Statistics: Estimation

Estimation is a classical problem in statistics. We have access to samples from a random vector $\mathbf{Y}$ and want to estimate the corresponding samples of a related random variable X (X could also be a random vector, but we consider a single random variable to simplify the exposition). If we assume that we know the joint distribution of X and $\mathbf{Y}$, then we can derive estimators that are optimal in the sense that they minimize error metrics of interest.

1 Continuous random variables

If X is a continuous random variable, then a popular measure of the error is the mean square error, which we already encountered when we considered the problem of fitting the parameter of a parametric model. In fact, if we interpret X as a parameter, estimation is equivalent to parameter fitting in a Bayesian setting.

Definition 1.1 (Mean square error)**.**

$$\mathrm{E}\left(\left(X - g\left(\boldsymbol{Y}\right)\right)^2\right). \tag{1}$$

As we saw for parametric estimation, the optimal estimator in terms of MSE is the conditional expectation of X given $\mathbf{Y}$.

Theorem 1.2 (The conditional mean minimizes the MSE)**.** *For a random variable X and a random vector $\boldsymbol{Y}$ belonging to the same probability space, the conditional mean of X given $\boldsymbol{Y}$ is the minimum MSE estimator of X given $\boldsymbol{Y}$.*

$$\mathrm{E}\big(X \,|\, \boldsymbol{Y}\big) = \arg\min_{g} \mathrm{E}\left(\left(X - g\left(\boldsymbol{Y}\right)\right)^2\right). \tag{2}$$

Proof. The proof is identical to that of Theorem 1.6 in Lecture Notes 5. □

Example 1.3 (Gangue)**.** In mining, gangue is worthless rock or material that is mixed with the ore. Imagine that we have a model which indicates that every day a mine produces an amount of ore that is uniformly distributed between 0 and 1 metric ton, and also an amount of gangue that is uniformly distributed between 0 and 1 metric ton. Both amounts are modeled as independent. If the weight of the mixture of ore and gangue is equal to y, what is the best estimate of the amount of ore in terms of MSE?

Let the amount of ore be modeled as a random variable Y. By Theorem 1.2, the best estimate is $E(X|Y=y)$. If we fix $X=x$ then Y is uniformly distributed between 0 and 1, so

$$f_{Y|X}(y|x) = \begin{cases} 1 & \text{if } x \le y \le x+1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We now compute the marginal pdf of Y. If $y<0$ or $y>2$, $f_Y(y)=0$ because the amount of ore is bounded between 0 and 2. If $0 \le y \le 1$,

$$f_Y(y) = \int_{x=0}^{y} f_X(x) f_{Y|X}(y|x)\, \mathrm{d}x \tag{4}$$

$$= y. \tag{5}$$

If $1 \le y \le 2$,

$$f_Y(y) = \int_{x=y-1}^{1} f_X(x) f_{Y|X}(y|x)\, \mathrm{d}x \tag{6}$$

$$= 2-y. \tag{7}$$

Now we compute the conditional pdf of Y given X. If $0 \le y \le 1$,

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} \tag{8}$$

$$= \frac{1}{y} \quad \text{for } 0 \le x \le y. \tag{9}$$

This means that conditioned on the event $\{Y=y\}$ where $0 \le y \le 1$, X is uniformly distributed between 0 and y, so $\mathrm{E}_{X|Y}(X|Y=y) = y/2$.

If $1 \le y \le 2$

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} \tag{10}$$

$$= \frac{1}{2-y} \quad \text{for } y-1 \le x \le 1. \tag{11}$$

Conditioned on the event $\{Y=y\}$ where $1 \le y \le 2$, X is uniformly distributed between $y-1$ and 1, so $\mathrm{E}_{X|Y}(X|Y=y) = \frac{1-(1-y)}{2} = y/2$.

The best estimate in terms of MSE is $y/2$.

---

2 Discrete random variables

If the quantity X that we want to estimate is discrete and can only take a small number of values $x_1, \ldots, x_m$, the MSE is not a very reasonable error metric. Indeed, the conditional mean is not necessarily restricted to $\{x_1, \ldots, x_m\}$. It seems to make more sense to choose the *best* value within $\{x_1, \ldots, x_m\}$.

If we only know the likelihood of the data given the signal of interest, a possibility is to choose the value with a highest likelihood, just as in parameter estimation within a frequentist setting.

