Adjusted empirical likelihood and its properties

By Jiahua Chen, Asokan Mulayath Variyath and Bovas Abraham

Department of Statistics and Actuarial Science, University of Waterloo

Waterloo, ON, N2L 3G1, Canada

jhchen@uwaterloo.ca mvasokan@uwaterloo.ca babraham@uwaterloo.ca

Summary

Computing profile empirical likelihood function is a key step in applications of empirical likelihood which involves constrained maximization. However, in some situations, solutions to the corresponding constraints may not exist. In this case, the convention is to assign a zero value to the profile empirical likelihood. This convention has at least two limitations. First, it is numerically difficult to determine the non-existence of any solution; Second, it provides no information on the relative plausibility of these parameter values. In this paper, we use a novel adjustment to the empirical likelihood so that the new method retains all the optimal properties, while guarantees a sensible value at any parameter point. Coupled with this adjustment, we introduce an iterative algorithm with guaranteed convergence. Our simulation indicates that the adjusted empirical likelihood is much faster to compute. The confidence regions constructed via the adjusted empirical likelihood are found to have closer to nominal coverage probabilities without resorting to more complex procedures such as Bartlett correction or bootstrap caliberation. Through some application examples, the method is also shown to be very effective in solving some practical problems associated with the use of empirical likelihood.

Some key words: Algorithm, confidence region, constrained maximization, coverage probability, variable selection.

1. Introduction

Since the pioneering work by Owen (1988), Qin & Lawless (1994) and others, the empirical likelihood methodology has quickly become a very powerful and widely applicable non-parametric and semi-parametric tool in statistical inference. In this approach, the parameters are usually defined as functionals of the population distribution. The profile empirical likelihood function of these parameters share many nice properties with the parametric likelihood functions without restrictive model assumptions.

Assume we have a set of independent and identically distributed (iid) vector valued observations $y_1, y_2, \ldots, y_n$ from an unknown distribution function $F(y)$. The problem of interest is to make inference on some q-dimensional parameter $\theta = \theta(F)$ defined as the unique solution to some estimation equation $E\{g(Y;\theta) : F\} = 0$ where $g(\cdot)$ is an $m \geq q$ dimensional function and the expectation is taken under the distribution F. For example, instead of assuming that F is a member of Poisson distribution family and θ is the corresponding mean, a semi-parametric model assumption takes F as a distribution having finite first two moments with equal mean and variance. Hence, the parameter in this semi-parametric model is specified by estimating functions

$$g_1(Y,\theta) = Y - \theta; \quad g_2(Y,\theta) = Y^2 - \theta - \theta^2.$$

In this example, we have $m = 2 > q = 1$.

The empirical likelihood function of F is defined as

$$L_n(F) = \prod_{i=1}^{n} p_i$$

with $p_i = F(\{y_i\}) = \Pr(Y_i = y_i)$ when there are no ties in observations. However, this definition is also applicable even when there are tied observations (Owen, 2001). Without any further information on the distribution F, the empirical likelihood is maximized when

F is the empirical distribution function

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(y_i \leq y)$$

where $I(\cdot)$ is the indicator function, and the inequality is interpreted component-wise. In general, it is more convenient to work with the logarithm of the empirical likelihood

$$l_n(F) = \sum_{i=1}^{n} \log(p_i). \tag{1}$$

We further require $\sum_{i=1}^{n} p_i = 1$, see also Owen (2001) for justifications.

Let θ be defined by the estimation equation $E\{g(Y;\theta) : F\} = 0$. In Qin & Lawless (1994) or Owen (2001), the profile log-empirical likelihood function of θ is defined as

$$l_{EL}(\theta) = \sup\{l_n(F) : \ p_i \geq 0, \ i = 1, \ldots, n; \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i g(y_i, \theta) = 0\}. \tag{2}$$

When the model is correct and some moment conditions on F are satisfied, the profile log-empirical likelihood function has many familiar optimal properties similar to its parametric sibling. In particular, when θ is defined by $g(Y, \theta) = Y - \theta$, the profile log-empirical likelihood functions $l_{EL}(\theta)$ can be used to construct asymptotic confidence regions for θ conveniently. Such confidence regions are valued for its data driven shape and range respecting properties.

