Analysis of time-series data

(based on the lecture of Prof. Vicent Lefieux, VIASM 2016)

Trịnh Quốc Anh

Bài giảng của DSLab

Viện nghiên cứu cao cấp về Toán (VIASM)

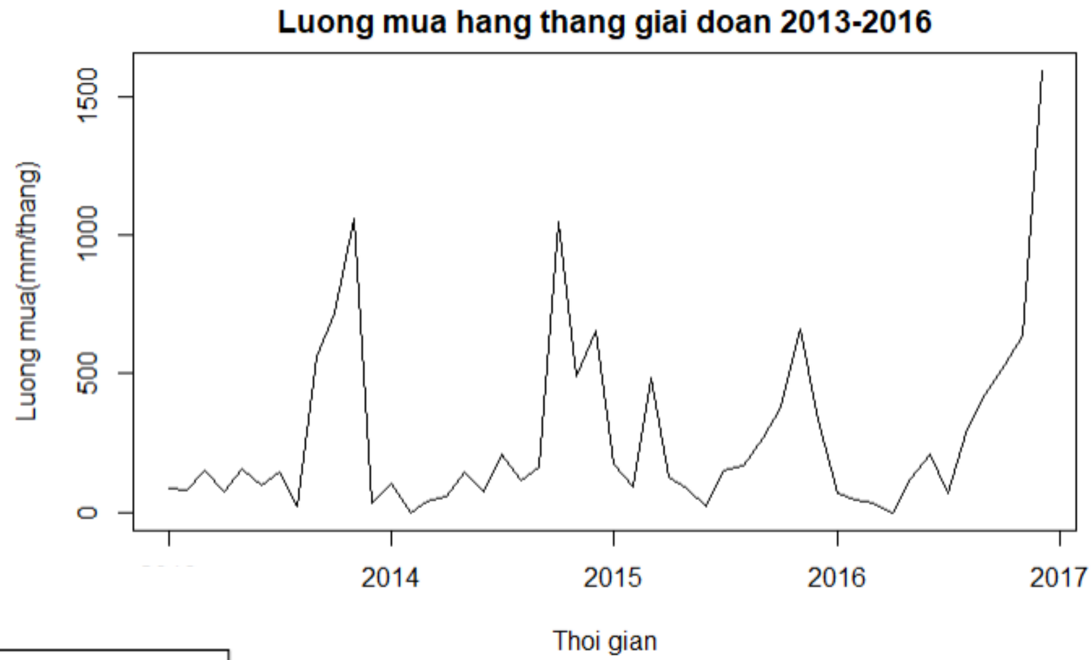
VIASM

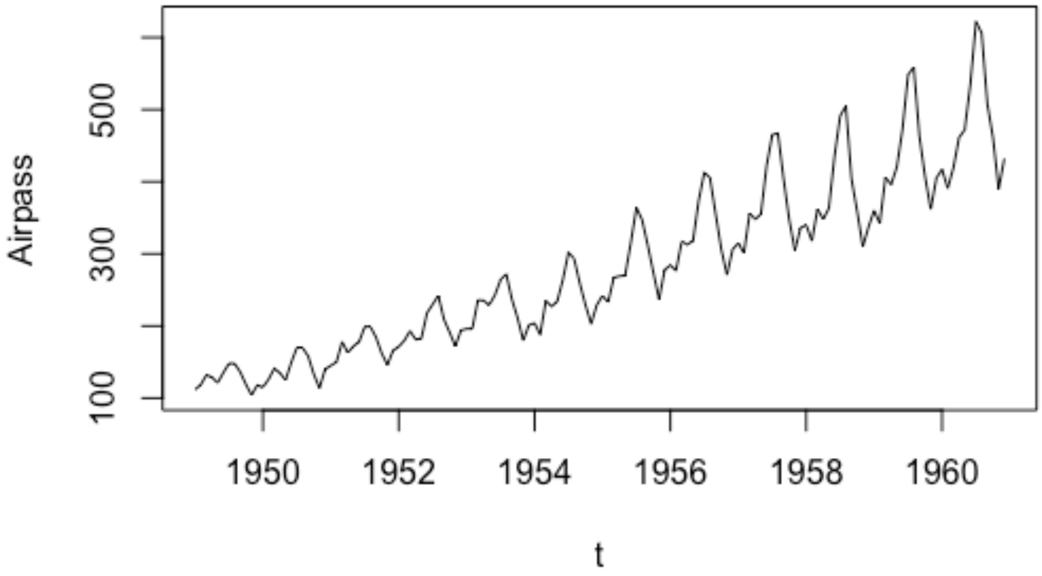
Vietnam Institute for Advanced Study in Mathematics

