Анализ автокорреляций временного ряда

А.Г. Трофимов
к.т.н., доцент, НИЯУ МИФИ

lab@neuroinfo.ru
http://datalearning.ru

Курс “Статистическая обработка временных рядов”

Сентябрь 2018

Box-Jenkins Methodology

The Box-Jenkins methodology (1970) is a procedure for identifying, selecting and estimating ARMA models for discrete univariate time series data

Step 1. Establish the stationarity of your time series. If it is non-stationary try to transform it to be stationary

Step 2. Identify a (stationary) ARMA model for your data

Step 3. Estimate the parameters of the chosen model

Step 4. Conduct goodness-of-fit checks to ensure the model describes your data adequately

Step 5. After choosing a model and checking its fit and forecasting ability you can use the model to forecast

Why Stationarity?

The first step in Box-Jenkins methodology is to determine if the time series is stationary

Why we need the time series to be stationary?

Any stationary process can be approximated with stationary ARMA process (by Wold's theorem). It is the reason why ARMA models are very popular and that is why we need to make sure that the series is stationary to use these models

The detection of non-stationarity includes:

- Graphical (qualitative) analysis
  Plotting data over time, the autocorrelation function (ACF) and the partial autocorrelation function (PACF)
- Statistical tests

If the time series is not stationary it should be transformed to stationary

Autocorrelation Function (ACF)

Definition

Autocorrelation function (ACF) of stationary process $\{Y_t\}$ is

$$\rho(\tau) = \frac{c(\tau)}{c(0)}, \quad \tau = 0, 1, ...$$

where $c(\tau) = \mathrm{M}[(Y_t - \mu)(Y_{t-\tau} - \mu)]$ is autocovariance function, $\mu = \mathrm{M}[Y_t]$ is expectation of the process $\{Y_t\}$

Properties of autocovariances:

- $c(0) = \mathrm{D}[Y_t] \geq 0$
- $|c(\tau)| \leq c(0)$
- $c(\tau) = c(-\tau)$

Properties of ACF:

- $\rho(0) = 1$
- $|\rho(\tau)| \leq 1$
- $\rho(\tau) = \rho(-\tau)$

Sample Autocorrelation Function

Definition

Let $y_1, ..., y_T$ be a realization of stationary process $\{Y_t\}$. Sample autocorrelation function (sample ACF) is

$$\tilde{\rho}(\tau) = \frac{\tilde{c}(\tau)}{\tilde{c}(0)}, \quad \tau = 0, 1, ...$$

where $\tilde{c}(\tau) = \frac{1}{T-\tau} \sum_{i=1}^{T-\tau} (y_i - \bar{y})(y_{i+\tau} - \bar{y})$ is sample autocovariance

A more useful estimator of autocovariance for time series prediction tasks: $\tilde{c}(\tau) = \frac{1}{T} \sum_{i=1}^{T-\tau} (y_i - \bar{y})(y_{i+\tau} - \bar{y})$

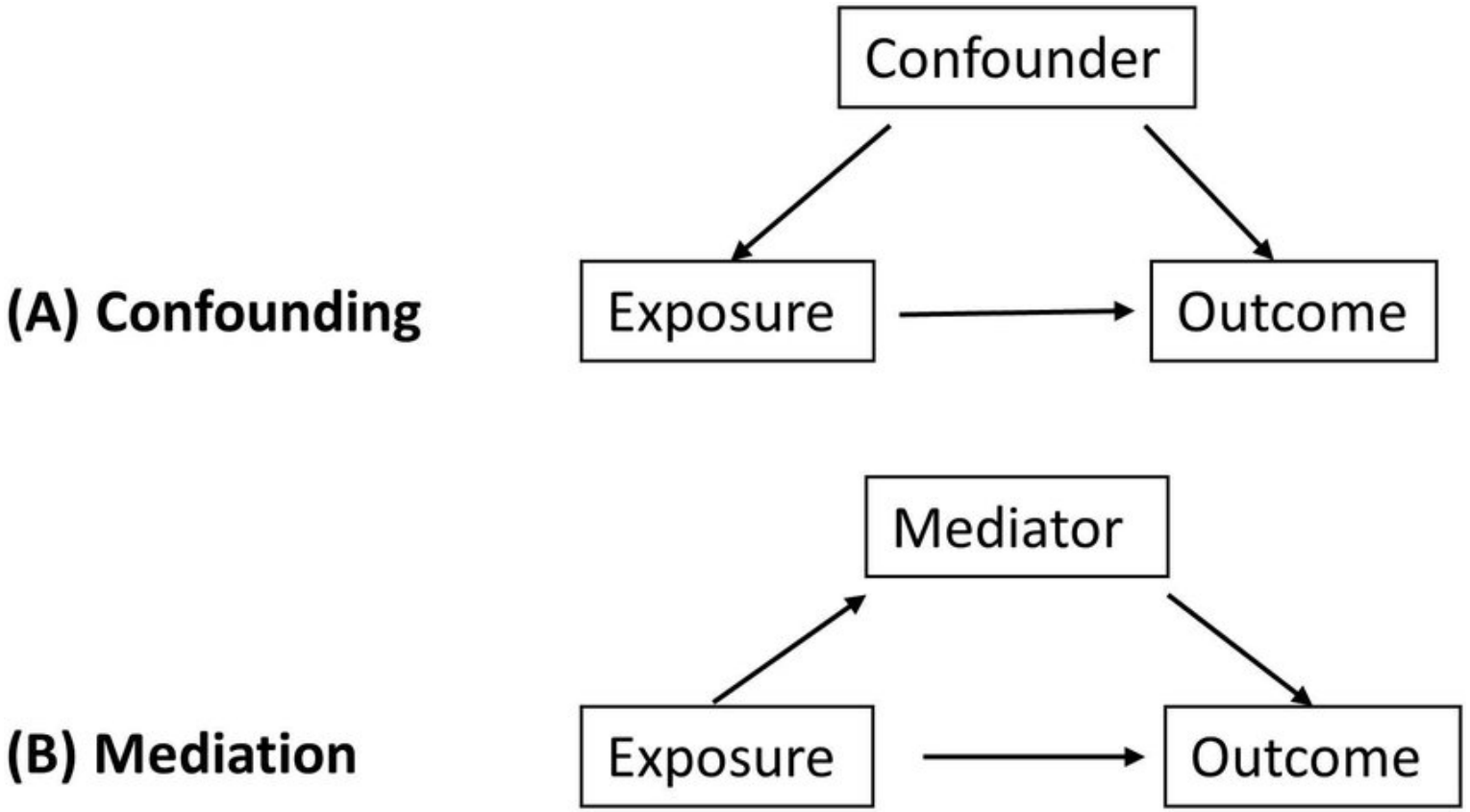
In practice, substitution $\frac{1}{T-\tau}$ by $\frac{1}{T}$ doesn't matter if $T \gg \tau$

Correlation and Relationship

Correlation coefficient ρ_{XY} represents the correlation (linear effect) but not the causation neither direct relationship between variables X and Y

- **Correlation $\neq$ causation**
  Assume that random variables X and Y are both affected by a some third factor Z (confounding variable) in the same manner. In this case, the correlation coefficient ρ_{XY} will be high but does it mean that there is a relationship between X and Y?
- **Correlation $\neq$ direct relationship**
  The correlation coefficient ρ_{XY} can be high because of indirect relationship mediated by factor Z (mediation variable, or mediator). Also ρ_{XY} can be insignificant because of masking true relationship by suppressor variable Z

