The Annals of Statistics
1983, Vol. 11, No. 2, 486–497

EFFICIENCY OF THE CONDITIONAL SCORE IN A MIXTURE SETTING[1]

BY B. G. LINDSAY

The Pennsylvania State University

The conditional score function is found to be generally fully informative concerning a parameter of interest when the conditioning statistic S is sufficient for the nuisance parameter and has an exponential family distribution. Information is here measured by assuming the nuisance parameter to have been generated by an unknown mixing distribution and then computing the minimal Fisher information. The solution depends upon a study of the geometry of centered likelihood ratios within the space of zero-unbiased functions of S. The two-by-two table model is considered in detail.

1. Introduction. Consider the following problem: We observe a sequence of random variables $X_1, X_2, X_3 \cdots$ where the ith observation comes from the two parameter density $f(\cdot; \theta, \phi_i)$. Here the real-valued parameter θ is of interest; the values $\phi_1, \phi_2, \phi_3, \cdots$ are regarded as nuisance parameters. In this setting Neyman and Scott (1948) showed that the usual asymptotic properties of maximum likelihood estimation fail to hold. In particular, $\hat{\theta}$ could fail to be consistent, or, even if it were consistent and asymptotically normal, it could fail to have lowest asymptotic variance among such estimators.

A resolution of this problem which is sometimes appropriate is to model the ϕ's themselves as independent and identically distributed random variables from an arbitrary unknown mixing distribution Q. The resulting marginal density for X_i is the mixture

$$f_Q(x; \theta) = \int f(x; \theta, \phi)\, dQ(\phi). \tag{1.1}$$

In this formulation the sequence $X_1, X_2, X_3, \cdots$ is now an i.i.d. sequence from the two parameter model (θ, Q). Kiefer and Wolfowitz (1956) showed that in this model the aforementioned inconsistency problem with maximum likelihood estimation of θ is alleviated. This theoretically important result seems to have had no practical impact at the time, as the new maximization problem is much more difficult. However, modern computational abilities have enabled recent use of this estimator (Heckman and Singer, 1982).

In this paper we are concerned with measuring the information about θ in the density (1.1). The measure used herein, called minimal Fisher's information, was suggested by Stein (1956) and adapted for the mixture problem by Lindsay (1980). The measure has been further developed by Begun, Hall, Huang, and Wellner (1983).

In an important class of models, there is a natural competitor to the Kiefer-Wolfowitz estimation method which has substantial computational advantages. In these cases there exists a conditional likelihood which seems to carry much of the information about θ. Discussion of the issue as to when the conditional likelihood has all the information about a parameter of interest has been presented from many points of view. This paper considers the Fisher information in i.i.d. replications from model (1.1). Although Basawa's (1981) results appear similar, in his model the entire string of observations $X_1, X_2, X_3, \cdots$ comes from a single realized value ϕ of the mixing distribution rather than from a sequence of realized values $\phi_1, \phi_2, \phi_3, \cdots$.

Before discussing the general nature of the results herein, we further introduce the

Received October 1981; revised October 1982.

[1] This research was partially supported by the National Science Foundation under Grant MCS-8003081.

AMS 1970 *subject classifications.* Primary 62F20; secondary 62G20.

Key words and phrases. Conditional score, Fisher's information, mixture, likelihood ratio.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

models of interest. Suppose that for density $f(\cdot; \theta, \phi)$ there is a minimal sufficient statistic $S = S(\theta)$ for ϕ when θ is fixed. Suppose $\log f$ is differentiable in θ and that $U = D_\theta \log f$. Then the *conditional score function* is defined to be

$$U^c = U - E(U \mid S). \tag{1.2}$$

Its value at θ_0 can also be derived as the θ-derivative of the conditional (given $S(\theta_0)$) log likelihood. We illustrate with two common exponential family structures: if

$$f(x; \theta, \phi) = \exp\{\theta y + \phi s - \kappa(\theta, \phi)\}, \tag{1.3}$$

then $U^c = y - E_\theta(Y \mid S = s)$. If

$$f(x; \theta, \phi) = \exp\{\phi s(\theta) - \kappa(\theta, \phi)\}, \tag{1.4}$$

then $U^c = \phi[S'(\theta) - E_\theta\{S'(\theta) \mid S(\theta)\}]$.

In the model (1.3), and quite generally when S is free of θ, the conditional score is free of the nuisance parameter. This model can arise as a result of paired comparisons, where the treatment effect θ is represented by a difference in natural parameters. Several examples are treated in Section 6. We now review briefly results for conditional score methods which can be found in E. B. Andersen (1973). The conditional score yields directly an estimating equation for θ:

$$\Sigma U_i^c = \Sigma\{Y_i - E_\theta(Y_i \mid S_i)\} = 0,$$

the solution to which is called the conditional maximum likelihood estimator. Under regularity conditions this estimator, normalized, has an unconditional asymptotic variance under the mixture model which is the inverse of the conditional score information $i_c = E(U^c)^2$. If, as will be shown, i_c is the minimal Fisher information for the density, then conditional maximum likelihood estimation is fully efficient.

In the model (1.4), the conditional score depends upon the nuisance parameter but only as a weight, and quite generally the conditional score eliminates potential score bias caused by the estimation of ϕ. Details can be found in Lindsay (1982). In this setting an alternative to the Kiefer-Wolfowitz joint maximization over (θ, Q) would be to estimate Q as $\hat{Q}(\theta)$ from maximization over Q of the marginal density of $(S_1(\theta), \cdots, S_n(\theta))$ for θ fixed, then estimating θ from the conditional score by

$$0 = \Sigma E_{\hat{Q}(\theta)}(\Phi \mid X_i)[S_i'(\theta) - E\{S_i'(\theta) \mid S_i(\theta)\}].$$

The results in this paper concerning i_c suggest that this method is potentially fully efficient.

