Time Series

Muchang Bahng

Spring 2025

Contents

1 Time Series Analysis

If we try sticking to linear algebra, we hope to model time series of the form

$$X_t = f(t) + w_t \tag{1}$$

so that we can decompose to a deterministic process followed by some white noise. There are several ways to approach this, including kernel smoothing, moving average smoothing, or cubic spline smoothing. However, this falls short when you look the residuals. They will follow some pattern that must be removed due to autocorrelation.

In linear regression, one of the fundamental assumptions was independence of errors. Ideally, we would also like independence of features, but this is usually not true (in fact, in extreme cases, multicollinearity can screw us up). The relaxation of these assumptions helps us transition from linear regression to time series analysis. Let's go over some basic things with new terms.

Definition 1.1 (Time Series)

A stochastic process

$$\{X_1, \dots, X_t, \dots\} \tag{2}$$

of random variables indexed by time t is a **time series**. The stochastic behavior of $\{X_t\}$ is determined by specifying the PDF/PMF

$$p(x_{t_1}, \dots, x_{t_m}) \tag{3}$$

for all finite collections of time indices

$$\{(t_1, \dots, t_m), m < \infty\} \tag{4}$$

i.e. all finite-dimensional distributions of X_t.

Definition 1.2 (White Noise)

White noise w_t is a random variable indexed by time t satisfying

1. $\mathbb{E}[w_t] = 0$
2. $\text{Var}[w_t] = \sigma^2$
3. $\text{Cov}[w_t, w_s] = 0$ for $s \neq t$. That is, they are uncorrelated but not necessarily independent.

Note that this third condition can be strengthened to independence or uncorrelated Gaussians, which automatically imply independence.

1.1 Properties of Processes

Now let's define some properties. We will start with the time series analogue of covariance and correlation.

Definition 1.3 (Autocovariance)

The **autocovariance** between two time steps t, s of process $\{X_t\}$ is defined

$$K_X(s,t) = \text{Cov}(X_t, X_s) \tag{5}$$

Definition 1.4 (Autocorrelation)

The **autocorrelation** is

$$\rho_X(s,t) = \frac{K_X(s,t)}{\sqrt{K_X(s,s)\, K_X(t,t)}} \tag{6}$$

Definition 1.5 (Cross Covariance)

Given two stochastic processes $\{X_t\}, \{Y_t\}$, the **cross covariance** is

$$K_{XY}(t,s) = \text{Cov}(X_t, Y_s) \tag{7}$$

and the **cross correlation** is

$$\rho_{XY}(t,s) = \frac{K_{XY}(t,s)}{K_X(t,s)\,K_Y(s,s)} \tag{8}$$

It is used to model the correlations between two related products with a certain time lag perhaps.

Definition 1.6 (Stationarity)

There are two types of stationarity.

1. A **weakly stationary** or **covariance stationary** process means that its mean and autocovariance are invariant to time shifts. That is, for all r,

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t+r}] = \mu \tag{9}$$
$$\text{Var}[X_t] = \text{Var}[X_{t+r}] = \sigma_X^2 \tag{10}$$
$$K_X(t,s) = K_X(t+r, s+r) \tag{11}$$
$$\tag{12}$$

2. A **strongly stationary** process means that any joint distribution function of a finite set of time steps is invariant to time shifts. That is, for any $r > 0$ and finite collection of time points $t_1, \dots, t_k$,

$$F(X_{t_1}, \dots, X_{t_k}) = F(X_{t_1+r}, \dots, X_{t_k+r}) \tag{13}$$

where F is the joint pdf and equality means almost everywhere equality.

Clearly, weakly stationary implies strongly stationary, and the difference is that weakly stationary has invariance in the first two moments while strongly stationary holds for all moments.

Theorem 1.1 ()

It immediately follows that for a stationary process X_t, the autocovariance function can be defined

$$K_X(s,t) = K_X(s-t, 0) = K_X(\tau) \tag{14}$$

for some difference between the time points, called the lag. From this, we can see that $\text{Var}[X_t] = K_X(0)$, so the autocorrelation can be defined as

$$\rho_X(\tau) = \frac{K_X(\tau)}{K_X(0)} \tag{15}$$

Stationary time series are very desirable, since if we do parameter estimation, we don't want to estimate parameters that are always changing. For example, in stationary processes, we know that the mean never changes, so we have a bunch of sample points to choose from, and if every wasn't stationary, then every X_t would have its own mean and we won't be able to estimate it. Similarly, we also know that for some fixed τ, the autocorrelation does not change, so we can estimate $K_X(\tau)$ with a bunch of fixed intervals of length τ. Therefore, if we want to test for stationary of a fixed time process, we want to conduct a test where we want to find whether the autocovariance is relatively invariant. This gives us a bit of intuition.

