16 Maximum Likelihood Estimates

Many think that maximum likelihood is the greatest conceptual invention in the history of statistics. Although in some high or infinite dimensional problems, computation and performance of maximum likelihood estimates (MLEs) are problematic, in a vast majority of models in practical use, MLEs are about the best that one can do. They have many asymptotic optimality properties which translate into fine performance in finite samples. We treat MLEs and their asymptotic properties in this chapter. We start with a sequence of examples, each illustrating an interesting phenomenon.

16.1 Some Examples

Example 16.1. In smooth regular problems, MLEs are asymptotically normal with a $\sqrt{n}$-norming rate. For example, if $X_1, \ldots, X_n$ are iid $N(\mu, 1), -\infty < \mu < \infty$, then the MLE of μ is $\hat{\mu} = \bar{X}$ and $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{L}} N(0,1), \forall \mu$.

Example 16.2. Let us change the problem somewhat to $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$ with $\mu \geq 0$. Then the MLE of μ is

$$\hat{\mu} = \begin{cases} \bar{X} & \text{if } \bar{X} \geq 0 \\ 0 & \text{if } \bar{X} < 0 \end{cases},$$

i.e., $\hat{\mu} = \bar{X} I_{\bar{X} \geq 0}$. If the true $\mu > 0$, then $\hat{\mu} = \bar{X}$ a.s. for all large n and $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{L}} N(0,1)$. If the true $\mu = 0$, then we still have consistency, in fact, still $\hat{\mu} \xrightarrow{\text{a.s}} \mu = 0$. Let us now look at the question of the limiting distribution of $\hat{\mu}$. Denote $Z_n = \sqrt{n}\bar{X}$, so that $\hat{\mu} = \frac{Z_n I_{Z_n \geq 0}}{\sqrt{n}}$.

Let $x < 0$. Then $P_0(\sqrt{n}\hat{\mu} \leq x) = 0$. Let $x = 0$; then $P_0(\sqrt{n}\hat{\mu} \leq x) = \frac{1}{2}$. Let $x > 0$. Then

$$\begin{aligned} P_0(\sqrt{n}\hat{\mu} \leq x) &= P(Z_n I_{Z_n \geq 0} \leq x) \\ &= \frac{1}{2} + P(0 < Z_n \leq x) \\ &\to \Phi(x). \end{aligned}$$

So

$$P(\sqrt{n}\hat{\mu} \leq x) \to \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{2} & \text{for } x = 0 \\ \Phi(x) & \text{for } x > 0 \end{cases}$$

The limit distribution of $\sqrt{n}\hat{\mu}$ is thus not normal; it is a mixed distribution.

Example 16.3. Consider the case when $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with μ known to be an integer. For the argument below, existence of an MLE of μ is implicitly assumed; but this can be directly proved by considering tail behavior of the likelihood function $l(\mu, \sigma^2)$.

Let $\hat{\mu}$ = MLE of μ. Then by standard calculus, $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \hat{\mu})^2$ is the MLE of σ^2.

Consider for integer μ, the ratio

$$\begin{aligned}
\frac{l(\mu, \sigma^2)}{l(\mu - 1, \sigma^2)} &= e^{\frac{1}{2\sigma^2}\left\{\sum(x_i - \mu + 1)^2 - \sum(x_i - \mu)^2\right\}} \\
&= e^{\frac{1}{2\sigma^2}\{n + 2\sum(x_i - \mu)\}} \\
&= e^{\frac{n}{2\sigma^2} + \frac{2n(\bar{X} - \mu)}{2\sigma^2}} \\
&= e^{\frac{n}{2\sigma^2}\{2(\bar{X} - \mu) + 1\}} \\
&\geq 1
\end{aligned}$$

iff $2(\bar{X} - \mu) + 1 \geq 0$ iff $\mu \leq \bar{X} + \frac{1}{2}$. In the interval $(\bar{X} - \frac{1}{2}, \bar{X} + \frac{1}{2}]$, there is a unique integer. It is the integer closest to $\bar{X}$. This is the MLE of μ.

Now let us look at the asymptotic behavior of the MLE $\hat{\mu}$.

$$\begin{aligned}
P(\hat{\mu} \neq \mu) &= P(\text{Integer closest to } \bar{X} \text{ is} \neq \mu) \\
&= P(\bar{X} > \mu + \frac{1}{2}) + P(\bar{X} < \mu - \frac{1}{2}) \\
&= 2P(\bar{X} > \mu + \frac{1}{2}) \\
&= 2P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}}{2\sigma}\right) \\
&= 2\left(1 - \Phi(\frac{\sqrt{n}}{2\sigma})\right) \\
&\sim 2\phi(\frac{\sqrt{n}}{2\sigma})\frac{2\sigma}{\sqrt{n}} \\
&= \frac{4\sigma}{\sqrt{2\pi n}} e^{-\frac{n}{8\sigma^2}}
\end{aligned}$$

For any $c > 0$, $\sum_n \frac{e^{-cn}}{\sqrt{n}} < \infty$. Therefore, $\sum_n P(\hat{\mu} \neq \mu) < \infty$ and so by the Borel-Cantelli lemma, $\hat{\mu} = \mu$ a.s. for all large n. Thus there is no asymptotic distribution of $\hat{\mu}$ in the usual sense.

Example 16.4. We do not need a closed form formula for figuring out the asymptotic behavior of MLEs. In smooth regular problems, MLEs will be jointly asymptotically normal with a $\sqrt{n}$-norming. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \Gamma(\alpha, \lambda)$ with density $\frac{e^{-\lambda x} x^{\alpha-1} \lambda^\alpha}{\Gamma(\alpha)}$. Then the likelihood function is

$$l(\alpha, \lambda) = \frac{e^{-\lambda \sum x_i} (\prod x_i)^\alpha \lambda^{n\alpha}}{(\Gamma(\alpha))^n}, \qquad \alpha, \lambda > 0$$

So

$$L = \log l(\mu, \sigma) = \alpha \log P - \lambda \sum x_i + n\alpha \log \lambda - n \log \Gamma(\alpha),$$

where $P = \prod x_i$.

The likelihood equations are

$$0 = \frac{\partial L}{\partial \alpha} = \log P + n \log \lambda - n\Psi(\alpha),$$

$$0 = \frac{\partial L}{\partial \lambda} = -\sum x_i + \frac{n\alpha}{\lambda},$$

where $\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digmma function. From solving $\frac{\partial L}{\partial \lambda} = 0$, one gets $\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}}$, where $\hat{\alpha}$ is the MLE of α. Existence of MLEs of $\hat{\alpha}, \hat{\lambda}$ can be directly concluded from the behavior of $l(\alpha, \lambda)$. Using $\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}}$, $\hat{\alpha}$ satisfies

$$\log P + n \log \hat{\alpha} - n \log \bar{X} - n\Psi(\hat{\alpha}) = \log P - n \log \bar{X} - n(\Psi(\hat{\alpha}) - \log \hat{\alpha}) = 0$$

The function $\Psi(\alpha) - \log \alpha$ is strictly monotone and continuous with range $\supset (-\infty, 0)$. So there is a unique $\hat{\alpha} > 0$ at which $n(\Psi(\hat{\alpha}) - \log \hat{\alpha}) = \log P - n \log \bar{X}$. This is the MLE of α. It can be found only numerically, and yet, from general theory, one can assert that $\sqrt{n}(\hat{\alpha} - \alpha, \hat{\lambda} - \lambda) \xrightarrow{\mathcal{L}} N(\underset{\sim}{0}, \Sigma)$ for some covariance matrix Σ.

