PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Nuisance Parameters, Mixture Models, and the Efficiency of Partial Likelihood Estimators

B. G. Lindsay

Phil. Trans. R. Soc. Lond. A 1980 **296**, 639-662
doi: 10.1098/rsta.1980.0197

NUISANCE PARAMETERS, MIXTURE MODELS, AND THE EFFICIENCY OF PARTIAL LIKELIHOOD ESTIMATORS†

By B. G. LINDSAY

Department of Statistics, Pennsylvania State University, PA 16802, *U.S.A.*

(*Communicated by D. R. Cox, F.R.S. – Received* 12 *June* 1979)

CONTENTS

† Prepared while a visitor at Imperial College, London.

Vol. 296. A 1427. 54 [Published 18 April 1980

This paper establishes lower bounds for estimation in parametric statistical models in which one wishes to estimate a real-valued parameter of interest in the presence of nuisance parameters which are accruing in number in direct proportion to the number of independent observations. The formal setting requires that the nuisance parameters be independent observations from an unknown distribution. In this setting an information measure analogous to the Fisher information is derived. It is then used to generate lower bounds for the variance of unbiased estimators and also for the asymptotic variance of consistent asymptotically normal estimators. Under certain conditions, consistent asymptotically normal estimators can be generated by maximizing factors of the complete likelihood, even though the maximum likelihood estimator is inconsistent. These estimators can be fully efficient in the sense of meeting the lower bounds despite their apparent wasteful use of the likelihood, as is demonstrated, in several important examples, by the use of a natural sufficient condition.

1. Introduction

One of the most important problems in statistics concerns estimation in the presence of nuisance parameters. Consider an experiment in which there are two treatments to be compared for efficacy in a target population. If that population is heterogeneous, then there will be other variables, both measured and not, that have a bearing on the outcome measure of the experiment. The statistical analysis of the experiment starts with a plausible model. It is common practice for the difference in treatment effects to be represented in the model by a single real-valued parameter and for the heterogeneity of the sampled population to be accounted for by other, nuisance, parameters. Within the model the user of statistics then seeks an estimator of the treatment difference parameter that is best for that model under some reasonable optimality criterion, thereby augmenting the persuasiveness of his results.

The discovery that the maximum likelihood estimator is, under one potent optimality criterion,

a best possible estimator is a profound achievement of the classical theory of statistics. Yet a substantial problem can arise because the classical theory does not incorporate nuisance parameters in a substantial manner. In fact, it will be demonstrated that in some nuisance parameter models the maximum likelihood estimator is very misleading. To remedy this situation this paper establishes for the following large class of nuisance parameter models an optimality theory analogous to the classical one and then demonstrates an optimal estimator for an important subclass of these models.

Suppose that $X_1, X_2, \ldots, X_n$ is a set of independent random variables with common range space $(\mathscr{X}, \mathscr{A})$ such that for each i X_i has the parametric density $f(\cdot\,; \theta, \phi_i)$ with respect to a σ-finite measure ν on $(\mathscr{X}, \mathscr{A})$. The pair (θ, ϕ_i) is assumed to be an element of a cross-product parameter space $\Lambda = \Theta\Phi$, where Θ is an open set of $\mathbb{R}$, the real numbers, and $(\Phi, \mathscr{B})$ is a measurable space. Denote the measure on $(\mathscr{X}^n, \mathscr{A}^n)$ induced by $(X_1, \ldots, X_n)$ as $\prod_{i=1}^{n} (\theta, \phi_i)$. This is a class of models characterized by the number of parameters increasing to infinity with the sample size n. (Hereafter, when limits are not expressed on Σ and Π, they will be understood to be index i from 1 to n.)

The general problem of this paper is that of estimating the parameter θ, called the *parameter of interest*, in the presence of the unknown ϕ-parameters, called the *nuisance parameters*. Neyman & Scott (1948) established that the maximum likelihood estimator (m.l.e.) of the parameter θ (which they called the structural parameter) is liable to be inconsistent as $n \to \infty$. This is demonstrated by their example (2), which follows.

For each i, $i = 1, \ldots, n$ let $X_i = (X_{i1}, \ldots, X_{iJ})$ be a J-vector of independent and identically distributed (i.i.d.) normal random variables, with mean ϕ_i and variance θ. The maximum likelihood estimate of θ,

$$\hat{\theta}_n = \sum_{i=1}^{n} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2/(nJ)$$

is not consistent as $n \to \infty$, whereas, in contrast, the bias-corrected m.l.e.,

$$S_n^2 = J(J-1)^{-1} \hat{\theta}_n, \tag{1.1.1}$$

is consistent as either $n \to \infty$ or $J \to \infty$. This model will be an illustrative example for many of the ideas in this paper and will be referred to as the one way analysis of variance (an.o.va.) model.

Neyman & Scott (1948) also demonstrated that, even in those nuisance parameter models in which the maximum likelihood estimator is consistent, it can fail to have minimal asymptotic variance in the class of asymptotical normal estimates. (See their example (1); another example is in § 9.6 of the present paper.)

These results, of course, thereafter weakened the credibility of the m.l.e. in models with many nuisance parameters. On the other hand, in models of the form $(\theta, \phi)^n$, where the number of nuisance parameters does not increase with n, an elegant structure has been created that justifies the use of maximum likelihood estimation in large samples by its minimal asymptotic variance; this is discussed further in § 5.1. This paper is a preliminary search for the corresponding structure in the $\Pi(\theta, \phi_i)$ model.

For example, is S_n^2 an optimal estimator as $n \to \infty$ in the one way an.o.va. model? It is consistent for θ and asymptotically normal (henceforth denoted c.a.n.), with asymptotic variance $2\theta^2/(J-1)$. Andersen (1970) derived a lower bound for the asymptotic variance of c.a.n. estimators in the model $\Pi(\theta, \phi_i)$ which, when applied to the Neyman–Scott example, gives the lower bound

of $2\theta^2/J$. This suggests an efficiency for S_n^2 of $(J-1)/J$. The startling implication is that, although S_n^2 is the standard estimator of variance for the one way an.o.va. model, there may be substantially better estimators asymptotically.

In the following pages, S_n^2 will be shown to be an asymptotically efficient estimator based on the following arguments. First, it will be argued that, for efficiency considerations, the model $\Pi(\theta, \phi_i)$ can sensibly be transformed into the mixture model of § 2. In § 3, an information measure for the mixture model is created, use of which is justified by the lower bound theorems of § 4 and 5. In § 6 a class of estimators that are based on likelihood factorizations and are called partial maximum likelihood estimators (p.m.l.es) are introduced; S_n^2 falls into this class. After a discussion of the general c.a.n. properties of this class in § 7, a simple criterion for their full mixture model efficiency is presented in § 8. The criterion is met by S_n^2. In § 9, four other examples are discussed. One example with a history of controversy appears in § 9.5, the pairs of Bernoulli variables with common log odds ratios. Another, in § 9.6, demonstrates a p.m.l.e. that is not fully mixture model efficient.

2. The mixture model

2.1. *The model*

Suppose that $\{\phi_1, \phi_2, \ldots\}$ is a sequence of i.i.d. random variables from an unknown probability measure Q on $(\Phi, \mathscr{B})$. Conditioned on the realized sequence $\{\phi_1, \phi_2, \ldots, \phi_n\}$, the random variables $X_1, \ldots, X_n$ are independent but are not identically distributed (unless $\phi_1 = \phi_2 = \ldots = \phi_n$). They have the distribution $\Pi(\theta, \phi_i)$. Viewed unconditionally, however, they are i.i.d. random variables from the *mixture density*,

$$\int f(x; \theta, \phi)\, \mathrm{d}Q(\phi) := f(x; \theta, Q).$$

(The symbol $:=$ here means 'is defined to be equal to'.) The corresponding measure on $(\mathscr{X}^n, \mathscr{A}^n)$ will be denoted by $(\theta, Q)^n$. Let $\mathscr{Q}$ be the family of all probability measures on $(\Phi, \mathscr{B})$.

Henceforth it is assumed that $\mathscr{B}$, the σ-algebra of Φ, contains all one-element sets $\{\phi\}$. Let $\delta(\phi)$ be the probability measure that puts mass one at $\{\phi\}$. If

$$\Lambda^* := \{(\theta, Q): \theta \in \theta, Q \in \mathscr{Q}\}$$

(hereafter called the *mixture parameter space*), then the parameter space Λ can be embedded in Λ^* by $(\theta, \phi) \to (\theta, \delta(\phi))$. The parameter point $(\theta, \delta(\phi))$ will be called a *fixed point*, there being no randomness in the choice of ϕ.

2.2. *Rationales for mixture model analysis*

The mixture model $(\theta, Q)^n$ of § 2.1 may indeed be the model initially postulated, in which case no rationale for interest in mixture model efficiency is needed. On the other hand, there may be some other structure on the nuisance parameters. For example, in the one way an.o.va. model the treatment groups may have been fixed in advance. For this problem, then, an i.i.d. specification for the ϕ's would seem quite incorrect. However, there are several reasons for the mixture model remaining a more relevant place than $\Pi(\theta, \phi_i)$ to measure information.

The first reason comes from the perspective of a fixed sample size. The information measure created for the mixture space Λ^* in § 3 is used in § 4 to give a lower bound for the variance for

unbiased estimates in $\Pi(\theta, \phi_i)$ for quite arbitrary sequences $\{\phi_1, \ldots, \phi_n\}$. It is superior to the Cramér–Rao lower bound, itself generated by the fixed sequence point of view.

Another reason comes from an asymptotic perspective. Consider the one-way an.o.va. with n fixed treatment groups. For an asymptotic approximation to the finite sample problem that is faithful to the idea that there are many nuisance parameters relative to observations, one needs an asymptotic generating scheme for the remainder $\{\phi_{n+1}, \ldots\}$ of the null parameter. For that asymptotic model to be a reasonable approximation to the finite sample model, the tail of the sequence must be similar in character to the initial part. If the treatment means $\{\phi_1, \phi_2, \ldots, \phi_n\}$ do not have an inherent pattern, then to assume that they are i.i.d. observations may be a reasonable way to create an asymptotic approximation.

