CHAPTER 10

The New Likelihoods and the Neyman–Scott Problems

10.1. Introduction. The traditional method of using the likelihood to make inference about the parameter of interest is to use the so-called profile likelihood, which is the likelihood maximized with respect to the nuisance parameters. It has been known for a long time that this is a wrong thing to do if there are many nuisance parameters. The Neyman–Scott examples provide a dramatic example of this. In one of them, maximizing the profile likelihood, which is the same thing as using the mle, provides an inconsistent estimate of the parameter of interest.

Two modifications of profile likelihood have been proposed recently. Conditional likelihood, owing to Cox and Reid (1987) and adjusted likelihood, due to McCullagh and Tibshirani (1990), both try to modify the profile likelihood so that it may be expected to behave more like an honest likelihood. Both have been tried on the two Neyman–Scott examples we discuss here.

In Section 10.2 we introduce and briefly study these new likelihoods by methods of higher order asymptotics. In Section 10.3 we introduce two Neyman–Scott examples, as well as a general formulation, and introduce estimates which are FOE in a sense appropriate for these problems. We then apply the new likelihoods to these examples and note that they fail to provide FOE estimates. We suggest that they are not the right answers to these problems. A modified version of the general Neyman–Scott problem is posed for which higher order asymptotics seems to be the right tool and the two new likelihoods may do better than profile likelihood.

10.2. Conditional and adjusted likelihood. We consider $\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is the nuisance parameter. We will require θ_1 and θ_2 to be orthogonal and so, as mentioned in Chapters 8 and 9, θ_1 will need to be real valued.

Let $L(\theta_1, \theta_2) = p(X_1, X_2, \dots, X_n | \theta_1, \theta_2)$. The profile likelihood is

$$L_p(\theta_1) = \sup_{\theta_2} L(\theta_1, \theta_2) = L\big(\theta_1, \hat{\theta}_2(\theta_1)\big), \tag{10.1}$$

where $\hat{\theta}_2(\theta_1)$ is the mle for θ_2 given θ_1 is known.

To introduce the conditional likelihood, introduce a statistic T (of dimension $n - d_2$, where d_2 is the dimension of θ_2) such that $(T, \hat{\theta}_2(\theta_1))$ provides a one-to-one transformation of $(X_1, X_2, \dots, X_n)$ for each θ_1. Let the density of $(T, \hat{\theta}_2(\theta_1))$ be denoted by $q(t, \hat{\theta}_2(\theta_1) | \theta_1, \theta_2)$. Let the conditional density of T given $\hat{\theta}_2(\theta_1)$ be $q(t | \hat{\theta}_2(\theta_1), \theta_1, \theta_2)$. Use of this tries to correct for the substitution of $\hat{\theta}_2(\theta_1)$ for θ_2 in the profile likelihood. The reason for orthogonality is to reduce the undesirable effect of changes in the conditioning statistic $\hat{\theta}_2(\theta_1)$ with θ_1.

By two applications of the magic formula (Chapter 8), Cox and Reid (1987) show the logarithm of conditional likelihood $q(t|$ etc.) can be approximated by

$$\log p\big(X_1, X_2, \dots, X_n | \theta_1, \hat{\theta}_2(\theta_1)\big) + \tfrac{1}{2} \log nb\big(\theta_1, \hat{\theta}_2(\theta_1)\big) = L_c(\theta_1), \tag{10.2a}$$

where

$$b(\theta_1, \theta_2) = - \left. \frac{\partial^2 \log L}{\partial \theta_2^2} \right|_{(\theta_1, \theta_2)}.$$

We refer to (10.2a) as the (approximate) conditional likelihood due to Cox and Reid.

McCullagh and Tibshirani (1990) also start with the profile likelihood, and begin by adjusting the corresponding score function

$$\frac{\partial}{\partial \theta_1} \log L_p(\theta_1) = U(\theta_1) \tag{10.2b}$$

