The Annals of Statistics
1981, Vol. 9, No. 4, 861–869

PSEUDO MAXIMUM LIKELIHOOD ESTIMATION: THEORY AND APPLICATIONS[1]

By Gail Gong and Francisco J. Samaniego

Stanford University and University of California, Davis

Let $X_1, \cdots, X_n$ be i.i.d. random variables with probability distribution $F_{\theta,p}$ indexed by two real parameters. Let $\hat{p} = \hat{p}(X_1, \cdots, X_n)$ be an estimate of p other than the maximum likelihood estimate, and let $\hat{\theta}$ be the solution of the likelihood equation $\partial/\partial\theta \ln L(\mathbf{x}, \theta, \hat{p}) = 0$ which maximizes the likelihood. We call $\hat{\theta}$ a pseudo maximum likelihood estimate of θ, and give conditions under which $\hat{\theta}$ is consistent and asymptotically normal. Pseudo maximum likelihood estimation easily extends to k-parameter models, and is of interest in problems in which the likelihood surface is ill-behaved in higher dimensions but well-behaved in lower dimensions. We examine several signal-plus-noise, or convolution, models which exhibit such behavior and satisfy the regularity conditions of the asymptotic theory. For specific models, a numerical comparison of asymptotic variances suggests that a pseudo maximum likelihood estimate of the signal parameter is uniformly more efficient than estimators proposed previously.

1. Introduction. Probability models abound for which the analytical derivation of the maximum likelihood estimate of model parameters is virtually impossible. For many such models, one among the wide variety of numerical algorithms available for approximating the MLE will prove satisfactory. For other models, numerical methods are unreliable or converge too slowly to be of use. There has been particular difficulty with the likelihood approach to estimation in the presence of nuisance parameters. Godambe (1974, 1977) has referred to this area as the major failure of the likelihood approach, and has developed the theory of estimating equations in part to fill this void. The difficulties in obtaining the MLE for models with nuisance parameters has led to the investigation of alternative estimation procedures which have the spirit of likelihood procedures, but are compromises due to the intractability of the preferred approach. Several likelihood-based procedures are discussed by Kalbfleisch and Sprott (1970). The method of maximum partial likelihood estimation was introduced and studied by Cox (1975). Optimality results for several such methods have been obtained by Andersen (1970), Godambe (1976) and by Liu and Crowley (1978).

A comprehensive review of the literature on estimation in the presence of nuisance parameters is given by Basu (1977). To the extent that approaches to estimation in nuisance parameter problems have focused on the elimination of nuisance parameters through conditioning or data reduction, the approaches have limited applicability. Many problems of practical importance do not give rise to convenient factorizations or to the existence of useful sufficient or ancillary statistics. The convolution models considered in this paper are models in which the approaches mentioned in the preceeding paragraph fail. It is precisely the characteristics of these models that has led us to the approach studied here.

Received September, 1978; revised September, 1980.

[1] This research was supported in part by the Air Force Office of Scientific Research under Contract AFOSR-77-3180.

AMS 1970 *subject classifications*. Primary 62F12; secondary 62A10.

Key words and phrases. Likelihood, estimation, asymptotics, consistency, relative efficiency, convolution.

In general, pseudo maximum likelihood estimation consists of replacing all nuisance parameters in a model by estimates and solving a reduced system of likelihood equations. The method is a reasonable one in problems in which lower dimensional maximum likelihood estimation is feasible while higher dimensional maximum likelihood estimation is intractable. As we shall see, the method is in this sense ideally suited for application to many convolution models. Pseudo maximum likelihood estimation has a natural counterpart in testing theory, namely, the $C(\alpha)$ tests introduced and studied by Neyman. In his fundamental paper on testing composite hypotheses, Neyman (1959) proposed a test statistic whose construction involves replacing nuisance parameters by $\sqrt{n}$-consistent estimates. The asymptotic distribution of the test statistic was given under conditions more stringent than we will impose. Specifically, a condition equivalent to Neyman's that the test statistic be uncorrelated (under the null hypothesis) with the logarithmic derivative of the likelihood with respect to nuisance parameters would guarantee in our context an efficient estimator of the structural parameter. We investigate here the asymptotic behavior of pseudo MLE's under less restrictive conditions, and indicate some applications of our results.