If the null ϕ-sequence has some non-random pattern in its asymptotic generation, then it is possible (see Lindsay 1978) to create lower bounds in $\Pi(\theta, \phi_i)$ that are again based on the mixture model information measure. However, the non-i.i.d. structure of the observations makes the results immensely more complicated and, it is now argued, those results can be misleading.

For example, suppose X_i is a normal random variable with mean θ and variance ϕ_i. Then if $\{a_1, a_2, \ldots\}$ is a sequence of positive constants, the estimator

$$T_n = \Sigma a_i X_i / \Sigma a_i \tag{2.2.1}$$

is c.a.n. for θ when $a_i = \phi_i^{-1}$, with minimal possible asymptotic variance in $\Pi(\theta, \phi_i)$. This estimator does quite well for one particular null sequence, at the price of poor performance along permuted versions of it.

The point is that using realized sequences for null points for lower bounds introduces asymmetry into the problem. That is, $(\theta, Q)^n$ provides a permutation symmetric distribution for $(X_1, \ldots, X_n)$, whereas $\Pi(\theta, \phi_i)$ does not. One result of this is that squared error loss in $(\theta, Q)^n$ penalizes estimators that are not permutation symmetric functions of $(X_1, X_2, \ldots, X_n)$, but such estimators may be optimal in $\Pi(\theta, \phi_i)$ along some null sequences. However, such estimators can only be *locally optimal*, as their average mean square error over all permuted versions of the null sequence cannot be less than that of a permutation symmetric estimator.

The asymptotic variance of a c.a.n. estimator is not necessarily equal to its asymptotic mean square error (although the former is a lower bound to the latter); indeed, the mean square error can be infinite, but the asymptotic variance not. However, Chernoff (1956) noted that if $n^{\frac{1}{2}}(T_n - \theta)$ has a limiting distribution with second moment σ^2, then

$$\lim_{k \to \infty} \lim_{n \to \infty} n E[\min\{(T_n - \theta)^2, k^2/n\}] = \sigma^2.$$

Since, for the mixture model, the expectation E is a mean over sequences, the following insights can be gained about c.a.n. estimators.

Suppose that one evaluates an estimator by seeing if it has minimal asymptotic variance in $(\theta, Q)^n$ among a class of c.a.n. estimators. If it is not optimal there, then one can be sure that in $\Pi(\theta, \phi_i)$ there are many sequences along which it is suboptimal, and there may be estimators uniformly better there. On the other hand, if it is optimal in $(\theta, Q)^n$, it still may not be optimal along all realized sequences (being beaten by estimators like (2.2.1)), but any estimator that is superior to its along some sequences *must* be inferior along others. Thus, a conservative procedure, if there is no *a priori* information about the sequence, is to use a permutation symmetric estimator optimal in $(\theta, Q)^n$.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Another reason for interest in Λ^* information is related to robustness. Suppose that one starts not with a nuisance parameter model but with an i.i.d. model of the form $(\theta, \phi)^n$. The mixture model $(\theta, Q)^n$ is then a means of generating a family of alternatives arbitrarily close to the original model. Moreover, θ is, typically, a well defined parameter in these models. These models are typified by having a greater variability in the sample than is generated by the model $(\theta, \phi)^n$: the variability in X is compounded by variability in ϕ. An example of this is the normal mixture, where X is a normal random variable with mean θ and variance ϕ. The normal mixtures family includes the t distributions and the Cauchy distributions.

Based on this argument, it is reasonable to ask how well an estimator can perform at a null $(\theta, \phi)^n$ and still be well behaved for nearby distributions in $(\theta, Q)^n$. This is exactly the question answered by the lower bound theorem of §§4 and 5 when they are evaluated at the null points $(\theta, \delta(\phi))$.

3. Minimal Fisher's information

3.1. *Definitions and properties*

For much of the ensuing discussion attention will be restricted to a null point $\lambda_0 = (\theta_0, Q_0) \in \Lambda^*$. Hereafter E_0, P_0, and var_0 will denote the values of these operators under the null distribution.

Let $\epsilon > 0$ and let $Q_\cdot$ be a function from $[0, \epsilon]$ to $\mathcal{Q}$ of the form $\tau \to Q_\tau$. For $\alpha = +1$ or -1, let $\lambda_\tau = (\theta_0 + \alpha\tau, Q_\tau)$. A *local family of alternatives* to the null λ_0 is a set $\mathcal{F}(\alpha, Q_\cdot) := \{\lambda_\tau : \tau \in [0, \epsilon]$ and $\lambda_\tau \in \Lambda^*\}$ with the property that, if L_τ is the *likelihood ratio* $f(X; \lambda_\tau)/f(X; \lambda_0)$, then L_τ is well defined, $E_0[L_\tau] = 1$, and

$$\mathrm{var}_0[L_\tau] \to 0 \quad \text{as} \quad \tau \to 0^+. \tag{3.1.1}$$

If L_τ satisfies, in addition, the Cramér type regularity conditions cited in appendix A, then $\mathcal{F}(\alpha, Q_\cdot)$ will be called a *regular* local family of alternatives to null λ_0. Under the smoothness conditions there exists a (left-hand) first derivative of L_τ at $\tau = 0^+$ denoted L_0', and a Fisher's information

$$\mathcal{I}(\lambda_0 \,|\alpha, Q_\cdot) = E_0[L_0']^2,$$

which is just the Fisher's information at $\tau = 0^+$ of the smooth parametric family $\mathcal{F}(\alpha, Q_\cdot)$.

The *minimal Fisher's information from above* is then defined to be

$$\mathcal{I}^+(\theta_0 \,|Q_0) = \inf\{\mathcal{I}(\lambda_0 \,| +1, Q_\cdot) : \mathcal{F}(+1, Q_\cdot) \text{ regular}\}. \tag{3.1.2}$$

For $\mathcal{I}^-$, the *minimal Fisher's information from below*, replace $\alpha = +1$ by $\alpha = -1$ in (3.1.2). The *minimal Fisher's information* is $\mathcal{I} = \min\{\mathcal{I}^+, \mathcal{I}^-\}$.

An antecedent to the above notion of computing information by considering the most difficult (least informative) one-dimensional subfamilies through a null point can be found in Stein (1956). One peculiarity of the formulation presented here is that the null point λ_0 is a boundary point of the local subfamily. This is done because in Λ^* it is not generally true that $\mathcal{I}^+ = \mathcal{I}^-$. This is shown in appendix B with use of the one way an.o.va. example. Thus, this definition of information yields more powerful results than requiring that the null point λ_0 be in the interior of the one-dimensional subfamily $\{\lambda_\tau\}$. The essential point here is that, even though the one-dimensional subfamily $\mathcal{F}(\alpha, Q_\cdot)$ can sometimes be extended analytically to $\tau < 0$, the densities so generated may no longer come from Λ^*.

Often, lower bound theorems require that the null point be in the interior of the parameter space. This requirement is generally not essential, as will be seen in the lower bounds of §§4 and 5.

3.2. *Relation with Fisher's information*

Suppose that Φ is an open subset of $\mathbb{R}$. It is clear that a minimal Fisher's information could be defined as well for $\Lambda = \Theta\Phi$ by considering one-dimensional subfamilies of Λ. Further, the embedding property required in §2.2 ensures that the minimal Fisher's information in Λ at (θ_0, ϕ_0), here denoted $\mathcal{J}_0(\theta_0|\phi_0)$, is larger than the Λ^* information at $(\theta_0, \delta(\phi_0))$,

$$\mathcal{J}_0(\theta_0|\phi_0) \geqslant \mathcal{J}(\theta_0|\delta(\phi_0)), \tag{3.2.1}$$

since the one-dimensional subfamilies of Λ can be embedded in Λ^*.

What is $\mathcal{J}_0$? Suppose that a partitioned Fisher's information matrix,

$$\mathcal{J} = \begin{pmatrix} \mathcal{J}_{\theta\theta} & \mathcal{J}_{\theta\phi} \\ \mathcal{J}_{\phi\theta} & \mathcal{J}_{\phi\phi} \end{pmatrix},$$

exists for density $f(x; \theta, \phi)$. The multivariate Cramér–Rao lower bound for unbiased estimates of the differentiable function $h(\theta)$ is

$$\operatorname{var}(T) \geqslant [h'(\theta)\ 0] \begin{pmatrix} \mathcal{J}_{\theta\theta} & \mathcal{J}_{\theta\phi} \\ \mathcal{J}_{\phi\theta} & \mathcal{J}_{\phi\phi} \end{pmatrix}^{-1} [h'(\theta)\ 0]^t = (h'(\theta))^2 [\mathcal{J}_{\theta\theta} - \mathcal{J}_{\theta\phi}\mathcal{J}_{\phi\phi}^{-1}\mathcal{J}_{\phi\theta}]^{-1}.$$

This suggests, by analogy with the one-dimensional Cramér–Rao bound, that

$$\mathcal{J}_0(\theta_0|\phi_0) = \mathcal{J}_{\theta\theta} - \mathcal{J}_{\theta\phi}\mathcal{J}_{\phi\phi}^{-1}\mathcal{J}_{\phi\theta}. \tag{3.2.2}$$

This is true, this minimal information, in fact, being generated by the one parameter family $\lambda_\tau = (\theta_0 + \tau, \phi_0 - \mathcal{J}_{\phi\phi}^{-1}\mathcal{J}_{\phi\theta}\cdot\tau)$. (See Stein 1956.)

4. Lower bounds for unbiased estimation

4.1. *Local versus global properties*

Suppose that $f(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}$, is a family of densities satisfying the usual regularity conditions, with Fisher's information $\mathcal{J}(\theta)$. Pick a null point θ_0. Then the Cramér–Rao lower bound for the variance of unbiased estimators T of θ,

$$\operatorname{var}_0(T) \geqslant \mathcal{J}^{-1}(\theta_0),$$

can be derived with use of only the weaker requirement on T that it be unbiased for θ on some arbitrarily small interval containing θ_0 ('locally unbiased'). As such, there exist superior lower bounds for globally unbiased estimators (Barankin 1949). It is important to note, though, that it is the information $\mathcal{J}$, derived from purely local properties of the likelihood, that provides a tight *asymptotic* lower bound.

