CHAPTER 6

Third Order Efficiency, Admissibility and Minimaxity

6.1. Third order efficiency in the general case. We assume regularity conditions on $p(x|\theta)$ so that Theorem 5.1d holds, and assume the following three conditions on estimates:

CONDITION 1. $E\{(T_n - \theta)^2|\theta\} = n^{-1}I^{-1}(\theta) + n^{-2}g(\theta) + o(n^{-2})$, uniformly in compact θ sets where $g(\theta)$ is a continuous function of θ.

CONDITION 2. $E\{(T_n - \theta)|\theta\} = n^{-1}b(\theta) + O(n^{-(1+\varepsilon)})$, uniformly in compact θ sets where $b(\theta)$ is continuously differentiable and $\varepsilon > 0$.

CONDITION 3. $\sup_{\theta \in [a,b]} E\{(T - \theta)^4|\theta\} \leq M_{a,b} < \infty$ bound intervals $[a, b]$.

We assume $\hat{\theta}$ satisfies these conditions also. That expectation of $\hat{\theta}$ and variance of $\hat{\theta}$ have expansions in powers of n^{-1} follows from the regularity conditions assumed earlier. The expansions agree with those obtained by the delta method. If b_0 and g_0 stand for b and g when T_n is replaced by $\hat{\theta}$, we do not need to make the additional assumption that b_0 is continuously differentiable and g_0 is continuous. The quantities b_0 and g_0 as calculated earlier for curved exponentials have the same expressions in the general case.

Fix θ_0 and introduce a sequence of priors $\pi_n \in D_\infty$, concentrating on the interval (a_n, b_n) with $a_n = \theta_0 - (\log n)^{1/4}$, $b_n + \theta_0 + (\log n)^{14}$. [In fact, choose the prior exhibited in (5.6).] It can be shown that Theorem 5.1d continues to hold if π is replaced by π_n. Write B_n', corresponding to π_n, in the form

$$B_n' = \hat{\theta} + d_n(\hat{\theta})/n. \tag{6.1}$$

Let

$$\hat{\theta}^* = \hat{\theta} - \frac{1}{n}\big(b_0(\hat{\theta}) - b(\theta)\big), \tag{6.2}$$

$$T_n' = T_n + \frac{1}{n}\big(b_0(\hat{\theta}) - b(\hat{\theta}) + d_n(\hat{\theta})\big), \tag{6.3}$$

$$T_n'' = \text{the natural truncation of } T_n' \quad [\text{see (5.21)}]. \tag{6.4}$$

Under the regularity conditions assumed,

$$E\Big\{(\hat{\theta}^* - \theta)^2 | \theta\Big\} = \frac{1}{nI(\theta)} + \frac{g_0(\theta)}{n^2} + o(n^{-2}). \tag{6.5}$$

We sketch a proof that

$$g(\theta_0) \geq g_0(\theta_0) \qquad \forall\, \theta_0. \tag{6.6}$$

A direct argument shows

$$\begin{aligned} &E\Big\{(\hat{\theta}^* - \theta)^2 | \theta\Big\} - E\big\{(B_n'' - \theta)^2 | \theta\big\} \\ &\quad = E\big\{(T_n - \theta)^2 | \theta\big\} - E\big\{(T_n'' - \theta)^2 | \theta\big\} + o(n^{-2}). \end{aligned} \tag{6.7}$$

It is easy to check this "formally" by the delta method. To see this simply note that T_n' is the same perturbation of T_n as B_n' is of $\hat{\theta}_n^*$.

Hence

$$\begin{aligned} &\int \Big[E\Big\{(\hat{\theta}^* - \theta)^2 | \theta\Big\} - E\big\{(B_n'' + \theta)^2 | \theta\big\}\Big]\pi_n(\theta)\, d\theta \\ &\quad = \int \Big[E\big\{(T_n - \theta)^2 | \theta\big\} - E\big\{(T_n'' - \theta)^2 | \theta\big\}\Big]\pi_n(\theta)\, d\theta + o(n^{-2}). \end{aligned} \tag{6.8}$$

On the other hand, by the Bayes property of B_n'',

$$\begin{aligned} &\int \Big[E\big\{(B_n'' - \theta)^2 | \theta\big\}\Big]\pi_n(\theta)\, d\theta \\ &\quad \leq \int \Big[E\big\{(T_n'' - \theta)^2 | \theta\big\}\Big]\pi_n(\theta)\, d\theta + o(n^{-2}). \end{aligned} \tag{6.9}$$

By (6.8) and (6.9),

$$\begin{aligned} &\int_{a_n}^{b_n} \Big[E\Big\{(\hat{\theta}^* - \theta)^2 | \theta\Big\}\Big]\pi_n(\theta)\, d\theta \\ &\quad \leq \int_{a_n}^{b_n} \Big[E\big\{(T_n - \theta)^2 | \theta\big\}\Big]\pi_n(\theta)\, d\theta + o(n^{-2}), \end{aligned} \tag{6.10}$$

that is,

$$\int_{a_n}^{b_n} g_0(\theta)\pi_n(\theta)\, d\theta \leq \int_{a_n}^{b_n} g(\theta)\pi_n(\theta)\, d\theta + o(n^{-2}), \tag{6.11}$$

which implies (6.6). Instead of choosing a sequence of priors as above, one can also fix $a < \theta_0 < b$ and then choose a prior π on (a, b) belonging to D_s,

$11 < s \leq \infty$. Going through exactly the same steps as above, one would get, in place of (6.11),

$$\int_a^b g_0(\theta)\pi(\theta)\,d\theta \leq \int_a^b g(\theta)\pi(\theta)\,d\theta. \tag{6.11a}$$

Now shrinking the interval toward θ_0, one gets (6.6).