In Section 2, we develop the asymptotic theory of pseudo maximum likelihood estimators in problems with a single structural parameter. Under fairly standard regularity conditions, the consistency and asymptotic normality of pseudo MLE's are established there. In Section 3, we describe the special features of certain convolution models which identifies them to be a natural domain for application of the method of pseudo maximum likelihood estimation. We discuss the regularity of such models and the asymptotic relative efficiency of specific pseudo MLE's. We make some concluding general remarks in Section 4.

2. Asymptotics. Let $X_1, \cdots, X_n$ be a random sample from a member of the two parameter family $\mathscr{F} = \{F_{\theta,\pi}\}$ of distributions on the real line. We will assume throughout our development the existence, for every (θ, π), of the density or probability mass function $f(x \mid \theta, \pi)$ with respect to some sigma-finite measure μ on $\mathscr{R}$. The method of pseudo maximum likelihood estimation may be viewed as follows. Given a sample of size n from $F_{\theta,\pi}$, an estimate $\hat{\pi}_n$ is developed for the parameter π by some technique or approach other than maximum likelihood estimation. The pseudo MLE is then obtained by maximizing the log likelihood $\mathscr{L}_n(\theta, \hat{\pi}_n)$, viewed as a function of the single parameter θ. The pseudo MLE $\hat{\theta}_n$ should have good large sample properties when $\hat{\pi}_n$ does. The consistency of the pseudo MLE is expected when $\hat{\pi}_n$ is consistent, and is established here under simple and natural regularity conditions. The efficiency of $\hat{\theta}_n$ will of course depend on the relative efficiency of $\hat{\pi}_n$. The asymptotic distribution of $\hat{\theta}_n$ is derived under regularity conditions when the estimator $\hat{\pi}_n$ is $\sqrt{n}$-consistent and asymptotically normal. The asymptotic theory for pseudo MLE's is developed here for a two-parameter problem rather than more generally because of the resultant ease of exposition and simplicity of notation. We trust it will be apparent to the reader that our result easily generalizes to the case when π is vector valued. The extension to vector valued θ will not be pursued here.

We will make use of the standard symbols $o_p(\cdot)$ and $\mathscr{O}_p(\cdot)$ for convergence and boundedness in probability (see Mann and Wald (1943)). We make repeated use of the following elementary yet fundamental lemma.

LEMMA 2.1. *Let $X_1, \cdots, X_n$ be* i.i.d. *random variables from a distribution F_π on the real line, with $\pi \in \Pi \subseteq \mathscr{R}$. Let $\pi_0 \in \Pi$ be the true value of the parameter, and let $\hat{\pi}_n = \hat{\pi}_n(X_1, \cdots, X_n)$ be such that $\hat{\pi}_n \to \pi_0$ in probability. Let $\psi(x, \pi)$ be a differentiable function of π for $\pi \in B$, an open neighborhood of π_0, and for almost all x in the sample space $\mathscr{X}$, and suppose $E \mid \psi(X_1, \pi_0) \mid < \infty$. If*

$$\left| \frac{\partial}{\partial \pi} \psi(x, \pi) \right| \leq M(x) \tag{2.1}$$

for all $\pi \in B$, where $EM(X_1) < \infty$, then

(2.2) $$\frac{1}{n}\sum_{i=1}^n \psi(X_i, \hat{\pi}_n) \to E\psi(X_1, \pi_0) \text{ in probability.}$$

PROOF. Consider the Taylor series expansion

(2.3) $$\frac{1}{n}\sum_{i=1}^n \psi(X_i, \hat{\pi}_n) = \frac{1}{n}\sum_{i=1}^n \psi(X_i, \pi_0) + (\hat{\pi}_n - \pi_0)\cdot \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\pi}\psi(X_i, \tilde{\pi}_n) + o_p(1)$$

where $\tilde{\pi}_n$ is between π_0 and $\hat{\pi}_n$. By our hypotheses and the weak law of large numbers, it will suffice to show that $1/n \sum_{i=1}^n (\partial/\partial\pi\, \psi(X_i, \tilde{\pi}_n) = \mathcal{O}_p(1)$. To this end, let $\epsilon > 0$, $\delta > 0$, and define

$$S_n = \{\hat{\pi}_n \in B\}$$

and

$$T_n = \left\{\frac{1}{n}\sum_{i=1}^n M(X_i) < EM(X_1) + \delta\right\}.$$

There exists an integer $N = N(\epsilon, \delta)$ such that if $n > N$, $P(S_n) > 1 - \epsilon/2$ and $P(T_n) > 1 - \epsilon/2$. We then have $\forall\, n > N$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\pi}\psi(X_i, \tilde{\pi}_n)\right| \le EM(X_1) + \delta\right) \ge P(S_n \cap T_n) > 1 - \epsilon,$$

completing the proof.

