Statistica Sinica **10**(2000), 545-557

EMPIRICAL EXPONENTIAL FAMILY LIKELIHOOD USING SEVERAL MOMENT CONDITIONS

S. A. Corcoran*

University of Oxford

Abstract: We consider nonparametric likelihoods for the mean of an unknown distribution using estimating equations for moments of order one and greater. Although empirical likelihood is the same regardless of the number of estimating equations used, use of two or more such estimating equations with the empirical exponential family gives a likelihood that agrees with empirical likelihood to third order. We show that the empirical exponential family using an arbitrary number of moments is a least favorable family. Simulations indicate that empirical exponential family using estimating equations for moments of order one and greater is very close to empirical likelihood and that a Wald statistic constructed using the empirical exponential family gives good coverage.

Key words and phrases: Empirical exponential family, empirical likelihood, estimating equation, least favorable family, nonparametric likelihood, nuisance parameter.

1. Introduction

We consider the construction of a likelihood for a scalar mean θ from an unknown distribution F using two nonparametric approaches: empirical likelihood (Owen (1988)) and empirical exponential family likelihood (Efron (1981), DiCiccio and Romano (1990), Davison, Hinkley and Worton (1992)).

We first review a general method of constructing nonparametric likelihoods. Consider maximum likelihood estimation of a distribution F having mean θ, restricting attention to multinomial distribution functions F_p supported on the sample and having probabilities $\{p_i\}$ (in Section 5 we show that this does not make the problem artificially easier). The nonparametric likelihood is $L(\theta) = \prod p_i$, where, $p = \{p_i\}$ satisfies $E_{F_p}(X) = \theta$. Its normed version is $\bar{L}(\theta) = L(\theta)/L\left(\hat{\theta}\right) = \prod(p_i/\hat{p}_i)$, where, in addition to the constraint on p, $\hat{p}$ satisfies $E_{F_{\hat{p}}}(X) = \hat{\theta}$. The nonparametric maximum likelihood estimate of F is the empirical distribution function $\hat{F}$, which places probability mass $\{n^{-1}\}$ on each

*S. A. Corcoran died on 8 March 1996, while working towards a DPhil in Statistics at the University of Oxford; his work was supported by an EPSRC Research Studentship. This paper was found among his effects and edited by A. C. Davison, Swiss Federal Institute of Technology, Lausanne.

of the sample points. Thus $\hat{p} = \{n^{-1}\}$, the nonparametric maximum likelihood estimate of θ is $\hat{\theta} = \bar{X}$, and a normed nonparametric log likelihood for θ is

$$\ell(\theta) = \log \bar{L}(\theta) = \sum_{i=1}^{n} \log(np_i). \tag{1.1}$$

We require that F_p has mean θ and that $p = \{p_i\}$ is a probability distribution, implying the following constraints on the $\{p_i\}$,

$$E_{F_p}(X) = \sum p_i X_i = \theta, \tag{1.2}$$

$$\sum p_i = 1. \tag{1.3}$$

The forward Kullback-Leibler discrepancy measure (Kullback and Leibler (1951)) between discrete distributions $\{p_i\}$ and $\{q_i\}$ is $D_f = \sum q_i \log(q_i/p_i)$. The empirical likelihood method uses probabilities p_i which minimize the forward Kullback–Leibler discrepancy measure between probabilities p_i consistent with F having mean θ, and probabilities $q_i = n^{-1}$ consistent with F having mean $\hat{\theta} = \bar{X}$. The use of this discrepancy measure means that $\ell_E = -nD_f$, so that minimum discrepancy between F_p and $F_{\hat{p}}$ maximizes ℓ_E, the empirical log likelihood. The solution of the minimization problem is $p_i = n^{-1}\{1 + \alpha_\theta(X_i - \theta)\}^{-1}$ where we may determine the "tilt" parameter α_θ for each value of θ from (1.2); see Section 2. The probabilities $\{p_i\}$ are subject to two constraints, which means that we have $(n-2)$ nuisance parameters and only n sample values. Hence it is surprising that empirical likelihood parallels many aspects of parametric likelihood, for instance, χ^2 asymptotics of the likelihood ratio. Further discussion and references are given by Davison and Hinkley (1997, Chapter 10).

The empirical exponential family likelihood ℓ_{EEF} is a related nonparametric likelihood. The likelihood itself is given by (1.1), where the probabilities $\{p_i\}$ are given by (1.2) and (1.3). However, the empirical exponential family likelihood uses probabilities $\{p_i\}$ that minimize the backward Kullback–Leibler discrepancy measure between $\{p_i\}$ and $\{n^{-1}\}$, i.e., $D_b = \sum p_i \log(np_i)$, leading to $p_i \propto e^{\alpha_\theta(X_i - \theta)}$; see Section 3.

It may be shown that $\ell_E(\theta) - \ell_{EEF}(\theta) = O_p(n^{-1})$ as the sample size $n \to \infty$. This asymptotic order of agreement ensures that empirical and empirical exponential family likelihoods share second-order properties. DiCiccio and Romano (1989) show that one-sided confidence limits for parameters that are smooth functions of vector means constructed using either empirical or empirical exponential family likelihood have coverage error $O(n^{-1/2})$, and that two-sided limits have error $O(n^{-1})$; these equal the parametric error rates. DiCiccio and Romano also develop third-order expansions for nonparametric confidence limits for a scalar

mean. These results indicate that one-sided empirical likelihood confidence limits may be corrected to $O(n^{-3/2})$ and that the empirical likelihood ratio admits a Bartlett correction, giving two-sided confidence limits with error $O(n^{-2})$ (DiCiccio, Hall and Romano (1991)); neither of these last two results holds for $\ell_{EEF}(\theta)$ (Jing and Wood (1996)).

In this paper we consider constructing these two nonparametric likelihoods using estimating equations for moments of order one and greater, generalizing the above framework where the constraint (1.2) may be written as the single equation $\sum p_i(X_i - \theta) = 0$. It may seem strange to introduce moments of order two and higher as nuisance parameters when we wish to construct a likelihood for the mean θ. Consider constructing the empirical exponential family likelihood using two estimating equations for the first two moments. As we will show in Section 3, this gives probabilities $p_i \propto e^{\eta_1(X_i-\theta)+\eta_2(X_i-\theta)^2}$ that take account of both the linear *and* quadratic effects of the sample. This paper examines the form of two nonparametric likelihoods when we attempt to take account of such higher-order sample effects in the hope of getting better large-sample properties.

We do not consider the case when higher-order moments are *known* to depend on the mean θ, such as a mean-variance relationship of the form $Var(X) = \theta^2 = E(X)^2$. The construction of empirical likelihood which makes efficient use of extra estimating equations for this situation has been discussed by Qin and Lawless (1994).

We show that use of estimating equations for moments of order one and greater leaves empirical likelihood unchanged, and that if we use two or more moments about the mean in constructing empirical exponential family likelihood then the result agrees with empirical likelihood to third-order. This ensures that the resulting empirical exponential family likelihood shares all the asymptotic properties of empirical likelihood established in the literature so far. We present empirical evidence in Section 6, and discuss the generality of our results in Section 7.

2. Empirical Likelihood

The motivation for empirical likelihood (Owen (1988, 1990, 1991)) for a scalar mean θ starts from minimization of the forward Kullback–Leibler discrepancy between two distributions. This measures the discrepancy between a multinomial distribution placing mass p_i on the observed sample values $X_1, \ldots, X_n$ and the empirical distribution function which places probability mass n^{-1} on each of the X_i. The distributions $\{n^{-1}\}$ and $\{p_i\}$ correspond to nonparametric maximum likelihood estimates of $\hat{F}_{\hat{\theta}}(x) = \prod(n^{-1}\, 1_{\{X_i \le x\}})$ with mean $\hat{\theta} = \bar{X}$, and $\hat{F}_{\theta}(x) = \prod(p_i\, 1_{\{X_i \le x\}})$ with mean θ. We have constraints that the probabilities

p_i sum to one, and that the mean $\sum p_i X_i$ is equal to θ, where θ is the parameter value at which we wish to construct the likelihood. The constraints on $\{n^{-1}\}$ are automatically satisfied since $\sum n^{-1} = 1, \sum n^{-1} X_i = \bar{X}$. Hence we minimize

$$-\sum_{i=1}^{n} \frac{1}{n} \log(np_i) + \beta\Big(\sum_{i=1}^{n} p_i - 1\Big) + \alpha\Big\{\sum_{i=1}^{n} p_i(X_i - \theta)\Big\}, \tag{2.1}$$

where α and β are Lagrange multipliers. The empirical likelihood for θ is

$$\ell_E(\theta) = \ell_E\{\theta(\alpha_\theta)\} = -\sum_{i=1}^{n} \log\{1 + \alpha_\theta(X_i - \theta)\}$$

and the "tilt" parameter α_θ is determined by

$$\sum_{i=1}^{n} \frac{(X_i - \theta)}{1 + \alpha_\theta(X_i - \theta)} = 0. \tag{2.2}$$

Consider including extra information by accounting for higher sample moments using estimating equations. Suppose μ^r is the rth population moment about the mean. We construct a joint likelihood $\ell(\theta, \mu^2, \ldots, \mu^q)$, and then profile out $\mu^2, \ldots, \mu^q$ to obtain $\ell_q(\theta)$. The usual empirical likelihood is thus $\ell_E = \ell_1$. However it is clear that maximizing $\ell(\theta, \mu^2, \ldots, \mu^q)$ over $\mu^2, \ldots, \mu^q$ will give $\ell_E(\theta)$ also, because we are simply imposing additional restrictions on the log likelihood and then maximizing over them. Hence the empirical likelihood for the mean is unchanged by including estimating equations for higher moments.

3. Empirical Exponential Family Likelihood

The empirical exponential family likelihood arises from considering the backward Kullback–Leibler discrepancy between n^{-1} and p_i. The Lagrangian is

$$-\sum_{i=1}^{n} p_i \log(np_i) + \beta\Big(\sum_{i=1}^{n} p_i - 1\Big) + \alpha\Big\{\sum_{i=1}^{n} p_i(X_i - \theta)\Big\}, \tag{3.1}$$

where α and β are again Lagrange multipliers. A simple calculation shows that

$$p_i(\alpha_\theta) = \frac{e^{\alpha_\theta(X_i - \theta)}}{\sum e^{\alpha_\theta(X_j - \theta)}} \tag{3.2}$$

and the empirical exponential family log likelihood is

$$\ell_{EEF,1}(\theta) = \ell_{EEF,1}\{\theta(\alpha_\theta)\} = \alpha_\theta \sum_{i=1}^{n} (X_i - \theta) - n \log\Big(\frac{1}{n} \sum_{i=1}^{n} e^{\alpha_\theta(X_j - \theta)}\Big).$$

We drop the subscript EEF and consider the construction of a profile likelihood analogous to that in Section 2, i.e., adding the $q-1$ estimating equations $\sum p_i\{(X_i-\theta)^r-\mu^r\}=0$ for $r=2,\ldots,q$. If Z_i denotes the $q\times 1$ vector whose rth element is $Y_i^r-\mu^r=(X_i-\theta)^r-\mu^r$, say, then the moment constraints may be expressed as $\sum Z_i p_i=0$. On adding these to $\sum p_i=1$ and modifying (3.1) accordingly, we find that $p_i(\eta)=e^{\eta^T Z_i}/\sum e^{\eta^T Z_j}$, and the corresponding log likelihood is $\ell(\theta,\mu^2,\ldots,\mu^q)=\eta^T\sum_{i=1}^n Z_i-n\log\left(\frac{1}{n}\sum_{j=1}^n e^{\eta^T Z_j}\right)$. As $\mu^2,\ldots,\mu^q$ cancel from $p_i(\eta)$, the log likelihood depends only on θ and η. So if Y_i is a $q\times 1$ vector with rth element Y_i^r, then $p_i(\eta)=e^{\eta^T Y_i}/\sum e^{\eta^T Y_j}$ with log likelihood

$$\ell_q(\theta)=\eta^T\sum_{i=1}^n Y_i-n\log\Big(\frac{1}{n}\sum_{j=1}^n e^{\eta^T Y_j}\Big). \tag{3.3}$$

The tilt parameters η are determined by maximizing $\ell_q(\theta)$ subject to $\sum(X_i-\theta)p_i(\eta)=0$.

What value has this? Empirical exponential family likelihood with just one moment condition is much easier to compute than empirical likelihood, because the score equations satisfied by α_θ are the same as those for a log-linear model in which a vector of n zeros is regressed on the vector Y^1 whose ith component is $X_i-\theta$, with no constant in the model. As this calculation is readily performed in any package that can fit a Poisson regression model, with no special programming, $\ell_1(\theta)$ and statistics computed from such a fit are simpler to obtain than are corresponding quantities for empirical likelihood. It does not seem possible to squeeze ℓ_q with $q=2,\ldots$ into this framework, but as our simulations show, there would be no small-sample gain to doing so, as despite the good theoretical properties of ℓ_E, it gives worse coverage than a simple statistic based on ℓ_1.

4. Empirical Exponential Family Expansion

In this section we outline an asymptotic expansion of (3.3), the empirical exponential family log likelihood for a scalar mean constructed using q estimating equations for the first q moments about the mean. By comparing this expansion to the corresponding expression for empirical likelihood we show that the empirical exponential family log likelihood for $q\geq 2$ agrees with that for empirical log likelihood to third order. We use the convention that repeated indices $a,b,\ldots$ are summed over, from 1 to q.

The Lagrangian for determining the tilt parameters η in (3.3) is

$$L=\eta^T\sum_{i=1}^n Y_i-n\log\Big(\frac{1}{n}\sum_{j=1}^n e^{\eta^T Y_j}\Big)+\lambda\sum Y_i^1 e^{\eta^T Y_i}=C_1-nC_2+\lambda C_3, \tag{4.1}$$

say. We determine η in terms of θ by equating the derivatives of L to zero. Hence inversion of $\partial L/\partial\eta^r = 0$ implies that

$$\begin{aligned}&\eta^a d_{a,r} + \frac{1}{2}\eta^a\eta^b e_1(a,b,r) + \frac{1}{6}\eta^a\eta^b\eta^c e_2(a,b,c,r)\\ &= \lambda\Big(m_{r+1} + \eta^a m_{a+r+1} + \frac{1}{2}\eta^a\eta^b m_{a+b+r+1} + \frac{1}{6}\eta^a\eta^b\eta^c m_{a+b+c+r+1}\Big), \quad (4.2)\end{aligned}$$

where $m_a = n^{-1}\sum_i (X_i-\theta)^a$, $d_{a,r} = m_{a+r} - m_a m_r$, and

$$\begin{aligned}e_1(a,b,r) &= m_{a+b+r} - 2m_a m_{b+r} + 2m_a m_b m_r - m_{a+b}m_r,\\ e_2(a,b,c,r) &= m_{a+b+c+r} - 3m_a m_{b+c+r} + 6m_a m_b m_{c+r} - 3m_{a+b}m_{c+r}\\ &\quad -6m_a m_b m_c m_r + 3m_a m_{b+c} m_r + 3m_{a+b}m_c m_r - m_{a+b+c}m_r.\end{aligned}$$

In order to simplify the asymptotic expansions, we set $\theta = \bar{X} + n^{-1/2}\psi$, so that $\psi = O_p(1)$ and $m_1 = n^{-1}\sum(X_i-\theta) = -n^{-1/2}\psi$. Then

$$n^{-1/2}\psi = \eta^a m_{a+1} + \frac{1}{2}\eta^a\eta^b m_{a+b+1} + \frac{1}{6}\eta^a\eta^b\eta^c m_{a+b+c+1} + O_p(n^{-2}). \quad (4.3)$$

Hence, $\eta^a = O_p(n^{-1/2})$, and (4.2) gives $\lambda = O_p(n^{-1/2})$, since $m_a = O_p(1)$ for $a \geq 2$. We now solve (4.2) and (4.3) to obtain expansions for η^a; it is nontrivial to collect powers of $n^{-1/2}$ as $m_1 = O_p(n^{-1/2})$ but $m_a = O_p(1)$ for $a \geq 2$. The expressions for the η^a given below differ according to the number, q, of estimating equations used, but they all have error $O_p(n^{-2})$. If $q = 1$ we have

$$\eta^1 = n^{-1/2}\psi/m_2 - n^{-1}\psi^2 m_3/(2m_2^3) + n^{-3/2}\psi^3(3m_3^2 - m_2m_4)/(6m_2^5).$$

If $q = 2$ we have

$$\begin{aligned}\eta^1 &= n^{-1/2}\psi/m_2 - n^{-1}\psi^2 m_3/m_2^3\\ &\quad + n^{-3/2}\frac{\psi^3(5m_2^3m_3^2 + 6m_3^4 - 2m_2^4m_4 - 9m_2m_3^2m_4 + 2m_2^2m_4^2 + m_2^2m_3m_5)}{3m_2^5(m_2^3 + m_3^2 - m_2m_4)},\\ \eta^2 &= n^{-1}\psi^2/(2m_2^2) + n^{-3/2}\frac{\psi^3(4m_2m_3m_4 - 2m_2^3m_3 - 3m_3^3 - m_2^2m_5)}{3m_2^5(m_2^3 + m_3^2 - m_2m_4)}.\end{aligned}$$

For $q \geq 3$ we have

$$\begin{aligned}\eta^1 &= n^{-1/2}\psi/m_2 - n^{-1}\psi^2 m_3/m_2^3 + n^{-3/2}\psi^3(2m_3^2 - m_2m_4)/m_2^5,\\ \eta^2 &= n^{-1}\psi^2/(2m_2^2) - n^{-3/2}\psi^3 m_3/m_2^4,\end{aligned}$$

$\eta^3 = n^{-3/2}\psi^3/(3m_2^3)$, and $\eta^z = 0$ for $z \geq 4$.

We now expand the log likelihood (3.3) for θ and insert the above expansions for η^a. The resulting expansion depends on q. Omitting the details we find that

$$\begin{aligned}\ell_1(\theta) &= -\psi^2/(2m_2) + n^{-1/2}\psi^3 m_3/(3m_2^3) + n^{-1}\psi^4(m_2^3 - 3m_3^2 + m_2 m_4)/(8m_2^5) \\ &\quad + O_p(n^{-3/2}), \\ \ell_q(\theta) &= -\psi^2/(2m_2) + n^{-1/2}\psi^3 m_3/(3m_2^3) + n^{-1}\psi^4(m_4 m_2 - 2m_3^2)/(4m_2^5) \\ &\quad + O_p(n^{-3/2}), \qquad q \geq 2. \end{aligned} \tag{4.4}$$

Rearrangement of equation (7.5) of Davison, Hinkley and Worton (1992) reveals that empirical likelihood $\ell_E(\theta)$ has the same expansion as (4.4). Hence $\ell_E(\theta) - \ell_q(\theta) = O_p(n^{-3/2})$ for $q \geq 2$, but $\ell_E(\theta) - \ell_1(\theta) = O_p(n^{-1})$. Thus, for $q \geq 2$, $\ell_q(\theta)$ has the same third-order asymptotic properties as $\ell_E(\theta)$, although the expressions for the tilt parameters η are different for $q = 2$ and $q \geq 3$.

An intuitive interpretation of this result is as follows. Consider the simplest non-trivial case, where $n = 3$ and $X_1 < X_2 < X_3$; the probabilities lie on the simplex $p_1 + p_2 + p_3 = 1$ with $p_i \geq 0$. If the contours of the function $\sum \log p_i$ are inscribed on the simplex, then as θ passes from X_1 to X_3, both $\ell_E(\theta)$ and $\ell_1(\theta)$ correspond to paths from $(1, 0, 0)$ to $(0, 0, 1)$ that pass through the centre $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Under mild conditions the same is true of any multinomial distribution with a single tilting parameter that starts at $p_1 = 1$ and finishes at $p_3 = 1$. Imposition of the constraint $\sum p_i X_i = \theta$ means that the probabilities move on a line on the simplex, and this line intersects with both $\ell_E(\theta)$ and $\ell_1(\theta)$. The argument in Section 2 shows that $\ell_E(\theta)$ is the highest point on this line, by definition, while $\ell_1(\theta) < \ell_E(\theta)$. Adding the additional "tilt" parameter η_2 corresponding to the second moment equation $(X_i - \theta)^2 - \mu^2$ perturbs $\ell_1(\theta)$ to another point $\ell_2(\theta)$, which is sufficiently close to $\ell_E(\theta)$ to share its third-order properties. In this particular case $\ell_2(\theta) = \ell_E(\theta)$, though this will not be true for larger n. It is tempting to believe that addition of q estimating equations makes $\ell_E(\theta)$ and $\ell_q(\theta)$ agree to order $q + 2$, but a proof of this using the techniques above would be very tedious.

5. Least Favorable Family Interpretation

We consider the interpretation of empirical and empirical exponential family likelihoods as least favorable families (Stein (1956)). The problem of constructing a nonparametric likelihood is infinite-dimensional because an infinite number of distributions have the required parameter value. Both empirical likelihood and the standard formulation of empirical exponential family likelihood (DiCiccio and Romano (1990)) restrict consideration to multinomial distributions $\{p_i\}$ supported on the observed sample. If we use a single estimating

equation $\sum p_i(X_i - \theta) = 0$ this reduces the problem to a one-dimensional subfamily, indexed by the tilt parameter α — see (2.2) and (3.2). If, for a particular nonparametric likelihood ℓ, we have $E_{\hat{F}_n}\left(-\partial^2\ell/\partial\theta^2\right) = n/\hat{\sigma}^2$, where $\hat{\sigma}^2 = n^{-1}\sum(X_i - \bar{X})^2$, then the Cramér–Rao variance lower bound for estimating θ is the same as that obtained for a parametric family. Hence the reduction of the infinite-dimensional problem of nonparametric estimation of the mean to a one-dimensional subfamily has not been made artificially easier; we call such a (sub-)family "least favorable". Further discussion is given by DiCiccio and Romano (1990) and Efron and Tibshirani (1993).

In the case of empirical likelihood we have $\ell(\theta) = -\sum_{i=1}^n \log\{1 + \alpha(X_i - \theta)\}$, and a little calculation gives $\partial\ell/\partial\theta = n\alpha$ and $\partial^2\ell/\partial\theta^2 = n\partial\alpha/\partial\theta$. Differentiating (2.2) with respect to θ yields

$$\frac{\partial\alpha}{\partial\theta} = \left[\alpha\sum\frac{X_i - \theta}{\{1 + \alpha(X_i - \theta)\}^2} - n\right]\left[\sum\frac{(X_i - \theta)^2}{\{1 + \alpha(X_i - \theta)\}^2}\right]^{-1}.$$

If we take expectation of minus the second derivative of the log likelihood under the empirical distribution function $\hat{F}_n$, so that $\hat{\alpha} = 0$ and $\hat{\theta} = \bar{X}$, we obtain $n/\hat{\sigma}^2$. Hence empirical likelihood is a least favorable family indexed by the tilt parameter α.

We now outline the proof that the empirical exponential family for a scalar mean θ constructed using q estimating equations for the first q moments about the mean is a least favorable family indexed by $\eta^1, \ldots, \eta^q$. Using (3.3) it is easy to show that

$$E_{\hat{F}_n}\left(-\frac{\partial^2\ell_q(\theta)}{\partial\theta^2}\right) = n\,\hat{\phi}^T\hat{D}\hat{\phi}, \tag{5.1}$$

where $\hat{\phi}$ is the q-vector $(\hat{\phi}_a)$ with $\phi_a = \partial\eta_a/\partial\theta$, $\hat{D}$ is the $q \times q$-matrix with elements $(\hat{d}_{a,b}) = (\hat{m}_{a+b} - \hat{m}_a\hat{m}_b)$ and the circumflexes indicate evaluation at the maximum likelihood point $\hat{\eta} = 0, \hat{\lambda} = 0, \hat{\theta} = \bar{X}$. We now derive equations for $\hat{\phi}_a$. Differentiation of (4.1) with respect to $\{\eta_a\}$ and λ gives us $q + 1$ equations involving η and λ. Further differentiating these $q + 1$ equations with respect to θ gives us $q + 1$ equations involving $\{\phi_a\}$ and ξ, where $\xi = \partial\lambda/\partial\theta$. Finally setting $\hat{\eta} = 0, \hat{\lambda} = 0, \hat{\theta} = \bar{X}$ we will have a system of $q + 1$ equations for $\hat{\phi}_a$ and $\hat{\xi}$. In matrix form, we find

$$\begin{pmatrix} -\hat{D} & \hat{e} \\ \hat{e}^T & 0 \end{pmatrix}\begin{pmatrix} \hat{\phi} \\ \hat{\xi} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{5.2}$$

where $\hat{e} = (\hat{m}_2, \ldots, \hat{m}_{q+1})^T$. We now solve (5.2) for $\hat{\phi}$ and substitute the result into (5.1), obtaining

$$E_{\hat{F}_n}\left(-\frac{\partial^2\ell_q(\theta)}{\partial\theta^2}\right) = n\left(\hat{e}^T\hat{D}^{-1}\hat{e}\right)^{-1}. \tag{5.3}$$

To complete the argument we need to evaluate the scalar $\hat{z}_q = \hat{e}^T \hat{D}^{-1} \hat{e}$, where here and more generally the subscript q indicates that we are considering the empirical exponential family constructed using q estimating equations. For $q = 1$, the case considered by DiCiccio and Romano (1990), we have $\hat{D} = \hat{m}_2 - \hat{m}_1^2 = \hat{m}_2$ and $\hat{e} = \hat{m}_2$ so that $\hat{z}_1 = \hat{m}_2 = \hat{\sigma}^2$. Hence the empirical exponential family is least favorable for $q = 1$. For general q it is possible to show that

$$\hat{z}_{q+1} = \hat{z}_q + \hat{y}_{q+1}^{-1} (\hat{a}_q - \hat{m}_{q+2})^2, \tag{5.4}$$

where $\hat{x}_q^T = (\hat{m}_{q+2}, \hat{m}_{q+3} - \hat{m}_2 \hat{m}_{q+1}, \ldots, \hat{m}_{2q+1} - \hat{m}_q \hat{m}_{q+1})$, $\hat{y}_{q+1} = \hat{m}_{2q+2} - \hat{m}_{q+1}^2 - \hat{x}_q^T \hat{D}_q^{-1} \hat{x}_q$, and $\hat{a}_q = \hat{x}_q^T \hat{D}_q^{-1} \hat{e}_q$. Partition $\hat{D}_q$, $\hat{e}_q$ and $\hat{x}_q$ as follows

$$\hat{D}_q = \begin{pmatrix} \hat{m}_2 & \hat{u}_{q-1}^T \\ \hat{u}_{q-1} & \hat{W}_{q-1} \end{pmatrix}, \hat{u}_{q-1}^T = (\hat{m}_3, \ldots, \hat{m}_{q+1}), \hat{e}_q^T = \left(\hat{m}_2, \hat{u}_{q-1}^T\right), \hat{x}_q^T = \left(\hat{m}_{q+2}, \hat{s}_{q-1}^T\right),$$

where $\hat{W}_{q-1}$ is the $(q-1) \times (q-1)$-submatrix of $\hat{D}_q$ induced by the partition, and $\hat{s}_{q-1}^T = (\hat{m}_{q+3} - \hat{m}_2 \hat{m}_{q+1}, \ldots, \hat{m}_{2q+1} - \hat{m}_q \hat{m}_{q+1})$. If we invert the partitioned $\hat{D}_q$ we find $\hat{a}_q = \hat{m}_{q+2}$. Hence using (5.4) we have $\hat{z}_{q+1} = \hat{z}_q$, and so $\hat{z}_q = \hat{z}_1 = \hat{m}_2 = \hat{\sigma}^2$ and (5.3) reduces to $n/\hat{\sigma}^2$. Thus the empirical exponential family for a scalar mean θ constructed using q estimating equations is least favorable.

6. Simulations

Suppose the parameter of interest is a scalar mean θ. To compare the coverage accuracy of empirical and empirical exponential family likelihood statistics we simulated 50,000 samples from a number of distributions; only those for the $N(0,1)$ and $\log N(0,1)$ are reported here. We constructed two-sided confidence intervals for the mean with various nominal coverage levels α as indicated in Tables 1 and 2. We estimated the actual coverages for eight nonparametric likelihood statistics, all of which have an asymptotic χ_1^2 distribution with error $O(n^{-1})$, or $O(n^{-2})$ for those statistics admitting a Bartlett correction. The first five statistics are:(1) the log empirical likelihood ratio statistic, $W = -2\ell_E(\theta)$; (2) its Bartlett corrected version $W' = W/(1 + \hat{b}/n)$, where $\hat{b} = \hat{m_4}/(2\hat{m_2}^2) - \hat{m_3}^2/(3\hat{m_2}^3)$ is a $\sqrt{n}$-consistent estimate of the Bartlett factor b (DiCiccio, Hall and Romano (1991)); and empirical exponential family log likelihood ratio statistics for $q = 1, 2, 3$, i.e., (3) $X = -2\ell_1(\theta)$; (4) $Y = -2\ell_2(\theta)$; and (5) $Z = -2\ell_3(\theta)$.

The remaining three statistics are Wald statistics constructed using a robust "sandwich" estimate of variance of the tilt parameters η. Such a statistic represents the nonparametric analogue of the parametric Wald statistic recommended in the presence of possible model misspecification — see Kent (1982) and White

(1982). The tilt parameters for either empirical or empirical exponential family likelihood for the mean θ are determined by the score equation

$$\sum p_i(\eta)(X_i - \theta) = 0. \tag{6.1}$$

Table 1. Coverage results for confidence intervals for the mean of the $N(0,1)$ distribution.

	$\alpha = 0.85$		$\alpha = 0.90$		$\alpha = 0.95$	
	$n = 8$	$n = 15$	$n = 8$	$n = 15$	$n = 8$	$n = 15$
W	77.38	81.66	82.49	86.68	88.14	92.20
W'	79.87	83.26	84.59	88.05	89.74	93.17
X	76.29	80.75	81.19	85.78	86.70	91.17
Y	77.34	81.61	82.43	86.61	88.05	92.13
Z	77.38	81.65	82.48	86.68	88.14	92.20
C_E	73.91	78.21	78.71	83.07	84.22	88.32
C_1	81.53	84.55	87.34	89.90	93.91	95.39
C_2	73.68	80.91	77.96	85.40	83.00	90.49

Table 2. Coverage results for confidence intervals for the mean: Log $N(0,1)$ distribution.

	$\alpha = 0.85$		$\alpha = 0.90$		$\alpha = 0.95$	
	$n = 40$	$n = 60$	$n = 40$	$n = 60$	$n = 40$	$n = 60$
W	78.56	80.10	84.14	85.46	90.18	91.37
W'	79.86	81.28	85.21	86.41	91.03	92.05
X	77.93	79.54	83.41	84.78	89.38	90.67
Y	78.55	80.10	84.13	85.46	90.10	91.29
Z	78.56	80.12	84.15	85.47	90.17	91.36
C_E	76.60	78.34	81.20	83.11	86.36	88.18
C_1	80.94	82.00	87.07	87.87	93.15	93.71
C_2	80.25	81.95	84.84	86.54	89.75	90.99

Let $\hat{\eta} = \hat{\eta}(\theta)$ be the value of η solving (6.1) for given θ. We may integrate the score with respect to η and define a log likelihood for θ by $\ell(\theta) = \sum (X_i - \theta) \int p_i(\eta)\, d\eta = \sum \ell_i$, where ℓ_i is the log likelihood contribution from X_i. Since there is a unique solution $\hat{\eta}(\theta)$ for each θ, there is a q–1 correspondence between the q tilt parameters η and the parameter of interest θ and we may specify a test for θ in terms of a test for the tilt parameters η. Thus if

$$A = \left(\frac{\partial \ell}{\partial \eta}\right)^2 = \sum \left(\frac{\partial \ell_i}{\partial \eta}\right)^2 = \sum (X_i - \theta)^2 p_i(\eta)^2, B = -\frac{\partial^2 \ell}{\partial \eta \partial \eta^T} = \sum (X_i - \theta) \frac{\partial p_i(\eta)}{\partial \eta},$$

a robust estimate of the variance of $\hat{\eta}$ is $(\hat{B}\hat{A}^{-1}\hat{B})^{-1}$, where the circumflexes on A, B denote evaluation at $\hat{\eta}$, and $C = \hat{\eta}^T \left(\hat{B}\hat{A}^{-1}\hat{B}\right)\hat{\eta}$ is a Wald statistic

with an approximate χ^2_1 distribution. The last three statistics are: (6) a Wald statistic C_E for empirical likelihood, using $p_i(\eta) = n^{-1}\{1 + \eta(X_i - \theta)\}^{-1}$; (7) a Wald statistic C_1 for empirical exponential family likelihood for $q = 1$, using $p_i(\eta) \propto e^{\eta(X_i-\theta)}$; and (8) a Wald statistic C_2 for empirical exponential family likelihood for $q = 2$, using $p_i(\eta) \propto e^{\eta_1(X_i-\theta)+\eta_2(X_i-\theta)^2}$.

Tables 1 and 2 show that the empirical coverages of empirical exponential family log likelihood ratio statistics Y and Z for $q = 2$ and $q = 3$ are extremely close to that of the empirical log likelihood ratio W, as we may expect from the asymptotic results of Section 4. Further simulations, here unreported, show that the distribution of the empirical exponential family log likelihood ratios becomes increasingly close to that of the empirical log likelihood ratio W as we increase q, the number of estimating equations used; the distributions are already very close for $q = 2$ and $q = 3$.

The robust Wald statistics are included for comparison with the corresponding likelihood ratios. We have seen that the asymptotic expansions of empirical likelihood and empirical exponential family likelihood are in close agreement and the results in Tables 1 and 2 indicate that the small-sample performance of the likelihood ratio statistics is equally close. This agreement does not extend, however, to the Wald statistics. The statistic C_1 for the empirical exponential family for $q = 1$ is uniformly the best performing statistic; however, the robust Wald statistics for empirical likelihood and the empirical exponential family for $q = 2$ perform poorly.

7. Conclusions

We discussed in Section 5 how the construction of both empirical likelihood and empirical exponential family likelihood involves the reduction of the infinite-dimensional problem of nonparametric estimation of a likelihood for the mean of an underlying distribution F to a finite-dimensional one by considering multinomial distributions supported on the sample. In this section we show how empirical likelihood and empirical exponential family likelihood may be placed in the same general framework. If our sample is $\mathbf{X} = \{X_1, \ldots, X_n\}$ and a multinomial distribution supported on that sample is $\{p_i\}$, we estimate the distribution function F of X by $\hat{F}_p(x) = \prod p_i \, 1_{\{X_i \le x\}}$. If the parameter of interest is the mean θ, we require that $\{p_i\}$ satisfies (1.2), and a nonparametric likelihood for θ is then $L(\theta) = \prod p_i(\theta)$. Normalizing $L(\theta)$, we define a generalized empirical log likelihood by

$$\ell_{gen}(\theta) = \log\{L(\theta)\} - \log\left\{L(\hat{\theta})\right\} = \sum \log\{np_i(\theta)\}, \tag{7.1}$$

where $\{p_i(\theta)\}$ are defined as the solutions of (1.2). We need further information about the probabilities $\{p_i\}$ to determine such a solution. Empirical likelihood uses a one-dimensional subfamily $p_i(\alpha) = n^{-1}\{1 + \alpha(X_i - \theta)\}^{-1}$ which

ensures that the probabilities are uniquely determined for every θ in the range $\min\{X_i\} < \theta < \max\{X_i\}$ and a solution to (1.2) exists in this range. A motivation for using this subfamily is that such a family $\{p_i\}$ minimizes the forward Kullback–Leibler discrepancy $-\sum n^{-1}\log(np_i)$ between two multinomial distributions which are solutions of (1.2), $\{p_i\}$ at the value θ and $\{n^{-1}\}$ at the value $\hat{\theta}$. This procedure normalizes the log likelihood in a similar fashion to the way the normed parametric log likelihood compares the log likelihood value at θ with the maximum likelihood value at $\hat{\theta}$. This is an attractive choice of discrepancy measure since the maximum likelihood estimate is a natural estimate for parameters which minimize the Kullback–Leibler discrepancy when the underlying distribution is unknown (Akaike (1973)). Similarly, the empirical exponential family likelihood is obtained using the backward Kullback–Leibler discrepancy $-\sum p_i \log(np_i)$. Akaike (1985) discusses the choice of direction of the Kullback–Leibler discrepancy in parametric problems.

Although the Kullback–Leibler discrepancy is an attractive choice for obtaining probabilities $\{p_i\}$, it is also an essentially arbitrary choice. Indeed, the form of the empirical exponential family probabilities used in the papers by Efron (1981), DiCiccio and Romano (1990) and Davison, Hinkley and Worton (1992), $p_i \propto e^{\alpha(X_i-\theta)}$, is given *a priori* as a reasonable choice and is not motivated as minimizing the backward Kullback–Leibler discrepancy.

We have used empirical likelihood as a baseline for asymptotic comparison because the literature shows that it has attractive theoretical properties. However, its small-sample performance has not been shown to be optimal, and indeed a simple robust Wald statistic constructed using the empirical exponential family likelihood consistently gave better empirical coverage than other nonparametric likelihood statistics.

References

Akaike, H. (1973). Information theory and an extension of the likelihood principle. In *Proc. 2nd Inter. Symp. Inform. Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiado, Budapest.

Akaike, H. (1985). Prediction and entropy. In *A Celebration of Statistics - the ISI Centenary Volume* (Edited by A. C. Atkinson and S. E. Fienberg). Springer-Verlag, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

Davison, A. C., Hinkley, D. V. and Worton, B. J. (1992). Bootstrap likelihoods. *Biometrika* **79**, 113-130.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1991) Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053-1061.

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

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with Discussion). *Canad. J. Statist.* **9**, 139-172.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Jing, B.-Y. and Wood, A. T. A. (1996) Exponential empirical likelihood is not Bartlett correctable. *Ann. Statist.* **24**, 365-369.

Kent, J. T. (1982). Robust properties of the likelihood ratio test. *Biometrika* **69**, 19-27.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79-86.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-49.

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

Owen, A. B. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725-1747.

Qin, J. and Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-325.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proc. 3rd Berkeley Symp. Math. Statist. Probab.* (Edited by J. Neyman) **1**, 187-196. University of California Press, Berkeley, CA.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland.
E-mail: Anthony.Davison@epfl.ch

(Received November 1997; accepted May 1999)