The method of approach in this paper will be a geometric one. Thinking of conditional expectations as L_2 projections, we see that the conditional score $U^c = U - E(U \mid S)$ is the component of the θ-score U which is orthogonal to the space of S-functions. Full informativeness of U^c will devolve to the issue of determining whether $E(U \mid S)$ is in the subspace of the S-functions generated by the nuisance parameter scores.

The result is that in the natural interior of the parameter space the conditional information i_c quite generally equals the minimal Fisher information. If the null parameter point is on the boundary, then the answer depends on the structure of $E(U \mid S)$, but as discussed in Remark 7.1 a continuity extension of the information to the boundary would make the conditional score everywhere fully informative.

This paper has the following organization: Section 2 introduces directional score functions and minimal Fisher information, then argues their relevance to the measurement of information. Section 3 briefly presents a direct minimization approach to measuring minimal information, then identifies a critical decomposition of the directional score functions. In Section 4, the interior of the parameter space is identified and the informativeness of the conditional score is shown by appeal to the properties of convex sets in topological vector spaces. A famous example, the two-by-two table, is considered in depth in Section 5. The next-to-last section deals with parameter points on the boundary; it is followed in Section 7 by several remarks.

2. Minimal Fisher information. This section defines minimal Fisher information and outlines its relevance. The next two sections treat it in a general fashion, returning to the mixture application in Section 4. The fundamental idea can be found in Stein (1956) with extensions in Lindsay (1980). Let $f(\cdot; \theta, \psi)$ be a two parameter family of densities, θ real-valued, with a cross product parameter space $\theta \times \Psi$. In the mixture application, ψ will become the unknown mixing distribution Q. The problem of estimating θ at the null point $\omega_0 = (\theta_0, \psi_0)$ is at least as difficult as estimating τ at $\tau = 0$ in any one-dimensional parametric subfamily $\omega(\tau) = (\theta_0 + \tau, \psi_\tau)$, where τ is defined in a neighborhood of 0 and ψ_τ picks out an element of Ψ for each value of τ. Given a smooth family $\omega(\cdot)$ we define the *likelihood ratio function*

$$L_\omega(\tau) = L(\omega(\tau), \omega(0)) = f(X; \omega(\tau))/f(X; \omega(0)).$$

With a differentiability assumption we may define the *directional score statistics* corresponding to $\omega(\cdot)$,

$$U_\omega = D_\tau L_\omega|_\tau = 0;$$

and the *information* corresponding to $\omega(\cdot)$ at ω_0,

$$i_\omega(\omega_0) = E_0(U_\omega)^2.$$

The geometric term "directional" score is meant to refer to directions in the space of likelihood ratio functions. For example, if the sample space is $\mathcal{X} = \{x_1, \cdots, x_t\}$, then the function $L_\omega(\tau)$ on $\mathcal{X}$ can be coordinatized as a vector $(L_\omega(\tau)(x_1), \cdots, L_\omega(\tau)(x_t))$. In this case the score statistics $U_\omega = (U_\omega(x_1), \cdots, U_\omega(x_t))$ indicates the direction in Euclidean t-space from which $L_\omega(\tau)$ approaches $L_\omega(0) \equiv 1$.

Further regularity conditions of the Cramér-type are required on the family $\{L_\omega(\tau)\}$ to ensure that the information i_ω can be used in the usual fashion for lower bounds. Typical assumptions allowing the second order interchange of integration and differentiation yield

$$E_0(U_\omega) = 0 \quad \text{and} \quad i_\omega = -E_0\{D_\tau^2 \log L_\omega(\tau)\}|_{\tau=0}.$$

The *minimal Fisher information* $i^*(\omega_0)$ is defined to be the infimum of the informations $i_\omega(\omega_0)$ over all functions $\omega(\cdot)$ for which L_ω satisfies regularity conditions whose specification will occur later in this section.

If $\boldsymbol{\theta}$ is vector-valued, then one may define the minimal Fisher's information in the direction $\boldsymbol{\alpha}$ (a unit vector) to be the information in the least favorable one-dimensional family of the form $\omega(\tau) = (\boldsymbol{\theta}_0 + \tau\boldsymbol{\alpha}, \psi_\tau)$. We note that if ψ is real-vector valued and if

$$I = \begin{bmatrix} I_{\theta\theta} & I_{\theta\psi} \\ I_{\theta\psi}^t & I_{\psi\psi} \end{bmatrix}$$

is the Fisher information matrix for parameters (θ, ψ), then the minimal Fisher information equals the *marginal Fisher information* about θ:

$$i^* = I_{\theta\theta} - I_{\theta\psi} I_{\psi\psi}^{-1} I_{\theta\psi}^t \tag{2.1}$$

(see Stein, 1956; Lindsay, 1980, Section 3.2).

Lindsay (1980) noted that in a rich parametric setting such as the mixture model, one may be faced with unusual boundary problems in that there will exist smooth one-sided sequences of likelihood ratios $L_\omega(\tau)$, for $\tau > 0$ (or $\tau < 0$), whose analytic extension to $\tau < 0$ (or $\tau > 0$) define true likelihood ratios which do not, however, come from the parameter space. This one-sidedness arises naturally in the mixture setting from such families as $\{\omega(\tau) = (\theta_0 + \tau, (1 - \tau)Q_0 + \tau P), 0 \le \tau \le 1\}$.

