

Time Series Econometrics

Matthias R. Fengler

University of St. Gallen

Spring 2020

matthias.fengler@unisg.ch



Universität St.Gallen

Chapter 1:

Introduction to TSE



Why study time series econometrics?

TS data are pervasive in economics and finance. Essentially, all macro and finance data come along as TS data:

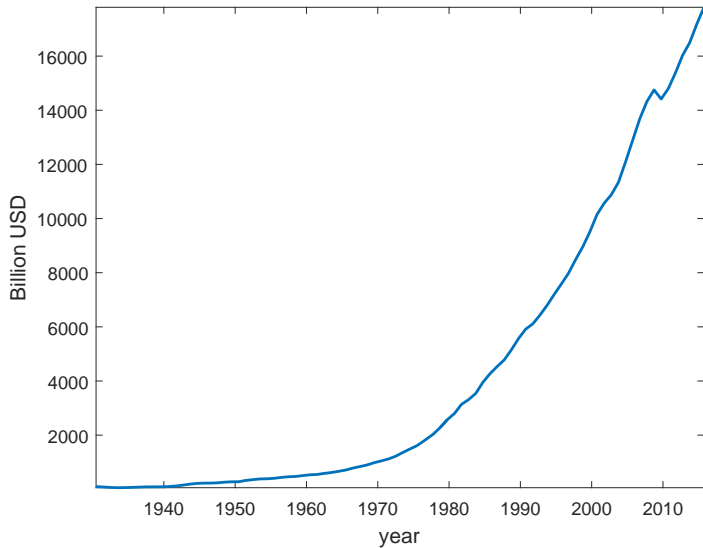
GDP, CPI, unemployment, housing, stock prices, but increasingly also micro economic data ...

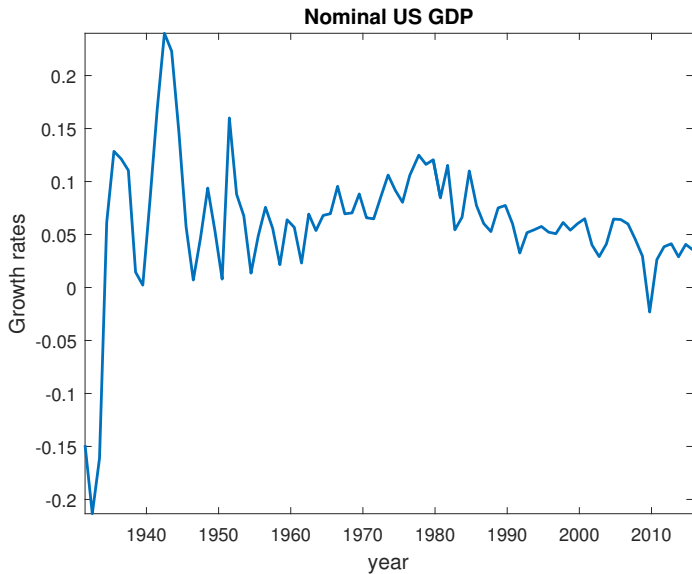
TS data are of paramount importance in biology, physics, medicine, climate science, ...

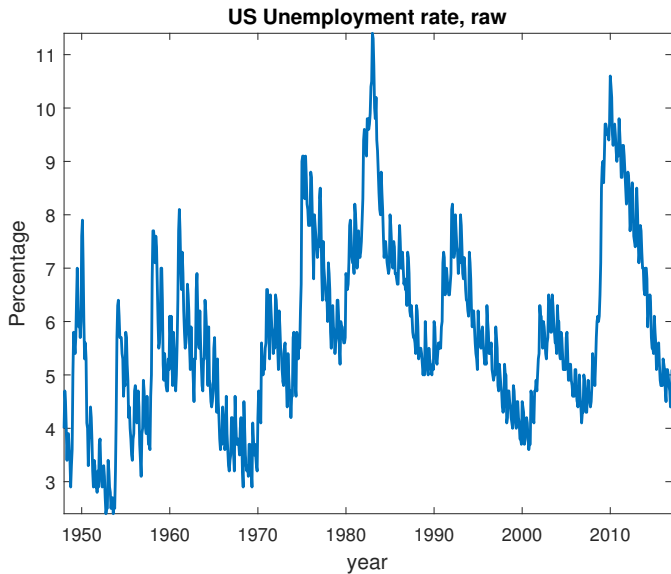
TS data have specific properties that need to be taken care for.



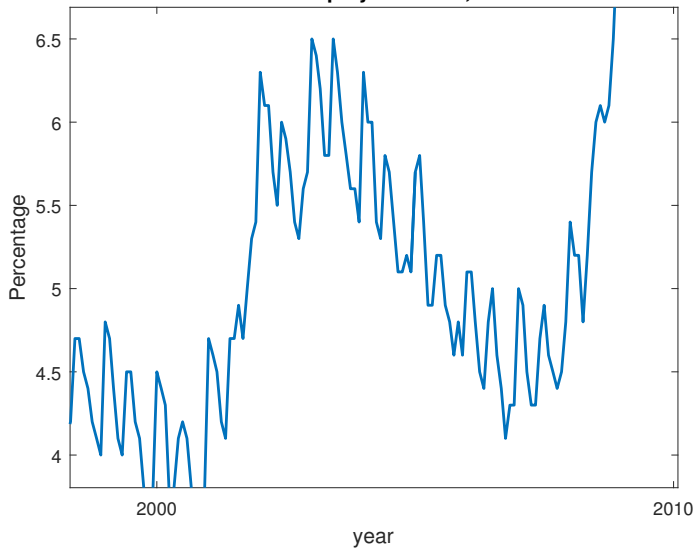
Nominal US GDP



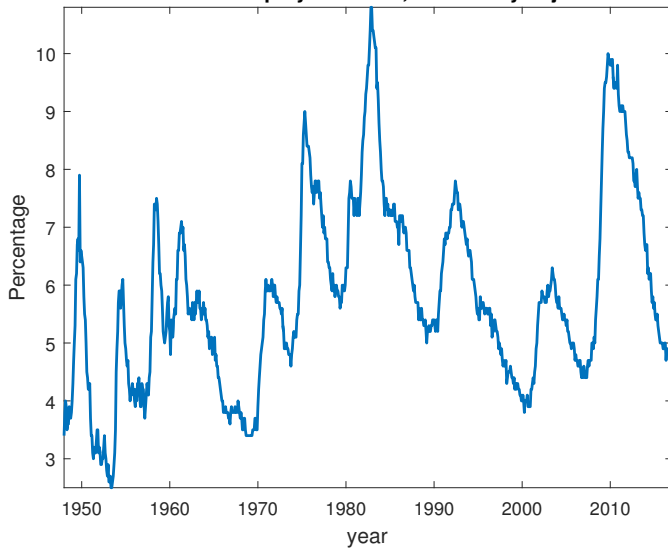




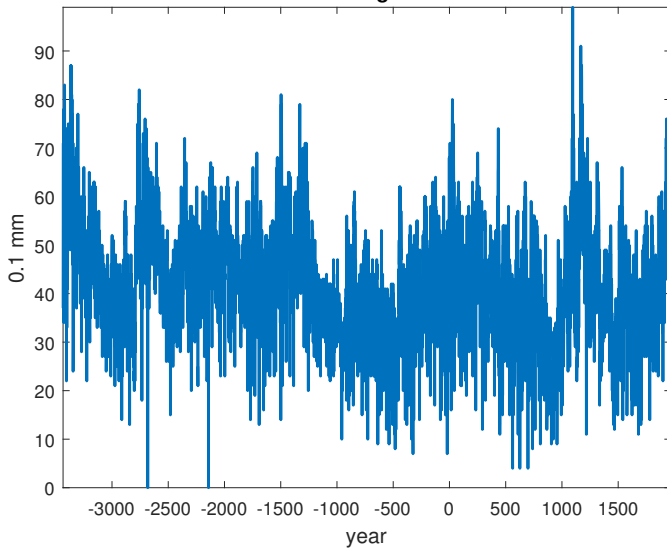
US Unemployment rate, raw

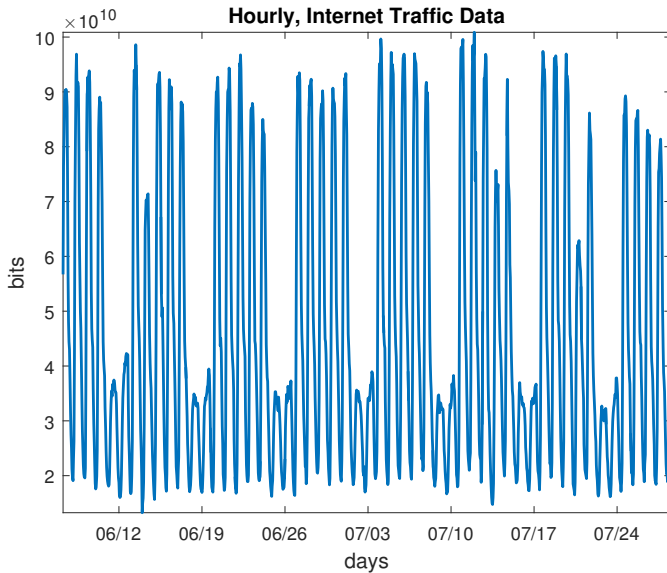


US Unemployment rate, seasonally adj.



Tree-Ring-Data





Properties of time series data:

1. the iid. assumption is problematic
 - ▶ data are serially dependent
 - ▶ data are not identically distributed
2. trends: deterministic or stochastic
3. patterns of seasonality: deterministic or stochastic



The dependence of ts data is both friend and foe:

It is our friend because

- it makes prediction possible;
- sometimes helps in estimation (“super consistency”);

but otherwise makes life more difficult, e.g., for the asymptotic theory.



Aims of times series analysis:

- ▣ descriptive analysis of intertemporal economic dynamics
- ▣ prediction
- ▣ testing of economic hypotheses in a times series context, such as the purchasing power parity, interest rate parities, or asset pricing
- ▣ simulation of economic processes, e.g., for studying counterfactuals of policy measures, for risk management and derivatives pricing ...



Elementary framework and basic techniques

Plots suggest

$$X_t = m_t + s_t + u_t$$

where m_t is a (deterministic) trend component, s_t is a (deterministic) seasonality component, u_t an irregular component with $E[u_t] = 0$ and $\text{Var}[u_t] = \text{const.}$ (i.e., covariance stationary).

Sometimes the multiplicative model

$$X_t = M_t S_t U_t$$

is more appropriate. Applying logs (if all elements are positive) reduces to the additive model.



If deterministic, typical trend specifications are

- polynomials:

$$m_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$$

- many others, e.g., logistic, exponentials...



For deterministic seasonality with periodicity S , one requires

$$s_t = s_{t-S}.$$

Achieved, e.g., by

- season dummies, i.e.,

$$s_t = \beta_1 d_{1,t} + \beta_2 d_{2,t} + \dots + \beta_S d_{S,t}$$

with seasonal dummies $d_{s,t}$, $s = 1, \dots, S$;

- trigonometric functions

$$s_t = \beta_1 \cos(\omega_1 t) + \gamma_1 \sin(\omega_1 t)$$

with $\omega_1 = 2\pi/S$; add further frequencies if necessary.



Can estimate deterministic components with ease; in particular

- use OLS or NLS, but may not be efficient
- standard errors need to be tailored to u_t
 - ▶ if iid., can use set-up of the “classical model”;
 - ▶ if heteroskedastic, use robust White estimator;
 - ▶ if heteroskedastic and serially correlated, use robust Newey-West estimator (to be discussed);
 - ▶ conduct additional model diagnostics (to be discussed).

One then continues the analysis on the estimated residuals

$$\hat{u}_t = X_t - \hat{m}_t - \hat{s}_t$$



What if trends and seasonal patterns are stochastic? E.g., could set $m_t = X_{t-1}$ leading to (with $s_t = 0$)

$$X_t = X_{t-1} + u_t ,$$

which yields the random walk model.

Or, set

$$s_t = s_{t-S} + \varepsilon_t$$

with $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$ and $m_t = 0$; yields a “seasonal random walk”.



In the latter cases, differencing is appropriate, i.e., compute

$$X_t - X_{t-1} = u_t .$$

or, using seasonal differencing,

$$X_t - X_{t-S} = s_t - s_{t-S} + u_t - u_{t-S} = \varepsilon_t + u_t - u_{t-S} .$$

Depending on true DGP, this yields a well-behaved error sequence that can be modeled further.

Sometimes both simple differencing and seasonal differencing may be necessary.



Data transformations

Data transformations are widely applied in TSE

A famous example is the log-transformation. As seen, this may be due to the multiplicative nature of trends and seasonality patterns.

Moreover, taking the logs often stabilizes the variance and/or makes the data look more “normal”. Or, it conforms with ideas of an exponential random walk

$$S_{t+1} = S_t e^{\varepsilon_{t+1}},$$

which is a standard asset pricing model, $e \approx 2.71828\dots$ is Euler's number



The Box-Cox transformation generalizes the log-transform:

$$u^{(\lambda)} = \begin{cases} \frac{u^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log u & \lambda = 0 \end{cases} .$$

If the data are confined to, say, the interval $[0, 1]$ the logistic transformation can be useful:

$$u^* = \log[u/(1 - u)] ,$$

after which u^* ranges from minus to plus infinity.



Filtering

Complementary to the component modeling is filtering.

Filtering

- extracts or eliminates certain features of the data
- transforms one times series into another one

Consider the moving average filter

$$Y_t = \frac{X_{t-1} + X_t + X_{t+1}}{3} \quad t = 2, 3, \dots, T - 1.$$



Example

Consider the filter

$$Y_t = \frac{1}{8}X_{t-2} + \frac{1}{4}X_{t-1} + \frac{1}{4}X_t + \frac{1}{4}X_{t+1} + \frac{1}{8}X_{t+2}$$

This filter can eliminate certain seasonal variations. Suppose, e.g., that X_t is a quarterly series generated by

$$X_t = \mu_t + u_t$$

where $\mu_t = \mu_j$, $j = 1, \dots, 4$ is constant, but differs from quarter to quarter, and u_t iid.



Applying the filter to X yields

$$Y_t = \frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4) + \frac{1}{8}u_{t-2} + \frac{1}{4}u_{t-1} + \frac{1}{4}u_t + \frac{1}{4}u_{t+1} + \frac{1}{8}u_{t+2}$$

Y has the constant mean $\frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4)$.

An issue with filtering is the boundary treatment because boundary values are lost. Boundary adjustments may be necessary, however, resulting into nonlinear, potentially complicated filters.



In general, a filter is a sequence of constants $\{w_i\}_{i=-\infty}^{\infty}$, such that we obtain a new process Y from some process X by computing

$$Y_t = \sum_{i=-\infty}^{\infty} w_i X_{t-i} .$$

This procedure is called filtering.

Example

For the first example, we have $w_i = \frac{1}{3}$, $i = -1, 0, 1$, and $w_i = 0$ otherwise.



A widely used filter is the **Hodrick-Prescott (HP) filter**.

The HP filter is defined by specifying the trend of a time series as a solver of the minimization problem

$$\min_{\{\mu_t\}_{t=1}^T} \sum_{t=1}^T [(x_t - \mu_t)^2 + \lambda \{(\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1})\}^2],$$

where $\lambda > 0$ is a smoothing parameter. Larger λ implies a smoother trend μ_t . The minimization problem can be uniquely solved.

Hodrick & Prescott (1997) suggest $\lambda = 100, 1600, 14400$ for yearly, quarterly, and monthly data, respectively.

The HP filter was criticed recently by [Hamilton \(2018\)](#).



Example

Is filtering an academic gimmick? No!!

The Swiss debt brake is institutionalized as a constitutional article (Art. 126). Its aim is not to directly reduce debt, but to keep debt constant over the business cycle. If the economy grows relative debt reduces and increases in bad times.

Idea: tie the maximal amount of fiscal expenditures A to estimated income E

$$A = k \times E$$

where k is a business cycle factor.



The business cycle factor is defined

$$k = \frac{\text{trend GDP}}{\text{current GDP}}.$$

We have $k > 1$ in times of crisis (higher deficit allowed) and $k < 1$ in boom times (must save).

How to determine the trend GDP? A **modified HP filter** is used!

Two official documents for download:

Botschaft zur Schuldenbremse vom 5. Juli 2000

Eine Neubewertung der Schuldenbremse, 2004



Summary

The objective of detrending, deseasonalizing, and filtering is

- to obtain a sequence of random variables
- without any apparent trends and seasonalities, a stationary process (to be defined).

One then studies the remaining dependence:

- if there is no dependence, we can stop;
- if there is a significant amount of dependence, we can start modeling the data with a more complex stationary time series model – the topics of the next chapters.



Software

There are many softwares that handle time series estimation and inference. A classical commercial time series package is Eviews. Also Matlab has a time series toolbox.

Free software can be found in the R project `tseries`.

For Matlab users, Kevin Sheppard's MFE toolbox is useful.

Also Python provides time series tools.



Textbooks

Recommendation:

- Brockwell, Peter J. and Richard A. Davis (2002): Introduction to Time Series and Forecasting, 2nd Edition, Springer.
- Enders, Walter (2010): Applied Econometric Time Series, 3rd Edition, Wiley.

PhD level book, but not too technical, very useful:

- Hamilton, James D. (1994): Time Series Analysis, Princeton.



Chapter 2:

Fundamental concepts of stochastic processes



Stochastic processes

Denote by (Ω, \mathcal{F}, P) a stochastic basis. A real-valued **stochastic process** $X = \{X_t, t \in \mathcal{T}\}$, is a family of random variables (rv) on $\Omega \times \mathcal{T}$ taking values in \mathbb{R} . Rv's are functions of the form

$$X_t(\omega) : \Omega \times \mathcal{T} \rightarrow \mathbb{R}$$

If $\mathcal{T} = \mathbb{Z}$, X is a discrete time process.

Can also have $\mathcal{T} = \mathbb{R}$, then X is a continuous time process.



I use $X = \{X_t, t \in \mathcal{T}\}$ to denote the entire process and X_t to denote an element of X occurring at time t .



For a fixed t , say $t = \bar{t}$, the set of values $\{X_{\bar{t}}(\omega), \omega \in \Omega\}$ represents the **set of possible states** at time \bar{t} :
“cross-sectional or ensemble view”

For a fixed ω , say $\bar{\omega}$, $\{X_t(\bar{\omega})\}_{t \in \mathcal{T}}$ is a **trajectory** or **path** of the process. It represents **one possible** evolution of the process.



Characterizing stochastic processes

We treat a stochastic process as a vector of rv. Hence we can characterize it by its **unconditional** distribution(s).

For some times $t_1, \dots, t_n \in \mathbb{N}$, the unconditional cdf of X is

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = P(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n)$$

If for any $t_1, \dots, t_n \in \mathbb{N}$, the cdf $F_{t_1, \dots, t_n}(x_1, \dots, x_n)$ is known, the process is **uniquely** determined.

If $F_{t_1, \dots, t_n}(x_1, \dots, x_n)$, for any $t_1, \dots, t_n \in \mathbb{N}$, is a multivariate normal, X is called a **Gaussian process**.



Stochastic processes can also be described by their **conditional** distributions.

The conditional cdf for X for $t_1, \dots, t_n \in \mathbb{N}$ and $t_1 < t_2 < \dots < t_n$ is defined as

$$F_{t_n|t_{n-1}, \dots, t_1}(x_n|x_{n-1}, \dots, x_1) = \\ P(X_{t_n} \leq x_n | X_{t_{n-1}} = x_{n-1}, \dots, X_{t_1} = x_1)$$

⚠ $F_{t_n|t_{n-1}, \dots, t_1}(x_n|x_{n-1}, \dots, x_1)$ being normal does not imply that X is Gaussian.



In many cases, $F_{t_n|t_{n-1},\dots,t_1}$ will not depend on the entire past, but only on a finite subset of the past $p < n$ observations:

$$F_{t_n|t_{n-1},\dots,t_1}(x_n|x_{n-1},\dots,x_1) = F_{t_n|t_{n-1},\dots,t_{n-p}}(x_n|x_{n-1},\dots,x_{n-p})$$

A process with this property called a **Markov process of order p** .

The Markov property is essential to deriving the ML function for parameter estimation.



