DS-GA 1018: Lecture 1 Notes

Sebastian Wagner-Carena

September 2024

1 Course Logistics

All of the relevant course logistics can be found on the syllabus posted on Brightspace.

2 Time Series Data Examples

Examples can be found in the Lecture 1 slides.

3 Introduction to Time Series

Definition 1 *A time series is a collection of random variables indexed by time:*

$$\textit{time series: } \{X_1, X_2, X_3, \cdots X_t \cdots\}, \tag{1}$$

where the subscript denotes the time index.

The above notation is common for discrete time intervals, but it is also possible for our time series to be continuous in time. For most of this class, we will be considering discrete time intervals. The process that generates the time series is denoted as a **stochastic process**. The data we observe is one / multiple realization of this stochastic process.

There are two properties of time series that deviate from the traditional case:

- **Ordering** – The data in a time series has a natural ordering established by the time index.

- **Not I.I.D.** – While our definition doesn't require that the data not be independently, identically distributed, we are interested in time series because there are some time dependencies in our data that we want to use for analysis or prediction.

We can fully specify our time series through the **joint cumulative distribution function (cdf)**:

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_t \leq x_t, \ldots). \tag{2}$$

This measures the probability that the values of our series are less than a series of constant points $x_1, x_2, \ldots x_t, \ldots$. Another way to quantify the statistics of our time series is through the **joint probability density function (pdf)**:

$$p(X_1 = x_1, X_2 = x_2, \ldots X_t = x_t, \ldots). \tag{3}$$

This measures the probability density that the values of our series have a specific value. Generically, it may be intractable to write out the full cdf / pdf. We will exploit knowledge of (or make assumptions about) the structure of the problem in order to make the math tractable.

To make this challenge clear, let's consider a sequence of random variables with values 0 or 1. Imagine I wanted to understand the distribution of a sequence of length 2: $\{X_1, X_2\}$. If I make no assumptions about the structure of my data, there are 4 possibilities, and 3 unique parameters for the distribution[1]. Now imagine I want to understand the distribution of a sequence of length 3: $\{X_1, X_2, X_3\}$. There are now 8 possibilities and 7 parameters. The number of parameters grows exponentially with the length of the sequence! For a series of length T, I would need to collect at least 2^T datapoints just to have seen every option. We will often be considering time series where the number of realizations is $\mathcal{O}(1)$.

3.1 Basic Statistics of a Time Series

Because the distribution functions will often be unwieldy objects to work with for a time series, we will take advantage of a number of statistical summaries of the process. These statistics will be useful for quantifying our assumptions, correcting trends in our time series, and building our models. The first is the mean of the time series:

$$\mu_X(t) = \mathbb{E}(X_t), \tag{4}$$

[1]The final value is constrained by the sum of the probabilities equaling 1.

where $\mathbb{E}$ denotes the expectation value. Another useful statistic is the (auto)-covariance function:

$$\gamma_X(s,t) = \mathbb{E}\left[(X_s - \mu_X(s))\,(X_t - \mu_X(t))\right]. \tag{5}$$

Sometimes, it will be useful to understand how correlated two points in time are independent of each parameter's intrinsic variance. In those cases, the auto-correlation function can be useful:

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

The auto-correlation function is a measure of how predictable each time point is of the other. Finally, we may want to understand how predictive one time series is of another. In these cases, it is useful to employ the cross-covariance and cross-correlation function:

$$\gamma_{X,Y}(s,t) = \mathbb{E}\left[(X_s - \mu_X(s))\,(Y_t - \mu_Y(t))\right] \tag{7}$$

$$\rho_{X,Y}(s,t) = \frac{\gamma_{X,Y}(s,t)}{\sqrt{\gamma_X(s,s)\gamma_Y(t,t)}}. \tag{8}$$