The information measure $\mathcal{J}$ of §3.1 similarly measures local properties of the likelihood in the mixture model. As such, lower bounds based on it will make assumptions about the local behaviour of the estimators, where the following is meant by local: A property, such as unbiasedness, which holds for τ sufficiently small in each and every local family of alternatives to λ_0 will be called a *local property at* λ_0. This definition retains from the one parameter model the intuitive notion that local likelihoods are those which are most difficult to distinguish from the null (see equation (3.1.1)).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

4.2. *Lower bounds*

The purpose of this section is to establish that the $\mathcal{J}$-information of Λ^* provides a measure of interest in finite samples from models of the form $\Pi(\theta, \phi_i)$, even when the sequence $\phi_1, \ldots, \phi_n$ is quite arbitrary. Before showing this, in corollary C, the following preparation is necessary. Let $(\phi_1, \ldots, \phi_n)$ be an observation from a product measure $Q_1 \times \ldots \times Q_n$, with each $Q_i \in \mathcal{Q}$. Now the model for $(X_1, \ldots, X_n)$ is $\prod_{i=1}^{n} (\theta, Q_i)$ rather than $(\theta, Q)^n$. In this case, one dimensional subfamilies are of the form $\Pi(\theta + \alpha\tau, Q_{i\tau})$. Let $\mathcal{J}_n^{+(-)}$ be the upper (lower) minimal Fisher's information for this space, $\mathcal{J}_1^{+(-)}$ thus being the measure for the original Λ^*.

LEMMA A. $\mathcal{J}_n^{+(-)}(\theta_0 | Q_{10} \times \ldots \times Q_{n0}) = \sum_{i=1}^{n} \mathcal{J}_1^{+(-)}(\theta_0 | Q_{i0})$.

Remark. The proof is in appendix C.

THEOREM B. If T is a locally unbiased estimator in $\Pi(\theta, Q_i)$ of a differentiable function $h(\theta)$ at null $\Pi(\theta_0, Q_{i0})$, then

$$\mathrm{var}_0(T) \geqslant [h'(\theta_0)]^2 / \mathcal{J}_n(\theta_0 | Q_{10} \times \ldots \times Q_{n0}).$$

Remark. The proof, a simple application of the Cauchy–Schwarz inequality, is found in appendix C. Lemma A provides a simplified means of computation of the bound.

COROLLARY C. If T is a globally unbiased estimator of $h(\theta)$ in $\Pi(\theta, \phi_i)$, then at null point $\Pi(\theta_0, \phi_{i0})$,

$$\mathrm{var}_0(T) \geqslant [h'(\theta_0)]^2 / \mathcal{J}_n(\theta_0 | \delta(\phi_{10}) \times \ldots \times \delta(\phi_{n0})).$$

Remark. The proof, found in appendix C, consists of showing that T is also locally unbiased in $\Pi(\theta, Q_i)$ at the null $\Pi(\theta, \delta(\phi_{i0}))$. Theorem B then applies. This result will be used to show that S_n^2 is the minimum variance unbiased estimate of the variance θ in the one way an.o.va. model. This result can be obtained by other means, but the form of the result ensures that functions $h(\theta)$, which can be more efficiently estimated than θ itself, do not exist. Notice that (3.2.1) ensures that the bound is superior to the Cramér–Rao lower bound.

5. LOWER BOUNDS FOR C.A.N. ESTIMATORS

5.1. *A review of asymptotic efficiency*

Before a lower bound for c.a.n. estimates is established, a brief review of the structure of such bounds is offered. For the purposes of this review, the variables $X_1, \ldots, X_n$ are presumed to be i.i.d. observations from a density $f(x; \theta)$ that satisfies the usual regularity conditions in θ, a real-valued parameter, and has Fisher's information $\mathcal{I}$.

Fisher (1925) proposed that the inverse of the Fisher's information was a lower bound to the asymptotic variance of consistent asymptotically normal estimators. The maximum likelihood estimators met that bound and so appeared supreme. However, by constructing 'superefficient' estimators, Hodges (see Le Cam 1953) revealed that there can be no lower bound, other than zero, to that asymptotic variance. Are superefficient estimators then preferable to the m.l.e.? The following results suggest that they are not.

First, any globally c.a.n. estimator $\{T_n\}$ of θ must have asymptotic variance at least as large as $\mathcal{I}^{-1}$ for almost all θ with respect to Lebesque measure (Le Cam 1953; also Bahadur 1964). For a smooth family, $\mathcal{I}(\theta)$ is continuous in θ, so that, if $\{T_n\}$ has a continuous asymptotic variance, it

can nowhere go beneath the lower bound. The continuity of the variance of an estimator is of undoubted utility if that variance is to be estimated.

A second result is that $\mathscr{I}^{-1}$ is a global lower bound for the asymptotic variance of estimators that are uniformly approaching normality on compact subsets of Θ (Rao 1963). The uniform approach to normality is essential if one wishes to use the asymptotic distribution for confidence intervals; see Wolfowitz (1965) and Roussas (1972) for stronger results of the same type.

Lastly, Le Cam (1953) has shown that superefficiency at one point entails bad risk values (asymptotically) in a vicinity of that point. From this work, as extended by Hajek (1972), it is clear that, under regularity, the maximum likelihood estimator minimizes the asymptotic local maximum risk.

Each of these results provides a different insight into the problem of optimal estimation. Yet a different approach will be used here, creating a result that looks much like an asymptotic version of the Cramér–Rao lower bound for locally unbiased estimates. To do so, the asymptotic version of 'locally unbiased' is developed.

5.2. *Local properties of estimators*

Let $\mathscr{N}(\cdot)$ denote the standardized normal distribution function. If $\{T_n\}$ is an estimator of θ such that for every local family $\{\lambda_\tau\}$ of alternatives to λ_0 there exists an interval $I = [0, \tau_0]$ and a function $\sigma(\tau)$, the *asymptotic variance*, such that for all $y \in \mathbb{R}$

$$\sup_{\tau \in I} \{|P[(T_n - (\theta_0 + \alpha\tau))/\sigma(\tau) \leqslant y; \lambda_\tau] - \mathscr{N}(y)|\} \to 0 \quad \text{as} \quad n \to \infty, \tag{5.2.1}$$

then $\{T_n\}$ will be called *locally uniformly asymptotically normal* (l.u.a.n.). It will be called *locally uniformly asymptotically median unbiased* (l.u.a.m.u.) if (5.2.1) holds at $y = 0$, in which case it may be rewritten, for $\theta_\tau = \theta_0 + \alpha\tau$, as

$$\sup_{\tau \in I} \{|P[T_n \leqslant \theta_\tau; \lambda_\tau] - \tfrac{1}{2}|\} \to 0.$$

If an estimator $\{T_n\}$ fails to be l.u.a.n., then the asymptotic distribution is no longer guaranteed to be a reasonable approximation for any large fixed sample size to the distribution of that estimator in any arbitrarily small local neighbourhood of the null. For such an estimator it is not clear that asymptotic variance is a reasonable measure of optimality. Moreover, the use of the asymptotic distribution for confidence intervals is impossible. Although the l.u.a.n. property is more obviously desirable, the weaker requirement of l.u.a.m.u. suffices for the lower bound result.

5.3. *The lower bound theorem*

Once again, the $X_1, \ldots, X_n$ are i.i.d. observations from (θ, Q).

THEOREM. If $\{T_n\}$ is c.a.n.-l.u.a.m.u. for θ at null $\lambda_0 = (\theta_0, Q_0)$, with asymptotic variance σ_0^2, then

$$\sigma_0^2 \geqslant \mathscr{I}^{-1}(\theta_0 | Q_0).$$

Remark. The proof is to be found in appendix D. It is a straightforward adaptation of results of Bahadur (1964) to the one-sided informations used herein. The results of Bahadur apply to more general sampling frameworks than i.i.d. sampling (see his § 4). Thus, this result could be extended to models in which the denxities f_i for X_i depend on i in some known fashion. This would clearly require modification of the notion of local property.

6. Likelihood factors

6.1. *Definition and properties*

Suppose that the random vector $\boldsymbol{X}_k = (X_1, \ldots, X_k)$ has a distribution $F(\cdot\,; \lambda)$ which depends on λ in parameter space Λ. (For the moment, Λ is arbitrary and the existence of a density is not required.) The function $g(\boldsymbol{X}; \lambda)$ will be called a *likelihood factor* if for each λ and λ_0 the *g-likelihood ratio*

$$L_g := L_g(\lambda, \lambda_0) = g(\boldsymbol{X}; \lambda)/g(\boldsymbol{X}; \lambda_0)$$

is well defined and

$$E_0[L_g] = 1. \tag{6.1.1}$$

The last equation, highly useful in establishing a score theory for likelihood factors, will be called the *factor equation.*

The work 'factor' is used because, if $\boldsymbol{X}_k$ has density f and likelihood factor p, then $r = f/p$, if well defined, is also a likelihood factor. Hence $f = pr$ is a factorization of f into two likelihood factors.

It is evident that marginal densities provide examples of likelihood factors. A simple check shows that if T and C are statistics such that T has a conditional density $g(t|c; \lambda)$ for each fixed $C = c$, then g is a likelihood factor. It is true, however, that there may exist likelihood factors that are neither marginal nor conditional densities. In particular, suppose that $\boldsymbol{X}_k$ has a density factorization of the form:

$$f(\boldsymbol{x}_k; \lambda) = f_1(x_1; \lambda) \prod_{i=2}^{k} f_i(x_i | \boldsymbol{x}_{i-1}; \lambda).$$

Here, $f_i(x_i | \boldsymbol{x}_{i-1}; \lambda)$ represents the conditional density of X_i, given that $\boldsymbol{X}_{i-1} := (X_1, \ldots, X_{i-1})$. Cox (1975) suggested that asymptotic inference (as $k \to \infty$) could be based on subproducts of the f_i terms, chosen for their simple structure. Such a subproduct, termed by him a partial likelihood, is an example of a likelihood factor.