In order to solve for the numerical problem related to $l_{EL}(\theta)$, a pre-requisite is that the convex hull of $\{g(y_i, \theta), i = 1, 2, \ldots, n\}$ must have vector 0 as its interior point. In the simple example of population mean, $l_{EL}(\theta)$ is well defined for all θ in the convex hull of $\{y_i, i = 1, 2, \ldots, n\}$. In some applications which we will discuss in more details, the dimension m of g can be much larger than the dimension q of the parameter θ. It can be difficult to determine the parameter region Θ over which the $l_{EL}(\theta)$ is well defined. In fact, it is also possible that Θ is an empty set. A convention is to define $l_{EL}(\theta) = -\infty$ for $\theta \notin \Theta$. However, this convention has two drawbacks: 1. it is often difficult to specify the

region Θ which is data specific; 2. When $\theta_j \notin \Theta$ for $j = 1, 2$, the values $l_{EL}(\theta_j)$ provide no information on their relative plausibility, even if one is very close to and the other is far away from the boundary of Θ. In particular, the second drawback makes it a challenge to find the maximum point of $l_{EL}(\theta)$. Even finding a proper initial value can be a difficult task in this case.

In this paper, we make a novel adjustment to the commonly used empirical likelihood. With this adjustment, the profile adjusted empirical likelihood is well defined for all parameter values. Hence, it eliminates the problem of determining whether $\theta \in \Theta$. Because of this, finding the maximum point of the adjusted $l_{EL}(\theta)$ becomes a much simpler problem. Further, the new solution also provides a good initial value when Θ for the original problem is non-empty. We also show that the asymptotic properties of the empirical likelihood are preserved. The adjusted empirical likelihood confidence region based on chi-square limiting distribution is generally larger. Hence, it has higher coverage probability particularly when the sample size is small and provides a promising solution to the small sample under-coverage problem discussed in Tsao (2004). As it will be seen in our simulation, the improved coverage probability is achieved without resorting to more complex techniques such as Bartlett-correction or bootstrap calibration. In fact, the algorithm for the adjusted empirical likelihood is much quicker to converge. Finally, a numerical algorithm of the profile adjusted empirical likelihood is given with guaranteed convergence.

The rest of the paper is organized as follows. In Section 2, we give additional necessary details of empirical likelihood and its extensions. In Section 3, we introduce our new method. The asymptotic properties of the new method are discussed in Section 4. In Section 5, we present a simple algorithm for computing the adjusted profile log-empirical likelihood ratio function. We demonstrate the usefulness of the new method by simulation and by some application examples. We draw some conclusions in Section 6. Some theoretical proofs are given in an Appendix.

2. Empirical likelihood

As in the introduction, let $y_1, y_2, \ldots, y_n$ be a set of independent and identically distributed (iid) vector valued observations. Let $g(Y, \theta)$ be the estimating function which defines the parameter θ of the population distribution F through $E\{g(Y, \theta)\} = 0$. The log-empirical likelihood $l_n(F)$ is defined by (1) and the profile log-empirical likelihood $l_{EL}(\theta)$ is defined as in (2). We denote Θ be the set of parameter values such that for each $\theta \in \Theta$, the solution to

$$\sum_{i=1}^{n} p_i g(y_i, \theta) = 0 \tag{3}$$

for p_i's exist. For each $\theta \in \Theta$, $l_n(F)$ under constraint (3) is maximized when

$$\hat{p}_i = \frac{1}{n\{1 + \lambda^\tau g(y_i, \theta)\}},$$

for $i = 1, 2, \ldots, n$ with the Lagrange multiplier λ being the solution of

$$\sum_{i=1}^{n} \hat{p}_i g(y_i, \theta) = 0.$$

Hence, we also have the following expression for profile log-empirical likelihood function,

$$l_{EL}(\theta) = -n \log(n) - \sum_{i=1}^{n} \log\{1 + \lambda^\tau g(y_i, \theta)\}.$$

We further define the profile empirical log-likelihood ratio function,

$$W(\theta) = \sum_{i=1}^{n} \log(n\hat{p}_i) = -\sum_{i=1}^{n} \log\{1 + \lambda^\tau g(y_i, \theta)\}.$$

For the special case when $g(y, \theta) = y - \theta$, Owen (1990) shows that, when θ_0 is the true population mean and under some moment conditions, $-2W(\theta_0) \to \chi^2_m$ in distribution as $n \to \infty$ similar to the result for the parametric likelihood (Wilks, 1938). This result is the foundation for hypothesis test on θ and the construction of confidence regions. An approximate $100(1 - \alpha)\%$ region is given by

$$\{\theta : -2W(\theta) \leq \chi^2_m(1 - \alpha)\} \tag{4}$$

where $\chi^2_m(1-\alpha)$ is the $(1-\alpha)^{th}$ quantile of the chi-square distribution with m degrees of freedom. Unlike the confidence region constructed based on normal approximation, the empirical likelihood regions have data driven shape, are range respecting, and often has better coverage properties (Owen, 1990; Chen et al., 2003).

