PSTAT 174/274: Time Series

Prof. Dr. Gareth W. Peters

(CStat-RSS, CMath-IMA, FIOR, SIRM, FRSS, FIMA, YAS-RSE)

Professor of Statistics for Risk and Insurance &
Janet and Ian Duncan Endowed Chair in Actuarial Science
University of California Santa Barbara

January 15, 2023

PART IV: Statistical Structure in Time Series and Decomposition.

Time Series Based on Signals in Noise

In time series modelling there are multiple approaches one can adopt:

- Assume a particular model structure (signal structure) for the time series data generation process. Then seek to learn the properties of this signal structure in the presence of noise.
  - This is often used in Physical modelling or Engineering settings where the system under study has physical laws it must obey that lead to expectation of typical signal structures.
- Data driven approaches or statistical structures - where one makes more general assumptions on the data generating process and seeks to learn a flexible model structure or Time Series Decomposition (not governed by a set of physical laws necessarily).

UC SANTA BARBARA UCSB
Prof. Gareth W. Peters

Simple Temporal Models in TS settings:

Example

A non-stationary process (simplest) is given by

$$Y_t = \kappa_t + \epsilon_t = \underbrace{\beta t}_{\text{deterministic component}} + \underbrace{\epsilon_t}_{\text{stochastic component}}, \quad \epsilon_t \sim WN$$

- *set deterministic process* κ_t *as linear time trend s.t.* $\mathbb{E}[Y_t] = \beta t$ *depends on* t
- *However,* $X_t = Y_t - \beta t$ *is covariance stationary.*

Example

Another example could be a signal in noise given by a model:

$$Y_t = f(t; \boldsymbol{\theta}) + \epsilon_t$$

Note - it could be linear or non-linear in parameters of model/signal $\boldsymbol{\theta}$.

Example of a signal in noise model where we consider κ_t as the realisations of some function at times $t = t_1, t_2, \ldots$ (typically default $t = 1, 2, 3, \ldots$) for instance given by:

$$Y_t = A\cos\left(\frac{2\pi t}{p} + \theta\right) + \epsilon_t$$

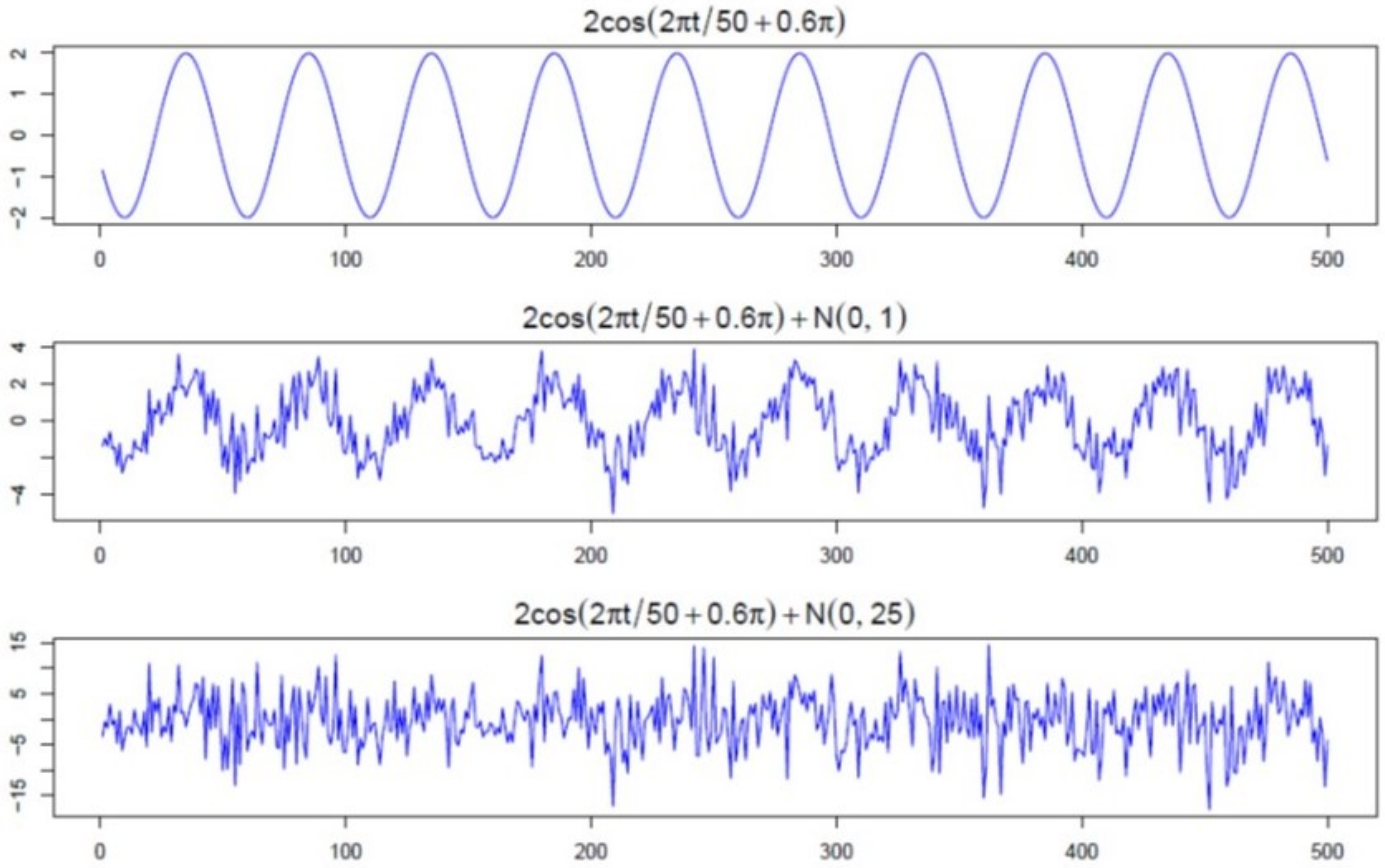
- the first component is deterministic and considered a "signal" and
- the second component is stochastic, denoted $\epsilon_t \sim WN(0, \sigma^2)$ and considered observation noise or innovation error or stochastic driver or risk driver - many names given in various literature's depending on interpretation and application context.

- NOTE: Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise.

TS models with time-trend

(source: Dewei Wang https://people.stat.sc.edu/wang52)

```
set.seed(100); cs = 2*cos(2*pi*1:500/50 + .6*pi); w = rnorm(500,0,1)
par(mfrow=c(3,1), mar=c(3,2,2,1), cex.main=1.5)
plot.ts(cs, main=expression(2*cos(2*pi*t/50+.6*pi)),col="blue")
plot.ts(cs+w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,1)),col="blue")
plot.ts(cs+5*w, main=expression(2*cos(2*pi*t/50+.6*pi) + N(0,25)),col="blue")
```



Statistical Model Structures in Time Series Data

UC SANTA BARBARA UCSB
Prof. Gareth W. Peters

Trend pattern exists when there is a long-term increase or decrease in the data.

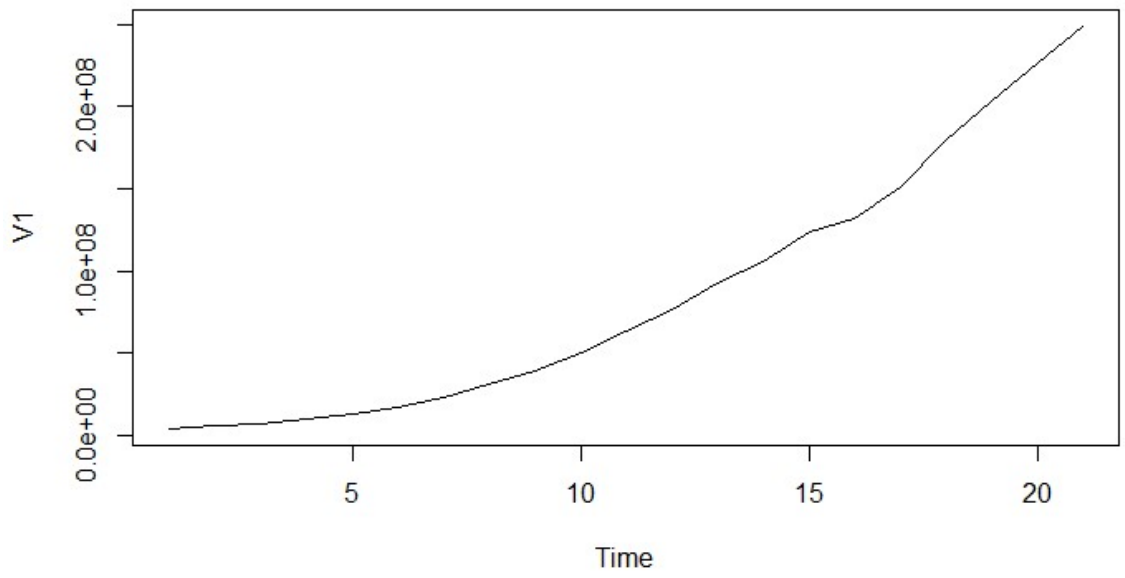
Seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).

Cyclic pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length;
- cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

US Population, 1790-1990, ten year intervals



X_t = population of US at year t (in millions):

$t = 1790,\ X_t = 3,929,214$

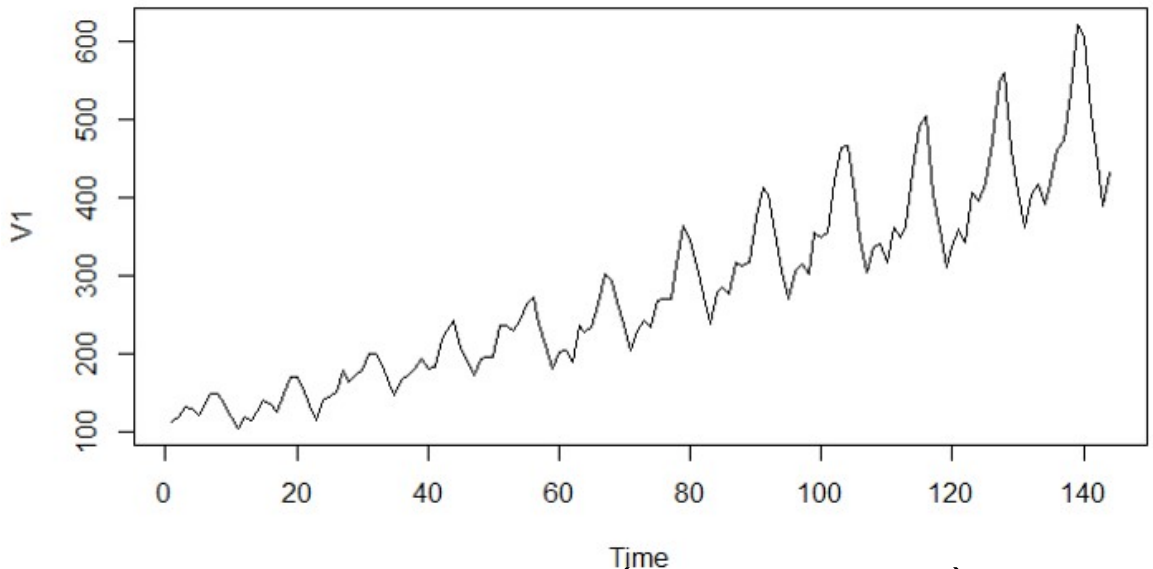
$t = 1800,\ X_t = 5,308,483$

...

$t = 1990,\ X_t = 248,709,873$

Note: looks like exponential trend. Non-stationary times series

International Airline Data. Monthly totals of international passengers (1/1949 –12/1960)

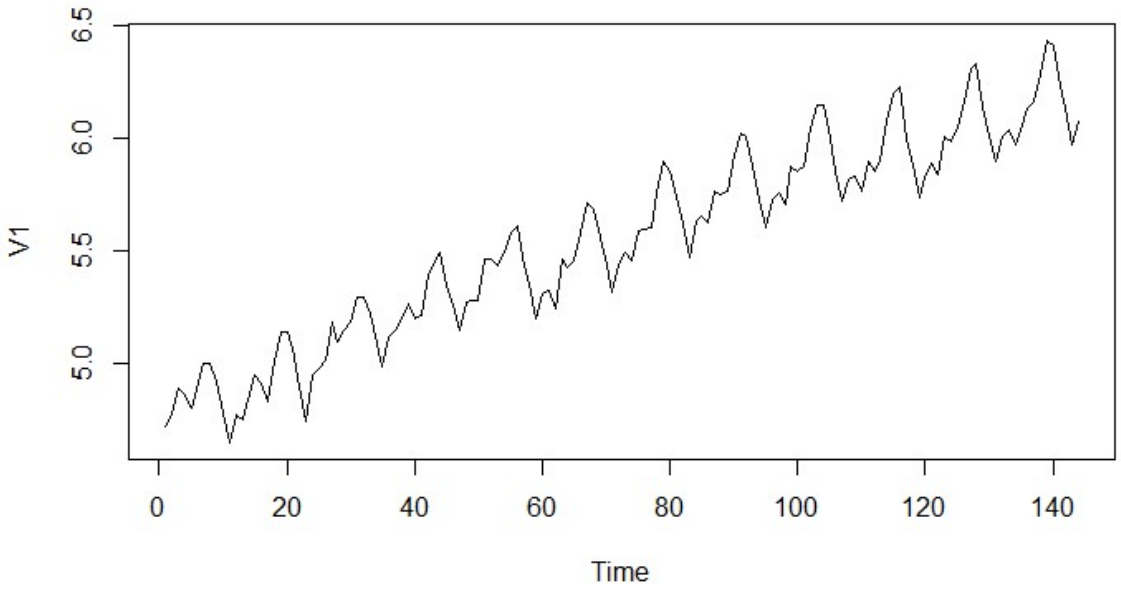


X_t = total number of passengers (in thousands) taking flights on the airline in a month t.

Non-stationary seasonal times series:

- upward trend (linear?)
- Seasonal (low in winter, high in summer)
- Variability increases with time

International Airline Data. Monthly totals of international passengers (1/1949 –12/1960)



$X_t =$ total number of passengers (in thousands) taking flights on the airline in a month t.

What happens if we transform the time series data? $V_t = \log(X_t)$, $t = 1, 2, \ldots, 144$, natural logarithm
$\Rightarrow$ stabilised volatility!

Additive or Multiplicative

Under an assumption of an additive decomposition, one can write

$$Y_t = S_t + T_t + R_t,$$

Under an assumption of a multiplicative decomposition, one can write

$$Y_t = S_t \times T_t \times R_t,$$

where

- y_t is the data
- S_t is the seasonal component
- T_t is the trend component
- R_t is the remainder component

Additive or Multiplicative

How should one decide between additive vs multiplicative?

- The additive decomposition is the most appropriate if the magnitude of the seasonal fluctuations, or the variation around the trend-cycle, does not vary with the level of the time series.
- When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative decomposition is more appropriate.

Additive or Multiplicative

NOTE: Multiplicative decompositions are common with economic time series.

- An alternative to using a multiplicative decomposition is to first transform the data until the variation in the series appears to be stable over time, then use an additive decomposition.
- eg. log transform, sqrt or Box-Cox transform (discussed more later)

When a log transformation has been used, this is equivalent to using a multiplicative decomposition because

$$Y_t = S_t \times T_t \times R_t,$$

is equivalent to

$$\ln Y_t = \ln S_t + \ln T_t + \ln R_t,$$

So mostly in practice use the additive form.

Trend and Seasonal Variation-summary

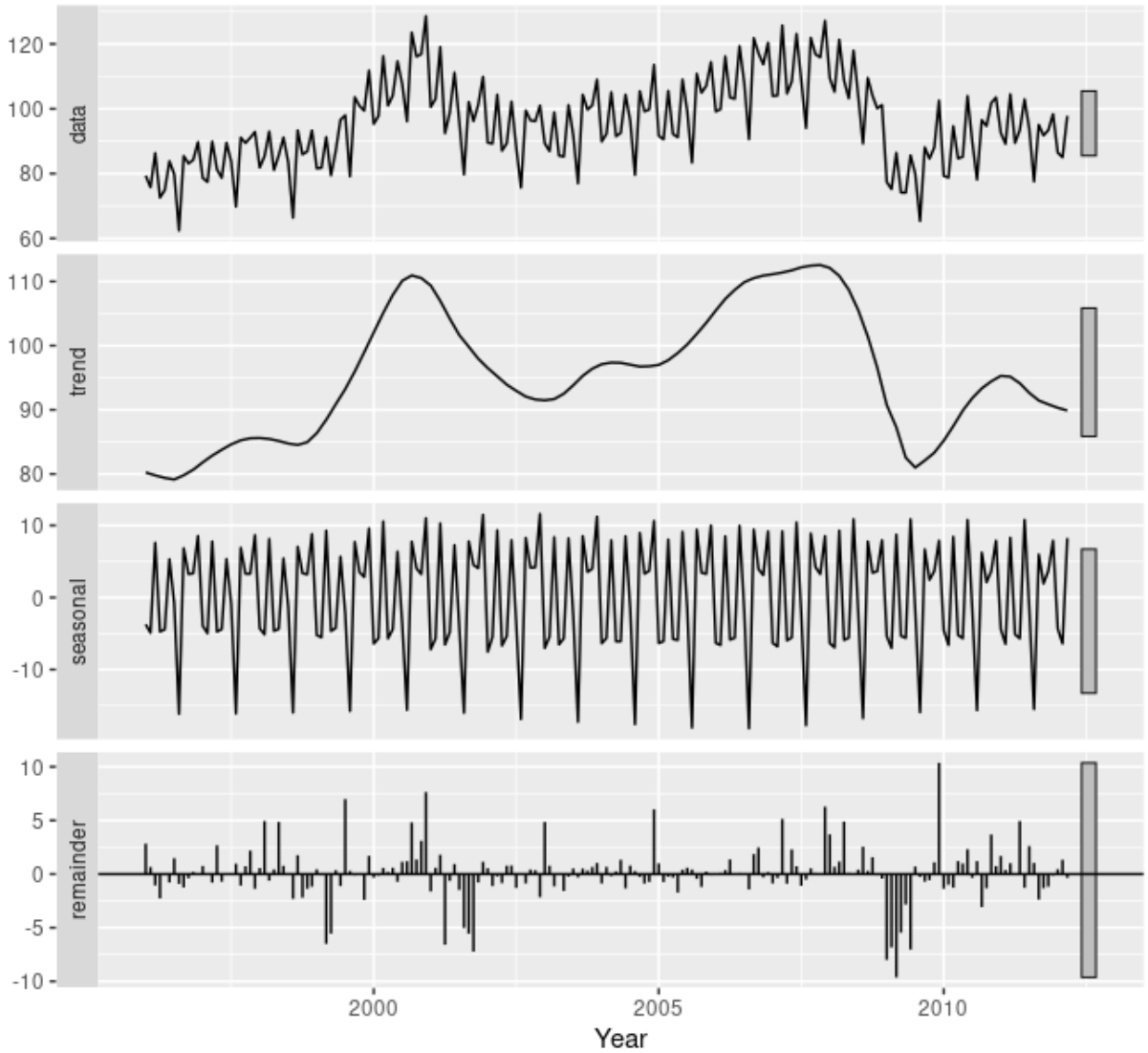


Figure: Source: Rob J. Hyndman - Forecasting: Principles and Practice (2nd ed)

If the time series is modelled additively

$$Y_t = T_t + S_t + \epsilon_t$$

where

- T_t : trend component (sometimes m_t);
- S_t : seasonal component;
- ϵ_t : zero-mean error (residual R_t assumed stochastic noise or error).

