Applied Time Series Analysis
SS 2013 – Week 04

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 11, 2013

Applied Time Series Analysis
SS 2013 – Week 04

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 2013 – Week 04

Autocorrelation

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

Def: **Autocorrelation**

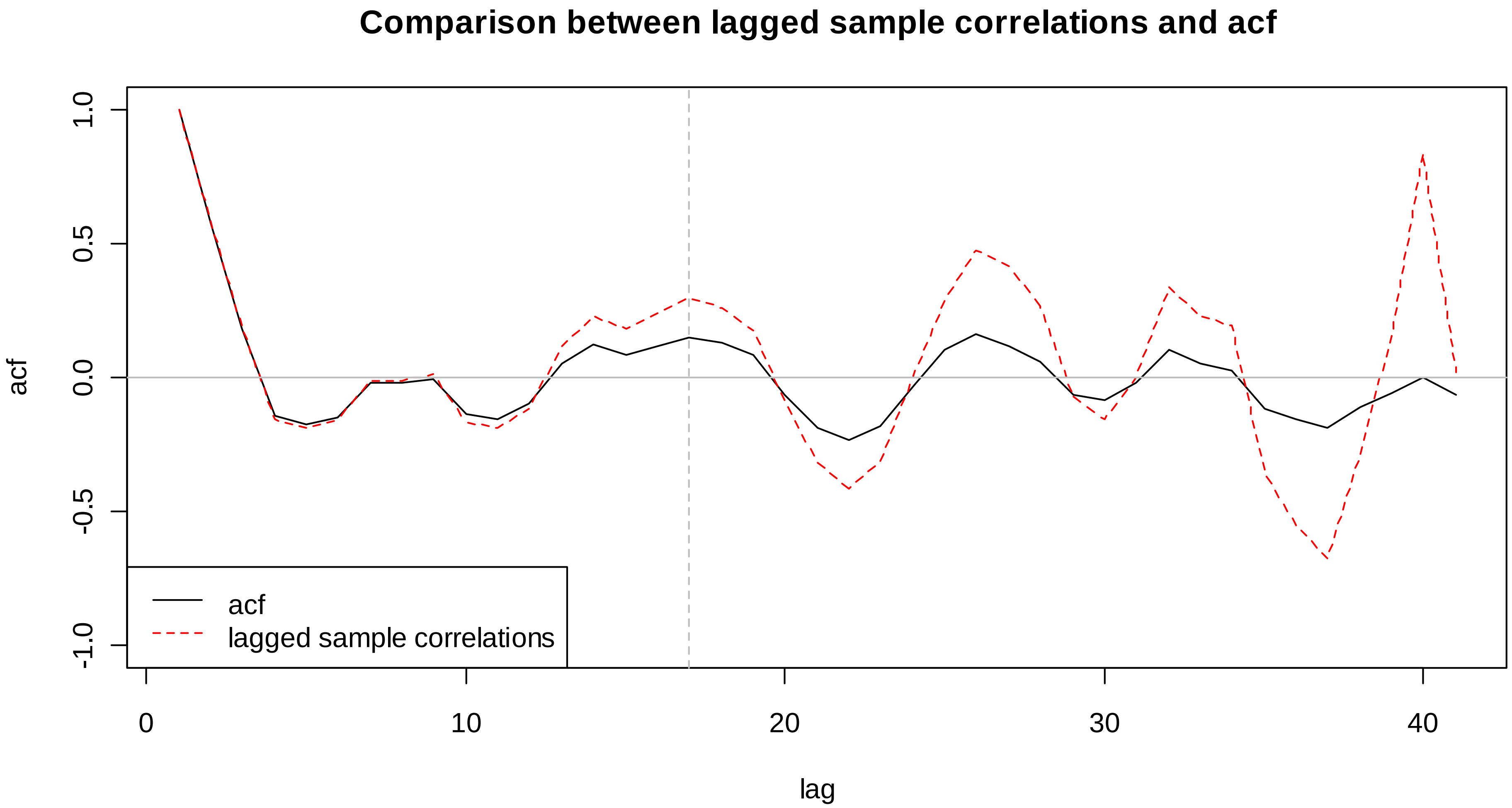
$$Cor(X_{t+k}, X_t) = \frac{Cov(X_{t+k}, X_t)}{\sqrt{Var(X_{t+k}) \cdot Var(X_t)}}$$

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 the blackboard for a sketch of the two approaches…*

Applied Time Series Analysis
SS 2013 – Week 04

Comparison Idea 1 vs. Idea 2



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}.

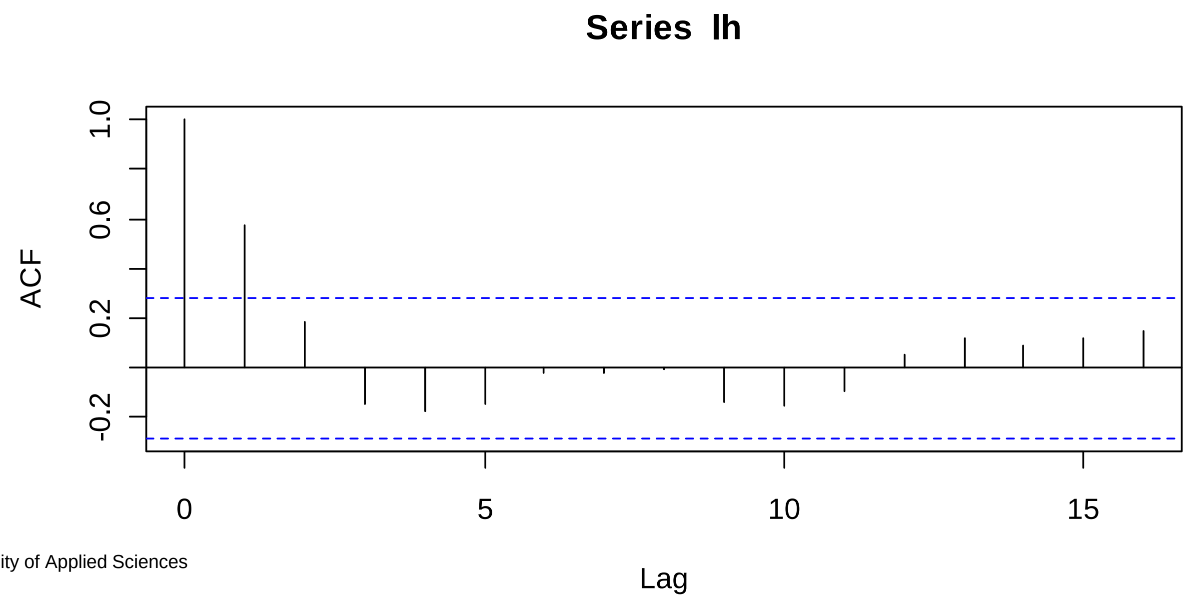
→ Thus, in our example, X_{t-1} accounts for roughly 49% of the variability observed in random variable X_t. Only roughly because the world is not linear.

→ 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.

