THE MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS CONTAINED WITHIN FINITELY BOUNDED COMPACT SETS: SOME PRELIMINARY RESULTS

by

James B. Ramsey

NO. 81-23 September 1981

THE MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS CONTAINED WITHIN FINITELY BOUNDED COMPACT SETS: SOME PRELIMINARY RESULTS

By

James B. Ramsey

New York University

THE MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS CONTAINED WITHIN FINITELY BOUNDED COMPACT SETS: SOME PRELIMINARY RESULTS*

Despite the great generality inherent in the definition of maximum likelihood estimators, the actual use and the derivation of "optimal" properties have been restricted, with few exceptions, to those cases where regularity conditions hold. The continued fascination of maximum likelihood estimators for econometricians lies in the challenge to weaken the regularity conditions and to derive properties of the estimators in more general situations than formerly. With few exceptions, Hammersly (1950), Hanson (1965), Hocking (1965), and Moran (1971), the attention of researchers has been restricted to the problems raised by the algebraic form of the parent cumulative distribution function (c.d.f.) or, when it exists, the probability density function (p.d.f.): see, for example, the latest efforts along these lines Huber (1967) and Weiss (1971). Almost all of these papers have ignored the topological structure of the parameter space, assuming invariably that it is an open set contained within a compact subset of the appropriately dimensioned Euclidean space. It is interesting to note that the scattered exceptions examined a number of special cases and none of the authors referred to the others.

---

*The research discussed in this paper was supported in part by the National Science Foundation Grant No. GS-3291 and by the Social Science Research Council (U.K.) Grant No. HR 2166/1. Their support is gratefully acknowledged.

In this paper I intend to concentrate on the effect of changing these "regularity" assumptions about the parameter space on the distribution of maximum likelihood estimators. More precisely, the basic assumption made in this paper is that the parameter space is compact and finitely bounded.[1] As will be shown, this seemingly minor change has extensive and significant effects on the distribution of maximum likelihood estimators and leads to definitions of alternative estimators with lower mean squared errors than the conventional maximum likelihood estimator.

The results in this paper are important to econometric research for three basic reasons:

(i) Economic theory frequently indicates the existence of inequality constraints on parameters and on the values which endogenous variables may assume,

(ii) In many important applications, theory indicates that parameters and endogenous variables are restricted to a number of isolated points, e.g. a dependent variable may only take on integer values,

(iii) The discipline of economics has in many areas progressed to the stage in which the imposition of inequality and other constraints is indicated by prior research efforts and it is suspected at least that the use of such empirical information would improve the quality of inferences.

So far in the econometric literature inequality and "isolated point" constraints have not been used as extensively (and intensively for that matter) as might have been the case. This is due in part to the absense of an adequate theoretical analysis of the properties of such estimators. This paper is one further step in the direction of providing such an analysis.

The outline of the remainder of the paper is quite straight forward. The first section states the problem formally, defines three alternative estimators, relates the current analysis to the existing literature, and discusses the economic relevance of the problem. The second section comments

on the respective statistical properties of the estimators, while the third section examines the results of some sampling runs which serve both to illustrate the theory in the first section and give some insights into the potential results of further and more extensive formal analysis. The work reported in this paper is to be regarded mianly as exploratory and preliminary, although I trust the reader will find it both interesting and potentially useful.

I. The Problem Stated and Some Alternative Estimators Defined

Some examples will be helpful in further illustrating the problem and to indicate the usefulness of considering parameters constrained to lie in compact sets. First, the obvious examples are those in which parameter values are constrained by inequalities, the most usual situation being that $0 \leq \theta_i \leq 1$, where θ_i denotes the ith parameter. Economic examples are marginal propensities to consume or export, the distributed lag parameter in adaptive expectations or in partial adjustment hypotheses, bounds on production coefficients, coefficients representing threshold effects or minimum consumption levels, etc.

With dependent variables it is often the case that not only is the range of the variable bounded, but that the boundary values are often taken with positive probability. This is opposed to the standard logistics curve approach in which it is assumed that the boundaries represent suprema (or infima) of functions so that the variable approaches the boundary asymptotically. Economic examples would involve fixed minimum size of purchase order, pricing decisions with price or wage controls, budget deficit limits, fixed capacity output limits, etc.

Many other situations involve restricting parameter or dependent variable values to a finite number of discrete values. For example a parameter may be defined by the ratio k/n where, k, n are integers, $k \leq n$, and k is unknown and to be estimated; or one wishes to estimate which one of a finite set of alternative strategies decision-makers will choose. Many variables should naturally be restricted to integer values; estimating the number of plants or machines in a firm, number of firms in an oligopolistic industry, number of aircraft in a run-way queue, etc.

Lastly, the suggestion has been frequently made that various inequality constraints obtained from prior research should be incorporated in later estimates and hypotheses tests in order to improve estimating and test efficiency. However intuitively appealing this recommendation is, the gains from such a procedure are not generally clear, especially in view of the increased cost in complexity of estimation thereby engendered. Theoretical analysis of the effects of incorporating inequality constraints into one's inferences will provide the tools necessary for answering such questions.

Let $L(\underline{\theta}|\underline{x})$ denote the likelihood function obtained from an assumed parent p.d.f. $f(X|\underline{\theta})$ and a random sample of size n on the random (vector) variable X whose n-fold realization is represented by $\underline{x}$. The sample space is a subset of R^n, n dimensional Euclidean space, and is assumed to be independent of θ. $\underline{\theta}$ is a p dimensional parameter vector assumed to be contained in a parameter space Θ which is itself assumed to a compact subset of R^p. Let $\ell(\underline{\theta}|\underline{x})$ denote the ln likelihood function obtained from $L(\underline{\theta}|\underline{x})$.

The maximum likelihood estimator $\hat{\underline{\theta}}$ is defined quite simply by:

$$\ell(\hat{\underline{\theta}}|\underline{x}) = \sup_{\underline{\theta} \in \Theta} \ell(\underline{\theta}|\underline{x}). \tag{1}$$

When $f(x|\underline{\theta})$ is continuous in $\underline{\theta}$ for all $\underline{x}$ and Θ is assumed to be compact, the supremum exists, so that we do not need to consider Rao's "near maximum likelihood" estimators.

The only property which follows from the definition (1) and some very weak regularity conditions is that of consistency as proved by Wald (1949).[2] Desirable small sample properties, if any exist with respect to a given p.d.f., have only been derived under the assumption that Θ is open and even under this assumption, the properties stem from the fact that maximum likelihood estimators are functions of the sufficient statistics when they exist, and sometimes can be shown to be minimally sufficient themselves. Asymptotic properties, more particularly, efficiency (in the sense of attaining the Cramer-Rao lower bound) and normality of the distribution require more strigent regularity conditions than does consistency. In the derivation of these properties, the openness of Θ is crucial, see, for example, Daniels (1960). Provided Θ is open, then for sufficiently large n, an open neighborhood about $\underline{\theta}_o$, the true parameter point, can be found such that within it the first two total derivatives of the likelihood with respect to $\underline{\theta}$ exist, and the first partial derivatives vanish except over sets of measure zero. When Θ is bounded and if $\underline{\theta}_o$ lies on the boundary, this regularity condition no longer holds, even asymptotically.

An important regularity condition on the compact parameter space Θ which will be imposed throughout the remainder of the paper is that Θ is convex or that it contains only a denumerably infinite number of points. In order to relate the estimators which incorporate knowledge of Θ to the unconstrained maximum likelihood estimator which does not, it is mathematically convenient to assume that $\Theta \subset \Theta_o$, where Θ_o is an open set, and that for all $\underline{x}$ (except over a set of measure zero--a condition hereafter referred

to as "almost everywhere") the supremum (over $\underline{\theta} \in \Theta_o$) of the likelihood function exists.

If $\underline{\lambda}$ denotes an r dimensional vector of Lagrange multipliers, it is notationally convenient to define $\Theta_\Lambda \subset R^{(p+r)}$, where Θ_Λ is the Cartesian product of the parameter spaces Θ and Λ, $\underline{\lambda} \in \Lambda$.

Three estimators will be considered: unconstrained maximum likelihood (UML), constrained maximum likelihood (CML), and the minimum distance estimator (MD).

The unconstrained estimator, to which the other estimators will be related, is defined in the conventional way for $\underline{\theta} \in \Theta_o$, Θ_o an open subset of R^p which contains Θ. Since, the intent of this paper is to concentrate on the effects of making Θ compact, it will be assumed that the full set of regularity conditions needed to ensure that the UML estimator is asymptotically efficient and normally distributed do in fact hold for all $\underline{\theta} \in \Theta_o$; a useful reference is Daniels (1960). Under these conditions, the UML estimator $\hat{\underline{\theta}}_u$ is defined by:

$$\ell(\hat{\underline{\theta}}_u | \underline{x}) = \max_{\underline{\theta} \in \Theta_o} \ell(\underline{\theta} | \underline{x}), \tag{2}$$

and $\hat{\underline{\theta}}_u$ can be obtained as a root of the normal equations $\partial \ell(\underline{\theta} | \underline{x}) / \partial \underline{\theta} = \emptyset$ almost every where for sufficiently large n.

The constrained maximum likelihood estimator (CML) $\hat{\underline{\theta}}_c$ is defined by:

$$\ell(\hat{\underline{\theta}}_c | \underline{x}) = \sup_{\underline{\theta} \in \Theta} \ell(\underline{\theta} | \underline{x}). \tag{3}$$

For parameter spaces which are at most piecewise continuous, like that shown in Figure 1, there are no simple analytical solutions and the process of obtaining an actual value for $\hat{\underline{\theta}}_c$ from a given sample reduces

essentially to some form of search procedure. However, if $\hat{\underline{\theta}}_u \in \Theta$, then $\underline{\theta}_c = \hat{\underline{\theta}}_u$, otherwise $\hat{\underline{\theta}}_c$ is found by searching along the boundary to θ.

If the boundary of θ can be characterized in terms of a set of r continuous differentiable functions $g_j(\underline{\theta})$, $j = 1, 2, \ldots r$, $r < p$, then the constrained optimization problem can be formulated as:

$$\max_{(\underline{\theta}, \underline{\lambda}) \in \Theta_\Lambda} \ell(\underline{\theta}|\underline{x}) - \underline{\lambda}'\underline{g}(\underline{\theta}) \quad , \tag{4}$$

where $\underline{g}(\underline{\theta})$ is the r dimensional vector of constraints defining the boundary of Θ and $\underline{\theta}$ must satisfy for each j, $j = 1, 2, \ldots r$, $g_j(\underline{\theta}) \leq 0$ and $\underline{\lambda}$ is the vector of Lagrange multipliers. If the inequalities are replaced by equalities, one has the Aitchison and Silvey (1958) problem. A slightly more general problem is specified by assuming that Θ imposes non-negativity (or equivalently non-positivity) constraints on $\underline{\theta}$.