Example 16.5. In non-regular problems, the MLE is not asymptotically normal and the norming constant is usually not $\sqrt{n}$. For example, if $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U[0, \theta]$, then the MLE $\hat{\theta} = X_{(n)}$ satisfies $n(\theta - \hat{\theta}) \xrightarrow{\mathcal{L}} \text{Exp}(\theta)$.

Example 16.6. This example shows that MLEs need not be functions of a minimal sufficient statistic. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U[\mu - \frac{1}{2}, \mu + \frac{1}{2}]$. Then the likelihood function is

$$l(\mu) = I_{\mu - \frac{1}{2} \le X_{(1)} \le X_{(n)} \le \mu + \frac{1}{2}} = I_{X_{(n)} - \frac{1}{2} \le \mu \le X_{(1)} + \frac{1}{2}}.$$

So any function of $X_1, \ldots, X_n$ that is in the interval $[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}]$ is an MLE, e.g., $e^{-\bar{X}^2}(X_{(n)} - \frac{1}{2}) + (1 - e^{-\bar{X}^2})(X_{(1)} + \frac{1}{2})$ is an MLE, but it is not a function of $(X_{(1)}, X_{(n)})$, the minimal sufficient statistic.

Example 16.7. This example shows that MLEs of different parameters can have limit distributions with different norming rates.

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\mu, \sigma)$ with density $\frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}}$, $x \geq \mu$. By simple calculus, the MLEs $\hat{\mu}, \hat{\sigma}$ are

$$\hat{\mu} = X_{(1)}$$

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

We can assume $\mu = 0$ and $\sigma = 1$ for the following calculations, from which the case of general μ, σ follows.

If $\mu = 0, \sigma = 1$, then $nX_{(1)} \sim \text{Exp}(1)$; thus for the general case, $n(\hat{\mu} - \mu) \xrightarrow{\mathcal{L}} \text{Exp}(\sigma)$. On the other hand, if $\mu = 0, \sigma = 1$, then $\sqrt{n}(\hat{\sigma} - 1) = \sqrt{n}(\bar{X} - X_{(1)} - 1) = \sqrt{n}(\bar{X} - 1) - \sqrt{n}X_{(1)} = \sqrt{n}(\bar{X} - 1) - \frac{nX_{(1)}}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, 1)$ by the CLT and Slutsky's theorem. Thus, for the general case, $\sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$. Note the different norming constants for $\hat{\mu}, \hat{\sigma}$.

16.2 Inconsistent MLEs

The first example of an MLE being inconsistent was provided by Neyman and Scott(1948). It is by now a classic example and is known as the Neyman-Scott example. That first example shocked everyone at the time and sparked a flurry of new examples of inconsistent MLEs including those offered by LeCam (1953) and Basu (1955). Note that what makes the Neyman-Scott example work is that, compared to the number of parameters, there isn't enough data to kill the bias of the MLE. It is possible to find adjustments to the MLE or suitable Bayesian estimates in many of these problems which do have the consistency property; see Ghosh (1994) for examples and also some general techniques.

Example 16.8. Let $X_{ij} \quad i = 1, 2, ..., n$ and $j = 1, 2, ..., k$ be independent with $X_{ij} \sim N(\mu_i, \sigma^2)$. Note that this is basically a balanced one-way ANOVA design where we assume k is fixed and $n \to \infty$. So the sample sizes of the groups are (probably) big, but the number of groups is bigger. We want to estimate the common variance of the groups. By routine calculus, the MLEs are

$$\hat{\mu}_i = \bar{X}_i \text{ and } \hat{\sigma}^2 = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_i)^2.$$

It is the MLE of σ^2 that is inconsistent. Indeed,

$$\hat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2 = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2\right) = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\sigma^2 W_i$$

where the W_i are independent χ^2_{k-1}. By the WLLN,

$$\frac{\sigma^2}{k}\frac{1}{n}\sum_{i=1}^{n}W_i \xrightarrow{\mathcal{P}} \frac{\sigma^2}{k}(k-1)$$

Hence, the MLE for σ^2 does not converge to σ^2! It is the bias that is making the estimate inconsistent; if we kill the bias by multiplying by $\frac{k}{k-1}$ the new estimator is consistent, i.e., if we "adjust" the MLE and use

$$\frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2$$

then we return to consistency. In these sorts of problems, where the number of observations and the number of free parameters grow at the same rate, maximum likelihood often runs into problems. However, these problems are hard for any school of thought.

16.3 MLEs in Exponential Family

It is part of the statistical folklore that MLEs cannot be beaten asymptotically. One needs to be careful in making such a statement. Under various conditions, MLEs are indeed asymptotically optimal and asymptotically normal. But it is important to remember that the conditions needed are NOT just on the probability model; there must be conditions imposed on the competing estimates also for optimality.

There are other potential problems. MLEs may not exist for all samples. In such cases, one can only talk about asymptotic behavior and optimality of estimates that are quasi-MLEs. Careful exposition of the technical issues and proofs may be seen in Perlman(1983), Bickel and Doksum (2001), Lehmann and Casella (1998) and Brown (1986). Computing the MLE can also be a difficult numerical exercise in general; the EM algorithm is a popular tool for this. See McLachlan and Krishnan (1997).

We start with a familiar model, namely exponential families; things are relatively uncomplicated in this case. For the sake of completeness, we state the definition and a few basic facts about the Exponential family.

Definition 16.1. Let $f(x|\theta) = e^{\theta T(x) - \psi(\theta)} h(x) d\mu(x)$, where μ is a positive σ-finite measure on the Real line, and $\theta \in \Theta = \{\theta : \int e^{\theta T(x)} h(x) d\mu(x) < \infty\}$. Then, f is said to belong to the one parameter Exponential family with natural parameter space Θ. The parameter θ is called the natural parameter of f.

The following are some standard facts about a density in the one parameter Exponential family.

Proposition (a) For $\theta \in \Theta^0$, the interior of Θ, all moments of $T(X)$ exist, and $\psi(\theta)$ is infinitely differntiable at any such θ. Furthermore, $E_\theta(T) = \psi'(\theta)$, and $var_\theta(T) = \psi''(\theta)$;

(b) Given an iid sample of size n from f, $\sum_{i=1}^n T(X_i)$ is minimal sufficient;

(c) The Fisher information function exists, is finite at all $\theta \in \Theta^0$, and equals $I(\theta) = \psi''(\theta)$;

(d) The following families of distributions belong to the one parameter Exponential family:

$$N(\mu, 1), N(0, \sigma^2), Ber(p), Bin(n, p), n \text{ fixed},$$

$$Poi(\mu), Geo(p), Exp(\lambda), Gamma(\alpha, \lambda), \alpha \text{ fixed}, Gamma(\alpha, \lambda),\ \lambda \text{fixed}.$$

Theorem 16.1. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x|\theta) = e^{\theta T(x) - \psi(\theta)} h(x) d\mu(x)$. Let the true $\theta = \theta_0 \in \Theta^o$, i.e., the interior of the natural parameter space. Assume $\psi''(\theta) > 0\ \forall \theta \in \Theta^o$. Then for all large n, w.p. 1, a unique MLE of θ exists, is consistent, and is asymptotically normal.