Random Series – Confidence Bands

If a time series is completely random, i.e. consists of i.i.d. random variables X_t, the (theoretical) autocorrelations $\rho(k)$ are equal to 0.

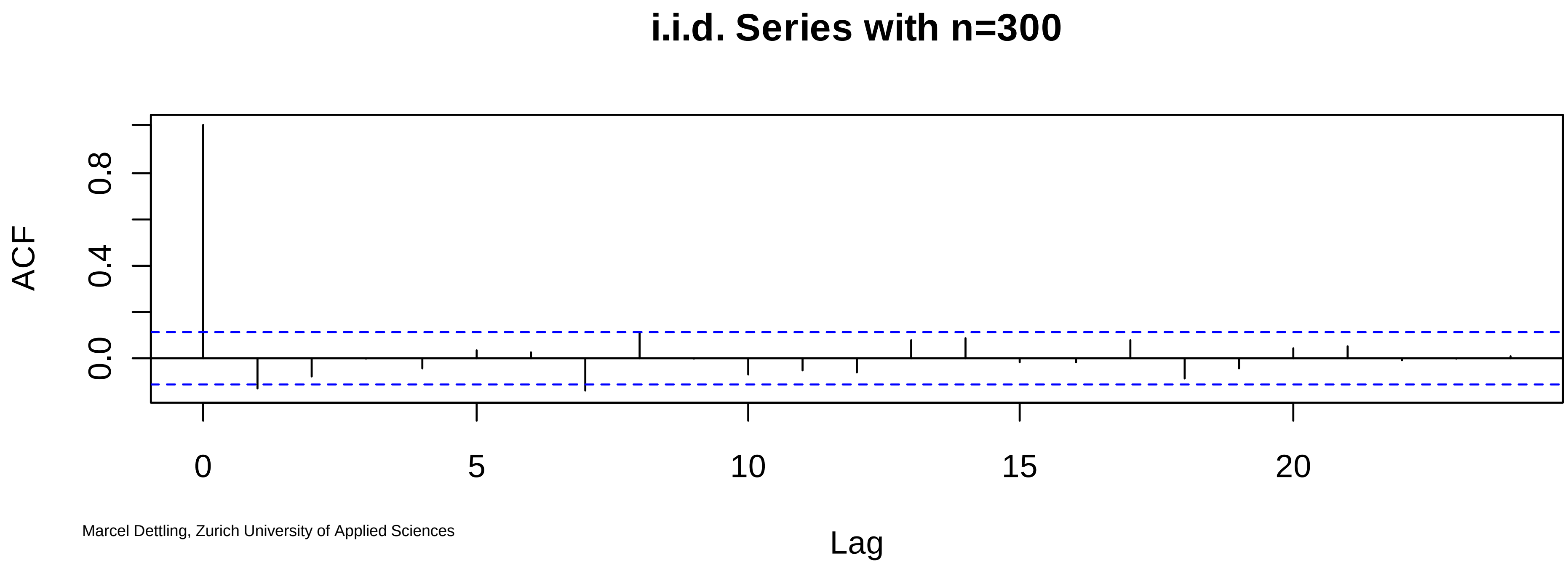
However, the estimated $\hat{\rho}(k)$ are not. We thus need to decide, whether an observed $\hat{\rho}(k) \neq 0$ is significantly so, or just appeared by chance. This is the idea behind the confidence bands.



Random Series – Confidence Bands

For long i.i.d. time series, it can be shown that the $\hat{\rho}(k)$ are approximately $N(0,1/n)$ distributed.

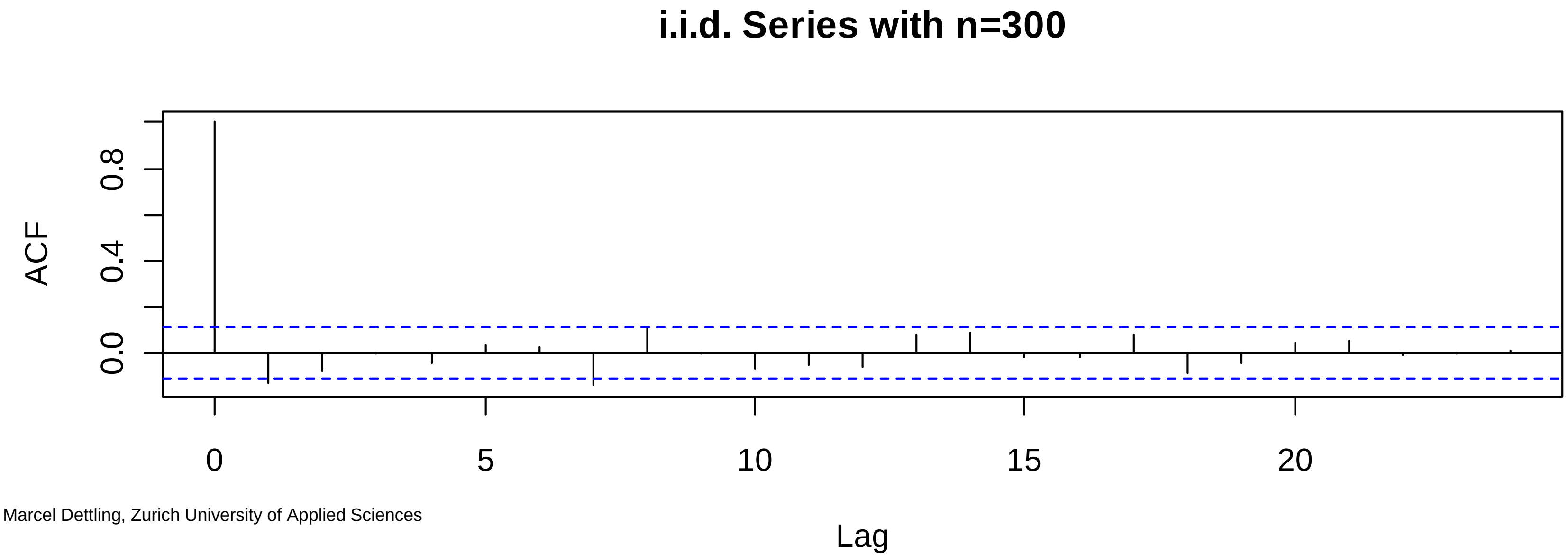
Thus, if a series is random, 95% of the estimated $\hat{\rho}(k)$ can be expected to lie within the interval $\pm 2/\sqrt{n}$