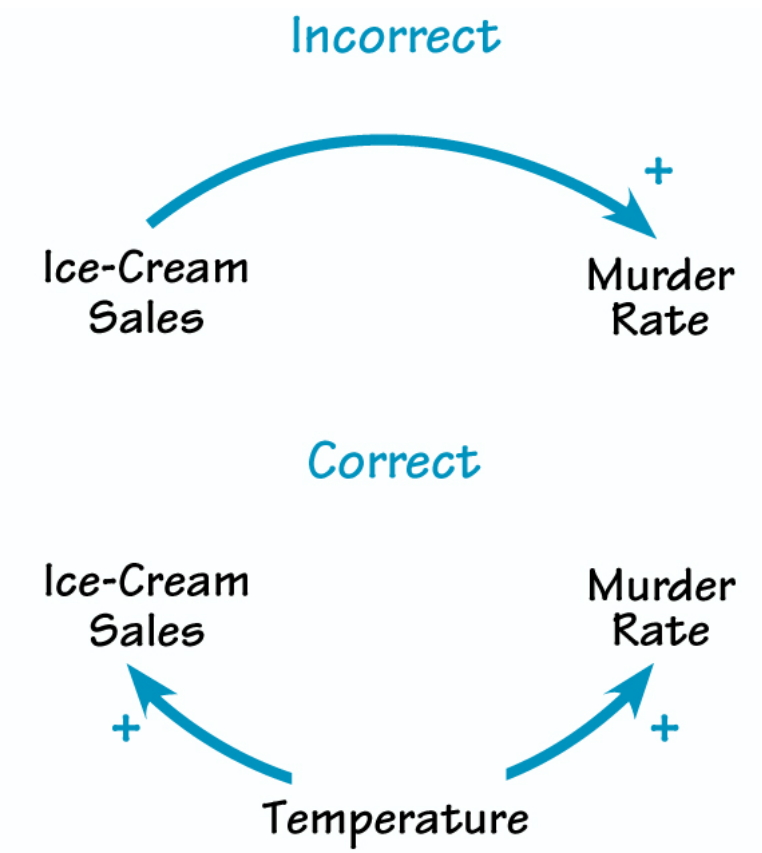
Confounding and Mediation



Confounding and mediation can lead to an underestimation (masking significant correlation) or an overestimation (spurious correlation) of the effect of X on Y

Example 1. Confounding Factor

There is high correlation between ice cream sales and homicides in New York. Does the consumption of ice cream causing the death of the people?

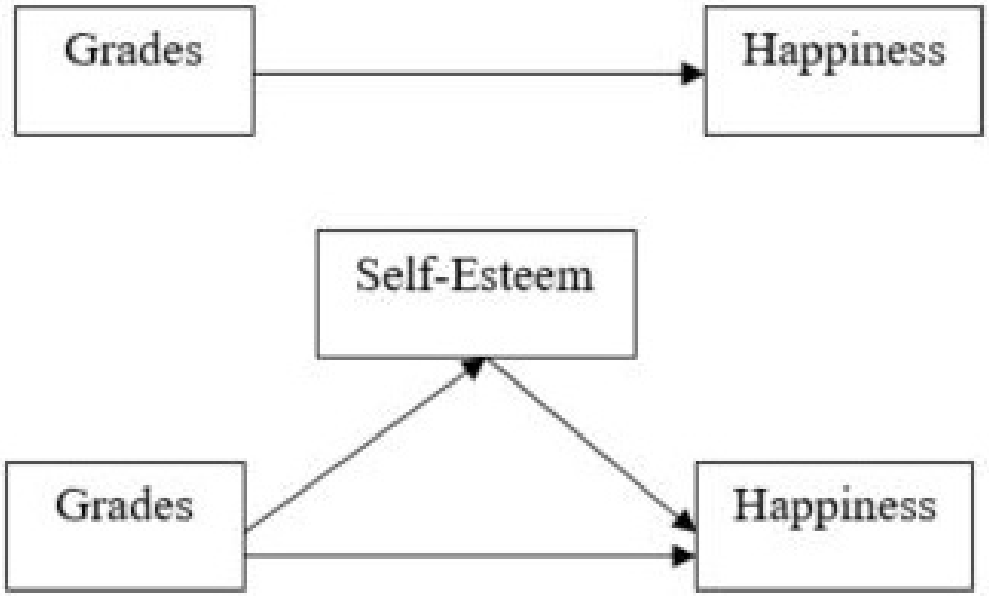


There is a hidden factor (weather) which is causing both the things. In summer people usually go out, enjoy nice sunny day and chill themselves with ice creams. So when it's sunny, wide range of people are outside and there is a wider selection of victims for predators

What correlation will be if we fix the weather?

Example 2. Mediation Factor

There is high correlation between grades in school and happiness. Does grades in school causing happiness?



The grades in school have a direct relationship on self-esteem, and then self-esteem has a direct relationship on happiness

What correlation will be if we fix the self-esteem?

Partial Correlation Coefficient

To measure the association between two random variables, with the effect of a set of controlling random variables removed, the partial correlation is used

Partial correlation coefficient $\pi_{XY|Z}$ between X and Y given controlling variables $Z = (Z_1, ..., Z_k)$ can be calculated by solving two associated linear regression problems (X on Z and Y on Z), get the residuals, and calculate the correlation between them:

$$\pi_{XY|Z} = corr(e_{XZ}, e_{YZ})$$

where $e_{XZ} = X - \mathrm{M}[X|Z]$ and $e_{YZ} = Y - \mathrm{M}[Y|Z]$ are residuals

If there is only one controlling variable Z ($k = 1$), then

$$\pi_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

Partial Autocorrelation Function (PACF)

Definition

Partial autocorrelation function (PACF) $\pi(\tau)$ of stationary process $\{Y_t\}$ is defined as

$$\pi(\tau) = \pi_{Y_t Y_{t-\tau} | Y_{t-1} \ldots Y_{t-\tau+1}}, \quad \tau = 0, 1, \ldots$$

$\pi(\tau)$ is a partial correlation coefficient between Y_t and $Y_{t-\tau}$ given controlling variables $Z = (Y_{t-1}, \ldots Y_{t-\tau+1})$

The PACF $\pi(\tau)$ measures the correlation between Y_t and $Y_{t-\tau}$ after eliminating the linear effects of intermediate variables $Y_{t-1}, \ldots Y_{t-\tau+1}$

ACF and PACF of AR(1) Process

AR(1) process:

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t, \quad (|\phi_1| < 1)$$

Autocovariance and ACF:

