The Annals of Statistics
2003, Vol. 31, No. 4, 1154–1169

SADDLEPOINT APPROXIMATIONS AND TESTS BASED ON MULTIVARIATE *M*-ESTIMATES

BY J. ROBINSON, E. RONCHETTI AND G. A. YOUNG

University of Sydney, University of Geneva and University of Cambridge

We consider multidimensional M-functional parameters defined by expectations of score functions associated with multivariate M-estimators and tests for hypotheses concerning multidimensional smooth functions of these parameters. We propose a test statistic suggested by the exponent in the saddlepoint approximation to the density of the function of the M-estimates. This statistic is analogous to the log likelihood ratio in the parametric case. We show that this statistic is approximately distributed as a chi-squared variate and obtain a Lugannani–Rice style adjustment giving a relative error of order n^{-1}. We propose an empirical exponential likelihood statistic and consider a test based on this statistic. Finally we present numerical results for three examples including one in robust regression.

1. Introduction. Let $X_1, \ldots, X_n$ be an independent, identically distributed sample of random vectors from a distribution F with density f on the sample space $\mathcal{X}$. Define the M-functional $\theta(F)$ to satisfy

$$E\{\psi(X;\theta)\} = 0, \tag{1.1}$$

where ψ is assumed to be a smooth function from $\mathcal{X} \times \mathbb{R}^d$ to $\mathbb{R}^d$ and the expectation is taken with respect to F. Suppose we wish to test a hypothesis concerning parameters defined by a smooth transformation $\eta = g(\theta)$, to a space of dimension $d_1 \leq d$. Consider test statistics based on $g(T_n)$, where T_n is the M-estimate of θ given by the solution of

$$\sum_{i=1}^{n} \psi(X_i; T_n) = 0. \tag{1.2}$$

When $d_1 = 1$ we can simply base the test on $g(T_n)$ and calculate the observed significance level or p-value $p = P(g(T_n) \geq g(t_n))$, where t_n is the observed value of T_n. Saddlepoint approximations with relative error of order n^{-1} are available for this case; see, for example, Tingley and Field (1990), Daniels and Young (1991), Jing and Robinson (1994), Fan and Field (1995), Davison, Hinkley and Worton (1995) and Gatto and Ronchetti (1996). In this special case a one-sided test is possible. However, when $d_1 > 1$, a single summary statistic, $h(g(T_n))$ of dimension 1 is needed to obtain the test. In classical parametric cases quadratic

Received February 2002; revised September 2002.

[1]Supported in part at the University of Sydney by an ARC Institutional Grant.

AMS 2000 subject classifications. Primary 62F11, 62F05; secondary 62G09.

Key words and phrases. Bootstrap tests, composite hypothesis, nonparametric likelihood, relative error, smooth functions of M-estimators.

forms in the mean scores or pseudo-likelihood statistics are competitors. Tests of the kind considered here arise naturally in, for example, the context of multiple regression, where interest lies in testing a hypothesis concerning a sub-vector of the vector of regression parameters, with the remaining parameters including the scale as nuisance parameters.

We consider the case when the cumulant generating function of the vector of scores, defined by

$$K_\psi(\lambda;\theta) = \log E\{e^{\lambda^T \psi(X;\theta)}\}, \tag{1.3}$$

exists. Then, under the assumption of the existence of a density for the M-estimates, discussed in Section 2, we obtain a saddlepoint approximation to the density of $g(T_n)$ of the form

$$f_{g(T_n)}(y) = r_n e^{-nh(y)}\gamma(y)\big(1 + O(n^{-1})\big),$$

where

$$h(y) = \inf_{\{\theta : g(\theta)=y\}} \sup_{\lambda}\{-K_\psi(\lambda;\theta)\}. \tag{1.4}$$

Thus we propose the test statistic $h(g(T_n))$ and obtain the p-value

$$p = P\big(h(g(T_n)) \geq h(g(t_n))\big).$$

In the parametric case F is a known distribution from the class of distributions satisfying (1.1) and under the null hypothesis the choice of θ is restricted to the set $\Theta_0 = \{\theta : g(\theta) = \eta_0\}$. Theorem 2 shows that the statistic $h(g(T_n))$ is asymptotically pivotal, since the asymptotic distribution does not depend on the choice of θ in Θ_0.

Using a proof modelled on Barndorff-Nielsen and Cox (1984), we show that

$$p = \bar{Q}_{d_1}(n\hat{u}^2) + n^{-1}c_n\hat{u}^{d_1}e^{-n\hat{u}^2/2}\left[\frac{G(\hat{u}) - 1}{\hat{u}^2}\right] + \bar{Q}_{d_1}(n\hat{u}^2)O(1/n), \tag{1.5}$$

where $\hat{u} = \sqrt{2h(g(t_n))}$,

$$c_n = \frac{n^{d_1/2}}{2^{d_1/2-1}\Gamma(d_1/2)},$$

and $Q_{d_1} = 1 - \bar{Q}_{d_1}$ is the distribution function of a chi-squared variate with d_1 degrees of freedom and G is a function defined in Theorem 1. We show in the proof of Theorem 2 that the error here is relative uniformly for $\hat{u} < \varepsilon$ for some $\varepsilon > 0$. This result holds only for the case of the particular summary statistic $h(g(T_n))$ defined in (1.4). In general G requires a numerical integration over a sphere of dimension d_1, but a simple Monte Carlo approximation to any degree of

accuracy required can be readily obtained. In addition, we show that $(G(u)-1)/u^2$ is bounded for u bounded and so we obtain the simpler approximation

$$p = \bar{Q}_{d_1}(n\hat{u}^2)\big(1 + O\big((1+n\hat{u}^2)/n\big)\big). \tag{1.6}$$

This simpler form does not have small relative error in the large deviation region.