Random Series – Confidence Bands

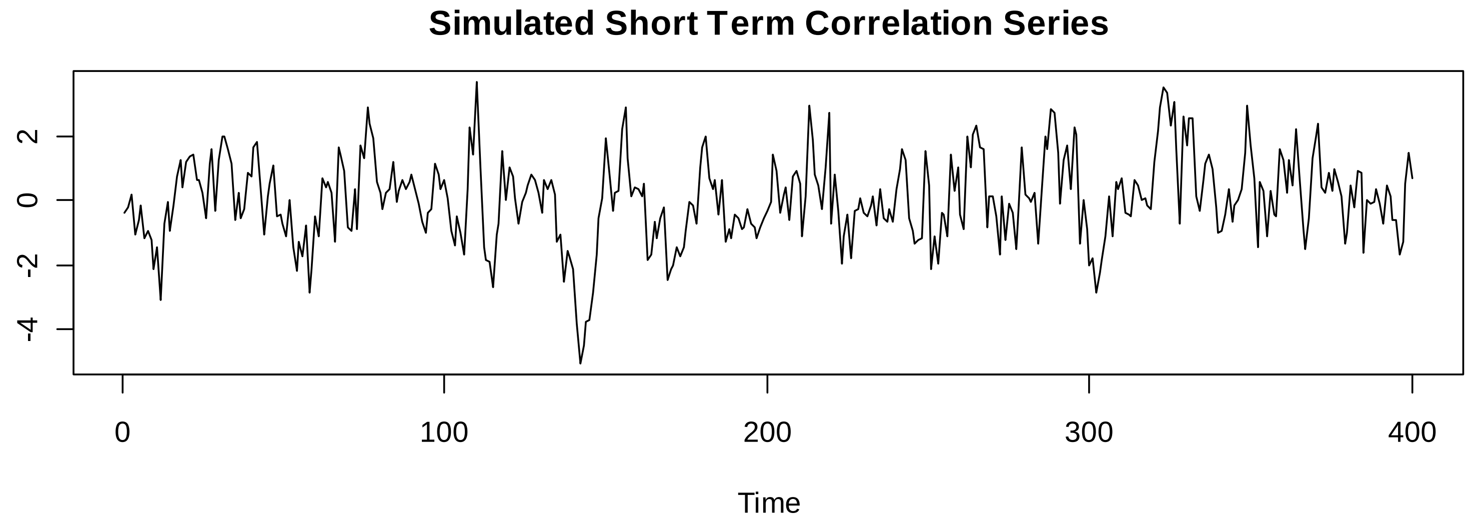
Thus, even for a (long) i.i.d. time series, we expect that 5% of the estimated autocorrelation coeffcients exceed the confidence bounds. They correspond to type I errors.

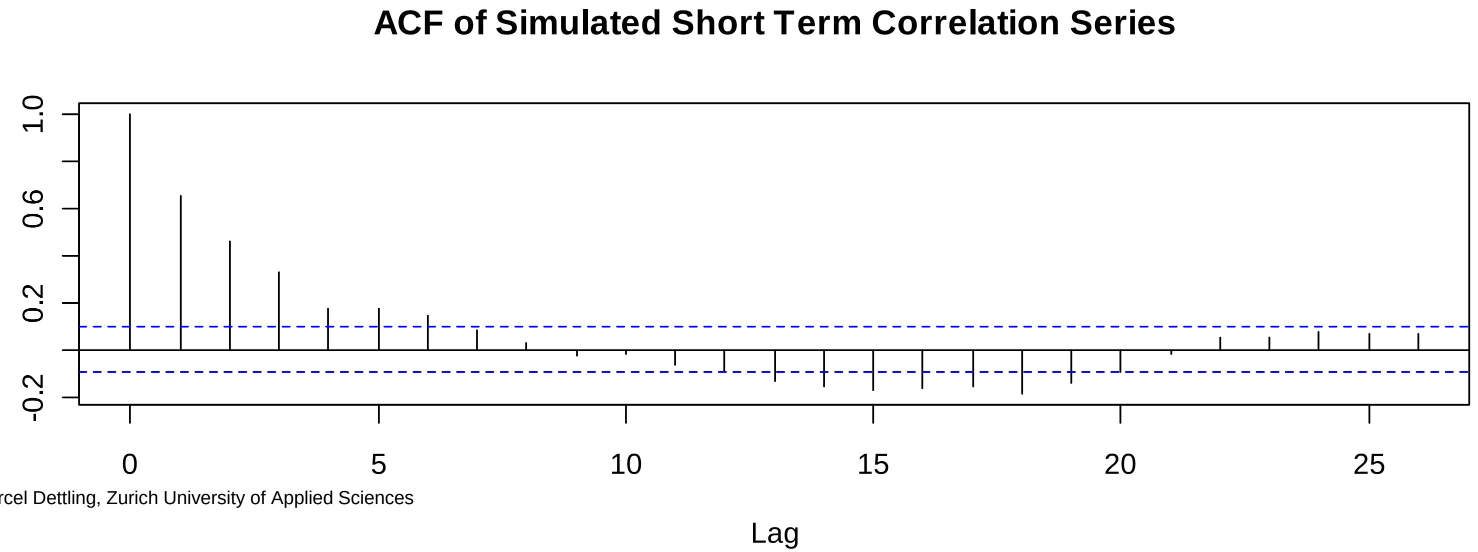
Note: the probabilistic properties of non-normal i.i.d series are much more difficult to derive.



Applied Time Series Analysis
SS 2013 – Week 04

Short Term Correlation





Short Term Correlation

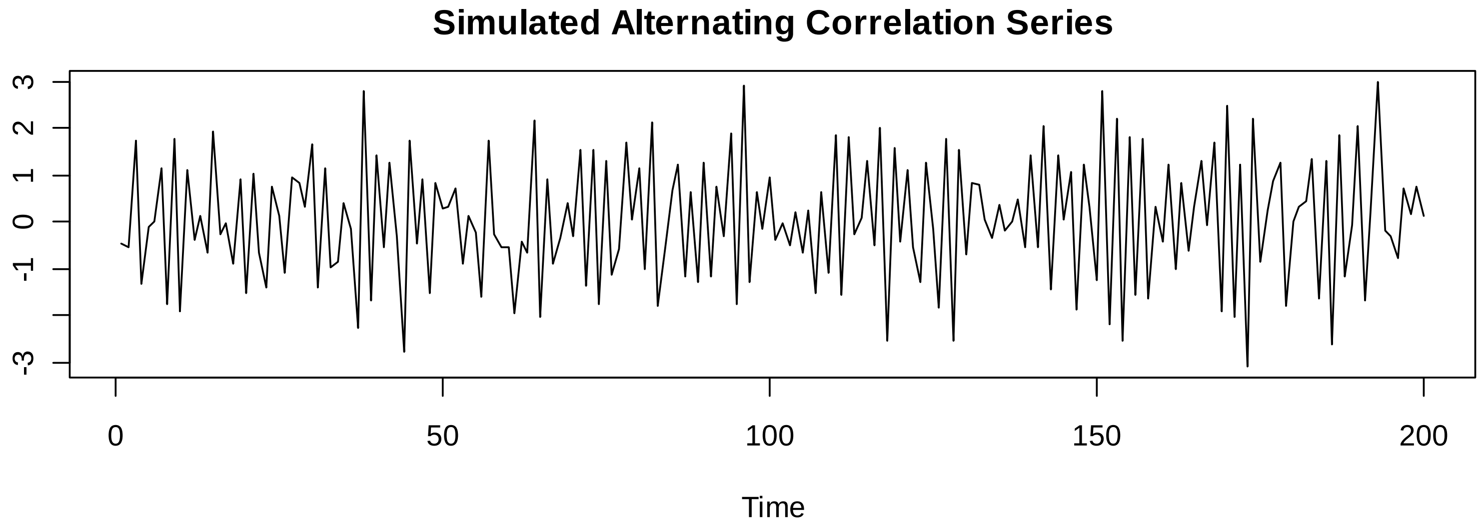
Stationary series often exhibit short-term correlation, characterized by a fairly large value of $\hat{\rho}(1)$, followed by a few more coefficients which, while significantly greater than zero, tend to get successively smaller. For longer lags k, they are close to 0.

