

# TIME SERIES ANALYSIS

## LECTURE 1

---

**[datascience@berkeley](mailto:datascience@berkeley)**

# Notion and Measure of Dependency

# General Description of a Stochastic Process

A complete description of a time series as a collection of  $k$  random variables requires a joint distribution function:

$$F(c_1, c_2, \dots, c_n) = P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_n} \leq c_n)$$

- Characterizing the joint distribution function in its most general form, such as the one written in the above form, is very challenging, if not impossible.
  - A single realization of a time series does not offer enough information to describe the entire underlying joint distribution function from which the realizations are generated.
  - Doing so requires many more assumptions about the joint distribution are imposed.
- In the context of time series analysis, one of the most important probabilistic features is **the dependency structure** embedded in the joint distribution.
- We will
  - discuss how dependency is defined, measured, and estimated,
  - illustrate these concepts using the foundational time series models, and
  - the mean and variance functions of a time series.

# Mean Function and Stationarity in the Mean

Before discussing the dependency structure of a time series, let's define the mean and variance functions of a stochastic process. The **mean function** is defined as

$$\mu_x(t) = E(x_t) = \int_{-\infty}^{+\infty} x_t f_t(x_t) dx_t$$

where  $E(.)$  denotes the expected value operator, and the expectation is taken over the *ensemble*, which consists of the entire population, of all possible time series that might have been produced by the underlying time series data generating process (DGP).

Note that the mean is a function of the index  $t$ . If the function is constant (i.e., it does not vary with  $t$ ), then the underlying stochastic process is said to be **stationary in the mean**.

# Variance Function and Stationarity in the Variance

The variance function of a time series model that is stationary in the mean is defined as

$$\sigma_x^2(t) = E(x_t - \mu)^2 = \int_{-\infty}^{+\infty} (x_t - \mu)^2 f_t(x_t) dx_t$$

where  $E(.)$  denotes the expected value operator, and the expectation is also taken over the *ensemble*. Note that the mean  $\mu$  is a constant not varying with time.

- Note that the variance function is also a function of the index .
- Yet, operation-wise, it is impossible to estimate different variance for different points in time, given only a single realization of the underlying stochastic process (i.e., a single time series).
- To make the concept operational, one has to impose more structure on the underlying stochastic process.
  - One popular assumption is **stationary in variance**, which, combining with stationarity in autocovariance to be introduced in the next slide, produce a large class of models called **stationary time series models**.

# Autocovariance and Autocorrelation Functions

Recall from DATASCI w203 that linear (or order) dependency is measured by covariance and correlation. In the context of time series analysis, we speak of the covariance and correlation between different random variables of the same series, and hence the term “autocovariance” and “autocorrelation”

The **autocovariance function (acvf)** is defined as

$$\gamma_x(s, t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \forall s, t$$

Two natural implications are (1)  $\gamma_x(s, t) = \gamma_x(t, s)$  (Exercise: verify it) and (2)  $\gamma_x(s, s) = \text{cov}(x_s, x_s) = E[(x_s - \mu_s)^2]$

- A correlation of a variable with itself at different times is known as *autocorrelation*.
- If a time series is second-order stationary (i.e. stationary in both mean and variance:  $\mu_t = \mu$  and  $\sigma_t^2 = \sigma^2$  for all  $t$ ), then an *autocovariance function* can be expressed as a function only of the time lag  $k$ :

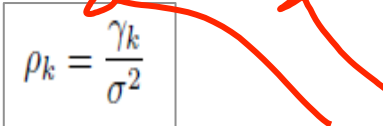
$$\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]$$

- Likewise, the **autocorrelation function (acf)** is defined as

When  $k = 0$ ,  $\rho_0 = 1$

$$\rho_k = \frac{\gamma_k}{\sigma^2}$$

## Autocovariance and Autocorrelation Functions (2)

$$\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]$$


A diagram illustrating the relationship between the autocovariance function  $\gamma_k$  and the autocorrelation function  $\rho_k$ . A box contains the equation  $\rho_k = \frac{\gamma_k}{\sigma^2}$ . Two red arrows originate from the box: one points to the  $\gamma_k$  term in the equation above, and the other points to the  $\mu$  term in the same equation.

- I want to emphasize the importance of the requirement that the series is second-order stationary before defining the autocovariance and autocorrelation functions.
- As you can see in these functions, the mean and variance are both constant.
- In other words, without a constant mean and variance, these functions are not well-defined!
- In time series, it is the dependence between the values of the series that is important to measure, so at a minimum, we want to estimate autocorrelation with precision.
- It would be difficult to measure the temporal dependence if the dependence structure change at every time point.
- I'd like you to keep this notion in mind because they will have important implications in the empirical work.

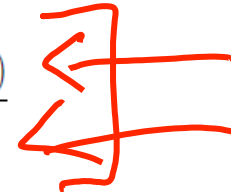
# Estimation of Autocovariance and Autocorrelation

Using the *moment principles*, the *acvf* and *acf* can be estimated from a time series by their sample equivalents. The sample *acvf* can be estimated using the following formula:

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \hat{x})(x_{t+k} - \hat{x})$$

Note that the sum is divided by  $T$  and not  $T-k$ .

The sample *acf* is defined as

$$\frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2}$$


In the next section, we will estimate and examine the autocovariance and autocorrelation of the real-world time series data and simulated time series.



# Partial Autocorrelation Functions

- The concept of **partial autocorrelation** in the context of time series is analogous to the **partial correlation** in the context of multiple linear regressions.
- It is a “conditional” correlation, conditional on other explanatory variables accounted for in a time series model.
- The partial autocorrelation of a process  $z_t$  at lag  $k$ ,  $\phi_{kk}$  is the (auto)correlation between the variable  $z_t$  and  $z_{t-k}$ , adjusting for the effects from variables  $z_{t-1}, z_{t-2}, \dots, z_{t-k+1}$ .
- In other words, it is the coefficient of  $z_{t-k}$  in a linear regression of  $z_t$  on  $z_{t-1}, z_{t-2}, \dots, z_{t-k+1}$ ; it is called **autoregression** because the variable  $z_t$  is regressed on its own lagged values.
- The difference between autocorrelation and partial autocorrelation is similar to the difference between the coefficients in simple linear regression (i.e. regression with 1 explanatory variables) and the coefficients in a multiple linear regression.
- Like the autocorrelation, the partial autocorrelation summarizes the dynamics of a process, and as you will see in the next lecture, partial autocorrelation is a power device for identifying the order of an AR(p) model

# Berkeley

SCHOOL OF  
INFORMATION