The solution to the most general problem is given by solving the Kuhn-Tucker conditions:

$$\begin{aligned}
&\partial\ell(\underline{\theta}|\underline{x})/\partial\underline{\theta} - \underline{\lambda}'\partial\underline{g}(\underline{\theta})/\partial\underline{\theta} \leq \underline{\emptyset} && \text{(a)}\\
&(\partial\ell(\underline{\theta}|\underline{x})/\partial\underline{\theta} - \underline{\lambda}'\partial\underline{g}(\underline{\theta})/\partial\underline{\theta})\underline{\theta} = 0 && \text{(b)}\\
&\underline{\theta} \geq \underline{\emptyset} && \text{(c)}\\
&\underline{g}(\underline{\theta}) \leq \underline{\emptyset} && \text{(d)}\\
&\underline{\lambda}'\underline{g}(\underline{\theta}) = 0 && \text{(e)}\\
&\underline{\lambda} \geq \underline{\emptyset} && \text{(f)}
\end{aligned} \tag{5}$$

If the inequality constraint (c) is not imposed then inequality (a) becomes an equality and equation (b) is redundant. If the inequality (d)

is in fact an equality, then equation (e) is redundant and the inequality (f) no longer applies. The Aitchison and Silvey problem can be characterized in terms of conditions (5) by stating that in their problem (c) does not apply and that (d) is an equality. Consequently, the solution is characterized by (a) being an equality and (f) not applying. Under these conditions, Aitchison and Silvey were able to show that the vector $(\hat{\underline{\theta}}_c, \hat{\underline{\lambda}})$, the constrainted maximum likelihood estimator of the vector $(\underline{\theta}, \underline{\lambda})$, is efficient and asymptotically distributed as (p + r)-variate normal. If inequality (5(c)) does not apply, then $\hat{\underline{\theta}}_c = \hat{\theta}_u$ if $\underline{\lambda} = \emptyset$, i.e., if $\underline{g}(\hat{\underline{\theta}}_c) < \emptyset$; otherwise $\hat{\underline{\theta}}_c$ lies on the boundary and is not equal to $\hat{\underline{\theta}}_u$.

In terms of the general formulation in equation (5) Hanson (1965) extended the Aitchison and Silvey results by giving the conditions required to prove the existence of the constrained maximum likelihood estimator and its convergence in probability to the unique solution of the system of equations (5). Moran (1971), apparently unaware of Hanson's work, derived under quite strong regularity conditions the asymptotic distribution of the constrained maximum likelihood estimators of a set of parameters θ_i, i = 1, 2, ..., of a continuous unimodal p.d.f. $f(X, \theta)$ defined with respect to a vector random variable X where $\theta_1^{\,o} = 0$ and $0 < \theta_i^{\,o} < b_i$, i = 2, ... k, $\underline{\theta}^o$ denotes the true parameter vector.

If $\underline{z}_n(\theta) = n^{1/2}(\hat{\underline{\theta}} - \underline{\theta})$, where $\hat{\underline{\theta}}$ is the constrained maximum likelihood estimator of $\underline{\theta}$ and the matrix $\{c_{ij}\}$ is defined by:

$$\{c_{ij}\} = -E_{\underline{\theta}}\left\{\frac{\partial^2 \ln f(X,\theta)}{\partial\theta_i \partial\theta_j}\right\} ,$$

then Moran's chief result is the following theorem; Moran (1971, 444).

Theorem.--Suppose that $\theta^1 = 0$, and $0 < \theta^i < b_i$ for $i = 2, \ldots, k$. Then the distribution function

$$\phi_n(\underline{t}, \underline{\theta}) = \text{Prob } (\underline{z}_n < \underline{t}, \underline{\theta})$$

converges, uniformly in $\underline{t}$ and $\underline{\theta}$, towards the mixture of distributions

$$\text{Prob } (\underline{z} < \underline{t}, \underline{\theta}) = \frac{1}{2} F_1(\underline{t}, \underline{\theta}) + \frac{1}{2} F_2(\underline{t}, \underline{\theta}),$$

where $F_1(\underline{t}, \underline{\theta})$ is a k-dimensional multivariate distribution defined on the region

$$t^1 > 0, \quad -\infty < t^i < \infty \qquad (i = 2, \ldots, k),$$

and having in this region a probability density equal to twice the density of a multivariate normal distribution with means zero, and covariance matrix equal to (σ_{ij}). $F_2(\underline{t}, \underline{\theta})$ is a $(k - 1)$-dimensional distribution concentrated on the subspace $t^1 = 0$, $-\infty < t^i < \infty$ for $i = 2, \ldots, k$, and such that the joint distribution of $z^2, \ldots, z^k$ is that of the quantities

$$z^i = \sum_{j=2}^{k} \sigma_{ij}^{(1)} y^j \qquad (i = 2, \ldots, k),$$

where $y^1, \ldots, y^k$ are jointly normally distributed with zero means and covariance matrix (c_{ij}), the distribution of $y^2, \ldots, y^k$ being taken conditional on the inequality

$$y^1 - \sum_{j=2}^{k} c_{1j} \sum_{s=2}^{k} \sigma_{js}^{(1)} y^s \leq 0,$$

where the $\sigma_{ij}^{(1)}$ are the elements of the matrix

$$(\sigma_{ij}^{(1)}) = \begin{pmatrix} c_{22} & \cdots & c_{2k} \\ \vdots & & \vdots \\ c_{k2} & & c_{kk} \end{pmatrix}^{-1}$$

Furthermore, the convergence to the limiting distribution is uniform in $\underline{t}$ and in the subset of Ω_1 given by $\theta_1 = 0$. $F_2(t, \underline{\theta})$ is not distributed as normal.

The last estimator to be considered is the minimum distance (MD) estimator $\hat{\underline{\theta}}_m$ defined by:

$$||\hat{\underline{\theta}}_u - \hat{\underline{\theta}}_m||^2 = \min_{\underline{\theta} \in \Theta} ||\hat{\underline{\theta}}_u - \underline{\theta}||^2, \qquad (6)$$

where $||\underline{a}||$ denotes the Euclidean norm of the vector $\underline{a}$. If $\hat{\underline{\theta}}_u \in \Theta$, then the norm is zero and $\hat{\underline{\theta}}_m = \hat{\underline{\theta}}_u$, otherwise the norm is non-zero and $\hat{\underline{\theta}}_m$ lies on the boundary. $\hat{\theta}_m$ is characterized by the relation that the vector $(\hat{\theta}_u - \hat{\theta}_m)$ is orthogonal to the tangent plane to Θ at the point $\hat{\underline{\theta}}_m$ and that $\hat{\underline{\theta}}_m$ is unique; these results follow when Θ is convex. If Θ is composed of a finite number of points, $\hat{\underline{\theta}}_m$ need not be unique.

Hammersly (1950) completed an intensive analysis of the distributions of two special uniparameter cases of the minimum distance estimator where it is formally identical to the constrained maximum likelihood estimator; these two situations may be regarded as examples of the use of Theorem 2 proved in the next section. The first and more interesting problem examined by Hammersly involved the estimation of the mean of a normal distribution with known variance σ^2 with a random sample of size n. The unknown mean is restricted to an integral lattice, i.e. $\mu \in \{0, \pm 1, \pm 2, \ldots\}$. If $\bar{x}$ is the sample mean, the minimum distance (constrained maximum likelihood) estimator

is given by m, $m = n.i.(\bar{x})$, where n.i.(.) indicates taking the nearest integer. The distribution obtained is a special case of that derived in Section II of this paper. Of particular interest is the asymptotic approximation to the sampling variance of m which is (Hammersly (1950, 192)):

$$\text{var}(m) \sim \left(\frac{8\sigma^2}{\pi n}\right)^{1/2} \text{Exp}\left(-n/(8\sigma^2)\right), \text{ as } n/\sigma^2 \to \infty.$$

The estimator converges very rapidly indeed to its unbiased expectation since the variance is decreasing to order e^{an}, where a is a positive constant! This property is illustrated in the examples discussed in Section III.

Hocking also considered the use of the minimum distance estimator in the case of a linear regression model with normally distributed disturbance terms with the parameter space constrained by a continuous everywhere differentiable boundary function. The result is a special case of the distributional results discussed in the next section of the paper.

Of some interest is the relationship between $\hat{\underline{\theta}}_c$ and $\hat{\underline{\theta}}_m$, one aspect of which is illustrated in Figure 1. One may informally characterize the difference between the estimators by stating that whereas $\hat{\underline{\theta}}_c$ maximizes the likelihood subject to the constraint, $\hat{\underline{\theta}}_m$ is the closest point in Θ to the unconstrained maximum. Still speaking loosely, it is clear that (in two dimensional space) the values taken by the two estimators will differ most markedly when the angle between the major "axes" of the likelihood contour and of the parameter space is $\pi/4$.

Clearly all three estimators are functions of "n," sample size, and strictly speaking $\hat{\underline{\theta}}_u$, $\hat{\underline{\theta}}_c$, $\hat{\underline{\theta}}_m$ should be written $\hat{\underline{\theta}}_u(n)$, $\hat{\underline{\theta}}_c(n)$, $\hat{\underline{\theta}}_n(n)$ to stress

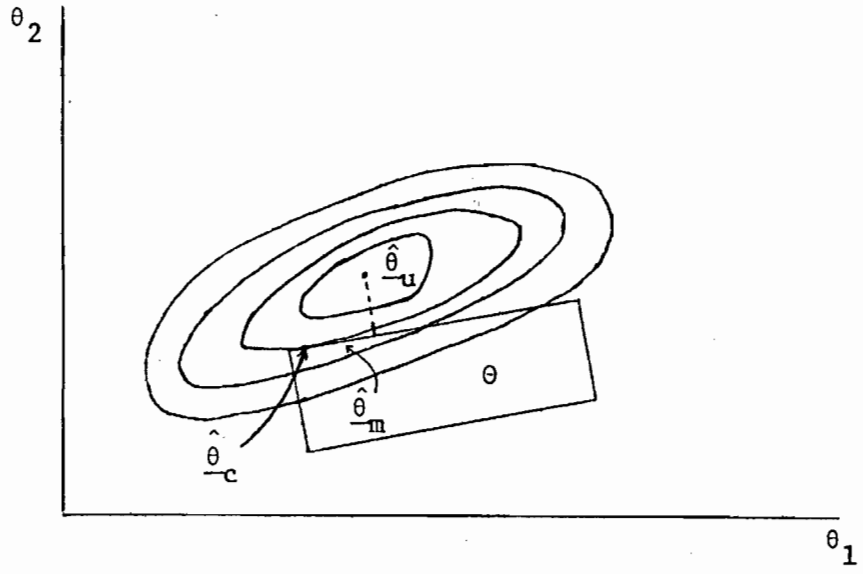

Figure 1.--Illustration of the Relationship Between $\hat{\theta}_u$, $\hat{\theta}_c$, and $\hat{\theta}_m$ when $\hat{\theta}_u \notin \Theta$.

that fact. However, it is expositionally convenient to suppress the notation on "n" which will be followed throughout the paper.

