ARIMA Notes

Team Fourier

November 2020

1 Time Series Data

Time series data comes in different forms, from prices to temperatures to people as a function of time. Recall one type of time series data relevant to the concept of system control, position over time. In the system identification problem, we have a state transition equation:

$$X_{t+1} = AX_t + B\mu_t \tag{1}$$

What is crucial to understand is that the only observable output from the system are the series of X_t's as well as the inputs placed in to the system, the μ's. In the system identification problem, we use a series of inputs and outputs to estimate and refine these estimations for the model parameters A and B.

We find that this modeling is not always fully applicable to all time series data. For instance, in modeling and forecasting stock prices over time, there is not a well-defined user input, μ_t. Also, there may be additional previous system states contributing to the current observation instead of only the immediately previous observation. In these instances, a more appropriate model is sought after to express the behavior of the time series data.

In these notes, we will explain how ARIMA and moving average models are more suitable to specific instances of time series data than the basic system identification model, and how the parameters can be estimated through a combination of least squares, used in the traditional system identification process, and iterative maximum likelihood expectation, the pattern used by k-means approach to clustering taught in earlier weeks of this course.

2 Trends

One commonly encountered time series component is a trend. Trends encompass datasets with non-repetitive generally increasing or decreasing values. An example of a trend is a child's weight as a function of time.

2.1 Trend Models

We can model these data by adding a deterministic trend function to white noise:

$$X_t = m_t + Z_t \tag{2}$$

where X_t is the observation, m_t is the deterministic trend function and Z_t is white noise. Random variables are white noise if they have zero mean and finite variance. This note will cover several methods for fitting 2 to data with trends.

2.1.1 Linear Regression

A simple technique would be to perform linear regression on the raw data or a feature augmented data set. If the underlying m_t trend is of a polynomial nature, using linear regression can be used to quite accurately fit the time series data. However, if this observation doesn't hold true, using the following moving average techniques can provide a better estimation for each m_t using neighboring values.

2.1.2 Simple Moving Average (SMA)

SMA averages the previous data points with the current sample to reduce noise. For a $2p$-sample moving average, we calculate:

$$\hat{m}_t = \frac{1}{2p+1} \sum_{i=-p}^{p} m_{t+i}{}^{1} \tag{3}$$

where $\hat{m}_t$ is the *simple moving average* of X_t. If m_t is linear over $[t-p, t+p]$, then:

$$\hat{m}_t = m_t + \frac{1}{2p+1} \sum_{i=-p}^{p} Z_{t+i} \approx m_t \tag{4}$$

We can generalize SMA as a weighted sum of the samples:

$$\hat{m}_t = \sum_{i=-p}^{p} \theta_j X_{t+i} \tag{5}$$

where θ_j are the weights. SMA is a special case of this weighted sum with constant weights of $\frac{1}{2p+1}$. The choice of p affects the bias-variance tradeoff, a property of models that determines sensitivity to noise. Read bias-variance-tradeoff for more information on this property. Selecting p often requires a hyperparameter search.

2.1.3 Exponential Moving Average (EMA)

EMA follows a similar approach to SMA, but instead of constant weights, EMA heavily emphasizes samples closer in time and depends only on previous observations:

$$\hat{m}_t = \frac{1-\phi}{\phi} \sum_{i=1}^{\infty} \phi^i X_{t-i} \tag{6}$$

Here the weights ϕ remain constant and samples further back in time are exponentially worth less. We multiply the sum by $\frac{1-\phi}{\phi}$ to ensure the weights sum to 1 since $\sum_{i=1}^{\infty} \phi^i = \frac{\phi}{1-\phi}$.

2.2 Detrending Data

To find patterns in trend models, it is often useful to study the stochastic portions of the models. This way we can exploit structure, if any, and create distributions to sample noise for forecasting. One natural approach to obtain the detrended data is differencing:

$$\nabla X_t = X_t - X_{t-1} \qquad \forall t \in 2, \ldots, n, \tag{7}$$

where ∇X_t is called a residual. Let $m_t = at + b$, a linear trend. Then

$$\begin{aligned} \nabla X_t &= a(t+1) + b + Z_{t+1} - (at + b + Z_t) \qquad \forall t \in 2, \ldots, n \\ &= a + Z_{t+1} - Z_t. \end{aligned} \tag{8}$$

[1] Note in financial applications, the averages are taken over the previous $2n+1$ samples rather than n samples on either side of m_t

If ∇X_t is white noise, we can simply predict X_{t+1} by forecasting ∇X_{t+1} as the sample mean $\overline{\nabla X} = \sum_{i=2}^{n} Y_i/(n-1)$. Then, rearranging (7) we get:

$$\begin{aligned} X_{t+1} &= \nabla X_{t+1} + X_t \qquad \forall t \in 2, \ldots, n \\ &= \overline{\nabla X} + X_t. \end{aligned} \tag{9}$$

Similarly, if there are trends after differencing, we can difference again:

$$\begin{aligned} \nabla^2 X_t &= \nabla(\nabla X_t) \qquad \forall t \in 3, \ldots, n \\ &= \nabla X_t - \nabla X_{t-1} \\ &= X_t - 2X_{t-1} + X_{t-2}. \end{aligned} \tag{10}$$

Second order differencing removes quadratic trends. A nice way to check understanding is to prove why ∇^n removes n^{th} order trends.

