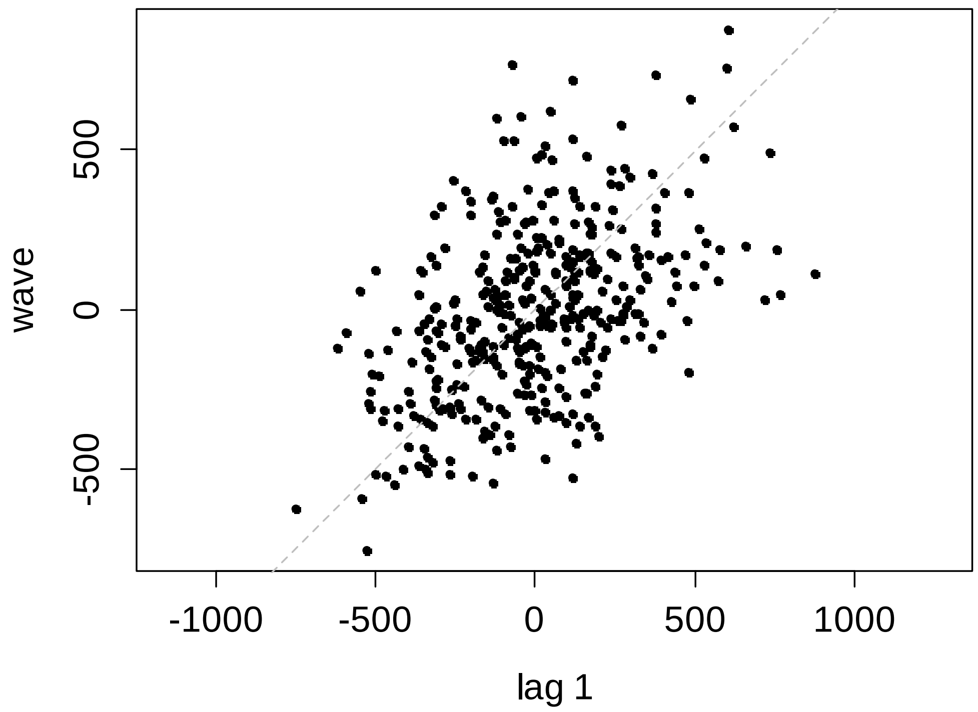
Example: Wave Tank Data



Lagged Scatterplot Approach

Generate a plot of (x_t, x_{t+k}) for all $t = 1, ..., n-k$ and compute the canonical Pearson correlation coefficient from these data pairs.

Lagged Scatterplot, k=1, cor=0.47




```
> lag.plot(wave, do.lines=FALSE, pch=20)
> title("Lagged Scatter, k=1, cor=0.47")
```

$$\tilde{\rho}(k) = \frac{\sum_{s=1}^{n-k}(x_{s+k} - \bar{x}_{(k)})(x_s - \bar{x}_{(1)})}{\sqrt{\sum_{s=k+1}^{n}(x_s - \bar{x}_{(k)})^2 \cdot \sum_{t=1}^{n-k}(x_t - \bar{x}_{(1)})^2}}$$

Plug-In Estimation

For obtaining an estimate of $\hat{\rho}(k)$, determine the sample covariance at lag k and divide by the sample variance.

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} = \frac{Cov(X_t, X_{t+k})}{Var(X_t)}$$

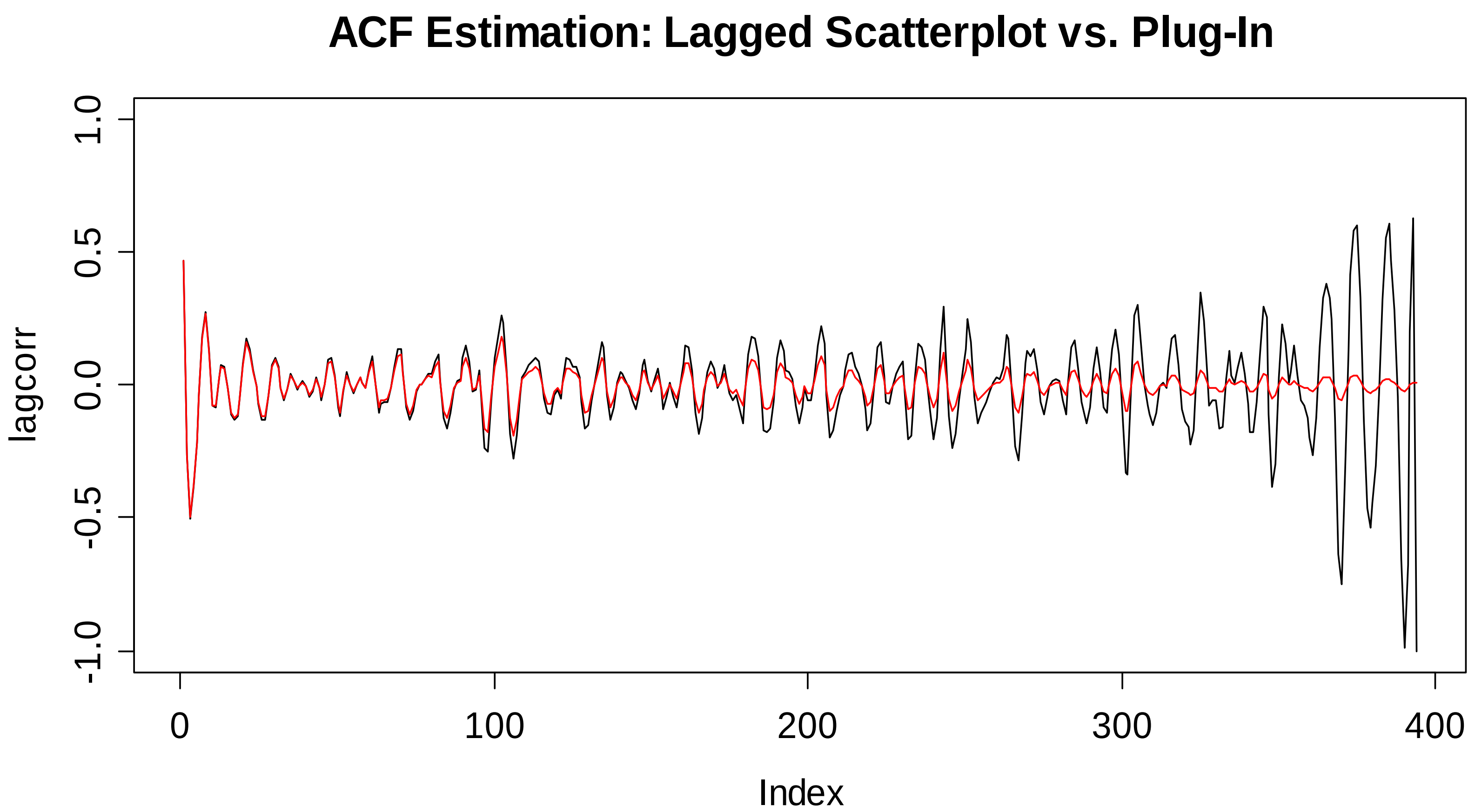
$$\hat{\gamma}(k) = \frac{1}{n} \sum_{s=1}^{n-k} (x_{s+k} - \overline{x})(x_s - \overline{x})$$

where $\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$

This is the standard approach for computing autocorrelations in time series analysis. It is better than the lagged scatterplot idea.

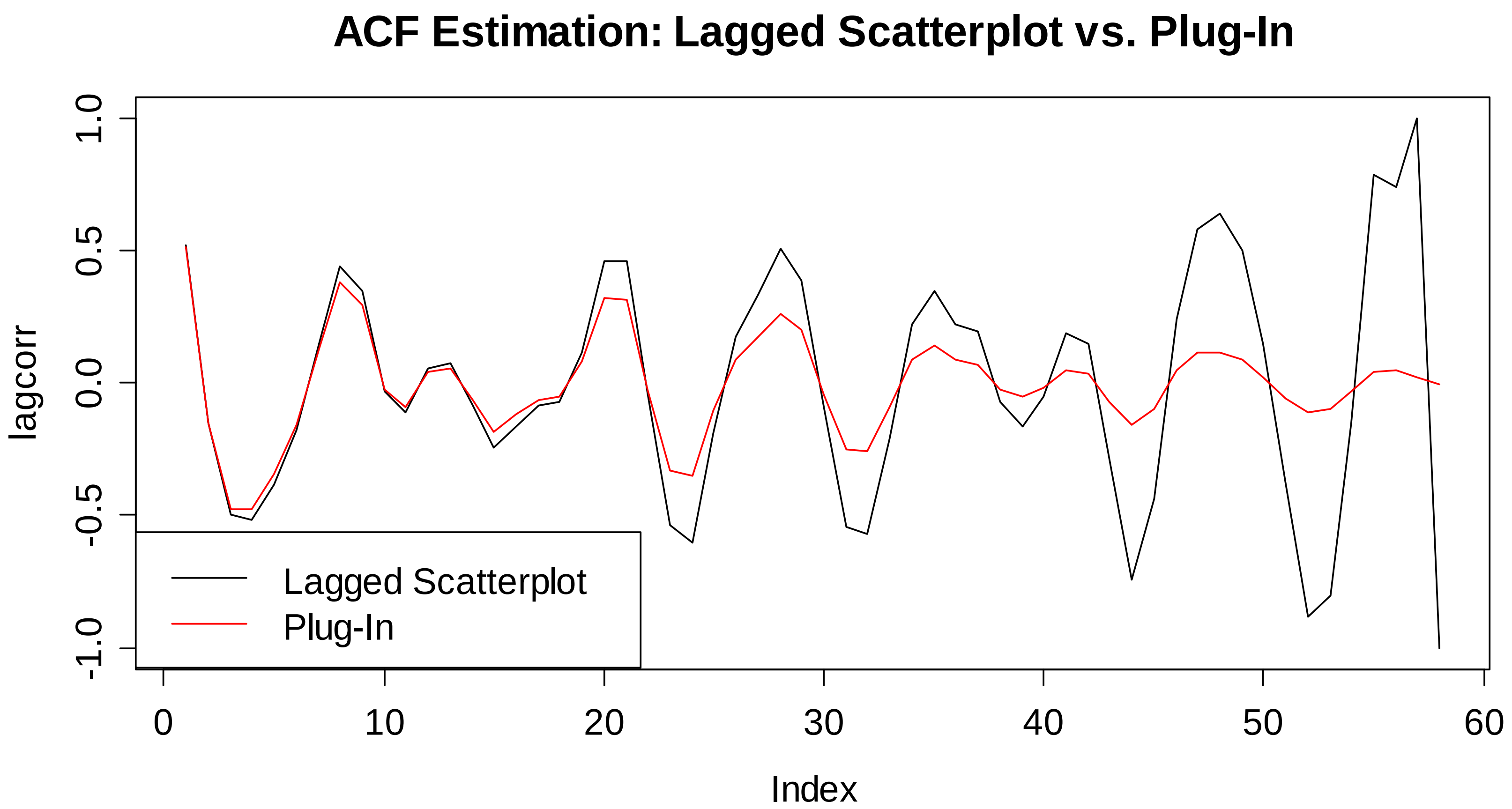
Applied Time Series Analysis
SS 2014 – Week 03

Comparison Idea 1 vs. Idea 2



Applied Time Series Analysis
SS 2014 – Week 03

Comparison Idea 1 vs. Idea 2



What is important about ACF estimation?

