Statistical Science
2000, Vol. 15, No. 4, 313–341

Eliminating Multiple Root Problems in Estimation

Christopher G. Small, Jinfang Wang and Zejiang Yang

Abstract. Estimating functions, such as the score or quasiscore, can have more than one root. In many of these cases, theory tells us that there is a unique consistent root of the estimating function. However, in practice, there may be considerable doubt as to which root is appropriate as a parameter estimate. The problem is of practical importance to data analysts and theoretically challenging as well. In this paper, we review the literature on this problem. A variety of examples are provided to illustrate the diversity of situations in which multiple roots can arise. Some methods are suggested to investigate the possibility of multiple roots, search for all roots and compute the distributions of the roots. Various approaches are discussed for selecting among the roots. These methods include (1) iterating from consistent estimators, (2) examining the asymptotics when explicit formulas for roots are available, (3) testing the consistency of each root, (4) selecting by bootstrapping and (5) using information-theoretic methods for certain parametric models. As an alternative approach to the problem, we consider how an estimating function can be modified to reduce the number of roots. Finally, we survey some techniques of artificial likelihoods for semiparametric models and discuss their relationship to the multiple root problem.

Key words and phrases: Bootstrapping, consistent root, estimating functions, likelihood, multiple roots, Newton–Raphson iteration, parameter, quasilikelihood.

1. INTRODUCTION

As it is usually defined, a *point estimator* is a function of a random sample which takes values within a parameter space. In practice, however, it is typically only in rather simple models, such as the linear model for regression or the exponential family model for parametric inference, that the best point estimators can be constructed explicitly as a function of the sample. For many other important models, the construction of a point estimator is more computationally intensive and involves an iteration to search for a solution to one or more *estimating equations* of the form $g(\theta) = 0$, where $g(\theta)$ is a function of the data. The construction of maximum likelihood estimators, where

$$g(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)$$

and L is the likelihood function, is a case in point. In most cases, the likelihood equations cannot be solved explicitly, and the investigator must resort to some numerical method to construct a point estimate. After the initial enthusiasm for the method of maximum likelihood proposed by Fisher (1925), questions arose as to the existence and uniqueness of a root of the likelihood equation and whether that root corresponded to a maximum of the likelihood function.

As Huzurbazar (1948) noted, proofs of the consistency and asymptotic efficiency of the maximum

Christopher G. Small is Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (E-mail: cgsmall@uwaterloo.ca). Jinfang Wang is Visiting Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo, and Assistant Professor, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan. Zejiang Yang is a graduate student, Department of Statistics and Actuarial Science, University of Waterloo, and Statistician, Kendle International, 55 Hatchets Hill Road, Old Lyme, Connecticut 06371.

likelihood estimator were more properly proofs of the existence of a consistent and asymptotically efficient root of the likelihood equations. However, such a root could also be a local minimum of the likelihood, unless proven otherwise. It was Huzurbazar (1948) who showed that a consistent root of the likelihood equation is asymptotically unique and corresponds to a local maximum of the likelihood function. More precisely, if $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ are two consistent roots of the likelihood equation for a sample of size n, then $P(\hat{\theta}_{1n} = \hat{\theta}_{2n}) \to 1$ as $n \to \infty$. The standard regularity due to Cramér is assumed. It should be noted that this result of Huzurbazar has no implications for the existence of extraneous inconsistent solutions of the likelihood equation.

Simple as this result of Huzurbazar would appear to be, the statement that consistent roots are asymptotically unique must be treated with care. This is because a root $\hat{\theta}_n$ is not merely a real number, but is most appropriately understood as a function $\hat{\theta}_n = \hat{\theta}_n(y_1, \ldots, y_n)$ with the property that

$$\left(\frac{\partial \log L}{\partial \theta}\right)_{\theta=\hat{\theta}_n} = 0.$$

So if the likelihood equation has multiple roots, then it has infinitely many roots if these are understood to be infinitely many distinct functions. In particular, it will have infinitely many distinct consistent roots, understood in the same sense. Of course, it was Huzurbazar's result that all such distinct consistent roots are asymptotically equivalent. Perlman (1983) provided a more precise formulation of Huzurbazar's result, by showing that for sufficiently small $\delta > 0$, with probability 1 there exists exactly one solution to the likelihood equation in the interval $[\theta_0 - \delta, \theta_0 + \delta]$ for all but finitely many n, where θ_0 is the true value of the parameter.

The numerical problem of finding all the roots of the likelihood equation was considered by Barnett (1966). In view of the advances in computation that have been made since the 1960's, some of the comments in Barnett's paper are now dated. However, many of the basic insights about iterative searches for the roots of estimating equations remain true today. Barnett considered five methods for iterating toward a root of the likelihood equation and performed a simulation study of the likelihood equation for the Cauchy location model, which we will consider in the next section. Barnett (1966) considered *Newton–Raphson iteration*

$$\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - g(\hat{\theta}^{(i)})/\dot{g}(\hat{\theta}^{(i)}), \tag{1}$$

where $\dot{g}$ represents the derivative of g with respect to θ and compared it to Fisher's *scoring of parameters* and the *fixed-derivative Newton method* as well as the *method of false positions* using the Cauchy location model as a test case. The method of false positions is of particular importance for locating a root in a single-parameter model, or in a model with several parameters in which all but one of the parameters can be solved for explicitly. While the procedure can be generalized into higher dimensions, the advantage of using the method of false positions is that for a continuous real-valued estimating function, the procedure is guaranteed to converge to a root—a guarantee provided by the intermediate value theorem.

A survey of the modern theory of efficient likelihood estimation for likelihoods with multiple local maxima can be found in Lehmann (1983, pages 420–427). In particular, if $\hat{\theta}^{(1)}$ is a $\sqrt{n}$-consistent estimator for θ, then the *one-step estimator* $\hat{\theta}^{(2)}$, found by applying Newton–Raphson iteration (1) to $\hat{\theta}^{(1)}$, will be asymptotically efficient. While this seems to provide a satisfactory theory for root approximation and selection, it raises the problem of how to appropriately select a good $\sqrt{n}$-consistent estimator. Asymptotic considerations can only take us so far: there are infinitely many asymptotically efficient estimators which can be constructed even in the most regular of models. How to choose among them remains a problem.

Recently, attention has turned to the problem of multiple roots which arise with more general estimating equations. An example is the *quasiscore*, defined by

$$g(\theta) = \sum_{j=1}^{n} \frac{\dot{\mu}_j(\theta)}{\sigma_j^2(\theta)}[Y_j - \mu_j(\theta)], \tag{2}$$

where $\mu_j(\theta) = E_\theta(Y_j)$, $\sigma_j^2(\theta) = \mathrm{Var}_\theta(Y_j)$, and once again the dot operator denotes differentiation with respect to θ. This, and many other such estimating functions, are *unbiased*, so that

$$E_\theta[g(\theta)] = 0 \tag{3}$$

for all $\theta \in \Theta$, and, additionally, *information unbiased*, so that

$$-E_\theta[\dot{g}(\theta)] = E_\theta[g^t(\theta)\, g(\theta)] \tag{4}$$

for all θ, where g is a $1 \times k$ row vector and $\dot{g}(\theta)$ is the $k \times k$ matrix of partial derivatives of g with respect to θ. However, while these are standard properties of estimating functions, neither property is essential for studying the phenomenon of multiple roots. Under mild regularity, an estimating function will have a consistent root. See Crowder (1986). In addition, under reasonable regularity, any consistent root of g is unique with probability tending to 1; see Tzavelas (1998) for a proof of uniqueness for quasiscore functions. Thus

in these, and many other cases, our central interest is how the consistent root of the estimating function can be determined.

While many of the multiple root issues for estimating functions are identical to multiple root problems for the likelihood equations, the major difference is that estimating functions cannot typically be represented as the derivative of an objective function, such as the log-likelihood. This means that it is not possible to distinguish between the roots of a general estimating function as a likelihood does among its relative maxima (McCullagh, 1991). An exception to this is the case where we have an estimating function for a real-valued parameter. If θ is a real-valued parameter, it is possible to artificially construct an objective function, say $\lambda(\theta)$, whose derivative is $g(\theta)$ by integrating the estimating function

$$\lambda(\theta) = \int_{\theta_0}^{\theta} g(\eta)\, d\eta. \tag{5}$$

Here θ_0 is arbitrary and may be chosen for computational convenience. If g is the score function, then, up to an additive constant, λ is the log-likelihood. On the other hand, if g is the quasiscore function, then λ, as defined by (5), is the *quasilikelihood*. See McCullagh and Nelder (1989). However, for estimating functions other than the score function, $\lambda(\theta)$ is not justified by the usual theoretical considerations which justify likelihoods directly as measures of agreement between parameters and data. Moreover, when θ is a vector-valued parameter, the estimating function g also becomes vector-valued, and the line integral

$$\lambda(\theta) = \int_{\theta_0}^{\theta} g(\eta)\, d\eta^t \tag{6}$$

is typically *path-dependent*. Therefore it is not well defined. If g is the score vector, this ambiguity is avoided, because the vector field defined by g on Θ is conservative, being the gradient vector field of the log-likelihood. If $g(\theta)$ is a conservative vector field, then (6) is path-independent. However, this is not the case for general estimating functions such as quasi-score functions and others.

This paper is organized as follows. In Section 2 we consider a number of examples of estimating functions with multiple roots. In Section 3, we consider methods to detect the presence and probability of multiple roots. In Section 4, we consider methods for choosing a root of an estimating function when more than one root is present. In Section 5 we consider an alternative to this: the modification of an estimating function with multiple roots so that the number of roots is reduced. Finally, in Section 6, we consider how to build objective functions, that is, analogs of likelihoods, so as to compare the plausibility of various roots as is accomplished by likelihood analysis.

2. EXAMPLES

2.1 Estimation of the Correlation Coefficient

Consider a set of independent bivariate observations (x_i, y_i), $i = 1, \ldots, n$, from a bivariate normal distribution which is standardized to have means $\mu_x = \mu_y = 0$ and variances $\sigma_x^2 = \sigma_y^2 = 1$. We assume that there is an unknown correlation coefficient ρ between any x_i and y_i. The likelihood equation $\dot{l}(\rho) = 0$, where $l(\rho) = \log L(\rho)$, reduces to

$$P(\rho) = \rho(1-\rho^2) + (1+\rho^2)\frac{\sum xy}{n} - \rho\left[\frac{\sum(x^2+y^2)}{n}\right] = 0, \tag{7}$$

which can have as many as three real roots in the interval $(-1, 1)$. If three roots are present, then these will correspond to two relative maxima and one relative minimum of the likelihood.

We can check to see when this cubic equation is monotone by investigating the number of distinct real solutions to the quadratic equation $\dot{P}(\rho) = 0$. The cubic $P(\rho)$ will be monotone, and therefore have a unique real root, when $\dot{P}(\rho) = 0$ has at most one real solution. In turn, this will be true if the discriminant of the quadratic

$$D = 4\left(\frac{\sum xy}{n}\right)^2 + 12\left[1 - \frac{\sum(x^2+y^2)}{n}\right]$$

is zero or strictly negative. From the law of large numbers, we see that D converges to $4\rho^2 - 12$ as $n \to \infty$. So with probability converging to 1, the likelihood equation will have a unique root for large sample sizes.

To analyze this cubic equation further, let us define $S_1 = \sum xy/n$ and $S_2 = \sum(x^2+y^2)/n$. The pair (S_1, S_2) forms a minimal sufficient statistic for the estimation of ρ. Next, we perform a location shift $z = \rho - S_1/3$. Equation (7) reduces to

$$z^3 + a(S_1, S_2)\, z + b(S_1, S_2) = 0. \tag{8}$$

We can study the multiple solutions to this equation by plotting in $\mathbb{R}^3$ all points (a, b, z), where z is a root of (8) with given coefficients a and b. Figure 1 shows the surface so obtained. The resulting surface is an example of the well-known *cusp catastrophe*. With this interpretation, the coefficients a and b represent *control parameters* for the cusp catastrophe. In the control space, the projection of the folds of the surface defines the *separatrix*, whose

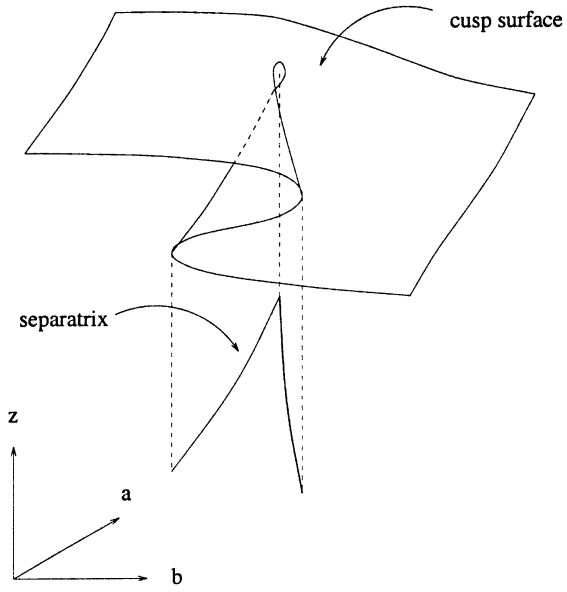

FIG. 1. *The cusp surface with the roots as functions of the control parameters.*

equations in (a, b) are $4a^3 + 27b^2 = 0$. The separatrix divides the control space into two regions. In the first region, $4a^3 + 27b^2 > 0$, and there is a single root. In the other region where $4a^3 + 27b^2 < 0$, the surface folds back on itself, so that there are three roots. The point $a = b = 0$ defines the control parameters of the cusp catastrophe where the two fold lines of the separatrix meet. See Figure 2. The reader is referred to Gilmore (1981, page 61) for more on the theory of the cusp catastrophe. The

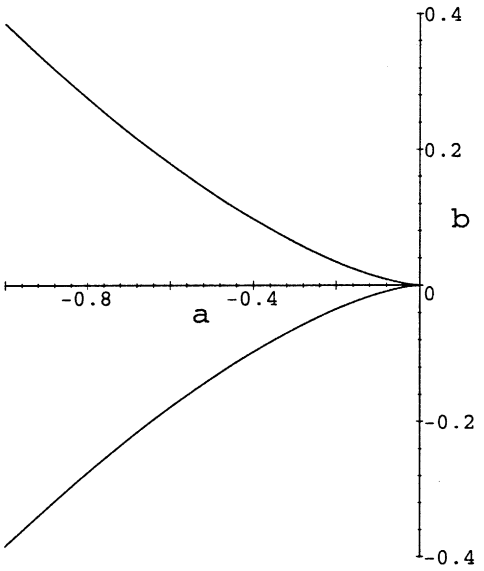

FIG. 2. *The separatrix dividing the control space into two regions.*

detailed analysis of the likelihood function and its extrema can be found in Stuart and Ord (1991).

2.2 Cauchy Location Model

The previous example may suggest that multiple root problems are small sample issues which will disappear for sufficiently large sample sizes. This commonly held belief is unfortunately too optimistic, as the following example illustrates. Suppose n variates are drawn independently from a Cauchy location model with common density function $f(y;\theta) = 1/\{\pi[1 + (y - \theta)^2]\}$. The likelihood equation becomes

$$\sum_{i=1}^{n} \frac{2(y_i - \theta)}{1 + (y_i - \theta)^2} = 0.$$

Upon taking a common denominator, the solution set for this equation is equivalent to that of

$$\sum_{i=1}^{n} \left\{ (y_i - \theta) \prod_{j \neq i} [1 + (y_j - \theta)^2] \right\} = 0,$$

which is a polynomial equation of degree $2n-1$. For extreme configurations of highly separated variates, the polynomial equation admits a full $2n-1$ distinct solutions, corresponding to n relative maxima and $n-1$ relative minima of the likelihood. Fortunately, as Reeds (1985) has shown, this extreme situation is rare. As $n \to \infty$, the asymptotic distribution of the number of relative maxima of the Cauchy likelihood converges to that of $1 + M$, where M has a Poisson distribution with mean $1/\pi$. A consequence of this is that the number of extraneous local maxima of the likelihood will be positive with an asymptotic probability given by $1 - e^{-1/\pi} \approx 0.2726$, which is less than one time in three. However, the probability that extraneous roots occur does not go to zero as the sample size gets large. See Figure 3.

A postscript to the investigation of the Cauchy distribution has been provided by Copas (1975), who showed that if the Cauchy location model is extended to include a scale parameter τ, then with probability 1 there is a unique solution to the simultaneous likelihood equations $\partial L/\partial\theta = 0$ and $\partial L/\partial\tau = 0$. The solution to these equations fails to be unique only if exactly 50% of the data values are coincident at some y_1 and the other 50% are coincident at some y_2. The other case worthy of special consideration occurs when more than 50% of the values are coincident at a point y. In this case there is no solution to the likelihood equations, and the likelihood is maximized on the boundary of the parameter space with $\hat{\theta} = y$ and $\hat{\tau} = 0$. Both of these special cases have probability 0.