3 Stationarity

Building off of the connection between ARIMA modeling and system identification, the underlying system that directs the observations has to satisfy some key conditions. In a kinematics model, these key conditions revolve around the laws of physics. In time series modeling, the parallel assumption is stationarity.

Stationary processes form the foundation for systematic time series analysis, especially for **ARMA** (autoregressive moving average) models. Most existing models require stationarity to function. Consequently, there are many methods to transform non-stationary data into stationary data. There are two standard definitions of stationarity.

3.1 Strong or Strict Stationarity

*A stochastic process $\{X_t\}$ is **strongly stationary** if the joint distribution of any set of samples $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is the same as the joint distribution of $(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_k+\tau}) \forall \tau, t_1, \ldots, t_n \in \mathbb{R}$ and $\forall n \in \mathbb{N}$.*

The intuition here is that a stationary time series can be fully modeled by looking at a sliding window across the data. There is no underlying structure to the data that cannot be accounted for without looking at the data as a whole.

This allows us to learn summary statistics of the process, such as means and variances. With this knowledge, we can then predict future observations. A joint distribution is the distribution of many variables. Since we will not use strict stationarity, it is fine to not understand joint distributions now.

3.2 Weak or Wide-Sense Stationarity

*A stochastic process $\{X_t\}$ is **weakly stationary** if*

1. $\mathbb{E}[X_t] = \mathbb{E}[X_{t+\tau}] \forall \tau \in \mathbb{R}$
2. $K_{XX}(t,s) = K_{XX}(t-s, 0) \forall \in \mathbb{R}$

where $K_{XX}(t,s)$ is the autocovariance function of X_t evaluated at time t and s.

The first condition of weak stationarity is that the mean of the data remains constant throughout the time series. This goes hand in hand with the idea of trendlessness, where there is no underlying direction to the data.

The second condition conveys the idea that the dependence of an observation on a past observation p steps in the past is the same regardless of where we look in the data. This allows us to formulate a model in which the current observation is some combination of past observations, repeatable at any time step.

Strong and weak stationarities do not imply each other. The rest of this note will refer to *weak stationarity* when discussing *stationarity*.

3.2.1 Autocovariance Function *(acvf)*

The *acvf* is defined as:

$$\begin{aligned} K_{XX}(t,s) &= \mathrm{Cov}[X_t, X_s] \\ &= \mathbb{E}[(X_t-\mu_t)(X_s-\mu_s)] \\ &= \mathbb{E}[X_t, X_s] - \mu_1\mu_2 \end{aligned}$$

From condition 2, we realize that the covariance of two random variables only only depends on τ, the time lag between them. Therefore, we can further simplify the notation of the *acvf*:

$$K_{XX}(\tau) = K_{XX}(t, t+\tau) \tag{11}$$

We know the covariances of uncorrelated samples:

$$\mathrm{Cov}(X_t, X_s) = \begin{cases} \sigma_z^2 & \text{if } t = s \\ 0 & \text{otherwise.} \end{cases}$$

Likewise, stationary white noise has *acvf*:

$$K_{XX}(\tau) = \begin{cases} \sigma_z^2 & \text{if } \tau = 0 \\ 0 & \text{otherwise.} \end{cases}$$

3.2.2 Autocorrelation Function *(acf)*

The *acf* is defined as:

$$\rho(t,s) = \frac{K_{XX}(t,s)}{\sqrt{\mathrm{Var}(X_t), \mathrm{Var}(X_s)}}$$

which can then be simplified to:

$$\rho(\tau) = \frac{K_{XX}(\tau)}{K_{XX}(0)}. \tag{12}$$

The *acf* of a white noise process is:

$$\rho(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ 0 & \text{otherwise.} \end{cases}$$

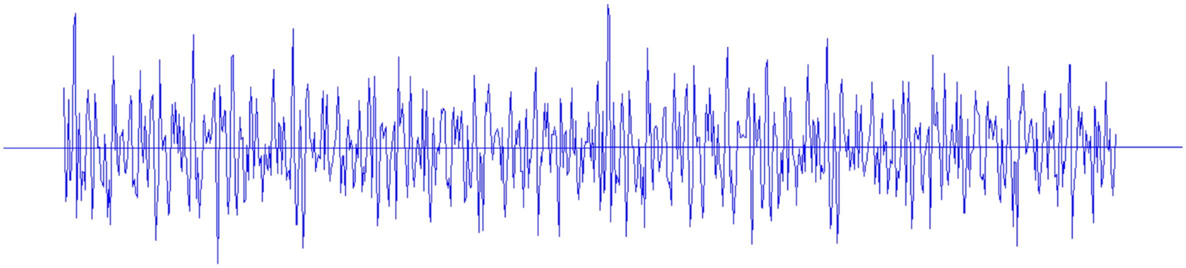

Figure 1: Example of stationary process

Notice the mean and variance of the data are consistent across time.