The seasonally adjusted series is given by

$$\widetilde{Y}_t := Y_t - S_t = T_t + \epsilon_t$$

Seasonally Adjusted

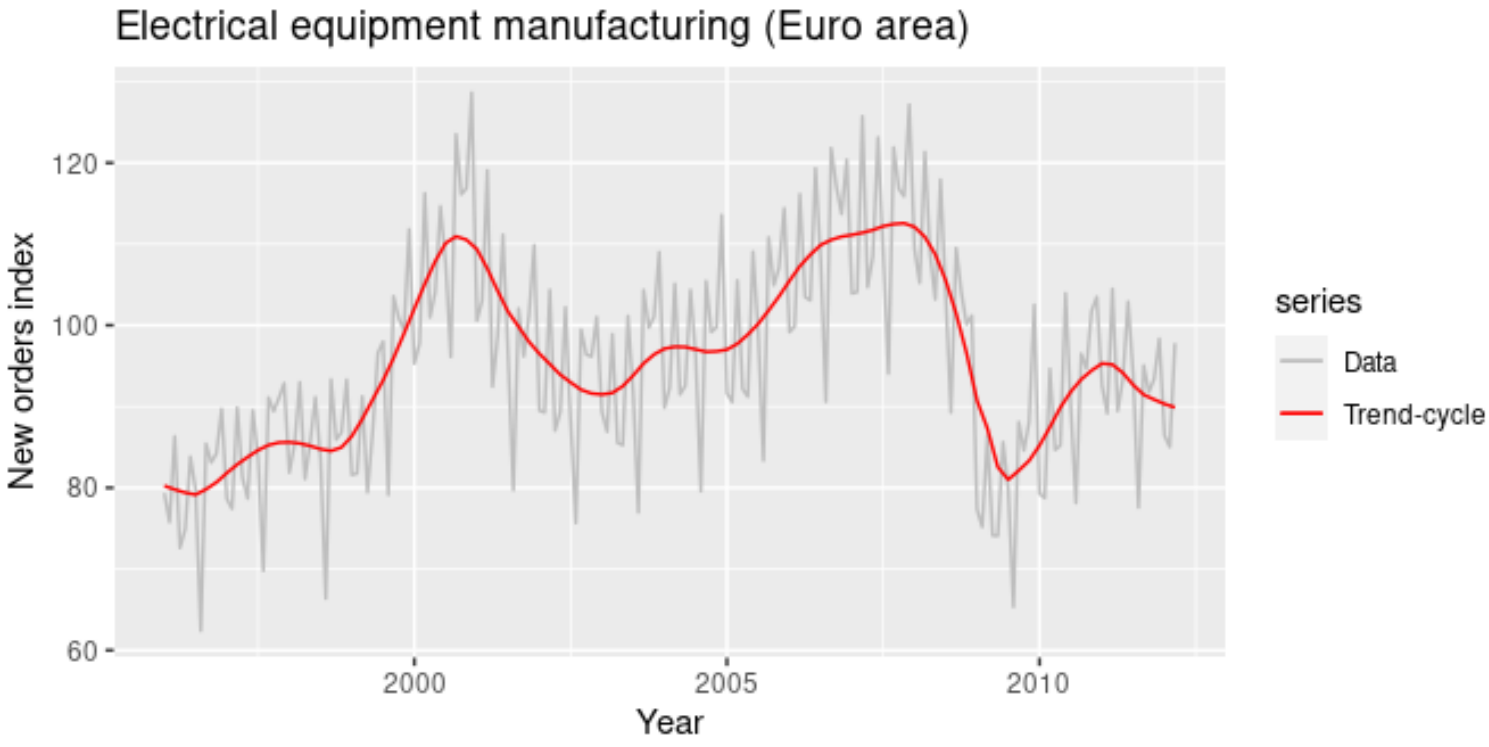


Figure: Source: Rob J. Hyndman - Forecasting: Principles and Practice (2nd ed)

Residual Additive Error

The residual term (R_t) will be treated as random stochastic noise often we will denote by ϵ_t and work specifically with special properties of this noise - which dictate the distribution of the data:

A **time series model** specifies the joint distribution of the sequence $\{X_t\}$ of random variables; e.g.,

$$P(X_1 \leq x_1, \ldots, X_t \leq x_t) \text{ for all } t \text{ and } x_1, \ldots, x_t.$$

where $\{X_1, X_2, \ldots\}$ is a stochastic process, and $\{x_1, x_2, \ldots\}$ is a single realization. Through this course, we will mostly restrict our attention to **the first- and second-order properties** only: $E(X_t), \text{Cov}(X_{t_1}, X_{t_2})$

Example of zero-mean error models (residual):

- Most basic TS model is comprised of a model with no trend, no seasonality and a zero-mean error model that produces the time series X_t as simply IID random variables. Such as sequence of zero-mean errors that produces this are referred to as White Noise (more discussions later). This produces a joint distribution for TS $\{X_t\}$ with simple structure

$$\mathbb{P}r\left(X_1 \leq x_1, \ldots, X_t \leq x_t\right) = \prod_t \mathbb{P}r\left(X_t \leq x_t\right) = \prod_t F_X(x_t)$$

  where $F(\cdot)$ is the cdf of each X_t. Note that $X_t | X_{t-1}, \ldots, X_1 = X_t$ in this case.

- A binary example of iid noise. Consider a binary process $\{X_t\}$ as a sequence of iid. r.v.s with

$$\mathbb{P}\left(X_t = 1\right) = 0.5, \quad \mathbb{P}\left(X_t = -1\right) = 0.5$$

- A continous example of iid noise. Consider a TS process $\{X_t\}$ as a sequence of iid. r.v.s with normal marginal distributions $X_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$

Trend and Seasonal Variation-summary

Trend: a systematic, non-periodic change in the time series.

Deterministic trends: caused by non-random phenomenon,

- can be predicted with certainty,
- can be modeled by regression

Linear trend: $X_t = a + bt + \epsilon_t$

Quadratic trend: $X_t = a + bt + ct^2 + \epsilon_t$

Stochastic trends: caused by random variation often induced by the dependence between adjacent variables
$\Rightarrow$Analysis uses autocorrelations

Seasonal Variation: any recurrent pattern, e.g., within each year in which the series is observed.

Error term: variation in the time series not explained by the trend or seasonality.
$\Rightarrow$ Unlike in regression, the error terms are depended random variables.

R CODE EXAMPLE: Additive Trend Structures in TS in R.Rmd
(See the Appendix for Code Example)

Trend and Seasonal Variation-summary

REMEMBER: **The main difference between time series and other statistical samples:**

- *dependent observations*

  that become available at

  - *equally spaced time intervals* **&**
    - *are time-ordered*

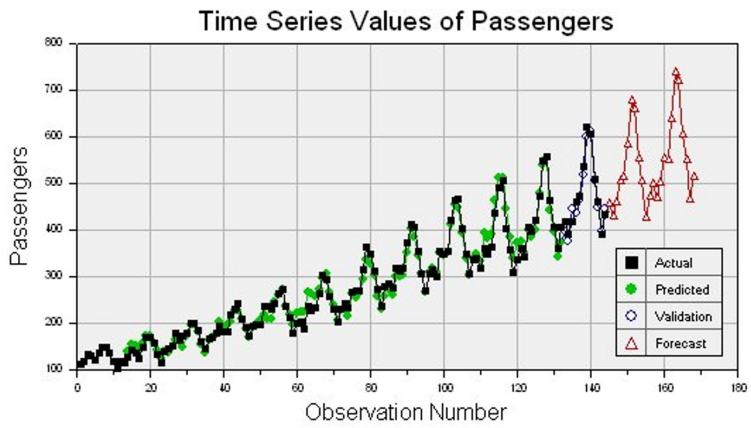


Figure: (Source: Dr. Raya Feldman, UCSB Time Series Lectures 2021)

Visual Characteristics to Identify Structures in TS

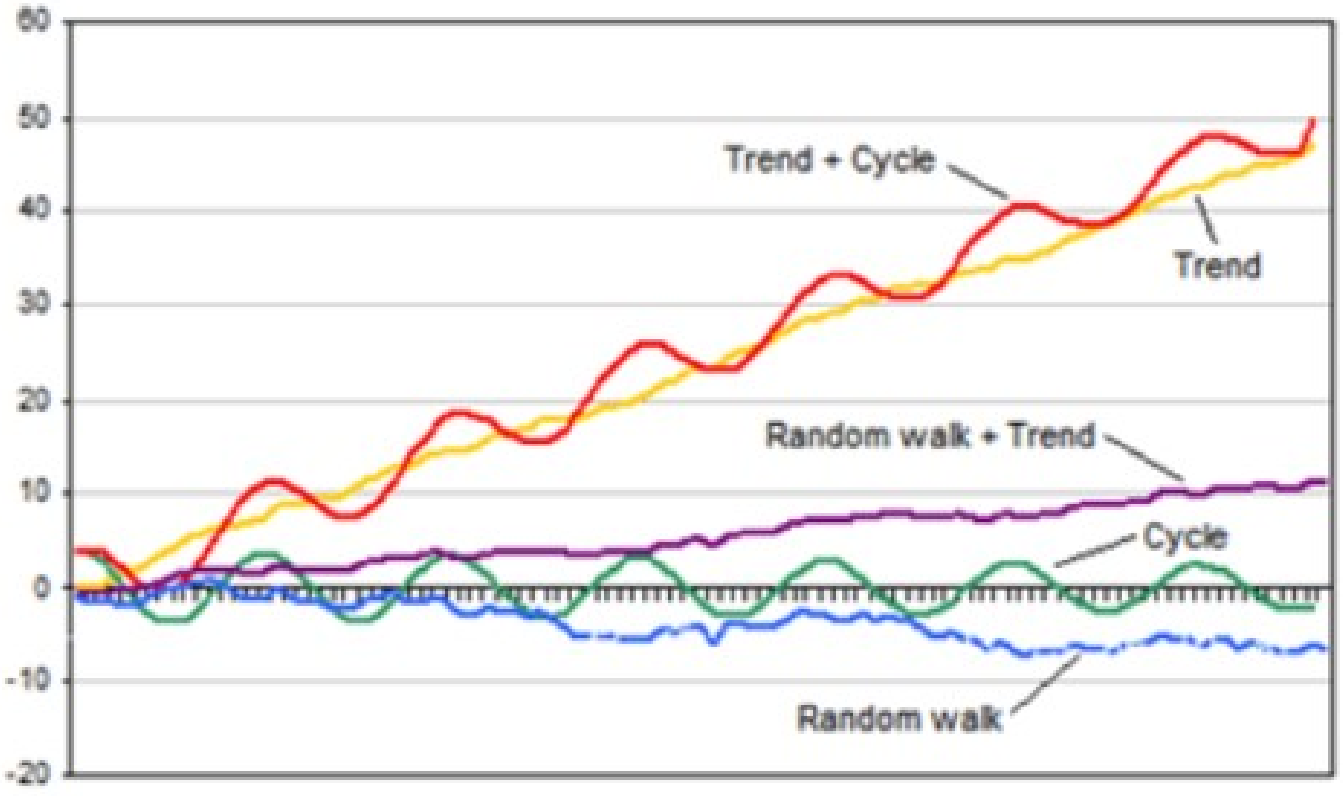


Figure: (Source: Dr. Raya Feldman, UCSB Time Series Lectures 2021)

EXAMPLE of X-11/X-12/X-13 and SEATS

- http://www.seasonal.website/

```
# data is two columns (column 1 is dates, column 2 is time
series values)
> data<-read.csv(".../data.csv", header=F)
# set frequency and start date
> data.ts<-ts(data[,2], frequency=m, start=c(YYYY,1))
# create a time series object
> data.de<-decompose(data.ts, type="additive")
> plot(data.de)
```

R CODE EXAMPLE: Decompositions in TS with R.Rmd
(See code and example html in Appendix)

Further examples provided with: SEATS, X11 and more sophisticated versions in
https://sylwiagrudkowska.github.io/JDemetra-documentation/

Statistical Population Characteristics and Descriptors of Time Series

A time series can be thought of as a stochastic process:

$$\{Y_t\} = \{Y_1, \ldots, Y_T\},$$

where Y_t are RVs and $t = 1, \ldots, T$ denotes (discrete) time. In practice, the observed data $\{y_1, \ldots, y_T\}$ is a realisation of $\{Y_t\}$.

Important points to note:

- the order of observations is determined by time
- it is usually (for the 'purposes of this immediate area of study') impossible to make multiple observations at one instant of time
- it is 'difficult' and 'often' not necessary to find joint distrib. of $\{Y_t\}$

i.e., for some fixed t_0 we cannot take lots of samples of Y_{t_0}.

- We will often switch between generic r.v. for TS $\{Y_t\}$ or $\{X_t\}$ $\rightarrow$ just two different time series.

Definition

Let $\{Y_t\}$ be a time series. The mean function of Y_t is defined by:

$$\mu_Y(t) := \mathbb{E}(Y_t)$$

The variance function of Y_t is defined by:

$$\sigma^2_Y(t) := \text{var}(Y_t)$$

The autocovariance function (acvf) of Y_t is defined by:

$$\gamma_Y(t,k) := \text{cov}(Y_t, Y_k)$$

The autocorrelation function (acf) of Y_t is defined by:

$$\rho_Y(t,k) := \text{cov}(Y_t, Y_k)$$

Remark

$$\text{var}(Y_t) = \gamma_Y(t,t)$$

Proposition (autocovariance properties)

1. $\gamma(0) \geq 0$
2. $|\gamma(k)| \leq \gamma(0)$ *for all* k
3. $\gamma(k) = \gamma(-k)$ *for all* k
4. $\gamma(\cdot)$ *is nonnegative definite, i.e., a real valued function* K *defined on the integers is nonnegative definite iff*

$$\sum_{i,j=1}^{n} a_i K(i-j) a_j \geq 0$$

for all postive integers n *and real vectores* $(a_1, \ldots, a_n) \in \mathbb{R}^n$.

Autocovariance function (B)

Proof.

- The first item is trivial since $\gamma_Y(0) = cov(Y_t, Y_t) = Var(Y_t)$
- The second item is based on the Cauchy-Schwarz inequality

$$|\gamma_Y(k)| = Cov(Y_{t+k}, Y_t) \leq \sqrt{Var\,(Y_{t+k})}\sqrt{Var\,(Y_t)} = \gamma_Y(0)$$

$\Rightarrow |\gamma_Y(k)| \leq Var(Y_t)$ since

$$\sqrt{Var\,(Y_{t+k})}\sqrt{Var\,(Y_t)} = \sqrt{\sigma_Y^2}\sqrt{\sigma_Y^2} = \sigma_Y^2 = Var(Y_t) = \gamma_Y(0)$$

- The third item is established by observing that

$$\gamma_y(k) = Cov(Y_{t+k}, Y_t) = Cov(Y_t, Y_{t+k}) = \gamma_Y(-k).$$

- The fourth statement is proven as follows

$$0 \leq Var\left(\boldsymbol{a}^T \boldsymbol{Y}_t\right) = \boldsymbol{a}^T \boldsymbol{\Gamma}_n \boldsymbol{a} = \sum_{i,j=1}^{n} a_i \gamma_Y(i-j) a_j$$

Proof.
where one can define the matrices:

$$\boldsymbol{\Gamma}_n = Var\left(\boldsymbol{Y}_n\right) = \begin{pmatrix} Cov(Y_n, Y_n) & Cov(Y_n, Y_{n-1}) & \cdots & Cov(Y_n, Y_2) & Cov(Y_n, Y_1) \\ Cov(Y_{n-1}, Y_n) & Cov(Y_{n-1}, Y_{n-1}) & \cdots & Cov(Y_{n-1}, Y_2) & Cov(Y_{n-1}, Y_1) \\ & & \vdots & & \\ Cov(Y_2, Y_n) & Cov(Y_2, Y_{n-1}) & \cdots & Cov(Y_2, Y_2) & Cov(Y_2, Y_1) \\ Cov(Y_1, Y_n) & Cov(Y_1, Y_{n-1}) & \cdots & Cov(Y_1, Y_2) & Cov(Y_1, Y_1) \end{pmatrix}$$

$$= \begin{pmatrix} \gamma_Y(0) & \gamma_Y(1) & \cdots & \gamma_Y(n-2) & \gamma_Y(n-1) \\ \gamma_Y(1) & \gamma_Y(0) & \cdots & \gamma_Y(n-3) & \gamma_Y(n-2) \\ & & \vdots & & \\ \gamma_Y(n-2) & \gamma_Y(n-3) & \cdots & \gamma_Y(0) & \gamma_Y(0) \\ \gamma_Y(n-1) & \gamma_Y(n-2) & \cdots & \gamma_Y(1) & \gamma_Y(0) \end{pmatrix}$$

□

Autocovariance function (Complex) (B)

Note, one can also define a complex valued process $\{X_t\}$ for $X_t = X_{t,1} + \imath X_{t,2} \in \mathbb{C}$ where $X_{t,1}$ is the real component and $\imath X_{t,2}$ is the imaginary component. Here, we will use notation $\bar{\cdot}$ to denote the complex conjugate. eg. $\overline{X}_t = X_{t,1} - \imath X_{t,2}$.
We can restate the autocovariance in the complex case as follows

Proposition (autocovariance properties)

1. $\gamma_X(0) \geq 0$
2. $|\gamma_X(k)| \leq \gamma_X(0)$ *for all* k
3. $\gamma_X(k) = \overline{\gamma_X(-k)}$ *for all* k
4. $\gamma_X(\cdot)$ *is Hermitian and nonnegative definite, i.e., a (possibly complex) valued function* K *defined on the integers is Hermitian and nonnegative definite iff* $K(n) = \overline{K(-n)}$ *and*

$$\sum_{i,j=1}^{n} a_i K(i-j)\bar{a}_j \geq 0$$

for all postive integers n *and complex vectors* $(a_1, \ldots, a_n) \in \mathbb{C}^n$.

This will be useful later in spectral representations of time series.

Autocorrelation function (B)

The autocorrelation function $\rho(\cdot)$ has all the properties of an autocovariance function and satisfies the additional condition $\rho(0) = 1$.