II. The Statistical Properties of $\hat{\underline{\theta}}_u$, $\hat{\underline{\theta}}_c$, $\hat{\underline{\theta}}_m$

As a first remark in this section, we should note the distinction between the theoretical analysis of the distributions of the estimators and the algorithms required to produce estimates from actual data. Although, as will be shown, the theoretical analysis is at times complex, the required algorithm for $\hat{\underline{\theta}}_m$ is relatively simple; in short, the theoretical justification for $\hat{\underline{\theta}}_m$ is much more difficult than the actual use of the estimator. The problem involved in finding suitable algorithms for obtaining estimates $\hat{\underline{\theta}}_c$ have been widely discussed elsewhere, see for example Judge and Takayama (1966).

Under the assumptions put forward in the first section, it is intuitively clear that no matter what the specification of the parameter space Θ, $\hat{\underline{\theta}}_u$ is asymptotically distributed as multi-variate normal and is efficient. This result follows from consistency, which is easily proved below, and the regularity conditions which are assumed to apply. It is further clear that the minimal sufficiency of $\hat{\underline{\theta}}_u$ is not affected by the specification of θ.

Consistency of the Estimators

The first step in the comparison of the three estimators is to demonstrate that both $\hat{\underline{\theta}}_c$ and $\hat{\underline{\theta}}_m$ are consistent. Wald's 1949 proof of consistency utilizes the fact that $\ell(\hat{\underline{\theta}}|\underline{x}) \geq \ell(\underline{\theta}_o|\underline{x})$ where $\underline{\theta}_o$ denotes the true parameter point. By using the strong law[3] of large numbers, Wald proved that for all $\underline{\theta} \in \Theta$:

$$\lim_{n\to\infty} \Pr(\ell(\hat{\underline{\theta}}|\underline{x}) < \ell(\underline{\theta}_o|\underline{x})) = 1, \tag{7}$$

and combining the two inequalities one obtains the result:

$$\lim_{n\to\infty} \Pr(\hat{\underline{\theta}} = \underline{\theta}_o) = 1. \tag{8}$$

Since Wald's proof relies directly on the compactness of Θ, $\hat{\underline{\theta}}_c$ is consistent. $\hat{\underline{\theta}}_u$ is consistant since $\ell(\hat{\underline{\theta}}_u|\underline{x}) \geq \ell(\underline{\theta}_o|\underline{x})$ holds *a fortiori*.

Because the space Θ is a metric space and $\hat{\underline{\theta}}_m$ is defined by $\min_{\underline{\theta} \in \Theta} ||\hat{\underline{\theta}}_u - \underline{\theta}||$, the consistency of $\hat{\underline{\theta}}_u$ implies that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(||\hat{\underline{\theta}}_u - \underline{\theta}_o|| < \varepsilon) = 1. \tag{9}$$

Since $||\hat{\underline{\theta}}_u - \hat{\underline{\theta}}_m|| \leq ||\hat{\underline{\theta}}_u - \underline{\theta}_o||$, it follows immediately that $\hat{\underline{\theta}}_m$ is consistent.

Some Mean Squared Error Properties

Before deriving the probability distribution functions (p.d.f.s) of the estimators defined in the previous section, it is useful to consider as generally as possible the degree of mean squared error gain from imposing inequality constraints. Although the mean squared error result for the single parameter situation is obviously a special case of the multi-parameter case, it is expositionally convenient to begin by proving some lemmas and theorems in the former case.

Mean Squared Error for a Single Parameter

In this situation Θ, the parameter space, is a compact sub-set of R and it is assumed that either (i) θ contains an infinite number of points and is convex, or (ii) θ contains at most a finite number of points; (the extension to a denumerably infinite set of points is straight forward).

Let t be any unconstrained estimator with finite second moment of $\theta_o \in \Theta$ with p.d.f. $f(t|\theta_o)$; t is said to be unconstrained if its range is $(-\infty, \infty)$, i.e. t belongs to a class of estimators not restricted to a proper sub-set of R. When Θ is a compact convex set it can be represented by the interval [a, b]. Let the minimum distance constrained estimator $\tilde{t}_m$ be defined by:

$$\tilde{t}_m = \begin{cases} t, & \text{if } t \in [a, b] \\ a, & \text{if } t \leq a \\ b, & \text{if } t \geq b \end{cases} \tag{10}$$

In the uniparameter case, the estimator $\tilde{t}_m$ has been recognised for some time; in particular Zellner suggested the use of this estimator several years ago, Zellner (1961).

Theorem 1.

$$MSE(\tilde{t}_m) \leq MSE(t)$$

Proof.--Let θ_o be the true value of the parameter, $\theta_o \in \Theta$ and let $\bar{\Theta}$ denote the complement of Θ in R, then:

$$MSE(t) = \int_{-\infty}^{\infty} (t - \theta_o)^2 f(t|\theta_o)dt \tag{11}$$

$$= \int_{\Theta} (t - \theta_o)^2 f(t|\theta_o)dt + \int_{\bar{\Theta}} (t - \theta_o)^2 f(t|\theta_o)dt,$$

and

$$MSE(\tilde{t}_m) = \int_{-\infty}^{\infty} (\tilde{t}_m - \theta_o)^2 f(t|\theta_o)dt \tag{12}$$

$$= \int_{\Theta} (t - \theta_o)^2 f(t|\theta_o)dt + \int_{-\infty}^{a} (a - \theta_o)^2 f(t|\theta_o)dt + \int_{b}^{\infty} (b - \theta_o)^2 f(t|\theta_o)dt.$$

Since $(t - \theta_o)^2 \geq (a - \theta_o)^2$ for all $t \leq a$ and $(t - \theta_c)^2 \geq (b - \theta_o)^2$ for all $t \geq b$ and the first two integrals in equations (11) and (12) are identical, the theorem follows immediately.

Now let t be the unrestricted maximum likelihood equation estimator of $\theta_o \in \Theta$, i.e. t is defined by the solution to the normal equation:

$$\frac{d\ell}{d\theta}(\theta|X) = 0, \tag{13}$$

where $\ell(\theta|X)$ denotes the log likelihood function and X a set of observed data points. We assume that the usual regularity conditions hold for $\ell(\theta|X)$ for all $\theta \in \Theta_o$, Θ_o an open set containing Θ; see for example Daniels (1960, p. 152), so that t is asymptotically distributed as normal with mean θ_o and variance given by the inverse of the information measure. The corresponding constrained maximum likelihood estimator $\tilde{t}$ is defined by:

$$\tilde{t} = \begin{cases} t, \text{ if } t \in [a, b] \\ a, \text{ if } t \notin [a, b] \text{ and } \ell(a|X) \geq \ell(b|X) \\ b, \text{ if } t \notin [a, b] \text{ and } \ell(a|X) < \ell(b|X) \end{cases} \tag{14}$$

The definition given in equation (14) is equivalent under the given conditions to the standard formulation, namely where t is defined by the solution to equation (15):

$$\ell(\tilde{t}|X) = \sup_{t \in \Theta} \ell(\theta|X). \tag{15}$$

Lemma 1.--Under condition I(3) (Daniels, 1960, p. 152), namely the condition that $\partial\ell(\theta|X)/\partial\theta$ is a nowhere increasing and somewhere decreasing function of θ, then, when $t \notin [a, b]$, $\ell(a|X) \geq \ell(b|X)$ is a necessary and

sufficient condition for $t < a$, and $\ell(a|X) < \ell(b|X)$ is a necessary and sufficient condition for $t > b$.

Proof.--Need only prove the result for $t < a$ since the proof for $t > b$ is similar. If $t < a$, then by the conditions of the lemma $\ell(a|X) \geq \ell(b|X)$. If the inequality $\ell(a|X) \geq \ell(b|X)$ holds, $\ell(a|X) > \ell(b|X)$ for otherwise by the conditions of the lemma $\ell(\theta|X)$ is constant over the interval [a, b] and $t \in [a, b]$ which contradicts the assumption that $t \notin [a, b]$. Further, if $\ell(a|X) > \ell(b|X)$, then given the assumed regularity condition that $\ell(\theta|X)$ is continuous in Θ_o, for almost all X, there exists an δ neighborhood of "a," such that $\ell(\theta|X)$ is a decreasing function of θ so that $\ell(\theta|X) > \ell(a|X)$ implies $\theta < a$ in the δ neighborhood, in short, $t < a$.

Theorem 2.--Under the conditions of Lemma 1 where t is an unconstrained maximum likelihood equation estimator, the constrained maximum likelihood estimator $\tilde{t}$ and the minimum distance estimator $\tilde{t}_m$ are equivalent in the sense that for any value of t, $\tilde{t} = \tilde{t}_m$. Further, the mean squared error of $\tilde{t}$ is equal to that of $\tilde{t}_m$ which is less than or equal to that of t.

The proof of this theorem follows in an obvious manner from Lemma 1 and Theorem 1.

The proof of Theorem 1 depends upon the restrictive assumption that for all $\theta \in \Theta_o$, $\ell(\theta|X)$ is concave in θ. One can replace the condition I(3) in Lemma 1 by a much weaker condition, but under such conditions the conclusions of Theorem 1 can be shown to hold only asymptotically. Accordingly, consider Lemma 2.

Lemma 2.--Following Daniels (1960, p. 155) assume: (i) At every $\theta \in \Theta_o$, $\partial \ell(\theta|X)/\partial\theta$ exists for almost all X and is not almost everywhere zero. $\ell(\theta|X)$ is discontinuous in θ at at most a finite number of points

at which points the discontinuities are finitely valued. (ii) The probability that the interval $(\theta, \theta + \Delta\theta)$ contains a discontinuity point of $\partial \ell(\theta|X)/\partial\theta$ is of order not exceeding $|\Delta\theta|$ for any true $\theta_o \in \theta_o$.

Under assumptions (i) and (ii) (together with some further minor regularity conditions, see Daniels (1960, p. 155)), and when t, the unconstrained maximum likelihood estimator, $\notin [a, b]$, then, as n (sample size) $\to \infty$, the probability that each of the following statements holds approaches 1:

(i) $\ell(t|X) \geq \ell(a|X)$ is a necessary and sufficient condition for $t < a$;

(ii) $\ell(t|X) \geq \ell(b|X)$ is a necessary and sufficient condition for $t > b$.