6.2. *Partial likelihood factorization*

Attention is now focused on the following problem. If $f(X; \theta, \phi_i)$ is the density for observation X_i, does there exist a non-trivial likelihood factorization of the form

$$f(x; \theta, \phi_i) = p(x; \theta)\, r(x; \theta, \phi_i)?$$

Notice that the factor p, hereafter called the *partial likelihood*, after Cox (1975), does not depend functionally on ϕ. The factor r will be called the *remainder likelihood.*

The existence of such partial likelihood factorizations is of key importance to the nuisance parameter problem because, as will be seen in § 7, it can provide a means of obtaining c.a.n. estimators in those models in which the m.l.e. fails, to wit, by maximization of

$$\prod_{i=1}^{n} p(X_i; \theta).$$

There does not appear to be a single direct method of finding such a factorization. Two methods of major importance will be briefly reviewed here. These can be viewed as the likelihood factorizations implied by certain hypothesis testing procedures.

6.3. *Conditional factors*

One possible approach is to use the conditioning techniques of the theory of unbiased testing (Lehmann 1959, ch. 4). Suppose that for each fixed θ there exists a minimal sufficient statistic C for the parameter ϕ, where C is not functionally dependent on θ. If the pair (T, C) is sufficient for the pair (θ, ϕ), then the conditional distribution of $T|C$ does not depend on ϕ. If the densities exist, a factorization of their joint density f,

$$f(t, c; \lambda) = p(t|c; \theta)\, r(c; \lambda),$$

will be possible.

This sort of factorization (and the asymptotic inference for it) is extensively treated by Andersen (1973). The one way an.o.va. model of §1 provides a demonstration of its use. The statistic $C_i = \bar{X}_i$ is sufficient for the mean ϕ_i when θ is fixed. The statistic

$$T_i = \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i.})^2,$$

together with C_i, forms a pair of statistics sufficient for (θ, ϕ_i). The factorization is

$$f(t, c; \theta, \phi_i) = p(t; \theta)\, r(c; \theta, \phi_i),$$

where r is the marginal density of $\bar{X}_i$ and, because of independence, p is the marginal density of T_i. In §§9.3 and 9.5 two further examples of this type of factorization are provided.

6.4. *Maximal invariants*

The second approach to factorization corresponds to the theory of invariant testing (Lehmann 1959, ch. 6). One examines the problem for a group of transformations $\mathcal{G}$ that act on the sample space in such a fashion that

(1) If X has density $f(\cdot\,; \lambda)$, then for each $g \in \mathcal{G}$, gX has density $f(\cdot\,; g^*\lambda)$ for some $g^*\lambda \in \Lambda$.

(2) The orbits of Λ induced by $\mathcal{G}^* := \{g^* : g \in \mathcal{G}\}$ correspond to the values of θ. (The orbits are equivalence classes of elements that can be carried into one another by $\mathcal{G}^*$.)

If the orbits of the sample space under $\mathcal{G}$ correspond to a statistic T, that statistic is called a maximal invariant. Its distribution depends on θ alone. If (T, C) are sufficient for (θ, ϕ_i), then the factorization

$$f(t, c; \lambda) = p(t; \theta)\, r(c|t; \lambda)$$

is possible if the densities exist.

In the one way an.o.va. model, the statistic T defined in (6.3.1) is a maximal invariant under the group of location transformations. The factorization (6.3.2) is thus also induced by invariance. Other examples of maximal invariant factorizations appear in §§9.3, 9.4 and 9.6.

6.5. *Information and likelihood factorizations*

A principal feature of this paper is that the informativeness of a conditioning or invariant procedure can be judged purely on the single principle of mixture model efficiency. There is an extensive literature dealing with the desirability and optimality of conditioning on 'ancillary' statistics or using 'marginally sufficient' (by invariance) statistics (see, for example, Basu 1977; Sprott 1975; Barndorff-Nielsen 1973). The approach herein is held advantageous because it provides an answer to the question, 'Can one find an appreciably more powerful procedure by relaxing similarity or invariance requirements?' Of the five examples of this paper (§9), four have fully efficient partial likelihoods and one has an inefficient invariant partial likelihood.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

7. Partial maximum likelihood estimators (p.m.l.e.)

7.1. *Preliminary remarks*

If $\Pi p(X_i;\theta)$ has a unique maximum at $\hat{\theta}_n$ as a function of θ, $\hat{\theta}_n$ will be called the *partial maximum likelihood estimator* (p.m.l.e). If p is a marginal density for statistic T, then the p.m.l.e. is just the maximum likelihood estimator for the i.i.d. sample $T_1,\ldots,T_n$. Hence, under regularity, the usual properties of maximum likelihood estimators hold. Andersen (1973) has shown that, if p is a conditional likelihood, then the resulting p.m.l.e. (which he called the conditional maximum likelihood estimator) has (under regularity) the desired properties of consistency and asymptotic normality, provided that the null sequence of nuisance parameters is either an i.i.d. sample from a distribution Q or drawn from a compact set Φ.

Because of these previous treatments, the thrust of this section is restricted to demonstrating the role of the factor equation (6.1.1) in obtaining asymptotic results.

7.2. *Regularity assumptions*

Fix null $\lambda_0 = (\theta_0, Q_0)$, let $L_p(\theta) := p(X;\theta)/p(X;\theta_0)$, and let $l_p(X;\theta) := \ln p(X;\theta)$. For the purposes of deriving an asymptotic theory, the following assumptions concerning L_p are made. Let dashes denote derivatives with respect to θ. The assumptions are stated in a form that makes explicit their relation with the factor equation (6.1.1).

Assumption A. The ratio $L_p(\theta)$ is twice differentiable in θ.

Assumption B. $E_0[l_p(X;\theta)] < \infty$ for θ in a neighbourhood of θ_0.

Assumption C. For all θ in a neighbourhood of θ_0,

$$P_0[L_p(\theta) = 1] < 1.$$

Assumption D. For θ_1 in a neighbourhood of θ_0, the following reversal of integral and limit takes place.

$$0 = \lim_{\theta\to\theta_1} E_0 \frac{L_p(\theta) - L_p(\theta_1)}{\theta - \theta_1} = E_0[L_p'(\theta_1)].$$

An implication of this assumption is that $E_0[l_p'(X;\theta_0)] = 0$.

Assumption E. A second order reversal of integral and limit holds:

$$0 = \lim_{\theta_1\to\theta_0} E_0 \frac{L_p'(\theta_1) - L_p'(\theta_0)}{\theta_1 - \theta_0} = E_0[L_p''(\theta_0)].$$

An implication of this assumption is that $-E_0[l_p''(X;\theta)] = E_0[l_p'(X;\theta)]^2$.

Assumption F. There exists a random variable $M(X)$ such that

$$|l_p''(X;\theta)| \leqslant M(X)$$

or all θ in some neighbourhood of θ_0, with $E_0[M(X)] < \infty$.
The function $E_0[l_p'(X;\theta_0)]^2 = \mathscr{I}_p(\lambda_0)$ will be called the *partial Fisher's information.*

7.3. *Asymptotic properties*

LEMMA. Suppose that for each n, $\Sigma\, l_p(X_i;\theta)$ has, almost surely, λ_0, a unique maximum $\hat{\theta}_n$ as a function of θ. Under assumptions A, B, and C of § 7.2, $\hat{\theta}_n \to \theta_0$, almost surely, λ_0.

Remark. The factor equation, assumption C, and Jensen's inequality lead to the conclusion that

$$E_0[l_p(X;\theta) - l_p(X;\theta_0)] < 0 \tag{7.3.1}$$

for θ in a neighbourhood of θ_0. The rest of the proof is standard and so is incorporated into appendix E.

Suppose $\infty > \mathscr{I}_p(\theta_0, Q_0) > 0$. If the p.m.l.e. is consistent, then it is clear from the assumptions D, E and F that a Taylor's expansion of $\Sigma l_p(X_i;\hat{\theta})$ about θ_0 will show that $\hat{\theta}_n$ is asymptotically normal, with asymptotic variance $\mathscr{I}_p^{-1}(\theta_0, Q_0)$. (See, for example, Andersen 1973.) Several additional regularity assumptions guarantee that $\hat{\theta}_n$ will be locally uniformly asymptotically median unbiased. The details are in appendix F.

The p.m.l.e. for θ in the one way an.o.va., for the factorization suggested in §§ 6.3 and 6.4, is S_n^2. It is clearly c.a.n. and l.u.a.n.

8. EFFICIENCY AND UNIQUENESS OF THE P.M.L.E.

8.1. *Locally fully informative factor*

An important criterion for the full asymptotic efficiency of the p.m.l.e. in the mixture model is now presented. A likelihood factor $p(X;\theta)$ will be called *locally fully informative* (l.f.i.) at $\lambda_0 = (\theta_0, Q_0)$ in Λ^* if there exists a local family of alternatives $\mathscr{F}(\alpha, Q_\cdot)$ such that the remainder $r = f/p$ satisfies

$$\int r(X;\theta_0+\alpha\tau,\phi)\,\mathrm{d}Q_\tau(\phi) = \int r(X;\theta_0,\phi)\,\mathrm{d}Q_0(\phi), \tag{8.1.1}$$

almost surely (λ_0). (Notice that the parameter τ is locally unidentifiable based on the factor r.) *Upper* and *lower* l.f.i. will mean that (8.1.1) holds for $\alpha = +1$ and $\alpha = -1$, respectively. The measure Q_τ will be called the *remainder eliminating distribution* because, for this local family, $L_\tau = L_p(\theta_0+\alpha\tau)$.

The following lemma simplifies the determination of l.f.i. The fixed points are, in a sense, the most difficult.

LEMMA. If p is upper (lower) l.f.i. at all fixed points $\lambda_0 = (\theta_0, \delta(\phi_0))$, then it is upper (lower) l.f.i. at all points (θ_0, Q_0).

Proof. If the measure $Q_\tau(\cdot\,|y)$ is a remainder eliminating distribution for null (θ_0, y), define measure Q_τ^* by

$$\int_B \mathrm{d}Q_\tau^*(\phi) = \int_B \mathrm{d}Q_\tau(\phi|y)\,\mathrm{d}Q_0(y).$$

This measure satisfies

$$\begin{aligned}\int r(X;\theta_0+\alpha_\tau,\phi)\,\mathrm{d}Q_\tau^*(\phi) &= \iint r(X;\theta_0+\alpha\tau,\phi)\,\mathrm{d}Q_\tau(\phi|y)\,\mathrm{d}Q_0(y)\\ &= \int r(X;\theta_0,y)\,\mathrm{d}Q_0(y),\end{aligned}$$

as required.

