Statistica Sinica 4(1994), 701-714

APPROXIMATE PIVOTS FROM M-ESTIMATORS

Christopher J. Lloyd

LaTrobe University

Abstract: We consider the problem of inference on an M-estimator of the location of a continuous but unknown density function from a single sample. Four approximate pivots are studied including one derived from the empirical likelihood of Owen (1988). It is argued that first order M-estimator inference is often correct, close to second order, and that two of the four pivots are distinctly better than the others in this regard.

Key words and phrases: Bootstrap likelihood, empirical likelihood, second order accuracy, edgeworth approximation, robustness.

1. Introduction

Let $X_1, \ldots, X_n$ be an identical independent sample on a continuous random variable X with absolutely continuous distribution function F and density dF. We suppose that F is unknown, the so-called distribution free assumption. In much of the robustness literature F is further assumed symmetric. Let $\theta(F)$ be a functional which in this paper will be a location parameter. Inference on θ can be based on a *pivot* which is a function $P(X_1, \ldots, X_n, t)$ with the property that the distribution of

$$P(X_1, \ldots, X_n, \theta(F))$$

does not depend on F and therefore on $\theta(F)$. When F is completely unknown then approximate pivots are more often used, exact pivots being rare or inefficient (e.g. sign test for a location parameter).

Commonly the pivot P has a distribution $H(F)$ which converges to standard normal as n diverges. Departures from the standard normal density function $\phi(x)$ can often be represented via an Edgeworth expansion which is a sequence of approximations to the density function h of P of the form

$$h(x) \approx \phi(x) \left\{ 1 + \frac{q(x)}{\sqrt{n}} \right\} + O(n^{-1}), \tag{1}$$

where $q(x)$ is a polynomial of degree 3 which typically adjusts for skewness and is called the *second order* term. The expansion requires that P has been standardized to have mean zero and variance 1 with error $O(n^{-1})$. If this is not the case then lower order terms are needed. For fixed x and large n the right hand side converges to, and provides a better approximation to, the left hand side than the standard normal density. When θ is a smooth function of means, sufficient conditions for this expansion are that F has an absolutely continuous component (i.e. is not discrete) and that the third absolute moment of P exists. A pivot whose scale must be estimated is called studentized. The fourth absolute moment of P must exist for a valid expansion of a studentized pivot. The distribution of a continuous transform of P may be expanded by tranforming a valid expansion of P. For an account of these results see Hall (1992, Chapter 2). Using his expression (2.55) and differentiating, an explicit form for the second order correction term in the studentized case is

$$q(x) = \frac{\rho_3}{6}(3x - 2x^3), \tag{2}$$

where ρ_3 is the standardized skewness of P. Multiplying the expansion (1) by x or x^3 and integrating implies that $E(P) = -\rho_3/(2\sqrt{n})$ and $\text{skew}(P) = -2\rho_3/\sqrt{n}$ with error $O(n^{-1})$. These results are less easily obtainable in other ways (see Section 5).

This paper concerns second order accurate inference on a location parameter θ. By second order accurate I mean equivalent in accuracy to estimating the second order term $q(x)$ in an Edgeworth expansion. A concrete statistical consequence of second order accuracy is that the coverage error of one-sided confidence intervals for θ will have error $O(n^{-1})$ rather than $O(n^{-1/2})$. By a location parameter I mean a functional satisfying the implicit equation

$$E_F\{\psi(X - \theta)\} = 0, \tag{3}$$

where it is assumed that $\theta(F)$ is uniquely defined by (3) for each F. For this it is sufficient that ψ is non-decreasing. This formulation includes the mean, the median and various robust measures of location such as trimmed means and was introduced by Huber (1964). The distribution free or 'bootstrap' estimator of $\theta(F)$ is given by the solution of

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}) = 0 \tag{4}$$

and is called an *M-estimator*. This class of estimators includes the sample mean, median and trimmed mean as well as several robust estimators of location. For a given sample there may not be a unique solution of (4) even though we suppose a unique solution of (3). In this case the multiple solutions are often substituted in some convex objective function and the minimizing value used. Asymptotically there is only one solution of (4).

The aim of this paper is to look at various pivots that might be constructed for θ and to examine the size of the second order correction terms. Two of the pivots are based directly on (4), one is based directly on $\hat{\theta}$ and its standard error and a fourth on an entirely different concept called empirical likelihood. Of course standard normality of a pivot is only one aspect of its statistical performance. A good pivot $P(X_1, \ldots, X_n, t)$ should also be monotonic in t and have an extreme distribution when t does not equal $\theta(F)$. Different pivots are hardly ever comparable in terms of their non-null distributions and this issue will not be addressed. The monotonicity properties of the different pivots will be briefly noted in passing.