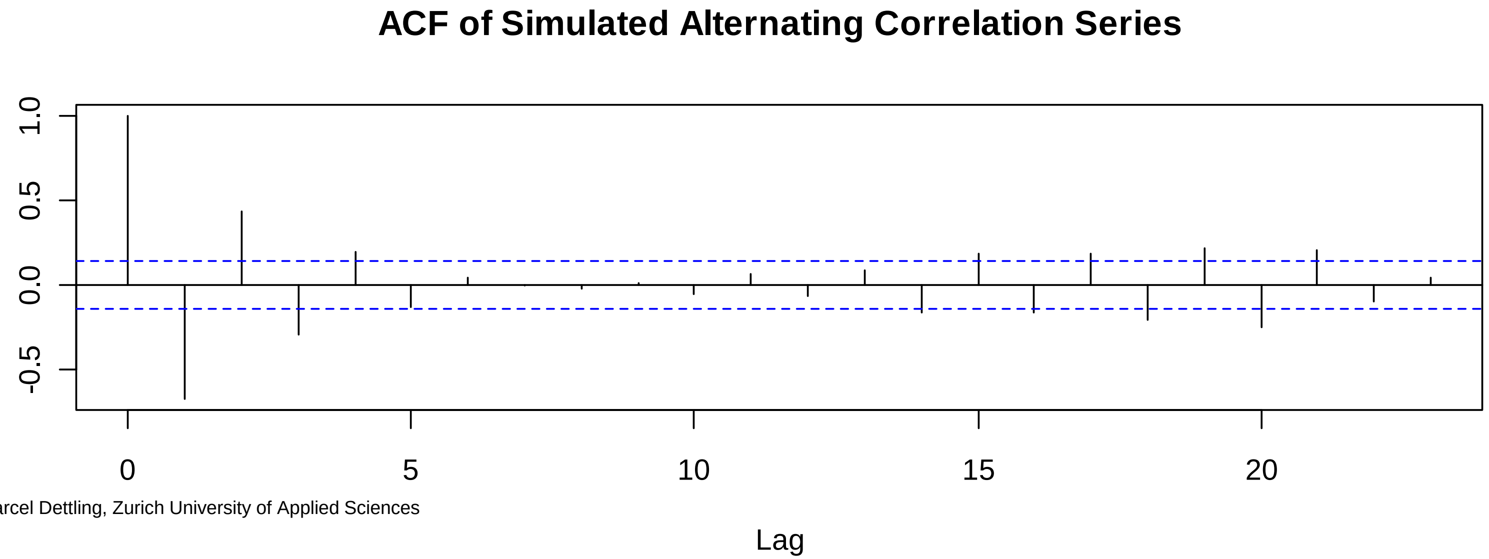
A time series which gives rise to such a correlogram, is one for which an observation above the mean tends to be followed by one or more further observations above the mean, and similarly for observations below the mean.