Definition

The autocorrelation function $\rho(k)$ is defined by:

$$\rho(k) := \frac{\gamma(k)}{\gamma(0)}$$

Proposition (autocorrelation properties)

1. $\rho(0) = 1$
2. $\rho(-k) = \rho(k)$
3. $|\rho(k)| \leq 1$
4. *Non-uniqueness:* $\{Y_t\} \neq \{X_t\} \not\Rightarrow \rho_Y \neq \rho_X$

Autocorrelation function (B)

Proof 1 (by definition) and 2 (by symmetry of the covariance operator) are trivial, we'll see an example of 4 later on in the course. For property 3: for $a, b \in \mathbb{R}$:

$$\begin{aligned} \text{var}(aY_t + bY_{t+k}) &\geq 0 \\ a^2 \,\text{var}\, Y_t + b^2 \,\text{var}\, Y_{t+k} + 2ab\,\text{cov}(Y_t, Y_{t+k}) &\geq 0 \\ (a^2 + b^2)\sigma_Y^2 + 2ab\gamma(k) &\geq 0 \end{aligned}$$

Choosing $a = b = 1$ gives:

$$\begin{aligned} 2\sigma_Y^2 + 2\gamma(k) &\geq 0 \\ 1 + \gamma(k)/\sigma_Y^2 &\geq 0 \\ \gamma(k)/\sigma_Y^2 &\geq -1 \end{aligned}$$

Choosing $a = 1, b = -1$ gives:

$$\begin{aligned} 2\sigma_Y^2 - 2\gamma(k) &\geq 0 \\ 1 - \gamma(k)/\sigma_Y^2 &\geq 0 \\ \gamma(k)/\sigma_Y^2 &\leq 1 \end{aligned}$$

Description of Time Series (B)

Probabilistic/Statistical Description of time series $Y_1, Y_2, Y_3, \ldots, Y_n$,

- Finite Dimensional Distributions: (f.d.d.) is the joint d.f. for random vector $(Y_{t_1}, \ldots, Y_{t_n})$:

$$F_{t_1 \ldots t_n}(y_1, \ldots, y_n) = \mathbb{P}r\left(Y_{t_1} \leq y_1, \ldots, Y_{t_n} \leq y_n\right), \forall t_1 < \cdots < t_n.$$

- First- and Second-order moments:
  - mean: $\mu_Y(t) = \mathbb{E}[Y_t]$
  - variance: $\sigma_Y^2(t) = \mathbb{E}\left[\left(Y_t - \mu_Y(t)\right)^2\right] = \mathbb{E}\left[Y_t^2\right] - \mu_Y(t)^2$
- Autocovariance function (ACVF):

$$\gamma_Y(t,s) = \text{cov}\left(Y_t, Y_s\right) = \mathbb{E}\left[\left(Y_t - \mu_Y(t)\right)\left(Y_s - \mu_Y(s)\right)\right] = \mathbb{E}\left[Y_t Y_s\right] - \mu_Y(t)\mu_Y(s)$$

  Note: this is an infinite matrix... in s,t
- Autocorrelation function (ACF):

$$\rho_Y(t,s) = Cor\left(Y_s, Y_t\right) = \frac{Cov\left(Y_t, Y_s\right)}{\sqrt{\text{var}\left(Y_t\right)\text{var}\left(Y_s\right)}} = \frac{\gamma_Y(t,s)}{\sigma_Y(t)\sigma_Y(s)}$$

  Properties of the process which are determined by the first- and second- order moments are called second-order properties.

R CODE EXAMPLE: White Noise Correlogram R.Rmd
(See code and example html in Appendix)

A time series is said to be completely random (or i.i.d.) if it consists of a series of independent observations having the same distribution.
Lets plot a completely random series and its correlogram.

```
> set.seed(1)
> x<-rnorm(400)
> par(mfrow=c(2,1), mar=c(3,4,3,4))
> plot(x, type="l", xlab="", ylab="")
> title(xlab="Time", ylab="Series", line=2, cex.lab=1.2)

> acf(x, ylab="", main="")
> title(xlab="Lag", ylab="ACF", line=2)
```

If a time series contains a trend, then the values of $r(k)$ (estimae of $\rho(k)$ will not come down to zero except for very large values of the lag. This is because an observation on one side of the overall mean tends to be followed by a large number of further observations on the same side of the mean because of the trend.

```
> set.seed(1)
> ts.sim3<-cumsum(rnorm(400))
> par(mfrow=c(2,1), mar=c(3,4,3,4))
> plot(ts.sim3, type="l", xlab="", ylab="")
> title(xlab="Time", ylab="Series", line=2, cex.lab=1.2)

> acf(ts.sim3, ylab="",main="")
> title(xlab="Lag", ylab="ACF", line=2)
```

Sample Estimation of Characteristics and Descriptors of Time Series

Sample Estimation of μ (B)

In practice, we do not know μ, γ, ρ, but rather we have data and must use estimates.

Definition

Let $\{y_t\} = \{y_1, \ldots, y_T\}$ be the observed values of a time series $\{Y_t\}$.

The sample mean of $\{y_t\}$ is defined as:

$$\hat{\mu} = \overline{y} := \frac{1}{T} \sum_{t=1}^{T} y_t$$

Question: Does the time series average converge to the same limit as the ensemble average?
Answer: Yes if Y_t is **stationary and ergodic**..

We will learn a lot about stationarity shortly....

Sample Estimation of μ (B)

Recall that Kolmogorov's Law of Large Numbers (LLN) states that if $Y_i \overset{iid}{\sim} F(\mu, \sigma^2)$ for $i = 1, \ldots, N$ then we have the following limit for the ensemble average

$$\overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i \to \mu$$

In time series, we have time series average and not ensemble average

- With one realisation observed of the TS then we can only average along this single realised path of the process to learn about population moment at a fixed time.

Sample Estimation of μ (B)

To explain the differences between ensemble average and time series average, consider the following experiment.

Example

Suppose we want to track the movements of some particles and draw inference about their expected position (suppose that these particles move on the real line).

- *If we have a group of particles (group size N), then we could track down the position of each particle and plot a distribution of their positions.*
- *The mean of this sample is called ensemble average.*
- *If all these particles are i.i.d., LLN tells that this average converges to its expectation as* $N \to \infty$.
- *However, with time series observations, we only have one history. That means, in this experiment, we only have one particle and its trajectory (path realisation)*
- *Then instead of collecting N particles, we can only track this single particle and record its position, say* y_t, *for* $t = 1, 2, ..., T$.
- *The mean we computed by averaging over time,* $1/T \sum_{t=1}^{T} y_t$ *is called time series average.*

Definition

In time series a stochasatic process $\{Y_t\}$ *is said to be ergodic if its statistical properties can be deduced from a single, sufficiently long* $T \to \infty$ *random sample (trajectory) of the process*

One can discuss the ergodicity of various statistics of a stochastic process.

- In general, ergodicity means that ensemble averages are equal to the time average.

Note: examples here.

Sample Estimation of μ (B)

Comment on Ergodicity:

- It is not immediately clear that one can obtain consistent estimates of the properties of a stationary process from a single realization.

If Y_t is stationary and ergodic with $\mathbb{E}[Y_t] = \mu$, then the time series average has the same limit as the ensemble average

$$\hat{\mu} = \overline{Y} := \frac{1}{T}\sum_{t=1}^{T} Y_t \to \mu, \quad T \to \infty.$$

Theorems, called ergodic theorems, have been proven, that show that for most stationary processes, encountered in practice, the sample moments of an observed time series do indeed converge to the corresponding population moments.

Remark on Ergodicity (B)

A sufficient condition for this to happen is that $\rho(k) \to 0$ as $k \to \infty$ and the TS process is then called "ergodic in the mean"

Under this requirement that the ACF decays with lag we can be sure that:

- An average over time for a single TS realisation, like $\overline{y} = \sum_{t=1}^{T} y_t / T$ can be used to estimate the ensemble properties of the underlying process at a particular time t, i.e. the properties of the series at time t are estimated using data collected at other time points.
- i.e. time averages like $\sum_{t=1}^{T} y_t / T$ converge to population quantities like $\mathbb{E}[Y_t]$ as $T \to \infty$.

We will see that the faster the ACF decays - the faster the rate of convergence of sample mean estimator for a TS.

Sample Estimation of μ (B)

A word of caution on using the sample mean of a time series:

- In time-series analysis there are some special problems that relate to using a time-series sample mean as an estimate of some underlying population mean.
- Although we will use the sample mean in computing the sample ACFV and ACF, we should note, that the sample mean is a potentially misleading summary statistic unless all systematic components of a TS have been removed.
- Thus the sample mean should only be considered as a summary statistic for data thought to have come from a stationary process. (To be defined & discussed next)
- Even when this is so, it is important to realize that the statistical properties of the sample mean are quite different from those that usually apply.

Sample Estimation of μ (B)

Suppose we have TS data $\{y_i\}$ for $i = 1, \ldots, T$ from a stationary process having mean μ and variance σ^2 and theoretical ACF $\rho(k)$.

Let $\overline{Y} = \sum_{i=1}^{T} Y_i$ denote the sample mean value expressed as a r.v.

The usual result of independent observations is that $Var(\overline{Y}) = \frac{1}{T}\sigma^2$.

However, for autocorrelated observations it can be shown that

$$Var(\overline{Y}) = \frac{1}{T}\sigma^2 \left[1 + 2\sum_{j=1}^{T-1}\left(1 - \frac{|j|}{T}\right)\rho(j)\right]$$

and this quantity can differ considerably from $\frac{1}{T}\sigma^2$ when autocorrelation is substantial.

- We will need to see what the ACF looks like in terms of model parameters for different TS models to explore this further
- In general for some TS models, depending on their parameter values (coefficients) there can be substantially less information in estimation of the mean than we would otherwise expect compared to the independent case (for given sample size Variance can be large)

Sample Estimation of μ (B)

So under an ergodicity assumption one can find that $\mathbb{E}\left[\overline{Y}\right] = \mu_Y$ and we can calculate its mean square error (in this case variance) as follows:

$$\begin{aligned}
Var\left(\overline{Y}\right) &= \mathbb{E}\left[\left(\overline{Y}_T - \mu_Y\right)^2\right] = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T} Cov\left(Y_i, Y_j\right) \\
&= \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\gamma_Y(i-j) \\
&= \frac{1}{T^2}\sum_{i-j=1-T}^{T-1}\left(T - |i-j|\right)\gamma_Y(i-j) \\
&= \frac{1}{T}\sum_{k=1-T}^{T-1}\left(1 - \frac{|k|}{T}\right)\gamma_Y(k) \\
&= \frac{\gamma_Y(0)}{T} + \frac{2}{T}\sum_{k=1}^{T-1}\left(1 - \frac{|k|}{T}\right)\gamma_Y(k)
\end{aligned}$$

where the first term $\frac{\gamma_Y(0)}{T}$ is the $Var(\overline{Y}_T)$ for the i.i.d. case for data observed from $\{Y_t\}$

Sample Estimation of μ (B)

Depending on the nature of the correlation structure, the standard error of $\overline{Y}_T$ may be smaller or larger than the i.i.d. case

Example

Consider a time series $Y_t = \epsilon_t - \theta_1 \epsilon_{t-1}$ *where we will set* $\epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$. *We can calculate* $\gamma_Y(k)$ *as follows (we will derive this in detail later when talking about MA(1) models):*

$$\begin{aligned}
\gamma_Y(k) = \text{cov}(Y_t, Y_{t-k}) &= \text{cov}(\epsilon_t - \theta_1\epsilon_{t-1}, \epsilon_{t-k} - \theta_1\epsilon_{t-k-1}) \\
&= \text{cov}(\epsilon_t, \epsilon_{t-k}) - \theta_1 \text{cov}(\epsilon_t, \epsilon_{t-k-1}) \\
&\quad -\theta_1 \text{cov}(\epsilon_{t-1}, \epsilon_{t-k}) + \theta_1^2 \text{cov}(\epsilon_{t-1}, \epsilon_{t-k-1}) \\
&= \begin{cases} \sigma^2(1+\theta_1^2), & k = 0 \\ -\theta_1\sigma^2, & |k| = 1 \\ 0, & \textit{oth.} \ (|k| \geq 2) \end{cases}
\end{aligned}$$

Sample Estimation of μ (B)

Note for instance:

$$\begin{aligned} k=0: \quad & \text{cov}(\epsilon_t,\epsilon_t) - \theta_1 \text{cov}(\epsilon_t,\epsilon_{t-1}) - \theta_1 \text{cov}(\epsilon_{t-1},\epsilon_t) + \theta_1^2 \text{cov}(\epsilon_{t-1},\epsilon_{t-1}) \\ = \quad & \text{cov}(\epsilon_t,\epsilon_t) + \theta_1^2 \text{cov}(\epsilon_{t-1},\epsilon_{t-1}) \\ = \quad & \sigma^2 + \theta_1^2\sigma^2 \end{aligned}$$

Confirm for yourself the cases $k=1$, $k>1$

Example

Then we can calculate the variance of the data $\{Y_t\}$ *as follows:*

$$Var\left(\overline{Y}\right) = \frac{\gamma_Y(0)}{T} + \frac{2}{T}\sum_{k=1}^{T-1}\left(1 - \frac{|k|}{T}\right)\gamma_Y(k)$$

This gives the formula simplified as follows:

$$Var\left(\overline{Y}\right) = \frac{\sigma^2(1+\theta_1^2)}{T} - \frac{2}{T}\left(1 - \frac{1}{T}\right)\theta_1\sigma^2$$

Sample Estimation of μ (B)

Example

$$\begin{aligned} Var\left(\overline{Y}\right) &= \frac{\gamma_Y(0)}{T} + \frac{2}{T}\sum_{k=1}^{T-1}\left(1-\frac{|k|}{T}\right)\gamma_Y(k) \\ &= \frac{1}{T}\sigma^2(1+\theta_1^2) - \frac{2}{T}\left(1-\frac{1}{T}\right)\theta_1\sigma^2 \end{aligned}$$

We can now clearly see that in this example - when $\theta_1 = 0$ *we get the standard i.i.d. data case with*

$$Var\left(\overline{Y}\right) = \frac{1}{T}\sigma^2$$

Then for instance if $\theta_1 > 0$ *we will have* $Var\left(\overline{Y}\right) > \frac{1}{T}\sigma^2$ *if* θ_1 *satisfies:*

$$\frac{1}{T}\sigma^2(1+\theta_1^2) - \frac{2}{T}\left(1-\frac{1}{T}\right)\theta_1\sigma^2 > \frac{1}{T}\sigma^2$$

$$\Rightarrow \theta_1 > 2\left(1-\frac{1}{T}\right)$$

and if $0 < \theta_1 < 2\left(1-\frac{1}{T}\right)$ *we will have* $Var\left(\overline{Y}\right) < \frac{1}{T}\sigma^2$.

Sample Estimation of μ (B)

We can also observe that if $\gamma_Y(k) \to 0$ as $k \to 0$ then we have

$$|Var\left(\overline{Y}_T\right)| \leq \frac{\gamma_Y(0)}{T} + 2\frac{\sum_{k=1}^{T-1}|\gamma_Y(k)|}{T} \to 0, \quad \text{as } T \to \infty.$$

Thus $\overline{Y}_T$ converges in mean square to μ.

Furthermore, if $\sum_{k=-\infty}^{\infty}|\gamma_Y(k)| < \infty$, then as $T \to \infty$

$$\begin{aligned} T\ Var\left(\overline{Y}_T\right) &= \gamma_Y(0) + 2\sum_{k=1}^{T-1}\left(1 - \frac{|k|}{T}\right)\gamma_Y(k) \\ &\to \gamma_Y(0) + 2\sum_{i=1}^{\infty}\gamma_Y(i) = \sum_{i=-\infty}^{\infty}\gamma_Y(i) = \gamma_Y(0)\sum_{i=-\infty}^{\infty}\rho_Y(i). \end{aligned}$$

Which allows us to make an interpretation that instead of $Var\left(\overline{Y}_T\right) \approx \frac{\gamma_Y(0)}{T}$, we have $Var\left(\overline{Y}_T\right) \approx \frac{\gamma_Y(0)}{T/\tau}$ with $\tau = \sum_{k=-\infty}^{\infty}\rho_Y(k)$.

The effect of the correlation is a reduction of sample size from T to T/τ!

Sample Estimation of μ (B)

To perform any form of inference about a population mean μ_Y such as null hypothesis that $\mathcal{H}_0 : \mu_Y = 0$ vs alternative $\mathcal{H}_Y \mu_Y \neq 0$, where the test statistic is based on the sample mean $\overline{Y}_T$, then it is necessary to know the asymptotic distribution of $\overline{Y}_T$.

If $\{Y_t\}$ is Gaussian stationary time series, then, for any sample size T,

$$\sqrt{T}\left(\overline{Y}_T - \mu_Y\right) \sim N\left(0, \sum_{k=1-T}^{T-1}\left(1 - \frac{|k|}{T}\right)\gamma_Y(k)\right)$$

Definition

Let $\{y_t\} = \{y_1, \ldots, y_T\}$ *be the observed values of a time series* $\{Y_t\}$.

The sample autocovariance function is:

$$c(k) = \hat{\gamma}(k) := \frac{1}{T} \sum_{t=k+1}^{T} (y_{t-k} - \hat{\mu})(y_t - \hat{\mu})$$

You may also see sample autocovariance function written as follows for observed data $\{y_t\} = \{y_1, \ldots, y_T\}$

$$c(k) = \hat{\gamma}(k) := \frac{1}{T} \sum_{t=1}^{T-|k|} (y_{t+|k|} - \hat{\mu})(y_t - \hat{\mu}) \quad \text{for } -T < k < T.$$

Proposition

For any sequence of a time series $y_1, \ldots, y_T$ the sample ACVF $\widehat{\gamma}_Y$ satisfies:

- *symmetry around the origin such that $\widehat{\gamma}_Y(k) = \widehat{\gamma}_Y(-k)$*
- *$\widehat{\gamma}_Y$ non-negative definite and hence*
- *$\widehat{\gamma}_Y(0) \geq 0$ and $|\widehat{\gamma}_Y(k)| \leq \widehat{\gamma}_Y(0)$.*

The first condition is trivial. Lets prove the second and third items which are equivalent to show that for each $s \geq 1$ the s-dimensional sample covariance $\widehat{\Gamma}_s$ given by matrix

$$\widehat{\Gamma}_s = \begin{pmatrix} \widehat{\gamma}_Y(0) & \widehat{\gamma}_Y(1) & \ldots & \widehat{\gamma}_Y(s-1) \\ \widehat{\gamma}_Y(1) & \widehat{\gamma}_Y(0) & \ldots & \widehat{\gamma}_Y(s-2) \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{\gamma}_Y(s-1) & \widehat{\gamma}_Y(s-2) & \ldots & \widehat{\gamma}_Y(0) \end{pmatrix}$$

is non-negative definite.

