Note Set 3: Models, Parameters, and Likelihood

Padhraic Smyth,
Department of Computer Science
University of California, Irvine
January 2024

1 Introduction

This is a brief set of notes covering basic concepts related to likelihood and maximum likelihood. The goal of this set of notes is to connect the types of probability models we have discussed in Notes 1 and 2 to observed data. Essentially this involves two steps:

1. Construct a *generative* or *forward* model M with parameters θ of how data D can be generated. We can think of this generative model as a stochastic simulator for the data, with parameters θ. We will assume for now that M, the structure or functional form of the model, is known, but that the parameters θ are unknown[1]. An example would be that the model M is a Gaussian (Normal) probability density function with unknown parameters $\theta = \{\mu, \sigma^2\}$.

2. Given the generative model for the data we then "work backwards" to make inferences about θ given observed data D. This is the essence of probabilistic learning (and much of statistics): going from observed data to inferences about unknown parameters that we are interested in, via a probabilistic model. In this set of Notes we will focus on so-called **point estimates** of parameters θ, denoted by $\hat{\theta}$. The idea is that this is our best guess, if forced to select a single number, of some true (but unknown) θ.

2 Likelihood

We define likelihood as the probability of observed data D given a model M where the model has parameters θ, i.e.,

$$L(\theta) \; = \; P(D|\theta, M)$$

- Likelihood is always defined relative to some model M. However, for our initial discussions at least, we will often drop the explicit reference to M in discussions below and just implicitly assume that there is some model M being conditioned on.

[1] Later in class we will discuss the situation where there are multiple candidate models $M_1, \ldots, M_K$ under consideration.

- We will refer to data sets as D. For 1-dimensional observations this will be a set of values $\{x_1, \ldots, x_n\}$. For d-dimensional vector observations $\underline{x}$ we have $D = \{\underline{x}_1, \ldots, \underline{x}_n\}$, where x_{ij} is the jth component of the ith observation, $1 \leq j \leq d, 1 \leq i \leq n$. We can also think of D as a data matrix with n rows indexed by i (each row is a data vector $\underline{x}_i$), and with d columns (variables) indexed by j.

- Likelihood is viewed as a function of θ conditioned on a fixed observed data set D. We are interested in how the likelihood changes as θ changes, where θ is usually real-valued. If a parameter θ_1 has higher likelihood $L(\theta_1)$ than the likelihood of another parameter θ_2, then $P(D|\theta_1) > P(D|\theta_2)$, i.e., the observed data is more likely given θ_1 than θ_2.

- This leads naturally to the concept of **maximum likelihood** (discussed below), i.e., finding the θ value that corresponds to the maximum of $L(\theta)$ (assuming a unique maximum exists).

- In defining the likelihood we can drop (ignore) any terms in $p(D|\theta)$ that don't involve θ, such as normalizing constants. What is usually important is the shape of the likelihood function, or the relative value of the likelihood, rather than the actual value of the likelihood.

- The likelihood function will typically be quite "wide" when we have relatively little data, and will "narrow" in shape as we get more data. (This is generally a good description of what happens for simple models, but is not necessarily true for more complex ones).

- The likelihood function can be defined on vectors of parameters $\underline{\theta}$, rather than just a single scalar parameter θ. For a parameter vector defined as $\underline{\theta} = (\theta_1, \ldots, \theta_p)$, $L(\underline{\theta})$ is a scalar function of p arguments. As with a multi-dimensional probability density function, we can think of the multi-dimensional likelihood function as a "surface" (non-negative) defined over p dimensions.

- As an example, for a Gaussian density model $p(x)$ for a one-dimensional continuous random variable X, the parameters are $\underline{\theta} = \{\theta_1, \theta_2\} = \{\mu, \sigma^2\}$, i.e., the unknown mean and variance. The likelihood $L(\underline{\theta}) = L(\mu, \sigma^2)$ is a scalar function over the two-dimensional μ, σ^2 space. Note that we could define θ_2 here as either σ or σ^2—either is fine, but it turns out that σ^2 will make the maximum likelihood analysis somewhat easier to work with. It can also sometimes be convenient to work with reparametrizations such as $\log \sigma$ or $\frac{1}{\sigma^2}$, depending on the context, rather than σ or σ^2 directly.

- The likelihood function can equally well be defined when the probability model is a distribution $P(D|\theta)$ (e.g., for discrete random variables) or a probability density function $p(D|\theta)$ (for continuous random variables), or for a combination of the two (e.g., $p(D_1|D_2, \theta_1)P(D_2|\theta_2)$) where D_1 models the variables that are real-valued using parameters θ_1, and D_2 models the variables that are discrete-valued with parameters θ_2.