Definition 2.1 (Maximum-likelihood estimator)**.** *Let X be a random variable with range restricted to the set $\{x_1, \ldots, x_m\}$ and let $\boldsymbol{Y}$ be a random vector belonging to the same probability space. The maximum-likelihood (ML) estimator of X given $\boldsymbol{Y}$ is*

$$g_{\mathrm{ML}}\left(\boldsymbol{y}\right) := \arg \max_{u \in \{x_1, \ldots, x_m\}} \mathcal{L}_{\boldsymbol{y}}\left(u\right) \tag{12}$$

$$= \arg \max_{u \in \{x_1, \ldots, x_m\}} \log \mathcal{L}_{\boldsymbol{y}}\left(u\right), \tag{13}$$

where the likelihood function $\mathcal{L}_{\boldsymbol{y}}\left(x\right)$ is defined as

$$\mathcal{L}_{\boldsymbol{y}}\left(x\right) := p_{\boldsymbol{Y}|X}\left(\boldsymbol{y}|x\right) \tag{14}$$

if $\boldsymbol{Y}$ is discrete and

$$\mathcal{L}_{\boldsymbol{y}}\left(x\right) := f_{\boldsymbol{Y}|X}\left(\boldsymbol{y}|x\right) \tag{15}$$

if $\boldsymbol{Y}$ is continuous.

The maximum of the likelihood function and that of the log-likelihood function are at the same location because the logarithm is monotone.

Note that we are not assuming that the entries of $\mathbf{Y}$ are necessarily iid, as we did in parameter estimation.

If the joint distribution of X and $\mathbf{Y}$ is known a natural metric is the **probability of error**, i.e. the probability that an estimate chosen from the possible values in $\{x_1, \ldots, x_m\}$ is wrong. Note that for continuous models this probability always equals zero, so it does not make sense to use probability of error as a criterion.

Intuitively, to minimize the probability of making a mistake we should choose the element of $\{x_1, \ldots, x_m\}$ that has the *highest probability conditioned on the data*. This corresponds to the *mode* of the posterior distribution of the signal X given $\mathbf{Y}$, which is known as the **maximum-a-posteriori** (MAP) estimator.

Definition 2.2 (Maximum-a-posteriori estimator)**.** *Let X be a random variable with range restricted to the set $\{x_1, \ldots, x_m\}$ and let $\boldsymbol{Y}$ be a random vector belonging to the same probability space. The maximum-a-posteriori (MAP) estimator of X given $\boldsymbol{Y}$ is*

$$g_{\mathrm{MAP}}(\boldsymbol{y}) := \arg \max_{u \in \{x_1, \ldots, x_m\}} p_{X|\boldsymbol{Y}}(u). \tag{16}$$

Our intuition can be made precise. If our aim is to minimize the probability of error, then the MAP estimator is optimal.

Theorem 2.3 (MAP estimator minimizes probability of error)**.** *Let X be a random variable with range restricted to the set $\{x_1, \ldots, x_m\}$ and let $\boldsymbol{Y}$ be a random vector belonging to the same probability space. Then*

$$g_{\mathrm{MAP}}(\boldsymbol{y}) = \arg \min_{g(\boldsymbol{Y})} \mathrm{P}(X \neq g(\boldsymbol{Y})). \tag{17}$$

Proof. We show that the probability that $g_{MAP}(\mathbf{Y})$ equals X is an upper bound for the probability that any arbitrary estimator $g(\mathbf{Y})$ equals X,

$$\begin{aligned} \mathrm{P}(X = g(\mathbf{Y})) &= \int_{\mathbf{Y}} f_{\mathbf{Y}}(\mathbf{y}) \, \mathrm{P}\left(X = g(\mathbf{y}) \,\middle|\, \mathbf{Y} = \mathbf{y}\right) \, \mathbf{dy} && (18) \\ &= \int_{\mathbf{Y}} f_{\mathbf{Y}}(\mathbf{y}) \, p_{X|\mathbf{Y}}(g(\mathbf{y}) \,|\mathbf{y}) \, \mathbf{dy} && (19) \\ &\leq \int_{\mathbf{Y}} f_{\mathbf{Y}}(\mathbf{y}) \, p_{X|\mathbf{Y}}(g_{\mathrm{MAP}}(\mathbf{y}) \,|\mathbf{y}) \, \mathbf{dy}, && (20) \end{aligned}$$

by the definition of the MAP estimator. This obviously implies that

$$\mathrm{P}(X \neq g(\mathbf{Y})) \geq \mathrm{P}(X \neq g_{\mathrm{MAP}}(\mathbf{Y})), \tag{21}$$

so the MAP estimator achieves the lowest probability of error. □

Example 2.4 (Sending bits)**.** We consider a very simple model for a communication channel in which we aim to send a signal X consisting of a single bit. Our prior knowledge indicates that the signal is equal to one with probability 1/4.