As an example consider the factorization (6.3.2) for the one way an.o.va. problem. It is claimed that p is lower l.f.i. Here, r is a normal density with mean ϕ and variance θ/J. Let $\lambda_0 = (\theta_0, \delta(\phi_0))$ be a fixed point of Λ^*. Let Q_τ be the normal distribution with mean ϕ_0 and variance τ/J. Then

$$\int r(X; \theta_0 - \tau, \phi)\, \mathrm{d}Q_\tau(\phi) = r(X; \theta_0, \phi_0)$$

because the left hand side is the convolution of two normal densities with respective means 0 and ϕ_0 and respective variances $(\theta_0 - \tau)/J$ and τ/J.

8.2. *Efficiency*

Suppose now that for a given density with l.f.i. partial likelihood p, the p.m.l.e. is c.a.n.-l.u.a.m.u. with asymptotic variance $\mathscr{I}_p^{-1}$. If the family $\mathscr{F}(\alpha, Q_\cdot)$ defined by the remainder eliminating distributions is regular (where now $L_\tau = L_p(\theta + \alpha\tau)$ and the regularity conditions of appendix A are quite similar to those of § 7.2), then the minimal chacter of $\mathscr{J}$ implies $\mathscr{J} \leqslant \mathscr{I}_p$. On the other hand, the lower bound theorem of § 5.3 gives the reverse inequality, so $\mathscr{J} = \mathscr{I}_p$. Hence the p.m.l.e. has minimal asymptotic variance. In particular, the estimator S_n^2 of the one way an.o.va. is fully mixture model efficient.

8.3. *Uniqueness*

If it were possible for there to be two l.f.i. partial likelihoods, then considerations of efficiency would not enable one to choose between them. The following lemma, again based on the factor equation, shows uniqueness of the likelihood ratios.

Lemma. If p is a l.f.i. partial likelihood, then any other partial likelihood p^* satisfies

$$E_0[\ln L_p(\theta_0 + \alpha\tau)] \leqslant E_0[\ln L_{p^*}(\theta_0 + \alpha\tau)]$$

with equality only if

$$L_p(\theta_0 + \alpha\tau) = L_{p^*}(\theta_0 + \alpha\tau),$$

almost surely (λ_0).

Proof. Because $f = pr$ and $f = p^*r^*$,

$$L_p(\theta_0 + \alpha\tau)\, L_r(\theta_0 + \alpha\tau, Q_\tau) = L_{p^*}(\theta_0 + \alpha\tau)\, L_{r^*}(\theta_0 + \alpha\tau, Q_\tau),$$

where Q_τ is the remainder eliminating distribution for partial p. Recalling that $L_r(\theta_0 + \alpha\tau; Q_\tau) = 1$, almost surely (λ_0), taking logarithms gives

$$\ln L_p(\theta_0 + \alpha\tau) - \ln L_{p^*}(\theta_0 + \alpha\tau) = \ln L_{r^*}(\theta_0 + \alpha\tau, Q_\tau).$$

Since r^* is a likelihood factor, the conclusion of the lemma follows by Jensen's inequality and the factor equation.

8.4. *Hypothesis testing*

Finally, the relation between the l.f.i. property and uniformly most powerful tests should be mentioned. Following Lehmann (1959, § 3.8), let the null hypothesis be $H: \theta = \theta_0 + \alpha\tau$, ϕ unspecified, and let the alternative be K: $\theta = \theta_0$, $\phi \sim Q_0$. For each mixing distribution Q_τ, one can form the most powerful size α test of $H_Q: \theta = \theta_0 + \alpha\tau$, $\phi \sim Q$ against K; it is based on the likelihood ratio. If one minimizes the power of the test over Q, one finds a *least favourable distribution* Q_τ (see Lehmann (1952) for existence) for which the ratio L_τ generates the most powerful size α test of H against K. If the partial likelihood is l.f.i., then the remainder eliminating distributions

are least favourable in terms of minimizing information and so are, at least, approximately least favourable in the Lehmann sense. Hence, the most powerful test of $H^-\colon \theta \leqslant \theta_0$ (for $\alpha = -1$) or $H^+\colon \theta \geqslant \theta_0$ (for $\alpha = +1$) against $K^-\colon \theta > \theta_0$ or $K^+\colon \theta < \theta_0$ is generated, approximately, by the p.m.l.e. For example, in the one way an.o.va., S_n^2 generates the most powerful test of $H\colon \theta \leqslant \theta_0$ against $K\colon \theta > \theta_0$ (see Lehmann 1959, §3.9.)

9. The examples

9.1. *Preliminary remarks*

The emphasis in this section is upon showing that the locally fully informative criterion can be a useful tool. The demonstration of this property in any particular factorization is not necessarily elementary, but, once done, corollary C of §4.2 and the theorem of §5.3 show that the partial likelihood is a sufficient source of information. The last example demonstrates that not all partial factorizations have this property.

9.2. *One way an.o.va.*

In §6.3 and 6.4 it was demonstrated that the density of the pair (T_i, C_i) couldbe factored into a partial likelihood (the marginal density of T_i) multiplied by a remainder likelihood (the marginal density of C_i) It was pointed out in §7.4 that S_n^2 is the p.m.l.e. and so, because of regularity, is c.a.n.-l.u.a.m.u. In §8.2 it was shown that p is locally fully informative. The conclusion is that

$$\mathscr{J}(\theta_0|Q_0) = \mathscr{J}_p(\theta_0|Q_0) = \frac{J-1}{2\theta_0^2}$$

and that S_n^2 is fully mixture efficient. Moreover, from corollary C of §4.2, S_n^2 is not only the minimum variance unbiased estimator, but also the most efficient unbiased estimator.

9.3. *Exponentials with unknown support*

Suppose that

$$X_{ij} = \theta Y_{ij} + \phi_i \quad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, J),$$

where the Y_{ij} are independent unit exponentials. In this model, the support of the density changes with ϕ_i. Let $X_{i(1)}, \ldots, X_{i(J)}$ be the order statistics for $X_{i1}, \ldots, X_{iJ}$. The random variables X_{ij} can be transformed to

$$Z_{i1} = JX_{i(1)},$$
$$Z_{ik} = (J-k+1)\,(X_{i(k)} - X_{i(k-1)}) \quad (k = 2, 3, \ldots, J).$$

Then $\{Z_{i1} - J\phi_i, Z_{i2}, \ldots, Z_{iJ}\}$ are distributed as i.i.d. exponentials with mean θ (see, for example, Johnson & Kotz 1970). In this transformed space it is easily seen that, since the m.l.e. of $J\phi_i$ is Z_{i1}, the m.l.e. of θ is

$$\hat{\theta}_n^* = \frac{1}{nJ} \sum_{i=1}^{n} \sum_{j=2}^{J} Z_{ij}.$$

The m.l.e. is thus inconsistent, converging to $[(J-1)/J]\,\theta$.

On the other hand (since Z_{i1} is clearly minimal sufficient for ϕ_i) the conditioning argument of §6.3 leads to the marginal density of $Z_{i2}, \ldots, Z_{iJ}$ as a partial likelihood. The same conclusion follows from using location transformations on the original X_{ij} variables and an invariance factorization. The remainder likelihood is the density of Z_{i1}.

It is now demonstrated that the partial likelihood is locally fully informative. Fix null $\lambda_0 = (\theta_0, \delta(\phi_0))$. First note that the requirement $E[L_r] = 1$ on the local family of alternatives $\mathscr{F}(\alpha, Q)$ implies that Q must have mass one on $[\phi_0, \infty)$. For fixed i, let $Y := Z_{i1}$ and $\phi_0 := J\phi_{i0}$. Let W be a unit exponential variable. If $\overset{\mathscr{L}}{=}$ denotes equality of probability distribution, then, under the null,

$$Y \overset{\mathscr{L}}{=} \theta_0 W + \phi_0.$$

Under an alternative (θ, Q), where Q puts mass 1 on $[\phi_0, \infty)$,

$$Y \overset{\mathscr{L}}{=} \theta W + U,$$

where U is a random variable with distribution Q, independent of W. If Q is to be remainder eliminating, then

$$\theta W + V \overset{\mathscr{L}}{=} \theta_0 W, \tag{9.3.1}$$

where $V = U - \phi_0$, a non-negative random variable. It is now shown that such a non-negative V exists for $\theta = \theta_0 - \tau > 0$.

Setting Laplace transforms equal in (9.3.1), one obtains

$$\begin{aligned} E[e^{-Vt}] &= E[e^{-\theta_0 Wt}]/E[e^{-\theta Wt}] \\ &= (1+\theta t)/(1+\theta_0 t) \\ &= \frac{\theta}{\theta_0} + \left(1 - \frac{\theta}{\theta_0}\right)\frac{1}{1+\theta_0 t}. \end{aligned}$$

Provided that $0 \leqslant \theta < \theta_0$, this is the Laplace transform of a random variable V, which is 0 with probability θ/θ_0 and exponential with mean parameter θ_0 with probability $(1-\theta/\theta_0)$.

The conclusion is that the p.m.l.e.

$$\hat{\theta}_n = \left(\sum_{i=1}^{n}\sum_{j=1}^{J} Z_{ij}\right)/[n(J-1)]$$

is a fully efficient unbiased estimator of θ and asymptotically fully mixture space efficient.

9.4. *Bernoulli pairs with invariant reversals*

For each i the pair (X_i, Y_i) are independent Bernoulli random variables with a common probability of success $\frac{1}{2} + \theta\phi_i$, where ϕ_i is $+1$ or -1 and $\theta \in [0, \frac{1}{2}]$. This model is an elementary version of one arising in the study of evolutionary trees (see Felsenstein 1973).

The transformation $(x, y) \to (1-x, 1-y)$, together with the identity, form a group of transformations as in §6.4. The orbits in the parameter space correspond to θ and the orbits in the sample space correspond to values of W_i, the indicator function for the set $\{X_i + Y_i = 1\}$. The partial likelihood so generated is then the marginal density of W_i. (Only the trivial $p \equiv 1$ is generated by the conditional approach of §6.3.) The variables $W_1, \ldots, W_n$ are independent Bernoullis with mean $2(\frac{1}{4} - \theta^2)$.

