System Understanding and Statistical Uncertainty Bounds from Limited Test Data

James C. Spall

In many DoD test and evaluation programs, it is necessary to obtain statistical estimates for parameters in the system under study. For these estimates to provide meaningful system understanding, uncertainty bounds (e.g., statistical confidence intervals) must be attached to the estimates. Current methods for constructing uncertainty bounds are almost all based on theory that assumes a large amount of test data. Such methods are not justified in many realistic testing environments where only a limited amount of data is available. This article presents a new method for constructing uncertainty bounds for a broad class of statistical estimation procedures when faced with only a limited amount of data. The approach is illustrated on a problem motivated by a Navy program related to missile accuracy, where each test is very expensive. This example will illustrate how the small-sample approach is able to obtain more information from the limited sample than traditional approaches such as asymptotic approximations and the bootstrap.

(Keywords: Confidence regions, Parameter estimation, Small-sample analysis.)

INTRODUCTION

A pervasive problem in defense test and evaluation (and other areas) is the need to make meaningful inference from a limited amount of data. This issue is especially critical as defense testing budgets are reduced, resulting in a need to extract as much information as possible from a limited sample. Such inference usually involves a statistical estimation process and an uncertainty calculation (e.g., confidence region). For many estimators used in system testing (e.g., least squares, maximum likelihood, and maximum *a posteriori*), there exists a large-sample theory that provides the basis for determining probabilities and confidence regions in large samples (see, e.g., Refs. 1 and 2). However, except for relatively simple cases, it is generally not possible to determine this uncertainty information in the small-sample setting. This article presents an approach to determining small-sample probabilities and confidence regions for a general class of multivariate M-estimators (M-estimators are those found as the solution to a system of equations, and include those common estimators just mentioned). Theory and implementation aspects will be presented. Three distinct examples will be presented to illustrate the broad potential applications of the approach: (1) a "signal-plus-noise" estimation problem that arises in applications such as weapon system accuracy analysis, small-area estimation from surveys, Kalman filter (state-space) model identification, and estimate combining; (2) a nonlinear regression setting; and (3) a problem in correlation analysis for time series.

The approach is based on a simple—but apparently unexamined—idea. Suppose that the model used in the identification problem is some ϵ distance (to be defined later) away from an idealized model, where the small-

sample distribution of the M-estimate for the idealized model is known. Then the known probabilities and confidence regions for the idealized model provide the basis for computing the probabilities and confidence regions in the actual model. The ϵ distance may be reflected in a conservative adjustment to the idealized quantities. This approach is fundamentally different from other finite-sample approaches, where the accuracy of the relevant approximations is tied to the size of the sample (versus the deviation from an idealized model).

The M-estimation framework is very general as it encompasses many (perhaps most) estimators of practical interest in system testing and allows us to develop concrete regularity conditions that are largely in terms of the score function (the score is typically the gradient of the objective function, which is being set to zero to create the system of equations that yields the estimate). One of the significant challenges in assessing the small-sample behavior of M-estimates is that they are usually nonlinear, implicitly defined functions of the data (determined by some numerical iteration procedure).

The problem of probability and confidence region calculation in small samples has, of course, been widely considered. Let us briefly highlight some of the most popular approaches that might apply to M-estimates. Perhaps the most popular current approach is resampling, most notably the bootstrap (e.g., Refs. 3–6). The main appeal of this approach is relative ease of use, even for complex estimation problems. Resampling techniques make few analytical demands on the user, instead shifting the burden to one of computation. However, the bootstrap can provide a highly inaccurate description of M-estimate uncertainties in small samples, as illustrated in the example discussed later (which considers the parametric bootstrap technique described in Efron and Tibshirani[3]). This poor performance is inherently linked to the limited amount of information about the small sample, with little improvement possible through a larger amount of resampling.

Other relatively popular methods for small-sample probability and confidence region calculation are those based on series expansions, particularly the Edgeworth and saddlepoint (e.g., Refs 7–10). However, as noted in Reid,[8] "saddlepoint approximations have not yet had much impact on statistical practice." The major limiting factor of these series-based methods is the cumbersome analytical form and numerical calculations involved. Namely, there is a requirement to compute a certain probabilistic generating function and associated inverse transformations (typically via numerical integration) to obtain the small-sample density and then (typically) perform additional numerical integrations to obtain the boundaries for a confidence region. Hence, most of the literature in this area focuses on the relatively tractable case of estimates that are a smooth function of a sample mean (e.g., Refs. 8, 11, 12, and 13), although some authors have considered more general cases such as M-estimates or quantities related to certain stochastic processes (e.g., Refs. 9, 14, and 15). The cumbersome calculations for these methods become particularly prominent in multivariate M-estimation problems, although some partial results for this setting are in Tingley and Field[14] and Field and Ronchetti[9] ("partial" in the sense that these results consider a scalar function of a multivariate M-estimate; they use the 1980 Lugannani–Rice formula[16] together with the bootstrap to calculate the resulting scalar tail probabilities and associated confidence interval endpoints). An alternate approach to the multivariate M-estimation setting, which is based on estimating various saddlepoint quantities, is given in Ronchetti and Welsh.[17] This approach seems most appropriate in sample sizes that are at least moderately large, where these estimated quantities would be reliable.

The essential relationship of the small-sample approach here to the analytical (saddlepoint) methods is as follows. The saddlepoint methods are very general in principle and may provide accurate probability and confidence region approximations. However, they make strong analytical and computational demands on the user and appear infeasible in most of the multivariate M-estimation problems encountered in practice (where the estimate is usually implicitly defined and must be found numerically). The approach here, on the other hand, is generally easy to use for multivariate M-estimates and also can provide accurate results. However, it requires that an idealized setting be identified from which to make adjustments, which may not be available in some problems. In the signal-plus-noise example considered here, where the data are nonidentically distributed, the idealized case is one where the data would be i.i.d (independent, identically distributed). A resulting fundamental distinction between the saddlepoint method and the method here is in the nature of the errors in the probability calculations. For the saddlepoint, these errors are in terms of the sample size n and are typically of order $1/n$; for the approach here, the error is in terms of the deviation from the idealized case. In particular, if $\epsilon > 0$ is some measure of the deviation (to be defined), then the error is of order ϵ for any n for which the estimate is defined. In addition, the implied constant of the order ϵ term can be explicitly bounded if needed.

PROBLEM FORMULATION

Suppose we have a vector of data X (representing a test sample of size n) whose distribution depends on a p-dimensional vector θ and a scalar ϵ, where θ is to be estimated by maximizing some objective (say, as in maximum likelihood) and ϵ represents a known parameter. The estimate $\hat{\theta}$ is the quantity for which we wish to

characterize the uncertainty when n is small. It is assumed to be found as the objective-maximizing solution to the score equation:

$$s(\theta; x, \epsilon) = 0, \tag{1}$$

where $s(\cdot)$ represents the gradient of a log-likelihood function with respect to θ when $\hat{\theta}$ represents a maximum likelihood estimate. Suppose further that if $\epsilon = 0$ (the idealized case), then the distribution for the estimate is known for the small n of interest. Our goal is to establish conditions under which probabilities for $\hat{\theta}$ with $\epsilon > 0$ (the real case) are close to the known probabilities in the idealized case. In particular, we show that the difference between the unknown and known probabilities for the estimates is proportional to ϵ when ϵ is small. This justifies using the known distribution for $\hat{\theta}$ when $\epsilon = 0$ to construct approximate confidence regions for $\hat{\theta}$ when ϵ is small. Further, when ϵ is not so small, we show how the difference in the real and idealized probabilities can be approximated or bounded.