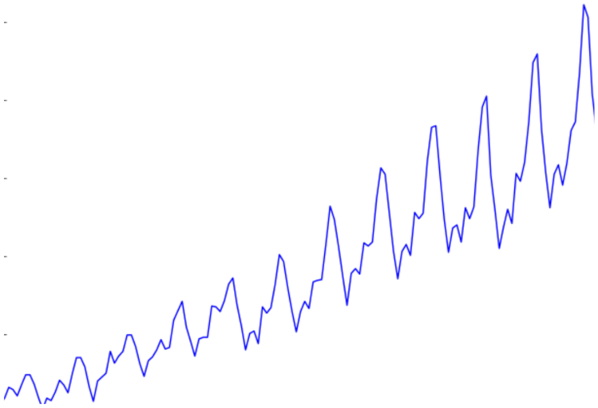

Figure 2: Example of nonstationary process

Notice how both the mean and variance of the sequence increase over time.

4 Moving Average (MA) Models [2]

Moving average models allow for forecasting based on past stochastic terms. They are one of the most widely used models to study time series.

Let $Z_t, Z_{t-1}, \ldots, Z_{t-q}$ denote a white noise sequence and μ denote the mean of the sequence. The ***moving average model*** *of order q,* ***MA(q)****, is defined as:*

$$\begin{aligned} X_t &= \mu + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q} \\ &= \mu + \sum_{j=0}^{q} \theta_j Z_{t-j} \qquad \textit{where } \theta_0 = 1 \\ &= \sum_{j=0}^{q} \theta_j Z_{t-j} \end{aligned} \tag{13}$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model.
We get (13) from the second line by assuming a zero mean process.

[2] Moving average models are different from the moving average (SMA).

4.1 Weak Stationarity of an MA(q) Model

4.1.1 *ACVF* and *ACF*

The *acvf* is derived as follows:

$$\begin{aligned} K_{XX}(\tau) &= \mathrm{cov}(\sum_{j=0}^{q} \theta_j Z_{t-j}, \sum_{k=0}^{q} \theta_j Z_{t+\tau-k}) \\ &= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \mathrm{cov}(Z_{t-j}, Z_{t+\tau-k}) \end{aligned}$$

$K_{XX}(h)$ is only non-zero when $t - j = t + \tau - k$, or when $k = j + \tau$, since Z_t are uncorrelated. We know k is bounded by 0 and q which further means j is bounded by 0 and $q - h$. We can rewrite the *acvf* as:

$$K_{XX}(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & \text{if } \ h \in 0, 1, \ldots, q \\ 0 & \text{if } h > q \end{cases} \tag{14}$$

With some pattern matching, we get the *acf*:

$$K_{XX}(h) = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2} & \text{if } \ h \in 0, 1, \ldots, q \\ 0 & \text{if } h > q \end{cases} \tag{15}$$

Notice that $K_{XX}(\tau)$ is independent of t. From this we deduce that:
If $\{X_t\}$ follows an MA(q) model, then $\{X_t\}$ is ***weakly stationary****.*

4.1.2 Backshift Operator

Let B denote the backshift operator. It is defined as:

$$BX_t = X_{t-1}, B^2 X_t = X_{t-2}, \ldots, B^n X_t = X_{t-n}$$

and

$$B^{-n} X_t = X_{t+n}, \ldots, B^{-1} X_t = X_{t+1}.$$

4.1.3 Moving Average Operator

For parameters $\theta_1, \theta_2, \ldots, \theta_q$ with $\theta_q \neq 0$, define the **moving average operator** of order q as:

$$\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q. \tag{16}$$

This allows us to succinctly write an MA(q) model from (13) as:

$$X_t = \theta(B) Z_t. \tag{17}$$

4.2 Invertibility

Consider an MA(1) model with acf:

$$\rho(1) = \frac{\theta}{1 + \theta^2}$$

For any value of θ, $1/\theta$ gives the same autocorrelation. Therefore, there is no unique solution for the parameters of an MA(1) model. To enforce a unique solution, we impose a theoretical restriction to only consider MA(1) models with $|\theta| < 1$. Essentially, θ must lie within the unit circle. This condition

is called *invertibility*. More broadly,
An MA(q) model $X_t = \theta(B)Z_t$ is invertible iff it can be written as

$$\begin{aligned} Z_t &= \pi(B)X_t \\ &= \sum_{j=0}^{\infty} \pi_j X_{t-j} \end{aligned} \tag{18}$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and $\pi_0 = 1$.