Example 1: Binomial Likelihood: Consider tossing a coin with probability θ of heads and $1-\theta$ of tails. This is the **Bernoulli model**. Now say we observe a sequence of tosses of the same coin. This set of outcomes represents our data D, where $D = \{x_1, \ldots, x_i\}$ and $x_i \in \{0, 1\}$ represents the outcome of the ith toss (e.g., with 1 corresponding to head and 0 to tails).
In defining a likelihood, we need to specify a probability model for multiple samples $\{x_1, \ldots, x_i\}$ rather than just for a single sample x_i. The standard assumption for coin-tossing (and many other phenomena that don't exhibit any "memory" in how individual data points are generated) is to assume that each observation x_i is conditionally independent of the other observations given the parameter θ, i.e.,

$$L(\theta) \;=\; P(D|\theta) \;=\; P(x_1, \ldots, x_n|\theta) \;=\; \prod_{i=1}^{n} P(x_i|\theta)$$

where $P(x_i|\theta) = \theta$ for $x_i = 1$ and $P(x_i|\theta) = 1-\theta$ for $x_i = 0$. This particular "coin-tossing" model, combining a Bernoulli with conditional independence of the x_i's is referred to as a **Binomial likelihood**.
The conditional independence assumption on the x_i's in the likelihood definition is sometimes (loosely) also referred to as the **IID assumption** (independent and identically distributed). The notion of exchangeability in statistics is essentially the same idea. Note this assumption allows for a tremendous simplification in our model: instead of dealing with the joint $P(x_1, \ldots, x_n|\theta)$ we can instead work with individual terms $P(x_i|\theta)$. Of course we have to be careful that this is a reasonable assumption. It is certainly a reasonable assumption in the case where the x_i's are coin tosses, or perhaps (and closer to the real-world) the case where X_i represents the ith Web surfer to arrive at an ecommerce Web site and x_i is a binary value indicating whether the Web surfer makes a purchase or not. But in other applications the x_i's may have some dependence on each other, e.g., if the x_i's represented the value of the stock market on different days or words in text. If such dependence was thought to exist then it should be modeled (see example below).

Continuing on with our binomial likelihood example, we can write

$$L(\theta) \;=\; \prod_{i=1}^{n} P(x_i|\theta) \;=\; \theta^r(1-\theta)^{n-r}$$

where r is the number of "heads" observed and $n-r$ is the number of tails. Note that we did not include the usual combinatorial (binomial) term in front of the expression above, i.e., $\binom{n}{r}$ to count the number of different ways that r heads could occur in n trials, since this term does not involve θ.

Figure 1 shows two examples of the binomial likelihood function for different data sets. In Figure 1(a) we have $r = 3$ and $n = 10$. The likelihood function is relatively wide and is maximized at 3/10 = 0.3, which makes sense intuitively. In Figure Figure 1(b) we have $r = 30$ and $n = 100$: here the likelihood is much narrower as we might expect and, as a result, the plausible values for θ are much narrower after seeing 100 observations compared with just 10.

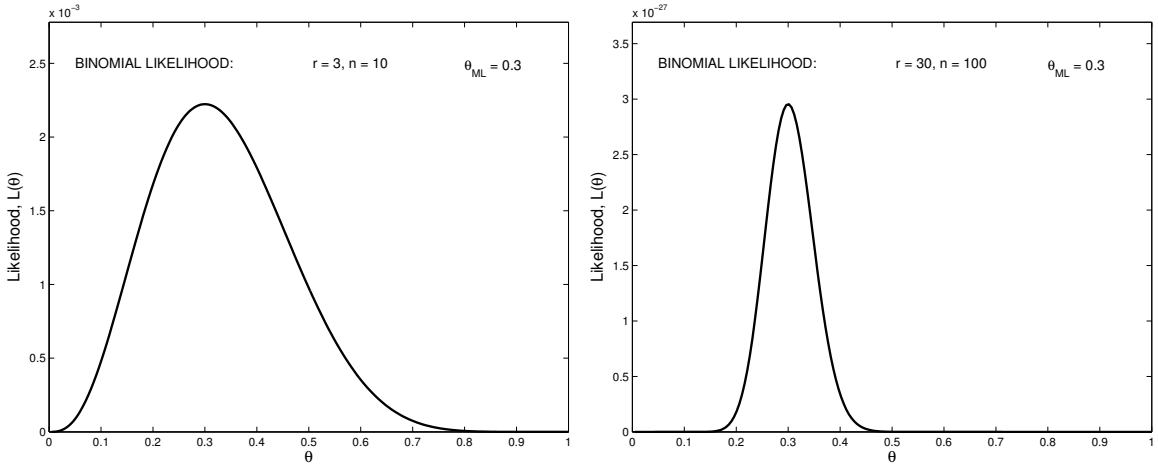

Figure 1: Binomial likelihood for (a) $r = 3, n = 10$, and (b) $r = 30, n = 100$.