The object of study

- A time series is set of observations recorded in time order.
  - history (e.g. industrial revolution),
  - geography (e.g. migratory flows),
  - demography (e.g. growth of a population),
  - economics (e.g. rate of inflation),
  - finance (e.g. stock price),
  - meteorology (e.g. temperatures),
  - medicine (e.g. electrocardiogram),
  - epidemiology (e.g. spread of a disease),
  - geophysics (e.g. earthquakes),
  - communication (e.g. digital television),
  - energy (e.g. load curve, wind and solar generation),
  - . . .

Samples: Rainfall vs AirPassengers





Vietnam Institute for
Advanced Study in Mathematics

Course aims
(Time-series analysis in practice)

- understand the stationarity of time dependent data,
- use R statistics tools to detect seasonality,
- be able to estimate model's parameters of time series,
- and forecast.

References

[1] Gwilym M. Box, George E. P.and Jenkins and Gregory C. Reinsel. *Time series. Theory and methods*. Wiley, Fourth edition, 2008.

[2] Peter J. Brockwell and Richard A. Davis. *Introduction to time series and forecasting*. Springer, Second edition, 2002.

[3] Peter J. Brockwell and Richard A. Davis. *Time series: Theory and methods*. Springer, Second edition, 2009.

[4] Michael Friendly. A brief history of data visualization. In Chun-houh Chen, Wolfgang Hardle, and Antony Unwin, editors, *Handbook of data visualization*, chapter II.1, pages 15{56. Springer, 2008.

[5] Alan Pankratz. *Forecasting with dynamic regression models*. Wiley, 1991.

[6] Robert H. Shumway and David S. Stoer. *Time series analysis and its applications. With R examples*. Springer Texts in Statistics. Springer, 3 edition, 2011.

Table of content

- Methodology
- Stationarity
- Order selection
- Estimation
- Diagnostic checking
- Model selection
- Forecasting

Methodology

Probabilistic model

Let $(x_t)_{t\in\mathcal{T}}$ be a sequence of observations (for example in the fields of economics, life sciences, physics. . .).

Each observation x_t, in $\mathbb{R}^d$, is recorded at a specific time $t \in \mathcal{T}$.

$(x_t)_{t\in\mathcal{T}}$ is called time series.

- Observation x_t is considered as the realization of a random variable X_t.
- Time series $(x_t)_{t\in\mathcal{T}}$ is considered as the realization of a stochastic process $(X_t)_{t\in\mathcal{T}}$.

Methodology

Stationary

- To obtain parsimony in a time series model we often assume some form of distributional invariance over time, or stationarity.
- For observed time series:
  - Fluctuations appear random.
  - However, same type of stochastic behavior holds from one time period to the next.
- For example, returns on stocks or changes in interest rates:
  - Individually, very different from the previous year.
  - But mean, standard deviation, and other statistical properties are often similar from one year to the next.

Methodology

Stationary processes

$(X_t)_{t\in\mathbb{Z}}$ is said to be strictly stationary if the joint distribution of $(X_{t_1},\dots,X_{t_k})$ is equal to the distribution of $(X_{t_1+h},\dots,X_{t_k+h})$, for $k\in\mathbb{N}^\star$, $(t_1,\dots,t_k)\in\mathbb{Z}^k$ and $h\in\mathbb{Z}$.

Methodology

(Weakly) Stationary processes

$(X_t)_{t\in\mathbb{Z}}$ is said to be a second-order process if:

$$\forall t \in \mathbb{Z} : \mathbb{E}\left(X_t^2\right) < +\infty.$$

A second-order process $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary, if the expectation $\mathbb{E}\left(x_t\right)$ and the (auto)covariances $\mathrm{Cov}(X_s, X_t)$ are time-shifted invariant:

- $\forall t \in \mathbb{Z} : \mathbb{E}\left(x_t\right) = \mu$
- $\forall \left(s, t\right) \in \mathbb{Z}^2, \forall h \in \mathbb{Z} :$

$$\mathrm{Cov}(X_s, X_t) = \mathrm{Cov}(X_{s+h}, X_{t+h}).$$

In this case we have:

$$\mathrm{Cov}\left(X_s, X_t\right) = \gamma\left(t - s\right).$$

Methodology

Weakly Stationary

Weakly stationary is also referred to covariance stationary.