To characterize probabilities for the estimate $\hat{\theta}$, we introduce two artificial estimators that have the same distribution as $\hat{\theta}$ when $\epsilon > 0$ and when $\epsilon = 0$, respectively. The two artificial estimators, say $\hat{\theta}_\epsilon$ and $\hat{\theta}_0$, are based, respectively, on fictitious vectors of data, y_ϵ and y_0, of the same dimension as x. To construct the two fictitious data vectors, we suppose there exists a random vector z (same dimension as x), with associated transformations T_ϵ and T_0 (T_0 is the same as T_ϵ at $\epsilon = 0$) such that

$$y_\epsilon = T_\epsilon(z, \theta), \tag{2}$$

$$y_0 = T_0(z, \theta), \tag{3}$$

and such that y_ϵ and y_0 have the same probability distribution as x for the chosen $\epsilon > 0$ and for $\epsilon = 0$, respectively. Then, from Eq. 1,

$$\hat{\theta}_\epsilon : s(\hat{\theta}_\epsilon ; y_\epsilon, \epsilon) = 0, \tag{4}$$

$$\hat{\theta}_0 : s(\hat{\theta}_0 ; y_0, 0) = 0. \tag{5}$$

The fundamental point in the preceding machinations is that the distributions of $\hat{\theta}_\epsilon$ and $\hat{\theta}_0$ are identical to the distributions of the estimate $\hat{\theta}$ under $\epsilon > 0$ and $\epsilon = 0$, even though the various quantities (z, y_ϵ, etc.) have nothing per se to do with the real data and associated estimate. Our goal in the "Main Results" section is to establish conditions under which probabilities for $\hat{\theta}_\epsilon$ are close to the known probabilities for $\hat{\theta}_0$, irrespective of the sample size n. This provides a basis for approximating (or bounding) the probabilities and confidence regions for $\hat{\theta}$ under $\epsilon > 0$ through knowledge of the corresponding quantities for $\hat{\theta}_0$. Throughout the remainder of this article, we use the order notation $O(\epsilon)$ and $o(\epsilon)$ to denote terms such that $O(\epsilon)/\epsilon$ and $o(\epsilon)/\epsilon$ are bounded and approach 0, respectively, as $\epsilon \to 0$.

THREE EXAMPLE PROBLEM SETTINGS

To illustrate the range of problems for which the small-sample approach is useful, this section sketches how the approach would be applied in three distinct M-estimation settings. Further detailed analysis (including numerical results) for the first of these settings is provided in the section entitled "Application in Signal-Plus-Noise and Related CEP Problem."

Example 1: Parameter Estimation in Signal-Plus-Noise Setting with Non-i.i.d. Data

Consider the problem of estimating the mean and covariance matrix of a random signal when the measurements of the signal include added independent noise with known distributional characteristics. In particular, suppose we have observations $\{x_1, x_2, \ldots, x_n\}$ distributed $N(\mu, \Sigma + Q_i)$, where the noise covariances Q_i are known and the signal parameters μ, Σ (for which the unique elements are represented in vector format by θ) are to be jointly determined using maximum likelihood estimation (MLE). From the form of the score vector, we find that there is generally no closed-form solution (and no known finite-sample distribution) for the MLE when $Q_i \neq Q_j$ for at least one $i \neq j$. This corresponds to the actual ($\epsilon > 0$) case of interest. We also found that the saddlepoint method was analytically intractable for this problem (because of the relative complexity of the score vector) and that the bootstrap method worked poorly in sample sizes of practical interest (e.g., $n = 5$).

Estimation problems of this type (with either scalar or multivariate data) have been considered in many different problem contexts, for example, Rao et al.[18] in the estimation of a random effects model; James and Venables[19] and the National Research Council[20] in a problem of combining independent estimates of coefficients; Shumway et al.[21] and Sun[22] in a Kalman filter (state-space) model identification problem; Ghosh and Rao[23] in small-area estimation in survey sampling; and Hui and Berger[24] in the empirical Bayesian estimation of a dose–response curve. One of the author's interests in this type of problem lies in estimating projectile impact means and covariance matrices from noisy observations of varying quality; these are then used in

calculating CEP values (the 50% circular quantile values) for measuring projectile accuracy, as in Spall and Maryak.[25] Finally, for general multivariate versions of this problem, Smith[26] presents an approach for ensuring that the MLE of the covariance matrix is positive semi-definite, and Spall and Chin[27] present an approach for data influence and sensitivity analysis.

Central to implementing the small-sample approach is the identification of the idealized ($\epsilon = 0$) case and definition of ϵ relative to the problem structure. We can write $Q_i^{-1} = Q^{-1} + \epsilon D_i$, where Q and D_i are known matrices. (We are using the inverse form here to relate the various matrices since the more natural parameterization in the score vector is in terms of $\{Q_i^{-1}\}$, not $\{Q_i\}$. However, this is not required as the basic ideas would also apply in working with the noninverse form.) If $\epsilon = 0$ (the idealized identical Q_i case), the distribution of the μ, Σ estimate is normal-Wishart for all sample sizes of at least two. For this application, the Theorem in the "Main Results" section provides the basis for determining whether confidence regions from this idealized distribution are acceptable approximations to the unknown confidence regions resulting from nonidentical Q_i. In employing the Theorem (via Eqs. 4 and 5), we let $y_{\epsilon,i} = (\sigma^2 + Q_i)^{1/2} z_i + \mu = [\sigma^2 + (Q^{-1} + \epsilon D_i)^{-1}]^{1/2} z_i + \mu$, where z_i is distributed according to an $N(0, I)$ distribution, where I represents the identity matrix, and $i = 1, 2, \ldots, n$. In cases with a larger degree of difference in the Q_i's (as expressed through a larger ϵ)—where this idealized approximation for the confidence regions may not be adequate—implied constants associated with the $O(\epsilon)$ bound of the Theorem provide a means of altering the idealized confidence regions (these implied constants depend on terms other than ϵ: Q, $\{D_i\}$, etc.).

This example illustrates the arbitrariness sometimes present in specifying a numerical value of ϵ (e.g., if the elements of D_i are made larger, then the value of ϵ must be made proportionally smaller to preserve algebraic equivalence). This apparent arbitrariness has no effect on the fundamental limiting process as it is only the relative values of ϵ that have meaning after the other parameters (e.g., Q, D_i, etc.) have been specified. In particular, the numerical value of the $O(\epsilon)$ bound does not depend on the arbitrary way in which the deviation from the idealized case is allocated to ϵ and to the other parameters; in this example, $O(\epsilon)$ depends on the products $\{\epsilon D_i\}$, which are certainly not arbitrary.

Example 2: Nonlinear Regression

Although the standard linear regression framework is appropriate for modeling input–output relationships in some problems, a great number of practical problems have inherent nonlinearities. In particular, suppose that data are modeled as coming from the relationship

$$x_i = f_i(\theta, \epsilon, \eta_i) , \tag{6}$$

where $f_i(\cdot)$ is nonlinear mapping and η_i is a normally distributed random noise term. Typically, least-squares, Bayesian, or MLE techniques are used to find an estimate of θ. In contrast to the linear setting (with normally distributed noise terms), the finite-sample distribution of $\hat{\theta}$ will rarely be known in the nonlinear setting. In particular, although the problem of estimating parameters in nonlinear regression models is frequently solvable using numerical optimization methods, the "situation is much worse when considering the accuracy of the obtained estimates."[28] The small-sample approach here is appropriate when the degree of nonlinearity is moderate; the corresponding idealized case is a linear regression setting that, in a sense illustrated later, is close to the actual nonlinear setting. Relative to Eqs. 4 and 5, it is natural to choose $y_\epsilon = [f_1(\theta, \epsilon, z_1)^T, f_2(\theta, \epsilon, z_2)^T, \ldots, f_n(\theta, \epsilon, z_n)^T]^T$, where $z = (z_1^T, z_2^T, \ldots, z_n^T)^T$ has the joint normal distribution of $(\eta_1^T, \eta_2^T, \ldots, \eta_n^T)^T$.