2. Some Approximate Inference Methods

One method of correcting a pivot for non-normality of its distribution is to estimate terms in a valid Edgeworth expansion and to use the estimated Edgeworth approximation instead of the standard normal distribution. Essentially the same result may be achieved by using the bootstrap technique. In essence the idea of the bootstrap is to replace F by the empirical distribution function

$$\hat{F}(x) = \#(X_i \leq x)/n, \qquad x \in \Re,$$

in estimating any functional of F. This generates distribution free estimators $\theta(\hat{F})$ such as (4) as well as estimators of the entire distribution function $H(F)$ of a pivot P. In many cases it is mathematically difficult to derive the bootstrap distribution $\hat{H} = H(\hat{F})$ and simulation is employed. Hall (1992) has demonstrated that using the bootstrap distribution $\hat{H}$ to generate confidence intervals from a pivot P gives the same theoretical coverage accuracy as estimating the second order term in the Edgeworth expansion of P, at least when P is a studentized pivot for a smooth functions of means. His smooth function of means model does not include most M-estimators however.

An entirely different approach to distribution free inference on a functional $\theta(F)$ is via *empirical likelihood* as introduced by Owen (1988, 1990). For the smooth function of means model, DiCiccio and Romano (1989) and DiCiccio, Hall and Romano (1989) developed various expansions for empirical likelihood and found that it leads to second order correct inference only after a location adjustment. While this is still true in multidimensions Hall (1990) found that empirical likelihood approximates inference based on a pivot which is not the usual student-t.

For an arbitrary functional $\theta(F)$, the empirical likelihood function based on data $X_1, \ldots, X_n$ is

$$\hat{L}(t) = \sup\left\{\prod_{i=1}^{n} p_i, \quad \theta(F_p) = t\right\}, \tag{5}$$

where the supremum is taken over all probability vectors $p = (p_1, \ldots, p_n)$ that put mass p_i at the observed point x_i and F_p is the corresponding discrete distribution function. When each $p_i = 1/n$ then F_p becomes the empirical distribution function $\hat{F}$. The maximum value $1/n^n$ of $\hat{L}(t)$ occurs at the bootstrap estimator $\theta(\hat{F})$. The best way to think of $\hat{L}$ is as the ordinary non-parametric likelihood for F but with $\theta(F)$ held fixed and profiled over all other aspects of F. The empirical log-likelihood ratio is defined as

$$\hat{Q}(t) = -2\log(n^n \hat{L}(t)) = -\sup\Big\{2\sum_{i=1}^{n}\log(np_i), \quad \theta(F_p) = t\Big\} \tag{6}$$

and the minimum value of this function is zero when $t = \hat{\theta}$. For the case of the location functional (3), Owen (1988) used calculus of variations to show that

$$\hat{Q}(t) = 2\sum_{i=1}^{n}\log(1 + \lambda\psi(X_i - t)), \tag{7}$$

where the multiplier $\lambda(t)$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\psi(X_i - t)}{1 + \lambda\psi(X_i - t)} = 0. \tag{8}$$

This must almost always be solved iteratively starting at $\lambda(\hat{\theta}) = 0$ which follows from (4). Owen further showed that comparing $\hat{Q}$ to the χ^2_1 distribution leads to first order correct inference.

3. Second Order Expansion of Empirical Likelihood

In this section an expansion of $\hat{Q}(t)$ is derived in an $O_p(n^{-1/2})$ neighbourhood of $\hat{\theta}$. Inferential accuracy will turn out to depend on the joint moments of the random variables

$$\xi_{1j}(X,t) = \psi^j(X - t), \quad \xi_{2k}(X,t) = \psi^{(k)}(X - t),$$

where the bracketted superscript denotes repeated derivative and we take $\psi^{(0)}$ to be identically 1. Define

$$M_{jk}(t) = E\{\xi_{1j}(X,t)\xi_{2k}(X,t)\}$$

assuming they exist. Provided $E(|\xi_{1j}\xi_{2k}|) < \infty$ these moments may be consistently estimated by

$$\tilde{M}_{jk}(t) = \frac{1}{n}\sum_{i=1}^{n}\psi^j(X_i - t)\psi^{(k)}(X_i - t). \tag{9}$$

For ease of notation denote M_{j0} by K_j and M_{0k} by D_k. The argument t will often be suppressed both in moments and estimators. Estimators $\tilde{M}_{jk}$ evaluated at $t = \hat{\theta}$ will be denoted $\hat{M}_{jk}$ and estimate $M_{jk}(\theta)$ under further regularity conditions to be specified.