Theorem 1.2 ()

Note the following properties.

1. $K_X(\tau) = K_X(-\tau)$
2. By Cauchy-Schwartz, $K_X(0)^2 = \text{Var}[X_t]\,\text{Var}[X_{t+r}] \geq \text{Cov}(X_t, X_{t+r}) = K_X(r)^2$, so $|K_X(\tau)| \leq K_X(0)$.

Therefore, we would like to decompose a general time series to a stationary component and a nonstationary simple component, and do some statistics on the stationary one.

Definition 1.7 (Joint Stationarity)

Two processes X_t, Y_t, are said to be jointly stationary if both are individually stationary and also if the cross covariance function is also stationary. That is, for all r,

$$K_{XY}(t, s) = K_{XY}(t + r, s + r) \tag{16}$$

Definition 1.8 (Backshift Operator)

The backshift operator B acts on time series by

$$BX_t = X_{t-1} \tag{17}$$

It can be iterated to get $B^k X_t = X_{t-k}$ and can also be inverted to get a forward shift $B^{-k} X_t = X_{t+k}$. We can just think of this as (not necessarily linear?) operators between the function space of X-measurable functions.

1.1.1 Estimation

We should now try to estimate some parameters of a weakly stationary process.

Theorem 1.3 (Sampling Distribution of Mean)

We can already estimate the mean. We should get the mean of the mean and the variance of the mean.

1. The mean is trivial, since by linearity of expectation we can get

$$\hat{\mu} = \bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t \tag{18}$$

2. The variance is a bit more involved since there are covariance terms, so

$$\begin{aligned} \text{Var}[\bar{X}] &= \text{Var}\left(\frac{1}{T} \sum_{t=1}^{T} X_t \right) && (19) \\ &= \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \text{Cov}(X_t, X_s) && (20) \\ &= \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} K_X(|t - s|) && (21) \\ &= \frac{1}{T} K_X(0) + \frac{2}{T} \sum_{z=1}^{T-1} \left(1 - \frac{z}{T}\right) K_X(z) && (22) \end{aligned}$$

In the unrealistic situation where the X_t's are uncorrelated, we have $K_X(0) = \sigma^2$ and $K_X(z) = 0$ for all $z > 0$, leaving us with σ^2/T.

Theorem 1.4 (Sampling Distribution of Autocovariance)

To estimate the autocovariance of a weakly stationary process, we can define the sample autocovariance function to be

$$\hat{K}_X(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \bar{X})(X_t - \bar{X}) \tag{23}$$

Note that we divide by T rather than $T-h$ so that this covariance is positive semidefinite. Note that as h gets bigger, the number of terms in the sum decreases giving less accurate estimation. Similarly, the sample autocorrelation function is

$$\hat{\rho}(h) = \frac{\hat{K}_X(h)}{\hat{K}_X(0)} \tag{24}$$

The sample cross covariance and cross correlation are

$$\hat{K}_{XY}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \bar{X})(Y_t - \bar{Y}) \text{ and } \hat{\rho}_{XY}(h) = \frac{\hat{K}_{XY}(h)}{\sqrt{\hat{K}_X(0)\,\hat{K}_Y(0)}} \tag{25}$$

Note that even though we can just plug these formulas and get the sample estimators for any time series, these don't mean anything if they are not stationary.

1.1.2 Detecting White Noise

Ultimately, the main goal of time series analysis is to transform the data into a white noise process. We want to first identify trends and patterns in the process, remove them, and hopefully get white noise. To actually detect if we have white noise, one way to do this is to look at the estimated autocorrelation function across h. Note that for white noise, we have a spike at $h = 0$ to be 1 (since it is just the correlation of a variable with itself), and then it drops to 0 immediately (since by definition, w_s, w_t are uncorrelated). We would like to see this behavior within a certain confidence interval.