If the underlying distribution of the observations belongs to a full exponential model with score statistic $\psi(x;\theta) = x - \theta$, where θ is the mean parameter, then the statistic defined in (1.4) is the log-likelihood ratio statistic [see, e.g., Barndorff-Nielsen and Cox (1984)]. The same holds for curved exponential models. In general parametric models, even when T_n is the maximum likelihood estimator, this is not necessarily the case.

If the underlying distribution of the observations does not belong to the model but is assumed to lie in a neighborhood, robust tests should be used. In this case the statistic $h(g(T_n))$ extends the notion of log-likelihood ratio and the test based on this statistic is asymptotically equivalent to first order to robust versions of score and Wald tests discussed in Heritier and Ronchetti (1994). In particular, by an appropriate choice of the function ψ these tests have robustness of validity and robustness of efficiency in a neighborhood of the model. These first order properties are shared by the test based on $h(g(T_n))$. In addition, the adjusted chi-squared approximation to the p-value of the test based on $h(g(T_n))$ is here shown to have relative error of order n^{-1} under the model. We cannot expect this second order relative error property to be maintained in a general neighborhood of the model.

In Section 2 we consider the special case of testing the hypothesis $H_0 : \theta = \theta_0$ in $\mathbb{R}^d$ and we show, in Theorem 1, that a Lugannani–Rice style adjustment to the chi-squared approximation has relative error $O(n^{-1})$. In Section 3 we consider the more general hypothesis $H_0 : g(\theta) = \eta_0$ and obtain a similar result in Theorem 2. A proof of this more general result is notationally complex but requires the same lines of argument as those used in Theorem 1, which is therefore proved in detail.

If the distribution of the observations is completely unspecified, we can use an empirical exponential family to approximate the distribution of the observations by $\hat{F}_0$, a tilted empirical distribution satisfying the null hypothesis, and use this to give $\hat{h}(g(T_n))$, an empirical version of the test statistic. If we sample from $\hat{F}_0$ then this gives an empirical exponential likelihood version of the test. The saddlepoint approximation to this probability might be expected to hold.

In Section 4 we consider empirical exponential likelihood and approximate tests based on this, noting that an extension of the theorems should show that a simple bootstrap approximation to these should have the saddlepoint approximation from the theorems. Section 5 contains two examples illustrating the accuracy of the chi-squared approximation in a parametric setting and in the bootstrap setting of Section 3. Also, in Section 5 a numerical example in the case of robust

regression compares the distribution of the test statistic from Section 4 and that of other available robust test statistics with the distribution obtained by Monte Carlo resampling.

2. Simple hypothesis. Consider the simple hypothesis $H_0 : \theta = \theta_0$ in $\mathbb{R}^d$. We derive an approximation, with relative error $O(n^{-1})$, to the p-value

$$p = P_{H_0}\{h(T_n) \geq h(t_n)\},$$

of the test based on the statistic $h(T_n)$, where T_n is defined in (1.2), t_n is its observed value with

$$h(y) = \sup_{\lambda}\{-K_\psi(\lambda; y)\},$$

and K_ψ is defined in (1.3). We assume the following:

(A1): The density of T_n exists and has the saddlepoint approximation

$$f_{T_n}(t) = (2\pi/n)^{d/2} e^{nK_\psi(\lambda(t);t)} |B(t)||\Sigma(t)|^{-1/2}\big(1 + O(n^{-1})\big), \tag{2.1}$$

where $\lambda(t)$ is the saddlepoint satisfying

$$K'_\psi(\lambda; t) \equiv \frac{\partial}{\partial\lambda} K_\psi(\lambda; t) = 0, \tag{2.2}$$

and $|\cdot|$ denotes the determinant; further, writing $\lambda \equiv \lambda(t)$,

$$B(t) = e^{-K_\psi(\lambda;t)} E\big\{\dot\psi(X; t) e^{\lambda^T \psi(X;t)}\big\} \tag{2.3}$$

and

$$\Sigma(t) = e^{-K_\psi(\lambda;t)} E\big\{\psi(X; t)\psi^T(X; t) e^{\lambda^T \psi(X;t)}\big\},$$

and $\dot\psi(X; t) = \frac{\partial}{\partial t}\psi(X; t)$.

The saddlepoint approximation (2.1) was given in Field (1982) and has subsequently been considered by Skovgaard (1990), Jensen and Wood (1998) and Almudevar, Field and Robinson (2000). Conditions which imply (A1) and cover, in particular, the case when ψ is not differentiable are given in Almudevar, Field and Robinson (2000).

THEOREM 1. *Under assumption* (A1), p *is given by* (1.5) *and* (1.6), *with* $d_1 = d$, *where*

$$G(u) = \int_{S_d} \delta(u, s)\, ds = 1 + u^2 k(u) \tag{2.4}$$

for

$$\delta(u, s) = \frac{\Gamma(d/2)|B(y)||\Sigma(y)|^{-1/2} J_1(y) J_2(y)}{2\pi^{d/2} u^{d-1}}, \tag{2.5}$$

where, for any $y \in \mathbb{R}^d$, (r, s) *are the polar coordinates corresponding to* y, $r = \sqrt{(y^T y)}$ *is the radial component and* $s \in S_d$, *the* d*-dimensional sphere of unit radius,* $u = \sqrt{2h(y)}$, $J_1(y) = r^{d-1}$ *and* $J_2(y) = ru/(h'(y)^T y)$, $\hat{u} = \sqrt{2h(t_n)}$ *and* $k(\hat{u})$ *is bounded and the order terms are uniform for* $\hat{u} < \varepsilon$ *for some* $\varepsilon > 0$.