$$c(\tau) = \phi_1^\tau \frac{\sigma^2}{1-\phi_1^2} = \phi_1 c(\tau - 1)$$

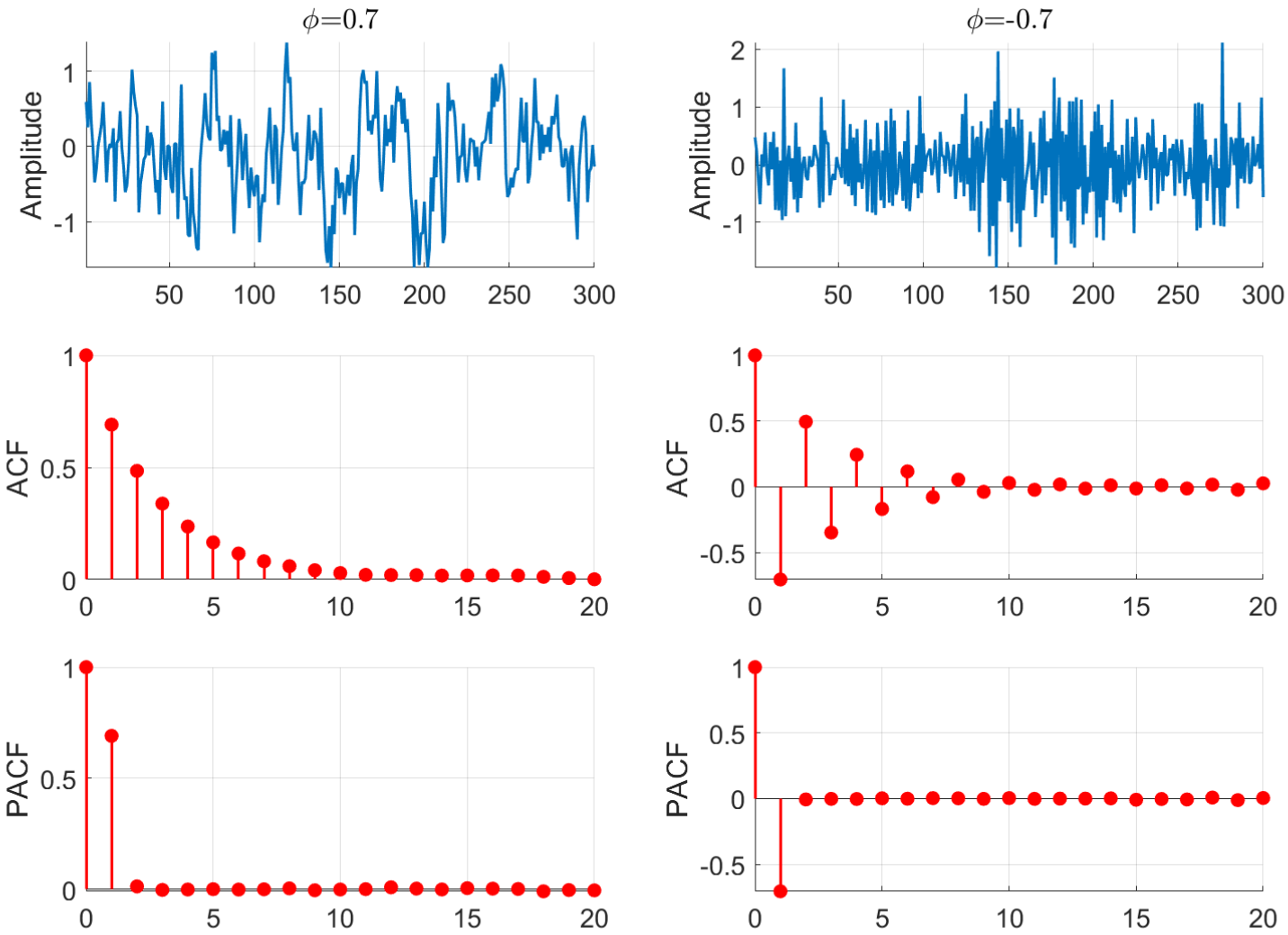
$$\rho(\tau) = \frac{c(\tau)}{c(0)} = \phi_1^\tau \frac{\sigma^2}{1-\phi_1^2} \Big/ \frac{\sigma^2}{1-\phi_1^2} = \phi_1^\tau, \quad \tau = 0, 1, ...$$

ACF decays exponentially and it changes sign every time moment if $\phi_1 < 0$

PACF: Y_t depends directly only on the previous variable Y_{t-1}, thus

$$\pi(\tau) = \begin{cases} 1, & \tau = 0, \\ \phi_1, & \tau = 1, \\ 0, & \tau > 1 \end{cases}$$

ACF and PACF of AR(1) Processes. Illustration



ACF and PACF of MA(1) Processes

MA(1) process:

$$Y_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t, \quad (|\theta_1| < 1)$$

Autocovariance and ACF (from Yule-Walker equations):

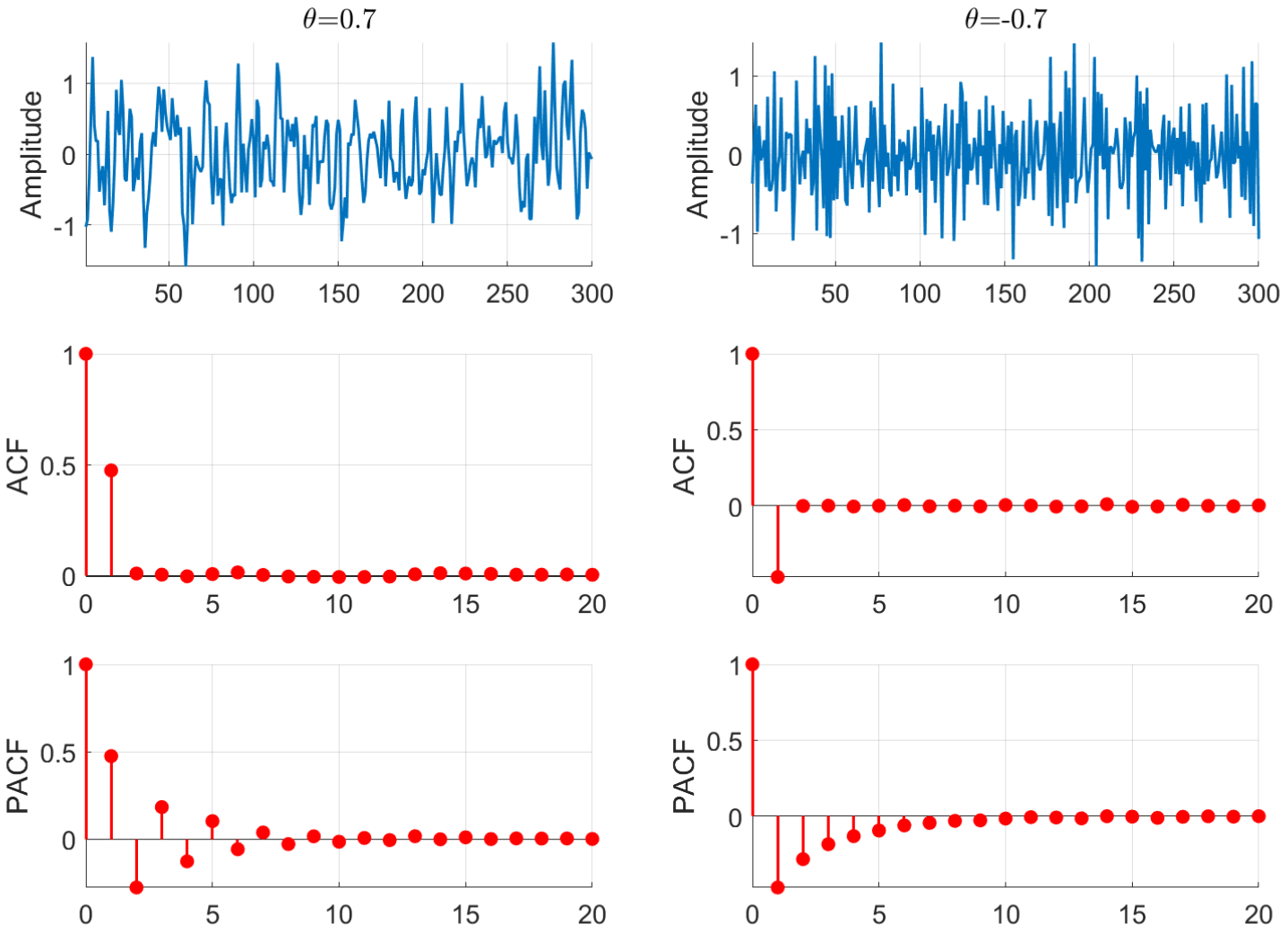
$$c(\tau) = \begin{cases} (1+\theta_1^2)\sigma^2, & \tau = 0, \\ \theta_1 \sigma^2, & \tau = 1, \\ 0, & \tau > 1 \end{cases} \qquad \rho(\tau) = \begin{cases} 1, & \tau = 0, \\ \frac{\theta_1}{1+\theta_1^2}, & \tau = 1, \\ 0, & \tau > 1 \end{cases}$$

AR(∞) form of invertible MA(1) process:

$$Y_t = \frac{c}{1+\theta_1} - \sum_{i=1}^{\infty} (-\theta_1)^i Y_{t-i} + \varepsilon_t$$

PACF: Y_t depends directly on all previous variables $Y_{t-1}, Y_{t-2}, \ldots$
Thus, PACF $\pi(\tau)$ decays exponentially and it changes sign every time moment if $\theta_1 > 0$

ACF and PACF of MA(1) Processes. Illustration



ACF and PACF of AR(p) Process

AR(p) process:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t$$

Autocovariance:

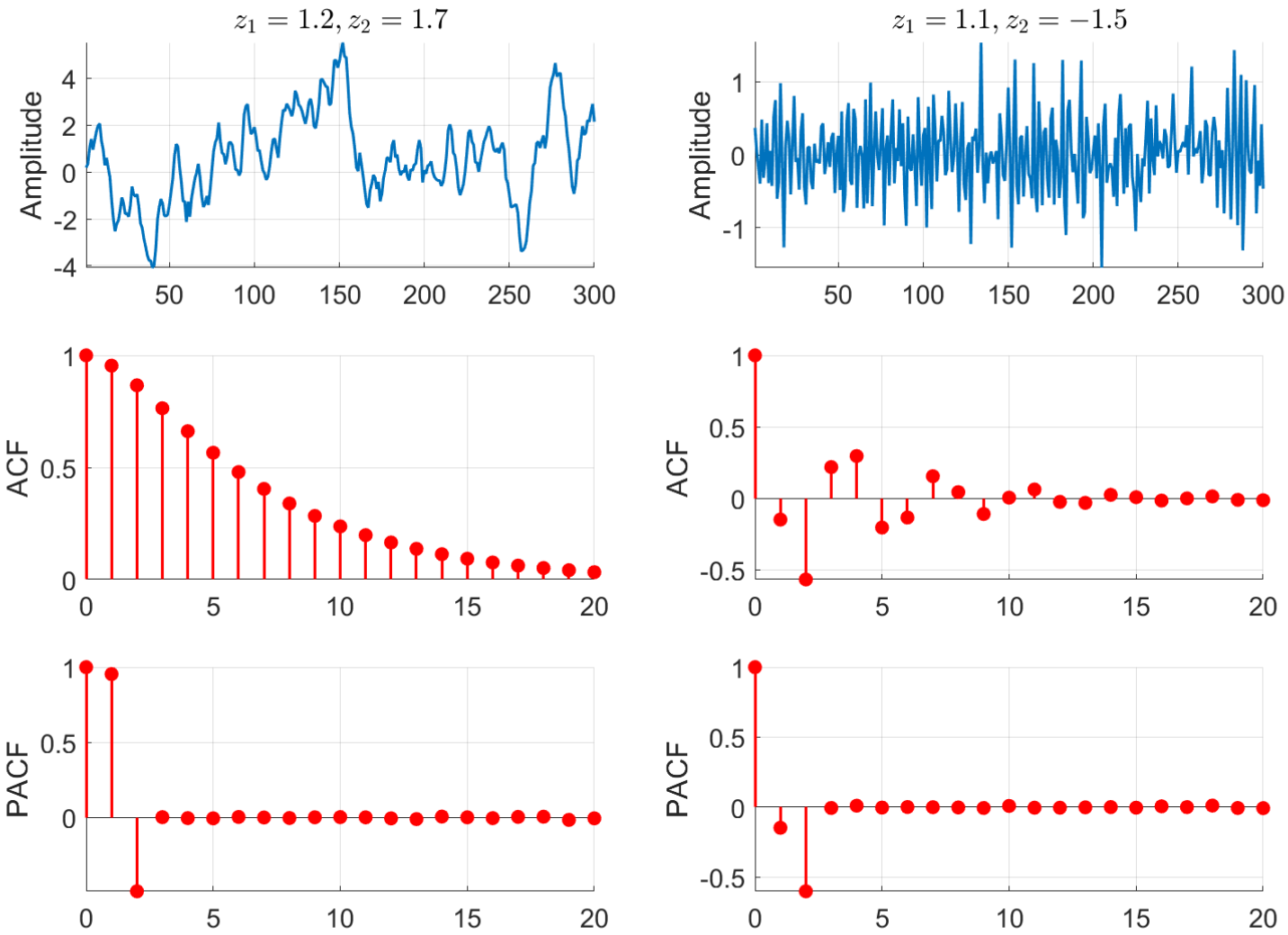
$$c(\tau) = \sum_{i=1}^{p} A_i \left(\frac{1}{|z_i|} \right)^{\tau}$$

where $A_1, ..., A_p$ are some constants and $z_1, ..., z_p$ are roots of characteristic polynomial $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$

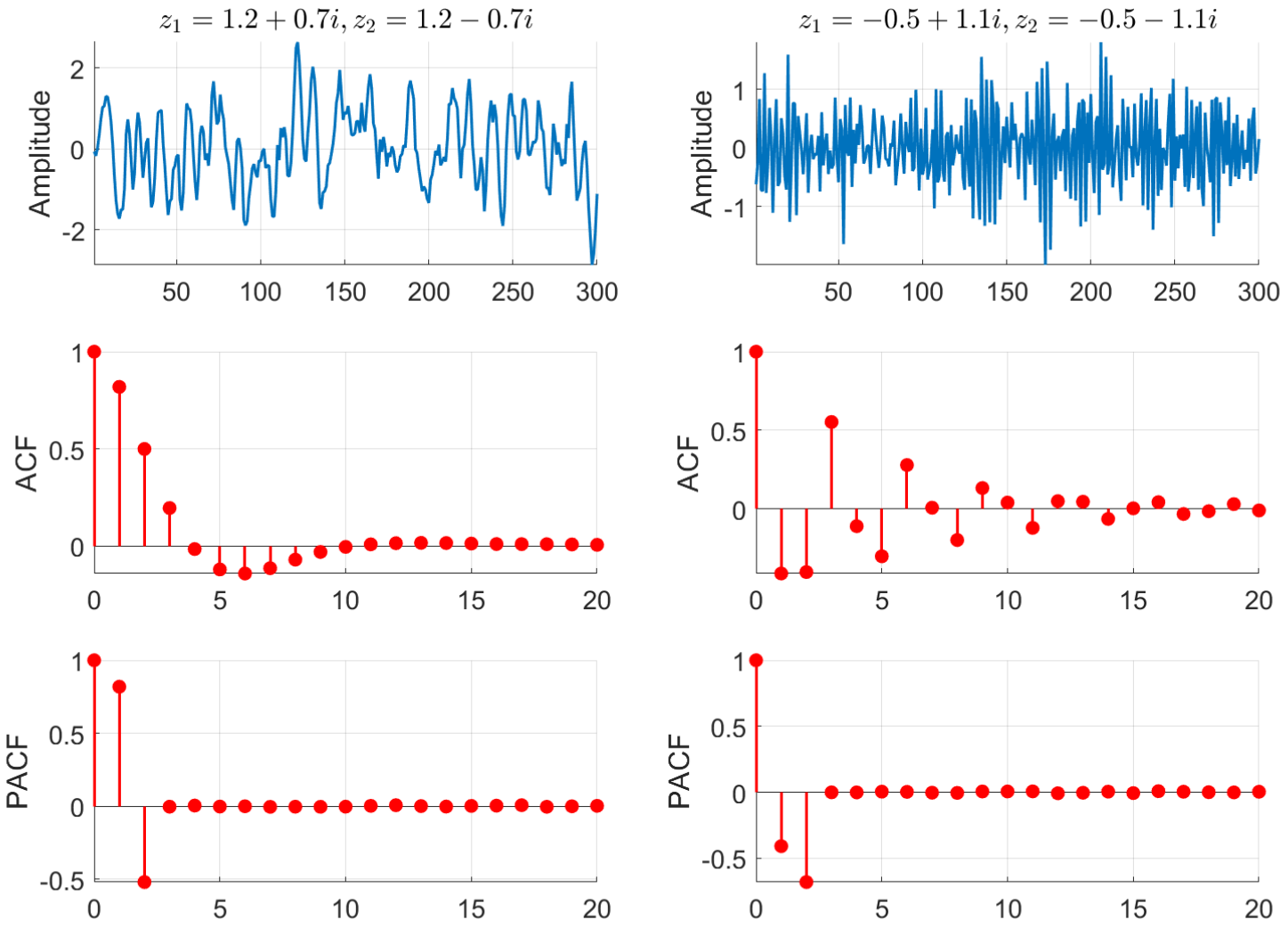
The AR(p) process is stable if all roots $z_1, ..., z_p$ are outside of unit circle. In this case $c(\tau) \to 0$ for $\tau \to \infty$

PACF: has only p non-zero values related to coefficients $\phi_1, ..., \phi_p$

ACF and PACF of AR(2) Processes. Illustration 1



ACF and PACF of AR(2) Processes. Illustration 2



ACF and PACF of MA(q) Process

MA(q) process:

$$Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

Autocovariance (from Yule-Walker equations):

