The Annals of Statistics
1985, Vol. 13, No. 3, 914–931

USING EMPIRICAL PARTIALLY BAYES INFERENCE FOR INCREASED EFFICIENCY[1]

BY BRUCE G. LINDSAY

The Pennsylvania State University

Empirical partially Bayes methods are considered as a means of improving efficiency in a class of problems in which the number of nuisance parameters increases to infinity. In the method used, the parameter of interest is estimated in an asymptotically unbiased way while James-Stein shrinkage is applied to the nuisance parameter estimates. When the shrinkage estimators are carefully chosen, this yields estimators generally more efficient than maximum likelihood. In the models considered, the conditional structure imposed allows construction of a simple estimator which is broadly consistent and efficient.

1. Introduction. The problem, in its simplest form, is to estimate θ, a scalar parameter, in the presence of a vector of nuisance parameters $\phi_1, \phi_2, \cdots, \phi_p$, where we presume there is relatively little information about the individual ϕ_i, but there is substantial information about θ. In particular, suppose there are p strata in the data, and that the vector X_i of observations from the ith stratum has a density $f_i(x_i; \theta, \phi_i)$, where f_i is a known density. Neyman and Scott (1948), in a classic paper on the limitations of likelihood methods, showed that the usual asymptotic results concerning the consistency and asymptotic efficiency of the maximum likelihood estimator of θ, could fail utterly when the number of nuisance parameters $p \to \infty$. This paper will show that the loss in efficiency is closely related to the James-Stein effect. Further background will be presented in Section 2. We first offer a simple example as a reference point for the discussion.

EXAMPLE A. Let X_i and Y_i be independent exponential random variables with means $1/\theta\phi_i$ and $1/\phi_i$, respectively, so that the parameter of interest, θ, is the ratio of hazards, which is constant through the strata. We can derive the maximum likelihood estimator as follows. The score function in θ for the ith stratum is

$$U_i = U_i(\theta, \phi_i) = D_\theta \log f_i(X_i, Y_i; \theta, \phi_i) = (1 - \theta\phi X)/\theta.$$

Given data $(x, y), \cdots, (x_p, y_p)$ one can maximize the ith stratum likelihood over ϕ_i for fixed θ to obtain

$$\hat{\phi}_i(\theta) = 2/(\theta x_i + y_i).$$

Received June 1984; revised February 1985.

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

AMS 1980 *subject classifications.* 62F10, 62F12.

Key words and phrases. Conditional score, empirical Bayes, James-Stein estimators, weighted means, nuisance parameters.

This yields the maximum likelihood estimator for θ as the solution to

$$\sum_{i=1}^{p} (y_i - \theta x_i)/\theta(\theta x_i + y_i) = 0. \tag{1.1}$$

If one proceeds through a conventional asymptotic analysis, using the appropriate element of the inverse of the Fisher's information matrix, one arrives at $2\theta^2$ as the asymptotic variance of this estimator as $p \to \infty$. (The inverse of this asymptotic variance term is sometimes called the marginal Fisher's information about θ in the presence of the nuisance parameters.)

Although (as soon will be apparent) this estimator is consistent and asymptotically normal as $p \to \infty$, this is not the correct asymptotic variance. One can observe from the estimating equation that the estimator depends solely on the ratios $t_i = y_i/x_i$. In fact, this estimating equation is also the maximum likelihood equation for the marginal distribution of the statistics $t_1, t_2, \cdots, t_p$. These statistics are i.i.d. with a distribution not depending on the ϕ_i. Indeed, these are the invariant statistics of the natural family of scale transformations which leave θ invariant. The usual asymptotic theory does hold here because of the i.i.d. structure and it follows from a Fisher's information calculation that the solution to (1.1) has asymptotic variance $3\theta^2$ rather than $2\theta^2$.

With this example in mind we can discuss the central question of this paper: is the above maximum likelihood estimator of θ fully efficient as $p \to \infty$? The answer given will be no, but it will depend on a rather intricate construction involving several important ideas. First, we will demonstrate that the conditional score function can be used to generate a large class of consistent estimators (Lindsay, 1982). Next, it will be argued that one natural approach to estimating the unknown nuisance parameters in the conditional score is a smooth empirical Bayes one. The empirical Bayes method will be chosen in such a way that the resulting estimator dominates the above m.l.e.; the device used here will be to maximize an empirical version of the asymptotic information.

The methods used here have extensions to other models than Example A. In particular, an exponential family structure will be identified in the next section for which the arguments hold. A second important example, the weighted means problem, will be considered in the last section.

This paper can be contrasted with Kumon and Amari (1984), who treat the above example (their Example 5) and conclude that the m.l.e. is fully efficient. The distinction between the two approaches is that Kumon and Amari restrict attention to a class of estimators they call uniformly informative (their C_2). This approach ignores the information available from an empirical Bayes treatment. Some intuition for this will be given in Section 3, which relates this problem to the James-Stein problem. Indeed, both are problems where invariance considerations lead to inefficient solutions. One key difference is that in the problem posed here, there are obvious repercussions for confidence statements, not just the mean square error of point estimates.

2. Background. There have been two important methodologies developed for inference in the presence of many nuisance parameters. One, initiated by Kiefer and Wolfowitz (1956), is to model the sequence $\phi_1, \phi_2, \cdots, \phi_p$ as an i.i.d.

sequence of random variables from an unknown distribution Q, and to use maximum likelihood estimation for the pair (θ, Q). Although they showed that this approach led much more generally to consistent estimates of θ, this method apparently was not applied until recently, when computational techniques (e.g., Laird, 1978) and characterizations of the solution (Lindsay, 1983a,c) made this feasible (e.g., Heckman and Singer, 1982).

Another track is to use a marginal or conditional likelihood function which depends solely on θ. This method is possible only for likelihoods which have a factorizable structure. A key paper in this vein is by Kalbfleisch and Sprott (1970), with further development of the conditional approach by Andersen (1973). Cox (1975a) extended this methodology to the partial likelihood.

This paper will in fact blend these two approaches, using conditioning to generate a class of consistent estimating equations, then using empirical Bayes to increase efficiency. We will use parametric empirical Bayes methods to model the unknown Q rather than nonparametric empirical Bayes methods (as in Kiefer and Wolfowitz) because in this manner simple estimators can be developed which will recoup much of the loss due to the unknown $\phi_1, \cdots, \phi_p$. We first present some of the relevant background on estimating equations.