Let us illustrate the ideas for two specific nonlinear cases. First, suppose that $f_i(\cdot)$ is a quadratic function $A_i + B_i\theta + \epsilon\theta^T C_i \theta + \eta_i$, where A_i, B_i, and C_i are vectors or matrices (as appropriate) of known constants. Such a setting might arise in an inversion problem of attempting to recover an unknown input value from observed outputs (as is, e.g., the main theme of fields such as statistical pattern recognition, image analysis, and signal processing). Clearly, for $\epsilon = 0$, we have the standard linear regression model. As with Example 1, the apparent arbitrariness in specifying ϵ is accommodated since the product ϵC_i is the inherent expression of nonlinearity appearing in the $O(\epsilon)$ bound. In the second case, suppose that $f_i(\cdot)$ represents the constant elasticity of substitution (CES) production function relating labor and capital inputs to production output within a sector of the economy (Kmenta[29] or Nicholson[30]). This model includes a "substitution parameter," which we represent by ϵ. After making a standard log transformation, the CES model has the form $f_i(\theta, \eta_i) = \theta_1 - (\theta_2/\epsilon)\log[\theta_3 CAP_i^{-\epsilon} + (1 - \theta_3) LAB_i^{-\epsilon}] + \eta_i$, where the three parameters within the θ vector represent parameters of economic interest, and CAP_i and LAB_i represent capital and labor input from firm i. As discussed in Kmenta[29] and Nicholson,[30] when $\epsilon = 0$ the CES function reduces to the well-known (log-linear) Cobb–Douglas production function, representing the idealized case here. Hence, confidence regions for the θ estimate in the CES model can be derived from the standard linear regression–based confidence regions for the Cobb–Douglas function through use of the Theorem.

Example 3: Estimates of Serial Correlation for Time Series

A basic problem in time series analysis is to determine whether a sequence of measurements is correlated over time, and, if so, to determine the maximum order of correlation (i.e., the maximum number of time steps apart for which the elements in the sequence are correlated). A standard approach for testing this hypothesis is to construct estimates of correlation coefficients for varying order correlations, and then to use the known distribution of the estimates to test against the hypothesis of zero correlation. Let us suppose that we construct MLEs of the jth-order correlation coefficients, $j = 1, 2, \ldots$. Our interest here is in the case where the data are non-normally distributed. This contrasts, for example, with the small-sample approach in Cook,[31] which is oriented to normal (and autoregressive) models. (By the result on pp. 220–221 of Bickel and Doksum,[32] we know that standard correlation estimate forms in, say, Section 6.1 of Anderson,[33] correspond to the MLE when the data are normally distributed.)

There are many ways, of course, in which one can model the non-normality in practical test and evaluation settings, but let us consider the fairly simple way of supposing the data are distributed according to a nonlinear transformation of a normal random vector. (Two other ways that may also be appropriate are (1) suppose that the data are distributed according to a mixture distribution where at least one of the distributions in the mixture is normal and where the weighting on the other distributions is expressed in terms of ϵ, or (2) suppose that the data are composed of a convolution of two random vectors, one of which is normal and the other non-normal with a weighting expressed by ϵ.) In particular, consistent with Eqs. 4 and 5, suppose that x has the same distribution as $y_\epsilon = T_\epsilon(z, \theta)$, where z is a normally distributed random vector and $T_\epsilon(\cdot)$ is a transformation, with ϵ measuring the degree of nonlinearity. Since $T_0(\cdot)$ is a linear transformation, the resulting artificial estimate $\hat{\theta}_0$ has one of the finite-sample distributions shown in Section 6.7 of Anderson[33] or Wilks[34] (the specific form of distribution depends on the properties of the eigenvalues of matrices defining the time series progression). Note that aside from entering the score function through the artificial data y_ϵ, ϵ appears explicitly (à la Eq. 1) through its effect on the form of the distribution (and hence likelihood function) for the data x or y_ϵ. Then, provided that ϵ is not too large, the Theorem (with or without the implied constant of the $O(\epsilon)$ bound, as appropriate) can be used with the known finite-sample distribution to determine set probabilities for testing the hypothesis of sequential uncorrelatedness in the non-normal case of interest.

MAIN RESULTS

Background and Notation

The following subsection presents the main result, showing how the difference in the unknown ($\epsilon > 0$) and known ($\epsilon = 0$) probabilities for $\hat{\theta}$ lying in a p-fold rectangle decreases as $\epsilon \to 0$. The computation of such probabilities is the critical ingredient in determining the small-sample confidence regions. In particular, we will be interested in characterizing the probabilities associated with p-fold rectangles:

$$\Omega_{a,h} = \left[a_1 - h_1, a_1 + h_1\right] \times \left[a_2 - h_2, a_2 + h_2\right] \times \ldots \times \left[a_p - h_p, a_p + h_p\right], \tag{7}$$

where $a = (a_1, a_2, \ldots, a_p)^T$, $h = (h_1, h_2, \ldots, h_p)^T$, and $h_j \geq 0 \; \forall \, j$. (Of course, by considering a union of arbitrarily small rectangles, the results here can be applied to a nonrectangular compact set subject to an arbitrarily small error.) As discussed earlier, we will use the artificial estimates $\hat{\theta}_\epsilon$ and $\hat{\theta}_0$ in analyzing this difference. The Theorem shows that the difference in probabilities is $O(\epsilon)$.

An expression of critical importance in the Theorem (and in the calculation of the implied constants for the $O(\epsilon)$ result in the Theorem) is the gradient $d\hat{\theta}_\epsilon/d\epsilon$. From the fact that $\hat{\theta}_\epsilon$ depends on y_ϵ and ϵ, we have

$$\frac{d\hat{\theta}_\epsilon}{d\epsilon} = \frac{\partial \hat{\theta}_\epsilon}{\partial \epsilon} + \frac{\partial \hat{\theta}_\epsilon}{\partial y_\epsilon^T} \frac{dy_\epsilon}{d\epsilon}. \tag{8}$$

When the score $s(\cdot)$ is a continuously differentiable function in a neighborhood of $\hat{\theta}_\epsilon$ and the y_ϵ, ϵ of interest, and when $(\partial s/\partial \theta^T)^{-1}$exists at these $\hat{\theta}_\epsilon$, y_ϵ, ϵ, the well-known implicit function theorem (e.g., Trench and Kolman[35]) applies to two of the gradients on the right-hand side of Eq. 8:

$$\frac{\partial \hat{\theta}_\epsilon}{\partial \epsilon} = -\left(\frac{\partial s}{\partial \theta^T}\right)^{-1} \frac{\partial s}{\partial \epsilon}, \tag{9}$$

$$\frac{\partial \hat{\theta}_\epsilon}{\partial y_\epsilon^T} = -\left(\frac{\partial s}{\partial \theta^T}\right)^{-1} \frac{\partial s}{\partial y_\epsilon^T}, \tag{10}$$

where the right-hand sides of the expressions in Eqs. 9 and 10 are evaluated at $\hat{\theta}_\epsilon$, $y_\epsilon = T_\epsilon(z, \theta)$, and the ϵ of

interest. (All references to $s = s(\cdot)$ here and in the Theorem correspond to the definition in Eq. 1, with y_ϵ replacing x as in Eq. 4.) The remaining gradient on the right-hand side of Eq. 8 is obtainable directly as $dy_\epsilon/d\epsilon = dT_\epsilon(z, \theta)/d\epsilon$. Note that y_ϵ (and its derivative in Eq. 8) is evaluated at the true θ in contrast to the other expressions in Eqs. 8–10, which are evaluated at the estimated θ. One interesting implication of Eqs. 8–10 is that $d\hat{\theta}_\epsilon/d\epsilon$ is explicitly calculable even though $\hat{\theta}_\epsilon$ is, in general, only implicitly defined. From the implicit function theorem, the computation of $d\hat{\theta}_\epsilon/d\epsilon$ for the important special case of $\epsilon = 0$ (see notation that follows) relies on the previously mentioned assumptions of continuous differentiability for $s(\cdot)$ holding for ϵ both slightly positive and negative.