(iii) In addition, $\tilde{t}_m$ as defined in equation (1) is equivalent to $\tilde{t}$, where $\tilde{t}$ is the constrained maximum likelihood estimator as defined in (17), and where equivalence is defined in Theorem 2.

Proof.--Under the assumptions of Lemma 2, Daniels showed that for given $\alpha > 0$, $\delta > 0$:

$$\lim_{n\to\infty} \Pr\left\{\theta_o - \alpha n^{-\frac{1}{2}+\delta} < t < \theta_o + \alpha n^{-\frac{1}{2}+\delta}\right\} = 1,$$

or $$\lim_{n\to\infty} \Pr\left\{t \in \ell_n | \theta_o\right\} = 1,$$

where $\ell_n = (\theta_o - \alpha n^{-\frac{1}{2}+\delta}, \theta_o + \alpha n^{-\frac{1}{2}+\delta})$. Further, for all $\theta \in \ell_n$, $\partial \ell(\theta|X)/\partial\theta$ lies within a narrow band of slope $-nI(\theta_o)$, where $I(\theta_o)$ is the information measure, with probability approaching 1 as $n \to \infty$. In short $\partial \ell(\theta|X)/\partial\theta$ can be approximated arbitrarily closely by a nowhere increasing and somewhere decreasing function. The meaning of (iii) is that under the given assumptions $\tilde{t}_m$ converges almost surely to $\tilde{t}$.

Suppose $\theta_o \in (a, b)$, then, given $\alpha > 0$, $\delta > 0$, as $n \to \infty$, $Pr\{t \in \ell_n\} < Pr\{t \in (a, b)\} \to 1$ and $Pr\{t \notin (a, b)\} \to 0$; i.e. $Pr\{\tilde{t} = a, \text{ or } b \mid \theta_o \in (a, b)\} \to 0$.

Suppose $\theta_o = a$, then from the results in the previous paragraph, as $n \to \infty$:

(a) $Pr\{t \notin m_\delta(a)\} \to 0$, where $m_\delta(a)$ denotes any neighborhood of "a" of radius δ.

(b) For $t \in \ell_n$, ℓ_n centered at "a," $\partial \ell(\theta/X)/\partial\theta$ can be approximated arbitrarily closely by a nowhere increasing and somewhere decreasing function.

$\tilde{t}_m$ is equal to $\tilde{t}$ for $t \in (a, b)$, by definition of $\tilde{t}_m$ and $\tilde{t}$. $\tilde{t}_m$ and $\tilde{t}$ can only differ for $t \notin (a, b)$ for if $\theta_o \in (a, b)$, then as $n \to \infty$, $Pr\{t \notin (a, b)\} \to 0$, so that $Pr\{|\tilde{t}_m - \tilde{t}| < \delta \mid \theta_o \in (a, b)\} \to 1$. Let $\theta_o = a$, since as $n \to \infty$, results (a) and (b) hold, then $Pr\{\tilde{t}_m = a\} \to 1$ and $Pr\{\tilde{t} = a\} \to 1$, so that $Pr\{|\tilde{t}_m - \tilde{t}| < \delta\} \to 1$. This concludes the proof of proposition (iii). Proposition (i) also follows from results (a), (b), and that $t \notin [a, b]$. Proposition (ii) follows from similar arguments when $\theta_o = b$.

Theorem 2'.--Under the conditions of Lemma 2, as $n \to \infty$, the probability limit is one that:

(a) $\tilde{t}$ is equivalent to $\tilde{t}_m$;

(b) the mean squared error of $\tilde{t}$ is equal to that of $\tilde{t}_m$ which is less that or equal to that of t.

The proof of Theorem 2' is an obvious extension of Lemma 2 and Theorem 2.

Given the simple nature of the above lemmas and theorems we might anticipate that the mean squared error gain[4] for constrained estimators in the case where Θ is composed of a finite number of points would be even

greater. However, despite Hammersly's results cited earlier, a general proof has not yet been found. The difficulty is illustrated in Figure 2 where we assume that $\Theta = \{a, b\}$. Let $c = (a+b)/2$ be the mid point of the interval. The minimum distance estimator $\tilde{t}_m$ takes on only two values; a if $t \leq c$, b if $t > c$, where t is some unconstrained estimator. The contributions to mean squared error for both estimators $(t, \tilde{t}_m)$ over the intervals $[-\infty, c]$, $[b, \infty]$ indicate that $\tilde{t}_m$ has smaller contribution to mean squared error over the indicated sub-region. However, little can be said in general, it seems, about the relative contributions to mean squared error over the interval [c, b]. Consequently, it appears that nothing can be said in general about the mean squared error gain for the minimum distance estimator. A similar tentative conclusion holds for the constrained maximum likelihood estimator for finite sample sizes.

Definite conclusions can be obtained asymptotically since Theorem 2' can be extended with only trivial modifications to handle the situation where Θ contains only a finite number of points. The reason is that for given $\alpha > 0$, $\varepsilon > 0$, one can obtain for sufficiently large n an interval ℓ_n which contains only the true value θ_o.

A special case can also be proven. Let us assume that the distribution of $\hat{\theta}_u$ is normal with mean θ_o and variance σ^2_n, where the subscript "n" stresses the fact that the sampling distribution of $\hat{\theta}_u$ depends on sample size n. Let us assume that $q = 2$ and without essential loss of generality that $\theta_o = a = 0$. Under these conditions, the expected value of θ_m is $b\pi_2$,

$$\pi_2 = \int_c^\infty \mathrm{Exp}(-u^2/2\sigma_n^2)/(2\pi)^{1/2}\sigma_n \,,$$

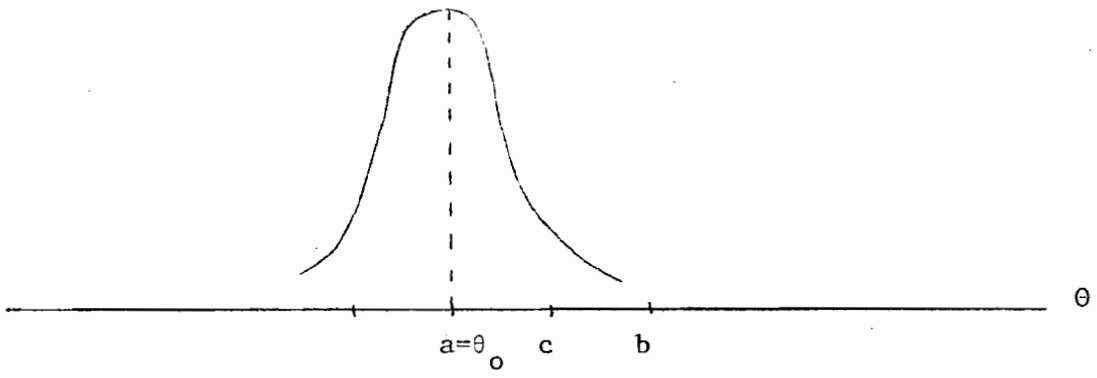

Figure 2.--Illustration of Constrained Estimator of Two Parameter Points Showing p.d.f. of $t|\theta_o = a$.

$c = (a + b)/2$ so that $\hat{\theta}_m$ is biased upward (but not asymptotically, since $\pi_2 \to 0$). The mean squared error of $\hat{\theta}_u$ is simply the variance of $\hat{\theta}_u$, σ^2_n. The mean squared error of $\hat{\theta}_m$, $E\{\hat{\theta}^2_m\}$ is given by:

$$E\{\hat{\theta}^2_m\} = a^2\pi_1 + b^2\pi_2 \tag{16}$$

$$= b^2\pi_2 = \frac{b^2}{\sqrt{2\pi}\sigma_n} \int_{b/2}^{\infty} \text{Exp}\{-u^2/(2\sigma^2_n)\}du,$$

and one wants to prove that $E\{\hat{\theta}^2_m\} < \sigma^2_n$ for all $b > 0$. An equivalent expression of the problem is to prove that:

$$b^2 \int_{b/2}^{\infty} \phi(u)du < 1, \text{ for all } b > 0, \tag{17}$$

where $\phi(u)du$ represents the p.d.f. of a normal variable with mean zero and unit variance.[5] Using an inequality cited in Rao (1965, p. 117), we have:

$$\int_{b/2}^{\infty} \phi(u)du \leq \frac{2}{b} \frac{\text{Exp}\{-b^2/8\}}{\sqrt{2\pi}}, \tag{18}$$

so that if the right hand side of the inequality in (18) can be shown to be less than b^{-2} for all b, the inequality in (17) follows. The inequality in (18) can be re-expressed as:

$$\frac{\sqrt{2\pi}}{2} \text{Exp}\{b^2/8\} > b. \tag{19}$$

It is easily shown that the function g(b) defined by:

$$g(b) = (\frac{\sqrt{2\pi}}{2} \text{Exp}\{b^2/8\} - b), \tag{20}$$

is everywhere strictly positive with a minimum at $b = 1.9675$ (minimum

$g(b) = 0.0658$). Consequently, the mean squared error for $\hat{\theta}_m$ is less for any alternative point in the two point problem. In passing, it is interesting to note from this analysis that the minimum gain in M.S.E. occurs at b = 1.9675 standard deviations from the true value. The extension to multiple points on the line is easily made.

Mean Squared Error in the Multiparameter Case

There are a number of alternative ways in which the concept of mean squared error can be extended to the situation in which we estimate a vector of parameters.

Following the discussion in the uniparameter case let us define Θ as the parameter space, where Θ is a compact subset of R^k so that elements of Θ can be represented as k dimensional vectors. We further assume that either Θ contains a finite number of points, or Θ contains a non-countably infinite number of points and is a convex set. Let $\underline{\theta}_o$ denote the true element of Θ and let $\underline{t}$, $\underline{t}^*$, be two alternative vector estimators of $\underline{\theta}_o$. We can now state some alternative definitions of "smaller" mean squared error.

(A) $\underline{t}^*$ has smaller mean squared error that $\underline{t}$ iff

$$E\{(t_i^* - \theta_{oi})^2\} \leq E\{(t_i - \theta_{oi})^2\}, \quad i = 1, 2, \ldots k.$$

(B) $\underline{t}^*$ has smaller mean squared error than $\underline{t}$ iff

$$E\{(\underline{t}^* - \underline{\theta}_o)'(\underline{t}^* - \underline{\theta}_o)\} \leq E\{(\underline{t} - \underline{\theta}_o)'(\underline{t} - \underline{\theta}_o)\}$$

(C) $\underline{t}^*$ has smaller mean squared error than $\underline{t}$ iff

$$E\{(\underline{t}^* - \underline{\theta}_o)'A(\underline{t}^* - \underline{\theta}_o)\} \leq E\{(\underline{t} - \underline{\theta}_o)'A(\underline{t} - \underline{\theta}_o)\},$$

for all positive semi-definite matrices A.