To see this, one has that, for $s \geq T$,

$$\widehat{\Gamma}_s = \frac{1}{T} M M^T$$

where matrix M is a $s \times 2s$ matrix with elements $\widetilde{Y}_i = Y_i - \overline{Y}_T$ for $i = 1, \ldots, T$ and $\widetilde{Y}_i = 0$ for $i = T+1, \ldots, s$, with structure

$$M = \begin{pmatrix} 0 & \ldots & 0 & 0 & \widetilde{Y}_1 & \widetilde{Y}_2 & \cdots & \widetilde{Y}_s \\ 0 & \ldots & 0 & \widetilde{Y}_1 & \widetilde{Y}_2 & \cdots & \widetilde{Y}_s & 0 \\ \vdots & \vdots & \vdots & \vdots & & & & \\ 0 & \widetilde{Y}_1 & \widetilde{Y}_2 & \cdots & \widetilde{Y}_s & 0 & \ldots & 0 \end{pmatrix}.$$

Note that, if Γ_m is non-negative definite, then all Γ_s's are non-negative definite for all $s < m$. □

Note, one can easily see that sample autocovariance estimated for samples $\{y_1, \ldots, y_T\}$

$$c(k) = \hat{\gamma}(k) := \frac{1}{T} \sum_{t=1}^{T-|k|} \left(y_{t+|k|} - \hat{\mu}\right)\left(y_t - \hat{\mu}\right) \quad \text{for } -T < k < T.$$

is biased, take for instance the value at $k = 0$ given by

$$c(0) = \widehat{\gamma}_Y(0) = \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \overline{Y}_T\right)^2$$

where even in an iid case, this is a biased estimator (sample variance is unbiased which has $(T-1)^{-1}$ instead of T^{-1}.

Note on the degrees of freedom and scaling of estimator:

- Alternative estimator of ACVF also occasionally used (more like linear regression variance estimator)

$$c^{(2)}(k) = \widehat{\gamma}(k) = \frac{1}{T-k} \sum_{t=1}^{T-k} \left(Y_t - \overline{Y}\right)\left(Y_{t+k} - \overline{Y}\right)$$

Although this definition may have less bias, the (1/T) formulation has some desirable statistical properties

Remark

Note: the non-negative definite property of $c(k)$ is not always true if $1/T$ is replaced with $1/(T-k)$ as in the estimator $c^{(2)}(k)$.

sample ACVFγ (B)

It can be shown that the estimator $c(k)$ is finite sample biased

$$\mathbb{E}[c(k)] = \frac{T-|k|}{T}\gamma_Y(k) \neq \gamma_Y(k),$$

and that the bias is of order (1/T) such that it is asymptotically unbiased as follows:

$$\lim_{T\to\infty} \mathbb{E}(c(k)) = \gamma(k).$$

Furthermore, it can be shown that

$$Cov(c(k), c(m)) \simeq \sum_{r=-\infty}^{\infty} \frac{1}{T}[\gamma(r)\gamma(r+m-k) + \gamma(r+m)\gamma(r-k)]$$

such that when $m = k$ this gives us the variance of $c(k)$ estimator. This result highlights the fact that successive values of $c(k)$ may be (highly) correlated and this increases the difficulty in accurate interpretation of the correlogram (in small samples).

- In large TS sample lengths T, these results are not substantially important in practice for most practitioners, which is why the first estimator is the typical default.

sample ACVFγ (B)

Regarding estimator $c^{(2)}(k)$ which has scale factor $1/(T-k)$ rather than T, there is a common misconception that this is unbiased...

Not always true that $c^{(2)}(k)$ is unbiased in general, furthermore, it may not always have a lower finite sample bias than $c(k)$

- $c^{(2)}(k)$ will typically be biased when a sample mean estimate is used rather than population mean which is in most practical settings.
- The estimator $c(k)$ leads to a sample ACVF having a useful property called positive semi-definiteness, which means that its finite Fourier transform is non-negative, among other consequences. (to be discussed later)
- This property is useful in estimating the spectrum and so we generally prefer to use $c(k)$, rather than $c^{(2)}(k)$.
- Note that when $k = 0$, we get the same estimate of variance using both formulae, but that this estimate, involving $\overline{x}$, will generally still be biased (Percival, 1993).

Sample Estimation of ACF $\widehat{\rho}$ and Correlogram Plot

Correlogram

Covariance and correlation: measure extent of linear relationship between two variables (y and X).

Autocovariance and autocorrelation: measure linear relationship between lagged values of a time series y.
We measure the relationship between:

- y_t and y_{t1}
- y_t and y_{t2}
- y_t and y_{t3}
- etc.

Correlogram

In TS data analysis one of the core plots used for empirical data analysis is known as the correlogram

The correlogram can help provide answers to the following questions:

- Is the data TS random or deterministically generated?
- Are observations in the TS related to adjacent observations in the TS and at what lags?
- Is the observed TS pure white noise? (to be defined formally later)
- Is the observed TS periodic in trend structure? (to be defined formally later)
- Is the observed TS autoregressive? (to be defined formally later)
- What might be an appropriate model for the observed TS?

- When data have a trend, the autocorrelations for small lags tend to be large and positive.
- When data are seasonal, the autocorrelations will be larger at the seasonal lags (i.e., at multiples of the seasonal frequency)
- When data are trended and seasonal, you see a combination of these effects.

Definition

The sample autocorrelation function is:

$$r(k) = \hat{\rho}(k) := \frac{c(k)}{c(0)} = \frac{\sum_{t=k+1}^{T} (y_{t-k} - \hat{\mu})(y_t - \hat{\mu})}{\sum_{t=1}^{T} (y_t - \hat{\mu})^2}$$

The properties of $r(k)$ are rather more difficult to find than those of $c(k)$ because it is the ratio of two random variables.

It can be shown that $r(k)$ is generally biased.

A general formula for the variance of $r(k)$ is given by (Kendall, 1983, See book: "Time-Series") and depends on all the autocorrelation coefficients of the process.

Note: the sample autocorrelation coefficient of lag k is equivalent to the usual correlation coefficient from linear regression given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Some alternative expressions for S_{xy}, S_{xx} (derive as an easy exercise):

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Also,

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}) x_i \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

for the data (x_i, y_i) for $i = 1, 2, ..., n + k$ pairs of data.

Estimating the Correlogram

Consider the data (x_i, y_i) for $i = 1, 2, ..., n + k$ pairs of data given by

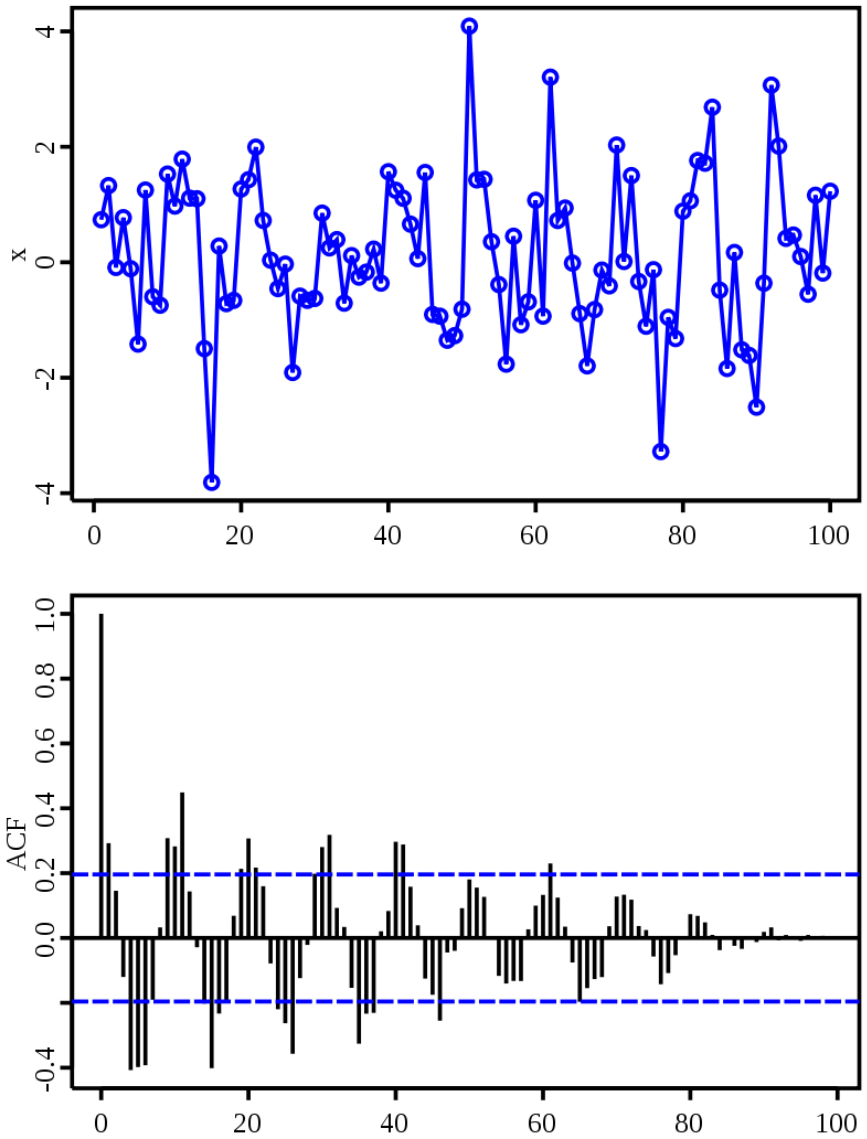
$$\begin{array}{cc} 0 & \\ \vdots & \vdots \\ 0 & \\ & \\ \vdots & \vdots \\ & \\ & 0 \\ \vdots & \vdots \\ & 0 \end{array}$$

Hence it is clear that $\hat{\rho}(k)$ has the following properties:

- $\hat{\rho}(0) = 1$
- $-1 \leq \hat{\rho}(k) \leq 1$ with $\hat{\rho}(k) = 1$ if $x(t + k)$ is perfectly linearly predictable from $x(t)$ - slope of scatter plot is positive
- with $\hat{\rho}(k) = -1$ if $x(t + k)$ is perfectly linearly predictable from $x(t)$ - slope of scatter plot is negative

Correlogram

A plot of $r(k)$ against lag k is called the correlogram.



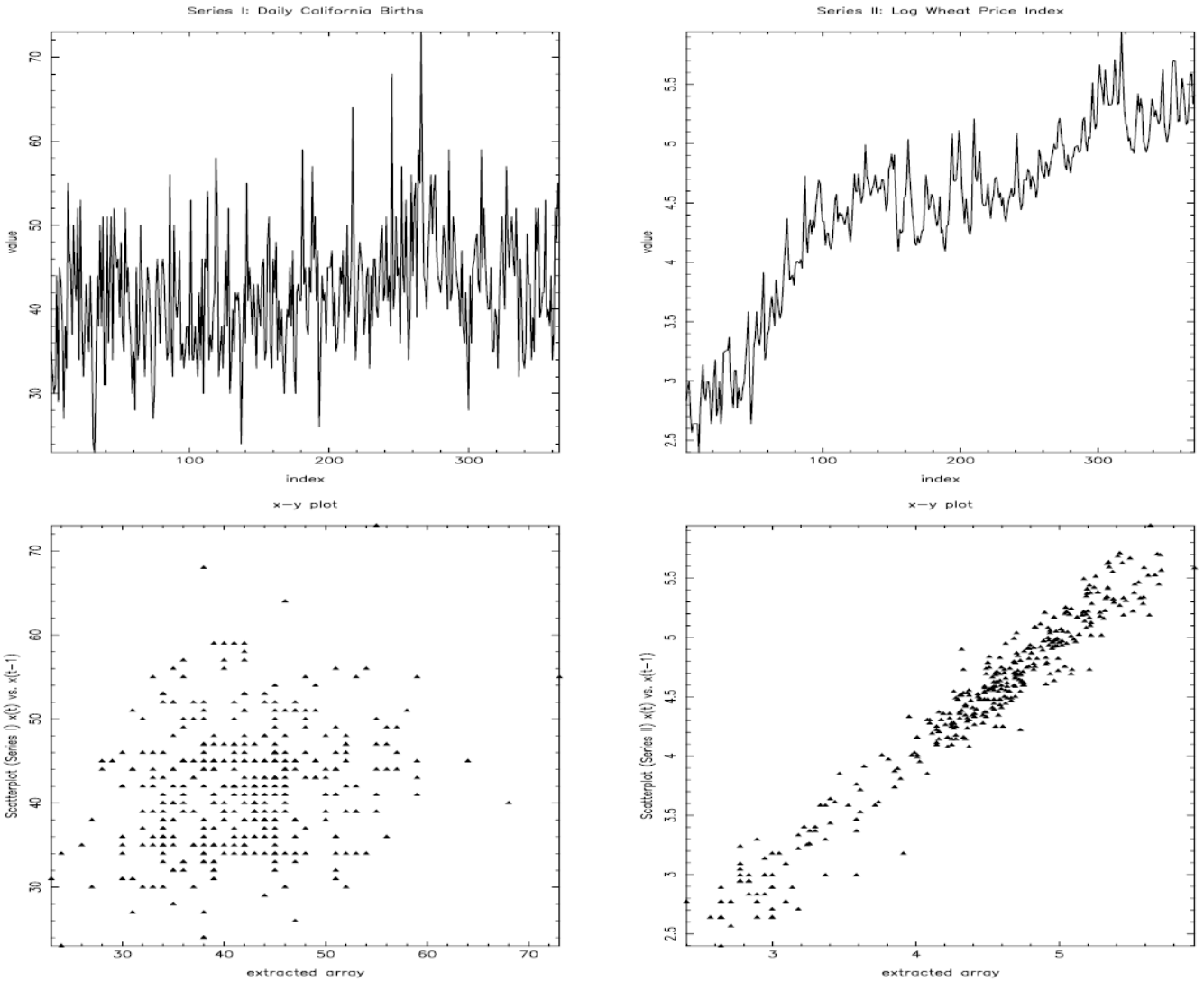
Fundamentals of the Sample Correlogram:

Time series often can exhibit correlation over time (sometimes known as serial correlation or autocorrelation). In the next figure a scatter plot of $y(t)$ vs. $y(t-k)$, i.e. y with itself lagged by some time unit (e.g. $k-1$), for two time series.

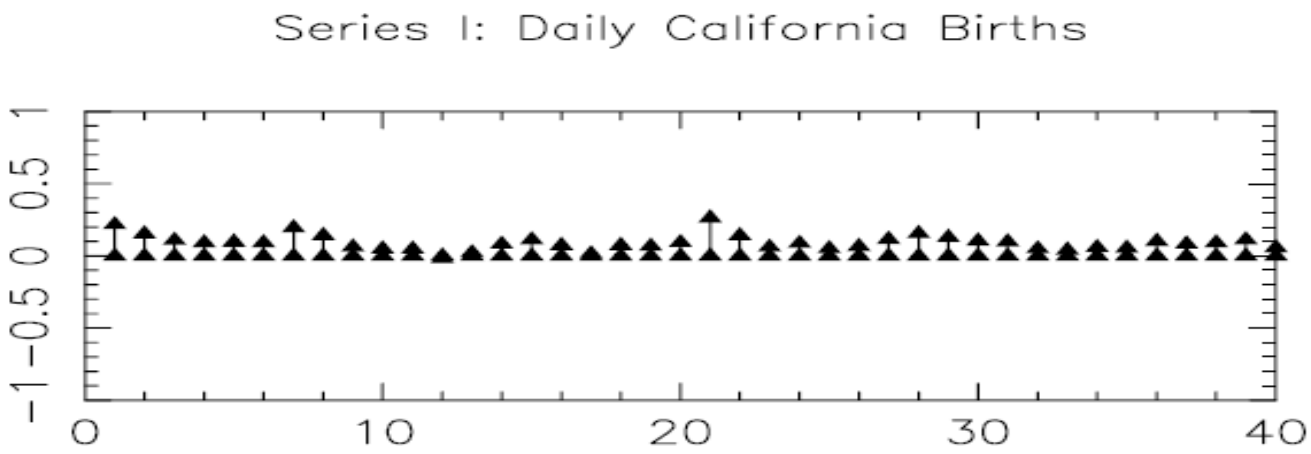
- The one on the left for "Daily Californian Births" demonstrates weak linear predictability; and
- The series on the right for "log Wheat Prices Index" demonstrates strong linear predictability.

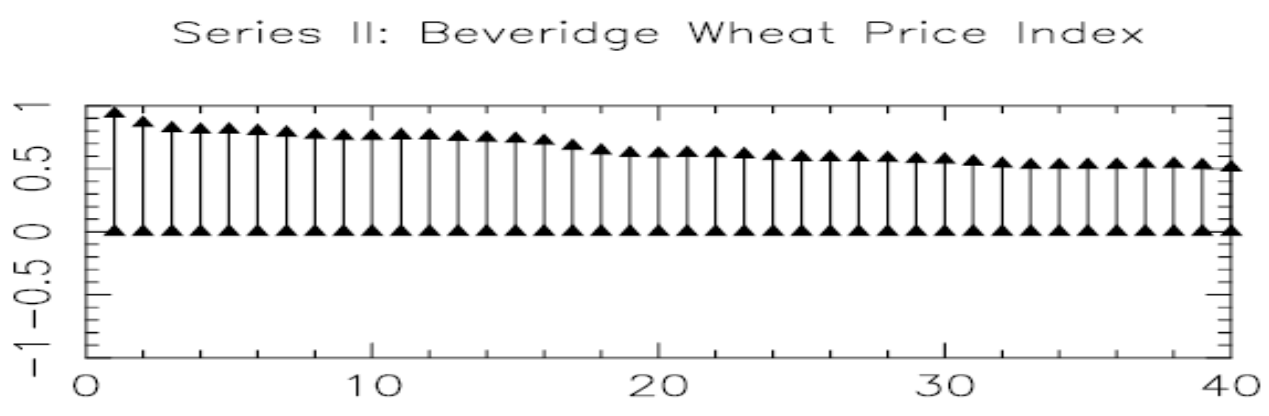
We can learn qualitative and quantitative information regarding linear predictability of a time series by looking at a correlogram (as demonstrated next)!

TS Plot and Scatter Plots (Y_t, Y_{t-k})



TS Correlograms





Sample Confidence Intervals of Correlogram and Inference on Significance of $\rho(k) \neq 0$

Confidence intervals on each r_k estimator (not often plotted in packages)

- On the correlogram (see grey C.I.'s in previous plot for r_k estimators), one can draw upper and lower bounds for autocorrelation with significance level α given by

$$\pm z_{1-\alpha/2} SE(r_k)$$

  with r_k as the estimated autocorrelation at lag k.

