7.6 Properties of Maximum Likelihood Estimators

Invariance

Ex 1: A reparameterization refers to a change in the indexing set for a model. Consider the family of exponential distributions, which is classically parameterized by the rate λ, where $X \sim \lambda$ means that the density of X is

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0$$

The mean of X is the reciprocal of the rate parameter: $E[X] = \frac{1}{\lambda}$.

But note that if we are told the value of the mean of X, we can determine the rate parameter and specify the distribution of X. For example if, $E[X] = 2$, then $\lambda = \frac{1}{2}$ and $X \sim \text{Expo}(1/2)$.

Or more generally, if $E[X] = \mu$, then $\lambda = \frac{1}{\mu}$ and $X \sim \text{Expo}(1/\mu)$. But this means that the family of exponential distributions can be parameterized by the mean μ: specifying a value of μ specifies the particular density function of X.

Note that while every exponential distribution can be specified either by the mean μ or by the rate λ, identical values of μ and λ do not give identical distributions. If $\mu = 2$, then the density for X is

$$f(x) = \frac{1}{2}e^{-x/2}$$

while if $\lambda = 2$, then

$$f(x) = 2e^{-2x}.$$

Maximum Likelihood Estimators have an important **invariance property**:

Thm: If $\hat{\theta}$ is the MLE of θ and g is a one-to-one function, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Proof. Let $\Gamma = g(\Omega)$. Since g is one-to-one, it has an inverse h on Γ. The likelihood function for $\psi = g(\theta)$ is conditional distribution of $\mathbf{x}$ given ψ. But we know the conditional distribution of $\mathbf{x}$ given θ: $f(\mathbf{x}|\theta)$, and since $\theta = h(\psi)$, the likelihood function for ψ is

$$f(\mathbf{x}|h(\psi))$$

This function is maximized when $\theta = \hat{\theta} = h(g(\hat{\theta}))$, and so is maximized when $g(\theta) = g(\hat{\theta})$. □

Essentially, what this property means is that the MLE estimator associated to a likelihood function is invariant under reparameterization.

Ex 2: Suppose $X_1, \ldots, X_n$ are conditionally iid Pois(θ). Find the MLE for $p = P(X_i = 0)$.

Solution. Note that $p = P(X_i = 0) = e^{-\theta}$, so by the invariance principal, it suffices to find the MLE for θ. Note that the likelihood and log likelihood functions for θ are

$$f(\mathbf{x}|\theta) \propto e^{-n\theta}\theta^{x_1+\cdots+x_n} \qquad \log f(\mathbf{x}|\theta) = c - n\theta + (x_1 + \cdots + x_n)\log\theta$$

Differentiating

$$\frac{\partial}{\partial\theta}\log f = -n + \frac{x_1 + \cdots + x_n}{\theta}$$

which has a zero at $\theta = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}$. Hence, the MLE for p is $\hat{p} = e^{-\bar{x}}$. ■

We would like to extend this result to arbitrary functions g. But one problem is that if $g(\theta)$ is not one-to-one, and the statistical model is parameterized by θ, then the likelihood function for $g(\theta)$ isn't well-defined. To rectify, we introduce a more general notion of likelihood function:

Def: Let $g(\theta)$ be an arbitrary function of the parameter, and let $\Gamma = g(\Omega)$. For each $\gamma \in \Gamma$, let $T_\gamma = g^{-1}(\gamma) = \{\theta : g(\theta) = \gamma\}$. Define the **induced log-likelihood function** $L^*(\gamma)$ by

$$L^*(\gamma) = \max_{\theta \in T_\gamma} \log f(\mathbf{x}|\theta)$$

Define the MLE of $g(\theta)$ to be $\hat{\gamma}$ where

$$\hat{\gamma} = \arg\max_{\gamma \in \Gamma} L^*(\gamma)$$

Thm: Let $\hat{\theta}$ be an MLE of θ and let $g(\theta)$ be a function of θ. Then an MLE of $g(\theta)$ is $g(\hat{\theta})$

Ex 3: Suppose $X_1, \ldots, X_n$ are a random sample from Bern(θ), a distribution with mean θ and variance $\nu = \theta(1-\theta)$. If $\hat{\theta}$ is the MLE for θ, then $\hat{\theta}(1-\hat{\theta})$ is an MLE for ν.

Ex 4: Previously, we showed the that the MLE for the parameter $\theta = (\mu, \sigma^2)$ in the model $X_i \sim N(\mu, \sigma)$ is

$$\hat{\theta} = \left(\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)$$

Let g be the function $g(\mu, \sigma^2) = \sigma^2$. Note that g is not invertible, since multiple different values of μ get sent to the same value of σ^2. However, if we are interested in the MLE of σ^2 alone (not the MLE of the pair of parameter $\theta = (\mu, \sigma^2)$) we can use the invariance property of the MLE to see that

$$\widehat{(\sigma^2)} = g(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Additionally, if we are interested in the standard deviation parameter $\sigma = \sqrt{\sigma^2}$, then the MLE of σ is

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

For this reason, when talking about MLEs, we'll write $\hat{\sigma}^2$ and note specify whether we first computed the MLE of σ and squared it, or computed the MLE of σ^2 directly.

Not every estimator has the invariance property.

