Bayesian Estimation

Consider the data points \((x_i, y_i)\) with \(i \in [1, m]\) generated from a linear model of the form:

\[
y_i = ax_i + n_i,
\]

where the model parameter is \(a\). We assume some amount of imperfection in the data or underlying model, which we approximate with additive noise \(n_i\). This model is just a line with slope \(a\) and zero intercept. Given \(m\) points \((x_i, y_i)\) we would like to estimate the model parameter \(a\). We could employ a least-squares estimation. This would entail finding the value of \(a\) that minimizes the least-squares error \(\sum_i (y_i - ax_i)^2\). This is a perfectly reasonable approach but may fail to take advantage of everything that we know about the data and model. Alternatively, we can take advantage of additional information by formulating this estimation within a probabilistic framework.

Maximum likelihood (ML): To begin, assume that we have perfect knowledge of the independent variable \(x_i\), and that the uncertainty \(n_i\) affects only the dependent variable \(y_i\). We then ask what is the probability of observing the dependent variables \(y_i\) given the model parameter \(a\). This probability – referred to as the likelihood – is \(P(Y \mid a)\), where \(Y = \{y_1, \ldots, y_m\}\). We can frame the estimation of the model parameter \(a\) as a maximization of the likelihood. If we assume that each data point is independent, then the
likelihood can be expressed as a product of individual probabilities:

\[ P(Y \mid a) = \prod_{i=1}^{m} P(y_i \mid a). \quad (2) \]

By taking the logarithm of the likelihood\(^1\) we can convert this product into a more manageable summation:

\[ \log(P(Y \mid a)) = \sum_{i=1}^{m} \log(P(y_i \mid a)). \quad (3) \]

This log likelihood can be maximized by differentiating with respect to the model parameter \(a\), setting the result equal to 0, and solving for \(a\). This is because for a unimodal probability function, the peak is simply the value of \(a\) that has the greatest likelihood – the value that we are trying to estimate. The peak of the probability function is also where the slope (derivative) of the function is 0.

Consider, for example, the case when the noise \(n_i\) is drawn from a Gaussian distribution\(^2\), which for simplicity we will assume has unit variance:

\[ P(n_i) \propto e^{-n_i^2}. \quad (4) \]

The probability of observing the data point \((x_i, y_i)\) conditioned on the model parameter \(a\) is equivalent to the noise value \(n_i\) being equal to \(y_i - ax_i\). As a result, the conditional probability is given by a Gaussian:

\[ P(y_i \mid a) \propto e^{-(y_i - ax_i)^2}. \quad (5) \]

By way of intuition, when \(y_i - ax_i = 0\), this probability is \(e^0 = 1\). That is, the probability of parameter \(a\) is 1 for points \(x_i\) and \(y_i\) that perfectly satisfy the model. As the difference \(y_i - ax_i\) deviates from 0, the probability decays according to a Gaussian. Substituting this probability into the log likelihood yields:

\[ \log(P(Y \mid a)) = \sum_{i=1}^{m} \log \left( e^{-(y_i - ax_i)^2} \right) \]

\[ = \sum_{i=1}^{m} -(y_i - ax_i)^2. \quad (6) \]

\(^1\)The log of the likelihood does not change the maximum of the function.

\(^2\)The normalizing scale factor on the Gaussian can be ignored because it has no impact on the maximization.
Differentiating the log likelihood with respect to the parameter $a$ yields:

$$
\frac{d \log(P(Y \mid a))}{da} = \sum_{i=1}^{m} 2x_i(y_i - ax_i). \tag{7}
$$

Setting the derivative equal to 0 and solving for $a$ yields:

$$
\sum_{i=1}^{m} 2x_i(y_i - ax_i) = 0
$$

$$
\sum_{i=1}^{m} 2x_iy_i - \sum_{i=1}^{m} 2ax_i^2 = 0
$$

$$
\sum_{i=1}^{m} ax_i^2 = \sum_{i=1}^{m} x_iy_i,
$$

$$
a = \frac{\sum_{i=1}^{m} x_iy_i}{\sum_{i=1}^{m} x_i^2}. \tag{8}
$$

This estimate of the model parameter $a$ is the result of maximizing the log likelihood and is referred to as the maximum likelihood (ML) estimator.

**Maximum a posteriori (MAP):** The ML estimation is strictly data driven in that it does not incorporate any knowledge of the model parameter that is not provided by the data. For example, we may expect that the slope $a$ will always be near unit-value. A maximum a posteriori (MAP) estimator allows for this type of prior knowledge to be incorporated into an estimator.