5 Autoregressive (AR) Models

Autoregressive models predict future observations directly using past observations and a stochastic term. These models form another set of widely used techniques.

*Let $Z_t, Z_{t-1}, \ldots, Z_{t-q}$ denote a white noise sequence and c a constant. The **autoregressive model** of order p, **AR(p)**, is defined as:*

$$\begin{aligned} X_t &= c + Z_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} \\ &= c + Z_t + \sum_{j=0}^{p} \phi_j X_{t-j} \qquad \textit{where } \phi_p \neq 0 \\ &= Z_t + \sum_{j=0}^{p} \phi_j X_{t-j} \qquad \textit{where } \phi_p \neq 0 \end{aligned} \tag{19}$$

*where $\phi_1, \ldots, \phi_p$ are parameters of the model and (19) is a zero mean **AR(p)** process.*

5.1 Autoregressive Operator

For parameters $\phi_1, \ldots, \phi_p$ with $\phi_p \neq 0$, define the **autoregressive operator** of order p as:

$$\phi(B) = 1 - \phi_1 B - \ldots \phi_p B^p. \tag{20}$$

This allows us to write an AR(p) model from (19) as:

$$\phi(B)X_t = Z_t \tag{21}$$

which has the polynomial form

$$\phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p. \tag{22}$$

5.2 Causality

Let us take the case of an AR(1) model:

$$X_t = Z_t + \phi X_{t-1}$$

We can rewrite (23) as:

$$\begin{aligned} X_t &= \frac{Z_t}{\phi(B)} \\ &= \frac{1}{1 - \phi B} Z_t \end{aligned} \tag{23}$$

Notice we have the sum of a geometric series for $|\phi| < 1$:

$$\begin{aligned} X_t &= (1 + \phi B + \phi^2 B^2 + \ldots) Z_t \\ &= \sum_{j=0}^{\infty} \phi^j Z_{t-j} \end{aligned} \tag{24}$$

Now consider $|\phi| > 1$. With some clever algebraic manipulation we can get:

$$\begin{aligned} X_t &= (-\frac{B^{-1}}{\phi} - \frac{B^{-2}}{\phi^2} - \ldots) Z_t \\ &= -\frac{Z_{t+1}}{\phi} - \frac{Z_{t+2}}{\phi^2} - \ldots \\ &= -\sum_{j=1}^{\infty} \frac{Z_{t+j}}{\phi^j}. \end{aligned} \tag{25}$$

We see for this case that X_t depends on future stochastic values, which usually does not work for forecasting. For an AR(1) model, a unique stationary solution is causal when $|\phi| < 1$ because it only depends on present and past values of $\{Z_t\}$. An AR(1) model with $|\phi| > 1$ is non-causal because it depends on future values. An exercise to ensure understanding is to show why no stationary solution exists for $|\phi| = 1$.

We define an AR(p) model as ***causal*** *if $\phi(z) \neq 0$ for $|z| > 1$.*

Essentially, z must lie outside the complex unit circle. Equivalently, ϕ must lie within the complex unit circle, similar to θ for an MA model. since $\phi(z)$ can be interpreted as a polynomial (22), z plays the role as the roots of the polynomial. If the roots of an AR(p) process are not 1, the process is stationary. Similar to invertibility for MA(q) models,

an AR(p) model is causal iff it can be written as

$$\begin{aligned} X_t &= \psi(B) Z_t \\ &= \sum_{j=0}^{\infty} \psi_j Z_{t-j} \end{aligned} \tag{26}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $\psi_0 = 1$.

5.3 Yule-Walker Equations

Given a zero mean and causal AR(p) process $\{X_t\}$ defined as (19), we can try to solve for its parameters. Let us begin by multiplying (19) by X_{t-1} and taking the expectation to obtain:

$$\begin{aligned} \mathbb{E}[X_t X_{t-1}] &= \sum_{j=1}^{p} \phi_j \mathbb{E}[X_{t-j} X_{t-1}] + \mathbb{E}[Z_t X_{t-1}] \\ &= \sum_{j=1}^{p} \phi_j \mathbb{E}[X_{t-j} X_{t-1}] \end{aligned} \tag{27}$$

where $\mathbb{E}[Z_t X_{t-1}]$ is 0 because Z_t is uncorrelated with previous values of the process. We have just derived the Yule-Walker equation for *lag 1.* We repeat this derivation for *lag 2, ..., lag p.* As an exercise, derive the Yule-Walker equation for *lag p*. The following is the generalized equation:

$$\begin{aligned}\mathbb{E}[X_t X_{t-k}] &= \sum_{j=0}^{p} \phi_j \mathbb{E}[X_{t-j} X_{t-k}] + \mathbb{E}[Z_t X_{t-k}] \\ &= \sum_{j=0}^{p} \phi_j \mathbb{E}[X_{t-j} X_{t-k}] \qquad (28)\end{aligned}$$