Proof: The likelihood function is

$$l(\theta) = e^{\theta \Sigma T(x_i) - n\psi(\theta)}$$

$$\Rightarrow L(\theta) = logl(\theta) = n[\theta \bar{T} - \psi(\theta)].$$

Therefore, the likelihood equation is

$$L'(\theta) = n[\overline{T} - \psi'(\theta)] = 0 \Leftrightarrow \overline{T} - E_\theta(T(X_1)) = 0.$$

Now $T(X_1)$ has a finite mean, and hence, by the SLLN,

$$\overline{T} \xrightarrow{\text{a.s}} E_{\theta_0} T(X_1) = \psi'(\theta_0).$$

Hence, for all large n, w.p. 1, $\overline{T}$ is in the interior of the range of the function $\theta \to \psi(\theta)$. On the other hand, $E_\theta(T(X)) = \psi'(\theta)$ is a strictly monotone increasing

function of θ because $\psi''(\theta) = Var_\theta(T(X)) > 0$. Therefore, for all large n, w.p. 1, there exists a unique θ such that $E_\theta T(X) = \overline{T}$. This is the MLE of θ and is characterized as the unique root of $\psi'(\theta) = \overline{T} \Leftrightarrow \hat{\theta} = (\psi')^{-1}(\overline{T})$.

By the Continuous mapping theorem, $\hat{\theta} \xrightarrow[\mathcal{P}_{\theta_0}]{a.s.} \theta_0$. By the Central Limit Theorem, $\overline{T}$ is asymptotically normal. By the Delta Theorem, a smooth function of an asymptotically normal sequence is also asymptotically normal. Indeed, since $\sqrt{n}(\bar{T} - \psi'(\theta_0)) \overset{\mathcal{L}}{\Rightarrow} N(0, \psi''(\theta_0))$, and since $\hat{\theta} = (\psi')^{-1}(\overline{T})$, a direct application of the the Delta theorem implies that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[\mathcal{P}_{\theta_0}]{\mathcal{L}} N(0, I^{-1}(\theta_0))$$

where $I(\theta_0) = \psi''(\theta_0)$.

Remark: So this is a success story for the MLE: strong consistency and asymptotic normality hold. Nevertheless, even in this successful case, it is not true that this estimate gives the uniformly best limit distribution. It is possible to find a competing estimate, $\hat{\hat{\theta}}$, that converges to some other limit distribution, which has a smaller variance for some particular θ. We will discuss these important subtleties in a later section.

16.4 More General Cases and Asymptotic Normality

In general, we may have problems with the existence of the MLE, even for large samples. What can we get in such cases? We will need a laundry list of assumptions. If we are satisfied with something that is merely consistent, the list is shorter. If we want something that is also asymptotically normal, then the list of assumptions gets longer. This list has come to be known as the Cramér-Rao conditions; see Lehmann and Casella (1998) and Lehmann (1999) for a proof of the next theorem.

Theorem 16.2. Cramér-Rao conditions Assume $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_\theta$ and $\frac{dP_\theta}{d\mu} = f(x|\theta)$ for some σ-finite μ (e.g., Lebesgue measure in the continuous case or counting measure in the discrete case).

Assume the conditions:

(A1) Identifiability, i.e., $P_{\theta_1} = P_{\theta_2} \Leftrightarrow \theta_1 = \theta_2$.

(A2) $\theta \in \Theta =$ an open interval in the Real line.

(A3) $S = \{x : f(x|\theta) > 0\}$ is free of θ .

(A4) $\forall x \in S, \frac{d}{d\theta} f(x|\theta)$ exists, i.e., the likelihood function is smooth as a function of the parameter.

Let $\theta_0 \in \Theta^0$ be the true value of θ. Then there exists a sequence of functions $\hat{\theta_n} = \hat{\theta_n}(X_1, \ldots, X_n)$ such that

(i) $\hat{\theta}_n$ is a root of the likelihood equation $L'(\theta) = 0$ for all large n,

where $L(\theta)$ is $\log l(\theta) = \Sigma \log f(x_i|\theta)$.

(ii) P_{θ_0}(the root $\hat{\theta}_n$ is a local maximum of $l(\theta)) \longrightarrow 1$ as $n \longrightarrow \infty$

(iii)$\hat{\theta_n} \xrightarrow{P_{\theta_0}} \theta_0$.

Remark: This theorem does not say which sequence of roots of $L'(\theta) = 0$ should be chosen to ensure consistency in the case of multiple roots. It does not even guarentee that for any given n, however large, the likelihood function $l(\theta)$ has any local maxima at all. This specific theorem is useful in ONLY those cases where $L'(\theta) = 0$ has a UNIQUE root for all n.

Since consistency is regarded as a weak positive property, and since in statistics one usually wants to make actual inferences such as confidence interval construction, it is important to have weak convergence results, in addition to consistency. As we remarked earlier, establishing weak convergence results requires more conditions.

The issues and the results in the multiparameter case are analogous to those in the one parameter case. As in the one parameter case, in general one can only assert consistency and asymptotic normality of suitable sequences of roots of the likelihood equation. We state here the asymptotic normality result directly for the multiparameter case, from which the one parameter case follows as a special case. For a complete list of the regularity conditions needed to prove the following theorem, and also for a proof, see Lehmann and Casella (1998). We refer to the list of all assumptions as *multiparameter Cramér-Rao conditions for asymptotic normality*. A problem in which these conditions are all satisfied is usually called a *regular parametric problem*.

Theorem 16.3. Under the multiparameter case Cramér-Rao conditions for asymptotic normality, there exists a sequence of roots of the likelihood equation which is consistent and which satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\theta_0))$, where $I(\theta) = ((I_{ij}(\theta)))$ with $I_{ij}(\theta) = -E_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(X|\theta)\right]$, is the Fisher information matrix.

Remark: As an example of what the conditions are, one of the conditions for the above theorem is that $\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\log f(x|\theta)$ exists for all x in $S = \{x : f(x|\theta) > 0\}$ and

$$\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\int_S \log f(x|\theta)dx = \int_S \left\{\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\log f(x|\theta)\right\}dx$$

Remark: The theorem applies to any distribution in the Exponential family for which $\psi''(\theta)$ is positive and finite for every θ in the interior of the natural parameter space. There are also *multiparameter Exponential families*, very similar to the one parameter version, for which the theorem holds, but we will not treat them here.

16.5 Observed and Expected Fisher Information

Consider the regular one parameter problem with iid observations from a density $f(x|\theta)$ wrt some dominating measure μ. According to the previous theorem, the "MLE" $\hat{\theta}_n$ is asymptotically normal with mean θ and variance $\frac{1}{nI(\theta)}$, where $I(\theta) = -E_\theta(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta))$.Since the observations X_i are iid, by Kolmogorov's SLLN, the average $\frac{1}{n}\sum_{i=1}^n -\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta) \stackrel{a.s.}{\Rightarrow} I(\theta)$. Thus, as a matter of providing an estimate of the variance of the MLE, it is very reasonable to provide the estimate $\frac{1}{\sum_{i=1}^n -\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)|_{\theta=\hat{\theta}}}$, where $\hat{\theta}$ is the MLE of θ. The quantity $\frac{1}{n}\sum_{i=1}^n -\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)$ is called the *Observed Fisher Information*. Its expectation, which is just the Fisher information function $I(\theta)$ is called the *Expected Fisher Information*. It is natural to ask which gives a better estimate of the true variance of the MLE : $\frac{1}{nI(\hat{\theta})}$, or $\frac{1}{\sum_{i=1}^n -\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)|_{\theta=\hat{\theta}}}$?