The following notation will be used in the Theorem conditions and proof:

- Consistent with preceding usage, a subscript i or j on a vector (say on $\hat{\theta}_\epsilon$, z, etc.) denotes the ith or jth component.
- A_ϵ represents the inverse image of $\Omega_{a,h}$ relative to $\hat{\theta}_\epsilon$, i.e., the set $\{z: a_j - h_j \le \hat{\theta}_{\epsilon,j} \le a_j + h_j \forall\, j = 1,2, \ldots, p\}$. Likewise, A_0 is the inverse image relative to $\hat{\theta}_0$.
- $\Delta(z) = [d\hat{\theta}_\epsilon/d\epsilon]_{\epsilon = 0}$.

Order Result on Small-Sample Probabilities

The main theoretical result of this article is presented in the Theorem. A proof is provided in Spall.[36] The Theorem regularity conditions are quite modest, as discussed in the remarks following their presentation in the Appendix and as demonstrated in the signal-plus-noise/CEP example discussed later. The regularity conditions pertain essentially to smoothness properties of the score vector and to characteristics of the distribution of z and would apply in almost all practical test and evaluation applications.

Theorem. Let $\hat{\theta}_\epsilon$ and $\hat{\theta}_0$ be as given in Eqs. 4 and 5, and let $a \pm h$ be continuity points of the associated distribution functions. Then, under regularity conditions C.1–C.5 in the Appendix,

$$P(\hat{\theta}_\epsilon \in \Omega_{a,h}) - P(\hat{\theta}_\epsilon \in \Omega_{a,h}) = O(\epsilon)\,. \qquad (11)$$

The Implied Constant of $O(\epsilon)$ Bound

Through the form of the calculations in the proof of the Theorem, it is possible to produce computable implied constants for the $O(\epsilon)$ bound, i.e., constants $c(a, h) > 0$ such that

$$\left|P(\hat{\theta}_\epsilon \in \Omega_{a,h}) - P(\hat{\theta}_0 \in \Omega_{a,h})\right| \le c(a,h)\epsilon + o(\epsilon)\,. \quad (12)$$

We present one such constant here; another is presented in Spall.[36] The constant here will tend to be conservative in that it is based on upper bounds to certain quantities in the proof of the Theorem. This conservativeness may be desirable in cases where ϵ is relatively large to ensure that the "$\le$" in Eq. 12 is preserved in practical applications [i.e., when the $o(\epsilon)$ term is ignored]. (The constant in Spall[36] is less conservative and is determined through a computer resampling procedure.)

The details behind the derivation of the bound here follow exactly as in Spall[36] and so will not be repeated here. The bound is

$$c(a,h) = 2\sum_{j=1}^{p} M_j P(a_i - h_i \le \hat{\theta}_{0,i} \le a_i + h_i \forall\, i \ne j) \times [p_j(\zeta_j^{(-)}) + p_j(\zeta_j^{(+)})], \qquad (13)$$

where M_j is an upper bound to $|\Delta_j(z)|$ for $z \in A_0$, $p_j(\cdot)$ is the marginal density function for $\hat{\theta}_{0,j}$, $\zeta_j^{(+)} \in [a_j + h_j - \epsilon M_j, a_j + h_j]$, and $\zeta_j^{(-)} \in [a_j - h_j, a_j - h_j + \epsilon M_j]$. From a practical point of view, $\zeta_j^{(+)}$ and $\zeta_j^{(-)}$ could be chosen as the midpoint of the (assumed small) intervals in which they lie.

APPLICATION IN SIGNAL-PLUS-NOISE AND RELATED CEP PROBLEM

Background

This section returns to Example 1 and presents an analysis of how the small-sample approach would apply in practical test and evaluation. In particular, consider independent scalar observations $\{x_1, x_2, \ldots, x_n\}$, distributed $x_i \sim N(\mu, \sigma^2 + Q_i)$, where the Q_i are known and $\theta = (\mu, \sigma^2)^T$ is to be estimated using maximum likelihood. We also consider the CEP estimate derived from the θ estimate. As mentioned in the discussion of Example 1, when $Q_i \ne Q_j$ for at least one i, j (the $\epsilon \ne 0$ actual case), no closed-form expression (and hence no computable distribution) is generally available for $\hat{\theta}$. When $Q_i = Q_j$ for all i, j (the $\epsilon = 0$ idealized case), the distribution of $\hat{\theta}$ is known (see Eqs. 18 and 19 in the next subsection).

This example is directly motivated by the author's work in accuracy analysis for naval missile systems. The parameters μ and σ^2 represent the impact mean (relative to the target point) and variance along either the downrange or crossrange direction. (The real implementation of the small-sample approach pertains to the simultaneous estimation of the downrange/crossrange parameters, but we do not present those results here in

order to better illustrate the fundamental issues without the encumbrances and additional notation of the multivariate analysis.) The Q_i terms represent the variance of the measurement error associated with the impact measurement instrumentation. After obtaining an estimate of θ (μ and σ^2), we then estimate the missile CEP in a manner analogous to Spall and Maryak.[25] In particular, if $R = R(\theta)$ represents the function translating the impact mean and variance into the CEP radius, the CEP estimate (MLE) is $\hat{R} = R(\hat{\theta})$ (i.e., a deterministic function of an MLE is also an MLE, as in, e.g., Bickel and Doksum,[32] p.111). One of the subsections that follow will elaborate on some of this in the context of presenting numerical results for CEP estimation.

For this estimation problem, the next subsection discusses the regularity conditions of the Theorem and comments on the calculation of the implied constant $c(a, h)$, and the following two subsections present some numerical results. This two-parameter estimation problem is one where the other analytical techniques discussed in the Introduction (i.e., Edgeworth expansion and saddlepoint approximation) are impractical because of the unwieldy calculations required (say, as related to the cumulant generating function and its inverse). The parametric bootstrap technique was also tested, but it performed poorly because of the small-sample size from which the resampling was performed, as discussed later.

When using the $\epsilon = 0$ distribution for $\hat{\theta}$ as an approximation to the actual $\epsilon \neq 0$ distribution (when justified by the Theorem), we choose a value of Q corresponding to the "information average" of the individual's Q_i's, i.e., Q is such that $Q^{-1} = n^{-1}\sum_{i=1}^{n} Q_i^{-1}$. (The idea of summing information terms for different measurements is analogous to the idea in Rao.[37]) As mentioned in the discussion of Example 1, deviations of order ϵ from the common Q are then naturally expressed in the inverse domain: $Q_i^{-1} = Q^{-1} + \epsilon D_i$, where the D_i's are some fixed quantities (discussed later). Working with information averages has proven desirable as a way of down-weighting the relative contribution of the larger Q_i's versus what their contribution would be, say, if Q were a simple mean of the Q_i's. (From Eq. 15 that follows, we see that the score expression also down-weights the data associated with larger Q_i.) A further reason to favor the information average is that the score is naturally parameterized directly in terms of Q_i^{-1} through use of the relationship $(\sigma^2 + Q_i)^{-1} = Q_i^{-1} - (1 + \sigma^2 Q_i^{-1})^{-1}\sigma^2 Q_i^{-2}$. Hence, Q^{-1} represents the mean of the natural nuisance parameters in the problem. Finally, we have found numerically that the idealized probabilities computed with the information average have provided more accurate approximations to the true probabilities when the Q_i's vary moderately than, say, idealized probabilities based on an average equal to the mean Q_i. Note, however, that any type of average Q_i will work when the Q_i's are sufficiently close since $Q_i^{-1} - Q^{-1} = O(\epsilon)$ if and only if $Q_i - Q = O(\epsilon)$ when $Q > 0$.

