Notes on Maximum Likelihood & Maximum *a Posteriori* Estimation

Paul F. Roysdon, Ph.D.

I. Derivations of the MLE and MAP

A. Estimator Background

When the PDF is viewed as a function of the unknown parameter (with $\boldsymbol{x}$ fixed) it is termed the "likelihood function". Given a generic model

$$\boldsymbol{x}[0] = \mathbf{A} + \boldsymbol{\eta}[0]$$

where

$$\boldsymbol{\eta}[0] \sim \mathcal{N}(0, \sigma^2)$$

and PDF

$$\mathrm{p}_i(\boldsymbol{x}[0]; \mathbf{A}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\left[-\frac{1}{2\sigma_i^2}(\boldsymbol{x}[0]-\mathbf{A})^2\right]}.$$

The log-likelihood simplifies the equation

$$\ln \mathrm{p}(\boldsymbol{x}[0]; \mathbf{A}) = -\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(\boldsymbol{x}[0] - \mathbf{A})^2.$$

The first-derivative is

$$-\frac{\partial \ln \mathrm{p}(\boldsymbol{x}[0]; \mathbf{A})}{\partial \mathbf{A}} = \frac{1}{\sigma^2}(\boldsymbol{x}[0] - \mathbf{A}).$$

The second-derivative is

$$-\frac{\partial^2 \ln \mathrm{p}(\boldsymbol{x}[0]; \mathbf{A})}{\partial \mathbf{A}^2} = \frac{1}{\sigma^2}.$$

The efficiency of an estimator is measured by the curvature of the estimator. The curvature can be computed by the variance of the estimator

$$\mathrm{var}\left\langle \hat{\mathbf{A}} \right\rangle = \frac{1}{-\frac{\partial^2 \ln \mathrm{p}(\boldsymbol{x}[0]; \mathbf{A})}{\partial \mathbf{A}^2}},$$

or more conveniently as the expected value

$$-\mathrm{E}\left\langle \frac{\partial^2 \ln \mathrm{p}(\boldsymbol{x}[0]; \mathbf{A})}{\partial \mathbf{A}^2} \right\rangle.$$

B. General Model

Assume a linear model

$$\boldsymbol{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\eta}$$

where $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, $\mathbf{H} \in \mathbb{R}^{n\times p}$ with $n > p$ and $\mathrm{rank}(\mathbf{H}) = p$, $\boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$, and $\boldsymbol{\eta} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.

The PDF of $\boldsymbol{x}$ is

$$\mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbf{C})}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\theta})^\intercal \mathbf{C}^{-1}(\boldsymbol{x}-\mathbf{H}\boldsymbol{\theta})\right]}$$

C. Cramer-Rao Lower Bound

The Cramer-Rao Lower Bound (CRLB) is the lower bound for all unbiased estimators, thus it is the Minimum Variance Unbiased Estimator (MVUE). Given the model in Section I-B, the CRLB is defined as

$$\frac{\partial \ln \mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\mathbf{H}^\intercal \mathbf{C}^{-1}\mathbf{H})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

where

$$\hat{\boldsymbol{\theta}}_{\mathrm{CRLB}} = (\mathbf{H}^\intercal \mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal \mathbf{C}^{-1}\boldsymbol{x}$$

Note if $\hat{\boldsymbol{\theta}}$ is the MVUE, it is an "efficient estimator", and the minimum variance is the diagonal entries of the covariance matrix

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = (\mathbf{H}^\intercal \mathbf{C}^{-1}\mathbf{H})^{-1}$$

D. Maximum Likelihood Estimator

Given the model in Section I-B, the Maximum Likelihood Estimator (MLE) maximizes the likelihood of the parameter $\boldsymbol{\theta}$ such that

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} &= \underset{\boldsymbol{\theta}}{arg\,max}\ \mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta}) \\ &= (\mathbf{H}^\intercal \mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^\intercal \mathbf{C}^{-1}\boldsymbol{x}\end{aligned}$$

If $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is "efficient", then $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} \mapsto$ CRLB and thus is the MVUE.

Note, in general the MLE is not optimal. However, as $n \to \infty$ then the MLE converges to the MVUE asymptotically.

E. Maximum A Posteriori Estimator

Given the model in Section I-B, the Maximum *a Posteriori* (MAP) Estimator maximizes the posterior PDF, such that

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \underset{\boldsymbol{\theta}}{arg\,max}\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}),$$

where

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\boldsymbol{x})}, \tag{1}$$

also known as "Bayes' Rule" or "Bayes' Risk". Therefore

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \underset{\boldsymbol{\theta}}{arg\,max}\ \Big(\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\Big) \\ &= \underset{\boldsymbol{\theta}}{arg\,max}\ \Big(\ln \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}) + \ln \mathrm{p}(\boldsymbol{\theta})\Big).\end{aligned}$$

Note, under normality constraints the MAP is equivalent to the Minimum Mean Square Error (MMSE).

