STAT 3503/8109 Lecture 7 Notes

Edoardo Airoldi Scribe: Srikar Katta*

Fall 2020: October 12, 2020

Introduction

Today, we will discuss the latent space model a little further, a time series model - the state space model - which will use the data generating process. The data generating process is the integration of the model statement that would allow us to algorithmically recreate our dataset. If you are not familiar with calculus, this is the part of the course material that you must master. Even data science managers who may not be fluent in statistics understand modeling and how to fit models to data. As a manager, while you may not need to know the exact calculations motivating data sciene, it is essential that you understand what the construction of the models, namely what the models is, the quantities involved in the model to help make decisions about what data to make available to the engineers or scientists and which may not be as necessary to help mitigate costs associated with data collection. So, we need to be extremely familiar with the

- quantities involved: constants, variables, and range of possible values these quantities can adopt
- data statement: observed and unobserved quantities, which is dependent on what can be realistically measured/recorded

Latent Space Models

Last week, we discussed latent space models - a technique that allows you to model the probability of a relationship between two people given their distance, which could be physical/geographic or similarity in behavior or similarity in demographics or anything else that can be compared via similarity/dissimilarity. We utilize this assumption (i.e., if two people are close together they are more likely to know one another) to rebuild social networks. For example, take the following as your latent space and values on this space:

Based on how close these two people, we can have an estimate of how likely they are to have some sort of relationship. Mathematically it can be expressed as follows. Suppose we have n people, then we can have two counters - $i = 1...n, j = 1...n, j > i$ - which simply means that one counter i exists for all of our people and the other counter j goes through all people that come after i. Now, we can represent the probability of two people being connected using log-odds of the probability that two people are connected (please review last week's lectures for an in depth introduction to log-odds). So, $\eta_{ij} = logodds(y_{ij}|x_1, x_2, y_1, y_2, \alpha, \beta)$, where

*Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

$x1, x2$ are the locations for person 1 and 2 on the horizontal axis and y_1, y_2 are the locations for person 1 and 2 on the vertical axis and α, β are unknown constants that allow us to measure the probability of two people being connected:

$$\begin{aligned}\eta_{ij} &= logodds(y_{ij} = 1|x_1, x_2, y_1, y_2, \alpha, \beta) \\ &log\left(\frac{\mathbb{P}(y_{ij} = 1|x_1, x_2, y_1, y_2, \alpha, \beta)}{\mathbb{P}(y_{ij} = 0|x_1, x_2, y_1, y_2, \alpha, \beta)}\right) \\ &= \alpha - \beta\left|\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right|.\end{aligned}$$

Here α just represents a baseline probability for two people being connected. So if the two people have the same location (i.e., $x_1 = x_2, y_1 = y_2$), then the probability that they are related is α. Now, β is a scaling factor that allows the distance between $(x_1, y_1), (x_2, y_2)$ to be represented as values between 0 and α, allowing the distance to be represented as probabilities.

Example 0.1. *Suppose* $\begin{pmatrix} x_i \\ y_i \end{pmatrix} \in \mathbb{R}^4 \to z_i$. *So, the two dimensional location of person i and the movement from the earlier time period (0 in the* x_1 *direction and 0 in the* x_2 *direction) is represented as a vector, which will be denoted as* z_i *and does not follow some sort of distribution. Also, the probability of two people being connected, which is dependent on some baseline probability,* α*, and some scaling factor for our distances,* β*. We will represent this as a log-likelihood:*

$$\begin{aligned}\eta_{ij} &= log\left(\frac{\mathbb{P}(y_{ij} = 1|z_1...z_n, \alpha, \beta)}{\mathbb{P}(y_{ij} = 0|z_1...z_n, \alpha, \beta)}\right) \\ &= \alpha - \beta|z_i - z_j|,\end{aligned}$$

where we are given the following information: $y_{12}...y_{(n-1)n}$ *but not* $z_1...z_n, \alpha, \beta$*. Assume* $y_{ij} \overset{\text{iid}}{\sim} Bernoulli(p)$*, where p is the probability of success. What is the log-likelihood?*

Fist, we need to find the 2x2 table and classify our quantities.

