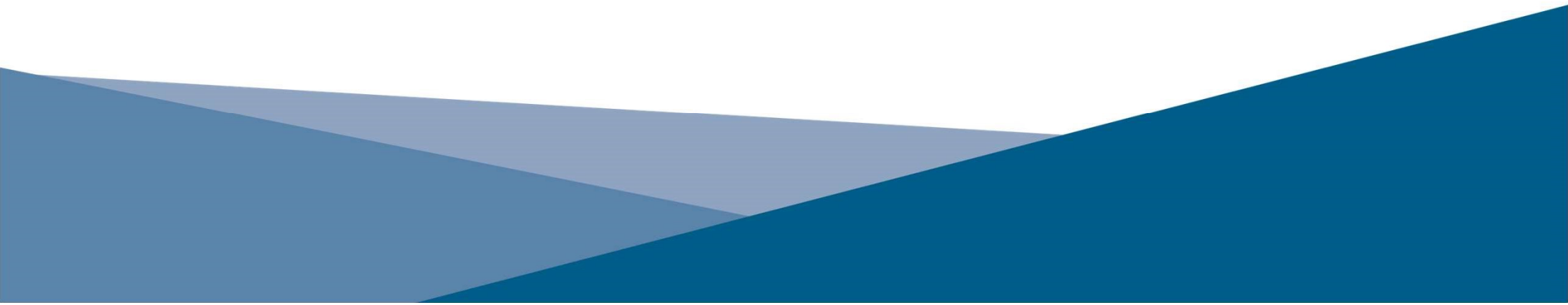
Univariate Time Series Models (ARMA)



Motivation

- Popularized by Box and Jenkings (1976)
- workhorse/benchmark models in forecasting time series data
- requires past to be similar to the future -> weak stationarity concept (discussed later)
- Statistical justification: «Any weakly stationary process can ALWAYS be represented as an (possibly infinite) sum of white-noise processes (zero mean, constant variance, uncorrelated over time; Wold decomposition)… in turn sum of white-noise processes can be approximated with a model in which a variable depends on a finite number of own lags plus a finite number of lags of a homoscedastic process (=ARMA model).»
- Bottom line: I you have a weekly stationary variable you can always model it with an ARMA model.
- Quick to estimate and simple to handle.

Topics Covered

AutoRegressive Moving Average model

- ARMA(p,q) model specification for a weakly stationary variable
- estimation
- ARIMA(p,q) model specification for a non-stationary variable
- Diagnostic checking
- Use for forecasting

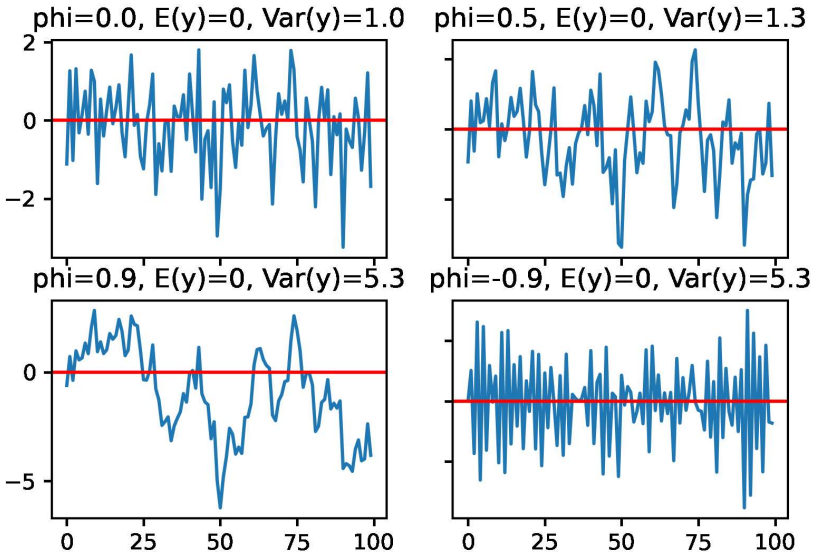
Lectures will closely follow:

“Applied Economic Forecasting using Time Series Methods”, by Eric Ghysels and Massimiliano Marcellino

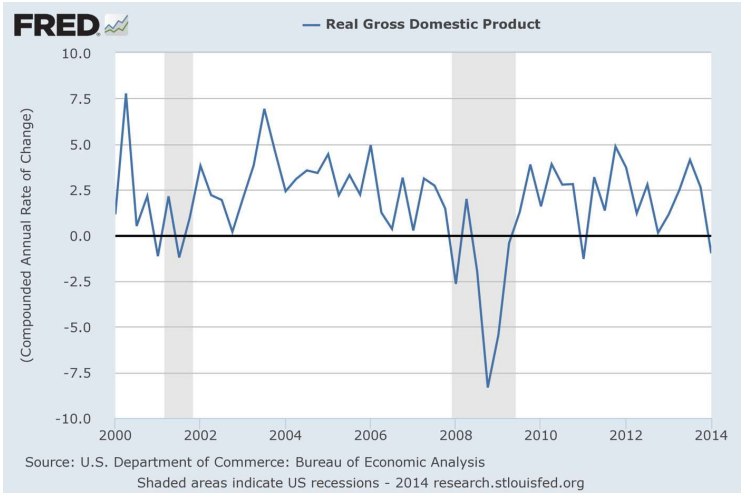
Further resources: Hamilton (1994) and Lutkepohl (2007).

Examples of a stationary time series

Simulated series



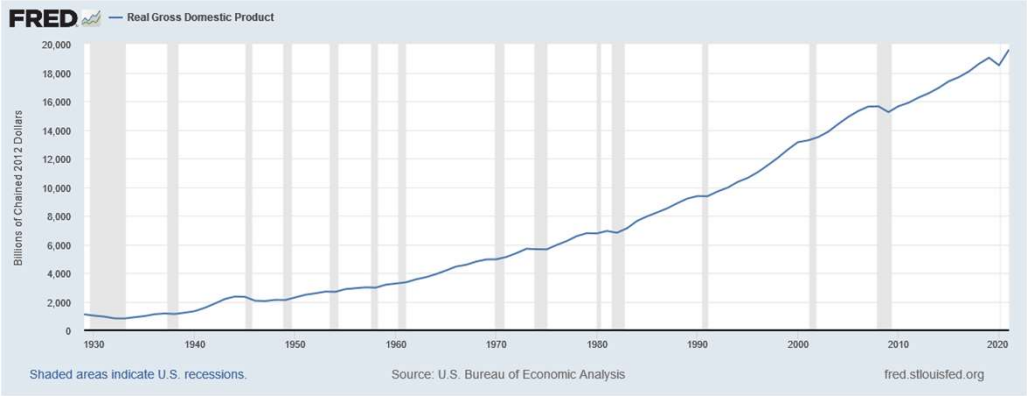
US rGDP growth



Source: FRED

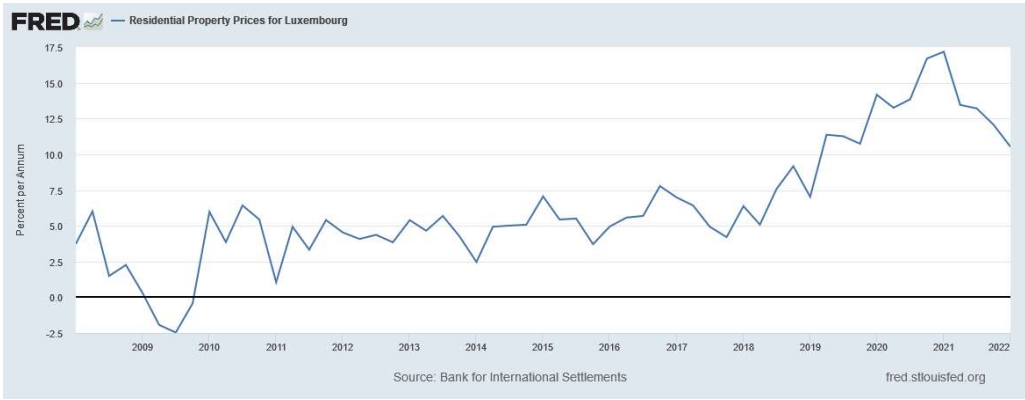
Examples of a non-stationary time series

US rGDP (level)



[Link to simulated non-stationary processes.]

Residential property price index for Luxemoburg



Source:
FRED

We start with “stationary” series

- Strictly stationarity TS:

$$F\{y_t, \dots, y_{t+T}\} = F\{y_{t+k}, \dots, y_{t+k+T}\}, \qquad \forall t, k, T$$

Where F is joint density of y.