II. Tutorial

Both the MLE and MAP can be used for *estimation* and *prediction*. We will consider both cases, and define how the MLE differs from MAP and Bayesian methods. Throughout this section we will use as example a "fair" coin, meaning that both heads and tails are equally likely.

To begin, let our *evidence*, $\boldsymbol{x}$, be a set of *independent* observations

$$\boldsymbol{x} = \{x_i\}_{i=1}^{|\boldsymbol{x}|}, \tag{2}$$

where each x_i is a realization (a.k.a., an observation or measurement), of a random variable x, and the notation $|\boldsymbol{x}|$ represents the cardinality, or total number, of observations in the set $\boldsymbol{x}$ in a multidimensional space.

Let $\boldsymbol{\theta}$ represent the *set of probability distribution parameters that best explain the evidence* $\boldsymbol{x}$.

- **Parameter Estimation:** Given $\boldsymbol{x}$, we can estimate $\boldsymbol{\theta}$ using Bayes' Rule (repeated here):

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\boldsymbol{x})},$$

  where $\mathrm{p}(A)$ is the probability of A, and $\mathrm{p}(A|B)$ is the *conditional* probability of A given B.
- **Data Prediction:** If we are given new observations $\tilde{\boldsymbol{x}}$, we can compute the probability of the new observation supported by, or given, the evidence

$$\mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{x}).$$

A. *Parameter Estimation*

Notice that Eqn. 1 can be interpreted as

$$posterior = \frac{likelihood \cdot prior}{evidence}.$$

Therefore, the likelihood function is

$$likelihood = \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}).$$

*1) **MLE of** $\boldsymbol{\theta}$:* To estimate $\boldsymbol{\theta}$, we seek the value for $\boldsymbol{\theta}$ that maximizes the likelihood. Let $\hat{\boldsymbol{\theta}}_{ML}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$.

From [1] we know that the joint probability of a set of *independent* random variables (rv's) is a *product* of the probabilities associated with the individual rv's in the set. Thus, because $\boldsymbol{x}$ is a set of *independent* observations, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, we seek the value $\boldsymbol{\theta}$ that maximizes

$$\prod_{x_i \in \boldsymbol{x}} \mathrm{p}(x_i|\boldsymbol{\theta}). \tag{3}$$

Because the logarithm is a monotonically increasing function of its argument, we can simplify Eqn. 3 by replacing the product with a sum, and using the notation $\mathcal{L}$ to represent the logarithm:

$$\mathcal{L} = \sum_{x_i \in \boldsymbol{x}} \log \mathrm{p}(x_i|\boldsymbol{\theta}). \tag{4}$$

Therefore, we seek the values for the parameters in $\boldsymbol{\theta}$ that maximize $\mathcal{L}$, defined as

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{arg\,max}\ \Big(\mathcal{L}\Big). \tag{5}$$

Eqn. 5 is solved by by setting the first derivative to zero, such that

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_i} = 0 \qquad \forall \boldsymbol{\theta}_i \in \boldsymbol{\theta}. \tag{6}$$

Note: The logarithm, specifically the *natural logarithm*, will further simplify our later examples because most distributions are from exponential families (see Section III), e.g., given γ^b, let

$$a = \gamma^b,$$

then the logarithm of base γ is

$$\log_\gamma(a) = b.$$

The natural logarithm for an exponential function $a = e^b$ is defined as

$$\ln(a) = \log_e(a).$$

Therefore,

$$\ln(a) = b.$$

Notice that if in Eqn. 4, $\mathrm{p}(x_i|\boldsymbol{\theta})$ is equal to an exponential family, e.g., $e^{b(x_i,\boldsymbol{\theta})}$, then the expression simplifies to

$$\mathcal{L} = \sum_{x_i \in \boldsymbol{x}} b(x_i, \boldsymbol{\theta}).$$

An application of this fact is provided in Sections I-B thru I-E, and [2]. For additional reference, see [3], [4].

*2) **MAP Estimate of** $\boldsymbol{\theta}$:* In maximum *a posteriori* estimation, we seek the value for $\boldsymbol{\theta}$ that maximizes the *posterior* of Eqn. 1, e.g., $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$, denoted as $\hat{\boldsymbol{\theta}}_{MAP}$.