We need to assume a central limit theorem for the random variables $\tilde{K}_j(t)$ which are averages of identical independent variables $\psi^j(X_i - t)$. For this it is certainly sufficient that the variance function $V_j(t) = K_{2j}(t) - K_j(t)^2$ exists. In this case write

$$\tilde{K}_j(t) = K_j(t) + \frac{Z_j(t)V_j^{1/2}(t)}{\sqrt{n}} + O_p(n^{-1}), \tag{10}$$

where $Z_j(t)$ is standard normal. We will also need to expand the estimators $\tilde{M}_{jk}(t)$ in a neighbourhood of the estimator $\hat{\theta}$. Assuming Condition (C) of Section 4 we easily derive the relations

$$\begin{aligned}\tilde{K}_1(t) &= (\hat{\theta} - t)\hat{D}_1 + (\hat{\theta} - t)^2\hat{D}_2/2 + O_p(n^{-3/2}),\\ \tilde{K}_2(t) &= \hat{K}_2 + 2(\hat{\theta} - t)\hat{M}_{11} + O_p(n^{-1}),\end{aligned} \tag{11}$$

for $t = \hat{\theta} + O_p(n^{-1/2})$, in particular at $t = \theta$ if $\hat{\theta}$ is consistent. Finally, suppose that $K_2(\theta) = D_1(\theta)$ which, provided neither equals zero, may be arranged simply by multiplying the function ψ by an appropriate function of θ. This is called Condition (A) in Section 4 and is analogous to assuming $\sigma = 1$ in case of the mean.

In order to expand $\hat{Q}(\hat{\theta}+\delta)$ about $\hat{Q}(\hat{\theta}) = 0$ for small δ we expand the solution $\lambda(t)$ of (8) about $\hat{\theta}$. The denominator of the summand may be expanded in a power series for small λ giving

$$0 = \tilde{K}_1(t) - \lambda\tilde{K}_2(t) + O(\lambda(t)^2).$$

At $t = \hat{\theta} + \delta$ and assuming Condition (C), we may substitute $\tilde{K}_1(t) = -\delta\hat{D}_1 + O(\delta^2)$ to obtain

$$\hat{K}_2\lambda = -\delta\hat{D}_1 + O(\delta^2)$$

using $\lambda = O(\delta)$. Under Condition (D) and the further assumption that $\hat{\theta}$ is consistent, $\hat{D}_1/\hat{K}_2 = 1 + O_p(n^{-1/2})$ and so, taking $\delta = O_p(n^{-1/2})$, we obtain $\lambda(t) = \hat{\theta} - t + O_p(n^{-1})$. To obtain a more refined expansion put $\lambda(t) = \hat{\theta} - t + \delta$ where δ is now $O_p(n^{-1})$. The other statistic which enters the expansion of the empirical likelihood beside $\hat{\theta}$ is

$$R = \hat{K}_2 - \hat{D}_1 = \sum_{i=1}^{n} \psi^2(X_i - \hat{\theta})/n - \sum_{i=1}^{n} \psi'(X_i - \hat{\theta})/n$$

which will be $O_p(n^{-1/2})$. We now expand (7) up to the cubic term and use the two relations in (11) to obtain

$$\begin{aligned}\hat{Q}(t) &\approx 2n\lambda\tilde{K}_1 - n\lambda^2\tilde{K}_2 + \frac{2n\lambda^3}{3}\tilde{K}_3 \\ &= n(\hat{\theta}-t)^2(\hat{D}_1 - R) + n(\hat{\theta}-t)^3\left(\hat{D}_2 - 2\hat{M}_{11} + \frac{2\tilde{K}_3}{3}\right) \qquad (12)\end{aligned}$$

with error $O_p(n^{-1})$. The regularity conditions required are consistency of $\hat{\theta}$ and conditions (A,C,D) of Section 4. Notice that the terms involving δ disappear to this order. Also note that, in the second $O_p(n^{-1/2})$ term, either the actual values D_2, K_3, M_{11} or sample estimates may be used to the same order of accuracy. In the special case that ψ is the identity function, $D_1 = \hat{D}_1 = 1$, $D_2 = \hat{D}_2 = 0$, $M_{11} = \hat{M}_{11} = 0$, K_3 is the skewness and R may be written as $\Sigma X^2/n - 1$ with error $O_p(n^{-1})$. Then the empirical likelihood is approximately

$$n(\bar{X}-\mu)^2(1-R) + \frac{2\rho_3}{3}n(\bar{X}-\mu)^3 + O_p(n^{-1})$$

in agreement with DiCiccio and Romano ((1989), Equation (3.6)).