We present two examples to help understand this question.

Example 16.9. Suppose $X_1, \ldots, X_n$ are iid from a distribution in the one parameter Exponential family with density $f(x|\theta) = e^{\theta T(x)-\psi(\theta)}h(x)(d\mu)$. Then, $\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = -\psi''(\theta)$. Thus, $I(\theta)$ and $\frac{1}{n}\sum_{i=1}^n -\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)$ are both equal to $\psi''(\theta)$, and so use of the observed or the expected Fisher information lead to the same estimate for the variance of $\hat{\theta}_n$.

Example 16.10. Suppose $X_1, \ldots, X_n$ are iid from the Cauchy distribution $C(\theta, 1)$. Then, $f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, and $\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = \frac{2((x-\theta)^2-1)}{(1+(x-\theta)^2)^2}$. On doing the necessary integration, $I(\theta) = \frac{1}{2}$. Thus the estimate of the variance of the MLE based on the expected information is $\frac{2}{n}$. However, it is clear that the observed information method would produce an estimated variance that depends on the actual observed

data. Over repeated sampling, it will have, typically, an asymptotically normal distribution itself, but its performance as an estimate of the true variance relative to the constant estimate $\frac{2}{n}$ can be accurately understood only by careful simulation.

Some interesting facts are revealed by a simulation. For $n = 20$, and the true θ value equal to 0, a simulation of size 500 was conducted to enquire into the performance of the variance estimates discussed above. The estimate based on the expected Fisher information is $\frac{2}{n} = .1$. The true variance of the MLE when $n = 20$ is .1225 according to the simulation. Thus the expected Fisher information method produces an underestimate of the variance by about 16%. The variance estimate produced by the observed information method gives an average estimate of .1071 over the 500 simulations. Thus, the bias is significantly lower. However, the variability in the variance estimate over the simulations is high. While the smallest variance estimate produced by the observed information method is .0443, the largest one is .9014. The heaviness of the Cauchy tail impacts the variance of the variance estimate as well. The estimate produced by the observed information method has a smaller bias than the one based on expected information, but can go wild from time to time, and is perhaps risky. The expected information estimate, on the other hand, is just a constant estimate $\frac{2}{n}$, and is not prone to fluctuations caused by a whimsical Cauchy sample. This example illustrates the care needed in assessing the accuracy of maximum likelihood estimates; the problem is harder than it is commonly believed to be.

16.6 Edgeworth Expansions for MLEs

The central limit theorem gives a first order approximation to the distribution of the MLE under regularity conditions. More accurate approximations can be obtained by Edgeworth expansions of higher order. In the Exponential family, where the MLE is a linear statistic, the expansion is a bit easier to state. For more general regular densities, the assumptions are complex and many, and the expansion itself is notationally more messy. References for these expansions are Pfanzagl (1973), Bhattacharya and Ghosh(1978), and Bai and Rao(1991). We present the expansions in two cases below, namely, the case of the Exponential family, and a more general regular case. Of these, in the Exponential family, the MLE of the mean function is the sample mean itself, and so an Edgeworth expansion follows from general expansions for sample means, as in Chapter 13; a specific reference for the next theorem is Pfanzagl(1973).

Theorem 16.4. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = e^{\theta x - \psi(\theta)} h(x) dx$. Consider estimation of $E_\theta(X) = \psi(\theta)$ and let $F_n(x) = P_\theta(\frac{\sqrt{n}(\bar{X} - \psi'(\theta))}{\sqrt{\psi''(\theta)}} \leq x)$. Then,

$$F_n(x) = \Phi(x) + \frac{p_1(x,\theta)\phi(x)}{\sqrt{n}} + \frac{p_2(x,\theta)\phi(x)}{n} + O(n^{-3/2}),$$

uniformly in x, where,

$$p_1(x,\theta) = c_1(1-x^2), p_2(x,\theta) = c_2(3x - x^3) + \frac{c_1^2}{72}(10x^3 - 15x - x^5),$$

with $c_1 = \frac{\psi^{(3)}(\theta)}{6(\psi''(\theta))^{3/2}}, c_2 = \frac{\psi^{(4)}(\theta)}{24(\psi''(\theta))^2}$.

Example 16.11. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = \theta e^{\theta x}, x < 0, \theta > 0$. Then, $\psi(\theta) = -\log\theta, c_1 = -\frac{1}{3}, c_2 = \frac{1}{4}$. Thus, the MLE $\bar{X}$ of $E_\theta(X) = -\frac{1}{\theta}$ satisfies the expansion

$$P_\theta(\sqrt{n}(\bar{X} + \frac{1}{\theta}) \leq \frac{x}{\theta}) = \Phi(x) - \frac{(1-x^2)\phi(x)}{3\sqrt{n}} + \frac{(\frac{3x - x^3}{4} + \frac{10x^3 - 15x - x^5}{648})\phi(x)}{n} + O(n^{-3/2}),$$

uniformly in x. For ease of reference, we will denote the two term expansion by $H(n,x)$. As a test of the expansion's numerical accuracy, suppose the true $\theta = 1$ and we want to approximate $P((\sqrt{n}(\bar{X} + \frac{1}{\theta}) \leq \frac{x}{\theta})$. Since $-\sum_{i=1}^n X_i$ is Gamma with shape parameter n and scale parameter 1, on computation, one finds the following exact values and approximations obtained from the above two term expansion; we use $n = 30$.

$$x = .5; exact = .675; H(n,x) = .679;$$

$$x = 2.0; exact = .988; H(n,x) = .986;$$

$$x = 3.0; exact = 1.000; H(n,x) = 1.0001.$$

Thus, the expansion is quite accurate at the sample size of $n = 30$. This example brings out an undesirable feature of Edgeworth expansions that they are not CDFs and can take values < 0 or > 1, as it does here when $x = 3.0$.

A general Edgeworth expansion for the MLE under a variety of regularity conditions is given in Pfanzagl(1973). The conditions are too many to state them here. However, the expansion is explicit. We give the expansion below. Pfanzagl(1973) gives examples of families of densities that satisfy the conditions required for the validity of his theorem. We first need some more notation.

For a given density $f_\theta(x)$, let

$$l_\theta = \log f_\theta(x), \dot{l}_\theta = \frac{\partial}{\partial\theta} l_\theta, \ddot{l}_\theta = \frac{\partial^2}{\partial\theta^2} l_\theta, l_\theta^{(3)} = \frac{\partial^3}{\partial\theta^3} l_\theta,$$

$$\rho_{20} = E_\theta[\dot{l}_\theta]^2, \rho_{11} = -E_\theta[\ddot{l}_\theta], \rho_{30} = -E_\theta[\dot{l}_\theta]^3, \rho_{12} = -E_\theta[l_\theta^{(3)}], \rho_{21} = 2E_\theta[\dot{l}_\theta \ddot{l}_\theta].$$

With this notation, we have the following expansion for the CDF of the MLE; for notational simplicity, we present only the one term expansion in the general case.

Theorem 16.5. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$. Under the regularity conditions on f_θ as in Pfanzagl(1973), the MLE $\hat{\theta}_n$ of θ satisfies $F_n(x,\theta) = P_\theta(\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\beta} \leq x) = \Phi(x) + \frac{q_1(x,\theta)\phi(x)}{\sqrt{n}} + o(\frac{1}{\sqrt{n}})$, uniformly in x and uniformly in compact neighborhoods of the given θ, where $q_1(x,\theta) = a_{10} + a_{11}x^2$, with $a_{10} = -\frac{\rho_{30}}{6\rho_{20}^{3/2}}, a_{11} = -a_{10} + \frac{\rho_{12}\sqrt{\rho_{20}}}{2\rho_{11}^2} - \frac{\rho_{21}}{2\sqrt{\rho_{20}}\rho_{11}}, \beta = \frac{\sqrt{\rho_{20}}}{\rho_{11}}$.