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{MAP} &= \underset{\boldsymbol{\theta}}{arg\,max}\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) \\
&= \underset{\boldsymbol{\theta}}{arg\,max}\ \left(\frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\boldsymbol{x})}\right) && (7) \\
&= \underset{\boldsymbol{\theta}}{arg\,max}\ \Big(\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})\Big) && (8) \\
&= \underset{\boldsymbol{\theta}}{arg\,max}\ \left(\prod_{x_i \in \boldsymbol{x}} \mathrm{p}(x_i|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})\right)
\end{aligned}$$

Note the denominator in Eqn. 7 is dropped in Eqn. 8 because it has no functional dependence on the maximization of $\boldsymbol{\theta}$.

Again, because because $\boldsymbol{x}$ is a set of *independent* observations, we can change the product to a sum, and we can simplify the expression by taking the logarithm of the posteriors:

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{arg\,max}\ \sum_{x_i \in \boldsymbol{x}} \log \mathrm{p}(x_i|\boldsymbol{\theta}) + \log \mathrm{p}(\boldsymbol{\theta}) \tag{9}$$

*3) **Bayesian Estimate of $\boldsymbol{\theta}$:*** Given the evidence $\boldsymbol{x}$, the MLE holds $\boldsymbol{\theta}$ constant and seeks a value for $\boldsymbol{\theta}$ that maximizes the belief in the evidence. Recall that there is no prior in MLE, therefore we cannot use prior belief in likely values for $\boldsymbol{\theta}$ in the estimation.

On the contrary, the MAP estimate of $\boldsymbol{\theta}$ provides a method to consider values from a distribution (not the distribution of values) that express our prior beliefs regarding the parameters, i.e., the MAP estimate returns the values of $\boldsymbol{\theta}$ where the probability $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$ is maximized.

By contrast, a Bayesian estimate calculates the full posterior *distribution* $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$. That is, all of the $\boldsymbol{\theta}$ values that are possible, we select a value for each element of the vector that is, by some definition, "*best*". For example, we may choose the expected value of $\boldsymbol{\theta}$, assuming its variance is sufficiently small.

Recall the definition of Bayes' Risk, Eqn. 1. In Bayesian estimation the denominator, the *probability of the evidence*, cannot be ignored. This can be reformulated into a integral

$$\mathrm{p}(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta})\ d\boldsymbol{\theta} \tag{10}$$

Note: for a given likelihood function, if we have a choice regarding how we express our prior beliefs, we must use the form that allows us to carry out the integration of Eqn 10. This is the notion of conjugate priors, see Sections III-F & III-G.

Finally, as with MAP estimation, Bayesian estimation also requires the prior belief in possible values of the parameter vector $\boldsymbol{\theta}$ to be expressed in the form of a distribution.

*4) **Example and Comparison**:* The MAP estimate differs from the MLE by allowing us to inject our prior beliefs of the parameter values in $\boldsymbol{\theta}$.

Consider a Bernoulli experiment of n trials flipping a fair coin. Let p_h be the probability of heads, and for each observation x_i, a scalar, the value of x_i is either heads or tails.

MLE Example: For the scalar parameter θ_h, let's calculate the MLE, $\hat{\theta}_{h(MLE)}$. The evidence, $\boldsymbol{x}$, is the set of *independent* observations

$$\boldsymbol{x} = \{x_i = |_{\text{tails}}^{\text{heads}}, \quad i = 1, \ldots, n\}$$

The log likelihood function is

$$\begin{aligned}\log \mathrm{p}(\boldsymbol{x}|\theta_h) &= \sum_{i=1}^{n} \log \mathrm{p}(x_i|\theta_h) \\ &= \sum_i \log \mathrm{p}(x_i = \text{heads}) + \sum_i \log \mathrm{p}(x_i = \text{tails}) \\ &= n_h \cdot \log \theta_h + (n - n_h) \cdot \log(1 - \theta_h),\end{aligned}$$

where n_h is the number of trials you predict will be heads. Let

$$\mathcal{L} = \log \mathrm{p}(\boldsymbol{x}|\theta_h).$$

The MLE for θ_h is solved by setting

$$\frac{\partial \mathcal{L}}{\partial \theta_h} = 0,$$

which yields the equation

$$\frac{n_h}{\theta_h} - \frac{(n - n_h)}{(1 - \theta_h)} = 0,$$

whose solution is the MLE

$$\hat{\theta}_{h(MLE)} = \frac{n_h}{n}.$$

If the number of trials, $n = 20$, and our prediction of heads, $n_h = 12$, the *maximum likelihood estimate of our parameter* θ_h is $\hat{\theta}_{h(MLE)} = 12/20 = \mathbf{0.6}$.