We now define the notation to be used in the sequel. Let $X_1, \cdots, X_n$ be i.i.d., each with density or probability mass function $f(\cdot \mid \theta_0, \pi_0)$ defined on the sample space $\mathcal{X} \subseteq \mathcal{R}$, where $\theta_0 \in A$, A open with $A \subseteq \Theta \subseteq \mathcal{R}$ and $\pi_0 \in B$, B open with $B \subseteq \Pi \subseteq \mathcal{R}$. Let

$$\Phi(x \mid \theta, \pi) \equiv \ln f(x \mid \theta, \pi)$$

$$\mathcal{L}_n(\theta, \pi) \equiv \ln \prod_{i=1}^n f(x_i \mid \theta, \pi) = \sum_{i=1}^n \Phi(x_i \mid \theta, \pi)$$

$$\bar{\mathcal{L}}_n(\theta, \pi) \equiv \frac{1}{n}\mathcal{L}_n(\theta, \pi).$$

We will occasionally suppress the subscript n in the latter expressions. Partial derivatives are denoted with subscript notation; for example, $\Phi_{\theta\pi} = (\partial^2/\partial\pi\partial\theta)\Phi$. Throughout this section, we view θ as a structural parameter and π as a nuisance parameter. We will make use of the following regularity conditions.

(A1) For all x and for all $(\theta, \pi) \in A \times B$, the following partial derivatives exist: Φ_θ, $\Phi_{\theta\theta}$, $\Phi_{\theta\theta\theta}$, Φ_π, $\Phi_{\theta\pi}$, $\Phi_{\theta\theta\pi}$, $\Phi_{\theta\pi\pi}$.

(A2) Interchange of differentiation and integration of f is valid for first and second derivatives with respect to θ and for the mixed partial derivative with respect to θ and π.

(A3) $\mathcal{I}_{11}$ and $\mathcal{I}_{12}$ exist with $\mathcal{I}_{11} > 0$, where $\mathcal{I}_{11} = E_{\theta_0,\pi_0}(\Phi_\theta^2)$ and $\mathcal{I}_{12} = E_{\theta_0,\pi_0}(\Phi_\theta \cdot \Phi_\pi)$.

(A4) For all $(\theta, \pi) \in A \times B$ and for all $x \in \mathcal{X}$,

$$\left|\frac{\partial}{\partial\pi}\ln\frac{f(x \mid \theta, \pi)}{f(x \mid \theta_0, \pi)}\right| \le M(x, \theta)$$

where $EM(X_1, \theta) < \infty \quad \forall\theta \in A$.

(A5) The following third partial derivatives are bounded by integrable functions:

(i) $|\Phi_{\theta\theta\theta}(x \mid \theta, \pi)| \le M(x) \quad \forall(\theta, \pi) \in A \times B, \forall x$

(ii) $|\Phi_{\theta\theta\pi}(x \mid \theta_0, \pi)| \le M(x) \quad \forall\pi \in B, \forall x$

(iii) $|\Phi_{\theta\pi\pi}(x \mid \theta_0, \pi)| \le M(x) \quad \forall\pi \in B, \forall x,$

where $EM(X_1) < \infty$.

(A6) For any $(\theta, \pi) \neq (\theta_0, \pi_0)$,

$$P_{\theta_0,\pi_0}\{f(X_1 \mid \theta, \pi) = f(X_1 \mid \theta_0, \pi_0)\} < 1.$$

We first establish the consistency of the pseudo MLE.

THEOREM 2.1. *Let* $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F_{\theta_0,\pi_0}$, *and let* $\hat{\pi}_n = \hat{\pi}_n(X_1, \cdots, X_n)$ *be a consistent estimate of* π_0. *Suppose that regularity conditions* (A1), (A4) *and* (A6) *hold. For* $\epsilon > 0$, *let* $A_n(\epsilon)$ *be the event that there exists a root* $\hat{\theta}_n$ *of the equation*

$$\frac{\partial}{\partial\theta} \mathscr{L}_n(\theta, \hat{\pi}_n) = 0$$

for which $|\hat{\theta}_n - \theta_0| < \epsilon$. *Then, for any* $\epsilon > 0$, $P\{A_n(\epsilon)\} \to 1$.