- The mean and variance do not change with time
- The covariance between two observations depends only on the lag, the time distance $|t - s|$ between observations, not on the indices t or s.

Autocovariance function

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process.
The autocovariance function of X is:

$$\forall h \in \mathbb{Z} : \gamma(h) = \mathrm{Cov}\left(X_t, X_{t-h}\right).$$

- $\gamma(0) \geq 0$
- $\forall h \in \mathbb{Z} : |\gamma(h)| \leq \gamma(0)$.
- γ is even:
$$\forall h \in \mathbb{Z} : \gamma(-h) = \gamma(h).$$
- γ is a nonnegative definite function:
$$\forall n \in \mathbb{N}^*, \forall (a_i)_{i\in\{1,\dots,n\}} \in \mathbb{R}^n : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(i-j) \geq 0.$$

Autocorrelation function

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process.
We call (simple) autocorrelation function of X the following function ρ:

$$\forall h \in \mathbb{Z} : \rho(h) = \mathrm{Corr}(X_t, X_{t-h}) = \frac{\gamma(h)}{\gamma(0)}.$$

We have $\rho(0) = 1$.

Autocorrelation matrix

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process.
The autocorrelation matrix of $(X_t, \dots, X_{t-h+1})$ is:

$$R_h = \begin{bmatrix} 1 & \rho(1) & \dots & \rho(h-1) \\ \rho(1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho(1) \\ \rho(h-1) & \dots & \rho(1) & 1 \end{bmatrix}.$$

Notice that:

$$R_h = \left[\begin{array}{ccc|c} & & & \rho(h-1) \\ & R_{h-1} & & \vdots \\ & & & \rho(1) \\ \hline \rho(h-1) & \dots & \rho(1) & 1 \end{array}\right]$$

Estimation of the autocorrelation functions

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process.
Based on $(X_1,\ldots,X_T)$, $\overline{X}_T$ is a consistent and unbiased estimator of $\mathbb{E}(X)=\mu$:

$$\overline{X}_T=\frac{1}{T}\sum_{t=1}^{T}X_t.$$

$\forall h\in\{1,\ldots,T-1\}$:

$$\widehat{\rho}(h)=\frac{\sum_{t=h+1}^{T}\left(X_t-\overline{X}_T\right)\left(X_{t-h}-\overline{X}_T\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}_T\right)^2}.$$

Test of randomness

Portmanteau test

Portmanteau test

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process.
Consider the test:

$$\begin{cases} H_0 : (X_t)_{t\in\mathbb{Z}} \text{ is a white noise} \\ H_1 : (X_t)_{t\in\mathbb{Z}} \text{ isn't a white noise} \end{cases} .$$

Based on $(X_1, \dots, X_T)$, the Portmanteau statistic is:

$$Q_k = T \sum_{h=1}^{k} \widehat{\rho}^2(h)$$

Under H_0: $Q_k \xrightarrow{d} \chi_k^2$.
So we reject the null hypothesis at the α level if $Q_k > \chi_k^2(1-\alpha)$.

Test of randomness

Shapiro-Wilk test

$$\begin{cases} H_0 : (X_1, \ldots, X_n) \text{ comes from a normal distribution} \\ H_1 : (X_1, \ldots, X_n) \text{ doesn't come from a normal distribution} \end{cases}$$

The Shapiro-Wilk statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}$$

where $X_{(i)}$ is the i-th order statistic. Coefficients $(a_i)_{i\in\{1,\ldots,n\}}$ are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution.
We reject the null hypothesis at the α level if:

$$W < W_{n,\alpha}^{threshold}.$$

$W^{threshold}$ is obtained with Monte Carlo simulations

AR process

Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a white noise of variance σ^2.

$(X_t)_{t\in\mathbb{Z}}$ is said to be an autoregressive process or a AR process of order p, written AR(p), if:

- $(X_t)_{t\in\mathbb{Z}}$ is stationary,
- $$\forall t \in \mathbb{Z} : X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

  where $(\varphi_1, \dots, \varphi_p) \in \mathbb{R}^p$ and $\varphi_p \neq 0$.