- If the autocorrelation is higher (lower) than this upper (lower) bound, the null hypothesis that there is no autocorrelation at and beyond a given lag is rejected at a significance level of α.
- Approximate test assuming the TS is Gaussian WN or CLT argument (since uses z-score).
- $z_{1\alpha/2}$ is the quantile of the normal distribution & SE is the standard error of ACF estimator.

One can estimate $SE(r_k)$ as follows for an ARIMA model for the TS process:

- one typically uses an $MA(q)$ model to approximately compute the s.e. via Bartlett's formula (Autoregressive Integrated Moving Averages described in detail later):

$$SE(r_1) = \frac{1}{\sqrt{T}}$$

$$SE(r_k) = \sqrt{\frac{1 + 2\sum_{i=1}^{k-1} r_i^2}{T}}, \quad k > 1.$$

We will see a more general theorem for this result stated in the linear process section of notes!

sample ACF for WN

Now lets consider the properties of $r(k)$ when sampling from a purely random process, when all the theoretical autocorrelation coefficients are zero except at lag zero.

These results help us to decide if the observed values of $r(k)$ from a given time series are significantly different from zero.

Theorem

For an IID process $\{\epsilon_t\}$, *if* $\mathbb{E}\left[\epsilon_t^4\right] < \infty$, *we have*

$$\widehat{\rho}_\epsilon(k) = \begin{pmatrix} \widehat{\rho}_\epsilon(1) \\ \vdots \\ \widehat{\rho}_\epsilon(k) \end{pmatrix} \sim N\left(0, T^{-1}\mathbb{I}_k\right), \quad \text{as } T \to \infty$$

Note the condition on the moment can be relaxed to for instance use Lindeberg–Lévy condition or Lyapunuv conditions for the CLT.

Correlogram

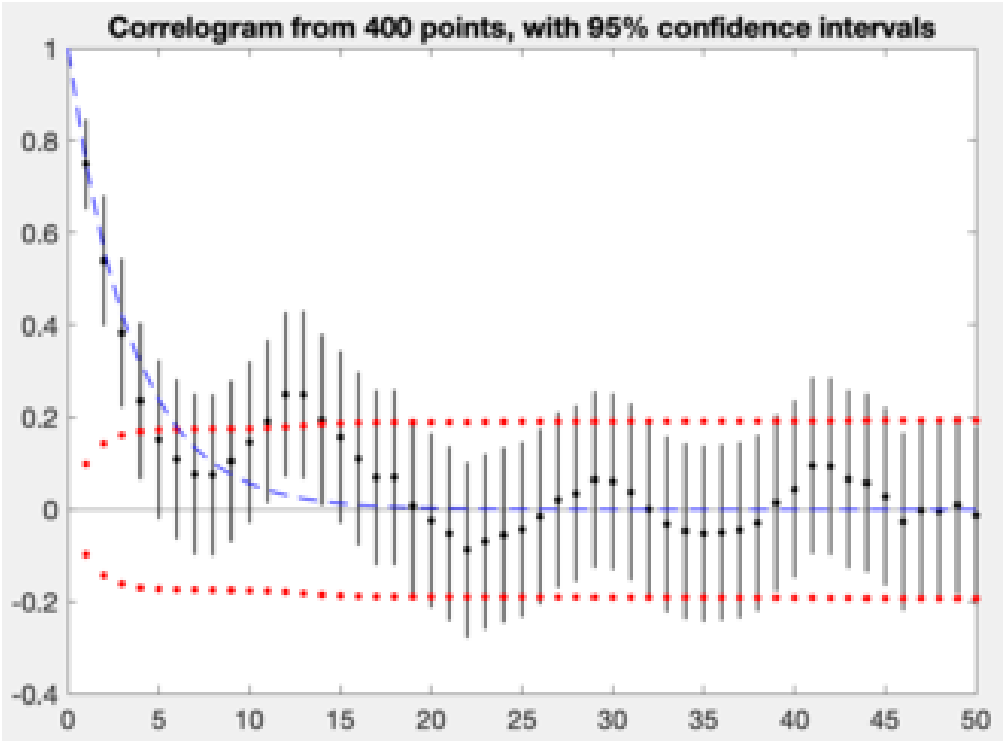


Figure: Correlogram example using 400 samples of an AR(1) process with 0.75 correlation of adjacent points, along with the 95% confidence intervals.

Common practice to plot the confidence intervals in red for the statistical test on whether the estimator at lag k is statistically different from 0. (more after define White Noise).

This theorem yields numerous procedures to test the hypothesis

$$\mathcal{H}_0 : \text{ iid} \quad vs \quad \mathcal{H}_1 : \text{ not iid}$$

Per-Lag z-test: Based on values of sample ACF, if for one k one has that

$$\widehat{\rho}_Y(k) \pm z_{\alpha/2}/\sqrt{T}$$

does not contain zero, then reject $\mathcal{H}_0$.

Portmanteau tests: instead of checking $\widehat{\rho}_Y(k)$ each k consider statistics

- $Q = T\sum_{j=1}^{k} \widehat{\rho}_Y^2(j)$ where under $\mathcal{H}_0$, $Q \sim \chi_k^2$ which gives a decision (rejection region) $Q > \chi_k^2(1-\alpha)$
- Ljung-Box test $Q_{LB} = T(T+2)\sum_{j=1}^{k} \widehat{\rho}_Y^2(j)/(T-j)$ which is better approximated by χ_k^2 under $\mathcal{H}_0$.

Remarks on Ljung-Box Test

- Ljung GM, Box GE. On a measure of lack of fit in time series models. Biometrika. 1978 Aug 1;65(2):297-303.

It is common to use a Ljung-Box test to check that the residuals from a time series model resemble white noise.

- Very little practical advice around about how to choose the number of lags for the test. i.e. selection of k.

Recommended using $k = 10$ for non-seasonal data and $k = 2m$ for seasonal data, where m is the period of seasonality.

- These suggestions are based on power considerations.
- We want to ensure that k is large enough to capture any meaningful and troublesome correlations.
- For seasonal data, it is common to have correlations at multiples of the seasonal lag remaining in the residuals, so we wanted to include at least two seasonal lags.

R CODE EXAMPLE: illustrate the Ljung-Box Test

White Noise Core Stochastic Component of Time Series Model Development

White Noise

White noise is a type of noise that is produced by combining all different frequencies together.

e.g. if you took all of the imaginable tones that a human can hear and combined them together, you would have white noise.

Because white noise contains all frequencies, it is frequently used to mask other sounds.

Take a listen to white noise at
https://www.youtube.com/watch?v=nMfPqeZjc2c

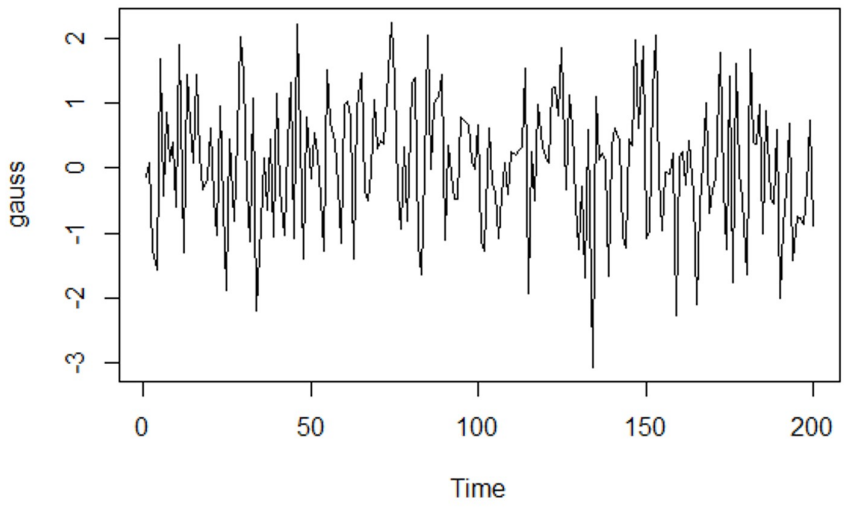
UC SANTA BARBARA UCSB
Prof. Gareth W. Peters

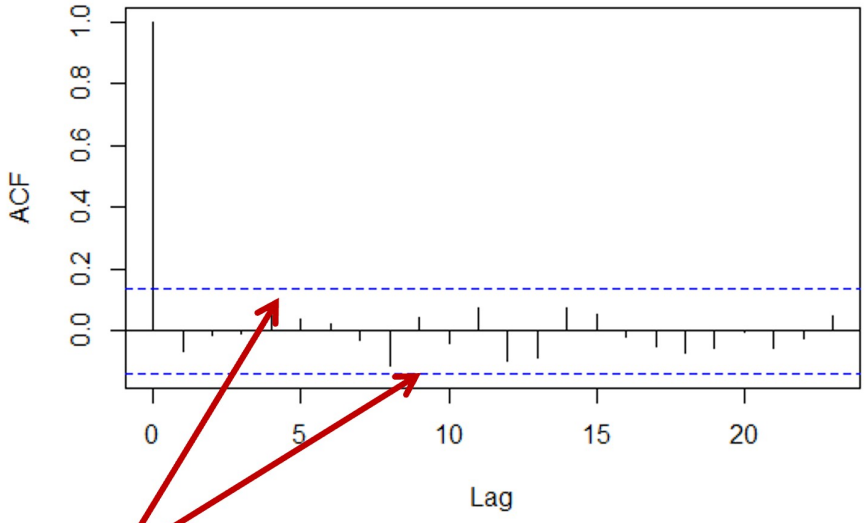
Definition (White Noise)

Z_t are mean zero, constant variance and uncorrelated: $\mu_Z(t) = \mathbb{E}(Z_t) = 0$, $\gamma_Z(t,s) = \text{cov}(Z_t, Z_s) = 0$ *if* $t \neq s$ *(uncorrelated) and* $\gamma_Z(t,t) = \text{cov}(Z_t, Z_t) = \sigma_Z^2 = \text{var}(Z_t)$ *and the law of white noise is given by* $Z_t \sim WN(0, \sigma_Z^2)$.

Various notations often used for White Noise : $\{Z_t\}$ or $\{W_t\}$ or $\{\epsilon_t\}$ it will be made clear by a statement such as $Z_t \sim WN(0, \sigma^2)$ or $\epsilon_t \sim WN(0, \sigma^2)$ etc.

White Noise





No trend or seasonality. Very choppy.
ACF = 0 for lags k= |t-s| > 0.

95% confidence interval based on Normal distribution.

Figure: (Source: Dr. Raya Feldman, UCSB Time Series Lectures 2021)

Example (White noise)

Consider $\{\epsilon_t\} \sim \mathcal{WN}(0, \sigma^2)$. *Then*

$$\rho(k) = \begin{cases} 1, & \text{for } k = 0 \\ 0, & \text{otherwise} \end{cases}$$

Remark

It can be shown that, for $k \neq 0$ *and a large sample size* T*:*

$$r_\epsilon(k) \sim \mathcal{N}(0, 1/T)$$

$\Rightarrow$ 95% *confidence interval is* $(-1.96/\sqrt{T}, 1.96/\sqrt{T})$. *I.e.,*

- *for each* k*, we expect* 95% *of realisations of this time series to have* $r_\epsilon(k)$ *inside interval*
- *observed values of* $r_\epsilon(k)$ *that fall outside these limits are considered 'significantly' different from zero at the 5% level*
- *expect to get* 5% *of* $r_\epsilon(k)$ *coeffs outside the* 95% *confidence limits*
- *if you plot* 20 *values of* $r_\epsilon(k)$ *you would expect to see one value outside the limits by chance*

Based on this result, one can use this in the correlogram to determine if the sample ACF is that arising from WN or from a different signal.

Examples of WN (B)

We say that every $IID(0, \sigma^2)$ sequence TS is $WN(0, \sigma^2)$ but not conversely.

An example of WN that is not IID.

Example

Define a r.v. $Y_t = \epsilon_t$ *when* t *is odd, and* $Y_t = \sqrt{3}\epsilon_{t-1}^2 - 2/\sqrt{3}$ *when* t *is even, where* $\{\epsilon_t\}$ *is an iid sequence from distribution with p.m.f.* $f(-1) = 1/3$, $f(0) = 1/3$, $f(1) = 1/3$.

Is $\{Y_t\}$ *a WN process?*

Is the mean 0?

$$\begin{aligned}
\mathbb{E}[Y_t] &= \begin{cases} \mathbb{E}\left[\sqrt{3}\epsilon_{t-1}^2 - 2/\sqrt{3}\right], & t \text{ even} \\ \mathbb{E}[\epsilon_t], & t \text{ odd} \end{cases} \\
&= \begin{cases} \sqrt{3}\left(1/3(-1)^2 + 1/3(0)^2 + 1/3(1)^2\right) - 2/\sqrt{3}, & t \text{ even} \\ 0, & t \text{ odd} \end{cases}, \\
&= 0
\end{aligned}$$

Example

Is $\{Y_t\}$ *a WN process?*

Is variance constant?

$$
\begin{aligned}
Var[Y_t] &= \mathbb{E}\left[Y_t^2\right] - \mathbb{E}[Y_t]^2 \\
&= \begin{cases} \mathbb{E}\left[\left(\sqrt{3}\epsilon_{t-1}^2 - 2/\sqrt{3}\right)^2\right], & t \text{ even} \\ \left(1/3(-1)^2 + 1/3(0)^2 + 1/3(1)^2\right), & t \text{ odd}, \end{cases} \\
&= \begin{cases} \mathbb{E}\left[3\epsilon_{t-1}^4 - 4\epsilon_{t-1}^2 + 4/3\right], & t \text{ even} \\ 2/3, & t \text{ odd}, \end{cases} \\
&= 2/3
\end{aligned}
$$

$Cov(Y_{t_1}, Y_{t_2}) = 0$ *for all* $t_1 \neq t_2$, *since*

$$Cov\left(\epsilon_t, \sqrt{3}\epsilon_{t-1}^2 - 2/\sqrt{3}\right) = \sqrt{3}Cov\left(\epsilon_t, \epsilon_{t-1}^2\right) \propto \sqrt{Var(\epsilon_t)}\sqrt{Var(\epsilon_{t-1}^2)}Corr(\epsilon_t, \epsilon_{t-1}^2)$$

since ϵ_t *are i.i.d. so uncorrelated then* $Cov\left(\epsilon_t, \epsilon_{t-1}^2\right) = 0$

Example

Is $\{Y_t\}$ *a WN process?*

However, $\{Y_t\}$ *is not an iid sequence.*

Can see this since $Y_{2k+1} = \epsilon_{2k+1}$ *is determined fully by* ϵ_{2k} *such that*

$$\epsilon_{2k} = 0 \Rightarrow \epsilon_{2k+1} = -2/\sqrt{3}, \tag{0.1}$$

$$\epsilon_{2k} = \pm 1 \Rightarrow \epsilon_{2k+1} = \sqrt{3} - 2/\sqrt{3}. \tag{0.2}$$

Illustrations of WN

An example realisation of WN but not IID noise continued:

```
set.seed(100); par(mfrow=c(1,2)); par(mar=c(4,4,2,.5))
t=seq(1,100,by=1); res=c(-1,0,1)
Zt=sample(res,length(t)/2,replace=TRUE); Xt=c()
for(i in 1:length(Zt)){
Xt=c(Xt,c(Zt[i], sqrt(3)*Zt[i]^2-2/sqrt(3)))}
plot(t,Xt,type="o",col="blue",xlab="t",ylab=expression(X[t]))
plot(c(0,t),c(0,cumsum(Xt)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
```

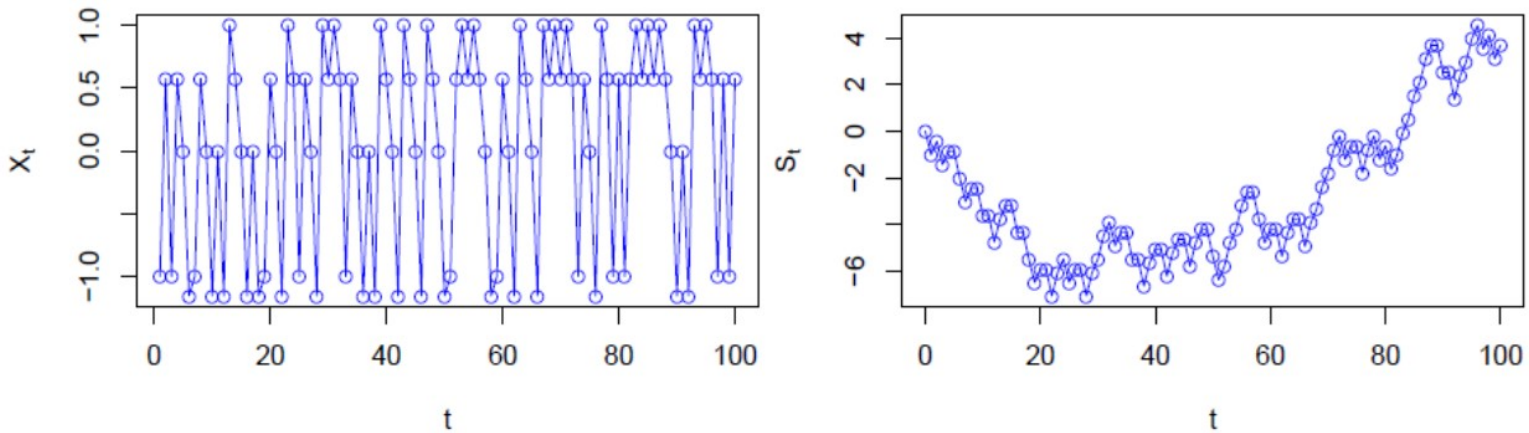


Figure: (source: Dewei Wang https://people.stat.sc.edu/wang52)

What are the most important concepts so far?

Hints:

- **What is time series?**
  - **order of observations** ▪ **in/dependence of observations**
- **Visible characteristics of time series**
  - **length** ▪ **trend** ▪ **seasonality** ▪ **scale**
- **Second-order probabilistic descriptions**
  - **mean** ▪ **variance** ▪ **autocovariance** ▪ **autocorrelation**
- **WN and Gaussian WN**

Notions of Stationarity of a Time Series

Definition

$\{Y_t\}$ *is strictly stationary (SS) if* $\forall k, m, t_1, \ldots t_m$*:*

$$(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_m}) \stackrel{d}{=} (Y_{t_1+k}, Y_{t_2+k}, \ldots, Y_{t_m+k})$$

Remark

Strict stationarity $\Rightarrow$ *time shift has no effect on any joint distributions.*

Remark

Put $m = 1$*. Then strict stationarity* $\Rightarrow Y_s \stackrel{d}{=} Y_t, \forall s, t$ *and*

$$\mathbb{E}(Y_s) = \mathbb{E}(Y_t), \quad \mu(t) = \mu, \quad \text{var}(Y_s) = \text{var}(Y_t), \quad \sigma^2(t) = \sigma^2$$

Remark

Put $m = 2$*. Then strict stationarity* $\Rightarrow (Y_t, Y_{t+k}) \stackrel{d}{=} (Y_s, Y_{s+k}) \forall k, s, t$ *and*

$$\text{cov}(Y_t, Y_{t+k}) = \text{cov}(Y_s, Y_{s+k})$$

Example of strictly stationary processes:

- Example 1: If $\{Y_t\}$ is an iid sequence, then it is strictly stationary.

