Lecture 11: February 7

Lecturer: Donlapark Pornnopparath

We consider the likelihood function in each case. When the coin's values are missing at random, we have

$$L(\phi, \psi|D) = \phi^H(1-\phi)^T\psi^{n(H)+n(T)}(1-\psi)^{n(?)},$$

which yields MLE $\hat{\phi} = H/(H+T)$ and $\hat{\psi} = (H+T)/(H+T+n(?))$. In contrast, when the chance of missing is influenced by the outcome, the likelihood function is

$$L(\phi, \psi|D) = \phi^H(1-\phi)^T\psi_{O_X|H}^{n(H)}\psi_{O_X|T}^{n(T)}(\phi(1-\psi_{O_X|H}) + (1-\phi)(1-\psi_{O_X|T}))^{n(?)},$$

which is much harder to find MLE, as there is no way to decouple parameter ϕ from ψ. One way to guarantee that the parameters decouple is to assume that the missing data is *missing completely at random* (MCAR) where we assume $X \perp\!\!\!\perp O_X$ as in the first example. Nonetheless, this assumption is not necessary for the decomposition of the likelihood function as we shall see in the following example.

Example 11.1. Suppose that we successively toss two coins X_1 and X_2 in a way that the outcome of the first coin X_1 can always be observed while the outcome of the second coin X_2 can be missing depending on X_1. If we look at the corresponding graphical model, we see that $(O_{X_2} \perp\!\!\!\perp X_2|X_1)$. In other words, the outcomes of both coins are independent of whether X_2 is observed or not. We show that the likelihood function can be decomposed by considering all possible scenarios.

- Both coins are observed. For example, if $Y_1 = H$, $Y_2 = T$, then the probability of this observation is
$$P(X_1 = H, X_2 = T, O_{X_1} = o^1, O_{X_2} = o^1) = \phi_{X_1}(1-\phi_{X_2})\psi_{O_{X_2|H}}$$
- Only X_1 is observed. For example, if $Y_1 = H$, $Y_2 =?$, then the probability of this observation is
$$P(X_1 = H, O_{X_1} = o^1, O_{X_2} = o^0) = \phi_{X_1}(1-\psi_{O_{X_2|H}}).$$

If we did not assume the conditional independence, then the terms ϕ_{X_2} and $1-\phi_{X_2}$ would have appeared and entangled this product just like the previous example. Considering the product of all possible cases, we can easily see that all parameters are separated from each others. ◇

We can see that the conditional independence is sufficient to decompose the parameters in the likelihood function. In fact, it is also necessary for such condition to hold.

Definition 11.2. Let X_{obs} be random variables that are completely observed and X_{hid} be the ones that are not. Then we say that the data of these variables are missing at random (MAR) if

$$(O_X \perp\!\!\!\perp X_{hid}|X_{obs})$$

is true for $X = X_{obs}$ and $X = X_{hid}$.

For an example of a situation where the MAR assumption holds, consider a variable X which indicates the record of an X-ray scan for bone fractures of a single patient in a hospital. If the patient did not have any bone fracture, it is likely that X is missing, and so X and O_X are not independent. However, if we added another variable Y which records all tests that have been performed on the patient, then O_X would be independent of X given Y, and O_Y would be independent of Y given X. Thus, we can see that even if MAR does not hold, it might hold on a larger set of variables.

Theorem 11.3. If data with an underlying distribution satisfies MAR, then the likelihood function $L(\phi, \psi|D)$ can be written as a product of two likelihood functions:

$$L(\phi, \psi|D) = L(\phi|D)L(\psi|D)$$

Proof: Suppose that we observe $X_{obs} = x_{obs}$ and $O_X = o_X$. If all variables are observed, then

$$P(X_{obs} = x_{obs}, O_X = O_X) = P(o_X|x_{obs})P(x_{obs})$$

which is clearly a product of functions of ϕ and ψ. On the other hand, if there are some missing variables X_{hidden} then

$$\begin{aligned} P(X_{obs} = x_{obs}, O_X = o_X) &= P(o_X|x_{obs})P(x_{obs}) \\ &= \sum_{x_{hid}} P(o_X|x_{obs}, x_{hid})P(x_{obs}, x_{hid}) \\ &= \sum_{x_{hid}} P(o_X|x_{obs})P(x_{obs}, x_{hid}) \\ &= P(o_X|x_{obs}) \sum_{x_{hid}} P(x_{obs}, x_{hid}) \\ &= P(o_X|x_{obs})P(x_{obs}), \end{aligned}$$

which is again a product of functions of ϕ and ψ. ■

11.1 Likelihood function

In this section we assume the MAR assumption and address some issue regarding the likelihood function. Let $\boldsymbol{O}^(n)$ and $\boldsymbol{H}^n$ be the set of observed and missing variables of the n–th instance, respectively. Then, as in the proof of Theorem 11.3, we compute the likelihood function by marginalizing out the hidden variables

$$L(\boldsymbol{\theta}|D) = \prod_n P(\boldsymbol{o}^{(n)}|\boldsymbol{\theta}) = \prod_n \sum_{\boldsymbol{h}^{(n)}} P(\boldsymbol{o}^{(n)}, \boldsymbol{h}^{(n)}|\boldsymbol{\theta}).$$

However, we can see in previous examples that the parameter $\phi_{\boldsymbol{O}}$ of the observed variables and $\psi_{\boldsymbol{H}}$ of the missing variables are mixed together in the summation, which would make MLE inference intractable. This is different than when complete data are observed, in which case the parameters can be decoupled and the MLE is easy to compute.

Another issue of missing data is that MLE estimator might not be unique, as we will see in the following example.

Example 11.4. Suppose that there are two coins of different size and assume that the experimenter can randomly choose to toss between the two coins. However, due to a miscommunication, the experimenter recorded only the outcome, but not the size of the coins. Let X indicates the outcome and $M \in \{m^1, m^2\}$ indicates which coin was being tossed. Then the parameters consist of $\theta_M, \theta_{X|m^1}$ and $\theta_{X|m^2}$. Since the data is MCAR, we have the following likelihood function

$$L(\boldsymbol{\theta}|D) = P(H)^{n(H)}P(T)^{n(T)}$$

where

$$P(H) = \theta_M\theta_{X|m^1} + (1 - \theta_M)\theta_{X|m^2}, \quad P(T) = 1 - P(H).$$

We notice that there will be infinitely many combination of parameters that maximize L. For example, if $\theta_M = 0.5, \theta_{X|m^1} = 0.5$ and $\theta_{X|m^2} = 0.5$, then so is $\theta_M = 0.5, \theta_{X|m^1} = 0.8$ and $\theta_{X|m^2} = 0.2$. $\diamond$

From this example, we see that we cannot always build a model by just learning from the data. Statisticians thus introduce the notion of *identifiability* to avoid this issue.

Definition 11.5. Suppose that we have a parametric model $P(\boldsymbol{X}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$ over variables $\boldsymbol{X}$. A choice of $\boldsymbol{\theta} \in \Theta$ is identifiable if there is no $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ such that $P(\boldsymbol{X}|\boldsymbol{\theta}') = P(\boldsymbol{X}|\boldsymbol{\theta})$. A model is identifiable if $\boldsymbol{\theta}$ is identifiable for all $\boldsymbol{\theta} \in \Theta$.