STA355H1
Excerpt From Personal Notes

Omerzahid Ali

November 16, 2024

Contents

1 Data Reduction

1.1 Maximum Likelihood Estimators

1.1.1 Definition

Suppose $(X_1, \ldots, X_n)$ are random variables with joint pdf/pmf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k)$ where $\theta_1, \ldots, \theta_k$ are unknown. Then, given data $x_1, \ldots, x_n$, we can define the likelihood function.

Definition 1.1 (Likelihood Function)**.**

$$\mathcal{L}(\theta_1, \cdots, \theta_k) = f(\underbrace{x_1, \cdots, x_n}_{data}; \theta_1, \cdots, \theta_k)$$

We commonly refer to the natural log of the likelihood, $\ln \mathcal{L}(\theta)$, as the log-likelihood.

Example 1.1. **Model**: $X_1, \cdots, X_n$ independent random variables with pdf

$$f(x; \theta) = \frac{|x - \theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x - \theta|\right).$$

The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left\{ \frac{|x_i - \theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x_i - \theta|\right) \right\}.$$

□

Definition 1.2 (Maximum Likelihood Estimator)**.** Suppose that for each $\mathbf{x} = (x_1, \ldots, x_n)$, $(T_1(x), \ldots, T_k(x))$ maximize $\mathcal{L}(\theta_1, \ldots, \theta_k)$. Then the maximum likelihood estimators (MLEs) of $\theta_1, \ldots, \theta_k$ are

$$\hat{\theta}_j = T_j(X_1, \ldots, X_n) \text{ for } j = 1, \ldots, k.$$

To compute such an estimator, it depends on our parameter space Θ. If $\mathcal{L}(\theta)$ is differentiable, Θ is an open set, and an MLE exists, it satisfies the **likelihood equation**:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(\hat{\theta}) = 0$$

This may not always be the case, as Θ could be closed, potentially leaving $\hat{\theta}$ to be a boundary point. Another possibility is that $\hat{\theta}$ may be an extremum of the data (Like $\hat{\theta} = X_{(n)}$). If this happens, we must directly maximize $\mathcal{L}(\theta)$.

Lemma 1 (MLE Invariance)**.** *If $\hat{\theta}$ is an MLE of θ and $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is an MLE for $u(\theta)$.*

1.1.2 Consistency

Recall the definition of consistency.

Definition 1.3 (Consistent Estimator)**.** An estimator $\hat{\theta}$ is consistent if it approaches the true value θ_0 as more data is observed. That is, for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left[|\hat{\theta}_n - \theta_0| > \epsilon \right] = 0.$$

The MLE is consistent when particular conditions are met:

1. The model must be identifiable.
2. The parameter space Θ must be compact.
3. The density function must be continuous.
4. The log-likelihood must converge uniformly:

$$\sup_{\theta \in \Theta} ||\ell_n(\theta) - \ell(\theta)|| \xrightarrow{p} 0$$

   where $\ell(\theta)$ is the expected log-likelihood, $\ell(\theta) = \mathbb{E}[\log f(X_i, \theta)]$.

These conditions are somewhat mild, and hold for most i.i.d. samples with most common distributions, but still must not be assumed.

Example 1.2. **Neyman-Scott Problem**

Model: $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent pairs of independent Normal random variables, where for all fixed i, $X_i, Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
Note that X_i, Y_i are independent measurements of the same quantity, and $\mu_1, \ldots, \mu_n$ are unknown. Then, the likelihood function would be,

$$\mathcal{L}(\mu_1, \ldots, \mu_n, \sigma) = \prod_{i=1}^{n} \left[\frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{2\sigma^2} \right) \right]$$

Maximizing this over σ^2 and $\{\hat{\mu}_i\}$, you get MLEs:

$$\hat{\mu}_i = \frac{X_i + Y_i}{2},\ i = 1, \ldots, n.$$

$$\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2.$$

By WLLN, we know

$$\frac{1}{n} \sum_{i=0}^{n} (X_i - Y_i) \xrightarrow{p} 2\sigma^2$$

Thus, for our MLE of σ^2, our limiting behaviour ends up being,

$$\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2 \xrightarrow{p} \sigma^2/2$$

Thus, $\hat{\sigma}^2$ is not consistent.

□

This happens because we're using $2n$ observations to estimate n means, so we overfit.

1.1.3 Fisher Information

Definition 1.4 (Fisher Information)**.** Suppose $\ln \mathcal{L}(\theta) = \log f(X;\theta)$ is the log-likelihood function. Then, we say that

$$\mathcal{I}(\theta) = E\left[\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta)\right]^2$$

is the fisher information. Under appropriate conditions on $f(x;\theta)$ and $\mathcal{I}(\theta)$, it may also (and more commonly) be expressed as:

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln\mathcal{L}(\theta)\right]$$

Note that the first equation may also be expressed as the variance of the **score**. However, oftentimes this is not readily available as we work with samples, suggesting we need an analogous sample version.