- Example 2: Non iid sequence, let $\{Y_t\}$ be an iid sequence and let $X \sim N(0,1)$ independent of $\{Y_t\}$. Define $Z_t = Y_t + X$. The sequence $\{Z_t\}$ is not an independent sequence (because of common X) but it is identically distributed sequence and is strictly stationary.

Visual Analysis of Strict Stationarity

UC SANTA BARBARA UCSB
Prof. Gareth W. Peters

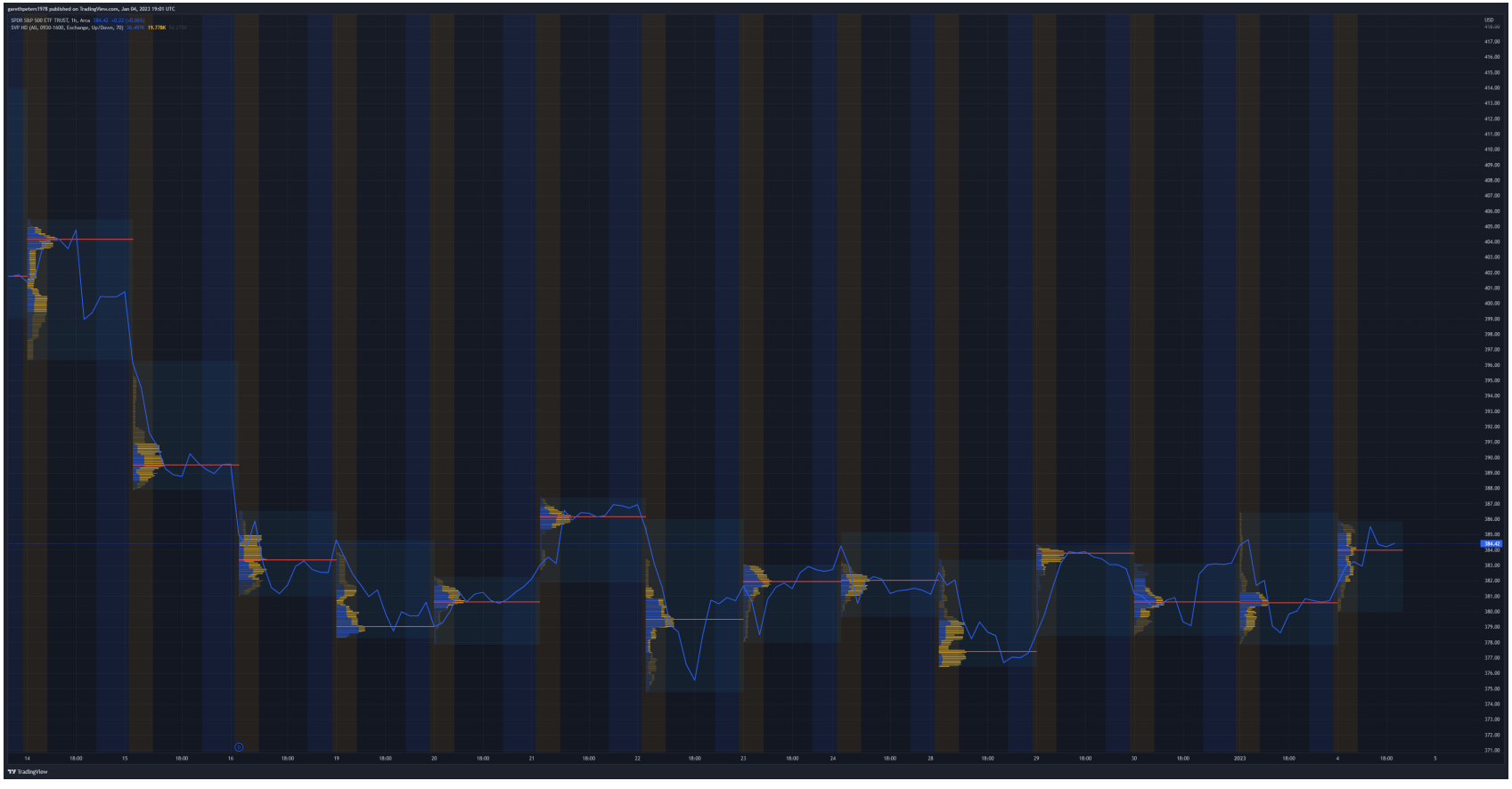


Figure: SPY ETF

Definition

$\{Y_t\}$ *is weakly stationary (WS) if* $\forall k, s, t$*:*

1. $\mathbb{E}(Y_s) = \mathbb{E}(Y_t)$, *i.e.* $\mu(t) = \mu$ *(a constant)*
2. $\text{cov}(Y_t, Y_{t+k}) = \text{cov}(Y_s, Y_{s+k})$ *i.e.*
   $\gamma_Y(t, s) = \gamma_Y(t + k, s + k) = \gamma_Y(t - s)$ *for all* $t, s, k \in T$

Weak stationarity generally does not imply strict stationarity, howerver:

Corollary

strict stationarity + finite moments $\Rightarrow$ *weak stationarity*

Remark

Weak Stationarity $\Rightarrow \gamma(t, t + k) = \gamma(s, s + k)$ *is a function of lag* k *and we write (define)*

$$\gamma(k) := \text{cov}(Y_t, Y_{t+k})$$

In particular, ws $\Rightarrow \text{var}(Y_t) = \gamma(0)$, *i.e.* $\sigma_Y^2(t) = \sigma_Y^2$ *(const.)*

R CODE EXAMPLE:

We can use feasts package:

```
var_tiled_var(x, .size = NULL, .period = 1)
var_tiled_mean(x, .size = NULL, .period = 1)
```

It will return a sequence of windows of data mean and variance you can plot to see if likely that weak sense stationarity is present (at least in mean and variance aspect).

For ACF - can plot this over time using package: runner
https://cran.r-project.org/web/packages/runner/runner.pdf

Caveat: 'Second order stationarity'

Some texts say weak stationarity is also called second-order stationarity. Unfortunately, the literature is sometimes inconsistent.

Definition ('Second order stationary')

$\{Y_t\}$ *is sometimes called* *second order stationary* *or* *stationary to the 2nd order* *if* $\forall k, s, t,$

$$(Y_t) \stackrel{d}{=} (Y_{t+k}) \quad \textit{and} \quad (Y_t, Y_s) \stackrel{d}{=} (Y_{t+k}, Y_{s+k})$$

We will use the 'weakly stationary' definition (Defn. previous slide).

Definition (Covariance-Stationary)

Process $\{Y_t\}$ *is covariance-stationary (weak stationary) if*

1. $\mathbb{E}[Y_t] = \mu, \quad \forall t$
2. $\mathbb{V}\mathrm{ar}[Y_t, Y_{t-j}] = \mathbb{E}[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j, \quad \forall t, j$

This is weak stationarity because it only relates to the first two moments, higher moments can be time varying.

- Example 1: $Y_t \overset{iid}{\sim} WN(0, \sigma^2) \Rightarrow \{Y_t\}$ white noise (WN)
- Example 2: $Y_t \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow$ Gaussian white noise (GWN)

Example

Consider process $\{Y_t = Acos(\theta t) + Bsin(\theta t)\}$ *with* A, B *independent random variables with zero mean and unit variance. Let* $\theta \in [-\pi, \pi]$.

Is Y_t **weakly stationary?**

Is the mean constant?

$$\mathbb{E}[Y_t] = cos(\theta t)\mathbb{E}[A] + sin(\theta t)\mathbb{E}[B] = 0$$

Is the ACVF just a function of time lag?

$$\begin{aligned}
\gamma_Y(t+k,t) &= \mathbb{E}[Y_{t+k}Y_t] \\
&= \mathbb{E}[(Acos(\theta t+\theta k) + Bsin(\theta t+\theta k))(Acos(\theta t) + Bsin(\theta t))] \\
&= cos(\theta t+\theta k)cos(\theta t)\mathbb{E}[A^2] + sin(\theta t+\theta k)sin(\theta t)\mathbb{E}[B^2] \\
&\quad (\text{note } A \text{ indep. } B \Rightarrow \mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B] = 0) \\
&= cos(\theta t+\theta k)cos(\theta t)\sigma_A^2 + sin(\theta t+\theta k)sin(\theta t)\sigma_B^2 \\
&= cos(\theta t+\theta k)cos(\theta t) + sin(\theta t+\theta k)sin(\theta t) \\
&\quad (\text{cosine identity } cos(x-y) = cos(x)cos(y) + sin(x)sin(y)) \\
&= cos(\theta k)
\end{aligned}$$

which is free of t *so* $\{Y_t\}$ *is weakly stationary.*

Useful properties of a Strictly Stationary time series $\{Y_t\}$

Theorem

A strictly stationary time series $\{Y_t\}$ satisfies:

- *Y_t's are from the same distribution.*
- *$(Y_t, Y_{t+k}) \stackrel{d}{=} (Y_1, Y_{1+k})$ for all t, k integers*
- *$\{Y_t\}$ is weakly stationary if $\mathbb{E}\left[Y_t^2\right] < \infty$ for all t.*
- *Weak stationarity does not imply strict stationarity*
- *An iid sequence is strictly stationary.*

Proofs are direct from definition application or simple HW exercise.

Another useful property of strict stationary is that it is preserved under general transformations.

Theorem
Let $\{Y_t\}$ be strictly stationary and let $g(\cdot)$ be any function of the elements in $\{Y_t\}$. Then $\{g(Y_t)\}$ is also strictly stationary.

For example, if $\{Y_t\}$ is strictly stationary then $\left\{Y_t^2\right\}$ and $\{Y_t Y_{t-1}\}$ are also strictly stationary.

Example (q-dependent strictly stationary time series)

A simple way to construct a time series $\{Y_t\}$ *that is strictly stationary is to "filter" an iid sequence. Let* $\{\epsilon_t\} \sim IID(0, \sigma^2)$, *then define*

$$Y_t = g\left(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-1}\right)$$

for some real-valued function g.

Then $\{Y_t\}$ *is strictly stationary and also q-dependent such that* Y_s *and* Y_t *are independent only whenever* $|t - s| > q$.

Example of WN Smoothing (B)

Smoothing of WN$(0, \sigma^2)$ that induces q-dependence
Goal: Calculate AVCF of the smoothed Gaussian WN:
$Y_t = \frac{1}{3}(Z_{t-1} + Z_t + Z_{t+1})$

$$\begin{aligned}
\mathbb{E}(Y_t) &= \frac{1}{3}(\mathbb{E}(Z_{t-1}) + \mathbb{E}(Z_t) + \mathbb{E}(Z_{t+1})) = 0 \\
\gamma_Y(t,s) &= \text{cov}(X_t, X_s) = \mathbb{E}(X_t X_s) \\
&= \frac{1}{9}\mathbb{E}\left[(Z_{t-1} + Z_t + Z_{t+1})(Z_{s-1} + Z_s + Z_{s+1})\right] \\
&= \frac{1}{9}\mathbb{E}\left[(Z_{t-1}Z_{s-1} + Z_t Z_s + Z_{t+1}Z_{s+1}) + (Z_{t-1}Z_s + Z_t Z_{s+1})\right. \\
&+ \left.(Z_t Z_{s-1} + Z_{t+1}Z_s) + (Z_{t-1}Z_{s+1} + Z_{t+1}Z_{s-1})\right] \\
&= \begin{cases} \frac{3}{9}\sigma_Z^2 & \text{if } |t-s| = 0 \\ \frac{2}{9}\sigma_Z^2 & \text{if } |t-s| = 1 \\ \frac{1}{9}\sigma_Z^2 & \text{if } |t-s| = 2 \\ 0 & \text{if } |t-s| > 2 \end{cases}
\end{aligned}$$

Summary to Remember

Stationary Time Series

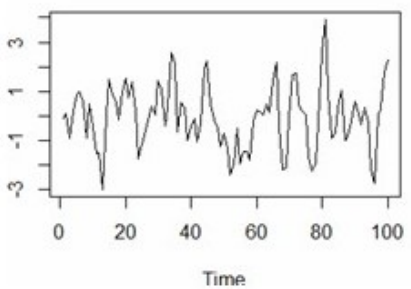
- **No trend:** Mean $\mu_X(t) = E(X_t) = \mu_X$ (constant)
- **No change of variance:** Variance $\sigma_X^2(t) = Var(X_t) = \sigma_X^2$ (constant)
- **No seasonality, no sharp change of behavior**

- **ACVF is a function of a lag:**

$$\gamma_X(t, s) = Cov(X_t, X_s) = \gamma_X(t - s) = \gamma_X(k) \text{ with } |t - s| = k;$$

- **ACF is a function of a lag:**

$$\rho_X(t, s) = Cor(X_t, X_s) = \rho_X(k) \text{ with } |t - s| = k.$$



- **Some Formulas to Remember:**

$$\rho_X(k) = \gamma_X(k) / \gamma_X(0); \qquad \gamma_X(0) = Cov(X_t, X_t) = Var(X_t) = \sigma_X^2;$$
$$\gamma_X(k) = \gamma_X(-k); \qquad \rho_X(k) = \rho_X(-k)$$

Best to plot the time series and look for structures in the trace plot (time series plot) of observed samples that indicate any of the following

- Any time series with a non-constant trend is not stationary.
- Any time series with a seasonal, cyclical pattern is not stationary.
- Any time series with non-constant variance is not stationary.

Remarks: Detecting Non-Stationarity

Example *Deterministically trending process*

Suppose $\{Y_t\}_{t=0}^{\infty}$ is generated according to the deterministically trending process

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, \ \varepsilon_t \sim WN(0, \sigma_\varepsilon^2),$$
$$t = 0, 1, 2, \ldots$$

Then $\{Y_t\}_{t=0}^{\infty}$ is nonstationary because the mean of Y_t depends on t :

$$E[Y_t] = \beta_0 + \beta_1 t \ \text{ depends on } t.$$

Figure 1.3 shows a realization of this process with $\beta_0 = 0, \beta_1 = 0.1$ and $\sigma_\varepsilon^2 = 1$ created using the R commands:

```
> set.seed(123)
> e = rnorm(250)
> y.dt = 0.1*seq(1,250) + e
> ts.plot(y.dt, lwd=2, col="blue", main="Deterministic Trend + Noise")
> abline(a=0, b=0.1)
```

Remarks: Detecting Non-Stationarity

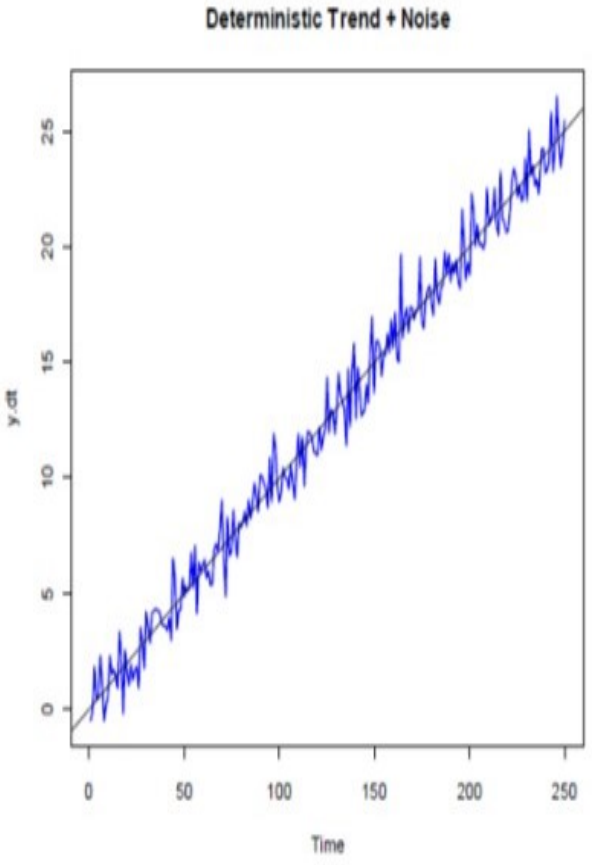


Figure Deterministically trending nonstationary process $Y_t = 0.1t + \varepsilon_t, \varepsilon_t \sim N(0,1)$

Here the non-stationarity is created by the deterministic trend $\beta_0 + \beta_1 t$ in the data. The non-stationary process $\{Y_t\}_{t=0}^{\infty}$ can be transformed into a stationary process by simply subtracting off the trend:

$$X_t = Y_t - \beta_0 - \beta_1 t = \varepsilon_t \sim \text{WN}(0, \sigma_\varepsilon^2).$$

The detrended process $X_t \sim \text{WN}(0, \sigma_\varepsilon^2)$. ■

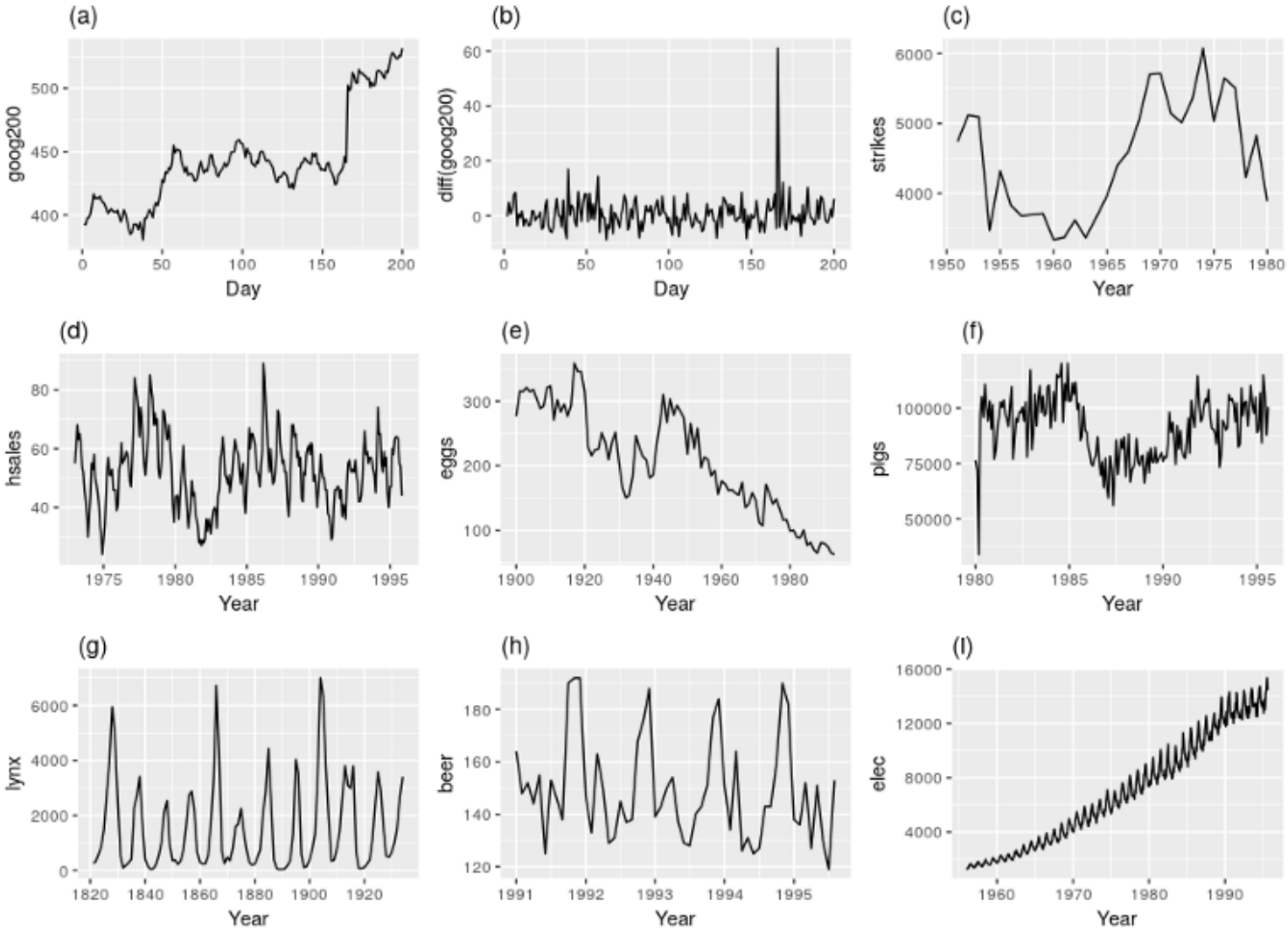
Inference Procedures for Stationarity

Stationarity and Gaussian Distribution

- The most basic methods for stationarity detection rely on plotting the data, and visually checking for trend and seasonal components.
- Trying to determine whether a time series was generated by a stationary process just by looking at its plot is a dubious task. However, there are some basic properties of non-stationary data that we can look for.
- Let's take an example the following nice plots from [Hyndman Athanasopoulos, 2018]:

  Nine examples of time series data; (a) Google stock price for 200 consecutive days; (b) Daily change in the Google stock price for 200 consecutive days; (c) Annual number of strikes in the US; (d) Monthly sales of new one-family houses sold in the US; (e) Annual price of a dozen eggs in the US (constant dollars); (f) Monthly total of pigs slaughtered in Victoria, Australia; (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada; (h) Monthly Australian beer production; (i) Monthly Australian electricity production.