[see (10.1)] to a new function of θ_1 so as to have mean zero (as a score function derived from an honest likelihood function should). This is done by subtracting the expectation of $U(\theta_1)$ under (θ_1, θ_2). Finally, there is an adjustment for the variance also, the need for which is less clear. After these adjustments, we end up with

$$\begin{aligned} &V(X_1, X_2, \dots, X_n, \theta_1) \\ &\quad = \big\{U(\theta_1) - E\big(U(\theta_1) | \theta_1, \hat{\theta}_2(\theta_1)\big)\big\} w(\theta_1), \end{aligned} \tag{10.3}$$

where

$$\begin{aligned} w(\theta_1) = \Bigg[&-E\left\{ \frac{\partial^2}{\partial \theta_1^2} \log L_p(\theta_1) | \theta_1, \hat{\theta}_2(\theta_1) \right\} \\ &+ \frac{\partial}{\partial \theta_1} E\big\{U(\theta_1) | \theta_1, \hat{\theta}_2(\theta_1)\big\} \Bigg] \\ &\times \Big[\mathrm{Var}\big\{U(\theta_1) | \theta_1, \hat{\theta}_2(\theta_1)\big\} \Big]^{-1}. \end{aligned} \tag{10.4}$$

The integral

$$\int_{\theta_{10}}^{\theta_1} V(X_1,\dots,X_n,t)\,dt = \log L_{ap}(\theta_1) \tag{10.5}$$

is the new adjusted (log profile) likelihood of McCullagh and Tibshirani. If we maximize the profile likelihood, we get the mle $\hat{\theta}_1$, and if we test H_0: $\theta_1 = \theta_{10}$ by $2\{\log L_p(\hat{\theta}_1) - \log L_p(\theta_{10})\}$, we get the likelihood ratio test. If we replace the profile likelihood by L_c, we get a maximum conditional likelihood estimate $\hat{\theta}_{1c}$ and a conditional likelihood ratio test. The estimate $\hat{\theta}_a$ and the adjusted likelihood ratio test are similarly defined.

Under regularity conditions [see Mukerjee (1992) and Ghosh and Mukerjee (1994)], the following facts have been proved by the delta method in the cited references. We recall that θ_1 and θ_2 are orthogonal, as may be assumed without loss of generality for scalar θ_1.

1. The conditional likelihood ratio test has the same power as the likelihood ratio test with a known nuisance parameter up to $o(n^{-1/2})$ for the local (Pitman) alternatives of the form $\theta_1 = \theta_{10} + n^{-1/2}\delta_1$. This is not true for the usual (profile) likelihood ratio test.
2. The conditional likelihood ratio test admits of Bayesian and frequentist Bartlett correction, and matching of probabilities as in Section 8.4 can be done with the conditional likelihood ratio statistic replacing the likelihood ratio statistic. For Example 8.2, the right invariant Haar measure still satisfies the resulting equation for the prior.
3. The adjusted likelihood ratio test admits of Bartlett correction, but, in general, we may not be able to define an adjusted likelihood if θ_1 is multidimensional since the differential equations arising from adjustment over different components of θ_1 will not, in general, be consistent.
4. Adjusted likelihood and conditional likelihood are indistinguishable at the level of second order asymptotics. In particular, the adjusted likelihood ratio test has the optimum property mentioned in paragraph 1 and $n(\hat{\theta}_c - \tilde{\theta}_1) \to_p 0$.
5. Paragraphs 3 and 4 remain true if in the definition of adjusted likelihood we do not adjust for variance, that is, we do not divide by $w(\theta_1)$ in the definition of V.

10.3. Neyman–Scott problems. In the Neyman and Scott (1948) problems there is a parameter of interest, called the structural parameter, and many nuisance parameters. In Section 10.2, the number of nuisance parameters is held fixed, but in the Neyman–Scott examples the number of nuisance parameters grows very fast, at the same rate as the sample size. The Neyman–Scott problems are the simplest examples where, because of a high dimensional parameter space, a classical procedure like the mle fails dramatically.

EXAMPLE 10.1. Consider r.v.'s X_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$. The r.v.'s are independent but not identically distributed. For fixed i, $X_{i1}, \ldots, X_{ik}$ are i.i.d. $N(\mu_i, \sigma^2)$. Here σ^2 is the parameter of interest and the μ_i's are nuisance parameters. In the asymptotics k is held fixed, $n \to \infty$. The maximum likelihood (and hence maximum profile likelihood) estimate of σ^2 is

$$\hat{\sigma}^2 = \sum \sum \left(X_{ij} - \overline{X}_i\right)^2 / nk \to \frac{k-1}{k} \sigma^2 \quad \text{a.s.}$$

(Here $\overline{X}_i = k^{-1} \Sigma_i X_{ij}$.) If $k = 2$, $\hat{\sigma}^2 \to \sigma^2/2$ a.s. It is clear that the mle misses so badly because the profile likelihood makes no adjustment for replacing unknown parameters by estimates and in the process gets the degrees of freedom wrong. In this example the correct d.f. is $n(k-1)$ and the "right" estimate is $\tilde{\sigma}^2 = \Sigma\Sigma(X_{ij} - \overline{X}_i)^2/n(k-1)$.