An interesting side-note with the example above is that *conditional* independence plays a key role in our definition of likelihood in the binomial model. In fact the x_i's are not marginally independent, but only conditionally independent. Why? If θ is unknown (remember that θ is the probability of heads) then the x_i's carry information about each other. As an example, say $\theta = 0.999$ but we don't know this. So we will tend to see a lot of heads showing up and very rarely a tail showing up. Having seen such a sequence of x_i's with many more heads than tails, this data is informative about the next coin toss. Of course, if someone were to tell us the true value of θ then the previous x_i values have no information at all in terms of predicting the next x value, since we have all the information we need in θ.

Example 2: Likelihood with Memory: In the previous binomial example, if instead of modeling coin tosses we were modeling the occurrence of rain on day i in Irvine (x_i indicates whether it rains or not on day i), then we would want to consider abandoning the IID assumption and introducing some dependence among the x_i values (since we will tend to get "runs" of wet days and dry days). For example, we could make a Markov assumption (Note Set 2) and assume that x_{i+1} on day $i+1$ is conditionally independent of x values on days $i-1, i-2, \ldots, 1$, given the value of x_i. Accordingly the likelihood would be defined as:

$$L(\underline{\theta}) \;=\; P(x_1, \ldots, x_n|\underline{\theta}) = P(x_1|\theta_1) \cdot \prod_{i=1}^{n-1} P(x_{i+1}|x_i, \underline{\theta}_2)$$

where $\theta_1 = p(x_1 = 1)$ and $\underline{\theta}_2$ is a parameter vector representing a 2×2 Markov transition matrix of parameters (the conditional probabilities of rain or not-rain, conditioned on rain or not-rain the day before).

Example 3: Gaussian Likelihood: Consider a data set $D = \{x_1, \ldots, x_n\}$ where the x_i's are real-valued scalars and are samples from a random variable X. Assume we wish to model the x_i values with a Gaussian density function. The Gaussian has two parameters μ and σ^2. Treating these two parameters as unknown, and referring to them as $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ we can write the likelihood as:

$$p(D|\underline{\theta}) \;=\; p(x_1, \ldots, x_n|\underline{\theta}) = \prod_{i=1}^{n} p(x_i|\underline{\theta})$$

where here we make the assumption that the x_i's are conditionally independent given $\underline{\theta}$ (for a real problem we would want to convince ourselves that this is reasonable to do).

The individual terms in our likelihood are by definition Gaussian density functions, each evaluated at x_i:

$$p(x_i|\underline{\theta}) \;=\; p(x_i|\mu, \theta) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} exp^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}.$$

Taking the product of these terms, and then taking the log (to the base e for convenience) we arrive at the log-likelihood

$$\log L(\underline{\theta}) = l(\underline{\theta}) \;=\; -\frac{n}{2}\log(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}\Big(x_i - \theta_1\Big)^2.$$

Imagine that $\theta_2 = \sigma^2$ is fixed (assume for example that it is known). Then $l(\theta_1)$ (viewed as a function of θ_1 only) is proportional to a 2nd order polynomial involving x_i's and θ_1, i.e.,

$$l(\theta_1) \;\propto\; -\sum_{i=1}^{n}(x_i - \theta_1)^2$$

from which we see that $l(\theta_1)$ is larger if $\sum_{i=1}^{n}(x_i - \theta_1)^2$ is smaller, i.e., $l(\theta_1)$ will be larger for values of $\theta_1 = \mu$ that are closer to x_i's on average (since this is a sum of squared errors between the observed set of x_i values and a single scalar $\theta_1 = \mu$).

Figure 1 shows some examples of the Gaussian log-likelihood function $l(\mu)$ (treating μ as unknown, but assuming that σ^2 is known) being plotted for different sized data samples, where the data was simulated from a known Gaussian density function with $\mu = 5$ and $\sigma^2 = 1$. Again as n increases we see that the likelihood begins to narrow in around the true value of $\mu = 5$.