The results on population mean are more general. Similar conclusions are found true for linear models, generalized linear models, models defined by estimating equations and many others (Owen 1991; Kolaczyk 1994; Qin & Lawless 1994). In applications, we must solve the problem of computing $W(\theta)$ at various θ values.

3. The constraint problem and the adjustment

Computing $W(\theta)$ numerically is usually done by solving

$$\sum_{i=1}^{n} \frac{g(y_i,\theta)}{1+\lambda^\tau g(y_i,\theta)} = 0 \tag{5}$$

for the Lagrange multiplier λ. The solution of λ we look for must satisfy $1+\lambda^\tau g(y_i,\theta) > 0$ for all $i = 1, 2, \ldots, n$. A necessary and sufficient condition for its existence is that the vector '0' is an inner point of the convex hull of $\{g(y_i,\theta), i = 1, 2, \ldots, n\}$.

By definition, the true parameter value θ_0 is the unique solution of $E\{g(Y;\theta) : F\} = 0$. Hence, under some moment conditions on $g(Y,\theta)$ (Owen, 2001), the convex hull $\{g(y_i,\theta_0), i = 1, 2, \ldots, n\}$ contains 0 as its inner point with probability 1 as $n \to \infty$. When θ is not close to θ_0, or when n is small, there is a considerable chance that the solution to (5) does not exist. This can be a serious limitation in some applications as shown in examples to be presented in the next section. In this section, we propose the following adjusted empirical likelihood.

Denote $g_i = g_i(\theta) = g(y_i;\theta)$ and $\bar{g}_n = \bar{g}_n(\theta) = n^{-1}\sum_{i=1}^n g_i$ for any given θ. For some positive constant a_n, define

$$g_{n+1} = g_{n+1}(\theta) = -\frac{a_n}{n}\sum_{i=1}^{n} g_i = -a_n\bar{g}_n.$$

We now adjust the profile empirical log-likelihood ratio function to

$$W^*(\theta) = \sup\{\sum_{i=1}^{n+1} \log[(n+1)p_i] : \ p_i \geq 0, \ i = 1, \ldots, n+1; \ \sum_{i=1}^{n+1} p_i = 1; \ \sum_{i=1}^{n+1} p_i g_i = 0\}, \quad (6)$$

Since the convex hull of $\{g_i, i = 1, 2, \ldots, n, n+1\}$ for any given θ contains 0, $W^*(\theta)$ is well defined without exceptions.

It is seen that if $\bar{g}_n(\theta) = 0$, we have $W^*(\theta) = W(\theta) = 0$. For θ values such that $\bar{g}_n \approx 0$, we have $W(\theta) \approx W^*(\theta)$. When $\theta \notin \Theta$, we have $W(\theta) = -\infty$ while $W^*(\theta)$ is still well defined, and is likely to assume a negative value with its magnitude depending on how far θ deviates from Θ. Hence, $W^*(\theta)$ is informative for searching Θ in which $W(\theta)$ is finite.

The value of a_n should be chosen to fit the problem of the user's particular application. In theory, the first order asymptotic property of $W(\theta_0)$ is unchanged for $W^*(\theta_0)$ as long as $a_n = o_p(n^{2/3})$. When $a_n = O_p(n^{1/2})$ and $\theta = \theta_0$, we have $g_{n+1} = O_p(1)$ under mild moment conditions. Thus, the effect of our adjustment is very mild: it is equivalent to adding a few artificial but comparable observations to the set of n real observations. In general, we recommend the use of a_n with smaller magnitude.

When θ is far from θ_0, some g_i's values can be much larger than the rest of g_i values. In this case, the $\bar{g}_n$ can substantially deviate from 0 and distort the true likelihood configuration around θ. When the semi-parametric model is correct, this problem will not occur if a good initial estimate of θ_0 is available. The profile likelihood ratio function will be very low around θ compared to the neighborhood of θ_0. When the model is not correct, which happens in the model selection example as in next section, we do not have a hypothetical θ_0 to rely on. In this case, our strategy is to replace $\bar{g}_n$ by the median or trimmed mean of g_i's. The particular form of $\bar{g}_n$ can be chosen by the user.

In most applications, our focus is on θ in a small neighborhood of θ_0 so that $\bar{g}_n$ is of mild size. Combining these considerations together, our general recommendation is to have $a_n = \max(1, \log(n)/2)$ coupled with trimmed version of $\bar{g}_n$ when appropriate. Investigation

of optimal a_n under various practically importance situations is still underway. This choice works well in our simulations to be presented. When the sample size n increases, our estimate of θ_0 will get into $n^{-1/2}$ range, hence with this choice, we have $a_n\bar{g}_n = o_p(1)$. The effect of this adjustment is well below the order of $n^{-1/2}$. Yet when n is small, this adjustment is effective at improving the coverage probability of confidence regions.