If we maximize the adjusted likelihood, we have to solve

$$U(\sigma^2) - E\left(U(\sigma^2) | \sigma^2, \overline{X}_i, i = 1, \ldots, n\right) = 0,$$

where

$$U(\sigma^2) = \frac{d \log L_{ap}(\sigma^2)}{d\sigma^2},$$

$$\log L_{ap}(\sigma^2) = \frac{-nk \log \sigma^2}{2} - \frac{1}{2\sigma^2} \sum \sum \left(X_{ij} - \overline{X}\right)^2.$$

Hence $\hat{\theta}_a^2 = \tilde{\sigma}^2$, the "right" estimate. One can check that the same is true of $\hat{\theta}_c^2$ obtained by maximizing the conditional likelihood.

EXAMPLE 10.2. X_{ij}'s are as in Example 10.1, but the roles of mean and variance are interchanged. Thus X_{ij} is $N(\mu, \sigma_i^2)$. It is sometimes called the problem of common mean. (One may compare with the Behrens–Fisher problem of Chapters 2 and 3, which is superficially similar, but $n = 2$, $k \to \infty$, so that the asymptotics is quite different.) In this case,

$$\log L_p(\mu) = -\frac{k}{2} \sum_{i=1}^{n} \log \hat{\sigma}_i^2(\mu) - \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{\left(X_{ij} - \mu\right)^2}{2\hat{\sigma}_i^2(\mu)},$$

where

$$\hat{\sigma}_i^2(\mu) = \frac{1}{k} \sum_1^k \left(X_{ij} - \overline{X}_i\right)^2 + \left(\overline{X}_i - \mu\right)2.$$

Assuming σ_i's are bounded above and bounded below by a positive number, one can show both the mle $\hat{\theta}$ and the grand mean $\overline{\overline{X}} = \Sigma\Sigma(X_{ij})/nk$ are consistent and asymptotically normal with mean θ. It is also possible to show, using the symmetry of the normal, that $\hat{\theta}_c$ and $\hat{\theta}_a$ are consistent and asymptotically normal with mean θ. However, none of these four estimates are "right" in the sense that there is an estimate which is asymptotically normal with mean θ and smaller variance. In fact, in a sense to be explained a little later, this last estimate is FOE in this problem and none of the two new likelihoods can find it.

We now introduce the general form of the Neyman–Scott problem and the natural class of estimates associated with them.

General problem. Let X_{ij}'s, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, k$, be independent and for each i let $X_{ij}, X_{i2}, \dots, X_{in}$ be i.i.d. with density $p(x|\theta_1, \theta_{2i})$. Note that the form of the density and the value of the parameter of interest θ_1 remain the same for all i, while the value of the nuisance parameters θ_2 changes with i. We assume θ_{2i}'s lie in a compact set, as in Example 10.2.

An estimate for θ_1 is obtained by solving an equation of the form

$$\sum_{i+1}^{n} \psi(X_{i1}, X_{i2}, \dots, X_{ik}, \theta_1) = 0, \tag{10.6}$$

where ψ is continuously differentiable in θ_1 and

$$E_{\theta_1,\theta_2} \psi(X_{i1}, \dots, X_{ik}, \theta_1) = 0 \qquad \forall\, (\theta_1, \theta_2). \tag{10.7}$$

As in the case of the likelihood equation, or more generally, Huber's M estimates, one can easily show by Taylor expansion arguments that there is a solution T_ψ which converges in probability to θ_1 and T_ψ is A.N. $(\theta_1, \sigma^2_{\psi,n}/n)$, where

$$\begin{aligned} \sigma^2_{\psi,n} &= \frac{A_{1n}}{A_{2n}}, \\ A_{1n} &= \frac{1}{n} \sum_{i=1}^{n} E_{\theta_1,\theta_{2i}} \psi^2(X_{i1}, \dots, X_{ik}, \theta_1), \\ A_{2,n} &= \left(\frac{1}{n} \sum_{i=1}^{n} E_{\theta_1,\theta_{2i}} \psi_{\theta_1}(X_{i1}, \dots, X_{ik}, \theta_1) \right)^2, \end{aligned} \tag{10.8}$$

and $\psi_{\theta_1} = (\partial/\partial\theta_1)\psi$.