This argument first appears informally in Ghosh and Subramanyam (1974). The rigorous version for squared error loss is developed in Ghosh, Sinha and Joshi (1982). This chapter is based on that paper, but the tedious details have all been omitted. Bickel, Götze and van Zwet (1985) develop the same argument for vector θ and general loss functions, but using perturbations of loss functions rather than perturbations of estimates. The technique of perturbed loss functions is interesting for its own sake and is explained in the next chapter.

Another Bayesian proof for general loss functions is given in Ghosh, Sinha and Wieand (1980). This is much less technical (and tedious!) than Ghosh, Sinha and Joshi (1982), but does not extend to the multiparameter case.

6.2. Remarks on general third order efficiency results and proofs. As mentioned earlier a proof for general bowl-shaped symmetric loss functions (not necessarily smooth) and multiparameter problems with regularity conditions but permitting dependence and not requiring Edgeworth expansions is given in Bickel, Götze and van Zwet (1985). The argument is similar to that of Section 6.1 except that, instead of using perturbations of estimates, the loss is cleverly perturbed. Unfortunately the proof is still messy and not all the details are given. Except for the result mentioned in Section 6.5, this appears to be the most general result on third order efficiency.

For the one parameter case, under the assumption of valid Edgeworth expansions (uniformly on compact θ-sets) or smooth loss functions (and some uniformity), a proof with complete details is given in Ghosh, Sinha and Wieand (1980). This seems to be the cleanest proof of a fairly general version of the result. The argument is Bayesian but avoids using the expansion of Chapter 5 by reducing the comparison of estimates to comparison of tests based on them, as in Rao (1963). For this reason it does not seem to extend to the multiparameter case.

The fact that a general third order efficiency result holds, without the restriction to curved exponentials or squared error loss or Fisher consistent estimates, was first conjectured in Ghosh and Subramanyam (1974). They also outlined a heuristic argument for such a conjecture, which forms the basis of proof in Section 6.1 as well as Bickel, Götze and van Zwet (1985).

6.3. Median unbiasedness and matching bias. We first note a fact about Edgeworth expansions. Suppose $\sqrt{nI}(T - \theta) = Y_1$ has valid Edgeworth expansion up to $o(n^{-1})$. Let $\sqrt{nI}(T' - \theta) = Y_2$, where $T' = T + c(T)/n$ and $c(\cdot)$ is continuously differentiable. We have seen in Chapter 2 that Y_2 has a

valid Edgeworth expansion up to $o(n^{-1})$ and the Edgeworth expansion associated with

$$Y_2 - c(\theta)\sqrt{I/n} = \sqrt{nI}\,(T' - \theta - c(\theta)/n)$$

is identical with that of $\sqrt{nI}(T - \theta)(1 + c'(\theta)/n)$. Since the last random variable is a linear transformation of Y_1, we now see how the cumulants of Y_2 can be obtained from those of Y_1. In particular, recalling the structure of cumulants of Y_1, at least for Fisher consistent estimates $T = H(\bar{Z})$ [(2.6)], it is clear that Y_1 and Y_2 have the same third and fourth cumulants up to $o(n^{-1})$. The first two cumulants of Y_2 differ from those of Y_1 as follows [keep in mind the structure given by (2.6)]:

$$\text{first cumulant of } Y_2 = (\text{first cumulant of } Y_1) + \frac{c\sqrt{I}}{\sqrt{n}} + o(n^{-1}), \tag{6.12a}$$

$$\text{second cumulant of } Y_2 = (\text{second cumulant of } Y_1) + \frac{2c'(\theta)}{n} + o(n^{-1}). \tag{6.12b}$$

We now choose $c(\theta)$ so that

$$\sqrt{n}\left\{\left|P_\theta\{\sqrt{nI}\,(T' - \theta) \geq 0\} - \tfrac{1}{2}\right| + \left|\sqrt{n}\,P_\theta\{\sqrt{nI}\,(T' - \theta) \leq 0\} - \tfrac{1}{2}\right|\right\} \to 0. \tag{6.13a}$$

It is clear that (6.13a) implies

$$n\left\{\left|P_\theta\{\sqrt{nI}\,(T' - \theta) \geq 0\} - \tfrac{1}{2}\right| + \left|\sqrt{n}\,P_\theta\{\sqrt{nI}\,(T' - \theta) \leq 0\} - \tfrac{1}{2}\right|\right\} \to 0. \tag{6.13b}$$

We refer to (6.13a) and (6.13b) as median unbiasedness up to $o(n^{-1/2})$and $o(n^{-1})$, respectively. It is clear from the remarks made earlier about Y_1 and Y_2, that a $c(\cdot)$ satisfying (6.13a) and hence (6.13b) can be found. It is clear from (2.21) that $c(\cdot)$ has to be chosen such that $\kappa'_{31} = \kappa_{31}$ and $\kappa'_{11} = \kappa_{11} + c\sqrt{I}$ have to satisfy a linear relation for median unbiasedness. Moreover, Y_1 and Y_2 will have the same third and fourth cumulants up to $o(n^{-1})$.