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

Applied Time Series Analysis
SS 2013 – Week 04

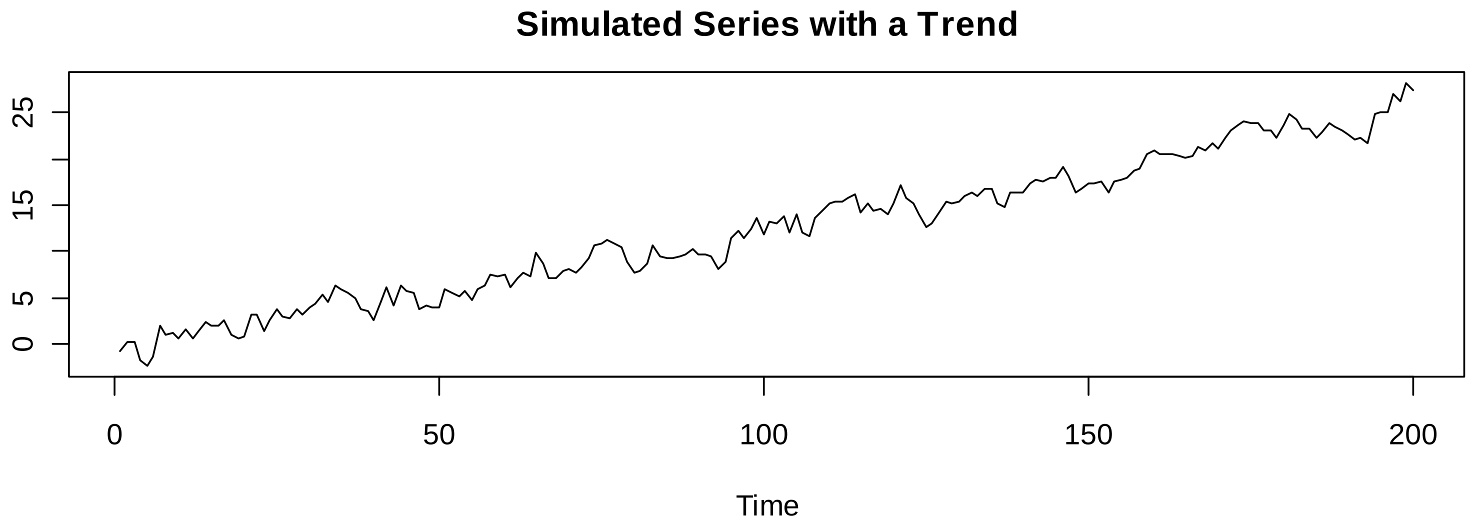
Alternating Time Series

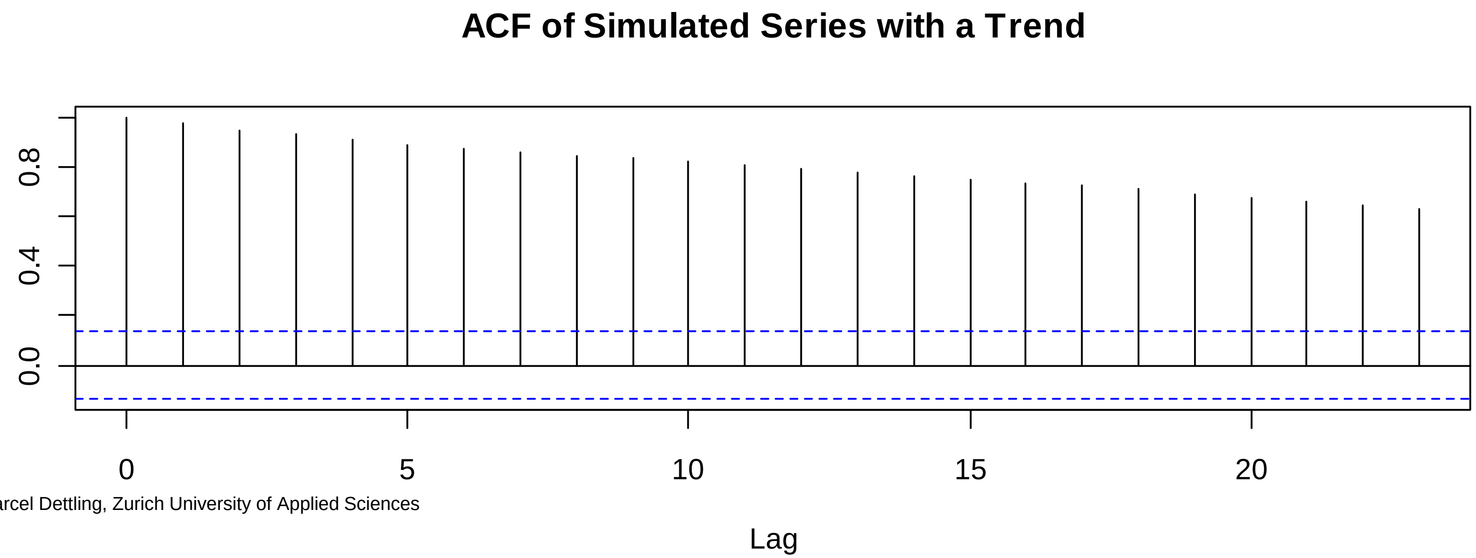




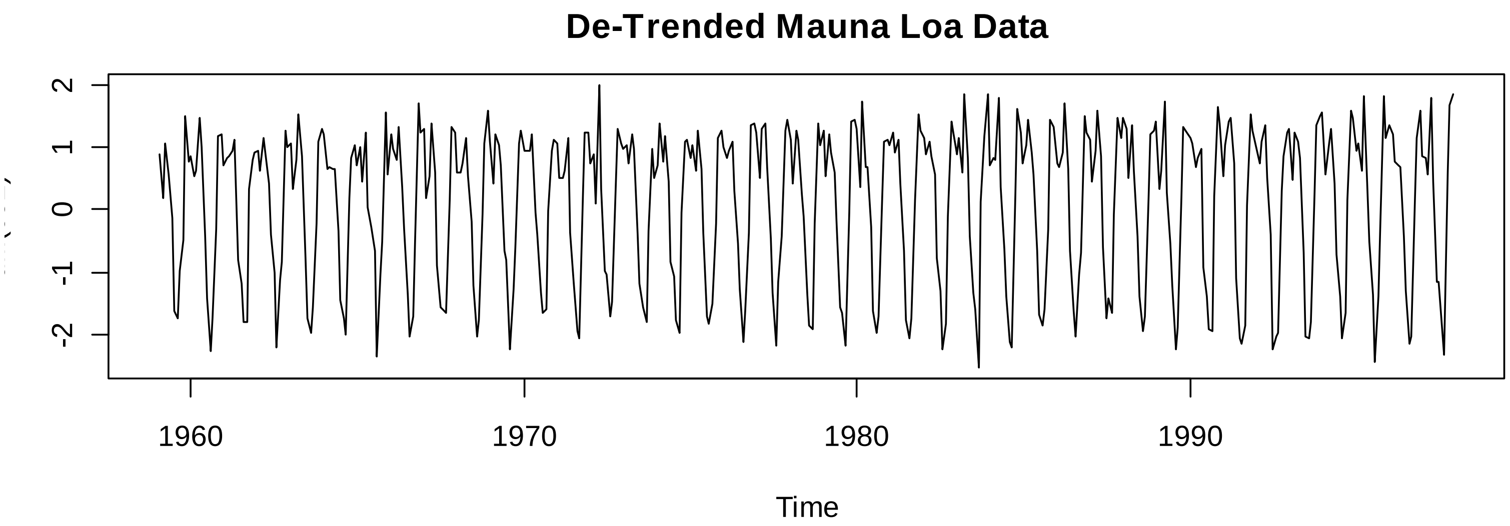
Applied Time Series Analysis
SS 2013 – Week 04