Example 16.12. Consider maximum likelihood estimation of the location parameter of a Cauchy distribution. The regularity conditions needed for an application of Theorem 16.5 are met; see Pfanzagl(1973). We have $f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$. Because of the fact that the density is symmetric, the coefficients $\rho_{12}, \rho_{21}, \rho_{30}$ (i.e., those whose subscripts add to an odd integer) are all zero. Therefore, $a_{10} = a_{11} = 0$, and so it follows that $F_n(x,\theta) = \Phi(x) + o(\frac{1}{\sqrt{n}})$, uniformly in x, i.e., the CLT approximation is *second order accurate*; this is interesting, and is a consequence of the symmetry.

16.7 Asymptotic Optimality of the MLE and Superefficiency

It was first believed as a folklore that the MLE under regularity conditions on the underlying distribution is asymptotically the best for every value of θ, i.e. if a MLE $\hat{\theta}_n$ exists and , $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} N(0, I^{-1}(\theta))$, and if another competing sequence T_n satisfies $\sqrt{n}(T_n - \theta) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} N(0, V(\theta))$, then for every θ, $V(\theta) \geq \frac{1}{I(\theta)}$. It was a major shock when in 1952 Hodges gave an example that destroys this belief and proved it to be false even in the normal case. Hodges (1952) produced an estimate T_n that beats the MLE $\overline{X}$ locally at some θ, say, $\theta = 0$. The example can be easily refined to produce estimates T_n that beat $\overline{X}$ at any given finite set of values of θ. Later, in a very insightful result, LeCam(LeCam(1953)) showed that this can happen only on Lebesgue-null sets of θ. If, in addition, we insist on using only such estimates T_n that have a certain smoothness property (to be made precise later), then the inequality $V(\theta) < \frac{1}{I(\theta)}$ cannot materialize at all. So to justify the folklore that MLEs are

asymptotically the best, one not only needs regularity conditions on $f(x|\theta)$, but one also must restrict attention to only those estimates that are adequately nice (and Hodges' estimate is not). An excellent reference for this topic is van der Vaart (1998).

Example 16.13. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Define an estimate T_n as

$$T_n = \begin{cases} \bar{X} & \text{if } \left|\bar{X}\right| > n^{-\frac{1}{4}} \\ a\bar{X} & \text{if } \left|\bar{X}\right| \leq n^{-\frac{1}{4}} \end{cases}$$

where $0 \leq a < 1$.

To derive the limiting distribution of T_n, notice that

$$P_\theta(\sqrt{n}\,|T_n - \theta| \leq c)$$

$$= P_\theta(\sqrt{n}\,|T_n - \theta| \leq c, \left|\overline{X}\right| \leq n^{-\frac{1}{4}}) + P_\theta(\sqrt{n}\,|T_n - \theta| \leq c, \left|\overline{X}\right| > n^{-\frac{1}{4}}).$$

If $\theta = 0$ then the second term goes to zero and so the limit distribution is determined from the first term. For $\theta \neq 0$, the situation reverses. It follows that $\sqrt{n}(T_n - \theta) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} N(0,1)$, if $\theta \neq 0$ and $\xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} N(0, a^2)$, if $\theta = 0$. Thus if we denote by $V(\theta)$ the asymptotic variance of T_n, then $V(\theta) = \frac{1}{I(\theta)}$, for $\theta \neq 0$, $V(\theta) = \frac{a^2}{I(\theta)}$, for $\theta = 0$. Therefore, $V(\theta) \leq \frac{1}{I(\theta)}$ for every θ and $V(\theta) < \frac{1}{I(\theta)}$ at $\theta = 0$.

Remark: The Hodges estimate T_n is what we call a *shrinkage estimate* these days. Because $V(\theta) \leq \frac{1}{I(\theta)}\,\forall\theta$ and $V(\theta) < \frac{1}{I(\theta)}$ at $\theta = 0$, the asymptotic relative efficiency (ARE) of T_n with respect to $\overline{X}$ is $\geq 1, \forall\theta$ and > 1 when $\theta = 0$. Such estimates, which have a smaller asymptotic variance than the MLE locally at some θ, and never a larger asymptotic variance at any θ, are called *superefficient*.

It is clear however that T_n has certain undesirable features. First, as a function of $X_1, ..., X_n$, T_n is not smooth. Second $V(\theta)$ is not continuous in θ. However , what transpires is that something is wrong with T_n very seriously. The mean squared error of T_n behaves erratically. For given n, as a function of θ, $nE_\theta(T_n - \theta)^2$ sharply leaps over $nE(\overline{X} - \theta)^2 = 1$ for values of $\theta \approx 0$. The values of θ at which this occurs change with n. At any given θ, the leaps vanish for large n. But for any n, the leaps reoccur at other values of θ close to 0. Thus the superefficiency of T_n is being purchased at the cost of a sharp spike in the mean squared error at values of θ very very close to 0. DasGupta (2004) shows that

$$\liminf_{n\to\infty} \sup_{|\theta| \leq n^{-\frac{1}{4}}} \sqrt{n} E_\theta(T_n - \theta)^2 \geq \frac{1}{2}.$$

Notice the $\sqrt{n}$ norming in the result, as opposed to the norming by n for the equalizer minimax estimate $\bar{X}$.

16.8 Hajek-Le Cam Convolution Theorem

The superefficiency phenomenon, it turns out, can only happen on Lebesgue-null subsets of Θ. It cannot happen at all if furthermore attention is restricted to estimators that are distributionally smooth in the following sense.

Definition 16.2. Let T_n be an estimate sequence for a vector function $\psi(\theta)$ such that $\sqrt{n}(T_n - \psi(\theta)) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} \mu_\theta$; T_n is called a regular estimating sequence if for all finite h,

$$\sqrt{n}(T_n - \psi(\theta + \frac{h}{\sqrt{n}})) \xrightarrow[\mathcal{P}_{\theta+\frac{h}{\sqrt{n}}}]{\mathcal{L}} \mu_\theta.$$

Remark: Thus, for a regular estimating sequence T_n, changing the parameter ever so slightly would not change the limit distribution at all. Among such estimates we cannot find one that is superefficient.

Theorem 16.6. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_\theta \ll \mu, \theta \in \Theta$. Suppose $f(x|\theta) = \frac{dP_\theta}{d\mu}(x) > 0$ for every x, θ, and $\nabla_\theta f(x|\theta)$ exists for every x, θ. Suppose also that

$$0 < I_{ij}(\theta) = E_\theta \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(X|\theta) \right] < \infty$$

for every θ, is continuous at every θ and $I^{-1}(\theta)$ exists for every θ. Let $\psi(\theta)$ be any differentiable function of θ with gradient vector $\nabla\psi(\theta)$. Let T_n be any regular estimate sequence of $\psi(\theta)$ with $\sqrt{n}(T_n - \psi(\theta)) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} \mu_\theta$. Then there exists a (unique) probability distribution ν_θ such that μ_θ admits the convolution representation $\mu_\theta = N(0, (\nabla\psi)I^{-1}(\theta)(\nabla\psi)') * \nu_\theta$. In particular, if μ_θ has a covariance matrix, say $\sum_\theta$, then $\sum_\theta \geq (\nabla\psi)I^{-1}(\theta)(\nabla\psi)'$ in the sense $\sum_\theta - (\nabla\psi)I^{-1}(\theta)(\nabla\psi)'$ is n.n.d. at every θ.