4. Student-Like Pivots

For the remainder of the paper assume that the estimator defined by (4) is asymptotically normal with mean θ and variance $K_2/(D_1^2 n)$. Sufficient conditions for this are given by Huber (1964) both for continuous and discontinuous functions ψ. This implies that $\hat{\theta}$ is consistent for θ from which it differs by $O_p(n^{-1/2})$. For estimating means the classical pivot is the student-t or Wald statistic. In the context of M-estimators, there are at least three natural analogues, namely

$$\begin{aligned} P1(t) &= \frac{\sqrt{n}(\hat{\theta}-t)}{\hat{\sigma}(\hat{\theta})} = \frac{\sqrt{n}\hat{D}_1(\hat{\theta}-t)}{\hat{K}_2^{1/2}}, \\ P2(t) &= \frac{\sum_{i=1}^n \psi(X_i - t)}{\sqrt{\sum_{i=1}^n \psi^2(X_i-\hat{\theta})}} = \frac{\sqrt{n}\tilde{K}_1}{\hat{K}_2^{1/2}}, \\ P3(t) &= \frac{\sum_{i=1}^n \psi(X_i - t)}{\sqrt{\sum_{i=1}^n \psi^2(X_i - t)}} = \frac{\sqrt{n}\tilde{K}_1}{\tilde{K}_2^{1/2}}. \qquad (13)\end{aligned}$$

Under mild conditions each of these pivots has asymptotic standard normal distribution. Note that while $P1$ is the simplest function and $P3$ the most complex, it turns out that $P3$ is the simplest to work with theoretically and $P1$ the most complex. Boos (1980) suggested $P2$ when the ψ function is monotonic, which excludes many robust estimators and compared $P2$ to $P1$ in power. To first order the two pivots are equivalent but simulation evidence suggests that $P2$ gives intervals with slightly narrower average width. While $P1$ is always a linear function of t, when ψ is a redescending function $P2$ and $P3$ may both be non-monotonic and may indeed have multiple zeros. I do not pursue monotonicity issues here but note that non-monotonicity complicates but does not invalidate the use of a pivot.

If K_2 were known then instead of $P2$ or $P3$ one could use the unstudentized pivot $P = \sqrt{n}\tilde{K}_1/K_2^{1/2}$ which differs from $Z_1(t)$ in (10) by $O_p(n^{-1/2})$. Apparently, $P3$ is more similar to P than are either $P2$ or $P1$ and this is why it is theoretically easier to work with. The properties of the three pivots are studied by expanding $P1$ about $P2$, $P2$ about $P3$ and $P3$ about P the last of whose null cumulants are simply given by

$$E(P) = 0,\ \mathrm{Var}(P) = 1,\ \mathrm{Skew}(P) = K_3/(\sqrt{n}K_2^{3/2}). \tag{14}$$

Since P is a standardized sum of identical independent variables $\psi(X_i - t)$ its distribution is asymptotically standard normal provided K_2 exists and a valid Edgeworth expansion exists if K_3 exists. These conditions are stronger than necessary (see for instance Shiryayev (1984, p.326)).

I next list some further regularity conditions to be referred to where needed. These fall into two categories: (i) statistical regularity conditions for consistency and normality of estimators of the moments $M_{jk}(t)$, (ii) mathematical regularity conditions on estimators $\tilde{M}_{jk}(t)$ as functions of t.

(A) $0 < D_1 = K_2 < \infty$,
(B) K_4 exists,
(C) the second derivatives of $\tilde{K}_1$ and $\tilde{K}_2$ exist and are bounded in a neighbourhood of $t = \hat{\theta}$,
(D) D_1, K_2 are continuous at $t = \theta$ and $\tilde{D}_1$, $\tilde{K}_2$ are continuous at $t = \hat{\theta}$,
(E) D_2, M_{11} are continuous at $t = \theta$ and $\tilde{D}_2$, $\tilde{M}_{11}$ are continuous at $t = \hat{\theta}$.

Theorem 1. *If conditions* (A) *and* (B) *hold, then, for* $t = \hat{\theta} + O_p(n^{-1/2})$

$$P3(t) = P(t) - \frac{\sqrt{n}\tilde{K}_1(\tilde{K}_2 - K_2)}{2K_2^{3/2}} + O_p(n^{-1}). \tag{15}$$

Proof. If K_4 exists, then, since $\tilde{K}_2$ is an identical independent average we have the stochastic expansion (10) which becomes

$$\tilde{K}_2 = K_2 + \frac{Z_2(t)V_2^{1/2}}{\sqrt{n}} + O_p(n^{-1}).$$

If $K_2 > 0$ then $\tilde{K}_2 = O_p(1)$ and this may be formally rearranged to give

$$\tilde{K}_2^{-1/2} = K_2^{-1/2} - \frac{\tilde{K}_2 - K_2}{2K_2^{3/2}} + O_p(n^{-1}).$$

Multiplying by $\sqrt{n}\tilde{K}_1$ gives (15).

