Review and continuation from last week

Properties of MLEs

As we have mentioned, MLEs have a nice intuitive property, and as we have seen, they have a certain equivariance property.

We will see later that they also often have good asymptotic properties.

We now consider some other properties; some useful and some less desirable.
Relation to Sufficient Statistics

Theorem 1
If there is a sufficient statistic and an MLE exists, then an MLE is a function of the sufficient statistic.

Proof.
This follows directly from the factorization theorem.
Relation to Efficient Statistics

Given the three Fisher information regularity conditions, we have defined “Fisher efficient estimators” as unbiased estimators that achieve the lower bound on their variance.

**Theorem 2**

Assume the FI regularity conditions for a family of distributions \( \{P_\theta\} \) with the additional Le Cam-type requirement that the Fisher information matrix \( I(\theta) \) is positive definite for all \( \theta \). Let \( T(X) \) be a Fisher efficient estimator of \( \theta \). Then \( T(X) \) is an MLE of \( \theta \).
Proof.

Let $p_\theta(x)$ be the PDF. We have

$$\frac{\partial}{\partial \theta} \log(p_\theta(x)) = I(\theta)(T(x) - \theta)$$

for any $\theta$ and $x$. Clearly, for $\theta = T(x)$, this equation is 0 (hence, $T(X)$ is an RLE). Because $I(\theta)$, which is the negative of the Hessian of the likelihood, is positive definite, $T(x)$ maximizes the likelihood.

Notice that without the additional requirement of a positive definite information matrix, Theorem 2 would yield only the conclusion that $T(X)$ is an RLE.
Equivariance of MLEs

If \( \hat{\theta} \) is a good estimator of \( \theta \), it would seem to be reasonable that \( g(\hat{\theta}) \) is a good estimator of \( g(\theta) \), where \( g \) is a Borel function.

“Good”, of course, is relative to some criterion.

If the criterion is UMVU, then the estimator in general will not have this equivariance property; that is, if \( \hat{\theta} \) is a UMVUE of \( \theta \), then \( g(\hat{\theta}) \) may not be a UMVUE of \( g(\theta) \).

Maximum likelihood estimators almost have this equivariance property, however.

Because of the word “almost” in the previous sentence, which we will elaborate on below, we will just define MLEs of functions so as to be equivariant.
Definition 1 (functions of estimands and of MLEs)
If $\hat{\theta}$ is an MLE of $\theta$, and $g$ is a Borel function, then $g(\hat{\theta})$ is an MLE of the estimand $g(\theta)$.

The reason that this does not follow from the definition of an MLE is that the function may not be one-to-one.

We could just restrict our attention to one-to-one functions.

Instead, however, we will now define a function of the likelihood, and show that $g(\hat{\theta})$ is an MLE w.r.t. this transformed likelihood, called the induced likelihood.

Note, of course, that the expression $g(\theta)$ does not allow other variables and functions, e.g., we would not say that $E((Y - X\beta)^T(Y - X\beta))$ qualifies as a $g(\theta)$ in the definition above. (Otherwise, what would an MLE of $\sigma^2$ be?)
Definition 2 (induced likelihood)
Let \( \{p_\theta : \theta \in \Theta\} \) with \( \Theta \subset \mathbb{R}^d \) be a family of PDFs w.r.t. a common \( \sigma \)-finite measure, and let \( L(\theta) \) be the likelihood associated with this family, given observations. Now let \( g \) be a Borel function from \( \Theta \) to \( \Lambda \subset \mathbb{R}^{d_1} \) where \( 1 \leq d_1 \leq d \). Then
\[
\tilde{L}(\lambda) = \sup_{\{\theta : \theta \in \Theta \text{ and } g(\theta) = \lambda\}} L(\theta)
\]
is called the \textit{induced likelihood function} for the transformed parameter.