PROOF. Let $\theta \in A$ be fixed and define

$$\psi(x \mid \pi) = \Phi(x \mid \theta, \pi) - \Phi(x \mid \theta_0, \pi)$$
$$= \ln \frac{f(x \mid \theta, \pi)}{f(x \mid \theta_0, \pi)}.$$

By (A1), $\psi(x \mid \pi)$ is differentiable for $\pi \in B$, and by (A4),

$$|\psi_\pi(x \mid \pi)| \leq M(x) \qquad \forall \pi \in B, \forall x.$$

Thus, by lemma 2.1,

$$\mathscr{L}_n(\theta, \hat{\pi}_n) - \mathscr{L}_n(\theta_0, \hat{\pi}_n) \to \mathrm{E}_{\theta_0,\pi_0} \ln \frac{f(X_1 \mid \theta, \pi_0)}{f(X_1 \mid \theta_0 \pi_0)} \quad \text{in probability.} \tag{2.4}$$

It is easy to show that $E_{(\theta_0,\pi_0)}\psi(X_1 \mid \pi_0)^+ < \infty$ in general, and we thus claim that $\mathscr{L}_n(\theta, \hat{\pi}_n) - \mathscr{L}_n(\theta_0, \hat{\pi}_n)$ converges to a negative number (possibly $-\infty$) if $\theta \in A - \{\theta_0\}$. This claim follows from Jensen's inequality and condition (A6) since

$$E_{\theta_0,\pi_0} \ln \frac{f(X_1 \mid \theta, \pi_0)}{f(X_1 \mid \theta_0, \pi_0)} < \ln E_{\theta_0,\pi_0} \frac{f(X_1 \mid \theta, \pi_0)}{f(X_1 \mid \theta_0, \pi_0)} = 0.$$

Because of the convergence demonstrated above, we may find, for any ϵ, $\delta > 0$ for which $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset A$, an integer $N = N(\epsilon, \delta)$ such that $n > N$ implies that $P(\hat{\pi}_n \in B) > 1 - \delta$, and that for $\theta = \theta_0 \pm \epsilon$, $P(\mathscr{L}_n(\theta), \hat{\pi}_n) < \mathscr{L}_n(\theta_0, \hat{\pi}_n)) > 1 - \delta$. Thus, for $n > N$, $P(\mathscr{L}_n(\theta, \hat{\pi}_n)$ has a local maximum $\hat{\theta}_n \in (\theta_0 - \epsilon, \theta_0 + \epsilon)) > 1 - 3\delta$. By (A1), $\hat{\theta}_n$ satisfies the equation $\frac{\partial}{\partial\theta} \mathscr{L}_n(\theta, \hat{\pi}_n) = 0$, completing the proof.

Theorem 2.1 establishes only that the pseudo maximum likelihood equation has a consistent root. In all applications considered in this paper, however, the pseudo maximum likelihood equation has a unique solution and the pseudo MLE is indeed consistent. This follows from the fact that the logarithmic derivative of the pseudo likelihood in each of these applications is a decreasing function of the structural parameter. We turn to the asymptotic distribution of the pseudo MLE.

THEOREM 2.2. *Let* $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F_{\theta_0,\pi_0}$, *and let* $\hat{\pi}_n = \hat{\pi}_n(X_1, \cdots, X_n)$ *be such that* $(\hat{\pi}_n - \pi_0) = O_p(1/\sqrt{n})$, *and suppose*

$$\sqrt{n} \begin{bmatrix} \mathscr{L}_n(\theta_0, \pi_0) \\ \hat{\pi}_n - \pi_0 \end{bmatrix} \to_{\mathscr{L}} \mathscr{N}\left(\mathbf{0}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ & \Sigma_{22} \end{bmatrix}\right) \tag{2.5}$$

Then, under regularity conditions (A1)–(A6), *the pseudo* MLE $\hat{\theta}_n$ *is asymptotically* *normal, that is,* $\sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma$ *converges in law to a standard normal variable, where*

$$\sigma^2 = \frac{1}{\mathscr{I}_{11}} + \frac{\mathscr{I}_{12}}{\mathscr{I}_{11}^2}(\Sigma_{22}\mathscr{I}_{12} - 2\Sigma_{12}). \tag{2.6}$$