Example 3.1 *The statistical properties of white noise.*

A stochastic process that will show up repeatedly in this course is **white noise**. Within a white noise process, the distribution of each random variable is defined by a Gaussian distribution with mean zero and variance σ_w^2:

$$p(X_t) = \mathcal{N}(\mu = 0, \sigma = \sigma_w). \tag{9}$$

By construction, the mean of the white noise process is zero:

$$\mu_X(t) = 0 \tag{10}$$

The covariance is also fairly quick to calculate:

$$\gamma_X(s,t) = \mathbb{E}\left[(X_s - \mu_X(s))\,(X_t - \mu_X(t))\right] \tag{11}$$

$$= \int (X_s X_t - X_s\mu_X(t) - X_t\mu_X(s) + \mu_X(s)\mu_X(t))\, p(X_s, X_t)\; dX_s\; dX_t \tag{12}$$

$$= \int X_s X_t\; p(X_s, X_t)\; dX_s\; dX_t - \mu_X(s)\mu_X(t) \tag{13}$$

$$= \begin{cases} s = t & \sigma_w^2 \\ s \neq t & 0 \end{cases}. \tag{14}$$

3.2 Structure and Regularities

Let's start by breaking our joint cumulative distribution function down into a number of conditional distributions:

$$P(X_1, X_2, \dots X_T) = P(X_1)P(X_2|X_1), P(X_3|X_{1:2}) \dots P(X_T|X_{1:T-1}). \quad (15)$$

One simplifying structure we can impose is to assume that only the previous random variable (or the current state) will influence the next random variable. This is often called the Markov assumption:

$$P(X_1, X_2, \dots X_T) = P(X_1)P(X_2|X_1), P(X_3|X_2) \dots P(X_T|X_{T-1}). \quad (16)$$

We can generalize this assumption to include the k previous random variables.

Another common set of assumptions involve the stationarity of the system. The most restrictive is strong / strict stationarity.

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 strongly stationary if the two time series have the same joint cumulative distribution for all t, h, k. Any statistic of the stationary series will be independent of the time interval.*

While this is a powerful assumption, we will rarely want to be so restrictive. Instead, we will often only assume weak stationarity[2]

Definition 3 *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.*

Weak stationarity significantly reduces the parameters we need to describe our system. So long as the variance is finite, strong stationarity yields weak stationarity. In cases where we assume that the underlying stochastic

[2]The exact definition of weak stationarity can vary depending on the source. This is the one we will employ in this course.

process is Gaussian, weak stationarity is equivalent to strong stationarity. **Throughout this course, if we say a process is 'stationary' we are referring to weak stationary.**

Definition 4 *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.*

With trend stationary processes we will often attempt to partition the data into a time-dependent mean and a weak stationary process.

4 Bayesian Inference and Graphical Models

Before we introduce our Bayesian inference and graphical modeling concepts, it's worth highlighting why these models are still valuable in the deep learning era.

- Simpler, probabilistic models can excel in the small data regime.
- They offer interpretability to the results – can be a useful check even when deep learning models are appropriate.
- Helps us quantify our domain knowledge through priors and causal structure.
- Gives us an explicit understanding of the uncertainty of our predictions.
- Can be complementary to deep learning models, either by inspiring architectures, using deep learning for parameter estimation, or making the deep learning method itself Bayesian.

In the context of time series, we can break our use of the Bayesian framework into three buckets:

- **Representation:** How we represent our prior knowledge of the problem using probabilistic language.

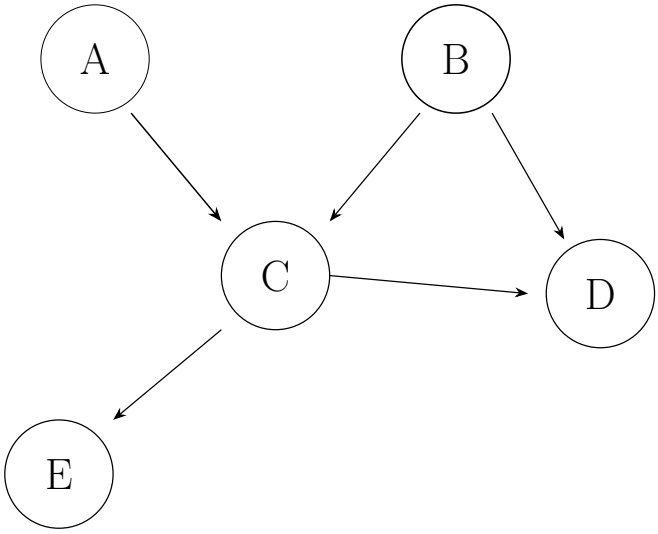

Figure 1: An example of a graphical model. This graphical model is a directed acyclic graph.

- **Learning:** How we learn update our knowledge of the structure of the problem given our data (often in the context of model parameters).
- **Inference:** How we use our model to make predictions about information we don't have (future / missing data / underlying representations).