Suppose that from data $\mathbf{x}$ we form an estimator by solving for θ in an estimating equation

$$H_p(\mathbf{x}; \theta) = 0.$$

If H_p is itself a sum of independent components, as in (1.1), then one can often argue that a necessary condition for the consistency of the estimator is zero-unbiasedness:

$$E_\theta[H_p(\mathbf{X}; \theta)] = 0. \tag{2.1}$$

By mimicking the derivation of the asymptotic distribution of the maximum likelihood estimator, one arrives at asymptotic normality and a formula for the asymptotic variance, the inverse of which will be called the information in H:

$$i_H = (E[H_p'(\mathbf{X}; \theta)])^2 / E[H_p^2(\mathbf{X}; \theta)]. \tag{2.2}$$

It is clear that zero-unbiasedness of H need not imply consistency, nor except for regularity need i_H be the inverse of an asymptotic variance. However, a simple and elegant path is to use (2.1) as a requirement for estimators and judge them by (2.2). See Godambe (1976, 1980) for these developments. In this paper, (2.1) and (2.2) will be the focal point, but the asymptotic results will be considered the primary motivation and so will receive attention in Section 4. Indeed, the information given for the estimators to be derived will be in effect an asymptotic approximation to (2.2).

Turning now to the conditional score function, it will be presumed throughout the following discussion that for each density $f_i(x_i; \theta, \phi_i)$ there exists a complete and sufficient statistic $S_i(\theta)$ for ϕ_i when θ is fixed. From this statistic one can create a *conditional score function* W_i for θ from the score function $U_i(\theta, \phi_i) = D_\theta \log f_i(X_i; \theta, \phi_i)$ by the operation

$$W_i(\theta, \phi_i) = U_i(\theta, \phi_i) - E_\theta[U_i(\theta, \phi_i) \mid S_i(\theta)].$$

Lindsay (1982) introduced this representation to derive estimating equations and later (1983b) presented these scores from a geometrical viewpoint to calculate efficiencies.

The interpretation of the conditional score W_i depends on whether S_i is truly a function of θ. If it is not, then the conditional score function is the derivative with respect to θ of the conditional density of the data given S_i. Since by sufficiency that density is free of ϕ_i, the function W_i is also. Thus we can estimate θ from the equation

$$\sum W_i(\theta) = 0,$$

obtaining the conditional maximum likelihood estimator. This estimator is quite generally consistent. The asymptotic variance for this estimator is the inverse of the *conditional information*

$$i_c = (1/p) \sum E_i\{W_i^2\}.$$

See Andersen (1973) for a development of the asymptotics and examples.

If S_i does depend on θ, then one can interpret the conditional score at θ_0 as follows: it is the derivative with respect to θ, evaluated at θ_0, of the log conditional density of the data given $S(\theta_0)$. Although the conditional density $p(x \mid s(\theta); \theta)$ is free of ϕ, when θ is fixed at θ_0 in the conditioning statistic, we get $p(x \mid s(\theta_0); \theta, \phi_i)$ where ϕ_i need disappear only when $\theta = \theta_0$. It follows that the conditional score may depend on ϕ_i, but, as shown in Lindsay (1982), the dependence is weakened.

In particular, the example introduced in Section 1 is in a class of models in which the conditional score has a simple structure. Suppose that the density of x has the following exponential family form:

$$f(x; \theta, \phi) = h(x)\exp(\phi s(\theta) - b(\theta, \phi)). \tag{2.3}$$

Notice that $S(\theta)$ is the complete and sufficient statistic for the nuisance parameter. In this case, the conditional score has the form:

$$W(\theta, \phi) = \phi[S'(\theta) - E(S'(\theta) \mid S(\theta))] =_{\text{def}} \phi\Delta.$$

In the setting of the paired exponential example, this conditional score is

$$W(\theta, \phi) = \phi(Y - \theta X)/2\theta. \tag{2.4}$$

From this we see that the ith conditional score, $W_i = \phi_i\Delta_i$, is the product of an unknown weight times a function with zero expectation. In particular, note that if one estimates the weights ϕ_i by maximum likelihood in (2.4) and sums over i, one arrives at the maximum likelihood equation (1.1).

However, the consistency of an estimator formed as the solution to $\sum \tilde{\phi}_i\Delta_i = 0$, with data-dependent weights $\tilde{\phi}_i$, does not depend on using the maximum likelihood weights $\hat{\phi}_i$. For the weights, one could use any sequence of constants, or more generally, use for $\tilde{\phi}_i$ any function $\delta_i(\theta, S_i(\theta))$ and still retain the fundamental zero-unbiased property $E[\delta_i\Delta_i] = 0$, as we have $E[\Delta_i \mid S_i] = 0$.

We are now in a position to provide an overview of the remainder of the paper. The central emphasis will be on models of the form (2.3). At this point we have

identified a class of estimating equations of the form

$$\sum \delta_i \Delta_i = 0 \tag{2.5}$$

with data-dependent weights δ_i. In the next section, it will be shown that efficiency considerations lead one to consider using weights "shrunken" under an empirical Bayes formulation. In Section 4, attention is given to the correct asymptotic distribution of an estimator with "estimated" weights δ_i. It is then shown in Section 5 that by maximizing the information in an estimating equation we can achieve an estimator with high global efficiency. Section 6 presents a special class of "linear" empirical Bayes solutions for which the information can be explicitly calculated. Finally, in Section 7, varying sample sizes and the weighted-means problem are considered.

3. Efficiency considerations. The question of efficiency as p, the number of nuisance parameters, goes to ∞ presents several delicate issues in terms of modelling and interpretation. We first follow Godambe (1976) in using the simple but instructive approach of leaving p fixed and using the information i_H in (2.2) as a means of comparing estimating functions.

If the conditioning statistic S_i is free of θ, then Godambe (1976) showed that the conditional estimating function $\sum W_i(\theta)$ was the best possible in the sense that, for any other unbiased estimating function H, we have $i_c \geq i_H$. That is, it is uniformly maximal-information over the parameter space.

With considerably more effort (Lindsay, 1980, 1983b), it is often possible to show that the conditional maximum likelihood estimator is best asymptotically normal under restrictions on the regularity of the class of estimators and assumptions on the limiting behavior of the sequence $\phi_1, \phi_2, \cdots, \phi_p, \cdots$. Since the conditional estimators are also generally first-order efficient as the information per stratum increases with p held fixed (Liang, 1984), the problem seems well solved to first order.