The log-likelihood function, $L(\theta; x, \epsilon)$, for the estimation of $\theta = (\mu, \sigma^2)^T$ is

$$L(\theta; x, \epsilon) = -\sum_{i=1}^{n}\left[\log(\sigma^2 + Q_i) + (\sigma^2 + Q_i)^{-1}(x_i - \mu)^2\right] + \text{constant}, \tag{14}$$

where $Q_i = Q_i(\epsilon) = (Q^{-1} + \epsilon D_i)^{-1}$, from which the score expression $s(\theta; x, \epsilon) = \partial L/\partial\theta$ is found:

$$s(\theta; x, \epsilon) = \begin{bmatrix} 2\sum_{i=1}^{n}(\sigma^2 + Q_i)^{-1}(x_i - \mu) \\ \sum_{i=1}^{n}[-(\sigma^2 + Q_i)^{-1} + (\sigma^2 + Q_i)^{-2}(x_i - \mu)^2] \end{bmatrix}. \tag{15}$$

Since $\sigma^2 \geq 0$, we will consider only those sets of interest (i.e., $\Omega_{a,h} = [a_1 - h_1, a_1 + h_1] \times [a_2 - h_2, a_2 + h_2]$) such that $a_2 - h_2 \geq 0$. This does not preclude having a practical estimate $\hat{\sigma}^2 < 0$ come from $s(\theta; x, \epsilon) = 0$ (in which case one would typically set $\hat{\sigma}^2$ to 0); however, in specifying confidence sets, we will only consider those points in σ^2 space that make physical sense. (Note that if n is reasonably large and/or the $\{Q_i\}$ are reasonably small relative to σ^2, then $\hat{\sigma}^2$ from $s(\theta; x, \epsilon) = 0$ will almost always be positive.)

Theorem Regularity Conditions and Calculation of Implied Constant

The first step in checking the conditions for the Theorem is to define the artificial data sequences, $\{y_{\epsilon,i}\}$, $\{y_{0,i}\}$, and associated artificial estimators $\hat{\theta}_\epsilon$ and $\hat{\theta}_0$. From the definitions in the "Problem Formulation" section, the two artificial MLEs are

$$\hat{\theta}_\epsilon = \left\{\theta : s(\theta; y_\epsilon, \epsilon) = 0\right\}, \tag{16}$$

$$\hat{\theta}_0 = \begin{bmatrix} \hat{\mu}_0 \\ \hat{\sigma}_0^2 \end{bmatrix} = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} y_{0,i} \\ n^{-1}\sum_{i=1}^{n}(y_{0,i} - \bar{y})^2 - Q \end{bmatrix}, \tag{17}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_{0,i}$. As required to apply the Theorem, $\hat{\theta}_0$ has a known distribution (the same, of course, as for the $\hat{\theta}$ of interest from Eq. 15 when $Q_i = Q_j \; \forall\, i,j$). In particular, $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ satisfy

$$\hat{\mu}_0 \sim N[\mu, (\sigma^2 + Q)/n] \quad \text{(normal)}, \tag{18}$$

$$n\frac{\hat{\sigma}_0^2 + Q}{\sigma^2 + Q} \sim \chi^2_{n-1} \quad \text{(chi-squared)}. \tag{19}$$

Spall[36] includes a verification of the regularity conditions C.1–C.4 in the Appendix (C.5 is immediate by the definition of $\{z_i\}$). We assume that $Q > 0$; then $Q_i > 0$ for all ϵ in a neighborhood of 0 (i.e., Q_i is well defined as a variance for all ϵ near 0, including $\epsilon < 0$, as required by the implicit function theorem in computing $\Delta(z)$, as discussed previously).

Now consider the calculation of the constant $c(a, h)$ introduced earlier. Although an analytical form is available for $\Delta(z)$, it may be easier in practice to approximate $\max\{|\Delta_j(z)|: z \in A_0\}$ for each j by randomly sampling $\Delta(z)$ over $z \in A_0$. This yields estimates of M_1 and M_2 and is the procedure used in computing $c(a, h)$ in the subsection that follows. The probabilities $P(a_j - h_j \le \hat{\theta}_{0,j} \le a_j + h_j)$ for $j = 1, 2$ are readily available by the normal and chi-squared distributions for $\hat{\mu}_0$ and $\hat{\sigma}_0^2$. Likewise, the density-based values $p_j(\zeta_j^{(\pm)})$ are easily approximated by taking $\zeta_j^{(\pm)}$ as an intermediate (we use mid) point of the appropriate interval $[a_j + h_j - \epsilon M_j, a_j + h_j]$. This provides all the elements needed for a practical determination of $c(a, h)$, as illustrated in the next subsection.

Numerical Results for Signal-Plus-Noise Problem

This subsection presents results of a numerical study of the preceding MLE problem. Our goals here are primarily to compare the accuracy of confidence regions based on the small-sample theory with the actual (empirically determined) regions. We also briefly examine the performance of the bootstrap technique and asymptotic MLE theory. Computations for this subsection were performed on an IBM mainframe with IMSL subroutines DRNNOR to generate normal random variables and DNEQNJ to find the solution of the MLE score equations. (The high variability in the small-sample estimates requires that very large Monte Carlo studies be performed here; these studies were beyond the capability of the Pentium-based PCs available to the author.) In this study, we took $n = 5$ and generated data according to $x_i \sim N(0, 1 + Q_i)$ with Q_i such that $Q_i^{-1} = 0.04^{-1} + \epsilon D_i$, $D_1 = D_2 = D_3 = 40$, $D_4 = D_5 = -60$ (so the average Q_i, in an information sense, is 0.04 according to the earlier discussion). As discussed previously, we estimate $\theta = (\mu, \sigma^2)^T$ and are interested in confidence regions for $\hat{\theta}_1 = \hat{\mu}$ and $\hat{\theta}_2 = \hat{\sigma}^2$. For ease of presentation and interpretation, we will focus largely on the marginal distributions and confidence regions for each of $\hat{\mu}$ and $\hat{\sigma}^2$; this also is partly justified by the fact that $\hat{\mu}$ and $\hat{\sigma}^2$ are approximately independent (i.e., when either $\epsilon = 0$ or n is large, $\hat{\mu}$ and $\hat{\sigma}^2$ are independent). We will report results for $\epsilon = 0.15$ and $\epsilon = 0.30$, which correspond to values of $\{Q_1, Q_2, Q_3, Q_4, Q_5\}$ equal to $\{0.0323, 0.0323, 0.0323, 0.0625, 0.0625\}$ and $\{0.0271, 0.0271, 0.0271, 0.143, 0.143\}$, respectively. Results for these two values of ϵ are intended to represent the performance of the small-sample theory for a small- and a moderate-sized ϵ.

Before proceeding, let us briefly discuss our experience with the bootstrap method mentioned in the Introduction. Since our data are non-i.i.d., we used a parametric bootstrap approach, as discussed in Sections 2 and 7 in Efron and Tibshirani.[3] Essentially, the bootstrap confidence region is determined by first estimating θ from the given sample of size 5, and then, following Efron and Tibshirani,[3] generating 1000 bootstrap data sets (also of size 5) from the distributions $N(\hat{\mu}, \hat{\sigma}^2 + Q_i)$, $i = 1$ to 5. These bootstrap data sets are used to produce 1000 new estimates $\hat{\mu}^*$, $\hat{\sigma}^{2*}$, which are then ranked and sorted to determine quantile points (and the associated confidence regions). This procedure was repeated for 10 original samples of size 5 to produce 10 different bootstrap confidence regions in the $\epsilon = 0.15$ case. (Of course, in practice, only one sample is available.) These 10 confidence regions varied considerably: the width of a 95% (marginal) confidence interval for $\hat{\mu}$ varied from 0.39 to 4.91, whereas for $\hat{\sigma}^2$ it varied from 0.09 to 16.58 (the true widths are 1.79 and 2.22, respectively, as considered later). This unacceptable variation is inherently a result of the small sample size, and no real improvement was seen with larger (e.g., 10,000) bootstrap data sets. Hence, we rule out the bootstrap method from further consideration in this small-sample setting. These results also suggest that one should use the bootstrap with extreme caution in other small-sample test and evaluation settings.