	Observed	**Unobserved**
Variable	$y_{12}...y_{(n-1)n}$	NA
Constant	NA	$z_1...z_n, \alpha, \beta$

So,

$$\begin{aligned}likelihood &= \mathbb{P}(observed\ rv|constants) \\ &= \mathbb{P}(y_{12}...y_{(n-1)n}|z_1...z_n, \alpha, \beta) \\ &= \prod_{1\le i<j\le n} \mathbb{P}(y_{ij}|\alpha, \beta, z_i, z_j) \\ &= \prod_{1\le i<j\le n} [\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)]^{y_{ij}} [\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)]^{1-y_{ij}}.\end{aligned}$$

Notie, if person i and j are connected, then $y_{ij} = 1$, so the probability of this happening would be $\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j) = \mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)^{y_{ij}}$. But if they are not connected, then $y_{ij} = 0$, and the probability of this happening would have been $\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)^{1-y_{ij}}$. So, using those exponents, we can represent $\mathbb{P}(y_{ij}|\alpha, \beta, z_i, z_j))$ as

$$[\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)]^{y_{ij}} [\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)]^{1-y_{ij}} .$$

To review this material, please revisit lecture notes from the beginning of the semester. So, the likelihood would be

$$likelihood = \prod_{1\leq i<j\leq n} [\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)]^{y_{ij}} [\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)]^{1-y_{ij}} .$$

Then, the log-likelihood would be

$$\begin{aligned} log(likelihood) &= log\left(\prod_{1\leq i<j\leq n} [\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)]^{y_{ij}} [\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)]^{1-y_{ij}}\right) \\ &= \sum_{1\leq i<j\leq n} log\left([\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)]^{y_{ij}} [\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)]^{1-y_{ij}}\right) \\ &= \sum_{1\leq i<j\leq n} [y_{ij} log\mathbb{P}(y_{ij} = 1|\alpha, \beta, z_i, z_j)] + [(1 - y_{ij})log\mathbb{P}(y_{ij} = 0|\alpha, \beta, z_i, z_j)] . \end{aligned}$$

Now, when we go through the calculations, we will eventually end up with

$$logLikelihood = \sum_{1\leq i<j\leq n} [\eta_{ij} y_{ij} - log(1 + e^{\eta_{ij}})] .$$

Example 0.2. *Suppose we have n people, $i = 1...n$. Now, we are going to have our distance distributed normally, so essentially people are on a line: $z_i \sim Normal(\mu, \sigma^2)$. And then, it will be our usual representation of this as log-odds ratios:*

$$\begin{aligned} \eta_{ij} &= log\left(\frac{\mathbb{P}(y_{ij} = 1|z_1...z_n, \alpha, \beta)}{\mathbb{P}(y_{ij} = 0|z_1...z_n, \alpha, \beta)}\right) \\ &= \alpha - \beta|z_i - z_j|, \end{aligned}$$

where $y_{ij} = 1$ represents a relationship between two people, α is some baseline probability of people being connected and β is some scaling factor for the distance. Suppose y_{ij} follows a Bernoulli distribution. The probability of success would be $\frac{1}{1+e^{-\alpha+\beta|z_i-z_j|}}$ (which you can rederive yourself using the relationship between y_{ij} and $\alpha-\beta|z_i-z_j|$ if you so choose. Also, y_{ij} is given but nothing is else is observed. Find the log-likelihood.

Notice, in this problem, z_i is now distributed normally, whereas it was constant in example 0.1. So, instead of having to estimate n unknowns about their locations, we are now only having to estimate two - μ, σ^2, a major shift[1]. First, let us create our 2x2 table for our quantities: $y_{ij}, \alpha, \beta, z_i, z_j, \mu, \sigma^2$. First, notice that there are probabilities associated with y_{ij}, suggesting that it follows some theoretical distribution. Also, this is the only information we are given, so it is an observed variable. Also, $z_1...z_n$ are normally distributed but not given, so they are latent random variables. And $\alpha, \beta, \mu, \sigma^2$ are all constants but not given, so they would be unknown constants. Using this, we can create our 2x2 table:

	Observed	**Unobserved**
Variable	$y_{12}...y_{(n-1)n}$	$z_1...z_n$
Constant	NA	$\alpha, \beta, \mu, \sigma^2$

.

So, the complete likelihood for this model would look as follows:

$$\begin{aligned}
completeLikelihood &= \mathbb{P}(y_{12}...y_{n(n-1)}, z_1...z_n|\alpha, \beta, \mu, \sigma^2) \\
&= \prod_{1\leq i<j\leq n} \mathbb{P}(y_{ij}, z_i, z_j|\alpha, \beta, \mu, \sigma^2) \\
&= \prod_{1\leq i<j\leq n} \mathbb{P}(y_{ij}|\alpha, \beta, z_i, z_j)\mathbb{P}(z_i, z_j|\mu, \sigma^2), \text{ by Bayes' Rule} \\
&= \prod_{1\leq i<j\leq n} \mathbb{P}(y_{ij}|\alpha, \beta, \mu, z_i, z_j)\mathbb{P}(z_i|\mu, \sigma^2)\mathbb{P}(z_j|\mu, \sigma^2), \text{ because } z_i \text{ independent of } z_j.
\end{aligned}$$