- Correlations are never to be trusted without a visual inspection with a scatterplot.

- The bigger the lag k, the fewer data pairs remain for estimating the acf at lag k.

- Rule of the thumb: the acf is only meaningful up to about

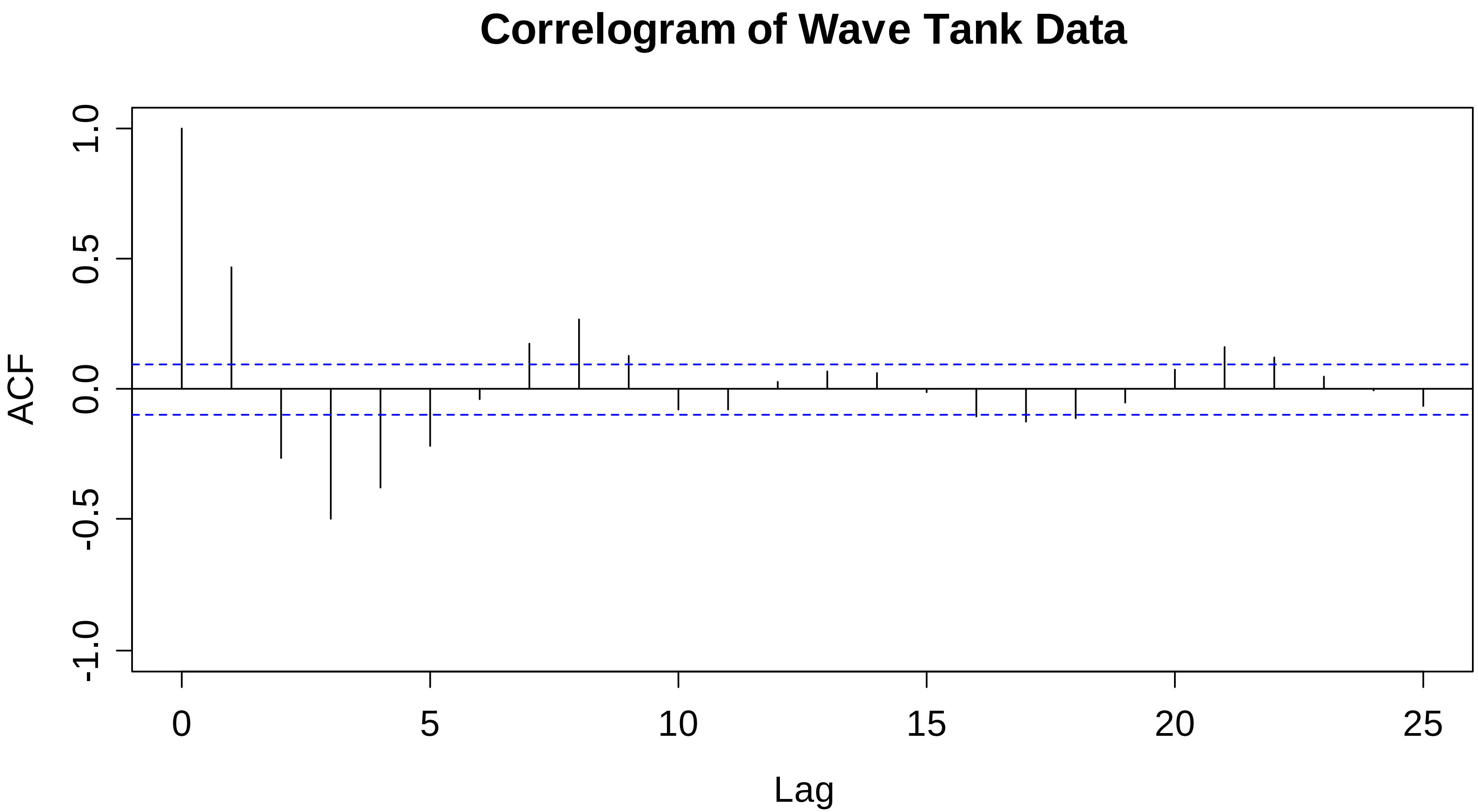
  a) lag $10*\log_{10}(n)$
  b) lag $n/4$

- The estimated sample ACs can be highly correlated.

- The correlogram is only meaningful for stationary series!!!

Correlogram

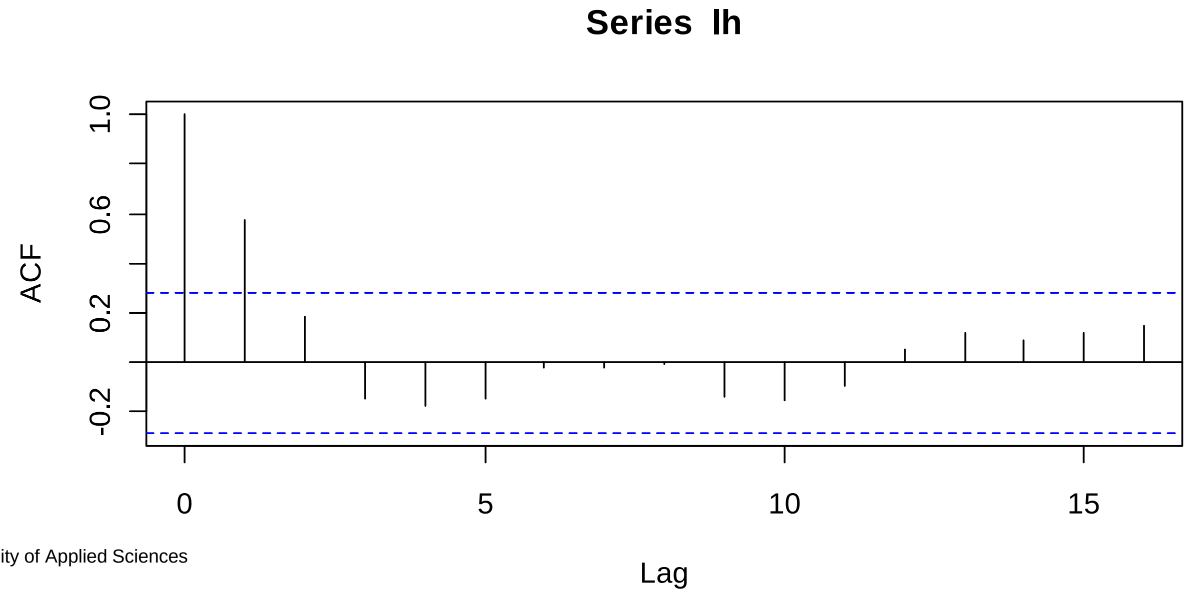
```
> acf(wave, ylim=c(-1,1))
```



Random Series – Confidence Bands

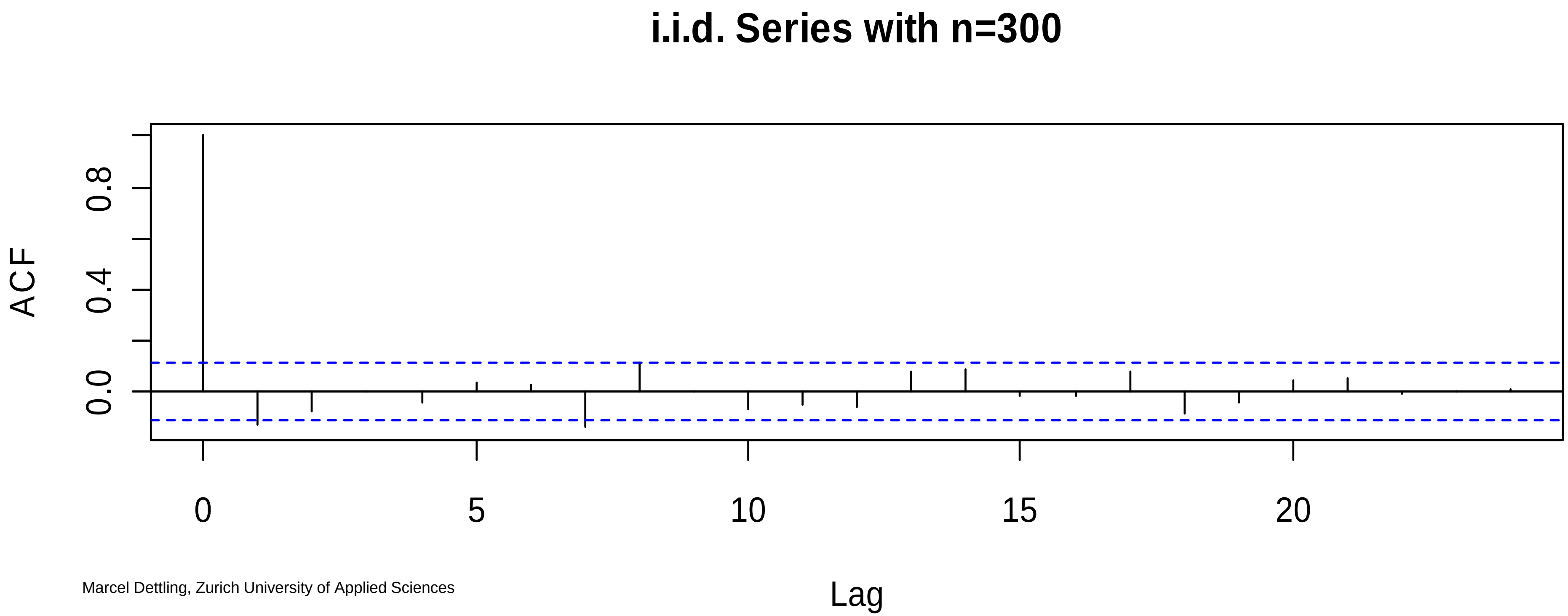
If a time series is White Noise, i.e. consists of iid random variables X_t, the (theoretical) autocorrelations $\rho(k)$ are all 0.

However, the estimated $\hat{\rho}(k)$ are not. We thus need to decide, whether an observed $\hat{\rho}(k) \neq 0$ is significantly so, or just appeared by chance. This is the idea behind the confidence bands.