(D) $\underline{t}^*$ has "smaller" second moment matrix that that of $\underline{t}$ iff the matrix $N(\underline{t}^*, t) = M(\underline{t}) - M(\underline{t}^*)$ is positive semi-definite, where

$$M(\underline{t}^*) = E\{(\underline{t}^* - \underline{\theta}_o)(\underline{t}^* - \underline{\theta}_o)'\}, \quad M(\underline{t}) = E\{(\underline{t} - \underline{\theta}_o)(\underline{t} - \underline{\theta}_o)'\}.$$

(E) $\underline{t}^*$ has smaller mean squared error than $\underline{t}$ iff

$|M(\underline{t}^*)| \leq |M(\underline{t})|$, where $|A|$ represents the determinant for any square matrix A.

(F) $t^* = \underline{a}'\underline{t}^*$ has smaller mean squared error than $t = \underline{a}'\underline{t}$, iff $E\{(t^* - \phi_o)^2\} \leq E\{(t - \phi_o)^2\}$, where $\underline{a}'$ is any k dimensional vector and $\phi_o = \underline{a}'\underline{\theta}_o$.

A last alternative which is useful to consider for mean squared error for linear combinations of coefficients is to define the space $\Phi \subset R$, which is a mapping from Θ, and is defined by $\Phi = \{\phi | \phi = \underline{a}'\underline{\theta}, \underline{\theta} \in \Theta\}$. Define the unconstrained estimator t by $t = \underline{a}'\underline{t}$, where $\underline{t}$ is the unconstrained estimator for $\underline{\theta} \in \Theta$ and the constrained estimator t^* by

$$t^* = \begin{cases} t, & \text{if } t \in \Phi \\ a, & \text{if } t \leq a \\ b, & \text{if } t \geq b \end{cases}, \qquad (21)$$

where $a = \inf\{\phi \in \Phi\}$ and $b = \sup\{\phi \in \Phi\}$. In short, the constrained estimator is defined in terms of the constraints <u>after</u> transformation from Θ space to Φ space.

The various definitions for smaller mean squared error of an estimator are not as disparate as would at first sight appear. Definitions (C) and (D) are equivalent in that the inequality in (C) is a necessary and sufficient condition for $N(\underline{t}^*, \underline{t})$ to be positive semi-definite, for proof see Theobald (1974). Definition (D) (and hence (C) as well) implies (E), see Rao (1965, p. 267). Definition (E) implies (B) since letting the matrix A in (C) be the identity matrix yields the condition in (B). Similarly; (C) implies the condition in (A) (by suitable choice of the matrix A) and the condition in (F), by rewriting each term in the inequality as a quadratic

form in the matrix $A = \underline{a}\,\underline{a}'$. Finally, condition (A) implies (B). Consequently, the condition in (B) is necessary, but not sufficient, for the conditions in (A), (C), (D), and (F).

With these preliminaries completed some mean squared error results can be obtained in the multiparameter situation. Consider first the situation in which $\Theta \subset R^k$ is a convex set. Let $\underline{t}$ be any unconstrained vector estimator of $\underline{\theta}_o \in \Theta$. Let $\tilde{\underline{t}}$ be the corresponding minimum distance estimator defined by:

$$\tilde{\underline{t}} = \begin{cases} \underline{t}, \text{ if } \underline{t} \in \Theta, \\ \underline{t}^* \ni ||\underline{t} - \underline{t}^*|| = \min_{\theta \in \Theta} ||\underline{t} - \underline{\theta}||, \ \underline{t} \notin \Theta, \ \underline{t}^* \in \Theta. \end{cases} \tag{22}$$

Theorem 3.--Under definition (B) above $\tilde{\underline{t}}$ has smaller mean squared error than $\underline{t}$.

Proof.--If $\underline{t} \in \Theta$, $\tilde{\underline{t}} = \underline{t}$, so need only consider $\underline{t} \notin \Theta$. If $\underline{t} \notin \theta$, then through the point $\tilde{\underline{t}}$ lies a separating hyperplane between $\underline{t}$ and Θ; this follows from the convexity of Θ. Consider the line defined by the two points $\underline{t}$ and $\tilde{\underline{t}}$. Suppose $\underline{\theta}_o$ lies on this line, then clearly $||\underline{t} - \underline{\theta}_o||^2 > ||\tilde{\underline{t}} - \underline{\theta}_o||^2$, since $\tilde{\underline{t}}$ minimizes the Euclidean distance from $\underline{t}$ to Θ. Suppose $\underline{\theta}_o$ does not lie on the line, so that there exists on that line a point $\underline{c}$ which minimizes the distance from $\underline{\theta}_o$ to the line defined by $\underline{t}$ and $\tilde{\underline{t}}$. Clearly $\underline{c}$ lies on the same side of the separating hyperplane as $\underline{\theta}_o$. Since the squared Euclidean distance from any point $\underline{x}$ on the line defined by $\underline{t}$, $\tilde{\underline{t}}$ to the point $\underline{\theta}_o$ is given by $||\underline{\theta}_o - \underline{x}||^2$ which is equal to $||\underline{\theta}_o - \underline{c}||^2 + ||\underline{c} - \underline{x}||^2$, it follows immediately that $||\underline{\theta}_o - \tilde{\underline{t}}||^2 < ||\underline{\theta}_o - \underline{t}||^2$ for all $t \notin \theta$. This geometric result is sufficient for showing that:

$$E\{(\tilde{\underline{t}} - \underline{\theta}_o)'(\tilde{\underline{t}} - \underline{\theta}_o)\} \leq E\{(\underline{t} - \underline{\theta}_o)'(\underline{t} - \underline{\theta}_o)\}.$$

Unfortunately, one cannot prove in general for any estimators $\underline{t}$ and $\tilde{\underline{t}}$ as defined above that $\tilde{\underline{t}}$ has smaller mean squared than $\underline{t}$ in the sense of definition (C). The problem is best illustrated in terms of definition (F) which is a special case of (C); see Figure 4. ϕ_o is the infimum and ϕ_1 is the supremum of the projections of vectors $\underline{\theta} \in \Theta$. Over the domains A and C shown in Figure 3, $\tilde{\underline{t}}$ clearly contributes less to mean squared error than $\underline{t}$. However, over the domain B $\tilde{\underline{t}}$ obviously does not have smaller mean squared error than $\underline{t}$ for any value of $\underline{t}$, as is illustrated in the figure, so that without further information about the distribution of $\underline{t}$, one cannot conclude that $\tilde{\underline{t}}$ has smaller mean squared error than $\underline{t}$ in the sense of (F) or (C). The problem is that $\tilde{\underline{t}}$ is defined in terms of the Euclidean distance from $\underline{t}$.

Further insight into this problem can be gained by considering a generalization of the minimum distance estimator. Thus, let us define $\tilde{\underline{t}}_a$ by:

$$\tilde{\underline{t}}_a = \begin{cases} \underline{t}, \text{ if } \underline{t} \in \Theta, \\ \tilde{\underline{t}}, \text{ if } \underline{t} \in A, C \\ \underline{t}^* \ni ||\underline{t} - \underline{t}^*|| = \min||\underline{t} - \underline{\theta}|| \ \forall \ \underline{\theta} \\ \ni \underline{\theta} \in \Theta \text{ and } \underline{a}'(\underline{t} - \underline{\theta}) = 0, \text{ if } \underline{t} \in B \end{cases} \tag{23}$$

Essentially, $\tilde{\underline{t}}_a$ for $\underline{t} \notin \Theta$ is that point on the boundary of Θ which is "nearest" to $\underline{t}$ among the set of vectors with the same projection as $\underline{t}$ onto the line "a," when $\underline{t}$ is contained in the domain B; otherwise $\tilde{\underline{t}}_a$ is equal to $\tilde{\underline{t}}$.

If $\underline{t} \in$ domain A or C, then

$$\underline{a}'(\underline{t} - \underline{\theta}_o) \geq \underline{a}'(\tilde{\underline{t}}_a - \underline{\theta}_o);$$

and if $t \in$ domain B,

$$\underline{a}'(t - \underline{\theta}_o) = \underline{a}'(\tilde{\underline{t}}_a - \underline{\theta}_o)$$

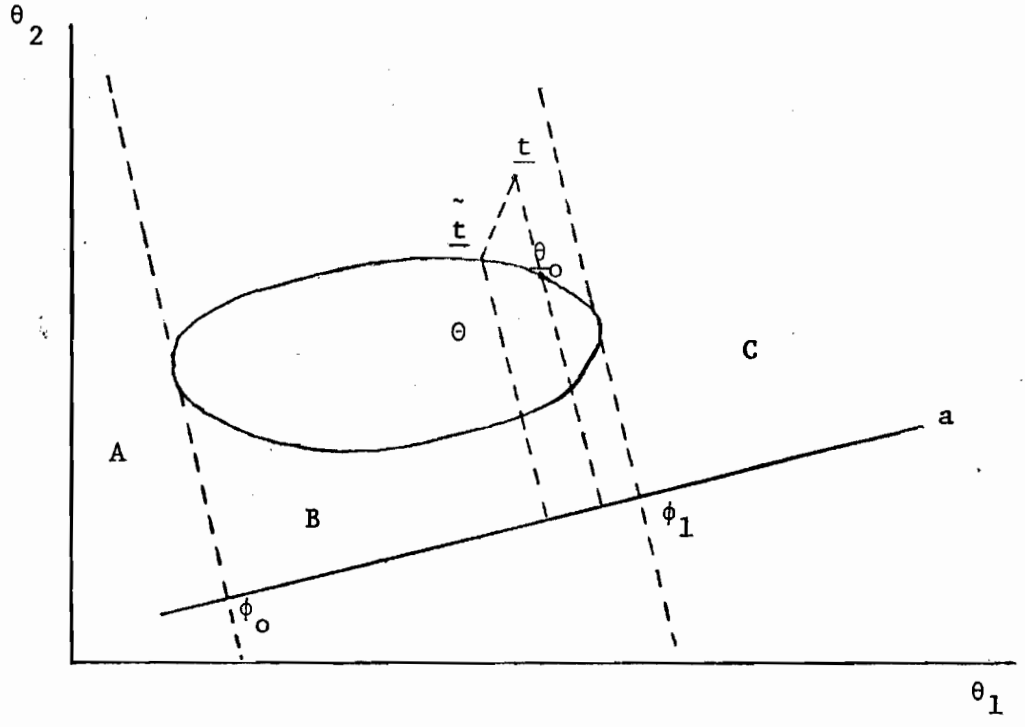

Figure 3.--Two Dimensional Parameter Space and Projections Onto an Arbitrary Line "a."

by construction. Consequently, for a given vector "a" in definition (F), $\tilde{\underline{t}}_a$ has smaller mean squared error than that of $\underline{t}$.

