DS-GA 1018: Lecture 3 Notes

Sebastian Wagner-Carena

September 2023

1 Review

We reviewed the following concepts the beginning of class:

Definition 1 *An* ***autoregressive model of order p – AR(p)*** *– is a random process with the form:*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + W_t \tag{1}$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$ and $\phi_1, \ldots, \phi_p$ are constants. For the model to be order p, it must be true that $\phi_p \neq 0$. X_t is stationary, and the mean of X_t is 0. For cases where the mean of X_t is nonzero, we can recast our AR(p) relation as:

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots \phi_p(X_{t-p} - \mu) + W_t \tag{2}$$

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + W_t \tag{3}$$

$$\alpha = \mu(1 - \phi_1 - \ldots - \phi_p). \tag{4}$$

Definition 2 *A* ***moving average model of order p – MA(p)*** *– is a random process with the form:*

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_p W_{t-p} \tag{5}$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$ and $\theta_1, \ldots, \theta_p$ are constants. For the model to be order p, it must be true that $\theta_p \neq 0$.

Definition 3 *An Autoregressive Moving Average process of order p,q - ARMA(p,q) is a process with the form:*

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q} \tag{6}$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$, $\phi_1, \ldots, \phi_p$ are constants, $\theta_1, \ldots, \theta_q$ are constants, and both $\theta_q \neq 0$ and $\phi_p \neq 0$.

2 Backshift Operators, Causality, and Invertability

Now that we have our full model (we will introduce a further extension later in this lecture), we want to understand how to quantify the statistics, measure the model parameters on data, and make new data predictions given the parameters of the model. In this pursuit it will be useful to reframe our AR and MA processes slightly. Let's return the the definition of our AR(p) process:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + W_t. \tag{7}$$

We can trivially rearrange the terms as follows:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots \phi_p X_{t-p} = W_t. \tag{8}$$

We will also want to introduce the **backshift operator**, B. We will define our backshift operator such that:

$$BX_t = X_{t-1}. \tag{9}$$

This leads us to the **autoregressive operator**.

Definition 4 *The autoregressive operator, $P(B)$, for an AR(p) model is defined as:*

$$P(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \tag{10}$$

such that:

$$P(B)X_t = W_t. \tag{11}$$

Remember that last week we derived the expectation value and covariance of an AR(1) process by expanding out the AR(1) equation:

$$\begin{align} X_t &= \phi X_{t-1} + W_t \tag{12} \\ &= W_t + \phi W_{t-1} + \phi^2 W_{t-2} + \ldots. \tag{13} \end{align}$$

The equation on the second line looks like an MA process. In fact, it is an MA process of infinite order. This means that we can write:

$$\begin{align} P(B)X_t &= W_t \tag{14} \\ X_t &= \psi(B)W_t \tag{15} \\ P(B)^{-1} &= \psi(B). \tag{16} \end{align}$$

Last week we found that for $|\phi| < 1$ our AR(1) process was causal. It's natural to wonder how we can extend this notion of causality to a higher order process or an ARMA process. As we will see soon, $P(B)^{-1}$ will be vital to this definition. It also confuses the definitions of AR and MA to say that we can write one in terms of the other. To make matter more commplicated we can go down the same rabbit hole with our MA process:

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_p W_{t-p}. \tag{17}$$

We can also write this in terms of backshift operators and introduce the **moving average operator**:

Definition 5 *The moving average operator for an MA(p) process is defined as:*

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_p B^p, \tag{18}$$

such that:

$$X_t = \Theta(B) W_t. \tag{19}$$

As with the AR process, we can reframe our MA process as:

$$X_t = \Theta(B) W_t \tag{20}$$
$$\Theta(B)^{-1} X_t = W_t. \tag{21}$$

It turns out that just like we were concerned with causality for our AR process, we can be concerend with invertability for our MA process. Specifically, consider two MA(1) processes:

$$X_t = W_t + \frac{1}{5} W_{t-1}, \quad p(W_t) = \mathcal{N}(0, 25) \tag{22}$$
$$Y_t = V_t + 5 V_{t-1}, \quad p(V_t) = \mathcal{N}(0, 1). \tag{23}$$