- In practice we are content with weak stationarity defined as:

$$E(y_t) = E(y_{t+k}), \qquad \forall t, k$$
$$Var(y_t) = Var(y_{t+k}), \qquad \forall t, k$$
$$Cov(y_t, y_{t-m}) = Cov\left(y_{t+k, y_{t-m+k}}\right), \qquad \forall t, m, k$$

- Strict stationarity -> weak stationarity
- Reverse does not always hold (unless the process is Gaussian).

- A weakly stationary process can always be represented as (Wald decomposition theorem):

$$y_t = \epsilon_t + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + \cdots =$$

$$= \sum_{i=0}^{\infty} c_i\epsilon_{t-i} = \sum_{i=0}^{\infty} c_i L^i \epsilon_t = c(L)\epsilon_t$$

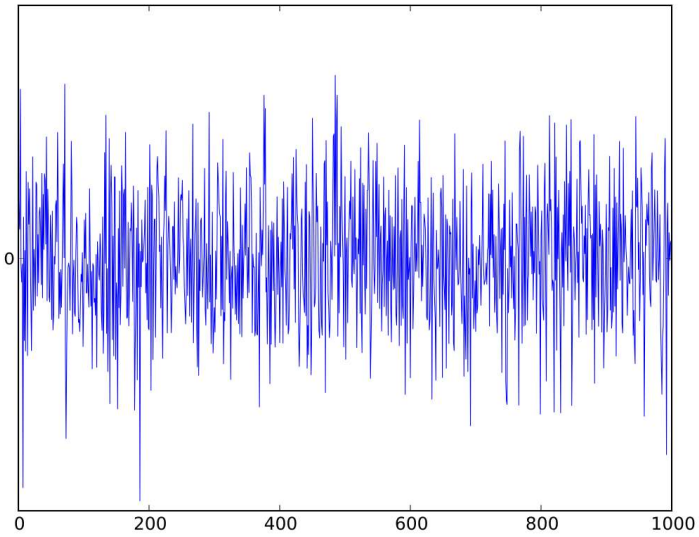
Where

- L is lag operator: $L\epsilon_t = \epsilon_{t-1}$ and $L^i\epsilon_t = \epsilon_{t-i}$
- Error process (ϵ_t) is a white-noise process

(zero mean, cons. var. and uncorrelated) →

$$\epsilon_t \sim WN(0, \sigma^2)$$

- We abstract from the constant ($E(y_t) = 0$).



Source: Wikipedia

- c(L) is an infinite order polynomial -> not very practical
- However, we can approximate c(L) with a ratio of two finite order polynomials:

$$c(L) = \frac{\psi(L)}{\phi(L)} \quad \text{eq. (1)}$$

- We can then re-express y_t (if $\phi(L)$ is invertible – weak stationarity):

$$\phi(L)y_t = \psi(L)\epsilon_t \quad \text{eq. (2)}$$

- Which is the same as (unwrap lag polynomials):

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \psi_1 \epsilon_{t-1} + \cdots + \psi_q \epsilon_{t-q}$$

- We can expressed a generic (weakly stationary) y_t process with a finite order ARMA(p,q) model. (statistical justification for these models)

AUTOREGRESSIVE MODEL (AR(p))

- Assume $c(L)$ in (1) is invertible (c(z) = 0 = $\sum_{j=0}^{\infty} c_j z^j$) ["has all the roots inside the unit circle", condition for the stability of a dynamic model)] and $\psi(L) = 1$ [no error dynamics], then we write (2) as:

$$y_t = \sum_{j=1}^{\infty} \phi_j y_{t-j} + \epsilon_t, \qquad \epsilon_t \sim WN(0, \sigma^2) \quad (3)$$

- Weak stationarity implies that the effect of y_{t-k} on y_t vanishes with increase in k, hence in practice we approximate (3) with a finite order AR(p) model

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad (4)$$

- If y_t is weakly stationary then $\phi(L)$ in (4) can be inverted and AR(p) model admits a MA(∞) representation (will be convenient later):

$$y_t = \frac{1}{\phi(L)} \epsilon_t$$

Example: AR(1)

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

- y_t is determined by a single own lag (y_{t-1}) and some random noise (ϵ_t) [purely random part that cannot be modelled].
- characteristic polynomial can is expressed as

$$\phi(z) = 1 - \phi_1 z = 0 \;\rightarrow\; z = \frac{1}{\phi_1}$$

- Weak stationarity requires $|\phi_1| < 1$ (“has all the roots outside the unit circle”).
- In practical terms, $|\phi_1| < 1$assures that y_t is a stable process. If not ($|\phi_1| \geq 1$), then y_t doesn’t have a stable long-term mean and/or variance, it can even be explosive, and we can not model it with an AR(p) model.

Introduction of autocorrelation function

- Autocorrelation function (AC)

$$Cov(y_t, y_{t-1}) = \gamma(1)$$
$$Cov(y_t, y_{t-2}) = \gamma(2)$$
$$Cov(y_t, y_{t-k}) = \gamma(k)$$

$$AC(k) = \frac{Cov(y_t, y_{t-k})}{\sqrt{Var(y_k)}\sqrt{Var(y_{t-k})}} = \frac{\gamma(k)}{\gamma(0)}$$

AC function of order k summarizes the strength of co-movement between y_t and y_{t-k} ($AC(k) \in [-1,1]$).

- Partial autocorrelation function

PAC(k) measures correlation between y_t and y_{t-k} conditional on ("controlling for") $y_{t-1}, \ldots, y_{t-k+1}$.

PAC coincide with (multiple) regression coefficients:

PAC(1) – coefficient on y_{t-1} from regressing y_t on y_{t-1}

PAC(2) – coefficient on y_{t-2} from regressing y_t on y_{t-1} and y_{t-2}

$\vdots$

PAC(k) – coefficient on y_{t-k} from regressing y_t on $y_{t-1}, .., y_{t-k}$

AC and PAC functions can guide selection of ARMA(p,q) autoregressive order p and moving average order q.

Example: AR(1) continued

- Assuming $|\phi_1| < 1$ holds, one can show that the mean, variance autocovariance of an AR(1) model are (steps: convert AR into MA form and take expectations):