Definition 1.5 (Observed Fisher Information)**.** Suppose $\ell(\hat{\theta}) = \ln\mathcal{L}(\hat{\theta})$ is the observed log-likelihood function. Then, we say that

$$I(\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta^2}\ln\mathcal{L}(\hat{\theta})$$

is the observed fisher information for a particular observation. Note that for an i.i.d. sample, the fisher information produced is identical, and thus the observed fisher information of the entire sample is $nI(\theta)$.

From basic calculus techniques, it is easy to observe that the observed Fisher information is simply the (absolute) curvature of the log-likelihood function at its maximum. The greater the curvature, the more well-defined the maximizer $\hat{\theta}$ is.
This suggests that as the observed Fisher information increases, the more prominent our $\hat{\theta}$, resulting with a better estimator and less uncertainty. This is the motivation behind our standard error estimate,

$$\widehat{\mathrm{se}}(\hat{\theta}) = \left\{-\frac{\mathrm{d}}{\mathrm{d}\theta^2}\ln\mathcal{L}(\hat{\theta})\right\}$$

1.1.4 Exponential MLE Convergence

Some common distributions that can be parameterized "nicely" are distributions from the exponential class or family. These include the normal distribution, bernoulli, poisson, and more. Because of their special parameterizations, it is often easier to prove their convergence properties in a more general fashion.

Definition 1.6 (One-parameter exponential family)**.** Assume $f(x;\theta)$ is the pmf of a random variable X. If f has the form,

$$f(x;\theta) = \exp[c(\theta)T(x) - d(\theta) + h(x)] \text{ for } x \in A$$

then X is considered to be from the one-parameter exponential family of distributions.

Lemma 2 (Exponential MLE Consistency)**.** *If $\hat{\theta}_{ML}$ is the MLE of a one-parameter exponential family, then:*

$$\hat{\theta}_{ML} \xrightarrow{p} \theta$$

Proof. For this model, the log-likelihood becomes

$$\ln\mathcal{L}(\theta) = \sum_{i=1}^{n}[c(\theta)T(x_i) - d(\theta) + h(x_i)]$$

and its derivative is

$$\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta) = \sum_{i=1}^{n}[c'(\theta)T(x_i) - d'(\theta)]$$

Thus, the MLE $\hat{\theta}_n$ satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n}T(x_i) = \frac{d'(\hat{\theta}_n)}{c'(\hat{\theta}_n)}$$

We now want to show that

$$\mathbb{E}_\theta\left[T(X_i)\right] = \frac{d'(\theta)}{c'(\theta)}$$

and

$$\mathrm{Var}_\theta[T(X_i)] = \frac{d''(\theta)c'(\theta) - c''(\theta)d'(\theta)}{[c'(\theta)]^3}$$

Subproof: Expectation / Variance of $T(X_i)$.

Let $m_\theta(s)$ be the mgf of $T(X_i)$.

$$m_\theta(s) = \mathbb{E}_\theta\left[\exp(sT(X_i))\right]$$

$$= \int_A \exp[sT(x)] \cdot \exp[c(\theta)T(x) - d(\theta) + h(x)]\,dx$$

$$= \int_A \exp[(s + c(\theta))T(x) - d(\theta) + h(x)]\,dx$$

Let $c(\theta') := s + c(\theta)$. If $\theta' \in \Theta$, then, by integrating $f(x;\theta')$

$$\implies \int_A \exp[c(\theta')T(x) - d(\theta') + h(x)]\,dx = 1$$

$$\implies \int_A \exp[c(\theta')T(x) + h(x)] = \exp[d(\theta')]$$

$$\implies \int_A \exp[c(\theta')T(x) - d(\theta) + h(x)] = \exp[d(\theta') - d(\theta)]$$

Thus, substituting back $c(\theta')$, then differentiating with respect to s,

$$m_\theta(s) = \exp[d(\theta') - d(\theta)]$$

$$\implies m'_\theta(s) = \exp[d(\theta') - d(\theta)] \cdot \frac{\partial d(\theta')}{\partial s}$$

By the chain rule,

$$\frac{\partial d(\theta')}{\partial s} = \frac{\partial d(\theta')}{\partial \theta'} \cdot \frac{\partial \theta'}{\partial s}$$

Recall that $c(\theta') = s + c(\theta)$, meaning that $c'(\theta')$ with respect to s is 1. Thus, by chain rule,

$$1 = \frac{\partial c(\theta')}{\partial s} = \frac{\partial c(\theta')}{\partial \theta'} \cdot \frac{\partial \theta'}{\partial s}$$

$$\implies \frac{\partial \theta'}{\partial s} = \frac{1}{\partial c(\theta')/\partial \theta'} = \frac{1}{c'(\theta')}$$

$$\implies \frac{\partial d(\theta')}{\partial s} = d'(\theta') \cdot \frac{1}{c'(\theta')}$$

Thus,

$$\implies m'_\theta(s) = \exp[d(\theta') - d(\theta)] \cdot \frac{d'(\theta')}{c'(\theta')}$$