MAP Example: For the scalar parameter θ_h, let's calculate the MAP estimate, $\hat{\theta}_{h(MAP)}$. The MAP requires a prior belief distribution for the parameter θ_h, with the following constraints:

- The prior for θ_h must be zero outside the interval $[0, 1]$.
- Within the interval $[0, 1]$, θ_h may be any real value.
- A distribution that peaks within the interval $[0, 1]$ is desirable, e.g., the *beta distribution*.

The beta distribution, $Beta(\mathrm{p}|\alpha, \beta)$, has "shape" parameters α and β (we also refer to these as *hyper-parameters* to the parameter θ_h):

$$\mathrm{p}(\theta_h|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta_h^{\alpha-1} (1 - \theta_h)^{\beta-1}, \tag{11}$$

where $B = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function. For $\alpha, \beta > 0$ the mode (maximum value) of the beta distribution is located at

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

Returning to our coin example, assume that prior experience with the same coin yielded a prior distribution for θ_h with a peak at 0.5, i.e., a fair coin. Setting $\alpha = \beta$ results in a distribution with a peak in the center of the interval $[0, 1]$.

The variance of the beta distribution is given by

$$\text{var}\,\langle Beta \rangle = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Again, based on prior experience, let $\alpha = \beta = 5$ Therefore, the variance, $\text{var}\,\langle Beta \rangle \approx 0.025$, and standard deviation, $\text{std}\,\langle Beta \rangle \approx 0.16$.

Eqn. 9 can be rewritten to be

$$\begin{aligned}\hat{\theta}_{h(MAP)} = \underset{\theta_h}{\boldsymbol{arg\,max}} \Big(& n_h \cdot \log \theta_h + (n - n_h) \\ & \cdot \log(1 - \theta_h) + \log \mathrm{p}(\theta_h) \Big).\end{aligned}$$

The MAP for θ_h is solved by substituting the beta distribution for $\mathrm{p}(\theta_h)$ and setting the first derivative equal to zero:

$$\frac{n_h}{\theta_h} - \frac{n - n_h}{1 - \theta_h} + \frac{\alpha - 1}{\theta_h} - \frac{\beta - 1}{1 - \theta_h} = 0$$

whose solution is the MAP

$$\hat{\theta}_{h(MAP)} = \frac{n_h + \alpha - 1}{n + \alpha + \beta - 2} = \frac{n_h + 4}{n + 8}$$

Based on prior experience, $\alpha = \beta = 5$, if the number of trials, $n = 20$, and our prediction of heads, $n_h = 12$, the *maximum a posteriori estimate of our parameter* θ_h is $\hat{\theta}_{h(MAP)} = (12 + 4)/(20 + 8) = \mathbf{0.571}$.

Bayesian Estimation Example: Again consider the Bernoulli trial example of a fair coin and our measurement, and prediction, of heads and tails. Our prior for that example is given by the beta distribution, Eqn. 11. Note the explicit dependence of the prior on the hyper-parameters α and β. Inserting Eqn. 11 into Eqn. 10,

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}) &= \int_0^1 \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}_h)\ \mathrm{p}(\boldsymbol{\theta}_h)\ d\boldsymbol{\theta}_h \\ &= \int_0^1 \left(\prod_{i=1}^{n} \mathrm{p}(x_i|\boldsymbol{\theta}_h) \right)\ \mathrm{p}(\boldsymbol{\theta}_h)\ d\boldsymbol{\theta}_h \\ &= \int_0^1 \left(\boldsymbol{\theta}_h^{n_h} (1 - \boldsymbol{\theta}_h)^{n - n_h} \right)\ \mathrm{p}(\boldsymbol{\theta}_h)\ d\boldsymbol{\theta}_h. \end{aligned}$$

Notice that when we multiply a beta distribution with either a power of $\boldsymbol{\theta}_h$ of a power of $(1 - \boldsymbol{\theta}_h)$, the result is a different beta distribution. Thus, the probability of the evidence $\boldsymbol{x}$ is equivalent to a constant, z, whose value is dependent on the values α, β and measurement n_h. Reformulating Eqn. 1, we have