$$E(y_t) = 0$$

$$Var(y_t) = \frac{\sigma^2}{1 - \phi_1^2}$$

$$\gamma(1) = \phi_1 \gamma(0)$$

$$\vdots$$

$$\gamma(k) = \phi_1^k \gamma(0)$$

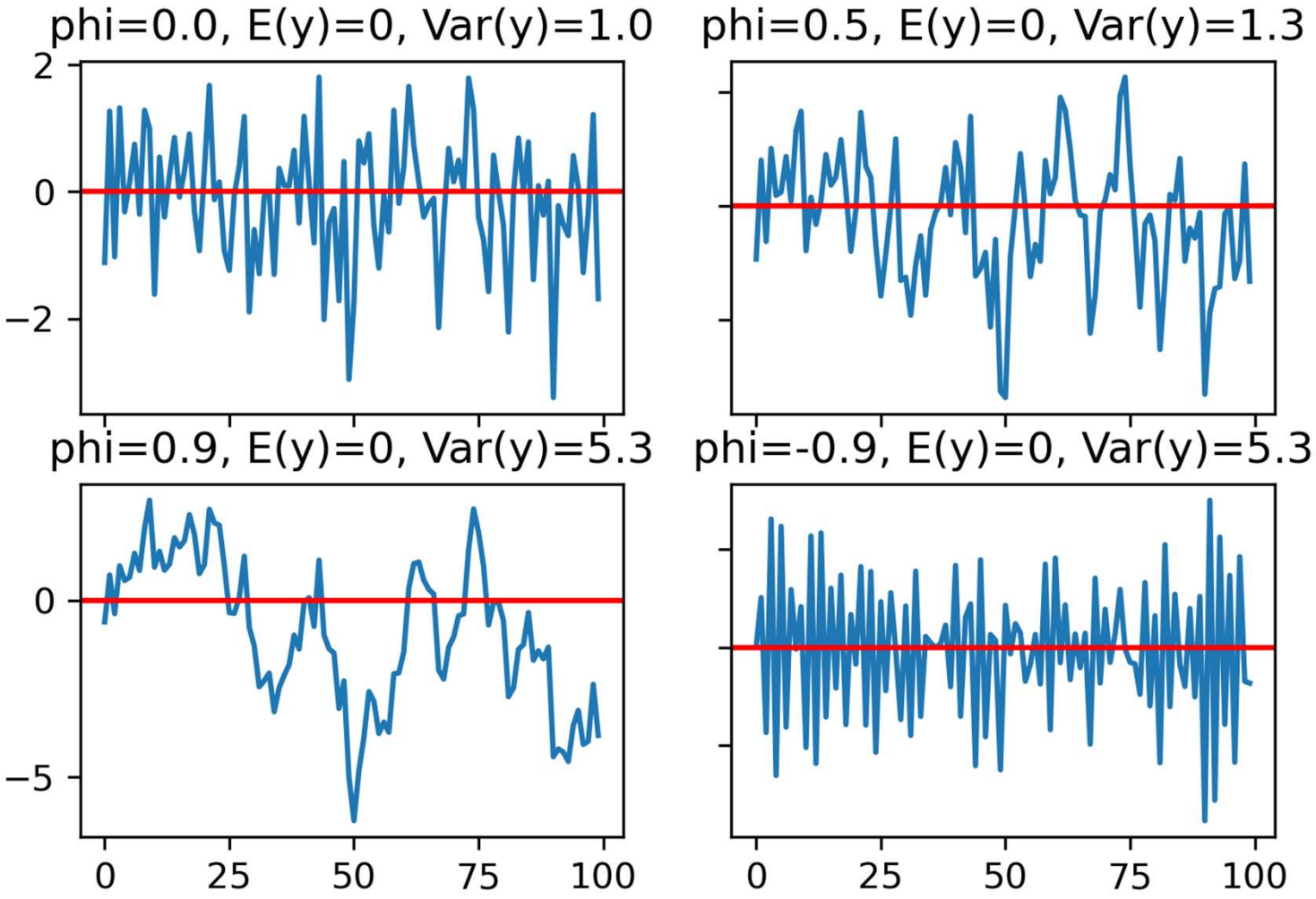
"long run" values

- And AC and PAC are:

$$AC(j) = \phi_1^j, \qquad PAC(1) = \phi_1, PAC(j > 1) = 0$$

- Point: This also means that if we take y_t and plot its AC&PAC functions and observe the patterns described with equations above, we confidently determine that y_t should be modeled in an AR(1) model (same as ARMA(1,0)). We will do this later.

AR(1) – simulated, different ϕ_1, $\epsilon_t \sim N(0,1)$



Link to MA.

Example: AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

- Weak stationarity conditions (derived from $\phi(z) = 0$, Harvey (1993))

$$\phi_1 + \phi_2 < 1$$
$$\phi_2 - \phi_1 < 1$$
$$-\phi_2 < 1$$

- AC and PAC function:

$$AC(1) = \frac{\phi_1}{1 - \phi_2}, \cdots, AC(k) = \phi_1 AC(k-1) + \phi_2 AC(k-2)$$

$$PAC(j) = 0 \; for \; j > 2$$

- Note again that AC function declines with k and PAC drops 0 for j>p.

MOVING AVERAGE MODEL (MA(q))

$$y_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \cdots + \psi_q \epsilon_{t-q} = \psi(L)\epsilon_t, \qquad \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

MA(q) process is always stationary with mean, variance, and autocovariance:

$$E(y_t) = 0$$

$$Var(y_t) = \left(1 + \psi_1^2 + \cdots + \psi_q^2\right)\sigma_\epsilon^2 = \gamma(0)$$

$$\gamma(k) = \begin{cases} \left(-\psi_k + \psi_{k+1}\psi_1 + \cdots + \psi_q \psi_{q-k}\right)\sigma_\epsilon^2, k = 1, .., q \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , k > q \end{cases}$$

- A nice MA process is invertible (if all the roots of $\psi(z) = 0$ are larger than 1 in absolute value), in which case we can express MA(q) process as an AR(∞):

$$\frac{1}{\psi(l)} y_t = \epsilon_t$$

- This property is useful to derive AC&PAC function of an MA process since it coincides with AC&PAC function of an AR(∞) process.

- PAC(k) of an AR(p) process was zero for k>p. Since an MA process can be represented as an AR(∞) process, an MA process PAC is always different than zero (except in the limit). It slowly decays but only touches zero in the limit. This helps identifying MA process by plotting its AC and PAC.

Example: MA(1)

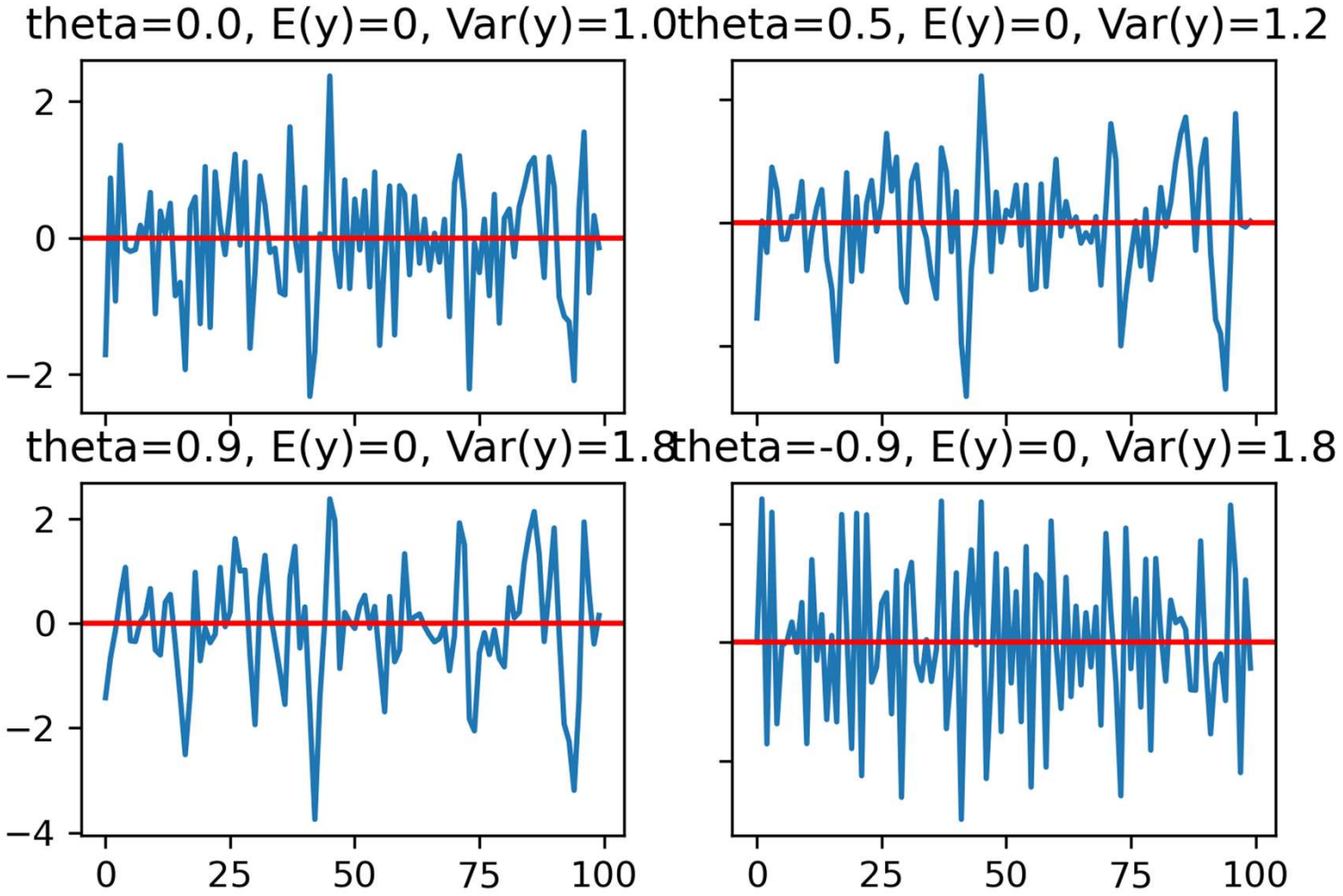
$$y_t = \epsilon_t + \psi_1 \epsilon_{t-1}$$

- MA(1) is invertible if $|\psi_1| < 1$
- Mean, variance and AC function

$$E(y_t) = 0$$
$$Var(y_t) = \sigma_\epsilon^2 (1 + \psi_1^2)$$
$$\gamma(1) = -1\psi_1 \sigma_\epsilon^2$$
$$\gamma(k) = 0, k > 1$$
$$AC(1) = -\frac{\psi_1}{1 + \psi_2} \quad and \quad AC(k) = 0 \; if \; k > 1$$

- Note: For the AR(p) process the AC function decays but only touches 0 in the limit. For the MA(q) process the AC function is 0 from lag q onward. Converse holds for the PAC function. For the AR(p) process it is zero from lag p onward but for MA(q) process it slowly decays but never touches 0. All this implies that when deciding on ARMA(p,q) model, p and q can be selected by plotting AC and PAC function of y_t.

MA(1) – simulated, different ψ_1, $\epsilon_t \sim N(0,1)$



Link to AR.