Random Series – Confidence Bands

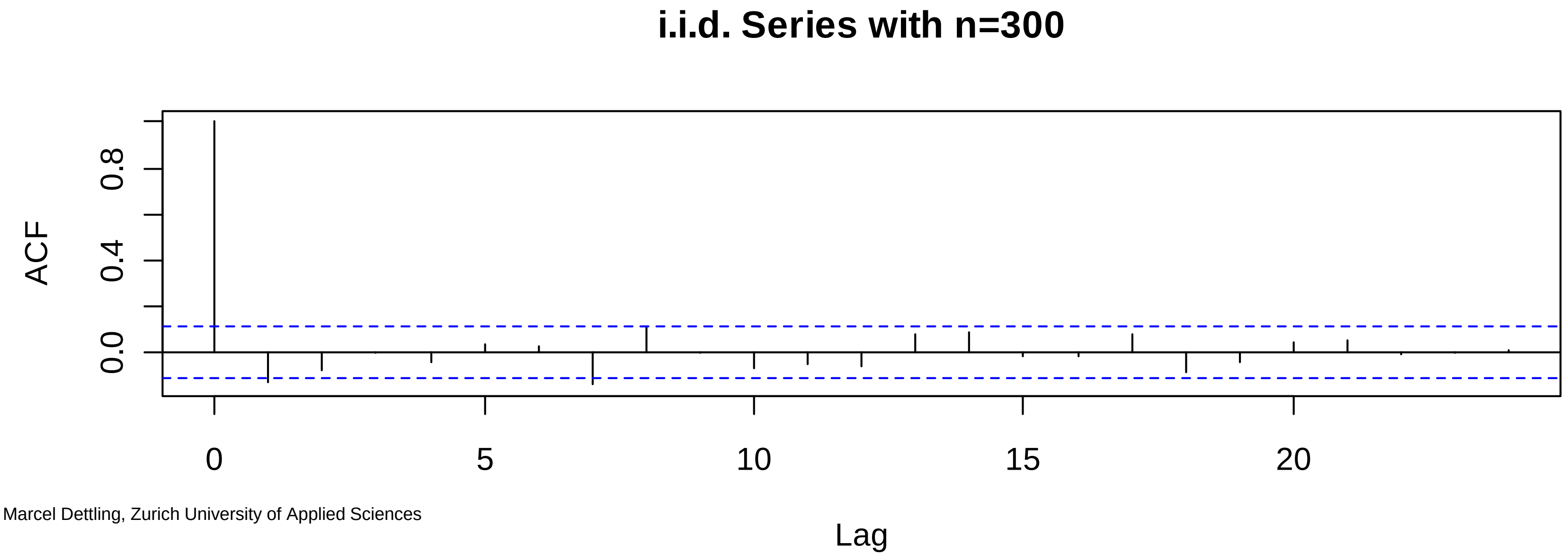
For long iid series, it can be shown that $\hat{\rho}(k)$ is approximately $N(0,1/n)$. Thus, under the null hypothesis that a series is iid and hence $\rho(k)=0$, the 95% acceptance region for the null is given by the interval $\pm 2/\sqrt{n}$.



Random Series – Confidence Bands

Thus, even for a (long) i.i.d. time series, we expect that 5% of the estimated autocorrelation coeffcients exceed the confidence bounds. They correspond to type I errors.

Note: the probabilistic properties of non-normal i.i.d series are much more difficult to derive.



Ljung-Box Test

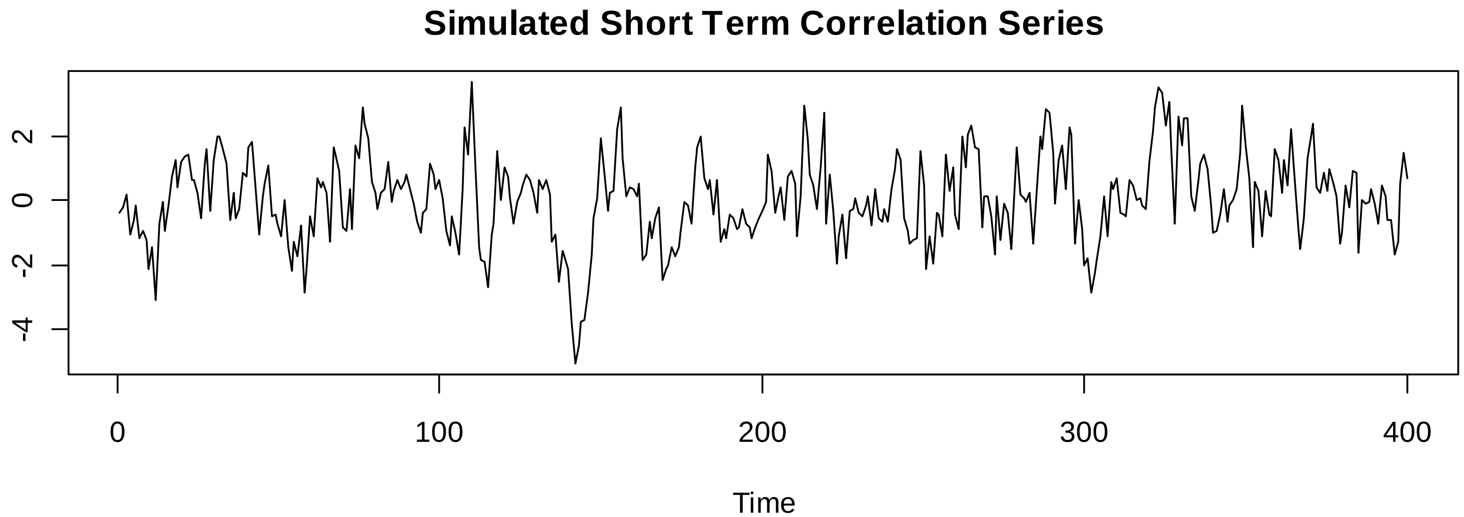
The Ljung-Box approach tests the null hypothesis that a number of autocorrelation coefficients are simultaneously equal to zero. Thus, it tests for significant autocorrelation in a series. The test statistic is:

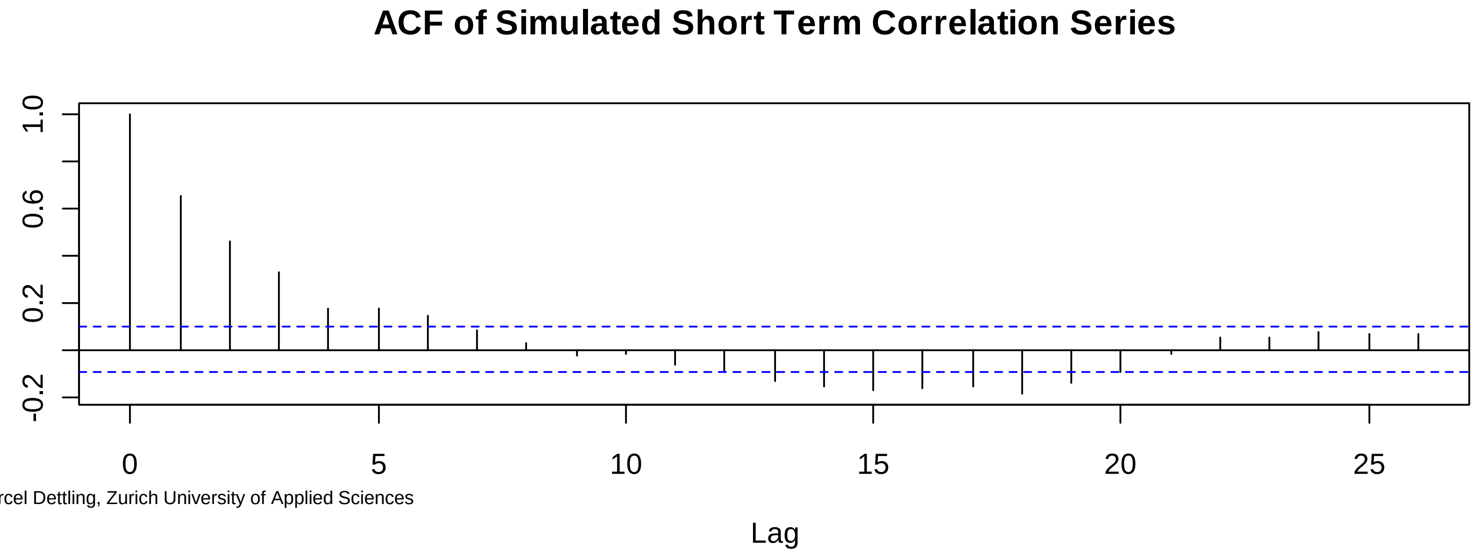
$$Q(h) = n \cdot (n+2) \cdot \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \sim \chi_h^2$$

In R:

```
> Box.test(wave, lag=10, type="Ljung-Box")
Box-Ljung test
data: wave
X-squared = 344.0155, df = 10, p-value < 2.2e-16
```