Theorem 2. *If conditions* (A, C, D) *hold, then, for* $t = \hat{\theta} + O_p(n^{-1/2})$

$$P2(t) = P3(t) + \frac{P1(t)P2(t)\tilde{M}_{11}}{\hat{D}_1\hat{K}_2^{1/2}\sqrt{n}} + O_p(n^{-1}). \tag{16}$$

Proof. By (C) we may expand $\tilde{K}_2(t)$ about $t = \hat{\theta}$ in a Taylor's series which gives the second relation in (11). Since $K_2 > 0$ by (A), $\tilde{K}_2$ is $O_p(1)$ and under (D) $\hat{K}_2$ is consistent for $K_2(\theta) > 0$. Thus $\hat{K}_2 = O_p(1)$ and so the expansion for $\tilde{K}_2$ can be rearranged as

$$\tilde{K}_2^{-1/2} = \hat{K}_2^{-1/2} - \frac{(\hat{\theta} - t)\hat{M}_{11}}{\hat{K}_2^{3/2}} + O_p(n^{-1}).$$

Multiplying both sides by $\sqrt{n}\tilde{K}_1$ and using the definitions of $P1$ and $P2$ gives (16).

Theorem 3. *If conditions* (A, C, D) *hold, then, for* $t = \hat{\theta} + O_p(n^{-1/2})$

$$P1(t) = P2(t) - \frac{P1(t)^2\hat{K}_2^{1/2}\hat{D}_2}{2\hat{D}_1^2\sqrt{n}} + O_p(n^{-1}). \tag{17}$$

Proof. Assuming (C), $\tilde{K}_1$ may be expanded about $t = \hat{\theta}$ in a Taylor's series which is the first relation of (11) and rearranges as

$$\hat{D}_1(\hat{\theta} - t) = \tilde{K}_1(t) - (\hat{\theta} - t)^2\hat{D}_2/2 + O_p(n^{-3/2}).$$

Under (A) $\tilde{K}_2$ is consistent for $K_2 > 0$ and under (D) this implies that $\hat{K}_2$ is consistent for $K_2(\theta)$. Multiplying by the $O_p(\sqrt{n})$ quantity $\sqrt{n}\hat{K}_2^{-1/2}$ and using the definitions of $P1$ and $P2$ gives (17) with the stated error.

Theorems 2 and 3 show that to order $O(n^{-1})$ the three pivots are functionally related. Thus inferences based on the three pivots will be second order equivalent provided second order accurate approximations to their distribution are used. Such an approximation would be the second order Edgeworth expansion in (1) with the polynomial $q(x)$ estimated. A valid Edgeworth expansion for P exists if

$M_3 < \infty$ and $\psi(X-\theta)$ is not discrete. This expansion transforms to an expansion for $P1 - P3$ under the conditions of the above theorems. In the next section we investigate the size of the second order adjustment by deriving expressions for the bias, variance and skewness of the three pivots.

5. Cumulants of the Pivots

Under Condition (D) we may replace estimators $\hat{K}_2$ and $\hat{D}_1$ by K_2 with error $O_p(n^{-1/2})$, and so the $O_p(n^{-1/2})$ terms in Theorems $1-3$ simplify slightly. Further, since all the pivots differ by $O_p(n^{-1/2})$ the term $P1(t)P2(t)$ in (16) can be replaced by $P3(t)^2$ and the term $P1(t)^2$ in (17) by $P2(t)^2$. Under (E) the estimators of M_{11} and D_2 can be replaced by their asymptotic values. Each of the expansions in Theorems $1-3$ are of the same basic form, namely

$$P_i = P_j + \frac{X}{\sqrt{n}} + O_p(n^{-1}), \tag{18}$$

where $X = O_p(1)$. In the sequel we compute approximations to cumulants of the pivots at $t = \theta$, i.e. their 'null' cumulants. It will turn out that all the pivots have null mean, variance and skewness of orders $O(n^{-1/2}), 1 + O(n^{-1}), O(n^{-1/2})$. Then, collecting terms of the same order in (18) it follows that

$$\begin{aligned} E(P_i) &= E(P_j) + \frac{E(X)}{\sqrt{n}} + O(n^{-1}), \\ \mathrm{Var}(P_i) &= \mathrm{Var}(P_j) + \frac{2\mathrm{Cov}(P_j, X)}{\sqrt{n}} + O(n^{-1}), \\ \mathrm{Skew}(P_i) &= \mathrm{Skew}(P_j) + \frac{3\mathrm{Cov}(P_j^2, X)}{\sqrt{n}} + O(n^{-1}). \end{aligned} \tag{19}$$