JOIN AR AND MA MODEL: ARMA(p,q)

$$\phi(L)y_t = \psi(L)\epsilon_t$$

or

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \psi_1 \epsilon_{t-1} + \cdots + \psi_q \epsilon_{t-q}$$

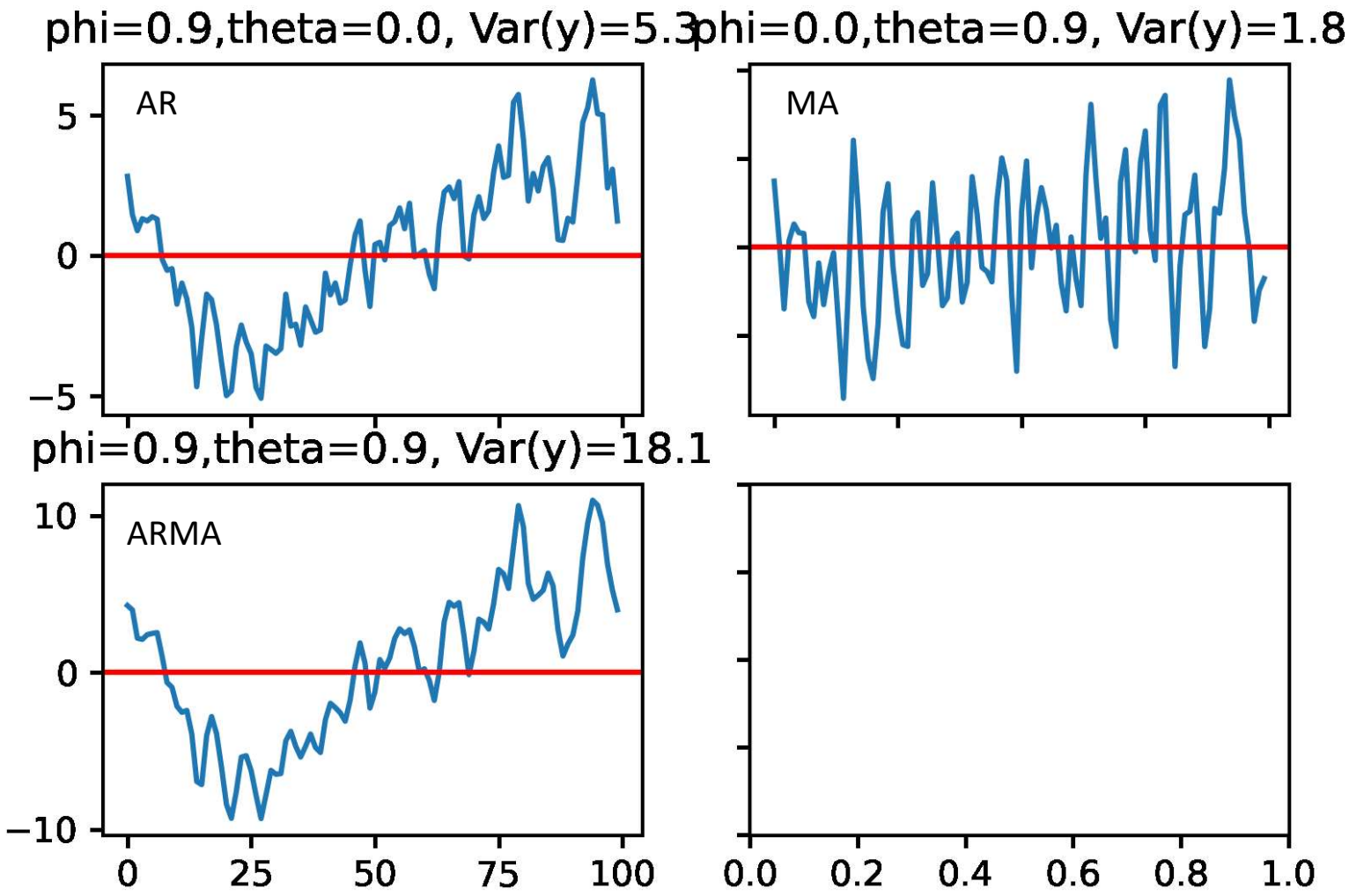
- weak stationarity condition: $\phi(z) = 0 \rightarrow |z_i| > 1,\ i = 1,..,p$
- Invertibility condition: $\psi(z) = 0 \rightarrow |z_i| > 1,\ i = 1,..,q$
- ARMA model can be written as an MA model: $y_t = \psi^{-1}(L)\phi(L)\epsilon_t = c(L)\epsilon_t$
- Therefore the first two moments are: $E(y_t) = 0\ and\ Var(y_t) = \sigma^2 \sum_{i=0}^{\infty} c_i^2$

Example: ARMA(1,1)

$$y_t = \phi_1 y_{t-1} + \epsilon_t + \psi_1 \epsilon_{t-1}$$

- $E(y_t) = 0, \;\; Var(y_t) = \gamma(0) = \frac{(1+\psi_1^2+2\phi_1\psi_1)\sigma_\epsilon^2}{(1-\phi_1^2)}$
- $Cov(y_t, y_{t-k}) = \phi_1 \gamma_{k-1}$
- $AC(k) = \frac{\gamma_k}{\gamma_0}$
- AC and PAC functions are like AR(∞) and MA(∞) (both fade away slowly).

ARMA(1,1) – simulated, $\epsilon_t \sim N(0,1)$



NON-STATIONARY & INTEGRATED PROCESSES

So far we always assumed that y_t is stationary. If it is not stationary we call it an integrated process:

"An integrated process y_t is a non-stationary process such that $(1-L)^d y_t$ is stationary, where d is the order of integration. Such process is labeled as $I(d)$." (Ghysels&Marcellino, (2018))

- Non-stationary processes' mean and/or variance vary over time. They are said to be unstable and are sometimes called unit root process (at least one root in $\phi(z) = 0$ is exactly 1).

- If you estimate a model on a non-stationarity series than the usual statistical tests are not valid (will be discussed later).

Example of non-stationary process: Simple Random Walk (RW)

$$y_t = y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

- The most common non-stationary process.
- In fact, it often does rather well in forecasting non-stationary processes.
- RW is integrated of order 1, $I(d) = I(1)$, since it can be made stationary by multiplying it with $(1-L)^d = (1-L)^1$:

$$(1-L)y_t = y_t - y_{t-1} = \Delta y_t = \epsilon_t$$

- That is, a non-stationary series y_t can be made stationary by taking first differences.

- RW also be expressed as:

$$y_t = y_{t-1} + \epsilon_t = \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} + \epsilon_{t-3} + \cdots$$

- For a RW process the effect of shocks does not decay, e.g. even ϵ_{t-100} affects its value.
- Its mean and variance are

$E(y_t) = 0$ (consistent with stationarity)

$Var(y_t) = Var(\sum_{k=0}^{\infty} \epsilon_{t-k}) \to \infty$ (unstable variance)

- Like variance it's AC is not properly defined. If computed empirically it does not decay but stays close to one for all lags.

- Slight change, RW with drift:

$$y_t = \mu + y_{t-1} + \epsilon_t$$

- Repeated substitution:

$y_t = \mu + \epsilon_t + \mu + \epsilon_{t-1} + \mu + \epsilon_{t-2} + \ldots$

- The moments change drastically:

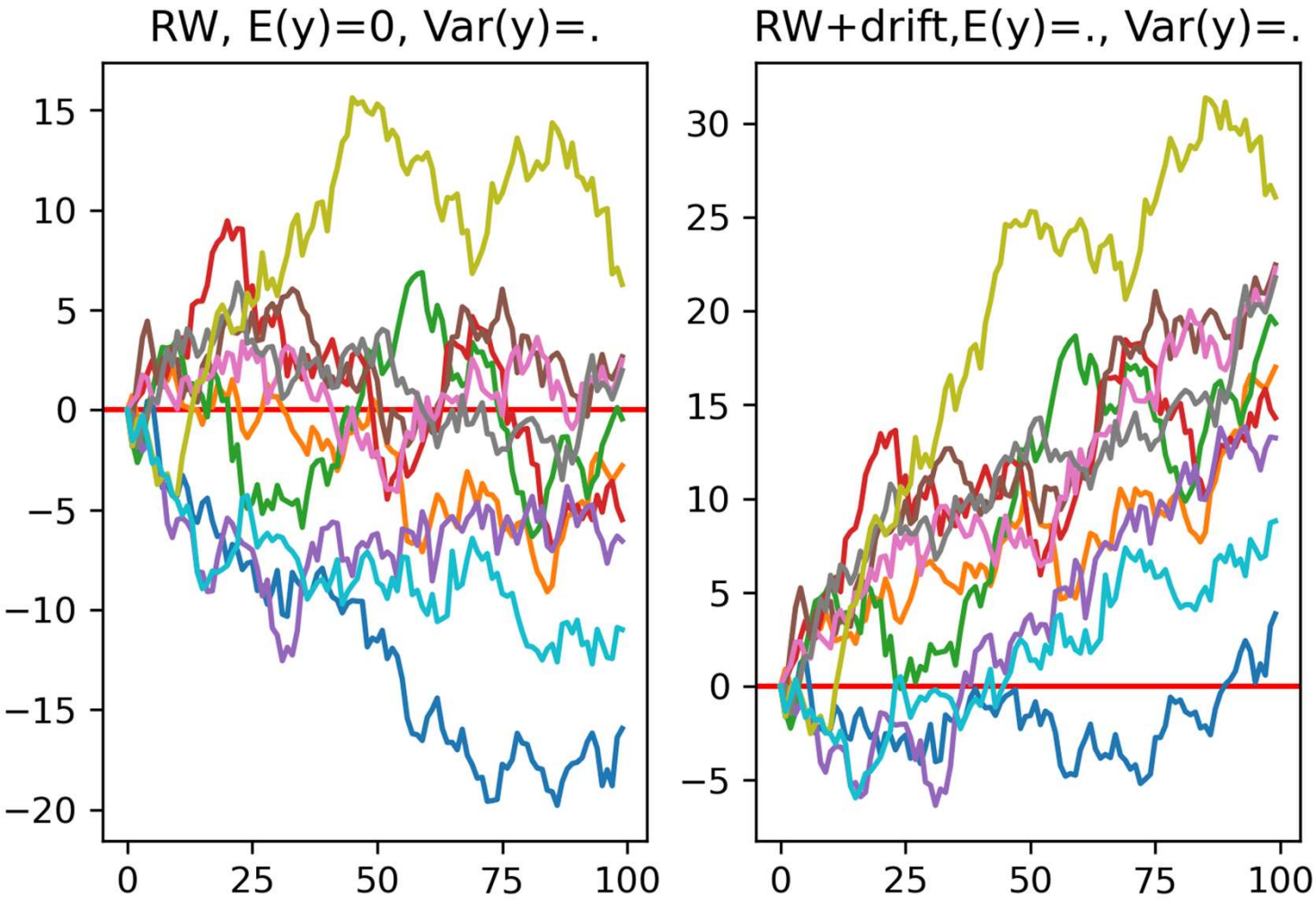
$$E(y_t) \to \infty$$
$$Var(y_t) \to \infty$$

- Differencing solves this issue (e.g.):

$$\Delta y_t = y_t - y_{t-1} = y_{t-1} + \epsilon_t - y_{t-1} = \epsilon_t$$

$$E(\Delta y_t) = 0$$
$$Var(\Delta y_t) = Var(\epsilon_t)$$

Simulated RW and RW+drift (10x)



Recap: ARIMA(p,I,q) model

- $\phi(L)\Delta^d y_t = \psi(L)\epsilon_t$
- Where:
- $\Delta^d = (1-L)^d$ - this just means that y_t needs to be differenced d-times to make it stationary
- $\phi(L) = (1 + \phi_1 L + \phi_2 L^2 + \cdots)$ – AR polynomial
- $\psi(L) = (1 + \psi_1 L + \psi_2 L^2 + \cdots)$ – MA polynomial

MODEL SPECIFICATION

We need to decide on:

- p – number of AR lags; AC&PAC, tests or IC
- q – number of MA lags; AC&PAC, tests or IC
- d – order of integration; NEW

Selection of p, q (we will deal with ***d*** later)

We first assume y_t is stationary and inspect how to decide on p&q.

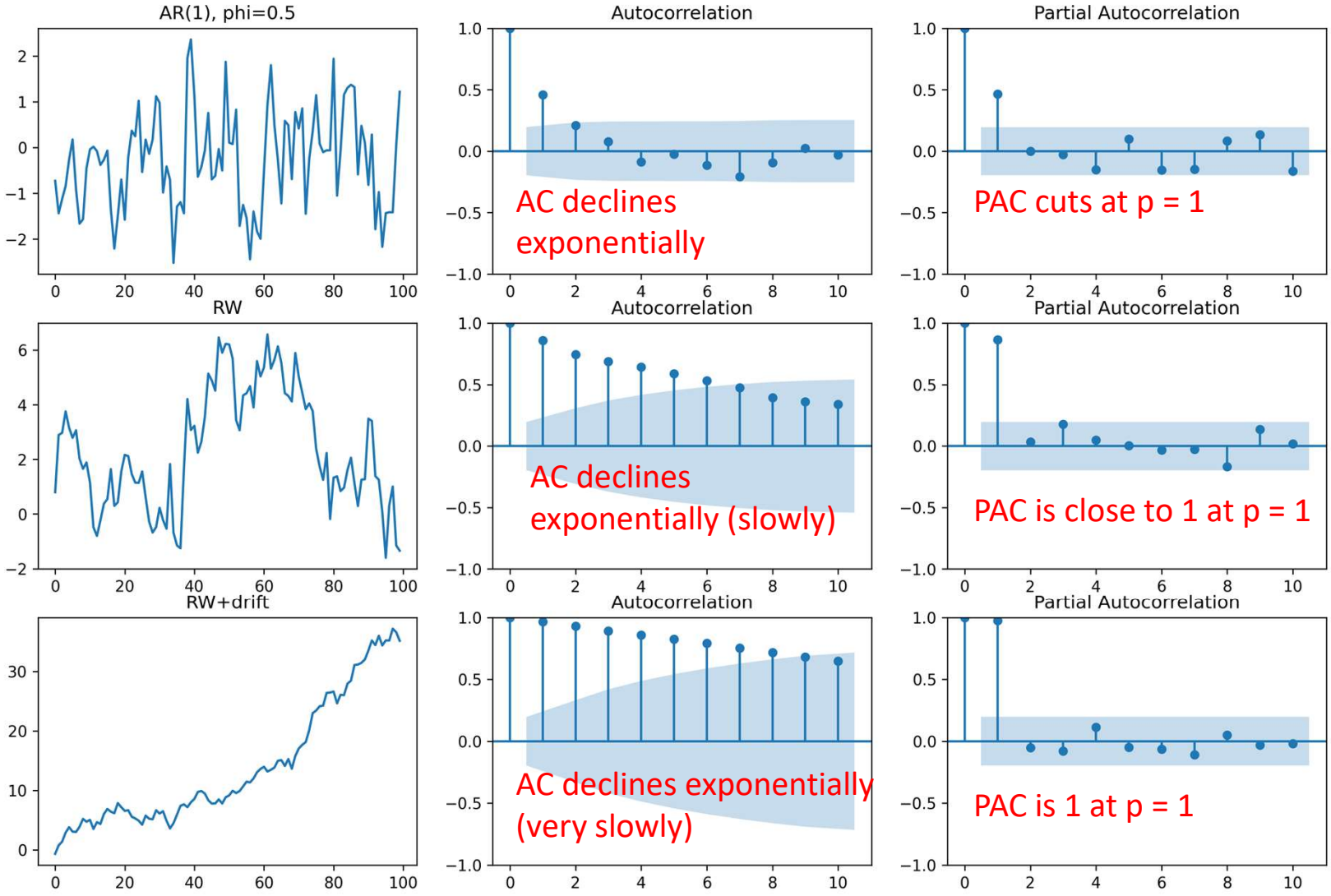
1) Plot AC and PAC functions:

- Take y_t and estimate AC(k) and PAC(k) for different values of k.
- Plot the two functions together with their associated $(1-\alpha)\%$ confidence bands, where α is typically 5%.