In the absence of the regularity of the estimate sequence T_n, we can assert something a bit weaker.

Theorem 16.7. Assume the conditions in the previous theorem on P_θ ,$I(\theta)$, and $\psi(\theta)$. Suppose $\sqrt{n}(T_n - \psi(\theta)) \xrightarrow[\mathcal{P}_\theta]{\mathcal{L}} \mu_\theta$. Then for almost all θ (Lebesgue), μ_θ admits

the convolution representation $\mu_\theta = N(0, (\nabla\psi)I^{-1}(\theta)(\nabla\psi)') * \nu_\theta$.

Remark: These theorems are collectively known as the Hajek-Le Cam convolution theorem. See van der Vaart (1998) for greater details and proofs. The second theorem says that even without regularity of the competing estimates T_n, superefficiency can occur only on sets of θ of Lebesgue measure 0. This result of Lucien Le Cam is regarded as one of the most insightful results in theoretical statistics.

We give an example of a nonregular estimate for illustration.

Example 16.14. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N_p(\theta, I)$. The MLE of θ is $\overline{X}$. In 1961, James and Stein showed that $T_n = (1 - \frac{p-2}{n\|\overline{X}\|^2})\overline{X}$ has a smaller mean squared error than $\overline{X}$ at every θ, provided $p \geq 3$, i.e. $E_\theta\|T_n - \theta\|^2 < E_\theta\|\overline{X} - \theta\|^2 = \frac{p}{n}, \forall\theta$. The James-Stein estimate T_n has the property that $E_{\frac{h}{\sqrt{n}}}\|T_n - \frac{h}{\sqrt{n}}\|^2 < E_{\frac{h}{\sqrt{n}}}\|\overline{X} - \frac{h}{\sqrt{n}}\|^2, \forall h$. It follows that the limit distribution of $\sqrt{n}(\overline{X} - \frac{h}{\sqrt{n}})$ does not have a smaller covariance matrix than that of $\sqrt{n}(T_n - \frac{h}{\sqrt{n}})$. The James-Stein estimate does not have the property of regularity. And it is exactly at $\theta = 0$ that the estimate T_n is nonregular; i.e. $\sqrt{n}(T_n - \frac{h}{\sqrt{n}}) \underset{\mathcal{P}_{\frac{h}{\sqrt{n}}}}{\xrightarrow{\mathcal{L}}} \mu_h$ for some distribution μ_h that really does depend on h. In fact one can describe μ_h. It is the same as the distribution of $(1 - \frac{p-2}{\|Z\|^2})Z - h$, where $Z \sim N_p(0, I)$.

16.9 Loss of Information and Efron's Curvature

In Exponential families, the maximum likelihood estimate based on n iid observations is itself a sufficient statistic. Since we think of sufficient statistics as capturing all the information about the parameter present in the sample, it would mean that the loss of information caused by summarizing the full data into the MLE is zero in Exponential families. How does one formalize this question for nonexponential families and give a quantification of the loss of information suffered by the MLE and relate it to something of actual statistical relevance ? Efforts to show that the maximum likelihood estimate leads to the least amount of information lost by a one dimensional summary started with the seminal second order efficiency theory of Rao(1961,1962,1963). More recently, Efron(1975) gave a theory of curvature of parametric families which attempts to connect the information loss question with how nonexponential a family is. The idea is that the more *curved* a parametric family is, the greater is the information loss suffered by the MLE. We present a few results in this direction below.

Definition 16.3. Let $P_\theta << \mu$ be a family of dominated measures with corresponding densities $f_\theta(x)$ in an Euclidean space. Assuming all the required derivatives and the expectations exist, let

$$l_\theta(x) = \log f_\theta(x), \nu_{11}(\theta) = E_\theta[\frac{\partial}{\partial\theta} l_\theta \frac{\partial^2}{\partial\theta^2} l_\theta], \nu_{02}(\theta) = E_\theta[\frac{\partial^2}{\partial\theta^2} l_\theta]^2 - I^2(\theta),$$

where $I(\theta)$ denotes the Fisher information at θ. The curvature of $\{P_\theta\}$ at θ is defined as

$$\gamma_\theta = \sqrt{\frac{\nu_{02}(\theta)}{I^2(\theta)} - \frac{\nu_{11}^2(\theta)}{I^3(\theta)}}$$

.

Remark: A detailed geometric justification for the name *curvature* is given in Efron(1975). The *curvature* γ_θ defined above works out to zero in the regular Exponential family, which acts like a straight line in the space of all probability distributions on the given Euclidean space. Nonexponential families have nonzero (at some values of θ) γ_θ and act like curves in the space of all probability distributions. Hence the name *curvature*. Before explaining a theoretical significance of γ_θ in terms of information loss suffered by the MLE, let us see a few examples.

Example 16.15. Suppose $f_\theta(x)$ is a member of the Exponential family with $f_\theta(x) = e^{\theta T(x) - \psi(\theta)} h(x)(d\mu)$. Then, $l_\theta(x) = \theta T(x) - \psi(\theta) + \log h(x)$, and hence, $\frac{\partial}{\partial\theta} l_\theta = T(x) - \psi'(\theta), \frac{\partial^2}{\partial\theta^2} l_\theta = -\psi''(\theta)$. Therefore, the Fisher information function $I(\theta) = \psi''(\theta)$. On the other hand, $\nu_{02}(\theta) = E_\theta[\frac{\partial^2}{\partial\theta^2} l_\theta]^2 - I^2(\theta) = 0$, and also, $\nu_{11}(\theta) = E_\theta[\frac{\partial}{\partial\theta} l_\theta \frac{\partial^2}{\partial\theta^2} l_\theta] = -\psi''(\theta) E_\theta[T(X) - \psi'(\theta)] = 0$, as $E_\theta[T(X)] = \psi'(\theta)$. It follows from the definition of the curvature that $\gamma_\theta = 0$.

Example 16.16. Consider a general location parameter density $f_\theta(x) = g(x - \theta)$, with support of g as the entire real line. Then, writing $\log g(x) = h(x), l_\theta = h(x - \theta)$, and by direct algebra, $I(\theta) = \int \frac{g'^2}{g}, \nu_{02}(\theta) = \int g h''^2 - (\int \frac{g'^2}{g})^2, \nu_{11}(\theta) = -\int h' h'' g$. All these integrals are on $(-\infty, \infty)$, and the expressions are independent of θ. Consequently, the curvature γ_θ is also independent of θ.

For instance, if $f_\theta(x)$ is the density of the central t distribution with location parameter θ and m degrees of freedom, then, on the requisite integrations, the different quantities are :

$$I(\theta) = \frac{m+1}{m+3}, \nu_{02}(\theta) = \frac{m+1}{m+3}[\frac{(m+2)(m^2+8m+19)}{m(m+5)(m+7)} - \frac{m+1}{m+3}], \nu_{11}(\theta) = 0.$$

On plugging into the definition of γ_θ, one finds that $\gamma_\theta^2 = \frac{6(3m^2+18m+19)}{m(m+1)(m+5)(m+7)}$; see Efron(1975). As $m \to \infty, \gamma_\theta \to 0$, which one would expect, since the t distribution converges to the Normal when $m \to \infty$, and the normal has zero curvature by the previous example. For the Cauchy case corresponding to $m = 1$, γ_θ^2 works out to 2.5. The curvature across the whole family as m varies between 1 and ∞ is a bounded decreasing function of m. The curvature becomes unbounded when $m \to 0$.

