DS-GA 1018: Lecture 2 Notes

Sebastian Wagner-Carena

September 2024

1 Review

We reviewed the following concepts at the beginning of class:

Definition 1 *Consider two time series drawn from our stochastic process, $\{X_t, \dots X_{t+k}\}$ and $\{X_{t+h}, \dots X_{t+h+k}\}$. The stochastic process is weakly stationary if:*

i *The mean value of the process does not depend on time: $\mu_X(t) = constant$*

ii *The auto-correlation function only depends on the absolute difference between the time indices: $\rho_X(t,s) = \rho_X(|t-s|)$*

iii *The variance is finite.*

Definition 2 *Consider two time series drawn from our stochastic process, $\{X_t, \dots X_{t+k}\}$ and $\{X_{t+h}, \dots X_{t+h+k}\}$. The stochastic process is trend stationary if:*

i *The mean value of the process has a time dependence: $\mu_X(t) = f(t)$*

ii *The auto-correlation function only depends on the absolute difference between the time indices: $\rho_X(t,s) = \rho_X(|t-s|)$*

iii *The variance is finite.*

Definition 3 *The Makrov boundary for X is the union of its parents, it children, and the parents of children. If a variable is conditioned on its Markov boundary, it becomes independent of all other random variables.*

See last week's notes for more information on these topics. As an additional review, let us use last week's graphical model to examine the Markov boundary.

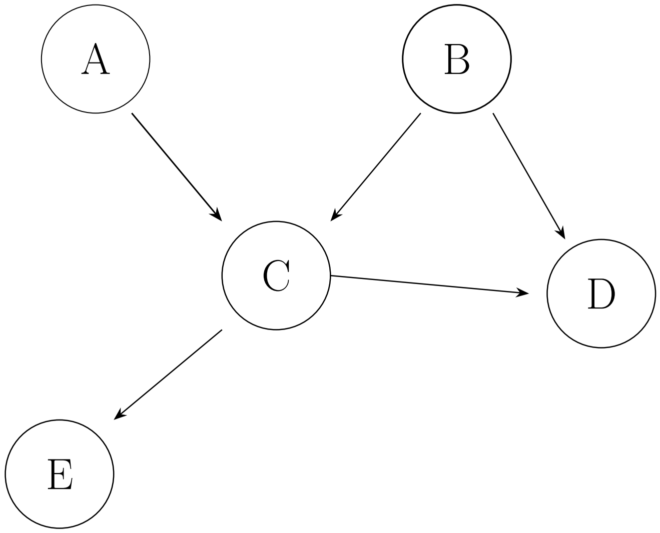

Figure 1: An example of a graphical model. This graphical model is a directed acyclic graph.

Example 1.1 *Consider the graphical model given in Figure 1. What is the Markov boundary of node D and node A?*

The Markov boundary is the union of the parents, children, and the parents of children of a node. For node D that is the set $\{B, C\}$. For node A we need to include its child C and the parents of its child B. The Markov boundary is not the only set of variables that create independence between two subsets of our graph. In fact, it can be useful to think about the Markov blanket of a variable.

Definition 4 *A Markov blanket of X is a set of variables that, when conditioned on, make X independent of all other random variables. The Markov boundary is the minimal version of the Markov blanket.*

2 Baysian Inference Continued

In statistical inference, there are two useful properties we often exploit. We will write these rules in terms of the probability distribution functions p, but we can also express them in terms of the cumulative density functions

P (see last week's notes). The first is the **sum rule**:

$$p(x) = \int p(x, y)dy, \tag{1}$$

where $p(x)$ is called the marginal distribution of x. The process of applying the sum rule to get the marginal distribution is often referred to as **marginalizing** over the variable y. The second is the **product rule**:

$$p(x, y) = p(x)p(y|x). \tag{2}$$

These concepts will show up in times series analysis repeatedly. For example, given a distribution $p(X_{1:t+1})$, we will often want to make predictions of the next time step, $t+1$, given an observation of a times series up to time index t:

$$p(X_{t+1}|X_{1:t}) = \frac{p(X_{1:t+1})}{p(X_{1:t})}. \tag{3}$$

In this case, to get the bottom term, we will have to take advantage of the sum rule.

We can use the product rule to prove Bayes' theorem:

$$\begin{align} p(A, B) &= p(B|A)p(A) \tag{4}\\ p(A, B) &= p(A|B)p(B) \tag{5}\\ p(A|B)p(B) &= p(B|A)p(A) \tag{6}\\ p(A|B) &= \frac{p(B|A)p(A)}{p(B)}. \tag{7} \end{align}$$

In most Bayesian inference, we are interested in learning the parameters of some model, θ, given some dataset $\mathcal{D}$. Our model will often provide us with the probability distribution $p(\mathcal{D}|\theta)$, and we will use Bayes' theorem to calculate our desired distribution:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}. \tag{8}$$

In Bayesian inference, we update our prior beliefs based off of the data that we collect. For this reason, $p(\theta)$ is called our **prior**. It represents our prior beliefs about the parameters of our model. The distribution $p(\mathcal{D}|\theta)$ is called **likelihood**, and it represents how likely our data is given a specific set of parameters for our model. The distribution $p(\theta|\mathcal{D})$ is called the **posterior**.