So, then the proper likelihood would be

$$\begin{aligned}
likelihood &= \int_{latent\ rv} completeLikelihoodd(latent_rv) \\
&= \int_{z_1} ... \int_{z_n} \prod_{1\leq i<j\leq n} \mathbb{P}(y_{ij}|\alpha, \beta, \mu, \sigma^2, z_i, z_j)\mathbb{P}(z_i|\mu, \sigma^2)\mathbb{P}(z_j|\mu, \sigma^2)dz_n...dz_1 \\
&= \int_{z_1} ... \int_{z_n} \prod_{1\leq i<j\leq n} Bernoulli(y_{ij}|\alpha, \beta, z_i, z_j)Normal(z_i|\mu, \sigma^2)Normal(z_j|\mu, \sigma^2)dz_n...dz_1.
\end{aligned}$$

Now, we can find the log-likelihood($logL$):

[1]This is a simplified example of what might happen in the real world, where we would not assume the same distribution for every individual. Rather, we would assume subsets of our population follow their own distributions

$$
\begin{aligned}
logL &= log \int_{z_1} ... \int_{z_n} \prod_{1 \le i < j \le n} Bernoulli(y_{ij}|\alpha, \beta, z_i, z_j) Normal(z_i|\mu, \sigma^2) Normal(z_j|\mu, \sigma^2) dz_n ... dz_1 \\
&= \int_{z_1} ... \int_{z_n} log \left(\prod_{1 \le i < j \le n} Bernoulli(y_{ij}|\alpha, \beta, z_i, z_j) Normal(z_i|\mu, \sigma^2) Normal(z_j|\mu, \sigma^2) \right) dz_n ... dz_1 \\
&= \int_{z_1} ... \int_{z_n} \sum_{1 \le i < j \le n} log \left(Bernoulli(y_{ij}|\alpha, \beta, z_i, z_j) Normal(z_i|\mu, \sigma^2) Normal(z_j|\mu, \sigma^2) \right) dz_n ... dz_1 \\
&= \sum_{1 \le i < j \le n} \int_{z_1} ... \int_{z_n} logBern(y_{ij}|\alpha, \beta, z_i, z_j) + logNormal(z_i|\mu, \sigma^2) + logNormal(z_j|\mu, \sigma^2) dz_n ... dz_1.
\end{aligned}
$$

This last line of the integration is all that is required. Performing the extra steps is just all computation that a computer program can handle for you.

State Space Model

We just finished discussing techniques for network analysis. We will now consider a model for time series, namely the state space model. First, let us review some basic physics/calculus that can help motivate the intuition for this technique. Then, we will explain the model thoroughly with an example.

Time, Position, and Derivatives

Suppose you have some variable, call it a, that moves and has a position associated with it at each time point. So, we can plot x on a plot where time is on the x-axis and position is on the y-axis.

We can also represent this as a function $f(t) = x$, where t is time and x is the position of a at time t. From this function, we can also calculate how much a's position changed over time. That is, what was the rate of change in the a's position, moving from time t_0 to t_1. Notice, this is a rate of change. So, we can find the instantaneous rate of change in a's position, which is known as the derivative, denoted as $\frac{d\ position}{dt}$. Now, since we found the instantaneous rate of change, we can model a's position over time as a recursive function. Let x_t represent the position of a at time t and x_{t-1} represent the position of a at time $t-1$. Then, the change in position from x_{t-1} to x_t would be $\frac{dx_{t-1}}{dt}(dt) = dx_{t-1}$. So, if we start at time $t = 1$ at position x_1, then

$$x_t = x_{t-1} + dx_{t-1}.$$

Location Tracking

If you go to your smartphone and open up a GPS app of some sort, you would notice an estimate of your location, but this is not a specific measurement. It simply approximates your current geographic location. As

you move over time - perhaps when driving or walking - your phone would record these changes, but it is still an estimate of your location. As data scientists, our goal is to estimate the user's location as best as possible. One method, would be to track the location at the smallest time interval possible - perhaps ever nanosecond. But this would take up a lot of the phone's energy. Instead, perhaps it would be wiser to take a reading every second or two seconds instead and interpolate/estimate the rest of the movement in between, which is the goal of the state space model.