1.2 Autoregressive (AR) Processes

The assumptions are:

1. the data must be stationary (though it is not always stationary as it may contain a unit root)
2. the relationship between the variables and their lagged values must be linear (nonlinear gives large language models like LSTMs)
3. the error term should be white noise

Definition 1.9 (Autoregressive Process)

An **AR(p)** process encodes causality[a] into the white noise process. It is a stochastic process with mean 0 and of the form

$$X_t = w_t + \sum_{i=1}^{p}\phi_i X_{t-i} \tag{26}$$

where p is the hyperparameter of steps to look back, w_t is white noise with variance σ^2, and ϕ_i are constants $\neq 0$. Using the backshift operator B, we can write the AR(p) process as

$$\Phi(B)X_t = w_t \tag{27}$$

where

$$\Phi(B) = \left(1 - \sum_{i=1}^{p} \phi_i B^i\right) \tag{28}$$

In fact, we have already seen this process many times.

Example 1.1 (AR(p) Processes)

Consider the following.

1. AR(O) is simply a white noise process
$$X_t = w_t \tag{29}$$
2. AR(1) with $\theta = 1$ gives us the formula
$$X_t = X_{t-1} + w_t \tag{30}$$
which is a random walk. It is also a Markov process and a martingale.
3. AR(1) of the form
$$X_t = a + X_{t-1} + w_t \tag{31}$$
is a random walk with drift.
4. AR(2) can be of form
$$X_t = X_{t-1} - 0.2X_{t-2} + w_t \tag{32}$$
5. AR(3) can be of form
$$X_t = X_{t-1} - 0.2X_{t-2} + 0.13X_{t-3} + w_t \tag{33}$$

Occasionally, it may be hard to determine the difference between the difference of AR(p) processes.

Example 1.2 (AR(1) Processes)

Let's focus on the AR(1) process. Later on in linear processes, we see that the AR(1) process has a causal representation as a linear process.

$$X_t = \phi_1 X_{t-1} + w_t = \sum_{i=0}^{\infty} \phi_1^i w_{t-i} \tag{34}$$

This is stationary under certain conditions.

1. If $\phi < 1$, then the series is stationary.
2. If $\phi = 1$, this is a random walk which is not stationary.
3. If $\phi > 1$, then this process grows exponentially fast.

Now to determine weak stationarity, let's go back to the equation. Talk about unit root test.

Definition 1.10 (Augmented Dicky-Fuller Test)

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a time series is stationary or not. Here's a step-by-step explanation of how the ADF test is typically implemented:

1. **Model Specification**. The ADF test is based on an autoregressive model. The general form is:
$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \cdots + \delta_{p-1} \Delta Y_{t-p+1} + \varepsilon_t \tag{35}$$
Where:

[a] on how a random variable Y is *caused* by another RV X.

- ΔY_t is the first difference of the series at time t
- α is the constant term
- βt is the time trend term
- γ and δ are coefficients
- ε_t is the error term
- p is the lag order

2. **Determine the lag order (p):**
   - This can be done using information criteria like AIC or BIC
   - Or by starting with a maximum lag and testing down
3. **Estimate the model:**

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta X_{t-i} + \varepsilon_t \tag{36}$$

Where $\Delta X_t = X_t - X_{t-1}$ is the first difference of the series. To apply OLS, we rewrite this in matrix form:

$$Y = X\beta + \varepsilon \tag{37}$$

Where:
- Y is an $(n-p) \times 1$ vector of ΔX_t values
- X is an $(n-p) \times (p+2)$ matrix of explanatory variables
- β is a $(p+2) \times 1$ vector of coefficients $(\alpha, \beta, \gamma, \delta_1, \ldots, \delta_{p-1})$
- ε is an $(n-p) \times 1$ vector of error terms
- n is the number of observations
- p is the lag order

The OLS estimator for β is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{38}$$