It represents our updated beliefs about the value of the parameters in our model. The denominator, $p(\mathcal{D})$ is the marginal distribution of the data. As you can see, it does not depend on the value of the parameters θ, so we can think of it as a normalizing factor. It is possible to conduct Bayesian inference even when $p(\mathcal{D})$ cannot be calculated. When a distribution either has no analytical form or is too complex to calculate, we will refer to it as **intractable**.

3 Multivariate Gaussian Review

Before we introduce our first major time series model, it is important for us to review the basics of multivariate Gaussian distributions. As we will see, they will show up repeatedly in our inference. The multivariate Gaussian distribution is an extension of the Gaussian distribution to multiple dimensions. As with the univariate Gaussian, it is fully described by a vector of means, $\boldsymbol{\mu}$, and a matrix of (co)variances, $\boldsymbol{\Sigma}$. Written in the language of last week's lecture, if we consider the vector $\boldsymbol{X} = \begin{bmatrix} X_1, X_2, \dots X_n \end{bmatrix}$ then:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{X_1} & \mu_{X_2} & \dots & \mu_{X_n} \end{bmatrix} \tag{9}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \gamma_{X_1,X_1} & \gamma_{X_1,X_2} & \cdots & \gamma_{X_1,X_n} \\ \dots & \dots & \dots & \dots \\ \gamma_{X_n,X_1} & \gamma_{X_n,X_2} & \cdots & \gamma_{X_n,X_n} \end{bmatrix}. \tag{10}$$

I will try to bold variables when they represent a vector / matrix instead of a single value. The probability density functions of the multivariate Gaussian distributions is:

$$\mathcal{N}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right]. \tag{11}$$

In Equation 11, d represents the number of dimensions, $|\boldsymbol{\Sigma}|$ represents the determinant of the covariance matrix, and $\boldsymbol{x}$ is the value at which we are evaluating the probability density function of our vector of random variables $\boldsymbol{X}$. When we are working with vector and matrices, we will often use the shorthand $p(\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x})$ so that we can keep our vectors lowercased and our matrices uppercased.

3.1 Marginal and Conditional Distributions

We will frequently take advantage of the marginal and conditional distributions of a multivariate Gaussian. In these cases, we will want to divide our larger vector of random variables $\boldsymbol{X}$ into two subsets $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$. Note that we have made no assumption to the size of the subsets $|a|$ and $|b|$ beyond that $|a|, |b| \geq 1$ and that $|a|+|b| = |d|$, where $|d|$ is the dimensionality of our original vector of random variables $\boldsymbol{X}$. We can divide our pdf from Equation 11 accordingly:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_a \\ \boldsymbol{x}_b\end{bmatrix} \mid \begin{bmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right). \tag{12}$$

We have assumed that the subset b comes after the subset a in our vector for convenience. We can always reorder the variables in our vector and covariance matrix just by swapping indices (we will do this in the lab this week). To make things, more explicit, let's consider the three-dimensional Gaussian with mean and covariance given by:

$$\boldsymbol{\mu}_{\text{example}} = \begin{bmatrix}\mu_1 \\ \mu_2 \\ \mu_3\end{bmatrix} \tag{13}$$

$$\boldsymbol{\Sigma}_{\text{example}} = \begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2} & \Sigma_{1,3} \\ \Sigma_{2,1} & \Sigma_{2,2} & \Sigma_{2,3} \\ \Sigma_{3,1} & \Sigma_{3,2} & \Sigma_{3,3}\end{bmatrix}. \tag{14}$$

If we take the subset a to be the first two variables and the subset b to be the last, then we can have that the mean values in Equation 12 are:

$$\boldsymbol{\mu}_a = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} \tag{15}$$

$$\boldsymbol{\mu}_b = \begin{bmatrix}\mu_3\end{bmatrix} \tag{16}$$

and the covariance matrix values in Equation 12 are:

$$\boldsymbol{\Sigma}_{aa} = \begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2}\end{bmatrix} \tag{17}$$

$$\boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^T = \begin{bmatrix}\Sigma_{1,3} \\ \Sigma_{2,3}\end{bmatrix} \tag{18}$$

$$\boldsymbol{\Sigma}_{bb} = \begin{bmatrix}\Sigma_{3,3}\end{bmatrix}. \tag{19}$$

Given the factorization in Equation 12, we can now easily define the marginal and conditional probability density functions. The marginal probability density function is given by:

$$p(\boldsymbol{x}_a) = \mathcal{N}\left(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right) \tag{20}$$

$$p(\boldsymbol{x}_b) = \mathcal{N}\left(\boldsymbol{x}_b|\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb}\right). \tag{21}$$

The conditional probability density function is a bit more complicated:

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}\left(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right) \tag{22}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \tag{23}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \tag{24}$$

We can reverse a and b in the equation above to get the other equation. Note that the conditional covariance matrix does not depend on the value of the conditioned variable $\boldsymbol{x}_b$. In lab we will play around with the conditional and marginal distributions to gain some intuition.