Testing for Stationarity



Properties:

- Seasonality can be observed in series (d), (h), and (i)
- The trend can be observed in series (a), (c), (e), (f), and (i)
- Series (b) and (g) are stationary

Statistical Tests:

Some statistical tests which we will be discussing are

- Augmented Dickey-Fuller (ADF) Test
- Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The Augmented Dickey-Fuller test is a type of statistical test called a unit root test.

- A unit root is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving time series models.
- The unit root is non-stationary but does not always have a trend component.

ADF test is conducted with the following assumptions.

- Null Hypothesis ($\mathcal{H}_O$): Series is non-stationary or series has a unit root.
- Alternate Hypothesis ($\mathcal{H}_A$): Series is stationary or series has no unit root.

If the null hypothesis is failed to be rejected, this test may provide evidence that the series is non-stationary.

Conditions to Reject Null Hypothesis ($\mathcal{H}_O$):

- If Test statistic $<$ Critical Value i.e p-value $<$ 0.05
- – Reject Null Hypothesis ($\mathcal{H}_O$) i.e., time series does not have a unit root, meaning it is stationary.
- It does not have a time-dependent structure.

Basics of the Dickey-Fuller and (Augmented version) Test:
Consider the model (more about model structure later)

$$Y_t = \mu + \beta t + \alpha Y_{t-1} + \sum_{j=1}^{p} \phi_j \triangle Y_{t-j} + \epsilon_t$$

- A Dickey-Fuller test is a unit root test that tests the null hypothesis that $\alpha = 1$ with $p = 0$.
- alpha is the coefficient of the first lag on Y and Y_{t-1} is the first lag, with $\triangle Y_{t-j}$ first difference in series at time $t - j$.
- ADF test has $p > 0$

The Augmented Dickey-Fuller test evolved based on the above equation and is one of the most common form of Unit Root Test.

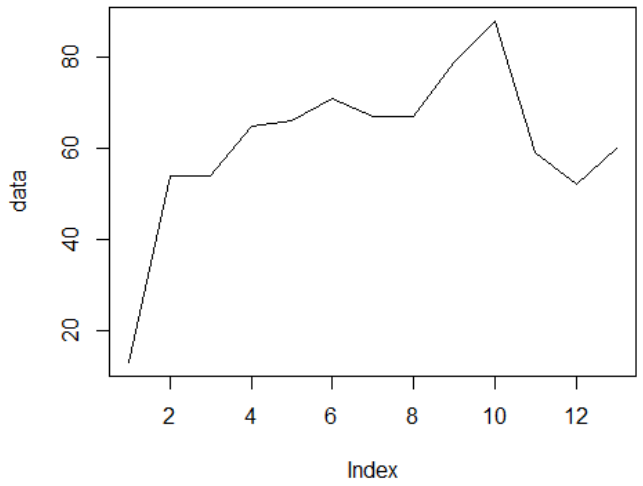
- The ADF test expands the Dickey-Fuller test equation to include high order regressive process in the model $p > 0$

A key point to remember: since the null hypothesis assumes the presence of unit root, that is $\alpha = 1$, the p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis. Thereby, inferring that the series is stationary.

Example in R

```
library(tseries)
data <-
c(13, 54, 54, 65, 66, 71, 67, 67, 79, 88, 59, 52, 60)
plot(data, type='l')
adf.test(data)
```

Outcome is



Augmented Dickey-Fuller Test
data: data
Dickey-Fuller = -1.6549, Lag order = 2, p-value = 0.7039
alternative hypothesis: stationary
We cannot reject the null hypothesis because the p-value is not smaller than 0.05.

- Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54 (1-3): 159-178.

The KPSS test, short for, Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a type of Unit root test that tests for the stationarity of a given series around a deterministic trend.

- A common misconception, however, is that it can be used interchangeably with the ADF test. This can lead to misinterpretations about the stationarity, which can easily go undetected causing more problems down the line.
- A major difference between KPSS and ADF tests is the capability of the KPSS test to check for stationarity in the 'presence of a deterministic trend'.
- What that effectively means to us is, the test may not necessarily reject the null hypothesis (that the series is stationary) even if a series is steadily increasing or decreasing.

A KPSS test used to determine if a time series is trend stationary.

This test uses the following null and alternative hypothesis:

- $\mathcal{H}_O$: The time series is trend stationary.
- $\mathcal{H}_A$: The time series is not trend stationary.

If the p-value of the test is less than some significance level (e.g. $\alpha = .05$) then we reject the null hypothesis and conclude that the time series is not trend stationary.

- Otherwise, we fail to reject the null hypothesis.

Reading (Advanced) for more details:

- Lima LR, Neri B. A test for strict stationarity. InUncertainty Analysis in Econometrics with Applications 2013 (pp. 17-30). Springer, Berlin, Heidelberg.

KPSS Test assumes that the time series can be divided into a deterministic trend, a random walk and a stationary error.

$$Y_t = \beta t + R_t + \epsilon_t$$

with βt deterministic trend, R_t random walk and ϵ_t stationary error with zero mean.

$$R_t = R_{t-1} + W_t$$

with R_0 the intercept and $W_t \sim iid(0, \sigma_w^2)$.

- Null hypothesis implies that $\sigma_w^2 = 0$ which results in a time series Y_t being trend stationary (stationary around a trend).
- KPSS then tests if there is a unit root in R_t when β is non-zero.

The test statistic is the one-sided Lagrange Multiplier

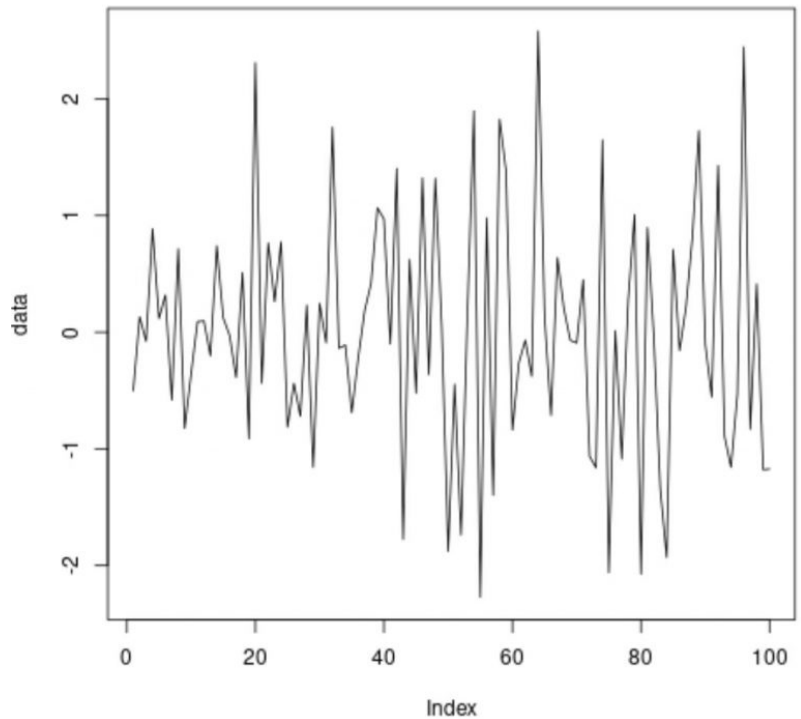
$$LM = T^{-2} \frac{\sum_{t=1}^{T} S_t^2}{\widehat{\sigma}_\epsilon^2}$$

with S_t the sum of the residuals and $\widehat{\sigma}_\epsilon^2$ estimated error variance.

R Example of KPSS test.

```
make this example reproducible
set.seed(100)
create time series data
data<-rnorm(100)
plot time series data as line plot
plot(data, type='l')
```

UC SANTA BARBARA UCSB
Prof. Gareth W. Peters



```
library(tseries)
perform KPSS test
kpss.test(data, null="Trend")
KPSS Test for Trend Stationarity
data: data
KPSS Trend = 0.034563, Truncation lag parameter = 4,
p-value = 0.1
Warning message:
In kpss.test(data, null = "Trend") : p-value greater than
printed p-value
```

White Noise, Multivariate Gaussian and Gaussian Processes

Remark

For Gaussian Time Series i.e. linear time series with WN process generated by a Gaussian distribution:

$$stationarity = strict\ stationarity$$

Why?
Ans: think Sufficient Statistics

Recall:

Definition

A statistic $T(\boldsymbol{X})$ is **sufficient** *for model parameters $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on $\boldsymbol{\theta}$.*

Note: Gaussian distributions are determined by sufficient statistics given by mean and covariance.

To prove the sufficient statistics of a Guassian are the mean and covariance - one can use the Fisher-Neyman Factorisation theorem.

Theorem (Fisher-Neyman Factorisation)

Let $X_1, X_2, \ldots, X_n$ be a random asmple with joint density $f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta})$. A statistic $T(\boldsymbol{X}) = r(X_1, \ldots, X_n)$ is sufficient if and only if the joint density can be factored as follows:

$$f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = a(x_1, \ldots, x_n)b\left(r(x_1, x_2, \ldots, x_n), \boldsymbol{\theta}\right)$$

where a and b are non-negative functions and the function a can depend on the full random sample $x_1, \ldots, x_n$, but not on unknown parameters $\boldsymbol{\theta}$. The function b can depend on $\boldsymbol{\theta}$ but only on the sample through the statistic value $r(x_1, x_2, \ldots, x_n)$.

Properties of Gaussian Distributions

Facts about bivariate Gaussian distribution

- Continuous r.v.s X_1 and X_2 have a bivariate normal distribution with the parameters $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and covariance

$$\begin{aligned}\Sigma &= \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_2 - \mathrm{E}[X_2])] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\end{aligned}$$

with $\rho \in [-1, 1]$ and $\sigma_i^2 \geq 0$ if the joint p.d.f. $f(x_1, x_2)$ is given by

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

Facts about bivariate Gaussian distribution continued

- the marginal distributions for X_1 and X_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively
- the $\text{corr}(X_1, X_2) = \rho$
- X_1 and X_2 are independent if $\rho = 0$
- Conditional dist. of X_2 given $X_1 = x$ is

$$N\left(\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1), (1 - \rho^2)\sigma_2^2\right)$$

- Conditional dist. of X_1 given $X_2 = y$ is

$$N\left(\mu_1 + \rho(\sigma_1/\sigma_2)(y - \mu_2), (1 - \rho^2)\sigma_1^2\right)$$

Sufficiency for d-variate Guassian random vector.

Example (Gaussian) $X \sim \mathcal{N}(\mu, \Sigma)$ *is d-dimensional.*
$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$; $\theta = (\mu, \Sigma)$

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi^d |\Sigma|}} e^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$
$$= 2\pi^{-nd/2} |\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$

Define sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and sample covariance

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Sufficiency for d-variate Guassian random vector.

$$\begin{aligned}
\exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) &= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu}+\hat{\mu}-\mu)^T\Sigma^{-1}(x_i-\hat{\mu}+\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(x_i-\hat{\mu}) - \sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(\hat{\mu}-\mu) - \frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(x_i-\hat{\mu}))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}tr(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}tr(\Sigma^{-1}(n\hat{\Sigma})))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu))
\end{aligned}$$

Note that the second term on the second line is zero because $\frac{1}{n}\sum_i x_i = \hat{\mu}$. *For any matrix* B, $tr(B)$ *is the sum of the diagonal elements. On the fourth line above we use the* trace *property,* $tr(AB) = tr(BA)$.

$$p(x_1,\ldots,x_n|\theta) = \underbrace{2\pi^{-nd/2}|\Sigma|^{-n/2}\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu))\exp(-\frac{1}{2}tr(\Sigma^{-1}n\hat{\Sigma}))}_{b(\hat{\mu},\hat{\Sigma},\theta)}\cdot\underbrace{1}_{a(x_1,\ldots,x_n)}$$

Example

White noise $Z_t \sim WN(0, \sigma^2)$ *is a TS s.t.*

$$\begin{aligned} \mathbb{E}[Z_t] &= 0 \\ \gamma(s,t) = \mathbb{E}[Z_t Z_s] &= \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Question: is it stationary? Strictly stationary?

Does knowing that Z_t is Gaussian distributed make a difference?

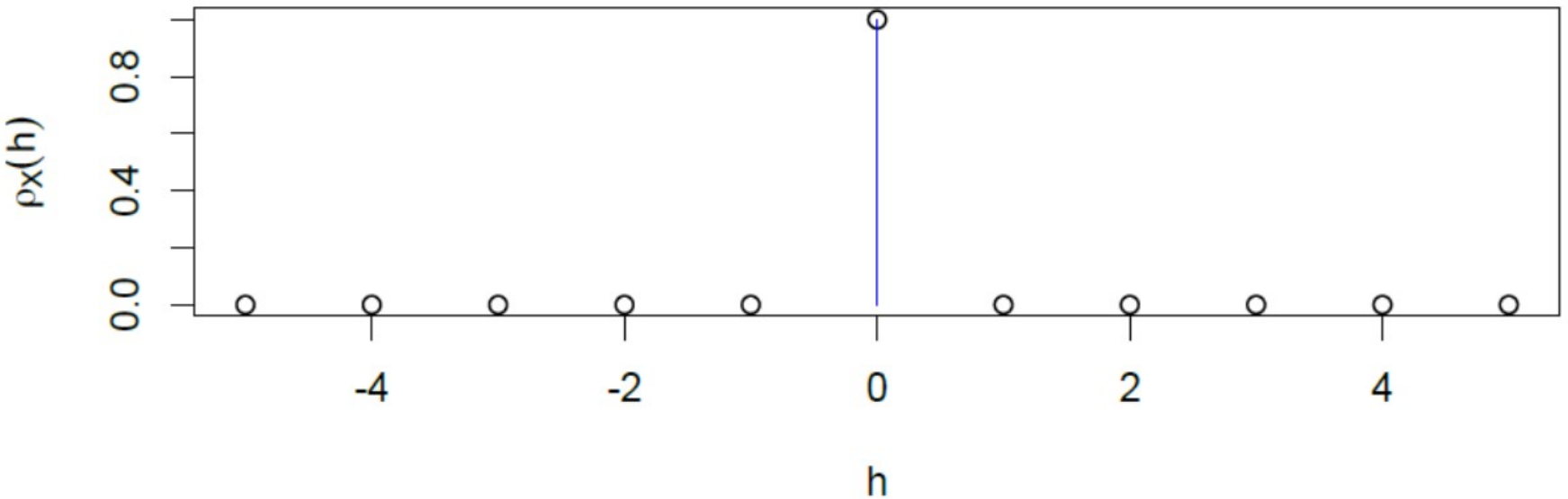
Example

White noise $Z_t \sim N(0, \sigma^2)$ *for all* t*, is a TS.*

Question: is it stationary? Strictly stationary?