$$p_X(1) = \frac{1}{4}, \qquad p_X(0) = \frac{3}{4}. \tag{22}$$

Due to the presence of noise in the channel, we send the signal repeatedly. At the receptor we observe

$$Y_i = X + Z_i, \quad 1 \leq i \leq n, \tag{23}$$

where $Z_1, Z_2, \ldots, Z_n$ are iid Gaussian random variables with mean zero and unit variance. Modeling perturbations as Gaussian is a popular choice in communications. It is justified by the Central Limit Theorem, the assumption being that the noise is a combination of many small effects that are approximately independent.

We will now compute and compare the ML and MAP estimators of X given the observations.

The likelihood is equal to

$$\mathcal{L}_{\mathbf{y}}(x) = \prod_{i=1}^{n} f_{Y_i|X}(y_i|x) \tag{24}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-x)^2}{2}}. \tag{25}$$

It is easier to deal with the log-likelihood function,

$$\log \mathcal{L}_{\mathbf{y}}(x) = -\sum_{i=1}^{n} \frac{(y_i - x)^2}{2} - \frac{n}{2} \log 2\pi. \tag{26}$$

Since X only takes two values, we can compare directly. We will choose $g_{\mathrm{ML}}(\mathbf{y}) = 1$ if

$$\log \mathcal{L}_{\mathbf{y}}(1) = -\sum_{i=1}^{n} \frac{y_i^2 - 2y_i + 1}{2} - \frac{n}{2} \log 2\pi \tag{27}$$

$$\geq -\sum_{i=1}^{n} \frac{y_i^2}{2} - \frac{n}{2} \log 2\pi \tag{28}$$

$$= \log \mathcal{L}_{\mathbf{y}}(0). \tag{29}$$

Equivalently,

$$g_{\mathrm{ML}}(\mathbf{y}) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n} y_i > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

The rule makes a lot of sense: if the sample mean of the data is closer to 1 than to 0 then our estimate is equal to 1. By the Law of Total Probability, the probability of error of this estimator is equal to

$$\begin{aligned} \mathrm{P}\left(X \neq g_{\mathrm{ML}}(\mathbf{y})\right) &= \mathrm{P}\left(X \neq g_{\mathrm{ML}}(\mathbf{y}) \,\middle|\, X = 0\right) \mathrm{P}(X = 0) + \mathrm{P}\left(X \neq g_{\mathrm{ML}}(\mathbf{y}) \,\middle|\, X = 1\right) \mathrm{P}(X = 1) \\ &= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} y_i > \frac{1}{2} \middle| X = 0\right) \mathrm{P}(X = 0) + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} y_i < \frac{1}{2} \middle| X = 1\right) \mathrm{P}(X = 1) \\ &= Q\left(\sqrt{n}/2\right), \end{aligned} \tag{31}$$

where the last equality follows from the fact that the sample mean is Gaussian with variance σ^2/n and mean equal to the value taken by X by Theorem 2.21 in Lecture Notes 3.

To compute the MAP estimate we compare the posterior distributions. Similarly to before, we compare the logarithm of these functions, which is equivalent as the logarithm is a monotone function.

$$\log p_{X|\mathbf{Y}}(x|\mathbf{y}) = \log \frac{\prod_{i=1}^n f_{Y_i|X}(y_i|x)\, p_X(x)}{f_{\mathbf{Y}}(\mathbf{y})} \tag{32}$$

$$= \sum_{i=1}^n \log f_{Y_i|X}(y_i|x)\, p_X(x) - \log f_{\mathbf{Y}}(\mathbf{y}) \tag{33}$$

$$= -\sum_{i=1}^n \frac{y_i^2 - 2y_i x + x^2}{2} - \frac{n}{2}\log 2\pi + \log p_X(x) - \log f_{\mathbf{Y}}(\mathbf{y}). \tag{34}$$

