Estimation and Optimality

Estimation of features of interest of many populations is very important in many areas. Average or median family income, average yield of an agricultural crop, proportion of eligible voters who favour a certain candidate and so on are some examples. In many fields this is done without making use of probabibility models for the data. With the use of appropriate models, efficient procedures can be developed for this. The first thing to realize then is that it becomes a model fitting problem where unknown parameters of the model are to be determined. This is parametric estimation. In other words, we assume that the data $\mathbf{X}$ comes from the model $\{P_\theta, \theta \in \Theta\}$. The first project is *model fitting*, which means we want to fit the best model to the data: choose $\theta \in \Theta$ which best describes the realization of $\mathbf{X}$. This is also known as point estimation to distinguish it from other procedures. The setup is as follows.

Point Estimation. Consider $X_1, \ldots, X_n$ i.i.d from P_θ. Estimate θ or $q(\theta)$. This is the simplest setup, and we will also consider $\mathbf{X} \sim P_\theta$ when it is not necessarily a random sample or i.i.d.

1. Method of moments. Let the population moments be
$\mu_r(\theta) = E_\theta(X^r)$, $r = 1, 2, \ldots$ and the sample moments be
$\hat{\mu}_r(\theta) = \frac{1}{n}\sum_{j=1}^n X_j^r$, $r = 1, 2, \ldots$. Suppose
$q(\theta) = g(\mu_1(\theta), \ldots, \mu_k(\theta))$ for $k \geq 1$, and g is continuous. Then the method of moments estimate of $q(\theta)$ is
$\widehat{q(\theta)} = g(\hat{\mu}_1(\theta), \ldots, \hat{\mu}_k(\theta))$.

Example. $\sigma^2 = Var_\theta(X) = E_\theta(X^2) - \{E_\theta(X)\}^2 = \mu_2(\theta) - \mu_1^2(\theta) = g(\mu_1(\theta), \mu_2(\theta))$ where $g(x, y) = y - x^2$ is continuous in (x, y). The method of moments estimate of $Var_\theta(X)$ is

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{j=1}^n X_j^2 - \left(\frac{1}{n}\sum_{j=1}^n X_j\right)^2 \\ &= \frac{1}{n}\left(\sum_{j=1}^n X_j^2 - n\bar{X}^2\right) = \frac{1}{n}\sum_{j=1}^n (X_j - \bar{X})^2, \end{aligned}$$

which is the sample variance with divisor n.

Example. $X_1, \ldots, X_n$ i.i.d Poisson(λ). Here $\theta = \lambda > 0$. Then $\mu_1(\theta) = \lambda$ and $\mu_2(\theta) = \lambda + \lambda^2$. Two different method of moments are readily available for λ. The one using only the first moment gives $\hat{\lambda}_1 = \hat{\mu}_1(\theta) = \bar{X}$, and

another using the first two gives $\hat{\lambda}_2 = \hat{\mu}_2(\theta) - \hat{\mu}_1^2(\theta) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2$. Normally one would use the first one, unless there was a need to check how good the Poisson model would be for the given data. Note that the mean and the variance are equal for the Poisson model. In many applications over-dispersion (i.e., variance larger than mean) is common suggesting other possibilities such as the negative binomial model.

It can be readily seen that the method of moments is basically a substitution method, where population moments are substituted by the corresponding sample moments. The idea of fitting a model is not stressed there. The idea of model fitting forms an important basis for the following method.

2. Maximum likelihood estimation

This requires consideration of a concept of fundamental importance called the *likelihood function*.

Likelihood function. Let $\mathbf{x}$ be the observed data; $\{P_\theta, \theta \in \Theta\}$ with density $f(\mathbf{x}|\theta)$ is the model under consideration for model fitting. Then the function $L(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)$, regarded as a function of θ for fixed $\mathbf{x}$ is called the likelihood function. Often $\mathbf{x}$ is suppressed and f is taken as the likelihood function and written $L(\theta)$.

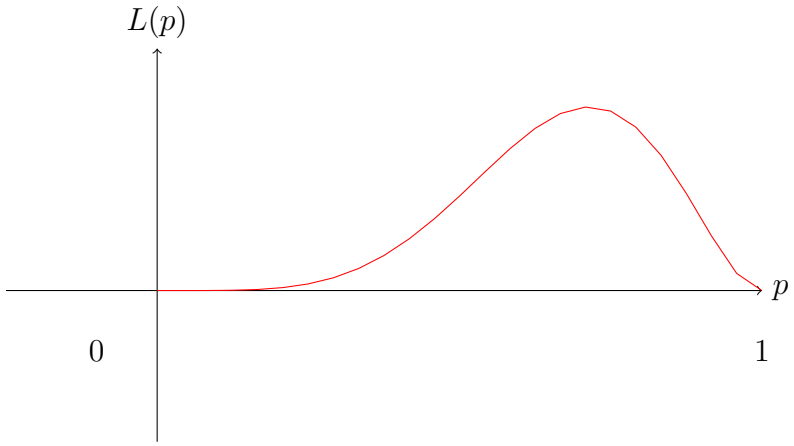
Interpretation of the likelihood function as relevant for inference about θ is the following. The data, $\mathbf{x}$, has been observed already, so θ is the only unknown. Then it makes sense to assume (according to a principle called *likelihood principle*), that all information about θ is contained in $L(\theta)$ for the observed $\mathbf{x}$. Since $f(\mathbf{x}|\theta)$ measures how likely $\mathbf{x}$ is if θ is the true parameter, observing $\mathbf{x}$ must then provide information through $L(\theta)$, on how to regard θ as the true parameter. The likelihood function is not unique in that for any $c(\mathbf{x}) > 0$ that may depend on $\mathbf{x}$ but not on θ, $c(\mathbf{x})f(\mathbf{x}|\theta)$ is also a likelihood function. What is unique are the likelihood ratios $L(\theta_\mathbf{2})/L(\theta_\mathbf{1})$, which indicate how plausible is $\theta_\mathbf{2}$, relative to $\theta_\mathbf{1}$, in the light of the given data $\mathbf{x}$. In particular, if the ratio is large, we have a lot of confidence in $\theta_\mathbf{2}$ relative to $\theta_\mathbf{1}$ and the reverse situation holds if the ratio is small.