Alternatively, a mean squared error gain for a given vector $\underline{a}$ can be obtained in the following way. Let $\phi_o = \underline{a}'\underline{\theta}_o$, $\underline{\theta} \in \Theta$ and define the convex set Φ by:

$$\Phi = \{\phi | \phi = \underline{a}'\underline{\theta},\ \underline{\theta} \in \Theta\}.$$

The convexity of Φ follows from that of Θ. The unconstrained estimator of ϕ is $t = \underline{a}'\underline{t}$, where $\underline{t}$ is any unconstrained estimator of $\underline{\theta}_o$. Let us now define the constrained (minimum distance) estimator in Φ space, instead of in Θ space, in short:

$$\tilde{t} = \begin{cases} t, & \text{if } t \in \Phi, \\ \phi_o, & \text{if } t \leq \phi_o, \\ \phi_1, & \text{if } t \geq \phi_1, \end{cases} \qquad (24)$$

where ϕ_o and ϕ_1 are as defined above.

Since the multiparameter problem has been transformed into a uniparameter problem, all of the results derived in the first part of this section for the minimum distance estimator now apply, so that $\tilde{t}$ has smaller mean squared error than t.

No useful general results have yet been directly obtained for the constrained maximum likelihood estimator in any of the above situations even though the required moments exist under the assumptions of the problem.

With respect to the situation in which Θ contains only a finite number of points no general finite sample size results have yet been obtained on mean squared error properties. However, under suitable regularity conditions mean squared error gains can be shown to hold asymptotically by a minor extension of Theorem 2' to the multivariate case and in various specific cases, see for example Hammersly (1950).

Finite and Asymptotic Distributions of $\hat{\underline{\theta}}_c$ and $\hat{\underline{\theta}}_m$

Finite Sample Results

In the uniparameter situation there are two subcases; those where Θ is composed of a finite set of points and those where it is not. In the former case, let $\Theta = \{\lambda_1, \lambda_2, \ldots \lambda_q\}$, $\lambda_i \in R^1$, $\lambda_i < \lambda_{i+1}$, and we define R_i, $i = 1, 2, \ldots q$, by:

$$\begin{aligned} R_1 &= (-\infty, 1/2(\lambda_1 + \lambda_2)), \\ R_2 &= [1/2(\lambda_1 + \lambda_2), 1/2(\lambda_2 + \lambda_3)), \\ &\cdots \\ R_q &= [1/2(\lambda_{q-1} + \lambda_q), \infty). \end{aligned} \tag{25}$$

If $f(\hat{\theta}_u)\, d\hat{\theta}_u$ denotes the p.d.f. of the random variable $\hat{\theta}_u$, the discrete distribution of $\hat{\theta}_m$ is given by:

$$\hat{\theta}_m = \begin{cases} \lambda_1, \text{ with probability } \pi_1 \\ \lambda_2, \text{ with probability } \pi_2 \\ \cdots \\ \lambda_q, \text{ with probability } \pi_q, \end{cases} \tag{26}$$

where:

$$\pi_i = \int_{R_i} f(\hat{\theta}_u) d\hat{\theta}_u, \quad i = 1, 2, \ldots q. \tag{27}$$

In the latter case, where Θ is convex, θ can be represented by a closed interval $[\lambda_1, \lambda_2]$, $\lambda_1, \lambda_2 \in R^1$, $\hat{\theta}_m$ is given by:

$$\hat{\theta}_m = \begin{cases} \lambda_1, \text{ if } \hat{\theta}_u \leq \lambda_1 \\ \lambda_2, \text{ if } \hat{\theta}_u \geq \lambda_2 \\ \hat{\theta}_u, \text{ otherwise,} \end{cases} \tag{28}$$

Setting:

$$\pi_1 = \int_{\hat{\theta}_u \leq \lambda_1} f(\hat{\theta}_u)d\hat{\theta}_u, \quad \pi_2 = \int_{\hat{\theta}_u \geq \lambda_2} f(\hat{\theta}_u)d\hat{\theta}_u. \tag{29}$$

The distribution of $\hat{\theta}_m$ is clearly of the mixed discrete/continuous type whose specific form is:

$$g(\hat{\theta}_m) = \begin{cases} \pi_1, \ \hat{\theta}_m = \lambda_1, \\ \pi_2, \ \hat{\theta}_m = \lambda_2, \\ f(\hat{\theta}_u), \ \lambda_1 < \hat{\theta}_m < \lambda_2. \end{cases} \tag{30}$$

The multi-variate extension of the finite number of points is to consider that $\Theta = \{\underline{P}_1, \underline{P}_2, \ldots \underline{P}_q\}$, where $\underline{P}_i$ is a point in s-dimensional Euclidean space. The definition of the minimum distance estimator leads directly to the definition of regions in the whole of R^s such that if R_i is the appropriately defined subset of R^s, then:

$$\pi_i = Pr\{\hat{\underline{\theta}}_m = \underline{P}_i\} = \int_{R_i} \phi_{\underline{\theta}_o} d\underline{u}, \tag{31}$$

where $\phi_{\underline{\theta}_o} d\underline{u}$ denotes an s-variate p.d.f. of the vector estimator $\hat{\underline{\theta}}_u$ with mean vector $\underline{\theta}_o$. The discrete distribution is given by:

$$\hat{\underline{\theta}}_m = \begin{cases} \underline{P}_1, \text{ with probability } \pi_1 \\ \underline{P}_2, \text{ with probability } \pi_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \underline{P}_q, \text{ with probability } \pi_q \end{cases} \tag{32}$$

To illustrate, consider the situation in which $q = s = 2$ and $\underline{\theta}_o = \emptyset$ as shown in Figure 4. Under these conditions $E(\hat{\underline{\theta}}_m) = \pi_2\underline{P}$, so that the M.D. estimator is biased, though consistent since $\pi_2 \to 0$ as $n \to \infty$ for any $\underline{P} \neq \emptyset$.

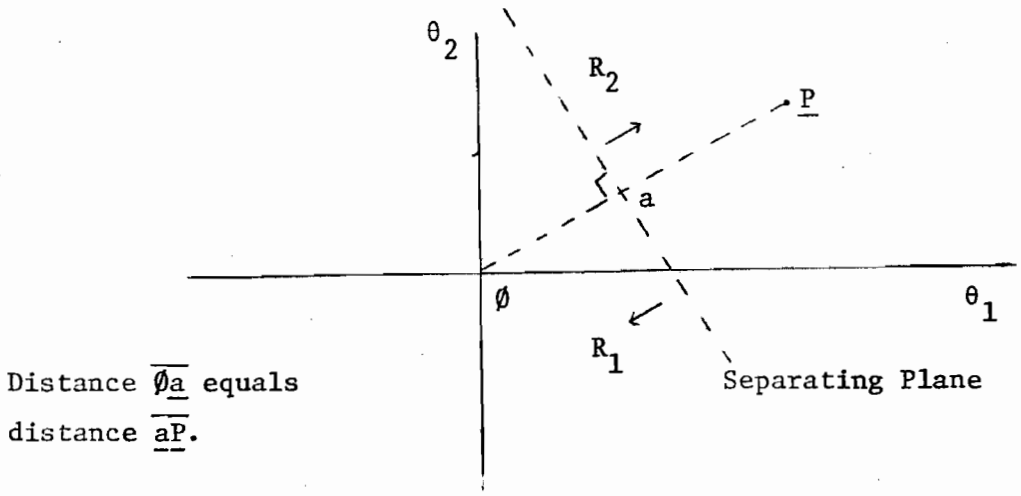

Figure 4.--M.D. Estimator in the Two Dimensional Case Illustrating the Integrating Regions.

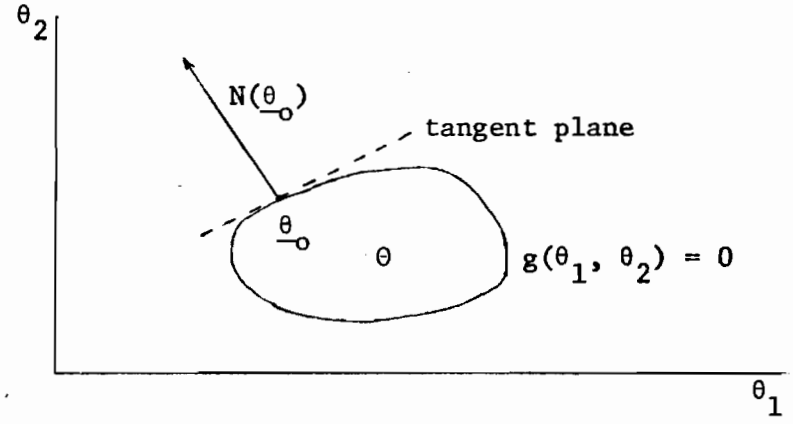

Figure 5.--Illustration of a Compact Two Dimensional Parameter Space with Continuous Differentiable Boundary.

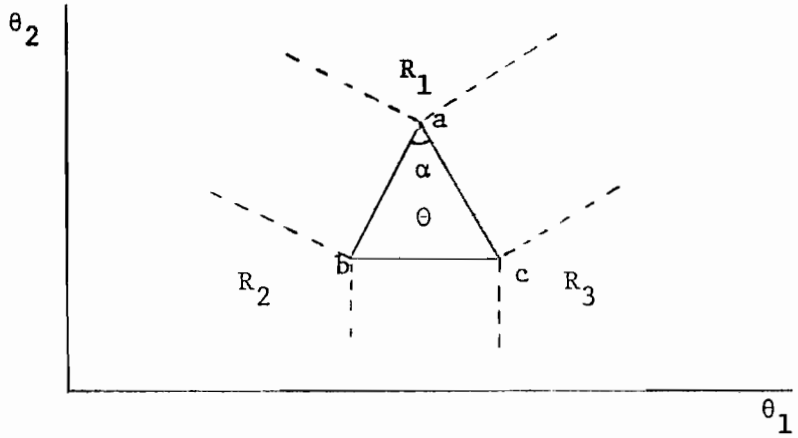

Figure 6.--Illustration of a Compact Two Dimensional Parameter Space with Only Piece-wise Continuous Boundary.