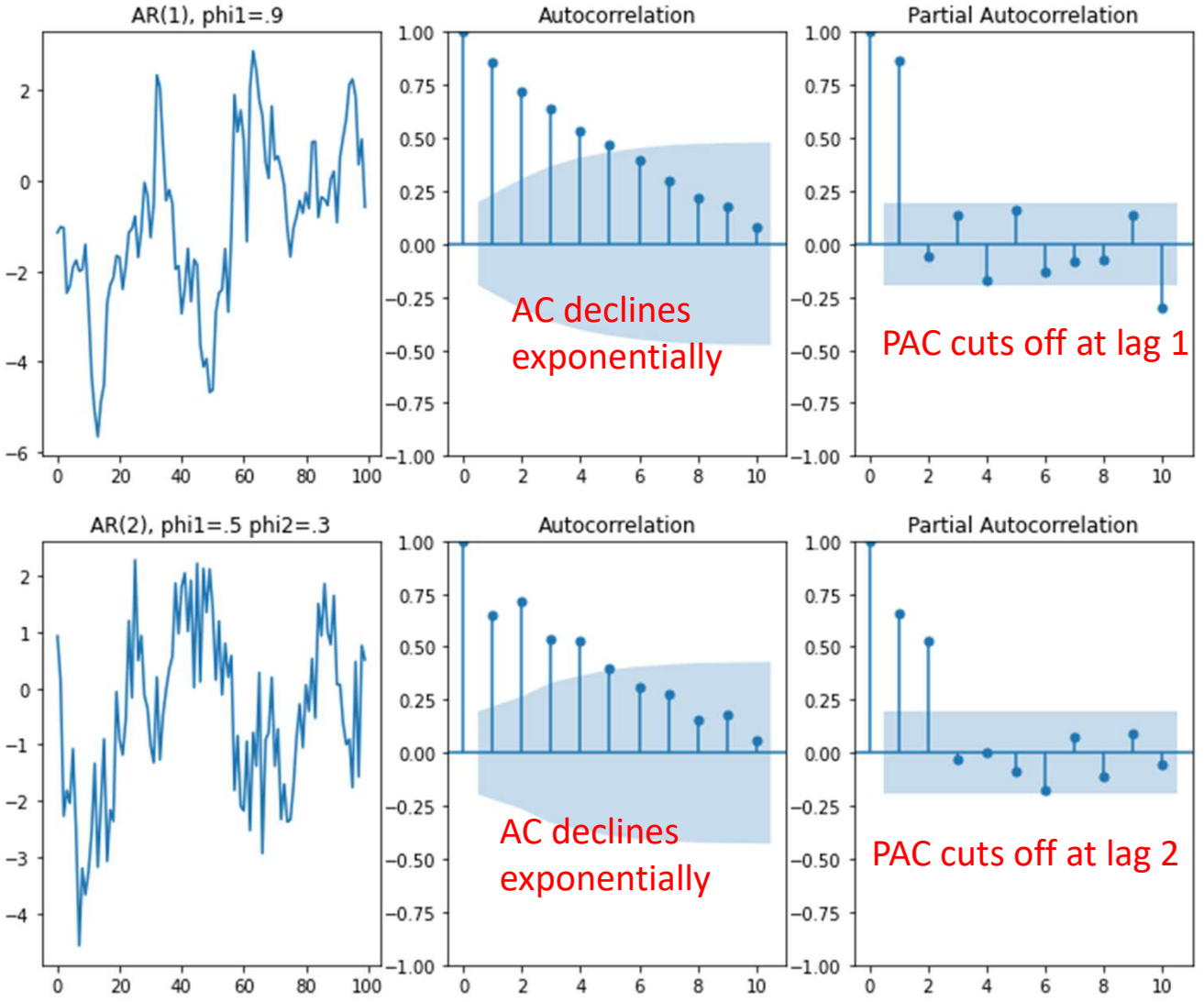
We first need to decide on d (we skip this part since one can tests for this):

- AC&PAC decrease when k increases → (evidence) y_t is stationary, set d = 0
- AC decreases VERY slowly, and PAC(1) is close to 1 → (evidence) y_t is non-stationary, set d = 1 (=difference y_t) and work with Δy_t