We first use these relations to derive cumulants of $P3(\theta)$ from the known cumulants of $P(\theta)$ in (14). In (15) we have

$$X = -\frac{n\tilde{K}_1(\tilde{K}_2 - K_2)}{2K_2^{3/2}}$$

and since $E(\tilde{K}_1(t)) = 0$ at $t = \theta$ it is easy to show that $E(X) = -K_3/(2K_2^{3/2})$. Next the variance of $P(\theta)$ is exactly 1 and

$$\begin{aligned} \mathrm{Cov}(P, X) &= \mathrm{Cov}\left(Z_1 + O_p(n^{-1/2}), -\frac{Z_1 Z_2 V_2^{1/2}}{2K_2} + O_p(n^{-1/2})\right) \\ &= -\frac{V_2^{1/2}}{2K_2}\mathrm{Cov}(Z_1, Z_1 Z_2) + O(n^{-1/2}). \end{aligned}$$

Because of the oddness of $Z_1^2 Z_2$ we conclude that $\text{Cov}(P, X) = O(n^{-1/2})$, and so $\text{Var}(P3) = 1 + O(n^{-1})$. For the skewness of $P3$ we require $\text{Cov}(P^2, X)$. The difficult term is

$$P^2 X = -\frac{n^2 \tilde{K}_1^3 (\tilde{K}_2 - K_2)}{2K_2^{5/2}}$$

and the expectation of the product of four sums is dominated by the term $3\sum_{i \neq j} \psi_i^2 \psi_j (\psi_j^2 - K_2)$ each term of which has expectation $K_2 K_3$. Substituting this and $E(X)$ into the skewness expression in (19) gives

$$\text{Skew}(P3) = -\frac{2K_3}{K_2^{3/2}\sqrt{n}}$$

which agrees with the skewness calculation given after (2). Next we obtain cumulants of $P2$ from those of $P3$. In (16) we have $X = (P3)^2 M_{11}/K_2^{3/2}$ and we quickly obtain

$$E(P2) = \frac{2M_{11} - K_3}{2K_2^{3/2}\sqrt{n}}.$$

It again turn outs that $\text{Cov}(P3, X) = O(n^{-1/2})$ because of the oddness of a polynomial and so again $\text{Var}(P2) = 1 + O(n^{-1})$. The skewness of $P2$ is

$$\text{Skew}(P3) + \frac{3M_{11}}{K_2^{3/2}\sqrt{n}} \text{Cov}((P3)^2, (P3)^2) = \frac{6M_{11} - 2K_3}{K_2^{3/2}\sqrt{n}}$$

with error $O(n^{-1})$, and using $\text{Cov}((P3)^2, (P3)^2) = 2 + O(n^{-1/2})$. Lastly the cumulants of $P1$ are obtained from those of $P2$. The mean is

$$E(P2) - \frac{D_2 E(P2^2)}{2D_1 K_2^{1/2}} = \frac{2M_{11} - K_3 - D_2}{2K_2^{3/2}\sqrt{n}}.$$

We again find that $\text{Var}(P1) = 1 + O(n^{-1})$ and the skewness is

$$\text{Skew}(P2) - \frac{3D_2}{2D_1 K_2^{1/2}\sqrt{n}} \text{Cov}((P2)^2, (P2)^2) = \frac{6M_{11} - 2K_3 - 3D_2}{K_2^{3/2}\sqrt{n}}$$

with error $O(n^{-1})$, and using $\text{Cov}((P2)^2, (P2)^2) = 2 + O(n^{-1/2})$.

6. The Empirical Likelihood Pivot

Using the definitions of $P1(t)$ and R one can show that

$$P1(t)^2 = n(\hat{\theta} - t)^2 (\hat{D}_1 - R) + O_p(n^{-1})$$

which is the leading term in expansion (12), which hence becomes

$$\hat{Q}(t) = P1(t)^2 \left\{ 1 + \frac{P(t)}{\sqrt{n}} \left(\frac{D_2 - 2M_{11} + 2K_3/3}{K_2^{3/2}} \right) \right\}.$$

The accuracy of the χ_1^2 approximation to the distribution of the empirical likelihood can be investigated by looking at the cumulants of the signed square root statistic $P4(t) = \text{sign}(t)\sqrt{\hat{Q}(t)}$ which will be approximately standard normal at $t = \theta$. Using the relations of the three pivots in Theorems $1 - 3$ we obtain

$$P4(t) = P3(t) + \frac{P(t)^2 K_3}{3K_2^{3/2}}. \tag{20}$$

Using the same techniques as were used in the last section we find that

$$E(P4) = -\frac{K_3}{6K_2^{3/2}} + O(n^{-1}), \ \text{Var}(P4) = 1 + O(n^{-1})$$

and the skewness turns out to vanish to order $n^{-1/2}$. Thus the empirical likelihood pivot has standard variance and skewness to second order but there is an $O(n^{-1/2})$ bias which we note is three times smaller than the bias of $P3$. This indicates that using $P4$ as if it were standard normal, or equivalently comparing $\hat{Q}(t)$ to the χ_1^2 distribution, would gives results very close to second order accurate.