PROOF. Without loss of generality we assume $\theta_0 = 0$ and $h''(0) \equiv \frac{\partial^2}{\partial y\, \partial y^T} h(y)|_{y=0} = I$. Otherwise, transform $\psi(X_i; \theta)$ to

$$\tilde{\psi}(X_i; \tilde{\theta}) = \psi\big(X_i; h''(0)^{-1/2}(\theta - \theta_0)\big).$$

The proof follows by integrating (2.1) to get the p-value. Writing $h(y) = -K_\psi(\lambda(y); y)$, we have

$$p = \int_A \frac{e^{-nh(y)}}{(2\pi/n)^{d/2}} |B(y)||\Sigma(y)|^{-1/2}\big(1 + O(n^{-1})\big)\, dy,$$

where $A = \{y : h(y) \geq h(t_n)\}$. We may consider the order term to be uniform in y, since we can consider the approximation obtained by integrating over $A \cap B^c$, where $B = \{y : h(y) \geq h(t_n) + \varepsilon\}$ and $P(T_n \in B) = P(T_n \in A) O(e^{-n\varepsilon})$.

In order to integrate this to find p we perform two transformations, the first the polar transformation $y \to (r, s)$ and the second $(r, s) \to (u, s)$, where $u = \sqrt{2h(y)}$. The Jacobians of these transformations are respectively $J_1 = r^{d-1}$ and $J_2 = ru/(h'(y)^T y)$.

Following these transformations we have

$$p = \int_{\hat{u}}^{(\hat{u}^2 + 2\varepsilon)^{1/2}} c_n u^{d-1} e^{-nu^2/2} \left\{ \int_{S_d} \delta(u, s)\big(1 + O(n^{-1})\big)\, ds \right\} du. \tag{2.6}$$

Now expanding each term of $\delta(u, s)$ we have

$$|B(y)| = |B(0)|\big(1 + r\xi_1(s) + r^2 \gamma_1(r, s)\big) \tag{2.7}$$

and

$$|\Sigma(y)|^{-1/2} = |\Sigma(0)|^{-1/2}\big(1 + r\xi_2(s) + r^2 \gamma_2(r, s)\big), \tag{2.8}$$

where $\xi_1(s)$, $\xi_2(s)$ are linear combinations of the components of s, and γ_1 and γ_2 are uniformly bounded for r bounded. Also

$$u = \sqrt{2h(y)} = r\big(1 + r\rho(s) + r^2 \gamma_3(r, s)\big), \tag{2.9}$$

where $\rho(s)$ is a linear combination of terms of the form $s_i s_j s_k$ and γ_3 is uniformly bounded for r bounded. Combining (2.7)–(2.9) in (2.5) and using

$$|B(0)||\Sigma(0)|^{-1/2} = |h''(0)|^{1/2} = 1,$$

we have

$$\delta(u, s) = \big(1 + ub(s) + u^2 \gamma_4(u, s)\big)\Gamma(d/2)/2\pi^{d/2}, \tag{2.10}$$

where $b(s)$ is an odd function, $b(s) = -b(-s)$ and $\gamma_4(u, s)$ is uniformly bounded when u is bounded, since $h(0) = 0$, $h'(0) = 0$ and, by assumption, $h''(0) = I$. Hence the second equality in (2.4) follows and similarly

$$G'(u) = uk^*(u),$$

where $k(u)$ and $k^*(u)$ are bounded for u bounded. So

$$\begin{aligned} p &= \int_{\hat{u}}^{\sqrt{\hat{u}^2+2\varepsilon}} c_n u^{d-1} e^{-nu^2/2} G(u)\, du \,\bigl(1 + O(n^{-1})\bigr) \\ &= \int_{\hat{u}}^{\infty} c_n u^{d-1} e^{-nu^2/2}\, du \,\bigl(1 + O(n^{-1})\bigr) \\ &\quad + \frac{c_n}{n} \int_{\hat{u}}^{\sqrt{\hat{u}^2+2\varepsilon}} u^{d-2}\bigl(G(u) - 1\bigr)\frac{d}{du}\bigl[-e^{-nu^2/2}\bigr]\, du \,\bigl(1 + O(n^{-1})\bigr), \end{aligned} \tag{2.11}$$

and by integrating by parts,

$$\begin{aligned} p &= \left\{ \bar{Q}_d(n\hat{u}^2) + n^{-1} c_n \hat{u}^d e^{-n\hat{u}^2/2} \frac{G(\hat{u}) - 1}{\hat{u}^2} \right\} \\ &\quad + \frac{c_n}{n} \int_{\hat{u}}^{\sqrt{\hat{u}^2+2\varepsilon}} \bigl[(d-2)u^{d-3}\bigl(G(u) - 1\bigr) + u^{d-2} G'(u)\bigr] e^{-nu^2/2}\, du \\ &\qquad \times \bigl(1 + O(n^{-1})\bigr), \\ &= \bar{Q}_d(n\hat{u}^2) + n^{-1} c_n \hat{u}^d e^{-n\hat{u}^2/2} \frac{G(\hat{u}) - 1}{\hat{u}^2} + \bar{Q}_d(n\hat{u}^2) O(1/n). \end{aligned} \tag{2.12}$$

The simpler form is obtained immediately from (2.4). □

REMARK. The second term of (1.5) is very much like the second term in the Lugannani–Rice formula. When $\sqrt{n}\hat{u}$ is bounded this term is of order n^{-1}, but for $\hat{u}$ bounded, that is in the large deviation region, this term is not of order n^{-1} relative to the first term.

In the special case where $T_n = \bar{X}$, the assumptions of the theorem reduce to assuming the existence of a density for X and the existence of a cumulant generating function $K(\lambda) = \log E e^{\lambda^T X}$, with $K(\lambda) < C$ for $\|\lambda\|_\infty < a$, for some $0 < a < \infty$ and $0 < C < \infty$, where $\|\cdot\|_\infty$ denotes the sup norm.

It is possible to extend the result of Theorem 1 to the case when X_i are not identically distributed or when T_n is defined by the more general estimating equation

$$\sum_{i=1}^{n} \psi_i(X_i; T_n) = 0.$$

To do this we need to generalize the results of Field (1982) as in Section 4.5.c of Field and Ronchetti (1990).

The expansion (2.9) shows that $2nh(T_n)$ is asymptotically equivalent to first order to the Wald and the score tests based on the M-estimator T_n. In particular, these tests have the same influence function. Therefore, by appropriately choosing a bounded function ψ we can define a test which is asymptotically first order robust, that is its asymptotic level and asymptotic power remain stable when the distribution of the observations does not belong to the model but lies in a neighborhood of it.