We compare the value of this function for $x = 0$ and $x = 1$. Note that the denominator does not depend on x so we can ignore it. We choose $g_{\text{MAP}}(\mathbf{y}) = 1$ if

$$\log p_{X|\mathbf{Y}}(1|\mathbf{y}) + \log f_{\mathbf{Y}}(\mathbf{y}) = -\sum_{i=1}^n \frac{y_i^2 - 2y_i + 1}{2} - \frac{n}{2}\log 2\pi - \log 4 \tag{35}$$

$$\geq -\sum_{i=1}^n \frac{y_i^2}{2} - \frac{n}{2}\log 2\pi - \log 4 + \log 3 \tag{36}$$

$$= \log p_{X|\mathbf{Y}}(0|\mathbf{y}) + \log f_{\mathbf{Y}}(\mathbf{y}). \tag{37}$$

Equivalently,

$$g_{\text{MAP}}(\mathbf{y}) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^n y_i > \frac{1}{2} + \frac{\log 3}{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

The MAP estimate takes into account that the signal is equal to zero more often. The correction term tends to zero as we gather more evidence. The probability of error of the estimator is equal to

$$\mathrm{P}\left(X \neq g_{\text{MAP}}(\mathbf{y})\right) = \mathrm{P}\left(X \neq g_{\text{MAP}}(\mathbf{y})\,|X=0\right)\mathrm{P}(X=0) + \mathrm{P}\left(X \neq g_{\text{MAP}}(\mathbf{y})\,|X=1\right)\mathrm{P}(X=1)$$

$$= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n y_i > \frac{1}{2} + \frac{\log 3}{n}\middle| X=0\right)\mathrm{P}(X=0) \tag{39}$$

$$+ \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n y_i < \frac{1}{2} + \frac{\log 3}{n}\middle| X=1\right)\mathrm{P}(X=1)$$

$$= \frac{3}{4}Q\left(\sqrt{n}/2 + \frac{\log 3}{\sqrt{n}}\right) + \frac{1}{4}Q\left(\sqrt{n}/2 - \frac{\log 3}{\sqrt{n}}\right). \tag{40}$$

We compare the probability of error of the ML and MAP estimates in Figure 1.

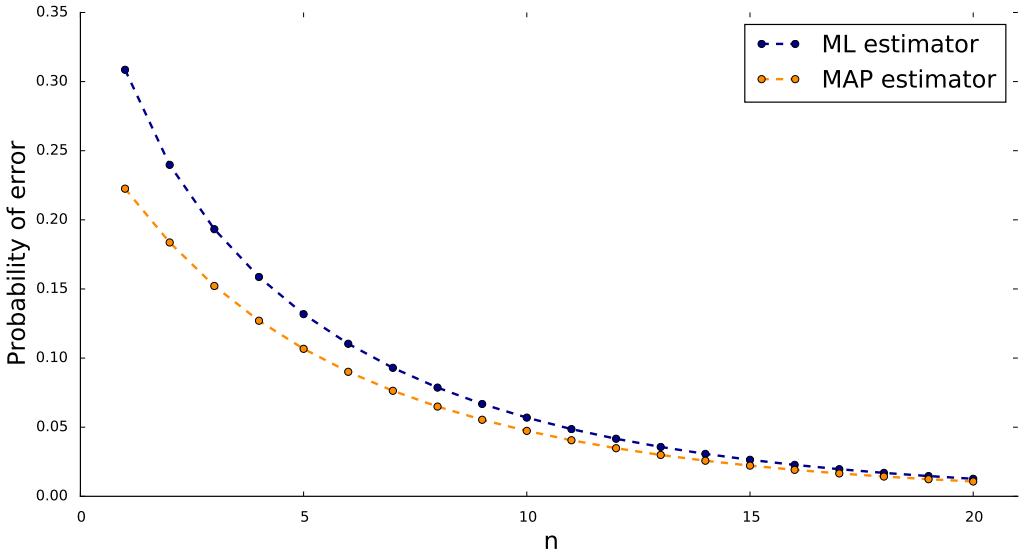

Figure 1: Probability of error of the ML and MAP estimators in Example 2.4 for different values of n.