Example of ACF for WN:

```
rho=function(h,theta){I(h==0)*1}
h=seq(-5,5,1); s=1:length(h); y=rho(h,.6)
plot(h,y,xlab="h",ylab=expression(rho[X](h)))
segments(h[s],y[s],h[s],0,col="blue")
```



Example

Consider a smoothed Gaussian WN given by

$$X_t = \frac{1}{3}(Z_{t-1} + Z_t + Z_{t+1})$$

We know from previously that the mean and ACVF are given by

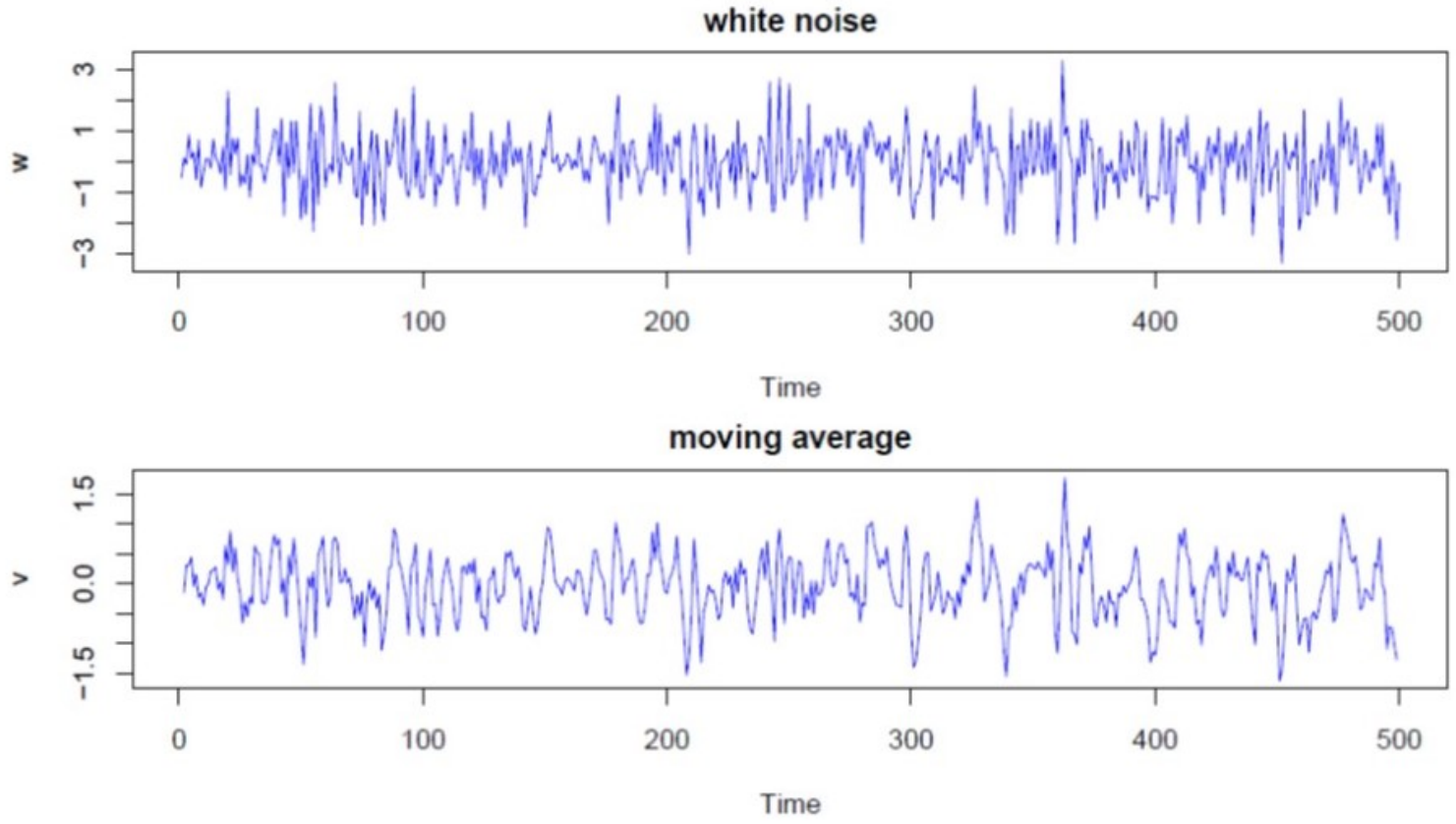
$$\begin{aligned}
\mathbb{E}(X_t) &= \frac{1}{3}(\mathbb{E}(Z_{t-1}) + \mathbb{E}(Z_t) + \mathbb{E}(Z_{t+1})) = 0 \\
\gamma_X(t,s) &= \begin{cases} \frac{3}{9}\sigma_Z^2 & \text{if } |t-s| = 0 \\ \frac{2}{9}\sigma_Z^2 & \text{if } |t-s| = 1 \\ \frac{1}{9}\sigma_Z^2 & \text{if } |t-s| = 2 \\ 0 & \text{if } |t-s| > 2 \end{cases}
\end{aligned}$$

Question: is the time series stationary? Strictly stationary?

Calculate the ACF and plot the correlogram for the TS of X_t.

Examples: WN Smoother


```
set.seed(100); w = rnorm(500,0,1) # 500 N(0,1) variates
v = filter(w, sides=2, rep(1/3,3)) # moving average
par(mfrow=c(2,1)); par(mar=c(4,4,2,.5))
plot.ts(w, main="white noise",col="blue")
plot.ts(v, main="moving average",col="blue")
```



Examples: Random walk

Consider $Z_t \sim IID(0, \sigma_Z^2)$ be I.I.D. WN.

Example

Random Walk (RW) is defined as $X_t = Z_1 + \ldots + Z_t$
Calculate summary of representation of RW:

$$\mathbb{E}[X_t] = \mathbb{E}[Z_1 + \ldots + Z_t] = \sum_{i=1}^{t} \mathbb{E}[Z_i] = 0$$

Take $s = t + k$, $k \geq 0$. *Then*

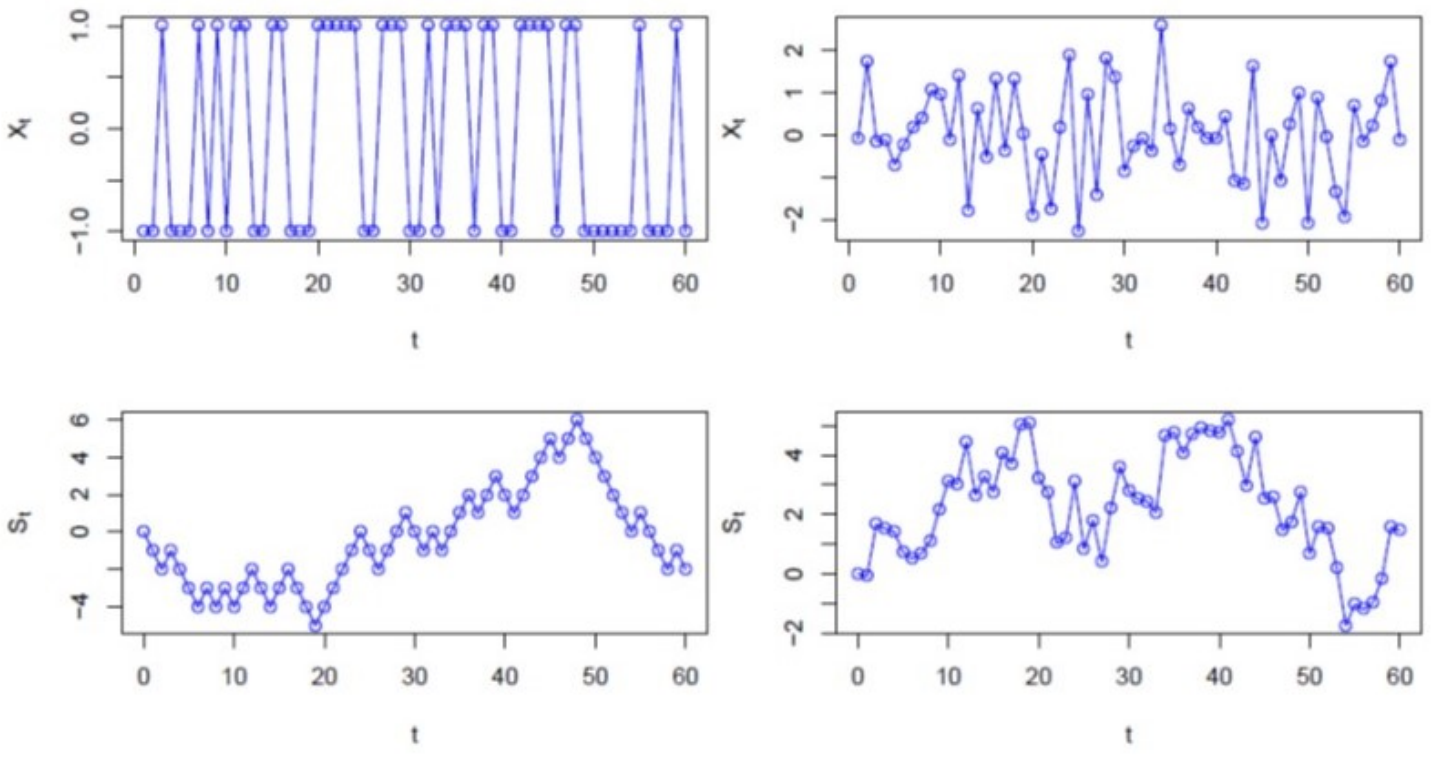
$$\begin{aligned}
\gamma_X(t,s) &= \mathbb{E}[X_t X_s] = \mathbb{E}[(Z_1 + \cdots + Z_t)(Z_1 + \ldots + Z_t + Z_{t+1} + \cdots + Z_{t+k})] \\
&= \mathbb{E}\left[(Z_1 + \cdots + Z_t)^2\right] + \mathbb{E}[(Z_1 + \ldots + Z_t)(Z_{t+1} + \cdots + Z_{t+k})] \\
&= \text{var}(Z_1 + \cdots + Z_t) + 0 = t\sigma^2.
\end{aligned}$$

Summary:$\gamma_X(t,s) = \sigma^2 \min(t,s)$ *and in particular* $\sigma_X^2 = \gamma_X(t,t) = \sigma^2 t$

Question: is RW stationary? Strictly stationary?

Examples: Random walk


```
set.seed(100); par(mfrow=c(2,2)); par(mar=c(4,4,2,.5))
t=seq(1,60,by=1); Xt1=rbinom(length(t),1,.5)*2-1
plot(t,Xt1,type="o",col="blue",xlab="t",ylab=expression(X[t]))
t=seq(1,60,by=1); Xt2=rnorm(length(t),0,1)
plot(t,Xt2,type="o",col="blue",xlab="t",ylab=expression(X[t]))
plot(c(0,t),c(0,cumsum(Xt1)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
plot(c(0,t),c(0,cumsum(Xt2)),type="o",col="blue",xlab="t",ylab=expression(S[t]))
```



Top: One realization of a binary process (left) and a Gaussian noise (right). Bottom: the corresponding random walk

Definition (Gaussian Processes)

A process, $\{Y_t\}$ is said to be a **Gaussian process** *if all n-dimensional vectors ($n \in \mathbb{N}$) given by $\boldsymbol{Y} = (Y_{t_1}, \ldots, Y_{t_n})$, for any collection of time points $t_1, \ldots, t_n$, have a multivariate Gaussian distribution.*

That is all Finite Dimensional Distributions (F.D.D.'s) are multivariate Gaussian.

Lemma

For Gaussian processes, weakly stationary is equivalent to strictly stationary.

Proof.

It suffices to show that every weakly stationary Gaussian process $\{X_t\}$ is strictly stationary. Suppose it is not true (by contradiction), then there must exist (t_1, t_2) and $(t_1 + h, t_2 + h)$ such that (X_{t_1}, X_{t_2}) and (X_{t_1+h}, X_{t_2+h}) have different distributions, which contradicts the assumption of weakly stationary and definition of a Gaussian process. □

Theorem

A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and non-negative.

Proof.

Ony need to prove that for any even and non-negative definite function $K(\cdot)$, we can find a stationary process $\{Y_t\}$ such that $\gamma_X(h) = K(h)$ for any integer h. It is trivial to chose $\{Y_t\}$ to be a Gaussian process such that $Cov(Y_i, Y_j) = K(i-j)$ for any i and j. □

Wold Decomposition Theorem & Time Series Models Built from WN

The Wold decomposition theorem states that any covariance stationary process can be decomposed into two mutually uncorrelated component processes,

- one a linear combination of lags of a white noise process; and
- and the other a process, future values of which can be predicted exactly by some linear function of past observations.

As we will see, one reason for the popularity of the class of linear time series models we will develop, known as the ARIMA family of models, derives from Wold's Theorem.

Definition

Let $\{Y_t\}$ *be a covariance-stationary process. The projection random variable*

$$P\left[Y_{t+h}|Y_{t-1},\ldots,Y_{t-N}\right]=\beta_0^N+\beta_1^N Y_{t-1}+\cdots+\beta_N^N Y_{t-N}$$

where coefficients $\beta_0^N,\beta_1^N,\ldots,\beta_N^N$ *are such that*

$$S\left(\beta_0^N,\beta_1^N,\ldots,\beta_N^N\right)=\mathbb{E}\left[Y_{t+h}-P\left[Y_{t+h}|Y_{t-1},Y_{t-N}\right]\right]^2$$

is minimum, is called the **orthogonal projection** *of* Y_{t+h} *on past* $Y_{t-1},\ldots,Y_{t-N}$.

Introduction to Wold Decomposition (B)

The orthogonal projection of Y_{t+h} on infinite past $Y_{t-1}, Y_{t-2}, \ldots$, denoted $P[Y_{t+h}|Y_{t-1}, Y_{t-2}, \ldots]$ is defined by

$$P[Y_{t+h}|Y_{t-1}, Y_{t-2}, \ldots] = \lim_{N\to\infty} P[Y_{t+h}|Y_{t-1}, \ldots, Y_{t-N}]$$

Definition

A covariance-stationary process $\{Y_t\}$ is called linearly deterministic if

$$P[Y_t|Y_{t-1}, Y_{t-2}, \ldots] = Y_t$$

We have then that a stationary process $\{Y_t\}$ is deterministic if Y_t can be predicted correctly (with zero error) using the entire past $Y_{t-1}, Y_{t-2}, \ldots$.

For a deterministic process the one-step prediction error is zero.

Lets take an example. Let $\{Y_t\}$ be a stochastic process defined by

$$Y_t = A\cos(t) + B\sin(t)$$

for A and B independent standard normal random variables. **This process is linearly deterministic**. In fact it is possible to show that

$$Y_t = \frac{\sin(2)}{\sin(1)} Y_{t-1} - Y_{t-2}$$

and this gives

$$P\left[Y_t | Y_{t-1}, Y_{t-2}, \ldots\right] = \frac{\sin(2)}{\sin(1)} Y_{t-1} - Y_{t-2} = Y_t$$

It should be noted in this context that deterministic does not mean that Y_t is non-random, but rather completely determined in regard to this projection on infinite past.

Theorem (Wold's Decomposition Theorem)

Any zero-mean nondeterministic covariance-stationary process $\{Y_t\}$ *can be decomposed as*

$$Y_t = \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j} + \kappa_t$$

where

- $\beta_0 = 1$ *and* $\sum_{j=0}^{\infty} \beta_j^2 < \infty$,
- $\epsilon_t \sim WN(0, \sigma^2)$,
- $\{\beta_j\}$ *and* $\{\epsilon_t\}$ *are unique,*
- $\{\kappa_t\}$ *is deterministic (as defined previously)*
- ϵ_t *is the limit of linear combinations of* Y_s, $s \leq t$,
- $\mathbb{E}[\kappa_t \epsilon_s] = 0, \quad \forall t, s.$

The Wold representation is the **unique** linear representation where the innovations are linear forecast errors.

Definition

A zero-mean non-deterministic covariance-stationary process $\{Y_t\}$ *is called purely non-deterministic (or regular) if* $\kappa_t = 0$

This means that if the process $\{Y_t\}$ is purely non-deterministic then by Wold Decomposition Theorem:

$$Y_t = \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j}$$

- $\beta_0 = 1$ and $\sum_{j=0}^{\infty} \beta_j^2 < \infty$,
- $\epsilon_t \sim WN(0, \sigma^2)$,
- ϵ_t is the limit of linear combinations of Y_s, $s \leq t$,

Remark

The Wold Decomposition Theorem plays a critical role in time series analysis:

It implies that the dynamic of any purely non-deterministic covariance-stationary process can be arbitrarily well approximated by an ARMA process.

Furthermore, any purely non-deterministic covariance-stationary process can be written as a linear combination of lagged values of a White Noise process ($MA(\infty)$ representation)

Hence, we will focus largely for the rest of the course on linear Autoregressive Moving Average (ARMA) models.

This theorem can be considered as an existence theorem: any stationary process has this seemingly special representation.

All linear stationary models can be written in a generic model representation due to Wolds Decomposition!

Tells us that at least in the class of linear stationary processes the core ingredient to represent or construct any linear model can be achieved by combinations of lagged values from a WN process...

$\Rightarrow$ WN is the building block for constructing a range of different TS models.

It will be useful to calculate the first and second order moments of Y_t (assume w.l.o.g. $\kappa_t = \mu$).

Corollary (Wold Decomposition: ACVF & ACF)

$$E(Y_t) = \mu \tag{0.3}$$

$$\gamma(k) = \sigma^2 \sum_{j=0}^{\infty} \beta_j \beta_{j+k} \tag{0.4}$$

$$\rho(k) = \frac{\sum_{j=0}^{\infty} \beta_j \beta_{j+k}}{\sum_{j=0}^{\infty} \beta_j^2}. \tag{0.5}$$

Proof For (0.3):

$$Y_t = \mu + \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j}, \quad \Rightarrow \quad \mathbb{E}(Y_t) = \mu + \sum_{j=0}^{\infty} \beta_j \cancelto{0}{\mathbb{E}(\epsilon_{t-j})} = 0 \Rightarrow \text{Eqn 0.3}$$

Proof For (0.4):

$$\gamma(k) = \text{cov}\left(\mu + \sum_{j=0}^{\infty}\beta_j\epsilon_{t-j}, \mu + \sum_{i=0}^{\infty}\beta_i\epsilon_{t+k-i}\right)$$

$$= \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\beta_j\beta_i\,\text{cov}(\epsilon_{t-j}, \epsilon_{t+k-i}) \qquad \text{(from cov property 7)}$$

Now, $\{\epsilon_t\}$ has uncorrelated terms, i.e., recall:

$$\text{cov}(\epsilon_t, \epsilon_{t+k}) = \delta_{0,k}\sigma^2$$

Hence,

$$\text{cov}(\epsilon_{t-j}, \epsilon_{t+k-i}) = \delta_{t-j,t+k-i}\sigma^2 = \delta_{-j,k-i}\sigma^2 = \delta_{i,j+k}\sigma^2$$

$$\therefore \gamma(k) = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\beta_j\beta_i\sigma^2\delta_{i,j+k} = \sigma^2\sum_{j=0}^{\infty}\beta_j\beta_{j+k} \Rightarrow \text{Eqn 0.4}$$

For (0.5): $\rho(k) = \gamma(k)/\gamma(0)$. I.e.

$$\gamma(0) = \gamma(k)|_{k=0} = \left(\sigma^2 \sum_{j=0}^{\infty} \beta_j \beta_{j+k} \right) \Bigg|_{k=0}$$

$$= \sigma^2 \sum_{j=0}^{\infty} \beta_j^2$$

$$\therefore \rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\cancel{\sigma^2} \sum_{j=0}^{\infty} \beta_j \beta_{j+k}}{\cancel{\sigma^2} \sum_{j=0}^{\infty} \beta_j^2} \Rightarrow Eqn0.5 \quad \blacksquare$$

Remark

Given a (non-stationary) time series $\{\tilde{Y}_t\}$ with non-zero constant mean μ, simply consider $\{Y_t\} = \{\tilde{Y}_t - \mu\}$ with zero mean. I.e. for convenience we will, for now, assume $\mathbb{E}(Y_t) = 0$ (without loss of generality).