Spall[36] shows that for the μ portion of θ, the small-sample confidence intervals differ little from the true intervals or those obtained by asymptotic theory. Hence, we focus here on the σ^2 part of θ. Figure 1 depicts three density functions for $\hat{\sigma}^2$ for each of the $\epsilon = 0.15$ and $\epsilon = 0.30$ cases: (1) the "true" density based on the marginal histogram constructed from 2.5×10^6 estimates of θ determined from 2.5×10^6 independent sets of $n = 5$ measurements (a smoother in the SAS/GRAPH software system[38] was used to smooth out the small amount of jaggedness in the empirical histogram), (2) the small-sample density from Eq. 19 (corresponding to the idealized $O(\epsilon) = 0$ case), and (3) the asymptotic-based normal density with mean = 1 and variance given by the appropriate diagonal element of the inverse Fisher information matrix for $\hat{\theta}$. We see that with $\epsilon = 0.15$, the true and small-sample densities are virtually identical throughout the domain while the

asymptotic-based density is dramatically different. For $\epsilon = 0.30$, there is some degradation in the match between the true and idealized small-sample densities, but the match is still much better than between the true and asymptotic-based densities. (Of course, it is the purpose of the $O(\epsilon)$ adjustment based on $c(a, h)$ to compensate for such a discrepancy in confidence interval calculation, as discussed below.) Note that the true densities illustrate the frequency with which we can expect to see a negative variance estimate, which is an inherent problem due to the small size of the sample (the asymptotic-based density significantly overstates this frequency). Because of the relatively poor performance of the asymptotic-based approach, we focus here on comparing confidence regions from only the true distributions and the small-sample approach.

Figure 2 translates the preceding situation into a comparison of small-sample confidence regions with the true regions. Included here are regions based on the $O(\epsilon)$ term of the Theorem when quantified through use of the constant, $c(a, h)$. The indicated interval endpoints were chosen based on preserving equal probability (0.025) in each tail, with the exception of the conservative $\epsilon = 0.30$ case; here the lower bound went slightly below 0 using symmetry, so the lower endpoint was shifted upward to 0 with a corresponding adjustment made to the upper endpoint to preserve at least 95% coverage. (Spall[36] includes more detail on how the $O(\epsilon) = c(a, h)\epsilon$ probability adjustment was translated into a confidence interval adjustment.) For $\epsilon = 0.15$, we see that the idealized small-sample bound is identical to the true bound. (This, of course, is the most desirable situation since there is then no need to work with the $c(a, h)$-based adjustment.) As expected, the confidence intervals with the conservative $c(a, h)$-based adjustments are wider. For $\epsilon = 0.30$, there is some degradation in the accuracy of coverage for the idealized small-sample interval, which implies a greater need to use the conservative interval to ensure the intended coverage probability for the interval.

The preceding study is fully representative of others that we have conducted for this estimation framework (e.g., nominal coverage probabilities of 90% and 99% and other values of $0 < \epsilon \leq 30$). They illustrate that with relatively small values of ϵ, the idealized confidence intervals are very accurate, but that with larger values (e.g., $\epsilon = 0.25$ or 0.30), the idealized interval becomes visibly too short. In these cases, the $c(a, h)$-based adjustment to the idealized interval provides a means for broadening the coverage to encompass the true confidence interval.

Numerical Results for CEP Problem

We now illustrate how the results in the preceding subsection translate into confidence intervals for the weapon system CEP estimate. As mentioned earlier, the MLE for CEP is $\hat{R} = R(\hat{\theta})$, where $R(\cdot)$ represents the nonlinear CEP function relating the impact means and variances to the radius such that there is a 50% probability of an impact landing within a circle of this radius about the target point. The function $R(\cdot)$ is presented, for example, in Spall and Maryak[25] and Shnidman.[39] The CEP value for a weapon system is typically the single most important measure of accuracy for the system, and is a critical number in making strategic decisions related to targeting and number of missiles to fire. Hence, it is important to make accurate statements about likely maximum and minimum values for the CEP based on the small number of tests available.

Figure 3 compares the 90% confidence intervals for $\hat{R}$ resulting from an application of the small-sample approach of this article and from the classical asymptotic approach discussed in the preceding subsection.

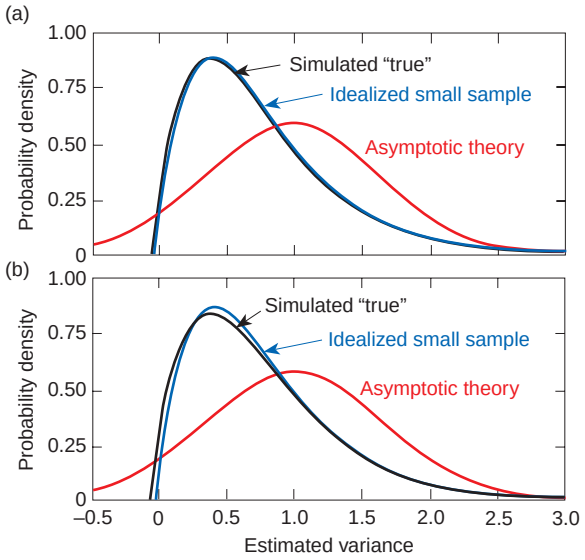

Figure 1. Comparison of true, idealized small-sample, and asymptotic density functions for $\hat{\sigma}^2$ when (a) $\epsilon = 0.15$ and (b) $\epsilon = 0.30$.

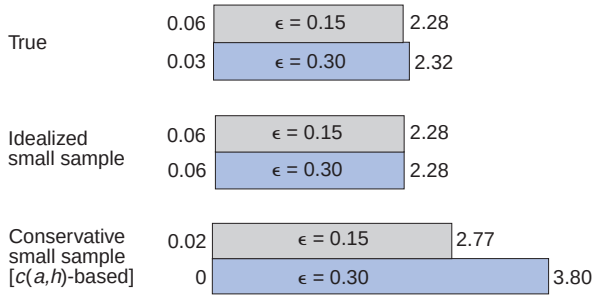

Figure 2. True and small-sample 95% confidence intervals for $\hat{\sigma}^2$ when $\epsilon = 0.15$ and $\epsilon = 0.30$.

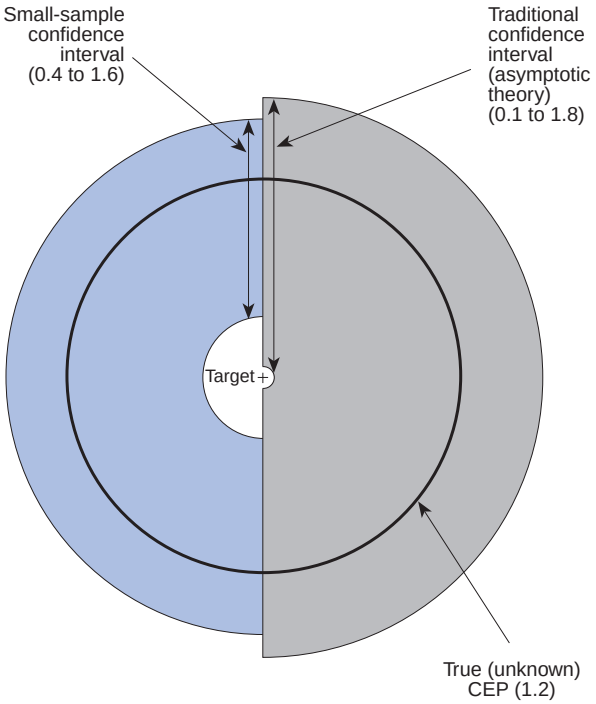

Figure 3. 90% confidence intervals on CEP from 5 test flights.