3. Composite hypothesis. Consider now the composite hypothesis H_0: $g(\theta) = \eta_0$, for a smooth function g from $\mathbb{R}^d$ to $\mathbb{R}^{d_1}$. As in Section 2, we consider the approximation to the p-value

$$p = P_{H_0}\{h(g(T_n)) \geq h(g(t_n))\}$$

of the test based on the statistic $h(g(T_n))$. As before, T_n is the M-estimator satisfying (1.2), t_n is the observed value and now

$$h(y) = \inf_{\{t\,:\,g(t)=y\}} \{-K_\psi(\lambda(t); t)\}, \tag{3.1}$$

where K_ψ is defined by (1.3) and $\lambda(t)$ satisfies (2.2).

THEOREM 2. *Assume* (A1) *in Theorem* 1 *and*

(A2): *The transformation* $t \to (y = g(t), z = g_1(t))^T$, *for* g_1 *of dimension* $d - d_1$, *has continuous second derivatives and has nonzero Jacobian at the solution* t *of* (1.1).

Then p *is given by* (1.5) *and* (1.6), *where* (r, s) *are the polar coordinates corresponding to* y, $u = \sqrt{2h(y)}$, $\hat{u} = \sqrt{2h(g(t_n))}$ *and* $G(u)$ *is given by* (2.4) *with*

$$\delta(u, s) = \frac{\Gamma(d_1/2)|B(\tilde{t})||\Sigma(\tilde{t})|^{-1/2} J_0(\tilde{t}) J_1(y) J_2(y)}{2\pi^{d_1/2} u^{d_1-1} |L_{22}(y, \tilde{z})|^{1/2}}, \tag{3.2}$$

where $t(y, z)$ *is the inverse of the transformation in* (A2), $\tilde{t} = t(y, \tilde{z})$ *is such that* $h(y) = K_\psi(\lambda(\tilde{t}); \tilde{t})$, $L_{22}(y, z) = \partial^2 K_\psi(\lambda(t(y, z)); t(y, z))/\partial z^2$, $J_0(t)$ *is the Jacobian of the transformation* $t \to (y, z)$, $J_1(y) = r^{d_1-1}$ *and* $J_2(y) = ru/(h'(y)^T y)$ *and* $k(\hat{u})$ *is bounded and the order terms are uniform for* $\hat{u} < \varepsilon$ *for some* $\varepsilon > 0$.

PROOF. We first obtain, by Laplace's method, an approximation to the d_1-dimensional density of $g(T_n)$ as in Jing and Robinson (1994). The result is then obtained by the same techniques as those used in the proof of Theorem 1. We transform the density (2.1) of T_n to obtain the joint density of $g(T_n)$ and

$g_1(T_n)$. The marginal density of $g(T_n)$ is obtained by integrating out $g_1(T_n)$. Using Laplace's method, this is seen to have the form

$$f_{g(T_n)}(y) = (2\pi/n)^{d_1/2} e^{-nh(y)} \gamma(y)\big(1 + O(n^{-1})\big),$$

where $h(y)$ is as given by (3.1) and

$$\gamma(y) = \frac{|B(\tilde{t})||\Sigma(\tilde{t})|^{-1/2} J_0(\tilde{t})}{|L_{22}(y, \tilde{z})|^{1/2}}. \tag{3.3}$$

At this point we can apply the same arguments as used in the proof of Theorem 1 to approximate the p-value

$$p = \int_A r_n e^{-nh(y)} \gamma(y)\big(1 + O(n^{-1})\big)\, dy,$$

where $A = \{y : h(g(y)) \geq h(g(t_n))\}$. In order to obtain the expressions (2.10) and (2.4) we need to prove that $h'(\tilde{y}) = 0$ and $h''(\tilde{y})$ is positive definite, where $\tilde{y}$ is the unconstrained minimiser of $h(y)$.

Note that $h(\tilde{y}) = 0$. Let the Lagrangian be

$$L(\theta; \beta) = -K_\psi\big(\lambda(\theta); \theta\big) + \beta^T \big(g(\theta) - y\big).$$

Then

$$\begin{aligned}\frac{\partial}{\partial\theta} L(\theta; \beta) &= -K'_\psi\big(\lambda(\theta); \theta\big)\lambda'(\theta) - \dot{K}_\psi\big(\lambda(\theta); \theta\big) + g'(\theta)^T \beta \\ &= -B(\theta)\lambda(\theta) + g'(\theta)^T \beta,\end{aligned}$$

since $K'_\psi(\lambda(\theta); \theta) = 0$ and where $B(\theta)$ is defined by (2.3), $g'(\theta) = dg(\theta)/d\theta$, $\lambda'(\theta) = d\lambda(\theta)/d\theta$ and

$$\dot{K}_\psi(\lambda; \theta) = \frac{\partial}{\partial\theta} K_\psi(\lambda; \theta).$$

It follows that $h(y) = -K_\psi(\lambda(\theta), \theta)$, where $\theta \equiv \theta(y)$, $\lambda(\theta) \equiv \lambda(\theta(y))$ and $\beta \equiv \beta(y)$ satisfy the $2d + d_1$ constraints

$$g(\theta) = y, \tag{3.4}$$

$$K'_\psi\big(\lambda(\theta); \theta\big) = 0, \tag{3.5}$$

$$g'(\theta)^T \beta = B(\theta)\lambda(\theta). \tag{3.6}$$

Writing $\lambda' = \lambda'(\theta)$ and $\theta' = d\theta(y)/dy$, we have

$$\begin{aligned}h'(y) &= -(\theta')^T (\lambda')^T K'_\psi\big(\lambda(\theta); \theta\big) - (\theta')^T \dot{K}_\psi\big(\lambda(\theta); \theta\big) \\ &= -(\theta')^T \dot{K}_\psi\big(\lambda(\theta); \theta\big)\end{aligned}$$

by (3.5).