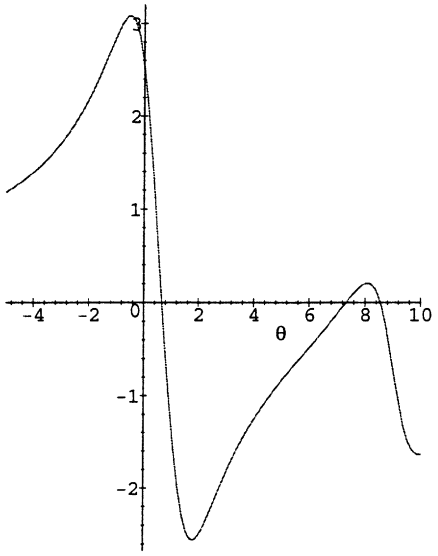

FIG. 3. *The score function for the Cauchy location model with an outlier producing extraneous roots.*

2.3 An Inconsistent Global Maximum of the Likelihood

In the previous example, the problem of multiple solutions for the likelihood equation did not disappear asymptotically. Nevertheless, we could reasonably expect to order the roots of the score function as estimators by calculating the likelihood at each root. The most appropriate root under this criterion would be that which globally maximizes the likelihood.

However, the usual Cramér conditions that are imposed for the asymptotic efficiency of the maximum likelihood estimate only ensure that the consistent root of the likelihood equations is efficient; there is no guarantee that the global maximum of the likelihood corresponds to a consistent root. So, in the absence of any regularity on the model, it is possible for this strategy to come undone for some parametric models. Examples due to Kraft and Le Cam (1956), Le Cam (1979), Bahadur (1958) and Ferguson (1982) illustrate that the global maximum of the likelihood can correspond to an inconsistent root of the score function; at the same time, some other root of the score function is a consistent estimator for the parameter. See also Le Cam (1990) and Example 3.1 in Lehmann (1983), Chapter 6. While inconsistent maximum likelihood estimates are well known from the examples of Neyman and Scott (1948), the examples due to Kraft and Le Cam and others are more problematic for likelihood methodology. This is because they do not involve the use of nuisance parameters, and satisfy the regularity conditions of Cramér (1946); at the same time, the global maximum likelihood estimator (MLE) is inconsistent. While it is possible to invoke regularity conditions such as those of Wald (1949) to ensure that the global maximum likelihood estimate is consistent, the conditions are difficult to check for models involving multiple roots. As the examples above show, the Wald conditions can fail for models which are, in other respects, regular.

2.4 Estimating the Normal Mean in Stratified Sampling

Suppose that random variables y_{ij} are independent $N(\xi, \sigma_i^2)$ and are divided into m strata, the ith stratum consisting of the variables $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Let $\bar{y}_i = \sum_j y_{ij}/n_i$ and $s_i^2 = \sum_j (y_{ij} - \bar{y}_i)^2/n_i$ denote the sample mean and variance, respectively, in stratum i. Suppose we are interested in estimating the common mean ξ based on $y_{i1}, \ldots, y_{in_i}$ for $i = 1, \ldots, m$. The maximum likelihood estimator for ξ is consistent but not efficient. For instance, it is less efficient than the estimator derived from the estimating equation

$$\sum_{i=1}^{m} w_i \frac{\bar{y}_i - \xi}{s_i^2 + (\bar{y}_i - \xi)^2} = 0 \tag{9}$$

with $w_i = n_i - 2$ as advocated by Bartlett (1936) and Neyman and Scott (1948). A profile likelihood theory (Barndorff-Nielsen, 1983) leads to the same estimating equation (9) with $w_i = n_i$. Sufficiency and ancillarity arguments (Kalbfleisch and Sprott, 1970) also leads to the same equation (9) but with $w_i = n_i - 1$.

Chaubey and Gabor (1981) noted that the profile likelihood may well be multimodal. Moreover, the class of estimating functions defined in (9) generally admits multiple roots (Barndorff-Nielsen, 1983). The geometric structure of roots to this class of estimating functions is essentially the same as the problem in the Cauchy location model. While the estimating equations are formally similar, the Cauchy and stratified normal models are quite different; hence the probabilities of multiple roots arising in the two models are different.

2.5 Regression with Measurement Error

Stefanski and Carroll (1987) have considered generalized linear models in which the covariates cannot be observed directly, but can only be measured with a certain amount of measurement error. Suppose that Y has density

$$f_Y(y; \alpha, \beta, \phi, u) = \exp\left[\frac{y(\alpha + \beta u) - b(\alpha + \beta u)}{a(\phi)} + c(y, \phi)\right], \tag{10}$$

where α and ϕ are unknown parameters, β is a row vector of parameters, u is a column vector of covariates, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known real-valued functions. Suppose that we cannot observe u, but can observe $X = u + \varepsilon$, etc. where ε has zero as its mean vector and covariance matrix $a(\phi)\Omega$, where Ω is known. A sample (x_i, y_i), $i = 1, \ldots, n$ of independent observations is given. For convenience we represent the full vector of parameters (α, β, ϕ) by θ.

In this model, the covariates $u_1, \ldots, u_n$ act as nuisance parameters for the problem of estimating θ. Stefanski and Carroll (1987) eliminated the nuisance parameters from the model by conditioning the joint density

$$\prod_{i=1}^{n} f_{X,Y}(x_i, y_i; \theta, u_i)$$

on a complete sufficient statistic for $(u_1, \ldots, u_n)$, namely $\delta = x + y\Omega\beta^t$. So an estimating equation for θ has the form

$$g(\theta) = \sum_{i=1}^{n} r(x_i, y_i; \theta) = 0,$$

where

$$r(x, y; \theta) = \frac{\partial}{\partial\theta} \log f_{Y|\delta}(y|\delta; \theta)$$

is the score function for the conditional model of Y given δ, which does not depend upon u.

Stefanski and Carroll considered the special case where Y has a normal distribution with mean $\alpha+\beta u$ and variance σ^2, and found that in general the estimating equation $g(\theta) = 0$ has multiple solutions. Stefanski and Carroll also reported that a similar problem of multiple roots arises in logistic regression with errors in covariates. In this case, Y is assumed to be a binary random variable with mean p, which relates to (α, β) through the canonical link

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta u.$$

We also assume the additive error model $X = u + \varepsilon$. Further analysis of this model can be found in Hanfelt and Liang (1995, 1997), where an objective function is constructed by a path-dependent integration approach.

2.6 Weighted Likelihood Equations

Markatou, Basu and Lindsay (1998) have proposed a modification of the likelihood method for data in which there is some suspicion that some observations may not be correctly modeled. They have introduced a weighting function to the likelihood equation which adaptively downweights those observations which appear to be inconsistent with the model. The *weighted likelihood equations* proposed take the form

$$\sum_{j=1}^{n} w_j(y_j, \beta, \widehat{F}) g(\theta, y_j). \tag{11}$$

Here $y_1, \ldots, y_n$ are assumed to be a random sample from the distribution $F = F_\theta$, and $\widehat{F}$ denotes the empirical distribution function. The function g could be an appropriately chosen estimating function for a single observation from F_θ. However, the authors specialize to the case where g is the (vector-valued) score function for θ based upon a single observation from F.

The weight function w is assumed to take values in the interval [0, 1]. In particular, $w(y_j, \beta, \widehat{F})$ will be close to 1 provided that in a neighborhood of the variable y_j the empirical distribution $\widehat{F}$ is concordant with the model distribution associated with the given value β of the parameter. The value of w is close to 0 when there is a large discordance between $\widehat{F}$ and the model in a neighborhood of y_j. In the case where the data are discrete, Markatou, Basu and Lindsay define the "neighborhood about y_j" to be the point y_j itself. A degree of concordance between the model and the data can be constructed by from a *Pearson residual*, defined for a random sample of size n as $\delta(y_j)$, where

$$\delta(t) = \frac{n^{-1}\#\{y_k\colon y_k = t\}}{P_\beta(y_k = t)} - 1. \tag{12}$$

The weight function w is then defined as

$$w = 1 - \frac{\delta^2}{(\delta+2)^2}.$$

In the continuous case, definition (12) is replaced by

$$\delta(t) = \frac{\int k(y;t)\, d\widehat{F}(y)}{\int k(y;t)\, dF_\beta(y)} - 1,$$

where k is some smooth kernel appropriate for kernel density estimation.

Markatou, Basu and Lindsay noted that such weighted equations can admit multiple roots. For example, they considered data of Lubischew (1962) describing bivariate measurements of two species of beetles. There were 21 bivariate observations for the species *Chaetocnema concinna* and 22 such observations for the species *Chaetocnema heptapotamica*. To test the method, the two species were artificially pooled, and a weighted likelihood estimate for the location of a bivariate normal distribution was conducted. The results were in agreement with the data: the weighting successfully separated the data by providing two roots as location estimates, one for each species.

3. DIAGNOSING MULTIPLE ROOT PROBLEMS

How can the researcher detect the possibility that an estimating equation can have multiple roots? When an equation can have multiple roots, do multiple root problems arise with a reasonably large probability? Having found one or more roots of an estimating equation, how can we be sure that there are no additional roots that remain undetected? Finally, without actually solving an equation, can we compute the distribution of multiple roots? In this section, we shall answer these questions in turn.

3.1 Detecting the Presence of Multiple Roots

First, we consider when an estimating function can have multiple roots. Perhaps the most common way of proving that the likelihood equations have a unique root is to show that the Hessian matrix

$$\ddot{l}(\theta) = \left(\frac{\partial^2 l(\theta)}{\partial \theta_j \, \partial \theta_k} \right) \tag{13}$$

of the log-likelihood $l(\theta)$ is negative definite for all values of θ. However, this is more than needs to be proved. In fact, we need only show that the Hessian matrix is negative definite *at the stationary points of the log-likelihood*. In particular, suppose that the parameter space Θ is an open, connected subset of $\mathbb{R}^k$. Let l: $\Theta \to \mathbb{R}$ be a twice continuously differentiable log-likelihood function on Θ such that the global maximum of $l(\theta)$ is achieved at some point in Θ. Next, we assume that $l(\theta) \to -\infty$ as θ goes to the boundary of Θ. (Note that the boundary may include the points at infinity if Θ is unbounded.) Finally, suppose that for all $\hat{\theta} \in \Theta$ satisfying $\dot{l}(\hat{\theta}) = 0$, the Hessian matrix $\ddot{l}(\hat{\theta})$ is negative definite. Then the equation $\dot{l}(\hat{\theta}) = 0$ will have a unique solution in Θ.

An illustration of this idea can be found in Figure 4. If two local maxima were to exist, then there would also be a saddle point of the likelihood surface where the likelihood equations would be satisfied, but the Hessian matrix would not be negative definite. Note that the concavity of the log-likelihood is not invariant under a reparametrization of the parameter space. Thus the Hessian matrix of a model may not be negative definite while the Hessian matrix of the likelihood under reparametrization may be negative definite. However, if the Hessian matrix is negative definite at any solution to the likelihood equations, it will also be negative definite at a solution in a reparametrized model.

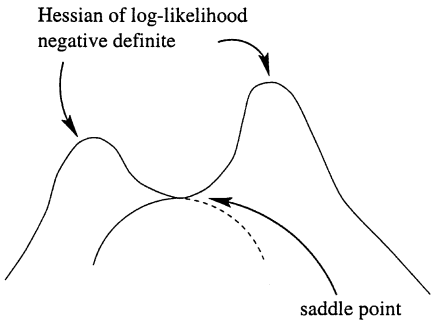

FIG. 4. *A likelihood with two or more maxima must have a saddle point where the Hessian matrix is not negative definite.*

An argument based upon $\ddot{l}(\hat{\theta})$ was used by Copas (1975) to show that the Cauchy location–scale model has a unique MLE. In another example, Huzurbazar (1948) showed that linear exponential families have unique solutions to their likelihood equations under mild regularity. This follows from a special identity for exponential families, namely $\ddot{l}(\hat{\theta}) = -I(\hat{\theta})$, where $I(\theta)$ is the expected information matrix. Under standard conditions, the expected information matrix is positive definite. So it follows that the root of the likelihood equation for an exponential family is unique. Note that if θ is the natural parameter of the exponential family, then the identity $\ddot{l}(\theta) = -I(\theta)$ holds for all θ. However, the equation $\ddot{l}(\hat{\theta}) = -I(\hat{\theta})$ holds even when θ is not the natural parameter, because the condition is invariant under smooth reparametrizations of the parameter space.

In models where the Hessian matrix fails to be negative definite, multiple root problems need investigation. However, the researcher should avoid the trap of presuming that the negative definiteness of the Hessian matrix is necessary for the uniqueness of the root. An interesting case in point is provided by the *Tobit model*, where

$$y_i^* = \beta \, x_i^t + \varepsilon_i,$$
$$y_i = \max\{0, \; y_i^*\}$$

for $i = 1, \ldots, n$. In this model, x_i is a vector of covariates, β is a coefficient vector and $\varepsilon_1, \ldots \varepsilon_n$ are independent $N(0, \sigma^2)$ error terms. The random variable y_i^* is not observed directly. Rather, we observe y_i. That is, y_i^* is observed only if it is nonnegative. Amemiya (1973) noticed that the Hessian matrix for the parameter vector $\theta = (\beta, \sigma^2)$ is not negative definite. Thus the question of multiple roots arose. However, Olsen (1978) showed that by letting $\zeta = \beta/\sigma$ and $\xi = 1/\sigma$ the Hessian is negative definite in the new parametrization with $\theta = (\zeta, \xi)$. So multiple roots cannot occur. See Amemiya (1973), Greene (1990), Olsen (1978), Orme (1990) and Iwata (1993) for discussion of this model. Burridge (1981) discusses the concavity of

the log-likelihood function in the case of regression with grouped data. See also Pratt (1981).

The task of detecting multiple roots for estimating functions in general is more problematic than that for the likelihood equations, as there may not exist a statistically meaningful objective function whose stationary points correspond to roots of the estimating equation. Some geometrical insight into the nature of an estimating function can be obtained by interpreting a vector-valued estimating function $g(\theta)$ as a vector field on the parameter space Θ. There are two possibilities:

1. *The matrix $\dot{g}(\theta)$ is symmetric for all θ and all samples $y_1, \ldots, y_n$.* In this case the vector field is conservative so that there exists a real-valued function $\lambda(\theta)$ such that $g(\theta) = \nabla\lambda(\theta)$. The function λ could be a log-likelihood or, in the case where the estimating function g is both unbiased as in (3) and information unbiased as in (4), may share some of the properties that are typically associated with log-likelihoods. In the case where Θ is one-dimensional, the symmetry condition is trivially satisfied.
   As $\dot{g}$ is symmetric, its eigenvalues will all be real. Those points $\hat{\theta} \in \Theta$ at which λ has a local maximum will correspond to points where the vector field g vanishes and the eigenvalues of $\dot{g}$ are all negative. Similarly, points at which λ is locally minimized will correspond to $\hat{\theta} \in \Theta$ where g vanishes and the eigenvalues will all be positive. Saddle points of λ will occur where g vanishes and the eigenvalues are mixtures of positive and negative quantities.
2. *The matrix $\dot{g}(\theta)$ is not symmetric in general.* In this case, there will be no objective function whose gradient is $g(\theta)$. The points in Θ where the vector field vanishes will correspond to the roots of $g(\theta)$.
   Despite the absence of an objective function, we can nevertheless determine roots of $g(\theta)$ which are analogs of local maxima and other roots which are analogs of local minima. To do this we investigate the eigenvalues of $\dot{g}(\theta)$. Let $\kappa_1(\theta), \kappa_2(\theta), \ldots, \kappa_k(\theta)$ be the eigenvalues of $\dot{g}(\theta)$, in arbitrary order, where $k = \dim(\Theta)$. As $\dot{g}$ is not symmetric, these eigenvalues will generally be complex-valued. Therefore, we can write each eigenvalue in terms of its real and imaginary parts as $\Re[\kappa_j(\theta)] + \sqrt{-1}\,\Im[\kappa_j(\theta)]$. A point $\hat{\theta} \in \Theta$ where $g(\hat{\theta}) = 0$ and where

$$\Re[\kappa_1(\hat{\theta})],\ \Re[\kappa_2(\hat{\theta})],\ \ldots,\ \Re[\kappa_k(\hat{\theta})] \tag{14}$$

   are all negative is a *sink* for the flow of the vector field determined by $g(\theta)$. Figure 5 shows a nonconservative vector field with two sinks. Similarly, if the k eigenvalues have real parts that are all positive, the point $\hat{\theta}$ is a *source* of the flow determined by the vector field.
   Although an objective function does not exist, there may exist analogs of maxima and minima among the roots of the estimating function. A sink of a vector field corresponds to a local maximum whereas a source corresponds to a local minimum. So it is natural to consider whether there is an immediate generalization of the uniqueness result illustrated in Figure 4 to the case where the vector field defined by $g(\theta)$ is not conservative. In other words, if a vector field is such that all points at which it vanishes are sinks, is it true that there can be at most one such point? Unfortunately, this generalization is false as Figure 5 demonstrates. Because there is a rotational component to the vector field, there does not exist an analog of a saddle point at which the vector field vanishes in the parameter space.

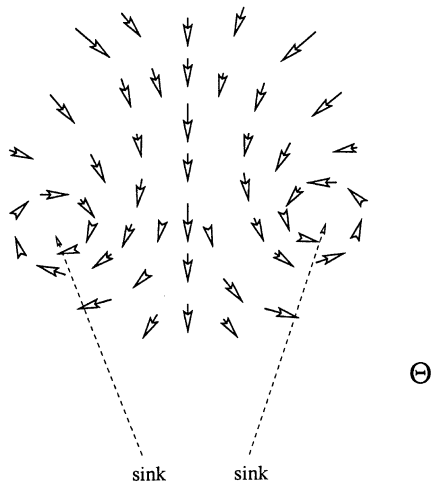

FIG. 5. *A nonconservative vector field which vanishes at exactly two distinct points, both of which are sinks.*

3.2 Finding All the Roots

Next, we turn our attention to the problem of detecting all roots for estimating functions which admit the possibility of multiple roots. In principle, a careful search in the parameter space should uncover all the roots of any given estimating function. However, in practice, this may be far too time-consuming, especially if the parameter space is of high dimension.