We now give a simple example to illustrate the convex hull problem and our adjustment. We generate 50 observations from independent bivariate standard normal distribution. We compute the profile likelihood at $(\mu_1, \mu_2) = (2, 2)$. Figure 1 (left) gives the plot of g values and it is seen that the convex hull does not contain 0. By adding an artificial observation $g_{n+1} = -a_n\bar{g}_n$ with $a_n = \log(n)/2$, the convex hull is expanded and 0 is now an inner point as in Figure 1 (right). Since $(2, 2)$ is way out of the convex hull, the adjustment in this case appears to be substantial. However, the log-empirical likelihood is only adjusted from $-\infty$ to some finite big negative number, and it does not really have a big impact. At the same time, one will not notice anything substantial on the adjustment of empirical likelihood at $(\mu_1, \mu_2) = (0, 0)$.

4. Asymptotic properties

The most impressive result in the content of empirical likelihood is the asymptotic limiting distribution of the log-empirical likelihood ratio function at the true parameter value. We now show that the proposed adjusted profile empirical likelihood has the same asymptotic properties as the unadjusted empirical likelihood.

Theorem 1 *Let $y_1, y_2, \ldots, y_n$ be a set of independent and identically distributed vector observations of dimension q from some unknown distribution F_0. Let θ_0 is the true parameter that satisfies $E\{g(Y, \theta) : F_0\} = 0$ where g is a vector valued function with dimension m. Assume further that $Var\{g(Y, \theta) : F_0\}$ is finite and has rank $m > q$. Let $W^*(\theta)$ be the adjusted log-empirical likelihood ratio function defined by (4) and $a_n = o_p(n^{2/3})$. As $n \to \infty$,*

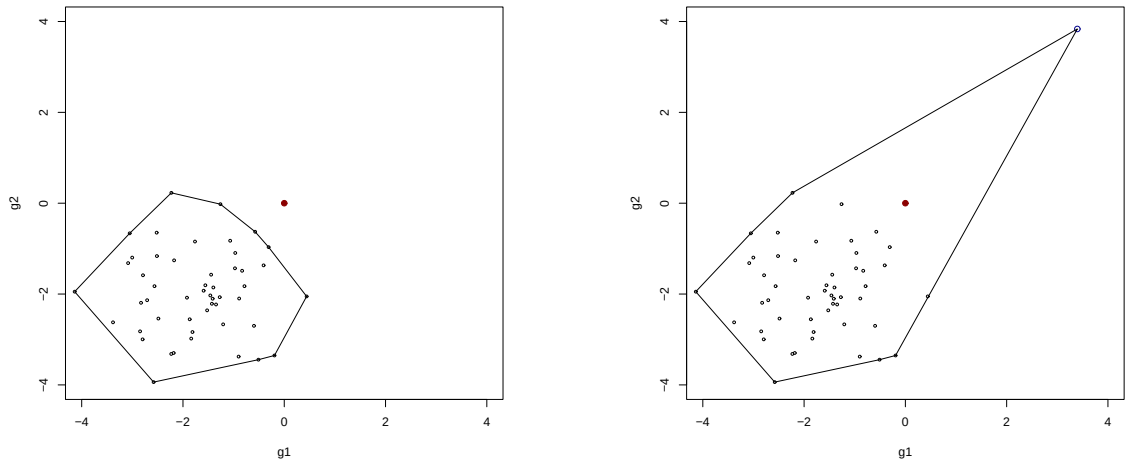

Figure 1: Convex hull (left) and adjusted convex hull with $a_n = \log(n)/2$ (right). The shaded circle is (0,0)

we have

$$-2W^*(\theta_0) \to \chi^2_m$$

in distribution.

If the rank of $\mathrm{Var}\{g(Y,\theta) : F_0\}$ is lower than m, then some components of the estimating function g can be removed from the constraint set and the above conclusion can be revised accordingly. This result can be compared to Theorem 3.4 in Owen (2001). Based on this result, an empirical likelihood based confidence interval (region) as discussed in the introduction can be easily constructed. When θ is the population mean, the confidence region using adjusted profile empirical likelihood is no longer confined within the convex hull of the observed values. Thus, the new method has the potential to improve the under-coverage problem due to small sample size or high dimension effectively. The proof of this theorem is given in Appendix.

We are naturally interested in knowing the asymptotic behavior of $W^*(\theta)$ when $\theta \neq \theta_0$, and that of $W(\theta)$. Interestingly, this problem failed to draw much attention in the literature

so far.