In every statistic we can measure on X_t and Y_t the two processes are indistinguishable[1]. The only way we would know the difference is to have access to their white noise process, and we do not have that. The only difference

[1] Convince yourself of this by plugging these two processes into our mean and covariance equations from last week.

is that one of these two processes can be written as an infinite AR process and the other cannot. Specifically, we can write:

$$X_t = W_t + \theta W_{t-1} \tag{24}$$

$$\sum_{j=0}^{\infty} (-\theta)^j X_{t-j} = W_t. \tag{25}$$

But the term on the right-hand side is divergent if $|\theta| > 1$. So only the model with $\theta = \frac{1}{5}$ is invertable. We will also define this more broadly using $\Theta(B)^{-1}$ in a moment. Before we do that, let me convince you of the value of these operators one more time. Imagine we have a simple white noise process:

$$X_t = W_t. \tag{26}$$

We can trivially also write:

$$0.5X_{t-1} = 0.5W_{t-1} \tag{27}$$

$$X_t - 0.5X_{t-1} = W_t - 0.5W_{t-1} \tag{28}$$

$$X_t = 0.5X_{t-1} + W_t - 0.5W_{t-1}. \tag{29}$$

This should be alarming, because I just made a white noise process look like an ARMA(1,1) process.

Let's resolve these issues in reverse order. First, we can write any ARMA process as:

$$P(B)X_t = \Theta(B)W_t. \tag{30}$$

We will demand that an ARMA(p, q) process has the property that $P(B)$ and $\Theta(B)$ do not share any roots. That is for all complex numbers z_i for which $P(z_i) = 0$ and for all complex numbers z_j for which $\Theta(z_j) = 0$ there does not exist i or j for which $z_i = z_j$. Returning to the white noise example we made look like ARMA(1,1) we have that:

$$P(B) = 1 - 0.5B \tag{31}$$

$$\Theta(B) = 1 - 0.5B. \tag{32}$$

We can simplify both equation by dividing by a common factor of $1 - 0.5B$ (eliminating the common root of 2) and get:

$$P(B) = 1 \tag{33}$$

$$\Theta(B) = 1, \tag{34}$$

thereby demonstrating that this process is in fact white noise.

Next we can return to the invertability issue for the MA process:

Definition 6 *An ARMA(p, q) process is invertable if the time series can be written as:*

$$\pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = W_t \tag{35}$$

with the infinite sum $\sum_{j=0}^{\infty} |\pi_j| < \infty$. We can determine $\pi(B)$ as:

$$\pi(B) = \frac{P(B)}{\Theta(B)}. \tag{36}$$

The conditions for invertability hold so long as the roots of $\Theta(z)$ lie outside the unit circle. See Shumway and Stoffer for the proof of this statement.

We can make a similar definition for causality:

Definition 7 *An ARMA(p, q) process is causal if the time series can be written as:*

$$X_t = \psi(B)W_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} \tag{37}$$

with the infinite sum $\sum_{j=0}^{\infty} |\psi_j| < \infty$. We can determine $\psi(B)$ as:

$$\psi(B) = \frac{\Theta(B)}{P(B)}. \tag{38}$$

The conditions for causality hold so long as the roots of $P(z)$ lie outside the unit circle. Again, see Shumway and Stoffer for the proof of this statement.

Let's consider the process defined by:

$$X_t = 0.4X_{t-1} + 0.45X_{t-2} + W_t + W_{t-1} + 0.25W_{t-2}. \tag{39}$$

We want to know what order ARMA process it is and whether or not it is causal and invertable. It looks like ARMA(2,2), but looks can be deceiving. Let's get started by finding our $P(B)$ and $\Theta(B)$ operators:

$$P(B) = 1 - 0.4B - 0.45B^2 \tag{40}$$
$$\Theta(B) = 1 + B + 0.25B^2. \tag{41}$$

If we factor both equation we find:

$$P(B) = (1 + 0.5B)(1 - 0.9B) \tag{42}$$
$$\Theta(B) = (1 + 0.5B)^2. \tag{43}$$

We can cancel out the shared root to get:

$$P(B) = (1 - 0.9B) \tag{44}$$

$$\Theta(B) = (1 + 0.5B). \tag{45}$$

and find that our process is ARMA(1,1) and can be written as:

$$X_t = 0.9X_{t-1} + W_t + 0.5W_{t-1}. \tag{46}$$

The process is invertible and causal because both roots (10/9, -2) are outside the unit circle.