We generally use the notation $\Phi(B) X_t = \varepsilon_t$ where:

$$\Phi(B) = I - \sum_{i=1}^{p} \varphi_i B^i.$$

Note that:

- Sometimes we find $\Phi(B) = I + \sum_{i=1}^{p} \varphi_i B^i$.
- If $\Phi(B)$ has a root on the unit circle then the process $(X_t)_{t\in\mathbb{Z}}$ isn't stationary, thus it isn't an AR process.

MA Process

Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a white noise of variance σ^2.

$(X_t)_{t\in\mathbb{Z}}$ is said to be a MA process of order q, written MA(q), if:

$$\forall t \in \mathbb{Z} : X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

where $(\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$ and $\theta_q \neq 0$.

Autocovariance and Autocorrelation

The function γ *(gamma) is called the* *autocovariance function.*

Note that $\gamma(h) = \gamma(-h)$. *Why?*

Assuming weak stationarity:

Correlation between Y_t *and* Y_{t+h} *is denoted by* $\rho(h)$.

The function ρ *(rho) is called the* *autocorrelation function.*

Note:

- $\gamma(0) = \sigma^2$ *(variance)*
- $\gamma(h) = \sigma^2 \rho(h)$ *(autocovariance)*
- $\rho(h) = \gamma(h)/\sigma^2 = \gamma(h)/\gamma(0)$ *(autocorrelation)*

Estimating Parameters of a Stationary Process

Suppose we observe $Y_1, \ldots, Y_n$ from a weakly stationary process.

Estimate the mean μ and variance σ^2 using:

- the sample mean $\overline{y}$ and sample variance s^2.

Estimate the autocovariance function using

- the sample autocovariance function

$$\widehat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (Y_{t+h} - \overline{y})(Y_t - \overline{y}) = n^{-1} \sum_{t=h+1}^{n} (Y_t - \overline{y})(Y_{t-h} - \overline{y}).$$

Estimating Autocorrelations of a Stationary Process

To estimate $\rho(\cdot)$, we use the sample autocorrelation function (sample ACF) defined as

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)},$$

for each lag h.

ACF Plot

R will plot a sample ACF with test bounds.

- Bounds test the null hypothesis that an autocorrelation coefficient is 0.
- The null hypothesis is rejected if the sample autocorrelation is outside the bounds.
- The usual level of the test is $\alpha = 0.05$
- We expect 1 out of 20 sample autocorrelations outside the test bounds simply by chance.

ACF plot example

Inflation rates and changes in the inflation rate—sample ACF plots

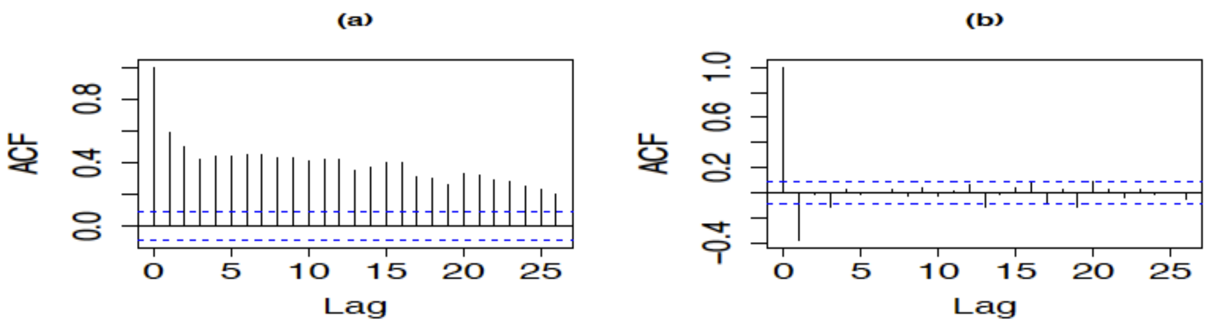


Figure: *Sample ACF plots of the one-month inflation rate (a) and changes in the inflation rate (b).*

```
data(Mishkin, package = "Ecdat")
y = as.vector(Mishkin[,1])
par(mfrow=c(1,2))
acf(y)
acf(diff(y))
```

Vietnam Institute for Advanced Study in Mathematics

Autoregressive (AR) processes

Let $\epsilon_1, \epsilon_2, \ldots$ be White Noise(0,σ_ϵ^2) innovations, with variance σ_ϵ^2