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta}_h|\boldsymbol{x}) &= \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}_h)\ \mathrm{p}(\boldsymbol{\theta}_h)}{\mathrm{p}(\boldsymbol{x})} \\ &= \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}_h)\ \mathrm{p}(\boldsymbol{\theta}_h)}{z} \\ &= \frac{1}{z}\ \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}_h)\ \mathrm{p}(\boldsymbol{\theta}_h) \\ &= \frac{1}{z} \left(\prod_{i=1}^{n} \mathrm{p}(x_i|\boldsymbol{\theta}_h) \right)\ \mathrm{p}(\boldsymbol{\theta}_h) \\ &= \frac{1}{z} \left(\boldsymbol{\theta}_h^{n_h} (1 - \boldsymbol{\theta}_h)^{n - n_h} \right)\ \mathrm{p}(\boldsymbol{\theta}_h) \\ &= Beta\left(\boldsymbol{\theta}_h \middle| \alpha + n_h, \beta + n - n_h \right). \end{aligned}$$

The last step is obtained from the fact of products or powers of distributions, mentioned in Section II-A.3. The result is a closed form expression for the posterior distribution for the parameter vector $\boldsymbol{\theta}_h$ to be estimated. If instead we seek a single value for $\boldsymbol{\theta}_h$, the expected value of the distribution can be calculated.

Recall from the example ,based on prior experience, let $\alpha = \beta = 5$, then

$$\begin{aligned} \hat{\theta}_{h(Bayesian)} &= \mathrm{E}\left\langle \boldsymbol{\theta}_h | \boldsymbol{x} \right\rangle \\ &= \frac{\alpha + n_h}{\alpha + \beta + n} \\ &= \frac{5 + n_h}{10 + n}. \end{aligned}$$

If the number of trials, $n = 20$, and our prediction of heads, $n_h = 12$, the *Bayesian estimate of our parameter* θ_h is $\hat{\theta}_{h(Bayesian)} = (5 + 12)/(10 + 20) = \mathbf{0.567}$.

A benefit of the Bayesian estimation is the ability to also calculate the variance of the estimate. Using the standard form for the variance of the beta distribution, the variance of the Bayesian estimate is 0.0079.

Comparison: We notice a few things:

- While the MLE only uses our prediction, the MAP and Bayesian estimates use our prediction and "pulls" the estimate toward our *prior* experience. Therefore a larger prior will more strongly influence the MAP and Bayesian estimates. Compare, $\hat{\theta}_{h(MLE)} = 0.6$ vs. $\hat{\theta}_{h(MAP)} = 0.571$ vs. $\hat{\theta}_{h(Bayesian)} = 0.567$.
- Larger, but equal, values for α and β, narrow the peak of the beta distribution around $\theta_h = 0.5$, also moving the MAP and Bayesian estimates toward the prior.
- The values for α and β tend to "smooth" the result of the measurement n_h.

While we used the Beta distribution in this example, we can certainly use, or assume, other distributions, e.g., in [2] both a Gaussian and Laplacian distribution are used.

B. Data Prediction

Given the evidence, Eq. 2, and new measurements $\tilde{\boldsymbol{x}}$, we want to know how much our new measurements are supported by the evidence, i.e.,

$$\mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{x}). \tag{12}$$

In other words, give the evidence, can we *predict* the new measurements. Prediction, also called *regression*, can be solved by either the MLE, MAP or Bayesian estimators.

*1) **MLE Prediction of** $\tilde{\boldsymbol{x}}$:* Expanding Eq. 12,

$$\begin{aligned} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) &= \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \\ &\approx \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\hat{\boldsymbol{\theta}}_{MLE})\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \\ &= \mathrm{p}(\tilde{\boldsymbol{x}}|\hat{\boldsymbol{\theta}}_{MLE}). \end{aligned}$$

Therefore, the probability for the new measurements $\tilde{\boldsymbol{x}}$ is the same for all prior measurements, based on the evidence $\boldsymbol{x}$ and probability model $\hat{\boldsymbol{\theta}}_{MLE}$.

*2) **MAP Prediction of** $\tilde{\boldsymbol{x}}$:* Expanding Eq. 12,

$$\begin{aligned} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) &= \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \\ &\approx \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\hat{\boldsymbol{\theta}}_{MAP})\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \\ &= \mathrm{p}(\tilde{\boldsymbol{x}}|\hat{\boldsymbol{\theta}}_{MAP}). \end{aligned}$$

Therefore, the probability for the new measurements $\tilde{\boldsymbol{x}}$ is the same for all prior measurements, based on the evidence $\boldsymbol{x}$ and probability model $\hat{\boldsymbol{\theta}}_{MAP}$.