for $k \in [1, \ldots, p]$. Note $r_{-l} = r_l$. Using r_l for $\mathbb{E}[X_{t-l}X_{t-k}]$, (28) simplifies to:

$$r_l = \sum_{j=1}^{p} \phi_j r_{j-l} \qquad (29)$$

We can then reconfigure the p equations into a $p \times p$ matrix multiplication:

$$\underbrace{\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{p-1} \\ r_p \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} r_0 & r_1 & \ldots & r_{p-2} & r_{p-1} \\ r_1 & r_2 & \ldots & r_{p-3} & r_{p-2} \\ & & \vdots & & \\ r_{p-2} & r_{p-3} & \ldots & r_2 & r_1 \\ r_{p-1} & r_{p-2} & \ldots & r_1 & r_0 \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p-1} \\ \phi_p \end{bmatrix}}_{\Phi} \qquad (30)$$

We can simplify further using compact matrix notation:

$$\mathbf{R}\Phi = \mathbf{r}. \qquad (31)$$

We know **R** is full rank and symmetric, so we can invert **R** to solve for Φ:

$$\hat{\Phi} = \mathbf{R}^{-1}\mathbf{r}. \qquad (32)$$

R is also known as the **covariance matrix** of X.

6 ARMA Models

Let $Z_t, Z_{t-1}, \ldots, Z_{t-q}$ denote a white noise sequence. The ***AutoRegressive Moving Average model*** *of order (p, q),* ***ARMA(p, q)*** *is defined as:*

$$\begin{aligned}\phi(B)X_t &= \theta(B)Z_t \\ X_t - \phi_1 X_{t-1} - \ldots - \phi_p(X_{t-p}) &= Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \qquad (33)\end{aligned}$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA operators from (20) and (16) with parameters $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ where $\phi_p, \theta_q \neq 0$.

ARMA(p, q) processes allow us to model more complex data sets while retaining the properties of the individual processes. We can rearrange (33) to isolate X_t:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p(X_{t-p}) + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}. \qquad (34)$$

These are the parameters to obtain models discussed earlier:

1. **White Noise**: ARMA(0, 0) $\implies \phi(z) = 1$ and $\theta(z) = 1$

2. **Moving Average**: ARMA(0, q) $\implies \phi(z) = 1$ and $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$

3. **Autoregressive**: ARMA(p, 0) $\implies \phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and $\theta(z) = 1$

6.1 Invertibility

Let us take a closer look at (24). We notice that it resembles an ARMA(0, ∞) model with $\theta_j = \phi^j$:

$$\sum_{j=0}^{\infty} \phi^j Z_{t-j} = \sum_{j=0}^{\infty} \theta_j Z_{t-j} \tag{35}$$

when $|\phi| < 1$. Therefore, we can invert an ARMA(1, 0) process to an ARMA(0, ∞) process. More generally,

an ARMA(p, 0) process, $\phi(B)X_t = Z_t$, can be inverted to an ARMA(0, ∞) process, $X_t = \theta(B)Z_t$ if all λ_i in

$$\begin{aligned} 1 - \phi_1 B - \ldots - \phi_p B^p &= (1 - \lambda_1 B)(1 - \lambda_2 B)\ldots(1 - \lambda_p B) \\ &= \theta_1(B)\theta_2(B)\ldots\theta_p(B) \end{aligned} \tag{36}$$

have magnitudes less than 1 *or lie within the complex unit circle.*

For an ARMA(2, 0) process, we can then solve:

$$\begin{aligned} \theta_1(B)\theta_2(B)\ldots\theta_p(B) &= (1 + \lambda_1 B + \lambda_1 B^2 + \ldots)(1 + \lambda_2 B + \lambda_2 B^2 + \ldots) \\ &= 1 + (\lambda_1 + \lambda_2)B + (\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2)B^2 + \ldots \\ &= \sum_{k=0}^{\infty}(\sum_{j=0}^{k} \lambda_1^j \lambda_2^{k-j})B^k \qquad (37) \\ &= \psi(B) \\ & \qquad (38) \end{aligned}$$

where $\psi_k = \sum_{j=0}^{k} \lambda_1^j \lambda_2^{k-j}$. To derive a general method for ARMA(p, q) to ARMA(0, q) inversion, we look at (35). Manipulating this, we get the unique solution form of an ARMA(p, q) process:

$$\begin{aligned} X_t &= \frac{\theta(B)Z_t}{\phi(B)} \\ &= \psi(B)Z_t \end{aligned} \tag{39}$$

where we have a new ARMA(0, q) process with coefficients ψ_k. To calculate the values of ψ_k, we rearrange (39) to obtain $\theta(B) = \phi(B)\psi(B)$. We can then match the corresponding coefficients for the same powers of B. To test for understanding, calculate the inversion for an ARMA(1, 1) process. Invertible processes are useful because we can invert an MA process to an AR process to find Z_t, which are non-observable, using the past values of X_t.