Then, $Y_1, Y_2, \ldots$ is an AR process if for some constants μ and ϕ,

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t$$

- We focus on 1st order case, the simplest AR process

Autoregressive (AR) processes

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t$$

- μ is the mean of the $\{Y_t\}$ process
- If $\phi = 0$, then $Y_t = \mu + \epsilon_t$, such that Y_t is White Noise(μ, σ_ϵ^2)
- If $\phi \neq 0$, then observations Y_t depend on both ϵ_t and Y_{t-1}
- And the process $\{Y_t\}$ is autocorrelated
- If $\phi \neq 0$, then $(Y_{t-1} - \mu)$ is fed forward into Y_t
- ϕ determines the amount of feedback
- Larger values of $|\phi|$ result in more feedback

AR Processes: Properties

If $|\phi| < 1$, then

$$\begin{aligned} E(Y_t) &= \mu \\ \mathrm{Var}(Y_t) &= \sigma_Y^2 = \frac{\sigma_\epsilon^2}{1-\phi^2} \\ \mathrm{Corr}(Y_t, Y_{t-h}) &= \rho(h) = \phi^{|h|} \quad for\ all\ h \end{aligned}$$

- If $\mu = 0$ and $\phi = 1$, then

$$Y_t = Y_{t-1} + \epsilon_t$$

which is a random walk process, and $\{Y_t\}$ is NOT stationary

Simple moving average (MA) processes

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

- μ is the mean of the $\{Y_t\}$ process
- If $\theta = 0$, then $Y_t = \mu + \epsilon_t$, such that Y_t is White Noise(μ, σ_ϵ^2)
- If $\theta \neq 0$, then observations Y_t depend on both ϵ_t and ϵ_{t-1}
- And the process $\{Y_t\}$ is autocorrelated
- If $\theta \neq 0$, then ϵ_{t-1} is fed forward into Y_t
- θ determines its impact
- Larger values of $|\theta|$ result in greater impact

MA Processes: Autocorrelations

$$\begin{aligned} \text{Corr}(Y_t, Y_{t-1}) &= \rho(1) = \frac{\theta}{1+\theta^2} \\ \text{Corr}(Y_t, Y_{t-h}) &= \rho(h) = 0 \qquad \textit{for all } h > 1 \end{aligned}$$

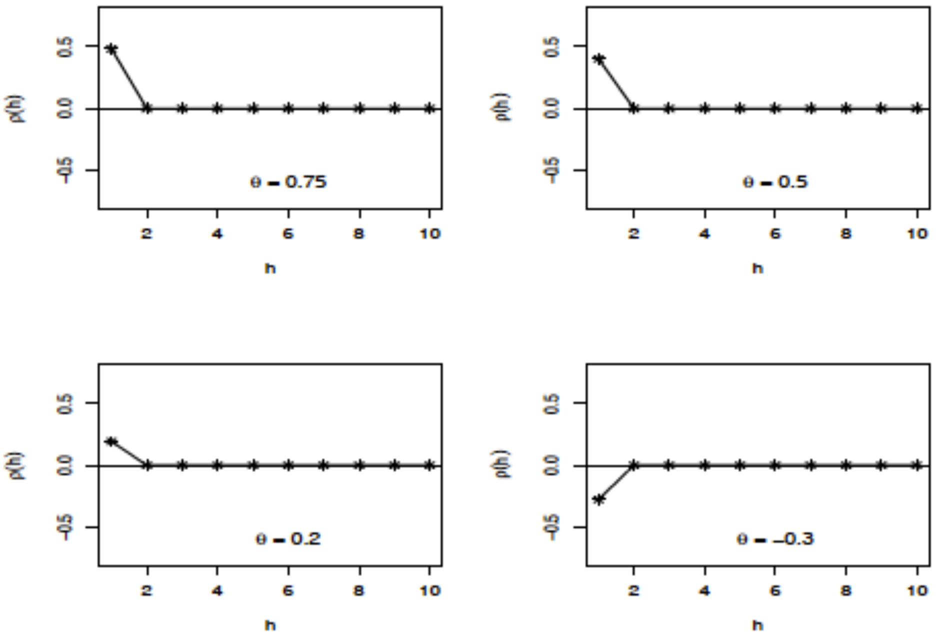


Figure: *Autocorrelation functions of MA processes with* θ *equal to* 0.75, 0.5, 0.2, *and* -0.3.

Time-series analysis with R

- Simulation
- AirPassenger data
- Rainfall data