*3) **Bayesian Prediction of $\tilde{\boldsymbol{x}}$:*** As before, we begin with Eq. 12, but now we must use Bayes' Rule for the posterior $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$:

$$\begin{aligned}\mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) &= \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})\ \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \mathrm{p}(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})\ \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}_h)\ \mathrm{p}(\boldsymbol{\theta}_h)}{\mathrm{p}(\boldsymbol{x})}\ d\boldsymbol{\theta}.\end{aligned}$$

III. Exponential Families

Many *exponential family* distributions are used in both MLE and MAP estimation. This section defines the most common distributions, derives the *natural parameters* and both the *first* and *second derivatives* for use in numerical (software) applications. This section closes with a brief explanation of *conjugate priors*.

A. Background

Consider the general exponential function

$$\mathrm{p}(\boldsymbol{x}) = h(\boldsymbol{x})e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})-A(\boldsymbol{\theta})},$$

where $\boldsymbol{\theta}$ is a parameter vector, $T(\boldsymbol{x})$ is a vector of "sufficient statistics", $A(\boldsymbol{\theta})$ is a cumulant generating function, and a general input function $h(\boldsymbol{x})$.

A normalized distribution for any $\boldsymbol{\theta}$ is obtained by taking the integral with respect to $\boldsymbol{x}$

$$\int \mathrm{p}(\boldsymbol{x})d\boldsymbol{x} = e^{-A(\boldsymbol{\theta})} \int h(\boldsymbol{x})e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})}d\boldsymbol{\theta} = 1,$$

therefore

$$e^{A(\boldsymbol{\theta})} = \int h(\boldsymbol{x})e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})}d\boldsymbol{\theta}.$$

Note: when $T(\boldsymbol{x}) = \boldsymbol{x}$, then $A(\boldsymbol{\theta})$ is the log of the Laplace transform of $h(x)$.

B. Examples

Gaussian:

$$\mathrm{p}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\|x-\mu\|^2/(2\sigma^2)} \quad x \in \mathbb{R}$$

Bernoulli:

$$\mathrm{p}(x) = \alpha^x(1-\alpha)^{1-x} \quad x \in \{0,1\}$$

Binomial:

$$\mathrm{p}(x) = \binom{n}{k}\alpha^x(1-\alpha)^{n-x} \quad x \in \{0,1,2,\ldots,n\}$$

Multinomial:

$$\begin{aligned}\mathrm{p}(x) &= \frac{n!}{x_1!x_2!\ldots x_n!}\prod_{i=1}^{n}\alpha_i^{x_i} \\ x &\in \{0,1,2,\ldots,n\}, \sum_i x_i = n\end{aligned}$$

Exponential:

$$\mathrm{p}(x) = \lambda e^{-\lambda x} \quad x \in \mathbb{R}^+$$

Poisson:

$$\mathrm{p}(x) = \frac{e^{-\lambda}}{x!}\lambda^x \quad x \in \{0,1,2,\ldots,n\}$$

Dirichlet:

$$\mathrm{p}(x) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i x_i^{\alpha_i-1} \quad x \in [0,1], \sum_i x_i = n$$

C. Natural Parameter Forms

*1) **Bernoulli:*** Recall the general equation

$$\mathrm{p}(\boldsymbol{x}) = h(\boldsymbol{x})e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})-A(\boldsymbol{\theta})},$$

then

$$\begin{aligned}\mathrm{p}(\boldsymbol{x}) &= \boldsymbol{\alpha}^{\boldsymbol{x}}(1-\boldsymbol{\alpha})^{1-\boldsymbol{x}} \\ &= \exp\left[\log(\boldsymbol{\alpha}^x(1-\boldsymbol{\alpha})^{1-\boldsymbol{x}})\right] \\ &= \exp\left[\boldsymbol{x}\log\boldsymbol{\alpha} + (1-\boldsymbol{x})\log(1-\boldsymbol{\alpha})\right] \\ &= \exp\left[\boldsymbol{x}\log\frac{\boldsymbol{\alpha}}{1-\boldsymbol{\alpha}} + \log(1-\boldsymbol{\alpha})\right] \\ &= \exp\left[\boldsymbol{x}\boldsymbol{\theta} - \log(1+e^{\boldsymbol{\theta}})\right],\end{aligned}$$

where $T(\boldsymbol{x}) = \boldsymbol{x}$, $\boldsymbol{\theta} = \log\frac{\boldsymbol{\alpha}}{1-\boldsymbol{\alpha}}$, and $A(\boldsymbol{\theta}) = \log(1+e^{\boldsymbol{\theta}})$.