Markatou, Basu and Lindsay (1998) have incorporated a bootstrap root search into their analysis of the roots of weighted likelihood equations. As this

method seems applicable to a wide variety of estimating functions, we now give a brief description of the method. We begin by noting that the roots of an estimating function can often be divided into *reasonable roots*, which, upon examination, can be considered as candidates for estimation, and *unreasonable roots*, which arise in the estimating function for incidental reasons that have little to do with estimation. For example, the local maxima of the likelihood for the Cauchy location model can all be regarded as reasonable in this sense. However, the local minima are unreasonable and only arise from the fact that two local maxima must have a local minimum between them. We note also that, in many cases, all reasonable roots of an estimating function are those which are supported by some subset of the data. For example, in the case of the Cauchy location model, each relative maximum either is "caused" by a visual outlier (i.e., supported by an outlying observation) or is the consistent root that is supported by the majority of the observations which lie in the center of the Cauchy distribution. The unreasonable local minima are not supported by observations or subsets of observations in this sense.

Suppose that $y_1, \ldots, y_n$ are n independent observations from some distribution with parameter θ. Let $m \leq n$ be the minimum number needed for the equation

$$g(\hat{\theta};\, y_{i_1}, y_{i_2}, \ldots, y_{i_m}) = 0$$

to have a solution for all subsets $y_{i_1}, y_{i_2}, \ldots, y_{i_m}$ of size m with probability 1. Typically, m will be the dimension of the parameter space, although counterexamples to this can be found.

Markatou, Basu and Lindsay proposed that bootstrap samples $y_1^*, \ldots, y_m^*$ of size m be constructed by sampling m distinct elements of the data set $y_1, \ldots, y_n$. For each such bootstrap sample, the root θ^* found by solving

$$g(\theta; y_1^*, \ldots, y_m^*) = 0$$

is to be used as a starting point for an appropriate algorithm which iterates to a root of the equation $g(\theta; y_1, \ldots, y_n) = 0$.

So for estimating the location parameter θ of the Cauchy location model using the likelihood equations, we will have $m = 1$. As the MLE for θ based on a sample of size 1 is the observation $\theta^* = y_j$, itself, the set of roots obtained would be those found by using iteration from the n original sample observations. In those estimation problems where $\binom{n}{m} \leq 100$ it is possible to do an exhaustive systematic search of all such starting points by using all subsets. For $\binom{n}{m} > 100$, Markatou, Basu and Lindsay reported that randomization with 100 bootstrap samples is sufficient, in the cases they considered, to ensure that all reasonable roots are detected.

An approach to root detection by placing a probability distribution on the parameter space has been proposed by Finch, Mendell and Thode (1989). Their method provides a way to estimate the probability that an iterative search from a random starting point (RSP) will find a root not observed in previous searches from RSP's. Suppose that some probability distribution π is placed upon the parameter space. We begin by generating a random sample of size r from the distribution π, and useing each of these RSP's as the r initial values of an algorithm, such as Newton–Raphson, which searches for roots. In general, these r iterative trials will converge to a number of roots of the estimating function which we can write as $\hat{\theta}_1, \ldots, \hat{\theta}_K$, where $K \leq r$ is a random variable. For each $j = 1, \ldots, K$, let D_j be the domain of convergence of the algorithm to $\hat{\theta}_j$. The number of undetected roots cannot be estimated from $\hat{\theta}_1, \ldots, \hat{\theta}_K$. However, it is possible to estimate

$$U_r = 1 - \sum_{j=1}^{K} \pi(D_j). \tag{15}$$

As D_j is a random set, U_r is itself a random variable. Based upon a suggestion of Good (1953), Finch, Mendell and Thode (1989) suggested that U_r be estimated by $V_1 = S/r$, where S is the number of observed $\hat{\theta}_j$ to which only one of the r RSP's converged. This estimate can be generalized to

$$V_t = \sum_{i=1}^{t} \left[\frac{\binom{t-1}{i-1}}{\binom{r}{i}} \right] Q_i, \tag{16}$$

where Q_i is the number of roots among the k discovered to which exactly i RSP's converged. It can be shown that $E(V_t) = E(U_{r-t})$.

Through the use of statistics such as V_t, we can estimate the probability of detecting new roots with such a random search. So we can use such a measure to determine whether to continue searching further from additional RSP's. However, this does not tell us whether there are roots which are extremely unlikely to be detected because the choice of distribution π puts low probability on the domain of convergence of some root. The major hope for solving this problem may lie in bootstrap searches such as that of Markatou, Basu and Lindsay (1998) mentioned above.

3.3 Distributions of Roots

In some cases, it is possible to determine the number of roots that an estimating function has without

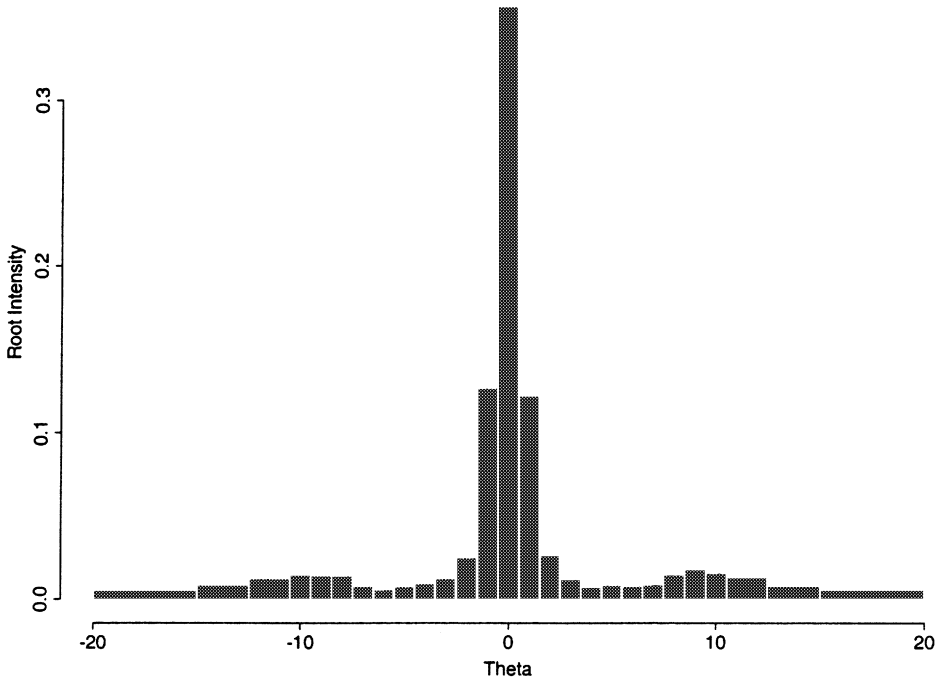

FIG. 6. *Histogram of the root intensity for the Cauchy location model.*

directly finding the roots themselves. This is well known for polynomial estimating equations, where *Sturm's theorem* can be used to determine the number of roots in an interval.

Suppose that θ is a real-valued parameter and that $g(\theta)$ is a polynomial in θ. We construct a *Sturm chain* for g by defining $g_0 = g$, $g_1 = \dot{g}$ and, for $n \geq 0$,

$$g_n(\theta) = g_{n+1}(\theta)\, q_{n+1}(\theta) - g_{n+2}(\theta), \tag{17}$$

where $\deg g_{n+2} < \deg g_{n+1}$. That is, $-g_{n+2}$ is the remainder term obtained when we divide g_n by g_{n+1}. Eventually, a remainder of zero is reached as the degree of the polynomials decreases in n. Next, we define a random variable $N(\theta)$ for every $\theta \in \Theta$, where $g(\theta) \neq 0$, by letting $N(\theta)$ be the number of sign changes of the sequence $g_0(\theta)$, $g_1(\theta)$, $g_2(\theta), \ldots$. Then Sturm's theorem says that the number of roots in the interval from a to b is exactly $N(a) - N(b)$, where $g(a) \neq 0$, $g(b) \neq 0$. In counting the number of roots, the theorem requires that a root with multiplicity greater than 1 be counted only once.

Sturm's theorem can be used in simulation studies of the distributions of roots. If a simulation is to be performed for more than 10,000 trials, it is usually too slow to investigate the exact locations of all roots in each trial. An alternative is to count the number of roots in successive intervals by constructing the Sturm chain and tabulating the results in a histogram. Software designed for symbolic computation with polynomials is particularly convenient for such simulations because the symbolic algebra required to construct the Sturm chain can be called as a subroutine.

Figure 6 shows the results of a simulation study using MAPLE V for 5,000 trials of sample size $n = 5$ for the Cauchy location model. Here the true value of θ was chosen to be zero. The area of each histogram bar represents the average number of roots found in each interval among the trials. Certain features are evident from the histogram. First of all, the presence of the consistent root close to zero is clear from the large mode at zero. The curious secondary modes on each side of the primary mode can be explained by the fact that, when the likelihood equation has multiple solutions, it will have local minima between the maxima. A more detailed investigation of the roots shows that the upcrossings of the score function are the principal cause of the secondary modes.

If we let the bin widths of the histogram go to zero, then the limiting form of the histogram will be a *root intensity function* $\psi\colon \Theta \to \mathbb{R}$. The characterizing property of ψ is that, for each measurable set $A \subset \Theta$,

$$E(N_A) = \int_A \psi(\theta)\, d\theta, \tag{18}$$

where N_A is the number of roots that g has in A. In those cases where g has a unique root with probabil-

ity 1, N_A will be an indicator random variable, and the left-hand side of (18) will reduce to $P(\hat{\theta} \in A)$. So for such estimating functions, ψ will be the density function of $\hat{\theta}$. Like a density function, ψ is defined except for a set of measure 0 by its integral over all measurable sets.

The root intensity function is similar to the intensity function of the point process of local maxima examined by Skovgaard (1990). The only difference here is that we do not restrict ourselves to roots that are associated with local maxima. The idea of applying the root intensity as a diagnostic tool for the multiple root problem was suggested by Small and Yang (1999). For many estimating functions, it is not necessary to find the roots explicitly in order to compute the root intensity function. Under certain regularity conditions, the root intensity ψ can be computed using the random vector $Z = g\dot{g}^{-1}$. Suppose the vector Z has density function $f_Z(z)$, where $z \in \mathbb{R}^k$. Then it can be shown that $\psi = f_Z(0)$. Note that dependence of ψ and Z upon the parameter θ is suppressed in this notation. The regularity conditions necessary to validate this formula are quite technical. See Skovgaard (1990) for a rigorous formulation of the regularity, and Small and Yang (1999) for more discussion in connection with the multiple root problem. The regularity conditions required for Skovgaard's formula were generalized by Jensen and Wood (1999).

4. METHODOLOGIES FOR ROOT SELECTION

We shall now turn to the problem of selecting a root for estimation when an estimating equation has multiple solutions. The first method we shall consider has its roots in Fisher's scoring of parameters and has been developed by C. R. Rao (1973) and Lehmann (1983).

4.1 Iterating from Consistent Estimators

For estimating functions with standard regularity, there is a unique consistent root which is isolated with high probability. More precisely, a neighborhood of the true value of θ can be found which contains exactly one root of g with a probability converging to 1 as $n \to \infty$. So an obvious strategy for selecting a root of g is to construct a consistent, albeit inefficient, estimator $\tilde{\theta}$ and to choose that root of g which is closest to the consistent estimator. The concept of the *closest root* could be measured by Euclidean distance in Θ. However, such a strategy does not pick a root in a parametrization-equivariant way. An alternative definition of the closest root to $\tilde{\theta}$ is that obtained by Newton–Raphson iteration with $\tilde{\theta}$ as a starting point.

An estimator $\tilde{\theta}$ is said to be $\sqrt{n}$-consistent for θ provided that $\sqrt{n}(\tilde{\theta} - \theta)$ is bounded in probability for every $\theta \in \Theta$. We define the *one-step estimator* for θ to be

$$\tilde{\theta}^* = \tilde{\theta} - g(\tilde{\theta})\dot{g}^{-1}(\tilde{\theta}). \tag{19}$$

For many estimating functions we can write

$$\sqrt{n}(\tilde{\theta}^*-\theta)=-\sqrt{n}\,g(\theta)\,\dot{g}^{-1}(\tilde{\theta})+[\sqrt{n}(\tilde{\theta}-\theta)]o_p(1). \tag{20}$$

The regularity conditions required for (20) to hold are similar to those for likelihood estimation found in Lehmann (1983, page 422), suitably modified for a general estimating function. As $\sqrt{n}(\tilde{\theta} - \theta) = O_p(1)$, by definition, the asymptotic distribution of the standardized statistic $\sqrt{n}(\tilde{\theta}^* - \theta)$ is the same as that of the first term in (20). For appropriately regular estimating functions, this will converge in distribution to $N(0, nI_g^{-1})$, where $I_g = I_g(\theta)$ is the *Godambe information* of the estimating function g defined by

$$I_g = E\dot{g}(Eg^t g)^{-1}E\dot{g}^t. \tag{21}$$

When g is the score function, the Godambe information reduces to the Fisher information.

This method of one-step estimation essentially replaces the problem of root selection with the problem of selecting an appropriate $\sqrt{n}$-consistent estimator. Unfortunately, the class of $\sqrt{n}$-consistent estimators is large. So the choice of $\tilde{\theta}$ within this class is critically important to the success of one-step iteration, as asymptotic considerations may not hold until n is very large. For example, if $\tilde{\theta}$ is $\sqrt{n}$-consistent, then so is $\tilde{\theta} + n^{-1}a$, no matter how large the constant a is. Clearly, in practice, if $a = 10^8$, we would could not consider $\tilde{\theta}$ and $\tilde{\theta}+n^{-1}a$ both appropriate.

4.2 Selecting Roots with Explicit Formulas

Heyde (1997) and Heyde and Morton (1998) have suggested that we can select among several roots in certain cases by finding explicit formulas for each root and examining the asymptotics for each formula. Such a method will be particularly useful if the estimating function is a polynomial of degree less than five, for all sample sizes. This straightforward approach to root selection is generally excellent and leads to sensible estimators. However, even when analytic formulas are available for all roots, it is not automatically true that the formula with the right asymptotic properties must be used.

A case in point is the estimation of the correlation coefficient using the cubic estimating function of (7). As the estimating equation is of degree 3, we

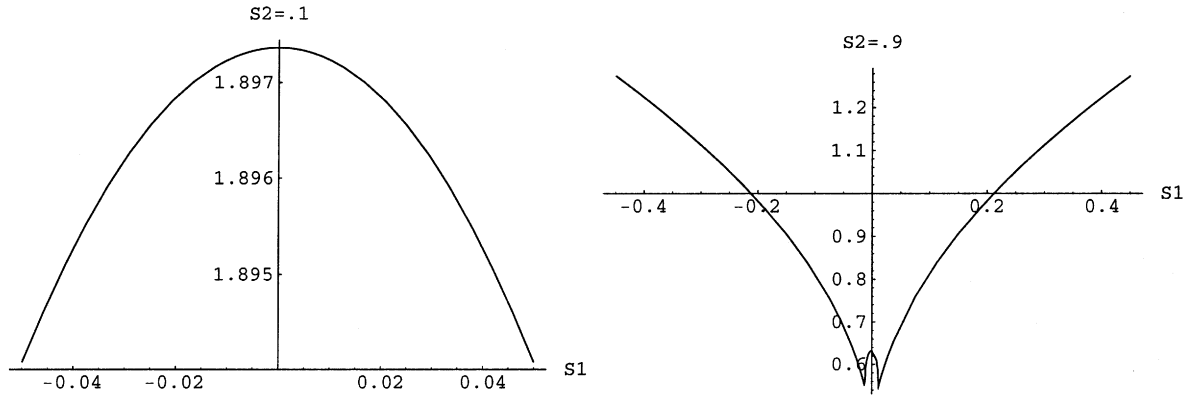

FIG. 7. *Plots of* $\Re[\hat{\rho}_1(S_1, S_2) + \hat{\rho}_1(-S_1, S_2)]$ *for two values of* S_2.

can find explicit formulas for the three roots using Cardano's formula:

$$\hat{\rho}_1 = \frac{1}{6} A_1 - 6 A_2 + \frac{1}{3} S_1, \tag{22}$$

$$\hat{\rho}_2 = \frac{-1 + \sqrt{-3}}{12} A_1 + 3(1 + \sqrt{-3}) A_2 + \frac{1}{3} S_1, \tag{23}$$

$$\hat{\rho}_3 = \frac{-1 - \sqrt{-3}}{12} A_1 + 3(1 - \sqrt{-3}) A_2 + \frac{1}{3} S_1, \tag{24}$$

where $S_1 = (\sum xy)/n$, $S_2 = \sum(x^2 + y^2)/n$,

$$A_1^3 = 12\,(-12 + 36S_2 + 132S_1^2 - 36S_2^2 - 48S_2S_1^2 + 12S_1^4 + 12S_2^3 - 3S_2^2S_1^2)^{-1/2} + 144S_1 - 36S_1S_2 + 8S_1^3$$

and $A_2 = (-3 + 3S_2 - S_1^2)/(9A_1)$.