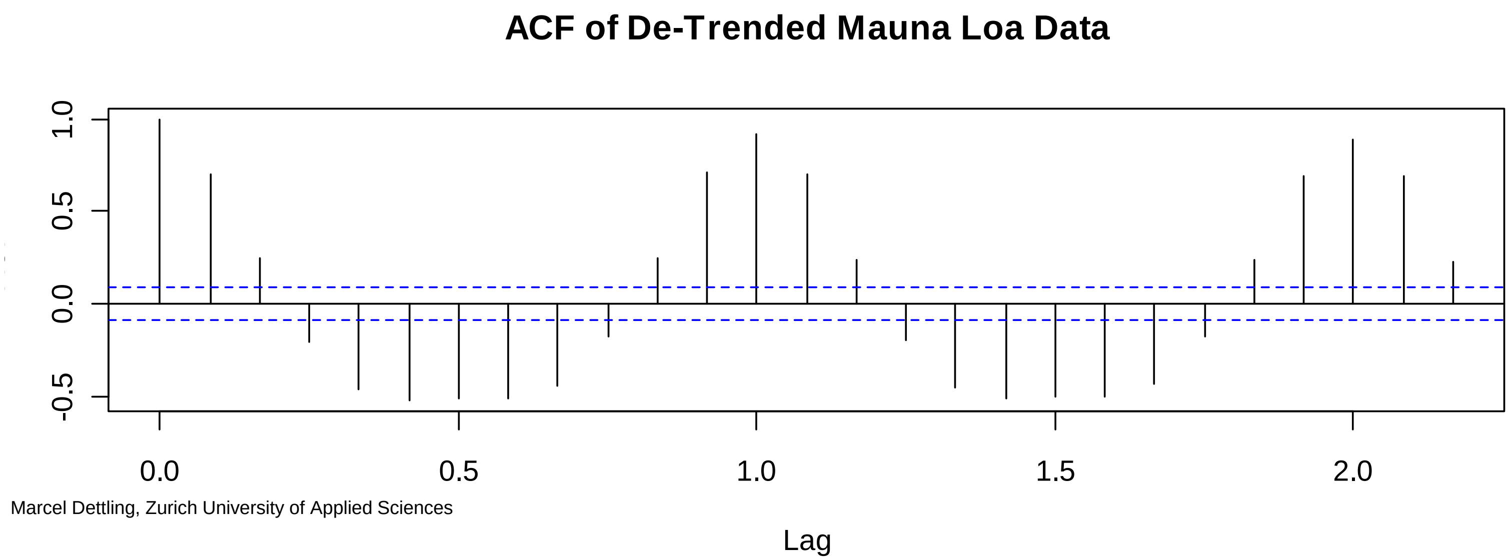
Non-Stationarity in the ACF: Trend



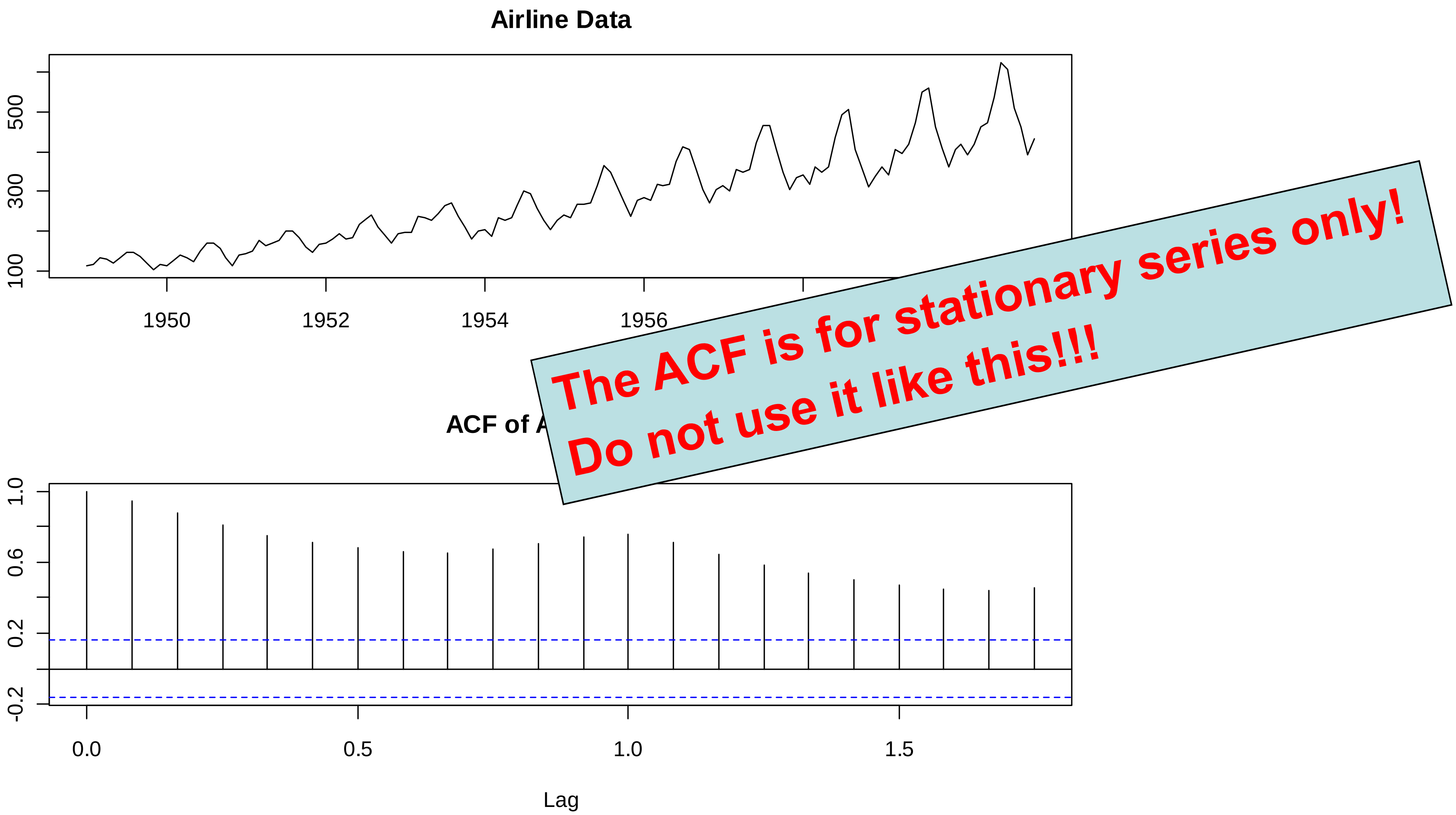


Non-Stationarity in the ACF: Seasonal Pattern





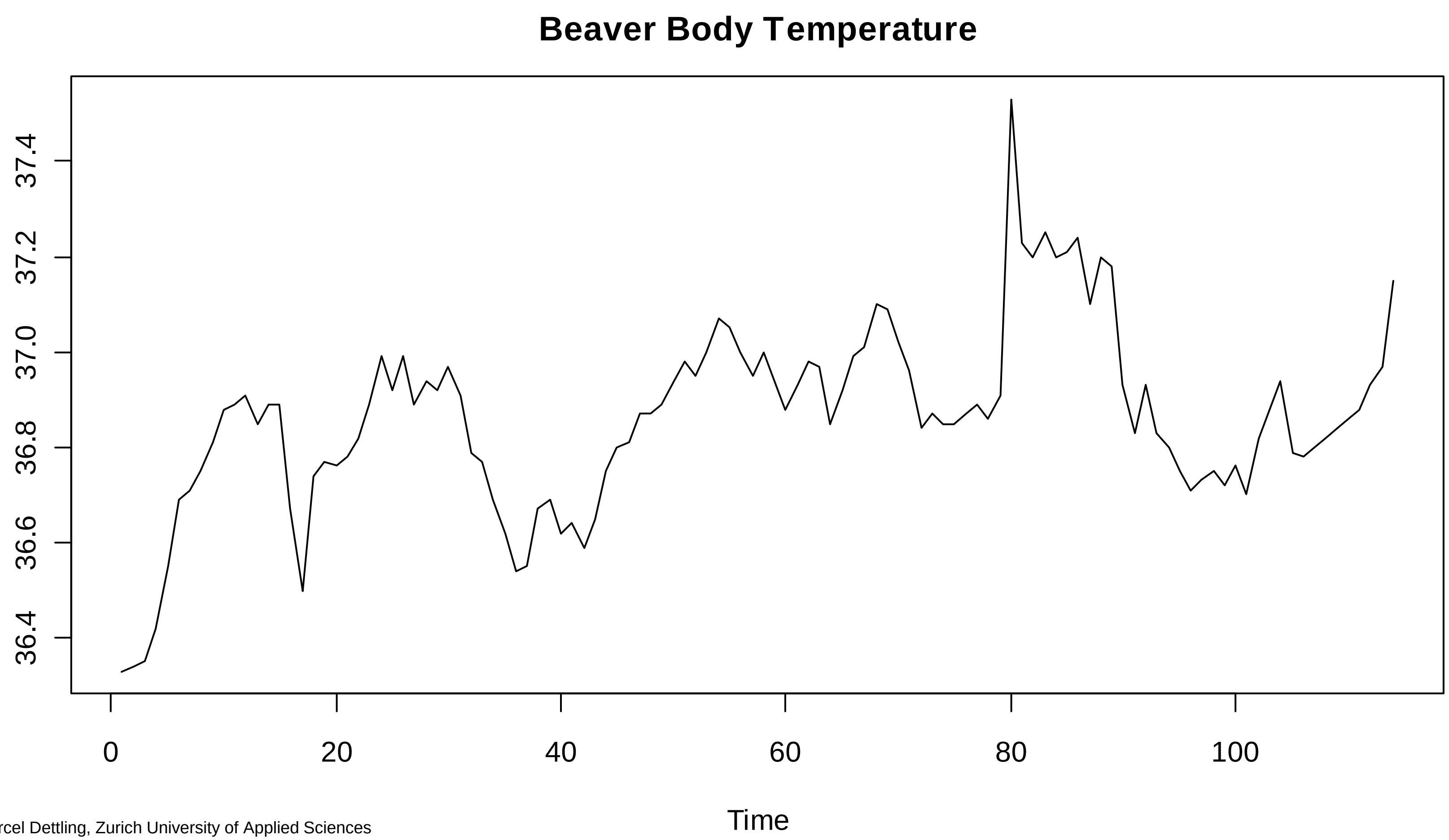
ACF of the Raw Airline Data



Applied Time Series Analysis
SS 2013 – Week 04

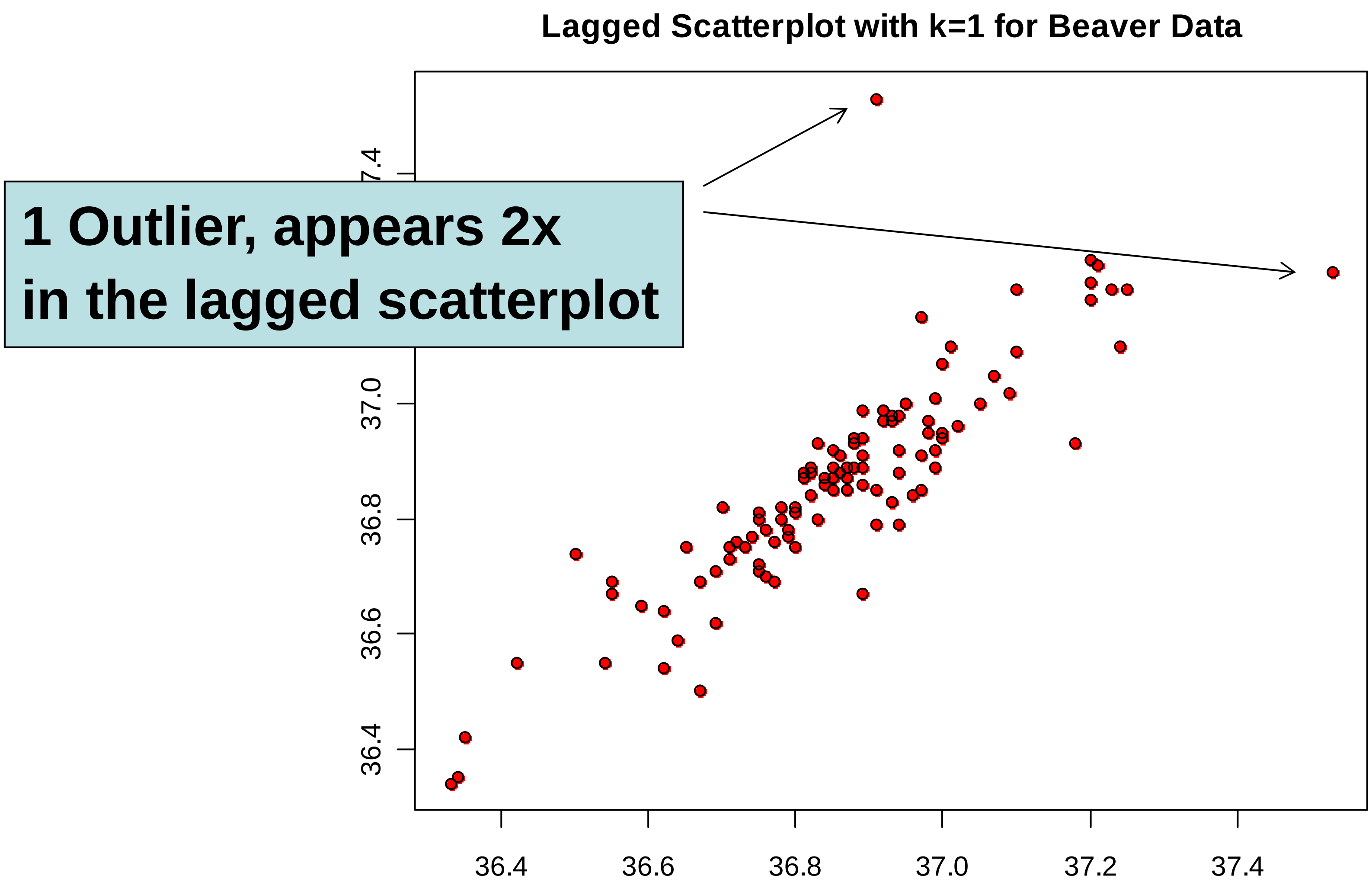
Outliers and the ACF