PROOF. By condition (A1) and the consistency of $\hat{\theta}_n$ and $\hat{\pi}_n$, we may expand $\bar{\mathscr{L}}_\theta(\hat{\theta}_n, \hat{\pi}_n)$ about θ_0 as follows:

$$\begin{aligned} 0 = \sqrt{n}\bar{\mathscr{L}}_\theta(\hat{\theta}_n, \hat{\pi}_n) &= \sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \hat{\pi}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)\bar{\mathscr{L}}_{\theta\theta}(\theta_0, \hat{\pi}_n) \\ &\quad + \tfrac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta_0)^2\bar{\mathscr{L}}_{\theta\theta\theta}(\tilde{\theta}_n, \hat{\pi}_n) + o_p(1), \end{aligned} \tag{2.7}$$

where $\tilde{\theta}_n$ lies between θ_0 and $\hat{\theta}_n$. We may rewrite (2.7) as

$$-\sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \hat{\pi}_n) = \sqrt{n}(\hat{\theta}_n - \theta_0)[\bar{\mathscr{L}}_{\theta\theta}(\theta_0, \hat{\pi}_n) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)\bar{\mathscr{L}}_{\theta\theta\theta}(\tilde{\theta}_n, \hat{\pi}_n)] + o_p(1), \tag{2.8}$$

We examine several terms in equation (2.8) separately; using conditions (A1), (A2), (A3) and (A5) and the consistency of $\hat{\theta}_n$ and $\hat{\pi}_n$, we establish the following three identities:

(a) $\sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \hat{\pi}_n) = \sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \pi_0) - \sqrt{n}(\hat{\pi}_n - \pi_0)\mathscr{I}_{12} + o_p(1)$. To see this, we expand $\bar{\mathscr{L}}_\theta(\theta_0, \hat{\pi}_n)$ about π_0, yielding

$$\begin{aligned} \sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \hat{\pi}_n) &= \sqrt{n}\bar{\mathscr{L}}_\theta(\theta_0, \pi_0) + \sqrt{n}(\hat{\pi}_n - \pi_0)\bar{\mathscr{L}}_{\theta\pi}(\theta_0, \pi_0) \\ &\quad + \sqrt{n}(\hat{\pi}_n - \pi_0)^2\bar{\mathscr{L}}_{\theta\pi\pi}(\theta_0, \tilde{\pi}_n) + o_p(1). \end{aligned} \tag{2.9}$$

where $\tilde{\pi}_n$ is between π_0 and $\hat{\pi}_n$. Arguing as in Lemma 2.1, one can show that $\bar{\mathscr{L}}_{\theta\pi\pi}(\theta_0, \tilde{\pi}_n)$ is bounded in probability. This establishes that $\sqrt{n}(\hat{\pi}_n - \pi_0)^2\bar{\mathscr{L}}_{\theta\pi\pi}(\theta_0, \tilde{\pi}_n) = o_p(1)$, which together with the fact that $\sqrt{n}(\hat{\pi}_n - \pi_0)[\bar{\mathscr{L}}_{\theta\pi}(\theta_0, \pi_0) + \mathscr{I}_{12}] = o_p(1)$, proves identity (a).

(b) $\bar{\mathscr{L}}_{\theta\theta}(\theta_0, \hat{\pi}_n) + \mathscr{I}_{11} = o_p(1)$. This follows easily from condition (A5), (ii), and Lemma 2.1.

(c) $\tfrac{1}{2}(\hat{\theta}_n - \theta_0)\bar{\mathscr{L}}_{\theta\theta\theta}(\tilde{\theta}_n, \pi_n) = o_p(1)$. To see this, it suffices to show that $\bar{\mathscr{L}}_{\theta\theta\theta}(\tilde{\theta}_n, \hat{\pi}_n)$ is bounded in probability. This follows from condition (A5), (i), and an argument similar to that made in proving Lemma 2.1. Applying the three identities above, we see that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically equivalent to $\{\sqrt{n}\bar{\mathscr{L}}(\theta_0, \pi_0) - \sqrt{n}(\hat{\pi}_n - \pi_0)\mathscr{I}_{12}\}/\mathscr{I}_{11}$. This variable converges in distribution to $\mathscr{N}(0, \sigma^2)$, where σ^2, given in (2.6), may be obtained by noting that $\Sigma_{11} = \mathscr{I}_{11}$.