To investigate the asymptotics of the root associated with each formula we can apply the limits $S_1 \to \rho$, $S_2 \to 2$, as $n \to \infty$. After some tedious algebra, we find that $\hat{\rho}_1 \to \rho$. The roots $\hat{\rho}_2$ and $\hat{\rho}_3$ converge to $+\sqrt{-1}$ and $-\sqrt{-1}$, respectively, as $n \to \infty$. This suggests that $\hat{\rho}_1$ is the root we want. As every cubic has at least one real root, it is tempting to suppose that $\hat{\rho}_1$ is real. However, the situation is not that simple. For example, when $S_1 = -0.1$ and $S_2 = 1.0$ then we have $\hat{\rho}_1 \approx 0.2 - 0.4\sqrt{-1}$.

An additional problem with $\hat{\rho}_1$ is that, unlike the maximum likelihood estimate, $\hat{\rho}_1$ is *not equivariant* with respect to reflections of the data through the x- and y-axes. To be equivariant, an estimator $\hat{\rho} = \hat{\rho}(S_1, S_2)$ should satisfy the equation $\hat{\rho}(-S_1, S_2) = -\hat{\rho}(S_1, S_2)$. However, this is not the case for $\hat{\rho}_1$. This problem arises because Cardano's method is not itself equivariant under these reflections. See Figure 7.

So even when it is possible to find an analytic formula for a consistent root of an estimating equation it is naive to assume that this formula is the obvious estimator. As we noted in Section 1, the uniqueness of the consistent root is an asymptotic result. Therefore two consistent root selection mechanisms may have radically different properties for a given sample size.

This example points out a difficulty with the use of explicit formulas for roots. However, in most contexts, the use of such root is recommended. If a root can be determined by an analytic formula and shown to be asymptotically consistent, then it is usually an excellent choice for estimation. Heyde and Morton (1998) have shown the efficacy of this method for a variety of multiple root problems.

4.3 Testing the Consistency of Roots

Suppose that there are roots $\hat{\theta}_1, \ldots, \hat{\theta}_m$ to estimating equation $g(\theta) = 0$. The problem of choosing a correct root may be formally studied as one of testing consistency for each root $\hat{\theta}_i$. Namely, we wish to consider testing the null hypotheses

$$H_0^{(i)}\text{: } \hat{\theta}_i \text{ is } \sqrt{n}\text{-consistent}$$

for $i = 1, \ldots, m$. The alternative to $H_0^{(i)}$ is that $\hat{\theta}_i$ is not $\sqrt{n}$-consistent. To make this hypothesis meaningful, we assume that the root $\hat{\theta}_i$ is defined simultaneously for all sample sizes, so that its asymptotic properties can be examined. To select a root we would find that root $\hat{\theta}_i$ at which a test statistic is least significant.

Heyde (1997) and Heyde and Morton (1998) have proposed two additional methods for testing roots:

1. picking the root $\hat{\theta}_i$ for which $\dot{g}(\theta)$ behaves asymptotically like $E_\theta \dot{g}(\theta)$ when evaluated at $\theta = \hat{\theta}_i$;
2. using a least squares or goodness-of-fit criterion to select the best root.

Both of these methods can be interpreted as selecting a root using a test statistic at each root. In the first case, the hypothesis

$$H'(\theta)\text{: } \dot{g}(\theta)[E_\theta \dot{g}(\theta)]^{-1} \to 1_{k \times k},$$

where $1_{k\times k}$ is the $k \times k$ identity matrix, is examined at $\theta = \hat{\theta}_1, \ldots, \hat{\theta}_m$ to determine at which root the hypothesis seems to be the most satisfactory. A test statistic for $H'(\theta)$ is based upon a comparison of $\dot{g}(\theta)$ and $E_\theta\, \dot{g}(\theta)$. The root which minimizes the value of the test statistic is chosen as the desired root of the estimating function.

In the second case, the test is constructed more directly from the data by examining the hypotheses

$$H''(\theta)\colon Y \sim P_\theta$$

at $\theta = \hat{\theta}_1, \ldots \hat{\theta}_m$. Here Y represents the observed data set and P_θ its distribution under the assumption that θ is the true value of the parameter. The test statistic for $H''(\theta)$ can be constructed by *partitioning* the sample space into cells and constructing a *chi-square* goodness-of-fit test or using a *weighted least-squares* criterion such as

$$\sum_{i=1}^{n} w_i(\theta)[Y_i - E_\theta(Y_i)]^2 \tag{25}$$

with some appropriate weighting function $w(\theta)$.

Naturally, we should not expect that a single testing procedure will work for all models. So the test statistic should be tailored to the particular features of the problem. For example, Singh and Mantel (1998) have noted that the choice of the least squares criterion can be critical to the success of the test. They have proposed modification of the least squares test statistic in (25) to a test statistic of the form

$$\left\{\sum_{i=1}^{n} w_i(\theta)[Y_i - E_\theta(Y_i)]\right\}^2. \tag{26}$$

4.4 A Bootstrap Method

It makes sense to focus attention on test statistics which are computationally convenient and fast to implement, at the same time applicable to a wide variety of problems. For this reason tests based upon resampling or cross-validation become particularly appealing. Suppose that $y = (y_1, \ldots, y_n)$ is a vector of observations. Let $Y^* = (Y_1^*, \ldots, Y_n^*)$ be a *bootstrap sample* drawn with replacement and with probability $1/n$ from the n observations $y_1, \ldots, y_n$. (The Y_i^*'s are iid even though the y_i's may not be.) Suppose g is an estimating function that is both unbiased and information-unbiased. A natural approach to testing is to use the bootstrap statistic

$$\gamma^*(\hat{\theta})_i = g(\hat{\theta}_i;\, Y^*)J^{-1}(\hat{\theta}_i;\, Y^*)g^t(\hat{\theta}_i; Y^*), \tag{27}$$

where

$$-2\widehat{J}_i = -2J(\hat{\theta}_i;\, y) = \dot{g}(\hat{\theta}_i;\, y) + \dot{g}^t(\hat{\theta}_i;\, y). \tag{28}$$

The motivation for this statistic is that if g is the score function, then $J(\hat{\theta}_i;\, y) = -\dot{g}(\hat{\theta}_i;\, y)$ is the observed information. When g is the score function, then $\dot{g}$ will be symmetric, and g will define a conservative vector field on Θ. In general g is not conservative and it is desirable to symmetrize $\dot{g}$. One immediate consequence of such symmetrization is that $J(\hat{\theta}_i;\, y)$ will have positive eigenvalues if $\hat{\theta}_i$ is consistent and n is large.

Next, let us suppose that g is an additive estimating function in the sense that we can write $g(\theta;\, y) = \sum_j h_j(\theta;\, y_j)$. Now for any root $\hat{\theta}_i$, the bootstrap distribution of $g(\hat{\theta}_i;\, Y^*)$ has expectation 0 and variance

$$\widehat{\Sigma}_i = \Sigma(\hat{\theta}_i) = \sum_{j=1}^{n} h_j(\hat{\theta}_i;\, y_j)\, h_j^t(\hat{\theta}_i;\, y_j).$$

When the sample size gets large, it may be reasonable to assume that the bootstrap distribution of $g(\hat{\theta}_i,\, Y^*)$ is approximately normal. If so, then $g(\hat{\theta}_i;\, Y^*) \sim N(0,\, \widehat{\Sigma}_i)$, approximately as $n \to \infty$. Therefore (27) would have a bootstrap distribution that is approximately

$$\gamma^*(\hat{\theta}_i) \sim \sum_{j=1}^{k} \lambda_j Z_j^2, \tag{29}$$

where the Z_j's are independent unit normal variates, and the λ_j's are the eigenvalues of $\widehat{\Sigma}_i \widehat{J}_i^{-1}$. Suppose, in addition, that $\hat{\theta}_i$ is $\sqrt{n}$-consistent. Then $\widehat{\Sigma}_i \widehat{J}_i^{-1} = 1_{k\times k} + O_p(n^{-1/2})$, because g is information-unbiased. It follows that $\lambda_j = 1 + O_p(n^{-1/2})$. So the bootstrap distribution of $\gamma^*(\hat{\theta}_i)$ is asymptotically $\chi^2(k)$. This result suggests a method for root selection. That root $\hat{\theta}_i$ whose *bootstrap distribution is closest to* $\chi^2(k)$ is a natural choice for $H_0^{(i)}$.

Note that the information-unbiasedness of g is a stronger property than is necessary. It suffices to assume that $g(\theta)$ is information-unbiased up to order $O(n^{1/2})$. That is, $E[\dot{g}(\theta) + g(\theta)g^t(\theta)] = O(n^{1/2})$.

In summary, we choose the root such that the bootstrap distribution of (27), or its asymptotic approximation (29), is closest to that of a chi-square variate with degrees of freedom $k = \dim \Theta$.

A simulation study was conducted to check the asymptotic approximation for γ^* and to determine its utility in selecting roots. The unbiased conditional score function derived for the logistic regression with measurement error discussed in Section 2.5 was used for the simulation study. We chose $k = 2$ with parameters α and β, and a sample of size $n = 100$. The standard deviation for the measurement error was chosen to be 0.8. For a sample generated using $\alpha = -1.4$ and $\beta = +1.4$,

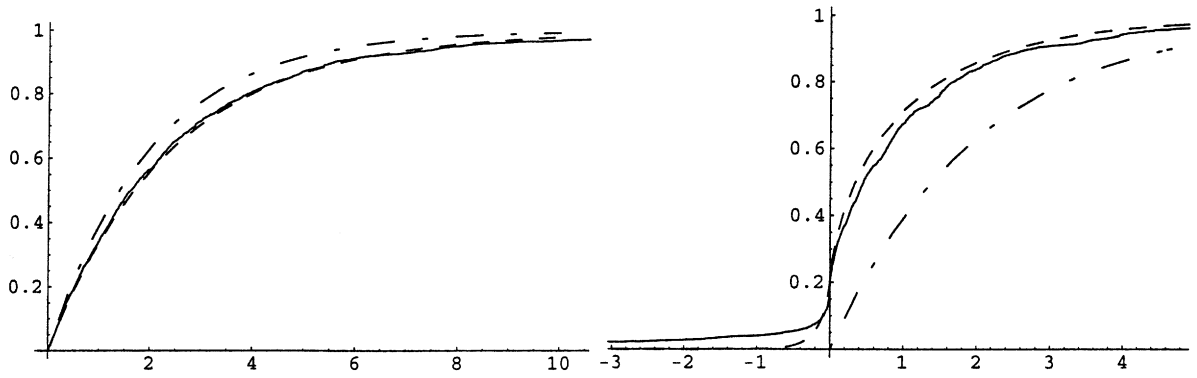

FIG. 8. *The bootstrap distribution of semiparametric score statistics in logistic regression with measurement error model.*

we found two roots $(\hat{\alpha}_1, \hat{\beta}_1) = (-1.05, 2.33)$ and $(\hat{\alpha}_2, \hat{\beta}_2) = (-0.96, 8.83)$. Figure 8 shows the bootstrap distribution of $\gamma^*(\hat{\alpha}_1, \hat{\beta}_1)$ and $\gamma^*(\hat{\alpha}_2, \hat{\beta}_2)$ based on 2,000 resamples at each of the two roots. On the left side are the simulated cdf's at $(\hat{\alpha}_1, \hat{\beta}_1)$ and on the right side are the corresponding cdf's at $(\hat{\alpha}_2, \hat{\beta}_2)$. The solid line shows the bootstrap distribution in each case. At $(\hat{\alpha}_2, \hat{\beta}_2)$ the bootstrap distribution has support on negative values. This follows from the fact that the matrix $\widehat{J}$ is not positive definite in a resampling neighborhood of the data. So the bootstrap distribution can be used to check the stability of the signs of the eigenvalues of $\widehat{J}$. In each diagram the line with alternating short and long dashes is the cdf for $\chi^2(2)$. As can be seen, the bootstrap distribution better approximates $\chi^2(2)$ at $(\hat{\alpha}_1, \hat{\beta}_1)$ than at $(\hat{\alpha}_2, \hat{\beta}_2)$. The plain dashed line in each plot shows the cdf of $\sum_i \lambda_i Z_i^2$, which is provided for the purposes of comparison.

Thus we choose $(\hat{\alpha}_1, \hat{\beta}_1) = (-1.05, 2.33)$ as our estimate for the true parameter $(\alpha, \beta) = (-1.4, 1.4)$.

4.5 Root Selection Based on Information

It is reasonable to expect that for certain statistical problems the consistent root of the estimating function will be *more informative* than other extraneous roots. In this section, we shall consider an information-based criterion for root selection in location models.

Let $Y_1, Y_2, \ldots, Y_n$ be iid random variables from the location model $f(y-\theta)$. So the score estimating function has the form $g(\theta) = \sum h(y_i - \theta)$, where $h(y) = -f'(y)/f(y)$.

We define the *shifted information function* to be

$$I(\theta_0, \theta) = \frac{[E_{\theta_0} h'(Y-\theta)]^2}{E_{\theta_0} h^2(Y-\theta)}. \tag{30}$$

In particular, $I(\theta, \theta)$ is the Godambe information at θ. See Godambe (1960). Now let us assume that, for all $t \in R$,

$$\lim_{|y|\to\infty} \frac{f(y+t)f'(y)}{f(y)} = 0. \tag{31}$$

Note that this condition is satisfied for a number of distributions including the Cauchy and normal. Using condition (31) and integrating by parts, we get

$$E_{\theta_0} h'(Y-\theta) = \int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)}\, dy$$

and

$$E_{\theta_0} h^2(Y-\theta) = \int_{-\infty}^{+\infty} \left[\frac{f'(y)}{f(y)}\right]^2 f(y+t)\, dy,$$

where $t = \theta - \theta_0$. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\left[\int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)}\, dy\right]^2 \\
&\quad = \left[\int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)f(y+t)}\, f(y+t)\, dy\right]^2 \\
&\quad \le \int_{-\infty}^{+\infty} \left[\frac{f'(y)}{f(y)}\right]^2 f(y+t)\, dy \int_{-\infty}^{\infty} \frac{[f'(y)]^2}{f(y)}\, dy,
\end{aligned}$$

which is equivalent to

$$I(\theta_0, \theta_0) \ge I(\theta_0, \theta). \tag{32}$$

To use this inequality as a method for root selection, we need to estimate the function $I_{\theta_0}(\theta) = I(\theta_0, \theta)$, where θ_0 is the true value of the parameter. We may use the sample mean as an estimate of the corresponding expectation. So

$$\widehat{I}_n(\theta) = \frac{[\sum_{j=1}^n h'(y_j - \theta)]^2}{n \sum_{j=1}^n h^2(y_j - \theta)} \tag{33}$$

can be regarded as an estimate of $I_{\theta_0}(\theta)$. As we shall see in greater detail in Section 5.3, some distributions are prone to visual outliers. For these cases, it may be more appropriate to replace the means

in 33 by trimmed means, along the lines suggested in Section 5.3.

The inequality (32) is similar in some respects to the inequality for the Kullback–Leibler information,

$$K(\theta_0, \theta) = E_{\theta_0} \log[L(\theta)/L(\theta_0)] \leq 0,$$

where equality holds if $\theta = \theta_0$. This suggests that, for certain models, an empirical estimate of the information in an estimating function can replace the likelihood as an objective function to be maximized. Therefore, we could choose the root $\hat{\theta}$ at which $\widehat{I}_n$ or a trimmed analog of $\widehat{I}_n$ is maximum. For example, a simulation of 2,000 trials for samples of size $n = 10$ found that the root which maximized the trimmed version of $\widehat{I}_n$ was the global maximum of the likelihood approximately 97% of the time.

Maximizing an information function may be most helpful when the estimating function is not the score function. For the method to be practicable, it is necessary that the inequality (32) hold for an additive estimating function, where we would define more generally

$$I(\theta_0, \theta) = \frac{\left[E_{\theta_0} \dot{h}(\theta)\right]^2}{E_{\theta_0} h^2(\theta)}.$$

Location parameters can also be estimated robustly by M-estimators. In those cases where the estimating function is redescending, multiple roots can arise. Clarke (1991) has given a method for selecting a robust root in such cases. Markatou, Basu and Lindsay (1998) have proposed the use of parallel disparity as a method of root selection.

5. MODIFYING ESTIMATING FUNCTIONS

5.1 A Warning about Modification

In this section, we shall consider how to modify an estimating function to eliminate or reduce the number of roots. However, before discussing such methods, a warning is in order. There are cases where the existence of multiple roots is informative in itself. For example, the presence of multiple roots in mixture models can serve as a diagnostic tool for the presence of different interpretations of the data. For an illustration of multiple roots in a mixture model, see Basford and McLachlan (1985). Markatou (1999a) has considered a problem in which the presence of multiple roots indicates more than one multiple mixture model fit. Formal test procedures for mixture model selection can be found in Markatou (1999b). Before eliminating multiple roots, the researcher should consider what information these roots provide about the fit of different models to the data.

5.2 Smoothing the Likelihood Function

Multiple root problems can be regarded as examples of excess variation in estimating functions. In several areas of statistics, a standard tool used to reduce variation is smoothing through the use of a moving average for a function, be it discrete or continuous. Daniels (1960) proposed the use of such a moving average to "reduce the chance of selecting one of the erratic cusps of the likelihood function" (Daniels, 1960, page 162). Daniels' smoothed likelihood was applied to the Cauchy location model by Barnett (1966, page 164).