Outliers in the time series strongly affect the ACF estimation!



Applied Time Series Analysis
SS 2013 – Week 04

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 an outlier: 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…

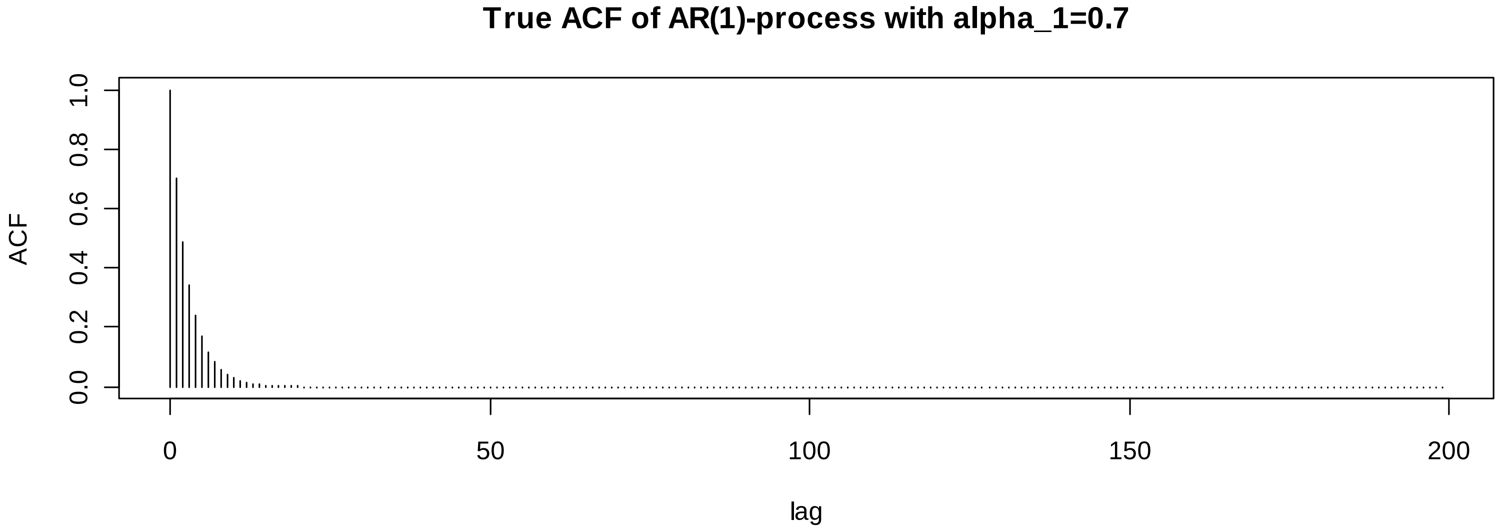
→ ACF estimation is not easy, and interpretation is tricky.