6.2 Causality

A process has a stationary solution of the form ARMA(p, q) iff (22), $\phi(z)$, is not equal to 0 *for all* $|z| = 1$

Similar to an ARMA(p, 0) model, *ARMA(p, q) models describe a causal process if (26) and its constraints hold true.*

6.3 Redundant Parameters

Cancelling out common factors is crucial because we might otherwise fit an ARMA(1, 1) model to white noise, or as in the following example, fit a higher order model to a simpler process.

6.3.1 Example

Consider a process:

$$\begin{aligned} X_t - 0.12X_{t-1} - 0.36X_{t-2} &= Z_t - 1.6Z_{t-1} + 0.6Z_{t-2} \\ (1 - 0.1B - 0.42B^2)X_t &= (1 + 1.1.5B + 0.54B^2)Z_t. \end{aligned} \tag{40}$$

Naturally we think this is an ARMA(2, 2) process. However, we need to check for redundant parameters in our models. Therefore, we factor (40) the operator polynomials:

$$\begin{aligned} (1 - 0.7z)(1 + 0.6z) &= (1 + 0.9z)(1 + 0.6z) \\ (1 - 0.7z) &= (1 + 0.9z) \end{aligned} \tag{41}$$

where we divided by $(1 + 0.6z)$, the common factor. It turns out our process is actually an ARMA(1, 1) model.

Our model is causal because:

$$\phi(z) = 1 - 0.7z = 0 \qquad \text{when } z = 10/7, \tag{42}$$

which has magnitude larger than 1 or $\lambda_1 < 1$. This model is also invertible because:

$$\theta(z) = 1 + 0.9z = 0 \qquad \text{when } z = -10/9, \tag{43}$$

which also has a magnitude larger than 1 or $\lambda_2 < 1$.

To find the unique solution, we must invert the ARMA(1, 1) model into an ARMA(0, ∞) model. Consequently, we need to calculate the coefficients ψ_k. For this ARMA(1, 1) process, we have:

$$\begin{aligned} 1 + 0.9z &= (1 - 0.7z)(\psi_0 + \psi_1 z + \ldots) \\ &= \psi_0 + (\psi_1 - 0.7\psi_0)z + (\psi_2 - 0.7\psi_1)z^2 + \ldots \end{aligned}$$

Now we match coefficients for z^k:

$$\begin{aligned} 1 &= \psi_0 \\ 0.9 &= \psi_1 - 0.7\psi_0 \\ 0 &= \psi_j - 0.7\psi_{j-1} \qquad \text{for } j \geq 2. \end{aligned}$$

We can then solve this system of equations using matrices, but we will manually solve in this example.

$$\begin{aligned} &\psi_0 = 1 \\ &\psi_1 = 0.7 + 0.9 = 1.6 \\ &\psi_j = (0.7)^{j-1}(0.7 + 0.9) = (0.7)^{j-1}(1.6) \qquad \text{for } j \geq 2 \end{aligned}$$

We now have the unique, stationary solution for $\{X_t\}$:

$$\begin{aligned} X_t &= \sum_{j=0}^{\infty} \psi_j Z_{t-j} \\ &= Z_t + 1.6\sum_{j=0}^{\infty}(0.7)^{j-1} Z_{t-j}. \end{aligned}$$

6.4 *ACVF* and *ACF* of ARMA Processes

From now on, we will use invertible, causal, and stationary ARMA processes when referring to ARMA processes. We will cover several approaches to compute the *acvf* and *acf* of ARMA processes.

6.4.1 Polynomial Division

Remember an ARMA process can be written as (39), which expands to:

$$X_t = \psi_0 Z_t + \psi_1 Z_{t-1} + \ldots \tag{44}$$

We can directly calculate the *acvf* as:

$$\begin{aligned} K_{XX}(\tau) &= cov(X_t, X_{t+\tau}) \\ &= \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+\tau} \qquad \text{for } \tau \geq 0 \end{aligned} \tag{45}$$

where σ_Z^2 is the variance of Z_t. The *acf* is then:

$$\rho(\tau) = \frac{K_{XX}(\tau)}{K_{XX}(0)}. \tag{46}$$

6.4.2 Difference Equations

Difference equations represent recursive functions or processes. In fact, we have seen them before in these notes, e.g. in Yule-Walker equations. Now let us derive the *acvf* for an ARMA(p, q) process. For any $k \geq 0$:

$$\begin{aligned} \text{Cov}(\theta(B)Z_t, X_{t-k})z &= \text{Cov}(\phi(B)X_t, X_{t-k}) && (47) \\ &= \text{Cov}(X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}, X_{t-k}) \\ &= \text{Cov}(X_t, X_{t-k}) - \phi_1 \text{Cov}(X_{t-1}, X_{t-k}) - \ldots - \phi_p \text{Cov}(X_{t-p}, X_{t-k}) \\ &= K_{XX}(k) - \phi_1 K_{XX}(k-1) - \ldots - \phi_p K_{XX}(k-p). && (48) \end{aligned}$$