Letting $\tilde{\theta} = \theta(\tilde{y})$, $\tilde{\theta}' = \theta'(\tilde{y})$, $\tilde{\lambda} = \lambda(\theta(\tilde{y}))$ and $\tilde{\beta} = \beta(\tilde{y})$, we have

$$h'(\tilde{y}) = -(\tilde{\theta}')^T B(\tilde{\theta})\tilde{\lambda} = -(\tilde{\theta}')^T g'(\tilde{\theta})^T \tilde{\beta} = -\tilde{\beta}$$

by (3.6) and on noting that, from (3.4), $g'(\tilde{\theta})\tilde{\theta}' = I$. Since $\tilde{y}$ is the unconstrained minimizer of $h(y)$, $\tilde{\beta} = 0$, $\tilde{\lambda} = 0$ and $h'(\tilde{y}) = 0$.

Noting that $K'_\psi(\lambda(\theta(y)); \theta(y)) = 0$ for any y and $\dot{K}_\psi(\tilde{\lambda}; \tilde{\theta}) = 0$, we obtain

$$h''(\tilde{y}) = (\tilde{\theta}')^T\big(-\ddot{K}_\psi(\tilde{\lambda}; \tilde{\theta})\big)(\tilde{\theta}'),$$

where $-\ddot{K}_\psi(\tilde{\lambda}; \tilde{\theta}) = -\ddot{K}_\psi(0; \tilde{\theta}) = \{E\dot{\psi}(X; \tilde{\theta})\}^T\{E\psi(X; \tilde{\theta})\psi^T(X; \tilde{\theta})\}^{-1} \times E\dot{\psi}(X; \tilde{\theta})$ the inverse of the asymptotic covariance matrix of the M-estimator T_n.

Since $h'(\tilde{y}) = 0$ and $h''(\tilde{y})$ is positive definite, the proof of the result is completed by arguments the same as those used in the proof of Theorem 1. □

The discussion following the proof of Theorem 1 also applies here.

4. Empirical exponential likelihood tests. In practice, the distribution F underlying the data sample $X_1, \ldots, X_n$ may be unknown. In these circumstances an empirical exponential likelihood may be used to provide empirical versions of the tests. To do this for a hypothesis $H : g(\theta) = \eta_0$, we need to consider the empirical exponential family and take

$$\hat{F}_0(x) = \sum_{i=1}^n e^{\beta(\eta_0)^T \psi(x_i; \theta(\eta_0))} \mathbb{1}\{x_i \le x\} \Big/ \sum_{i=1}^n e^{\beta(\eta_0)^T \psi(x_i; \theta(\eta_0))}, \tag{4.1}$$

where $\beta = \beta(\eta_0)$, $\theta = \theta(\eta_0)$ and the Lagrange multiplier $\gamma = \gamma(\eta_0)$, and the solutions of the equations

$$\kappa'(\beta; \theta) = 0, \tag{4.2}$$

$$g(\theta) = \eta_0, \tag{4.3}$$

$$\dot{\kappa}(\beta; \theta) = \gamma^T g'(\theta), \tag{4.4}$$

where

$$\kappa(\beta; \theta) = \log\left[\frac{1}{n}\sum_{i=1}^n e^{\beta^T \psi(x_i; \theta)}\right], \quad \kappa'(\beta; \theta) = \frac{\partial \kappa(\beta; \theta)}{\partial \beta}, \quad \dot{\kappa}(\beta; \theta) = \frac{\partial \kappa(\beta; \theta)}{\partial \theta},$$

are chosen to minimise the backward Kullback–Leibler distance between the empirical distribution and the tilted empirical distribution subject to

$$E_F \psi(X; \theta) = 0,$$

as in the $\mathbf{F}_2$ family of DiCiccio and Romano (1990). The construction of $\hat{F}_0$, as described by the solution of (4.2)–(4.4), is performed using standard numerical packages.

Now consider the cumulant generating function of $\psi(X^*;\theta)$ when X^* is drawn from $\hat{F}_0$:

$$K_\psi^\dagger(\lambda;\theta) = \log\left[\sum_{i=1}^n e^{\beta(\eta_0)^T\psi(x_i;\theta(\eta_0))+\lambda^T\psi(x_i;\theta)} \Big/ \sum_{i=1}^n e^{\beta(\eta_0)^T\psi(x_i;\theta(\eta_0))}\right]. \tag{4.5}$$

Then

$$\hat{h}(y) = \inf_{\{\theta:\, g(\theta)=y\}} \sup_\lambda\big[-K_\psi^\dagger(\lambda;\theta)\big] = -K_\psi^\dagger\big(\lambda(y);\vartheta(y)\big), \tag{4.6}$$

where $\lambda(y)$, $\vartheta(y)$ and the Lagrange multiplier $\delta(y)$ are obtained from

$$K_\psi^{\dagger\prime}\big(\lambda(y);\vartheta(y)\big) = 0, \tag{4.7}$$

$$g\big(\vartheta(y)\big) = y, \tag{4.8}$$

$$\dot{K}_\psi^\dagger\big(\lambda(y);\vartheta(y)\big) = \delta(y)^T g'(\vartheta(y)), \tag{4.9}$$

where K' and $\dot{K}$ are defined as for κ.

Now we obtain $\hat{h}(g(t_n))$ for t_n the solution of

$$\sum_{i=1}^n \psi(x_i;t_n) = 0$$

and $\hat{h}(g(T_n^*))$ for T_n^* the solution of

$$\sum_{i=1}^n \psi(X_i^*;T_n^*) = 0,$$

where $X_1^*,\dots,X_n^*$ is a sample from $\hat{F}_0$. The p-value based on this empirical exponential likelihood statistic is

$$p^* = P\big(\hat{h}(g(T_n^*)) \ge \hat{h}(g(t_n))\big). \tag{4.10}$$

Of course, to obtain a $1-\alpha$ confidence region for $g(\theta)$ we invert this procedure by finding the set of values of η_0 such that $p^* \ge \alpha$.