THEOREM **2** *Assume that all the conditions as in Theorem 1 and that for some* $\theta \neq \theta_0$

$$\|E\{g(Y,\theta)\}\| > 0.$$

Then, we have both $-2n^{-1/3}W^*(\theta) \to \infty$ *and* $-2n^{-1/3}\tilde{W}(\theta) \to \infty$ *in probability as* $n \to \infty$.

The proof of Theorem 2 is also given in Appendix. Based on Theorems 1 and 2, it is easily seen that the non-parametric maximum empirical likelihood estimator is consistent with some minor additional conditions. In addition, when θ is not the true value, the empirical likelihood ratio statistic tends to infinity at the rate of at least $n^{1/3}$. Qin & Lawless (1994) showed that there exists a local maximum of $W(\theta)$ in an $n^{-1/3}$ neighborhood of θ_0. The same result can be proved here. Since the proof of similar results for $W^*(\theta)$ does not contain new techniques, we skip the proof and only selectively present a result as follows.

THEOREM **3** *In addition to all the conditions as in Theorem 1, we assume that* $\frac{\partial^2 g(y,\theta)}{\partial\theta\partial\theta^\tau}$ *is continuous in the neigbourhood of* θ_0. *Also we assume that* $\|g(y,\theta)\|^3$ *and* $\|\frac{\partial^2 g(y,\theta)}{\partial\theta\partial\theta^\tau}\|$ *can be bounded by some integral function* $G(y)$ *in the neigbourhood of* θ_0. *Let* $\hat{\theta}$ *be a local maximum of* $W^*(\theta)$ *in an* $n^{-1/3}$ *neighborhood of* θ_0. *As* $n \to \infty$, *we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma)$$

in distribution, where

$$\Sigma = \{E(\frac{\partial g}{\partial\theta})^\tau \ (Egg^\tau)^{-1} \ E(\frac{\partial g}{\partial\theta})\}^{-1}.$$

5. NUMERICAL ALGORITHM, SIMULATIONS AND APPLICATION EXAMPLES

Since the constraints in the adjusted empirical likelihood are always satisfied, the numerical computation of the profile adjusted empirical likelihood can be done with any existing

algorithms. In particular, we suggest the use of modified Newton-Raphson algorithm proposed by Chen et al. (2001). To maximize $W^*(\theta)$ with respect to θ, we recommend the simplex method introduced by Nelder & Mead (1965). This method is numerically very stable which is important in this application. Most optimization softwares include built-in functions for this method.

For given θ and a chosen a_n, we compute $W^*(\theta)$ as described in the following pseudo code:

1. Compute $g_i = g(y_i, \theta)$ for $i = 1, \ldots, n$ and $g_{n+1} = -a_n \bar{g}_n$. This sample mean $\bar{g}_n$ may be replaced by trimmed mean or any other robust substitutes.

2. Set initial value for the Lagrange multiplier $\lambda^0 = 0$. Initialize iteration number $k = 0$, and let $\gamma = 1$ and $\varepsilon = 10^{-8}$ for the reference of step size in iteration, and the tolerance level.

3. Compute the first and second partial derivatives of

$$R(\lambda) = \sum_{i=1}^{n+1} \log(1 + \lambda^\tau g_i)$$

with respect to λ evaluated at λ^k. Let them be $\dot{R}$ and $\ddot{R}$ and further compute $\Delta = -\ddot{R}^{-1}\dot{R}$.

If $\|\Delta\| < \varepsilon$ stop the iteration, report λ^k and go to Step 6.

4. Compute $\delta = \gamma\Delta$. If any $1 + (\lambda^k - \delta)^\tau g_i \leq 0$ or $R(\lambda^k - \delta) < R(\lambda^k)$, let $\gamma = \gamma/2$, and repeat this step. Otherwise, continue to the next step.

5. Let $\lambda^{k+1} = \lambda^k - \delta$ and $\gamma = (k+1)^{-1/2}$. Increase the count k by 1. Return to Step 3.

6. Report λ^k and the value of $W^*(\theta) = -\sum_{i=1}^{n+1} \log(1 + \lambda^k g_i)$.

The convergence of the iteration is guaranteed because the existence of a solution is assured. We refer to Chen et al., (2001) for the proof of the algorithmic convergence. The above operations result in the value $W^*(\theta)$. We can then use the simplex method to optimize this function. The only remaining consideration is to give a set of good initial values of θ. In general, one can try to have a rough estimate of θ using other methods. To be more concrete as well as to demonstrate the usefulness of the new method, we give some application examples.