Suppose θ is a real-valued parameter. Let $u_n: \mathbb{R} \to \mathbb{R}$ be a nonnegative function such that

$$\int_{-\infty}^{+\infty} u_n(y)\, dy = 0$$

and

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty} y^2 u_n(y)\, dy = 1.$$

The smoothed likelihood, with kernel u_n for sample size n, is defined to be

$$\bar{l}_n(\theta) = \int_{-\infty}^{+\infty} l_n(\theta - y) u_n(y)\, dy, \tag{34}$$

where $l_n(\theta) = \log L_n(\theta)$. There exist obvious extensions of (34) for multiparameter models. The parameter value $\bar{\theta}_n$ which maximizes $\bar{l}_n(\theta)$ is called a *smoothed maximum likelihood estimator*. We can also write $\bar{\theta}_n$ as the root of the smoothed score provided we can interchange derivatives and integrals:

$$\bar{g}_n(\theta) = \int_{-\infty}^{+\infty} \dot{l}_n(\theta - y) u_n(y)\, dy. \tag{35}$$

A particularly simple choice of u_n is $u_n(y) = 1/(2\varepsilon_n)$ for $|y| \leq \varepsilon_n$ and $u_n(y) = 0$ for $|y| > \varepsilon$. Then $\bar{\theta}_n$ will be a solution to the equation $L(\theta - \varepsilon_n) = L(\theta + \varepsilon_n)$. A simulation study by Barnett (1966) found that for the Cauchy location model it is possible to obtain efficiencies for $\bar{\theta}_n$ which exceed the efficiency of the MLE. In particular, a best improvement on the MLE was obtained by choosing $\varepsilon = 2.0$ for a sample of size $n = 5$. In this case the improvement in efficiency was found to be approximately 10%.

The value of ε_n needs to be chosen sufficiently large so that the equation $L(\theta + \varepsilon_n) = L(\theta - \varepsilon_n)$ has only one solution. However, we must also have $\varepsilon_n \to 0$ so that $\bar{\theta}_n$ is asymptotically efficient. Making ε_n large for a continuous likelihood function with finitely many relative extrema will ensure at most one solution, as Figure 9 illustrates. Ensuring that the equation has only one solution in this manner will require that ε be data-dependent. On

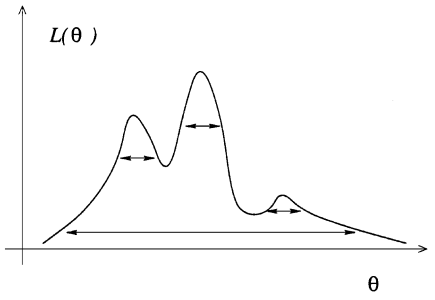

FIG. 9. *The equation $L(\theta+\varepsilon)=L(\theta-\varepsilon)$ for two values of ε.*

the other hand, to ensure efficiency we must examine the asymptotics. We can write

$$l_n(\theta\pm\varepsilon)=l_n(\theta)\pm\varepsilon\dot{l}_n(\theta)+\frac{\varepsilon^2}{2}\ddot{l}_n(\theta)\pm\frac{\varepsilon^3}{6}\dddot{l}_n[\theta+\xi(\pm\varepsilon)]$$

under smoothness conditions, where $0<\xi(\varepsilon)<\varepsilon$ and $-\varepsilon<\xi(-\varepsilon)<0$. So we have

$$\frac{l_n(\theta+\varepsilon)-l_n(\theta-\varepsilon)}{2\varepsilon}=\dot{l}_n(\theta)+\frac{\varepsilon^2}{12}\left\{\dddot{l}_n[\theta+\xi(\varepsilon)]+\dddot{l}_n[\theta+\xi(-\varepsilon)]\right\}. \tag{36}$$

The left-hand side of (36) can be regarded as an estimating function with $\bar{\theta}_n$ as a root. As such, it is generally a biased estimating function. However, if we focus our primary attention on those models for which $E_\theta l_n(\theta-\varepsilon)=E_\theta l_n(\theta+\varepsilon)$, then the essstimating function is unbiased. A location model with a symmetric density (such as the Cauchy) will satisfy this property. On the right-hand side, the first term is the score function, which is also unbiased under standard regularity conditions. Under the unbiasedness assumption, the term in brackets in (36) will be unbiased, and therefore will be of order $O_p(\sqrt{n})$. However, the asymptotic efficiency of a consistent root of (36) will be the asymptotic correlation between the left-hand side of (36) and the score function. So, if this is to be one, then the second term on the right-hand side of (36) must be of smaller order than the first term. As the first term is $O_p(\sqrt{n})$, we must have $\varepsilon=o_p(1)$.

Choosing $\varepsilon_n=o_p(1)$ runs counter to the requirement that ε_n is large enough to guarantee a unique root. A possible compromise is to use $\bar{\theta}_n$ as a starting point for Newton–Raphson iteration to a root of the score function or to define a one-step estimator. For this to work, the estimator $\bar{\theta}_n$ need only lie in a $\sqrt{n}$-neighborhood of the true parameter value and need not be efficient.

If a smoothed likelihood is used in a k-parameter model, then the score function becomes vector-valued, and the weight function u_n is a real-valued function of k variables. Once again, a uniform weight function can be used. However, in k dimensions, there are many types of regions available for smoothing.

5.3 Trimming the Estimating Function

Let us reconsider the Cauchy score function $g(\theta)=\sum_{j=1}^n h(y_j-\theta)$, where $h(x)=2x/(1+x^2)$. An example with four observations and multiple roots was given in Figure 3. This plot was generated by $y_1=1.1$, $y_2=0.5$, $y_3=9.0$ and $y_4=-0.3$. The extraneous roots are produced by the observation $y_3=9.0$, a visual outlier to which the Cauchy distribution is prone. Figure 10 shows the empirical distribution of $h(y_1-\theta),\ldots,h(y_4-\theta)$.

We see that the outlying observation y_3 produces extraneous roots because $h(y_3-\theta)$ has a large influence on the sum of $h(y_1-\theta),\ldots,h(y_4-\theta)$. So we can sometimes determine that a root is extraneous by examining the empirical distribution of the estimating function at the root. Figure 10 also suggests that robust estimators of location could be adapted to estimating functions to reduce the possibility of multiple roots. For example, we could replace the estimating function $\sum_j h(y_j-\theta)$ by the trimmed sum

$$g_{\mathrm{tr}}(\theta)=\sum_{j\neq i(\theta)} h(y_j-\theta), \tag{37}$$

where, for each θ, the value of $i(\theta)$ is such that $h[y_{i(\theta)}-\theta]$ is the most outlying observation among the values $h(y_1-\theta),\ldots,h(y_n-\theta)$. One way to define this value is to choose $i=i(\theta)$ so that

$$\left|h(y_i-\theta)-\frac{1}{n}\sum_{j=1}^n h(y_j-\theta)\right|$$

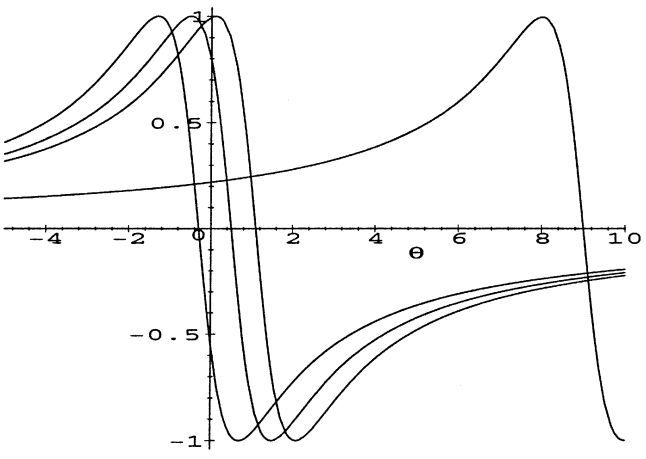

FIG. 10. *The empirical distribution of the Cauchy score function from Figure 3.*

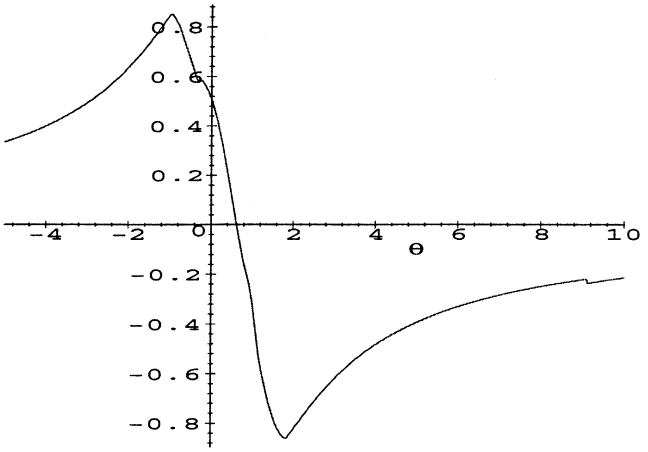

FIG. 11. *The median score of the Cauchy example from Figure* 3.

is maximum. This procedure was found to be particularly effective when used with the empirical methods discussed in Section 4.5, to which we refer the reader.

We can trim more than one value. If we trim to the middle of the empirical distribution, we obtain the *median score*, defined by

$$g_{\text{med}}(\theta) = \text{med}[h(y_1 - \theta), \ldots, h(y_n - \theta)]. \tag{38}$$

As shown in Figure 11, the median score has a unique root for our particular example. However, this estimating function is not differentiable. So the search for a root will require a slower but more reliable algorithm such as a binary search or the method of secants.

6. BUILDING AN OBJECTIVE FUNCTION

6.1 Motivation

We have seen that root selection provides a considerable challenge even for likelihood inference. It becomes even more involved when there does not exist an objective scalar function $\lambda = \lambda(\theta)$ such that $\nabla\lambda(\theta) = g(\theta)$. This case arises when the vector field on Θ defined by g is not conservative. In this section, we consider how to build an objective function and we discuss the connection with root selection.

There are several advantages to constructing an objective function. For example, an objective function allows us to *weight* roots of an estimating function and thereby to provide an objective ordering on the roots to assist root selection. Additionally, an objective function may allow us to *pool information* from the data with prior information, be it subjective or empirical. For example, if expert opinion were elicited to obtain probability weights $\pi_1, \ldots, \pi_m$ for each of m roots $\hat\theta_1, \ldots, \hat\theta_m$, then an objective function λ which behaves like a log-likelihood could be combined with the expert weights to produce posterior weights of the form $\pi_1\Lambda(\hat\theta_1), \ldots, \pi_m\Lambda(\hat\theta_m)$, where $\Lambda = \exp(\lambda)$.

Hanfelt and Liang (1995) have suggested using path-dependent integrals of the form (6) as objective functions and argue that for an optimal estimating function (McCullagh and Nelder, 1989, page 341) locally the dependence on the path should be small. Detailed simulation results are given for the logistic regression with measurement error model in which the root maximizing (6) is chosen. See Hanfelt and Liang (1995, 1997). In practice one has to compute many carefully chosen paths to compare the results. For the dependence on path to be practically negligible the estimating function under consideration must be very close to be conservative.

In this section we review two methods: projection in Section 6.2 and vector field decomposition in Section 6.4.

6.2 Projection

The quasiscore, given in (2), is the *projection* of the score function onto the space of linear and unbiased estimating functions spanned by the functions $y_1 - \mu_1(\theta), \ldots, y_n - \mu_n(\theta)$, where $\mu_i = E_\theta(Y_i)$. See McLeish and Small (1988), among others. As such, the quasiscore has a number of properties in common with a genuine score function. For instance, it inherits the first two Bartlett identities from the score function: it is unbiased and information-unbiased. For more on the quasi-score, see Wedderburn (1974) and McCullagh and Nelder (1989, Chapter 9).

So it is natural to seek a projection of the likelihood function or the likelihood ratio into a properly chosen space of functions. The elements of this space are functions of data and the parameter of interest. For such a method to be useful, we should be able to define the projection in terms of the low-order moments of the underlying variables. Suppose $y_1, \ldots, y_n$ are independent observations from unknown distributions having means $\mu_i(\theta)$, $i = 1, \ldots, n$, and variances $\sigma_i^2(\theta)$, $i = 1, \ldots, n$, respectively. For semiparametric models such as this, McLeish and Small (1992) have proposed an analog of the likelihood ratio $L(\eta)/L(\theta)$, namely,

$$\Lambda(\theta, \eta) = \prod_{i=1}^{n} \left\{ 1 + \frac{[\mu_i(\eta) - \mu_i(\theta)]}{\sigma_i^2(\theta)} [y_i - \mu_i(\theta)] \right\}. \tag{39}$$

See also Small and McLeish (1994). The function $\Lambda(\theta, \eta)$ can be obtained by projecting the likelihood ratio into a subspace which is the tensor product of the spaces generated by the n basis functions $y_i - \mu_i(\theta)$. This space enjoys certain kind of maximality

property, as outlined in McLeish and Small (1992). It is tangent to the quasiscore in the sense that

$$\frac{\partial}{\partial\eta}\Lambda(\theta,\eta)\bigg|_{\eta=\theta} = \sum_{i=1}^{n} \frac{\dot{\mu}_i(\theta)}{\sigma_i^2(\theta)}[y_i - \mu_i(\theta)].$$

Since $E_\theta\Lambda(\theta,\eta) = 1$, the function $\log \Lambda(\theta,\eta)$ is an exact local log-density (Severini, 1998). Moreover, $E_\eta\,\Lambda(\theta,\eta) \geq 1$, with equality if and only if $\mu_i(\eta) = \mu_i(\theta)$ for all $i = 1,\ldots,n$.

In one important respect, however, $\Lambda(\theta,\eta)$ does not resemble a likelihood ratio. It is not antisymmetric: $\Lambda(\theta,\eta) \neq \Lambda^{-1}(\eta,\theta)$ in general. This is due to the asymmetric roles that η and θ have in the projection. For the purposes of computing the projection, the parameter θ is assumed to be the true value of the parameter, and η is assumed to lie in a neighborhood of θ. An additional problem with Λ is that it is limited to independent observations, as can be seen from its multiplicative form.

The failure of Λ to be antisymmetric prompted Li (1993) to consider another type of projected likelihood that restores this property. He assumed a semiparametric model with first and second moment conditions as well. However, in this model, the independence assumptions on the variates is relaxed. Instead, we suppose that Y is vector-valued, with mean $\mu(\theta)$ and covariance matrix $\Sigma(\theta)$. Based on the approximation of $l(\eta) - l(\theta) = \log L(\eta)/L(\theta)$ by

$$l(\eta) - l(\theta) \approx \frac{L(\eta) - L(\theta)}{2L(\theta)} + \frac{L(\eta) - L(\theta)}{2L(\eta)}, \tag{40}$$

Li (1993) used a different projection argument to obtain the objective function

$$\begin{aligned}\lambda(\theta,\eta) = {} & \tfrac{1}{2}[\mu(\eta) - \mu(\theta)]\Sigma^{-1}(\theta)[y - \mu(\theta)]^t \\ & + \tfrac{1}{2}[\mu(\eta) - \mu(\theta)]\Sigma^{-1}(\eta)[y - \mu(\eta)]^t.\end{aligned} \tag{41}$$

It is straightforward to show that $\lambda(\theta,\eta)$ is tangent to quasi-score. Li (1993) also shows that under certain conditions

$$\mathrm{P}_\theta[\lambda(\theta,\xi) > \lambda(\eta,\xi)] \to 1 \quad \text{as } n \to \infty,$$

where $\eta \neq \theta$ and $\xi = \theta$ or η. We also refer the reader to Li and McCullagh (1994).

6.3 Projected Likelihoods and Root Selection

Projected likelihoods have been motivated by a need to construct confidence regions from estimating equations with multiple roots. See McCullagh (1991) and Li (1993). In this section, we shall discuss their connection with root selection. To be useful in this task, a projected likelihood must provide a consistent partial ordering of the parameter space. The following conditions are sufficient and would appear to be natural:

1. The ordering should be antisymmetric: $\lambda(\theta,\eta) = -\lambda(\eta,\theta)$ and $\Lambda(\theta,\eta) = \Lambda^{-1}(\eta,\theta)$.
2. The ordering should be transitive: if $\lambda(\theta,\eta) \geq 0$ and $\lambda(\eta,\xi) \geq 0$, then $\lambda(\theta,\xi) \geq 0$. Similar remarks hold for Λ.

The projected likelihood of Small and McLeish fails the antisymmetry condition. By contrast, Li's projected likelihood is constructed to be antisymmetric. However, it is easy to find examples where Li's likelihood is not transitive. For example, suppose Y is real-valued, and $\mu(\theta) = 0$, $\mu(\eta) = 1$ and $\mu(\xi) = 3$. Set $\sigma(\theta) = 1$ and $\sigma(\eta) = \sigma(\xi) = 0.1$. If $Y = 2.1$, say, then $\lambda(\theta,\eta)$, $\lambda(\eta,\xi)$ and $\lambda(\xi,\theta)$ are all positive. So neither type of projected likelihood can guarantee a consistent ordering of the parameter space.

Nevertheless, we do not need to compare all points in the parameter space to select roots. For example, if an estimating function has only two roots, then the issue of the transitivity of the projected likelihood is irrelevant to the problem of choosing one of these two values. In addition, the projected likelihood may well be transitive on the roots while not being transitive on the parameter space as a whole. Finally, even if it is not transitive or antisymmetric on the full set of roots, a projected likelihood may help us eliminate certain roots from consideration. This topic should provide a promising future research direction.

In the next section we consider a type of artificial likelihood which avoids these problems.

