Lecture 22: MLE of Vector parameters

17 April, 2018

Sunil Kumar Gauttam

Department of Mathematics, LNMIIT

Thus far we have been concerned with finding a MLE of single parameter. We encountered two parameters n, p when estimating the number of trials n of a binomial distribution, but we did not try to estimate both of those parameters simultaneously. Since we usually wish to estimate all the unknown parameters that determine a distribution, we need to study the problem of how find MLE of a vector parameter that contains at least two components.

Suppose the random variable X possesses a density that depends upon k parameters $\theta_1, \cdots, \theta_k$. We write $\theta = (\theta_1, \cdots, \theta_k)$. Let this density be denoted by $f(x|\theta_1, \cdots, \theta_k)$. Then the likelihood function corresponding to the random sample values $x_1, \cdots, x_n$ is defined by

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta_1, \cdots, \theta_k).$$

The maximum likelihood estimate of the vector parameter θ is the set of values $\hat{\theta}_1, \cdots, \hat{\theta}_k$ that maximizes $L(\theta)$. Since these values will depend upon the particular x_i's obtained in the sample, they are functions of the x_i's. The corresponding maximum likelihood estimators will therefore be denoted by

$$\begin{aligned} \hat{\theta}_1 &= \hat{\theta}_1(X_1, X_2, \cdots, X_n) \\ \vdots &= \vdots \\ \hat{\theta}_n &= \hat{\theta}_n(X_1, X_2, \cdots, X_n). \end{aligned}$$

As in the case of single parameter estimation, maximum likelihood estimates of vector parameters can often be obtained by calculus techniques, that is, by equating the partial derivatives of likelihood function to zero. However, since it is considerably more difficult in higher dimensions to determine whether a calculus critical point is a maximizing point for the function, we shall often dispense with second derivative calculus tests and use other arguments to justify the claim that a unique critical point obtained from the likelihood equations is a maximizing point for the function.

As an illustration of the maximum likelihood method for vector parameter estimation, consider the problem of estimating the mean and variance of a normal density.

Example 22.1 *Let $X_1, X_2, \cdots, X_n$ be iid $N(\mu, \sigma^2)$, with both μ and σ^2 are unknown. Then* *likelihood function is*

$$L(\mu,\sigma^2|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}e^{-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}}$$

Since we want to estimate σ^2 and hence need to differentiate w.r.t. σ^2, therefore we absorb σ in the square root sign and get σ^2 term.

$$L(\mu,\sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}}.$$

For convenience of notation let the variance σ^2 be denoted by θ. Then the likelihood function is

$$L(\mu,\theta) = \frac{1}{(2\pi\theta)^{\frac{n}{2}}}e^{-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\theta}}.$$

Taking logarithm

$$\log L(\mu,\theta) = \frac{-n}{2}\log(2\pi\theta) - \frac{\sum_{i=1}^n(x_i-\mu)^2}{2\theta} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta - \frac{\sum_{i=1}^n(x_i-\mu)^2}{2\theta}.$$

The admissible values of the vector parameter (μ,θ) is $(-\infty,\infty)\times(0,\infty)$. Computing the critical points of $\log L$

$$\frac{\partial L}{\partial\mu} = \frac{\sum_{i=1}^n(x_i-\mu)}{\theta},\ \frac{\partial L}{\partial\mu} = 0 \implies \sum_{i=1}^n(x_i-\mu) = 0 \implies \sum_{i=1}^n x_i = n\mu \implies \mu = \bar{x}$$

$$\frac{\partial L}{\partial\theta} = -\frac{n}{2\theta} + \frac{\sum_{i=1}^n(x_i-\mu)^2}{2\theta^2},\ \frac{\partial L}{\partial\theta} = 0 \implies \sum_{i=1}^n(x_i-\mu)^2 = n\theta \implies \theta = \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x})^2$$

To show that these values do maximize $\log L$*, it suffices to show that they minimize* $n\log\theta + \frac{1}{\theta}\sum_{i=1}^n(x_i-\mu)^2$ *(Because a maximizer of f would be a minimizer of $-f$ and vice-versa).*

Lemma 22.2 *Let $x_1, x_2, \cdots, x_n$ be any real numbers and $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$. Then*

$$\min_{a\in\mathbb{R}}\left\{\sum_{i=1}^n(x_i-a)^2\right\} = \sum_{i=1}^n(x_i-\bar{x})^2.$$