Short Term Positive Correlation





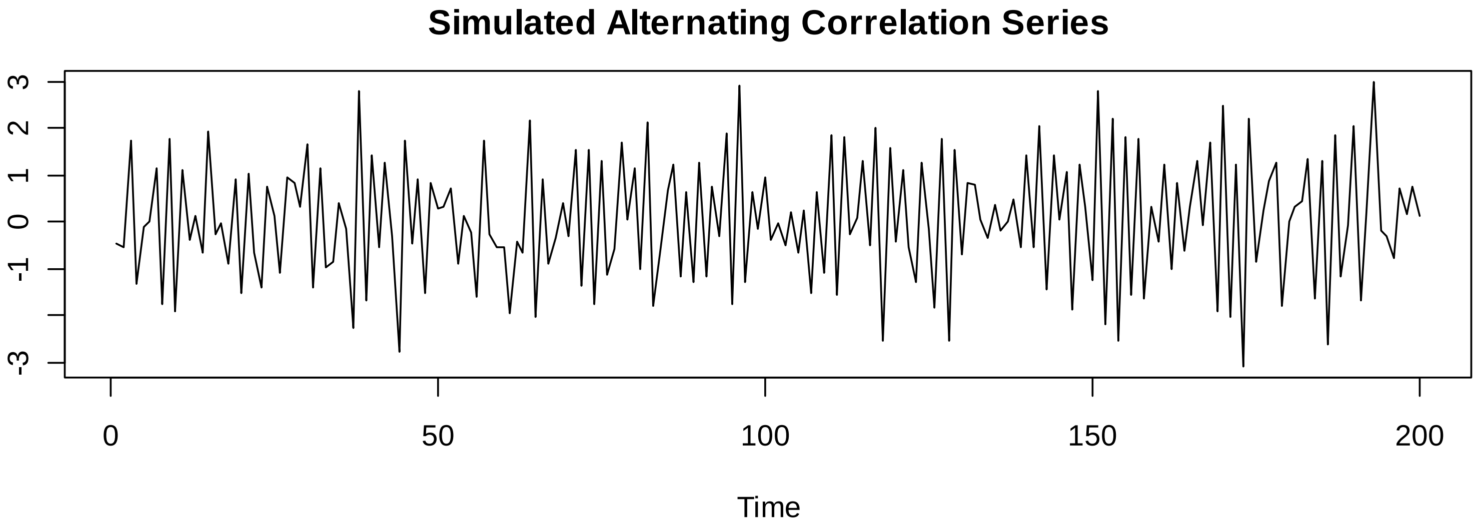
Short Term Positive Correlation

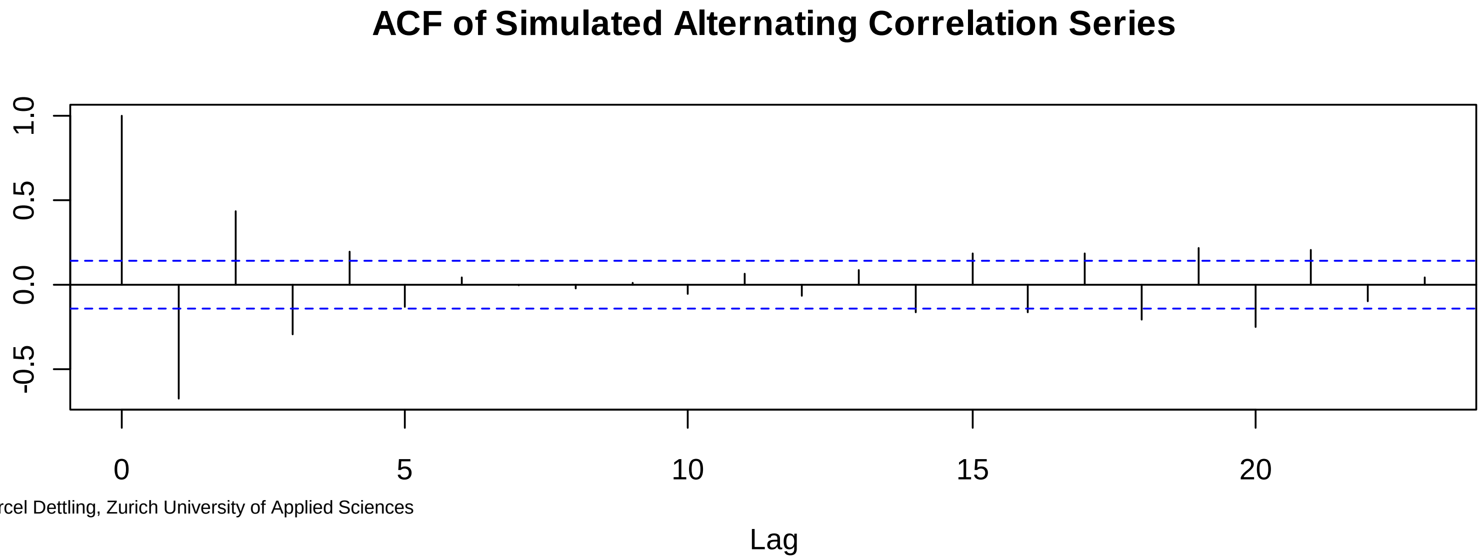
Stationary series often exhibit short-term correlation, characterized by a fairly large value of $\hat{\rho}(1)$, followed by a few more coefficients which, while significantly greater than zero, tend to get successively smaller. For longer lags k, they are close to 0.

A time series which gives rise to such a correlogram, is one for which an observation above the mean tends to be followed by one or more further observations above the mean, and similarly for observations below the mean.

A model called an autoregressive model may be appropriate for series of this type.