5.1. Confidence region

Constructing the confidence region for population means received primary attention in the pioneering paper of Owen (1988). The empirical likelihood confidence regions have data driven shape, are range respecting, and Bartlett-correctable (DiCiccio et al., 1991). When the sample size is not large, one may also use bootstrap calibration to replace the calibration based on chi-square distribution, to improve the accuracy of the coverage probability. As discussed in Tsao (2004) however, the confidence region constructed based on the unadjusted empirical likelihood is by definition confined inside the convex hull formed by the observed values, which is not affected by Bartlett correction or by bootstrap calibration. When the sample size is small, even the convex hull may not have a large enough coverage probability. In comparison, $W^*(\theta)$ is finite for all θ. Thus, the confidence region is no longer confined within the convex hull of the data. It is hence better fitted to solve the under-coverage problem.

For illustration, we applied the new method to construct confidence intervals for examples discussed in DiCiccio et al., (1991). We followed their, parameter, and simulation set-up. The coverage probabilities are based on 5000 simulations. The data were drawn from the standard normal distribution with sample sizes 10 and 20, a $\chi^2_{(1)}$ distribution with sample sizes 20 and 40, and a $t_{(5)}$ distribution with sample sizes 15 and 30. We let

$a_n = \log(n)/2$ in the definition of g_{n+1}. Our simulations show that value corresponding to the unadjusted EL are very close to those in DiCiccio et al., (1991). We use the Bartlett-corrected results from their paper for comparison. The coverage probabilities of intervals based on unadjusted, Bartlett-corrected, and our new method are given in Table 1 for nominal levels of 80, 90, 95 and 99 percent. It is clear that the coverage probabilities from the new method are closer to target values for the sample sizes and population distributions considered. The results are particularly impressive since the new method does not involve any complex theory or computational procedures. We also conducted simulations with relatively large sample sizes ($n = 100, 200, 500$), and the coverage probabilities were close to the target values for all methods.

We simulated the coverage probabilities of a bivariate population mean. We repeated the simulation 5000 times with $a_n = \log(n)/2$ for bivariate distribution of the standard normal, $\chi^2_{(1)}$ and $t_{(5)}$ with independent components. The coverage probabilities are reported in Table 2. It is clear that the adjusted empirical likelihood compares very favourably with the unadjusted usual empirical likelihood. The coverage probabilities are substantially improved with the adjusted EL method. Our simulations also reveal that the adjusted empirical likelihood took much less time to complete. The saving was found to be in the order of 5-fold to 7-fold.

To provide a more concrete comparison between the confidence regions constructed based on the unadjusted and the adjusted empirical likelihood, we considered a data set given in Owen (2001, p. 31). The original source of this data set is Iles (1993). This data set consists of four types of prey (Caddis fly larvae, Stonefly larvae, Mayfly larvae, and other invertebrates) of Dippers (Cinclus cinclus) found at 22 different sites along the river Wye and its tributaries in Wales. We constructed a 95% confidence region for the mean of (caddis fly larvae, stonefly larvae) and (mayfly larvae, other invertebrates) based on the adjusted and the unadjusted EL. These are given in Figure 2 together with a scatter

Table 1: Coverage Probabilities of Population Mean

Normal Data

	$n = 10$				$n = 20$			
NV	0.80	0.90	0.95	0.99	0.80	0.90	0.95	0.99
EL	0.7396	0.8318	0.8940	0.9526	0.7802	0.8756	0.9284	0.9794
TB	0.7796	0.8706	0.9182	0.9650	0.8006	0.8962	0.9416	0.9844
EB	0.7938	0.8802	0.9246	0.9696	0.8034	0.8980	0.9424	0.9848
AEL	0.7964	0.8892	0.9444	0.9962	0.8138	0.9028	0.9522	0.9898

$\chi^2_{(1)}$ Data

	$n = 20$				$n = 40$			
EL	0.7332	0.8354	0.8928	0.9524	0.7682	0.8640	0.9170	0.9742
TB	0.7872	0.8772	0.9262	0.9706	0.7910	0.8896	0.9418	0.9800
EB	0.7634	0.8546	0.9034	0.9616	0.7804	0.8789	0.9334	0.9774
AEL	0.7714	0.8652	0.9168	0.9660	0.7930	0.8810	0.9330	0.9818

$t_{(5)}$ Data

	$n = 15$				$n = 30$			
EL	0.7544	0.8504	0.9098	0.9674	0.7784	0.8834	0.9338	0.9812
TB	0.8266	0.9106	0.9544	0.9862	0.8114	0.9042	0.9496	0.9866
EB	0.7898	0.8884	0.9348	0.9794	0.7954	0.8928	0.94222	0.9832
AEL	0.7986	0.8944	0.9418	0.9876	0.8098	0.9070	0.9500	0.9874

NV = nominal value; EL = empirical likelihood; TB = EL with theoretical Bartlett correction; EB = EL with estimated Bartlett correction; AEL = Adjusted EL.