Following Amari and Kumon (1984), we call them C_1 estimates. It is clear that an mle satisfies an equation like (10.6), but in the absence of (10.7), will not be consistent in general. In fact, consistency as in Example 10.2 is an exception rather than the rule. The same is true of $\hat{\theta}_c\ \hat{\theta}_a$. Each satisfies an equation like (10.6), but (10.7) is an exception rather than the rule.

Even when $\hat{\theta}_a$ or $\hat{\theta}_c$ are consistent, they are not FOE in the sense explained below.

Let the distribution function of $\theta_{21}, \dots, \theta_{2n}$ be defined as

$$G_n(y) = \{\#\theta_{2i}\text{'s} \le y\}/n.$$

Of course since the θ_{2i}'s are unknown, so is G_{2n}. Note that the asymptotic variance of $\sqrt{n}(T_\psi - \theta_1)$ is the following functional in G_n:

$$\sigma^2_\psi(\theta_1, G_n) = \frac{\int \{E_{\theta_1,\theta_2} \psi^2(X_{i1}, \dots, X_{ik}, \theta_1)\} G_n(d\theta_2)}{\left[\int E_{\theta_1,\theta_2} \psi_{\theta_1}(X_{i1}, \dots, X_{ik}, \theta_1) G_n(d\theta_2)\right]^2}. \tag{10.9}$$

If we try to use the "Cramér–Rao" bound based on nk observations, say, $I^{11}(\theta_1, \theta_{21}, \dots, \theta_{2n})$, it does not work because, in general, the bound is not sharp globally, even asymptotically. Here $[I^{ij}] = [I_{ij}]^{-1}$. We need to develop a sharper bound, making use (among other things) of the fact that the estimates are invariant under permutation of i.

Lindsay (1980), in a pioneering paper, has noted that we may interpret this functional as a variance in what is called the mixture or empirical Bayes setup of Robbins.

Mixture setup. Consider the general formulation of the Neyman–Scott problem, but assume θ_{2i}'s are i.i.d. r.v.'s taking values in a compact set Θ_2 with common distribution function G. The object is still to estimate θ_1, when both θ_1 and G are unknown. Using the theory of semiparametric inference, one can find an analogue of Fisher's information which we denote as $I(\theta_1, G)$. As in Chapter 1, an estimate T_n of θ_1 is FOE or simply efficient if $\sqrt{n}\,(T_n - \theta_1)$ is A.N. $(0, (I(\theta_1, G))^{-1})$ uniformly on compact θ_1-sets and uniformly in G.

If we use an estimate T_ψ in the mixture setup, the $\sqrt{n}\,(T_\psi - \theta_1)$ is still A.N. $(0, \sigma_\psi^2(G))$, where $\sigma_\psi^2(\cdot)$ is the functional defined in (10.9). Hence (under uniformity of asymptotic normality with respect to G and θ_1 in compact sets),

$$\sigma_\psi^2(\theta_1, G) \geq I^{-1}(\theta_1, G). \tag{10.10}$$

Consequently, a T_ψ for which (10.10) is an equality for all θ_1, G may be called FOE. The "right" estimate in Example 10.1 is FOE [but does not attain even asymptotically the simple minded Cramér–Rao lower bound $I^{11}(\theta_1, \theta_{21}, \dots, \theta_{2n})$].

Unfortunately, in Example 10.2, and generally, even this better bound is not attained within the class of estimates T_ψ. Amari and Kumon (1984) have restricted attention to a subclass of the estimates T_ψ and, using a covariant derivative, have found a lower bound to asymptotic variance which can be attained. While their analysis is elegant, there does not seem to be any compelling reason to confine attention to their subclass. Following Bickel and Klaassen (1986), we prefer to enlarge the class of estimates to include all estimates which are regular in the following sense; see Bhanja and Ghosh (1992). An estimate T_n is regular if the following happen:

1. In the Neyman–Scott setup $\sqrt{n}\,(T_n - \theta_1)$ is A.N. $(0, \sigma^2(\theta_1, G_n))$ uniformly in $\theta_{2i} \in \Theta_2$ and compact θ_1-sets.
2. In the mixture setup $\sqrt{n}\,(T_n - \theta_1)$ is A.N. $(0, \sigma^2(\theta_1, G)$ uniformly in G and compact θ_1-sets.