Theorem 3
Suppose \( \{p_\theta : \theta \in \Theta\} \) with \( \Theta \subset \mathbb{R}^d \) is a family of PDFs w.r.t. a common \( \sigma \)-finite measure with associated likelihood \( L(\theta) \). Let \( \hat{\theta} \) be an MLE of \( \theta \). Now let \( g \) be a Borel function from \( \Theta \) to \( \Lambda \subset \mathbb{R}^{d_1} \) where \( 1 \leq d_1 \leq d \) and let \( \tilde{L}(\lambda) \) be the resulting induced likelihood. Then \( g(\hat{\theta}) \) maximizes \( \tilde{L}(\lambda) \).
Example 1 MLE of the variance in a Bernoulli distribution
Consider the Bernoulli family of distributions with parameter $\pi$. The variance of a Bernoulli distribution is $g(\pi) = \pi(1 - \pi)$.

Given a random sample $x_1, \ldots, x_n$, the MLE of $\pi$ is

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} x_i / n,$$

hence the MLE of the variance is

$$\frac{1}{n} \sum x_i \left(1 - \frac{\sum x_i}{n}\right).$$

Note that this estimator is biased and that it is the same estimator as that of the variance in a normal distribution:

$$\sum (x_i - \bar{x})^2 / n.$$

The UMVUE of the variance in a Bernoulli distribution is,

$$\frac{1}{n - 1} \sum x_i \left(1 - \frac{\sum x_i}{n}\right).$$
The difference in the MLE and the UMVUE of the variance in the Bernoulli distribution is the same as the difference in the estimators of the variance in the normal distribution.

How do the MSEs of the estimators of the variance in a Bernoulli distribution compare? (Exercise.)

Whenever the variance of a distribution can be expressed as a function of other parameters $g(\theta)$, as in the case of the Bernoulli distribution, the estimator of the variance is $g(\hat{\theta})$, where $\hat{\theta}$ is an MLE of $\theta$.

The MLE of the variance of the gamma distribution, for example, is $\hat{\alpha}\hat{\beta}^2$, where $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs.

The plug-in estimator of the variance of the gamma distribution is different. Given the sample, $X_1, X_2 \ldots, X_n$, as always, it is

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$
Other Properties

Other properties of MLEs are not always desirable.

First of all, we note that an MLE may be biased. The most familiar example of this is the MLE of the variance.

Another example is the MLE of the location parameter in the uniform distribution.

Although the MLE approach is usually an intuitively logical one, it is not based on a formal decision theory, so it is not surprising that MLEs may not possess certain desirable properties that are formulated from that perspective.

There are other interesting examples in which MLEs do not have desirable (or expected) properties.
• An MLE may be discontinuous as, for example, in the case of $\epsilon$-mixture distribution family with CDF

$$P_{x,\epsilon}(y) = (1 - \epsilon)P(y) + \epsilon I_{[x,\infty]}(y),$$

where $0 \leq \epsilon \leq 1$.

• An MLE may not be a function of a sufficient statistic (if the MLE is not unique).

• An MLE may not satisfy the likelihood equation as, for example, when the likelihood function is not differentiable at its maximum.

• An MLE may differ from a simple method of moments estimator; in particular an MLE of the population mean may not be the sample mean.
The likelihood equation may have a unique root, but no MLE exists.

While there are examples in which the roots of the likelihood equations occur at minima of the likelihood, this situation does not arise in any realistic distribution (that I am aware of).

Romano and Siegel (1986) construct a location family of distributions with support on $\mathbb{R} - \{x_1 + \theta, x_2 + \theta : x_1 < x_2\}$, where $x_1$ and $x_2$ are known but $\theta$ is unknown, with a Lebesgue density $p(x)$ that rises as $x \nearrow x_1$ to a singularity at $x_1$ and rises as $x \searrow x_2$ to a singularity at $x_2$ and that is continuous and strictly convex over $]x_1, x_2[$ and singular at both $x_1$ and $x_2$.

