Data, Estimation and Inference

Michael A. Osborne
mosb@robots.ox.ac.uk
http://www.robots.ox.ac.uk/~mosb/aims_cdt

Michaelmas 2015
Topic 4: Maximum Likelihood Estimation
Recall Bayes’ rule.

\[ p(f_* | \mathcal{D}) = \frac{p(\mathcal{D} | f_*) p(f_*)}{p(\mathcal{D})} \]

Given data \( \mathcal{D} \), and variable of interest (predictant) \( f_* \),

1. \( p(f_* | \mathcal{D}) \) is called the posterior for \( f_* \); this is our goal.
2. \( p(f_*) \) is called the prior for \( f_* \).
3. \( p(\mathcal{D} | f_*) \) is called the likelihood of \( f_* \), and is considered as a function of \( f_* \), \( L(f_*) = p(\mathcal{D} | f_*) \). Note that \( \int L(f_*)df_* \neq 1! \)
4. \( p(\mathcal{D}) \) is called the evidence, or marginal likelihood.

Note that posterior for \( f_* \) is proportional to its likelihood and prior; the evidence has no \( f_* \)-dependence and is just a normalisation factor.
It’s usual that the **prior is broad and the likelihood is narrow**.

Put another way, we begin highly ignorant, but the data we receive is very informative. Hence the product of prior and likelihood, \( \propto \) the posterior, is closer to the likelihood than the prior.

\[
p(x_1 | x_2 = -5) \quad p(x_1) \quad p(x_1, x_2)
\]
Consider a more complex scenario in which we have parameters \( \theta \).

As before, we want \( p(f_\star | \mathcal{D}) \). However, it’s impossible to evaluate \( p(f_\star, \mathcal{D}) \) without \( \theta \), which are nuisance variables.

**Example:** Physics tells us that resistance \( f \) varies linearly with temperature \( x \) given parameters \( \theta \). That is, \( f = \theta_1 x + \theta \).

We assume our measurements of \( x \) are noiseless, but we have only noisy sensor measurements, \( z \), of \( f \). Given measurements \( \mathcal{D} = \{ (x_i, z_i) \} \), we wish to estimate the resistance \( f_\star \) at known temperature \( x_\star \).
As always, we start by writing the probability of everything, also called the **generative model**.

With parameters, this model would be $p(f_*, D, \theta)$. Then

$$p(f_* \mid D) = \frac{p(f_*, D)}{p(D)} = \frac{\int p(f_*, D, \theta) \, d\theta}{p(D)} = \int \frac{p(f_* \mid D, \theta) \, p(D \mid \theta) \, p(\theta)}{p(D)} \, d\theta$$

1. $p(f_* \mid D)$ is called the **posterior** for $f_*$; this is our goal.
2. $p(f_* \mid D, \theta)$ are the **predictions** given $\theta$.
3. $p(\theta)$ is called the **prior** for $\theta$.
4. $p(D \mid \theta)$ is called the **likelihood** of $\theta$.
5. $p(D) = \int p(D \mid \theta) \, p(\theta) \, d\theta$ is called the **evidence**, or marginal likelihood.

The **predictions** are averaged over, weighted by their **likelihood** and **prior**.
Likelihoods are often a product of independent terms.

Henceforth, we’ll assume independent variables like \( x \) in \( f(x) \) are always known, and drop them, as \( p(a \mid x) = p(a) \).

Often the parameters are completely informative about the signal (as in the linear resistance example). If we know the parameters, and hence the true signal, the only variation left is due to noise, which we take as independent (and usually identically distributed). That is, the data become independent,

\[
p(D \mid \theta) = p(z_1, z_2, z_3, \ldots \mid \theta) = \prod_i p(z_i \mid \theta).
\]
Maximum likelihood approximates by ignoring the prior, and taking the likelihood as a delta function at its maximum.

\[
p(\theta) = 1
\]
\[
p(\mathcal{D} \mid \theta) = \delta(\theta - \hat{\theta})
\]
\[
\hat{\theta} = \arg\max_{\theta} p(\mathcal{D} \mid \theta).
\]

We’ll often work with log-likelihoods. As \(\log(\cdot)\) is monotonically increasing, the maximum of the log-likelihood is also the maximum of the likelihood, and if the log-likelihood is a delta function, so is the likelihood.
This approximation is justified when we have much data by the fact that likelihoods tend to become narrower with more data.