3 ACF and CCF

We have a strong definition of our ARMA process now, but we'll also want to quantify its statistics and parameters given data. First, let's talk about how to estimate our two main statistics of interest: the mean and the covariance function. Throughout we will assume the process is stationary. The sample mean is the easiest, it can be estimated by:

$$\hat{\mu} = \frac{1}{T} \sum_{t=0}^{T} X_n, \tag{47}$$

where T is the length of our sample. The sample covariance is a bit more involved:

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (x_{t+h} - \hat{\mu})(x_t - \hat{\mu}). \tag{48}$$

Note that the effective size of our sample is smaller when we calculate the sample covariance because we require an offset of h. Regardless we still divide by T because this estimator has a few nicer properties than the estimator we get if we divide by $T - h$ (see Shumway and Stoffer). We can also calculte the sample correlation from the sample covariance:

$$\hat{\rho}(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{49}$$

These estimators are great, but unless $T \to \infty$ they are not going to give us the true values. It's useful to understand the expected error of these estimators to have a handle on the significance of the measurement. For our correlation function, the best metric is a comparison to white noise (which we know has a correlation coefficient of 0 for all $h > 0$). For a large sample and finite h, the sample ACF of white noise is normally distributed with mean 0 and variance:

$$\sigma^2_{\hat{\rho}(h)} = \frac{1}{T}. \tag{50}$$

It is common to declare a measurement of $\hat{\rho}(h)$ significant if:

$$|\hat{\rho}(h)| > \frac{2}{\sqrt{T}}. \tag{51}$$

This is equivalent to the 95$^{\text{th}}$ percentile.

4 Partial Auto-Correlation Function

Let's imagine we have a process we suspect is a MA process. If we have a reliable estimate of the ACF we can immediately tell its order by the largest non-zero value. For an AR process that's a lot less clear. We have already shown that for an AR(2) process the ACF is non-zero for all separations. More generically, even if X_{t+h} does not directly depend on X_t in the equation that defines the process, it can still be correlated through some intermediary X_{t+k} with $k < h$. What we would like to know is the correlation between X_{t+h} and X_t conditioned on $X_{t+1:t+h-1}$:

$$\gamma_{X_{t+h},X_t|X_{t+1:t+h-1}} \tag{52}$$

If we assume that our process is Gaussian, between X_{t+h} and X_t after having subtracted the optimal linear predictor using $X_{t+1:t+h-1}$. We will call these linear predictor $\hat{X}_{t+h}$ and $\hat{X}_t$. This is known as the partial auto-correlation function:

Definition 8 *The partial auto-correlation function (PACF) for a stationary process, X_t, is defined as:*

$$\phi_{11} = \gamma(t+1,t) = \gamma(1) \tag{53}$$

$$\phi_{hh} = \mathbb{E}[(X_{t+h} - \hat{X}_{t+h})(X_t - \hat{X}_t)], \tag{54}$$

where $\hat{X}_{t+h}$ and $\hat{X}_t$ are the regressions on both variables using the variables $X_{t+1:t+h-1}$ that minimizes mean squared error.

We will show this in the next lecture, but for now we will assert that the PACF of an AR(p) process is 0 for $h > p$. Depending on interest, we may discuss the Durbin–Levinson Algorithm for calculating the PACF next week.

5 Estimating Parameters

Once we've selected what model to use on our data, we will want to estimate its parameters. For now, let's introduce the simplest choice we can make in Bayesian inference: the maximum a posteriori (MAP) estimate. Assume we have some data X and some model parameters θ. The MAP estimate is given by:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|X) \tag{55}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \frac{p(X|\theta)p(\theta)}{p(X)}. \tag{56}$$

We already have all the tools we need to evaluate this on our ARMA processes. We will see an example of prior choices in lab, and we can write the likelihood from the equations that define the process. For an AR(1) process we have two parameters, ϕ_1 and σ_w. If we set the boundary condition $X_1 = W_1$, we can write:

$$p(X_{1:t}|\phi_1, \sigma_w) = p(X_1|\phi_1, \sigma_w)p(X_2|X_1, \phi_1, \sigma_w)\times$$
$$p(X_3|X_{1:2}, \phi_1, \sigma_w)\dots p(X_t|X_{1:t-1}, \phi_1, \sigma_w). \tag{57}$$