6.4 Artificial Likelihoods from Vector Field Decompositions

The question still remains: What is a proper objective function for an arbitrary estimating function $g(\theta)$? The function $g(\theta)$ may use semiparametric information about more than the first two moments and in general need not be conservative. Since $g(\theta)$ may be viewed as a vector field defined in the parameter space, the nonexistence of an objective function implies that the vector field curls or rotates. Wang (1999) shows that, for any $k = \dim\,\Theta = \dim\, g \geq 2$, one can write $g = g_c + g_r$, where g_c is conservative (or irrotational) and g_r is divergence-free (or solenoidal). This is a generalization of the well-known Helmholtz decomposition for $k = 3$.

Having found such a decomposition, we can construct an objective function by discarding the divergence-free component and integrating the conservative part:

$$\lambda(\theta,\eta) = \int_\theta^\eta g_c(\xi)\,d\xi^t. \tag{42}$$

A problem with this definition of $\lambda(\theta, \eta)$ is that the decomposition is not unique. For example, suppose g is conservative to begin with. Then we would normally expect that the divergence-free part would be zero. However, with no additional guidance as to how to perform the decomposition, there is no reason for this to be true. Since we intend to discard the divergence-free part, it should contain as little statistical information as possible. This suggests two principles for the construction of the decomposition:

1. If $\dot{g}$ is symmetric, then $g_r = 0$.
2. If $E_\theta \dot{g}(\theta)$ is symmetric, then $E_\theta g_r(\theta) = 0$.

The solution to this problem involves the use of the geometry of differential forms—a topic that we shall not discuss here. See Darling (1994). A form of the decomposition which would satisfy these two principles is

$$g = d\lambda + *\,d\,[\alpha * dg],$$

where $\alpha = \alpha(\theta)$ is an appropriate scalar function of θ only, $*$ is the Hodges star operator and d is the exterior derivative. The vector field $g_r^o = *\,d\,[\alpha * dg]$ vanishes if g is conservative. Similarly, $E_\theta\, g_r^o = 0$ if $E_\theta \dot{g}(\theta) = 0$ for all θ. In general, there is no guarantee that such a scalar function $\alpha(\theta)$ exists. However, Wang (1999) shows that if g is linear in θ, then there exists such a scalar function $\alpha(\theta)$ so that the decomposition $g = d\lambda + g_r^o$ holds. He also gave an explicit formula for $\alpha(\theta)$, namely,

$$\alpha(\theta) = (-1)^k \tfrac{1}{4}\theta\theta^t. \tag{43}$$

See Wang (1999) for further justification for the choice of (43). The above arguments suggest that we use a linear approximation to g locally around the parameter values of interest. Such points of interest include, but are not restricted to, the set of roots of g. Using (43), we obtain the objective (potential) function

$$\lambda(\theta, \eta) = \tfrac{1}{2}(\eta - \theta)J(\theta)(\eta - \theta)^t + g(\theta)(\eta - \theta)^t, \tag{44}$$

where $2J(\theta) = \dot{g}(\eta) + \dot{g}^t(\eta)$ is the symmetrized Hessian.

That the objective function (44) is a semiparametric likelihood function can be argued along the lines of Severini (1998). For the quasiscore (44) becomes

$$\begin{aligned}\lambda(\theta, \eta) = &-\tfrac{1}{2}[\mu(\eta) - \mu(\theta)]\Sigma^{-1}(\theta)[\mu(\eta) - \mu(\theta)]^t \\ &+ [\mu(\eta) - \mu(\theta)]\Sigma^{-1}(\theta)[Y - \mu(\theta)]^t,\end{aligned}$$

a form closely resembling (39) and (41). From (44), we can formally define an artificial likelihood ratio statistic and study its properties. Indeed, root selection based on (27) is a bootstrap version of the artificial likelihood ratio.

REFERENCES

AMEMIYA, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* **41** 997–1016.

BAHADUR, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20** 207–210.

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

BARNETT, V. D. (1966). Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **53** 151–165.

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

BASFORD, K. E. and MCLACHLAN, G. J. (1985). Likelihood estimation with normal mixture models. *J. Roy. Statist. Soc. Ser. C* **34** 282–289.

BURRIDGE, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B* **43** 41–45.

CHAUBEY, Y. P. and GABOR, G. (1981). Another look at Fisher's solution to the problem of the weighted mean. *Comm. Statist. A* **10** 1225–1237.

CLARKE, R. B. (1991). The selection functional. *Probab. Math. Statist.* **11** 149–156.

COPAS, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62** 701–704.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

CROWDER, M. (1986). On consistency and inconsistency of estimating equations. *Econom. Theory* **2** 305–330.

DANIELS, H. E. (1960). The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 151–163. Univ. California Press, Berkeley.

DARLING, R. W. R. (1994). *Differential Forms and Connections*. Cambridge Univ. Press.

FERGUSON, T. S. (1982). An inconsistent maximum likelihood estimate. *J. Amer. Statist. Assoc.* **77** 831–834.

FINCH, S. J., MENDELL, N. R. and THODE, H. C. (1989). Probability measures of adequacy of a numerical search for a global maximum. *J. Amer. Statist. Assoc.* **84** 1020–1023.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

GILMORE, R. (1981). *Catastrophe Theory for Scientists and Engineers*. Dover, New York.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31** 1208–1212.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264.

GREENE, W. (1990). Multiple roots of the Tobit log-likelihood. *J. Econometrics* **46** 365–380.

HANFELT, J. J. and LIANG, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82** 461–477.

HANFELT, J. J. and LIANG, K.-Y. (1997). Approximate likelihood for generalized linear errors-in-variables models. *J. Roy. Statist. Soc. Ser. B* **59** 627–637.

HEYDE, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer, New York.

HEYDE, C. C. and MORTON, R. (1998). Multiple roots in general estimating equations. *Biometrika* **85** 954–959.

HUZURBAZAR, V. S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. *Ann. Eugenics* **14** 185–200.

IWATA, S. (1993). A note on multiple roots of the Tobit log likelihood. *J. Econometrics* **56** 441–445.

JENSEN, J. L. and WOOD, A. T. A. (1999). Large deviation results for minimum contrast estimators. *Ann. Inst. Statist. Math.* To appear.

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

KRAFT, C. H. and LE CAM, L. M. (1956). A remark on the roots of the maximum likelihood equation. *Ann. Math. Statist.* **27** 1174–1177.

LE CAM, L. (1979). *Maximum Likelihood: An Introduction. Lecture Notes in Statist.* **18**. Springer, Berlin.

LE CAM, L. (1990). Maximum likelihood: an introduction. *Internat. Statist. Rev.* **58** 153–171.

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

LI, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80** 741–753.

LI, B. and MCCULLAGH, P. (1994). Potential functions and conservative estimating functions. *Ann. Statist.* **22** 340–356.

LUBISCHEW, A. (1962). On the use of discriminant functions in taxonomy. *Biometrics* **18** 455–477.

MARKATOU, M. (1999a). Mixture models, robustness and the weighted likelihood methodology. Technical report, Dept. Statistics, Stanford Univ.

MARKATOU, M. (1999b). Model selection based on disparity measures with applications to mixture models. Technical report, Dept. Statistics, Columbia Univ.

MARKATOU, M., BASU, A. and LINDSAY, B. G. (1998). Weighted likelihood equations with bootstrap root search. *J. Amer. Statist. Assoc.* **93** 740–750.

MCCULLAGH, P. (1991). Quasi-likelihood and estimating functions. In *Statistical Theory and Modelling, in Honour of Sir David Cox* (D. V. Hinkley, N. Reid and E. J. Snell, eds.) 267–286. Chapman and Hall, London.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MCLEISH, D. L. and SMALL, C. G. (1988). *The Theory and Applications of Statistical Inference Functions. Lecture Notes in Statist.* **44**. Springer, New York.

MCLEISH, D. L. and SMALL, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79** 93–102.

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

OLSEN, R. (1978). Note on the uniqueness of the maximum likelihood estimator of the Tobit model. *Econometrica* **46** 1211–1215.

ORME, C. (1990). On the uniqueness of the maximum likelihood estimator in truncated regression models. *Econom. Rev.* **8** 217–222.

PERLMAN, M. D. (1983). The limiting behavior of multiple roots of the likelihood equation. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday* (M. Rizvi, J. Rustagi and D. Siegmund, eds.) 339–370. Academic Press, New York.

PRATT, J. W. (1981). Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76** 103–106.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

REEDS, J. A. (1985). Asymptotic number of roots of Cauchy location likelihood equations. *Ann. Statist.* **13** 775–784.

SEVERINI, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85** 507–522.

SINGH, A. C. and MANTEL, H. J. (1998). Minimum chi-square estimating function and the problem of choosing among multiple roots. *Proc. Amer. Statist. Assoc.* To appear.

SKOVGAARD, I. M. (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18** 779–789.

SMALL, C. G. and MCLEISH, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. Wiley, New York.

SMALL, C. G. and YANG, Z. (1999). Multiple roots of estimating functions. *Canad. J. Statist.* **27** 585–598.

STEFANSKI, L. A. and CARROLL, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74** 703–716.

STUART, A. and ORD, J. K. (1991). *Kendall's Advanced Theory of Statistics* **2**. *Classical Inference and Relationship.* Arnold, London.

TZAVELAS, G. (1998). A note on the uniqueness of the quasi-likelihood estimator. *Statist. Probab. Lett.* **38** 125–130.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

WANG, J. (1999). Nonconservative estimating functions and approximate quasi-likelihoods. *Ann. Inst. Statist. Math.* **51** 603–619.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss–Newton method. *Biometrika* **61** 439–447.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for generalized linear models. *Biometrika* **63** 27–32.

Comment

John J. Hanfelt

I suspect that most statisticians doubt the proposition that multiple roots pose a serious problem in data analyses. Certainly, very few published data analyses in scientific journals mention the existence of multiple roots or describe what methods were used to select among them. This state of affairs is perhaps not so surprising, given two unfortunate facts:

1. The presence of multiple roots is likely to go undetected in a data analysis unless the researcher specifically examines the issue;
2. the theory for dealing with multiple roots is underdeveloped, and the few existing statistical methods are not widely known.

Small, Wang and Yang (SWY) are to be commended for raising awareness of the problem of multiple roots; SWY provide an interesting theoretical discussion of multiple roots. Especially valuable is their comprehensive review of methods to locate and select among multiple roots of an estimating function. One reservation is that the majority of examples in SWY do not have much practical importance.

For applied statisticians to better appreciate the significance of the problem, it would be interesting to hear more from SWY about which statistical models in common use today are prone to generating multiple roots.

My encounters with multiple roots have been limited to applications involving regression with measurement error (such as discussed in Sections 2.5 and 4.4 of SWY) and mixture models (Lindsay, 1983; Everitt and Hand, 1981). To illustrate, consider a latent class analysis of the relationships among nine criteria (each coded as 1 = present, 0 = absent) used in the diagnosis of schizotypal personality disorder (Nestadt et al., 1994). For simplicity, here we confine our attention to a model with only two latent classes and $n = 479$ individuals with a family history of schizophrenia. The EM algorithm, as described by Everitt and Hand, was used to identify local maxima of the likelihood, of which two, $\hat{\theta}_1$ and $\hat{\theta}_2$, are shown in Table 1. Neither solution can be eliminated as scientifically implausible. Fortunately, here we are working within a fully parametric framework and have available the log-likelihood ratio, $\log\{L(\hat{\theta}_1)/L(\hat{\theta}_2)\} = 9.68$, indicating that $\hat{\theta}_1$ is better supported than $\hat{\theta}_2$. A more detailed analysis by Nestadt et al. (1994), based on a larger sample, revealed that the best solution actually has four latent classes.

An important, unresolved issue is how to construct an appropriate objective function $\lambda(\theta, \eta)$ for distinguishing among multiple roots of a general estimating function $g(\theta)$. The projected likelihoods discussed in Section 6.2 and 6.3 of SWY have some attractive properties, but are limited to estimating functions of the form $g(\theta) = \sum w_i(\theta)\{y_i - E_\theta(Y_i)\}$. Hanfelt and Liang (1995) proposed a generalized version of SWY's (41), namely,

$$\lambda(\theta, \eta) = \tfrac{1}{2}E_\eta\{g(\theta)|A_\eta\}\ \mathrm{var}_\theta^{-1}\{g(\theta)|A_\theta\}g(\theta)^t - \tfrac{1}{2}E_\theta\{g(\eta)|A_\theta\}\ \mathrm{var}_\eta^{-1}\{g(\eta)|A_\eta\}g(\eta)^t,$$

where A_θ is an optional conditioning argument, but noted that this projection approach does not exist for certain applications. Another limitation is that all versions of projected likelihood require the correct modelling of $\mathrm{var}_\theta\{g(\theta)|A_\theta\}$, which often is unrealistic in applications. Small, Wang and Yang suggest a different objective function based on a vector field decomposition approach. Their method is interesting, but seems to require that $g(\theta)$ be linear in θ, which limits its versatility. Clearly, more research is needed on artificial likelihoods to accompany estimating functions.

A connection between goodness-of-fit criteria and artificial log-likelihoods is provided by the concept of *artificial deviance*. Suppose we can write the estimating function as

$$g(\theta) = \sum_{i=1}^{n} \dot{\psi}_i(\theta)^t g_i\{\psi_i(\theta), y_i\},$$

where g_i and natural parameter $\psi_i = \psi_i(\theta)$ are column vectors of length $q_i \geq 1$, say, θ is a vector of length $p < n$ and $\dot{\psi}_i(\theta)$ is a $q_i \times p$ matrix of first partial derivatives. Suppose also that we have available an artificial log-likelihood $\lambda\{\psi(\theta), \psi(\eta)\}$, where $\psi(\theta) = \{\psi_1(\theta)^t, \ldots, \psi_n(\theta)^t\}^t$, to accompany $g(\theta)$. The idea is to measure the discrepancy of fit

John J. Hanfelt is Assistant Professor, Department of Biostatistics, Emory University, Atlanta, Georgia 30322.

TABLE 1
Two local maxima of a latent class model for the relationships among criteria used to diagnose schizotypal personality disorder. Entries denote the conditional probability of a criterion being present given membership in latent class I or II

	$\hat{\theta}_1$		$\hat{\theta}_2$	
	Latent classes		**Latent classes**	
	I	**II**	**I**	**II**
Prevalence	0.065	0.935	0.124	0.876
Criteria				
Ideas of reference	0.340	0.003	0.192	0.002
Social anxiety	0.537	0.077	0.483	0.053
Odd beliefs	0.429	0.033	0.246	0.032
Unusual perceptions	0.554	0.020	0.314	0.018
Odd behavior	0.097	0.000	0.051	0.000
No confidants	0.710	0.219	0.686	0.189
Odd speech	0.032	0.000	0.017	0.000
Inappr. affect	0.165	0.022	0.253	0.000
Suspiciousness	0.640	0.018	0.414	0.008

between the parsimonius model $g(\theta)$ and a saturated model via the artificial deviance

$$D(\theta) = 2\lambda\{\psi(\theta),\ \tilde{\psi}\},$$

where $\tilde{\psi} = \{\tilde{\psi}_1^t, \ldots, \tilde{\psi}_n^t\}^t$ is a function of the data such that $g_i(\tilde{\psi},\ y_i) = 0$, $i = 1, \ldots, n$. For example, in the special case where the estimating function is the quasiscore

$$(1) \qquad g(\theta) = \sum_{i=1}^{n} \dot{\mu}_i(\theta)^t \ \mathrm{var}^{-1}\{y_i; \mu_i(\theta)\}\{y_i - \mu_i(\theta)\},$$

and the choice of objective function $\lambda\{\mu(\theta),\ \mu(\eta)\}$ is given by SWY's (41), the artificial deviance reduces to the weighted least squares criterion

$$D(\theta) = \sum_{i=1}^{n}\{y_i - \mu_i(\theta)\}^t \ \mathrm{var}^{-1}\{y_i; \mu_i(\theta)\}\{y_i - \mu_i(\theta)\}.$$

The concept of artificial deviance might be useful in constructing appropriate goodnes-of-fit criteria for estimating functions $g(\theta)$ more general than quasiscore (1). See related work by Qian, Gabor and Gupta (1996) and Baggerly (1998). Note that if objective function $\lambda\{\psi(\theta),\ \psi(\eta)\}$ does not satisfy the transitive property (Section 6.3 of SWY), then inferences based on $D(\theta)$ might not be consistent with those based on direct comparisons of multiple roots; that is, we might not have the relation

$$D(\theta) > D(\eta) \quad \Rightarrow \quad \lambda(\theta,\ \eta) > 0.$$

A final comment on SWY: the discussion after (29) is rather puzzling. The authors seem to claim that if root $\hat{\theta}_i$ is $\sqrt{n}$-consistent, then $g(\theta)$ is necessarily information-unbiased and $\chi^2(k)$ is always the appropriate reference distribution for test statistic $\gamma^*(\hat{\theta}_i)$. However, clearly this claim is wrong: $g(\theta)$ does not have to be information-unbiased, even approximately, to generate a $\sqrt{n}$-consistent root, and so the reference distribution for $\gamma^*(\hat{\theta}_i)$ generally is given by (29). A case in point is the measurement error problem examined at the end of Section 4.4 of SWY, where the conditional score function is not information-unbiased. Perhaps the authors, in advocating reference to a $\chi^2(k)$ distribution, have in mind testing whether root $\hat{\theta}_i$ satisfies certain asymptotic efficiency properties? Some clarification would be helpful here.