Although one might define the smooth interior of the parameter space based on the extendability of the likelihood ratios, in an information sense it is more productive to focus on the possible directional score functions; that is, the sets of directions from which $L_\omega(\tau)$ may approach $L_\omega(0) = 1$. These sets may differ depending on whether $\tau \downarrow 0$ or $\tau \uparrow 0$.

Correspondingly, we define an *upper directional score* $U_{\omega+}$ to be the right hand derivative of $L_{\omega+}(\tau)$ at $\tau = 0$ for a smooth one-sided parameterization $\{\omega^+(\tau) = (\theta_0 + \tau, \psi_\tau) : \tau \geq 0\}$; a *lower directional score* $U_{\omega-}$ corresponds to a left-sided parameterization $\{\omega^-(\tau) = (\theta_0 + \tau, \psi_\tau) : \tau \leq 0\}$. The corresponding informations are $i_{\omega+}$ and $i_{\omega-}$. (Extensions to dim $\theta > 1$ are treated in Remark 7.2.) We will say that $\omega_0 = (\theta_0, \psi_0)$ is in the *symmetric score interior* of the parameter space if the set of upper scores with finite information and the set of lower scores with finite information are nonempty and equivalent in the sense that each is dense in the other with respect to the L_2-metric $\{E_0(g - h)^2\}^{1/2}$. A *symmetric score boundary point* fails to have this property. The scores are the directions from which $L_\omega(\tau)$ approaches 1; at a symmetric score interior point the set of directions are independent of τ's sign. In an undimensional problem, score symmetry corresponds to the equality of the left and right derivatives of the log likelihood.

We define the *upper* and *lower* minimal Fisher informations to be

$$i^{*+} = \inf\{i_{\omega+}(\omega_0)\}, \qquad i^{*-} = \inf\{i_{\omega-}(\omega_0)\},$$

where the infima are taken over one-sided families of likelihood ratios satisfying the following regularity conditions:

$$(1)\ E_0\{L_\omega(\tau)\} = 1, \qquad (2)\ D_\tau E_0\{L_\omega(\tau)\}\big|_{\tau=0^\pm} = E_0(U_\omega^\pm) = 0, \qquad (3)\ D_\tau^2 E_0\{L_\omega(\tau)\}_{\tau=0^\pm} = 0.$$

We define

$$i^{**} = \min(i^{*+}, i^{*-})$$

to be the modified minimal Fisher information. By working with i^{**} rather than i^*, we will avoid the need to consider any but one-sided approaches to the null point ω_0.

A question remains as to whether i^{**} is relevant as a measure of information. The answer as provided by Lindsay (1980) is yes on several counts. The inverse information $(i^{**})^{-1}$ provides a Cramer-Rao type lower bound for unbiased estimation of θ. It also provides a lower bound for the asymptotic variance of those consistent asymptotically normal estimators $\{T_n\}$ which are *smooth* in the sense that they are uniformly median unbiased. That is, $P_\omega[T_n < \theta(\omega)] \to \frac{1}{2}$ uniformly for ω in a neighborhood of ω_0. This is, of course, a weaker condition than uniform approach to asymptotic normality.

In effect one can generally use i^{**} just as one would use Fisher information for generating lower bounds. Of course the importance of any such bound largely lies in its attainment by some estimator. In the problems here discussed there is a conditional likelihood score $U^c = U - E[U \mid S]$ with information $i_c = E_0(U^c)^2$, so that when in Section 4 it is demonstrated that $i_c = i^{**}$, we will have asymptotic lower bounds which are generally attained by the conditional maximum likelihood estimators.

The most obvious and useful applications of this approach arise from assuming that S has an exponential family density in ϕ for each fixed θ_0. Looking ahead to Corollary 4.4 and Theorem 5.1, we find that the efficiency of the conditional approach is one if the null mixing distribution Q_0 is sufficiently "diffuse." If S has an absolutely continuous distribution, this would mean any Q_0 which is not concentrated on a topologically discrete set. If S has a finite number $K + 1$ of points of positive mass, diffuseness means that Q_0 has more than $K/2$ points of support. Corollary 5.2 applies this result to the two-by-two table. In Section 6, consideration is given to the problem's nature and solution when Q_0 does not satisfy the diffuseness criterion.

3. Minimization and decomposition. We first describe in outline a direct minimization approach to finding minimal Fisher's information. Utilizing the fact that the informations $i_{\omega\pm}$ are the second derivatives at $\tau = 0^\pm$ of $E_0[\log L_{\omega\pm}(\tau)]$, one may be able to generate "least favorable" directional scores by defining for each τ

$$\omega^\pm{}_{LF}(\tau) = (\theta_0 + \tau, \psi_\tau^*)$$

where ψ_τ^* minimizes $-E_0[\log L(\theta_0 + \tau, \psi; \theta_0, \psi_0)]$ over ψ in Ψ. Then the minimal Fisher

informations are

$$i^{*\pm} = i_{\omega\pm LF} = E_{\omega_0}(U_{\omega\pm LF})^2.$$

Although this approach can be generally useful to find minimal information and indeed generates the correct solution for the mixture problem, the geometric arguments to follow provide a greater insight.

We first decompose the scores U_ω into a θ-component and a ψ-component, which yields a geometric structure for identifying whether the conditional score U^c can be found as a directional score U_ω.