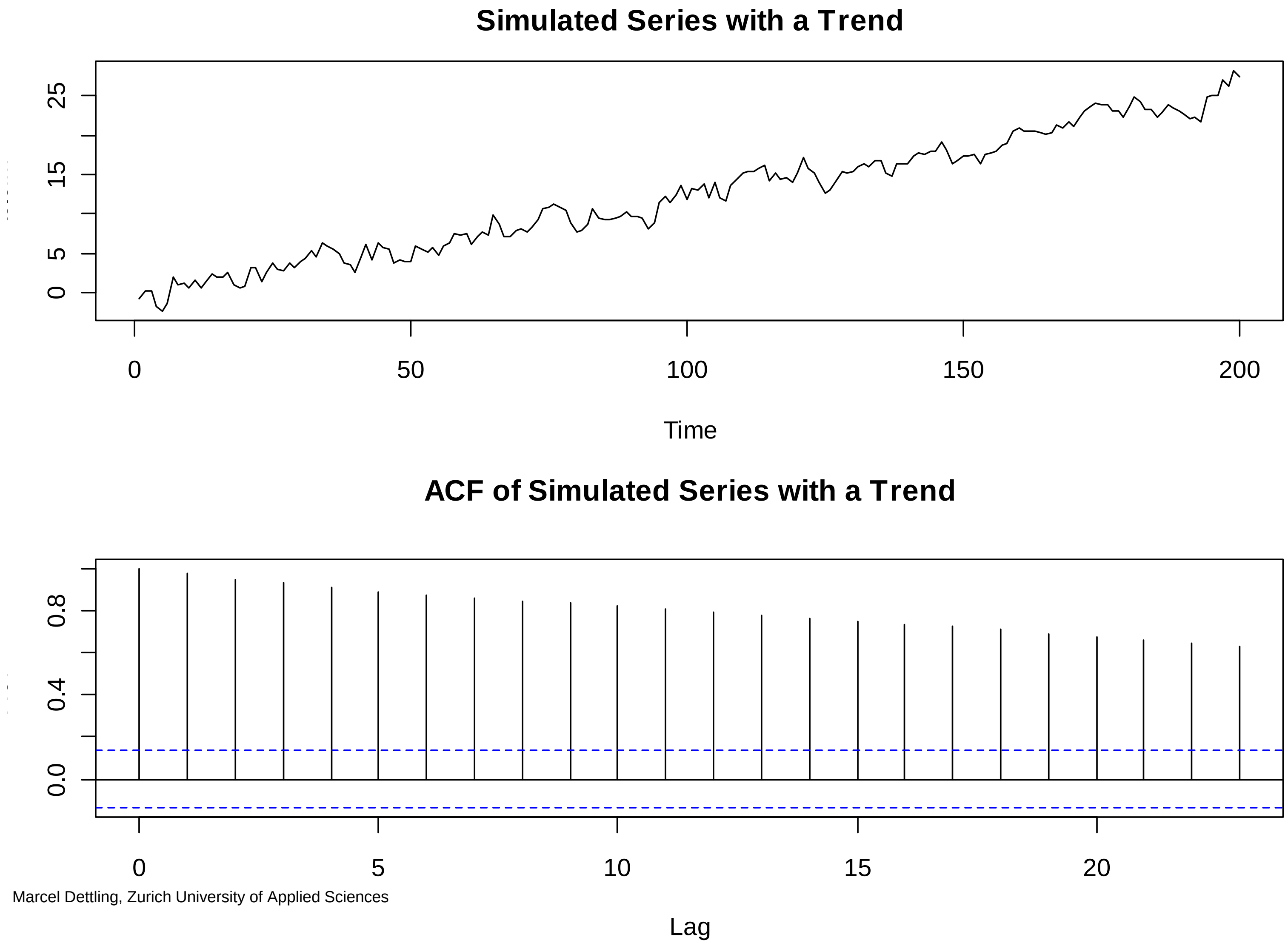
Alternating Time Series



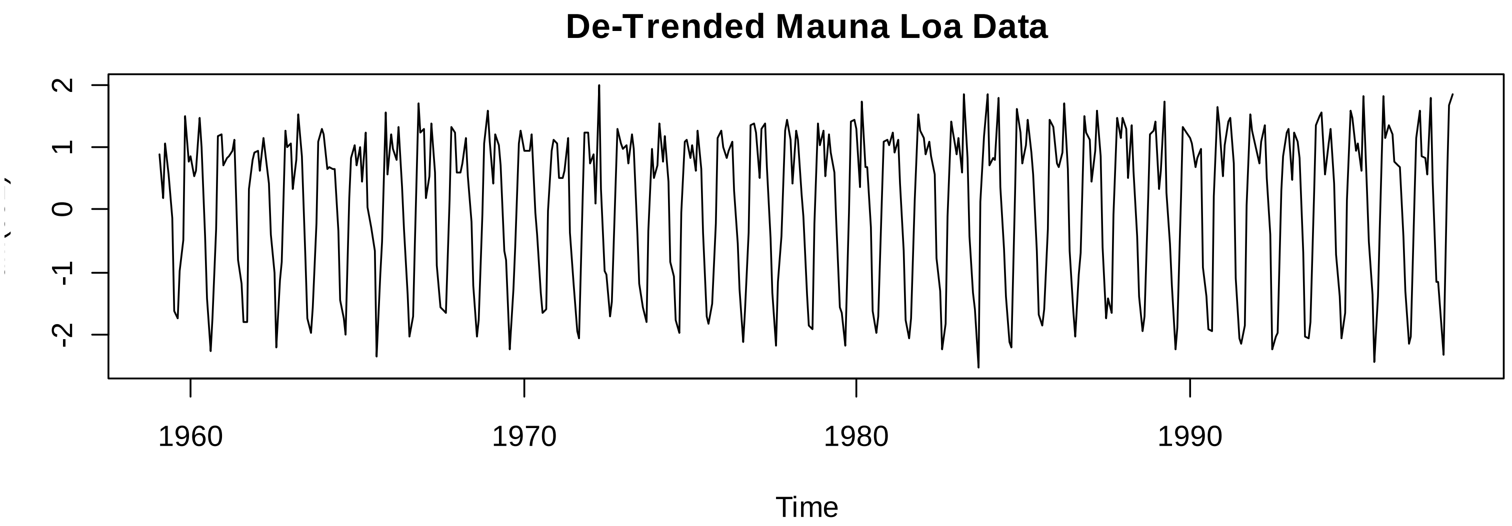


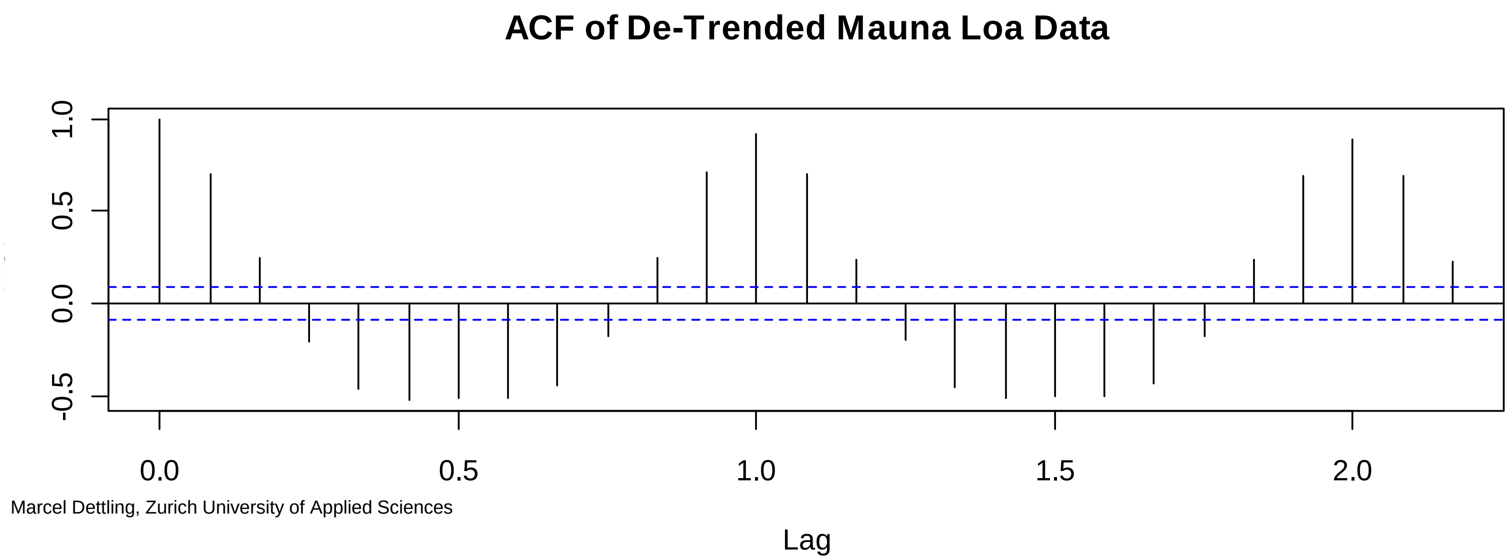
Applied Time Series Analysis
SS 2014 – Week 03

Non-Stationarity in the ACF: Trend



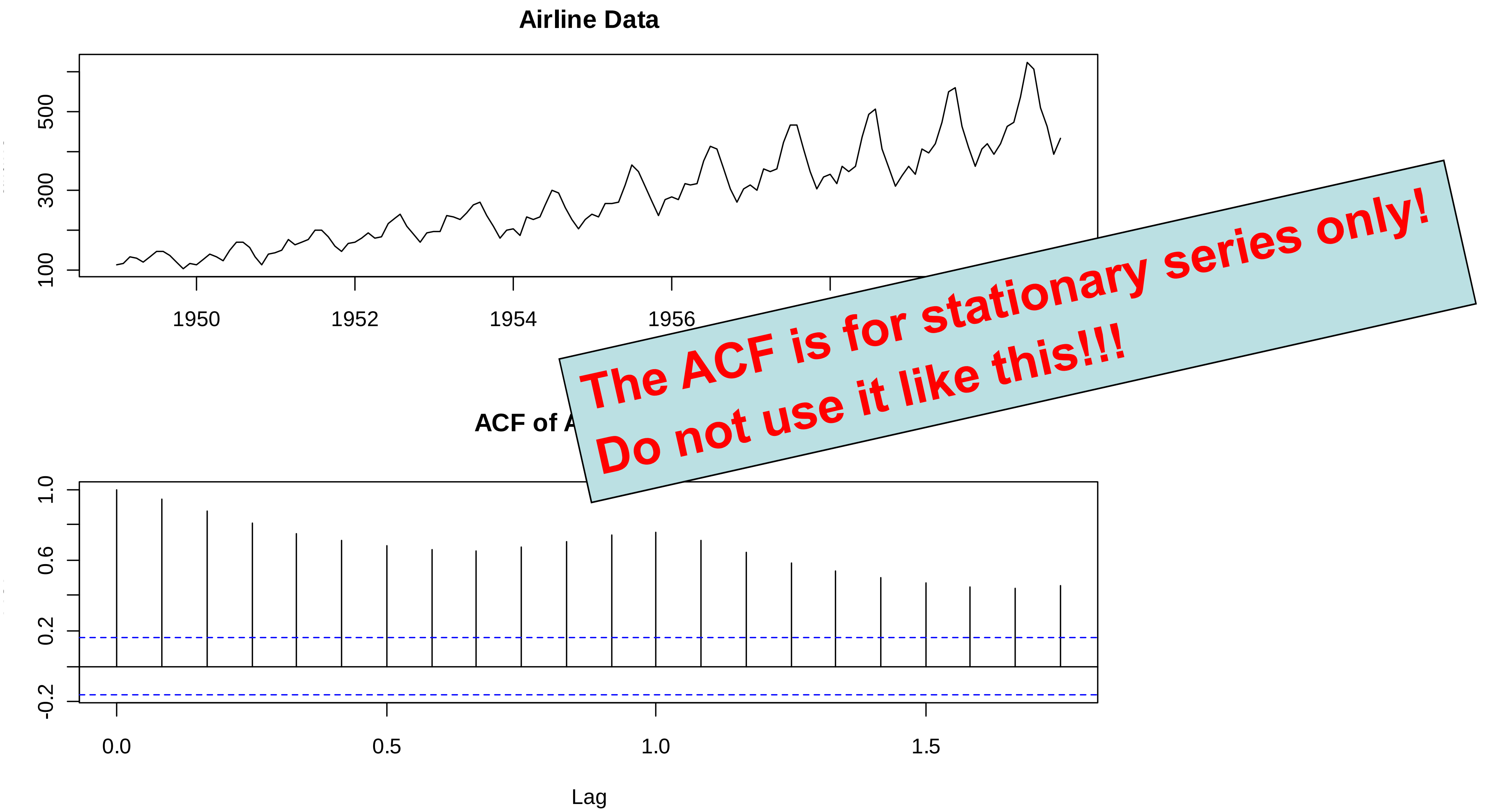
Non-Stationarity in the ACF: Seasonal Pattern





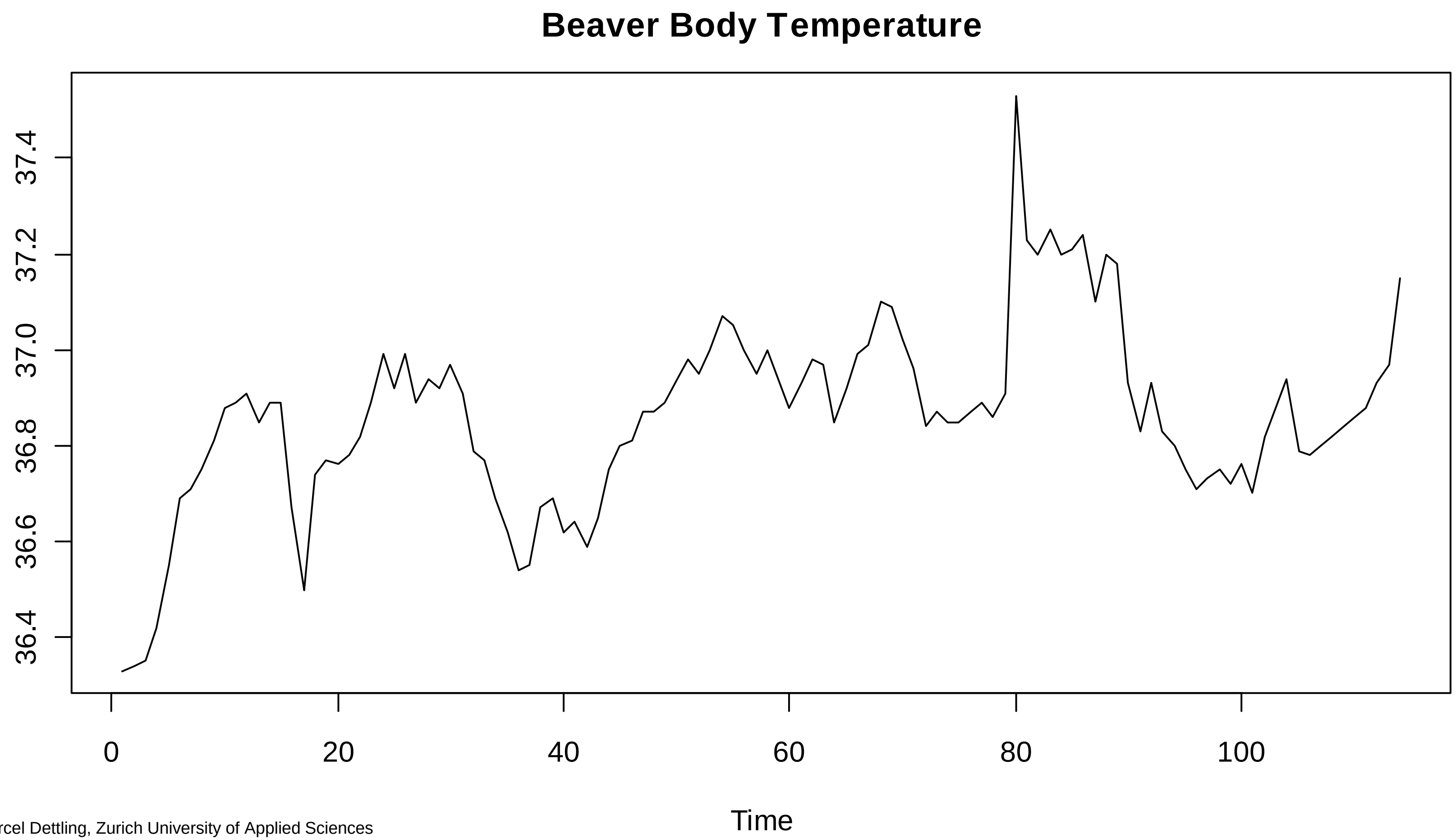
Applied Time Series Analysis
SS 2014 – Week 03