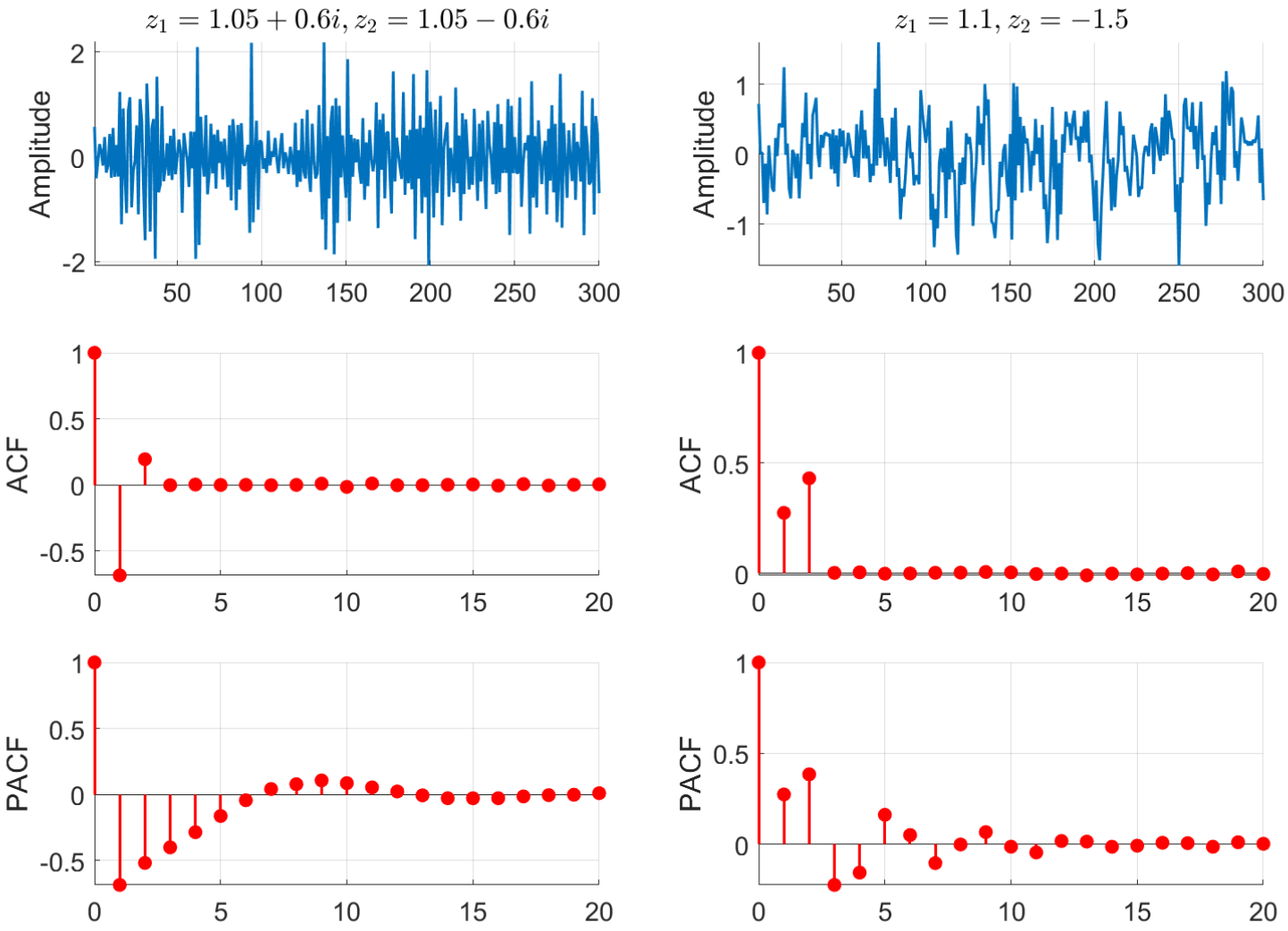
$$c(\tau) = \begin{cases} (1 + \theta_1^2 + ... + \theta_q^2)\sigma^2, & \tau = 0 \\ (\theta_\tau + \theta_{\tau+1}\theta_1 + ... + \theta_q \theta_{q-\tau})\sigma^2, & \tau = 1, ..., q \\ 0, & \tau > q \end{cases}$$

ACF has only q non-zero values related to coefficients $\theta_1, ..., \theta_q$

PACF: is non-zero for all τ as soon as invertible MA(q) process can be represented as AR(∞) process

Non-invertible MA(q) process can be represented as invertible MA(q) process by changing the coefficients and variance of innovations

ACF and PACF of MA(2) Processes. Illustration



ACF and PACF of ARMA(p, q) Process

ARMA(p,q) process:

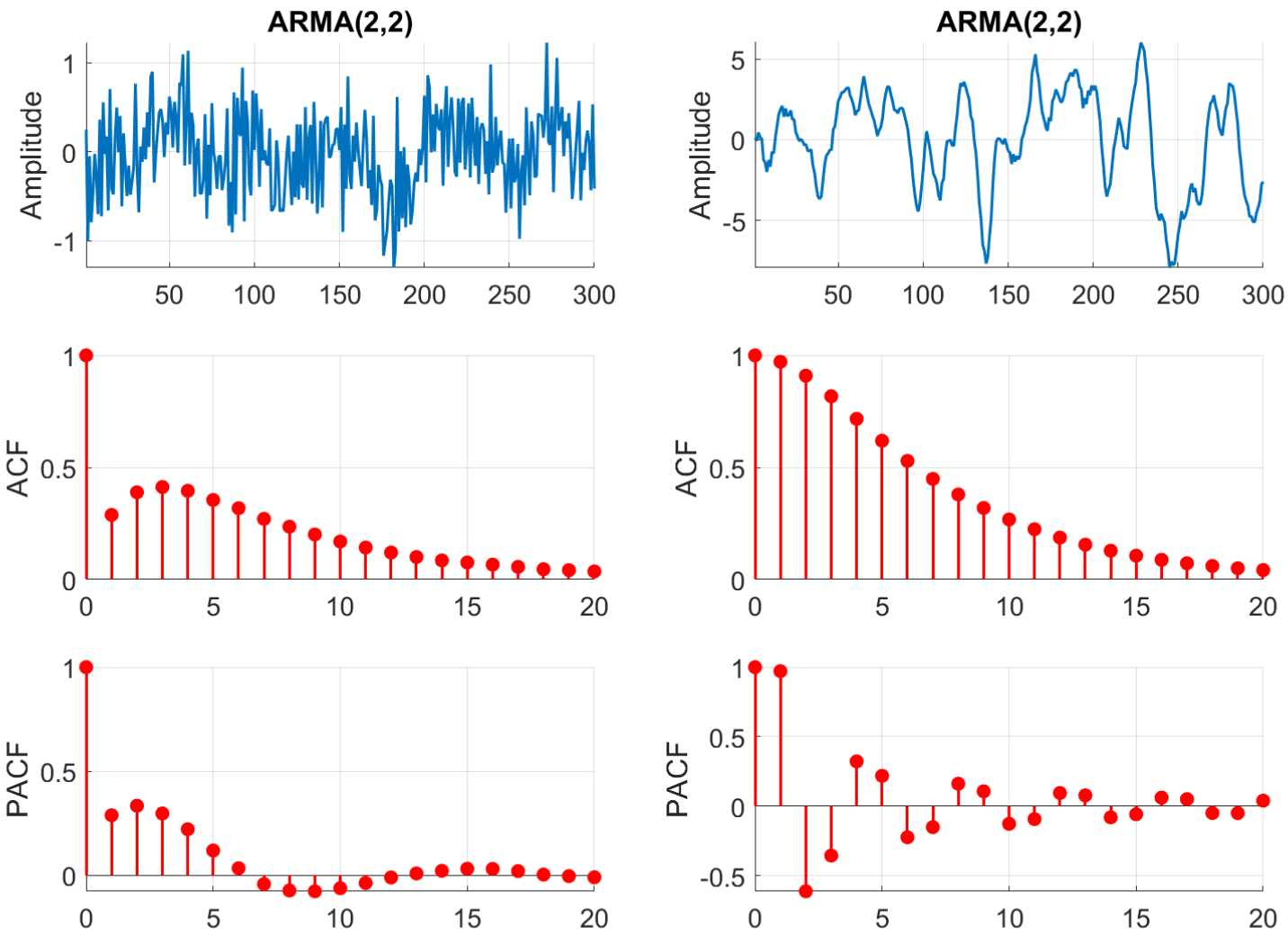
$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

