IEOR 165 – Lecture 6
Maximum Likelihood Estimation

1 Motivating Problem

Suppose we are working for a grocery store, and we have decided to model service time of an individual using the express lane (for 10 items or less) with an exponential distribution. An exponential service time is a common assumption in basic queuing theory models. Recall that: A random variable X with exponential distribution is denoted by $X \sim \mathcal{E}(\lambda)$, where $\lambda > 0$ is the *rate*; in our context, the rate is better understood to be the *service rate*. The exponential distribution has pdf

$$f_X(u) = \begin{cases} \lambda \exp(-\lambda u), & \text{if } u \geq 0, \\ 0 & \text{otherwise} \end{cases}.$$

In our motivating problem, suppose the data X_i for $i = 1, \ldots, n$ are assumed to be iid measurements of the service time.

1.1 Method of Moments Estimator

A first idea is to use the method of moments estimator, which we derived in the second lecture. The intuition behind the method of moments estimator is that the mean of an exponential distribution is $1/\lambda$, and so we estimate the service rate using the equation

$$\hat{\lambda} = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^{-1}.$$

This is justified by the law of large numbers, which states that the sample average converges to the expectation as we get more data. And since the service rate is assumed to be $\lambda > 0$ in the definition of an exponential random variable, it turns out that the inverse of the sample average will converge to the service rate λ. The potential weakness of this approach is that though it will eventually give us the correct answer, we are not using information we know about the underlying distribution of the data; it may be possible to get a more accurate estimate of the service rate using less data with a more sophisticated approach.

1.2 Alternative Approach

We first define a parametric distribution $\mathbb{P}_\theta$ with pdf $f_\theta(u)$, to be a distribution that depends on a vector of parameters $\theta \in \mathbb{R}^p$. Observe that the notation is slightly different here, in that we use a subscript θ to denote the parameters in the pdf. We have already seen three examples of a parametric distribution:

- $X \sim \mathcal{N}(\theta_1, \theta_2)$, where θ_1 is an unknown mean of the Gaussian and θ_2 is the unknown variance of the Gaussian;
- $X \sim \mathcal{E}(\theta_1)$, where θ_1 is the rate parameter of the exponential distribution.
- X has a gamma distribution, which has a pdf of

$$f_{(\theta,k)}(u) = \frac{1}{\Gamma(k)\theta^k} u^{k-1} \exp(-u/\theta),$$

  where $\Gamma(\cdot)$ is the Gamma function, and θ, k are the parameters of this distribution.

We define the likelihood function for a parametric distribution $\mathbb{P}_\theta$ with pdf $f_\theta(u)$ as the substitution of measured data x_i for $i = 1, \ldots, n$ into the joint pdf for the random variables X_i for $i = 1, \ldots, n$. However, when the X_i are iid, the joint pdf simplifies into the product of the individual pdf of each random variable X_i. More precisely, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

One alternative idea for constructing an estimator is to choose the value of θ that maximizes this likelihood. In order to be able to solve such likelihood maximization problems, we have to first discuss general optimization problems.

2 Nonlinear Programming

Consider the following optimization problem (P):

$$\begin{aligned} \min\ & f(x) \\ \text{s.t. } & x \in \mathbb{R}^n \\ & g_i(x) \leq 0, \forall i = 1, \ldots, m \\ & h_i(x) = 0, \forall i = 1, \ldots, k \end{aligned}$$

where $f(x), g_i(x), h_i(x)$ are continuously differentiable functions. We call the function $f(x)$ the objective, and we define the feasible set $\mathcal{X}$ as

$$\mathcal{X} = \{x \in \mathbb{R}^n \ : \ g_i(x) \leq 0 \ \forall i = 1, \ldots, m \wedge h_i(x) = 0 \ \forall i = 1, \ldots, k\}.$$

Note that this formulation also incorporates maximization problems such as $\max\{f(x) \mid x \in \mathcal{X}\}$ through rewriting the problem as $\min\{-f(x) \mid x \in \mathcal{X}\}$.