On the other hand when S_i depends on θ, the constraint of zero-unbiasedness no longer suffices to give a single uniformly maximal-information estimating equation (Lindsay, 1982). In particular, for the exponential family model (2.3) the optimal estimating equation depends on the true (but unknown) sequence of nuisance parameters and is the conditional score

$$W = \textstyle\sum_{i=1}^{p} \phi_i \Delta_i.$$

This is, of course, not typically optimal at nearby sequences. However, its full informativeness suggests that it be used as a building block, say by estimating the weights ϕ_i.

The following heuristic motivation may offer some illumination on why treating the ϕ_i as a sequence of i.i.d. random variables, an empirical Bayes formulation, provides an increase in efficiency. Let us suppose that the density f_i does not depend on i (other than through ϕ_i), as in (2.3). It then seems logical that we should estimate the ϕ_i in U_c with a function $\delta(\theta, S_i(\theta))$ depending on i only through the sufficient statistic for ϕ_i. Indeed, the m.l.e. weights $\hat{\phi}_i$ are such functions. Any function δ otherwise depending on i seems to presuppose special

knowledge about that stratum. More formally, this gives an estimation method invariant under permutation of the subscripts.

Thus let us consider the optimal information in estimating equations of the form

$$W^*(\theta) = \sum_i \delta(\theta, S_i(\theta))\Delta_i = 0.$$

Let Q^* be the empirical probability measure for the true unknown sequence $\phi_1, \cdots, \phi_p$. Noting that $EW^* = 0$, we may compute the information as

$$i^* = \frac{E^2(D_\theta W^*)}{E(W^{*2})} = \frac{(\int E_{\theta,\phi}[D_\theta \delta(S(\theta))\Delta]\, dQ^*(\phi))^2}{(\int E_{\theta,\phi}[\delta(S(\theta))\Delta]^2\, dQ^*(\phi)}. \tag{3.1}$$

Here we are using the fact that $\delta_i\Delta_i$ depends on i only through the data. The key insight is now this: if instead of modelling the observations as being from the density $f(x_1; \theta, \phi_1) \times \cdots \times f(x_p; \theta, \phi_p)$ we had used the i.i.d. density $\int f(x_1; \theta, \phi)\, dQ^*(\phi) \times \cdots \times \int f(x_p; \theta, \phi)\, dQ^*(\phi)$, we would have arrived at the same information. The class of W^* functions has a maximal information element, as the methods of Lindsay (1982) carry over to the mixed density, $\int f(x; \theta, \phi)\, dQ^*(\phi)$; the optimal estimating function in the larger class of all zero-unbiased functions for a density is the conditional score, which can be calculated as

$$W(\theta, Q^*) = \sum_i E_{Q^*}[\phi \mid S_i]\Delta_i.$$

It follows that the optimal weighting function for W^* is

$$\delta^*(\theta, s_i) = E_{Q^*}[\phi \mid s_i]. \tag{3.2}$$

Since Q^* is unknown, we still do not have a completely solved problem. However, if the sequence $\phi_1, \cdots, \phi_p$ were well-behaved, using an empirical estimate of Q^* in δ^* might well improve on estimation over using the m.l.e. weights.

There is a very important analogy here. In James-Stein estimation, if one is estimating means $\mu_1, \cdots, \mu_p$ from independent normal observations $x_1, \cdots, x_p$, if one is using squared error loss, and if one wishes to estimate μ_i uisng $\delta(x_i)$, then the risk is

$$\sum_i E_{\mu_i}[(\delta(X_i) - \mu_i)^2] = p \int E_\mu(\delta(X) - \mu)^2\, dQ^*(\mu),$$

where Q^* is the empirical c.d.f. of $\mu_1, \cdots, \mu_p$. This class of estimates includes the optimal invariant estimator $\delta_0(x) = x$. This risk minimization problem has the well-known Bayes solution

$$\delta(x) = E_{Q^*}(\mu \mid X = x), \tag{3.3}$$

which invites comparison with (3.2). Once again the optimal solution depends on the unknown Q^*. However, it is known that we can profitably estimate (3.3) and reduce risk over $\delta_0(x) = x$. This James-Stein solution has several surprises. One is that $p \geq 3$ is adequate for this improvement; the other is that one can obtain this improvement over δ_0 by simply modelling Q^* as a normal distribution and estimating it; one comes out ahead even if Q^* is not remotely normal. It will be

seen that in our case also that there is value in shrinking, even if the parametric choice for Q^* is quite wrong.

We now adopt as an operating assumption that $\phi_1, \phi_2, \cdots, \phi_p$ are an i.i.d. sequence from an unknown distribution Q, and hence $X_1, \cdots, X_p$ are i.i.d. observations from the mixed density $\int f(x; \theta, \phi)\, dQ(\phi)$. This assumption makes asymptotic calculations straightforward, and we have argued that invariance considerations in the model where $\phi_1, \cdots, \phi_p$ is a nonrandom sequence make this a plausible approach there also; just interpret Q as Q_p^*, the empirical c.d.f. of the sequence. Under an asymptotic formulation where Q_p^* converges to a distribution Q, the asymptotic distribution of an estimator of the form (2.5) will generally be the same as under the i.i.d. formulation above (subject to some technicalities needed to obtain appropriate convergence of the information (3.1), central limit theorems, etc.).

This leaves remaining the question of how well one can do in estimating θ in this setting. As noted above, the best estimating function in terms of information is based on the conditional score $W_i(\theta, Q) = E_Q[\phi \mid S_i]\Delta_i$.

If Q were known, one could use $W(\theta, Q)$ directly for estimation and achieve the asymptotic variance lower bound i_c^{-1}. To illustrate, suppose that in the paired exponential example the unknown Q is gamma (α, λ). Then the asymptotic variance attained by the conditional score estimator is

$$i_c^{-1} = (E[W^2])^{-1} = (\alpha + 3)/(\alpha + 2)2\theta^2.$$

We note that as $\alpha \to 0$ this approaches the upper bound of $3\theta^2$, the asymptotic variance of the m.l.e. On the other hand, for $\alpha \to \infty$ we approach the inverse of the marginal Fisher's information, $2\theta^2$. We conclude there is some hope of improving upon the m.l.e., but the variability in ϕ prevents attainment of the Cramér-Rao lower bound. In fact, for this problem one can explicitly write the information in the presence of mixing as

$$(1/2\theta^2)[1 - \tfrac{1}{6} E\{\mathrm{Var}_Q(\phi \mid S)S^2\}],$$

which explicitly shows the loss due to variation in ϕ.