In the particular case when $\psi(x;\theta) = x-\theta$ and $g(\theta)=\theta$, we have $\theta_0=\eta_0$ and we solve

$$\kappa'\big(\beta(\theta_0),\theta_0\big) = 0. \tag{4.11}$$

Then

$$K_\psi^\dagger(\lambda,\theta) = \log\left[\frac{1}{n}\sum_{i=1}^n e^{\beta(\theta_0)^T(x_i-\theta_0)+\lambda^T(x_i-\theta)-\kappa(\beta(\theta_0),\theta_0)}\right]. \tag{4.12}$$

If $\lambda(\theta)$ is the solution of $K_\psi^{\dagger\prime}(\lambda,\theta)=0$, then we see, taking $\theta=\bar{x}$, that $\lambda(\bar{x}) = -\beta(\theta_0)$ and so

$$\hat{h}(\bar{x}) = -\beta(\theta_0)^T(\bar{x}-\theta_0) + \kappa\big(\beta(\theta_0);\theta_0\big). \tag{4.13}$$

We can obtain $\lambda(\bar{x}^*)$ and then show that

$$\hat{h}(\bar{x}^*) = -\kappa\big(\beta(\theta_0) + \lambda(\bar{x}^*); \bar{x}^*\big) + \kappa\big(\beta(\theta_0); \bar{x}^*\big). \tag{4.14}$$

The p-value p^* might be estimated by a Monte Carlo simulation. A series of B bootstrap samples is drawn from $\hat{F}_0$. If $T_n^{*(b)}$ denotes the M-estimator for the bth such sample, $b = 1, \ldots, B$, then p^* is approximated by $[1 + \sum_{b=1}^{B} I\{\hat{h}(g(T_n^{*(b)})) \geq \hat{h}(g(t_n))\}]/(B+1)$, where $I(\cdot)$ denotes the indicator function. Alternatively, the bootstrap p-value p^* might be approximated directly by the chi-squared distribution on d_1 degrees of freedom, instead of by a Monte Carlo simulation. Since in this case a density of T_n^* does not exist, Theorems 1 and 2 cannot be applied and we are unable to prove that the relative error of the chi-squared approximation is of order given in (1.6) using the methods of this paper. However, we might expect that the approximation will still hold in this case with the same relative errors if the original sample is drawn from a distribution satisfying the conditions of Theorems 1 and 2. This is demonstrated numerically in the second example in Section 5 in the cases $d = 3$ with $\psi(x) = x$ and X drawn from independent exponential distributions with $n = 20$ and $d = 3$.

Note that the statistic $\hat{h}(g(T_n^*))$ defined by (4.6) can be viewed as a nonparametric likelihood with exponential weights. This differs from Owen's (1988) empirical likelihood which in turn is equivalent to Mykland's (1995) dual likelihood. A comparison between these nonparametric likelihoods in the case of a simple hypothesis is provided in Monti and Ronchetti (1993).

5. Numerical examples. We give three examples. The first is a parametric case when we can get analytic results for h, the second is a simple example of Section 4 and the third is a robust regression of a more realistic nature. The first example demonstrates the accuracy of the approximation of Theorem 1 in a simple parametric case. Another example for this case can be found in Gatto (2000). The second shows that accurate approximations are also given by Theorem 1 in the empirical exponential likelihood case. We give more extensive simulations for the third case which compares the accuracy of the chi-square approximation and the bootstrap approximation of Section 4.

EXAMPLE 1. Consider the method of Section 2 with $d = 3$, $\psi(x; \theta) = x - \theta$ and assume that X is distributed as a vector of three independent exponential variables with means 1. Elementary calculations give

$$h(y) = \sum_{j=1}^{3} [(y_j - 1) - \log y_j].$$

For $n = 20$ in this case $n\bar{X}$ is distributed as a vector of three independent gamma

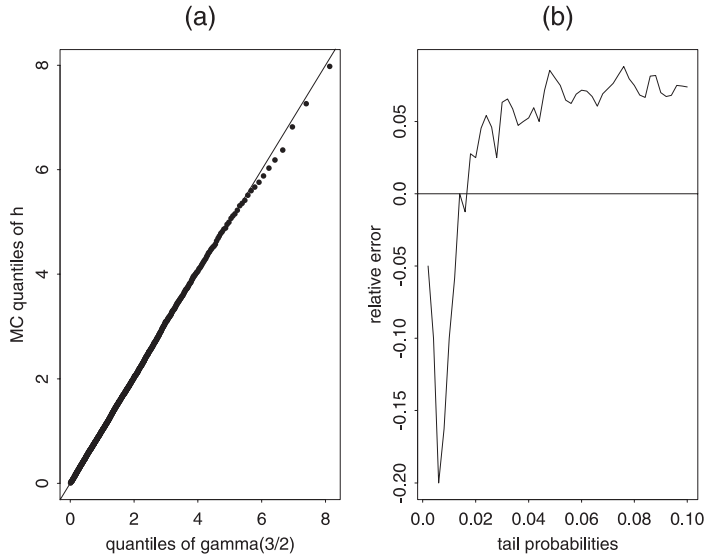

FIG. 1. (a) *Q–Q plot for* $h(\bar{X})$ *against theoretical quantiles of* χ^2_3; (b) *Relative errors of* χ^2_3 *approximation for* 10,000 *samples of size* 20 *from a vector of three independent exponential variables.*

variates with shape parameter n. So we can generate 10,000 Monte Carlo replicates of $2nh(\bar{X})$ and compare these to the approximating χ^2_3 distribution. Figure 1(a) gives a Q–Q plot of 10,000 Monte Carlo samples of $nh(\bar{X})$ with the theoretical quantiles (taking each 100th quantile in the plot) and Figure 1(b) plots the relative errors of the tail probabilities from 10,000 Monte Carlo trials compared to the χ^2_3 approximation. The relative error is $(P(2nh(\bar{X}) > v_\alpha) - \alpha)/\alpha$, where $P(\chi^2_3 > v_\alpha) = \alpha$, for $\alpha = 0.02, 0.04, \ldots, 0.1$. The approximation is very good except for the last 10 points where the Monte Carlo values are the cause of the variation.