2.1 Unconstrained Optimization

First consider the special case where $f : \mathbb{R} \to \mathbb{R}$. When (P) does not have any constraints, we know from calculus (specifically Fermat's theorem) that the global minimum must occur at points where either (i) the slope is zero $f'(x) = 0$, (ii) at $x = -\infty$, or (iii) at $x = \infty$. More rigorously, the theorem states that if $f'(x) \neq 0$ for $x \in \mathbb{R}$, then this x is not a local minimum. This result extends naturally to the multivariate case where $f : \mathbb{R}^n \to \mathbb{R}$, by insteading looking for points where the gradient is zero $\nabla f(x) = 0$.

This result is useful because it gives one possible approach to solve (P) in the case where there are no constraints: We can find all points where the gradient is zero, evaluate the function in the limits as x tends to $x = -\infty$, or $x = \infty$, and then select the minimum amongst these points.

One natural question to ask is how can we extend this approach to the general case (P)? The most general case is more complex than is needed for this class, and so we look at some special cases of constrained optimization. In particular, we consider two types of convex optimization.

2.2 Convex Optimizaton

Convex optimization problems are a useful framework because these problems can often be solved in polynomial time. Below, we discuss two common and important types of convex optimization problems: linear programs (LP's) and quadratic programs (QP's). These two classes of problems are important because there exist many software packages that can numerically solve these problems in polynomial time. An important point to note is that it is often the case that an optimization problem does not originally look like an LP or a QP, but that it can be rewritten as an LP or QP by rearranging the terms of the optimization problem.

2.2.1 Linear Programs

If f, g_i, h_i are affine functions (i.e., of the form $c'x + d$ where $x \in \mathbb{R}^n$ is the input, $c \in \mathbb{R}^n$ is a constant vector, and $d \in \mathbb{R}$ is a constant), then the optimization problem (P) is a linear program:

$$\begin{aligned} \min\ & c'x + d \\ \text{s.t. } & x \in \mathbb{R}^n \\ & g_i'x + e_i \leq 0, \forall i = 1, \ldots, m \\ & h_i'x + f_i = 0, \forall i = 1, \ldots, k \end{aligned}$$

2.2.2 Quadratic Programs

If (i) f is a convex quadratic function (i.e., of the form $f(x) = x'Qx + c'x + d$, where $x \in \mathbb{R}^n$ is the input, $Q \in \mathbb{R}^{n \times n} \geq 0$ is a positive semidefinite matrix, and g_i, h_i are affine functions), then

the optimization problem (P) is a quadratic program:

$$\begin{aligned} \min\ & x'Qx + c'x + d \\ \text{s.t. } & x \in \mathbb{R}^n \\ & g_i'x + e_i \leq 0, \forall i = 1, \ldots, m \\ & h_i'x + f_i = 0, \forall i = 1, \ldots, k \end{aligned}$$

It is important that the matrix Q be positive semidefinite (i.e., $Q = Q'$ and $x'Qx \geq 0$ for all $x \in \mathbb{R}^n$) because otherwise the optimization problem is NP-hard if Q is not positive semidefinite.

3 Maximum Likelihood Estimation

The idea of maximum likelihood estimation (MLE) is to generate an estimate of some unknown parameters by solving the following optimization problem

$$\max L(\theta) = \max \prod_{i=1}^{n} f_\theta(x_i).$$

The MLE approach is best explained by considering some examples.

3.1 Example: Service Rate of Exponential Distribution

To get a better understanding of how this approach works, consider the case of using MLE to estimate the service rate of an exponential. The likelihood is

$$L(\lambda) = \prod_{i=1}^{n} \left(\lambda \exp(-\lambda x_i)\right) = \lambda^n \cdot \exp(-\lambda \textstyle\sum_{i=1}^{n} x_i).$$

And we would like to solve the problem

$$\max\ \lambda^n \cdot \exp(-\lambda \textstyle\sum_{i=1}^{n} x_i).$$

Next, suppose we find the points at which the gradient is zero:

$$n\lambda^{n-1} \exp(-\lambda \textstyle\sum_{i=1}^{n} x_i) + -\lambda^n \exp(-\lambda \textstyle\sum_{i=1}^{n} x_i) \cdot \textstyle\sum_{i=1}^{n} x_i = 0.$$