The second and third buckets can both be considered applications of Bayesian inference.

One of the best ways to represent our knowledge about the structure of the problem is through the use of graphical models. Figure 1 shows an example of a probabilistic graphical model. This particular graph is a directed acyclic graph because all the relationship between random variables is directional and there are no cycles in the graph. Imagine that we now have a time series that goes $\{A, B, C, D, E\}$. This graphical representation of the structure of our time series allows us to factorize our joint distribution function. If we take the example in Figure 1 we can write:

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|C, B)P(E|C) \tag{17}$$

This factorization stems from a series of structural assumptions about the conditional dependencies in the problem. If we think back to the example

of binary variables, we can see how the number of parameters required to describe the joint distribution has been significantly reduced. Note that this is true for both the cdf and pdf.

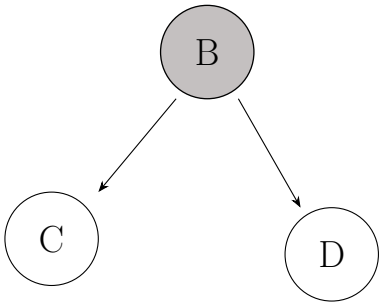

Figure 2: A graphical model where the 'parent' of two nodes is conditioned on.

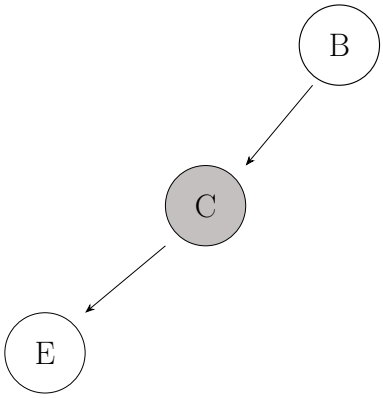

Figure 3: A graphical model where the 'child' of one node and the parent of another is conditioned on.

With our graphical model in hand, we can introduce two relevant concepts: d-separability and the Markov blanket:

Definition 5 *A set of random variables $\mathcal{V}$ is said to d-seperate the set of variables $\mathcal{X}$ from $\mathcal{Y}$ if every path between $\mathcal{X}$ and $\mathcal{Y}$ is blocked by $\mathcal{V}$. A path is blocked by $\mathcal{V}$ if there exists a node W along the path for which one of the two following conditions are met:*

- *W has converging arrows along the path and neither W nor its descendant are contained in $\mathcal{V}$.*

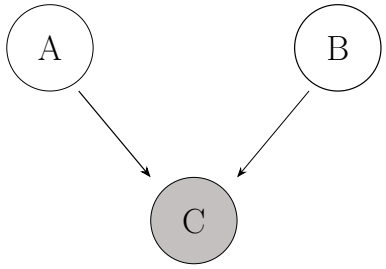

Figure 4: A graphical model where the 'child' of two nodes is conditioned on.

- *W does not have converging arrows and it is contained in $\mathcal{V}$.*

Definition 6 *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.*

We will not prove the definitions here, but we will highlight their importance with three examples.

The first example is conditioning on the parent of two nodes. We show this case in Figure 2. Before we condition on B, the nodes C and D are dependent on each other because they share a parent B. However, when we condition on B we break that dependence. If we know the value of B, then the value of D is no longer informative to the value of C. Written out mathematically:

$$P(C, D|B) = \frac{p(B, C, D)}{p(B)} \tag{18}$$

$$= \frac{p(B)p(C|B)p(D|B)}{p(B)} \tag{19}$$

$$= p(C|B)p(D|B). \tag{20}$$

Another example is conditioning on the child of one node and the parent of the next. We show an example in Figure 3. Before conditioning on C, E is dependent on B through C. If the value of B changes, it will affect the value of C, which will affect the value of E. Once we condition on C, that dependency is broken and E becomes independent of B.

The last example is the child of two parents. An example is shown in Figure 4. Before conditioning on C, A and B are independent of each other. However, once we have conditioned on C, B and A become dependent on

each other. To help build intuition for this property, imagine that C is positively correlated with A and B. That is as A or B goes up, so does C. Now imagine I have observed that C is higher than average (i.e. I have conditioned on a high C). Now knowing something about A/B tells me something about B/A. For example, if A is low then B must be particularly high so that the resulting C is high. This is a fairly qualitative example, but we will work through much more quantitative examples later in the course.