Suppose now we choose $d(\cdot)$ such that $\theta' = \theta + d(\theta)/n$ matches bias (or equivalently expectation) of T' up to $o(n^{-1})$. Then both $\sqrt{nI}(T' - \theta)$ and $\sqrt{nI}(\hat{\theta}' - \theta)$ have the same κ_{11}.

It is proved in Ghosh, Sinha and Wieand (1980) that having the same κ_{11} and first order efficiency of T_n entails $\sqrt{nI}(T' - \theta)$ and $\sqrt{nI}(\hat{\theta}' - \theta)$ have the same κ_{31}. Thus $\sqrt{nI}(\hat{\theta}' - \theta)$ must also be median unbiased if the bias of $\hat{\theta}'$ matches that of T'_n up to $o(n^{-1})$.

Hence, a median unbiased [up to $o(n^{-1/2})$ or, equivalently, $o(n^{-1})$] $\hat{\theta}'$ is third order better than a median unbiased [up to $o(n^{-1/2})$ or, equivalently, $o(n^{-1})$] FOE T'_n [assuming valid Edgeworth expansions exist for $\sqrt{n}(\hat{\theta} - \theta)$ and $\sqrt{n}(T - \theta)$].

Third order results of Akahira and Takeuchi, and Pfanzagl are usually stated for median unbiased estimates.

6.4. First order efficiency implies second order efficiency. Suppose we are given FOE T_1, T_2, and T_1', T_2' are perturbations of the form $T_i + c_i(T_i)/n$, c_i continuously differentiable, so that $\sqrt{nI}(T_i' - \theta)$ have the same bias up to $o(n^{-1/2})$ or they are unbiased up to $o(n^{-1/2})$ or they are median unbiased up to $o(n^{-1/2})$. (in each case this is equivalent to matching bias up to $o(n^{-1})$, but this fact is not relevant here]. Then they both have the same κ_{31} as that of a matching $\sqrt{nI}(\hat{\theta}' - \theta)$. By (2.6) and (2.21), both $\sqrt{nI}(T_i' - \theta)$ have the same Edgeworth expansion up to $o(n^{-1/2})$. This is the fact that is expressed by saying FOE implies second order efficiency (SOE). An excellent recent treatment is available in Bhattacharya and Denker (1990).

Since FOE implies SOE, it is clear we have to go one step further to third order asymptotics to distinguish between FOE estimates.

6.5. Strongest third order efficiency theorems for Fisher consistent, first order efficiency estimates and curved exponentials. It is shown in Ghosh, Sinha and Subramanyam (1979) that matching bias up to $o(n^{-1})$ entails that the third and fourth cumulants for all FC, FOE estimates agree up to $o(n^{-1})$. Combining this with Theorem 3.1, part (iv), one gets

$$E\big(l\big(\sqrt{n}\,(T - \theta)\big)\theta\big) \geq E\big(l\big(\sqrt{n}\,(\hat{\theta}' - \theta)\big)\theta\big) + o(n^{-1}) \tag{6.14}$$

where $\hat{\theta}'$ matches bias of T up to $o(n^{-1})$, and $l(0) = 0$, $l(y) > l(x)$ if $y > x \geq 0$ or $y < x \leq 0$. *The function l need not be symmetric*, but one either requires valid Edgeworth expansions for $\sqrt{n}\,(T - \theta)$ and $\sqrt{n}\,(\hat{\theta} - \theta)$ or that l is smooth, and so on, so that one can apply (2.10c). Because of Fisher consistency, uniformity is not needed.

6.6. Uniformity to third order and third order superefficiency. Except when estimates T are FC, we require either valid Edgeworth expansions for $\sqrt{n}\,(T - \theta)$ *uniformly on compact θ-sets* as in Ghosh, Sinha and Wieand (1980) or uniformity to third order in a different sense as in Conditions 1 to 3 of Section 6.1 for TOE results to hold. Bickel, Götze and van Zwet (1985) do not need uniformity because they state their results in terms of Le Cam's local minimax criterion: see Section 1.2.

Without uniformity to third order can there be superefficiency to third order but not superefficiency to first order ? In other words, can there exist an estimate T such that (i) $\sqrt{n}\,(T - \theta)$ is A.N. $(0, 1/I(\theta))$ uniformly on compact θ sets (see Assumption A2 of Chapter 1), (ii) $\sqrt{n}\,(T - \theta)$ has valid Edgeworth expansion up to $o(n^{-1})$ for all θ with $\kappa_{11} = 0$ [i.e., T is unbiased up to $o(n^{-1})$] and (iii) under $\theta = 0$, T is third order better than $\hat{\theta}^*$ of the following paragraph? The answer is yes. We produce such an example below, with a minor modification of the Hodges example of superefficiency.