Autocovariance: is a superposition of exponentially decaying autocovariances from AR part and q non-zero autocovariances from MA part

PACF: is a superposition of exponentially decaying autocovariances from MA part and p non-zero partial autocovariances from AR part

For ARMA processes neither ACF nor PACF have a cutoff, they are both non-zero for all τ

ACF and PACF of ARMA(2,2) Processes. Illustration



Confidence Bounds for ACF and PACF

The sample ACF and PACF may differ from the theoretical ones especially for small T

The confidence bounds (significance thresholds) are used to detect a significant deviation of autocorrelation sequence from zero (in assumption that observed process is a white noise):

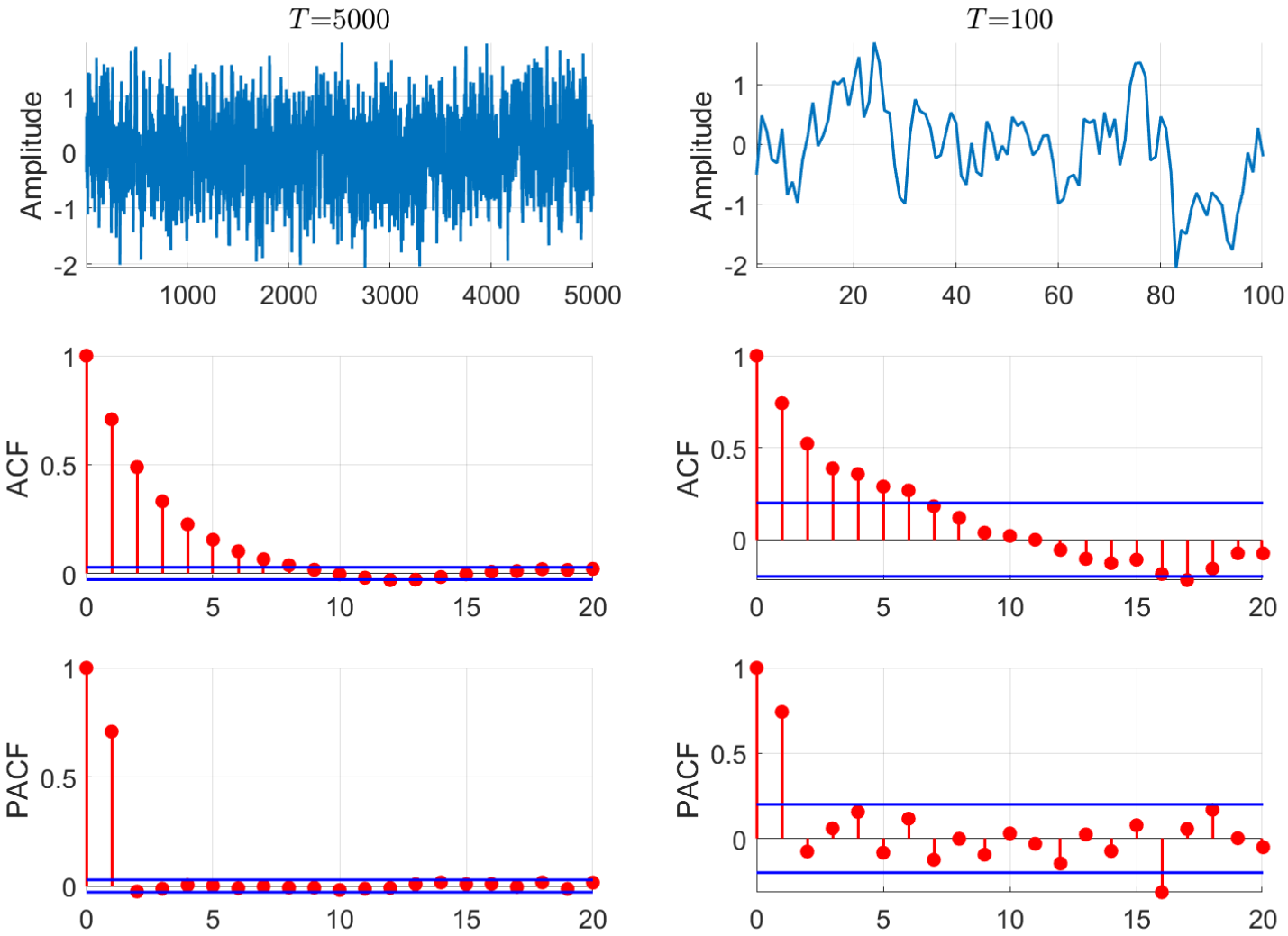
$$\pm u_{1-\alpha/2}\sigma[\tilde{\rho}(\tau)] = u_{1\pm\alpha/2}\frac{1}{\sqrt{T}}$$

where α is a significance level

If it's known that the observed process is MA(q) process:

$$\pm u_{1-\alpha/2}\sigma[\tilde{\rho}(\tau)] = u_{1\pm\alpha/2}\sqrt{\frac{1}{T}\left(1+2\sum_{\tau=1}^{q}\tilde{\rho}(\tau)\right)}$$

Sample ACF and PACF of AR(1) Processes. Illustration



Ljung-Box Q-Test

The Ljung-Box Q-test is a quantitative way to test for autocorrelation at multiple lags jointly

Null hypothesis $H_0: \ \rho(1) = \rho(2) = ... = \rho(m) = 0$

Test statistic:

$$Q(m) = T(T+2)\sum_{\tau=1}^{m}\frac{\tilde{\rho}(\tau)^2}{T-\tau}$$

$Q(m)|_{H_0} \sim \chi^2(m)$, the critical region is right-sided

Ljung-Box test statistic is a modified Box-Pierce test statistic:

$$Q_{BP}(m) = T\sum_{\tau=1}^{m}\tilde{\rho}(\tau)^2, \quad Q_{BP}(m)|_{H_0} \sim \chi^2(m)$$

It is shown, that $Q(m)|_{H_0}$ is better approximated by $\chi^2(m)$ distribution than $Q_{BP}(m)|_{H_0}$

Ljung-Box Q-Test. Notes

- For zero-mean Gaussian processes the Ljung-Box Q-test is a test for independence (a white noise test)
- The confidence bounds check the null hypothesis that a single autocorrelation coefficients are equal to zero independently. The Ljung-Box Q-test checks that all autocorrelation coefficients up to lag m are 0 simultaneously
- If m is too small, then the test does not detect high-order autocorrelations. If m is too large, then the test loses power when a significant correlation at one lag is washed out by insignificant correlations at other lags
- Test has a better power for $m \simeq \ln T$*
- Ljung-Box Q-test is also called as modified Box-Pierce test

*Tsay R. S. Analysis of Financial Time Series. 2nd Ed. Hoboken, NJ: John Wiley Sons, Inc., 2005.