The ML estimator maximized the likelihood probability distribution. The MAP estimator, on the other hand, finds the maximum value of the posterior probability distribution. The posterior distribution is the conditional probability of a model parameter $a$ given the data, $P(a \mid Y)$. According to Bayes’ rule, the posterior distribution can be expressed as:

$$
P(a \mid Y) = \frac{P(Y \mid a)P(a)}{P(Y)}, \tag{9}
$$

where, $P(Y \mid a)$ is the now familiar likelihood, $P(a)$ is the prior, and $P(Y)$ is the evidence (the data). The evidence does not depend on the model parameter $a$, so for the purpose of maximization it can be ignored. This yields the following posterior distribution to be maximized:

$$
P(a \mid Y) = P(Y \mid a)P(a). \tag{10}
$$
The MAP estimator maximizes the product of the likelihood and the prior. If the prior is uniform over all parameters $a$, then $P(a)$ is a constant, and the MAP estimator reduces to an ML estimator.

As with the ML estimator, we assume that each data point is independent so that the posterior can be expressed as a product of individual probabilities:

$$P(a \mid Y) = \prod_{i=1}^{m} P(y_i \mid a) P(a).$$

(11)

This product can again be simplified by considering the logarithm of the posterior:

$$\log(P(a \mid Y)) = \sum_{i=1}^{m} \log(P(y_i \mid a)) + \log(P(a)).$$

(12)

This log posterior can be maximized by differentiating with respect to the model parameter $a$, setting the result equal to 0, and solving for $a$.

To do so we need a probability distribution for both the likelihood and the prior. As with the ML example above, let’s assume that the noise in the model is a unit-variance Gaussian. As before, the conditional probability is given by a Gaussian:

$$P(y_i \mid a) \propto e^{-(y_i - ax_i)^2}.$$ 

(13)

For this example let’s assume that the prior is that the model parameter $a$ has a value near 1 and that the prior distribution is also a unit-variance Gaussian:

$$P(a) \propto e^{-(a-1)^2}.$$ 

(14)

Substituting these probabilities into the log posterior yields:

$$\log(P(a \mid Y)) = \sum_{i=1}^{m} \log(e^{-(y_i - ax_i)^2}) + \log(e^{-(a-1)^2})$$

$$= \sum_{i=1}^{m} -(y_i - ax_i)^2 - (a - 1)^2.$$ 

(15)

Differentiating the log posterior with respect to the parameter $a$ yields:

$$\frac{d \log(P(a \mid Y))}{da} = \sum_{i=1}^{m} 2x_i(y_i - ax_i) - 2(a - 1).$$ 

(16)
Setting the derivative equal to 0 and solving for $a$ yields:

\[
\sum_{i=1}^{m} 2x_i(y_i - ax_i) - 2(a - 1) = 0
\]

\[
\sum_{i=1}^{m} 2x_i y_i - \sum_{i=1}^{m} 2ax_i^2 - 2(a - 1) = 0
\]

\[
a + \sum_{i=1}^{m} ax_i^2 = 1 + \sum_{i=1}^{m} x_i y_i
\]

\[
a \left(1 + \sum_{i=1}^{m} x_i^2 \right) = 1 + \sum_{i=1}^{m} x_i y_i
\]

\[
a = \frac{1 + \sum_{i=1}^{m} x_i y_i}{1 + \sum_{i=1}^{m} x_i^2}. \quad (17)
\]

Notice the similarity of this MAP estimator to the ML estimator, Equation (8). The additive factor of 1 in the numerator and denominator biases the estimate of the slope $a$ towards a value of 1, as specified by the prior. For example, if the data consists of a single point $x_i = y_i = 0$, then the MAP estimator will be 1, whereas the ML estimator will be undefined.

**Bayesian estimation:** The ML and MAP estimators each maximize a probability distribution (the likelihood or the posterior). It can be beneficial to provide not just the maximum of a distribution, but the entire distribution. For example, knowing that a distribution is peaked rather than broad gives us more confidence in the estimator.

Our third and final estimator is a Bayes’ estimator in which we consider the full posterior distribution. While often difficult to compute, this is the most powerful estimator. We begin with the posterior expressed according to Bayes’ rule:

\[
P(a \mid Y) = \frac{P(Y \mid a)P(a)}{P(Y)}, \quad (18)
\]

As with the ML and MAP estimators, it is necessary to specify the likelihood and prior probability functions. Because the Bayes’ estimator involves the full posterior distribution, the evidence $P(Y)$ in the denominator cannot be ignored as it was in the MAP estimator; its distribution must be specified.
as well. The Bayes’ estimator that minimizes the least-squares error\(^3\) is the mean of the posterior:

\[
\int da \ P(a \mid Y) a. 
\]  

(19)

The difficulty with a Bayes’ estimator, as compared to the MAP estimator, is that the inclusion of the evidence in the denominator can complicate the integration of the posterior. When needed, however, this integration can be performed numerically.

\(^3\)The Bayes’ estimator minimzes the least-squares error in the parameter space \((\hat{a} - a)^2\), where \(\hat{a}\) is the estimated and \(a\) is the true model parameter.