ACKNOWLEDGMENT

The author thanks Ann E. Pulver, D.Sc., for kindly sharing her data on schizotypal personality disorder.

Comment

C. C. Heyde

Multiple roots of estimating equations is a topic which is interesting, important and often perceived as a source of serious complication (e.g., McCullagh 1991; Stefanski and Carroll, 1987, Section 2.3). It is of considerable value to have the overview of methods for treating this problem which is provided in this paper. However, it seems to me that the Small, Wang and Yang (SWY) have rather overemphasized the likelihood case and have not sought to demystify the topic as constructively as they might have done.

From the outset a tone is adopted which makes the subject seem more complicated than it really

Chris Heyde is Professor, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia, and Department of Statistics, Columbia University, 2990 Broadway, Mail Code 4403, New York, New York 10027.

is in practice. The discussion about infinitely many roots early in the Introduction of SWY's paper is a case in point. Another example occurs later in the Introduction in the context of multiple roots for estimating equations, where the possible problems with a likelihood analog approach are outlined, but the existence of simple and reasonable alternative ways to treat the problem are not explicitly mentioned, despite their treatment later in the paper.

It is my contention that multiple root problems in which a satisfactory and unambiguous choice cannot readily be found are rare. In illustration of this view I shall briefly discuss just two of the major examples of the paper.

First, take the case of the Cauchy location model discussed in Section 2.2. It has been noted that the likelihood estimating equation in the case of a sample of size n is a polynomial of degree $2n-1$ and the number of relative maxima among the solutions converges as $n \to \infty$ to $1+M$, where M has a Poisson distribution with mean $1/\pi$. However, the sample median is a convenient consistent estimator of the location parameter and a reasonable practical approach could be to use the root of the estimating equation which is closest to the median.

The next example concerns the estimation of the correlation coefficient from a bivariate normal sample (x_i, y_i), $i = 1, 2, \ldots, n$ which is treated in Sections 2.1 and 4.2. In this setting, with variables having means 0 and variances 1, the likelihood equation reduces to the cubic

$$\rho(1-\rho^2) + (1+\rho^2)S_1 - \rho S_2 = 0, \tag{1}$$

where $S_1 = \sum xy/n$, $S_2 = \sum(x^2+y^2)/n$.

The roots of (1) can be written down explicitly using Cardano's formula, but this is not very helpful. To discuss the solution we replace ρ by $z = \rho - S_1/3$ in (1) to obtain the equation

$$z^3 + az + b = 0, \tag{2}$$

where $a = (S_2 - 1 - S_1^2/3)$ and $b = -S_1(4 - S_2 + 2S_1^2/3^2)/3$. Equation (2) has one real and two complex roots if $\Delta = a^3/27 + b^2/4 > 0$, a condition which is easy to check in practice. Note that this condition applies more generally than the condition

$$D = 4S_1^2 + 12(1 - S_2) \le 0$$

(i.e., $a \ge 0$) which is discussed by SWY in their Section 2.1 as a sufficient condition for a single real root.

Using the strong law of large numbers, we have almost surely that $a \to 1 - \rho^2/3$ and $b \to -(2\rho/3)(1+\rho^2/9)$ as $n \to \infty$. It is then evident that Δ can be expected to be positive in most cases where it realistic to use this methodology. There should be no need in practice to deal with the case $\Delta < 0$ in which there are three real roots of (2), but if there were, a not unreasonable choice for the estimator of ρ would be the root of (1) which is closest to S_1, since S_1 is a consistent estimator of ρ.

For $\Delta > 0$, the real root can be written down using Cardano's formula. It is

$$S_1/3 + (-b/2 + \Delta^{1/2})^{1/3} + (-b/2 - \Delta^{1/2})^{1/3},$$

and it is easily checked that this tends to ρ as $n \to \infty$.

Small, Wang and Yang quote, in their Section 4.2, the example where $S_1 = -0.1$ and $S_2 = 1$ as a problem case, but it is straightforward to treat. We have $\Delta > 0$, but note that $D > 0$, and the roots of the likelihood equation (1), which is now

$$\rho^3 + (1+\rho^2)/10 = 0,$$

are easily checked to be $-0.5, 0.2 \pm 0.4\sqrt{-1}$. The example is, however, not very realistic since the value of $S_2 = 1$ is well removed from its limit of 2, and there is a substantial disparity between the sample correlation $S_1 = -0.1$ and the estimated ρ of -0.5.

It should be remarked that the principle of investigating the asymptotics as sketched in Heyde (1997) and Heyde and Morton (1998), and referred to by the authors in their Section 4.2, is intended to advocate the use of the asymptotics in a flexible exploratory way, as in the discussion above, and not just for checking on explicit formulae for the roots, when these are available. Multiple root problems cover a very diverse spectrum and it seems to me that they should be approached quite pragmatically.

Comment

Bing Li

Multiple roots problems have wide practical implications and are theoretically subtle and challenging. The paper of Small, Wang and Yang (SWY) surveys the various situations in which multiple roots can arise, the issues they bring about and the methods to deal with them. Much progress has recently been made toward solving this problem and I find this paper timely and helpful. Rather than discuss generally, I will focus on two points: (1) how to generalize the Wald approach to find consistent solutions of estimating equations; (2) how to construct a likelihood ratio that is antisymmetric and additive.

What makes the problem challenging is that, for estimating equations, the estimators are defined as the solutions of equations rather than the maximizers of objective functions, or "log-likelihoods." So when there are multiple solutions we cannot identify a consistent one as we do with a likelihood, whose maximizer is consistent under conditions such as assumed in Wald (1949). However, if we view the maximizer of the likelihood as the minimax point of the log-likelihood ratio, then we can generalize the Wald approach to nonconservative estimating equations. I will call this general approach the minimax principle and describe it in Section 1.

As surveyed by SWY, much has been done to develop a likelihood theory for nonconservative estimating equations, either to distinguish between solutions or to test hypotheses. One type of construction, such as in McLeish and Small (1992), Li (1993) and Wang (1999), is to make the objective function tangent to the quasiscore locally at a hypothesized parameter value. However, objective functions thus defined must depend on two points in the parameter space, one being the hypothesized value, the other an alternative value. Consequently they are not additive (the "log-likelihood ratio" between A and C is not the sum of those between A and B and B and C) and may not even be transitive. In Section 2, I describe how to apply the minimax principle to construct a log-likelihood that only depends on one parameter value, which implies that the corresponding log-likelihood ratio is antisymmetric and additive.

The minimax principle in Section 1 and the minimax likelihood in Section 2 were both introduced in Li (1996), but were not spelt out very explicitly there. The asymptotic distribution in Section 2 is a new result. In Section 3 I list some properties of the minimax likelihood as evidence that it may be a natural and competent candidate as a likelihood for nonconservative estimating equations.

1. MINIMAX PRINCIPLE AS A GENERALIZED WALD APPROACH

The logic of Wald's approach is this: since the log-likelihood converges to the expected log-likelihood, the maximum point of the log-likelihood should converge to the maximum point of the expected log-likelihood, which, by Jensen's inequality, is the true parameter value. However, an estimating equation is in general not the derivative of any function, and hence nothing plays the role of the log-likelihood. However, what drives the Wald argument is that the uniform convergence of a sequence of functions implies the convergence of some features of those functions. This basic idea can be generalized.

We can redefine the maximum likelihood estimator as the minimax point, or saddle point, of the log-likelihood ratio, and the true parameter value as the minimax of the expected log-likelihood ratio. Specifically, let $X = (X_1, \ldots, X_n)^T$ be the observations, and let $\theta \in \Theta \subset R^p$ be the parameter. Let $L(\theta; X) \equiv L(\theta)$ be the log-likelihood. Let $\hat{\theta}$ be the maximum likelihood estimator. Then

$$\begin{aligned} 0 = \inf_{\theta\in\Theta}(0) &\le \inf_{\theta\in\Theta}\sup_{\eta\in\Theta}\{L(\eta) - L(\theta)\} \\ &\le \sup_{\eta\in\Theta}\{L(\eta) - L(\hat{\theta})\} \\ &= L(\hat{\theta}) - L(\hat{\theta}) = 0. \end{aligned}$$

Hence $(\theta, \eta) = (\hat{\theta}, \hat{\theta})$ is the minimax point of the function $L(\eta) - L(\theta)$. Now let $R(\theta) = EL(\theta)$, and let θ_0 be the true parameter value under which we take the expectation E. Then, by Jensen's inequality, $R(\theta_0) > R(\theta)$ for all $\theta \neq \theta_0$, and hence, by the same argument as above, (θ_0, θ_0) is the minimax point of the function $R(\eta) - R(\theta)$. Under conditions similar to those of the Wald theorem,

$$n^{-1}\{L(\eta) - L(\theta)\} - n^{-1}\{R(\eta) - R(\theta)\} \xrightarrow{p} 0,$$

Bing Li is Associate Professor, Department of Statistics, Pennsylvania State University, 410 Thomas Building, University Park, Pennsylvania 16802.

uniformly over $(\theta, \eta) \in \Theta \times \Theta$. From here it is easy to deduce that $(\hat{\theta}, \hat{\theta}) \to_p (\theta_0, \theta_0)$, or $\hat{\theta} \to_p \theta_0$.

The function $L(\eta) - L(\theta)$ can be generalized to quasilikelihood equations by a projection argument, and the generalization sufficiently resembles it to make the minimax argument work (Li, 1996). Let μ_θ and V_θ be the mean vector and the covariance matrix of X. Define

$$L(\theta, \eta) = (1/2)\{(\mu_\eta - \mu_\theta)^T V_\theta^{-1}(X - \mu_\theta) + (\mu_\eta - \mu_\theta)^T V_\eta^{-1}(X - \mu_\eta)\}$$

and $R(\theta, \eta) = EL(\theta, \eta)$. Note that both $L(\theta, \eta)$ and $R(\theta, \eta)$ are antisymmetric, even though they cannot be detached as the difference between a function of η and a function of θ, as is the log-likelihood ratio. Antisymmetry plays a key role in generalizing the minimax argument. Under regularity conditions not unlike those used by in Wald (1949), Li (1996) demonstrates that (1) the minimax of $L(\theta, \eta)$ is asymptotically equivalent to any consistent solution of the quasilikelihood equation (in fact, under some conditions it is a consistent solution to that equation), (2) the minimax point of $R(\theta, \eta)$ is the true parameter value and (3) the difference between $n^{-1}L(\theta, \eta)$ and $n^{-1}R(\theta, \eta)$ converges to 0 uniformly over $\Theta \times \Theta$. Hence $(\hat{\theta}, \hat{\theta}) \to_p (\theta_0, \theta_0)$, or $\hat{\theta} \to_p \theta_0$.

This result leaves no ambiguity when the quasilikelihood equation has multiple roots, because whichever root is the minimax of $L(\theta, \eta)$ is consistent. The result is simplified in Li (1997) as follows: if a root of the quasilikelihood equation is the minimax among the collection of roots, then it is consistent. Thus we only need to identify the minimax point among roots.

As SWY point out, when $L(\theta, \eta)$ is used as "log-likelihood ratio" for hypothesis testing, as was done in Li (1993), it has the drawback of being intransitive. However, the minimax argument itself suggests a way to avoid this problem, as we will see in the next section.

2. MINIMAX QUASILIKELIHOOD

According to the minimax principle we estimate θ_0 by minimizing $\sup_{\eta\in\Theta} L(\theta, \eta)$, or maximizing $-\sup_{\eta\in\Theta} L(\theta, \eta)$. Therefore it is reasonable to define

$$\ell(\theta) = -\sup_{\eta\in\Theta} L(\theta, \eta) \tag{1}$$

as the "log-likelihood." Since ℓ is a function of θ alone, the "log-likelihood ratio" defined as its difference is antisymmetric, additive and transitive. Also note that $\ell(\theta)$ is not a potential function of the quasiscore, and yet it is tangent to the quasiscore at the minimax point. I will now outline the derivation of the asymptotic distribution of the "log-likelihood ratio" derived from this definition.

Suppose that the p-dimensional parameter θ is composed of an r-dimensional parameter of interest ψ and an s-dimensional nuisance parameter λ and that we are to test H_0: $\psi = \psi_0$ against H_1: $\psi \neq \psi_0$. Let $\hat{\theta}$ be the maximizer of $\ell(\theta)$, and let $\tilde{\theta} = (\psi_0, \tilde{\lambda})$ be the maximizer of $\ell(\theta)$ subject to the constraint $\psi = \psi_0$. I will argue that, under H_0,

$$2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\} \xrightarrow{p} \chi_r^2.$$

To avoid writing too many subscripts I will drop the index 0 and denote the true parameter value θ_0 by $\theta = (\psi, \lambda)$. As usual, assume that the quasi-Fisher information $\dot{\mu}_\theta^T V_\theta^{-1} \dot{\mu}_\theta / n$ converges to a constant matrix $I(\theta)$ as $n \to \infty$. Denote the standardized quasiscore $n^{-1/2}\dot{\mu}_\theta^T V_\theta^{-1}(X - \mu_\theta)$ by $S(\theta, X)$. By Taylor expansion,

$$\ell(\hat{\theta}) = \ell(\theta) + \dot{\ell}(\theta)(\hat{\theta} - \theta) + (\hat{\theta} - \theta)^T \ddot{\ell}(\theta)(\hat{\theta} - \theta)/2 + O_p(n^{-1/2}),$$

where $\dot{\ell}(\theta)$ and $\ddot{\ell}(\theta)$ are the first and second partial derivatives of ℓ with respect to θ.

We now approximate $\dot{\ell}(\theta)$ and $\ddot{\ell}(\theta)$. Let η_θ be the maximizer of $-L(\theta, \eta)$ over $\eta \in \Theta$. Then, $\ell(\theta) = -L(\theta, \eta_\theta)$. Let the partial derivatives of $L(\theta, \eta)$ with respect to θ or η be denoted by L indexed by 1 or 2; for example, $L_1(\theta, \eta)$ stands for $\partial L(\theta, \eta)/\partial\theta$ and $L_{12}(\theta, \eta)$ stands for $\partial^2 L(\theta, \eta)/\partial\theta\,\partial\eta$. By the chain rule,

$$\begin{aligned}\dot{\ell}(\theta) &= -L_1(\theta, \eta_\theta) - L_2(\theta, \eta_\theta)\dot{\eta}_\theta = -L_1(\theta, \eta_\theta) \\ &\quad + L_2(\theta, \eta_\theta)\{L_{22}(\theta, \eta_\theta)\}^{-1} L_{21}(\eta, \eta_\theta),\end{aligned} \tag{2}$$

where, for the last equality, we have used the fact $\dot{\eta}_\theta = -\{L_{22}(\theta, \eta_\theta)\}^{-1} L_{21}\ (\theta, \eta_\theta)$, which is derived from the equation $\dot{\ell}(\eta_\theta) = 0$. By Li (1996); see the discussion below Theorem (3b), η_θ differs from θ by $O_p(1/\sqrt{n})$. Therefore the following approximations hold: mathtight

$$\begin{aligned}L_1(\theta, \eta_\theta) &= L_1(\theta, \theta) + L_{12}(\theta, \theta)(\eta_\theta - \theta) + O_p(1), \\ L_2(\theta, \eta_\theta) &= L_2(\theta, \theta) + L_{22}(\theta, \theta)(\eta_\theta - \theta) + O_p(1), \\ L_{21}(\theta, \eta_\theta) &= L_{21}(\theta, \theta) + O_p(\sqrt{n}), \\ L_{22}(\theta, \eta_\theta) &= L_{22}(\theta, \theta) + O_p(\sqrt{n}).\end{aligned} \tag{3}$$

By taking derivatives and expectations, it is easy to verify that

$$\begin{aligned}L_1(\theta, \theta) &= -L_2(\theta, \theta) = -\sqrt{n}S(\theta, X), \\ L_{12}(\theta, \theta) &= 0, \\ E\{L_{11}(\theta, \theta)\} &= -E\{L_{22}(\theta, \theta)\} = \dot{\mu}_\theta V_\theta^{-1} \dot{\mu}_\theta \\ &= nI(\theta) + o(n) = O(n).\end{aligned} \tag{4}$$

Hence, in (3), $L_2(\theta, \eta_\theta)$ and $L_{21}(\theta, \eta_\theta)$ are of the order $O_p(\sqrt{n})$, $L_{22}(\theta, \eta_\theta)$ is of the order $O_p(n)$ and $L_1(\theta, \eta_\theta) = -\sqrt{n}S(\theta, X) + O_p(1)$. Substituting these approximations into expansion (2), we find

$$\dot{\ell}(\theta) = \sqrt{n}S(\theta, X) + O_p(1). \tag{5}$$

The approximation of $\ddot{\ell}(\theta)$ involves $\ddot{\eta}_\theta$ and taking higher derivatives of $L(\theta, \eta)$, but follows the same spirit. The result is

$$\begin{aligned}\ddot{\ell}(\theta) &= -\dot{\mu}_\theta^T V_\theta^{-1}\dot{\mu}_\theta + O_p(\sqrt{n}) \\ &= nI(\theta) + o_p(n).\end{aligned} \tag{6}$$

Now substituting the approximations of $\dot{\ell}(\theta)$ and $\ddot{\ell}(\theta)$ into the expansion of $\ell(\hat{\theta})$, we have

$$\begin{aligned}\ell(\hat{\theta}) = \ell(\theta) &+ S(\theta, X)^T\sqrt{n}(\hat{\theta} - \theta) \\ &- \sqrt{n}(\hat{\theta} - \theta)^T I(\theta)\sqrt{n}(\hat{\theta} - \theta)/2 + o_p(1).\end{aligned}$$