Let $\omega(\tau) = (\theta_0, \psi_\tau)$ be defined for $\tau \geq 0$, and let

$$V_\omega = \frac{d}{d\tau} L(\omega(t), \omega_0)\,|_{\tau=0^+}.$$

The upper and lower directional scores corresponding to $\omega^+(\tau) = (\theta_0 + \tau, \psi_\tau)$, $\tau \geq 0$, and $\omega^-(\tau) = (\theta_0 + \tau, \psi_{-\tau})$, $\tau \leq 0$, are

$$U_{\omega+} = U + V_\omega, \qquad U_{\omega-} = U - V_\omega \tag{3.1}$$

respectively. This gives i^{*+} and i^{*-} as the infima of $E_0(U + V_\omega)^2$ and $E_0(U - V_\omega)^2$ over the class of ψ-scores V_ω. Under regularity assumptions, the functions V_ω lie in the space of functions of the minimal ϕ-sufficient statistics S which satisfy $E_0(V) = 0$. Another element of this space is $E[U \mid S]$, which is the projection of U onto that space. Since it minimizes $E_0(U - V)^2$ over V in that space, it follows that if there exists $\omega(\cdot)$ such that $\pm V_\omega$ is arbitrarily close to $E(U \mid S)$, then $i_c = E_0\{U - E(U \mid S)\}^2 = i^{*\pm}$. Thus the problem devolves into a geometric one involving the Hilbert space of zero-unbiased functions of S with covariance inner product.

4. The geometry of the likelihood ratios. For the mixture model, with the mixing distribution Q playing the role of nuisance parameter ψ, a large class of parametric families are generated at null (θ_0, Q_0) by the function

$$\omega(\tau) = (\theta_0, (1 - c\tau)Q_0 + c\tau P), \qquad c \geq 0, \qquad 0 \leq c\tau \leq 1, \tag{4.1}$$

which in correspondence with (3.1) yields upper and lower scores of the form, for $c \geq 0$,

$$U_{\omega\pm} = U \pm c[L\{(\theta_0, P), (\theta_0, Q_0)\} - 1].$$

For fixed null point $\omega_0 = (\theta_0, Q_0)$, with $\delta(\phi)$ denoting point mass at ϕ, define the *centered likelihood ratios* to be

$$V(\phi) = L\{(\theta_0, \delta(\phi)), \omega_0\} - 1$$

and

$$V(Q) = L\{(\theta_0, Q), \omega_0\} - 1.$$

For the following treatment, we treat the space of finite variance (under ω_0) functions of S as an L_2 – space with covariance inner product. The subspace $\mathscr{L}$ is defined to be the linear subspace consisting of finite variance functions of S with zero mean (ω_0). We assume that $V(\phi)$ is in $\mathscr{L}$ for every ϕ, hence so are the $V(Q)$ when they have finite variance, which will certainly be true when Q has finite support. Let $\mathscr{H}$ be the convex hull in $\mathscr{L}$ of the set $\{V(\phi) : \phi \in \Phi\}$. The closure of $\mathscr{H}$, written cl ($\mathscr{H}$), equals the closure of the set $\{V(Q) \in L_2 : Q \text{ a mixing distribution}\}$. The closure of the set of positive rays from 0 through points in $\mathscr{H}$, here written

$$\mathscr{C} = \text{cl}\{cq \in \mathscr{L} : c \geq 0, q \in \mathscr{H}\}, \tag{4.2}$$

is the closed convex cone with apex 0 which represents all possible mixture scores generated by (4.1), plus their limit points. The point $0 = V(Q_0)$ is a *support point* of $\mathscr{C}$ if there exists

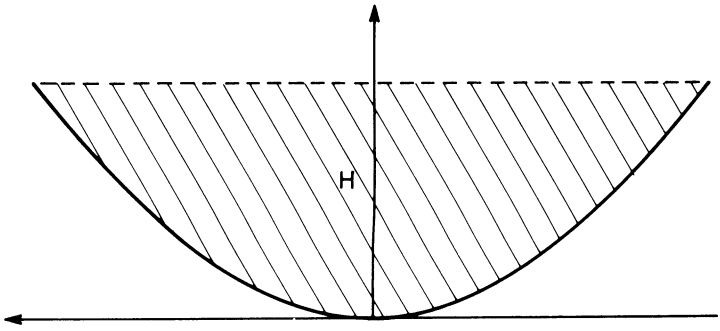

FIG. 1. *The function $V(\phi)$ traces out the bowl shape. The set $\mathscr{H}$ consists of the bowl and its interior. The set $\mathscr{C}$ is the half plane containing $\mathscr{H}$.*

a function $W \in \mathscr{L}$, $W \not\equiv 0$, such that $E_0(WH) \leq 0$ for all $H \in \mathscr{C}$. An equivalent requirement is that $E_0\{WV(\phi)\} \leq 0$ for all $\phi \in \Phi$. In Euclidean space, the support points of a closed convex set are those which have support hyperplanes to the set passing through them and so are equivalent to the usual boundary points. The following theorem is a generalization of the finite dimensional result that the union of the rays from a point **p** of a convex set through other points of the set are either the whole space, if **p** is interior to the set, or contained in a half space, if **p** is a boundary point.

THEOREM 4.1. (Klee, 1969, page 244). *If V is a point of a convex set $\mathscr{C}$ in a locally convex space E, then V is a support point of $\mathscr{C}$ if and only if the union of all rays from V through the various points of $\mathscr{C}$ fails to be dense in E.*

COROLLARY 4.2. *If 0 is not a support point of the convex set $\mathscr{C}$ defined in* (4.2), *then (θ_0, Q_0) is a symmetric score interior point of the mixture parameter space and*

$$i^{*-} = i^{*+} = i^{**} = i_c. \tag{4.3}$$

PROOF. Note that $-\mathscr{H}$ and $-\mathscr{C}$ are the convex hull and closed convex cone corresponding to the lower score functions $-V(\phi)$, and that Theorem 4.1 holds for $-\mathscr{C}$ also. Clearly 0 is a support point for one of $\mathscr{C}$ or $-\mathscr{C}$ if and only if it is a support point for both. If it is not a support point, then the limit sets of upper and lower scores, being of the form $\{U + V: V \in \mathscr{C}\}$ and $\{U + V: V \in -\mathscr{C}\}$, both equal $\{U + V: V \in \mathscr{L}\}$, and so (θ_0, Q_0) is a symmetric score interior point of the mixture parameter space. It also follows that $E(U \mid S)$ is in $\pm\mathscr{C}$, so (4.3) holds.