With a single observation, the likelihood equation has a root at the minimum of the convex portion of the density between $x_1 + \theta$ and $x_2 + \theta$, but likelihood increases without bound at both $x_1 + \theta$ and $x_2 + \theta$. 
Nonuniqueness

There are many cases in which the MLEs are not unique (and I’m not just referring to RLEs). The following examples illustrate this.

Example 2 likelihood in a Cauchy family

Consider the Cauchy distribution with location parameter $\theta$. The likelihood equation is

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}.$$ 

This may have multiple roots (depending on the sample), and so the one yielding the maximum would be the MLE. Depending on the sample, however, multiple roots can yield the same value of the likelihood function.
Another example in which the MLE is not unique is $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

Example 3 likelihood in a uniform family with fixed range

The likelihood function for $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ is

$$I_{x(n)-1/2,x(1)+1/2}[\theta].$$

It is maximized at any value between $x(n) - 1/2$ and $x(1) + 1/2$. 

Nonexistence and Other Properties

We have already mentioned situations in which the likelihood approach does not seem to be the logical way, and have seen that sometimes in nonparametric problems, the MLE does not exist.

This often happens when there are more “things to estimate” than there are observations.

This can also happen in parametric problems.

In this case, some people prefer to say that the likelihood function does not exist; that is, they suggest that the definition of a likelihood function include boundedness.
MLE and the Exponential Class

If $X$ has a distribution in the exponential class and we write its density in the natural or canonical form, the likelihood has the form

$$L(\eta; x) = \exp(\eta^\top T(x) - \zeta(\eta)) h(x).$$

The log-likelihood equation is particularly simple:

$$T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0.$$

Newton's method for solving the likelihood equation is

$$\eta^{(k)} = \eta^{(k-1)} - \left( \frac{\partial^2 \zeta(\eta)}{\partial \eta(\partial \eta)^\top} \bigg|_{\eta = \eta^{(k-1)}} \right)^{-1} \left( T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} \bigg|_{\eta = \eta^{(k-1)}} \right)$$

Note that the second term includes the Fisher information matrix for $\eta$. (The expectation is constant.) (Note that the FI is not for a distribution; it is for a parametrization.)
We have

\[ V(T(X)) = \frac{\partial^2 \zeta(\eta)}{\partial \eta(\partial \eta)^T} |_{\eta=\eta}. \]

Note that the variance is evaluated at the true \( \eta \).

If we have a full-rank member of the exponential class then \( V \) is positive definite, and hence there is a unique maximum.

If we write

\[ \mu(\eta) = \frac{\partial \zeta(\eta)}{\partial \eta}, \]

in the full-rank case, \( \mu^{-1} \) exists and so we have the solution to the likelihood equation:

\[ \hat{\eta} = \mu^{-1}(T(x)). \]

So maximum likelihood estimation is very nice for the exponential class.
EM Methods

Although EM methods do not rely on missing data, they can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing.

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded.

In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units.

The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.
Another common example that motivates the EM algorithm is a finite mixture model.

Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator.

The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods involve missing-data problems, but this is not necessary.
Example 4 MLE in a normal mixtures model

A two-component normal mixture model can be defined by two normal distributions, \( \text{N}(\mu_1, \sigma_1^2) \) and \( \text{N}(\mu_2, \sigma_2^2) \), and the probability that the random variable (the observable) arises from the first distribution is \( w \).

The parameter in this model is the vector \( \theta = (w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \).

(Note that \( w \) and the \( \sigma \)s have the obvious constraints.)

The pdf of the mixture is

\[
p(y; \theta) = wp_1(y; \mu_1, \sigma_1^2) + (1 - w)p_2(y; \mu_2, \sigma_2^2),
\]

where \( p_j(y; \mu_j, \sigma_j^2) \) is the normal pdf with parameters \( \mu_j \) and \( \sigma_j^2 \).
We denote the “complete” data as \( C = (X, U) \), where \( X \) represents the observed data, and the unobserved \( U \) represents class membership.