This estimator minimizes the sum of squared residuals:

$$\sum_{t=p+1}^{n} \varepsilon_t^2 = (Y - X\beta)'(Y - X\beta) \tag{39}$$

   - Use Ordinary Least Squares (OLS) to estimate the coefficients of the model
4. **Calculate the test statistic:**
   - The test statistic is the t-statistic for γ:

$$t = \frac{\hat{\gamma} - 0}{SE(\hat{\gamma})} \tag{40}$$

   Where $\hat{\gamma}$ is the estimated coefficient and $SE(\hat{\gamma})$ is its standard error
5. **Determine the critical values:**
   - These depend on the sample size and the model specification (whether it includes a constant and/or trend)
   - They're typically obtained from statistical tables or through simulation
6. **Compare the test statistic to the critical values:**
   - If the test statistic is less than (more negative than) the critical value, reject the null hypothesis
   - The null hypothesis is that the series has a unit root (is non-stationary)
7. **Interpret the results:**
   - If we reject the null, we conclude the series is stationary
   - If we fail to reject the null, we cannot conclude the series is stationary

Once this is settled, our job is now to estimate the parameters. We can use MLE.

1.3 Moving Average (MA) Processes

The key assumptions are:

1. The random shocks are white noise, mutually independent and coming from the same distribution with mean 0 and constant variance.

Definition 1.11 (Moving Average Process)

The **MA(q)** process is a smoother type of noise than the white noise process. It is expressed by the formula

$$X_t = \sum_{j=1}^{q} \phi_j w_{t-j} + w_t \tag{41}$$

for $\phi_j \in \mathbb{R}$. Compared to the AR formula, the MA formula averages over the noise terms w_t. It focuses on the ripples of the process; if there is a shock to the process w_{t-1}, then that shock is still felt at time t by the term $\phi_1 t_{t-1}$.
Alternatively, the MA model can be written as an overall average of both the past and future white noise.

$$X_t = \sum_{j=-q/2}^{q/2} \phi_j w_{t+j} \tag{42}$$

Theorem 1.5 ()

A nice property of MA(q) is that autocovariance vanishes beyond a certain point. More specifically, it decays *linearly* and vanishes after q steps behind.

1.4 Linear Processes

Many time series fall under the category of linear processes.

Definition 1.12 (Linear Processes)

A **linear process** is defined as

$$X_t = \mu + \sum_{j=-\infty}^{+\infty} \theta_j w_{t-j} \tag{43}$$

which means that every X_t is a linear combination of the terms in the white noise process with some mean μ added on. To ensure that this series doesn't blow up, we add the constraint that

$$\sum_j \theta_j^2 < \infty \tag{44}$$

However, since we are more interested in causal inference, to use the past to predict the future, we use the form

$$X_t = \mu + \sum_{j=0}^{\infty} \theta_j w_{t-j} \tag{45}$$

In fact, some AR processes are linear processes.

Example 1.3 (AR(1) as a Linear Process)

Note that AR(1) has a causal representation as a linear process. We can use the formula $X_t = \theta X_{t-1} + w_t$ and recursively define

$$X_t = \theta(\theta X_{t-2} + w_{t-1}) + w_t = \ldots = \sum_{j=0}^{\infty} \theta^j w_{t-j} \tag{46}$$

Going back to analysis, infinite series are just limits.

$$\lim_{N \to \infty} \sum_{j=0}^{N} \theta^j w_{t-j} \tag{47}$$

So this sum may not converge. Letting $S_N(\theta)$ be defined as above, we can compute that

$$\mathbb{E}[S_N(\theta)] = 0 \text{ and } \text{Var}[S_N] = \sigma^2 \sum_{j=0}^{N} \theta^{2j} = \sigma^2 \left(\frac{1 - \theta^{2N+2}}{1 - \theta^2} \right) \tag{48}$$

Thus, if $|\theta| < 1$, then $\text{Var}[S_N(\theta)] \to \sigma^2/(1-\theta^2)$, and if w_t is Gaussian noise, then

$$S_N(\theta) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2/(1-\theta^2)\big) \tag{49}$$

If $|\theta| = 1$, the series does not converge and is not stationary, and if $|\theta| > 1$, then the random talk will grow exponentially fast.

1.5 ARMA

We can combine both the AR and MA processes to make a more sophisticated model.