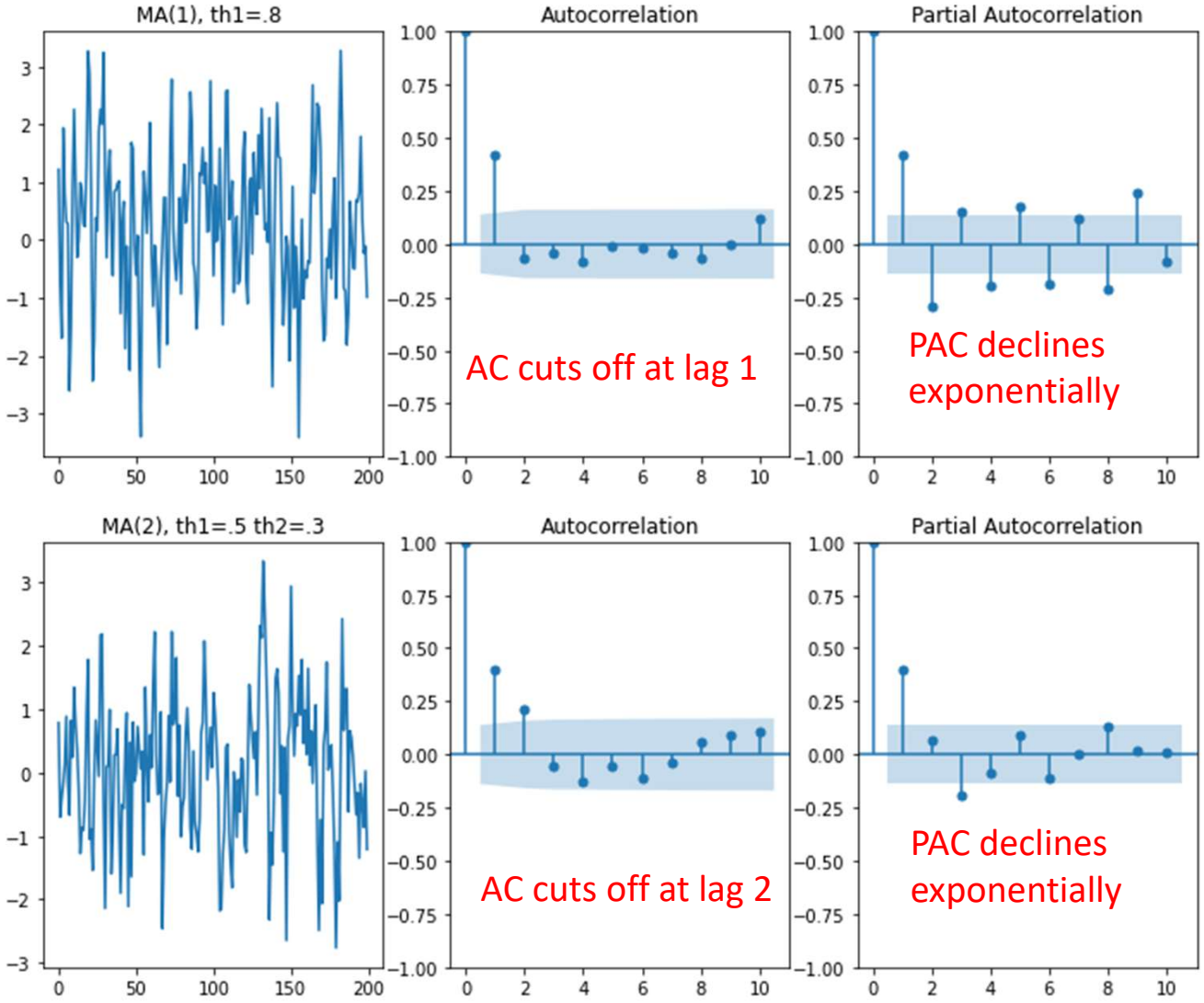
AC and PAC for AR(1), RW, RW+drift



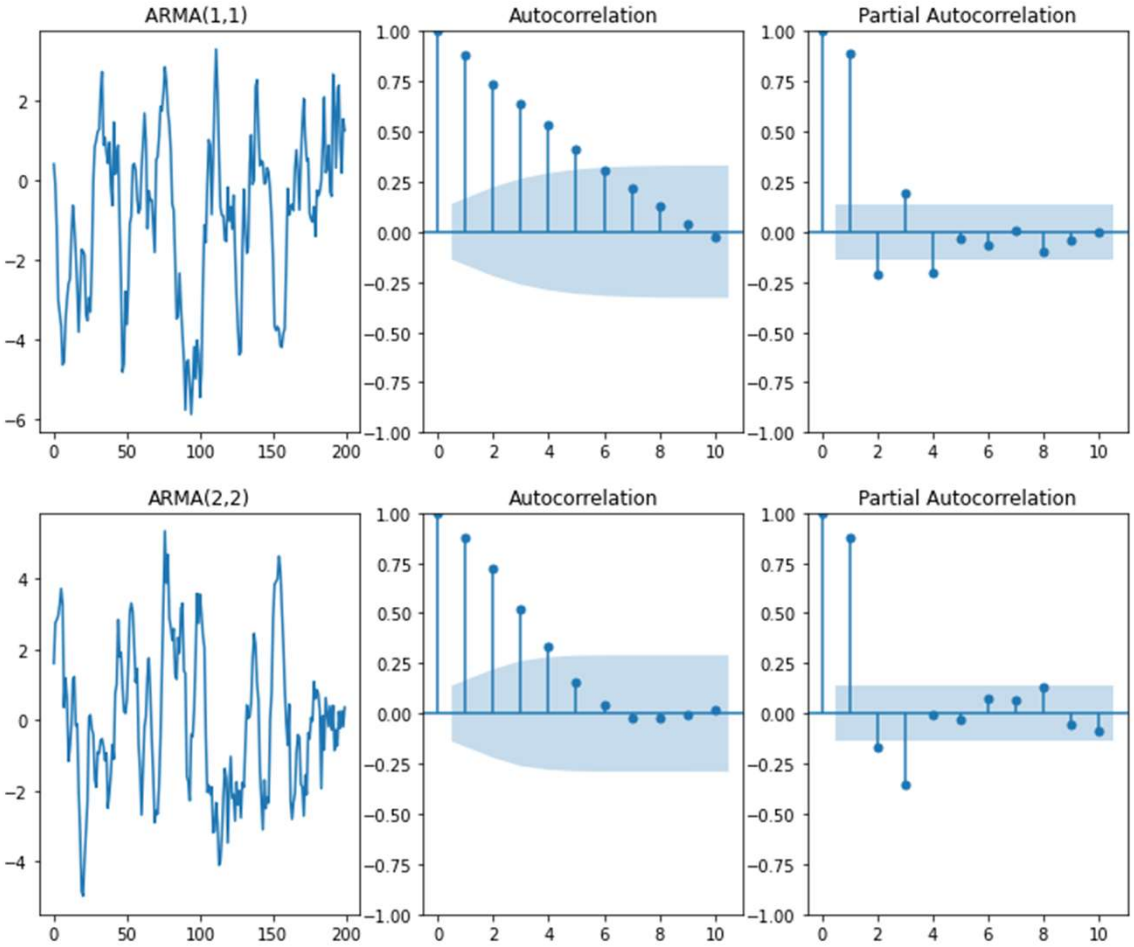
AC and PAC for AR(1) and AR(2)



AC and PAC for MA(1) and MA(2)



AC and PAC for ARMA(1,1) and ARMA(2,2) more complex



Instead of conjecturing from AC&PAC make a guess on p&q, check statistical significance of coefficients and whiteness of the residuals (next slide). If residuals are not “WN-like” increase p or q.

2) Testing based specification

- make a guess on p&q (TS frequency, previous work, AC&PAC functions)
- If the guess is correct we expect estimated residuals ($\hat{\epsilon}_t$) to be uncorrelated
- autocorrelation can be tested with Ljung-Box Q test (Ljung and Box (1978)) or Box-Pierce test (Box and Pierce (1970)); both are called *Portmanteau* tests
- Example: LB-test (standard output of TS regression commands)

$$Q_{BP} = T(T+1)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{T-k} \quad ; \quad \hat{Q}_{BP} \sim \chi^2_{(h)}$$

$$H_0: no\ AC\ of\ order\ k\ or\ less$$
$$H_1: evidence\ of\ AC\ of\ order\ k\ or\ less$$

3) Information Criteria

- Information Criteria is a statistic that help select models.
- They combine “goodness of fit” and “model complexity” into one number (general form):

$$IC = \log(\sigma^2) + g(k, T)$$

where:

σ_ϵ^2 - model error variance, goodness of fit

$g(k, T)$ - function increasing in the number of parameters (k) relative to the number of observation (T), complexity

- Compromise between fit and complexity/parsimony:
  - More parameters -> better in-sample fit
  - Less parameters -> lower variance of the estimates -> better forecasts

- worse fit: $\uparrow \sigma^2 \rightarrow \uparrow IC$
- more parameters increases the penalty function: $\uparrow k \rightarrow \uparrow g(k,T) \rightarrow \uparrow IC$

- Steps:
  - Estimate IC for all models up to $ARMA(p_{max}, q_{max})$ where $p_{max} > p_{true}$ and $q_{max} > q_{true}$
  - Select model with min. IC.

- Caution: ICs should be estimated over the same period (think lags!).

Popular ICs

- Akaike (AIC):

$$g(k,T) = \frac{2k}{T}$$

- Schwarz of Bayesian IC (BIC):

$$g(k,T) = \frac{k\log(T)}{T}$$

- Hannan-Quinn IC (HIC):

$$g(k,T) = \frac{2k\ log\ (\log(T))}{T}$$

When $T \to \infty$, IC select correct model with probability approaching 1 (BIC,HIC).

In finite samples no uniformly valid ranking of ICs exist.

In practice and in small samples BIC is preferred over AIC since it select more parsimonious models (better for forecasting, why?).

p, d, q in practice (cookbook)

- d – selected with unit root tests (we show them later). Unit-root test have low power in small samples. Results can be confusing. Consequences of over-differencing are less severe than failing to difference the data (from a testing point of view) but less relevant if model is to be used for forecasting.
- p, q – most often IC is used to select $p\&q$ with preference being given to BIC since it selects more compact models.
- It is important to test for autocorrelation. Residual autocorrelation implies that errors contain patterns so that forecasts power can be improved upon.
- Other assumptions, while important for statistical tests, are less important for forecasting.
- If $p\&q$ are too high the forecasts will be poor.
- If one has enough data available, a pseudo out-of-sample forecasting exercise can be used to validate or select a model.
- In case of conflicting suggestions (e.g. BIC selects AR(1) and AIC ARMA(1,1)) the two models will typically have similar forecasting performance. The data are not informative enough to discriminate among the two (or more) models.
- AR(p) are easier to estimate than models that include MA(q) term. In addition, If MA(q) is invertible, then ARMA(p,q) or MA(q) models can be approximated within AR(p*) (where p*>p). However, forecasts might be better with MA(q) component since parsimonious models tend to be better at forecasting.

Estimation: with OLS

ARIMA(p,d,q):

$$\phi(L)\Delta^d y_t = \psi(L)\epsilon_t$$

- If $y_t \neq I(0)$, that is if $d \neq 0$, then difference the series y_t until the resulting series w_t $(= \Delta^d y_t)$ is stationary replace y_t with w_t (estimate the model using $\phi(L)w_t = \psi(L)\epsilon_t$).
- If $\psi(L) = 1$ (= no MA components) the model can be estimated with OLS.
- If $\psi(L) \neq 1$ we use maximum likelihood approach to estimate the model.

Maximum Likelihood approach to estimation

- Our basic assumption was that $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$ (intuition: all important variables are included in the model).
- Add distributional assumption $\epsilon_t \sim iid\ N(0, \sigma_\epsilon^2)$ and maximize the log-likelihood function (numerical solution as opposed to analytical):

$$L\left(y_{p+1}, \dots, y_T; \theta\right) = -\frac{T-p}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_\epsilon^2) - \sum_{t=p+1}^{T}\frac{\epsilon_t^2}{2\sigma^2}$$

- If MA terms are present it requires an additional step (EM algorithm)

Unit root tests – to decide on d

- Unit root test are used to test if series y_t has a unit-root.
- If y_t has a unit-root, we usually model the differenced series $\Delta^d y_t$
- d is determined by how many times do we need to difference the series until we can reject the presence of unit-root (or cannot rejected a no unit-root hypothesis)
- In practice, testing for unit-root is the first in modelling.
- The most standard test for stationarity is the Dickey and Fuller (1979) test. This test is reported by most statistical software.
- To gain intuition, we describe the logic of the DF-test. Several improvements and alternative tests are available.

Dickey-Fuller test

- Assume the following process:

$$y_t = \rho y_{t-1} + \epsilon_t \qquad \text{eq. (1)}$$

- We know from before:

$|\rho| < 1$ – weakly stationary

$\rho = 1$ – process is not stationary or it has a unit-root

- We wish to test:

$$H_0: \rho = 1 \rightarrow y_t \; is \; not \; weakly \; stationary$$
$$H_1: |\rho| < 1 \rightarrow y_t \; is \; stationary$$

- Re-write the process:

$$\Delta y_t = (\rho - 1) y_{t-1} + \epsilon_t \qquad \text{eq. (2)}$$

- We are interested in testing if process has a unit-root ($\rho = 1$, y_t is $I(1)$). This is the same as $\rho - 1 = 0$ in eq. (2). Normally we would test it with a simple t-test.
- However, if y_t is $I(1)$ neither the coefficient $(\rho - 1)$ or t-test are normally distributed.
- The limiting distribution of t when $\rho = 1$ is called the Dickey-Fuller distribution.