Start with TOE $\hat{\theta}^*$ of Chapter 3, the perturbation of $\hat{\theta}$ to remove bias up to $o(n^{-1})$. Assume it has a valid Edgeworth expansion uniformly on compact θ-sets. This requires application of a uniform version of Theorem 2.1, which is

available in Bhattacharya and Ghosh (1978). Now define

$$T = \hat{\theta}^* \qquad \text{if } |\hat{\theta}^*| > (\log n)/\sqrt{n}$$
$$= (1 - c/n)\hat{\theta}^* \quad \text{if } |\hat{\theta}^*| \le \log n/\sqrt{n},$$

where $c > 0$. Then $\sqrt{n}\,|(T - \hat{\theta}^*)| \le (c/n)|\sqrt{n}\,\hat{\theta}^*|$, which has probability tending to zero uniformly on compact θ-sets. Since $\sqrt{n}\,(\hat{\theta}^* - \theta)$ satisfies Assumption A2 of Chapter 1, it is now clear that so does $\sqrt{n}\,(T - \theta)$. On the other hand, under $\theta = 0$, $\sqrt{n}\,(T - \theta)$ has the same first, third and fourth cumulant as $\sqrt{n}\,(\hat{\theta}^* - \theta)$ up to $o(n^{-1})$ and κ_{22} for T is (κ_{22} of $\hat{\theta}^*$) $- c$, that is, smaller than κ_{22} *of* $\hat{\theta}^*$. Also under any $\theta \neq 0$, $P_\theta\{T \neq \hat{\theta}^*\} = o(n^{-1})$, and so by Lemma 2.1, $\sqrt{n}\,(T - \theta)$ has the same Edgeworth expansion as $\sqrt{n}\,(\hat{\theta}^* - \theta)$ up to $o(n^{-1})$.

6.7. Third order admissibility. Third order efficiency of $\hat{\theta}$ may be interpreted as a complete class theorem in decision theory. Given a FOE T_n, satisfying some additional regularity conditions, there exists a countinuously differentiable $c(\cdot)$ such that $(\hat{\theta} + c(\hat{\theta})/n)$ is better than T_n to the third order. In this context it is natural to ask whether the complete class is actually a minimal complete class, that is, whether each $\hat{\theta} + c(\hat{\theta})/n$ is admissible to third order or not. In this section we completely characterize third order admissibility, showing in the process the class of third order admissible estimates $\hat{\theta} + c(\hat{\theta})/n$ is a proper subset of the whole class and it is complete, so it is minimal complete. This section is based on Ghosh and Sinha (1981). In the following $E_\theta = E\{|\theta\}$. Consider $\hat{\theta} + c(\hat{\theta})/n$ and $\hat{\theta} + d(\hat{\theta})/n$, where $c(\cdot)$ and $d(\cdot)$ are continuously differentiable:

$$E_\theta\big(\hat{\theta} + d(\hat{\theta})/n - \theta\big)^2 - E_\theta\big(\hat{\theta} + c(\hat{\theta})/n\big)$$
$$= \{g^2(\theta) + 2g(\theta)b(\theta) + 2g'(\theta)/I(\theta)\}/n^2 + o(n^{-2}) \tag{6.15}$$

where $(b(\theta))/n$ is the "bias" of $\hat{\theta} + (c(\hat{\theta}))/n$ up to $o(n^{-1})$ and $g(\theta) = d(\theta) - c(\theta)$. To prove (6.15), note

$$E_\theta\left(\hat{\theta} + \frac{d(\hat{\theta})}{n} - \theta\right)^2$$
$$= E_\theta\left(\hat{\theta} + \frac{c(\hat{\theta})}{n} + \frac{g(\hat{\theta})}{n} - \theta\right)^2$$
$$= E_\theta\left\{\left(\hat{\theta} + \frac{c(\hat{\theta})}{n} - \theta\right) + \frac{g(\theta)}{n} + g'(\theta)\frac{(\hat{\theta} - \theta)}{n}\right.$$
$$\left. + \text{smaller order terms}\right\}^2$$

$$= E_\theta\left(\hat\theta + \frac{c(\hat\theta)}{n} - \theta\right)^2 + \frac{g^2(\theta)}{n^2} + 2g(\theta)\frac{b(\theta)}{n}$$
$$+ 2g'(\theta)\frac{1}{n^2 I(\theta)} + o(n^{-2}),$$

where we have used

$$E_\theta\left(\hat\theta + \frac{c(\hat\theta)}{n} - \theta\right)(\hat\theta - \theta) = E_\theta(\hat\theta - \theta)^2(1 - o(1))$$
$$= \frac{1}{nI(\theta)} + o(n^{-1}).$$

Say $\hat\theta + c(\hat\theta)/n$ is third order inadmissible (TOI) if $\exists$ continuously differentiable $d(\theta)$ such that

$$\{g^2(\theta) + 2g(\theta)b(\theta) + 2g'(\theta)/I(\theta)\} \le 0 \qquad \forall\,\theta \tag{6.16}$$

with at least one strict inequality. Otherwise, $\hat\theta + c(\hat\theta)/n$ is third order admissible (TOA). Let $b_0(\theta)/n$ be the bias of $\hat\theta$ up to $o(n^{-1})$. Assume $b_0(\theta)$ and $I(\theta)$ are continuous, and $I(\theta) > 0\ \forall\,\theta$.