4. Smooth empirical Bayes. Having adopted the point of view that

$$\bar{W} = (1/p) \sum E_Q(\phi \mid s_i)\Delta_i$$

will be a basis for inference, we next turn to the estimation of the weights $\delta(s) = E_Q(\phi \mid s)$. One possibility, used in the James-Stein setting by Laird (1982) and Leonard (1984), is to estimate Q nonparametrically. Indeed, this appears to be the only approach with hope for attaining the lower bound of Section 3 for every possible Q. On the other hand, this form of estimation has an undeveloped asymptotic theory, so it is difficult to say if $E_{\hat{Q}}(\phi \mid s)$ would converge to the correct weighting function $\delta(s)$ sufficiently fast to attain the lower bound. Here we adopt a rather more elementary point of view. Surely if in James-Stein estimation the arbitrary use of the normal distribution to estimate Q^* yields positive results for every distribution Q^* on μ, one can by choice of a reasonable parametric family $\{Q_\gamma\}$ of distributions on ϕ attain improvement in our setting for every Q.

Hence let $W_i(\theta, \gamma) = E_\gamma[\phi \mid S_i(\theta)]\Delta_i$ for any particular choice $\{Q_\gamma\}$. We consider estimating equations of the form

$$\bar{W}(\theta, \hat{\gamma}(\theta)) = (1/p) \sum W_i(\theta, \hat{\gamma}(\theta)) = 0, \tag{4.1}$$

where $\hat{\gamma}(\theta)$ is an estimator of γ in $\{Q_\gamma\}$, later to be chosen so as to optimize the information in (4.1).

We note that Cox (1975b) used estimated "partially Bayes" weights in a class of normal theory models. This paper differs in extending the distributional class of models via the conditional score and in giving expanded consideration to the appropriate choice of Q_γ and $\hat{\gamma}$. Note also that an important early analysis of the problem of weighting over heterogeneous strata can be found in Yates and Cochran (1938, Section 6).

There are several important considerations in our choice of $\{Q_\gamma\}$ and our estimation of γ:

Consistency: We desire to retain consistency for all reasonable sequences ϕ_1, ϕ_2, ϕ_3, $\cdots$, not just those generated from Q_γ.

Local full efficiency: We desire full efficiency in the estimation of θ when the nuisance parameter sequence is from Q_γ. It should be noted that this is full efficiency in the sense of attaining the information upper bound for the nonparametric Q problem, not for the fully parametric (θ, γ) problem. Estimators which are consistent for the (θ, γ) problem may well be inconsistent in the (θ, Q) problem.

Global high efficiency: We desire high efficiency for arbitrary ϕ sequences and, in particular, if there exists a consistent m.l.e. for the problem, we wish to dominate it.

The remainder of this section will be devoted to sketching how to use relevant properties of the conditional score to attain consistency and local full efficiency. In doing so, it will be possible to retain sufficient flexibility so as to develop a method for attaining high efficiency globally. This will be developed in the following section. Regularity issues will be deemphasized so as to focus attention on more important aspects of the problem.

Consider first an estimating equation of the form

$$\sum W_i(\theta, \gamma) = 0,$$

where W_i are i.i.d. summands under the mixture model. To mimic a standard proof of the existence of a consistent sequence of roots to the likelihood equation, (as in Lehmann, 1983, page 413) where W_i plays the role of the usual score function, one only needs

$$\begin{aligned} &\text{(a) } E_{\theta_0}[W_i(\theta_0)] = 0, \\ &\text{(b) } E_{\theta_0}[W_i(\theta_0 - a)] > 0, \\ &\text{(c) } E_{\theta_0}[W_i(\theta_0 + a)] < 0, \end{aligned} \tag{4.2}$$

for all a in some neighborhood of zero. These properties follow easily here. For (a) we simply first compute the expectation given $S_i(\theta)$. For (b) and (c) we note

that differentiation of (a) with respect to θ yields (under regularity, all evaluations at θ_0)

$$E_{\theta_0}[D_\theta(W_i)] = E_{\theta_0}[-W_i D_\theta \log f],$$

which by preconditioning on S_i in the expectation yields $-E\{E_\gamma[\phi \mid S_i] \cdot E_Q[\phi \mid S_i]\Delta_i^2\}$. Thus when the ϕ-parameters are nonnegative, as they are in the considered examples, (b) and (c) follow from an interchange of limit and integral.

Next, in the method considered in (4.1), there is the additional nuisance that γ is to be estimated, so the weights $\hat{\delta}_i = E_{\hat{\gamma}}[\phi \mid s_i]$ will depend on the whole sequence $s_1, \cdots, s_p$ through $\hat{\gamma}$ and so $\{W_i(\theta_0, \hat{\gamma}(\theta_0))\}$ are not i.i.d. (The expectation properties above still hold.) This complication can be repaired using smoothness of the functions in γ and root-p convergence of $\hat{\gamma}$, properties which will be needed again shortly.

The next consideration to be developed is that of estimating γ in such a way that full efficiency is attained when the nuisance parameters are generated from Q_γ, but at the same time retaining sufficient flexibility to tackle global efficiency. The key idea here is due to Neyman (1959) in his work on $c(\alpha)$ tests of hypotheses. It has the following rough description: an asymptotically normal test function $H(\theta_0, \gamma)$ has the same asymptotic distribution as $H(\theta_0, \hat{\gamma})$ (that is, there is no loss in estimating the true parameters) provided that $\hat{\gamma}$ is any root-p consistent estimator and H is "orthogonal" to the tangent space of the nuisance parameter. With some slight extensions, this idea can be used here to find the asymptotic distributions of the estimators under consideration when $\hat{\gamma}(\theta)$ converges at root-p to γ^*, a parameter to be defined by the procedure determining $\hat{\gamma}$. Looking ahead, γ^* will be γ when Q_γ is the correct distribution on ϕ. Otherwise it will be an asymptotically optimal function of the unknown Q.