These results are summarized in Table 1, which displays the theoretical simplicity of $P4$ over $P3$ over $P2$ over $P1$. As a numerical check on these results I generated 30,000 random samples of size $n = 100$ from the exponential distribution with mean 1. With the downweighting function $\psi(x) = x \exp\{-|x|\}$, the location parameter (3) is $\theta = 1/\sqrt{2}$ and

$$D_1 = 0.3486, \ D_2 = 0.2052, \ K_2 = 0.0715, \ K_3 = 0.0013, \ M_{11} = -.0255.$$

Table 1. Second order coefficients for three pivots

	$2\sqrt{n}K_2^{3/2}E(P)$	$\sqrt{n}(\text{Var}(P) - 1)$	$\sqrt{n}K_2^{3/2}\text{Skew}(P)$
$P1$	$-K_3 + 2M_{11} - D_2$	0	$-2K_3 + 6M_{11} - 3D_2$
$P2$	$-K_3 + 2M_{11}$	0	$-2K_3 + 6M_{11}$
$P3$	$-K_3$	0	$-2K_3$
$P4$	$-K_3/3$	0	0

All coefficients in Table 1 require $D_1 = K_2$ which can be arranged by multiplying ψ by $c = D_1/K_2 = 4.925$. Let us denote the coefficients on this standardized scale by $*$ superscript. Then it simply follows that $D_j^* = cD_j$, $K_j^* = c^j K_j$ and for the present case

$$D_1^* = 1.694, \ D_2^* = 1.002, \ K_2^* = 1.694, \ K_3^* = 0.151, \ M_{11}^* = -.608.$$

Notice that since $K_3^*/K_2^{*3/2} = .068$ is small the mean and skewness of the pivots $P3, P4$ are almost zero and so CLT approximations to their distributions would be close to second order accurate. This is despite the extreme skewness of the exponential distribution. Table 2 compares the theoretical values given in Table 1 based on the simulated values for $P1, P2, P3$ with simulation standard errors in parentheses.

Table 2. Simulated versus theoretical moments

	Simulated (B=30,000)			Theoretical		
	$\sqrt{n}$mean	var	$\sqrt{n}$skew	$\sqrt{n}$mean	var	$\sqrt{n}$skew
$P1$	−.584(.06)	1.06(.01)	−2.93(.2)	−0.537	1.00	−3.15
$P2$	−.365(.06)	1.03(.01)	−1.69(.2)	−0.310	1.00	−1.79
$P3$	−.090(.06)	1.00(.01)	−0.08(.2)	−0.034	1.00	−0.14

7. Almost Second Order Correct Inference

The constants listed in Table 1 represent second order departure of the four pivots from standard normality and therefore the size of the errors incurred in applying the central limit theorem. In the case of the mean ($\psi(x) = x$) we have $D_2 = M_{11} = 0$ and so the first three pivots have identical bias and skewness coefficients. The empirical pivot has smaller bias and zero skewness to this order. On the other hand, it is a simple mathematical exercise to show that the coefficients K_3, M_{11} and D_2 all become smaller as F becomes more symmetric provided ψ is an odd function. In this case, central limit theorem inference based on any of the four pivots would be almost second order correct.

Table 3 lists the bias and skewness coefficients for the three pivotals $P1, P2, P3$. The left figure of each pair is the bias and the ith pair is for pivot Pi. When divided by $\sqrt{n}$ these figures are theoretical approximations to the actual bias and skewness. Four distributions F are considered ranging from slightly to very asymmetric. The location functional is defined by the downweighting function

$$\psi(x) = xe^{-|x|/k},$$

where k is chosen to equal $1, 2, 10, \infty$, the latter corresponding to the mean. The conclusions from the table are as follows: (i) the constants decrease from right to left i.e. as we move to more heavily downweighted location estimators, (ii) the effect in (i) seems to decrease with more asymmetric contamination, (iii) the absolute values of the coefficients seem to be smallest for $P3$ and largest for $P1$ and this pattern is clearest for heavy downweighting ($k = 1$). In the robustness literature 5% contamination and downweighting at around two standard deviations is a common benchmark. The corresponding cell of Table 3 shows that the coefficients for $P3(.004, .017)$ are around 20 times smaller than for $P1(.072, .425)$. Recall that the skewness coefficient for $P4$ is zero and the bias coefficient three times smaller than for $P3$.