We can use (39) to expand the left hand side of (47):

$$\begin{aligned} \text{Cov}(\phi(B)X_t, X_{t-k}) &= \text{Cov}(Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \psi_0 Z_{t-k} + \psi_1 Z_{t-k-1} + \ldots) \\ &= \begin{cases} (\psi_0 \theta_k + \psi + \theta_{k+1} + \ldots + \psi_{q-k}\theta_q)\sigma_Z^2 & \text{if } k \leq q \\ 0 & \text{if } k > q. \end{cases} \end{aligned} \tag{49}$$

We then plug in (48) and (49) to (47) to get:

$$K_{XX}(k) - \phi_1 K_{XX}(k-1) - \ldots - \phi_p K_{XX}(k-p) = c_k \qquad \forall k \geq 0 \tag{50}$$

where

$$c_k = \begin{cases} (\psi_0 \theta_k + \psi + \theta_{k+1} + \ldots + \psi_{q-k}\theta_q)\sigma_Z^2 & \text{if } k \leq q \\ 0 & \text{if } k > q. \end{cases} \tag{51}$$

Remember that $K_{XX}(-k) = K_{XX}(k)$, allowing us to build (30). A benefit of difference equations over polynomial division is that we only calculate q parameters compared to j parameters.

6.5 Parameter Estimation

We will now discuss the over-looming question at the back of our heads. How do we estimate the parameters θ_k and ψ_k? Here, we will present an algorithm for solving the parameters using conditional least squares in an iterative fashion very reminiscent of the optimization algorithm for k-means clustering, the problem in which we divide a set of points into k distinct clusters.

6.6 Conditional Least Squares

As described in the previous subsection, we will use conditional least squares iteratively in order to reach an optimal solution. Our iteration goes as follows:

1. Given current estimations for the parameters $(\phi_1, \phi_2, ...\theta_1, \theta_2)$, calculate the current residuals from the time series data.

2. Calculate each residual as follows:
$Z_t = X_t - \mu - \phi_1(X_{t-1} - \mu) - ... - \phi_p(X_{t-p} - \mu) - \theta_1(Z_{t-1}) - ... - \theta_q(Zt - q)$

3. Solve for new optimal parameters, $(\phi_1^*, \phi_2^*, ...\theta_1^*, \theta_2^*)$, conditioned on the calculated residuals and observations using least squares, stacking known values into a data matrix and then solving using typical least squares techniques.

We see that this process can be performed iteratively to further fine tune the residuals and the parameter estimates. Therefore, our entire process begins with initializing some random weights for the ϕ's and θ's, and performing this process repeatedly until convergence. This process is extremely similar to K-means clustering, involving a step fixing some variables, then solving for another parameter that maximizes a value conditioned on those fixed variables. The fixing of variables in K-means clustering is the assignment of points to clusters. The fixing of variables in the ARMA parameter solving is the calculation of residuals. The maximization in K-means clustering is the assignment of cluster centers as the centroid of the points in the cluster. The maximization in ARMA parameter solving is the solving of parameters ϕ's and θ's given the current residuals.

7 Extensions of ARMA

7.1 ARIMA

After deriving ARMA, auto-regressive and moving average, models, we are finally ready to extend to ARIMA. The jump between the two is actually quite small, including only one new step, differencing.

$$ARIMA(p, d, q) = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 ... - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \theta_2 B^2 ... + \theta_q B^q)\epsilon_t \tag{52}$$

The purpose of differencing is to eliminate trends and enforce stationarity in the data, one of the prerequisites for the ARMA model to fit properly to the time series data. Conveniently, we can perform the differencing prior to the ARMA parameter fitting, meaning we make no changes to the estimation of the model parameters, we simply operate on new differenced data.

7.2 SARIMA

Sometimes, our data has a seasonality, or cyclical behavior. In this case, ARIMA models are not expressive enough in order to truly model the behavior of the time series. In these instances, we add seasonal terms in order to fully express the time series model.

7.2.1 Seasonality

Similar to trends, seasonality exists in time series datasets. The simplest model for seasonal data is:

$$X_t = s_t + Z_t \tag{53}$$

where s_t is a periodic function with period d, i.e. $s_{t+d} = s_t \forall t$. Like with trends, there are several ways to transform seasonal data into stationary data.

- **Parametric Fitting**:
  We can fit a seasonal dataset with sin and cos functions. Let a and b represent amplitudes, $\frac{f}{d}$ represent frequency, and $\frac{d}{f}$ represent the period:

  $$s_t = a_0 + b_0 + \sum_{j=1}^{k}(a_j \cos \frac{2\pi f j}{d} + b_j \sin \frac{2\pi f j}{d}) \tag{54}$$

  The period is identical to the period in oscillations and the frequency is identical to the frequency in oscillations. As f increases, the number of oscillations increases within a set amount of time. We choose k through hyperparameter search.