ACF of the Raw Airline Data



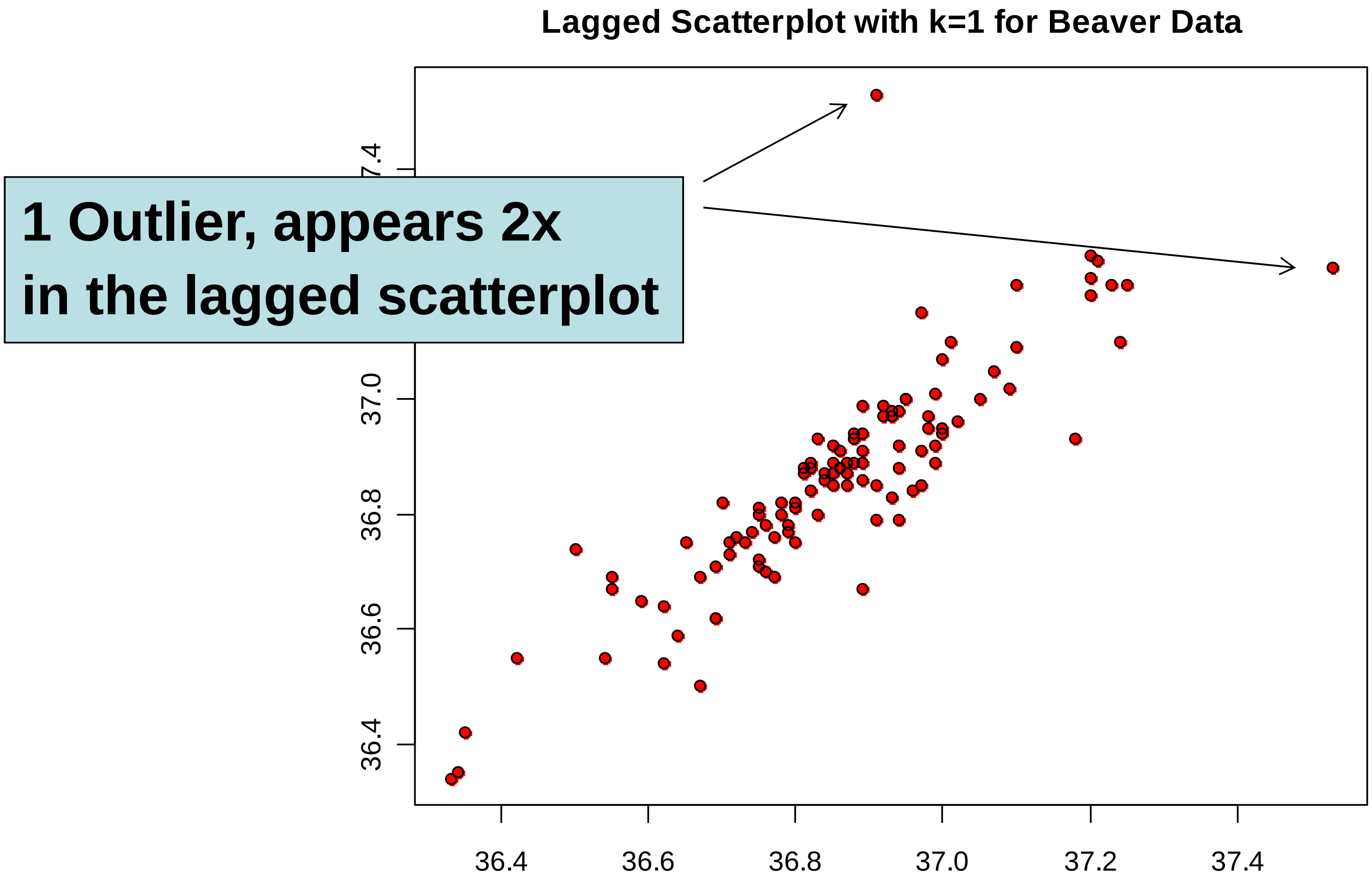
Outliers and the ACF

Outliers in the time series strongly affect the ACF estimation!



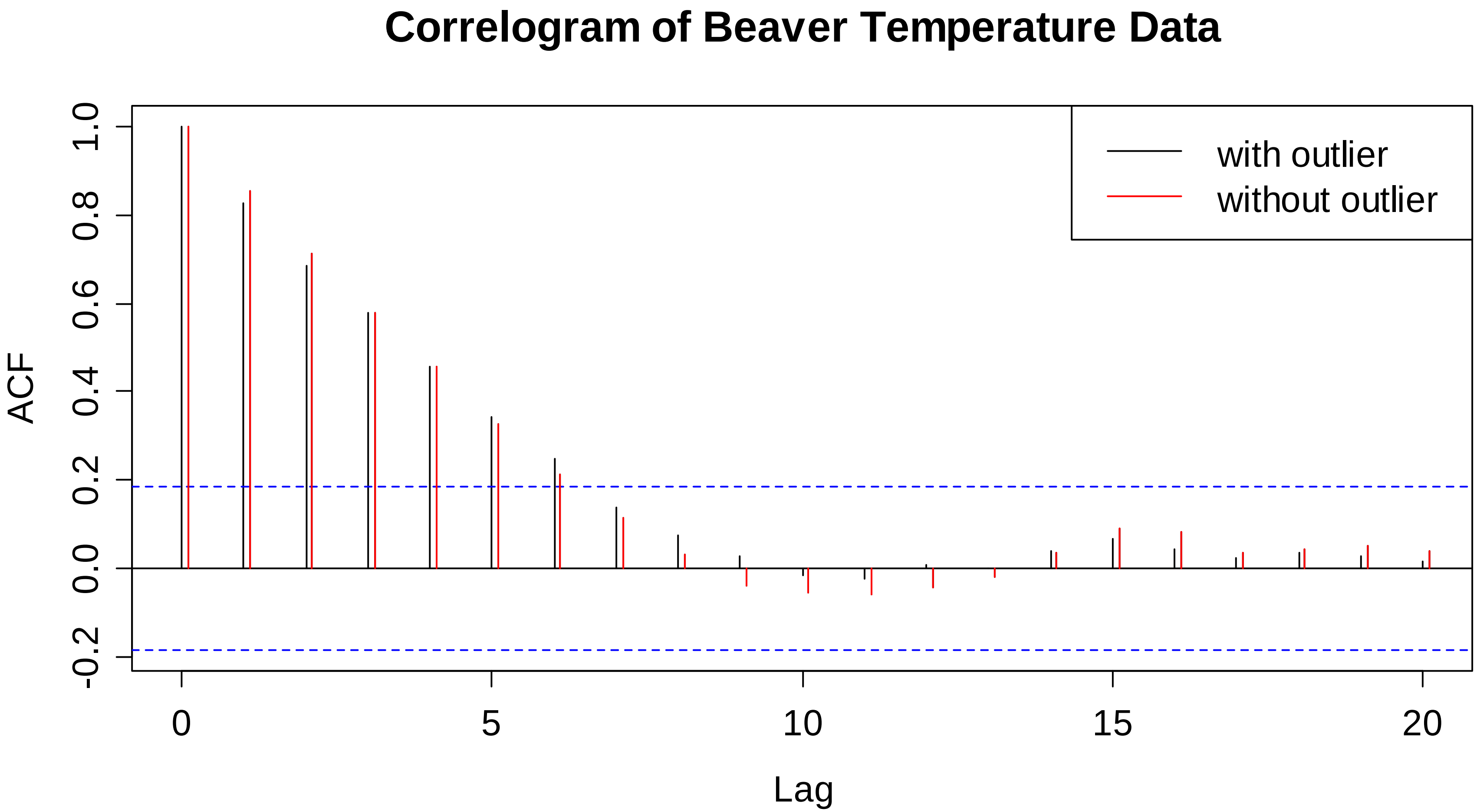
Applied Time Series Analysis
SS 2014 – Week 03

Outliers and the ACF



Applied Time Series Analysis
SS 2014 – Week 03

Outliers and the ACF



Outliers and the ACF

The estimates $\hat{\rho}(k)$ are very sensitive to outliers. They can be diagnosed using the lagged scatterplot, where every single outlier appears twice.

Strategy for dealing with outliers:

- if it is bad data point: delete the observation
- replace the now missing observations by either:

  a) global mean of the series
  b) local mean of the series, e.g. +/- 3 observations
  c) fit a time series model and predict the missing value

General Remarks about the ACF

a) Appearance of the series => Appearance of the ACF
Appearance of the series ~~<=~~ Appearance of the ACF

b) Compensation

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

All autocorrelation coefficients sum up to -1/2. For large lags k, they can thus not be trusted, but are at least damped. This is a reason for using the rule of the thumb.

How Well Can We Estimate the ACF?

What do we know already?

- The ACF estimates are biased
- At higher lags, we have few observations, and thus variability
- There also is the compensation problem…

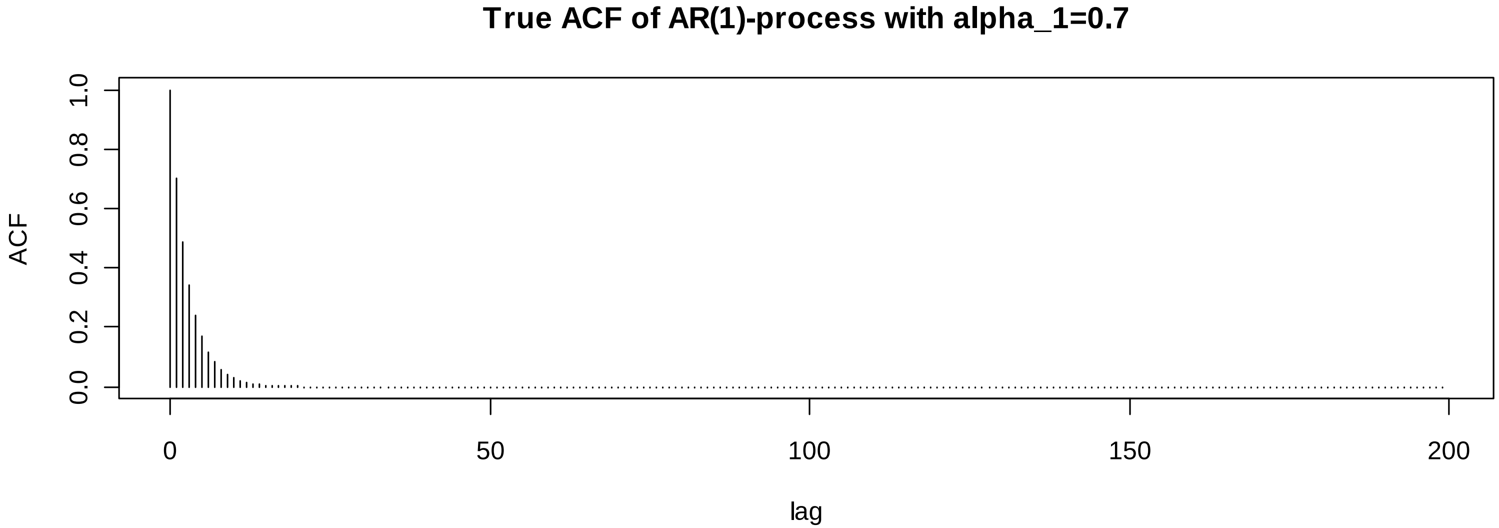
→ ACF estimation is not easy, and interpretation is tricky.