REMARKS. The regularity conditions under which the results of this section are proven can undoubtedly be weakened. We have made no effort to do this, but are satisfied with conditions that are fairly standard and are reasonably easy to check. Moreover, the applications of interest to us satisfy the stated conditions. We note that if $\mathscr{I}_{22} = E_{\theta_0,\pi_0}(\Phi_\pi^2)$ exists and is positive, and if $\hat{\pi}_n$ is asymptotically equivalent to the MLE of π_0, then $\Sigma_{12} = 0$, $\Sigma_{22} = \mathscr{I}_{11}/(\mathscr{I}_{12}\mathscr{I}_{22} - \mathscr{I}_{12}^2)$ and $\sigma^2 = \mathscr{I}_{22}/(\mathscr{I}_{11}\mathscr{I}_{22} - \mathscr{I}_{12}^2)$. Thus, in this case, $\hat{\theta}_n$ is asymptotically efficient.

3. Applications to signal plus noise models. Let X be a random variable whose distribution is that of the sum of independent variables Y and Z. The distribution of X is thus the convolution of F_Y and F_Z, and may be thought of as a model for signals in additive noise. Data obtained by a Geiger counter may serve as the prototypical example of signal plus noise data since these observations may be viewed as sums of counts due to the presence of a radioactive substance and counts due to noise or static. The wide applicability of continuous and discrete convolutions is discussed in Sclove and Van Ryzin (1969) and in Samaniego (1976).

Estimation problems for specific signal plus noise distributions have been examined by several authors, notably by Gaffey (1959) and by Sclove and Van Ryzin (1969). Due to the cumbersome nature of the likelihood function for convolution models, maximum likelihood estimation has met substantial resistance. We describe below a characteristic shared by a number of convolution models that facilitates maximum likelihood estimation in one-parameter problems and pseudo maximum likelihood estimation in multiparameter problems.

Let $f_\theta(x)$ represent the probability mass function or density of the variable X, where θ is a real valued parameter. While $f_\theta(x)$ may depend on other unknown parameters, such dependence, if any, is suppressed in the present discussion. Suppose $f_\theta(x)$ satisfies the system of differential equations

$$\frac{\partial}{\partial\theta} f_\theta(x) = f_\theta(x-1) - f_\theta(x) \tag{3.1}$$

for all appropriate x and θ. It follows that the likelihood equation $\mathscr{L}_\theta(\theta) = 0$ may be written as

$$\textstyle\sum_{i=1}^n \frac{f_\theta(x_i - 1)}{f_\theta(x_i)} = n, \tag{3.2}$$

where $\mathbf{x}$ is the vector of observations. It thus becomes clear that the behavior of probability ratios of the form

$$f_\theta(a)/f_\theta(b), \tag{3.3}$$

for $a < b$ is relevant to this estimation problem. The distribution of X is said to have *parametric monotone decreasing ratio* (PMDR) in θ if ratios (3.3) are decreasing in θ for all $a < b$ in the support of the distribution of X. PMDR in θ is well defined in multiparameter problems if we consider all parameters other than θ to be fixed. If a one-parameter model satisfies (3.1) and has PMDR, then the likelihood equation (3.2) has at most one solution, and the MLE is easily found numerically. The PMDR property is easily shown to be equivalent to monotone increasing likelihood ratio, but is a more convenient formulation of the notion in our context.

A number of signal plus noise models have the character described in the preceeding paragraph. For example, discrete convolutions of the Poisson or the binomial distribution satisfy the system (3.1) (see Samaniego (1976), (1980)), and many such convolutions have the PMDR property in the structural parameter. Several continuous convolution models have these same features (see Gong and Samaniego (1978)). The solvability of one-parameter estimation problems such as those described above provide the key to the feasibility of pseudo maximum likelihood estimation in convolution models, since for many convolution models, the pseudo MLE of a single parameter is easily found when all other parameters have been replaced by estimates.

Demonstrations that a given model satisfies a series of technical regularity conditions are inherently tedious and uninteresting. We therefore omit the details of such regularity checks and state without proof that the following convolution models satisfy the regularity conditions under which the asymptotic theory of pseudo maximum likelihood estimates has been developed: Poisson signals in binomial or normal noise, binomial signals in Poisson or normal noise. We refer the interested reader to Gong and Samaniego (1978) for a detailed investigation of regularity.

For the signal plus noise models we have discussed, we have established the consistency and asymptotic normality of the pseudo maximum likelihood estimate of the signal parameter, provided the noise parameter is estimated appropriately. The method of moments, for example, yields estimates of the noise parameter that satisfy the requirements of Theorems 2.1 and 2.2. The efficiency of the resulting pseudo maximum likelihood estimates will now be examined. We present below evidence in support of our conjecture that these pseudo MLE's are uniformly more efficient asymptotically than the method of moments estimators of the signal parameter. We study here the asymptotic relative efficiency of PMLE's both for Poisson signals in binomial noise and for binomial signals in

Poisson noise. Moment estimators of the binomial and Poisson parameters were proposed by Sclove and Van Ryzin (1969).