Ex 5: In general, the Bayes estimator with squared loss does not have the invariance property. Let θ be a parameter and let g be a convex function (for example, suppose $g(x) = x^2$). Let $\psi = g(\theta)$. By definition, the Bayes estimators for θ is $E[\theta|\mathbf{X}]$ and for ψ is $E[\psi|\mathbf{X}] = E[g(\theta)|\mathbf{X}]$. By Jensen's inequality, since g is a convex function, then

$$E[g(\theta)|\mathbf{X}] \geq g\left(E[\theta|\mathbf{X}]\right)$$

where equality is attained if and only if $g(\theta)$ is a linear function for fixed value of X.

Hence, in most cases, the Bayes estimator does not have the invariant property.

Consistency

Suppose we are interested in estimating a parameter θ using a particular sampling framework. Ideally, our estimate of θ should get closer to θ as the sample size increases. We might even say that in the limit, the estimator should equal the parameter. However, since the observed data is random, we need to be careful how we describe this limit.

Def: A sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a number c if and only if for every $\delta > 0$,

$$\lim_{n\to\infty} P(|X_n - c| > \delta) = 0$$

We've seen this definition once before, in the Weak Law of Large Numbers:

Thm: If $X_1, \ldots, X_n$ are iid with mean μ, then the sequence of sample means $\bar{X}_1, \bar{X}_2, \ldots$ converges in probability to μ.

Def: A sequence of estimators $\delta_n(\mathbf{X})$ of θ is **consistent** provided the sequence converges in probability to θ.

Thm: Let $X_1, X_2, \ldots$ be iid Bern(θ). For each n, let $\mathbf{X}_n = (X_1, \ldots, X_n)$ and let $\delta_n(\mathbf{X}_n)$ be the MLE of θ. Then the sequence $\delta_1(\mathbf{X}_1), \delta_2(\mathbf{X}_2), \delta_3(\mathbf{X}_3) \ldots$, is a consistent sequence of estimators.

Proof. Recall that the MLE of θ with n observations is

$$\delta_n(\mathbf{X}) = \frac{X_1 + \cdots + X_n}{n}$$

the sample mean. By the Weak Law of Large Numbers, we know that the sequence of sample means converges in probability to the mean of the data $E[X_1] = \theta$, which shows that the sequence of sample means is consistent. But we can also show this directly for extra practice. Let $\delta > 0$. Then

$$P(|\delta_n(\mathbf{X_n}) - \theta| > \theta) = P(|\bar{X}_n - \theta| > \theta) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = \frac{\theta(1-\theta)}{n\delta^2}$$

But this latter expression goes to 0 as $n \to \infty$, which shows that $\delta_n(\mathbf{X}_n)$ converges in probability to θ. □

Bias

While under reasonable assumptions, the maximum likelihood estimator is consistent and has the invariance property, it can be biased:

Def: An estimator $\delta(\mathbf{X})$ of $g(\theta)$ is **unbiased** if

$$E[\delta(\mathbf{X})] = g(\theta)$$

Ex 6: The MLE for variance in a Normal distribution is biased.

Solution. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Recall that the MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

Additionally, recall the square decomposition identity:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = n(\bar{X} - \mu)^2 + \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

which can be restated as

$$n\hat{\sigma}^2 = \sum (X_i - \bar{X})^2 = \left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\bar{X} - \mu)^2$$

By taking expectation of both sides and using linearity of expectation, along with the definition of variance, we have

$$\begin{aligned} E[n\hat{\sigma}^2] &= \left(\sum_{i=1}^{n} E\left[(X_i-\mu)^2\right]\right) - nE[(\bar{X}-\mu)^2] \\ &= \left(\sum_{i=1}^{n} \text{Var}(X_i)\right) - n\text{Var}(\bar{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2. \end{aligned}$$

Therefore,

$$E[\hat{\sigma}^2] = \frac{n_1}{n}\sigma^2 < \sigma^2$$

showing that $\hat{\sigma}^2$ is a biased estimator of σ^2. ■

In particular, this result shows that the sample variance $\hat{\sigma}^2$ tends to underestimate the population variance σ^2. Why? Because the data in the same tend to be closer on average to the sample mean, than data in the population are to the population mean.

Consider again the square decomposition identity:

$$\frac{1}{n}\sum(X_i-\mu)^2 = (\bar{X}-\mu)^2 + \frac{1}{n}\sum(X_i-\bar{X})^2$$

which implies that

$$\frac{1}{n}\sum(X_i-\mu)^2 \geq \frac{1}{n}\sum(X_i-\bar{X})^2$$

with equality attained only when $\bar{X} = \mu$.

Is it possible to find an unbiased estimator of the population variance? Yes, using what's called the **Bessel's Correction**.

So, should this mean we should use s^2 instead of $\hat{\sigma}^2$ to estimate σ^2? Probably not.

1. While s^2 is unbiased, it is not the MLE, and therefore, does not have the nice theoretical properties and justification as $\hat{\sigma}^2$.
2. In HW 4, you will show that the mean squared error of s^2 is actually higher than the mean squared error of $\hat{\sigma}^2$. This means that while on average, the value of s^2 is σ^2, it will still tend to be further away from σ^2 than $\hat{\sigma}^2$.
3. For n relatively large, the difference between $\hat{\sigma}^2$ and s^2 is very small, and so it doesn't matter in practice which you use. However, if n is small, you will usually have bigger problems in estimation than choosing whether to use the biased or unbiased estimator.

Finally, note that while $\hat{\sigma}^2$ is a biased estimator, it is still consistent:

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum(X_i-\mu)^2\right) - (\bar{X}-\mu)^2$$

The left expression converges in probability to σ^2 and the right expression converges in probability to 0, by the weak law of large numbers.