Basics on time series analysis

Yuta Koike

Mathematics and Informatics Center, The University of Tokyo
CREST JST

June 25, 2019

YSS2019

Time series analysis

- Time series analysis aims at analyzing **time series data**
- A main feature of time series data is that they are recorded along time
  - The order of observations plays a role
  - The independence assumption is often NOT reasonable; it is important to model the dependence structure btw different time points
- This session briefly describes basic models and tools to analyze time series
  - We focus on discrete-time stochastic processes, but many concepts used there are extended to continuous-time stochastic processes
  - R has the class `ts` to handle time series data observed at equi-spaced time points

Auto-covariance and auto-correlation functions

- $X = (X_1, \ldots, X_T)$: Time series data
  - X_t denotes the data observed at the time t
  - Each X_t is assumed to be a random variable
- A standard way to measure the dependence btw X_t and X_s is to evaluate their covariance

$$\mathrm{Cov}[X_t, X_s] = \mathrm{E}[(X_t - \mathrm{E}[X_t])(X_s - \mathrm{E}[X_s])]$$

- It would be natural to expect that this dependence is related to how the time t is far from the time s
  $\Longrightarrow$ This suggests us to model the above covariance as a function of $|t - s|$

Auto-covariance and auto-correlation functions

- The **auto-covariance function (ACF)** of X is a function $\gamma(h)$ such that

$$\mathrm{Cov}[X_t, X_s] = \gamma(|t-s|)$$

  for all t, s
- In this case, the variance of X_t is given by $\mathrm{Var}[X_t] = \gamma(0)$, so it does not depend on t
- Thus, the correlation coefficient btw X_t and X_s is given by

$$\mathrm{Corr}[X_t, X_s] = \gamma(|t-s|)/\gamma(0)$$

- The function $\rho(h) = \gamma(h)/\gamma(0)$ is called the **auto-correlation function** of X

Auto-covariance and auto-correlation functions

- How to estimate the ACF from data?
- One would naturally use the following sample counterpart:

$$\hat{\gamma}(h) := \frac{1}{T}\sum_{t=1}^{T-h}(X_t - \bar{X})(X_{t+h} - \bar{X}),$$

  where

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$$

  is the sample mean
- R has the function `acf()` to compute the ACF: `acf.r`

Weak stationarity

- In order that $\hat{\gamma}(h)$ is a reasonable estimator for $\gamma(h)$, we need (at least) the following conditions:
  1. The ACF really exists
  2. $\bar{X}$ is a reasonable estimator for $\mathrm{E}[X_t]$ for **all** t
- For this reason, we focus on the following class of stochastic processes:

Definition 1 (Weak stationarity)

The process X is said to be **weakly stationary** if it satisfies the following conditions:

1. The ACF of X exists, i.e. $\mathrm{Cov}[X_t, X_s]$ depends only on $|t-s|$.
2. $\mathrm{E}[X_t]$ deos not depend on t.

Weak stationarity

- Since non-stationary time series have time-varying characteristics such as mean and variance, it is not straightforward to fit an adequate model
- In practice, even if the time series data X itself is non-stationary, its differences $X_t - X_{t-1}$ or logarithms $\log X_t$ are often stationary
- Ex. X is called a **random walk** if it is of the form $X_t = X_{t-1} + \epsilon_t$ with ϵ_t's being centered i.i.d. variables
  - X is non-stationary because $\text{Var}[X_t] = t$
  - Meanwhile, the difference process is $X_t - X_{t-1} = \epsilon_t$ and thus weakly stationary

Weak stationarity

- A standard procedure to fit a model to the time series data X:
  1. Check whether X is stationary or not
  2. If X is stationary, replace X by its difference and return to Step 1; otherwise, go to the next step
  3. Fit an adequate model to X
- How to check whether X is stationary or not?
  1. Correlogram (ACF plot): A non-stationary process typically has (very) slowly decaying ACF
  2. Unit root test: statistical hypothesis testing for (non-)stationarity
     - ⋆ Ex.: **Phillips-Perron test** for the null hypothesis that X has a unit root (a random walk like structure) against a stationary alternative (`PP.test()` in R)
     - ⋆ The `tseries` package has other unit root tests `adf.test()` and `kpss.test()`
- **R example**: `stationary.r`

Autoregressive (AR) model

- The **autoregressive model of order** p is the time series model of the form

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + \epsilon_t,$$

  where $a_1, \dots, a_p$ are some constants and ϵ_t's are centered i.i.d. variables
- This model is usually referred to as the **AR(p) model**
  - The past p states $X_{t-1}, \dots, X_{t-p}$ linearly affects the present state X_t
- A random walk is a special case of AR(1) with $a_1 = 1$
- **R example**: `ar.r`

Moving average (MA) model

- The **moving average model of order** q is the time series model of the form

$$X_t = b_1\epsilon_{t-1} + \cdots + b_q\epsilon_{t-q} + \epsilon_t,$$

  where $b_1, \dots, b_q$ are some constants and ϵ_t's are centered i.i.d. variables
- This model is usually referred to as the **MA(q) model**
- **R example**: `ma.r`

ARMA model

- The **autoregressive moving average model of order** (p, q) is the time series model of the form

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q} + \epsilon_t, \qquad (1)$$

  where $a_1, \dots, a_p, b_1, \dots, b_q$ are constants and ϵ_t's are centered i.i.d. variables
- This model is usually referred to as the **ARMA(p,q) model**
  - Hybrid of AR and MA models
  - The ARMA model is one of the most widely used models for time series data; it is simple but can describe various ACF shapes
- **R example**: `arma.r`

ARIMA model

- As mentioned, when time series data seem to be non-stationary, one typically takes the differences until they seem to be stationary
- After that, we fit a stationary ARMA model to the differentiated data
- Such a time series model is called the ARIMA model; X is the **ARIMA(p,d,q)** model if $\Delta^d X$ is the ARMA(p,q) model, where $\Delta^0 X_t := X_t$ and

$$\Delta^d X_t := \Delta^{d-1} X_t - \Delta^{d-1} X_{t-1}, \qquad d = 1, 2, \ldots$$

  - d is the parameter of how many times one takes differences

Fitting ARIMA models

- We can use the R function `arima()` to fit a stationary ARMA model to time series possibly after differentiation
- When we fit the ARMA model (1) to the time series X, we usually check the ACF of the residuals, which are recursively defined by

$$e_t := X_t - (a_1 X_{t-1} + \cdots + a_p X_{t-p} + b_1 e_{t-1} + \cdots + b_q e_{t-q})$$

  (the initial values $e_1, \dots, e_q$ are set to 0 or computed by backward prediction)
  - If (1) is an adequate model, the residuals e_t should behave like white noise
  - The R function `tsdiag()` is helpful to assess this behavior
- The AIC can be used to compare several fitted models (`AIC()` in R)
- **R example**: `fit-arma.r`