These results are derived from the $\epsilon = 0.15$ case (so the small-sample confidence interval based on the idealized case is essentially identical to the true confidence interval). As a reflection of the limited information in only five test flights, the confidence intervals are relatively wide. However, the figure shows that the interval based on asymptotic theory is over 40% wider than the (true) small-sample interval. Hence, by more properly characterizing the distribution of the underlying parameter estimates, the small-sample approach is able to extract more information about the CEP quantity of interest from the limited number of test flights.

SUMMARY AND CONCLUSIONS

Making statistical inference in small samples is a problem encountered in many applications. Although techniques such as the bootstrap and saddlepoint approximation have shown some promise in the small-sample setting, there remain serious difficulties in accuracy and feasibility for the type of multivariate M-estimation problems frequently encountered in practical test and evaluation applications.

For a range of problem settings, the approach described here is able to provide accurate information about the estimate uncertainties. The primary restriction is that an idealized case must be identified (where the estimate uncertainty is known for the given sample size) with which the estimate uncertainty for the actual case is compared. A Theorem was presented that provides the basis for comparing the actual and idealized cases. The Theorem provides a bound on the difference between the cases. Implementations of the approach were discussed for three distinct well-known settings to illustrate the range of potential applications. These were a signal-plus-noise estimation problem (see the Appendix), a general nonlinear regression setting, and a problem in time series correlation analysis.

In illustrating the small-sample approach, this article focused mainly on a signal-plus-noise problem arising in a Navy test and evaluation program for missile accuracy analysis. It was shown that the small-sample approach yields a significant improvement over the conventional method based on large-sample theory. In particular, when translating the results into the CEP estimate confidence intervals, it was found that the small-sample approach yielded intervals that were significantly tighter and more precise than those resulting from the previously used large-sample approximations. This finding represents an increased understanding of the weapon system performance with no expenditure for additional tests. Hence, the methodology of this article is one example of how improved analytical techniques may be able to compensate for reductions in DoD test and evaluation budgets.

Although the approach here was developed for M-estimates (largely for purposes of identifying explicit regularity conditions in terms of the score function), this restriction is not necessary. In other small-sample settings, it appears that the ideas would also apply provided that an idealized case can be identified. Although such extensions are desirable, the analysis here shows that the approach is broadly applicable to small-sample problems of practical interest.

APPENDIX: THEOREM REGULARITY CONDITIONS

Regularity conditions C.1–C.5 for the Theorem are as follows:

C.1. Let $S_j^{(\pm)} \equiv \{z: \hat{\theta}_{0,j} - a_j \pm h_j = 0\} \cap A_0$ be a bounded set $\forall\, j = 1, 2, \ldots, p$ ($S_j^{(\pm)} = \varnothing$ is valid). Further, if $S_j^{(\pm)} = \varnothing$, suppose that $\hat{\theta}_{0,j} - a_j \pm h_j$ is uniformly bounded away from 0 on A_0. If $S_j^{(\pm)} \neq \varnothing$, then, except in an open neighborhood of $S_j^{(\pm)}$ (i.e., a region such that for some radius > 0, an n-ball of this radius around any point in $S_j^{(\pm)}$ is contained within this region), we have $\hat{\theta}_{0,j} - a_j \pm h_j$ uniformly bounded away from 0 on A_0.

C.2. Except on a set of P_z-measure 0, $\Delta = \Delta(z)$ exists on A_0. Further, for each $j = 1, 2, \ldots, p$, $\Delta_j \neq 0$ on A_0 almost surely (P_z) and $P(|\Delta_j|^{-1} \leq c\epsilon, A_0) = o(\epsilon)\ \forall\, 0 < c < \infty$.

C.3. For each $j = 1, 2, \ldots, p$, when $S_j^{(\pm)} \neq \varnothing$, suppose that there exists an open neighborhood of $S_j^{(\pm)}$ (see C.1) such that $d\hat{\theta}_0/dz$ and $d\Delta_j/dz$ exist continuously in the neighborhood. Further, for each j and sign $\pm$, there exists some scalar element in z, say $z_{kj(\pm)}$, such that $\forall\, z \in S_j^{(\pm)}$ we have $d\hat{\theta}_{0,j}/dz_{kj(\pm)} \neq 0$.

C.4. Let $\tilde{\theta}_j = \hat{\theta}_{\epsilon,j} - \hat{\theta}_{0,j}$. Then $\forall\ j = 1, 2, \ldots, p$,

$$P(-|\tilde{\theta}_j| \le \hat{\theta}_{0,j} - a_j + h_j \le 0, A_\epsilon) - P(0 \le \hat{\theta}_{0,j} - a_j + h_j \le |\tilde{\theta}_j|, A_0) = o(\epsilon),$$

$$P(0 \le \hat{\theta}_{0,j} - a_j - h_j \le |\tilde{\theta}_j|, A_\epsilon) - P(-|\tilde{\theta}_j| \le \hat{\theta}_{0,j} - a_j - h_j \le 0, A_0) = o(\epsilon).$$

C.5. For each $j = 1, 2, \ldots, p$, the distribution of z is absolutely continuous in an open neighborhood of $S_j^{(\pm)}$ (see C.1).

Remarks on C.1–C.5. In checking C.1, note that A_0 will often be a bounded set, which automatically implies that $S_j^{(\pm)}$ is bounded $\forall\ j$. The other requirement on $\hat{\theta}_{0,j} - a_j \pm h_j$ being bounded away from 0 is straightforward to check since $\hat{\theta}_{0,j}$ has a known distribution. Given the form of the score $s(\cdot)$ and transformation $T_\epsilon(\cdot)$, Eqs. 8–10 readily suggest when C.2 will be satisfied. Note that when A_0 is bounded and Δ_j is continuous on A_0, the last part of the condition will be automatically satisfied since Δ_j will be bounded above. The conditions in C.3 pertaining to $\hat{\theta}_0$ can generally be checked directly since $\hat{\theta}_0$ is typically available in closed form (its distribution is certainly available in closed form); the condition on $d\Delta_j/dz$ can be checked through use of Eqs. 8–10 (as in checking C.2). As illustrated earlier, to satisfy the near-symmetry condition, C.4, it is sufficient that $\hat{\theta}_{0,j}$ have a continuous density near $a_j \pm h_j\ \forall\ j$ and that $d\hat{\theta}_\epsilon/d\epsilon$ (from Eq. 8) be uniformly bounded on $A_0 \cup A_\epsilon$. (Of course, this condition can also be satisfied in settings where $d\hat{\theta}_\epsilon/d\epsilon$ is not bounded in this way.) Finally, C.5 is a straightforward condition to check since the analyst specifies the distribution for z. Note that none of the conditions impose any requirements of bounded moments for $\hat{\theta}_\epsilon$ (or equivalently, for θ), which would be virtually impossible to check for most practical M-estimation problems.

REFERENCES

[1] Ljung, L., "Convergence Analysis of Parametric Identification Methods," *IEEE Trans. Autom. Control* AC-23, 770–783 (1978).

[2] Serfling, R. J., *Approximation Theorems of Mathematical Statistics*, Wiley, New York (1980).

[3] Efron, B., and Tibshirani, R., "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy (with discussion)," *Stat. Sci.* 1, 54–77 (1986).