Table 2: Coverage Probabilities of Bivariate Population Mean

	$n = 20$				$n = 50$			
				Normal Data				
NV	0.80	0.90	0.95	0.99	0.80	0.90	0.95	0.99
EL	0.7466	0.8536	0.9104	0.9674	0.7836	0.8878	0.9400	0.9868
AEL	0.7998	0.8986	0.9458	0.9882	0.8080	0.9072	0.9528	0.9922
				$\chi^2_{(1)}$ Data				
EL	0.6702	0.7785	0.8449	0.9188	0.7476	0.8524	0.9106	0.9682
AEL	0.7248	0.8290	0.8836	0.9462	0.7764	0.8746	0.9256	0.9748
				$t_{(5)}$ Data				
EL	0.7150	0.8214	0.8862	0.9600	0.7576	0.8680	0.9300	0.9826
AEL	0.7762	0.8750	0.9344	0.9886	0.7854	0.8880	0.9442	0.9884

NV = nominal value; EL = empirical likelihood; AEL = Adjusted EL.

plot of the original data points. It can be seen that the confidence regions based on the adjusted empirical likelihood with $a_n = \log(n)/2$ retain the data-driven shape and contain the confidence regions based on the unadjusted empirical likelihood.

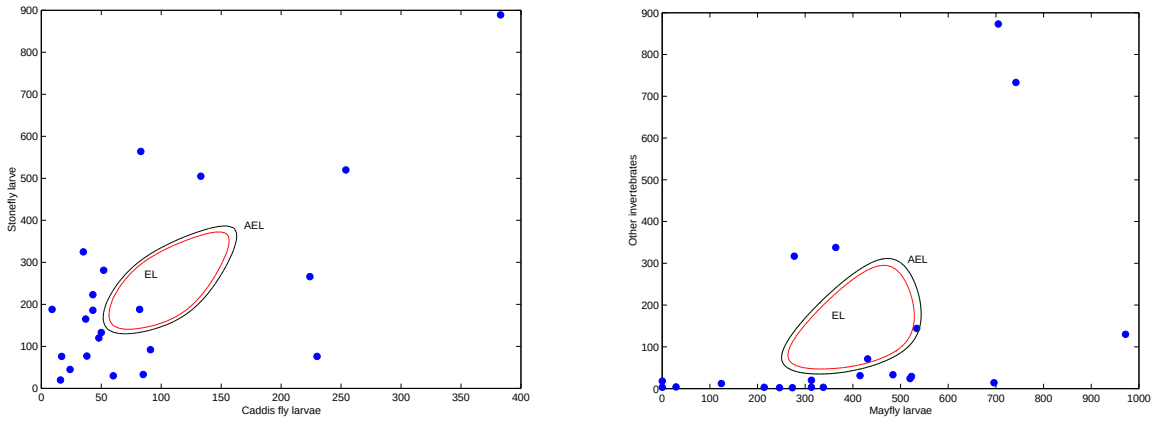

Figure 2: Comparison of confidence regions of based on EL and AEL

5.2. Estimation of a covariance matrix with known zero entries

Chaudhuri et al., (2006) discussed the problem of estimating covariance matrix of a random vector with many known zero entries. Such restrictions appear in many applications, Grzebyk et al., (2004). We refer the readers to Chaudhuri et al., (2006) for motivation and more references.

Suppose we have a random vector $Y = (Y_1, Y_2, Y_3, Y_4)^\tau$ whose covariance matrix Σ has the form

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & \sigma_{13} & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} \\ \sigma_{13} & 0 & \sigma_{33} & \sigma_{34} \\ 0 & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix}.$$

Under normality assumption, the mean $\mu = E\{Y\}$ and the variance can both in principle be estimated by the maximum likelihood estimator. However, maximizing the likelihood

function under the constraint of some covariaces being zero is not a simple problem. Chaudhuri et al., (2006) presented many numerical algorithms for this problem. One method is to ignore the normality assumption and to use the empirical likelihood instead. Let y_{ij}, $i = 1, \ldots, n$ and $j = 1, 2, 3, 4$ be a set of n observations from a distribution F with mean μ and covariance matrix Σ. They proposed computing the profile empirical likelihood at a given feasible μ and Σ:

$$\begin{aligned} l_{EL}(\mu, \Sigma) &= \sup\{\sum_{i=1}^{n} \log p_i : \ p_i > 0; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i (y_{ij} - \mu_j) = 0; \\ &\quad \sum_{i=1}^{n} p_i (y_{ij} - \mu_j)(y_{jk} - \mu_k) = 0; (j,k) = (1,2), (2,4), (3,4)\}. \end{aligned}$$

The maximum empirical likelihood estimates of μ and Σ are taken as the maximum points of $l_{EL}(\mu, \Sigma)$. It was found that when the normal model is true, the maximum empirical likelihood estimator has almost the same efficiency as to the parametric maximum likelihood estimator. When the normality is violated, the maximum empirical likelihood estimator is more efficient.