Now assuming $\lambda > 0$, we can simplify this to

$$n + -\lambda \cdot \textstyle\sum_{i=1}^{n} x_i = 0 \Rightarrow \lambda^* = \left(\frac{1}{n} \textstyle\sum_{i=1}^{n} x_i\right)^{-1}.$$

Since the random variables satisfy $x_i > 0$ when drawn from an exponential distribution, this estimate is strictly positive. The objective at $\lambda = 0$ is zero (i.e., $L(0) = 0$), and the objective as $\lambda \to \infty$ is also zero (i.e., $\lim_{\lambda \to \infty} L(\lambda) = 0$). Consequently, the above solution must be the global maximum.

Observe that in this case the MLE estimate is equivalent to the inverse of the sample average. However, in other cases the MLE estimate can differ from the estimates we previously considered.

3.2 Transformation of the Objective Function

In some cases, it can be easier to solve a transformed version of the optimization problem. For instance, two transformations that are useful in practice are:

1. instead of maximizing the likelihood, we could equivalently minimize the negative of the likelihood;

2. instead of maximizing the likelihood, we could equivalently maximize the likelihood composed with a strictly increasing function.

In fact, one very useful transformation is the minimize the negative log-likelihood. The utility of such transformations is most clearly elucidated by an example.

3.3 Example: Mean and Variance of Gaussian

Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$ are iid measurements from a Gaussian with unknown mean and variance. Then the MLE estimate is given by the solution to

$$\max\ (2\pi\sigma^2)^{-n/2} \exp\big(\textstyle\sum_{i=1}^n -(x_i-\mu)^2/(2\sigma^2)\big).$$

If we instead minimize the negative log-likelihood, then the problem we would like to solve is

$$\min\ (n/2)\log(2\pi\sigma^2) + \textstyle\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2).$$

If we consider the objective function $h(\mu, \sigma^2) = (n/2)\log(2\pi\sigma^2) + \sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)$ (that is we are treating the σ^2 as a single variable), then setting its gradient equal to zero gives

$$\begin{bmatrix} h_\mu \\ h_{\sigma^2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n 2(x_i-\mu)/(2\sigma^2) \\ n/(2\sigma^2) - \sum_{i=1}^n (x_i-\mu)^2/(2(\sigma^2)^2) \end{bmatrix} = 0,$$

where h_μ and h_{σ^2} denote the partial derivatives of $h(\mu, \sigma^2)$ with respect to μ and σ^2, respectively. To solve this, we first begin with

$$h_\mu = \textstyle\sum_{i=1}^n 2(x_i-\mu)/(2\sigma^2) = 0 \Rightarrow \sum_{i=1}^n (x_i-\mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i,$$

where we have made use of the fact that $\sigma^2 > 0$. Next, note that

$$h_{\sigma^2} = n/(2\sigma^2) - \textstyle\sum_{i=1}^n (x_i-\hat{\mu})^2/(2(\sigma^2)^2) = 0 \Rightarrow n\sigma^2 = \sum_{i=1}^n (x_i-\hat{\mu})^2 \Rightarrow \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})^2,$$

where we multiplied by $(\sigma^2)^2$ on the second step.

The estimate of the mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$ is simply the sample average; however, the estimate of the variance $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})^2$ is a biased estimator of the variance (cf. the unbiased estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i-\hat{\mu})^2$).

3.4 Example: Linear Function of Normal Distribution

As another example, suppose that $y_i = x_i'\beta + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid with unknown σ^2. The key observation is that $\epsilon_i = y_i - x_i'\beta$, and so the $y_i - x_i'\beta$ are iid $\mathcal{N}(0, \sigma^2)$.