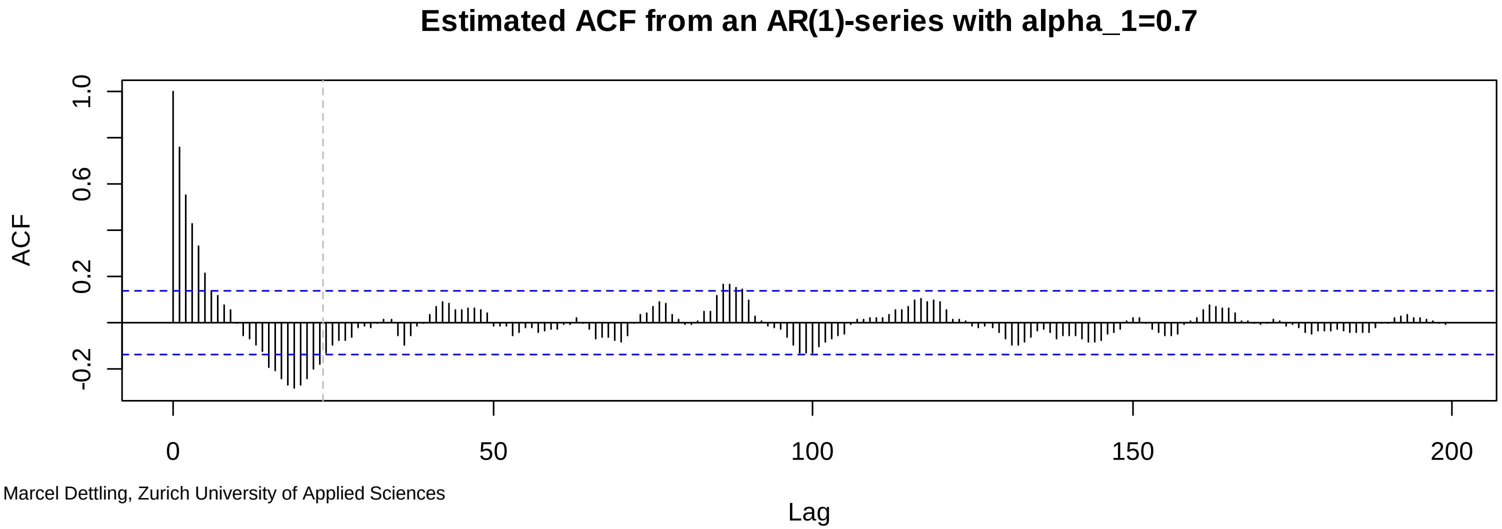
For answering the question above:

- For an AR(1) time series process, we know the true ACF
- We generate a number of realizations from this process
- We record the ACF estimates and compare to the truth

Applied Time Series Analysis
SS 2014 – Week 03

Theoretical vs. Estimated ACF





How Well Can We Estimate the ACF?

A) For AR(1)-processes we understand the theoretical ACF

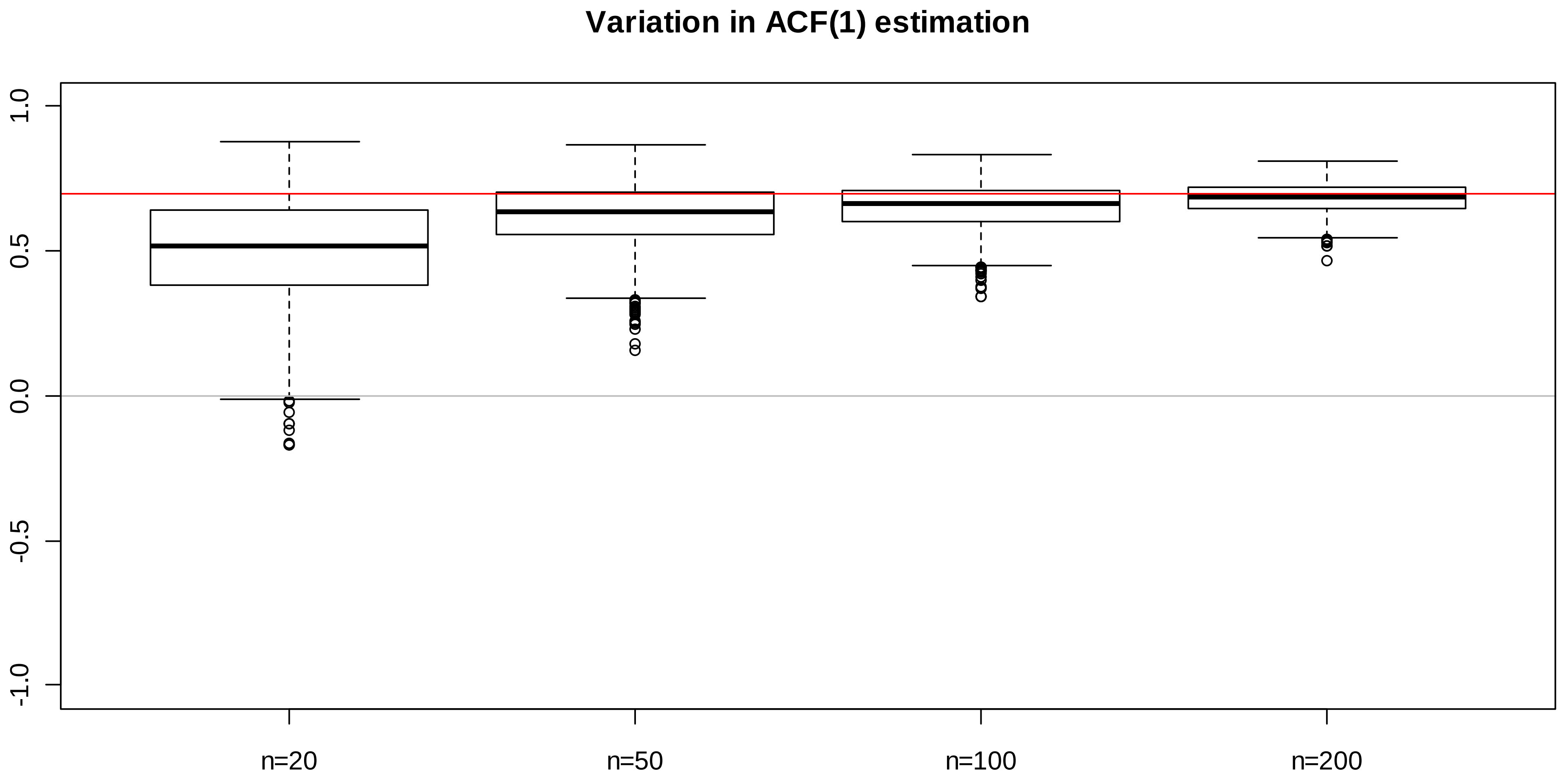
B) Repeat for i=1, …, 1000

Simulate a **length n** AR(1)-process
Estimate the ACF from that realization

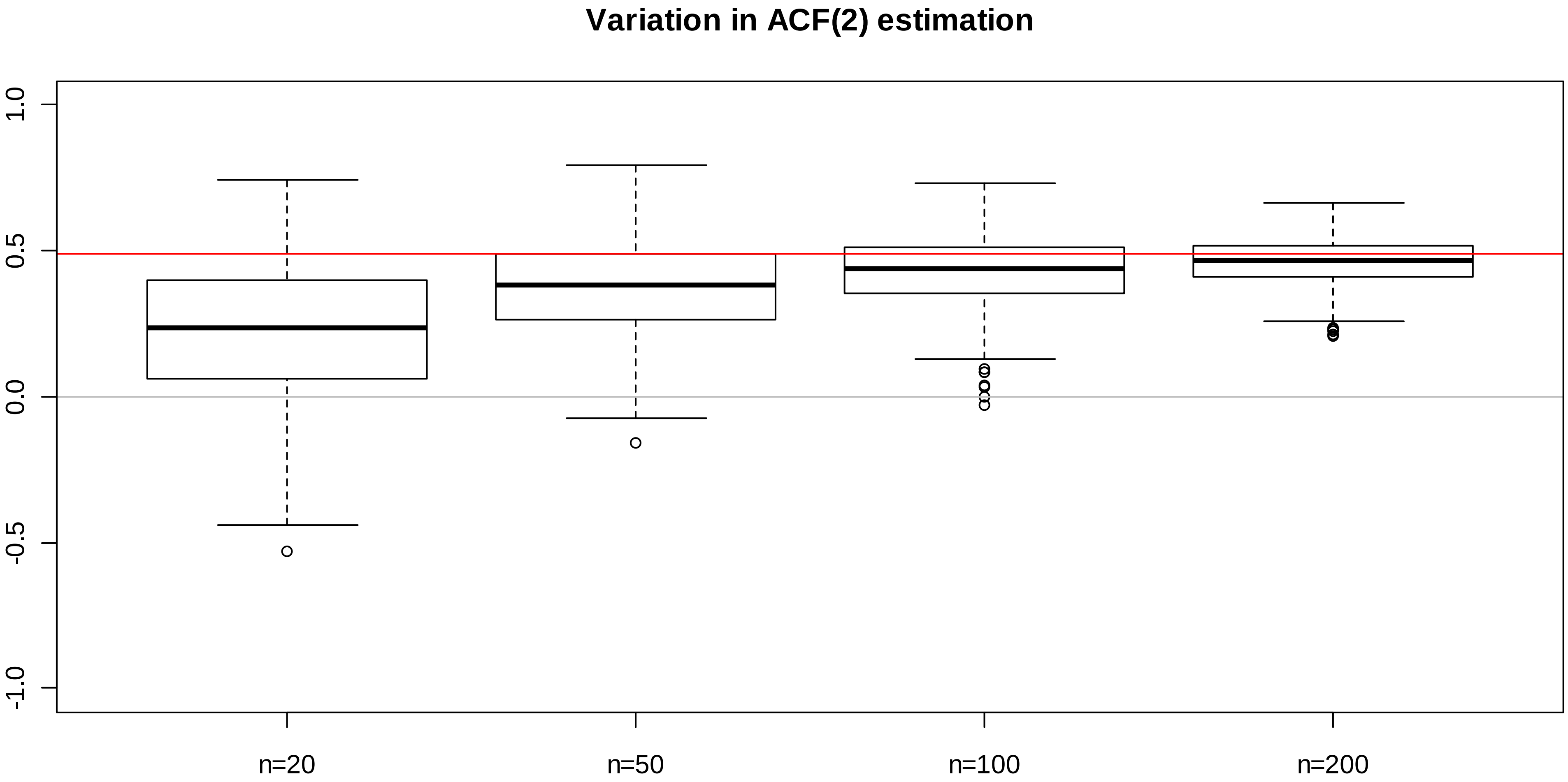
End for

C) Boxplot the (bootstrap) sample distribution of ACF-estimates
Do so for different **lags k** and different series **length n**