The multi-variate analogue to the closed interval is a more complicated case to analyze. The basic situation is most easily discussed in terms of two dimensional space and two important sub-cases are illustrated in Figures 5 and 6. Let us consider first the case in which the boundary of the parameter space is defined by a continuous differentiable function $g(\theta_1, \theta_2) = 0$ which defines a closed curve in R^2, see Figure 5.

The main problem is to define the p.d.f. of $\hat{\underline{\theta}}_m$ along the closed curve $g(\theta_1, \theta_2) = 0$. Let the point p.d.f. of $\hat{\underline{\theta}}_u$ be denoted by $f_{\theta_o}(\hat{\theta}_1, \hat{\theta}_2)d\hat{\theta}_1 d\hat{\theta}_2$ defined over all of R^2.

The density of $\hat{\underline{\theta}}_m$ satisfying $g(\hat{\underline{\theta}}_m) = 0$ at (θ_1^o, θ_2^o) is denoted by $P(\underline{\theta}^o)$ and is defined by:

$$P(\underline{\theta}^o) = \int_o^\infty f_{\theta_o}(\hat{\theta}_1, \hat{\theta}_2)ds, \tag{33}$$

where

$$ds = (g_1^2 + g_2^2)^{1/2}dt, \tag{34}$$

is the rate of change (with respect to a parameter t) in arc length of the normal to the tangent plane to the curve $g(\underline{\theta}) = 0$. The partial derivatives g_1, g_2 evaluated at $\underline{\theta}^o$ give the direction numbers of the normal ($N(\underline{\theta}^o)$ in Figure 5) to the tangent plane and the parametric representation of the normal line is:

$$\begin{aligned} \hat{\theta}_1 &= \theta_1^o + g_1 t \\ \hat{\theta}_2 &= \theta_2^o + g_2 t. \end{aligned} \tag{35}$$

The term $P(\underline{\theta})$ can be regarded as defining a "marginal" p.d.f. which has been rotated out of parallel with the axes and is truncated at the point $\underline{\theta}_o$, for

example if Θ in Figure 6 were to be reduced to a straight line parallel to the θ_1 axis (i.e. $\theta_2 = \theta_2^o$ is known), $P(\underline{\theta}^o)$ produces (by integrating in both directions from $\underline{\theta}^o$) the marginal distribution of $\hat{\theta}_{u1}$ by integrating out $\hat{\theta}_{u2}$.

The p.d.f. of $\hat{\underline{\theta}}_m$ can now be written down:

$$h(\hat{\underline{\theta}}_m) = \begin{cases} P(\hat{\underline{\theta}}_m), & g(\hat{\underline{\theta}}_m) = 0 \\ f(\hat{\underline{\theta}}_m), & g(\hat{\underline{\theta}}_m) < 0 \end{cases} \tag{36}$$

The mean vector of $\hat{\underline{\theta}}_m$ is obtained by:

$$\underline{\eta} = \int_c (\hat{\underline{\theta}}_m) P(\hat{\underline{\theta}}_m) ds + \iint_\Theta (\hat{\underline{\theta}}_m) f(\hat{\underline{\theta}}_m) d\hat{\underline{\theta}}_m, \tag{37}$$

where the first integral is evaluated along the length of the closed curve $g(\hat{\underline{\theta}}) = 0$ and the second integral is an ordinary multiple integral over the domain Θ.

Figure 6 illustrates the problem of point discontinuities in the first differential of a piecewise continuous boundary function. The regions R_1, R_2, R_3, are defined by the cones with apex at a, b, and c respectively and sides determined by the normals to the intersecting tangent planes at the point of intersection. For example, if at the point "a" in Figure 6 the angle subtended by the boundary of θ is α radians, the cone defining R_1 has angle $(\pi - \alpha)$ radians between its boundaries. In this situation, the distribution function of $\hat{\underline{\theta}}_m$ is defined by probabilities at certain points, line integrals, and ordinary multiple integrals. Thus, as in this example, the distribution of $\hat{\underline{\theta}}_m$ is:

$$h(\hat{\underline{\theta}}_m) = \begin{cases} \pi_1, \text{ for } \hat{\underline{\theta}}_m = a \\ \pi_2, \text{ for } \hat{\underline{\theta}}_m = b \\ \pi_3, \text{ for } \hat{\underline{\theta}}_m = c \\ P(\hat{\underline{\theta}}_m), \text{ for } g(\hat{\underline{\theta}}_m) = 0 \\ f(\hat{\underline{\theta}}_m), \text{ for } g(\hat{\underline{\theta}}_m) < 0, \end{cases} \tag{38}$$

where:

$$\pi_i = \int_{R_i} f(\hat{\underline{\theta}}_u) d\hat{\underline{\theta}}_u, \tag{39}$$

and $P(\hat{\underline{\theta}}_m)$ is defined as in the previous situation.

With respect to the distribution of $\hat{\underline{\theta}}_c$, it was shown in the previous section on mean squared error properties that under the conditions imposed by Lemma 1 for the uniparameter case where Θ is a closed interval that the distribution of $\hat{\theta}_c$ was the same as that for $\hat{\theta}_m$. Unfortunately, so far this is the only general finite sample size result for the distribution of $\hat{\underline{\theta}}_c$ when $\hat{\underline{\theta}}_c$ is not equivalent to $\hat{\underline{\theta}}_m$.

Asymptotic Results

By the definition of the parameter space Θ and the definitions of $\hat{\underline{\theta}}_c$ and $\hat{\underline{\theta}}_m$, it is easy to show that both estimators are convergent to $\underline{\theta}_o$ almost surely; i.e.:

$$\Pr(\lim_{n\to\infty} \hat{\underline{\theta}}_c = \underline{\theta}_o) = 1 \quad \text{and} \quad \Pr(\lim_{n\to\infty} \hat{\underline{\theta}}_m = \underline{\theta}_o) = 1.$$

By the definitions of Θ and of $\hat{\underline{\theta}}_c$, $\hat{\underline{\theta}}_m$ respectively, there exist vectors $\underline{a}$, $\underline{b}$ such that (in the vector sense) $\underline{a} \leq \hat{\underline{\theta}}_c \leq \underline{b}$ and similarly for $\hat{\underline{\theta}}_m$, consequently the variances of the components of $\hat{\underline{\theta}}_c$ and of $\hat{\underline{\theta}}_m$ are uniformly bounded, since the bounding vectors $\underline{a}$, $\underline{b}$ depend only upon Θ. Thus, by a corollary to the Kolmogorov Theorem on the strong law of large numbers $\hat{\underline{\theta}}_c$ and $\hat{\underline{\theta}}_m$ are convergent

almost surely to $\underset{\sim}{\theta}_o$, see, for example Fisz (1963, pp. 220-224). More importantly, $\hat{\underset{\sim}{\theta}}_c$ and $\hat{\underset{\sim}{\theta}}_m$ are almost surely convergent to $\underset{\sim}{\theta}_o$ even when $\hat{\underset{\sim}{\theta}}_u$ is not itself almost surely convergent, see for example the exchange between Wald (1949) and Wolfowitz (1949). In such cases, that is, cases where $\hat{\underset{\sim}{\theta}}_u$ although convergent in probability is not almost surely convergent, it will follow that the series

$$\sum_{n=1}^{\infty} \Pr(\hat{\underset{\sim}{\theta}}_c \neq \hat{\underset{\sim}{\theta}}_u) \quad \text{and} \quad \sum_{n=1}^{\infty} \Pr(\hat{\underset{\sim}{\theta}}_m \neq \hat{\underset{\sim}{\theta}}_u)$$

are divergent, even though by consistancy $\lim_{n\to\infty} \Pr\{|\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_u| < \varepsilon\} = \lim_{n\to\infty} \Pr\{|\hat{\underset{\sim}{\theta}}_m - \hat{\underset{\sim}{\theta}}_u| < \varepsilon\} = 1$; see Fisz (1963, p. 226).

One other important asymptotic result can be easily demonstrated and is posed as Lemma 3.

Lemma 3.--The sequence $\underset{\sim}{y}_n = (\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_m)$ is convergent in probability to $\emptyset$ so that $\hat{\underset{\sim}{\theta}}_c$ is convergent in probability to $\hat{\underset{\sim}{\theta}}_m$.

Proof.--Since the analysis is being conducted in terms of a metric space the convergence in probability of $\underset{\sim}{y}_n$ can be characterized by:

$$\lim_{n\to\infty} \Pr\{||\underset{\sim}{y}_n|| < \varepsilon\} = 1.$$

Since $||\underset{\sim}{y}_n||^2 \leq ||\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_u||^2 + ||\hat{\underset{\sim}{\theta}}_m - \hat{\underset{\sim}{\theta}}_u||^2$, then

$$\Pr\{||\underset{\sim}{y}_n|| < \varepsilon\} \geq \Pr\{||\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_u|| + ||\hat{\underset{\sim}{\theta}}_m - \hat{\underset{\sim}{\theta}}_u|| < \varepsilon\} \qquad (40)$$

The right hand side of inequality (40) is equal to:

$$\Pr\{||\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_u|| < \varepsilon\} + \Pr\{||\hat{\underset{\sim}{\theta}}_m - \hat{\underset{\sim}{\theta}}_u|| < \varepsilon\} - \Pr\{||\hat{\underset{\sim}{\theta}}_c - \hat{\underset{\sim}{\theta}}_u|| < \varepsilon, \ ||\hat{\underset{\sim}{\theta}}_m - \hat{\underset{\sim}{\theta}}_u|| < \varepsilon\} \qquad (41)$$

but

$$\Pr\{||\hat{\underline{\theta}}_c - \hat{\underline{\theta}}_u|| < \varepsilon,\ ||\hat{\underline{\theta}}_m - \hat{\underline{\theta}}_u|| < \varepsilon\} \leq \Pr\{||\hat{\underline{\theta}}_c - \hat{\underline{\theta}}_u|| < \varepsilon\},$$

therefore expression (41) is greater than or equal to $\Pr\{||\hat{\underline{\theta}}_m - \hat{\underline{\theta}}_u|| < \varepsilon\}$. But, have already shown that:

$$\lim_{n\to\infty} \Pr\{||\hat{\underline{\theta}}_m - \hat{\underline{\theta}}_u|| < \varepsilon\} = 1,$$

so that $\lim_{n\to\infty} \Pr\{||y_n|| < \varepsilon\} = 1$ and the lemma has been proved.