- Re-write the process:

$$\Delta y_t = (\rho - 1) y_{t-1} + \epsilon_t \qquad \text{eq. (2)}$$

- We are interested in testing if process has a unit-root ($\rho = 1$, y_t is $I(1)$). This is the same as $\rho - 1 = 0$ in eq. (2). Normally we would test it with a simple t-test.
- Intuition: if y_t is $I(0)$ it has a tendency to return to some constant mean. The level of the series will be a significant predictor of the next period's change.
- However, if y_t is $I(1)$ neither the coefficient $(\rho - 1)$ or t-test are normally distributed.
- The limiting distribution of t when $\rho = 1$ is called the Dickey-Fuller distribution (tabulated or numerically approximated, simulated).
- This is why one cannot simply estimate (2) and use the t-value from the regression output to test for $\rho - 1 = 0$ $(\rightarrow \rho = 1)$.
- If you instead use DF-test in statistical package it already report the correct tabulated values.

- Eq. presented the most simple case.
- DF-test with drift

$$\Delta y_t = v_0 + (\rho - 1) y_{t-1} + \epsilon_t$$

$H_0: y_t \; is \; I(1) \; with \; drift$

$H_1: y_t \; is \; I(0) \; with \; non - zero \; mean$

- DF-test with time-trend

$$\Delta y_t = v_0 + v_1 t + (\rho - 1) y_{t-1} + \epsilon_t$$

$H_0: y_t \; is \; I(1) \; with \; drift$

$H_1: y_t \; is \; I(0) \; with \; deterministic \; trend$

Improvements and alternatives

- Augmented DF-test: add lags of Δy_t to the test. This is done to account for the (potential) presence of serial correlation in eq. (2).

$$H_0: y_t \; is \; I(1)$$
$$H_1: y_t \; is \; I(0) \; or \; I(0) with \; time \; trend$$

- Augmented DF-test is perhaps the most popular unit-root test
- Phillips-Perron (1988) test - also accounts for the presence of serial correlation but in different way than ADF (null hypothesis is the same as in ADF test, $H_0: y_t \; is \; I(1)$)
- Kwiatkowski–Phillips–Schmidt–Shin (KPSS, *1992*) test. This test reverses the null hypothesis. In KPSS-test: $H_0: trend - stationary$
- DF-GLS test (Elliott, Rothenberg, and Stock (1996)) – improved small sample properties.
- Tests in the presence of breaks and outliers, etc..

Unit-root testing in practice

- Unit-root tests have low power in small samples.
- Best to start with the most complex model (e.g. ADF test with constant and time trend) and reduce it to the simplest case (e.g. ADF test without drift or time trend). If you cannot reject the null for any of the ADF tests you conclude that the series needs to be differenced.
- Once the decision is made, try also the other tests.
- Tests are sensitive to outliers and parameter breaks -> plot the data & test for these things.
- If in doubt, google for related papers and inspect how previous work modelled pertinent series.
- Small sample size and power of the test are usually quite bad.

FORECASTING

- Assume for now that there is no parameter uncertainty.
- Assume y_t is I(0) (if not, replace y_t with $w_t = \Delta^d y_t$)

$y_T = \phi_1 y_{T-1} + \cdots + \phi_p y_{T-p} + \epsilon_T + \psi_1 \epsilon_{T-1} + \psi_q \epsilon_{T-q}$; end of sample (T)

$y_{T+1} = \phi_1 y_T + \cdots + \phi_p y_{T-p+1} + \epsilon_{T+1} + \psi_1 \epsilon_T + \psi_q \epsilon_{T-q+1}$; T+1

- Take E():

$\hat{y}_{T+1} = E(y_{T+1}|I_t) = \phi_1 y_T + \cdots + \phi_p y_{T-p+1} + \psi_1 \epsilon_T + \psi_q \epsilon_{T-q+1}$; T+1

- Expressions for T+2,… can be obtained with simple sequential substitution (see next slide)

Forecasting with AR(1)

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

Forecast:	*Forecast error variance:*
$\hat{y}_{T+1} = \phi_1 y_t$	$Var(e_{T+1} = \epsilon_{T+1}) = \sigma_\epsilon^2$
$\hat{y}_{T+2} = \phi_1 y_{T+1} = \phi_1^2 y_T$	$Var(e_{T+2} = \epsilon_{T+2} + \epsilon_{T+1}) = (1 + \phi_1^2)\sigma_\epsilon^2$
$\vdots$	
$\hat{y}_{T+k} = \phi_1^k y_t$	$Var(e_{T+k}) = (1 + \phi_1^2 + \cdots + \phi_1^{2k-2})\sigma_\epsilon^2$

Note:

$\lim_{h \to \infty} \hat{y}_{T+h} = E(y_t) = 0$ and $\lim_{h \to \infty} Var(\hat{y}_{T+h}) = \frac{\sigma_\epsilon^2}{(1-\phi_1^2)} = Var(y_t)$

Forecasting with MA(1)

$$y_t = \epsilon_t + \psi_1 \epsilon_{t-1}$$

Forecast:

$\hat{y}_{T+1} = \psi \epsilon_T$

$\hat{y}_{T+2} = 0$

$\vdots$

$\hat{y}_{T+k} = 0$

Forecast error variance:

$Var(e_{T+1} = \epsilon_{T+1}) = \sigma_\epsilon^2$

$Var(e_{T+2} = \epsilon_{T+2} - \psi_1 \epsilon_{T+1}) = (1 + \psi_1^2)\sigma_\epsilon^2$

$Var(e_{T+k}) = (1 + \psi_1^2)\sigma_\epsilon^2$

Note:

$\lim_{h \to \infty} \hat{y}_{T+h} = E(y_t) = 0$ and $\lim_{h \to \infty} Var(\hat{y}_{T+h}) = (1 + \psi_1^2)\sigma_\epsilon^2 = Var(y_t)$

- Forecast expressions for ARMA(p,q) model are derived similar…

RANDOM WALK

$$y_t = y_{t-1} + \epsilon_t$$

$\hat{y}_{T+h} = y_T$ $\qquad$ $Var(e_{T+h} = \epsilon_{T+1} + \cdots + \epsilon_{T+h}) = h\sigma_\epsilon^2$

Note:

$\lim_{h\to\infty} \hat{y}_{T+h} = y_T$ and $\lim_{h\to\infty} Var(\hat{y}_{T+h}) = \infty$

Point: Due to the unit-root variance (non-stationarity) of the forecast error grows linearly with time. In the stationary model case it converges to the unconditional variance of the variable.

How to construct forecast confidence intervals?

- We need to assume a distribution for ϵ_t.
- Most often we assume ϵ_t is Gaussian in which case $(1-\alpha)\%$ interval is:

$$\left(\hat{y}_{T+h} - c_{\frac{\alpha}{2}}\sqrt{Var(e_{T+h})}\,;\ \hat{y}_{T+h} + c_{\frac{\alpha}{2}}\sqrt{Var(e_{T+h})}\right)$$

Where $c_{\frac{\alpha}{2}}$ are critical values from N(0,1). (e.g. $c_{\frac{5}{2}} = 1.96, c_{\frac{10}{2}} = 1.64$)

- Assume now that parameters are unknown but consistently estimated.
- Variance of the forecast errors increase due to parameter uncertainty. However, the increase is minor if sample is large enough.
- "AR(1) with constant" example from Ghysels and Marcellino (2018):

$$y_t = \mu + \alpha y_{t-1} + \epsilon_t$$

Forecast error variance becomes

$$Var(e_{T+1}) = \sigma_\epsilon^2 + x_T' Var\left(\hat{\theta}\right) x_T$$

Where $Var\left(\hat{\theta}\right) = Var\begin{pmatrix}\hat{\mu}\\ \hat{\alpha}\end{pmatrix}$.

See Lutkepohl (2007) for a general treatment.

Direct forecasts

- Before we used iterated forecasts.
- E.g. for AR(1) we derived the forecast "$\hat{y}_{T+k} = \phi_1^k y_t$" by iteratively replacing unknown y_{T+h} with past values (y_{T+h-j}) until we arrived at an expression that only contains known y_t.
- In a direct forecast we instead directly estimate the parameters that will be used to construct the forecast (e.g.):

$$y_t = \phi_h y_{t-h} + \epsilon_t$$

- The forecast ($\tilde{y}_{T+h}$) is then simply:

$$\tilde{y}_{T+h} = \phi_h y_T$$

Iterated ($\hat{y}_{T+k}$) vs. direct ($\tilde{y}_{T+h}$) forecast

- Correct model specification: both estimators for ϕ are consistent but $\hat{\phi}$ is more efficient than $\tilde{\phi}$. Since parameter uncertainty enters forecast error variance, iterated forecast will be better.
- However, in practice we are unsure of the correct model.
- In the presence of incorrect model specification (all models are merely approximations of the true underlying model) direct forecast can perform better (bias-variance trade-off).
- Typically direct forecasts perform well when the model choice is uncertain.
- Typically direct forecasts perform worse when forecast horizon increases (Marcellino, Stock, and Watson (2006)).