EXAMPLE 2. Consider the method of Section 4 and draw a sample of 20 from a three-dimensional distribution of independent exponential variables with mean 1. From (4.13) obtain $\hat{h}(\bar{x})$ and for each of 10,000 bootstrap samples from $\hat{F}_0$ obtain $\hat{h}(\bar{x}^*)$ from (4.14). As in Example 1 we give a Q–Q plot in Figure 2(a) and we obtain an approximation to the relative error for tail areas of the chi-square approximation as $(P(2n\hat{h}(\bar{X}^*) \geq v_\alpha) - \alpha)/\alpha$, where $P(\chi^2_3 > v_\alpha) = \alpha$, for $\alpha = 0.02, 0.04, \ldots, 0.1$ and plot these in Figure 2(b). Again the approximation is very good.

EXAMPLE 3. Now we consider a more realistic example to illustrate the results of Section 4. Consider the model

$$y = x^T\theta + e, \tag{5.1}$$

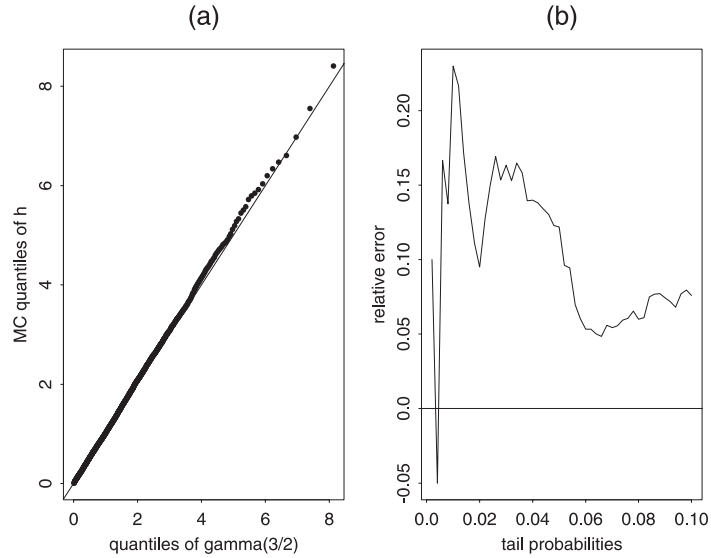

FIG. 2. (a) *Q–Q plot for* $\hat{h}(\bar{X}^*)$ *against theoretical quantiles of* χ^2_3; (b) *Relative errors of* χ^2_3 *approximation to tail probabilities of* $\hat{h}(\bar{X}^*)$ *from* 10,000 *bootstrap samples from a sample of size* 20 *from a vector of three independent exponentials.*

where $x = (1, x^{(2)}, x^{(3)})$ and $\theta = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$. We have taken the values of $x^{(2)}, x^{(3)}$ to be independent uniform on $(0, 1)$ and we consider the hypothesis $H_0 : \theta^{(2)} = \theta^{(3)} = 0$. The errors e are from the distribution $(1 - \varepsilon)\Phi(t) + \varepsilon\Phi(t/s)$ with settings of ε, s as in two of the settings in Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 379]; the other settings gave very similar results. The M-estimator of T_n satisfies

$$\sum_{i=1}^{n} \psi(y_i; T_n) = 0, \tag{5.2}$$

where

$$\psi(y; \theta) = \psi_c\left(\frac{y - x^T\theta}{\sigma}\right)x, \tag{5.3}$$

for $\psi_c(r) = \min\{c, \max(-c, r)\}$ and $c = 1.5$. The scale parameter σ is fixed at the value estimated by Huber's Proposal 2.

In addition to the empirical likelihood statistic $2n\hat{h}(g(T_n))$ of Section 4, we considered the Wald test statistic, the score test statistic and the likelihood ratio test statistic given in Welsh [(1996), Section 5.6.1]. We obtained 10,000 Monte Carlo samples of size $n = 20$. For the 25 values of $\alpha = 1/250, 2/250, \ldots, 25/250$, we obtained the proportion of times out of 10,000 that the statistic, S_n say, exceeded v_α, where $P(\chi^2_2 \geq v_\alpha) = \alpha$. Further, for each Monte Carlo sample we

obtained 299 bootstrap samples and calculated a bootstrap p-value, the proportion of the 299 bootstrap samples giving a value S_n^* of the statistic exceeding S_n. The bootstrap test of nominal level α rejects H_0 if the bootstrap p-value is less than α.

The results are plotted in Figure 3. In Figure 3(a) and (b) we plot the actual size against the nominal size for tests based on both the chi-square approximation and bootstrap approximation for $\hat{h}(g(T_n))$ and the three other statistics in the case $\varepsilon = 0$ and $s = 1$ and in (c) and (d) in the case $\varepsilon = 0.1$ and $s = 5$. It is clear that