Behaviour of ACF and PACF for Stationary ARMA Processes

Process	ACF	PACF
White noise	zero after lag 0	zero after lag 0
AR(p)	decays toward zero exponentially	zero after lag p
MA(q)	zero after lag q	decays toward zero exponentially
ARMA(p, q)	exponential decay after few lags	exponential decay after few lags

ACF and PACF of stationary ARMA process both decay to zero

The ARMA lags cannot be selected solely by looking at the ACF and PACF but their maximum number can be roughly estimated visually

ACF and PACF for Non-Stationary Processes

For non-stationary process autocovariance function $c(\tau)$ depends on τ differently in different fragments of time series. It means that the sample ACF and PACF cannot converge to the unique population ACF and PACF

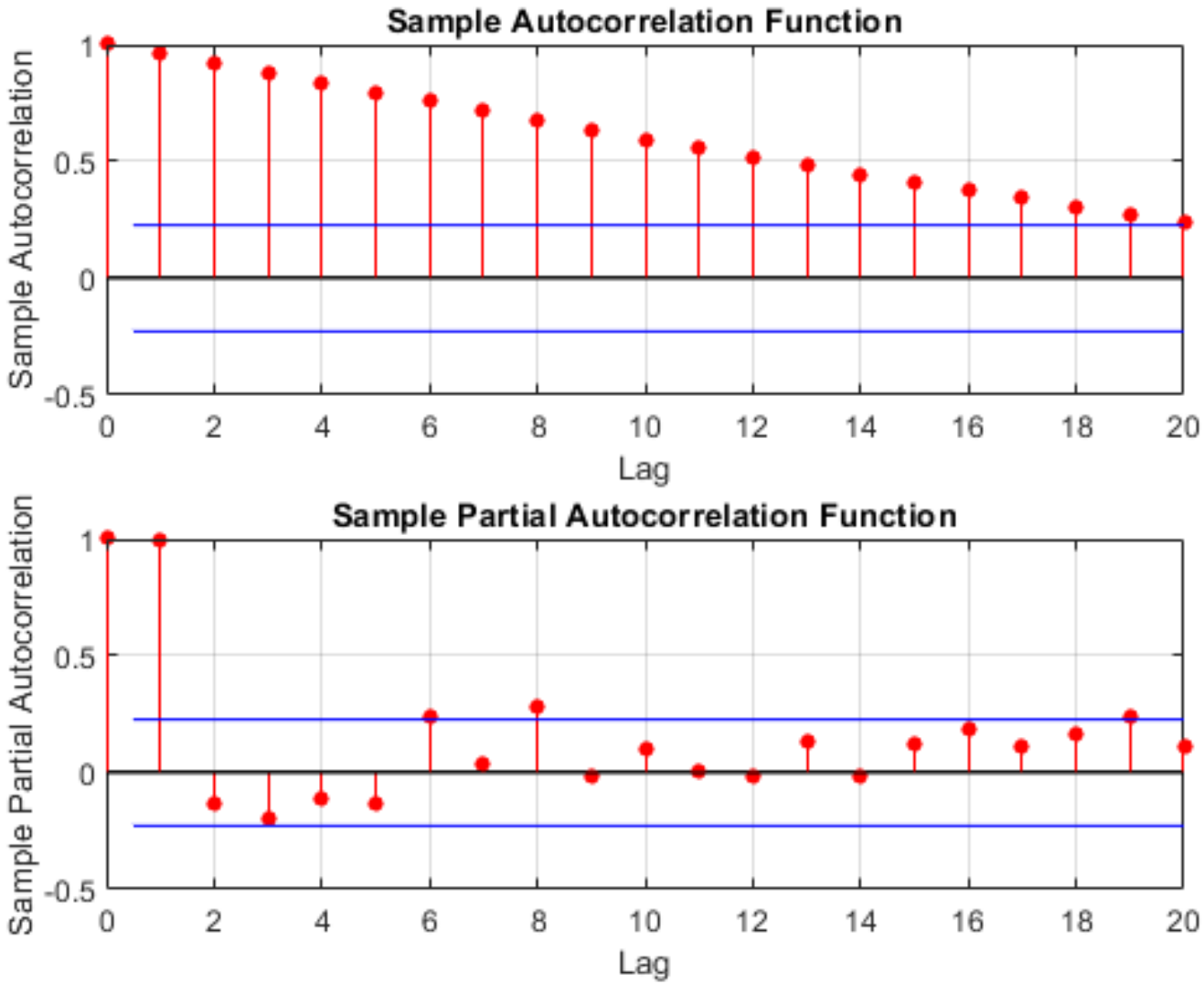
Signs of non-stationarity:

- ACF does not decrease to zero or has a very slow decay
- ACF has long downward sloping crossing zero line and continuing decay
- Linear decay of ACF or PACF

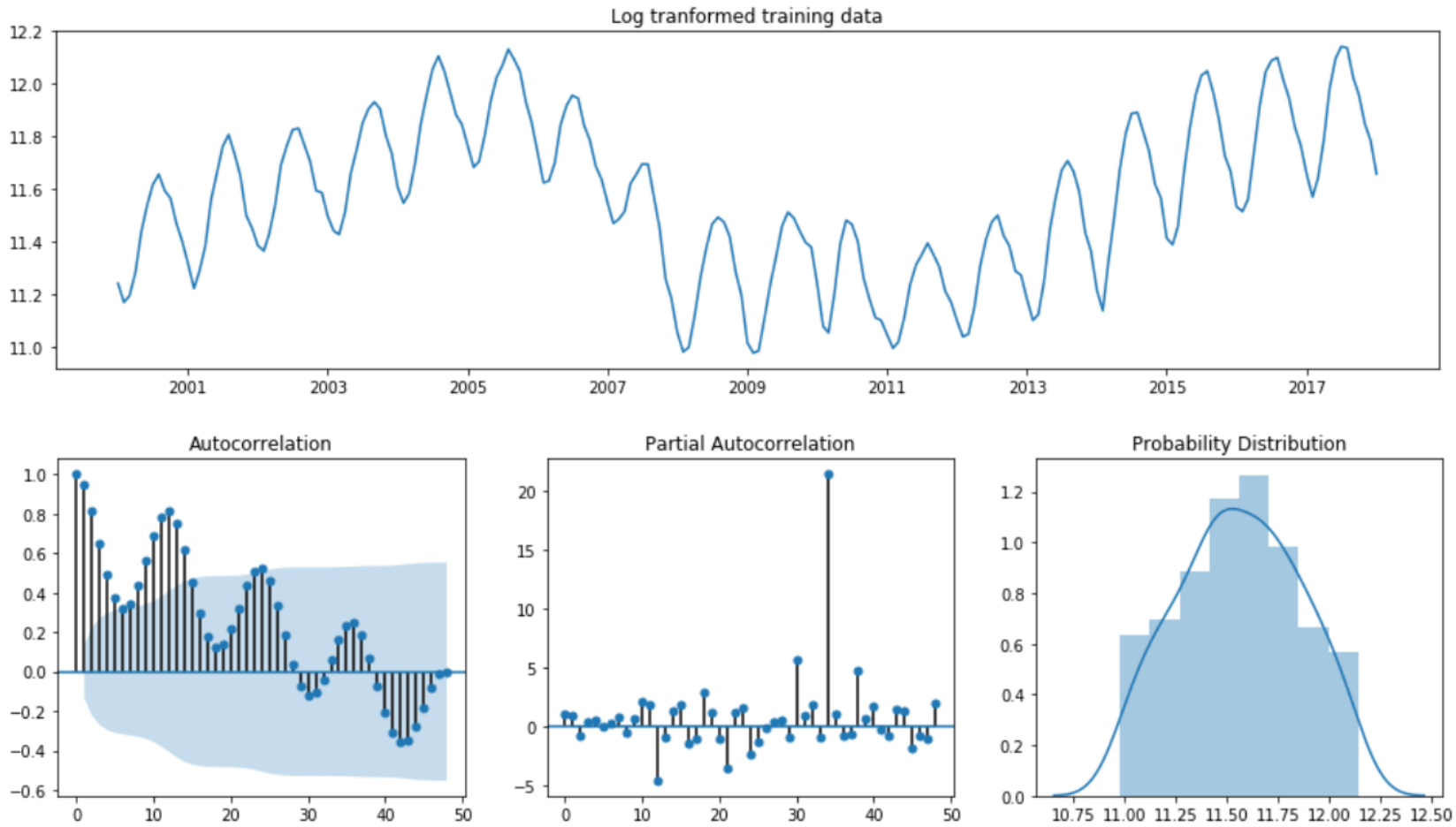
Common violations of stationarity are trending mean and seasonality. Usually they can be determined visually from time series plot

Some other types of non-stationarity are heteroscedasticity and structural breaks