But we know that for an AR(1) process X_t is conditionally independent of $X_{1:t-2}$ given X_{t-1} so we can simplify this equation to:

$$p(X_{1:t}|\phi_1, \sigma_w) = p(X_1|\phi_1, \sigma_w)p(X_2|X_1, \phi_1, \sigma_w)\times$$
$$p(X_3|X_2, \phi_1, \sigma_w)\dots p(X_t|X_{t-1}, \phi_1, \sigma_w). \tag{58}$$

The likelihood of X_t given X_{t-1} and the parameters is just a Gaussian distribution with mean $\phi_1 X_{t-1}$ and variance σ_w^2:

$$p(X_t|X_{t-1},, \phi_1, \sigma_w) = \mathcal{N}(\phi_1 X_{t-1}, \sigma_w) \tag{59}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left[\frac{(X_t - \phi_1 X_{t-1})^2}{2\sigma_w^2}\right] \tag{60}$$

You will need to extend this to AR(2) for lab this week, but I will spoil the answer and tell you that it's just a matter of adding $\phi_2 X_{t-2}$ to the mean.

One final note – when we take the product of probability distributions we quickly get vanishingly small numbers. From an implementation standpoint, this will run into the numerical precision of our computer. To avoid this, we often operate with log pdfs instead of the pdf itself and take advantage of the property that:

$$\log\left[p(x)p(y)\right] = \log\left[p(x)\right] + \log\left[p(y)\right]. \tag{61}$$

6 Difference Equations

As we wrap up, let's combine two of the concepts from this week and see how the backshift operator allows us calculate the correlation coefficient for an AR(2) process. Let's start with our equation:

$$X_{t+h} = \phi_1 X_{t+h-1} + \phi_2 X_{t+h-2} + W_{t+h}. \tag{62}$$

Multiply through by X_t and taking an expectation:

$$\begin{aligned}
\mathbb{E}[X_t X_{t+h}] &= \phi_1 \mathbb{E}[X_{t+h-1}X_t] + \phi_2 \mathbb{E}[X_{t+h-2}X_t] + \mathbb{E}[W_{t+h}X_t] && (63)\\
\gamma(h) &= \phi_1\gamma(h-1) + \phi_2\gamma(h-2) && (64)\\
P(B)\gamma(h) &= 0. && (65)
\end{aligned}$$

This is known as a **difference equation**. The roots of our backshift operator are the same roots of our difference equation. The solutions of a difference equation of order p:

$$u_n - \alpha_1 u_{n-1} - \ldots - \alpha_p u_{n-p} = 0 \tag{66}$$

are given by:

$$u_n = z_1^{-n}P_1(n) + z_2^{-n}P_2(n) + \ldots + z_r^{-n}P_r(n), \tag{67}$$

where r is the number of distinct roots and $P_j(n)$ is a polynomial in n of degree $m_j - 1$ where m_j is the multiplicity of the root z_j. For AR(2) we have a difference equation of order 2. If there are two distinct root, $z_1 \neq z_2$, then the solution has the form:

$$\rho(h) = c_1 z_1^{-|h|} + c_2 z_2^{-|h|}. \tag{68}$$

If instead the two roots are equal then the solution has the form:

$$z_1^{-|h|}(c_1 + c_2|h|). \tag{69}$$

For higher orders there is a wider variety of solutions.

7 ARIMA

Finally, it's worth introducing one extension to our ARMA model by introducing a non-stationary mean. In short, we can break our time series into:

$$X_t = \mu_t + Y_t \tag{70}$$

where Y_t is stationary and μ_t is a time-dependent mean. We can extend ARMA to **integrated ARMA (ARIMA)**:

Definition 9 *A process is said to be ARIMA(p, d, q) if:*

$$\nabla^d X_t = (1 - B)^d X_t, \tag{71}$$

is ARMA(p, q). This means that we can write the process as:

$$P(B)(1 - B)^d X_t = \Theta(B) W_t. \tag{72}$$

We will touch on this a bit more in the next lecture