A model of low dimension for which the ideas may be pictorially represented is the paired Bernoulli model. Let Y and Z be independent $0 - 1$ random variables with probabilities of success

$$p_y = \exp(\theta + \phi)/\{1 + \exp(\theta + \phi)\}$$

(4.4) and

$$p_z = \exp(\phi)/\{1 + \exp(\phi)\}$$

respectively. This is a model of form (1.3), with $S = Y + Z$ as a complete and sufficient statistic for ϕ when θ is fixed. We may graphically represent a function V in $\mathscr{L}$ as a three-vector with coordinates $(V(0), V(1), V(2))$. The constraint $E_0\{V(S)\} = 0 = \sum_{s=0}^{2} P_0(S = s)V(s)$ defines the two-dimensional space $\mathscr{L}$ pictured in Figure 1, in this case with $\omega_0 = (\theta_0, Q_0) = (0, \delta(0))$. The set $\mathscr{H}$ is the convex hull of the bowl-shaped curve $\{V(\phi): -\infty < \phi < \infty\}$. The cone $\mathscr{C}$ is the closed half-space containing $\mathscr{H}$. Its boundary consists of constant multiples of the ϕ-score function $V' = D_\phi V(\phi) \mid_{\omega_0} = D_\phi \log f(X; \theta, \phi) \mid_{\omega_0}$. In

Figure 1, it can be seen that $V(\omega_0) = 0$ is a support point. We note, however, that if Q_0 were a two point mixture, say $\alpha\delta(\phi_0) + (1 - \alpha)\delta(\phi_1)$, the picture would differ in that the point $V(\omega_0) = 0$ of $\mathscr{L}$ would be on the line joining the vectors $V(\phi_0)$ and $V(\phi_1)$ and so be in the interior of $\mathscr{H}$, thus making $\mathscr{C} = \mathscr{L}$. Thus Corollary 4.2 applies to two point mixtures in this case. This result will be generalized to larger sample sizes in Section 5, after developing some analytic techniques to identify null points ω_0 with $\mathscr{C}(\omega_0) = \mathscr{L}(\omega_0)$.

From Corollary 4.2 we see that the critical issue is to determine if 0 is a support point of $\mathscr{C}$. This may be replaced with a probabilistic criterion as follows.

THEOREM 4.3. *If the statement*

$$E_{(\theta_0,\phi)}(W) = 0 \quad \text{a.e.} \quad [Q_0]$$

implies that $W = 0$, *then* $\mathscr{C}(\omega_0) = \mathscr{L}(\omega_0)$.

PROOF. Suppose that $\mathscr{C} \neq \mathscr{L}$, so that by Theorem 4.1 there exists $W \neq 0$ such that

$$E_0\{WV(\phi)\} \leq 0 \quad \text{for all} \quad \phi \in \Phi.$$

When we integrate the left side $dQ_0(\phi)$, we get $E_0\{WV(Q_0)\} = E_0(W \cdot 0) = 0$. It follows that $E_{\theta_0,\phi}(W) = 0$ a.e. $[Q_0]$.

We conclude this section with a corollary which indicates that in an exponential family setting, if the true mixing distribution is sufficiently rich, then the conditional score is fully informative. In the next section this result will be amplified for a specific example.

COROLLARY 4.4. *Suppose for fixed* θ_0 *the statistic* S *has a univariate exponential family density with natural parameter* ϕ. *If the mixing distribution* Q_0 *is not a discrete measure with topologically discrete support points, then the conclusions of Corollary* 4.2 *hold.*

PROOF. The function $E_{\theta,\phi}\{W(S)\}$ is analytic in ϕ and so when $W \neq 0$ it can have only a finite number of zeroes in any interval. Thus it can be zero almost everywhere Q_0 only if Q_0 is a discrete measure with topologically discrete support points.

5. The two-by-two table. A classic problem asks if the marginal totals of a 2×2 table are noninformative for the log odds ratio of the table. The model is as follows: let (Y, Z) be independent binomial random variables with respective parameters (m, p_y) and (n, p_z) with p_y and p_z as in (4.4). The parameter of interest is the log odds ratio θ. The question, as posed here, has the following expression. Given a sequence of pairs (Y_i, Z_i), $i = 1, \cdots, N$ generated from a common θ but with ϕ_i varying as if generated by a random sample from an unknown mixing distribution Q, does the conditional distribution of $\{Y_i; i = 1, \cdots, N\}$ given the complete and sufficient statistics $\{S_i = Y_i + Z_i, i = 1, \cdots, N\}$ contain all the information about θ? The answer is that if the true mixture Q_0 has more than $(m + n)/2$ points of support in $\mathbb{R}$, then the null point is in the symmetric score interior and the conditional information is both the upper and the lower minimal Fisher information. On the other hand, if Q_0 has $(m + n)/2$ points of support or fewer, then, as will be seen in Section 6, exact computation of the minimal Fisher informations $i^{*\pm}$ is not elementary. However, see Remark 7.1.

THEOREM 5.1. *Suppose that statistic* S *has an exponential family density under parameter* θ *with* $K + 1$ *discrete points of support. If* $W = W(S)$ *satisfies* $E_\phi(W) \leq 0$ *and, in addition,*

$$E_\phi(W) = 0 \tag{5.1}$$

for more than $K/2$ distinct points ϕ in the interior of the parameter space, then $W = 0$ with probability one.