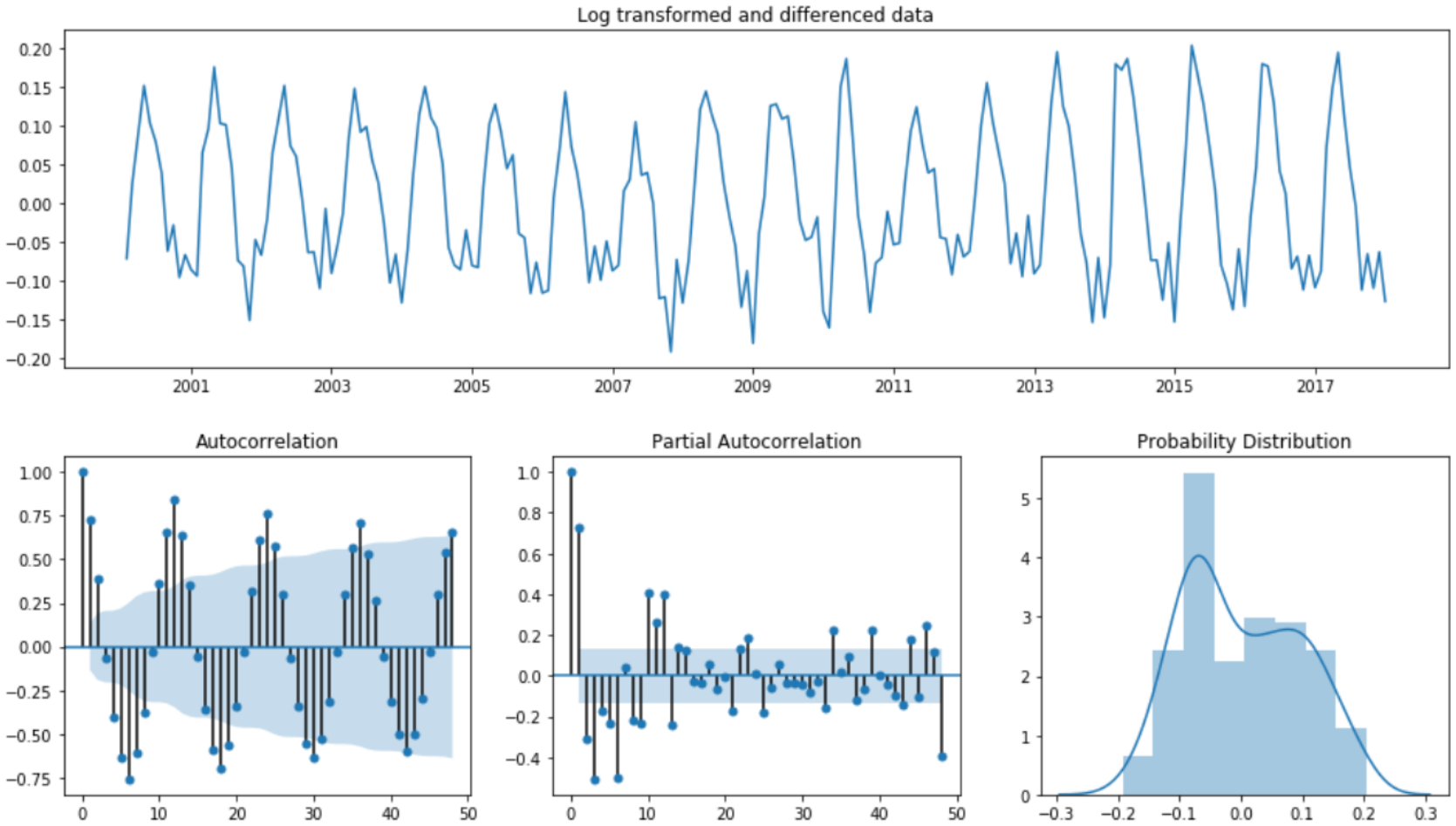
ACF and PACF for Non-Stationary Processes. Illustration 1



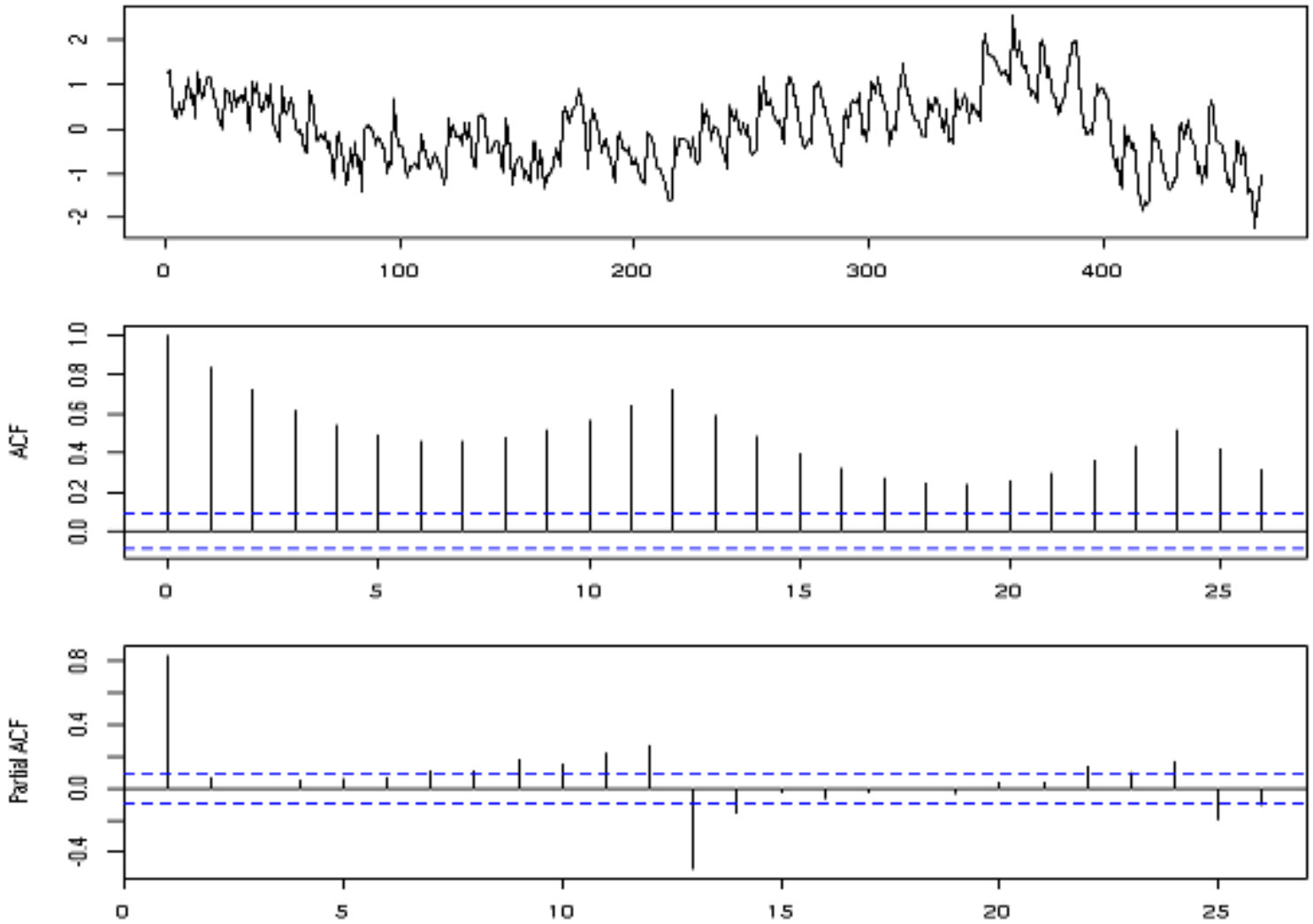
ACF and PACF for Non-Stationary Processes. Illustration 2



ACF and PACF for Non-Stationary Processes. Illustration 3



ACF and PACF for Non-Stationary Processes. Illustration 4



ACF and PACF for Non-Stationary Processes. Illustration 5