THEOREM 6.1. (i) *$\hat\theta + c(\hat\theta)/n$ is TOA if and only if for some $-\infty < \theta_0 < \infty$,*

$$\int_{\theta_0}^{\infty} I(\theta)\exp\left\{-\int_{\theta_0}^{\theta} b(u)I(u)\,du\right\} d\theta = \infty \tag{6.17}$$

and

$$\int_{-\infty}^{\theta_0} I(\theta)\exp\left\{\int_{\theta}^{\theta_0} b(u)I(u)\,du\right\} d\theta = \infty. \tag{6.18}$$

(ii) *If $\hat\theta + c(\hat\theta)/n$ is TOI, then one can find $d(\theta)$ explicitly such that $\hat\theta + d(\hat\theta)/n$ is TOA and better than $\hat\theta + c(\hat\theta)/n$ up to $o(n^{-2})$.*

The condition for TOA suggests that the bias term $b(\theta)$ should be negative as $\theta \to \infty$ and positive as $\theta \to -\infty$, that is, the estimate should behave like a shrinker at least as far as the bias is concerned. Intuitively this seems a good thing.

6.8. Berkson's example revisited. Let $\tilde\theta$ denote the Rao–Blackwellized minimum logit χ^2 estimate introduced in Section 2.4. We take $\alpha \equiv \theta$. Let the bias terms of $\tilde\theta$ and $\hat\theta$ be denoted by $b(\cdot)$ and $b_0(\cdot)$. Then [see Ghosh and Sinha (1981), page 1337]

$$I = \sum \pi_i(1 - \pi_i), \qquad b_0 = \sum \pi_i(1 - \pi_i)(2\pi_i - 1)/2I^2, \tag{6.19}$$

$$b = \sum \pi_i(1 - \pi_i)(2\pi_i - 1)/I^2 - \sum(2\pi_i - 1)/2I. \tag{6.20}$$

Since $I \sim \text{const.}\exp(-|\theta|)$ as $\theta \to \pm\infty$, $b_0 I \to \pm 1/2$ as $\theta \to \pm\infty$, it follows from Theorem 6.1 that $\hat\theta$ is inadmissible. Similarly, since $bI \to \mp(k-2)/2$ as $\theta \to \pm\infty$, $\tilde\theta$ is TOI (TOA) if $k < 4$ ($k > 4$). By similar analysis, one can show $\tilde\theta$ is TOA for $k = 4$.

Berkson (1980) and Amemiya (1980) found from numerical calculations of the mean squares of the minimum logit χ^2 estimate and $\hat{\theta}$ obtained by the delta method up to $o(n^{-2})$ that the mean square for $\tilde{\theta}$ is smaller in cases studied. They wondered if this was true for all θ.

Since $\tilde{\theta}$ is better to third order than Berkson's minimum logit χ^2, one can ask the same question in relation to $\tilde{\theta}$ and $\hat{\theta}$. The comparison between $\tilde{\theta}$ and $\hat{\theta}$ is analytically much easier than that between the original estimate and $\hat{\theta}$.

We have only to compare

$$A(\theta) = \{b^2(\theta) + 2b'(\theta)/I(\theta)\}I^4(\theta) \tag{6.21}$$

and

$$B(\theta) = \{b_0^2(\theta) + 2b_0'(\theta)/I(\theta)\}I^4(\theta). \tag{6.22}$$

Note that

$$\begin{aligned} A(\theta) &= \Big\{\sum(\pi_i - 1)\Big\}^2(6 + k^2/2 - 4k)/2\{1 + o(1)\} \quad \text{as } \theta \to +\infty \\ &= \Big\{\sum \pi_i\Big\}^2(6 + k^2/2 - 4k)/2\{1 + o(1)\} \quad \text{as } \theta \to -\infty \end{aligned} \tag{6.23}$$

and

$$\begin{aligned} B(\theta) &= \Big\{\sum(\pi_i - 1)\Big\}^2(5/4)\{1 + o(1)\} \quad \text{as } \theta \to +\infty \\ &= \Big\{\sum \pi_i\Big\}^2(5/4)\{1 + o(1)\} \quad \text{as } \theta \to -\infty. \end{aligned} \tag{6.24}$$

This shows that if k (the number of dose levels) is greater than or equal to 8, there exists an interval (θ_1, θ_2) such that for $\theta \notin (\theta_1, \theta_2)$, $\hat{\theta}$ is better than $\tilde{\theta}$, while for $k \leq 7$, $\tilde{\theta}$ is better outside a certain interval. In paticular, the answer to Berkson's question is no. In this connection also see Kariya, Sinha and Subramanyam (1984) and Davis (1984, 1985).