The more data we have, the more certain we are about parameters. The more data, the greater the number of independent terms in the likelihood,

$$p(\mathcal{D} \mid \theta) = \prod_{i} p(z_i \mid \theta),$$

each of which narrows it a little further.
Maximum likelihood gives a simple form for the posterior,

\[
p(f^* | D) = \frac{\int p(f^* | D, \theta) \delta(\theta - \hat{\theta}) \, 1 \, d\theta}{\int \delta(\theta - \hat{\theta}) \, 1 \, d\theta}
\]

\[
= p(f^* | D, \hat{\theta}).
\]

Without the maximum likelihood approximations, the integrals we would be required to perform are often intractable.
Maximum likelihood is not appropriate for multi-modal likelihoods, where there are multiple interpretations of the data.

Note that maximum likelihood is reliant upon being able to find the maximum of the likelihood: this is difficult in high dimension!
Consider an arbitrary likelihood function \( \mathcal{L}(\theta) = p(D \mid \theta) \) for a parameter \( \theta \).

The MLE \( \hat{\theta} \) satisfies

\[
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}} = 0
\]

or, equivalently,

\[
\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}} = 0
\]

and

\[
\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} \bigg|_{\theta=\hat{\theta}} < 0.
\]

\[
\frac{\partial^2 \log \mathcal{L}(\theta)}{\partial \theta^2} \bigg|_{\theta=\hat{\theta}} < 0.
\]
Let’s focus on the straight line problem, with independent Gaussian measurements.

1. Physics gives us

\[ p(f \mid \theta) = \delta(f - (\theta_0 + \theta_1 x)). \]

2. Measurements give

\[ p(z_i \mid f_i) = \mathcal{N}(z_i; f_i, \sigma^2) \]

1. Hence we have predictions

\[ p(f_* \mid \mathcal{D}, \theta) = p(f_* \mid \theta) = \delta(f_* - (\theta_0 + \theta_1 x_*)), \]

2. likelihood

\[
p(\mathcal{D} \mid \theta) = \prod_i \int p(z_i \mid f_i, \theta)p(f_i \mid \theta)df_i = \prod_i p(z_i \mid F_i = \theta_0 + \theta_1 x_i) = \prod_i \mathcal{N}(z_i; \theta_0 + \theta_1 x_i, \sigma^2).\]

Henceforth we’ll simply define \( f_i = \theta_0 + \theta_1 x_i \).
Now we adopt a maximum likelihood approach.

\[
\hat{\theta} = \arg \max_{\theta} p(\mathcal{D} \mid \theta) = \arg \max_{\theta} \left( \exp \sum_i \left( -\frac{1}{2} \frac{(f_i - z_i)^2}{\sigma^2} \right) \right)
\]

As before, taking logs does not affect the maximum’s location, and the min of the negative is equal to the max.

The maximum likelihood estimator (MLE) for \(\theta\) is therefore

\[
\hat{\theta} = \arg \min_{\theta} \left( \sum_i \frac{1}{2} \frac{(f_i - z_i)^2}{\sigma^2} \right) = \arg \min_{\theta} \left( \sum_i (f_i - z_i)^2 \right)
\]
Hence for independent Gaussian measurements, maximum likelihood is identical to least squares estimation.

\[ \hat{\theta} = \arg \min_\theta \left( \sum_i \frac{1}{2} \frac{(f_i - z_i)^2}{\sigma^2} \right) = \arg \min_\theta \left( \sum_i (f_i - z_i)^2 \right) \]
Suppose now that each observation has a different standard deviation $\sigma_i$.