Definition 1.13 (ARMA)

The time series X_t is an ARMA(p, q) process if X_t has 0-mean and if we can write it as

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j} \tag{50}$$

where w_t is white noise with variance σ^2 and $\boldsymbol{\phi}, \boldsymbol{\theta}$ do not have any zero elements. Using the backshift operator, we can write it as

$$\Phi(B) X_t = \Theta(B) w_t \tag{51}$$

where

$$\Phi(B) = \left(1 + \sum_{i=1}^{p} \phi_i B^i \right) \text{ and } \Theta(B) = \left(1 + \sum_{j=1}^{q} \theta_j B^j \right) \tag{52}$$

1.6 ARIMA

1.7 Other

Theorem 1.6 (Wold Representation Theorem)

Any 0-mean covariance stationary time series $\{X_t\}$ can be decomposed into two time series

$$X_t = V_t + S_t \tag{53}$$

where

1. V_t is a linear combination of past variables of V_t with constant coefficients.
2. $S_t = \sum_{i=0}^{\infty} \psi_i \eta_{t-i}$ is an infinite moving average process of error terms, where
   (a) $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.
   (b) $\{\eta_t\}$ is linearly unpredictable white noise, i.e.

$$\mathbb{E}[\eta_t] = 0 \tag{54}$$
$$\mathbb{E}[\eta_t^2] = \sigma^2 \tag{55}$$
$$\mathbb{E}[\eta_t \eta_s] = 0 \text{ for } s \neq t \tag{56}$$

and η_t is uncorrelated with $\{V_t\}$.

$$\mathbb{E}[\eta_t V_s] = 0 \text{ for all } t, s \tag{57}$$

Example 1.4 (Construction on Dataset)

Say that we have data $\{X_t\}_{t=1}^T$ that we want to model and we have evidence that it is covariance stationary. We can do the following.

1. Initialize a parameter p, the number of parameters in the linearly deterministic term of the Wold decomposition of $\{X_t\}$.
2. By assumption we would like to estimate the linear projection of X_t on $(X_{t-1}, X_{t-2}, \dots, X_{t-p})$.

Therefore, let us index the n subseries of length $p+1$ by y and we can write the OLS equation

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & y_0 & y_{-1} & \cdots & y_{-(p-1)} \\ 1 & y_1 & y_0 & \cdots & y_{-(p-2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{bmatrix} \tag{58}$$

and we apply OLS to the problem $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta}$ to give

$$\hat{\mathbf{y}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}\mathbf{y} \tag{59}$$
$$= \hat{P}(Y_t \mid Y_{t-1}, \dots, Y_{t-p}) \tag{60}$$
$$= \hat{\mathbf{y}}^{(p)} \tag{61}$$

We can compute the projection residuals

$$\boldsymbol{\epsilon}^{(p)} = \mathbf{y} - \hat{\mathbf{y}}^{(p)} \tag{62}$$

and apply time series analysis to the sequence $\boldsymbol{\epsilon}^{(p)} = \{\epsilon_t^{(p)}\}$ to specify a moving average model.

$$\epsilon_t^{(p)} = \sum_{i=0}^{\infty} \psi_i \eta_{t-i} \tag{63}$$

yielding $\{\hat{\psi}_j\}$ and $\{\hat{\eta}_t\}$ estimates of parameters and innovations. We then check these estimates and see if they are consistent with the model assumptions. If not, we can add additional legs or modify p.

Theoretically, as we increase p, the projection of Y_t over the past pth history should approach the true linear projection Y_t over the whole history.

$$\lim_{p\to\infty} \hat{\mathbf{p}}^{(p)} = \hat{\mathbf{y}} \tag{64}$$

But if p is too large compared to n, you run out of freedom to estimate your models. You generally want to have more data than the number of parameters.

Definition 1.14 (Lag Operator)

The **lag operator** L^k simply maps

$$L^k(X_t) = X_{t-k} \tag{65}$$

Inverses also exist, so $L^{-k}(X_t) = X_{t+k}$.

Therefore, the Wold representation for a covariance stationary time series $\{X_t\}$ can be expressed as

$$X_t = \sum_{i=0}^{\infty} \psi_i \eta_{t-i} + V_t \tag{66}$$

$$= \sum_{i=0}^{\infty} \psi_i L^i(\eta_t) + V_t \tag{67}$$

$$= \psi(L)\eta_t + V_t \tag{68}$$

where $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$.

1.8 Components of time series

1.9 Stationarity and tests for stationarity (including ADF test)

1.10 Autoregressive (AR) models

1.11 Moving average (MA) models

1.12 ARIMA models

1.13 Forecasting techniques

References