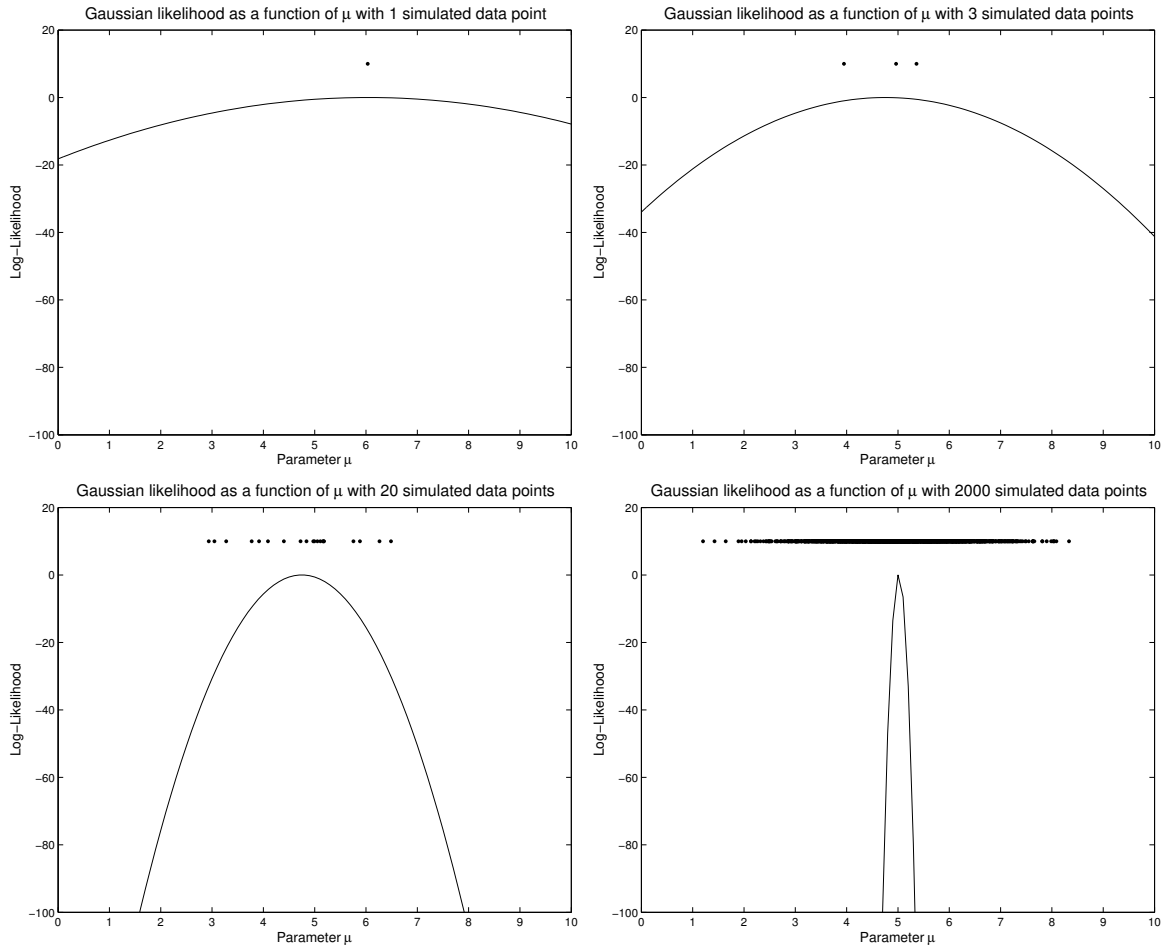

Figure 2: Log-likelihood for 4 different sample sizes, as a function of parameter $\theta = \mu$, with data simulated from a Gaussian with true $\mu = 5$ and $\sigma = 1$ (simulated data shown as dots horizontally at the top of the plot).

3 The Principle of Maximum Likelihood

The principle of maximum likelihood follows naturally from what we have discussed above, namely that if we had to summarize our data by selecting only a single parameter value $\hat{\theta}$, and if we only have the observed data and the likelihood available and no other information, then it is reasonable to argue that the value of θ that we should select is the one that maximizes the likelihood $L(\theta)$. Or, more formally:

$$\hat{\theta}_{ML} \;=\; \arg\max_{\theta} L(\theta) \;=\; \arg\max_{\theta} P(D|\theta).$$

The subscript "ML" denotes "maximum likelihood" since we will later discuss other types of estimates for which we will use other subscripts. The "hat" notation, $\hat{\theta}$, denotes an **estimate** of some unknown (true) quantity θ.

Example 4: Maximum Likelihood Estimate for the Binomial Model: From earlier, the binomial likelihood can be written as:

$$L(\underline{\theta}) \;=\; P(x_1, \ldots, x_n|\underline{\theta}) = P(x_1|\theta_1) \cdot \prod_{i=1}^{n-1} P(x_{i+1}|x_i, \underline{\theta}_2)$$

where r is the number of successes in n trials. We can easily find the maximum likelihood estimate of θ as follows. First lets work with the log-likelihood since the log-likelihood is a little easier to work with[a] .

$$\log L(\theta) \;=\; l(\theta) \;=\; r \log \theta + (n - r) \log(1 - \theta).$$