1. Physics gives us $p(f \mid \theta) = \delta(f - (\theta_0 + \theta_1 x))$.
2. Measurements give $p(z_i \mid f_i) = \mathcal{N}(z_i; f_i, \sigma_i^2)$.

Hence we have

1. predictions $p(f_\star \mid \mathcal{D}, \theta) = \delta(f_\star - (\theta_0 + \theta_1 x_\star))$,
2. likelihood $p(\mathcal{D} \mid \theta) = \prod_i \mathcal{N}(z_i; f_i, \sigma_i^2)$.

Hence the MLE is

$$\hat{\theta} = \arg \max \theta \prod_i \left( \exp \left( -\frac{1}{2} \frac{(f_i - z_i)^2}{\sigma_i^2} \right) \right)$$

Take logs ...  

$$\hat{\theta} = \arg \min \theta \left( \sum_i \frac{(f_i - z_i)^2}{\sigma_i^2} \right)$$

This appears then as a “weighted” optimization, with MLE telling you that the weight of each datum should be $w_i = 1/\sigma_i^2$. 
Recall the **Poisson** \( \mathcal{P}o(x; \mu) = \frac{\mu^x}{x!} e^{-\mu} \).

Given the observation 3, which of the following might be the likelihood function?

1. ![Graph 1](image1)
2. ![Graph 2](image2)
3. ![Graph 3](image3)
4. ![Graph 4](image4)
Recall the Poisson $\mathcal{P}o(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}$.

Given the observation 3, which of the following might be the likelihood function?

1. 

2. 

3. 

4. 
Recall the Poisson $\mathcal{P}o(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}$.

Given the observations 3, 1 and 5, which of the following might be the likelihood function?
Recall the **Poisson** $\mathcal{P} o(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}$.

Given the observations 3, 1 and 5, which of the following might be the likelihood function?
If our system has no latent variables $f$, we can use a relaxed version of maximum likelihood.

Without $f$, we can use Bayes’ rule directly,

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) p(\theta)}{\int p(\mathcal{D} \mid \theta) p(\theta) \, d\theta}.$$ 

As before, let’s assume a flat prior, $p(\theta) = 1$, valid when the likelihood is much narrower than the prior. This gives

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta),$$

where the posterior is just the normalised likelihood. Note that in this case we need not make the restrictive assumption that the likelihood is a delta function.
Now we repeatedly measure a fixed distance.

Here the only unknown is the distance, $\theta$. Assuming the $n$ measurements are independent,

$$p(D \mid \theta) = \prod_{i=1}^{n} p(z_i \mid \theta) = K_1 \prod_{i=1}^{n} \exp \left( -\frac{1}{2} \frac{(z_i - \theta)^2}{\sigma^2} \right)$$

$$- \ln p(D \mid \theta) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - \theta)^2 + K_2$$

$$= \frac{1}{2\sigma^2} \left( n\theta^2 - 2\theta \sum_{i=1}^{n} z_i + \sum_{i=1}^{n} z_i^2 \right) + K_2$$

$$= \frac{1}{2\sigma^2} \left( n\theta^2 - 2\theta n\bar{z} + \sum_{i=1}^{n} z_i^2 \right) + K_2$$

$$= \frac{1}{2\sigma^2} \left( n\theta^2 - 2\theta n\bar{z} + n\bar{z}^2 - n\bar{z}^2 + \sum_{i=1}^{n} z_i^2 \right) + K_2$$

$$= \frac{n}{2\sigma^2} (\theta - \bar{z})^2 + K(z_1, \ldots, z_n) + K_2$$
Now we repeatedly measure a fixed distance.

\[- \ln p(D \mid \theta) = \frac{n}{2\sigma^2} (\theta - \bar{z})^2 + K(z_1, \ldots, z_n) + K_2\]