The conditional cdf or density gives rise to the notion of the **conditional expectation**. Denote by

$$f_{t_n|t_{n-1}}(x_n|x_{n-1})$$

the conditional density of X_{t_n} given x_{n-1} . The conditional expectation of some function g of X_{t_n} is

$$E[g(X_{t_n})|X_{t_{n-1}} = x_{n-1}] = \int_{-\infty}^{\infty} g(x_n) f_{t_n|t_{n-1}}(x_n|x_{n-1}) dx_n$$

assuming that X_t is continuous.



The conditional expectation is a random variable. It depends on the conditioning variables (or the conditioning set). If X_{t_n} is independent of $X_{t_{n-1}}$ it holds

$$E[g(X_{t_n})|X_{t_{n-1}}] = E[g(X_{t_n})]$$

Because conditional expectation is a random variable we can take expectation of it. By the **law of iterated expectations (LIE)**, it holds

$$E[E[g(X_{t_n})|X_{t_{n-1}}]] = E[g(X_{t_n})]$$



You will know a conditional expectation of the form

$$E[X|Y].$$

In TSE it is common practice to write the conditional expectation w.r.t. an increasing family of events called **information set** or **filtration**, denoted by $\{\mathcal{F}_t, t \in \mathcal{T}\}$.

One writes

$$E[X_{t+1}|\mathcal{F}_t].$$



An information set is increasing if

$$s \leq t \Rightarrow \mathcal{F}_s \subseteq \mathcal{F}_t$$

One must specify what the information set contains. Given a process X , one often associates with \mathcal{F}_t the filtration generated by X itself (**natural filtration**).

A process X is called **adapted to** $\{\mathcal{F}_t\}$, if “knowledge” of \mathcal{F}_t is sufficient to determine the outcome of X_t for any t ; more technically speaking: X_t is \mathcal{F}_t -**measurable**.

X is always adapted to the natural filtration.



Law of iterated expectations (tower law)

Iterated conditioning comes along in various guises:

$$E\{E[X|Y]\} = E[X]$$

$$E\{E[X_{t+1}|\mathcal{F}_t]\} = E[X_{t+1}]$$

$$E\{E[X_{t+2}|\mathcal{F}_{t+1}|\mathcal{F}_t]\} = E[X_{t+2}|\mathcal{F}_t] \quad \text{with} \quad \mathcal{F}_t \subset \mathcal{F}_{t+1}$$

All these cases are valid applications of iterated conditioning!



Mean and autocovariance function

The mean of a process is X is defined as

$$\mu_t = E[X_t]$$

and the autocovariance

$$\gamma_{t,s} = E[(X_t - \mu_t)(X_s - \mu_s)]$$

For $s = t$, we get the variance

$$\gamma_{t,t} = \text{Var}[X_t] .$$



Ensemble averages



The unconditional mean and variance of X_t have to be understood as **ensemble averages**, i.e., for a fixed t and different ω . You can imagine this as follows:

Suppose a computer generates a sequence of some process

$$\{x_t\}_{t=-\infty}^{\infty} = \{\dots, x_{-1}, x_0, x_1, \dots\}$$

this sequence is one single realization of the process (“one ω ”).



Produce independently N such realizations of X :

$$\{x_t^{(1)}\}_{t=-\infty}^{\infty}, \{x_t^{(2)}\}_{t=-\infty}^{\infty}, \{x_t^{(3)}\}_{t=-\infty}^{\infty}, \dots$$

and single out the observations with a fixed date t :

$$\{x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, \dots, x_t^{(N)}\}$$

These are N independent realizations of X_t with cdf F_t . The unconditional mean of X_t is

$$E[X_t] = \int_{\mathbb{R}} x \, dF_t(x)$$

and it can be seen as the probability limit of the ensemble average:

$$E[X_t] = p\text{-}\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N X_t^{(i)}$$



Stationarity

A process X is called **covariance** or **weakly stationary** if

1. $\mu_t = \mu$ and
2. $\gamma_{t,t} < \infty \quad \forall t$
3. $\gamma_{r,s} = \gamma_{r+t,s+t} \quad \forall r, s, t$

Setting $t = -s$ in 3 shows $\gamma_{r,s} = \gamma_{r-s,0}$. So for a stationary process we can redefine

$$\gamma_h \equiv \gamma_{h,0} ,$$

wherer h is the “lag” or “lead” order.

Mean and (autoco)variance(s) of a weakly stationary process do not depend on time.



A process X is called **strictly stationary** if for any t_1, \dots, t_n and for all $n, s \in \mathbb{Z}$

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = F_{t_1+s, \dots, t_n+s}(x_1, \dots, x_n)$$

The multivariate distributions do not depend on time.

It is clear that if the first two moments exist,

strict stationarity \Rightarrow weak stationarity.

If not, a process may be strictly stationary without being weakly stationary.



Autocorrelation function (acf)

The autocorrelation function is defined by

$$\rho_h = \frac{\gamma_h}{\gamma_0} .$$

For autocorrelation to be defined X must be covariance stationary.

A plot of ρ_h as function of h is called **correlogram**. It is an important device to study the linear dependence structures of X .



Properties of acf of a stationary process

If X is stationary, the acf has the following properties:

1. $\gamma_0 > 0$
2. $|\gamma_h| \leq \gamma_0 \quad \forall h$
3. γ_h is even, i.e., $\forall h$

$$\gamma_h = \gamma_{(-h)} .$$

The autocorrelation function has the same properties, and additionally $\rho_0 = 1$.



Ergodicity

- In TSA, we aim to make inference about the properties of the process we observe, e.g., about the unconditional mean.
- Ideally, we'd wish to estimate the mean as an *ensemble average*, but this is not possible: we usually observe one single realization of a time series; the best we can do is computing the *time average*!
- This raises an important question: Under which circumstances can we expect that

$$\bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t \rightarrow E[X_t] = \mu$$

and in which sense does the convergence take place?



- at an intuitive level, for this to happen, the sample which we observe must be “representative” for the stochastic process
- it may seem that stationarity does the job, but this is wrong! Stationarity only ensures that it does not matter which “section” of the sample path we analyze. It does not ensure that the observed sample is representative for any possible realization of the process
- beyond stationarity, we must therefore ask for an additional property: **ergodicity**



A covariance stationary process X is called **ergodic for the mean** if

$$E[(\bar{X}_T - \mu)^2] \rightarrow 0$$

as T tends to infinity (convergence in mean square) where $E[X_t] = \mu$.

Because convergence in mean square implies convergence in probability, the time average is then a consistent estimator for the mean.



What conditions do we need to ask for?

Let X be a covariance stationary process. Then \bar{X}_T converges to μ in mean square sense, if and only if

$$\lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=0}^H \gamma_h = 0 .$$

Thus, a necessary and sufficient condition for ergodicity for the mean is *asymptotic average uncorrelatedness* in this sense.

Intuitively, for ergodicity to happen, outcomes of the process that are far apart in time must become less and less dependent.



Ergodicity – further remarks

- the ergodicity property for covariance stationary processes I present (ergodicity for the mean) is widespread in the econometrics/economics literature
- often the literature cites stronger conditions than the one I discussed: $\sum_{h=0}^{\infty} |\gamma_h| < \infty$
- sometimes, we require ergodicity to hold for higher order moments, e.g., ergodicity for second-order moments
- in statistics, a different notion of ergodicity is employed, referring to strictly stationary processes and attached to a different mode of convergence (almost sure convergence)

See [Hassler \(2017\)](#) for a nice overview.



Stationarity and ergodicity

A covariance stationary process may not be ergodic for the mean.

Example

Let $X_t = \eta + \varepsilon_t$ where $\eta \sim \mathcal{N}(0, 1)$ and $\varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$.

Direct verification shows $p\text{-lim } \frac{1}{T} \sum_{t=1}^T X_t = \eta \neq E[X_t] = 0$.

Convince yourself that the condition of asymptotic average uncorrelatedness is violated!



In process definition, η is drawn only once (not every t).



Basic processes

White noise (WN).

X is called white noise if

1. X is covariance stationary,
2. uncorrelated, i.e., $\rho_h = 0, h > 0$, and
3. has mean zero, i.e., $\mu = 0$.

X is called **strict white noise** if the X_t are additionally iid.

Example

The process

$$X_t = \varepsilon_t$$

with $\varepsilon_t \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ is a Gaussian strict white noise.



Let X be a white noise with $E[X_t^2] = \gamma_0 = \sigma^2$.

Then, the acf of X is

$$\gamma_h = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h > 0 \end{cases}$$



Martingale difference sequence (MDS).

The MDS is always defined in relation to an (increasing) information set $\{\mathcal{F}_t\}$.

X is called an MDS relative to $\{\mathcal{F}_t\}$ if

1. X is adapted to $\{\mathcal{F}_t\}$
2. $E[X_{t+1}|\mathcal{F}_t] = 0$ (a.s.)

As immediate consequences we have

$$E[X_t] = 0 \quad \text{and} \quad E[X_t X_{t+j}] = 0, \text{ if } j \neq 0$$



- An MDS is weaker than an iid. process, because there could be dependence in higher order moments of X_t .
- An MDS weaker than WN in that it is defined only via the first conditional moment.
- An MDS with finite variance is WN.
- Yet an MDS is stronger than WN in that an MDS cannot be forecast by both any linear and nonlinear combinations of past values of X_t , whereas a WN cannot be forecast by linear combinations of past values of X_t .

Useful LLNs and CLTs have been established for MDS.



Example

Let $\mathcal{F}_t = \{\varepsilon_t, \varepsilon_{t-1}, \dots\}$ and define the process

$$X_t = \varepsilon_t + \varepsilon_{t-1}\varepsilon_{t-2},$$

where ε is strict WN.

X is (weak) WN, but not an MDS because

$$E[X_t | \mathcal{F}_{t-1}] = \varepsilon_{t-1}\varepsilon_{t-2}$$



Another important stochastic process is the **random walk**.

X follows a random walk if it has the representation

$$X_t = c + X_{t-1} + \varepsilon_t ,$$

for some constant c and white noise ε with variance σ_ε .

If $c \neq 0$ we speak of a **random walk with drift**.



Assume $X_0 = 0$, $c = 0$, and let ε be iid.

Repeated substitution shows

$$X_t = \varepsilon_t + \varepsilon_{t-1} + \dots + \varepsilon_1 .$$

Thus

$$\text{Var}[X_t] = \text{Var} \left[\sum_{i=1}^t \varepsilon_i \right] = \sum_{i=1}^t \text{Var}[\varepsilon_i] = t\sigma_\varepsilon^2$$

and

$$\text{E}[X_t] = 0 .$$



Finally, we see

$$\gamma_{t,s} = \text{Cov}[X_t, X_s] = \text{Cov}\left[\sum_{i=1}^t \varepsilon_i, \sum_{i=1}^s \varepsilon_i\right] = \min(t, s) \sigma_\varepsilon^2$$

Clearly, the random walk is **not** stationary.



Linear processes

A process X is said to be **linear** if it has the representation

$$X_t = \sum_{i=-\infty}^{\infty} \psi_i \varepsilon_{t-i} ,$$

where ε is WN with $\text{Var}[\varepsilon_t] = \sigma_\varepsilon^2$ and $\{\psi_i\}$ is a sequence of constants such that

$$\sum_{i=-\infty}^{\infty} |\psi_i| < \infty \quad (\text{absolut summability})$$

Note that $\{\psi_i\}$ is a filter.

Includes autoregressive moving-average models (see later).



A linear process X is covariance stationary with

•

$$\mu = E[X_t] = 0$$

•

$$\gamma_0 = \text{Var}[X_t] = \sigma_\varepsilon^2 \sum_{i=-\infty}^{\infty} \psi_i^2;$$

• moreover,

$$\gamma_h = \sigma_\varepsilon^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h}$$



Remark 1:

$$\sum_{i=-\infty}^{\infty} |\psi_i| < \infty \Rightarrow \sum_{i=-\infty}^{\infty} \psi_i^2 < \infty .$$

Property $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$ is referred to as **stability** (stable filter).

Remark 2:

One calls the operation of taking an “average” of another series in this fashion **application of a filter**.

Many procedures in time series analysis are applications of filters, such as detrending, computing moving averages, exponential smoothing.



Important result on filters:

If X is stationary with acf γ and a process Y is obtained via

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_t$$

where ψ_j are absolutely summable, i.e., $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then Y is stationary and has acf

$$\bar{\gamma}_h = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_{h-j+k} .$$



The lag operator

We define the **lag (or backshift) operator** as

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

This extends to powers such that

$$L^2X_t = L(LX_t) = LX_{t-1} = X_{t-2}$$

and, generally,

$$L^kX_t = X_{t-k}$$



It will be useful to form polynomials of lag operators:

$$a(L) = a_0 + a_1L + a_2L^2 + a_3L^3 + \dots + a_pL^p$$

and

$$a(L)X_t = a_0X_t + a_1X_{t-1} + a_2X_{t-2} + a_3X_{t-3} + \dots + a_pX_{t-p}$$

Lag polynomials can be manipulated like polynomials of real variables:

e.g., multiplication is commutative:

$$a(L)b(L) = b(L)a(L)$$



Sometimes, we can invert them. Suppose $a(L) = 1 - aL$. We define $(1 - aL)^{-1}$ by the equality

$$(1 - aL)(1 - aL)^{-1} \equiv 1.$$

Well defined iff $|a| < 1$. Then

$$(1 - aL)^{-1} = \sum_{i=0}^{\infty} a^i L^i.$$



Checking the inequality shows

$$\begin{aligned}(1 - aL) \sum_{i=0}^{\infty} a^i L^i &= \sum_{i=0}^{\infty} a^i L^i - \sum_{i=1}^{\infty} a^i L^i \\ &= a^0 L^0 = 1\end{aligned}$$

The condition $|a| < 1$ is needed to ensure that the operator remains valid in the sense that it maps stationary processes on stationary processes, i.e., we want that if X is stationary then also $(1 - aL)^{-1}X$ is stationary.



The difference operator

The **difference operator** is defined as

$$\Delta X_t = (1 - L)X_t = X_t - X_{t-1}.$$

It can be applied iteratively:

$$\begin{aligned}\Delta^2 X_t &= \Delta(\Delta X_t) \\ &= \Delta X_t - \Delta X_{t-1} \\ &= X_t - 2X_{t-1} + X_{t-2}.\end{aligned}$$



Basic asymptotic theory: LLNs and CLTs

Suppose that X is a stationary process, and we observe a sample $\{x_t\}_{t=1}^T$. Can we estimate its mean μ ?

A natural estimator is the sample mean:

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

It is an unbiased estimator, because

$$E[\bar{X}_T] = T^{-1} \sum_{t=1}^T E[X_t] = \mu.$$



Moreover, the MSE is

$$\begin{aligned}
 E[(\bar{X}_T - \mu)^2] &= \text{Var}[\bar{X}_T] \\
 &= T^{-2} \sum_{i=1}^T \sum_{j=1}^T \text{Cov}[X_i, X_j] \\
 &= T^{-2} [T\gamma_0 + (T-1)\gamma_1 + (T-1)\gamma_{-1} \\
 &\quad + (T-2)\gamma_2 + (T-2)\gamma_{-2} \dots] \\
 &= T^{-1} \sum_{h=-(T-1)}^{T-1} \left(1 - \frac{|h|}{T}\right) \gamma_h \\
 &\leq T^{-1} \sum_{h=-(T-1)}^{T-1} |\gamma_h|
 \end{aligned}$$

Thus, if $\sum_{-\infty}^{\infty} |\gamma_h| < \infty$ as $h \rightarrow \infty$, then $E[(\bar{X}_T - \mu)^2] \rightarrow 0$ and \bar{X}_T converges in mse sense and therefore is consistent.



The **law of large numbers** for **covariance stationary processes**:

If X is a cov.-stationary process with mean μ and autocovariance function γ that is absolutely summable, then

$$\bar{X}_T \rightarrow \mu \text{ in mse sense and hence } p\text{-}\lim_{T \rightarrow \infty} \bar{X}_T = \mu$$



A CLT for covariance-stationary processes

Let X be cov.-stationary such that

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

where ε is **strict** WN and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$.

Then

$$\sqrt{T}(\bar{X}_T - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}),$$

where $\mathcal{I} = \sum_{h=-\infty}^{\infty} \gamma_h$ (**long-run variance**).



The asymptotic variance is different from the iid. case. In particular, it can be larger or smaller, depending on the autocorrelation structure of X .



A LLN for MDS:

Let X be a strictly stationary MDS with $E|X_t| < \infty$. Then

$$p\text{-}\lim_{T \rightarrow \infty} \bar{X}_T = 0$$

See Lütkepohl (2007), p.690.



A CLT for MDS:

Let X be a strictly stationary, second-order ergodic MDS with $E[X_t^2] = \sigma^2 < \infty$. Then

$$\sqrt{T}\bar{X}_T \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$



Summary

- A stochastic process is a sequence of rv ordered in time.
- The unconditional cdf characterizes the process completely .
- A process is weakly stationary if its mean and autocorrelation function are not functions of time.
- A process strictly stationary if its distribution does not depend on the time subscript.
- Important processes: WN, MDS, the random walk, the linear process.
- Do not confuse stationarity and ergodicity.



Chapter 3:

ARMA processes



ARMA processes: definition and notation

X is called an $\text{ARMA}(p, q)$ process (an autoregressive moving average process of order p and q) if X is stationary and if, for every t ,

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

where ε is white noise, and $\phi_i, i = 1, \dots, p$ and $\theta_i, i = 1, \dots, q$ are real parameters.

c is a constant. W.l.o.g., I set now $c = 0$.



Using lag polynomial notation we write more compactly

$$\phi(L)X_t = \theta(L)\varepsilon_t,$$

where

$$\begin{aligned}\phi(L) &= 1 - \phi_1 L - \phi_2 L^2 \dots - \phi_p L^p \\ \text{and} \quad \theta(L) &= 1 + \theta_1 L + \theta_2 L^2 \dots + \theta_q L^q\end{aligned}$$

We now study conditions for stationary solutions, prediction, estimation, and inference for ARMA processes.



MA processes

Setting $\phi(L) \equiv 1$ yields the MA(q) process

$$X_t = \theta(L)\varepsilon_t .$$

Consider the MA(1):

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} .$$

This is the unique solution to this “difference equation”. Moreover,

$$E[X_t] = 0$$

$$E[X_t^2] = (1 + \theta_1^2)\sigma_\varepsilon^2$$

$$E[X_t X_{t+1}] = \theta_1 \sigma_\varepsilon^2$$

$$E[X_t X_{t+h}] = 0 \quad \text{for } |h| > 1$$

Hence the MA is stationary for any value of θ_1 .



Clearly, it holds that

$$\lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=0}^{\infty} \gamma_h = \lim_{H \rightarrow \infty} \frac{1}{H} \left\{ (1 + \theta_1^2) \sigma_\varepsilon^2 + \theta_1 \sigma_\varepsilon^2 + 0 + 0 + \dots \right\} = 0$$

so asymptotic average uncorrelatedness holds and the process is ergodic for the mean.



The general MA(q)

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

is not more difficult. We have

$$E[X_t] = 0$$

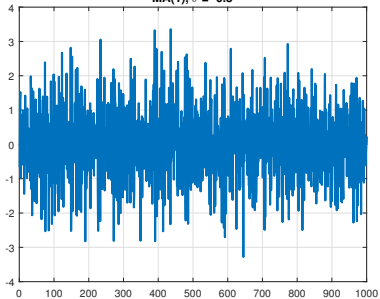
$$E[X_t X_{t+h}] = \sigma_\varepsilon^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|} \quad \text{for } |h| \leq q$$

$$E[X_t X_{t+h}] = 0 \quad \text{for } |h| > q$$

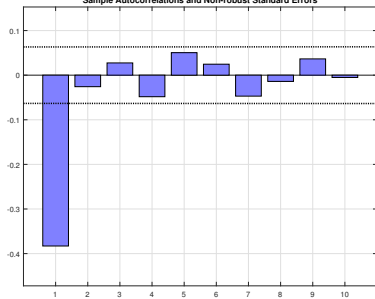
and the MA(q) is stationary and ergodic for the mean for any set of parameters θ_i , $i = 1, \dots, q$.