LEMMA 4.1. *Suppose that under* (θ_0, Q)

$$\sqrt{p}(\hat{\gamma}(\theta_0) - \gamma^*) = O_p(1).$$

Then under regularity conditions given in situ

$$\sqrt{p}[\bar{W}(\theta_0, \hat{\gamma}(\theta_0)) - \bar{W}(\theta_0, \gamma^*)] \to_p 0.$$

PROOF. Under the assumption that $E_\gamma(\phi \mid s)$ is twice differentiable in γ, one can write a Taylor series expansion

$$p^{1/2}(\bar{W}(\theta_0, \hat{\gamma}(\theta_0)) - \bar{W}(\theta_0, \gamma^*)) = p^{1/2}(\hat{\gamma}(\theta_0) - \gamma^*)D_\gamma \bar{W}(\theta_0, \gamma^*) + \text{Remainder}.$$

Under standard bounding conditions the Remainder term will converge to zero in probability. The convergence condition on $\hat{\gamma}$ implies that it suffices to show that $D_\gamma \bar{W}(\theta, \gamma^*) \to_p 0$. This follows from the law of large numbers since

$$D_\gamma W_i = D_\gamma[E_{\theta,\gamma^*}(\phi \mid S_i(\theta))]\Delta_i(\theta),$$

which can be shown to have expectation zero (if it exists) by first computing the expectation given $S_i(\theta)$.

The following theorem gives the asymptotic distribution of a consistent sequence of roots to our estimating equation. Of importance is the fact that if the ϕ-sequence is truly from Q_γ, then any root-p consistent estimator of γ will give an equation which achieves the information upper bound in the nonparametric Q problem.

THEOREM 4.2. *Given the conditions of Lemma* 4.1 *and some further regularity conditions indicated in the proof, a consistent solution* $\hat{\theta}$ *to* $\bar{W}(\hat{\theta}, \hat{\gamma}(\hat{\theta})) = 0$ *satisfies*

$$\sqrt{p}(\hat{\theta} - \theta_0) \to_p N(0, V),$$

where

$$V = E[W^2(\theta, \gamma^*)]/E^2[D_\theta W(\theta, \gamma^*)]. \tag{4.3}$$

In particular, if the true mixing distribution is Q_γ, *and* $\gamma^* = \gamma$, *then* $V = i_c^{-1}$.

PROOF. Consider the following first-order expansion:

$$0 = \sqrt{p}\bar{W}(\hat{\theta}, \hat{\gamma}(\hat{\theta})) = A + BC + DE + \text{Remainder},$$

where

$$A = \sqrt{p}\bar{W}(\theta_0, \hat{\gamma}(\theta_0))$$

$$B = \sqrt{p}(\hat{\theta} - \theta_0)$$

$$C = D_\theta \bar{W}(\theta_0, \hat{\gamma}(\theta_0))$$

$$D = \sqrt{p}(\hat{\gamma}(\hat{\theta}) - \hat{\gamma}(\theta_0))$$

$$E = D_\gamma \bar{W}(\theta_0, \hat{\gamma}(\theta_0)).$$

The terms can be dealt with as follows: First, from Lemma 4.1 and the central limit theorem $A \to N(0, E\ W^2(\theta, \gamma^*))$. Second,

$$C \cong D_\theta \bar{W}(\theta_0, \gamma^*) + D_\gamma D_\theta \bar{W}(\theta_0, \gamma^*)[\hat{\gamma}(\theta_0) - \gamma^*] \to_p E[D_\theta W(\theta_0, \gamma^*)] + 0.$$

Third, $D \cong \sqrt{p}(\hat{\theta} - \theta_0)\hat{\gamma}'(\theta_0)$ and $E \cong D_\gamma \bar{W}(\theta_0, \gamma^*) + (\hat{\gamma}(\theta_0) - \gamma^*)D_\gamma 2\bar{W}(\theta_0, \gamma^*)$, so that $BC + DE$ equals

$$\sqrt{p}(\hat{\theta} - \theta_0)[E(D_\theta W) + o_p(1)].$$

Hence it follows that under suitable regularity conditions

$$\sqrt{p}(\hat{\theta} - \theta_0) \to N(0, V).$$

5. Maximum information estimators. The twin goals of global consistency and parametric model efficiency have come at a relatively low cost due to the unique orthogonality properties of the conditional score function. What

remains is to exploit the weak requirements on $\hat{\gamma}$ to gain high efficiency globally. This last part of the approach will be illustrated through the paired exponential example.

EXAMPLE. Suppose that one has chosen as a parametric model for ϕ the gamma (α, λ) distribution. The estimating equation for θ is then

$$\sum E[\phi \mid s_i(\theta)] \cdot \frac{y_i - \theta x_i}{2\theta} = \sum \left(\frac{\hat{\alpha} + 2}{\hat{\lambda} + s_i(\theta)} \right) \frac{y_i - \theta x_i}{2\theta}, \tag{5.1}$$

where $\hat{\alpha}$ and $\hat{\lambda}$ are functions of $s_1(\theta), \cdots, s_p(\theta)$. Suppose in lieu of using $(\hat{\alpha}, \hat{\lambda})$ one used fixed constants (α, λ). The corresponding asymptotic variance formula for the resulting θ-estimator is the inverse of the information:

$$I_Q(\theta, \alpha, \lambda) = \frac{E^2\left[D_\theta \left\{ \frac{\alpha + 2}{\lambda + S(\theta)} \frac{Y_i - \theta X_i}{2\theta} \right\} \right]}{E\left(\frac{\alpha + 2}{\lambda + S(\theta)} \frac{Y_i - \theta X_i}{2\theta} \right)^2}. \tag{5.2}$$

Here E is computed under the true mixing distribution Q. As it comes from an unbiased estimating equation, this information is necessarily bounded above by the true conditional information, where $E_Q(\phi \mid s)$ replaces $(\alpha + 2)/(\lambda + s)$. In this i.i.d. setting, the information ratio does not depend on α, nor does the estimate of θ (see (5.1)) and one can define $\lambda^* = \lambda^*(\theta, Q)$ as the value of λ which maximizes this ratio. The information upper bound shows that if Q is gamma (α, λ), then λ^* equals λ. Thus by the theorems of Section 4, if we can find an estimator $\hat{\lambda} = \hat{\lambda}(\mathbf{s})$ such that $\sqrt{p}(\hat{\lambda} - \lambda^*) = O_p(1)$, we will attain the maximum information possible for the given form of the estimating equation, regardless of whether the data is gamma.