We now present an elegant result connecting curvature to the loss of information suffered by the MLE when f_θ satisfies certain structural and regularity assumptions. The density f_θ is assumed to belong to the *Curved Exponential Family*, as defined below.

Definition 16.4. Suppose for $\theta \in \Theta \subseteq \mathcal{R}$, $f_\theta(x) = e^{\eta' T(x) - \psi(\eta)} h(x)(d\mu)$, where $\eta = \eta(\theta)$ for some specified function from Θ to an Euclidean space $\mathcal{R}^k$. Then f_θ is said to belong to the Curved Exponential Family with carrier μ.

Remark: If η varies in the entire set $\{\eta : \int e^{\eta' T(x)} h(x) d\mu < \infty\}$, then the family would be a member of the Exponential family. By making the natural parameter η a function of a common underlying parameter θ, the Exponential family density has been restricted to a subset of lower dimension. In the curved Exponential family, the different components of the natural parameter vector of an Exponential family density are tied together by a common underlying parameter θ.

Example 16.17. Consider the $N(\theta, \theta^2)$ density, with $\theta \neq 0$. These form a subset of the two parameter $N(\mu, \sigma^2)$ densities, with $\mu(\theta) = \theta$ and $\sigma^2(\theta) = \theta^2$. Writing out the $N(\theta, \theta^2)$ density, it is seen to be a member of the curved Exponential family with $T(x) = (x^2, x)$ and $\eta(\theta) = (-\frac{1}{2\theta^2}, \frac{1}{\theta})$.

Example 16.18. Consider Gamma densities for which the mean is known to be 1. They have densities of the form $f_\theta(x) = \frac{e^{-x/\theta} x^{1/\theta - 1}}{\theta^{1/\theta} \Gamma(\frac{1}{\theta})}$. This is a member of the curved Exponential family with $T(x) = (x, \log x)$ and $\eta(\theta) = (-\frac{1}{\theta}, \frac{1}{\theta})$. Here is the principal theorem on information loss by the MLE in curved Exponential families.

Theorem 16.8. Suppose $f_\theta(x)$ is a member of the curved Exponential family and suppose the characteristic function $\psi_\theta(t)$ of f_θ is in $\mathcal{L}_p$ for some $p \geq 1$. Let $\hat{\theta}_n$ denote the MLE of θ based on n iid observations from f_θ, $I(\theta)$ = the Fisher information based on f_θ, and $I_{n,0}(\theta)$ the Fisher information obtained from the exact sampling distribution of $\hat{\theta}_n$ under θ. Then, $\lim_{n\to\infty}(nI(\theta) - I_{n,0}(\theta)) = I(\theta)\gamma_\theta^2$. In particular, the limiting loss of information suffered by the MLE is finite at any θ at which the curvature γ_θ is finite.

Remark: This is the principal theorem in Efron (1975). Efron's interpretation of this result is that the information obtained from n samples if one uses the MLE would equal the information obtained from $n - \gamma_\theta^2$ samples if the full sample is used. The interpretation hinges on use of Fisher information as the criterion. However, γ_θ has other statistical significances, e.g., in testing of hypothesis problems. In spite of the controversy about whether γ_θ has genuine inferential relevance, it seems to give qualitative insight into the wisdom of using methods based on the maximum likelihood estimate, when the minimal sufficient statistic is multidimensional.

16.10 Exercises

Exercise 16.1. * For each of the following cases, write or characterize the MLE and describe its asymptotic distribution and consistency properties:

(a) $X_1, \ldots, X_n$ are iid with density

$$f(x|\sigma_1, \sigma_2) = \begin{cases} ce^{-\frac{x}{\sigma_1}}, & x > 0 \\ ce^{\frac{x}{\sigma_2}}, & x < 0 \end{cases},$$

each of σ_1, σ_2 being unknown parameters;

REMARK: This is a standard way to produce a skewed density on the whole real line.

(b) $X_i, 1 \leq i \leq n$ are independent Poi(λx_i), the x_i being fixed covariates;

(c) $X_1, X_2, \cdots, X_m$ are iid $N(\mu, \sigma_1^2)$ and $Y_1, Y_2, \cdots, Y_n$ are iid $N(\mu, \sigma_2^2)$, and all $m + n$ observations are independent;

(d) m classes are represented in a sample of n individuals from a multinomial distribution with an unknown number of cells θ, and equal cell probabilities $\frac{1}{\theta}$.

Exercise 16.2. Suppose $X_1, \ldots, X_n$ are p - vectors uniformly distributed in the ball $B_r = \{x : ||x||_2 \leq r\}; r > 0$ is an unknown parameter. Find the MLE of r and its asymptotic distribution.

Exercise 16.3. * Two independent proof readers A and B are asked to read a manuscript containing N errors; $N \geq 0$ is unknown. n_1 errors are found by A alone, n_2 by B alone, and n_{12} by both. What is the MLE of N? What kind of asymptotics are meaningful here?

Exercise 16.4. * (Due to C. R. Rao) In an archaeological expedition, investigators are digging up human skulls in a particular region. They want to ascertain the sex of the individual from the skull and confirm that there is no demographic imbalance. However, determination of sex from an examination of the skull is inherently not an error free process.

Suppose they have data on n skulls, and for each one, they have classified the individual as being a male or female. Model the problem, and write the likelihood function for the following types of modelling:

(a) The error percentages in identifying the sex from the skull are assumed known;

(b) The error percentages in identifying the sex from the skull are considered unknown, but are assumed to be parameters independent of the basic parameter p, namely, the proportion of males in the presumed population;

(c) The error percentages in identifying the sex from the skull are considered unknown, and they are thought to be functions of the basic parameter p. The choice of the functions is also a part of the model.

Investigate, under each type of modelling, existence of the MLE of p, and write a formula, if possible under the particular model.

Exercise 16.5. * (Missing data) The number of fires reported in a week to a city fire station is Poisson with some mean λ. The city station is supposed to report the number each week to the central state office. But they do not bother to report it if their number of reports is less than 3.

Suppose you are employed at the state central office and want to estimate λ. Model the problem, and write the likelihood function for the following types of modelling:

(a) You ignore the weeks on which you did not get a report from the city office;

(b) You do not ignore the weeks on which you did not get a report from the city office, and you know that the city office does not send its report only when the number of incidents is less than 3;

(c) You do not ignore the weeks on which you did not get a report from the city office, and you do not know that the city office does not send its report only when the number of incidents is less than 3.

Investigate, under each type of modelling, existence of the MLE of λ, and write a formula, if possible under the particular model.

Exercise 16.6. * Find a location-scale parameter density $\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ for which the MLE of σ is $\frac{1}{n}\sum|X_i - M|$, where M is the median of the sample values $X_1, \ldots, X_n$. Find the asymptotic distribution of the MLE under this f (challenging!).

Exercise 16.7. * Consider the polynomial regression model $y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_i^j + \sigma e_i$, where e_i are iid $N(0,1)$. What is the MLE of m?