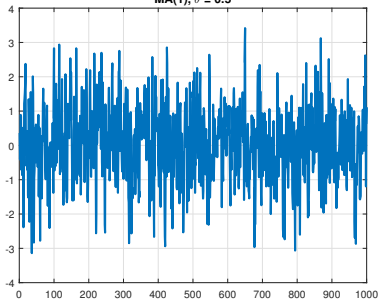
MA(1), $\theta = -0.5$



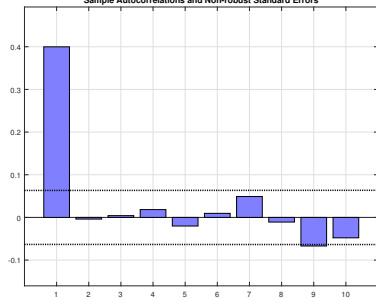
Sample Autocorrelations and Non-robust Standard Errors



MA(1), $\theta = 0.5$



Sample Autocorrelations and Non-robust Standard Errors



AR processes

Setting $\theta(L) \equiv 1$ yields the $AR(p)$ process

$$\phi(L)X_t = \varepsilon_t .$$

Consider the $AR(1)$:

$$X_t - \phi_1 X_{t-1} = \varepsilon_t .$$

This process deserves more attention.



Iterating backward shows

$$\begin{aligned}X_t &= \phi_1 X_{t-1} + \varepsilon_t \\&= \phi_1^2 X_{t-2} + \varepsilon_t + \phi_1 \varepsilon_{t-1} \\&\dots \\&= \phi_1^{k+1} X_{t-k-1} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \dots + \phi_1^k \varepsilon_{t-k}\end{aligned}$$

If $|\phi_1| < 1$ and X is stationary this suggests that

$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}$$

is the solution to the AR difference equation. It is an $MA(\infty)$ process.



The solution we obtain if $|\phi_1| < 1$ and X is stationary is the same as if we had inverted the lag polynomial

$$X_t - \phi_1 X_{t-1} = \varepsilon_t$$

$$(1 - \phi_1 L)X_t = \varepsilon_t$$

$$X_t = (1 - \phi_1 L)^{-1} \varepsilon_t$$

$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}$$



Is the AR(1) stationary? Yes, if $|\phi_1| < 1$.

$$E[X_t] = \sum_{j=0}^{\infty} \phi_1^j E[\varepsilon_{t-j}] = 0$$

for $h = 0, \pm 1, \pm 2, \dots$

$$\begin{aligned} \text{Cov}[X_t X_{t+h}] &= \lim_{n \rightarrow \infty} E \left[\sum_{j=0}^n \phi_1^j \varepsilon_{t-j+h} \sum_{k=0}^n \phi_1^k \varepsilon_{t-k} \right] \\ &= \sigma_\varepsilon^2 \phi_1^{|h|} \sum_{j=0}^{\infty} \phi_1^{2j} \\ &= \sigma_\varepsilon^2 \phi_1^{|h|} / (1 - \phi_1^2) \\ \rho_h &= \phi_1^{|h|} \end{aligned}$$



Causal versus noncausal processes

What if $|\phi_1| > 1$? Clearly,

$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}$$

would not converge. But we can rewrite the AR(1) equation as

$$X_t = \phi_1^{-1} X_{t+1} - \phi_1^{-1} \varepsilon_{t+1}.$$

Proceeding as before leads to the (stationary!) solution

$$X_t = - \sum_{j=0}^{\infty} \phi_1^{-j} \varepsilon_{t+j}$$



The process

$$X_t = - \sum_{j=0}^{\infty} \phi_1^{-j} \varepsilon_{t+j}$$

with $|\phi_1| > 1$ is called non-causal because today's values depend on the future ε_{t+j} , which is typically considered unnatural.

Therefore one only studies processes $|\phi_1| < 1$, because then

$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}.$$

Such processes are called **causal** (future-independent).



One omits processes $|\phi_1| > 1$. This is not a problem, for one can show that for any non-causal AR(1) process with $|\phi_1| > 1$ there exists a causal AR(1) process with $|\phi_1| < 1$ and a new white noise sequence such that the AR equation is satisfied. Hence nothing is lost.



For $|\phi_1| = 1$ no stationary solution can be found: we obtain the so called **unit root process**, aka. random walk.



Now consider the AR(2):

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \varepsilon_t .$$

It is tempting to find the solutions as

$$\begin{aligned} X_t &= (1 - \phi_1 L - \phi_2 L^2)^{-1} \varepsilon_t \\ &= \psi_0 + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} \dots \\ &= \psi(L) \varepsilon_t \end{aligned}$$

How make sense of $(1 - \phi_1 L - \phi_2 L^2)^{-1}$ and determine the ψ_j ?



One idea is to find numbers ξ_1 and ξ_2 such that

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \xi_1 L)(1 - \xi_2 L)$$

where $\xi_1 + \xi_2 = \phi_1$ and $\xi_1 \xi_2 = -\phi_2$. Provided that $|\xi_1| < 1$ and $|\xi_2| < 1$ we could conclude that

$$(1 - \phi_1 L - \phi_2 L^2)^{-1} = (1 - \xi_1 L)^{-1}(1 - \xi_2 L)^{-1}$$

where we know how to deal with $(1 - \xi_1 L)^{-1}$ and $(1 - \xi_2 L)^{-1}$.



How to find ξ_1 and ξ_2 such that

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \xi_1 L)(1 - \xi_2 L) ?$$

This holds whenever

$$(1 - \phi_1 z - \phi_2 z^2) = (1 - \xi_1 z)(1 - \xi_2 z),$$

where $z \in \mathbb{C}$.

This factorized polynomial clarifies that the problem is equivalent to finding the roots of the complex polynomial $(1 - \phi_1 z - \phi_2 z^2)$: The relationship must hold for any $z \in \mathbb{C}$, thus also for its roots $z_1^* = 1/\xi_1$ and $z_2^* = 1/\xi_2$.



From this, the following important conclusion follows:

If the roots of the complex polynomial $(1 - \phi_1 z - \phi_2 z^2)$ lie **outside the unit circle**, then $|\xi_1| < 1$ and $|\xi_2| < 1$, and the lag operator $(1 - \phi_1 L - \phi_2 L^2)$ is invertible because $(1 - \xi_1 L)^{-1}$ and $(1 - \xi_2 L)^{-1}$ are well defined.

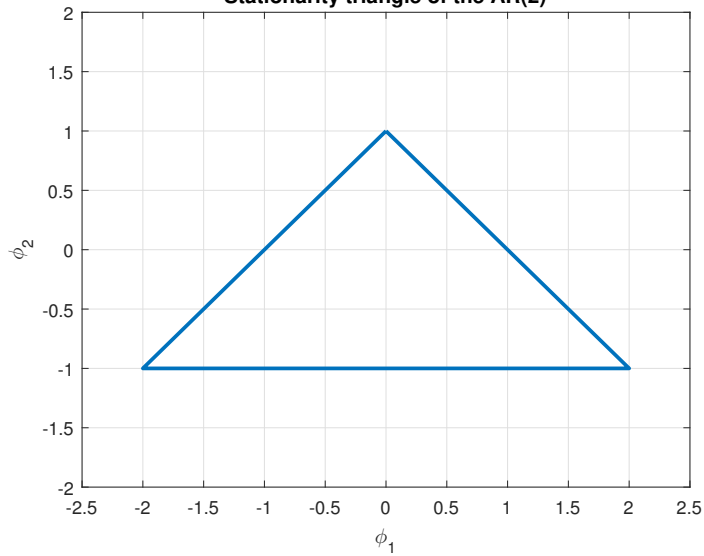
In consequence,

$$X_t = (1 - \phi_1 L - \phi_2 L^2)^{-1} \varepsilon_t = (1 - \xi_1 L)^{-1} (1 - \xi_2 L)^{-1} \varepsilon_t$$

represents the stationary solution to the AR(2).



Stationarity triangle of the AR(2)



How to find ψ_j such that

$$(1 - \phi_1 L - \phi_2 L^2)(\psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 \dots) = 1?$$

By factoring out and matching coefficients:

$$\psi_0 = 1 \quad \Rightarrow \quad \psi_0 = 1$$

$$\psi_1 - \phi_1 \psi_0 = 0 \quad \Rightarrow \quad \psi_1 = \phi_1 \psi_0 = \phi_1$$

$$\psi_2 - \phi_1 \psi_1 - \phi_2 \psi_0 = 0 \quad \Rightarrow \quad \psi_2 = \phi_1 \psi_1 + \phi_2 \psi_0 = \phi_1^2 + \phi_2$$

$$\begin{aligned} \psi_3 - \phi_1 \psi_2 - \phi_2 \psi_1 = 0 \quad \Rightarrow \quad \psi_3 = \phi_1 \psi_2 + \phi_2 \psi_1 = \\ \phi_1(\phi_1^2 + \phi_2) + \phi_2 \phi_1 \end{aligned}$$

\vdots

$$\psi_k = \phi_1 \psi_{k-1} + \phi_2 \psi_{k-2}$$



For the first two moments we find

$$E[X_t] = 0$$

$$\gamma_0 = \left(\frac{1 - \phi_2}{1 + \phi_2} \right) \frac{\sigma_\varepsilon^2}{(1 - \phi_2)^2 - \phi_1^2}$$

$$\gamma_1 = \frac{\phi_1}{1 - \phi_2} \gamma_0$$

$$\gamma_h = \phi_1 \gamma_{h-1} + \phi_2 \gamma_{h-2} \text{ for } h \geq 2$$

Thus the acf follows the same 2nd order difference equations as does the process.



For the general case, $AR(p)$

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p} = \varepsilon_t$$

the following result applies:

$$X_t = \psi(L)\varepsilon_t,$$

where $\psi(L) = 1/\phi(L)$, is the stationary solution of the $AR(p)$ if the roots of the complex polynomial

$$1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p = 0$$

lie **outside the unit circle**.



The acf of the $AR(p)$ follows a p th order difference equation:

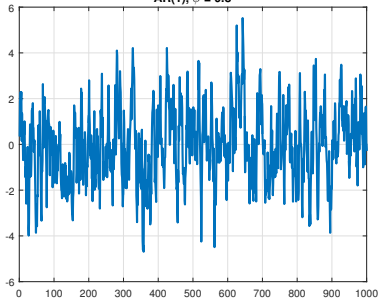
$$\gamma_h = \begin{cases} \phi_1\gamma_1 + \phi_2\gamma_2 + \dots + \phi_p\gamma_p + \sigma_\varepsilon^2 & \text{for } h = 0 \\ \phi_1\gamma_{h-1} + \phi_2\gamma_{h-2} + \dots + \phi_p\gamma_{h-p} & \text{for } h > 0 \end{cases}$$

Dividing by γ_0 yields a recursion known as **Yule-Walker** equations:

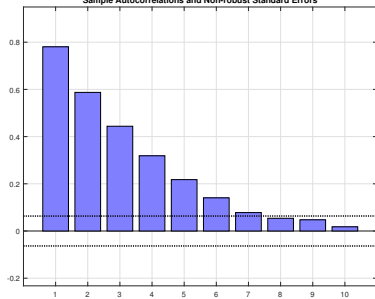
$$\rho_h = \phi_1\rho_{h-1} + \phi_2\rho_{h-2} + \dots + \phi_p\rho_{h-p} \text{ for } h > 0$$



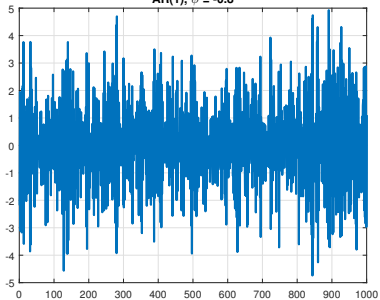
AR(1), $\phi = 0.8$



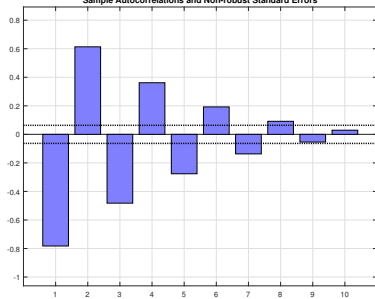
Sample Autocorrelations and Non-robust Standard Errors



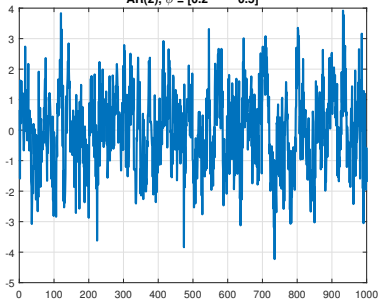
AR(1), $\phi = -0.8$



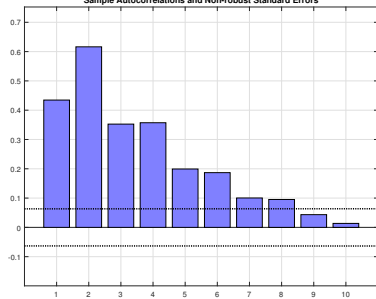
Sample Autocorrelations and Non-robust Standard Errors



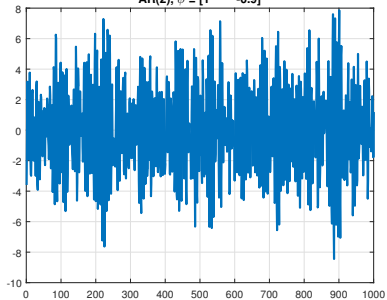
AR(2), $\phi = [0.2 \quad 0.5]$



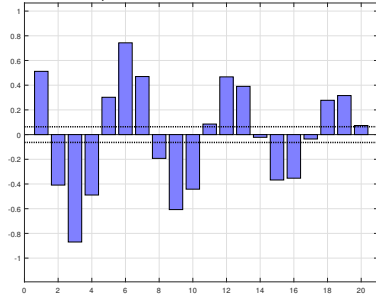
Sample Autocorrelations and Non-robust Standard Errors



AR(2), $\phi = [1 \quad -0.9]$



Sample Autocorrelations and Non-robust Standard Errors



ARMA processes

We tackle now the ARMA(p, q) process.

$$\phi(L)X_t = \theta(L)\varepsilon_t.$$

We directly conclude that the stationary solution is

$$X_t = \frac{\theta(L)}{\phi(L)}\varepsilon_t,$$

provided that the roots of the complex polynomial

$$1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p = 0$$

lie **outside the unit circle**. The ARMA is then causal.
Stationarity depends only on the AR part!



Yet there are two issues of indeterminacy in

$$X_t = \frac{\theta(L)}{\phi(L)} \varepsilon_t$$

common roots and **invertibility**.

Consider the WN:

$$X_t = \varepsilon_t .$$

Multiply that with $(1 - aL)$, for some $a \neq 0$, to receive

$$(1 - aL)X_t = (1 - aL)\varepsilon_t .$$

Magic: we made an ARMA(1,1) from a WN! This creates difficulties, because both versions are observationally equivalent.



This issue also occurs in the general ARMA (p, q) case:

$$\begin{aligned} X_t &= \frac{\theta(L)}{\phi(L)} \varepsilon_t \\ &= \frac{1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p} \varepsilon_t \\ &= \frac{(1 - \eta_1 L)(1 - \eta_2 L) \cdots (1 - \eta_q L)}{(1 - \xi_1 L)(1 - \xi_2 L) \cdots (1 - \xi_p L)} \varepsilon_t, \end{aligned}$$

where $\xi_j^{-1}, j = 1, \dots, p$ are the roots of $\phi(z) = 0$ and $\eta_i^{-1}, i = 1, \dots, q$ are the roots of $\theta(z) = 0$.

Suppose now that some roots coincide, i.e., $\eta_i = \xi_j$ for some i and j . We could cancel these roots without any implication to the representation of X , i.e., gives an ARMA $(p - 1, q - 1)$.



In consequence, we restrict attention to solutions

$$X_t = \frac{\theta(L)}{\phi(L)} \varepsilon_t,$$

where the polynomials associated to the two lag operators $\phi(L)$ and $\theta(L)$ **do not share common roots**.



There is another source of indeterminacy. Consider the MA(1)

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

and the second MA(1)

$$Y_t = \tilde{\varepsilon}_t + \tilde{\theta}_1 \tilde{\varepsilon}_{t-1} ,$$

where $\tilde{\theta}_1 = 1/\theta_1$ and $\text{Var}[\tilde{\varepsilon}_t] = \sigma_{\tilde{\varepsilon}}^2 = \sigma_{\varepsilon}^2 \theta_1^2$.

It is easy to see that

$$E[X_t X_{t+h}] = E[Y_t Y_{t+h}] .$$

Thus both processes are observationally second-order equivalent.



To resolve this problem, one chooses the **invertible** MA representation, i.e., the one for which $\theta(L) = 1 + \theta L$ can be inverted to such that

$$\varepsilon_t = \frac{1}{\theta(L)} X_t = \sum_{j=0}^{\infty} \psi_j X_{t-j}.$$

This is a causal $\text{AR}(\infty)$ representation. For this to hold the MA lag polynomial must admit an inverse, i.e., in $\theta(L) = 1 + \theta_1 L = 1 - (-\theta_1)L$ we need $|\theta_1| < 1$.

For the $\text{MA}(q)$, for **invertibility**, we require that all roots of $\theta(z)$ lie **outside the unit circle**.



Summarizing,

$$X_t = \frac{\theta(L)}{\phi(L)} \varepsilon_t,$$

is the stationary solution of $\phi(L)X_t = \theta(L)\varepsilon_t$ if and only if

- $\phi(z)$ has roots outside the unit circle;
- $\theta(z)$ has roots outside the unit circle;
- and $\phi(z)$ and $\theta(z)$ do not share any common roots.



Let's inspect the ARMA(1,1):

$$X_t - \phi_1 X_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Solution with MA(∞) representation:

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

To find the ψ_j , we proceed as before by coefficient matching

$$\psi_0 = 1 \quad \Rightarrow \quad \psi_0 = 1$$

$$-\phi_1 \psi_0 + \psi_1 = \theta_1 \quad \Rightarrow \quad \psi_1 = \theta_1 + \phi_1 \psi_0 = \phi_1 + \theta_1$$

$$-\phi_1 \psi_1 + \psi_2 = 0 \quad \Rightarrow \quad \psi_2 = \phi_1 \psi_1 = \phi_1(\phi_1 + \theta_1) = \phi_1^2 + \phi_1 \theta_1$$

$$\vdots$$

$$\psi_k = \phi_1 \psi_{k-1} = (\phi_1 + \theta_1) \phi_1^{k-1} \quad \text{for } k \geq 2.$$



Mean and moments are

$$E[X_t] = 0$$

$$\begin{aligned}\gamma_0 &= \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma_\varepsilon^2 \left(1 + \sum_{j=1}^{\infty} (\phi + \theta)^2 \phi^{2(j-1)} \right) \\ &= \sigma_\varepsilon^2 \left(1 + (\phi + \theta)^2 \sum_{j=0}^{\infty} \phi^{2j} \right) \\ &= \sigma_\varepsilon^2 \left(1 + \frac{(\phi + \theta)^2}{1 - \phi^2} \right) \\ &= \sigma_\varepsilon^2 \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}\end{aligned}$$



and for $h \geq 1$

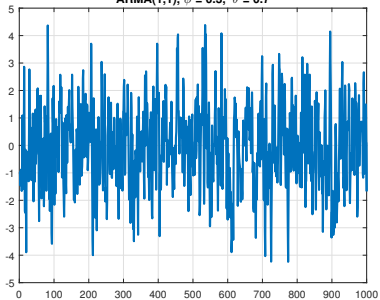
$$\gamma_h = \sigma_\varepsilon^2 \frac{(\phi + \theta)(1 + \theta\phi)}{1 - \phi^2} \phi^{h-1}$$

$$\rho_h = \sigma_\varepsilon^2 \frac{(\phi + \theta)(1 + \theta\phi)}{1 + 2\theta\phi + \theta^2} \phi^{h-1}$$

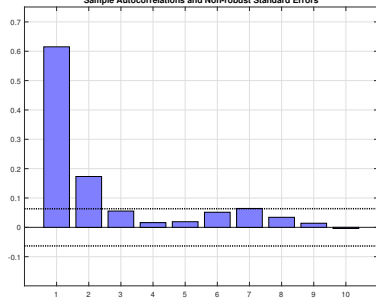
Note that it will be almost impossible to distinguish an ARMA(1,1) from the AR(1) by studying the acf!