As Ghosh (1980) has indicated, it was the controversy surrounding this example which made him take up higher order efficiency to resolve these questions. What have we learnt? As in all real life stories, the lessons are mixed. While Berkson's belief in the global superiority of his estimate over $\hat{\theta}$ [up to $o(n^{-2})$] is wrong, there is something to be said for using its Rao–Blackwellized version in preference to $\hat{\theta}$ because $\hat{\theta}$ is always TOI whereas $\tilde{\theta}$ is TOA for $k > 4$. Higher order theory also suggests how we can improve $\hat{\theta}$ always, even though this cannot be done by using Berkson's estimate.

6.9. Third order minimaxity. After third order efficiency and admissibility, it is natural to think of third order minimaxity. In fact, this question was raised in Ghosh and Subramanyam (1974). We follow Ghosh and Mukerjee (1993c).

Let $\mathscr{C}$ be the class of estimates of the form

$$T = \hat{\theta} + c(\hat{\theta})/n,$$

$c(\cdot)$ continuously differentiable. Then

$$nI(\theta)E_\theta(T-\theta)^2 = 1 + a_T(\theta)/n + o(n^{-1}). \tag{6.25}$$

An explicit expression for $a_T(\theta)$ will be provided shortly.

DEFINITION 6.1. An estimate T_0 in $\mathscr{C}$ is third order minimax (TOM) if

$$\sup_\theta a_{T_0}(\theta) \le \sup_\theta a_T(\theta) \qquad \forall\, T \in \mathscr{C}.$$

In general, minimaxity for all n does not imply TOM, nor does third order minimaxity imply minimaxity for sufficiently large n. A similar lack of relation between admissibility and asymptotic admissibility (in the sense of first order admissibility) was noted by Hájek (1972). This remains true if we replace first order admissibility by third order admissibility.

A sufficient condition for a minimax estimate to be approximable [in the sense of (5.20)] by an estimate in $\mathscr{C}$ is that for all $n > n_0$, there exists a fixed least favorable prior π which satisfies Johnson's (1970) conditions for all θ. If further a TOM exists and for both the TOM and the minimax estimates the risk expansions are uniform over Θ, then both expansions coincide up to $o(n^{-1})$. In location and scale problems, both estimates turn out to be equivariant and so uniformity holds. The assumption about a least favorable prior also holds. It is assumed the minimax estimate is Bayes with respect to the least favorable prior.

We now compute a_T explicitly. Take $T = \hat\theta + c(\hat\theta)/n$. Define $\lambda(\theta)$ by

$$c(\theta) = \lambda(\theta) - I^{-2}\left(\tfrac{1}{2}\bar L_{001} + \bar L_{11}\right),$$

where

$$\bar L_{ijkl} = E_\theta\left[\left(\frac{d\log p(x_1|\theta)}{d\theta}\right)^i\left(\frac{d^2\log p}{d\theta^2}\right)^j\left(\frac{d^3\log p}{d\theta^3}\right)^k\left(\frac{d^4\log p}{d\theta^4}\right)^l\right], \tag{6.26}$$

$$\bar L_{ijk} = \bar L_{ijko}, \qquad \bar L_{ij} = \bar L_{ijo}, \qquad \bar L_i = \bar L_{io}. \tag{6.27}$$

Then

$$nE_\theta\left[(T-\theta)^2 I(\theta)\right] = 1 + \left\{I\{\lambda(\theta)\}^2 + 2\lambda'(\theta) + \psi(\theta)\right\}/n, \tag{6.28}$$

where

$$\psi(\theta) = I^{-2}\bar L_{02} + I^{-3}\left(\tfrac{3}{2}\bar L_{001}^2 + \bar L_{11}^2 + 5\bar L_{11}\bar L_{001} - \bar L_{001}\bar L_3\right) - 1. \tag{6.29}$$

The above provides an expression for $a_T(\theta)$.

EXAMPLE 6.1 (Location family). Let $p(x_1|\theta) = g(x_1 - \theta)$. Consider

$$\begin{aligned} T_0 &= \hat\theta - n^{-1}I^{-2}\left(\tfrac{1}{2}\bar L_{001} + \bar L_{11}\right) \\ &= \hat\theta - n^{-1}c, \end{aligned} \tag{6.30}$$

where c is free of θ. (We assume tacitly $\bar{L}_{001}$, etc. are finite.) If we have a symmetric density like $N(\theta, 1)$ or a Cauchy with a location parameter θ, then $c = 0$ and $T_0 = \hat{\theta}$. We may also note that T_0 is identical with the Bayes estimate B_n' in Chapter 5 if one takes π to be the improper prior which is uniform over the whole real line.

We assume $\hat{\theta}$ is equivariant in the sense of

$$\hat{\theta}(x_1 + a, \ldots, x_n + a) = \hat{\theta}(x_1, \ldots, x_n) + a.$$

Note $\psi(\theta)$ is free of θ and

$$a_{T_0}(\theta) = \psi_0 \tag{6.30a}$$

for all θ.

We will show T_0 is TOM.

Given the expressions for a_{T_0} and a_T [see (6.29) and (6.30a)], it is enough to show

$$\sup_{\theta} \left[I\{\lambda(\theta)\}^2 + 2\lambda'(\theta) \right] \geq 0 \tag{6.31}$$

for all continuously differentiable λ.

We introduce an auxiliary problem in which $N(\theta, \sigma^2)$ with $\sigma^2 = 1/I$ is $p(x_1|\theta)$. Consider $T_1 \equiv \bar{X}$, which is TOA in the auxiliary problem by Theorem 6.1. Compare T_1 with $T_\lambda \equiv \bar{x} + \lambda(\bar{x})/n$. Note

$$nE\{(\bar{x} - \theta)^2|\theta\} = \frac{1}{I}, \tag{6.32}$$

$$nE\{(T_\lambda - \theta)^2|\theta\} = \frac{1}{I} + \frac{1}{I}\frac{\{\lambda^2 I + 2\lambda'\}}{n} + o(n^{-1}). \tag{6.33}$$

If (6.31) were false, T_λ would be better to third order than $\bar{x}$, making $\bar{x}$ TOI in the auxiliary problem!

EXAMPLE 6.2 (Scale Problem). Let $p(x|\theta) = \theta^{-1}g(x|\theta)$, $\theta > 0$. Here $I = \rho/\theta^2$, $\bar{L}_{11}L_{02} = l_{02}/\theta^4$, $\bar{L}_{11}L_{11} = l_{11}/\theta^3$ and $\bar{L}_{11}\ L_{001} = l_{001}/\theta^3$ for some constants, $\rho, l_{02}, l_{11}, l_{001}$ (we tacitly assume all the expectations involved are finite). It follows that $\psi(\theta) = \psi_0$, a constant.

Let

$$T_0 = \hat{\theta}(1 - c/n), \tag{6.34}$$

where $c = 1/\rho + (1/\rho^2)(\frac{1}{2}l_{001} + l_{11})$.

We show T_0 is TOM. Proceed as before, and to check, it suffices to prove only $\lambda \equiv 0$ solves

$$\sup_{\theta} \left[\frac{\rho}{\theta^2}\{\lambda(\theta)\}^2 + 2\lambda'(\theta) \right] > -1/\rho. \tag{6.35}$$

Introduce an auxiliary gamma scale family

$$p(x_1|\theta) = \text{const.} \exp\{-x|\theta\} x^{\rho-1}. \tag{6.36}$$

Use Theorem 6.1 again to check that in the auxiliary problem, $c_n\hat{\theta}$ is TOA, where $c_n\hat{\theta}$ is the best estimate for θ in the auxiliary problem with respect to squared error loss, among estimates of the form $\hat{\theta}(1 + d/n)$, where $d > 0$ is a constant and $\hat{\theta}$ is the mle of θ in the auxiliary problem. Now compare $c_n\hat{\theta}$ with $\hat{\theta} + \lambda(\hat{\theta})/n$. A nonzero solution λ of (6.35) would show $c_n\hat{\theta}$ is TOI, which would be a contradiction.

We choose a third example which has some interest of its own and does not belong to a location or scale family. It will also bring out some of the difficulties with these estimates.

EXAMPLE 6.3. $X_i = (U_i, V_i)$, $i = 1, 2, \ldots, n$, are bivariate normal with zero means, standard deviation (s.d.) equal to 1, and correlation coefficient $\rho \equiv \theta$. This example was introduced by Stewart to show naive substitution of known values of parameters for their estimates in a statistic may lead to an inferior estimate. In the problem of the bivariate normal, if the means and standard deviations were all unknown, then the sample correlation coefficient r is the natural estimate and it is also then the mle. Under the present setup, one is tempted to replace $\bar{X}, \bar{Y}$ by 0 and S_X, S_Y by unity in

$$r = \frac{(1/n)\sum(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

where $S_X^2 = (1/n)\Sigma(X_i - \bar{X})^2$ and $S_Y^2 = (1/n)\Sigma(Y_i - \bar{Y})^2$. The resulting estimate

$$T = \frac{1}{n}\sum X_i Y_i$$

is worse than r in the first order for all values of θ. The right way to use the information is to choose the consistent solution $\hat{\theta}$ of the likelihood equation which is a cubic in θ under the present assumptions. The mle $\hat{\theta}$ is better in the first order than r, as expected.

We confine ourselves, as before, to estimates $\hat{\theta} + c(\hat{\theta})/n$. Here

$$\psi(\theta) = 4(1 - \theta^2)^2/(1 + \theta^2)^3. \tag{6.37}$$

For convenience we exclude $\theta = 0$ from the parameter space (making the problem somewhat artificial) and restrict ourselves to a subclass of $\mathscr{C}$. We permit only the following λ_T's.