To second order, the non-normality of $P3$ and $P4$ is governed by $K_3/K_2^{3/2}$ which is the standardized skewness of $\psi(X - \theta)$. Table 4 lists this coefficient for the same four distributions. The upper figure is the Huber estimator with truncation parameter k and the lower the redescending estimator defined above. Both estimators approach the mean for large k. As in Table 3 the main feature that stands out is that for even moderately downweighted estimators the constants involved are *much* smaller than for the mean.

Table 3. Non-normality coefficients for three pivots

	$k = 1$	$k = 2$	$k = 10$	$k = \infty$
$.99N(0,1) + .01\exp(3)$	.009 .056	−.013 −.074	−.085 0.361	−.468 −1.872
	.004 .025	−.008 −.042	.084 −.354	−.468 −1.872
	.000 .000	−.002 −.007	−.076 −.302	−.468 −1.872
$.95N(0,1) + .05\delta_3$	−.037 −.252	−.072 −.425	−.200 −.869	−.226 −.915
	−.005 −.054	−.048 −.280	−.192 −.823	−.226 −.915
	.012 .054	−.004 −.017	−.166 −.665	−.226 −.915
$.9N(0,1) + .1\chi_2^2$	.044 .248	−.007 −.055	−.262 −1.120	−.856 −3.427
	.021 .115	−.005 −.044	−.257 −1.093	−.856 −3.427
	.008 .033	−.007 .028	−.227 −.909	−.856 −3.427
$\exp(1)$	−.537 −3.126	−.526 −2.779	−.752 −3.213	−1.000 −4.000
	−.310 −1.773	−.410 −2.088	−.727 −3.065	−1.000 −4.000
	−.034 −.138	−.187 −.750	−.649 −2.599	−1.000 −4.000

Both M-estimators examined here have continuous and differentiable ψ-functions. If ψ is discontinuous at zero then (C) will fail as $\tilde{K}_1$ and $\tilde{K}_2$ will be discontinuous when t equals any of the sample values X_i. This invalidates the expansions in (11) and therefore the relations of Theorems 2 and 3, as well as Equation (20).

Why is K_3 small for robust estimators? Robust estimation is usually indicated where F is contaminated and likely highly asymmetric. In downweighting extreme observations we not only reduce the influence of outliers but virtually ensure that $\psi(X - \theta)$ is much more symmetric than $X - \theta$. The only pivots which utilize this in their own distributions are $P3$ and $P4$. The distribution of $\hat{\theta}$ on the other hand is not symmetrized to the same extent.

Finally, notice that in Table 4, there appears to be a value of k for which the coefficient $K_3/K_2^{3/2}$ vanishes. This suggests specifically choosing the ψ-function so that $\psi(X - \theta)$ has skewness as small as possible by empirically choosing a value of k. More generally, we might choose a function such that the sample skewness of $\psi(x_i - \hat{\theta})$ is zero. Provided that the skewness of $\hat{\psi}(X - \theta)$ is $O(n^{-1/2})$, ordinary central limit theorem inference using $P3$ will be second order correct achieved without any bootstrap resampling, without any analytic computation of moments or without non-linear optimization such as is necessary to compute the empirical likelihood. This choice also has a neat statistical interpretation, namely that the estimator identifies and downweights outliers differently on the two tails depending on how the data point stands out from that particular tail. This would seem a statistically sensible thing to do from the point of view of robustness.

Table 4. Standardized skewness for two robust estimators

	$k = .25$	$k = .5$	$k = .75$	$k = 1.0$	$k = 2.0$	$k = \infty$
$.99N(0,1) + .01\exp(3)$	$-.000$	$-.000$	.000	.001	.016	.936
	.001	.000	$-.001$	.000	.003	.936
$.95N(0,1) + .05\delta_3$	.001	.003	.007	.015	.148	.457
	.000	.003	$-.014$	$-.027$	.008	.457
$.9N(0,1) + .1\chi^2_2$	$-.001$	$-.003$	$-.002$	.001	.081	1.714
	.020	$-.004$	$-.007$	$-.016$	$-.014$	1.714
exp(1)	.011	.059	.191	.412	1.022	2.000
	$-.178$	0.113	$-.021$	.069	.375	2.000

References

Boos, D. D. (1980). A new method for constructing approximate confidence intervals from M estimates. *J. Amer. Statist. Assoc.* **75**, 142-145.

DiCiccio, T. J. and Romano, J. P. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika* **76**, 447-456.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika* **76**, 465-476.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Ann. Statist.* **18**, 121-140.

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

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Statist.* **35**, 73-101.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Shiryayev, A. N. (1984). *Probability.* Springer-Verlag, New York.

Department of Statistics, LaTrobe University, Bundoora, Victoria 3083, Australia.

(Received January 1992; accepted March 1994)