Let \( U = 1 \) if the observation is from the first distribution and \( U = 0 \) if the observation is from the second distribution.

The unconditional \( E(U) \) is the probability that an observation comes from the first distribution, which of course is \( w \).

Suppose we have \( n \) observations on \( X, x_1, \ldots, x_n \).

Given a provisional value of \( \theta \), we can compute the conditional expected value \( E(U|x) \) for any realization of \( X \). It is merely

\[
E(U|x, \theta^{(k)}) = \frac{w^{(k)} p_1(x; \mu_1^{(k)}, \sigma_1^{2(k)})}{p(x; w^{(k)}, \mu_1^{(k)}, \sigma_1^{2(k)}, \mu_2^{(k)}, \sigma_2^{2(k)})}
\]
The M step is just the familiar MLE of the parameters:

\[ w^{(k+1)} = \frac{1}{n} \sum E(U|x_i, \theta^{(k)}) \]

\[ \mu_1^{(k+1)} = \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})x_i \]

\[ \sigma_1^{2(k+1)} = \frac{1}{nw^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_1^{(k+1)})^2 \]

\[ \mu_2^{(k+1)} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})x_i \]

\[ \sigma_2^{2(k+1)} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_2^{(k+1)})^2 \]

(Recall that the MLE of \( \sigma^2 \) has a divisor of \( n \), rather than \( n - 1 \).)
Asymptotic Properties of MLEs, RLEs, and GEE Estimators

The argmax of the likelihood function, that is, the MLE of the argument of the likelihood function, is obviously an important statistic.

In many cases, a likelihood equation exists, and often in those cases, the MLE is a root of the likelihood equation. In some cases there are roots of the likelihood equation (RLEs) that may or may not be an MLE.
Asymptotic Distributions of MLEs and RLEs

We recall that asymptotic expectations are defined as expectations in asymptotic distributions (rather than as limits of expectations).

The first step in studying asymptotic properties is to determine the asymptotic distribution.
Example 5 asymptotic distribution of the MLE of the variance in a Bernoulli family

In Example 1 we determined the MLE of the variance

\[ g(\pi) = \pi (1 - \pi) \]

in the Bernoulli family of distributions with parameter \( \pi \).

The MLE of \( g(\pi) \) is \( T_n = \bar{X} (1 - \bar{X}) \).

Now, let us determine its asymptotic distributions.
From the central limit theorem, $\sqrt{n}(\bar{X} - \pi) \to \text{N}(0, g(\pi))$.

If $\pi \neq 1/2$, $g'(\pi) \neq 0$, we can use the delta method and the CLT to get

$$\sqrt{n}(g(\pi) - T_n) \to \text{N}(0, \pi(1 - \pi)(1 - 2\pi)^2).$$

(I have written $(g(\pi) - T_n)$ instead of $(T_n - g(\pi))$ so the expression looks more similar to one we get next.)

If $\pi = 1/2$, this is a degenerate distribution.

(The limiting variance actually is 0, but the degenerate distribution is not very useful.)
Let’s take a different approach for the case $\pi = 1/2$. We have from the CLT,

$$\sqrt{n}(\overline{X} - 1/2) \rightarrow N\left(0, \frac{1}{4}\right).$$

Hence, if we scale and square, we get $4n(\overline{X} - 1/2)^2 \xrightarrow{d} \chi^2_1$, or

$$4n(g(\pi) - T_n) \xrightarrow{d} \chi^2_1.$$

This is a specific instance of a second order delta method.
Following the same methods as for deriving the first order delta method using a Taylor series expansion, for the univariate case, we can see that

\[ \sqrt{n}(S_n - c) \rightarrow N(0, \sigma^2) \]

\[ g'(c) = 0 \]

\[ g''(c) \neq 0 \]

\\[
\begin{align*}
\implies 2n \frac{g(S_n) - g(c)}{\sigma^2 g''(c)} & \overset{\text{d}}{\rightarrow} \chi_1^2.
\end{align*}
\]