- **Smoothing**:
  Similar to trends, we can smooth seasonal data. Because s_t only depends on d values $s_1, s_2, \ldots, s_d$, with n samples, s_t can be estimated by:

  $$s_t = \frac{1}{n}\sum_{j=0}^{n} X_{i+jd}. \tag{55}$$

  That is, we estimate the current s_t based on the previous samples of the same period.

- **Seasonal Differencing**: Based on (53), we can difference the data based on period:

  $$X_t - X_{t-d} = s_t - s_{t-d} + Z_t - Z_{t-d} = Z_t - Z_{t-d}. \tag{56}$$

  This lag-d differenced data does not show any seasonality, similar to how differencing in trend data resulted in trendless data.

Now we can tackle data with both trend and seasonal patterns. We can show SARIMA models as follows:

$$\begin{aligned} SARIMA(p,d,q)(P,D,Q)_m = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 ... - \phi_p B^p) \\ (1 - \Phi_1 B^m - \Phi_2 B^{2m} - \Phi_3 B^{3m} ... - \Phi_P B^{Pm})(1-L)^d \\ (1-B)^{Dm} y_t = c + (1 + \theta_1 B + \theta_2 B^2 ... + \theta_q B^q) \\ (1 + \Theta_1 B^m + \Theta_2 B^{2m} ... + \Theta_Q B^{Qm})\epsilon_t \end{aligned} \tag{57}$$

The SARIMA terms compare current observations against observations one period ago, instead of one time step ago. This applies for both AR terms, I terms, and MA terms. The parameter estimation happens in the same fashion, just with different observations being depended on for a current observation.

8 Model Selection

There are several methods to determine the best model among a family of alternatives, i.e. they help us choose p, q that is appropriate for a data set. Like with linear regression, we can fit a polynomial of degree 19 for a training dataset of 20 points - this model will perfectly fit the training dataset but will most likely perform poorly on future values. Similar to ridge regression, in the sense of using fewer parameters, more robust models will use lower degree polynomials. Consequently, the 19-degree polynomial will most likely be more sensitive to noise than a 2-degree polynomial. Therefore, we aim to choose the minimum amount of model parameters to fit the data well without losing too much generalization.

8.1 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) estimates the quality of models relative to other models of the same data set. It evaluates the log likelihood, ℓ of the data based on the number of parameters. The following the the likelihood function:

$$\begin{aligned} L(X_1, \ldots, X_k; \beta, \sigma^2) &= f(X_1, \ldots, X_k; \beta, \sigma^2) \\ &= f(X_1; \beta, \sigma^2) \prod_{i=1}^{k} f(X_i | X_{i-1}; \beta, \sigma^2). \end{aligned} \tag{58}$$

We take the log of L to get:

$$\ell(\beta, \sigma^2) = \log f(X_1; \beta, \sigma^2) + \sum_{i=1}^{k} f(X_i | X_{i-1}; \beta, \sigma^2) \tag{59}$$

The AIC is then defined as:

$$AIC = -2 \log(\ell) + 2k \tag{60}$$

where $k = p + q + 2$. The 2 comes from the mean and noise variance. The first term measures the fit of the model, which is counteracted by k because increasing the number of parameters almost always increases model fit. We choose the model that has the lowest AIC score. Where with ridge regression having too large a weight vector is penalized, the AIC penalizes our model for having too large a moving window.

8.2 Bayesian Information Criterion (BIC)

The Bayesion information criterion performs a similar comparison - it minimizes:

$$BIC = -2 \log(\ell) + k \log n \tag{61}$$

where n is the number of data points. We notice this penalty for more parameters is larger than the penalty of AIC. Consequently, BIC chooses sparser models compared to AIC.

9 References

https://en.wikipedia.org/wiki/White_noise
https://www.machinelearningplus.com/machine-learning/bias-variance-tradeoff/
https://people.duke.edu/~rnau/411diff.htm
https://www.stat.tamu.edu/~suhasini/teaching673/chapter3.pdf
http://www.maths.qmul.ac.uk/~bb/TimeSeries/TS_Chapter6_1.pdf
https://www.stat.berkeley.edu/~aditya/resources/LectureTWO.pdf
https://www.asc.ohio-state.edu/de-jong.8/note2.pdf
https://math.unice.fr/~frapetti/CorsoP/Chapitre_4_IMEA_1.pdf
https://www-jstor-org.libproxy.berkeley.edu/stable/pdf/4615673.pdf?refreqid=excelsior%3A640726649d0edf661655f0d5d3a3e167
https://courses.maths.ox.ac.uk/node/view_material/924