*2) **Gaussian:***

$$\begin{aligned}\mathrm{p}(\boldsymbol{x}) &= \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}^2}}e^{-\|\boldsymbol{x}-\boldsymbol{\mu}\|^2/(2\boldsymbol{\sigma}^2)} \\ &= \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}^2}}\exp\left(-\log\boldsymbol{\sigma} - \frac{\boldsymbol{x}^2}{2\boldsymbol{\sigma}^2} + \frac{\boldsymbol{\mu}\boldsymbol{x}}{\boldsymbol{\sigma}^2} - \frac{\boldsymbol{\mu}^2}{2\boldsymbol{\sigma}^2}\right) \\ &= \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}^2}}\exp\left(\boldsymbol{\theta}^\intercal T(\boldsymbol{x}) - \log\boldsymbol{\sigma} - \frac{\boldsymbol{\mu}^2}{2\boldsymbol{\sigma}^2}\right)\end{aligned}$$

where $h(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}^2}}$, and $A(\boldsymbol{\theta}) = \log\boldsymbol{\sigma} - \frac{\boldsymbol{\mu}^2}{2\boldsymbol{\sigma}^2}$. Then

$$T(\boldsymbol{x}) = \begin{pmatrix}\boldsymbol{x} \\ \boldsymbol{x}^2\end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix}\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2} \\ -\frac{1}{2\boldsymbol{\sigma}^2}\end{pmatrix},$$

and

$$\begin{aligned}A(\boldsymbol{\theta}) &= \frac{\boldsymbol{\mu}^2}{2\boldsymbol{\sigma}^2} + \log\boldsymbol{\theta} \\ &= -\frac{[\boldsymbol{\theta}]_1^2}{4[\boldsymbol{\theta}]_2^2} - \frac{1}{2}\log(-2[\boldsymbol{\theta}]_2).\end{aligned}$$

*3) **Multivariate Gaussian:***

$$\mathrm{p}(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}$$

where $h(\boldsymbol{x}) = (2\pi)^{D/2}$, and

$$T(\boldsymbol{x}) = \begin{pmatrix}\boldsymbol{x} \\ \boldsymbol{x}\boldsymbol{x}^\intercal\end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\end{pmatrix},$$

D. First & Second Derivatives

1) First Derivative: Let,

$$A(\boldsymbol{\theta}) = \log\left[\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} d\boldsymbol{x}\right],$$

where $Q(\boldsymbol{\theta}) = \int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} d\boldsymbol{x}$, then

$$\begin{aligned}\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{Q(\boldsymbol{\theta})}\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{Q'(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \\ &= \frac{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} T(\boldsymbol{x}) d\boldsymbol{x}}{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} d\boldsymbol{x}} \\ &= \frac{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta})} T(\boldsymbol{x}) d\boldsymbol{x}}{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta})} d\boldsymbol{x}} \\ &= \mathrm{E}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T(\boldsymbol{x})\right\rangle.\end{aligned}$$

2) Second Derivative: Using the previous definition for $A(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta})$,

$$\begin{aligned}\frac{\partial^2 A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{Q'(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})}\right] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[Q'(\boldsymbol{\theta})\frac{1}{Q(\boldsymbol{\theta})}\right] \\ &= \frac{Q''(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} - \frac{(Q'(\boldsymbol{\theta}))^2}{(Q(\boldsymbol{\theta}))^2},\end{aligned}$$

therefore

$$\begin{aligned}\frac{\partial^2 A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} &= \frac{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} T^2(\boldsymbol{x}) d\boldsymbol{x}}{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x})} d\boldsymbol{x}} - \left(\mathrm{E}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T(\boldsymbol{x})\right\rangle\right)^2 \\ &= \frac{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta})} T^2(\boldsymbol{x}) d\boldsymbol{x}}{\int h(\boldsymbol{x}) e^{\boldsymbol{\theta}^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta})} d\boldsymbol{x}} - \left(\mathrm{E}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T(\boldsymbol{x})\right\rangle\right)^2 \\ &= \mathrm{E}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T^2(\boldsymbol{x})\right\rangle - \left(\mathrm{E}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T(\boldsymbol{x})\right\rangle\right)^2 \\ &= \mathrm{Cov}_{\mathrm{p}_{\boldsymbol{\theta}}}\left\langle T(\boldsymbol{x})\right\rangle \succeq 0,\end{aligned}$$

where the symbol $\succeq$ defines a positive definite matrix, and $A(\boldsymbol{\theta})$ is convex.

E. Products of Distributions

Products of exponential family distributions are exponential family distributions, e.g.,

$$\begin{aligned}&\left(h(\boldsymbol{x}) e^{\boldsymbol{\theta}_1^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta}_1)}\right) \times \left(h(\boldsymbol{x}) e^{\boldsymbol{\theta}_2^\intercal T(\boldsymbol{x}) - A(\boldsymbol{\theta}_2)}\right) \\ &\quad = \tilde{h}(\boldsymbol{x}) e^{(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) T(\boldsymbol{x}) - \tilde{A}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)},\end{aligned}$$

however the parametric form is more complex; see [1]. The product of two Gaussians is *always* a Gaussian.