3.2 Gaussians Under Linear Operations

For any vector of random variables $\boldsymbol{X}$, matrix $\boldsymbol{A}$, and vector $\boldsymbol{y}$, we know that the mean and covariance:

$$\boldsymbol{\mu}_{(\boldsymbol{AX}+\boldsymbol{y})} = \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{y} \tag{25}$$

$$\boldsymbol{\Sigma}_{(\boldsymbol{AX}+\boldsymbol{y})} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{A}^T \tag{26}$$

Since our multivariate Gaussian is defined entirerly by a mean and covariance, for the vector of random variables $\mathbf{Z} = \boldsymbol{AX} + \boldsymbol{y}$ we can write:

$$p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}|\boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{y}, \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{A}^T\right) \tag{27}$$

A consequence of this is that **Gaussians are closed under linear operations.** That means that, under a linear operations, a Gaussian will be mapped to another Gaussian. This is important because we will often start from white noise assumptions and then use linear operations to build more complex models.

4 Introduction to ARMA

With the fundamentals in place, we can start digging into our first model, the Autoregressive Moving Average Model (ARMA).

4.1 Auto-regressive Process

Consider a random process of the form:

$$X_t = \phi X_{t-1} + W_t, \tag{28}$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$ and $|\phi| < 1$. This is known as an autoregressive process because each value of the random variable in our series depends on another random variables in the series. This autoregressive process is of order 1 because the current random variable only has a direct dependence on the previous time-step's random variable. Let's calculate the statistics of this random process by expanding the recursion:

$$\begin{aligned} X_t &= \phi X_{t-1} + W_t && (29) \\ &= W_t + \phi W_{t-1} + \phi^2 W_{t-2} + \dots. && (30) \end{aligned}$$

For now, we are ignoring the boundary conditions of our problem (i.e. what happens when we get to $t = 0$). The mean of this process is given by:

$$\begin{aligned} \mu_X &= \mathbb{E}\left[\sum_{h=0}^{\infty} \phi^h W_{t-h}\right] && (31) \\ &= 0. && (32) \end{aligned}$$

The covariance is given by:

$$\begin{aligned} \gamma_X(h) &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \phi^j W_{t+h-j}\right)\left(\sum_{k=0}^{\infty} \phi^k W_{t-k}\right)\right] && (33) \\ &= \sigma_W^2 \sum_{j=0}^{\infty} \phi^{h+j} \phi^j && (34) \\ &= \sigma_W^2 \phi^{|h|} \frac{1}{1-\phi^2}. && (35) \end{aligned}$$

Notice that this process is stationary.

We can generalize this into the broader concept of an AR(p) process:

Definition 5 *An **autoregressive model of order p – AR(p)** – is a random process with the form:*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots \phi_p X_{t-p} + W_t \qquad (36)$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$ and $\phi_1, \dots, \phi_p$ 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:

$$\begin{aligned} X_t - \mu &= \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \dots \phi_p(X_{t-p} - \mu) + W_t && (37) \\ X_t &= \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots \phi_p X_{t-p} + W_t && (38) \\ \alpha &= \mu(1 - \phi_1 - \dots - \phi_p). && (39) \end{aligned}$$

Note that, by construction, we assume that our autoregressive process is causal: a random variable only depends on random variables that came before it. In the slides you can find some example of AR processes.

4.2 Moving Average

Consider a random process of the form:

$$X_t = W_t + \theta W_{t-1}, \tag{40}$$

where W_t is drawn from $\mathcal{N}(0, \sigma_W^2)$ and $|\theta| < 1$.

This is known as a moving average process, with this particular process being order 1. Let's once again calculate our statistics of interest, starting with the mean:

$$\begin{align} \mu_X &= \mathbb{E}\left[W_t + \theta W_{t-1}\right] \tag{41} \\ &= 0. \tag{42} \end{align}$$

The covariance is given by:

$$\begin{align} \gamma_X(h) &= \mathbb{E}\left[\left(W_{t+h} + \theta W_{t+h-1}\right)\left(W_t + \theta W_{t-1}\right)\right] \tag{43} \\ &= \begin{cases} \sigma_W^2(1+\theta^2) & h = 0 \\ \sigma_W^2 \theta & |h| = 1 \\ 0 & |h| > 1 \end{cases} . \tag{44} \end{align}$$

The covariances are much more localized than in the AR case. The is also an example of a stationary process.

We can generalize this into the broader concept of an MA(p) process.

Definition 6 *A* ***moving average model of order p – MA(p)*** *– is a random process with the form:*

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_p W_{t-p} \tag{45}$$

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 $\theta_p \neq 0$.

4.3 Autoregressive Moving Average Model

We can combine these two concepts together to build what's known as an Autoregressive Moving Average Model (ARMA):

Definition 7 *An Autoregressive Moving Average process of order p,q - ARMA(p,q) is a process with the form:*

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q} \tag{46}$$

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* $\phi_p \neq 0$.

Next week we will dive more deeply into this model.