Let $X_1, \cdots, X_n$ be a random sample from $\mathscr{P}(\theta_0) * \mathscr{B}(N, p_0)$ with N known, and suppose we are primarily interested in estimating the (signal) parameter θ_0. It is easy to show that the moment estimators of p_0 and θ_0 are $\tilde{p}_n = \sqrt{(\bar{X} - s^2)/N}$ and $\tilde{\theta}_n = \bar{X} - N\tilde{p}_n$, where $\bar{X}$ and s^2 are the sample mean and variance, provided $\bar{X} \geq s^2$, which occurs with limiting probability one as $n \to \infty$. It is shown in Gong and Samaniego (1978) that

$$\sqrt{n}\begin{bmatrix} \bar{\mathscr{L}}_\theta(\theta_0, p_0) \\ \tilde{p}_n - p_0 \end{bmatrix} \to_{\mathscr{L}} \mathscr{N}(\mathbf{0}, \Sigma) \tag{3.4}$$

where

$$\Sigma = \begin{bmatrix} \mathscr{I}_{11} & 0 \\ 0 & t^2(\Gamma_{22} - 2\Gamma_{23} + \Gamma_{33}) \end{bmatrix}$$

with $t = 1/2Np_0$, $\Gamma_{22} = \theta_0 + Np_0(1 - p_0)$, $\Gamma_{23} = \theta_0 + Np_0(1 - p_0)(1 - 2p_0)$ and $\Gamma_{33} = \theta_0 + 2\{\theta_0 + Np_0(1 - p_0)\}^2 + Np_0(1 - p_0)\{1 - 6p_0(1 - p_0)\}^2$. Note that

$$\begin{bmatrix} \bar{\mathscr{L}}_\theta(\theta_0, p_0) \\ \tilde{p}_n \end{bmatrix} = g\begin{bmatrix} \mathscr{L}_\theta(\theta_0, p_0) \\ \bar{X}_1 \\ s^2 \end{bmatrix},$$

where g, given by

$$g_1(t_1, t_2, t_3) = t_1$$

and

$$g_2(t_1, t_2, t_3) = \sqrt{\frac{t_2 - t_3}{N}},$$

is a totally differentiable transformation. We may thus define

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & t & -t \end{bmatrix}.$$

and obtain (3.4) from the δ-method theorem, with $\Sigma = A\Gamma A^T$. We now record the asymptotic variances of the maximum likelihood estimate (MLE), the pseudo maximum likelihood estimate (PMLE) and the method of moments estimate (MME) of the signal parameter θ_0:

$$\sigma^2_{\text{MLE}} = \frac{\mathscr{I}_{22}}{\mathscr{I}_{11}\mathscr{I}_{22} - \mathscr{I}^2_{12}} \tag{3.5}$$

$$\sigma^2_{\text{PMLE}} = \frac{1}{\mathscr{I}_{11}} + \frac{\mathscr{I}^2_{12}}{\mathscr{I}^2_{11}} t^2(\Gamma_{22} - 2\Gamma_{23} + \Gamma_{33}) \tag{3.6}$$

$$\sigma^2_{\text{MME}} = (1 - Nt)^2\Gamma_{22} + 2Nt(1 - Nt)\Gamma_{23} + (Nt)^2\Gamma_{33}. \tag{3.7}$$

We have not been successful in comparing these expressions analytically. For a specific range of parameters, we have approximated $\mathscr{I}_{ij}$ with errors no greater than 10^{-7}, and have used these approximations in computing the asymptotic efficiency of the PMLE relative to the MLE and of the MME relative to the PMLE. The parameter values examined are the following: $\theta_0 = .1, .25, .5, 1(1)10$; $p_0 = .1(.1).9$; $N = 1(1)10$. For the parameter values above, we find

$$.57288 \leq ARE(\text{PMLE/MLE}) < 1$$

$$.42712 \leq ARE(\text{MME/PMLE}) < 1.$$

From this computation, we are led to conjecture that this pseudo maximum likelihood estimate of θ is uniformly more efficient asymptotically than the moment estimator of θ.