PROOF. Let $\{s_0, \cdots, s_K\}$ be the support points of the exponential family density, when expressed in the canonical form $\exp\{\phi s - \kappa(\phi)\}$. We may write

$$E_\phi(W) = \sum_{t=0}^{K} w(s_t)\exp\{\phi s_t - \kappa(\phi)\}$$

which has the same sign and zeros as

$$H(\phi) = \sum w(s_t)\exp(\phi s_t).$$

It can easily be shown that if $W \not\equiv 0$ then any such function has at most K zeroes counting multiplicities. We cite, for example, Karlin and Studden (1966, page 10). The points ϕ satisfying (5.1) are maxima and so are even order zeros, leading to the conclusion of the theorem.

COROLLARY 5.2. *In the two-by-two table model, if Q_0 has at least $(m + n + 1)/2$ points of support, then $\omega_0 = (\theta_0, Q_0)$ is symmetric score interior point with*

$$i^{*+} = i^{*-} = i^{**} = i_c.$$

PROOF. In this example $S = Y + Z$ has, for θ fixed, an exponential family density of the form

$$P\{S = s; (\theta, \phi)\} = C(s)\, d(\theta, \phi)\exp(\phi s).$$

Theorem 5.1 together with Theorem 4.3 gives the result.

6. The symmetric score boundary. Although the symmetric score interior points (θ_0, Q_0) must have a symmetry of information, the same situation will arise if $E_0(U \mid S)$ falls in $\mathscr{C} \cap -\mathscr{C}$, in which case we still have $i^{**} = i^{*+} = i^{*-} = i_c$.

The objective of this section is to identify when this occurs and to give several examples. We also give an example of a point of information asymmetry. Consider the one point null mixing distribution $Q_0 = \delta(\phi_0)$. Suppose that $V(\phi)$ is differentiable (two-sided) in ϕ at ϕ_0, with a derivative V' which is in $\mathscr{L}$. Then clearly V' is in $\mathscr{C} \cap -\mathscr{C}$. More generally, if $Q_0 = \Sigma\pi_t\delta(\phi_t)$ is a k-point mixing distribution, the $2k - 1$ score functions corresponding to the parameters $(\pi_1, \cdots, \pi_{k-1}, \phi_1, \cdots, \phi_k)$ are all in $\mathscr{C} \cap -\mathscr{C}$. This implies that if $E_0(U \mid S)$ is expressible as a linear combination of the parametric score functions, the conditional score is best from above and from below. In this case the marginal Fisher information about θ (equation (2.1)) is also the conditional information.

The atomic mixing distributions $\delta(\phi)$ would appear to be the most likely candidates for finding asymmetry in the information, being in some sense the extreme points of the space of mixing distributions. The following theorem formalizes this intuition.

THEOREM 6.1. *Suppose $E_0\{U(\omega_0) \mid S\} \in \mathscr{C}(\omega_0)$ for every ω_0 of the form $(\theta_0, \delta(\phi))$. If $Q_0 = \Sigma\pi_j\delta(\phi_j)$ is a finite point mixing distribution, then for $\omega_0^* = (\theta_0, Q_0)$ we have*

$$E_0\{U(\omega_0^*) \mid S\} \in \mathscr{C}(\omega_0^*).$$

PROOF. Suppose first $E_0\{U(\theta_0, \delta(\theta_j)) \mid S\}$ has an explicit representation for each j

$$c(\phi_j)\left\{\frac{\int f(x; \theta_0, \tilde{\phi})\, dQ(\tilde{\phi} \mid \phi_j)}{f(x; \theta_0, \phi_j)} - 1\right\}$$

as an element of $\mathscr{C}(\theta_0, \delta(\phi_j))$. We note that for a finite range of j we may choose $Q(\cdot \mid \phi_j)$ in such a way that $c(\phi_j) \equiv c$, some constant greater than zero, and we do so.

Since

$$U(\omega_0^*) = \frac{\int U(\theta_0, \phi) f(x; \theta_0, \phi)\, dQ_0(\phi)}{\int f(x; \theta_0, \phi)\, dQ_0(\phi)}$$

we have

$$E_0\{U(\omega_0^*) \mid S\} = \frac{\int E_0\{U(\theta_0, \phi) \mid S\} f(x; \theta_0, \phi)\, dQ_0(\phi)}{\int f(x; \theta_0, \phi)\, dQ_0(\phi)}$$

$$= \frac{c \int \left\{ \int f(x; \theta_0, \tilde{\phi})\, dQ(\tilde{\phi} \mid \phi) - f(x; \theta_0, \phi) \right\} dQ_0(\phi)}{\int f(x; \theta_0, \phi)\, dQ_0(\phi)}$$

$$= c \left\{ \frac{\int f(x; \theta_0, \tilde{\phi})\, dQ^*(\tilde{\phi})}{\int f(x; \theta_0, \phi)\, dQ_0(\phi)} - 1 \right\},$$

where $Q^*(A) = \int Q(A \mid \phi)\, dQ_0(\phi)$. Thus we have an explicit representation of $E_0\{U(\omega_0^*) \mid S\}$ as an element of $\mathscr{C}(\omega_0^*)$.

To complete the proof, we note that even if explicit representations are impossible we can find a sequence $(c_m, Q_m(\cdot \mid \phi))$ which generates points in $\mathscr{C}(\phi_0, \delta(\phi))$ approaching $E_0\{U(\theta_0, \phi) \mid S\}$. They will yield a corresponding sequence (c_m, Q_m^*) which generates points in $\mathscr{C}(\omega_0^*)$ approaching $E_0\{U(\omega_0^*) \mid S\}$ in the $L_2(\omega_0^*)$ sense.