Note that at $s = 0$, $\theta = \theta'$. Using the property of moment generating functions,

$$\mathbb{E}\left[T(X_i)\right] = m'_\theta(0) = \exp(0) \cdot \frac{d'(\theta)}{c'(\theta)} = \frac{d'(\theta)}{c'(\theta)}$$

By the quotient and product rule,

$$m''_\theta(s) = m_\theta(s) \cdot \frac{d''(\theta') \cdot c'(\theta') - d'(\theta') \cdot c''(\theta')}{(c'(\theta'))^2} \cdot \frac{\partial \theta'}{\partial s} + m'_\theta(s) \cdot \frac{d'(\theta')}{c'(\theta')}$$

Similar to before, $s = 0 \implies \theta = \theta'$,

$$m''_\theta(0) = \exp(0) \cdot \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3} + \left(\frac{d'(\theta)}{c'(\theta)}\right)^2$$

And thus,

$$\text{Var}[T(X_i)] = \mathbb{E}\left[T(X_i)^2\right] - (\mathbb{E}\left[T(X_i)\right])^2 = m''_\theta(0) - m'_\theta(0)^2$$

$$= \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3} + \left(\frac{d'(\theta)}{c'(\theta)}\right)^2 - \left(\frac{d'(\theta)}{c'(\theta)}\right)^2$$

$$\implies \text{Var}[T(X_i)] = \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3}$$

Which concludes the subproof. ■

Now that we know the expectation and variance, let $h'(\theta) = \frac{d'(\theta)}{c'(\theta)}$. By CLT,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} T(X_i) - h'(\theta)\right] \xrightarrow{p} \mathcal{N}(0, h''(\theta)/c'(\theta)).$$

Define g to be the inverse of $h'(\theta)$, such that $g(h'(\theta)) = \theta$. Thus,

$$g'(h'(\theta))h''(\theta) = 1$$

$$\implies g'(h'(\theta)) = \frac{1}{h''(\theta)}$$

We can now use the Delta Method, and result with

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, [g'(h'(\theta))]^2 \cdot h''(\theta)/c'(\theta)\right)$$

$$\implies \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{c'(\theta) \cdot h''(\theta)}\right)$$

□

Now, we can use this to estimate the standard error. Using the previous result,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{n \cdot c'\theta \cdot h''(\theta)}\right)$$

This suggests that we can estimate the standard error by

$$\widehat{\text{se}}(\hat{\theta}) = \left[n \cdot c'(\theta) \cdot h''(\theta)\right]^{-1/2}$$

Define $\ell(x;\theta) = \ln f(x;\theta)$ and $l'(x;\theta)$, $l''(x;\theta)$ to be its first two partial derivatives with respect to θ. Then,

$$\text{Var}_\theta[\ell'(X_i;\theta)] = [c'(\theta)]^2 \cdot \text{Var}_\theta[T(X_i)] = c'(\theta) \cdot h''(\theta)$$

We then have that,

$$I(\theta) = c'(\theta) \cdot h''(\theta) = \left(\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\hat{\theta})\right)$$

Recall that $nI(\theta)$ is the **observed Fisher Information**, and the following holds,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right)$$

1.2 Sufficiency

Suppose $(X_1, \ldots, X_n)$ is a random sample with joint pdf/pmf $f(\mathbf{x}; \boldsymbol{\theta})$, where $\mathbf{x} = (x_1, \ldots, x_n)$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Often times, $\mathbf{X} = (X_1, \ldots, X_n)$ contains information not relevant to $\boldsymbol{\theta}$, which we can reduce in complexity. We would like to form a statistic, $T(\mathbf{X})$, in which it is a data reduction or data summary of $\mathbf{X}$, and ideally, $T(\mathbf{X})$ would map to a smaller dimension than $\mathbf{X}$.

1.2.1 Sufficient Statistics

Definition 1.7 (Sufficient Statistic)**.** A statistic $T(\mathbf{X})$ is a *sufficient statistic* for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

We notationally define $T = (T_1(\mathbf{X}), \ldots, T_m(\mathbf{X}))$.

Theorem 1 (Factorization Theorem)**.** *Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the joint pdf/pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if and only if there exist functions $g(t|\boldsymbol{\theta})$ and $h(\boldsymbol{x})$ such that for all $(\boldsymbol{x}, \boldsymbol{\theta})$,*

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = g(T(\boldsymbol{x})|\boldsymbol{\theta})h(\boldsymbol{x})$$

Note that the likelihood function, $\mathcal{L}(\boldsymbol{\theta})$, is defined as $f(\mathbf{x}; \boldsymbol{\theta})$. Thus, for the factorization, we end up with,

$$\mathcal{L}(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta}) = g(T(\mathbf{x}); \boldsymbol{\theta})h(\mathbf{x})$$

This tells us that $h(\mathbf{x})$ is only a multiplicative constant in $\mathcal{L}(\boldsymbol{\theta})$. Thus, maximizing $\mathcal{L}(\boldsymbol{\theta})$ is equivalent to maximizing $g(T(\mathbf{x}); \boldsymbol{\theta})$.