Within this class, it is still true that $\sigma^2(\theta_1, G) \geq (I(\theta_1, G))^{-1}$. Hence, by occurrence 2 one may call a regular estimate T_n FOE in the Neyman–Scott setup if equality is attained for all θ_1, G.

The general theory of such estimates, given in Bhanja and Ghosh (1992) is very technical. It involves also a continuity assumption which is difficult to check, in general, but holds for the two examples in this chapter. We have already indicated an FOE estimate for Example 10.1. We will now describe briefly an FOE for Example 10.2.

In the Neyman–Scott framework of Example 10.2, pretend that $\theta_{21}, \dots, \theta_{2n}$ are i.i.d. $\sim G_n$ as in the mixture setup. Then the integrated likelihood equation is

$$\frac{d}{d\theta_1} \sum_{i=1}^{n} \log p\big(X_{i1}, \dots, X_{ik} | \theta_1, \hat{G}_n\big) = 0, \tag{10.11}$$

where $p(X_{i1}, \dots, X_{ik} | \theta_1, G_n) = \int p(X_{i1}, \dots, X_{ik} | \theta_1, \theta_{2i}) G_n(\theta_{2i})$, $p(X_{i1}, \dots, X_{ik} | \theta, \hat{G}_n)$ is obtained by replacing G_n with $\hat{G}_n$ and $\hat{G}_n$ is a nonparametric mle. We believe this is not only a natural estimate, but also FOE. However, there are some technical difficulties in proving that it is FOE. We present, therefore, another estimate which is shown in Bhanja and Ghosh (1992) to be FOE.

Permute the i's at random and produce two sets of n_1 and n_2 vectors X_i, with $n_1 + n_2 = n$, $n_1/n_2 \to 1$ as $n \to \infty$. Call the permuted observations Y_{ij}'s. Note $Y_{ij} = X_{i'j}$, $j = 1, 2, \dots, k$, for some i'. Call the permuted θ_{2i}'s η_i's. Then $\eta_i = \theta_{2i'}$. We write $Y_i = (Y_{i1}, \dots, Y_{ik})$.

From the two sets, get consistent estimates $\hat{G}_{n1}, \hat{G}_{n2}$ of G_{n1}, G_{n2} where

$$G_{n1}(h) = \#\{\eta_i\text{'s}; 1 \le i \le \eta_1, \eta_i \le h\}/n_1$$

and G_{n2} is similarly defined. Now solve

$$\frac{d}{d\theta}\left(\sum_{i=1}^{n_1} \log f\big(Y_i, \theta, \hat{G}_{n2}\big) + \sum_{i=n_1+1}^{n_1+n_2} \log f\big(Y_i, \theta, \hat{G}_{n1}\big) \right) = 0 \tag{10.12}$$

by a one-step Newton–Raphson method. Independence of Y_i and $\hat{G}_{n2}$ in the first sum and Y_i and $\hat{G}_{n1}$ in the second sum makes (10.12) relatively easy to handle.

We explain how $\hat{G}_{in}, \hat{G}_{2n}$ are calculated for $k = 3$. Let $s_i^2 = \sum_{i=1}^{3} (Y_{ij} - \overline{Y}_i)^2$ and look at the empirical distribution of s_i^2's, $i = 1, 2, \dots, n_1$, that is, at

$$F_{n1}^{(h)} = \big(n_1^{-1}\big)\#\big\{s_i^2; 1 \le i \le n_1, s_i^2 \le h\big\},$$

which is a consistent estimate of its expectation. Call this expectation $A_n(h)$. Then $A_n(h)$ is a scale mixture of exponentials with G_{n1} as the mixing distribution. Hence the algorithm of Jewell (1982) for estimating G_{n1} can be used. This is $\hat{G}_{n1}$.

Simulations show the asymptotics provides good approximation for $n = 100$. Incidentally,

$$I(\theta_1, G_n) = E\left[\left\{ \frac{d \log p(X_{11}, \dots, X_{1k} | \theta_1\, G_n)}{d\theta_1} \right\}^2 \middle| \theta_1, G_n \right].$$

To sum up, while there are FOE estimates for Example 10.2, maximizing the profile, conditional or adjusted likelihood will not produce an FOE. However, it is likely that $\hat{\theta}_1$, $\hat{\theta}_{1a}$ and $\hat{\theta}_{1c}$ are all FOE when $k \to \infty$, as $n \to \infty$, possibly at a suitable rate. In such cases, higher order asymptotics would help discriminate among them and would probably show the superiority of the new likelihoods. This is a problem that needs attention.