COROLLARY 6.2. *If X has an exponential family density of the form*

$$f(x; \theta, \phi) = \exp\{\theta y + \phi s - \kappa(\theta, \phi)\} \tag{6.1}$$

and if $E_0(Y \mid S)$ is linear in S, then $E\{U(\omega_0) \mid S\}$ is in $\mathscr{C}(\omega_0) \cap -\mathscr{C}(\omega_0)$ for every $\omega_0 = (\theta_0, Q_0)$ where Q_0 is a finite mixing distribution, and

$$i^{**}(\omega_0) = i^{*+}(\omega_0) = i^{*-}(\omega_0) = i_c(\omega_0).$$

PROOF. In this model with $\omega_0 = (\theta_0, \delta(\phi))$ the projected θ-score is $E_0(U \mid S) = E_0(Y \mid S) - E_0(S)$ and the ϕ-score is $V' = S - E_0(S)$. Thus $E_0(U \mid S)$ is in $\mathscr{C}(\theta_0, \delta(\phi)) \cap -\mathscr{C}(\theta_0, \delta(\phi))$ if $E_0(Y \mid S)$ is linear in S. We extend to finite point distributions Q_0 by Theorem 6.1.

This gives us now a simple linear criterion for testing when there is no inherent unconditional information loss in conditioning regardless of the unknown mixing distribution. Two of the following examples have the linear structure, two do not.

EXAMPLE 6.1. If (Y, Z) are jointly independent normal with means $(\theta + \phi, \phi)$ and variances (1, 1), then $S = Y + Z$ and linearity holds:

$$E_0(Y \mid S) = \tfrac{1}{2}(S + \theta).$$

EXAMPLE 6.2. If (Y, Z) are jointly independent Poisson variables with means $(\exp(\theta + \phi), \exp \phi)$, then $S = Y + Z$ and linearity holds:

$$E_0(Y \mid S) = Se^{\theta}/(1 + e^{\theta}).$$

EXAMPLE 6.3. If (Y, Z) are jointly independent exponential variables with means $((\theta + \phi)^{-1}, \phi^{-1})$, then $S = Y + Z$ and

$$E_0[Y \mid S] = \begin{cases} S/2 & \text{for} \quad \theta = 0, \\ \{\exp(\theta S) - 1 - \theta S\}/[\theta\{\exp(\theta S) - 1\}] & \text{for} \quad \theta \neq 0. \end{cases}$$

EXAMPLE 6.4. In the paired binomial example of Section 5, the mean is a complicated nonlinear function of S except where $\theta = 0$, in which case $E_0(Y \mid S) = mS/(m + n)$. See Harkness (1965).

Now Examples 6.3 and 6.4 are candidates to be points of score asymmetry for which the conditional information is not both the lower and the upper information. We treat Example 6.4 further, extending slightly our geometric techniques for the purpose of illustration.

If $E_0(U \mid S)$ is not linear in V' at null point $\omega_0 = (\theta_0, \delta(\phi))$, we may next reduce the problem by one dimension by considering $\mathscr{L}_0$, the subspace of $\mathscr{L}$ generated by those functions of S which are uncorrelated with V'. In the exponential family model (6.1), this means functions g such that $\mathrm{Cov}\{g(S), S\} = 0$. Note that $V(Q)$ is in $\mathscr{L}_0$ if and only if $E_Q(S) = E_0(S)$. Define $\mathscr{C}_0$ and $\mathscr{H}_0$ to be the sets corresponding to restrictions of $\mathscr{C}$ and $\mathscr{H}$ to $\mathscr{L}_0$. Define $\rho = E(UV')/\{E(V')\}^2$. then $E(U \mid S) - \rho V'$ is the projection of U onto $\mathscr{L}_0$. If it is in $\mathscr{C}_0$ or $-\mathscr{C}_0$, then $E_0(U \mid S)$ is in $\mathscr{C}$ or $-\mathscr{C}$, and so the conditional score is correspondingly upper or lower fully informative.

EXAMPLE 6.5. Suppose that in the paired binomial model of Section 5, one has $\theta_0 = \log 2$, $Q_0 = \delta(0)$, $m = 2$, and $n = 1$, the sample sizes having been chosen to make $\mathscr{L}_0$ two-dimensional. In Figure 2 it is demonstrated how $\pm\mathscr{H}_0$ sits in $\mathscr{L}_0$, and the ray in $-\mathscr{C}_0$ upon which

$$E(U \mid S) - \rho V' = E(Y \mid S) - \frac{4}{3} - \frac{16}{25}\left(S - \frac{11}{6}\right)$$

may be found is illustrated. The conditional score is lower fully informative but not upper, so ω_0 is a point of score and information asymmetry.

7. Some remarks.

REMARK 7.1. When one has parameter point ω_0 which is in the symmetric score boundary of the parameter space and for which $E(U \mid S)$ is not in $\mathscr{C} \cap -\mathscr{C}$, then, as seen in Example 6.5, the information problem involves a careful analysis of the geometry of $\mathscr{C}_0$. If for some ω_0, $E(U \mid S)$ is not in $\mathscr{C} \cup -\mathscr{C}$, then in theory it may be possible to construct a smooth consistent asymptotically normal estimator $\{T_n\}$ which has a smaller asymptotic variance at ω_0 than the conditional maximum likelihood estimator. We do note, however, that such an improved estimator generally cannot have an asymptotic variance $V_T(\theta, Q)$ which is weak convergence continuous in Q. For example, in the exponential family case (6.1), one can always find symmetric score interior points (θ, P_n) with $P_n \to_w Q_0$. Since the lower bound for $V_T(\theta, P_n)$ is $i_c^{-1} = \{\int E_{\theta,\phi}(U^c)^2\, dP_n\}^{-1}$, the weak continuity of i_c ensures that a lower bound at ω_0 for continuous variances V_T is $i_c^{-1} = \{\int E_{\theta,\phi}(U^c)^2\, dQ_0\}^{-1}$. If V_T is discontinuous, we note it may be difficult to estimate sensibly. Thus it is held that a detailed analysis of the cones is usually unnecessary, with the conditional score being the only generally satisfactory base for inference in the mixture setting.