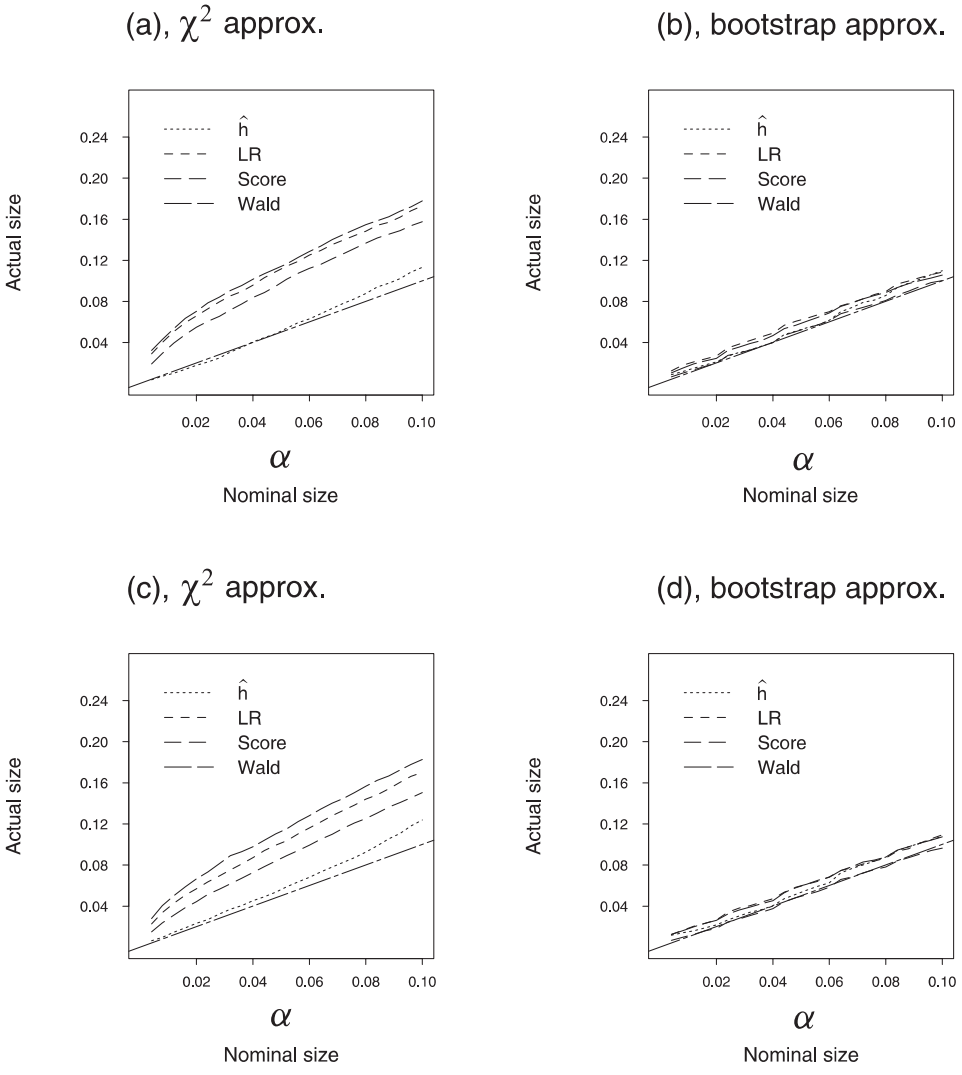

FIG. 3. *Actual size plotted against nominal size α for tests based on the statistic $\hat{h}(g(T_n))$ and the likelihood ratio, Wald and score tests.*

the chi-square approximation for $\hat{h}(g(T_n))$ is much better than the corresponding chi-square approximations for the other statistics. However, tests based on all the statistics are quite accurately approximated under the bootstrap and the bootstrap improves on the chi-square approximation in the case of the empirical exponential likelihood.

Acknowledgments. The authors are grateful for conversations with Ib Skovgaard on the order of the errors which led to the authors' development of the adjustment to the chi-squared approximation and to the referees for a careful reading of the paper which led to an improved version.

REFERENCES

ALMUDEVAR, A., FIELD, C. and ROBINSON, J. (2000). The density of multivariate M-estimates. *Ann. Statist.* **28** 275–297.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B*. **46** 483–495.

DANIELS, H. E. and YOUNG, G. A. (1991). Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika* **78** 169–179.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press.

DAVISON, A. C., HINKLEY, D. V. and WORTON, B. J. (1995). Accurate and efficient construction of bootstrap likelihoods. *Statistics and Computing* **5** 257–264.

DICICCIO, T. J. and ROMANO, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *Internat. Statist. Rev.* **58** 59–76.

FAN, R. Y. K. and FIELD, C. A. (1995). Approximations for marginal densities of M-estimators. *Canad. J. Statist.* **23** 185–197.

FIELD, C. A. (1982). Small sample asymptotic expansions for multivariate M-estimates. *Ann. Statist.* **10** 672–689.

FIELD, C. A. and RONCHETTI, E. (1990). *Small Sample Asymptotics*. IMS, Hayward, CA.

GATTO, R. (2000). Multivariate saddlepoint test for the wrapped normal model. *J. Statist. Comput. Simulation* **65** 271–285.

GATTO, R. and RONCHETTI, E. (1996). General saddlepoint approximations of marginal densities and tail probabilities. *J. Amer. Statist. Assoc.* **91** 666–673.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York.

HERITIER, S. and RONCHETTI, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89** 897–904.

JENSEN, J. L. and WOOD, A. T. A. (1998). Large deviation and other results for minimum contrast estimators. *Ann. Inst. Statist. Math.* **50** 673–695.

JING, B. and ROBINSON, J. (1994). Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Ann. Statist.* **22** 1115–1132.

MONTI, A. C. and RONCHETTI, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika* **80** 329–338.

MYKLAND, P. A. (1995). Dual likelihood. *Ann. Statist.* **23** 396–421.

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

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

TINGLEY, M. A. and FIELD, C. A. (1990). Small-sample confidence intervals. *J. Amer. Statist. Assoc.* **85** 427–434.

WELSH, A. H. (1996). *Aspects of Statistical Inference*. Wiley, New York.

J. ROBINSON
SCHOOL OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF SYDNEY
NEW SOUTH WALES 20006
AUSTRALIA
E-MAIL: johnr@maths.usyd.edu.au

E. RONCHETTI
DEPARTMENT OF ECONOMETRICS
UNIVERSITY OF GENEVA
BLV. PONT D'ARVE 40
CH-1211 GENEVA
SWITZERLAND
E-MAIL: Elvezio.Ronchetti@metri.unige.ch

G. A. YOUNG
STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: g.a.young@statslab.cam. ac.uk