Corollary.--Since by Lemma 3 $\hat{\underline{\theta}}_c$ converges in probability to $\hat{\underline{\theta}}_m$, and if $g(\hat{\underline{\theta}}_m)$ represents the asymptotic limiting cumulative distribution of $\hat{\underline{\theta}}_m$, and $Hn(\hat{\underline{\theta}}_c)$ the cumulative distribution corresponding to $\hat{\underline{\theta}}_c(n)$ in the sequence of random variables $\hat{\underline{\theta}}_c(n)$, n = 1, 2, ..., then:

$$\lim_{n\to\infty} Hn(\hat{\underline{\theta}}_c) = g(\hat{\underline{\theta}}_m)$$

at every continuity point of $g(\hat{\underline{\theta}}_m)$. The proof of this corollary is standard, see, for example, Fisz (1963, pp. 236-239).

The main result of the previous lemma and its corollary is that even though neither the finite sample nor asymptotic distribution of $\hat{\underline{\theta}}_c$ in general has been derived directly, we can say that the asymptotic distribution of $\hat{\underline{\theta}}_c$ is the same as that of $\hat{\underline{\theta}}_m$, the finite sample size distribution of which has been given above. The distribution of $\hat{\theta}_c$ in the uniparameter situation is simply a special case. The only finite sample results for the distribution of $\hat{\underline{\theta}}_c$ are in the uniparameter situation where Θ is an interval and the restrictive assumptions of Lemma 1 on the parent p.d.f. hold, or the situation examined by Hammersly (1950).

III. Illustration of the Three Estimators by Sampling Experiments

A number of simple models were run using random variables generated by computer routines in order to check the theory developed so far and to explore

the probable consequences of further analysis. In the single parameter situation $N(\theta, 1)$ variables were generated in sample sizes 30, 50, and 100. For various definitions of θ, UML, CML, and MD estimators were calculated for the mean; UML in this case is simply the sample mean. In the two parameter situation, two bivariate normal distributions were generated with mean vector $\underline{\theta}$; one with a unit scalar covariance matrix, and the other with a covariance matrix defined by $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 1/2$. In each of the two situations all three estimators were used to estimate the mean vector under a variety of constraints on θ. The experiments were each replicated 500 times and the Monte Carlo sample means and variances were calculated for each estimator. In addition Monte Carlo estimates of the standardized measures of skewness and kurtosis, γ_1, γ_2, were calculated where γ_1, γ_2 are defined by:

$$\gamma_1 = \mu_3/\mu_2^{3/2} \qquad \gamma_2 = \mu_4/\mu_2^2 - 3.0 \tag{42}$$

where μ_i denotes the ith moment about the mean. For the normal distribution (hence for the distribution of the UML estimator) $\gamma_1 = \gamma_2 = 0$.

The alternative specifications of the parameter space θ are:

Uniparameter (mean of a normal distribution)

I: $\Theta = \{4, 5\}$, $\theta_o = 4$

II: $\Theta = [4, 5]$, $\theta_o = 4$

III: $\Theta = [4, 5]$, $\theta_o = 4.333$

Biparameter (mean vector of a bivariate normal distribution)

I: $\Theta = \{(4, 2), (5, 3)\}$, $\underline{\theta}_o' = (4, 2)$

II: $\Theta = \{(4, 2), (5, 3), (5, 2)\}$, $\underline{\theta}_o' = (4, 2)$

III: Θ is defined by $\theta_1 + \theta_2 = 1$, $0 \leq \theta_i$, $i = 1, 2$; $\underline{\theta}_o' = (1, 0)$

IV: Same Θ as in III, $\underline{\theta}_o' = (0.5, 0.5)$

V: Θ is defined by $\theta_1^2 + \theta_2^2 = 6$, $\underline{\theta}_o' = (2.4495, 0.0)$

VI: Same Θ as in V, $\underline{\theta}_o' = (1.0, 1.0)$

VII: Θ is defined as a unit square with lower left co-ordinate of (4, 2), $\underline{\theta}_o' = (4, 2)$

VIII: Same Θ as in VII, $\underline{\theta}_o' = (4, 2.33)$

IX: Same Θ as in VII, $\underline{\theta}_o' = (4.318, 2.368)$.

The first question at issue is the difference in distribution between the CML and MD estimators. With two exceptions to be discussed below, the sample distributions of the two estimators appear to be almost identical in the models examined; this is so in terms of the means and variances, sampled values of γ_1 and γ_2, as well as in terms of the observed histograms.

The gain in mean squared error in using constrained over the unconstrained estimator can be considerable. Of course, if the true value of the parameter is strictly within the bounds, then for sufficiently large sample sizes all three estimators are asymptotically equivalent. In the single parameter case, model II, the M.S.E.$(\hat{\theta}_m)$ was less than half the variance of $\hat{\theta}_u$. In the biparameter models (III, V, VII, and VIII), the decrease in M.S.E. in using $\hat{\theta}_m$ instead of $\hat{\theta}_u$ was substantial; the ratio M.S.E.$(\hat{\theta}_m)$/Var$(\hat{\theta}_u)$ was typically about 0.5 and in one set of models about 0.13.

There are two exceptions to these results, the results for models V and VIII using the non-scalar covariance matrix. In these cases, but for only one of the coefficients, the ratio M.S.E.$(\hat{\theta}_m)$/Var$(\hat{\theta}_u)$ is only marginally under 1.0. Further the M.S.E. of $\underline{\hat{\theta}}_c$ seems to be considerably larger. These results are in broad agreement with the theoretical analysis in the previous sections.

The models involving the choice of isolated points show that the constrained estimators converge very strongly to the true parameter value. The only experiment for which the constrained estimator did not pick the true point every time in 500 trials was that for the univariate case, sample size 30, in which the true value was chosen 497 times.

Conclusion

In conclusion, we may say that the analysis in this paper indicates that the incorporation of inequality constraints in the estimation process through the use of the minimum distance estimator is advantageous. The estimator is relatively easy to calculate, leads to a considerable reduction in mean squared error, and has the same asymptotic distribution as the constrained maximum likelihood estimator. In addition, although the finite sample size distribution (given knowledge of the distribution of $\hat{\underline{\theta}}_u$) of $\hat{\underline{\theta}}_m$ is not always easily evaluated in actual situations, its analytical form is at least known.

The extremely rapid convergence of $\hat{\underline{\theta}}_m$ when Θ space is composed of at most a denumerably infinite number of points is well worth noting.

Further, the immediate extensions of the above analysis to forecasts of dependent variables leads to the broad conclusion that considerable gains in lowering the mean squared error of forecasts by using point or inequality constrained estimators can be achieved. Thus, the range of situations usefully analyzed by regression techniques has been extended by this analysis.

Finally, the work contained in this paper enables one to begin to evaluate the inferential gains from incorporating the results of previous research into subsequent analysis.

FOOTNOTES

[1]The additional and somewhat unnecessary phrase "finitely bounded" is added because of the common practice of extending compact spaces by "compactification," see Kingman and Taylor (1966).

[2]In addition to some obvious regularity conditions on the cumulative distribution function, the two most important assumptions from the viewpoint of this paper are:

(i) The parameter space is a metric space and every closed bounded subset is compact,

(ii) Expected values of the logarithms of the p.d.f.'s exist and are finite.

[3]Wolfowitz (1949), pointed out that only the weak law was needed to prove consistency.

[4]The required moments exist given the assumptions made for the previous theorems.

[5]If the right hand point were the true parameter point, the analysis would be essentially the same, except (17) would hold for all $b < 0$.

REFERENCES

Aitchison, and Silvey. "Maximum Likelihood Estimation of Parameters Subject to Restraints." Ann. Math. Statist., Vol. 29 (1958):813-828.

Brunk, H. D. "On the Estimation of Parameters Restricted by Inequalities." Annals of Math. Statist., Vol. 29 (1958):437-454.

Daniels, H. E. "The Asymptotic Efficiency of a Maximum Likelihood Estimator." Proc. of the Fourth Berkeley Symp. on Math. Statistics and Prob. Berkeley: University of California, 1960, pp. 151-163.

Fisz, M. Probability Theory and Mathematical Statistics. New York: Wiley, 1963.

Hammersly, T. M. "On Estimating Restricted Parameters." JRSS, Series B, Vol. XII (1950):192-229.

Hanson, M. A. "Inequality Constrained Maximum Likelihood Estimation." Annals of the Institute of Statistical Mathematics (Tokyo), Vol. 17 (1965):311-321.

Hocking, R. R. "The Distribution of a Projected Least Squares Estimator." Annals of the Institute of Statistical Mathematics (Tokyo), Vol. 17 (1965):357-362.

Huber, P. J. "The Behaviour of Maximum Likelihood Estimates Under Non-Standard Conditions." Proc. of the Fifth Berkeley Symp. on Math. and Prob. Berkeley: University of California, 1967, Vol. 1, pp. 221-233.

Judge, G. C., and Takayama, T. "Inequality Restrictions in Regression Analysis." JASA, Vol. 61 (1966):166-181.

Kingman, J. F. C., and Taylor, S. J. Introduction to Measure and Probability. Cambridge: Cambridge University, 1966.

LeCam, L. "On the Asymptotic Theory of Estimation and Testing Hypotheses." Proc. Third Berkeley Symp. on Math. Stat. and Prob. Berkeley: University of California, 1953, Vol. 1, pp. 129-156.

Moran, P. A. P. "Maximum-likelihood Estimation in Non-standard Conditions." Proc. Comb. Phil. Soc., Vol. 70 (1971):441-450.

Norden, R. H. "A Survey of Maximum Likelihood Estimation." Int. Stat. Rev., Vol. 40, Part 1, No. 3 (1972):324-354; Vol. 41, Part 2, No. 1 (1972): 39-58.

Rao, C. R. Linear Statistical Inference and Its Applications. New York: Wiley, 1965.

Ruben, H. "Probability Content of Regions Under Spherical Normal Distributions." Annals of Math. Statist., Vol. 31 (1960):598-618.

Theil, H., and Goldberger, A. S. "On Pure and Mixed Estimation in Economics." Int. Econ. Rev., Vol. 2 (1961):65-78.

Theil, H. Principles of Econometrics. New York: Wiley, 1971.

Wald, A. "Note on the Consistency of Maximum Likelihood Estimate." Ann. Math. Statist., Vol. 20 (1949):595-601.

Weiss, L. "Asymptotic Properties of Maximum Likelihood Estimators in Some Non-Standard Cases." JASA, Vol. 66 (1971):345-350.

Wolfowitz, J. "On Wald's Proof of the Consistency of the Maximum Likelihood Estimate." Ann. Math. Statist., Vol. 20 (1949):601-602.

Wolfowitz, J. "Estimation by the Minimum Distance Method in Non-Parametric Difference Equations." Annals of Math. Statist., Vol. 25 (1954).

Zellner, A. "Linear Regression with Inequalities Constraints on the Coefficients: An Application of Quadratic Programming and Linear Decision Rules." International Center for Management Science, Rotterdam, 1961.