Consequently, the MLE estimate of β is given by

$$\max\ (2\pi\sigma^2)^{-n/2} \exp\big(\textstyle\sum_{i=1}^n -(y_i - x_i'\beta)^2/(2\sigma^2)\big).$$

Taking the logarithm of the objective function, we get

$$\hat{\beta} = \arg\max\ \sum_{i=1}^n \Big(-(y_i - x_i'\beta)^2/(2\sigma^2) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi)\Big).$$

Note that the maximizer of this optimization problem does not depend on σ^2 or the constant $-\frac{1}{2}\log(2\pi)$. And so simplifying this, we have that

$$\hat{\beta} = \arg\max_\beta\ \sum_{i=1}^n -(y_i - x_i'\beta)^2 = \arg\min_\beta \sum_{i=1}^n (y_i - x_i'\beta)^2 = \arg\min_\beta \|Y - X\beta\|_2^2.$$

This is equivalent to the OLS estimate!

One dangling point remains: We still have to provide an estimate of σ^2. In some applications, we do not care about the variance of the noise. In these cases, this variance is known as a *nuisance parameter*. But suppose we have an application where we care about the variance of the noise. Then, we can finish computing the MLE estimate for the noise. In particular, suppose we set the derivative of the negative log-likelihood with respect to σ^2 (treated as a variable) equal to zero; then, this gives

$$\frac{n}{2\sigma^2} - \frac{\|Y - X\hat{\beta}\|_2^2}{2(\sigma^2)^2} = 0 \Rightarrow \hat{\sigma^2} = \frac{1}{n}\|Y - X\hat{\beta}\|_2^2.$$

3.5 Example: Linear Function of Laplacian Distribution

Gaussian noise is not the only possibility for the measurement error in a linear model. Some noise distributions are called "long-tailed" because the probability of seeing extreme values for these distributions is greater than that for the Gaussian distribution. A canonical example is the Laplacian distribution, which is defined as a random variable with pdf given by

$$f(u) = \frac{1}{2b}\exp\big(-|u - \mu|/b\big),$$

where μ is the *location* parameter and $b > 0$ is a *scale* parameter. The mean of this distribution is μ, and the variance is $2b^2$.

It is interesting to consider what the MLE estimate for the parameters of a linear model with Laplacian noise looks like. Specifically, this is the setting where the model is

$$y_i = x_i'\beta + \epsilon_i,$$

and the noise is described by a zero-mean Laplacian distribution with unknown scale parameter b. The likelihood for pairs of independent measurements (x_i, y_i) for $i = 1, \ldots, n$ is given by

$$L(\beta, b) = (2b)^{-n} \exp\big(-\textstyle\sum_{i=1}^n |y_i - x_i'\beta|/b\big).$$

So the MLE is given by the minimizer to the negative log-likelihood

$$\min \;\; n\log(2b) + \sum_{i=1}^n |y_i - x_i'\beta|/b.$$

Now observe that the minimum of $\sum_{i=1}^n |y_i - x_i'\beta|$ is independent of b. And so the MLE estimate of the parameters is equivalently given by the solution to

$$\hat{\beta} = \arg\min \sum_{i=1}^n |y_i - x_i'\beta|.$$

If we define the matrix $X \in \mathbb{R}^{n\times p}$ and a vector $Y \in \mathbb{R}^n$ such that the i-th row of X is x_i' and the i-th row of Y is y_i, then this problem can be rewritten as

$$\hat{\beta} = \arg\min \|Y - X\beta\|_1,$$

where $\|\cdot\|_1$ is the usual L^1-norm. (Recall that for a vector $v \in \mathbb{R}^k$ the L^1-norm is $\|v\|_1 = |v^1| + \ldots + |v^k|$.) There is no closed-form expression for the solution to this optimization problem. Fortunately, this optimization problem is a specific case of an LP, and so there are efficient algorithms to solve this problem.

Lastly, we can ask what is the estimate of the nuisance parameter b? Setting the derivative with respect to b of the negative log-likelihood equal to zero and then solving gives

$$\frac{n}{b} - \frac{\|Y - X\hat{\beta}\|_1}{b^2} = 0 \Rightarrow \hat{b} = \frac{1}{n}\|Y - X\hat{\beta}\|_1.$$