Graphical Representation of State Space Models

Let x_t denote someone's location at time t and y_t be our estimate of someone's location at time t. Suppose we are given x_0, that person's location at time 0 - the starting time. Then, from time 0 to time 1, x would evolve: $x_0 \rightarrow x_1$. However, our measurements do not show us x_1, rather an estimate for x_1, which we refer to as y_1. Now, x_1 will evolve to x_2, but our measurement will demonstrate how y_1 might evolve to y_2. Again, x_2 will evolve to x_3, but our measurement will yield the change from y_2 to y_3. We will do this for T time periods. So, in other words, $x_1...x_T$ are all unknown/latent variables, which are referred to as *states* in time series, while $y_1...y_T$ are all observed variables. In a state space model, $x_1...x_T$ evolves over time and follows the recursive algorithm we defined before:

$$x_t = x_{t-1} + dx_{t-1},$$

where x_t is the position at time t, x_{t-1} is the position at time $t-1$, and dx_{t-1} is change in position from time $t-1$ to time t.

Data Generating Process

Data scientists many times want to model the procedure by which data is created in the real world. For example, if we were trying to predict house prices, we are trying to recreate a model based on the features associated with house and prices and the values those features can take on. In essence, the data generating process is very similar to the model statement, except instead of just stating the distributions of quantities, we are explicitly stating the process by which certain quantities were synthesized together to create our variable of interest. For instance, how were house size and location and all the other features associated with house price combined to spit out the price of homes. Understanding this would enable us to algorithmically recreate the data we are interested in. So, in summary, the data generating process simply extends the model statement to depict not only which quantities are involved and their distributions, but also how those quantities are combined.

Data Generating Process of State Space Models

Here, We will create the data generating process for state space models.

Assume we have T total time periods. Every position is associated with an x-coordinate, denoted by x^1,

and a y-coordinate, denoted by x^2. In each time period, the position on the x axis is equal to the previous time period's position on the x axis plus the velocity and some random error: $x_t^1 = x_{t-1}^1 + dx_{t-1}^1 + w_t^{(x^1)}$. Also, the position on the y axis is equal to the position on the y axis in the last time period plus the velocity plus some random error: $x_t^2 = x_{t-1}^2 + dx_{t-1}^2 + w_t^{(x^2)}$. Also, we assume that the velocity dx_t^1 is equal to the velocity at the previous time period plus some random error $w_t^{(dx^1)}$, so $dx_t^1 = dx_{t-1}^1 + w_t^{(dx^1)}$. Now, we can also represent this in vector format:

$$\begin{pmatrix} x_t^1 \\ x_t^2 \\ dx_t^1 \\ dx_t^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1}^1 \\ x_{t-1}^2 \\ dx_{t-1}^1 \\ dx_{t-1}^2 \end{pmatrix} + \begin{pmatrix} w_t^{(x^2)} \\ w_t^{(x^2)} \\ w_t^{(dx^1)} \\ w_t^{(dx^2)} \end{pmatrix}, \text{ so}$$

$$\begin{aligned} x_t^1 &= x_{t-1}^1 + dx_{t-1}^1 w_t^{(x^1)}, \\ x_t^2 &= x_{t-1}^2 + dx_{t-1}^2 + w_t^{(x^2)}, \\ dx_t^1 &= dx_{t-1}^1 + w_t^{(dx^1)}, \\ dx_t^2 &= dx_{t-1}^2 + w_t^{(dx^2)}. \end{aligned}$$

Essentially, this matrix notation[2] allows us to compactly write everything about the evolution of x_t^1, x_t^2, dx_t^1, and dx_t^2 without writing out each line. And in order for this recursive algorithm to work, we need to define some starting position, x_0, so we must define this in our model.

Now, recall that in the state space model, x_t is unobserved. However, we do observe y_t, an estimate for x_t. Here, we do not define any velocity associated with y_t, so y_t is a two-dimensional vector: $y_t = \begin{pmatrix} y_t^1 \\ y_t^2 \end{pmatrix}$, where y_t^1 is the estimate of x_t on the horizontal axis and y_t^2 is the estimate of x_t on the vertical axis. Saying y_t^1 is an estimate for x_t^1 is essentially the same as writing $y_t^1 = x_t^1 + v_t^1$, where v_t^1 is some measurement error. And saying y_t^2 is an estimate for x_t^2 is essentially the same as writing $y_t^2 = x_t^2 + v_t^2$, where v_t^2 is some again measurement error but now for the vertical axis. Then, we can express the estimate of y_t in matrix notation as

$$\begin{pmatrix} y_t^1 \\ y_t^2 \end{pmatrix} = \begin{pmatrix} x_t^1 \\ x_t^2 \end{pmatrix} + \begin{pmatrix} v_t^1 \\ v_t^2 \end{pmatrix}.$$