The obvious approach to estimate the λ^* which maximizes the information is to maximize an empirical version of the information. In this regard, we note that the information can be written

$$I_Q(\theta, \gamma) = \frac{(E\{E[D_\theta(E_\gamma[\phi \mid S]\Delta(\theta)) \mid S]\})^2}{E\{E^2_\gamma(\phi \mid S)E(\Delta^2(\theta) \mid S)\}}$$

and hence has an empirical version

$$\hat{I}(\mathbf{s}, \theta, \lambda) = \frac{(\sum_i (1/p_i)E[D_\theta W_\lambda \mid s_i])^2(1/p)}{\sum_i (1/p)E^2_\lambda[\phi \mid s_i]E(\Delta^2 \mid s_i)}.$$

Thus for fixed θ, one could choose $\hat{\gamma}(\theta)$ to maximize $\hat{I}$; this estimator has the necessary property of depending on the data only through $s_1, \cdots, s_p$, and, under some constraints on $I_Q(\theta, \lambda)$, will be $\sqrt{p}$-consistent. For the paired exponential example, the empirical information is

$$\hat{I} = \frac{[3[\sum s_i/(\lambda + s_i)] - \sum s_i^2/(\lambda + s_i)^2]^2}{(\sum s_i^2/(\lambda + s_i)^2)(12)p\ \theta^2}.$$

The maximization over λ has no explicit solution, but it could be achieved with a Newtonian algorithm. Rather than pursue this line further, we offer in the next section a class of models where the maximum information problem has an explicit solution. However, we do note that we have achieved one of our objectives: finding an estimator which dominates the m.l.e. We know this without calculation because the case $\lambda = 0$ corresponds to the estimating equation for the m.l.e., and we have maximized information over λ. In the next section, we will find a dominating estimator whose information can be explicitly calculated.

REMARKS. (1) In this problem the information is concave as a function of $\mu(\lambda) = E[S/(S + \lambda)]$, strictly so if the measure corresponding to E is not degenerate, and so a unique solution is obtained in this problem, both for theoretical E and the empirical $\hat{E}$.

(2) Although the approach is asymptotically correct, it is somewhat naive for finite samples. For example, if one uses this approach in the James-Stein setting, one uses the shrinking factor $(1 - p/\sum x_i^2)$ in lieu of $(1 - (p - 2)/\sum x_i^2)$. The former shrinking factor offers improvement over $\delta(x) = x$ only when $p \geq 5$ (e.g., Lehmann, 1983, page 302).

(3) It should be noted that the estimated variance for $\hat{\theta}$ based on the maximized information is consistent, but it is likely to be overly optimistic in small samples.

(4) The consistency of the estimated variance for $\hat{\theta}$ means that one can construct asymptotic confidence limits for θ based on the limiting normal distribution provided that the convergence to normality of the standardized statistic is uniform on intervals of θ. It seems reasonable that the corresponding test procedure would have advantageous small sample similarity properties because the use of conditioning on the statistics S_i to form the estimated information yields a test procedure approximately correct in the conditional distribution of the data given S_i. This needs further investigation.

6. Linear empirical Bayes. Suppose that the model has the form

$$f(x; \theta, \phi) = h(x)\exp(\phi s(\theta) - a(\theta) - b(\phi)), \tag{6.1}$$

a special case of (2.1); the paired exponential example has this structure. Let $dP_\beta(\phi)$ be a family (over β) of infinitely divisible distributions, with moment generating function $\exp(\beta g(t))$. Next, define a new family of distributions with parameter β by reweighting these infinitely divisible distributions with weight $\exp(b(\phi))$:

$$dP_{\beta^*}(\phi) = \exp\, b(\phi)\, dP_\beta(\phi) \Big/ \left[\int \exp\, b(\phi)\, dP_\beta(\phi)\right].$$

Provided this family exists, one can generate a parametric mixture density family which retains the exponential family structure of f:

$$\int f(x; \theta, \phi)\, dP_{\beta^*}(\phi) = h(x)e^{\beta g(s(\theta)) - a(\theta)} \Big/ \int e^{b(\phi)}\, dP_\beta(\phi).$$

Lindsay (1984) discussed the creation and uses of such exponential families. By the closure of the infinitely divisible distributions under location changes and convolutions, one can generate a general mixed exponential family

$$\int f(x;\, \theta,\, \phi)\, dP_{\beta^*}(\phi) = c(\boldsymbol{\beta})h(x)e^{\beta_0 s(\theta)+\sum \beta_j g_j(s(\theta))}.$$

Here the $g_j(t)$ are log moment generating functions for infinitely divisible distributions. This is important in our context because for mixtures of the density in (6.1) we have

$$D_s \log\left[\int f(x;\, \theta,\, \phi)\, dQ(\phi)/h(x)\right] = E(\phi \mid s).$$

For the above choice of $Q = P_{\beta^*}$, this gives

$$E(\phi \mid s) = \beta_0 + \sum \beta_j g_j'(s(\theta)).$$

Letting $\mathbf{h}(s) = (1, g_1'(s), \cdots, g_k'(s))$, we write the above linear posterior mean relationship as

$$E(\phi \mid s) = \boldsymbol{\beta}^T \mathbf{h}(s).$$

If one uses this family of distributions, then the information function is

$$I_Q(\theta, \boldsymbol{\beta}) = \frac{E^2[\boldsymbol{\beta}^T \mathbf{h}(S) E(D_\theta \Delta \mid S) + \boldsymbol{\beta}^T E\{D_s \mathbf{h}(S)(D_\theta S)(\Delta) \mid S\}]}{E[\boldsymbol{\beta}^T \mathbf{h}(S)\}^2 E(\Delta^2 \mid S)]}$$

$$= (\boldsymbol{\beta}^T \mathbf{v})^2 / \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta},$$

where

$$\Sigma_{ij} = E[h_i(S) h_j(S) E(\Delta^2 \mid S)]$$

and $v_i = E[h_i(S)E(D_\theta \Delta \mid S) + (D_s h_i(S))E(\Delta D_\theta(S) \mid S)]$. Since Σ is clearly nonnegative definite, and typically positive definite, the information is maximized for

$$\boldsymbol{\beta} = \Sigma^{-1} \mathbf{v}$$

with value $I_{\max} = \mathbf{v}^T \Sigma^{-1} \mathbf{v}$. The maximum information estimators of $\boldsymbol{\beta}$ are therefore

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}^{-1} \hat{\mathbf{v}},$$

where $\hat{\Sigma}$ and $\hat{\mathbf{v}}$ are the corresponding moment matrices generated from the empirical distribution of $s_1, \cdots, s_p$.

