Note Set 3: Models, Parameters, and Likelihood

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

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 2 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 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: 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 with parameters θ, i.e.,

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

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

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

- 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 observations $\underline{x}$ we have $D = \{\underline{x}_1, \ldots, \underline{x}_n\}$, where x_{ij} is the jth component of the ith observation, $1 \leq j \leq d, 1 \leq i \leq n$.

- 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: if a parameter θ_1 has higher likelihood $L(\theta_1)$ than the likelihood of another parameter θ_2, then the observed data is more probable 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 terms in $p(D|\theta)$ that don't involve θ, such as normalizing constants. What is important is the shape of the likelihood function, not the 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 scalar parameters 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" defined over p dimensions. For example, for the Gaussian model $p(x)$ for scalar x on the real-line, 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 here that we could have defined θ_2 as 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 is also often convenient to work with $\log \sigma$ 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 $D = \{x_1, \ldots, x(n)\}$ and $x_i \in \{heads, tails\}$ represents the outcome of the ith toss.

As is standard in building likelihood models not only do we have to specify a probability model for a single "sample", x_i, we also need to specify a probability model for multiple samples. 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 $x_i = heads$ and $1-\theta$ for $x_i = tails$. This particular "coin-tossing" model, combining a Bernoulli with conditional independence of the x_i's, corresponds to the well-known **Binomial model**.

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). Note that it 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 (closer to the real-world) the case where X_i represents the ith Web surfer to arrive at Amazon's Web site and x_i is a binary value indicating whether the Web surfer make 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

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(x_i|\theta) \\ &= \theta^r (1-\theta)^{n-r} \end{aligned}$$

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 the convention is that we can ignore such terms in our definition of $L(\theta)$. Alternatively, we could assume that we were given the exact sequence of heads and tails that occurred (rather than just the numbers r and n), in which case no

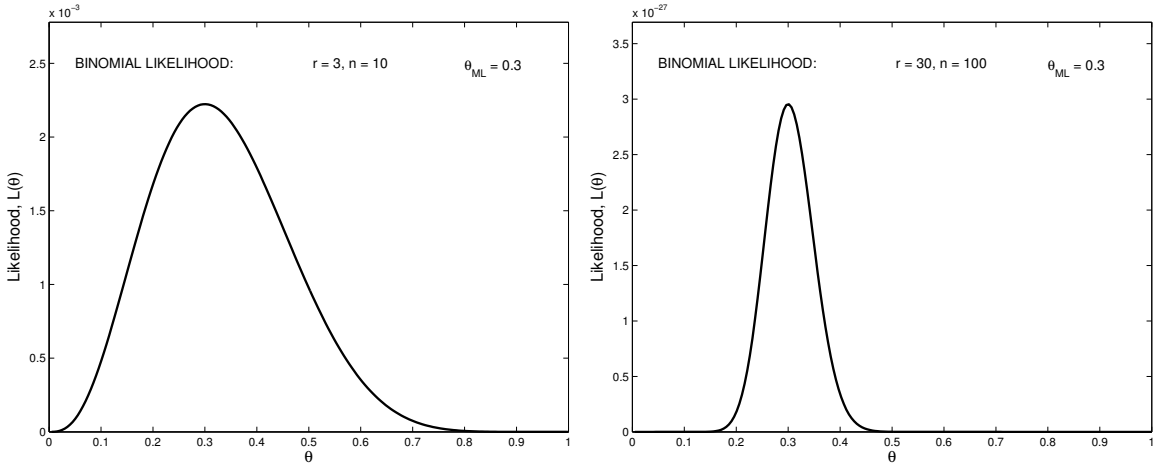

Figure 1: Binomial likelihood for (a) $r = 3, n = 10$, and (b) $r = 30, n = 100$.

combinatorial term is needed since the data D consists of the exact sequence rather than a report of how many heads and how many tails.

The figure shows two examples of the binomial likelihood function for different data sets. In the first 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. For the 2nd figure we have $r = 30$ and $n = 100$: here the likelihood is much narrower as we might expect. The plausible values for θ are much narrower after seeing 100 observations compared with just 10.