However, recall that x_t is a 4-dimensional column vector: $x_t = \begin{pmatrix} x_t^1 \\ x_t^2 \end{pmatrix}$, so we must convert x_t into just $\begin{pmatrix} x_t^1 \\ x_t^2 \end{pmatrix}$. We can do this using matrix notation:

[2]If you are unfamiliar with matrix notation, please review this Wiki article

$$\begin{pmatrix} x_t^1 \\ x_t^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_t^1 \\ x_t^2 \\ dx_t^1 \\ dx_t^2 \end{pmatrix}$$
$$= Ax_t,$$

where A is that matrix of 1s and 0s that allows us to isolate just x_t^1 and x_t^2. So, we can rewrite our matrix representation of our estimate of x_t as

$$y_t = Ax_t + v_t, \text{ where}$$
$$y_t = \begin{pmatrix} y_t^1 \\ y_t^2 \end{pmatrix}, A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, x_t = \begin{pmatrix} x_t^1 \\ x_t^2 \\ dx_t^1 \\ dx_t^2 \end{pmatrix}, v_t = \begin{pmatrix} v_t^1 \\ v_t^2 \end{pmatrix}.$$

This is the algorithm that allows us to model a person's movement over time. In other words, this is a representation of the data generating process for someone's movement. In order to complete this, the data scientist simply needs to understand what distributions these quantities - x_0, w_t, v_t - follow. We only need these three because the matrixes F and A will always be fixed, $x_t, t = 1...T$ comes from x_0 and w_t over time since it is recursively defined, so these quantities *need* to have distributions that are related to one another, and y_t is an estimate of x_t, so it needs to have a distribution that is related to x_0 and v_t. All of this will be given in the model statement. Let us try an example.

Example 0.3. *Suppose you are trying to create an app that tracks your dog's position over time using its collar. This is a perfect application of the state space model. Your manager offers you the following information:*

$x_0 \sim Normal(m, v), x_0 \in \mathbb{R}^4$, *where m is a 4-dim matrix of averages and v is a cov matrix*

for $t = 1...T$

$$x_t = Fx_{t-1} + w_t,$$
$$y_t = Ax_t + v_t,$$
$$w_t \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, Q\right), \text{ where } Q \text{ is the covariance matrix,}$$
$$v_t \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, R\right), \text{ where } R \text{ is the covariance matrix.}$$

Find the likelihood.

Notice, in the problem, the known and unknown quantities were never mentioned. That is because these are implied in the state space model: the only values that are known are $y_1...y_T, F, A$ - everything else is

unknown. So the 2x2 table would read as follows:

	Observed	**Unobserved**
Variable	$y_1....y_T$	$x_0, x_1...x_T, v_t...v_t, w_1...w_T$
Constant	F, A	m, v, Q, R

Now, because we have latent variables, we must first find the complete likelihood (CL) and integrate out the unknowns to find the proper likelihood:

$$\begin{aligned} CL &= \mathbb{P}(y_1...y_T, x_0, x_1...x_T, v_1...v_T, w_1...w_T|Q, R, m, v, F, A) \\ &= Normal(x_0|m, v)\prod_{t=1}^{T} Normal(w_t|Q)Normal(x_t|Fx_t, w_t)Normal(v_t|R)Normal(y_t|Ax_t, v_t), \end{aligned}$$

which we get from three properties: the sum of normally distributed variables is also normally distributed and an extended version of Bayes Rule and conditional independence. So, the likelihood(L) is just the integral of the complete likelihood over the latent variables:

$$L = \int_{latent\ rv} \mathbb{P}(y_1...y_T, x_0, x_1...x_T, v_1...v_T, w_1...w_T|Q, R, m, v, F, A)dw_T...dw_1dv_T...dv_1dx_T...dx_1dx_0,$$

so

$$L = \int_{latent\ rv} N(x_0|m, v)\prod_{t=1}^{T} N(w_t|Q)N(x_t|Fx_t, w_t)N(v_t|R)N(y_t|Ax_t, v_t)dw_T...dw_1dv_T...dv_1dx_T...dx_1dx_0$$

Conclusion

Today, we discussed the latent space model in further depth and offered another example. We also discussed the data generating process and introduced the state space model. These all depend on the 2x2 table and the likelihood, so thoroughly understanding those concepts is essential. In the future, we will discuss how to use estimation techniques to find unknown constants using these models that we detailed.