EXAMPLE. In the paired exponential example with posterior mean function $\beta_0 + \beta_1 h(s)$, we have

$$\Sigma = \frac{1}{12\theta^2} \begin{bmatrix} E[S^2] & E[S^2 h(S)] \\ E[S^2 h(S)] & E[S^2 h^2(S)] \end{bmatrix}$$

and

$$\mathbf{v}^T = (1/4\theta^2)(E(S),\quad E(Sh(S) + S^2 h'(S)/3)).$$

One special form for $h(s)$ which has several advantages in this problem is $h(s) = 2/s$. The function $g(s) = \beta \ln[\lambda/(\lambda + s)]$ is the log moment generating function of the infinitely divisible gamma distribution and so β/s is a limiting function of the posterior mean functions $\beta/(\lambda + s)$ generated by the gamma. One advantage to this form is that, since we are maximizing information over the weighting functions $\beta_0 + 2\beta_1/s$, we are sure to dominate the m.l.e., which has $\beta_0 = 0$. Another advantage to this form is that theoretical calculations of Σ and $\mathbf{v}$ for the gamma can be made explicitly.

Using this form the estimator of β is

$$\hat{\beta}^T = (\bar{s}, [\hat{\sigma}_s^2 - \tfrac{1}{2}\bar{s}^2]^+)(1/\hat{\sigma}_s^2),$$

where $\bar{s}$ and $\hat{\sigma}_s^2$ are the sample mean and variance of $s_1, \cdots, s_p$. For fixed mean $\bar{s}$, we note that if the sample variance is sufficiently small we use equal weights, and as the variance becomes infinite we approach the m.l.e.

The asymptotic variance of θ, the corresponding maximum information estimator of θ, is

$$\text{AV} = 3\theta^2\{1 + \tfrac{1}{4}E^2(S)/\text{Var } S\}^{-1},$$

which is superior to the asymptotic variance of the m.l.e. $(3\theta^2)$ unless Var $S = \infty$. In particular, suppose that Q is actually gamma (α, λ). We have previously seen that the optimal variance is $2\theta^2(\alpha + 3)/(\alpha + 2)$. The above estimator attains $2\theta^2 \min\{\tfrac{3}{2}, (\alpha + 1)/\alpha\}$, so the greatest loss in efficiency is for α small.

7. Varying sample sizes; weighted means problem. In the course of the preceding analysis, there have been several places where the assumption that density f_i depends on i only through ϕ_i has been important. The first case was in the heuristics of Section 3, where it enabled us to argue that the information function in the fixed sequence approach, with estimated weights, depended only on the empirical c.d.f. Q^* of the sequence, and in fact gave the information associated with treating Q^* as an unknown prior on the ϕ-sequence. Here the symmetry was used in the assumption that $\delta(\cdot)$, the estimated weighting function, should depend on i only through s_i; otherwise the information in a weighted conditional score would have depended on the ϕ-sequence ordering, not just Q^*. More generally, if f_i does depend on i, say through sample size n_i, there are other arguments (e.g., Lindsay, 1980, Section 2.2) which suggest that an optimal procedure in the class of mixed densities will have good properties when viewed along fixed sequences $\phi_1, \cdots, \phi_p$. For example, the mean square error of an estimator under the mixed model is simply an average (using the i.i.d. distribution of $\phi_1, \cdots, \phi_p$) of its mean square error along ϕ-sequences.

Another casualty to the dependence of f_i on i is the simple i.i.d. analysis used in the asymptotics. In particular, assumptions have to be made to ensure the asymptotic convergence of such quantities as the information.

To illustrate these difficulties, consider a problem where in the ith stratum one observes a sample of n_i variables with sufficient statistics y_i which have density $f(y_i; \theta, \phi_i, n_i)$. An estimating equation of the form $\sum H(n_i, y_i) = 0$ will

have average information

$$\frac{1}{p}\frac{\{E[\sum H'(n_i, Y_i)]\}^2}{E[\sum H(n_i, Y_i)]^2} = \frac{\{\int E_{\theta,\phi,N}[H'(N, Y)]\, dG^*(N, \phi)\}^2}{\int E_{\theta,\phi,N}[H^2(N, Y)]\, dG^*(N, \phi)},$$

where G is the empirical c.d.f. of the pairs (n_i, ϕ_i). That is, the limiting information now depends on the asymptotic relationship between the pair (n_i, ϕ_i), where one is observed and the other is not.

If we suppose there is asymptotic independence between n_i and ϕ_i, in the sense that as $p \to \infty$ one has $G^*(n, \phi) \to P(n)Q(\phi)$ for some distributions P and Q, then the limiting information could be written as

$$\{E[H'(N, Y)]\}^2/E[H^2(N, Y)],$$

where the expectation on (N, Y) is taken under joint density

$$p(n)\int f(y \mid n; \theta, \phi)\, dQ(\phi),$$

using $p(n)$ for the probability mass function of $P(n)$. In this density, it may be possible to extract a conditional score function, now by conditioning on N and $S(\theta)$. For example, if we observe n_i paired exponentials in the ith stratum, then this density is, with $y = \sum y_i$ and $x = \sum x_i$,

$$p(n)\int \theta^n\phi^n e^{-\theta\phi x}\phi^n e^{-\phi y}\, dQ(\phi).$$

Now N and $S(\theta) = \theta X + Y$ are jointly sufficient for ϕ (and Q) with θ fixed. The conditional score function is

$$W = E[\phi \mid S(\theta), n]((Y - \theta X_i)/2\theta).$$

Estimating the posterior mean in W leads to the technique described in the examples which follow. Provided there is no reason to presume a relationship between ϕ_i and n_i, this seems to be a natural approach.

In addition to losing simplicity in the varying sample size case, we note that the linear empirical Bayes method of Section 6 fails (to be Bayes) because the prior which gave a particular linear posterior mean $\beta^T\mathbf{h}$ depended implicitly on sample size; only under special circumstances would it give a linear posterior mean for any other sample size. On the other hand, if one views the empirical Bayes methods as being a scheme for choosing weights, one might still use the maximum information approach to find the best weighting function $\beta^T \cdot \mathbf{h}(s)$. Careful selection of $h(s)$ will ensure superiority to the maximum likelihood estimator. What is lost is a family of distributions $\{Q_\gamma\}$ for which full efficiency is guaranteed.

EXAMPLE A (continued). Suppose the ith stratum has n_i pairs of exponentials with common ϕ_i giving density

$$f(x, y; \theta, \phi) = (\theta\phi)^{n_i}e^{-\theta\phi x}\phi^{n_i}e^{-\phi y}.$$

The posterior mean function associated with the gamma (α, λ) prior is

$$E[\phi \mid s, n] = (\alpha + 2n_i)/(\lambda + s_i).$$

Thus the information function (5.2) would now be maximized over both α and λ. For fixed λ, the posterior mean is linear in α, and so the information can be maximized explicitly. For $\alpha = 0$, $\lambda = 0$, we recover the maximum likelihood equation. Another approach would be to use the linear weight $\beta_0 + \beta_1/\bar{s}_i$ and maximize information over this class (which also includes the m.l.e.).