By symmetry, it suffices to show that the partial likelihood is l.f.i. at null $(\theta_0, 1)$ for $\theta_0 \in [0, \frac{1}{2}]$. For fixed i, let $S = X_i + Y_i$. Since (S, W) is sufficient for (θ, ϕ), it suffices to show that there exists a remainder eliminating distribution Q on $\{-1, +1\}$ for the conditional density of S given W. Let

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

The density of S given W is defined by

$$\begin{aligned}
&P[S=1|W=1]=1,\\
&P[S=0|W=0]=a(\theta) \quad \text{if} \quad \phi=1,\\
&P[S=0|W=0]=1-a(\theta) \quad \text{if} \quad \phi=-1,\\
&P[S=1|W=0]=0,\\
&P[S=2|W=0]=1-a(\theta) \quad \text{if} \quad \phi=1,\\
&P[S=2|W=0]=a(\theta) \quad \text{if} \quad \phi=-1.
\end{aligned}$$

It follows that the single equation that a remainder eliminating distribution Q must satisfy, if $Q(\{1\})=q$, is

$$qa(\theta)+(1-q)\,(1-a(\theta))=a(\theta_0).$$

Solving for q gives

$$q=[a(\theta_0)+a(\theta)-1]/[2a(\theta)-1].$$

Given any θ_0 and any θ, if the last equation gives a value of q in $[0,1]$, then that value of q defines a remainder eliminating distribution Q by $Q(\{1\})=q$. Note first that $a(\theta)>\frac{1}{2}$ if $\theta>0$, so q is positive. Secondly, the function $a(\theta)$ is monotonely increasing for θ in $[0,\frac{1}{2}]$, so that $a(\theta_0)<a(\theta)$ is implied by $\theta_0<\theta$, which in turn implies that $q<1$. It follows that $\not{p}$ is upper l.f.i. Hence $n^{-1}\Sigma W_i$ is the (fully efficient) minimum variance unbiased estimator of $2(\frac{1}{4}-\theta^2)$ and also fully mixture space efficient asymptotically.

9.5. *Bernoulli pairs with common log odds ratio*

Of substantial theoretical and practical interest is the model in which the ith observation is a pair (X_i, Y_i) of independent Bernoulli variables with respective success parameters defined by

$$p_i=\exp(\theta+\phi_i)/(1+\exp(\theta+\phi_i))$$

and

$$q_i=\exp(\phi_i)/(1+\exp(\phi_i)).$$

The parameter θ, here constant over i, is the *log odds ratio* for the Bernoulli pair.

The minimal sufficient statistic for ϕ, when θ is fixed, is X_i+Y_i, so the argument of § 6.3 leads to the use of the conditional distribution of X_i, given that X_i+Y_i for inference about θ. There is no invariant partial likelihood.

It will here be demonstrated that the conditional likelihood is a locally fully informative partial likelihood, so that the p.m.l.e. is fully mixture efficient. Previous discussions from different points of view of this same model can be found in Barndorff-Nielsen (1973) and in Sprott (1975). Some initial results on the mixture efficiency of the conditional likelihood when X and Y are binomials can be found in Lindsay (1978).

The remainder likelihood is the marginal likelihood of $X+Y$ and is defined by the three probabilities:

$$P[X+Y=0]=\rho(\theta,\phi)=(1+e^{\phi})^{-1}(1+e^{\theta+\phi})^{-1};$$

$$P[X+Y=1]=\rho(\theta,\phi)\,e^{\phi}(1+e^{\theta});$$

and

$$P[X+Y=2]=\rho(\theta,\phi)\,e^{2\theta+2\phi}.$$

Hence a remainder eliminating distribution Q for the fixed point null $(\theta_0, \delta(\phi_0))$ must satisfy the following set of equations:

$$\left.\begin{aligned} &\int \rho(\theta,\phi)\,\mathrm{d}Q(\phi) = \rho(\theta_0,\phi_0);\\ &\int \rho(\theta,\phi)\,\mathrm{e}^{\phi}(1+\mathrm{e}^{\theta})\,\mathrm{d}Q(\phi) = \rho(\theta_0,\phi_0)\,\mathrm{e}^{\phi_0}(1+\mathrm{e}^{\theta_0});\\ \text{and}\quad &\int \rho(\theta,\phi)\,\mathrm{e}^{2\theta+2\phi}\mathrm{d}Q(\phi) = \rho(\theta_0,\phi_0)\,\mathrm{e}^{2\theta_0+2\phi_0}. \end{aligned}\right\} \tag{9.5.1}$$

If we write $\mathrm{d}\omega(\phi) = \dfrac{\rho(\theta,\phi)}{\rho(\theta_0,\phi_0)}\,\mathrm{d}Q(\phi)$, these may be rewritten as

$$\left.\begin{aligned} &\int \mathrm{d}\omega(\phi) = 1,\\ &\int \mathrm{e}^{\phi-\phi_0}\,\mathrm{d}\omega(\phi) = (1+\mathrm{e}^{\theta_0})/(1+\mathrm{e}^{\theta}),\\ \text{and}\quad &\int \mathrm{e}^{2(\phi-\phi_0)}\,\mathrm{d}\omega(\phi) = \mathrm{e}^{2(\theta_0-\theta)}. \end{aligned}\right\} \tag{9.5.2}$$

If a positive measure ω exists satisfying (9.5.2), then a positive measure Q can be constructed satisfying (9.5.1) by $\mathrm{d}Q(\phi) = \rho(\theta_0,\phi_0)/\rho(\theta,\phi)\,\mathrm{d}\omega(\phi)$. The measure Q so defined is positive (because ω is) and has mass one. (Add over the three equations in (9.5.1).)

If ϕ is a random variable with distribution ω, then the equations (9.5.2) are moment equations for the positive random variable $Z = \mathrm{e}^{\phi-\phi_0}$. Such a random variable Z exists if the mean $(1+\mathrm{e}^{\theta_0})/(1+\mathrm{e}^{\theta})$ and the variance,

$$\begin{aligned} \mathrm{var}\,(Z) &= \mathrm{e}^{2(\theta_0-\theta)} - (1+\mathrm{e}^{\theta_0})^2/(1+\mathrm{e}^{\theta})^2\\ &= (1+\mathrm{e}^{\theta_0})^2(\mathrm{e}^{-2\theta})\left[\left(\frac{\mathrm{e}^{\theta_0}}{1+\mathrm{e}^{\theta_0}}\right)^2 - \left(\frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}}\right)^2\right], \end{aligned}$$

are both positive. (The Γ distribution, for one, can fit any specified positive mean and variance.) They are positive if $\theta_0 > \theta$. Thus, the conditional likelihood is clearly lower locally fully informative.

9.6. *Paired exponentials with proportional hazards*

The final example demonstrates that there is no assurance that invariance considerations lead to a mixture efficient p.m.l.e. In this example there is a consistent asymptotically normal estimate which is strictly superior in asymptotic variance to the p.m.l.e. over part of the mixture space.

For each i let the pair (X_i, Y_i) be independent exponential random variables with a constant hazards ratio θ. That is, the density for the pair is

$$f(x,y;\theta,\phi) = \theta\phi^2 \exp[-(\theta x + y)\,\phi],$$

with positive valued parameters (θ,ϕ). The scale transformations $(x,y) \to (bx,by)$, b positive, transform the parameter space by $(\theta,\phi) \to (\theta, b^{-1}\phi)$, so the orbits correspond to θ. For these transformations the random variable $Z = X/Y$ is a maximal invariant. It has density

$$p(z;\theta) = (z\theta+1)^{-2} \quad (z \in [0,\infty)).$$

The information in this likelihood is

$$\mathscr{I}_p(\theta|Q) = 1/(3\theta^2)$$

and so, because of regularity, the p.m.l.e. has asymptotic variance $3\theta^2$. (Aside: For this model it is also true that the p.m.l.e. is the same as the m.l.e. for $\Pi(\theta, \phi_i)$. Since the Fisher's information for the full likelihood is $(2\theta^2)^{-1}$, this is a further example (see §1.2) of m.l.es that are consistent but not fully efficient.)

The maximum likelihood estimate for the fixed point model $(\theta, \phi)^n$ is

$$\hat{\theta}^*_n = \Sigma Y_i / \Sigma X_i.$$

If the null distribution is $(\theta, Q)^n$, then $\hat{\theta}^*$ is consistent for all measures Q such that $\int \phi^{-1}\,\mathrm{d}Q < \infty$, as

$$\hat{\theta}^*_n \to E_0 Y_i / E_0 X_i = \theta_0.$$

Provided that $\int \phi^{-2}\,\mathrm{d}Q_0 < \infty$, this estimator is also asymptotically normal, as

$$n^{\frac{1}{2}}(\hat{\theta}^*_n - \theta_0) = \frac{n^{-\frac{1}{2}}\Sigma(Y_i - \theta X_i)}{n^{-1}\Sigma X_i}.$$

Here the denominator converges almost surely to $\int (\theta_0 \phi)^{-1}\,\mathrm{d}Q_0$, and the numerator converges in law to a normal random variable with mean 0 and variance $2\int \phi^{-2}\,\mathrm{d}Q_0$. Hence

$$n^{\frac{1}{2}}(\hat{\theta}^* - \theta_0) \to N(0, 2\theta_0^2 c(Q_0)),$$

where

$$c(Q) = \frac{\int \phi^{-2}\,\mathrm{d}Q}{(\int \phi^{-1}\mathrm{d}Q)^2} \geqslant 1.$$

Thus, for all Q_0 such that $c(Q_0) < \frac{3}{2}$, the fixed point m.l.e. $\hat{\theta}^*$ is strictly superior to the p.m.l.e. in asymptotic variance.

10. Concluding discussion

This paper represents a preliminary exploration of the mixture space Λ^* and its information properties. The first conclusion is that partial likelihood techniques can provide fully mixture efficient estimators, but not invariably so.

The secondary explorations needed for this topic include, first, establishment of better techniques for computing the $\mathscr{J}$-information and, secondly, development of estimators that uniformly attain the lower bound of §5.3. In regard to the last point, it is possible that the m.l.e. for the mixture space of the pair (θ, Q), defined and discussed by Kiefer & Wolfowitz (1956), may provide such efficient estimators of θ. Another approach deserving exploration is the impact of various finite parameter representations of the mixture space upon overall mixture space consistency and efficiency.