By Theorem 2 of Li (1996), $\hat{\theta}$ is asymptotically equivalent (indeed identical in most cases) to a consistent solution of the quasilikelihood equation. Hence $S(\theta, X) = I\sqrt{n}(\hat{\theta} - \theta) + o_p(1)$. Consequently the above expansion reduces to

$$\begin{aligned}&2\{\ell(\hat{\theta}) - \ell(\theta)\} \\ &\quad = \sqrt{n}(\hat{\theta} - \theta)^T I(\theta)\sqrt{n}(\hat{\theta} - \theta) + o_p(1).\end{aligned} \tag{7}$$

Next, we deal with the restricted maximum $\ell(\tilde{\theta})$. Denote the derivatives of $\ell(\cdot)$ with respect to λ by $\dot{\ell}_\lambda(\theta)$ and $\ddot{\ell}_\lambda(\theta)$, the λ-component of $S(\theta, X)$ by $S_\lambda(\theta, X)$, and the four blocks of I by $I_{\psi\psi}$, $I_{\psi\lambda}$, $I_{\lambda\psi}$, $I_{\lambda\lambda}$. Expand $\ell(\psi, \tilde{\lambda})$ about θ:

$$\begin{aligned}\ell(\psi, \tilde{\lambda}) &= \ell(\theta) + \dot{\ell}_\lambda(\theta)(\tilde{\lambda} - \lambda) \\ &\quad + (\tilde{\lambda} - \lambda)^T\ddot{\ell}_\lambda(\theta)(\tilde{\lambda} - \lambda)/2 + O_p(n^{-1/2}) \\ &= \ell(\theta) + S_\lambda(\theta, X)^T\sqrt{n}(\tilde{\lambda} - \lambda) \\ &\quad - \sqrt{n}(\tilde{\lambda} - \lambda)^T I_{\lambda\lambda}(\theta)\sqrt{n}(\tilde{\lambda} - \lambda)/2 + o_p(1),\end{aligned} \tag{8}$$

where the second equality follows from the approximations (5) and (6). Since $\tilde{\lambda}$ maximizes $\ell(\psi, \cdot)$ over λ, it satisfies the equation

$$L_1(\tilde{\theta}, \eta_{\tilde{\theta}})(\partial\tilde{\theta}/\partial\tilde{\lambda}) + L_2(\tilde{\theta}, \eta_{\tilde{\theta}})(\partial\eta_{\tilde{\theta}}/\partial\tilde{\lambda}) = 0.$$

By the argument above expression (5) it can be shown that $\partial\eta_{\tilde{\theta}}/\partial\tilde{\lambda} = O_p(1/\sqrt{n})$, $L_2(\tilde{\theta}, \eta_{\tilde{\theta}}) = O_p(\sqrt{n})$ and $L_1(\tilde{\theta}, \eta_{\tilde{\theta}}) = L_1(\tilde{\theta}, \tilde{\theta}) + O_p(1)$. Hence the above equation is asymptotically equivalent to

$$L_1(\tilde{\theta}, \tilde{\theta})(\partial\tilde{\theta}/\partial\tilde{\lambda}) = -S_\lambda(\tilde{\theta}, X) = 0 \tag{9a}$$

and, consequently,

$$S_\lambda(\theta, X) = I_{\lambda\lambda}\sqrt{n}(\tilde{\lambda} - \lambda) + o_p(1). \tag{9b}$$

Combining (8) and (9b), we see that

$$\begin{aligned}2\{\ell(\tilde{\theta}) - \ell(\theta)\} &= \sqrt{n}(\tilde{\lambda} - \lambda)^T I_{\lambda\lambda}(\theta)\sqrt{n}(\tilde{\lambda} - \lambda) \\ &\quad + o_p(1).\end{aligned}$$

We now express $\sqrt{n}(\tilde{\lambda} - \lambda)$ as an approximate linear combination of $\sqrt{n}(\hat{\theta} - \theta)$. Recall that $S(\theta, X)$ is asymptotically equivalent to $I\sqrt{n}(\hat{\theta} - \theta)$. Therefore $S_\lambda(\theta, X)$ is asymptotically equivalent to $(I_{\lambda\psi}, I_{\lambda\lambda})\sqrt{n}(\hat{\theta} - \theta)$. Hence, by (9b),

$$\begin{aligned}\sqrt{n}(\tilde{\lambda} - \lambda) &= I_{\lambda\lambda}^{-1}S_\lambda(\theta, X) + o_p(1) \\ &= (I_{\lambda\lambda}^{-1}I_{\lambda\psi}, U)\sqrt{n}(\hat{\theta} - \theta) + o_p(1),\end{aligned}$$

where U denotes the $s \times s$ unit matrix. Therefore,

$$\begin{aligned}&2\{\ell(\tilde{\theta}) - \ell(\theta)\} \\ &\quad = \sqrt{n}(\hat{\theta} - \theta)^T(I_{\lambda\lambda}^{-1}I_{\lambda\psi}, U)^T \\ &\qquad \cdot I_{\lambda\lambda}(I_{\lambda\lambda}^{-1}I_{\lambda\psi}, U)\sqrt{n}(\hat{\theta} - \theta) + o_p(1).\end{aligned} \tag{10}$$

Subtracting (10) from (7) gives

$$\begin{aligned}2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\} &= \sqrt{n}(\hat{\psi} - \psi)^T(I_{\psi\psi} - I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi}) \\ &\quad \cdot \sqrt{n}(\hat{\psi} - \psi) + o_p(1).\end{aligned}$$

However, we know that $\sqrt{n}(\hat{\theta} - \theta) \to_{\mathcal{L}} N(0, I^{-1})$, and so $\sqrt{n}(\hat{\psi} - \psi)$ is asymptotically normal with mean 0 and variance $(I_{\psi\psi} - I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi})^{-1}$. Therefore the leading term on the right-hand side converges in distribution to χ_r^2.

3. CONCLUSION

I now list the properties of $\ell(\theta)$ that resemble a log-likelihood:

(i) *Wald consistency*—The maximizer $\hat{\theta}$ of $\ell(\theta)$ is a consistent estimator of θ_0 (Li, 1996).

(ii) *Invariance*—Let h: $\Theta \mapsto \Phi$ be a one-to-one transformation that maps $\theta \in \Theta$ to $\phi \in \Phi$, let $\tilde{\ell}$ be the likelihood for ϕ (i.e., $\tilde{\ell} = \ell \circ h^{-1}$) and let $\hat{\phi}$ be the maximizer of $\tilde{\ell}$. Then $\hat{\phi} = h(\hat{\theta})$.

(iii) *Efficiency*—Since $\hat{\theta}$ is asymptotically equivalent to a consistent solution to the quasilikelihood equation, $\hat{\theta}$ is efficient in the same sense that the quasilikelihood estimator is efficient. See, for example, Heyde (1997, Chapter 2) and Small and McLeish (1994, Chapter 4).

(iv) *Relation to quasiscore*—Under mild conditions [Li, 1996, condition (17)], it can be shown that

$$\dot{\ell}(\hat{\theta}) = \sqrt{n}S(\hat{\theta}, X) = 0$$

and

$$\ddot{\ell}(\hat{\theta}) = \sqrt{n}\,\partial S(\hat{\theta}, X)/\partial\theta.$$

(v) *Covariant tensor*—Again, suppose that condition (17) of Li (1996) holds. Then, under the transformation defined in (ii), the observed information $\ddot{\ell}(\hat{\theta})$ is a covariant tensor (McCullagh, 1987, page 6). That is,

$$\frac{\partial^2 \tilde{\ell}(\hat{\phi})}{\partial \phi^2} = \left\{ \frac{\partial h^{-1}(\hat{\phi})}{\partial \phi} \right\}^T \left\{ \frac{\partial^2 \ell(\hat{\theta})}{\partial \theta^2} \right\} \left\{ \frac{\partial h^{-1}(\hat{\phi})}{\partial \phi} \right\}.$$

(vi) χ^2*-distribution*—As argued in the last section, the likelihood ratio derived from ℓ has a χ^2 asymptotic distribution for testing composite hypotheses.

This list provides some evidence that $\ell(\theta)$ may be a natural and competent candidate for the definition of quasilikelihood when the quasiscore is not conservative. At least for establishing consistency, it gives as strong a result as does the Wald theorem under assumptions no more stringent than the latter. This seems to be unique among other candidates for quasilikelihood. Although I have restricted my discussion to the classical type of quasiscores, the minimax principle should be more generally applicable. For example, Li (1997) used the same principle to prove the consistency of Generalized Estimating Equations. We have also seen that ℓ may be used for testing composite hypothesis. In this respect, other methods such as the projected likelihood of McLeish and Small (1992), the path-dependent integral of quasiscore of Hanfelt and Liang (1995, 1997), the dual likelihood of Mykland (1995), the artificial likelihoods of Wang (1999), as well as the potential function of Li and McCullagh (1994), can also serve the purpose to varying degrees. In particular, the projected likelihood and the dual likelihood satisfies the Bartlett identities of all orders, whereas the information identity is satisfied for ℓ only approximately. This may give the former an advantage in χ^2-approximation (see Mykland, 1995). Nevertheless, ℓ disentangles, θ and η and thereby gives rise to an antisymmetric and additive log likelihood ratio.

Rejoinder

Christopher G. Small, Jinfang Wang and Zejiang Yang

We are grateful to Professors Hanfelt, Heyde and Li for their thoughtful comments on our paper.

WHY MULTIPLE ROOT PROBLEMS NOW?

It is difficult to overestimate the influence of modern computing technology upon statistical data analysis. The easy availability of computational resources permits the data analysts to study complex inferential issues, such as those arising in mixture modeling, as discussed by Professor Hanfelt. Nowadays, Fisher's likelihood methodology, and its generalization to the methods of optimal estimating functions, can be implemented for a large variety of models by accessing this computational technology. However, multiple roots can arise in the use of such methods for which estimates cannot be written in closed form. The roots of estimating functions should not be expected to be local extrema of *any* scalar objective function when the estimating function is nonconservative. To appreciate the relevance of the issue of nonconservativeness it suffices to note that, in the space of regular smooth estimating functions, conservative ones are "atypical" (Poston and Stewart, 1976, page 33). Therefore methods for root selection in the context of estimating functions are important and more difficult in nature than the cases based on parametric models. The construction of objective functions (Section 6) represents a major effort toward solving this problem. Section 4.4 provides an application. This view is also emphasized by Professors Hanfelt and Li.

AD HOC SOLUTIONS VERSUS OBJECTIVE METHODS

Intuitions often suggest reasonable answers in specific decision-making. In this aspect we agree with Professor Heyde, who points out that in a real situation a data analyst often can overcome the difficulty of multiple roots by problem-specific methods. It is the scientific justification and the general guidance that are required to be developed. Professor Heyde suggests comparing the roots of an estimating function with a consistent estimator such as the median when he discusses the Cauchy location model. However, in choosing the root that is closest to the median in some metric, we may lose the parametrization equivariance of the final estimator. One-step iteration from Fisher's scoring of course provides an alternative answer (Section 4.1), But now it leaves the essential problem of choosing an

appropriate consistent estimator unanswered. See Barendregt and van Pul (1995) for an interesting discussion of this problem. We agree with Professor Heyde that simple explorative methods such as the examination of the asymptotics are important when explicit formulae can be obtained. Unfortunately this only occurs in relatively simple situations. Even then care is still necessary (Section 4.2).

In the rest of our reply we shall focus on specific issues raised by the discussants.

Professor Hanfelt opens his discussion by examining why the subject has been overlooked or dismissed in the literature. We agree that multiple roots may well be overlooked in many data analyses. It is also true that the theory for treating multiple roots is undeveloped. So it may be that some researchers regard the problem as a can of worms that is better left unopened. Others, such as Professor Heyde, regard the problem as less serious than we have made it out to be. We shall examine his comments in greater detail below.

Professors Hanfelt and Li both emphasize the construction of artificial likelihoods to select among multiple roots. Both authors make some excellent arguments in favor of their methodologies. Since the subject suffers from too many ad hoc solutions, a great advantage of the artificial likelihood approach is that it avoids this problem in the same way that Fisher's likelihood methodology provided a unifying treatment of point estimation. However, artificial likelihoods are not without difficulties. For example, much of the justification for artificial likelihoods is asymptotic in nature. While multiple roots can exist asymptotically, a multiple root *selection* problem is only to be found for a fixed sample size. Asymptotically, all reasonable methods should select the same root. However, for a fixed sample size, the various methodologies for root selection can disagree with each other. Nevertheless, artificial likelihoods offer the most general tools we have for root selection in semiparametric cases. So if the problems can be overcome, the methodology should be quite powerful.

With this in mind, Professor Li's excellent analysis of intransitivity and antisymmetry within his minimax approach deserves careful consideration. By taking a supremum over η in the antisymmetric function $L(\theta, \eta)$, he reduces to the artificial likelihood $\ell(\theta)$ which provides a transitive partial ordering of the parameter space, much as an ordinary likelihood does. Building artificial likelihoods helps to bridge the gap between the multiple root problems of semiparametric analysis and those of parametric analysis. However, such artificial likelihoods cannot be a universal panacea for the problems of multiple roots any more than the parametric likelihood is.

Professor Hanfelt's confusion about our claims immediately after equation (29) is most likely due to our clumsy wording of the text. Of course, we are not saying that the $\sqrt{n}$-consistency of $\hat{\theta}_i$ implies that $g(\theta)$ is information-unbiased. Rather, our conclusions follow in cases where both conditions hold. The goal of the bootstrap methodology is similar to other goodness-of-fit methods. We construct a statistic which is asymptotically distribution free at a $\sqrt{n}$-consistent root. By examining the bootstrap distribution of the statistic, we hope to determine which root is the most reasonable. Once again, it is unfortunate that the justification is asymptotic in nature. However, this would seem to be unavoidable.

Professor Hanfelt also raised some concerns about the artificial likelihood discussed in Section 6.4. Linearity of $g(\theta)$ in θ is not required for the method to work, as may be appreciated by the example discussed in Section 4.4. The approximate (log) likelihood (44) has been derived from an optimal vector-field decomposition based on a locally linearized version of the estimating function g at a reference point. A nonlinear nonconservative estimating function is locally equivalent to a gradient system at a critical point if the quasi-Hessian matrix has unequal real eigenvalues. To further appreciate why the quadratic form is essential one may recall that a scalar function at a nondegenerate point can be put into a quadratic form by changing coordinates, a fact known as the Morse lemma (Poston and Stewart, 1978, page 54).

Unlike Professors Hanfelt and Li, Professor Heyde takes us to task for making more of multiple root problems than we should. Thus he falls into the category of statisticians described by Professor Hanfelt, who doubt that multiple roots pose a serious problem. The two examples presented by Professor Heyde illustrate that there are simple solutions to the problems of multiple roots in two of the cases we have considered. Our response to his objection is contained in the point stated above, namely, that most solutions to multiple roots are ad hoc. Naturally, we have no objections to the median as a location estimator for the Cauchy, or the root of the likelihood closest to the sample correlation coefficient for the standardized bivariate normal. However, these solutions do not extend to general methods that help us choose a root for other models. Our paper is an appeal for a general theory to treat such problems that is applicable to diverse models.

We must also respectfully disagree with Professor Heyde's statement that we unnecessarily complicate the issues of multiple roots. Our claim early in the paper that likelihood equations can formally be said to have infinitely many roots is not just a matter of mathematical precision. It shows a practical limitation on what can be accomplished by asymptotic arguments. In particular, it is possible for different statisticians to choose different roots of an equation while simultaneously justifying their choices as being the "unique consistent root." This problem lies at the heart of the problem of root selection and cannot be ignored.

We close our comments by reiterating the closing point made by Professor Heyde. Multiple root problems are indeed diverse and must be approached pragmatically. We hope that our paper will stimulate a discussion about the problems of multiple roots without losing sight of this fact.

ADDITIONAL REFERENCES

BAGGERLY, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85** 535–547.

BARENDREGT, L. G. and VAN PUL, M. C. (1995). On the estimation of the parameters for the Littlewood model in software reliability. *Statist. Neerlandica* **49** 165–184.

EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

LI, B. (1996). A minimax approach to consistency and efficiency for estimating equations. *Ann. Statist.* **24** 1283–1297.

LI, B. (1997). On the consistency of generalized estimating equations. In *Selected Proceedings of the Symposium on Estimating Functions* (I. V. Basawa, V. P. Godambe and R. L. Taylor, eds.) 115–136.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–94.

MCCULLAGH, P. (1996). *Tensor Method in Statistics*. Chapman and Hall, New York.

MYKLAND, P. A. (1995). Dual likelihood. *Ann. Statist.* **23** 386–421.

NESTADT, G., HANFELT, J., LIANG, K.-Y., LAMACZ, M., WOLYNIEC, P. and PULVER, A. E. (1994). An evaluation of the structure of schizophrenia spectrum personality disorders. *J. Personality Disorders* **8** 288–298.

POSTON, T. and STEWART, I. N. (1976). *Taylor Expansions and Catastrophe*. Pitman, London.

POSTON, T. and STEWART, I. N. (1978). *Catastrophe Theory and Its Applications*. Dover, New York.

QIAN, G., GABOR, G. and GUPTA, R. P. (1996). Generalised linear model selection by the predictive least quasi-deviance criterion. *Biometrika* **83** 41–54.