EXAMPLE B (weighted means). An important example which has received much treatment is the weighted means problem. (See, in particular, Bartlett, 1936; Neyman and Scott, 1948; and Cox, 1975b.) In its simplest version for each strata, one observes n_i independent normals $(x_{i1}, \cdots, x_{in_i})$, all X-variables having the same mean θ, but with a variance σ_i^2 which depends on the stratum. Neyman and Scott (1948) showed that the m.l.e. was inefficient, provided the n_i were not constant in i, and in particular if one considered the optimal estimating equation of the form

$$\sum w_i(n_i/s_i(\theta))n_i(\bar{x}_i - \theta) = 0,$$

where $s_i(\theta) = \sum_j (x_{ij} - \theta)^2$, then $w_i = (n_i - 2)/n_i$ dominated $w_i \equiv 1$. which is the maximum likelihood weighting system. (The paired exponential with varying n_i does not have this reweighting feature; see Lindsay, 1982.) The optimal system excludes all strata with just two observations, and with good reason, as the contribution to the asymptotic variance of these terms is infinite. This flaw in the weights of the m.l.e. is of a different nature than the empirical Bayes problem. It relates to an appropriate weighting of strata with different intrinsic information content. See Lindsay (1982).

This problem is easily put into our framework. The conditional score in θ is just the ordinary score function $n_i(\bar{X}_i - \theta)/\sigma_i^2$ and the complete and sufficient statistic for $\phi_i = \sigma_i^{-2}$ is $S_i(\theta)$ as defined above. If one uses a gamma prior on ϕ, one obtains the estimating equation given by Cox (1975b):

$$\sum ((\alpha + n_i/2)/(\lambda + s_i))n_i(\bar{x}_i - \theta) = 0. \tag{7.1}$$

The information in this case of varying sample sizes is

$$I = \frac{\left(\sum_i E\left[(\alpha + n_i/2)\left\{\frac{n_i}{\lambda + S_i} - \frac{2S_i}{(\lambda + S_i)^2}\right\}\right]\right)^2}{\sum_i E\left[\frac{2S_i(\alpha + n_i/2)^2}{(\lambda + S_i)^2}\right]}. \tag{7.2}$$

We consider how the maximum information estimator, which would maximize (7.2), with E's deleted, to obtain weights, compares with the Neyman-Scott solution. The answer is trivial for $n_i \equiv_i 2$, as Neyman and Scott's solution is vacuous, and the above equation (7.1) yields an estimator with finite asymptotic

variance for every $\lambda > 0$. Indeed, if the n_i are constant, the Neyman-Scott solution is the m.l.e., and one can maximize the information in (7.2) (α now drops out) to obtain a superior estimator, as $\lambda = 0$ recovers the m.l.e., and the information has positive slope at $\lambda = 0$ for any prior Q, provided $P[S_i(\theta) = 0] = 0$. (See also Remark (1) below.)

There appears to be a contradiction here for nonconstant n_i. For small values of α and λ, the empirical Bayes solution to (7.1) will be close to the maximum likelihood estimator ($\alpha = 0$, $\lambda = 0$) and be the best possible estimator for the gamma (α, λ) prior. On the other hand, the Neyman-Scott solution is strictly superior to the m.l.e. for *every* sequence $\phi_1, \cdots, \phi_p$, and corresponds to $\alpha = -1$, $\lambda = 0$ in (7.1). There is an interesting discontinuity in (α, λ) arising from the difference in behavior between $\lambda = 0$ and $\lambda > 0$. In the latter case, the weights are bounded above, while the possibility of very large weights in the former case apparently mandates a readjustment in the weights which depends on sample size and forces the solution away from the m.l.e. solution.

This still leaves open the question as to whether the gamma maximum information estimator dominates the Neyman-Scott estimator for *every* prior Q. From the above remarks, it is clear that this will be true if the values of α over which information is maximized are extended to $[-1, \infty)$, as then the class includes the Neyman-Scott solution. It is not apparent if this is a necessary step.

REMARKS. (1) It should be noted that if $s_i = 0$ for one particular i, then the corresponding ith summand in the denominator of the estimated information $\hat{I}$ is zero while the summand in the numerator is $\frac{1}{2}(n_i + 2\alpha)n_i/\lambda$, which goes to ∞ as $\lambda \to 0$. Thus the empirical information is infinite in this case, and the solution gives infinite weight to the ith stratum, yielding $\bar{x}_i$ as the estimator. This corresponds to the case $\sigma_i = 0$, where the mean is observed without error in the ith stratum.

(2) One linear pseudo-empirical-Bayes method would be to use weights $\beta_0 + \beta_1(n_i - 2)/s_i(\theta)$. This would ensure dominance of the Neyman-Scott solution, but be somewhat naive in weighting all strata with $n_i = 2$ equally.

REFERENCES

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

BARTLETT, M. S. (1936). The information available in small samples. *Proc. Camb. Philos. Soc.* **32** 560–566.

COX, D. R. (1975a). Partial likelihood. *Biometrika* **62** 269–276.

COX, D. R. (1975b). A note on partially Bayes inference and the linear model. *Biometrika* **62** 651–654.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63** 277–284.

GODAMBE, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67** 155–162.

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

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser.* B **32** 175–208.

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.

KUMON, M. and AMARI, S. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters. *Biometrika* **71** 445–460.

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LAIRD, N. M. (1982). Empirical Bayes estimates using the nonparametric estimate of the prior. *J. Statist. Comput. Simulation* **15** 211–220.

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LEONARD, T. A. (1984). Some data analytic modifications of Bayes-Stein estimation. *Ann. Inst. Statist. Math.* **36** 11–21.

LIANG, K.-Y. (1984). The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71** 305–314.

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

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

LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–95.

LINDSAY, B. G. (1983b). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

LINDSAY, B. G. (1983c). The geometry of mixture likelihoods: part II, the exponential family. *Ann. Statist.* **11** 783–792.

LINDSAY, B. G. (1984). Exponential family mixture models (with regression estimators). Pennsylvania State University, Department of Statistics Technical Report.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *The Harold Cramer Volume.* 213–234, Wiley, New York.

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

YATES, F. and COCHRAN, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28** 556–580.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802