When the sample size is small (n=10), they experienced problems with the empirical likelihood procedure due to the inability to find feasible starting values. With the adjusted empirical likelihood, however, this problem disappears immediately. According to our own simulation and also private communication, the minimizer of $W^*(\mu, \Sigma)$ provides initial values for maximizing $l_{EL}(\mu, \Sigma)$ by letting $a_n = n^{-1}$. One may also start with larger a_n and reduce its size gradually. This approach has been found to be effective in providing good initial values for the computation of $l_{EL}(\mu, \Sigma)$. More interestingly, the minimizer of $W^*(\mu, \Sigma)$ itself serves as an efficient estimate. We demonstrate this point in the following simulation study.

We simulated data sets $Y = (Y_1, Y_2, Y_3, Y_4)^\tau$ from the multivariate normal distribution

with zero mean vector and covariance matrix Σ

$$\Sigma = \begin{pmatrix} 1 & 0 & 0.375 & 0 \\ 0 & 1 & 0 & 0.165 \\ 0.375 & 0 & 1 & 0.65 \\ 0 & 0.165 & 0.65 & 1 \end{pmatrix}.$$

We let the sample size $n = 10$ and generated 1000 sets of random samples from the normal model. By using the sample mean as the initial value for μ, we failed to find a solution for the linear constraints in 75.9% of the samples. We next computed the adjusted empirical likelihood with $a_n = n^{-1}$. By definition, there is no problem with the existence of the solutions in this case. The maximum adjusted empirical likelihood mean estimator was then used to provide as initial values for the unadjusted EL. This time, we failed to find a solution in only 17.4% of the samples. Thus, the adjusted empirical likelihood is useful as an initial value locator even if one insists on using the original empirical likelihood.

Further, we examined the maximum adjusted empirical likelihood estimator by its own merit. We computed the total bias and root mean square error (RMSE) of the maximum unadjusted and adjusted empirical likelihood estimates according to the following definitions:

$$\text{Total bias} = \sum_{i<j} |\frac{1}{M} \sum_{k=1}^{M} (\hat{\sigma}_{ij}^{(k)} - \sigma_{ij})|$$

and

$$RMSE = \sqrt{\sum_{i<j} \frac{1}{M} \sum_{k=1}^{M} (\hat{\sigma}_{ij}^{(k)} - \sigma_{ij})^2}$$

where $\hat{\sigma}_{ij}^{(k)}$ is the estimate of σ_{ij} in the kth simulation data set and M is the total number of simulations. For this comparison, we set $n = 20$ and the number of simulations $M = 10000$. For the unadjusted EL with $a_n = n^{-1}$, there were no solutions in 2.1% of the simulations due to non-existence of solutions. Hence, total bias and root mean square error were computed based on the rest 97.9% samples.

The bias of the maximum unadjusted and the adjusted EL estimators for the standard deviations were found to be 0.98 and 1.00. Clearly, these two methods have very similar bias properties. The RMSEs for the maximum unadjusted and adjusted EL were found to be 0.87 and 0.81 which implies that the adjusted EL method has about a 13% gain in efficiency. We should certainly be cautious in generalizing this conclusion and more studies will be needed. The key advantage of the adjusted EL method in this example is computational; the solution is guaranteed to exist.

In most applications, having a sample size as low as $n = 10$ may not be likely. However, when the dimension of y increases, the problem of finding feasible starting values can remain a serious challenge even for moderate to large sample sizes. Our method can both be used in searching for feasible starting values, and to directly provide efficient estimates of unknown parameters in this and similar application examples.

5.3. Variable selection in regression analysis

Assume we have n independent observations described by the following linear model

$$y_i = \beta_0 + x_i^\tau \beta + \epsilon_i$$

$i = 1, 2, \ldots, n$ such that ϵ_i are independent errors with mean 0 and finite nonzero variance σ^2. We denote the dimension of x as m. In applications, the covariate x_i has high dimension and not all components of the regression coefficient β are significantly different from 0. Thus, a variable selection step is often applied to reduce the complexity of the model and hence reduce the variability in estimators.

There are many well known variable selection procedures. The Akaike information criterion (AIC) and the Bayes information criterion (BIC) are among the most investigated methods in the literature (Akaike, 1973; Schwarz, 1978; