Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 77 (2007) 104–110
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

A penalized version of the empirical likelihood ratio for the population mean

Francesco Bartolucci*

Dipartimento di Economia, Finanza e Statistica, Università di Perugia, 06123 Perugia, Italy

Received 29 June 2005; received in revised form 9 May 2006; accepted 23 May 2006
Available online 7 July 2006

Abstract

A penalized version of the empirical likelihood ratio test statistic for the population mean is proposed which may be computed even when this parameter does not belong to the convex hull of the data. Some theoretical results on the proposed test statistic are provided together with some guidelines on the choice of the penalization term. A double bootstrap procedure is also described which may be used for calibration when the sample size is small. The approach is illustrated by an example and a small simulation study.

Keywords: Constrained maximization; Convex hull condition; Double bootstrap; Non-parametric inference

1. Introduction

An effective non-parametric method for making inference on the mean μ of a p-variate distribution is the *empirical likelihood* (EL) method introduced by Owen (1988); see also Owen (1990, 2001). It is based on the non-parametric likelihood

$$L(\mu) = \max_{\pi \in \mathscr{S}_\mu(X)} \prod_i \pi_i,$$

where $X = (x_1, \ldots, x_n)$ denotes the sample and $\mathscr{S}_\mu(X)$ denotes the subset of the n-dimensional simplex $\mathscr{S}$ containing all the vectors $\pi = (\pi_1, \ldots, \pi_n)$ such that $\sum_i \pi_i x_i = \mu$. $L(\mu)$ is in practice the profile likelihood for μ of a multinomial model that places a mass probability π_i on any data point x_i. This function reaches its maximum value, n^{-n}, only at $\mu = \bar{x}$, with $\bar{x}$ denoting the sample mean. For making inference on μ, we can therefore use the *empirical likelihood ratio* (ELR) test statistic

$$R(\mu) = \frac{L(\mu)}{L(\bar{x})} = \max_{\pi \in \mathscr{S}_\mu(X)} \prod_i n\pi_i. \tag{1}$$

*Fax: +39 075 5855950.
E-mail address: bart@stat.unipg.it.

0167-7152/$ - see front matter
doi:10.1016/j.spl.2006.05.016

A very interesting result is that, under mild conditions, $-2\log\{R(\mu_0)\}$, with μ_0 denoting the true population mean, has asymptotic $\chi^2(p)$ distribution regardless of the distribution of the population (Owen, 1988). When the sample size is large enough, this extension of the Wilks (1938)'s theorem allows to compute very easily p-values and cutoff points for $R(\mu)$. For small samples, instead, a bootstrap calibration is required (Owen, 2001, Section 3.3); it is based on the computation of $R(\bar{x})$ for a suitable number of samples drawn with replacement from the observed sample.

A problem that may limit the applicability of the EL approach is that $L(\mu)$, and thus $R(\mu)$, may not be computed when μ does not belong to the *convex hull* of the data, $\mathscr{H}(X)$, defined as the subset of $\mathbb{R}^p$ containing all the p-dimensional vectors that may be expressed as $\sum_i \pi_i x_i$, $\pi \in \mathscr{S}$. The convention commonly followed in this case is to set $R(\mu) = 0$ (Owen, 2001, Section 3.14). In this way, however, the actual significance level of the test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ is at least equal to the probability that $\mu_0 \notin \mathscr{H}(X)$ and so it may be much higher than the nominal level α. Similarly, the actual coverage level of the confidence region for μ may be much smaller than the nominal level $1-\alpha$. Obviously, the probability that $\mu_0 \notin \mathscr{H}(X)$ is significative only when n is small with respect to p; the greater the skewness of the population is, the larger such a probability. The problem may also affect bootstrap calibration since there may exist bootstrap samples for which $R(\bar{x})$ may not be computed because their convex hull does not contain $\bar{x}$. If the proportion of these samples is larger than α, this calibration is not feasible (see Owen, 2001, Section 3.3).

To cope with the problem above, Owen (2001, Section 10.4) suggested to use the EL-t method (Baggerly, 1999) in which the constraint $\sum_i \pi_i x_i = \mu$ is replaced by a studentized version. Alternatively, we can use modified versions of EL that, as the *Euclidean likelihood*, are based on the maximization of functions of π that may be computed also when some π_i are negative (Baggerly, 1998; Owen, 2001, Sections 3.15 and 3.16). These solutions, however, are not considered completely satisfactory since they may lead to different inferential conclusions than that of EL when the latter may be applied without problems.

A different approach to deal with the convex hull condition is proposed in this paper. In this approach, the constrained maximization problem involved in the computation of the ELR is turned into an unconstrained maximization problem in which the multinomial likelihood in (1) is multiplied by a suitable penalization term. This term measures the discrepancy between $\sum_i \pi_i x_i$ and μ and depends on a positive scalar h in a way that forces the solution towards the EL solution as h approaches 0. The resulting likelihood ratio, denoted by $R^\dagger(\mu, h)$, may be computed for any μ and is always positive. Moreover, if h is small enough and $\mu \in \mathscr{H}(X)$, our approach leads to the same inferential conclusion on μ of the EL approach. If, instead, $\mu \notin \mathscr{H}(X)$, $R^\dagger(\mu, h)$ is still a measure of the evidence provided by the data in favor of μ for that bootstrap calibration is required.

The paper is organized as follows. The convex hull condition is illustrated, by using a simulated data set, in the remainder of this section. The proposed approach is described in Section 2, which also deals with the numerical computation of $\log\{R^\dagger(\mu, h)\}$ and its theoretical properties. The use of this statistic for making inference on μ is outlined in Section 3, where some guidelines for choosing h and bootstrap calibration are also illustrated. Finally, Section 4 proposes an empirical illustration based on the data set introduced below and on a small simulation study.

1.1. An example

The following sample consists of $n = 15$ observations generated from a vector of four independent random variables, each with $\chi^2(1)$ distribution.

0.65	0.00	0.75	0.07	0.64	0.20	0.01	0.08	0.21	0.40	0.27	0.09	1.28	0.33	0.08
2.43	0.15	4.65	1.43	0.04	0.00	0.23	0.01	0.44	0.79	0.15	5.66	0.24	1.21	0.84
0.23	1.31	0.00	1.45	0.65	0.00	0.12	0.00	0.65	0.40	4.22	0.00	0.22	1.16	0.00
0.07	0.82	0.12	8.27	0.18	1.46	1.04	0.05	0.04	0.03	1.60	1.17	0.01	0.21	0.99

Even though the true population mean is $\mu_0 = 1_4$, where 1_k denotes a column vector of k ones, $R(\mu_0)$ cannot be computed since μ_0 does not belong to the convex hull of the data. There does not exist in fact any convex linear combination of the 15 column vectors above which is equal to μ_0. Following the general convention

according to that we set $R(\mu)=0$ when $\mu \notin \mathscr{H}(X)$, we have to reject $H_0 : \mu = \mu_0$. In this way, however, the actual significance level for the test is at least equal to the probability of $\mu_0 \notin \mathscr{H}(X)$ which, by simulation, we estimated to be equal to 0.128.

Bootstrap calibration is also infeasible for the sample above. This is due to the fact that $\bar{x}$ does not belong to the convex hull of about 30% of the bootstrap samples. We cannot therefore construct a confidence region for μ with a coverage level larger than 0.7. A similar problem happens for most of the samples generated from the $\chi^2(1)$ population at issue. In particular, we observed that for only two of the 1000 samples generated from this population it is possible to perform a bootstrap calibration when α is equal to 0.05 and for only 44 samples when α is equal to 0.10.

2. Penalized EL

The proposed approach is based on the following penalized version of the EL

$$L^\dagger(\mu,h)=\max_{\pi\in\mathscr{S}}\left(\prod_i \pi_i\right)\mathrm{e}^{-n\delta(v-\mu)/(2h^2)},$$

where $v=\sum_i \pi_i x_i$ and $\delta(v-\mu)$ is a distance measure between μ and v. In particular, we will consider

$$\delta(v-\mu)=(v-\mu)'V^{-1}(v-\mu),$$

where V is the sample variance–covariance matrix; when V is singular, V^{-1} is substituted by $E\,\mathrm{diag}(v)^{-1}E'$, where v is the vector of the positive eigenvalues of V and E is the matrix of the corresponding eigenvectors.

An interesting property of $L^\dagger(\mu,h)$ is that, for any $h>0$, the maximum value of this function is attained only at $\mu=\bar{x}$ and this maximum is equal to n^{-n}. So, for making inference on μ, we can use the ratio

$$R^\dagger(\mu,h)=\frac{L^\dagger(\mu,h)}{L^\dagger(\bar{x},h)}=\max_{\pi\in\mathscr{S}}\left(\prod_i n\pi_i\right)\mathrm{e}^{-n\delta(v-\mu)/(2h^2)},$$

which is similar in the spirit to $R(\mu)$. The main difference between $R(\mu)$ and $R^\dagger(\mu,h)$ is that the latter escapes the convex hull condition and so may be computed for any value of μ. In practice, however, we will make use of

$$r^\dagger(\mu,h)=\log\{R^\dagger(\mu,h)\}=\max_{\pi\in\mathscr{S}}\sum_i \log(n\pi_i)-\frac{n}{2h^2}(v-\mu)'V^{-1}(v-\mu),$$

which is simpler to compute.

2.1. Computing $r^\dagger(\mu,h)$

First of all note that $r^\dagger(\mu,h)$ may be expressed in an equivalent form as

$$r^\dagger(\mu,h)=\max_{v\in\mathscr{H}(X)} g(\mu,v,h),$$

where

$$g(\mu,v,h)=r(v)-\frac{n}{2h^2}(v-\mu)'V^{-1}(v-\mu) \tag{2}$$

and $r(v)=\log\{R(v)\}$. Owen (2001, Section 3.14) showed that $r(v)=\sum_i \log\{n\hat{\pi}_i(v)\}$, where

$$\hat{\pi}_i(v)=\frac{1}{n}\frac{1}{1+\hat{\lambda}(v)'(x_i-v)},\quad i=1,\ldots,n,$$

and $\hat{\lambda}(v)$ is implicitly defined as the solution of the equation $t(\lambda,v)=0$, with

$$t(\lambda,v)=\sum_i \frac{x_i-v}{1+\lambda'(x_i-v)}.$$

Therefore, for given μ and h, $r^\dagger(\mu,h)$ may be computed by maximizing $g(\mu,v,h)$ with respect to v. This may be performed through a Newton–Raphson algorithm that consists of updating, until convergence, the vector v by adding to its current value, v_0, the vector

$$-g_{vv}(\mu,v_0,h)^{-1}g_v(\mu,v_0,h),$$

where $g_v(\mu,v,h)$ and $g_{vv}(\mu,v,h)$ denote, respectively, the first derivative vector and the second derivative matrix of $g(\mu,v,h)$ with respect to v. These are equal to

$$g_v(\mu,v,h) = r_v(v) - \frac{n}{h^2}V^{-1}(v-\mu),$$
$$g_{vv}(\mu,v,h) = r_{vv}(v) - \frac{n}{h^2}V^{-1},$$

where $r_v(v) = n\hat{\lambda}(v)$ and, according to the implicit derivative rule,

$$r_{vv}(v) = -nt_\lambda\{\hat{\lambda}(v),v\}^{-1}t_v\{\hat{\lambda}(v),v\}.$$

After some algebra, we can easily see that

$$t_v\{\hat{\lambda}(v),v\} = n^2\sum_i \hat{\pi}_i(v)^2(x_i-v)\hat{\lambda}(v)' - nI_p,$$
$$t_\lambda\{\hat{\lambda}(v),v\} = -n^2\sum_i \hat{\pi}_i(v)^2(x_i-v)(x_i-v)',$$

where I_p is the identity matrix of dimension p.

The point that maximizes $g(\mu,v,h)$ will be denoted by $\hat{v}(\mu,h)$ or simply by $\hat{v}$ when its arguments are clear from the context. Note that, since $g(\mu,v,h)$ is the sum of two concave functions in v, this solution is unique and the starting value of the algorithm above is not so relevant. However, we suggest to use $v=\mu$ as starting value for the algorithm when $\mu\in\mathscr{H}(X)$; we observed that in this way it converges very quickly: three or four iterations are usually enough. When instead $\mu\notin\mathscr{H}(X)$, some other starting point, such as $\bar{x}$, may be taken. In these circumstances, however, the algorithm is slower since $\hat{v}$ tends to be very close to the boundary of $\mathscr{H}(X)$. As will be explained below, this requires careful choice of h.

2.2. Properties of $r^\dagger(\mu,h)$

In the following, some results concerning $r^\dagger(\mu,h)$ are proved. These results allow to clarify the strong relation between this test statistic and $r(\mu)$.

Proposition 1. *For any sample X and any $h>0$, $r^\dagger(\mu,h)$ is a concave function of μ and reaches its maximum, equal to 0, only at $\mu=\bar{x}$.*

Proof. Consider two points $\mu^{(1)}$ and $\mu^{(2)}$ in $\mathbb{R}^p$ and the corresponding solutions in terms of v of $r^\dagger(\mu,h)$, $\hat{v}^{(1)}$ and $\hat{v}^{(2)}$. Consider also a third point in $\mathbb{R}^p$, $\bar{\mu}=\alpha\mu^{(1)}+(1-\alpha)\mu^{(2)}$, $0<\alpha<1$, and let $\bar{v}$ be defined in a similar way. We obviously have that $r^\dagger(\bar{\mu},h)=g(\bar{\mu},\hat{v},h)\geqslant g(\bar{\mu},\bar{v},h)$ which, in turn, is greater than $\alpha r^\dagger(\mu^{(1)},h)+(1-\alpha)r^\dagger(\mu^{(2)},h)$ for any $0<\alpha<1$ and so concavity of $r^\dagger(\mu,h)$ follows. This is because $\bar{v}-\bar{\mu}$ may be expressed as $\alpha(\hat{v}^{(1)}-\mu^{(1)})+(1-\alpha)(\hat{v}^{(2)}-\mu^{(2)})$ and for the concavity of both summands at the right-hand side of (2). For any $h>0$, the maximum of $r^\dagger(\mu,h)$ is reached only at $\mu=\bar{x}$ and $r^\dagger(\bar{x},h)=0$ since $L^\dagger(\mu,h)$ reaches its maximum only at $\mu=\bar{x}$ and $L^\dagger(\bar{x},h)=n^{-n}$. □

The proof of the following proposition is based on the first-step solution of the Newton–Raphson algorithm described in Section 2.1,

$$\tilde{v}(\mu,h) = \mu + \left(t_\lambda\{\hat{\lambda}(\mu),\mu\}^{-1}t_v\{\hat{\lambda}(\mu),\mu\} + \frac{1}{h^2}V^{-1}\right)^{-1}\hat{\lambda}(\mu), \tag{3}$$

which will denoted by $\tilde{v}$ when the arguments μ and h are clear from the context.

Proposition 2. *For any sample X and any μ, $r^\dagger(\mu,h)$ is a decreasing function of h. Moreover, if $\mu \in \mathscr{H}(X)$, $r^\dagger(\mu,h) \geqslant r(\mu)$ for any $h>0$, $\|\hat{v}-\mu\| = O(h^2)$ and so $\lim_{h\to 0} r^\dagger(\mu,h) = r(\mu)$. If instead $\mu \notin \mathscr{H}(X)$, $\|\hat{v}-\mu\|$ is bounded away from 0 for any $h>0$ and so $\lim_{h\to 0} r^\dagger(\mu,h) = -\infty$.*

Proof. The first derivative of $r^\dagger(\mu,h)$ with respect to h,

$$r^\dagger_h(\mu,h) = -\frac{n}{h^3}(\hat{v}-\mu)' V^{-1}(\hat{v}-\mu),$$

is always negative and so $r^\dagger(\mu,h)$ is decreasing in h. To prove that when $\mu \in \mathscr{H}(X)$, $r^\dagger(\mu,h) \geqslant r(\mu)$ for any $h>0$, simply consider that $r^\dagger(\mu,h) = g(\mu,\hat{v},h) \geqslant g(\mu,\mu,h) = r(\mu)$. To prove that, in the same situation, $\|\hat{v}-\mu\| = O(h^2)$, it is enough to consider that $\|\tilde{v}-\mu\| = O(h^2)$, where $\tilde{v}$ is the first-order solution defined in (3); by substitution, $\lim_{h\to 0} r^\dagger(\mu,h) = r(\mu)$. Finally, since $\hat{v} \in \mathscr{H}(X)$ for any μ and $h>0$, $\|\hat{v}-\mu\|$ is bounded away from 0 when $\mu \notin \mathscr{H}(X)$ and so the penalization term grows indefinitely as h approaches 0. □

3. Inference on the population mean

On the basis of penalized ELR introduced in the previous section, we can test a hypothesis of type $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ and construct confidence regions for μ. In particular, the critical region of a test of this type is

$$\{X : r^\dagger(\mu_0,h) < r_\alpha(h)\}, \tag{4}$$

where $r_\alpha(h)$ is a suitable cutoff point such that the probability that a random sample belongs to this region under H_0 is equal to the significance level α.

A confidence region at the level $1-\alpha$ for μ may be expressed as

$$\{\mu : r^\dagger(\mu,h) \geqslant r_\alpha(h)\}. \tag{5}$$

Convexity of such a region is ensured by Proposition 1.

Obviously, a crucial issue for this type of inference is the choice of h and $r_\alpha(h)$.

3.1. Choice of h

The criterion that we follow to choose h is based on the curvature of the function $r^\dagger(\mu,h)$ at $\mu = \bar{x}$ that may be measured through the determinant of its second derivative, i.e.

$$c(h) = |r^\dagger_{\mu\mu}(\bar{x},h)|.$$

This is because, provided that $r_\alpha(h)$ does not vary with h, the higher the curvature of $r^\dagger(\mu,h)$ is, the more powerful the test with critical region (4) and the lower the volume of the confidence region (5). In particular, using standard rules on the implicit derivatives, we have that

$$r^\dagger_{\mu\mu}(\bar{x},h) = g_{\mu\mu}(\bar{x},\hat{v},h) - g_{\mu v}(\bar{x},\hat{v},h) g_{vv}(\bar{x},\hat{v},h)^{-1} g_{v\mu}(\bar{x},\hat{v},h) = -\frac{n}{1+h^2} V^{-1},$$

which, in accordance with Proposition 1, is negative definite. So, we have that

$$c(h) = \left(-\frac{n}{1+h^2}\right)^p \frac{1}{|V|}$$

whose infimum is $c = (-n)^p/|V|$. Not surprisingly, this lower bound is equal to the determinant of $r_{\mu\mu}(\bar{x}) = -nV^{-1}$.

Since $\lim_{h\to 0} c^\dagger(h) = c$, according to our criterion we should choose a value of h very close to 0. To check if h is close enough to 0 we suggest to consider the ratio $c(h)/c = (1+h^2)^{-p}$; in particular, we can choose the optimal value of h by solving $c(h)/c = 1-\varepsilon/n$, where ε is a suitable tolerance level. The solution of this equation is simply

$$h = \sqrt{(1-\varepsilon/n)^{1/p} - 1}.$$

Note that, when h is very close to 0 and $\mu \notin \mathscr{H}(X)$, we may go into numerical problems in computing $r^\dagger(\mu, h)$ since $\hat{v}$ will be very close to the boundary of $\mathscr{H}(X)$ and both $r(\hat{v})$ and $r^\dagger(\mu, h)$ will be very large in absolute value. So, we suggest to avoid too stringent tolerance levels, i.e. less than $10^{-6}n$.

3.2. Computing $r^\dagger_\alpha$

First of all consider the following proposition concerning the asymptotic distribution of $r^\dagger(\mu, h_n)$, where h_n denotes the value of h when the sample size is n. Similarly, we will use V_n to denote the variance–covariance matrix of a sample of size n and $\hat{v}_n$ to denote $\hat{v}(\mu, h_n)$.

Proposition 3. *If the population has mean μ_0 and variance–covariance matrix Σ_0, finite and of full rank, and* $h_n = \mathrm{O}(n^{-1/2})$,

$$-2r^\dagger(\mu_0, h_n) = -2r(\hat{v}_n) + \frac{n}{h_n^2}(\hat{v}_n - \mu_0)' V_n^{-1}(\hat{v}_n - \mu_0) \tag{6}$$

has asymptotic $\chi^2(p)$ distribution.

Proof. According to Owen (2001, Section 11.2) we have that $\|\hat{\lambda}(\mu_0)\| = \mathrm{O_p}(n^{-1/2})$; this implies that $\|\tilde{v}_n - \mu_0\|$, and so $\|\hat{v}_n - \mu_0\|$, is $\mathrm{O_p}(n^{-3/2})$. Then, $-2r(\hat{v}_n)$ has asymptotic $\chi^2(p)$ distribution (see Owen, 1988, Corollary 1), whereas the second term at the right-hand side of (6) tends in probability to 0 and the result therefore follows. □

According to the above result, when the sample size is large enough and h is close to 0, both $r^\dagger(\mu, h)$ and $r(\mu)$ have approximately $\chi^2(p)$ distribution. This is not surprising because in these circumstances $r^\dagger(\mu, h)$ is usually very close to $r(\mu)$. This result, however, may not be applied when the sample size is small, a situation in which our approach is really advantageous with respect to the usual EL approach and that requires bootstrap calibration. This procedure consists of computing, for a suitable number of bootstrap samples, $X_1, \ldots, X_B$, the statistic $r^\dagger(\bar{x}, h)$. Then, following Davison and Hinkley (1997, Section 4.4) , the p-value for the observed value of $r^\dagger(\mu, h)$, $r^\dagger$, may be estimated as

$$p = \frac{1 + \sum_b I(r_b^\dagger \leqslant r^\dagger)}{B + 1},$$

where $I(\cdot)$ is the indicator function and $r_b^\dagger$ is the value of $r^\dagger(\bar{x}, h)$ for the bth bootstrap sample. Accordingly, $r_\alpha(h)$ may be found as the value of $r^\dagger$ such that $p = \alpha$. This may be obtained as the smallest $\hat{b}$th value of $r_1^\dagger, \ldots, r_B^\dagger$, denoted by $r^\dagger_{(\hat{b})}$, where $\hat{b} = \alpha(B + 1) - 1$.

The bootstrap procedure above may lead to p-values and cutoff points which are consistently biased. An improved procedure is based on a double bootstrapping (Davison and Hinkley, 1997, Section 4.5). It consists of drawing, from any bootstrap sample X_b, a suitable number of second-level bootstrap samples $X_{b1}, \ldots, X_{bC}$ and computing, for any of them, the statistic $r^\dagger(\bar{x}_b, h)$, where $\bar{x}_b$ is the mean of X_b. Then, the adjusted p-value is computed as

$$p_{\mathrm{adj}} = \frac{1 + \sum_b I(p_b \leqslant p)}{B + 1},$$

where

$$p_b = \frac{1 + \sum_c I(r_{bc}^\dagger \leqslant r_b^\dagger)}{C + 1} \tag{7}$$

is the p-value for $r_b^\dagger$, with $r_{bc}^\dagger$ denoting the value of $r^\dagger(\bar{x}_b, h)$ for the sample X_{bc}. The adjusted cutoff point $r_{\mathrm{adj},\alpha}(\mu)$ may be obtained as $r^\dagger_{(\hat{b}_{\mathrm{adj}})}$ where $\hat{b}_{\mathrm{adj}} = p_{(\hat{b})}(B + 1) - 1$. This procedure is very expensive from the computational point of view since it requires to BC maximizations of the type described in Section 2. Therefore, we substitute $r_b^\dagger$ and $r_{bc}^\dagger$ in (7) with approximated values obtained by a second-order expansion

Table 1
Cutoff points for the penalized ELR for the data in Section 1.1

α	$r_\alpha(h)$	$r_{\text{adj},\alpha}(h)$
0.01	-3.0231×10^7	-3.0231×10^7
0.05	-3.2119×10^6	-3.1742×10^6
0.10	-1.6335×10^6	-1.5587×10^6
0.25	-23.54	-1.4413×10^6

Table 2
Results of the simulation study

α	Bootstrap			Double bootstrap		
	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
0.01	0.003	0.000	0.006	0.019	0.011	0.027
0.05	0.030	0.019	0.041	0.052	0.038	0.066
0.10	0.059	0.044	0.074	0.088	0.070	0.106
0.25	0.211	0.186	0.236	0.245	0.218	0.272

$\hat{\alpha}$ is the actual significance level of the test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ and $(\hat{\alpha}_L, \hat{\alpha}_U)$ is the corresponding 95% confidence interval.

of $r^\dagger(\mu, h)$ around $\mu = \bar{x}$. According to this expansion we have that

$$r^\dagger(\mu, h) \approx -\frac{1}{2(1+h^2)}(\bar{x}-\mu)' V^{-1}(\bar{x}-\mu),$$

which may be computed very quickly.

4. Empirical illustration

To analyze the sample reported in Section 1.1 we used $h = 0.002$. With this value of h, the penalized ELR for the true value of the population mean, $\mu_0 = 1_4$, is $r^\dagger = -62,313$. With a bootstrap procedure based on $B = 199$ samples we obtained a p-value for $r^\dagger$ equal to $p = 0.225$, while the adjusted p-value based on the double bootstrap procedure based on $C = 199$ second-level samples is $p_{\text{adj}} = 0.265$. So, even if μ_0 does not belong to the convex hull of the data, there is not enough evidence against this value of μ. Using the approach described in Section 3.2 we can also compute the cutoff points $r^\dagger_\alpha(h)$ and $r^\dagger_{\text{adj},\alpha}(h)$. These are shown in Table 1 for some values of α.

To assess if these cutoff points are reliable, we carried out a simulation based on 1000 samples drawn from the same population from that the sample considered above has been drawn. The results of the simulation are shown in Table 2 where, for some values of α, the actual significance level and the corresponding 95% confidence interval are reported. From this table, it is possible to note that the actual significance level of the test based on $r^\dagger(\mu, h)$ is close to the nominal level when the double bootstrap procedure is used for calibration.

References

Baggerly, K.A., 1998. Empirical likelihood as a goodness-of-fit measure. Biometrika 85, 535–547.
Baggerly, K.A., 1999. Studentized empirical likelihood and maximum entropy. Technical Report, Department of Statistics, Rice University.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge.
Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75, 237–249.
Owen, A.B., 1990. Empirical likelihood ratio confidence regions. Ann. Statist. 18, 90–120.
Owen, A.B., 2001. Empirical Likelihood. Chapman & Hall, London.
Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist. 9, 60–62.