[Overview acf](#)

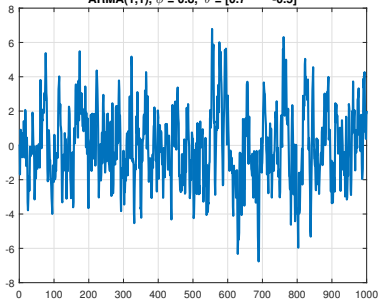
ARMA(1,1), $\phi = 0.3$, $\theta = 0.7$



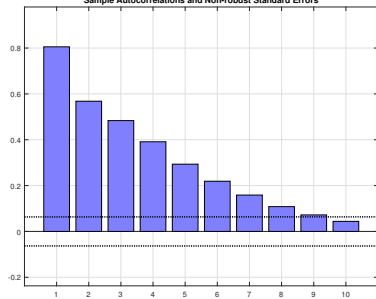
Sample Autocorrelations and Non-robust Standard Errors



ARMA(1,1), $\phi = 0.8$, $\theta = [0.7 \quad -0.5]$



Sample Autocorrelations and Non-robust Standard Errors



Forecasting

Prediction is very difficult, especially about the future.
Niels Bohr (attributed).

What is a good forecast? The short answer is: it depends.

It depends on the loss function that the economic agent uses to evaluate if the forecast is off by a particular amount. An optimal forecast is one that minimizes the agent's loss function.



A loss function $L(\cdot)$ is a function of the forecast error

$$e_{t+h} = X_{t+h} - \hat{X}_{t+h|t}$$

where X_{t+h} is an h -step ahead rv and $\hat{X}_{t+h|t}$ its time- t forecast.

A loss function has the properties:

- no error, no loss:

$$L(0) = 0$$

- nonnegativity:

$$L(u) \geq 0$$

- monotonically non-decreasing away from 0



Typical loss functions:

- quadratic loss

$$L(u) = u^2$$

- absolute loss

$$L(u) = |u|$$

- LINEX loss

$$L(u) = e^{au} - au - 1$$

with $a \neq 0$.

... and many more, depending on purpose.



The by far most frequent choice is $L(u) = u^2$, which leads to forecasts minimizing the MSE:

$$E[L(e_{t+h})] = E[(X_{t+h} - \hat{X}_{t+h|t})^2] .$$

The optimal prediction, given time- t information \mathcal{F}_t , is

$$\hat{X}_{t+h|t}^* = E[X_{t+h} | \mathcal{F}_t] ,$$

i.e. the conditional mean!



This is seen as follows

$$\begin{aligned} E[(X_{t+h} - \hat{X}_{t+h|t})^2] &= E[(X_{t+h} - E[X_{t+h}|\mathcal{F}_t] + E[X_{t+h}|\mathcal{F}_t] - \hat{X}_{t+h|t})^2] \\ &= E[(X_{t+h} - E[X_{t+h}|\mathcal{F}_t])^2] \\ &\quad + E[(E[X_{t+h}|\mathcal{F}_t] - \hat{X}_{t+h|t})^2] \end{aligned}$$

since the cross term is zero by the LIE.

The minimum is achieved for

$$\hat{X}_{t+h|t}^* = E[X_{t+h}|\mathcal{F}_t]$$

with

$$\text{MSE} = E[(X_{t+h} - E[X_{t+h}|\mathcal{F}_t])^2]$$



Linear forecasting rules and projection

In general, computing $E[X_{t+h}|\mathcal{F}_t]$ can be complicated! Therefore one often restricts attention to **linear forecasting rules**.

Suppose \mathbf{X}_t is vector that contains potential predictor variables including the constant 1. Further predictor values are current and past values, such as $X_t, X_{t-1}, X_{t-2}, \dots$

A linear forecasting rule is

$$\hat{X}_{t+h|t} = \alpha^\top \mathbf{X}_t$$

for some vector α .



How to choose α optimally?

The optimal forecast has a forecast error that is uncorrelated with the predictors:

$$E[(X_{t+h} - \alpha^\top \mathbf{X}_t) \mathbf{X}_t^\top] = 0,$$

in words: $E[\text{error} \times \text{predictor variable}] = 0$

If this holds, the forecast $\alpha^\top \mathbf{X}_t$ is called a **projection**.

Can also think of condition as the FOC of quadratic loss.



For forecasting rules of the form

$$\alpha^\top \mathbf{X}_t$$

which satisfy

$$E[(X_{t+h} - \alpha^\top \mathbf{X}_t) \mathbf{X}_t^\top] = 0$$

we will write

$$\text{proj}(X_{t+h} | \mathbf{X}_t) = \alpha^\top \mathbf{X}_t.$$



Projection minimizes MSE among all linear forecasting rules. But

$$\text{MSE}(\text{proj}(X_{t+h}|\mathbf{X}_t)) \geq \text{MSE}(E[X_{t+h}|\mathcal{F}_t]),$$

because the conditional expectation considers also nonlinear forecasting rules.

From the prediction equation, it is easily seen that

$$\alpha = \{E[(\mathbf{X}_t\mathbf{X}_t^\top)]\}^{-1}E[\mathbf{X}_tX_{t+h}]$$

This looks like OLS regression in population... Yet it is weaker: We only need second-order moments to exist, i.e., that \mathbf{X}_t is stationary and ergodic for second-order moments. We do not need that \mathbf{X} causes X_{t+h} in an economic (structural) sense.



Let X be a process (mean zero, for simplicity), and at time t we predict X_{t+h} , $h > 0$, using m values from the past which we stack in $\mathbf{X}_t = (X_t, X_{t-1}, \dots, X_{t-m})^\top$.

Then $\alpha = \{E[(\mathbf{X}_t \mathbf{X}_t^\top)]\}^{-1} E[\mathbf{X}_t X_{t+h}]$ has the succinct interpretation

$$\alpha = \Gamma_m^{-1} \gamma^{(m)}$$

where $\Gamma_m = [\gamma_{i-j}]_{i,j=1}^m$ and $\gamma^{(m)} = (\gamma_h, \gamma_{h+1}, \dots, \gamma_{h+m-1})^\top$.

Hence $\alpha = (\alpha_1, \dots, \alpha_m)^\top$ can be determined from the acf.



If X has a mean different from zero set $\mathbf{X}_t = (1, X_t, X_{t-1}, \dots, X_{t-m})^\top$.



Partial autocorrelation function

Set $h = 1$. Then we get

$$\alpha = \Gamma_m^{-1} \gamma^{(m)}$$

where $\Gamma_m = [\gamma_{i-j}]_{i,j=1}^m$ and $\gamma^{(m)} = (\gamma_1, \gamma_2, \dots, \gamma_m)^\top$.

In this setting, the m th entry of $\alpha = (\alpha_1, \dots, \alpha_m)^\top$, is called **partial autocorrelation (pacf)**.

Pacf measures the correlation between X_{t+1} and X_{t+1-m} while controlling for the in-between lagged values from t up to $t + 2 - m$, i.e., it gives the net effect of X_{t+1-m} on X_{t+1} .



In empirical analysis, one plots the pacf for $m = 1, 2, 3 \dots$

Depending on the process (MA, AR, ARMA), the pacf displays distinct patterns.

Thus pacf is an important diagnostic tool in time series modeling.



Consider the $AR(p)$. The AR has nonzero pacf for any $m \leq p$.
For $m = p$, the pacf is

$$\alpha_p = \phi_p$$

and

$$\alpha_m = 0$$

if $m > p$.

This follows from checking

$$E[(X_{t+1} - \alpha^\top \mathbf{x}_t) \mathbf{x}_t^\top] = 0.$$

In contrast to its acf, the pacf has a sharp cut-off.



The pacf of an MA process (and ARMA) dies off gradually, and derivation is more complicated.

E.g., for an MA(1)

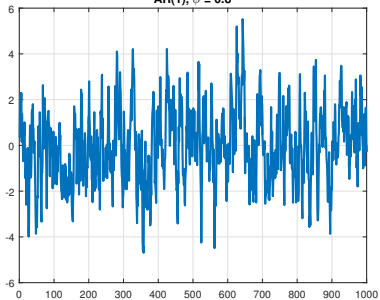
$$\alpha_m = -\frac{(-\theta)^m}{1 + \theta^2 + \dots + \theta^{2m}}$$



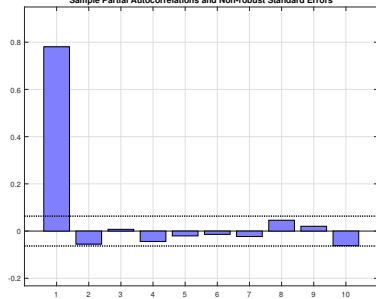
Like acf, pacf is used to detect ARMA components in a times series.

[Overview pacf](#)

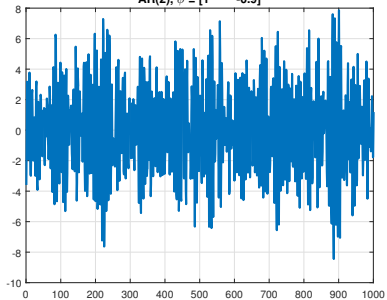
AR(1), $\phi = 0.8$



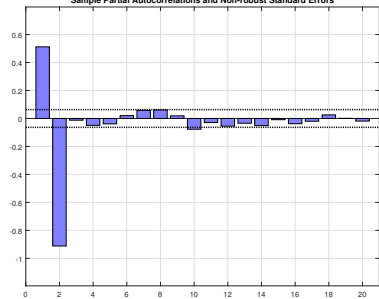
Sample Partial Autocorrelations and Non-robust Standard Errors



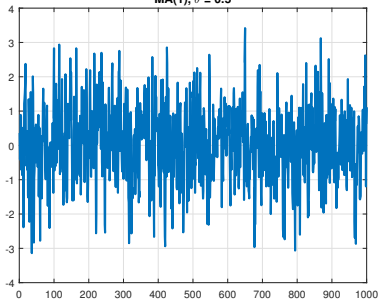
AR(2), $\phi = [1 \quad -0.9]$



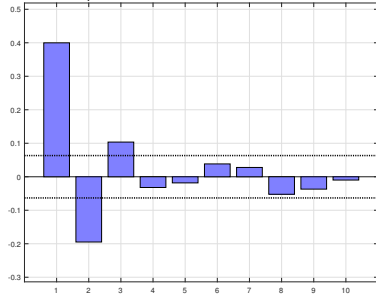
Sample Partial Autocorrelations and Non-robust Standard Errors



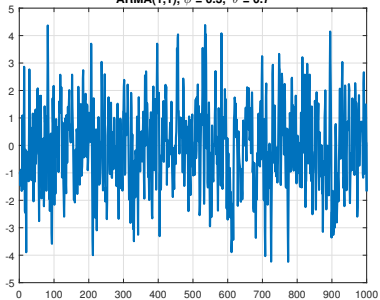
MA(1), $\theta = 0.5$



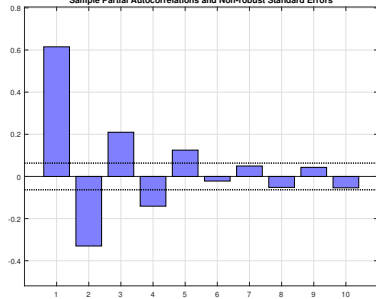
Sample Partial Autocorrelations and Non-robust Standard Errors



ARMA(1,1), $\phi = 0.3$, $\theta = 0.7$



Sample Partial Autocorrelations and Non-robust Standard Errors



Prediction of the MA(q)

If $h \leq q$

$$\begin{aligned} X_{t+h} &= \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + \dots \\ &\quad + \theta_{h-1} \varepsilon_{t+1} + \theta_h \varepsilon_t + \theta_{h+1} \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q+h} \end{aligned}$$

It is obvious that the optimal linear forecast is:

$$\hat{X}_{t+h|t} = \begin{cases} \theta_h \varepsilon_t + \theta_{h+1} \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q+h} & h \leq q \\ 0 & h > q \end{cases}$$



This optimal linear forecast is also the best MSE forecast!



The MSE is

$$\text{MSE} = \begin{cases} \sigma_{\varepsilon}^2 & h = 1 \\ (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_{h-1}^2) \sigma_{\varepsilon}^2 & h = 2, 3, \dots, q \\ (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2) \sigma_{\varepsilon}^2 & h = q + 1, q + 2, \dots \end{cases},$$

i.e., for $h > q$ the MSE is equal to the unconditional variance of the MA(1).

The MSE is the same as the forecast error variance because the forecast is unbiased ($E[e_{t+h}] = 0$):

$$\text{MSE} = E[(X_{t+h} - E[X_{t+h}|\mathcal{F}_t])^2] = \text{Var}[e_{t+h}]$$



Issue: not feasible because $\{\varepsilon_t\}$ is unknown.

Solutions:

1. adopt assumptions on initial values for $\{\varepsilon_t\}$ and then using observed X_t iterate forward through the MA difference equation to obtain a “reconstructed” error sequence $\{\hat{\varepsilon}_t\}$
2. base prediction on observed X_t by working with the $AR(\infty)$ representation – see discussion on ARMA processes!



Prediction of the AR(p)

If $p \leq h$

$$\begin{aligned} X_{t+h} = & \varepsilon_{t+h} + \phi_1 X_{t+h-1} + \dots \\ & + \phi_{h-1} X_{t+1} + \phi_h X_t + \phi_{h+1} X_{t-1} + \dots + \phi_p X_{t-p+h} \end{aligned}$$

The optimal linear forecast always has exactly p terms:

$$\hat{X}_{t+h|t} = \phi_1 X_{t+h-1} + \dots + \phi_p X_{t-p+h}$$



Again this is also the best MSE forecast!



Yet,

$$\hat{X}_{t+h|t} = \phi_1 X_{t+h-1} + \dots + \phi_p X_{t-p+h}$$

seems infeasible because future X_{t+h} are unknown. The fact the optimal prediction is minimal MSE allows us to work with a feasible version where future values are **recursively** replaced by their own forecast:

$$\hat{X}_{t+h|t} = \phi_1 \hat{X}_{t+h-1|t} + \dots + \phi_p \hat{X}_{t-p+h|t}$$

By properties of iterated projections, the forecast remains minimal MSE.



For the case if the AR(1), MSE is best computed from the MA(∞) representation. Then one immediately sees

$$\text{MSE} = \sigma_{\varepsilon}^2 \left(1 + \phi_1^2 + \phi_1^4 + \dots + \phi_1^{2(h-1)} \right)$$

so as $h \rightarrow \infty$

$$\text{MSE} = \sigma_{\varepsilon}^2 / (1 - \phi_1^2)$$

which is the unconditional variance of the AR(1).



Predicting ARMA processes

Poses similar difficulties as does the MA process. Two suitable approaches:

- put the ARMA in the $AR(\infty)$ representation (what do you need for this to work?).
- base predictions on the m past values of X_t , with m large
- solve the best linear prediction problem to obtain

$$\alpha = \Gamma_m^{-1} \gamma^{(m)}$$

where $\Gamma_m = [\gamma_{i-j}]_{i,j=1}^m$ and $\gamma^{(m)} = (\gamma_h, \gamma_{h+1}, \dots, \gamma_{h+m-1})^\top$.

- use $\alpha = (\alpha_1, \dots, \alpha_m)^\top$ for the predictions



OR:

1. adopt assumptions on initial values for $\{\varepsilon_t\}$ and then use observed X_t to iterate forward through the ARMA difference equation
2. obtain the “reconstructed” error sequence $\{\hat{\varepsilon}_t\}$
3. base predictions on observed $\{X_t\}$ and $\{\hat{\varepsilon}_t\}$



Wold's decomposition

All models which we considered so far can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ and $\psi_0 = 1$.

It may seem that this is because we restricted the discussion to a particular, convenient class of models. This is not true.



Wold's decomposition theorem:

Any covariance stationary process X can be represented uniquely as

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t,$$

where $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ and $\psi_0 = 1$.

- ε is a (weak) WN sequence
- κ is uncorrelated with ε and can be predicted perfectly by projection using infinitely many values of past X_t , i.e.,

$$\kappa_t = \text{proj}(\kappa_t | X_{t-1}, X_{t-2}, \dots)$$



In Wold's decomposition

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t,$$

□ $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ is called **linearly indeterministic component**;

□ κ_t is called **linearly deterministic component**;



linearly deterministic does not mean *non-random*!



Implications of the Wold theorem

- every second-order stationary process is either linear – or can be transformed into a linear process by subtracting a deterministic component;
- the representation is causal
- ARMA processes are purely indeterministic
- the class of linear processes, even when considering purely indeterministic processes, is larger than that of ARMA processes



Despite these facts,

- one justifies the concentration on ARMA processes by arguing that the $MA(\infty)$ representation of the Wold representation can be approximated well by a ratio of ARMA lag polynomials:

$$\psi(L) \approx \frac{\theta(L)}{\phi(L)}$$

\Rightarrow Box-Jenkins method



Example

Consider the following process:

$$X_t = 0.4 + 1.2X_{t-1} - 0.32X_{t-2} + u_t + 0.3u_{t-1} + 0.02u_{t-2},$$

where $u \sim (0, 1)$ is white noise.

- ▣ How is the model called? Write it in lag polynomial form.
- ▣ Discuss stationarity, invertibility.
- ▣ If stationary, find the first four terms of the $MA(\infty)$ representation.
- ▣ Find the unconditional mean.
- ▣ Find variance, and first, second, third autocovariance.
- ▣ Find the corresponding autocorrelations.
- ▣ Find 1 to 3 step ahead forecasts, forecast errors and forecast error variances. Assume $X_T = X_{T-1} = 1$ and $u_T = u_{T-1} = -1$.



Overview on (p)acf patterns, Source: Enders, Back acf Back pcf

Table 2.1 Properties of the ACF and PACF

Process	ACF	PACF
White noise	All $\rho_s = 0$ ($s \neq 0$)	All $\phi_{ss} = 0$
AR(1): $a_1 > 0$	Direct geometric decay: $\rho_s = a_1^s$	$\phi_{11} = \rho_1$; $\phi_{ss} = 0$ for $s \geq 2$
AR(1): $a_1 < 0$	Oscillating decay: $\rho_s = a_1^s$	$\phi_{11} = \rho_1$; $\phi_{ss} = 0$ for $s \geq 2$
AR(p)	Decays toward zero. Coefficients may oscillate.	Spikes through lag p . All $\phi_{ss} = 0$ for $s > p$.
MA(1): $\beta > 0$	Positive spike at lag 1. $\rho_s = 0$ for $s \geq 2$	Oscillating decay: $\phi_{11} > 0$.
MA(1): $\beta < 0$	Negative spike at lag 1. $\rho_s = 0$ for $s \geq 2$	Geometric decay: $\phi_{11} < 0$.
ARMA(1, 1) $a_1 > 0$	Geometric decay beginning after lag 1. Sign $\rho_1 = \text{sign}(a_1 + \beta)$	Oscillating decay after lag 1. $\phi_{11} = \rho_1$
ARMA(1, 1) $a_1 < 0$	Oscillating decay beginning after lag 1. Sign $\rho_1 = \text{sign}(a_1 + \beta)$	Geometric decay beginning after lag 1. $\phi_{11} = \rho_1$ and $\text{sign}(\phi_{ss}) = \text{sign}(\phi_{11})$.
ARMA(p, q)	Decay (either direct or oscillatory) beginning after lag q .	Decay (either direct or oscillatory) beginning after lag p .

Chapter 4:

Estimating Time Series Models



Outline

The following topics will be addressed:

- Diagnostics: how to detect serial dependence
- OLS estimation of AR models
- OLS estimation in the presence of serially correlated residuals
- Maximum Likelihood estimation of ARMA models



Estimation of the acf

A natural estimator of the acf is:

$$\hat{\gamma}_h = T^{-1} \sum_{i=1}^{T-h} (X_i - \bar{X}_T)(X_{i+h} - \bar{X}_T)$$

One prefers the normalization T^{-1} to $(T - |h|)^{-1}$ because then the matrix of autocovariances remains positive definite.