Exercise 16.8. * Suppose $X_1, \ldots, X_{m+n}$ are independent, with $X_1, \ldots, X_m \sim N(\mu_1, \sigma^2)$, $X_{m+1}, \ldots, X_{m+n} \sim N(\mu_2, \sigma^2)$, where $\mu_1 \leq \mu_2$ and σ^2 are unknown. Find the MLE of (μ_1, μ_2) and derive its asymptotic distribution when $\mu_1 < \mu_2, \mu_1 = \mu_2$.

Exercise 16.9. If $X_1, \ldots, X_n$ are iid $Poi(\lambda)$, show that $\frac{1}{\bar{X}+\frac{1}{n}}$ is second order unbiased for $\frac{1}{\lambda}$.

Exercise 16.10. * Find the limiting distribution of the MLE of (μ, σ, α) for the three-parameter gamma density

$$\frac{e^{-\frac{(x-\mu)}{\sigma}}(x-\mu)^{\alpha-1}}{\sigma^{\alpha}\Gamma(\alpha)}, \quad x \geq \mu, \quad \alpha, \sigma > 0, \quad -\infty < \mu < \infty.$$

Exercise 16.11. Suppose $X_1, \ldots, X_n$ are iid $Exp(\lambda)$. Find the MLE of the expected residual life $E(X_1 - t | X_1 > t)$ and its asymptotic distribution.

Exercise 16.12. * Suppose $X_1, \ldots, X_n$ are $BVN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, all five parameters being unknown. Find the MLE of $P(X_{11} > \mu_1, X_{12} > \mu_2)$, where $\binom{X_{11}}{X_{12}} = X_1$ and find its asymptotic distribution .

Exercise 16.13. * Derive a closed form expression for the mean squared error $R(\theta, T_n)$ of the Hodges superefficient estimate and show that $\limsup_{|\theta| \leq n^{-\frac{1}{4}}} nR(\theta, T_n) = \infty$.

Exercise 16.14. * Suppose $X_1, \ldots, X_n$ are iid with density

$$p\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}} + (1-p)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

where $0 < p < 1$ is known. Show that MLEs for μ, σ do not exist. How would you estimate μ, σ? What is the asymptotic distribution of your estimates?

Exercise 16.15. * Suppose X_i are iid $N(\mu, 1)$ where μ is known to be a positive integer. Let $g : \mathcal{R} \to \mathcal{R}$ be the function

$$g(x) = \begin{cases} x & \text{if } x \text{ is a prime} \\ -x & \text{if } x \text{ is not a prime} \end{cases}$$

(a) Is $\bar{X}$ consistent for μ?

(b) Is $g(\bar{X})$ consistent for $g(\mu)$?

Exercise 16.16. Suppose $X_i \overset{indep.}{\sim} Poi(\lambda^i)$ (thus X_i are not iid), $1 \leq i \leq n$,

(a) What is the MLE of λ?

(b) What is the asymptotic distribution of the MLE of λ?

Exercise 16.17. *

(a) Suppose X_i are iid $N(\mu, 1)$, but the collector rounds the X_i to Y_i, the nearest integer. Is $\bar{Y}$ consistent for μ?

(b) Find a consistent estimate for μ based on the Y_i.

Exercise 16.18. * Suppose X_i are iid nonnegative random variables and X_i are recorded as the integer closest to the X_i, say Y_i. Give a necessary and sufficient condition for $\bar{Y} \xrightarrow{\mathcal{P}} E(X_1)$.

Exercise 16.19. Suppose $X_i \overset{iid}{\sim} Poi(\lambda)$, where $\lambda > 0$ is known to be an integer.

(a) Find the MLE $\hat{\lambda}$ of λ.

(b) * What is $\lim_{n\to\infty} Var(\hat{\lambda})$?

Exercise 16.20. * Show that for iid $C(\theta, 1)$ data, the statistic $\frac{1}{\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)}$ is asymptotically normal. Find the appropriate centering, norming, and the variance of the asymptotic normal distribution.

Exercise 16.21. Compute the curvature of the $N(\theta, \theta^4)$ family.

Exercise 16.22. Compute the curvature of the family of Gamma densities with a known mean c.

Exercise 16.23. * For estimation of a Poisson mean λ, find the limiting information lost by s^2, the sample variance, compared to the information in the full sample. Is it finite ? Is it bounded ?

Exercise 16.24. Simulate the exact variance of the MLE of a double Exponential mean based on 20 iid samples, and compare the estimates based on expected and observed Fisher information with this exact value. Comment on the bias and the variability of these two estimates.

Exercise 16.25. * Is the central limit theorem for the MLE of a Logistic mean second order accurate ?

Exercise 16.26. * Derive a two term Edgeworth expansion for the MLE of the shape parameter of a Gamma distribution assuming the scale parameter is 1.

Exercise 16.27. * Derive a one term Edgeworth expansion for the MLE of θ in the $N(\theta, \theta^2)$ distribution.

16.11 References

Bai,Z.D. and Rao,C.R.(1991).Edgeworth expansion of a function of sample means, Ann.Stat., 19,3,1295-1315.

Basu, D. (1955). An inconsistency of the method of maximum likelihood, Ann. Math. Statist., 26, 144-145.

Bhattacharya,R.N. and Ghosh,J.K.(1978).On the validity of the formal Edgeworth expansion,Ann.Stat.,2,434-451.

Bickel,P.J. and Doksum,K.(2001).Mathematical Statistics, Basic Ideas and Selected Topics,Vol I,Prentice Hall,Upper Saddle River,NJ.

Brown,L.D.(1986).Fundamentals of Statistical Exponential Families,IMS Lecture Notes Monograph Ser. 9,Institute of Mathematical Statistics,Hayward,CA.

DasGupta,A.(2004).On the risk function of superefficient estimates,Preprint.

Efron,B.(1975).Defining the curvature of a statistical problem,with applications to second order efficiency,Ann.Stat.,3,6,1189-1242.

Ghosh, M. (1994). On some Bayesian solutions of the Neyman-Scott problem, Stat. Dec. Theory and Rel. Topics, V, 267-276, J. Berger and S.S. Gupta Eds., Springer-Verlag, New York.

Hodges,J.L.(1952).Private communication to Lucien LeCam

LeCam,L.(1953).On some asymptotic properties of maximum likelihood estimates and related Bayes estimates,Univ. of Calif. Publ.,1,277-330.

Lehmann,E.L. and Casella,G.(1998).Theory of Point estimation,2nd Edition,Springer, New York.

McLachlan, G. and Krishnan, T. (1997). The EM algorithm and Extensions, John Wiley, New York.

Neyman,J. and Scott,E.(1948).Consistent estimates based on partially consistet observations,Econometrica,16,1-32.

Perlman,M.D.(1983).The limiting behavior of multiple roots of the likelihood equation,in Recent Advances in Statistics,339-370,Academic Press,New York.

Pfanzagl,J.(1973).The accuracy of the normal approximation for estimates of vector parameters,Z.Wahr.Verw.Gebiete,25,171-198.

Rao,C.R.(1961).Asymptotic efficiency and limiting information,Proc. Fourth Berkeley Symp. Math.Statist. and Prob.,I,531-545,Univ.Calif.Publ.,Berkeley.

Rao,C.R.(1962).Efficient estimates and optimum inference procedures in large samples,JRSS,Ser. B,24,46-72.

Rao,C.R.(1963).Criteria of estimation in large samples,Sankhya Ser.A,25,189-206.

van der vaart,Aad(1998).Superefficiency,Festschrift for Lucien LeCam,397-410, Springer,New York.