A final point is that because of the asymmetry of the mixture space the measures defined in this paper apply only to scalar parameters. The extension to a multivariate parameter θ awaits further research.

References

Andersen, E. B. 1970 *Jl. R. statist. Soc.* B **32**, 283–301.
Andersen, E. B. 1973 *Conditional inference and models for measuring*. Copenhagen: Mentalhygienisk Forlag.
Bahadur, R. R. 1964 *Ann. math. Statist.* **38**, 303–324.
Barankin, E. 1949 *Ann. math. Statist.* **20**, 447–501.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Barndorff-Nielsen, O. 1973 *Biometrika* **60**, 447–455.
Basu, D. 1977 *J. Am. statist. Ass.* **72**, 355–367.
Chernoff, H. 1956 *Ann. math. Statist.* **27**, 1–22.
Cox, D. R. 1975 *Biometrika* **62**, 269–276.
Cramér, H. 1946 *Mathematical methods of statistics*. Princeton University Press.
Feller, W. 1966 *An introduction to probability theory and its applications*, vol. II (2nd edn). New York: John Wiley.
Felsenstein, J. 1973 *Syst. Zool.* **22**, 240–249.
Fisher, R. A. 1925 *Proc. Camb. phil. Soc.* **22**, 700–715.
Hajek, J. 1972 *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1**, 175–194.
Kiefer, J. & Wolfowitz, J. 1956 *Ann. math. Statist.* **27**, 887–906.
Johnson, N. L. & Kotz, S. 1970 *Continuous univariate distributions*. vol. I. *distributions in statistics*. New York: John Wiley.
Le Cam, L. 1953 *Univ. Calif. Publs Statist.* **1**, 277–329.
Lehmann, E. L. 1952 *Ann. math. Stat.* **23**, 408–416.
Lehmann, E. L. 1959 *Testing statistical hypotheses*. New York: John Wiley.
Lindsay, B. G. 1978 Doctoral dissertation, University of Washington.
Neyman, J. & Scott, E. L. 1948 *Econometrika* **16**, 1–32.
Rao, C. R. 1963 *Sankhyā* A **25**, 188–206.
Rao, C. R. 1973 *Linear statistical inference and its applications* (2nd edn). New York: John Wiley.
Roussas, C. G. 1972 *Continuity of probability measures: some applications in statistics*. Cambridge University Press.
Sprott, D. A. 1975 *Biometrika* **62**, 599–605.
Stein, C. 1956 *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 187–195.
Wolfowitz, J. 1965 *Theor. Probab. Applic.* **10**, 247–260.

Appendix A. Regularity conditions

For the family $\mathscr{F}(\alpha, Q)$ and L_τ as in §3.1, let $l_\tau = \ln L_\tau$ and let dashes denote derivatives with respect to τ. The following regularity conditions (A 1–A 4) are adaptations of the regularity conditions used by Bahadur (1964), which are themselves simplified versions of those of Cramér (1946). Those regularity conditions are needed for the lower bound theorems in §§4.2 and 5.3.

(A 1) The first two derivatives of l_τ exist and are continuous on $[0, \epsilon)$. Here l_0' and l_0'' will mean left-hand derivatives at $\tau = 0$.

(A 2) The following interchange of limit and integral takes place:

$$\lim_{\tau' \to \tau} E_0[(L'_{\tau'} - L_\tau)/(\tau' - \tau)] = E_0[L_\tau'] = 0, \quad \forall \tau \in [0, \epsilon).$$

This implies that $E[l_0'] = 0$.

(A 3) The following interchange of limit and integral takes place:

$$\lim_{\tau \to 0} E_0[(L_\tau' - L_0')/\tau] = E_0[L_0''] = 0.$$

This implies that $E_0[-l_0''] = E_0[l_0']^2 := \mathscr{I}(\lambda_0 | \alpha, Q)$. This last quantity is the Fisher's information at $\tau = 0$ for the one-dimensional subfamily.

(A 4) There exists a random variable $M(X)$ such that

$$|l_\tau''| \leqslant M(X)$$

for all $\tau \in [0, \epsilon)$, with $E_0[M(X)] < \infty$.

Adding the second set of regularity conditions (A 5–A 7) proves useful in finding the least favourable family $\mathscr{F}(\alpha, Q)$. (See appendix B and proofs for §4.2.)

(A 5) For $\tau \in [0, \epsilon)$, $E_0[-l_\tau]$ is finite and also twice differentiable in τ at $\tau = 0^+$, with

$$\frac{\partial}{\partial \tau} E_0[-l_\tau]\Big|_{\tau = 0^+} = E_0[l_0'] = 0$$

and

$$\frac{\partial^2}{\partial\tau^2}E_0[-l_\tau]\Big|_{\tau=0^+} = E_0[-l_0''] = \mathscr{I}(\lambda_0|\alpha, Q_\cdot).$$

(A 6) For $\tau\in[0,\epsilon]$, $E_0[L_\tau\, l_\tau]$ is finite and also twice differentiable in τ at $\tau = 0^+$, with

$$\frac{\partial}{\partial\tau}E_0[L_\tau\, l_\tau]\Big|_{\tau=0^+} = E_0[L_0\, l_0' + L_0'\, l_0] = 0$$

and

$$\frac{\partial^2}{\partial\tau^2}E_0[L_\tau\, l_\tau]\Big|_{\tau=0^+} = E_0[L_0\, l_0'' + 2L_0'\, l_0' + L_0''\, l_0] = \mathscr{I}(\lambda_0|\alpha, Q_\cdot).$$

(A 7) For $\tau\in[0,\epsilon)$, $E_0[\frac{1}{2}L_\tau]^2$ is finite and also twice differentiable in τ at $\tau = 0^+$, with

$$\frac{\partial}{\partial\tau}E_0[\tfrac{1}{2}L_\tau^2]\Big|_{\tau=0^+} = E_0[L_0\, L_0'] = 0$$

and

$$\frac{\partial^2}{\partial\tau^2}E_0[\tfrac{1}{2}L_\tau^2]\Big|_{\tau=0^+} = E_0[L_0\, L_0'' + (L_0')^2] = \mathscr{I}(\lambda_0|\alpha, Q_\cdot).$$

Appendix B. Asymmetry of minimal Fisher's information

The one way an.o.va. example demonstrates this. From the argument of § 8.3, the lower minimal Fisher's information for this model is

$$\mathscr{I}^-(\theta|Q) = \mathscr{I}_p(\theta|Q) = (J-1)/2\theta^2.$$

It will here be shown that for any ϕ,

$$\mathscr{I}^+(\theta|\delta(\phi)) = J/(2\theta^2),$$

giving the stated asymmetry.

First, some lemmas are given that reduce the minimal Fisher's information problem to a simpler form. For fixed α and τ, let

$$L_\tau(Q) = \frac{\int f(X;\theta_0+\alpha\tau,\phi)\,\mathrm{d}Q}{\int f(X;\theta_0,\phi)\,\mathrm{d}Q_0}.$$

Let $Q_{1\tau}$, $Q_{2\tau}$, and $Q_{3\tau}$ be the probability measures (if they exist) that respectively minimize

$$E_0[-\ln L_\tau(Q)],\ E_0[L_\tau(Q)\ln L_\tau(Q)] \quad\text{and}\quad E_0[\tfrac{1}{2}L_\tau^2(Q)]. \tag{B 1}$$

If the regularity conditions (A 1) to (A 7) are satisfied by any one of the families $\mathscr{F}(\alpha, Q_{1\cdot})$, $\mathscr{F}(\alpha, Q_{2\cdot})$, and $\mathscr{F}(\alpha, Q_{3\cdot})$, then conditions (A 5) to (A 7), respectively, show that

$$\mathscr{I}(\theta_0|Q_0) = \mathscr{I}(\lambda_0|\alpha, Q_{i\cdot})$$

for $i = 1$, 2, or 3. This said, it is clear that we wish to minimize the functions in (B 1) with respect to Q. The following lemmas are an aid in this minimization.

Lemma 1. The probability measure Q_τ minimizes $E_0[-\ln L_\tau(Q)]$ if and only if

$$E_0[L_\tau(\delta(\phi))/L_\tau(Q_\tau)] \leqslant 1$$

for all $\phi\in\Phi$.

LEMMA 2. The probability measure Q_τ minimizes $E_0[L_\tau(Q)\ln L_\tau(Q)]$ if and only if

$$E_0[L_\tau(Q_\tau)\ln L_\tau(Q_\tau)] \leqslant E_0[L_\tau(\delta(\phi))\ln L_\tau(Q_\tau)]$$

for all $\phi\in\Phi$.

LEMMA 3. The probability measure Q_τ minimizes $E_0[\frac{1}{2}L_\tau^2(Q)]$ if and only if

$$E_0[L_\tau^2(Q_\tau)] \leqslant E_0[L_\tau(\delta(\phi))\,L_\tau(Q_\tau)]$$

for all $\phi\in\Phi$.

The proof of each lemma is the same. Fix two probability measures $\mathscr{P}_1$ and $\mathscr{P}_0$. For $\rho\in[0,1]$, $\rho\mathscr{P}_1+(1-\rho)\mathscr{P}_0$ is also a probability measure. If $L_1:=L_\tau(\mathscr{P}_1)$ and $L_0:=L_\tau(\mathscr{P}_0)$, then $L_\tau(\rho\mathscr{P}_1+(1-\rho)\mathscr{P}_0)=\rho L_1+(1-\rho)L_0:=L_\rho$. We now minimize the respective functions of (B 1) with respect to the family of measures $\rho\mathscr{P}_1+(1-\rho)\mathscr{P}_0$.

For example, let

$$V(\rho)=E_0[L_\rho\ln L_\rho].$$

Then, dashes indicating derivatives with respect to ρ,

$$V'(\rho)=E_0[(L_1-L_0)\ln L_\rho]$$

and

$$V''(\rho)=E_0[(L_1-L_0)^2/L_\rho].$$

Hence $V(\rho)$ is convex in ρ and it is minimized at $\rho=1$ if and only if

$$V'(1)=E_0[(L_1-L_0)\ln L_1]\leqslant 0.$$

Now any probability measure $\mathscr{P}_1$ that minimizes $E_0[L_\tau(Q)\ln L_\tau(Q)]$ must minimize $V(\rho)$ for all ρ and all possible $\mathscr{P}_0$. This gives the criterion of lemma 2. The proofs of the other lemmas are similar.