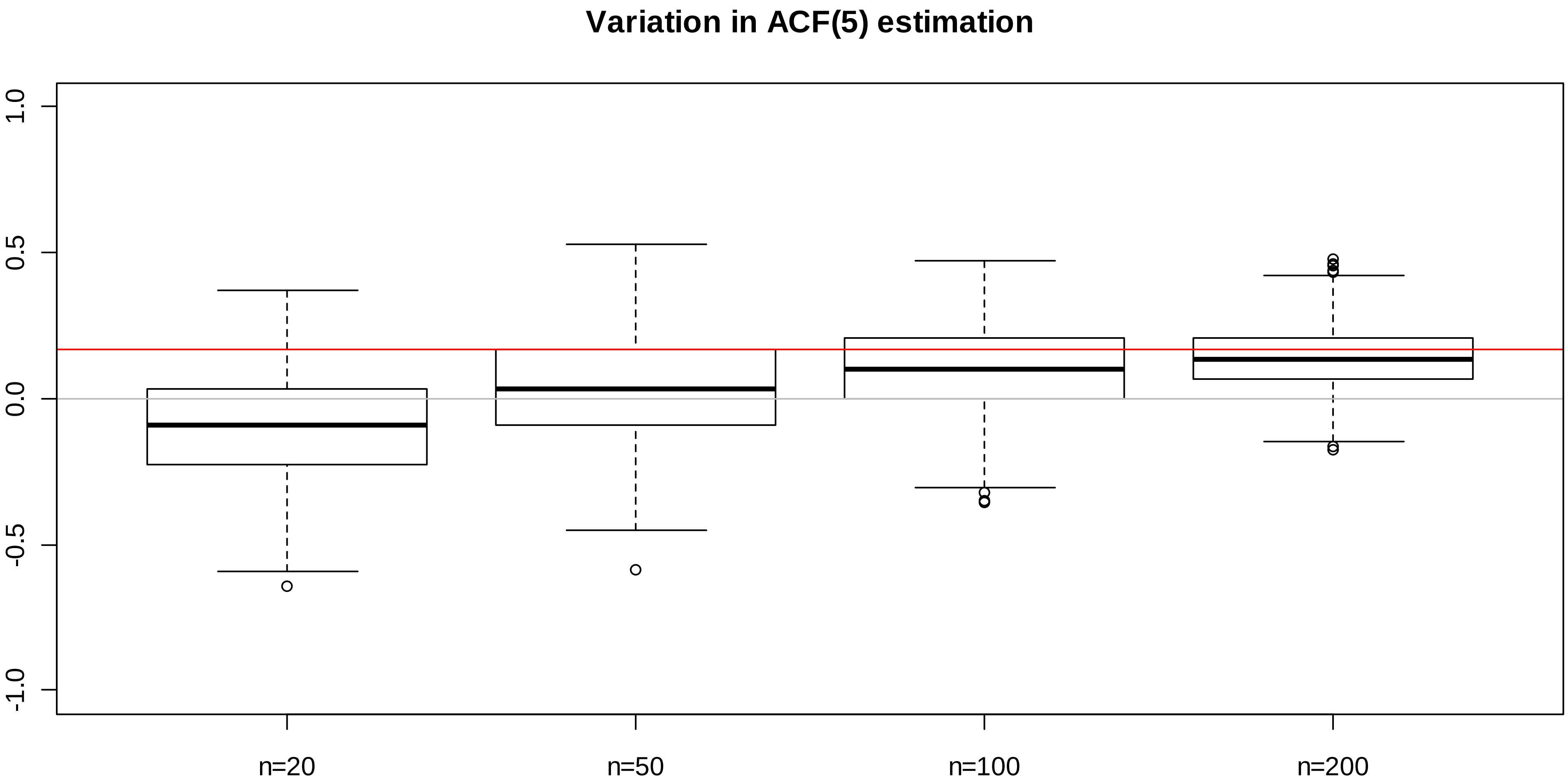
How Well Can We Estimate the ACF?



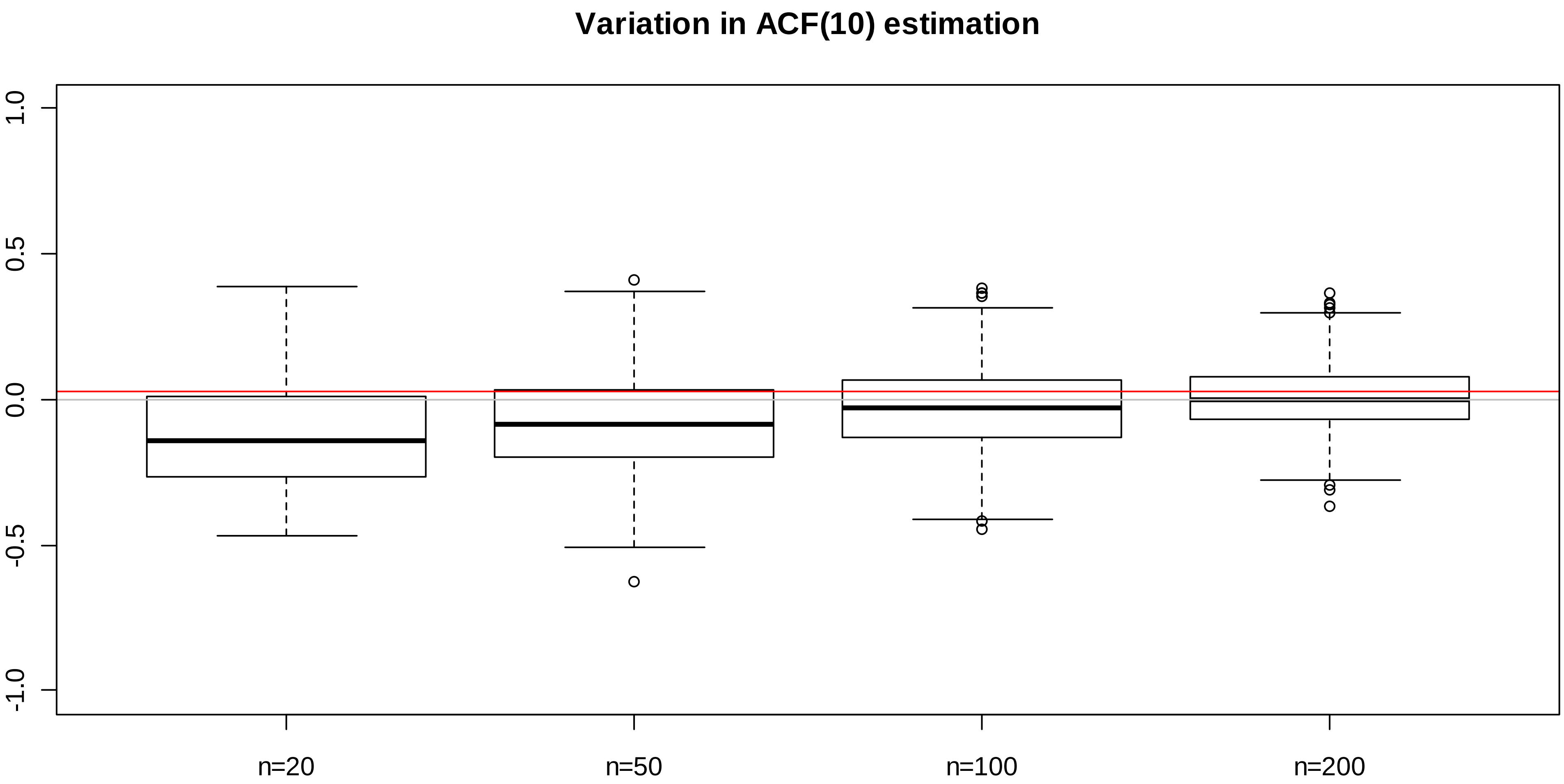
How Well Can We Estimate the ACF?



How Well Can We Estimate the ACF?



How Well Can We Estimate the ACF?



Trivia ACF Estimation

- In short series, the ACF is strongly biased. The consistency kicks in and kills the bias only after ~100 observations.
- The variability in ACF estimation is considerable. We observe that we need at least 50, or better, 100 observations.
- For higher lags k, the bias seems a little less problematic, but the variability remains large even with many observations n.
- The confidence bounds, derived under independence, are not very accurate for (dependent) time series.

→ ***Interpreting the ACF is tricky!***

Application: Variance of the Arithmetic Mean

Practical problem: we need to estimate the mean of a realized/ observed time series. We would like to attach a standard error.

- If we estimate the mean of a time series without taking into account the dependency, the standard error will be flawed.
- This leads to misinterpretation of tests and confidence intervals and therefore needs to be corrected.
- The standard error of the mean can both be over-, but also underestimated. This depends on the ACF of the series.

→ **For the derivation, see the blackboard…**

Partial Autocorrelation Function (PACF)

The k^{th} partial autocorrelation π_k is defined as the correlation between X_{t+k} and X_t, given all the values in between.

$$\pi_k = Cor(X_{t+k}, X_t \mid X_{t+1} = x_{t+1}, ..., X_{t+k-1} = x_{t+k-1})$$

Interpretation:

- Given a time series X_t, the partial autocorrelation of lag *k*, is the autocorrelation between X_t and X_{t+k} with the linear dependence of X_{t+1} through to X_{t+k-1} removed.
- One can draw an analogy to regression. The ACF measures the „simple“ dependence between X_t and X_{t+k}, whereas the PACF measures that dependence in a „multiple“ fashion.

Facts About the PACF and Estimation

We have:

- $\pi_1 = \rho_1$
- $\pi_2 = \dfrac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$ for AR(1) models, we have $\pi_2 = 0$, because $\rho_2 = \rho_1^2$
- For estimating the PACF, we utilize the fact that for any AR(p) model, we have: $\pi_p = \alpha_p$ and $\pi_k = 0$ for all $k > p$.

  Thus, for finding $\hat{\pi}_p$, we fit an AR(p) model to the series for various orders p and set $\hat{\pi}_p = \hat{\alpha}_p$

Facts about the PACF

- Estimation of the PACF is implemented in R.
- The first PACF coefficient is equal to the first ACF coefficient. Subsequent coefficients are not equal, but can be derived from each other.
- For a time series generated by an AR(p)-process, the p^{th} PACF coefficient is equal to the p^{th} AR-coefficient. All PACF coefficients for lags $k > p$ are equal to 0.
- Confidence bounds also exist for the PACF.