$\hat{\gamma}_h$ is biased, unless the mean is known¹, for

$$\mathbb{E}[\hat{\gamma}_h] = \left(1 - \frac{|h|}{T}\right) \gamma_h - \left(1 - \frac{|h|}{T}\right) \text{Var}[\bar{X}_T] + \mathcal{O}(T^{-2})$$

Positive autocovariances will be underestimated, but asymptotically $\lim_{T \rightarrow \infty} \mathbb{E}[\hat{\gamma}_h] = \gamma_h$.

¹Normalizing with $(T - |h|)^{-1}$ would not remove the bias!



$\text{Var}[\hat{\gamma}_h]$ is pretty ugly. It holds that $\lim_{T \rightarrow \infty} \text{Var}[\hat{\gamma}_h] = 0$ so that the estimator is consistent.

An estimator of autocorrelation is

$$\hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0} = \frac{\sum_{i=1}^{T-h} (X_i - \bar{X}_T)(X_{i+h} - \bar{X}_T)}{\sum_{i=1}^T (X_i - \bar{X}_T)^2}$$

Because $\hat{\gamma}_h$ is biased, so is $\hat{\rho}_h$:

$$E[\hat{\rho}_h] = \rho_h + \mathcal{O}(T^{-1})$$

As T tends to infinity the bias vanishes.



CLT for the acf:

Let X be cov.stationary and given by

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

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and ε is **strict** white noise with $E[\varepsilon_t^4] < \infty$ then we have

$$\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{\mathcal{L}} \mathcal{N}(0, W),$$

where $\hat{\rho} = (\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_h)^\top$ and where $W = [w]_{ij}$, $i, j = 1, \dots, h$ is given by **Bartlett's formula** (see textbooks).

If $\rho_i = 0$, $i = 1, \dots, h$, W collapses to the unit matrix.



TS diagnostics

From these results, we can derive typical diagnostic statistics which are frequently used in applied work.

They pertain to the question:

Is there autocorrelation in a given process X ?

X may also be regression residuals.



ACF of the WN process

From the Barlett's formula, we learn: if X is a strict WN with $E[X_t^4] < \infty$, then

$$\sqrt{T}\hat{\rho} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{I}_h)$$

where \mathbf{I}_h denotes the h -dimensional identity matrix.

A 95% confidence interval for $\hat{\rho}(h)$ is

$$\left[\hat{\rho}_h - \frac{1.96}{\sqrt{T}}; \hat{\rho}_h + \frac{1.96}{\sqrt{T}} \right]$$

Because $\hat{\rho}_h$ is biased this interval may not have correct coverage in small samples!



Example

Standard software packages usually display the acf with the confidence interval $\pm 1.96 T^{-1/2}$. Clearly, if ε is not iid. distributed these intervals are too small. Need robust standard error estimates.



See Francq and Zakoïan (JTSA, 2009) for a generalized Bartlett's formula and [Dalla et al. \(2019\)](#).



Portmanteau statistics

Based on these results, we might ask for a test under the null hypothesis of zero autocorrelation at all lags for $h > 0$.

A natural statistic is the **Box-Pierce statistic**

$$Q(h) = n \sum_{i=1}^h \hat{\rho}_i^2 ,$$

which is asymptotically $\chi^2(h)$ with h degrees of freedom.

This is standard output of ts software, but biased in small samples and not robust to heteroskedasticity.



Durbin-Watson statistic

A frequently employed test serial correlation at lag 1 is the Durbin-Watson statistic. It is often applied to regression residuals.

The DW statistic is

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

where T is sample size and e_t a regression residual.

We see

$$d \approx 2(1 - \rho_1)$$

from which we get $0 \leq d \leq 4$.



We also note from $d \approx 2(1 - \rho_1)$ and $0 \leq d \leq 4$:

- if $d \approx 2$, (almost) no autocorrelation
- if $d > 2$, successive errors are negatively correlated
- if $d < 2$, successive errors are positively correlated

One can run statistical test at significance level α , but critical values are nonstandard and need to be looked up in tables.

Implemented in most TS software.



Estimating AR models by OLS

Stationary $AR(p)$ models can also be estimated via OLS. Suppose an $AR(1)$

$$X_t = \phi X_{t-1} + \varepsilon_t$$

where ε is strict white noise and $|\phi| < 1$.

The OLS estimator is

$$\hat{\phi} = \frac{\sum_{t=1}^T X_{t-1} X_t}{\sum_{t=1}^T X_{t-1}^2}$$

What are its properties?



Properties of the OLS estimators

- OLS estimators are **biased**, because the **strict exogeneity** assumption of classical OLS, i.e., $E[X_t \varepsilon_s] = 0$, for any s , **fails**.
- Simulations show that the bias is the bigger the more persistent the process (the closer ϕ is to one).
- The constants are upward biased and the coefficients will in general be downward biased.



Under appropriate conditions, the OLS estimator is consistent:

$$\hat{\phi} = \frac{\sum_{t=1}^T X_{t-1} X_t}{\sum_{t=1}^T X_{t-1}^2} = \phi + \frac{\sum_{t=1}^T X_{t-1} \varepsilon_t}{\sum_{t=1}^T X_{t-1}^2}$$

We need

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 = \gamma_0$$

which holds under ergodicity (for second-order moments).



We need show that

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t = 0$$

Standard LLN do not apply here. But $X_{t-1}\varepsilon_t$ is an MDS wrt. \mathcal{F}_t (natural filtration), i.e.,

$$E[X_{t-1}\varepsilon_t | \mathcal{F}_{t-1}] = 0$$

with finite second moment

$$E[(X_{t-1}\varepsilon_t)^2] = E[X_{t-1}^2]E[\varepsilon_t^2] = \gamma_0\sigma_\varepsilon^2.$$

Moreover, $E[X_{t-1}\varepsilon_t] = 0$ and $\text{Cov}[X_{t-1}\varepsilon_t, X_{t-h-1}\varepsilon_{t-h}] = 0$.



Hence $X_{t-1}\varepsilon_t$ is a stationary sequence and we can apply our LLN for stationary time series.

We conclude

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1}\varepsilon_t = 0,$$

and by the usual Slutsky arguments we get consistency:

$$p\text{-}\lim_{T \rightarrow \infty} \hat{\phi} = \phi.$$



For the limit distribution we need to study

$$\sqrt{T}(\hat{\phi} - \phi) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} \varepsilon_t}{\frac{1}{T} \sum_{t=1}^T X_{t-1}^2}$$

As previously, $p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 = \gamma_0$.

We try to apply our CLT for the MDS with finite second-order moments to

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} \varepsilon_t$$



This ensures that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} \varepsilon_t \xrightarrow{\mathcal{L}} \mathcal{N}(0, E[(X_{t-1} \varepsilon_t)^2])$$

because

$$E[(X_{t-1} \varepsilon_t)^2] = \sigma_\varepsilon^2 \gamma_0 = \frac{\sigma_\varepsilon^4}{1 - \phi^2}$$

The issue to worry about is second-order ergodicity of $\{X_{t-1} \varepsilon_t\}$.



To establish that $X_{t-1}\varepsilon_t$ is second-order ergodic one checks

$$\frac{1}{T} \sum_{t=1}^T X_{t-1}^2 \varepsilon_t^2 = \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 (\varepsilon_t^2 - \sigma_\varepsilon^2) + \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 \sigma_\varepsilon^2.$$

Now $X_{t-1}^2 (\varepsilon_t^2 - \sigma_\varepsilon^2)$ is an MDS. Invoking our LLNs for MDS we get

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 (\varepsilon_t^2 - \sigma_\varepsilon^2) = 0.$$

Moreover, if X^2 is ergodic for the mean, i.e., X is second-order ergodic, it follows

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sigma_\varepsilon^2 X_{t-1}^2 = \sigma_\varepsilon^2 \mathbb{E}[X_{t-1}^2] = \sigma_\varepsilon^2 \gamma_0.$$



Summarizing, if X is a second-order ergodic AR(1) driven by strict white noise we deduce

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1 - \phi^2)$$

because $\gamma_0 = \sigma_\varepsilon^2 / (1 - \phi^2)$.

Further comments:

- ▣ multivariate extensions for the AR(p) apply accordingly.
- ▣ AR parameters of ARMA models cannot be estimated: OLS is inconsistent.



OLS with serially correlated residuals

Suppose you have the regression model

$$Y_t = \beta X_t + u_t$$

where X , Y are stationary, mean zero and $E[u_t|X_t] = 0$.
Your diagnostics suggest that $\{u_t\}$ is serially correlated.

What now?

Answer: OLS is consistent (but potentially biased); however, you may need to compute standard errors taking into account serial correlation. We skip consistency and look into standard errors.



The OLS estimator takes the form

$$\hat{\beta} = \frac{\sum_{t=1}^T X_t Y_t}{\sum_{t=1}^T X_t^2}$$

so

$$\sqrt{T}(\hat{\beta} - \beta) = \frac{\frac{1}{\sqrt{T}}}{\frac{1}{T}} \frac{\sum_{t=1}^T X_t u_t}{\sum_{t=1}^T X_t^2} = \frac{\frac{1}{\sqrt{T}}}{\frac{1}{T}} \frac{\sum_{t=1}^T \nu_t}{\sum_{t=1}^T X_t^2}$$

Under the usual assumptions the denominator poses no difficulties:

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_t^2 = \sigma_X^2$$



We find for the estimation error

$$T \operatorname{Var}[\hat{\beta}] = \frac{\frac{1}{T} \operatorname{Var}[\sum_{t=1}^T \nu_t]}{(\sigma_X^2)^2}$$

but

$$\operatorname{Var}\left[\sum_{t=1}^T \nu_t\right] = \sum_{t=1}^T \sum_{s=1}^T \operatorname{Cov}[\nu_t, \nu_s]$$

does not boil down to our usual formular because in general

$$\operatorname{Cov}[\nu_t, \nu_s] \neq 0$$

for $t \neq s$, if u_t is serially correlated and $\operatorname{Var}\left[\sum_{t=1}^T \nu_t\right] \neq T\sigma_\nu^2$



Instead, we find

$$\begin{aligned}\text{Var} \left[\sum_{t=1}^T \nu_t \right] &= \sum_{t=1}^T \sum_{s=1}^T \text{Cov}[\nu_t, \nu_s] \\ &= \sum_{t=1}^T \sum_{s=1}^T \text{E}[\nu_t, \nu_s] \\ &= \sum_{t=1}^T \sum_{s=1}^T \gamma_{t-s}\end{aligned}$$

where γ_h is the autocovariance function of $\{\nu_t\}$ – recall our CLT for covariance-stationary processes!



$$\begin{aligned}\text{Var} \left[\sum_{t=1}^T \nu_t \right] &= \sum_{t=1}^T \sum_{s=1}^T \gamma_{t-s} \\ &= \gamma_0 \sum_{t=1}^T \sum_{s=1}^T \rho_{t-s} \\ &= \gamma_0 \left(T + 2 \sum_{j=1}^{T-1} (T-j) \rho_j \right) \\ &= \sigma_\nu^2 \left(T + 2 \sum_{j=1}^{T-1} (T-j) \rho_j \right)\end{aligned}$$

because $\gamma_0 = \sigma_\nu^2$.



Summarizing we can write

$$\begin{aligned} T \operatorname{Var}[\hat{\beta}] &= \frac{\frac{1}{T} \operatorname{Var}[\sum_{t=1}^T \nu_t]}{(\sigma_X^2)^2} \\ &= \frac{\frac{1}{T} \sigma_\nu^2}{(\sigma_X^2)^2} \left(T + 2 \sum_{j=1}^{T-1} (T-j) \rho_j \right) \\ &= \frac{\sigma_\nu^2}{(\sigma_X^2)^2} \underbrace{\left(1 + 2 \sum_{j=1}^{T-1} \frac{(T-j)}{T} \rho_j \right)}_{\text{Newey-West correction}} \end{aligned}$$



Practically, we use

$$\text{Var}[\hat{\beta}] \approx \frac{\sigma_v^2}{T(\sigma_X^2)^2} \underbrace{\left(1 + 2 \sum_{j=1}^m \frac{(m-j)}{m} \rho_j \right)}_{\text{Newey-West correction}}$$

where $m \approx T^{1/3}$ or $m \approx T^{1/4}$ as a rule of thumb.

You do not have to compute this yourself: most software packages offer these corrected “Heteroscedasticity and autocorrelation-consistent standard errors (HAC)” due to Newey and West.



Maximum likelihood

We are given a data sample x_1, \dots, x_T , drawn from some distribution described by the density $f(x; \theta)$ parametrized with θ .

If x_t , $t = 1, \dots, T$ are **iid**, their joint density can be written as

$$f(x_1, \dots, x_T; \theta) = \prod_{i=1}^T f(x_t; \theta) = \mathcal{L}(\theta | x_1, \dots, x_T)$$

This joint density is the **likelihood function**, but the perspective changes: Instead of viewing \mathcal{L} as a function of data given parameters, we regard it as a function of parameters θ *given* data.



The ML estimator is the argument for which \mathcal{L} is maximal:

$$\hat{\theta}^{ML} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta | x_1, \dots, x_T) .$$

Taking logs results in

$$\log \mathcal{L}(\theta | x_1, \dots, x_T) = \sum_{t=1}^T \log f(x_t; \theta)$$

and – since the log is a monotone one-to-one function –

$$\hat{\theta}^{ML} = \arg \max_{\theta \in \Theta} \log \mathcal{L}(\theta | x_1, \dots, x_T) .$$

Nothing changes substantially.



The maximum of the function is attained when the "likelihood equation" attains zero (first order condition):

$$s(\hat{\theta}) = \frac{\partial \log \mathcal{L}(\hat{\theta} | x_1, \dots, x_T)}{\partial \theta} = 0$$

and when the Hessian

$$H(\hat{\theta}) = \frac{\partial^2 \log \mathcal{L}(\hat{\theta} | x_1, \dots, x_T)}{\partial \theta \partial \theta^\top} \text{ is negative definite}$$

(second order condition).

If the first order condition can be explicitly solved for θ we obtain a direct expression for $\hat{\theta}^{ML}$. If not, it must be solved numerically.



This derivation relies on the iid assumption, which not applicable for time series data. Thus

$$\mathcal{L}(\theta|x_1, \dots, x_T) = f(x_1, \dots, x_T; \theta) \neq \prod_{t=1}^T f(x_t; \theta)$$

We can then make use of iterative application of the rule

$$\text{joint pdf} = \text{conditional pdf} \times \text{marginal pdf}$$



Applying this formula yields (dependence on θ is suppressed):

$$\begin{aligned}f(x_1, \dots, x_T) &= f(x_2, \dots, x_T | x_1) f(x_1) \\&= f(x_3, \dots, x_T | x_1, x_2) f(x_2 | x_1) f(x_1) \\&= f(x_4, \dots, x_T | x_1, x_2, x_3) f(x_3 | x_1, x_2) f(x_2 | x_1) f(x_1) \\&= f(x_1) \prod_{t=2}^n f(x_t | x_1, \dots, x_{t-1})\end{aligned}$$

In the special case, if the conditional density $f(x_t | x_1, \dots, x_{t-1})$ depends only on the last realization x_{t-1} we get

$$f(x_1, \dots, x_T) = f(x_1) \prod_{t=2}^n f(x_t | x_{t-1}) .$$

This property is called **Markov property**, meaning that the process is memoryless and that x_t depends only on the current state summarized by x_{t-1} .



Properties of MLE:

- ML estimators are **consistent**, i.e.

$$p\text{-}\lim_{T \rightarrow \infty} \hat{\theta}^{ML} = \theta_0 .$$

- ML estimators might be **only asymptotically unbiased**, i.e.

$$E[\hat{\theta}^{(ML)}] = \theta_0 + \text{terms converging to zero as } T \rightarrow \infty$$

- Invariance:** Let $g(\theta)$ be a continuous function. Then the ML estimator of $g(\theta)$ is $g(\hat{\theta}^{ML})$.



- ML estimators are asymptotically normally distributed:

$$\sqrt{T}(\hat{\theta}^{ML} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

where

$$\mathcal{I}(\theta_0) = \lim_{T \rightarrow \infty} -E \left[\frac{1}{T} H(\theta_0) \right]$$

The variance of $\hat{\theta}^{ML}$ is the inverse of the **information matrix**.

- ML estimators are asymptotically efficient, i.e. the variance of $\hat{\theta}^{ML}$ achieves the minimal variance a consistent estimator can achieve, the so called **Cramér-Rao bound**.



ML estimation of ARMA models

Stationary, invertible ARMA models are usually estimated by means of Gaussian likelihood. Let's study the AR(1) first:

$$X_t = \phi X_{t-1} + \varepsilon_t$$

and assume $\varepsilon \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ and $|\phi| < 1$. The parameters to be estimated are $\theta = (\phi, \sigma_\varepsilon^2)$. We have

$$X_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2 / (1 - \phi^2))$$

$$\text{and} \quad X_t | X_{t-1} \sim \mathcal{N}(\phi X_{t-1}, \sigma_\varepsilon^2)$$



Accordingly, the full (log) likelihood is

$$\begin{aligned}\log \mathcal{L}(\theta) = & -\frac{1}{2} \log[\sigma_\varepsilon^2/(1 - \phi^2)] - \frac{x_1^2}{2\sigma_\varepsilon^2/(1 - \phi^2)} \\ & - \frac{T-1}{2} \log \sigma_\varepsilon^2 - \sum_{t=2}^T \frac{(x_t - \phi x_{t-1})^2}{2\sigma_\varepsilon^2}\end{aligned}$$

where the constants $-\frac{1}{2} \log(2\pi)$ and $-\frac{T-1}{2} \log(2\pi)$ are dropped ...

Optimization requires solving a system of nonlinear equations; closed-form solutions are not available.



As an alternative one studies the conditional likelihood where the initial value is treated as known. This is justified by noting that the influence of the initial value fades out under stationarity. Then

$$\log \mathcal{L}(\theta) = -\frac{T-1}{2} \log \sigma_\varepsilon^2 - \sum_{t=2}^T \frac{(x_t - \phi x_{t-1})^2}{2\sigma_\varepsilon^2}$$

Maximizing this is equivalent to OLS estimation: the same estimators result. Moreover, it follows that

$$\hat{\sigma}_\varepsilon^2 = \sum_{t=2}^T \frac{(x_t - \phi x_{t-1})^2}{T-1}$$



Let's inspect the MA(1) process. Assume

$$X_t = \varepsilon_t + \theta \varepsilon_{t-1}$$

with $\varepsilon \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$.

For the conditional likelihood, we exploit

$$X_t | \varepsilon_{t-1} \sim \mathcal{N}(\theta \varepsilon_{t-1}, \sigma_\varepsilon^2).$$

But it does not help to treat X_1 as known. Instead, we need to pretend to know ε_0 . Setting $\varepsilon_0 = 0$, we deduce the ε_t recursively

$$\varepsilon_t = \sum_{j=0}^{t-1} (-\theta)^j X_{t-j}$$

Clearly, for this to work the MA must be invertible!



Hence the log likelihood is

$$\log \mathcal{L}(\theta) = -\frac{T}{2} \log \sigma_{\varepsilon}^2 - \sum_{t=1}^T \frac{\varepsilon_t^2}{2\sigma_{\varepsilon}^2}$$

ε_t is a polynomial in θ , hence this is a highly non-linear estimation problem!



For the conditional likelihood of the ARMA model

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

one combines the previous strategies:

Condition on $X_1 = x_1$ and set $\varepsilon_0 = 0$. Reconstruct the ε_t recursively and the log likelihood becomes

$$\log \mathcal{L}(\theta) = -\frac{T-p}{2} \log \sigma_\varepsilon^2 - \sum_{t=p+1}^T \frac{\varepsilon_t^2}{2\sigma_\varepsilon^2}$$

Again ε_t is a polynomial in θ and ϕ , hence this is a highly non-linear estimation problem!



Most software packages compute $\text{ARMA}(p,q)$ from the conditional log-likelihood with normal errors, thus they to non-linear least squares estimation.

Potential estimation issues:

- ▣ problems are nonlinear, local (or nonsense) solutions possible
- ▣ good starting values needed!
- ▣ watch out for common roots; parameters are not identifiable
- ▣ watch out for non-invertible MA estimates



Model selection

Because economic theory does often not help determine a specific TS model, one employs statistical methods for model selection.

There are two avenues:

- by using test procedures offered by MLE, such as the Wald, the Lagrange Multiplier or the Likelihood ratio tests
- or by information criteria (in particular for order selection).