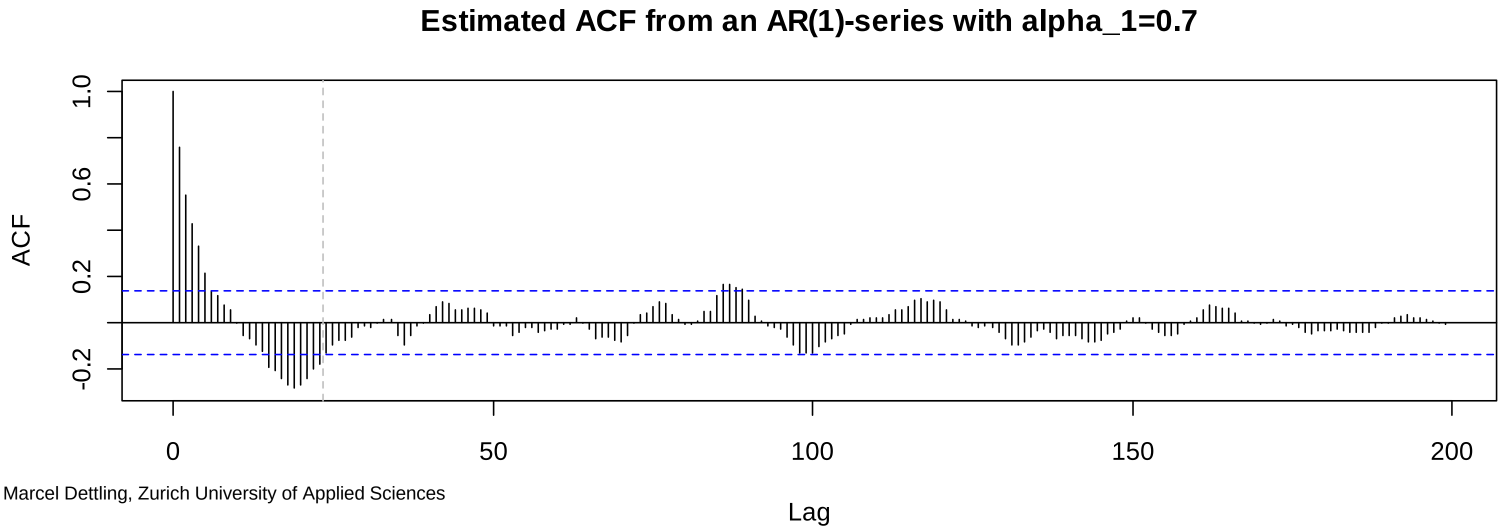
For answering the question above:

- 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 2013 – Week 04

Theoretical vs. Estimated ACF





How Well Can We Estimate the ACF?

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

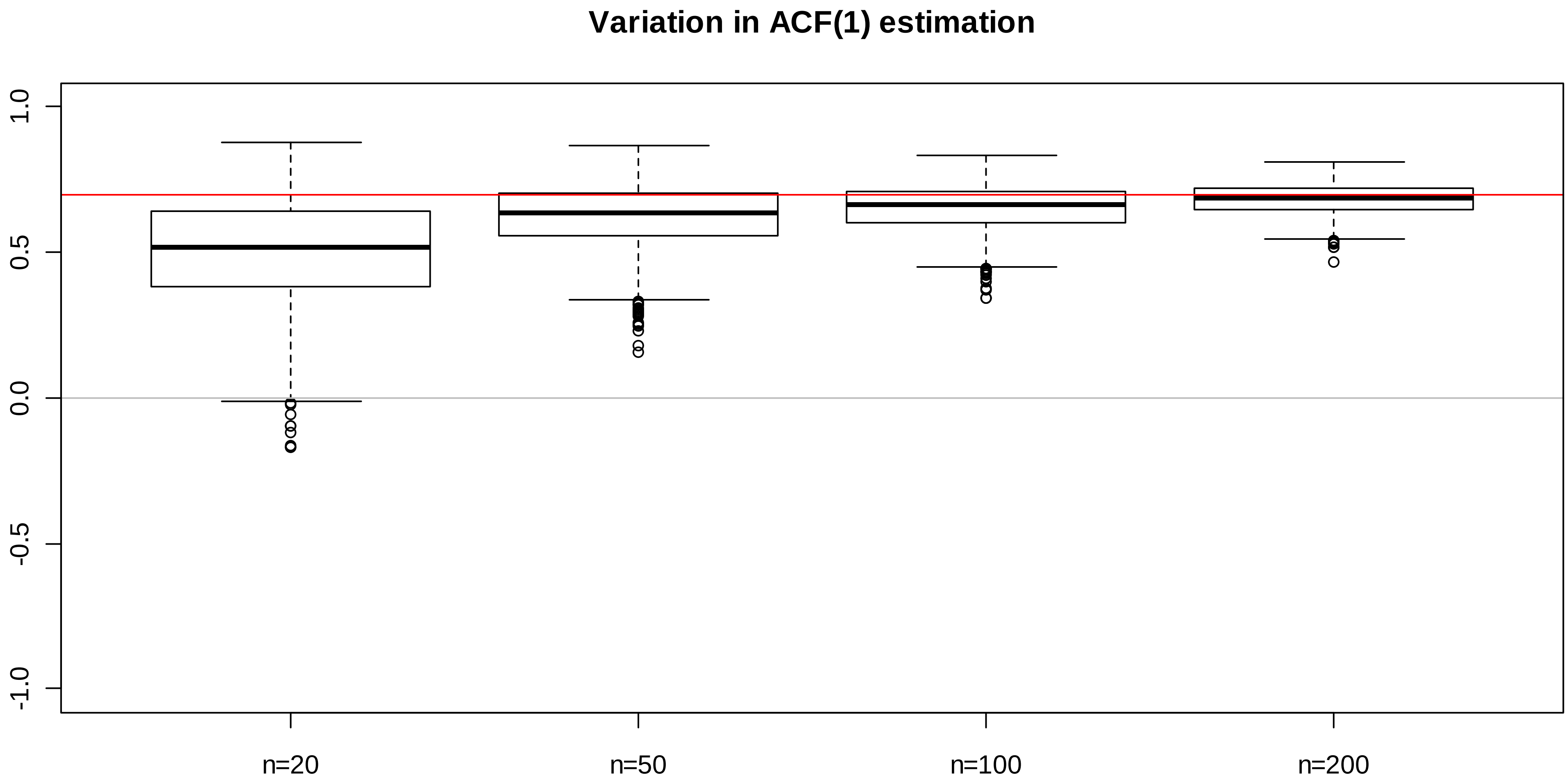
B) Repeat for i=1, …, 1000

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?



Applied Time Series Analysis
SS 2013 – Week 04

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.