REMARK 7.2. If the parameter of interest $\boldsymbol{\theta}$ is of dimension higher than one, then one should consider one dimensional families of the form $\omega(\tau) = (\boldsymbol{\theta}_0 + \tau\boldsymbol{\alpha}, \psi_\tau)$ for each unit vector $\boldsymbol{\alpha}$. The decomposed scores (3.1) are now of the form

$$\boldsymbol{\alpha} \cdot \mathbf{U} + V_\omega$$

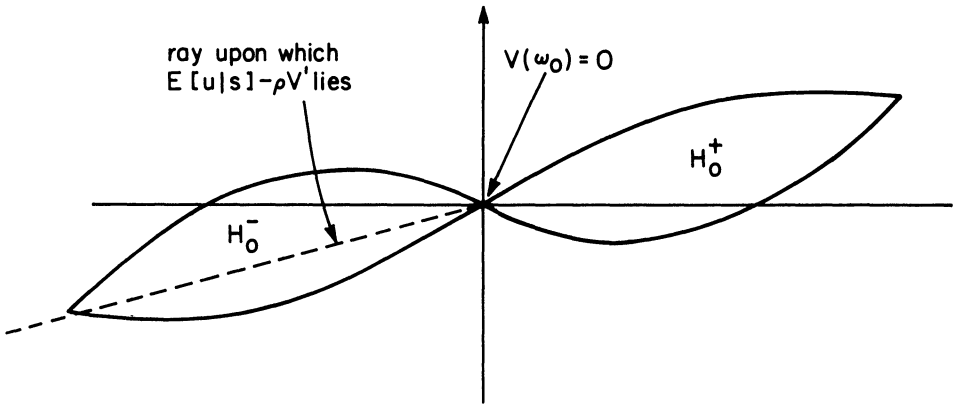

FIG. 2. *Functions of S which are orthogonal (E_0) to 1 and V′ lie in this two dimensional space. The function $E(U|S) - \rho V'$ lies in the cone generated by $\mathscr{H}_0^-$.*

where V_ω is generated by the likelihood ratios of $\{\omega(\tau) = (\boldsymbol{\theta}_0, \psi_\tau), \tau \geq 0\}$. As in the one dimensional case, the scores are symmetric at ω_0 if for each V_ω there exists ω^* with $V_\omega = -V_{\omega^*}^*$. Once again, the full informativeness of the conditional likelihood depends on whether $E(U|S)$ can be expressed as V_ω, with this result guaranteed in the mixture setting if the cone of centered likelihood ratios $\mathscr{C}$ is dense in the space $\mathscr{L}$ of zero-mean functions of the sufficient statistic S. Thus the conclusions in this setting match those of the unidimensional case.

REMARK 7.3. Two likelihood models are possible for a sequence of observations $X_1, \cdots, X_n$ from the mixture density. One is the mixture likelihood:

$$L(\theta, Q) = \Pi f(x_i; \theta, Q).$$

The other conditions upon the realized values $\phi_1, \cdots, \phi_n$ of the mixture Q, giving a likelihood of the Neyman-Scott type

$$L(\theta, \phi_1, \cdots, \phi_n) = \Pi f(X_i, \theta, \phi_i).$$

The mean square error of an estimator T_n under the mixture model is simply the average of its mean square error over samples $\phi_1, \cdots, \phi_n$ from the mixture Q. As such, one may conclude that an estimator which performs optimally for every sequence will certainly do as well in the mixture model. On the other hand, an estimator which is best in the mixture model may be far from the best along individual realized sequences; however, those estimators which beat it along one realized sequence must lose to it along other sequences in order for the mixture estimator to be superior on the average. More comments on this problem may be found in Lindsay (1980, 1982).

REFERENCES

ANDERSEN, E. B. (1973). *Conditional Inference and Models for Measuring.* Mentalhygienisk Forlag, Copenhagen.

BASAWA, I. V. (1981). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential families. *Biometrika* **68** 515–523.

BEGUN, J. M., HALL, W. J., HUANG, W.-M., and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

HARKNESS, W. L. (1965). Properties of the extended hypergeometric distribution. *Ann. Math. Statist.* **36** 938–945.

HECKMAN, J. J. and SINGER, B. (1982). Population heterogeneity in demographic models. *Multidimensional Mathematical Demographics*, ed. by K. C. Land and A. Rogers, Academic, New York.

KARLIN, S. and STUDDEN, W. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics.* Interscience, New York.

KLEE, V. (1969). Separation and support properties of convex sets—a survey. *Control Theory and the Calculus of Variations*, ed. by A. V. Balakrishnan, 235–303.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

LINDSAY, B. G. (1980). Nuisance parameter, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296** 639–665.

LINDSAY, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69** 503–512.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika* **16** 1–32.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–195.

DEPARTMENT OF STATISTICS
THE PENNSYLVANIA STATE UNIVERSITY
219 POND LABORATORY
UNIVERSITY PARK, PENNSYLVANIA 16802