A necessary condition to maximize $l(\theta)$ is that $\frac{d}{d\theta} l(\theta) = 0$, i.e., this condition must be satisfied at $\theta = \hat{\theta}_{ML}$. Thus, we calculate the derivative with respect to θ and set to 0, i.e.,

$$\frac{d}{d\theta} l(\theta) \;=\; \frac{r}{\theta} - \frac{n - r}{1 - \theta} = 0, \qquad \text{at } \theta = \hat{\theta}_{ML}$$

and after some rearrangement of terms we get

$$\hat{\theta}_{ML} \;=\; \frac{r}{n}$$

i.e., the standard intuitive frequency-based estimate for the probability of success given r successes in n trials. At this point it seems like we may not have gained very much with our likelihood-based framework since we arrived back at the "obvious" answer! However, the power of the likelihood (and related) approaches is that we can generalize to much more complex problems where there is no obvious "intuitive" estimator for a parameter θ. And if we think about it we should have expected to get this estimate for $\hat{\theta}_{ML}$ a priori. Had we gotten any other estimate we might have good cause for concern that our likelihood-based procedures did not match our intuition.

[a] Note that the value of θ that maximizes the log-likelihood is the same as the value of θ that maximizes the likelihood since log is a monotonic function.

Example 5: Maximum Likelihood Estimate for the Gaussian IID Model:
Consider the case where σ is known and μ is unknown. From Example 3 earlier we saw that for the Gaussian IID model we can write:

$$l(\theta) \;=\; -\sum_{i=1}^{n}(x_i-\theta)^2$$

where $\theta = \mu$ the unknown mean parameter. To maximize this as a function of θ we can use simple calculus, i.e., differentiate the right-hand side above with respect to θ, set to 0, and solve for θ. (Left as an exercise for the reader).

Example 6: Maximum Likelihood Estimation with Two Noisy Data Sources:
There are many problems in scientific data analysis where we need to combine multiple different data sets to make predictions about a single quantity of interest. The following example discusses such a problem and also illustrates a situation where the maximum likelihood approach leads to an estimate that is not obvious, i.e., the equation defining $\hat{\theta}_{ML}$ could not easily be guessed, at least not until we have an idea what the correct approach is.

Consider the following scenario. We are working with an astronomer monitoring a distant object in the sky with two different CCD cameras connected to 2 different telescopes in different parts of the world. Assume in this simplified example that each camera produces noisy estimates of the object's true brightness—we assume that there is a true constant brightness μ for the object but our cameras only get noisy measurements $x_1, x_2, \ldots$ (our astronomer can get multiple x_i measurements from each camera over multiple nights).

Say that camera 1 produces measurements that have a Gaussian distribution with mean μ and variance σ_1^2, and that camera 2 produces measurements with mean μ and variance σ_2^2. We are assuming that the true mean of the measurements from each individual camera is the same as the true brightness, but the variances are different, e.g., if σ_1^2 is much smaller than σ_2^2 this could be because camera 1 is connected to a much more accurate (newer, stronger) telescope. We will also assume (for simplicity) that the two variances are known (but that μ is unknown)—which is not unreasonable, since astronomers are often very good at coming up with techniques to calibrate the noise in their instruments.

The question is how to estimate μ given data D consisting of n_1 measurements from camera 1 and n_2 measurements from camera 2. A naive estimate of μ is simply the average over all of the measurements, i.e.,

$$\hat{\mu}_{naive} = \frac{1}{n_1+n_2}\sum_i x_i$$

where the sum ranges over all of the of measurements. But in constructing this simple estimate we are ignoring the fact that one camera is more accurate than the other, i.e., $\sigma_1^2 \neq \sigma_2^2$. The more different these two variances are, the more important it may be to account for measurements from the two data sets differently. In the extreme case, for example, we might have only 1 measurement in D_1 from camera 1 and (say) 1000 measurements from D_2 from camera 2, but say that camera 1 has 10 times less variance than camera 2. In

this case how should we combine the data to arrive at an estimate of μ? Intuitively we can imagine that some form of weighting scheme is probably appropriate, where we downweight measurements from the more noisy camera and upweight measurements from the more accurate one. But its not obvious what these weights should be.

This is the type of situation where formal probabilistic modeling (such as likelihood based methods) can be very useful. So lets see what the maximum likelihood estimator for μ is in this situation.

$$\begin{aligned} L(\mu) &= p(D|\mu) \\ &= p(D_1, D_2|\mu) \\ &= p(D_1|\mu)p(D_2|\mu) \\ &= \prod_{i=1}^{n_1} f\big(x_i; \mu, \sigma_1^2\big) \cdot \prod_{j=1}^{n_2} f\big(x_j; \mu, \sigma_2^2\big) \end{aligned}$$

where the first product is over the n_1 data points in data set D_1 and the second product is over the n_2 data points in data set D_2. The notation $f(x_i; \mu, \sigma_1^2)$ denotes a Gaussian (Normal) density function evaluated at x_i with mean μ and variance σ_1^2. We have also assumed IID measurements, which may be reasonable for example if the measurements were taken relatively far apart in time (e.g., on different nights). Taking logs and dropping terms that don't involve μ, we get

$$l(\mu) \;=\; -\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}\Big(x_i - \mu\Big)^2 - \frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}\Big(x_j - \mu\Big)^2.$$

Taking the derivative with respect to μ yields

$$\frac{d}{d\mu}l(\mu) \;=\; \frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(x_i - \mu) \;+\; \frac{1}{\sigma_2^2}\sum_{j=1}^{n_2}(x_j - \mu).$$

Setting this expression to 0, and rearranging terms we get that

$$\hat{\mu}_{ML}\left(\frac{n_1}{\sigma_1^2} \;+\; \frac{n_2}{\sigma_2^2}\right) \;=\; \frac{1}{\sigma_1^2}\sum_{i=1}^{n_1} x_i \;+\; \frac{1}{\sigma_2^2}\sum_{j=1}^{n_2} x_j.$$

Multiplying through by σ_1^2,

$$\hat{\mu}_{ML}\left(n_1 \;+\; n_2\frac{\sigma_1^2}{\sigma_2^2}\right) \;=\; \sum_{i=1}^{n_1} x_i \;+\; \frac{\sigma_1^2}{\sigma_2^2}\sum_{j=1}^{n_2} x_j,$$

yielding:

$$\hat{\mu}_{ML} \;=\; \left(n_1 \;+\; n_2\frac{\sigma_1^2}{\sigma_2^2}\right)^{-1}\left[\sum_{i=1}^{n_1} x_i \;+\; \frac{\sigma_1^2}{\sigma_2^2}\sum_{j=1}^{n_2} x_j\right].$$

We see that the relative weighting of the two data sets is controlled by the ratio $r = \frac{\sigma_1^2}{\sigma_2^2}$. If $r = 1$ (same variance in both cameras) we get the standard "unweighted" solution, i.e., the maximum likelihood estimate

of μ corresponds to the empirical average of all of the data points (as we would expect). If $\sigma_1^2 < \sigma_2^2$ (so the ratio $r < 1$) then the data points from camera 2 (with higher variance and more noise) are essentially being downweighted by a factor of $r = \frac{\sigma_1^2}{\sigma_2^2}$. Conversely, if $\sigma_1^2 > \sigma_2^2$ and the measurements from camera 2 are less noisy, then camera 2's measurements are upweighted by the factor $r > 1$.

We might have guessed at a similar solution in an ad hoc manner—but the likelihood-based approach provides a clear and principled way to derive estimators, and can be particularly useful in problems that are often much more complex than this example. For example, imagine K cameras, with different (possibly non-Gaussian) noise models for each and with various dependencies among the cameras. The noise characteristics for some cameras could be unknown but nonetheless may be known to be inter-dependent in some manner, e.g., two cameras have unknown variances but we know that the first camera has twice the variance of the other. Maximum likelihood gives us a principled way to address such problems.

4 Maximum Likelihood for Graphical Models

4.1 Basic Concepts: Two Random Variables

Consider two discrete random variables A and B each taking M values, with possible values a and b. Assume we already know the marginal probabilities $p(a)$ for variable A and we wish to learn the conditional probabilities $P(b|a)$. We will treat these unknown conditional probabilities as parameters $\underline{\theta}$. We can separate $\underline{\theta}$ into M different sets of conditional probability parameters, one for each value of A, i.e., $\underline{\theta} = (\underline{\theta}_1, \ldots, \underline{\theta}_M)$. Each set of parameters $\underline{\theta}_k$ contains M conditional probabilities that sum to 1, each conditioned on a particular value $A = k$, i.e., $\underline{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,M})$ where $\theta_{k,l} = P(B = l|A = k)$.

Sidenote on notation: we will use notation below such as $P(B = l|A = k, \theta_{k,l})$, which you can think of as saying "if we know $A = k$ and we know the value of $\theta_{k,l}$, then our conditional probability for $B = l$ given $A = k$ is itself the parameter $\theta_{k,l}$." This notation, where we put parameters like $\theta_{k,l}$ on the conditioning side of a conditional probability, is not very elegant, but it is convenient and useful in general (will be particularly useful when we discuss Bayesian learning later on).

Now say we have an observed data set $D = \{(a_i, b_i)\},\ 1 \leq i \leq N$, i.e., a set of N observations, with a_i and b_i denoting the value of A and the value of B respectively for each pair of observations. For example i might refer to an individual and A and B might be two discrete variables or attributes that we can measure for any individual.

If we assume the observations are IID conditioned on the unknown parameters $\underline{\theta}$, we can write the log-likelihood as

$$\begin{aligned} \log L(\underline{\theta}) &= \sum_{i=1}^{N} \log P(a_i, b_i|\underline{\theta}) \\ &= \sum_{i=1}^{N} \Big(\log P(b_i|a_i, \underline{\theta}) + \log P(a_i) \Big) \end{aligned}$$

We can drop the terms $\log P(a_i)$ from the likelihood since they are assumed here to be known and do not depend on $\underline{\theta}$. We can also simplify the log-likelihood expression by writing it out as a sum over the parameters for each of the different conditional probability vectors $\underline{\theta}_k$ for each value of A:

$$\begin{aligned}
\log L(\underline{\theta}) &= \sum_{i=1}^{N} \log P(b_i|a_i, \underline{\theta}) \\
&= \sum_{k=1}^{M} \Big(\sum_{i=1}^{N_k} \log P(b_i|a_i = k, \underline{\theta}_k) \Big) \\
&= \sum_{k=1}^{M} \log L(\underline{\theta}_k)
\end{aligned}$$

where N_k is the number of times that $a = k$ occurs in the data (here we have grouped the likelihood terms in correspondence with the M values of A). Since each of the terms $\log L(\underline{\theta}_k)$ involves different sets of parameters $\underline{\theta}_k$, we can maximize each one separately, i.e., estimate the maximum likelihood parameters (conditional probabilities) for each of the M different values of A. Thus, we have

$$\begin{aligned}
\log L(\underline{\theta}_k) &= \sum_{i=1}^{N_k} \log P(b_i|a_i = k, \underline{\theta}_k) \\
&= \sum_{l=1}^{M} \Big(\sum_{i:b_i=l} \log P(b_i = l|a_i = k, \underline{\theta}_k) \Big) \quad \text{by grouping terms with } b_i = l \\
&= \sum_{l=1}^{M} r_{k,l} \Big(\log P(b_i = l|a_i = k, \underline{\theta}_k) \Big) \\
&= \sum_{l=1}^{M} r_{k,l} \log \theta_{k,l}
\end{aligned}$$

where $r_{k,l}$ is a count of the number of times that $a = k$ and $b = l$ in the data D, and where $\sum_{l=1}^{M} r_{k,l} = N_k$. This is the same form as the multinomial problem (See lectures and/or homework 2). If we maximize this for each $\theta_{k,l}$, the solution is

$$\hat{\theta}_{k,l}^{ML} = \frac{r_{k,l}}{N_k} \qquad 1 \leq l, k \leq M$$

i.e., the maximum likelihood estimate of each conditional probability $\theta_{k,l} = P(B = l|A = k)$ is the number of times $r_{k,l}$ that $A = k$ and $B = l$ occur in the data, divided by the number of times N_k that $A = k$ occurs, i.e,. a standard frequency-based estimate for a conditional probability.

If we now have a more complicated graphical model, e.g., $A \rightarrow B \rightarrow C$, we can again factorize the likelihood into terms that only involve local conditional probability tables, with a local table for each variable conditioned on its parents. The maximum likelihood estimates of these conditional probabilities are the "local" frequency based estimates of how often both the parent-child combination of values occurs divided by the number of times the parent value occurs (see subsection below for details).

4.2 More General Graphical Models (Optional Reading)

We can generalize the ideas above to any arbitrary directed graphical model. Assume we have a set of d random variables where we know the structure of an associated graphical model, i.e., for each variable X_j we know the parent set $pa(X_j)$ in the graph. We also have available an $N \times d$ data matrix D consisting of independent random samples from the joint distribution $P(x_1, \ldots, x_d)$, where x_{ij} is the observed value for variable X_j in the ith random sample. For simplicity assume that each variable X_j is discrete and takes M values. Given the structure of the graphical model we would like to use the data D to estimate the CPTs for the model.

The parameters $\underline{\theta}$ can in general be defined as the set $\underline{\theta} = \{\underline{\theta}_j\}$ where the index $j = 1, \ldots, d$, i.e., j ranges over the variables $X_1, \ldots, X_d$ (note that this is a little different to the notation from earlier in this section). Each $\underline{\theta}_j$ is the set of relevant parameters for variable X_j, or more specifically, the set of parameters defining the CPT $P(x_j|pa(X_j))$. (Note again that when we say "parameters" here we mean conditional probabilities: we refer to them as parameters since they are unknown and we wish to estimate them from data).

It is straightforward to show that the overall likelihood can be decomposed into separate local likelihood terms, one per variable X_j, as follows:

$$\begin{aligned} L(\underline{\theta}) \;=\; P(D|\underline{\theta}) \quad &= \quad \prod_{i=1}^{N} P(\underline{x}_i|\underline{\theta}) \\ &= \quad \prod_{i=1}^{N} \Big(\prod_{j=1}^{d} P(x_{ij}|pa(X_j)_i, \underline{\theta}_j) \Big) \\ &= \quad \prod_{j=1}^{d} \Big(\prod_{i=1}^{N} P(x_{ij}|pa(X_j)_i, \underline{\theta}_j) \Big) \\ &= \quad \prod_{j=1}^{d} L(\underline{\theta}_j) \end{aligned}$$

where $L(\underline{\theta}_j) = \prod_{i=1}^{N} P(x_{ij}|pa(X_j)_i, \underline{\theta}_j)$ is the part of the likelihood only involving parameters $\underline{\theta}_j$ for variable X_j. (Here $pa(X_j)_i$ indicates the value(s) of the parents of X_j for the ith data point $\underline{x}_i$). Thus, the total likelihood decomposes into local likelihoods per node (or per variable).

We can write this in log-likelihood form as:

$$\log L(\underline{\theta}) \;=\; \sum_{j=1}^{d} \log L(\underline{\theta}_j)$$

where

$$\log L(\underline{\theta}_j) \;=\; \sum_{i=1}^{N} \log P(x_{ij}|pa(X_j)_i, \underline{\theta}_j)$$

We can maximize the full log-likelihood by independently maximizing each local log-likelihood $\log L(\underline{\theta}_j)$ as long as the $\underline{\theta}_j$ parameters for each variable X_j are not constrained or related (if they are then we would need to a joint maximization over the different terms). Thus, we have reduced the problem of finding the maximum likelihood parameters for a directed graphical model to d separate problems, where each problem corresponds to finding the maximum likelihood parameters for the conditional probability tables for child node X_j given its parents in the graphical model.

The parameters $\underline{\theta}_j$ can be defined as a set $\underline{\theta}_j = \{\underline{\theta}_{j,k}\}$, where $\underline{\theta}_{j,k} = \{\theta_{j,k,l}\}$ where each $\theta_{j,k,l} = P(x_j = l|pa(x_j) = k)$, i.e., these parameters are the conditional probabilities of X_j taking different values $l, 1 \leq l \leq M$, conditioned on a particular set of values k for the parent nodes. In the earlier subsection k ranged over M values (the values of variable A): but in the general case a node might have multiple parents, so k in general will range over all possible combinations of values of the parents.

Each local log-likelihood $\log L(\underline{\theta}_j)$ can be written as

$$\begin{aligned} \log L(\underline{\theta}_j) &= \sum_{i=1}^{N} \log P(x_{ij}|pa(X_j)_i, \underline{\theta}_j) \\ &= \sum_k \sum_l \left(r_{k,l} \log P(x_j = l|pa(X_j) = k, \underline{\theta}_{j,k} \right) \end{aligned}$$

where the two sums are over all possible values l and k of the child and parent variable(s) respectively, and where $r_{k,l}$ is the number of times that those particular combinations of parent-child values occur in the data. The sum over k has $M^{|pa(X_j|}$ different terms, where $|pa(X_j)|$ is the number of parents of variable x_j (for the special case in the model where all variables take the same number of values M). The innermost sum l is over the M possible values that each variable x_j can take, conditioned on some setting k of the values of the parent variables $pa(X_j)$.

It follows from the equation above that $\log L(\underline{\theta}_j)$ can be further broken down as sums of local likelihood terms

$$\log L(\underline{\theta}_j) = \sum_k \log L(\underline{\theta}_{j,k})$$

with a log-likelihood term $\log L(\underline{\theta}_{j,k})$ for each set of parameters $\underline{\theta}_{j,k}$, where each term $\log L(\underline{\theta}_{j,k})$ can be maximized separately from all the other terms. From a maximum likelihood perspective, each of these terms $\log L(\underline{\theta}_{j,k})$ corresponds (in general) to a different conditional distribution $\underline{\theta}_{j,k} = \{\theta_{j,k,l}\}$ with probabilities that sum to 1 (over l), and the maximum likelihood estimates for each such distribution (corresponding to a "variable and parent value" combination) is the standard multinomial estimate from earlier, i.e.,

$$\hat{\theta}_{j,k,l} = \frac{r_{j,k,l}}{N_k}$$

where N_k is the number of times the specific parent values corresponding to k occur in the data and $r_{j,k,l}$ is the number of times that variable X_j takes value l and that the parents $pa(X_j) = k$, with $j = 1, \ldots, d$, $l = 1, \ldots, M$, $k = 1, \ldots, M^{|pa(X_j)|}$.

In this manner our maximum likelihood problem for the graphical model reduces to $M^{|pa(x_j|}$ different maximum likelihood estimations of conditional distributions, repeated for each variable X_j in the model.