Proof: *Add and subtract $\bar{x}$ to get*

$$\begin{aligned}\sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}(x_i-a+\bar{x}-\bar{x})^2 &= \sum_{i=1}^{n}\left[(x_i-\bar{x})^2+(\bar{x}-a)^2+2(\bar{x}-a)(x_i-\bar{x})\right] \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-a)^2+2\sum_{i=1}^{n}(\bar{x}-a)(x_i-\bar{x}) \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-a)^2+2(\bar{x}-a)\sum_{i=1}^{n}(x_i-\bar{x}) \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-a)^2 \ (\because \textstyle\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_i^n x_i - n\bar{x} = n\bar{x}-n\bar{x} = 0)\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$. ■

Therefore we conclude that $\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2$ will be minimized when $\mu = \bar{x}$ regardless of the value of θ. It suffices therefore to choose θ to minimize $n\log\theta + \dfrac{c^2}{\theta}$, where $c^2 = \displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2$. The derivative of this function is $\dfrac{n}{\theta} - \dfrac{c^2}{\theta^2}$ which is negative for $\theta < \frac{c^2}{n}$ and positive for $\theta > \frac{c^2}{n}$, therefore $\theta = \frac{c^2}{n}$ minimizes this function. The pair of values

$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

consequently maximize the likelihood function.

Remark 22.3 *The above example tells us that a MLE is not necessarily an unbiased estimator, the MLE for variance is a biased estimator.*

Interval Estimation

So far we discussed point estimation of a real-valued parameter θ, where the inference is a guess of a single value as the value of θ. Now we discuss interval estimation. The inference in an interval estimation problem is the statement that "$\theta \in I$," where I subset of all admissible values of θ and $I = I(\mathbf{x})$ is an interval determined by the value of the data $\mathbf{X} = \mathbf{x}$ observed. Thus, instead of saying $\hat{\theta} = 34.24$ is the estimate of θ, we say that θ lies in the interval $[30.12, 37.98]$.

Definition 22.4 *An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \cdots, x_n)$ and $U(x_1, \cdots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in R(\mathbf{X})$ range of random sample $\mathbf{X}$. If $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an interval estimator.*

We write $[L(\mathbf{X}), U(\mathbf{X})]$ for an interval estimator of θ based on the random sample $\mathbf{X} = (X_1, \cdots, X_n)$ and $[L(\mathbf{x}), U(\mathbf{x})]$ for the realized value of the interval.

Example 22.5 *For a random sample X_1, X_2, X_3, X_4 from a $N(\mu, 1)$ distribution, an interval estimator of μ is $[\bar{X} - 1, \bar{X} + 1]$. This means that we will assert that μ in this interval.*

At this point, it is natural to inquire as to what is gained by using an interval estimator. Previously, we estimated μ with $\bar{X}$, and now we have the less precise estimator $[\bar{X} - 1, \bar{X} + 1]$. We surely must gain something! By giving up some precision in our estimate (or assertion about μ), we have gained some confidence, or assurance; that our assertion is correct.

When we estimate μ by $\bar{X}$ the probability that we are exactly correct, that is, $P(\bar{X} = \mu)$, is 0 $\left(\because \bar{X} = \dfrac{X_1 + \cdots + X_4}{4}\right.$ is sum of four independent normal random variables (because normality is preserved under linear transformation) hence $\bar{X}$ is $\left. N(\mu, \frac{1}{4}) \right)$. However, an interval estimator, we have a positive probability of being correct. The probability, that μ is covered by the interval $[\bar{X} - 1, \bar{X} + 1]$ can be calculated

$$P\left(\mu \in [\bar{X} - 1, \bar{X} + 1]\right) = P\left(\bar{X} - 1 \leq \mu \leq \bar{X} + 1\right) = P\left(-1 \leq \bar{X} - \mu \leq 1\right) = P\left(-2 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{1}{4}}} \leq 2\right)$$

$$= N(2) - N(-2) = N(2) - (1 - N(2)) = 2N(2) - 1 = 2 \times .9772 - 1 = 0.9544$$

Thus we have over a 95% chance of covering the unknown parameter with our interval estimator. Sacrificing some precision in our estimate, in moving from a point to an interval, has resulted in increased confidence that our assertion is correct.