We obtained similar results for the pseudo MLE of the binomial parameter using the moment estimator of the Poisson parameter. Specifically, for the same parameter set,

$$.63983 \leq ARE(\text{PMLE/MLE}) < 1$$

$$.34036 \leq ARE(\text{MME/PMLE}) < 1.$$

4. Discussion. This paper advances a method of estimating a subset of parameters in multiparameter models, and investigates its numerical and asymptotic characteristics. We have given general conditions under which the PMLE of a single structural parameter is consistent and asymptotically normal. For the signal plus noise models discussed in Section 3, we have demonstrated the numerical feasibility of the approach and have verified that the asymptotic properties of the general method obtain. Moreover, a numerical investigation suggests that pseudo maximum likelihood estimates lie strictly between the MLE and the method of moments estimate for the signal parameter in terms of asymptotic efficiency. In general, we view the process of solving a reduced system of likelihood equations as a technique which promises to improve the asymptotic behavior of estimates of specific parameters. The applicability of the method extends well beyond the convolution models we have discussed. Applications that seem promising include pseudo maximum likelihood estimation in mixture models, multiparameter models without a closed form MLE (for example, PMLE's for the scale parameter of the Weibull distribution have been recommended for use—see Johnson and Kotz (1970)), and "regular" nuisance parameter models.

It is worthwhile calling attention to the fact that the asymptotic theory developed here is not directly comparable to the asymptotic theory for maximum likelihood estimates or, more generally, for best asymptotically normal (BAN) estimates. In a sense, the requirements on the model are slightly less stringent (vis-à-vis the nuisance parameter) for pseudo maximum likelihood estimation than for these other methods. A full discussion of the differences in regularity conditions is somewhat academic, however, since the models we have examined are fully regular under either theory, as are many models of interest. It remains true that the method of pseudo maximum likelihood estimation may be applicable in situations where the standard theory for MLE's or BAN estimates breaks down. Be that as it may, it is clear that the main virtue of PMLE's is their tractability in problems in which optimal approaches are computationally unfeasible.

An iterative procedure based on the method of pseudo maximum likelihood estimation may provide more efficient estimates than the single iteration we have discussed. In a two parameter model $F_{\theta,p}$, for example, one might estimate p by the method of moments, and obtain alternately the pseudo MLE of θ, then of p, and so forth. This algorithm guarantees that the likelihood increases with each iteration. Under regularity conditions, the algorithm should converge to the MLE, but the speed of this convergence may preclude its use.

Added in proof: L. B. Klebanov has called our attention to his note with I. A. Melamed entitled "On a certain method of construction of statistical estimates of the parameters of families of distributions" (1977) published (in Russian) in *Bull Acad. Sci. Georgian SSR*, 87, 553-4. In that note, a result giving conditions for consistency and asymptotic normality of a solution of a certain pseudo estimating equation is stated without proof. Their regularity conditions differ from ours.

REFERENCES

ANDERSEN, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimates. *J. Roy. Statist. Soc. B* **32** 283-301.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355-66.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269-76.

GAFFEY, W. R. (1959). A consistent estimator of a component of a convolution. *Ann. Math. Statist.* **30** 198-205.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63** 277-84.

GODAMBE, V. P. (1977). Estimating equations. Invited paper, 40th annual meeting Inst. Math. Statist. Seattle, Washington.

GODAMBE, V. P. and THOMPSON, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.* **2** 568-71.

GONG, G. and SAMANIEGO, F. (1978). Pseudo maximum likelihood estimation: theory and applications. Tech. Report, Div. Stat., Univ. California, Davis.

JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in Statistics; Continuous Univariate Distributions-I*, Wiley, New York.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. B* **31** 175–208.

LIU, P.-Y. and CROWLEY, J. (1978). Large sample theory of the MLE based on Cox's regression model for survival data. Tech. Report, Wisconsin Clinical Cancer Center, Univ. Wisconsin, Madison.

MANN, H. B. and WALD, A. (1943). On stochastic limit and order relationships. *Ann. Math. Statist.* **14** 217–26.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics, The Harald Cramér Volume* (U. Grenander, ed.). Wiley, New York.

SAMANIEGO, F. J. (1976). A characterization of convoluted Poisson distributions with applications to estimation. *J. Amer. Statist. Assoc.* **71** 475–79.

SAMANIEGO, F. J. (1980). Maximum likelihood estimation for binomially distributed signals in discrete noise. *J. Amer. Statist. Assoc.* **75** 117–121.

SCLOVE, S. L. and VAN RYZIN, J. (1969). Estimating the parameters of a convolution. *J. Roy. Statist. Soc. B* **31** 181–91.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DIVISION OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA 95616