Given the data, $K()$ & $K_2$ are constant.
⇒ The minimum is at $\theta = \bar{z}$ (no surprise!!)
But the likelihood varies as

\[p(D \mid \theta) \propto \exp \left\{ -\frac{n}{2\sigma^2} (\theta - \bar{z})^2 \right\}\]

\[= \exp \left\{ -\frac{1}{2(\sigma/\sqrt{n})^2} (\theta - \bar{z})^2 \right\}\]
⇒ if the Gaussian sensor provides distance measurements with a variance of $\sigma^2$, and assuming $p(\theta) = 1$, $p(\theta \mid D)$ has the same variance as $p(D \mid \theta)$, $\sigma_n^2 = \sigma^2 / n$.
This agrees with the result of convolution (end of Topic 2).
Take two sensors with differing noise. These results generalise in an obvious way to more sensors.

\[- \ln p(\mathcal{D} \mid \theta) = \left( \frac{(z_1 - \theta)^2}{2\sigma_1^2} + \frac{(z_2 - \theta)^2}{2\sigma_2^2} \right) + K_2 \]

\[= \frac{1}{2} \left( \sigma_1^{-2}(z_1 - \theta)^2 + \sigma_2^{-2}(z_2 - \theta)^2 \right) + K_2 \]

\[= \frac{1}{2} \left( (\sigma_1^{-2} + \sigma_2^{-2})\theta^2 - 2(\sigma_1^{-2}z_1 + \sigma_2^{-2}z_2)\theta \right) + K(z_1, z_2, \sigma_1, \sigma_2) + K_2 \]

\[= \frac{1}{2}(\sigma_1^{-2} + \sigma_2^{-2}) \left( \theta - \frac{\sigma_1^{-2}z_1 + \sigma_2^{-2}z_2}{(\sigma_1^{-2} + \sigma_2^{-2})} \right)^2 + K(z_1, z_2, \sigma_1, \sigma_2) + K_2 \]

The MLE is hence

\[\hat{\theta} = \frac{\sigma_1^{-2}z_1 + \sigma_2^{-2}z_2}{(\sigma_1^{-2} + \sigma_2^{-2})}.\]

The MLE is a weighted mean, with weights \(w_i = 1/\sigma_i^2\).
Now we determine the variance of \( p(\theta \mid \mathcal{D}) \) given \( p(\theta) = 1 \).

Here \( p(\theta \mid \mathcal{D}) \) is proportional to \( p(\mathcal{D} \mid \theta) \), and

\[
- \ln p(\mathcal{D} \mid \theta) = \frac{1}{2} \left( \sigma_1^{-2} + \sigma_2^{-2} \right) \left[ \theta - \frac{\sigma_1^{-2} z_1 + \sigma_2^{-2} z_2}{(\sigma_1^{-2} + \sigma_2^{-2})} \right]^2 + \text{const.}
\]

So the likelihood varies as

\[
p(\mathcal{D} \mid \theta) \propto \exp \left\{ - \frac{(\theta - \hat{\theta})^2}{2 \sigma_D^2} \right\}
\]

where

\[
\sigma_D^{-2} = \sigma_1^{-2} + \sigma_2^{-2}.
\]
Example

Recall

\[ \hat{\theta} = \frac{\sigma_1^{-2} z_1 + \sigma_2^{-2} z_2}{\sigma_1^{-2} + \sigma_2^{-2}}. \quad \sigma_D^{-2} = \sigma_1^{-2} + \sigma_2^{-2} \]

Let the sensors be \( \mathcal{N}(z_1; \mu, 10^2) \), and \( \mathcal{N}(z_2; \mu, 20^2) \). We obtain sensor readings of \( z_1 = 130 \), and \( z_2 = 170 \). The MLE estimate is

\[ \hat{\theta} = \frac{130/10^2 + 170/20^2}{1/10^2 + 1/20^2} = 138.0 \]

closer to the lower variance sensor’s value.