Either (a) $I(\theta)\{\lambda_T^2(\theta)\} + 2\lambda_T'(\theta) \to \infty$ as $\theta \uparrow 1$ or $\downarrow -1$ or (b) $\lambda_T(\theta)/(1 - \theta)$ is bounded as $\theta \uparrow 1$ or (c) $\lambda_T(\theta)/(1 + \theta)$ is bounded as $\theta \downarrow -1$.

For example, if $\lambda_T(\theta) = p_1(\theta)/p_2(\theta)$, where $p_1(\cdot)$'s are polynomials in θ and $p_2(\theta) \neq 0$ for $\theta \in (-1, 0) \cup (0, 1)$, then λ_T belongs to our subclass. One can show

$$\inf_{\lambda \in C_1} \sup_{\theta} a_T(\theta) = \sup_{\theta} a_{T_0}(\theta) = -2,$$

where

$$T_0 = \hat{\theta} + \frac{c(\hat{\theta})}{n},$$

$$c(\theta) = \frac{1 - \theta^2}{\theta}.$$

The unpleasant fact that $c(\cdot)$ becomes unbounded as $\theta \to 0$ indicates the care with which such estimates need to be used in practice. The estimate T_0 will make sense and do better than $\hat{\theta}$ in the third order minimax sense if the parameter space were $\Theta = (-1, -\delta] \cup [\delta, 1)$ and n is sufficiently large to ensure $|\theta + c(\theta)/n|$ is bounded by 1, in Θ. Such a parameter space may be plausible if one knows some dependence must be present in the population.

6.10. Where do we go from here?

1. Given that FOE $\Rightarrow$ SOE, it is natural to conjecture TOE $\Rightarrow$ fourth order efficiency. The proof of that must be very messy. One may also ask if anything like TOE holds for $\hat{\theta}$ when we go to the fifth order. Ghosh and Sinha (1982) show with a counterexample that this is not possible.
2. The main advantage of these third order ideas (TOE, TOA, TOM) is, as we see it, that one has a set of optimality results which help one choose, within the frequentist paradigm, from the classical estimates for parametric families. Moreover given a pair of FOE estimates T_1, T_2 one can go over to $\hat{\theta} + c(\hat{\theta})/n$ and $\hat{\theta} + d(\hat{\theta})/n$ which are better than T_1, T_2, to third order with respect to all natural loss functions, and then compare the risks of $\hat{\theta} + c(\hat{\theta})/n$ and $\hat{\theta} + d(\hat{\theta})/n$. That is almost always much easier than comparing the risk of the original estimates. This is how Berkson's problem was solved. The most striking fact is that the same perturbation $\hat{\theta} + c(\hat{\theta})/n$ works for all natural, possibly nonconvex, loss functions, as if we are even better off than having a complete sufficient statistic. One of the reasons why this theory was developed is that classical frequentist statistics seems to oscillate between an asymptotic theory that holds for very large samples and an exact theory confined to a very small collection of parametric families like the normal. Higher order optimality provides approximations for moderate samples, presumably to the order of 10 or 15, and is applicable to quite general parametric families, satisfying regularity conditions. In particular, this can, in principle, open up the way to parametric studies of robustness, in which one would enlarge or embed a given oversimplified parametric family like the normal or exponential in one with more parameters and study the third order properties of proposed estimates.
3. There has been interesting new work by Yoshida (1992) on the application of Malliavin calculus to higher order asymptotics for diffusions. Of course, though we have worked within the self-imposed framework of i.i.d. r.v.'s, most of the ideas go over to dependent cases or for stochastic processes, but explicit calculations become more difficult.

4. The concepts of third order admissibility and minimaxity need more clarification. Can they lead to pathologies? For example, in the multiparameter case, we do not know of any estimate which is TOA. [In DasGupta and Ghosh (1983) several estimates are shown to be TOI in the multiparameter setting.] If all estimates in the multiparametric case turn out to be TOI, clearly the concept needs reexamination. Since the spherically symmetric case is often like the one-dimensional case, at least such problems ought to receive attention. The technical problem with the inequality defining third order admissibility is that the absence of a square of (g') in the defining inequality (6.16) permits g's with very large values of $|g'|$, that is g's with unacceptable oscillations. In a sense, it is also brought out in DasGupta and Ghosh (1983), where it is noted that the calculus of variation problem associated with the minimization of the integral of the left-hand side of (6.16) with respect to $\pi(\theta)$ leads to degenerate Euler equations. From a practical point of view as well as that of the validity of the asymptotics, one ought to put restrictions on oscillations and growth of g, but no natural way of doing this is clear yet. To some extent, the same problem occurs with third order minimaxity; see Example 6.3.
5. There is a theory of higher order inference for testing also, but it is more messy. For one-sided alternatives, see Pfanzagl (1979) or Akahira and Takeuchi (1981). For two-sided alternatives, most of the work has been done by Tapas Chandra and Rahul Mukerjee and their co-authors. Most of this work has provided complicated but fairly directly usable expressions for power against contiguous alternatives for a very large class of tests which includes Rao's, Wald's and the likelihood ratio tests. They have also elucidated a version of a conjecture of Rao. For a review, see Ghosh (1991).