Returning now to the normal example, let $\alpha=+1$ and let $Q_\tau:=\delta(\phi_0)$ for all $\tau\in[0,1]$. The criterion of lemma 2 is used to show that Q_τ minimizes $E_0[L_\tau(Q)\ln L_\tau(Q)]$ for the null $(\theta_0,\delta(\phi_0))$. Here

$$\ln L_\tau(Q_\tau)=\frac{J}{2}\ln\frac{\theta_0}{\theta_0+\tau}-\frac{1}{2}\sum_{j=1}^{J}(X_{ij}-\phi_0)^2\left[\frac{1}{\theta_0+\tau}-\frac{1}{\theta_0}\right].$$

The criterion of the lemma is equivalent to showing that $E[\ln L_\tau(Q);\,\theta+\tau,\phi]$ is minimized at $\phi=\phi_0$, which is clearly true. Use of $\mathscr{F}(1,\delta(\phi_0))$ gives the result $\mathscr{J}^+(\theta_0|\delta(\phi_0))=J/(2\theta_0^2)$, as required.

APPENDIX C. PROOFS FOR §4.2

LEMMA A (PROOF). Any regular upper (lower) one-dimensional subfamily $(\alpha, Q_{1\cdot}\times\ldots\times Q_{n\cdot})$ for null $\lambda_0=\Pi(\theta_0,Q_{i0})$ has a likelihood ratio

$$L_\tau^{(n)}=\prod_{i=1}^{n}L_{i\tau},$$

where

$$L_{i\tau}=f(X_i;\theta_0+\alpha\tau,Q_{i\tau})/f(X_i;\theta_0,Q_{i0}),$$

corresponding to a one-dimensional subfamily for null $\lambda_{i0}=(\theta_0,Q_{i0})$ in Λ^*. The regularity required by (A 5) assures that

$$\mathscr{I}(\lambda_0|\alpha,Q_{1\cdot}\times\ldots\times Q_{n\cdot})=\frac{\partial^2}{\partial\tau^2}E_0[-\ln L_\tau^{(n)}]\Big|_{\tau=0^+}=\frac{\partial^2}{\partial\tau^2}\Sigma E_0[-\ln L_{i\tau}]\Big|_{\tau=0^+}=\sum_{i=1}^{n}\mathscr{I}(\lambda_{i0}|\alpha,Q_{i\cdot}).$$

THEOREM B (PROOF). For any regular family the Cauchy–Schwarz inequality gives

$$\mathrm{var}_0[T]\,\mathrm{var}_0[L_\tau^{(n)}] \geqslant [h(\theta_0+\alpha\tau)-h(\theta_0)]^2.$$

Dividing through by $\tau^2 = \alpha^2\tau^2$ and letting $\tau \to 0^+$ gives

$$\mathrm{var}_0[T]\,\mathscr{I}(\lambda_0|\alpha, Q_{1.}\times\ldots\times Q_{n.}) \geqslant [h'(\theta_0)]^2$$

by regularity condition (A 7).

COROLLARY C (PROOF). If $\mathrm{var}_0\, T = \infty$, there is nothing to show. Otherwise, let $\mathscr{F}(\alpha, Q_{1.}\times\ldots\times Q_{n.})$ be a *local* family of alternatives to null $\Pi(\theta_0, \delta(\phi_{i0}))$. If the likelihood ratios are $L_\tau^{(n)}$, then, by the definition of local,

$$\mathrm{var}_0[L_\tau^{(n)}] < \infty,$$

for τ sufficiently small. By the Cauchy–Schwarz inequality, if E_τ is expectation under alternative τ, then

$$E_\tau^2[|T|] = E_0^2[|T|L_\tau^{(n)}] \leqslant E_0[T^2]\,E_0[L_\tau^{(n)}]^2 < \infty.$$

Hence, T is absolutely integrable under λ_τ. It follows by Fubini's theorem that $E_\tau[T] = \theta_0+\alpha\tau$. Hence, any estimator T unbiased (globally) in $\Pi(\theta, \phi_i)$ is locally unbiased in $\Pi(\theta, Q_i)$ at null $\Pi(\theta_0, \delta(\phi_{i0}))$, and so theorem B gives the necessary bound.

APPENDIX D. PROOF OF THEOREM OF §5.3

The proof consists of minor modifications of Bahadur (1964). Under the regularity conditions (A 1) to (A 4), the regular one-dimensional family $\mathscr{F}(\alpha, Q_.)$ satisfies Bahadur's regularity conditions (i)–(iv). Moreover, a careful reading of his proofs for lemmas 1 and 2 and for proposition 1 reveal that nowhere does he use his requirement that the null θ_0 be in the interior of Θ; all that is needed is a one-sided approach to the null by the alternatives. (The Taylor expansion in his equation (13) holds on the boundary by repeated applications of the mean value theorem.) Moreover, proposition 1 only requires the estimator T_n to be asymptotically normal at the null. This said, we may restate his proposition 1 in the notation of this paper.)

PROPOSITION. Let $\tau_n = n^{-\frac{1}{2}}$ and let $\mathscr{F}(\alpha, Q_.)$ be a regular family. If $\{T_n\}$ is c.a.n. for θ at $\lambda_0 = (\theta_0, Q_0)$, with asymptotic variance v, and if

$$\liminf_{n\to\infty} P[T_n < \theta_0+\alpha\tau_n; \lambda_{\tau_n}] \leqslant \tfrac{1}{2}, \tag{D 1}$$

then

$$v \geqslant \mathscr{I}^{-1}(\lambda_0|\alpha, Q_.).$$

The theorem of §5.3 is now an elementary corollary. The fact that T_n is l.u.a.m.u. gives

$$\sup\{|P[T_n < \theta_0+\alpha\tau; \lambda_\tau] - \tfrac{1}{2}| : \tau\in[0,\epsilon)\} \to 0 \quad \text{as} \quad n\to\infty,$$

which implies (D 1). The conclusion of the proposition can then be applied to arbitrary regular families to give the theorem.

Appendix E. Proof of lemma of §7.3

From equation (7.3.1) it follows that for $\delta > 0$ (sufficiently small) there exists an integer $n_0(\delta)$ such that

$$\max\left\{\sum_{i=1}^{n} l_p(X_i;\theta+\delta), \sum_{i=1}^{n} l_p(X_i;\theta-\delta)\right\} < \sum_{i=1}^{n} l_p(X_i;\theta_0),$$

almost surely λ_0, for $n \geqslant n_0(\delta)$. It follows by differentiability of l_p that the equation

$$\Sigma l_p'(X_i;\theta) = 0$$

almost surely has a solution in $(\theta_0-\delta, \theta_0+\delta)$ for $n \geqslant n_0(\delta)$. Since this solution must be $\hat{\theta}_n$ and since δ is arbitrarily small, $\hat{\theta}_n \to \theta_0$, almost surely λ_0.

Appendix F. The l.u.a.m.u. property of the p.m.l.e.

The following additional conditions are sufficient to ensure that the p.m.l.e. $\hat{\theta}_n$ is l.u.a.m.u.

(1) For each local family $\mathscr{F}(\alpha, Q_\cdot)$ there exists an integer n_0 and a real number $\epsilon_1 > 0$ such that for $n \geqslant n_0$ the equation $\sum_{i=1}^{n} l_p'(X_i;\theta) = 0$ has a single root $\hat{\theta}_n$, almost surely λ_τ for all $\tau \in [0, \epsilon_1]$.

(2) $E_0[l_p'(X;\theta)]^4$ and $E_0[l_p'(X;\theta)]^6$ are finite in a neighbourhood of θ_0 and continuous at θ_0.

Proof. Fix local family $\mathscr{F}(\alpha, Q_\cdot)$ and let $\theta_\tau = \theta_0 + \alpha\tau$. Assumption (1) ensures that $\Sigma l_p'(X_i;\theta)$ is negative for $\theta > \hat{\theta}_n$ and positive for $\theta < \hat{\theta}_n$, almost surely λ_τ for all $\tau \in [0, \epsilon]$. Hence it suffices to show that there exists an $\epsilon' > 0$ such that

$$\sup\{|P_\tau[\Sigma l_p'(X_i;\theta_\tau) \leqslant 0] - \tfrac{1}{2}| : \tau \in [0, \epsilon']\} \to 0 \quad \text{as} \quad n \to \infty.$$

The stronger result that $\Sigma l_p'(X_i;\theta_\tau)$ is locally uniformly asymptotically normal under λ_τ will be shown instead.

Because of the i.i.d. nature of the summation, asymptotic normality follows from finite variance which itself follows from assumption (2) and the definition of a local family:

$$E_\tau[l_p'(X;\theta_\tau)]^2 \leqslant E_0^{\frac{1}{2}}[l_p'(X;\theta_\tau)]^4 E_0^{\frac{1}{2}}[L_\tau^2] < \infty.$$

The uniform approach to normality is then guaranteed by the Berry–Esseen theorem (see, for example, Feller 1967, p. 542) if it is demonstrated that

$$\frac{E_\tau^2[|l_p'(X;\theta_\tau)|]^3}{E_\tau^3[l_p'(X;\theta_\tau)]^2} \tag{F 1}$$

is uniformly bounded for $\tau \in [0, \epsilon']$. The numerator is bounded above by assumption (2):

$$E_\tau^2[|l_p'(X;\theta_\tau|]^3 \leqslant E_0[l_p'(X;\theta_\tau)]^6 E_0[L_\tau]^2.$$

Lebesgue's extended dominated convergence theorem (see, for example, Rao 1973, p. 136), applied to the relation

$$0 \leqslant (l_p'(X;\theta_\tau))^2 L_\tau \leqslant \tfrac{1}{2}(l_p'(X;\theta_\tau))^4 + \tfrac{1}{2}L_\tau^2,$$

assures that the denominator of (F 1) is continuous in τ at $\tau = 0^+$. So, since the limit is $\mathscr{I}_p^3(\theta_0, Q_0) > 0$, the denominator is bounded away from zero for $\tau \in [0, \epsilon']$, some $\epsilon' > 0$. Thus (F 1) is uniformly bounded for $\tau \in [0, \epsilon']$, as required.