The variance is \( 1/(10^{-2} + 0.25 \times 10^{-2}) = 80 = 8.9^2 \), which is lower than the minimum variance of the two sensors. Once again, one sees that \( 1/\sigma^2 \) acts as a “weight”. 
We describe $1/\sigma^2$ as the precision in the measurement.

Multiple measurements then add to the precision

$$\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2} + \ldots$$

Suppose now the problem is a multivariate one. The analogue of the inverse variance is the inverse of the covariance matrix.

The precision matrix $S$ is defined as $S = \Sigma^{-1}$. 

Let’s generalise to the bivariate problem, where the unknown is a 2D position 
\[ \theta = (\theta_a, \theta_b) \].

We again take 2 independent measurements \( z_1 = (z_{1a}, z_{1b}) \) and \( z_2 = (z_{2a}, z_{2b}) \), where

\[ p(z_1 \mid \theta) = \mathcal{N}(z_1; \theta, \Sigma_1), \quad p(z_2 \mid \theta) = \mathcal{N}(z_2; \theta, \Sigma_2). \]

Then, briefly,

\[
- \ln p(D \mid \theta) \propto (z_1 - \theta)^\top \Sigma_1^{-1} (z_1 - \theta) + (z_2 - \theta)^\top \Sigma_2^{-1} (z_2 - \theta)
\]

\[
= \theta^\top (\Sigma_1^{-1} + \Sigma_2^{-1}) \theta - 2\theta^\top (\Sigma_1^{-1} z_1 + \Sigma_2^{-1} z_2) + \text{const.}
\]

\[
= (\theta - \hat{\theta})^\top \Sigma_\theta^{-1} (\theta - \hat{\theta}) + \text{const.}
\]

where

\[
\Sigma_\theta^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}
\]

\[
\hat{\theta} = \Sigma_\theta \left( \Sigma_1^{-1} z_1 + \Sigma_2^{-1} z_2 \right).
\]
If we ignore the prior by taking $p(\theta) = 1$, we can compute the posterior,

$$p(\theta | D) = \mathcal{N}(\theta; \hat{\theta}, \Sigma_\theta),$$

$$\Sigma_\theta^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}$$

$$\hat{\theta} = \Sigma_\theta \left( \Sigma_1^{-1}z_1 + \Sigma_2^{-1}z_2 \right).$$

Expressing using precision matrices $S = \Sigma^{-1},$

$$S_\theta = S_1 + S_2$$

$$\hat{\theta} = S_\theta^{-1} \left( S_1z_1 + S_2z_2 \right).$$

Large covariance means high uncertainty $\Rightarrow$ low precision

Small covariance means low uncertainty $\Rightarrow$ high precision
Example

[Q] Two independent measurements $z_1$ and $z_2$ are made of the position $\theta$ of an object, and the sensor errors may be modelled as $p(z_1|\theta) = \mathcal{N}(z_1; \theta, \Sigma_1)$ and $p(z_2|\theta) = \mathcal{N}(z_2; \theta, \Sigma_2)$. Determine the MLE of $\theta$ when

$$z_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad z_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

[A] The sensors are independent, so sum their precisions:

$$S_\theta = S_1 + S_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix} + \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1.25 & 0 \\ 0 & 1.25 \end{bmatrix}$$

$$\hat{\theta} = S_\theta^{-1} (S_1 z_1 + S_2 z_2)$$

$$= \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix} \left( \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1.2 \\ -0.6 \end{bmatrix}$$
Example

\[ z_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \]

\[ z_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \]

\[ \hat{\theta} = \begin{bmatrix} 1.2 \\ -0.6 \end{bmatrix}, \quad S_{\theta} = \begin{bmatrix} 1.25 & 0 \\ 0 & 1.25 \end{bmatrix} \]
In summary,

1. Maximum likelihood assumes the likelihood is a delta function at its maximum to resolve often intractable integrals.
2. This approximation is justified when we have lots of data.
3. Gaussian measurements are weighted proportional to their precision.