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 dependent, but only conditionally independent. Why? If θ is unknown (remember that θ is the probability of heads) then in this case 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 $x_i's$ (since we tend to get "runs" of wet days and dry days). For example, we could make a

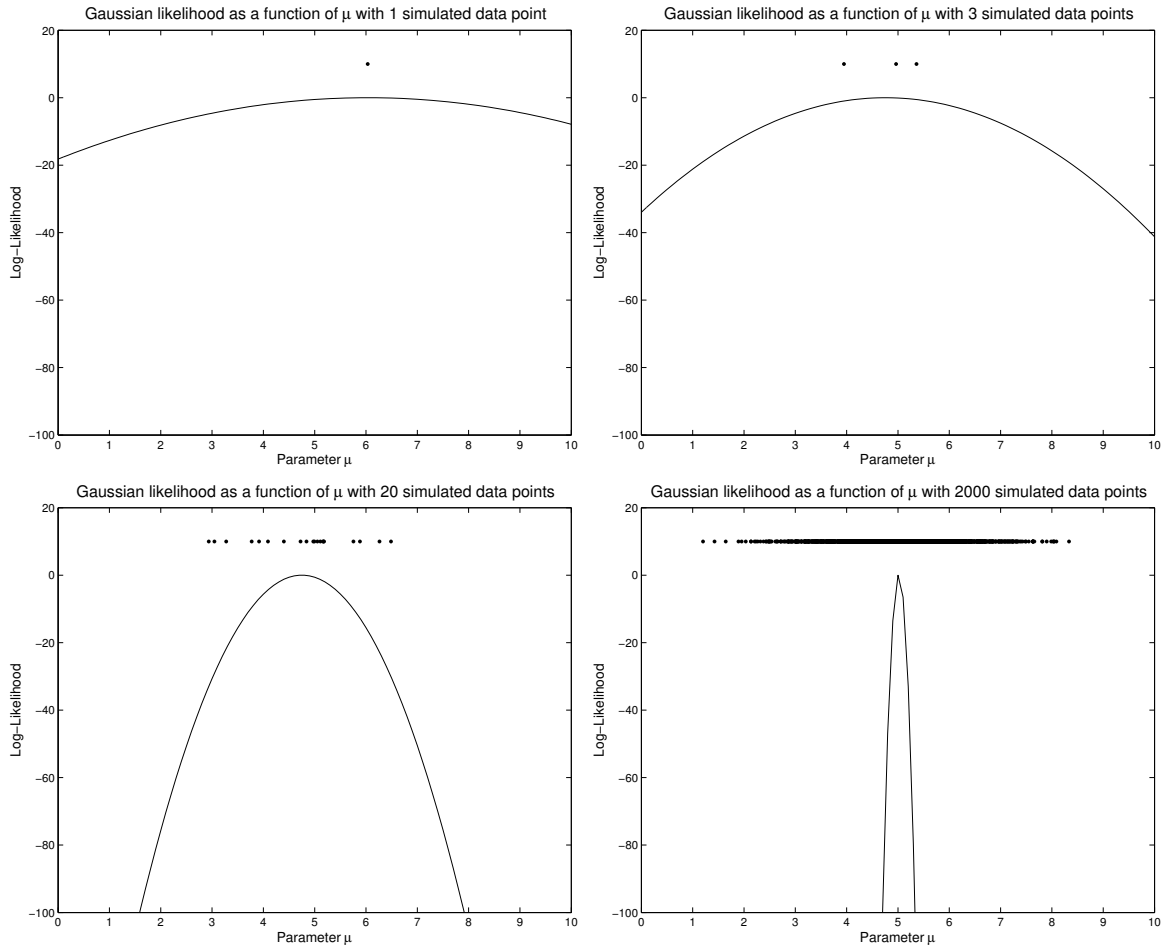

Figure 2: Log-likelihood for 4 different sample sizes, as a function of parameter $\theta = \mu$, with data simulated from a Gaussian with true $\mu = 5$ and $\sigma = 1$ (simulated data shown as dots horizontally at the top of the plot).

Markov assumption (Note Set 2) and assume that x on day $i+1$ is conditionally independent of x on days $i-1, i-2, \ldots, 1$, given 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/not-rain, given 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 say we wish to model these 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 $\underline{\theta}$ (and in 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 $\mu = 5$.

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:** We showed earlier that the binomial likelihood could be written as:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(x_i|\theta) \\ &= \theta^r (1-\theta)^{n-r} \end{aligned}$$

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[2] .

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

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.

EXAMPLE 5: **Maximum Likelihood Estimate for the Gaussian IID Model:**

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

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. See in-class notes for more details.

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. 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, corresponding to perhaps 2 different telescopes in different parts of the world for example. Assume in this simplified example that each camera produces noisy estimates of the object's 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[3], 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.,

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

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, if $\sigma_1^2 \neq \sigma_2^2$. The more different these two terms are, the more important it may be to account for measurements from the two data sets differently. In

[3] Another way to say this is that the camera measurements are **unbiased estimates** of the brightness.

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.

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

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 is reasonable for example if the measurements were taken relatively far apart in time. Taking logs and dropping terms that don't involve μ, we get

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

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.