Maximum likelihood estimate (MLE). MLE of θ is $\hat{\theta} = \hat{\theta}(x)$ where $L(\hat{\theta}, x) = \max_{\theta\in\Theta} L(\theta, x)$ if the maximum exists.
(i) MLE may not exist, or may not be unique; (ii) if $\hat{\theta}$ is the MLE of θ, then $q(\hat{\theta})$ is taken to be the MLE of $q(\theta)$.

Example. Let $X_1, \ldots, X_n$ be i.i.d Bernoulli(p). Then

$$L(p) = f(x_1, \ldots, x_n|p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

The MLE of p is $\hat{p} = \sum_{i=1}^{n} x_i/n$ as can be seen from the graph of $L(p)$ and using calculus:



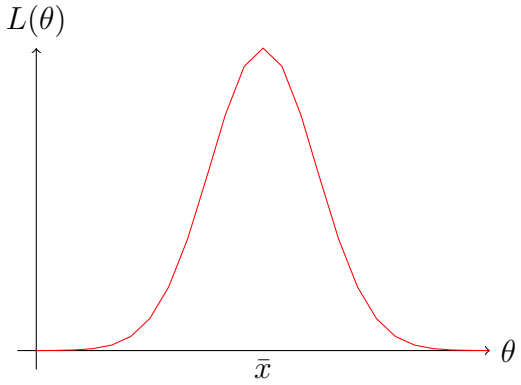
Result. MLE depends on $\mathbf{x}$ only through the sufficient statistics $T(\mathbf{x})$.

Proof. $L(\theta, x) = f(x|\theta) = g(T(x), \theta)h(x)$. Therefore, we have $L(\hat{\theta}(x), x) = \max_\theta g(T(x), \theta)h(x)$. Since $h(x) > 0$, we must have $L(\hat{\theta}(x), x) = h(x) \max_\theta g(T(x), \theta)$, where the maximization is on the part that invoves x through $T(x)$ only.

Example. Let $X_1, \ldots, X_n$ be i.i.d $N(\theta, 1)$. What is the MLE of θ? Sufficient statistic is $\bar{X} \sim N(\theta, 1/n)$. Then

$$L(\theta, x_1, \ldots, x_n) \propto f(\bar{x}|\theta) \propto \exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right),$$

which is maximized by $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$.



If the sample size n is large, usually the likelihood function has a sharp peak as shown in the following figure. This peak is at the maximum likelihood

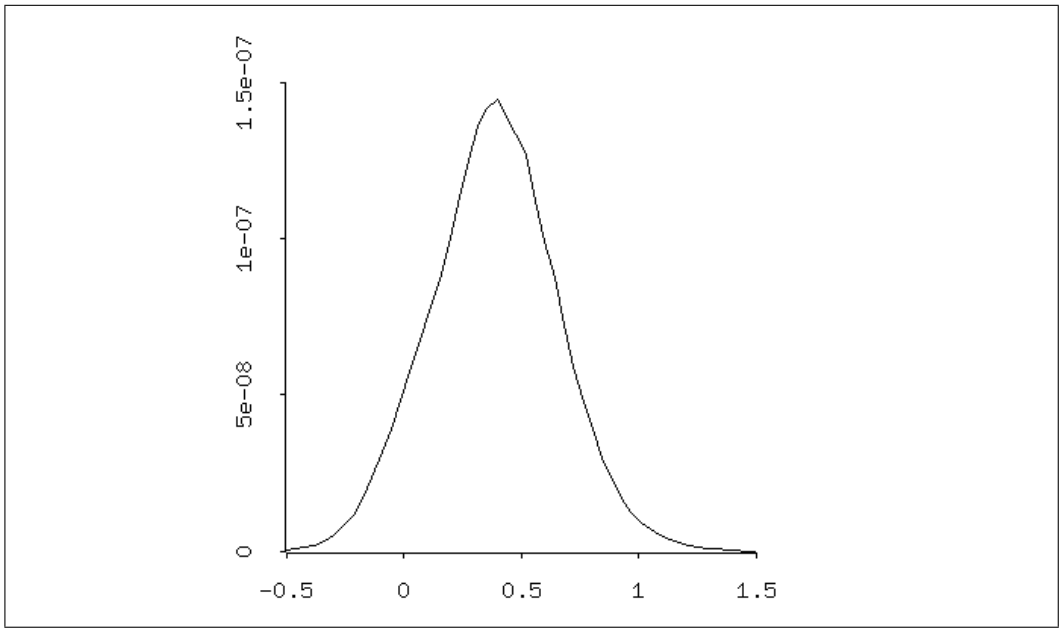


Figure 1: $L(\theta)$ for the double exponential model when data is normal mixture.

estimate (MLE) $\hat{\theta}$. In situations like this, one feels $\hat{\theta}$ is very plausible as an estimate of θ relative to any other points outside a small interval around $\hat{\theta}$. One would then expect $\hat{\theta}$ to be a good estimate of the unknown θ, at least in the sense of being close to it in some way.