Finally, one checks the residuals in order to validate whether the model achieves a good fit: residuals should be approximately white noise (Portmanteau statistics).



Information criteria

Here, I report the Eviews definitions. T is sample size.

□ Akaike:

$$\text{AIC} = \frac{2}{T} \times \text{number of parameters} - \frac{2}{T} \log \mathcal{L}(\hat{\theta})$$

□ Schwarz (or Bayesian):

$$\text{SC} = \frac{\log(T)}{T} \times \text{number of parameters} - \frac{2}{T} \log \mathcal{L}(\hat{\theta})$$



They all share the common structure

$$\frac{\phi(T)}{T} \times \text{number of parameters} - \frac{2}{T} \log \mathcal{L}(\hat{\theta})$$

where $\phi(T) = 2$ for AIC, and $\phi(T) = \log(T)$ for SC.

The selected model is the one which **minimizes** the criterion.

In general, AIC tends to pick rich models, SC smaller ones.



The ICs are based on likelihood theory. In the special case of LS estimation with normally distributed errors, they can be written as “LS analogues”. Then ‘ $-2 \log \mathcal{L}(\theta)$ ’ gets replaced by ‘ $T \log \hat{\sigma}^2$ ’ where $\hat{\sigma}^2$ is the (ML!) variance of the fitted regression model (sum of squared errors divided by sample size).

Thus we get, e.g.,

▣ Akaike:

$$\text{AIC} = \frac{2}{T} \times \text{number of parameters} + \log \hat{\sigma}^2$$



The number of parameters always includes the constant and the variance estimate!



Box-Jenkins' approach to ARMA modeling

1. choose the maximal p and q from studying ACF and PACF plots.
2. fit chosen model and record information criteria.
3. reduce the model size and check whether any important information is lost.
4. check residuals for serial correlation, nonnormality, heteroskedasticity, i.e. white noise properties.

The overall purpose is to find a parsimonious model. Possibly evaluate out of sample.



Forecast evaluation

Mincer-Zarnowitz (MZ) regressions are used to check the the forecast. Given forecasts $\hat{x}_{t+h|t}$ and realized values x_{t+h} , one performs the regression

$$x_{t+h} = \alpha + \beta \hat{x}_{t+h|t} + u_{t+h} .$$

An optimal forecast has $\beta = 1$ and $\alpha = 0$ which can be assessed by standard tests, such as the F -test, but note that u_{t+h} may display serial correlation!

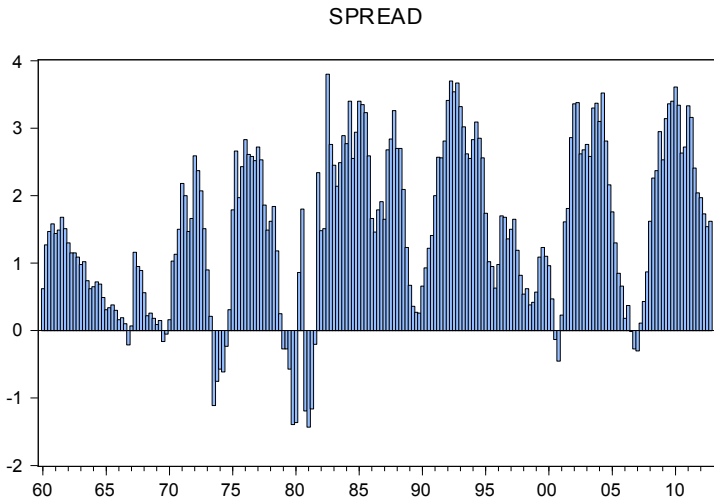
If you have two competing models, use the **Diebold-Mariano (DM) test** in addition.



Example

The following slides demonstrate the Box-Jenkins modeling approach on the SPREAD data; this is a quarterly spread defined as US-10yr-rate minus the T-bill rate. The data are included in from QUARTERLY.XLS downloaded from Enders' web site at <http://time-series.net/home>.













































Spread data look pretty stationary.

Correlogram of SPREAD

Date: 03/28/17 Time: 09:31
Sample: 1960Q1 2012Q4
Included observations: 212

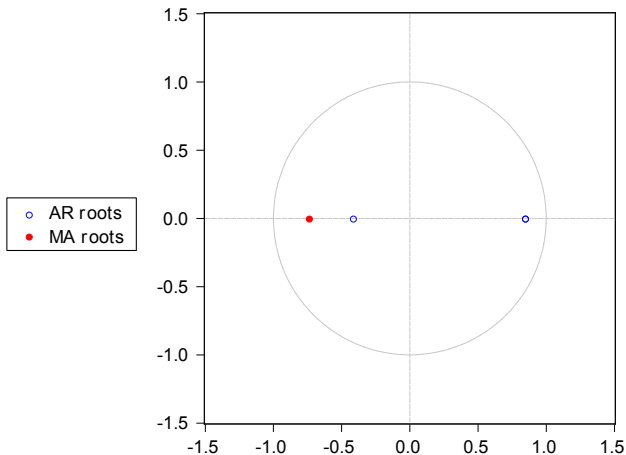
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.889	0.889	170.08	0.000
		2 0.744	-0.225	289.69	0.000
		3 0.626	0.078	374.77	0.000
		4 0.496	-0.179	428.53	0.000
		5 0.377	0.014	459.64	0.000
		6 0.247	-0.192	473.08	0.000
		7 0.144	0.104	477.69	0.000
		8 0.078	-0.006	479.05	0.000
		9 -0.004	-0.143	479.06	0.000
		10 -0.102	-0.140	481.39	0.000
		11 -0.183	-0.020	488.98	0.000
		12 -0.217	0.121	499.70	0.000
		13 -0.242	-0.102	513.03	0.000
		14 -0.261	0.046	528.68	0.000
		15 -0.255	0.014	543.69	0.000
		16 -0.227	0.039	555.65	0.000
		17 -0.183	0.004	563.42	0.000
		18 -0.116	0.172	566.59	0.000
		19 -0.049	0.014	567.15	0.000
		20 0.009	-0.056	567.16	0.000

Persistent correlations. ACF and PACF suggest 1-2 AR terms; alternating sign in PACF hints to positive MA(1).

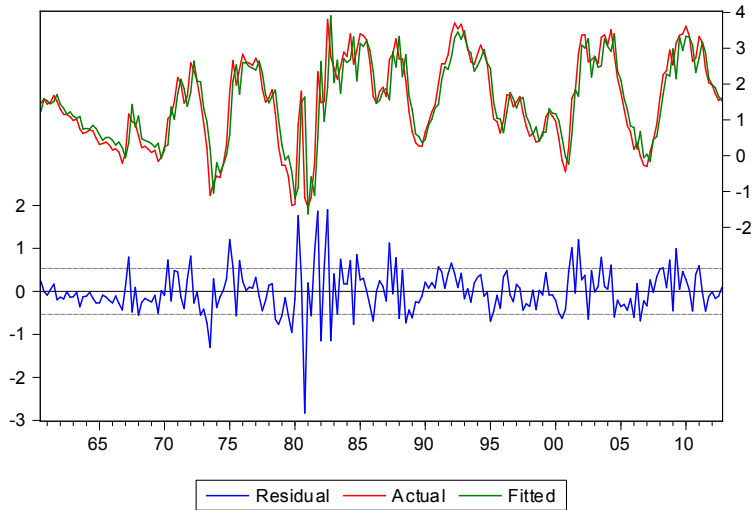
Dependent Variable: SPREAD Method: Least Squares Date: 03/28/17 Time: 09:56 Sample (adjusted): 1960Q3 2012Q4 Included observations: 210 after adjustments Convergence achieved after 33 iterations MA Backcast: 1960Q2				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.528631	0.308231	4.959367	0.0000
AR(1)	0.441202	0.142577	3.094481	0.0022
AR(2)	0.353201	0.135937	2.598263	0.0100
MA(1)	0.725486	0.109990	6.595960	0.0000
R-squared	0.814270	Mean dependent var	1.514143	
Adjusted R-squared	0.811565	S.D. dependent var	1.226845	
S.E. of regression	0.532563	Akaike info criterion	1.596632	
Sum squared resid	58.42634	Schwarz criterion	1.660386	
Log likelihood	-163.6463	Hannan-Quinn criter.	1.622405	
F-statistic	301.0451	Durbin-Watson stat	1.994311	
Prob(F-statistic)	0.000000			
Inverted AR Roots	.85	-.41		
Inverted MA Roots	-.73			

Indeed an ARMA(2,1) delivers the best information criteria. Estimated Model is stationary and invertible! See also next graph. Durbin-Watson looks good: no first-order serial correlation in residuals.

Inverse Roots of AR/MA Polynomial(s)



Visual confirmation of stationarity and invertibility.



Wondering about high R^2 ? Here is why: model fits very well.









































Correlogram of Residuals

Date: 03/28/17 Time: 10:08

Sample: 1960Q3 2012Q4

Included observations: 210

Q-statistic probabilities adjusted for 3 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.002	0.002	0.0011	
		2 0.042	0.042	0.3818	
		3 0.070	0.070	1.4364	
		4 -0.002	-0.004	1.4371	0.231
		5 0.071	0.065	2.5255	0.283
		6 -0.054	-0.060	3.1675	0.367
		7 -0.145	-0.152	7.8059	0.099
		8 0.109	0.108	10.437	0.064
		9 0.005	0.026	10.441	0.107
		10 -0.013	-0.008	10.476	0.163
		11 -0.218	-0.237	21.126	0.007
		12 -0.005	0.015	21.131	0.012
		13 -0.050	-0.060	21.698	0.017
		14 -0.092	-0.076	23.605	0.015
		15 -0.101	-0.072	25.946	0.011
		16 -0.053	-0.017	26.587	0.014
		17 -0.099	-0.126	28.835	0.011
		18 -0.003	-0.058	28.837	0.017
		19 0.030	0.111	29.049	0.024
		20 0.034	0.048	29.314	0.032

Are residuals WN? Initially ok, but from lag 11 on Q statistic rejects... How serious is that?



































Correlogram of Residuals Squared

Date: 03/28/17 Time: 10:08

Sample: 1960Q3 2012Q4

Included observations: 210

Q-statistic probabilities adjusted for 3 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.059	0.059	0.7455	
		2	0.292	0.289	18.953	
		3	0.149	0.131	23.714	
		4	0.332	0.268	47.491	0.000
		5	0.104	0.036	49.838	0.000
		6	0.093	-0.083	51.719	0.000
		7	0.306	0.236	72.249	0.000
		8	0.111	0.021	74.988	0.000
		9	0.108	-0.051	77.562	0.000
		10	0.005	-0.094	77.569	0.000
		11	0.034	-0.178	77.826	0.000
		12	-0.000	-0.052	77.826	0.000
		13	-0.013	0.003	77.864	0.000
		14	-0.025	-0.066	78.007	0.000
		15	-0.001	0.021	78.007	0.000
		16	0.012	0.049	78.038	0.000
		17	-0.042	0.011	78.445	0.000
		18	-0.035	0.047	78.733	0.000
		19	-0.010	0.043	78.757	0.000
		20	-0.019	0.003	78.842	0.000

As seen for squared residuals, there are potential heteroskedasticity effects. Nonrobust confidence intervals are misleading. Seems likely one cannot reject WN assumption using robust tests.

Chapter 5:

Unit root processes



Unit root processes

The case

$$X_{t-1} = \phi X_t + \varepsilon_t$$

where $\phi = 1$ and ε is WN is special:

1. no stationary solutions, neither causal no non-causal ones.
2. distributional result of the OLS estimator

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1 - \phi^2)$$

breaks down!

This is called a **unit root process**, a **random walk**, or a **stochastic trend model**.



If we know with certainty that the process has a unit root, we can remove it by first differences. Clearly

$$(1 - L)y_t \equiv \Delta y_t = \varepsilon_t$$

is a stationary.

It is not mandatory that ε is WN. Instead, ε could be a stationary ARMA process.

This gives to an extended class of models. We say X is

$$\mathbf{ARIMA}(p, d, q)$$

if d times differencing leads to a stationary ARMA(p,q) process.



X is $\text{ARIMA}(p, d, q)$ if the representation

$$(1 - L)^d \phi(L) X_t = \theta(L) \varepsilon_t$$

is a stationary $\text{ARMA}(p, q)$ process. Technically, we ‘factor’ the unit root from the overall lag polynomial.

We can also write

$$\Delta^d \phi(L) X_t = \theta(L) \varepsilon_t$$

We take d as an integer (mostly $d = 1$).



The bottom line is this:

If we know with certainty that there are d unit roots, we

- difference d times
- and model the remaining stationary parts components.

- For forecasting, we add the forecast of the stationary part to the forecast of the trend;
- the forecast error variances need to be adjusted such as to reflect the error from predicting the trend and the stationary part.

Practically, we are back to the previous chapter. All we need to know is how to detect a unit root....



OLS estimators and unit roots

Consider the simplest case

$$X_{t-1} = \phi X_t + \varepsilon_t$$

where $\varepsilon \sim (0, \sigma_\varepsilon^2)$ is strict Gaussian noise and $\phi = 1$.

To study the behavior of OLS estimator we study the behavior of

$$\hat{\phi} = \frac{\sum_{t=1}^T X_{t-1} X_t}{\sum_{t=1}^T X_{t-1}^2} = 1 + \frac{\sum_{t=1}^T X_{t-1} \varepsilon_t}{\sum_{t=1}^T X_{t-1}^2}$$



We would like to do this in two steps. We study $\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t$ and $\frac{1}{T} \sum_{t=1}^T X_{t-1}^2$.

Assume that $X_0 = 0$. Recall that

$$X_T = \sum_{t=1}^T \varepsilon_t.$$

This suggests that our previous results, such as

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_{t-1}^2 = \gamma_0$$

are doomed to fail. In particular, normalizing by $1/T$ is problematic. In fact, $1/T^2$ is more suggestive!



For analysis introduce the auxiliary random variable

$$Y_T(r) = \frac{1}{T} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t$$

which is indexed by $r \in [0, 1]$. Here $\lfloor u \rfloor$, $u \in \mathbb{R}$, denotes the largest integer less or equal to u (floor).

For any given realization, $Y_T(r)$ is a step function.



For the moment, assume that $r = 0, \frac{1}{T}, \frac{2}{T}, \dots, 1$. The following properties of $Y_T(r)$ are immediate:

1.

$$Y_T(0) = 0.$$

2. For $r_1 < r_2 < r_3$,

$$Y_T(r_2) - Y_T(r_1) \quad \text{is independent of} \quad Y_T(r_3) - Y_T(r_2)$$

3.

$$\sqrt{T}Y_T(r) \sim \mathcal{N}(0, r\sigma_\varepsilon^2)$$

4. For $r_1 < r_2$

$$\sqrt{T} (Y_T(r_2) - Y_T(r_1)) / \sigma_\varepsilon \sim \mathcal{N}(0, r_2 - r_1)$$



Now consider what happens as $T \rightarrow \infty$. $Y_T(r)$ still ranges in $[0, 1]$, but as more and more ε_t are drawn, it becomes more and more tightly packed.

Eventually the discrete jumps start to disappear, leaving a random function, which is continuous, but very rough in $r \in [0, 1]$.

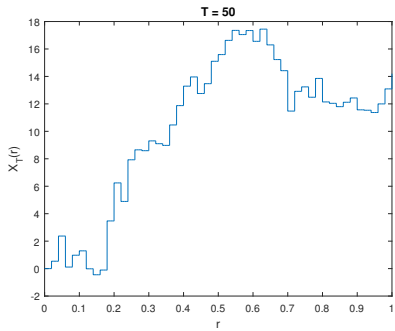
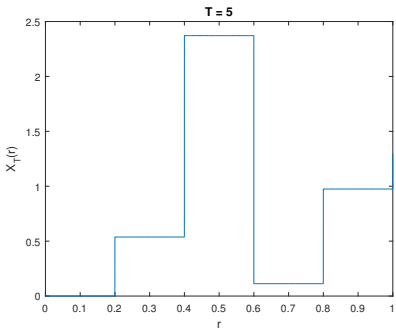
The limit can be characterized. It is a continuous time process on $[0, 1]$, known as the **Wiener process** (or Brownian motion), i.e.,

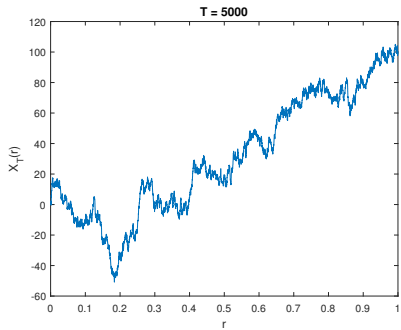
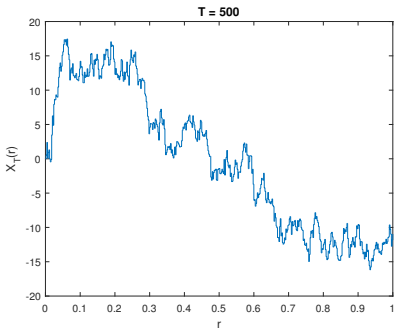
$$\sqrt{T}Y_T(r)/\sigma_\varepsilon \xrightarrow{\mathcal{L}} W_r$$

on $r \in [0, 1]$.

The Wiener process is a “continuous time random walk”.







From this, we can conclude

$$\sqrt{T}Y_T(1)/\sigma_\varepsilon = \frac{1}{\sqrt{T}\sigma_\varepsilon} \sum_{t=1}^T \varepsilon_t \xrightarrow{\mathcal{L}} W_1 \sim \mathcal{N}(0, 1).$$

... which does not really justify our efforts

But consider, e.g., evaluating $\int_0^1 [\sqrt{T}Y_T(r)]^2 dr$ as $T \rightarrow \infty$.

Because this is a continuous transformation, the **continuous mapping theorem** allows us to conclude

$$\int_0^1 [\sqrt{T}Y_T(r)]^2 dr \xrightarrow{\mathcal{L}} \sigma_\varepsilon^2 \int_0^1 (W_r)^2 dr$$



We evaluate the asymptotics of the OLS estimator, beginning with the denominator. Recall $Y_T(r) = \frac{1}{T} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t$ and $X_s = \sum_{t=1}^s \varepsilon_t$.

Define $S_T(r) = [\sqrt{T}Y_T(r)]^2$. Notice that

$$\int_0^1 S_T(r) dr = \frac{1}{T} \left(\frac{X_1^2}{T} + \frac{X_2^2}{T} + \dots + \frac{X_{T-1}^2}{T} \right)$$

Using the earlier result we have

$$\int_0^1 S_T(r) dr = \frac{1}{T^2} \sum_{t=1}^T X_t^2 \xrightarrow{\mathcal{L}} \sigma_\varepsilon^2 \int_0^1 W_r^2 dr$$

Only if we divide the squared sum of the random walk by T^2 we get a well-behaved convergence in law to a positive rv.



Let's tackle the numerator $\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t$. Note that

$$\begin{aligned} X_t &= \phi X_{t-1} + \varepsilon_t \\ \Rightarrow X_t^2 &= \phi^2 X_{t-1}^2 + 2\phi X_{t-1} \varepsilon_t + \varepsilon_t^2 \\ \text{or: } X_t^2 &= X_{t-1}^2 + 2X_{t-1} \varepsilon_t + \varepsilon_t^2 \end{aligned}$$

because $\phi = 1$ under the null.

With repeated substitution, dividing by T and $X_0 = 0$ we get

$$\frac{1}{T} X_T^2 = \frac{2}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t + \frac{1}{T} \sum_{t=1}^T \varepsilon_t^2$$

and thus

$$\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t = \frac{1}{2T} X_T^2 - \frac{1}{2T} \sum_{t=1}^T \varepsilon_t^2$$

.... but these are old friends!



We find

$$\frac{1}{2T}X_T^2 = \frac{1}{2}S(1) \xrightarrow{\mathcal{L}} \frac{1}{2}\sigma_\varepsilon^2 W_1^2$$

The remaining term is easy, for

$$p\text{-}\lim_{T \rightarrow \infty} \frac{1}{2T} \sum_{t=1}^T \varepsilon_t^2 = \frac{1}{2}\sigma_\varepsilon^2$$

which follows from the iid. assumption. Hence

$$\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t = \frac{1}{2T} X_T^2 - \frac{1}{2T} \sum_{t=1}^T \varepsilon_t^2 \xrightarrow{\mathcal{L}} \frac{1}{2}\sigma_\varepsilon^2 W_1^2 - \frac{1}{2}\sigma_\varepsilon^2$$



Summarizing we can conclude

$$\begin{aligned} p\text{-}\lim_{T \rightarrow \infty} \hat{\phi} &= 1 + p\text{-}\lim_{T \rightarrow \infty} \frac{\sum_{t=1}^T X_{t-1} \varepsilon_t}{\sum_{t=1}^T X_{t-1}^2} \\ &= 1 + p\text{-}\lim_{T \rightarrow \infty} \frac{\frac{1}{T^2} \sum_{t=1}^T X_{t-1} \varepsilon_t}{\frac{1}{T^2} \sum_{t=1}^T X_{t-1}^2} = 1 \end{aligned}$$

We have $p\text{-}\lim_{T \rightarrow \infty} \frac{1}{T^2} \sum_{t=1}^T X_{t-1} \varepsilon_t = 0$ because
 $\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t \xrightarrow{\mathcal{L}} \text{'random variable'}$.

Thus OLS estimator consistent.



We can also deduce the distributional law. We have

$$T(\hat{\phi} - 1) = \frac{\frac{1}{T} \sum_{t=1}^T X_{t-1} \varepsilon_t}{\frac{1}{T^2} \sum_{t=1}^T X_{t-1}^2} \xrightarrow{\mathcal{L}} \frac{\frac{1}{2} (\sigma_\varepsilon^2 W_1^2 - \sigma_\varepsilon^2)}{\sigma_\varepsilon^2 \int_0^1 W_r^2 dr} = \frac{\frac{1}{2} (W_1^2 - 1)}{\int_0^1 W_r^2 dr}$$

If $\phi = 1$ the OLS estimator of ϕ is nonnormally distributed.

Because the rate of convergence is T as opposed to the usual \sqrt{T} one says the OLS estimator is super-consistent.



Testing for a unit root

The Dickey-Fuller test for a unit roots tests $H_0 : \phi = 1$ against $H_A : \phi < 1$. The test statistic looks like a 't-statistic':

$$DF = \frac{T(\hat{\phi} - 1)}{\sqrt{\text{Var}[T\hat{\phi}]}}$$

where

$$\text{Var}[T\hat{\phi}] = \frac{\hat{\sigma}_\varepsilon^2}{\frac{1}{T^2} \sum_{t=1}^T X_{t-1}^2}$$

where $\hat{\sigma}_\varepsilon^2$ is a consistent estimator of σ_ε^2 , obtained, e.g., by imposing $\phi = 1$ and forming $\hat{\sigma}_\varepsilon^2 = \frac{1}{T} \sum_{t=1}^T (\Delta X_t)^2$



Thus

$$DF = \frac{(\hat{\phi} - 1) \sqrt{\sum_{t=1}^T X_{t-1}^2}}{\hat{\sigma}_\varepsilon^2}$$

The distribution of the test statistic is again non-standard: The denominator converges to σ_ε^2 , but the numerator is a product of two random variables. So the distribution differs from that of $T(\hat{\phi} - 1)$.



What happens if we add additional deterministic regressors although not present in the true DGP?

For iid. or stationary data we wouldn't worry. But with unit roots we have to! Consider the estimator with intercept

$$\begin{aligned} T(\hat{\phi} - \phi) &= \frac{T^{-1} \sum_{t=1}^T (X_{t-1} - \bar{X})(X_t - \bar{X})}{T^{-2} \sum_{t=1}^T (X_{t-1} - \bar{X})^2} \\ &= \frac{T^{-1} \sum_{t=1}^T X_{t-1} X_t - \bar{X}^2}{T^{-2} \sum_{t=1}^T X_{t-1}^2 - T^{-1} \bar{X}^2} \end{aligned}$$



In

$$T(\hat{\phi} - \phi) = \frac{T^{-1} \sum_{t=1}^T X_{t-1} X_t - \bar{X}^2}{T^{-2} \sum_{t=1}^T X_{t-1}^2 - T^{-1} \bar{X}^2}$$

$T^{-1} \sum_{t=1}^T X_{t-1} X_t$ and $T^{-2} \sum_{t=1}^T X_{t-1}^2$ behave as discussed.

Under stationarity we would conclude $p\text{-lim}_{T \rightarrow \infty} \bar{X}^2 = 0$ and nothing would change.

But if X_t is $I(1)$, \bar{X} (and hence \bar{X}^2) converges to a random variable.



So the distribution of $T(\hat{\phi} - \phi)$ changes depending on your including the intercept or not, even though it is not present in the DGP!

Also the distribution of the intercept itself is nonstandard and has a DF-type distribution .

This issue continues to exist with other deterministic regressors such as, time trends, dummy variables.

In each case the limiting distribution changes.



Suppose the true DGP is

$$X_t = \phi_0 + \phi_1 X_{t-1} + \varepsilon_t$$

with ε_t strict Gaussian noise, i.e., RW with drift .

In the regression model, you include the intercept.

⚠ Then **standard limit theory applies**, i.e., both estimates $\hat{\phi}_0$ and $\hat{\phi}_1$ are asymptotically normal with usual standard errors, but with rates $T^{1/2}$ and $T^{3/2}$, respectively ...

Why? The deterministic regressor, which has a slower rate of convergence, namely, $T^{1/2}$, dominates the overall convergence behavior



Unit roots and non-iid. normal noise

Non-normal noise is not a big deal. This is because averaged non-normal variables act as if they were normal: in

$$X_t = t \left(\frac{1}{t} \sum_{j=1}^t \varepsilon_j \right)$$

the term in brackets tends to be normally distributed under moderate restrictions of the existence of the moments of ε .

This is not a particular issue.



If ε is serially correlated, e.g., if ε is the AR(1) process

$$\varepsilon_t = \beta_1 \varepsilon_{t-1} + \epsilon_t$$

where ϵ is WN, we get

$$X_t = \phi_1 X_{t-1} + \beta_1 \varepsilon_{t-1} + \epsilon_t$$

Under the null, we have $\Delta X_t = \varepsilon_t$. This suggests the regression

$$X_t = \phi_1 X_{t-1} + \beta_1 \Delta X_{t-1} + \epsilon_t$$

and then testing $H_0 : \phi = 1$.

Here limit theory stays the same as in the strict WN case, up to a scaling factor involving β_1 . Yet this factor cancels when computing the 't ratio' and we can use the DF distribution for inference.



Procedures including regressions of the type

$$X_t = \phi_1 X_{t-1} + \beta_1 \Delta X_t + \epsilon_t$$

are called **augmented DF (ADF) methods**.

They continue to work if ε is an $AR(p)$.

If ε is an invertible MA process, then ε has an $AR(\infty)$ representation. One continues using ADF methods, but uses an $AR(p)$ where p is of the order $T^{1/3}$.



If ε is non-invertible, a more general approach is needed.



Unit root testing is confusing. As a practitioner we would like to have some type of a cooking book recipe of how to proceed.

Indeed such recipes do exist, see Enders, Chapter 4 or Campbell & Perron's famous '24 Rules' published in *Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots*.

Yet, still the matter remains confusing and, worse, involves many issues, in particular, the notorious lack of power of unit root tests and multiple testing problems.

It is always worthwhile combine common sense and economic theory to exclude certain implausible hypotheses.

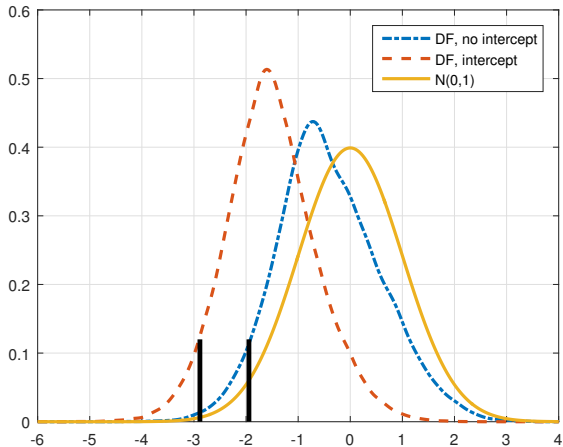


Table 4.2 Summary of the Dickey–Fuller Tests

Model	Hypothesis	Test Statistic	Critical Values for 95% and 99% Confidence Intervals
$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \varepsilon_t$	$\gamma = 0$	τ_τ	-3.45 and -4.04
	$\gamma = a_2 = 0$	ϕ_3	6.49 and 8.73
	$a_0 = \gamma = a_2 = 0$	ϕ_2	4.88 and 6.50
$\Delta y_t = a_0 + \gamma y_{t-1} + \varepsilon_t$	$\gamma = 0$	τ_μ	-2.89 and -3.51
	$a_0 = \gamma = 0$	ϕ_1	4.71 and 6.70
$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$	$\gamma = 0$	τ	-1.95 and -2.60

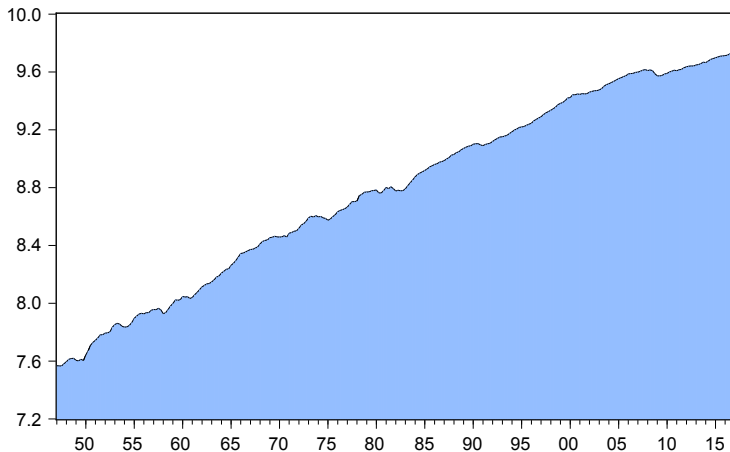
Note: Critical values are for a sample size of 100.

From Enders (2010), p. 208. τ denotes the various t -tests under the different settings; ϕ_i , $i = 1, 2, 3$, are F -tests for joint hypotheses. F -tests have a nonstandard distribution too.



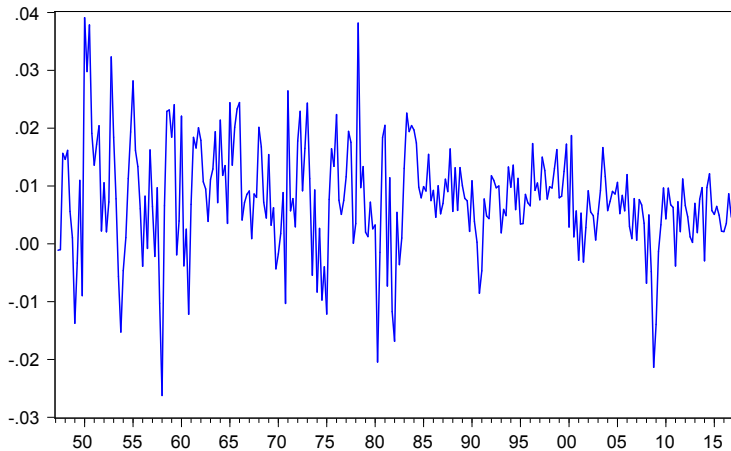
Simulated DF density with 5% critical values found at -1.944 and -2.886, respectively, based on 50 000 simulations.

LOG(GDP)



Log of Real GDP, Series ID: GDPC1, Source: U.S. Department of Commerce:
Bureau of Economic Analysis, Seasonal Adjustment: Seasonally Adjusted
Annual Rate, Frequency: Quarterly, Units: Billions of Chained 2005 Dollars

Growth rates



Log GDP is trending, growth rates are not. Combining this with economic plausibility checks we see: the only interesting cases are

1. a unit root with drift;
2. no unit root, but a trend component.

Thus we fit the model (now in Ender's notation)

$$\Delta X_t = a_0 + \gamma X_{t-1} + a_2 t + \varepsilon_t$$

and test

$$H_0 : a_2 = \gamma = 0$$

via an F -test. One could also only test $H_0 : \gamma = 0$ and – if we do not reject – conclude that $a_2 \neq 0$ because this is the only plausible alternative!



Dependent Variable: D(Y)

Method: Least Squares

Date: 03/22/17 Time: 18:04

Sample (adjusted): 1947Q3 2016Q4

Included observations: 278 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y(-1)	-0.010784	0.008222	-1.311577	0.1908
D(Y(-1))	0.361024	0.056604	6.378098	0.0000
C	0.089492	0.062905	1.422659	0.1560
@TREND	7.23E-05	6.61E-05	1.094394	0.2747
R-squared	0.157602	Mean dependent var		0.007780
Adjusted R-squared	0.148379	S.D. dependent var		0.009472
S.E. of regression	0.008741	Akaike info criterion		-6.627319
Sum squared resid	0.020935	Schwarz criterion		-6.575123
Log likelihood	925.1974	Hannan-Quinn criter.		-6.606379
F-statistic	17.08729	Durbin-Watson stat		2.071789
Prob(F-statistic)	0.000000			

Augmented Dickey-Fuller Unit Root Test on Y

Null Hypothesis: Y has a unit root Exogenous: Constant, Linear Trend Lag Length: 1 (Automatic - based on AIC, maxlag=15)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-1.311577	0.8829
Test critical values:	1% level	-3.991412		
	5% level	-3.426073		
	10% level	-3.136231		
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation Dependent Variable: D(Y) Method: Least Squares Date: 03/22/17 Time: 17:46 Sample (adjusted): 1947Q3 2016Q4 Included observations: 278 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y(-1)	-0.010784	0.008222	-1.311577	0.1908
D(Y(-1))	0.361024	0.056604	6.378098	0.0000
C	0.089492	0.062905	1.422659	0.1560
@TREND("1947Q1")	7.23E-05	6.61E-05	1.094394	0.2747
R-squared	0.157602	Mean dependent var	0.007780	
Adjusted R-squared	0.148379	S.D. dependent var	0.009472	
S.E. of regression	0.008741	Akaike info criterion	-6.627319	
Sum squared resid	0.020935	Schwarz criterion	-6.575123	
Log likelihood	925.1974	Hannan-Quinn criter.	-6.606379	
F-statistic	17.08729	Durbin-Watson stat	2.071789	
Prob(F-statistic)	0.000000			

ADF regression $\Delta X_t = a_0 + \gamma X_{t-1} + a_2 t + \varepsilon_t$ with DF t -test on $H_0 : \gamma = 0$ against $H_A : \gamma < 0$, i.e., test τ_τ in Ender's notation.

Chapter 6:

Multivariate time series models



VARs and VARMA

The theory of stationary ARMA models can be extended to vector-valued processes. This gives rise to vector-valued AR processes (VAR) and vector-valued ARMA processes (VARMA).

The topics of analysis stay the same: stationary solutions, ergodicity, invertibility, prediction, estimation, and inference.

VARMA's are used to study the dynamic relations of economic variables simultaneously, e.g., how shocks to one variable are transmitted to others.

VARs are found a lot, VARMA less so: they are difficult to estimate and complicated identifiability issues arise.



VAR

The k variable VAR is defined by

$$\Phi(L)X_t = \varepsilon_t$$

where

$$\Phi(L) = I_k - \Phi_1 L - \Phi_2 L^2 - \dots - \Phi_p L^p$$

and I_k is the identity matrix of dimension k , Φ_j , $j = 1, \dots, k$, are $k \times k$ coefficient matrices, ε_t vector valued $WN(0, \Sigma_\varepsilon)$, and

$$X_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{k,t} \end{pmatrix}, \varepsilon_t = \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{k,t} \end{pmatrix}$$



Stationarity

The VAR

$$\Phi(L)X_t = \varepsilon_t$$

is stationary if all $z \in \mathbb{C}$ satisfying

$$\Phi(z) = \left| I_k - \Phi_1 z - \Phi_2 z^2 - \dots - \Phi_p z^p \right| = 0$$

lie outside the unit circle. Here $|A|$ means determinant of matrix A .

Then $\Phi(L)$ is invertible and X has the VMA(∞) representation

$$X_t = \Psi(L)\varepsilon_t$$

with

$$\Psi(L) = I_k + \Psi_1 L + \Psi_2 L^2 \dots$$

and $\Psi(1)$ is absolutely summable (elementwise).



We have

$$E[X_t] = 0$$

and the matrix valued acf is

$$\Gamma_h \equiv \text{Cov}[X_t, X_{t-h}] = E[X_t X_{t-h}^\top]$$

Formulae for Γ_h can be derived ...

Note that $\Gamma_h \neq \Gamma_{-h}$ but

$$\Gamma_h^\top = \Gamma_{-h}$$

which follows from the definition of the acf.



Estimation

If ε_t is multivariate normal, ML estimates can be obtained by OLS regressions, *equation by equation*.

Estimates are consistent and asymptotically normal, but OLS regressions are only efficient if all variables appear in all equations.

If this is not the case, you can obtain efficient estimates using the seemingly unrelated regressions (SUR) approach.



VARs and policy analysis

VARs are used to study how shocks to the economic variables are absorbed and how they propagate through the system. To this end one studies the impulse response functions (IRF).

From the VMA(∞) representation

$$X_t = \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \Psi_2 \varepsilon_{t-2} \dots$$

we see that the Ψ_i have the interpretation

$$\frac{\partial X_{t+s}}{\partial \varepsilon_t^\top} = \Psi_s$$

i.e., $\psi_{ij,s}$, the row i , column j element of Ψ_s , summarizes how a 1-unit increase of $\varepsilon_{j,t}$ impacts $X_{i,t+s}$ holding all other $\varepsilon_{j',t}$, $j' \neq j$ constant for any t .



A plot of the row i -column j element of Ψ_s , i.e.,

$$\psi_{ij,s}$$

as a function of s is called **impulse response function (IRF)**.

This gives visual impression of the dynamic interrelationships within the system.

The IRF describes the response of $X_{i,t+s}$ to a one-time impulse in $X_{j,t}$ with all other variables dated in t or later held constant.



Does this interpretation make any sense? After all we are dealing with random variables which often are correlated!

For this reasons one prefers to think about the IRFs in terms of forecast revisions, i.e., optimal linear forecasts given the data history and given a shock in the variable of interest.

Under which conditions do we get this interpretation?

Only if ε_t is uncorrelated, i.e., $\Sigma_\varepsilon = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$.

Otherwise, if $\varepsilon_{1,t}$ receives a positive shock we could get useful information from the other shocks $\varepsilon_{2,t}, \dots, \varepsilon_{k,t}$.

With cross-sectionally correlated ε_t an IRF can be misleading.



How to get orthogonal errors? One way is to find a matrix P such that

$$\Sigma_{\varepsilon} = PP^{\top}$$

where P is a lower triangular matrix. This decomposition is called Cholesky decomposition.

Then we tweak the VMA(∞) representation such that

$$\begin{aligned} X_t &= \Psi(L)\varepsilon_t \\ &= \Psi(L)PP^{-1}\varepsilon_t \\ &= B(L)\epsilon_t \end{aligned}$$

where $B(L) = \Psi(L)P$ is a new matrix valued lag polynomial and $\epsilon_t = P^{-1}\varepsilon_t$.



Did we achieve our goal?

$$\begin{aligned} E[\epsilon_t \epsilon_t^\top] &= E[P^{-1} \epsilon_t (P^{-1} \epsilon_t)^\top] \\ &= P^{-1} E[\epsilon_t \epsilon_t^\top] (P^{-1})^\top \\ &= P^{-1} \Sigma_\epsilon (P^{-1})^\top \\ &= P^{-1} P P^\top (P^{-1})^\top = I_k \end{aligned}$$

Yes. Thus we build on $B(L) = \Psi(L)P$, i.e., we use

$$\begin{aligned} B_0 &= P \\ B_1 &= \Psi_1 P \\ B_2 &= \Psi_2 P \\ B_3 &= \Psi_3 P \\ &\dots \end{aligned}$$

to construct the IRF.



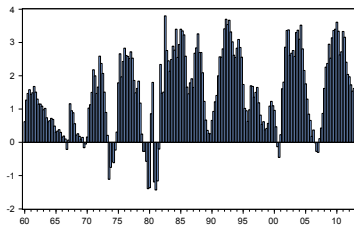
Example

The following pics show the regression results and graphs of a 3-variable VAR estimated on quarterly data of

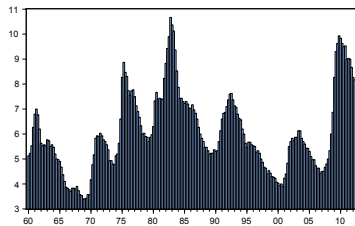
- ▣ the spread of the 10yr rate minus the T-Bill rate
- ▣ the differences log unemployment
- ▣ the differences log industrial production



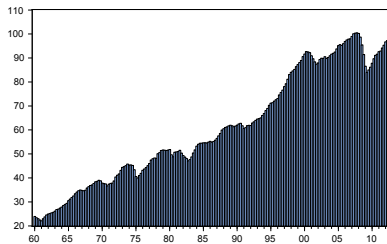
SPREAD



UNEMP



INDPROD

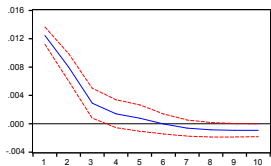


Date: 03/24/17 Time: 08:55
Sample (adjusted): 1961Q1 2012Q4
Included observations: 208 after adjustments
Standard errors in () & t-statistics in []

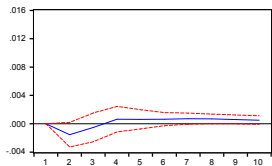
	INDPROD_DL	UNEMP_DLO	SPREAD
INDPROD_DLOG(-1)	0.525888 (0.09803) [5.36456]	-1.144112 (0.29096) [-3.93224]	-0.722572 (4.10225) [-0.17614]
INDPROD_DLOG(-2)	-0.143322 (0.10233) [-1.40056]	0.201906 (0.30373) [0.66477]	2.419961 (4.28228) [0.56511]
INDPROD_DLOG(-3)	0.123363 (0.09804) [1.25824]	-0.265016 (0.29100) [-0.91071]	-1.052442 (4.10286) [-0.25651]
UNEMP_DLOG(-1)	-0.061022 (0.03274) [-1.86401]	0.296267 (0.09716) [3.04912]	3.491498 (1.36994) [2.54865]
UNEMP_DLOG(-2)	0.019913 (0.03333) [0.59744]	0.012992 (0.09893) [0.13133]	-1.357446 (1.39481) [-0.97321]
UNEMP_DLOG(-3)	0.018998 (0.03026) [0.62788]	0.016266 (0.08980) [0.18113]	2.118226 (1.26617) [1.67295]
SPREAD(-1)	0.001334 (0.00169) [0.79077]	-0.003540 (0.00501) [-0.70708]	1.042248 (0.07060) [14.7635]
SPREAD(-2)	0.000805 (0.00241) [0.33418]	0.000839 (0.00715) [0.11738]	-0.314462 (0.10075) [-3.12106]
SPREAD(-3)	-0.000401 (0.00171) [-0.23418]	-0.007318 (0.00509) [-1.43906]	0.164659 (0.07170) [2.29659]
C	0.000980 (0.00187) [0.52260]	0.023952 (0.00556) [4.30426]	0.150620 (0.07846) [1.91976]
R-squared	0.402399	0.505470	0.829246
Adj. R-squared	0.375235	0.482992	0.821484
Sum sq. resids	0.030673	0.270210	53.71412
S.E. equation	0.012447	0.036942	0.520849
F-statistic	14.81384	22.48671	106.8402
Log likelihood	622.3379	396.0547	-154.3375
Akaike AIC	-5.887864	-3.712064	1.580169
Schwarz SC	-5.727405	-3.551606	1.740627
Mean dependent	0.007081	0.001068	1.514038
S.D. dependent	0.015747	0.051377	1.232746
Determinant resid covariance (dof adj.)	2.84E-08		
Determinant resid covariance	2.45E-08		
Log likelihood	937.0692		
Akaike information criterion	-8.721820		

Response to Cholesky One S.D. Innovations – 2 S.E.

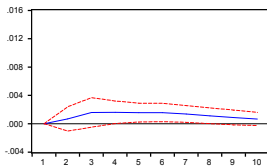
Response of INDPROD_DLOG to INDPROD_DLOG



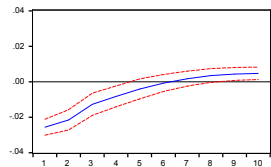
Response of INDPROD_DLOG to UNEMP_DLOG



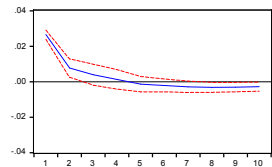
Response of INDPROD_DLOG to SPREAD



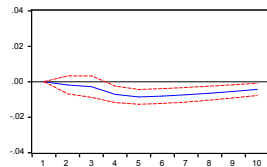
Response of UNEMP_DLOG to INDPROD_DLOG



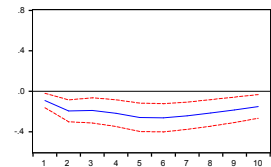
Response of UNEMP_DLOG to UNEMP_DLOG



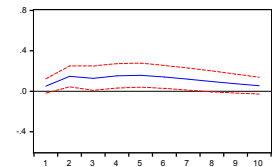
Response of UNEMP_DLOG to SPREAD



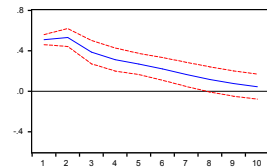
Response of SPREAD to INDPROD_DLOG



Response of SPREAD to UNEMP_DLOG



Response of SPREAD to SPREAD



IRFs may yield telling insights. Yet there is an issue.

We used the Cholesky decomposition to orthogonalize the errors. Because of the lower triangular structure (which is preserved in the inverse), this choice entails very specific assumptions on how errors affect the variables in the system.

In other words: the ordering of the variables in the VARs matters!

One needs to have strong economic grounds to justify a specific ordering choice.



But it comes even worse: Consider a matrix R with the property

$$R^{\top} = R^{-1}.$$

Then $RR^{\top} = I$. Such a matrix is called **rotation matrix**.

Reconsidering the decomposition and using R we get

$$\begin{aligned}\Sigma_{\epsilon} &= PP^{\top} \\ &= PRR^{\top}P^{\top} \\ &= \tilde{P}\tilde{P}^{\top}\end{aligned}$$

where $\tilde{P} = PR$. This also is a valid decomposition, which belongs to another orthogonal WN sequence $\tilde{\epsilon}_t$.



Unfortunately, infinitely many matrices of the type R exist. They just describe different rotations of the data space.

As consequence, IRFs are not identified.

Due to this fact, the literature nowadays focuses on developing structural VAR (SVAR) models. The aim is to find conditions, rooted in economic theory and/or additional external information, for instance, provided by additional data, to pin down the decompositions of the type

$$\Sigma_{\varepsilon} = \Sigma^{1/2} \Sigma^{1/2}$$

uniquely.



Granger causality

Often one likes to understand how variables interact in VARs using the *temporal* causality concept due to Clive Granger:

We say X **Granger-causes** Y if

$$P(Y_{t+1} \in A | \mathcal{F}_t) \neq P(Y_{t+1} \in A | \mathcal{F}_t^{-X})$$

where \mathcal{F}_t^{-X} is the information set without X . Thus, the cause

1. happens prior to its effect;
2. has unique information about the future values of its effect.



Causality in a *post hoc ergo propter hoc* sense.



One way to assess Granger causality within the VAR

$$\Phi(L)X_t = \varepsilon_t$$

is to test whether the off-diagonal lag polynomial have zero coefficients. So if a test on zero coefficients in $\phi_{12}(L)$ cannot be rejected, then $X_{2,t}$ does not Granger-cause $X_{1,t}$.

Can be done by means of the F -test

$$F = \frac{(RSS_R - RSS_{UR})/r}{RSS_{UR}/(T - r)}$$

where $RSS_{(U)R}$ is the residual sum of square of the (un)restricted model, r is the number of restrictions and T sample size.



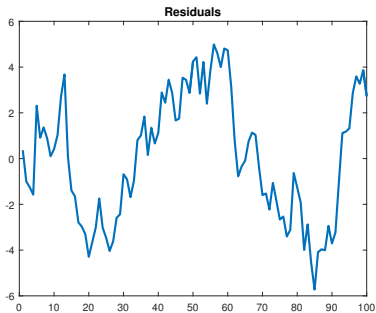
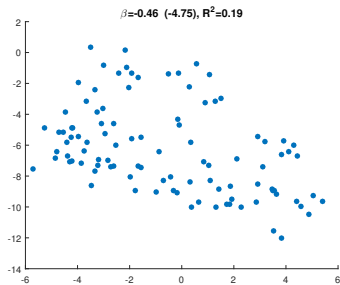
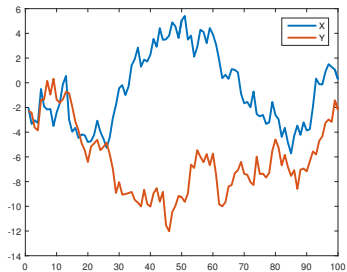
Regressions with integrated regressors

As long as sequences X and Y are both stationary, standard regression theory applies. So no particular issues arise in regressions of the type

$$X_t = \beta Y_t + \varepsilon_t$$

What happens if X and/or Y have a unit root?





Simulation and regression of two independent RW.

Observations:

- significant negative relationship
- high R^2
- residuals look nonstationary ... imposing $X_0 = Y_0 = 0$ shows why:

$$\varepsilon_t = X_t - \beta Y_t = \sum_{i=1}^T \varepsilon_{1,i} - \beta \sum_{i=1}^T \varepsilon_{2,t}$$

Results would even be more striking if the RWs had a drift.

Because the two RW were simulated independently from each other these regressions are meaningless; moreover, inference invalid due to nonstationary residuals:
spurious regressions problem.



If both X or Y are integrated of same order, one can difference at continue to work with the differenced series (which are stationary).

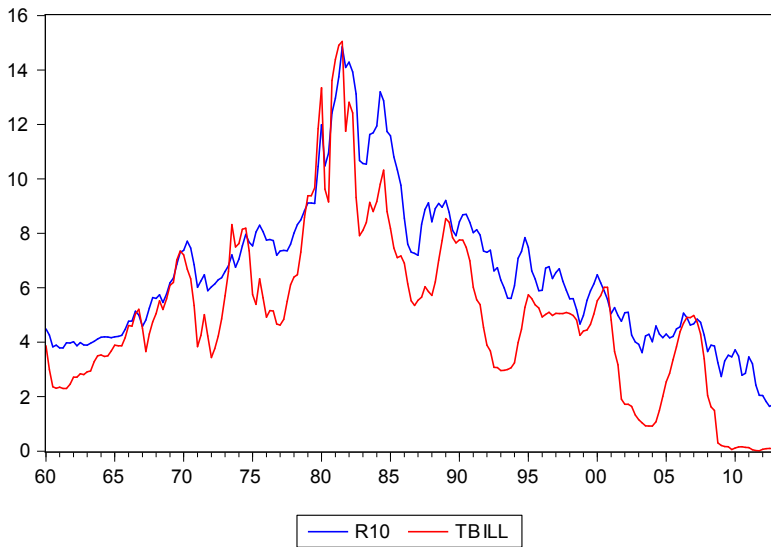
If either X or Y are stationary, any regressions are meaningless for the same reason. These are again spurious regressions.

There is fourth case:

Both X or Y are integrated of same order, but the regression residual sequence is stationary. This happens if X and Y form an economic equilibrium relationship. Both X and Y may trend, but given the equilibrium relationship, deviations from this equilibrium can only be of transitory nature.

X and Y are said to be **cointegrated**.





In fact, economic theory is rich of relationships, where variables are $I(1)$ but a linear combinations are $I(0)$

- permanent income model: consumption and income
- Fisher equation: nominal interest rates and inflation
- money demand: money, income, interest rates, prices
- term structure models: rates at different maturities
- purchasing power parity: fx rates and price levels of two countries



Definition of **cointegration**:

The components of the k -dimensional vector X are called co-integrated of order d, b , denoted $CI(d, b)$, if

1. all components of X are $I(d)$;
2. $\exists \beta \neq 0$ such that $\{Z_t = \beta^\top X_t, t \in \mathbb{Z}\}$ is $I(d - b)$ where $b > 0$.

The vector β is called the **co-integrating vector**.

The case most often seen in practice is $CI(1, 1)$.



- If β is a cointegrating vector, so is $\lambda\beta$, where $\lambda \in \mathbb{R}$.
Therefore one usually normalizes β such that

$$\beta = (1, \beta_2, \beta_3, \dots, \beta_k)^\top$$

- Two series of different orders of integration cannot be cointegrated.
- If X has k nonstationary components it can have up to $k - 1$ linearly independent cointegrating vectors. The number of cointegrating vectors is called the **cointegration rank** of X .
- Let r be the cointegration rank of X . The number of common stochastic trends (unit-roots) that remain is $k - r$.



X is said to have an **error correction representation** or **error correction model (ECM)** if it can be expressed as

$$\Phi(L)\Delta X_t = \alpha Z_{t-1} + \epsilon_t$$

where ϵ_t is stationary, $\Phi_0 = I_k$, $\Phi(1)$ finite, $Z_t = \beta^\top X_t$, and $\alpha \neq 0$.

The ECM formalizes how X_t changes in response to stationary stochastic shocks and in response to the previous period's deviation from the equilibrium captured in Z_{t-1}

α is the “speed of adjustment”: it tells us how quickly l.h.s. variables adjust to deviations from long-run path.



Example

Consider a case with $k = 2$, $r = 1$, $k - r = 1$ common trends.

$$X_{1,t} = \beta_2 X_{2,t} + u_{1,t}$$

$$X_{2,t} = X_{2,t-1} + u_{2,t}$$

where u_i is WN.

We see

- $X_{2,t}$ is the common stochastic trend (pure unit root process)
- $\beta = (1, -\beta_2)^\top$
- $X_{1,t}, X_{2,t}$ is $I(1)$, but $\beta^\top X_t = u_{1,t}$ is $I(0)$



To find the ECM of

$$X_{1,t} = \beta_2 X_{2,t} + u_{1,t}$$

$$X_{2,t} = X_{2,t-1} + u_{2,t}$$

subtract $X_{1,t-1}$ and add $+\beta_2 X_{2,t-1} - \beta_2 X_{2,t-1}$ to the first eq.

This yields

$$\Delta X_{1,t} = \beta_2 \Delta X_{2,t} - (X_{1,t-1} - \beta_2 X_{2,t-1}) + u_{1,t}$$

$$\Delta X_{2,t} = u_{2,t}$$

or

$$\Delta X_{1,t} = -(X_{1,t-1} - \beta_2 X_{2,t-1}) + u_{1,t} + \beta_2 u_{2,t}$$

$$\Delta X_{2,t} = u_{2,t}$$



The ECM can be written as

$$\begin{pmatrix} \Delta X_{1,t} \\ \Delta X_{2,t} \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} (1, -\beta_2)^\top \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} \xi_{1,t} \\ \xi_{2,t} \end{pmatrix}$$

with $\xi_{1,t} = u_{1,t} + \beta_2 u_{2,t}$ and $\xi_{2,t} = u_{2,t}$ and where $\beta = (1, -\beta_2)^\top$ and $\alpha = (-1, 0)^\top$.

From

$$\begin{pmatrix} -1 \\ 0 \end{pmatrix} (1, -\beta_2)^\top = \begin{pmatrix} -1 & \beta_2 \\ 0 & 0 \end{pmatrix}$$

it is evident that cointegration relationships imply certain rank restrictions on VAR models.



Engle-Granger 2-step approach to estimating cointegration relations

Assume: X, Y are $I(1)$, (X, Y) are $CI(1, 1)$. Know lhs., here Y .

1. Run the regression

$$Y_t = c + \beta_2 X_t + u_t.$$

2. Form residuals $\hat{u}_t = Y_t - \hat{c} - \hat{\beta}_2 X_t$.
3. Test for a unit root in \hat{u} using an (A)DF regression

$$\Delta \hat{u}_t = \rho \hat{u}_{t-1} + v_t.$$



Cannot use DF test critical values because \hat{u}_t are estimated! Prefer the *MacKinnon corrected* critical values.

If you reject $H_0 : \rho = 0$, conclude that u_t is $I(0)$. Cointegration relationship is given by $\hat{\beta} = (1, -\hat{\beta}_2)^\top$ plus \hat{c} .



4. \hat{u}_t is the disequilibrium error. Estimate the ECM with OLS

$$\Phi_y(L)\Delta Y_t = c_y + \alpha_y \hat{u}_{t-1} + v_{1,t}$$

$$\Phi_x(L)\Delta X_t = c_x + \alpha_x \hat{u}_{t-1} + v_{2,t}$$

5. Evaluate model adequacy: the estimated parameter (α_y, α_x) should be negative such that they can be interpreted as the speed of adjustment.

If, say, $\alpha_x = 0$, then X is said to be **weakly exogenous**. In this example, X collapses to a random walk model.

Disadvantage of the EG 2 step method: Can only test for one cointegration relationship. There are more powerful and sophisticated testing procedure (Johansen tests).



Engle-Granger Cointegration Test

Date: 03/25/17 Time: 16:08

Series: R10 TBILL

Sample: 1960Q1 2012Q4

Included observations: 212

Null hypothesis: Series are not cointegrated

Cointegrating equation deterministics: C

Automatic lags specification based on Schwarz criterion (maxlag=14)

Dependent	tau-statistic	Prob.*	z-statistic	Prob.*
R10	-3.888447	0.0120	-30.66695	0.0044
TBILL	-4.273403	0.0035	-36.82778	0.0010

*MacKinnon (1996) p-values.

Intermediate Results:

	R10	TBILL
Rho - 1	-0.114485	-0.134501
Rho S.E.	0.029442	0.031474
Residual variance	0.208848	0.294587
Long-run residual variance	0.339813	0.500814
Number of lags	1	1
Number of observations	210	210
Number of stochastic trends**	2	2

**Number of stochastic trends in asymptotic distribution

EG cointegration test on R10 and Tbill.

Vector Error Correction Estimates		
Date: 03/25/17 Time: 16:00		
Sample (adjusted): 1960Q4 2012Q4		
Included observations: 209 after adjustments		
Standard errors in () & t-statistics in []		
Cointegrating Eq:	CointEq1	
R10(-1)	1.000000	
TBILL(-1)	-0.977209 (0.08081) [-12.0921]	
C	-1.629222	
Error Correction:	D(R10)	D(TBILL)
CointEq1	-0.091939 (0.03000) [-3.06433]	0.028635 (0.04625) [0.61916]
D(R10(-1))	0.228168 (0.08817) [2.58781]	0.110791 (0.13591) [0.81518]
D(R10(-2))	-0.072465 (0.08825) [-0.82110]	-0.146716 (0.13604) [-1.07848]
D(TBILL(-1))	-0.003707 (0.05823) [-0.06367]	0.240969 (0.08975) [2.68478]
D(TBILL(-2))	-0.068424 (0.05946) [-1.15074]	-0.157469 (0.09166) [-1.71805]
C	-0.009498 (0.03220) [-0.29499]	-0.010863 (0.04963) [-0.21887]
R-squared	0.107679	0.108233
Adj. R-squared	0.085701	0.086269
Sum sq. resids	43.92279	104.3633
S.E. equation	0.465154	0.717011
F-statistic	4.899330	4.927611
Log likelihood	-133.5485	-223.9875
Akaike AIC	1.335392	2.200837
Schwarz SC	1.431345	2.296789
Mean dependent	-0.010144	-0.010861
S.D. dependent	0.486466	0.750096
Determinant resid covariance (dof adj.)	0.063220	
Determinant resid covariance	0.059642	
Log likelihood	-298.4899	
Akaike information criterion	2.990334	
Schwarz criterion	3.214222	

Summary

In addition to the topics of univariate TS modeling, such as stationarity, ergodicity, stationary solutions, prediction, new topics emerge in the multivariate case, in particular

- ▣ the identification of impulse responses: structural VARs
- ▣ determination and estimation of cointegration relationships in ECMs
- ▣ the identification of impulse responses in ECMs: SECMs