[4] Hall, P., *The Bootstrap and the Edgeworth Expansion*, Springer-Verlag, New York (1992).

[5] Hjorth, J. S. U., *Computer Intensive Statistical Methods*, Chapman and Hall, London (1994).

[6] Frickenstein, S., "Estimating Uncertainty Using the Bootstrap Technique," 63rd Military Operations Research Society Symposium, Final Program and Book of Abstracts, pp. 105–106 (1995).

[7] Daniels, H. E., "Saddlepoint Approximations in Statistics," *Ann. Math. Stat.* 25, 631–650 (1954).

[8] Reid, N., "Saddlepoint Methods and Statistical Inference (with Discussion)," *Stat. Sci.* 3, 213–238 (1988).

[9] Field, C., and Ronchetti, E., Chap. 6 in *Small Sample Asymptotics*, IMS Lecture Notes—Monograph Series, Vol. 13, Institute of Mathematical Statistics, Hayward, CA (1990).

[10] Ghosh, J. K., *Higher Order Asymptotics*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 4, Institute of Mathematical Statistics, Hayward, CA (1994).

[11] Kolassa, J. E., "Saddlepoint Approximations in the Case of Intractable Cumulant Generating Functions," *Selected Proceedings of the Sheffield Symposium on Applied Probability*, IMS Lecture Notes—Monograph Series, Vol. 18, Institute of Mathematical Statistics, Hayward, CA, pp. 236–255 (1991).

[12] Fraser, D. A. S., and Reid, N., "Third-Order Asymptotic Models: Likelihood Functions Leading to Accurate Approximations for Distribution Functions," *Stat. Sinica* 3, 67–82 (1993).

[13] Chen, Z., and Do, K.-A., "The Bootstrap Method with Saddlepoint Approximations and Importance Sampling," *Stat. Sinica* 4, 407–421 (1994).

[14] Tingley, M. A., and Field, C. A., "Small Sample Confidence Intervals," *J. Am. Stat. Assoc.* 85, 427–434 (1990).

[15] Wood, A. T. A., Booth, J. G., and Butler, R. W., "Saddlepoint Approximation to the CDF of Some Statistics with Nonnormal Limit Distributions," *J. Am. Stat. Assoc.* 88, 680–686 (1993).

[16] Lugannani, R., and Rice, S., "Saddlepoint Approximation for the Distribution of the Sum of Independent Random Variables," *Adv. Appl. Probab.* 12, 475–490 (1980).

[17] Ronchetti, E., and Welsh, A. H., "Empirical Saddlepoint Approximations for Multivariate M-estimators," *J. R. Stat. Soc. B* 56, 313–326 (1994).

[18] Rao, P. S. R. S., Kaplan, J., and Cochran, W. G., "Estimators for the One-Way Random Effects Model with Unequal Error Variances," *J. Am. Stat. Assoc.* 76, 89–97 (1981).

[19] James, A. T., and Venables, W. N., "Matrix Weighting of Several Regression Coefficient Vectors," *Ann. Stat.* 21, 1093–1114 (1993).

[20] National Research Council, *Combining Information: Statistical Issues and Opportunities for Research*, National Academy of Sciences, Washington, D.C., pp. 143–144 (1992).

[21] Shumway, R. H., Olsen, D. E., and Levy, L. J., "Estimation and Tests of Hypotheses for the Initial Mean and Covariance in the Kalman Filter Model," *Communications in Statistics—Theory and Methods* 10, 1625–1641 (1981).

[22] Sun, F. K., "A Maximum Likelihood Algorithm for the Mean and Covariance of Nonidentically Distributed Observations," *IEEE Trans. Autom. Control* AC-27, 245–247 (1982).

[23] Ghosh, M., and Rao, J. N. K., "Small Area Estimation: An Approach (with discussion)," *Stat. Sci.* 9, 55–93 (1994).

[24] Hui, S. L., and Berger, J. O., "Empirical Bayes Estimation of Rates in Longitudinal Studies," *J. Am. Stat. Assoc.* 78, 753–760 (1983).

[25] Spall, J. C., and Maryak, J. L., "A Feasible Bayesian Estimator of Quantiles for Projectile Accuracy from Non-i.i.d. Data," *J. Am. Stat. Assoc.* 87, 676–681 (1992).

[26] Smith, R. H., "Maximum Likelihood Mean and Covariance Matrix Estimation Constrained to General Positive Semi-Definiteness," *Communications in Statistics—Theory and Methods* 14, 2163–2180 (1985).

[27] Spall, J. C., and Chin, D. C., "First-Order Data Sensitivity Measures with Applications to a Multivariate Signal-Plus-Noise Problem," *Computational Statistics and Data Analysis* 9, 297–307 (1990).

[28] Pazman, A., "Small-Sample Distributional Properties of Nonlinear Regression Estimators (a Geometric Approach) (with discussion)," *Statistics* 21, 323–367 (1990).

[29] Kmenta, J., *Elements of Econometrics*, Macmillan, New York, pp. 462–464 (1971).

[30] Nicholson, W., *Microeconomic Theory*, Dryden, Hinsdale, IL, pp. 200–201 (1978).

[31] Cook, P., "Small-Sample Bayesian Frequency-Domain Analysis of Autoregressive Models," in *Bayesian Analysis of Time Series and Dynamic Models*, J. C. Spall (ed.), Marcel Dekker, New York, pp. 101–126 (1988).

[32] Bickel, P. J., and Doksum, K. A., *Mathematical Statistics*, Holden-Day, Oakland, CA (1977).

[33] Anderson, T. W., *The Statistical Analysis of Time Series*, Wiley, New York, Section 6.1 (1971).

[34] Wilks, S. S., *Mathematical Statistics*, Wiley, New York, pp. 592–593 (1962).

[35] Trench, W. F., and Kolman, B., *Multivariable Calculus with Linear Algebra and Series*, Academic, New York, p. 370 (1972).

[36] Spall, J. C., "Uncertainty Bounds for Parameter Identification with Small Sample Sizes," *Proc. IEEE Conf. on Decision and Control*, pp. 3504–3515 (1995).

[37] Rao, C. R., *Linear Statistical Inference and its Applications*, Wiley, New York, pp. 329–331 (1973).

[38] SAS Institute, *SAS/GRAPH Software: Reference, Version 6, First Edition*, SAS Institute, Cary, NC, p. 416 (1990).

[39] Shnidman, D. A., "Efficient Computation of the Circular Error Probable (CEP) Integral," *IEEE Trans. Autom. Control* 40, 1472–1474 (1995).

ACKNOWLEDGMENTS: This work was supported by U.S. Navy Contract N00039-95-C-0002. John L. Maryak of APL provided many helpful comments, and Robert C. Koch of the Federal National Mortgage Association (Fannie Mae) provided valuable computational assistance in carrying out the example. A preliminary version of this article was selected as the best paper in the Test and Evaluation Working Group at the 1995 Military Operations Research Society Annual Symposium.

THE AUTHOR

JAMES C. SPALL joined APL in 1983 and was appointed to the Principal Professional Staff in 1991. He also teaches in the JHU Whiting School Part-time Engineering Program. Dr. Spall has published over 80 articles in the areas of statistics and control and holds two U.S. patents. For the year 1990, he received the Hart Prize as principal investigator of the most outstanding Independent Research and Development project at APL. He is an Associate Editor for the *IEEE Transactions on Automatic Control* and a Contributing Editor for the *Current Index to Statistics*, and he served as editor and coauthor for the book *Bayesian Analysis of Time Series and Dynamic Models*. Dr. Spall is a senior member of IEEE, a member of the American Statistical Association and of Sigma Xi, and a fellow of the engineering honor society Tau Beta Pi. His e-mail address is James.Spall@jhuapl.edu.