F. Conjugate Priors

Given

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

notice the denominator is not a function of $\boldsymbol{\theta}$, rather it is a normalizing term. Thus, we have two parametric terms, $\mathrm{p}(\boldsymbol{\theta})$ and $\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\mathrm{p}(\boldsymbol{\theta}) \;\mapsto\; \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}) \;\mapsto\; \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}).$$

Conjugacy requires $\mathrm{p}(\boldsymbol{\theta})$ and $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$ to be of the same form, e.g.,

$$\mathrm{p}(\boldsymbol{\theta}) \;\mapsto\; \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}) \;\mapsto\; \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}),$$

where $\mathrm{p}(\boldsymbol{\theta})$ and $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$ are Dirichlet, and $\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})$ is Multinomial. Then, $\mathrm{p}(\boldsymbol{\theta})$ and $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x})$ are called conjugate distributions.

For example: the Dirichlet in $\boldsymbol{\theta}$, with $\Gamma(\boldsymbol{x}) = (\boldsymbol{x}-1)!$ is defined as

$$\mathrm{p}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_i \boldsymbol{\alpha}_i)}{\prod_i \Gamma(\boldsymbol{\alpha}_i)} \prod_i \boldsymbol{\theta}^{\boldsymbol{\alpha}_i - 1},$$

and Multinomial in $\boldsymbol{x}$ is defined as

$$\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{(\sum_i \boldsymbol{x}_i)!}{\boldsymbol{x}_1!\boldsymbol{x}_2!\ldots\boldsymbol{x}_n!} \prod_i^n \boldsymbol{\theta}_i^{\boldsymbol{x}_i}$$

Applying the definition above (repeated here)

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}) = \ldots \times \prod_i \boldsymbol{\theta}_i^{\boldsymbol{x}_i + \boldsymbol{\alpha}_i - 1},$$

which is again Dirichlet:

$$\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\Gamma(\sum_i \boldsymbol{\alpha}_i + \boldsymbol{x}_i)}{\prod_i \Gamma(\boldsymbol{\alpha}_i + \boldsymbol{x}_i)} \prod_i \boldsymbol{\theta}^{\boldsymbol{x}_i + \boldsymbol{\alpha}_i - 1}.$$

G. Conjugate Pairs

In the prior example we demonstrated conjugate pairs, now we define several common pairs.

Prior	Conditional
Gaussian: $e^{-\|\mu-\mu_0\|^2/(2\sigma^2)}$	Gaussian: $e^{-\|\mu-\mu_0\|^2/(2\sigma^2)}$
Beta: $\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\alpha^{r-1}(2-\alpha)^{s-1}$	Bernoulli: $\alpha^x(1-\alpha)^{1-x}$
Dirichlet: $\frac{\Gamma(\sum_i \boldsymbol{\alpha}_i)}{\prod_i \Gamma(\boldsymbol{\alpha}_i)} \prod_i \boldsymbol{\theta}^{\boldsymbol{\alpha}_i - 1}$	Multinomial: $\frac{(\sum_i \boldsymbol{x}_i)!}{\prod_i \boldsymbol{x}_i!} \prod_i^n \boldsymbol{\theta}_i^{\boldsymbol{x}_i}$
Inverse Wishart	Gaussian (cov.)

Note that conjugacy is mutual, e.g.,

$$\begin{aligned}&\text{Dirichlet} \;\mapsto\; \text{Multinomial} \;\mapsto\; \text{Dirichlet} \\ &\text{Multinomial} \;\mapsto\; \text{Dirichlet} \;\mapsto\; \text{Multinomial}\end{aligned}$$

IV. Multinomial Distributions for Machine Learning

In this section we will further examine Conjugate Pairs with Multinomial Distributions and explore their connection to modeling likelihoods for images and text data.

add text here...

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. I - Estimation Theory*. Prentice Hall PTR, 2013.

[2] P. Roysdon and J. Farrell, "Robust GPS-INS Outlier Accommodation using a Sliding Window Filter: A Comparative Study," 2017.

[3] P. F. Roysdon, *Math Handbook for Engineers and Scientists*. Fibonacci Press; 2nd edition, 2023.

[4] I. Bronshtein, K. Semendyayev, G. Musiol, and H. Muehlig, *Handbook of Mathematics*. Springer, 2007.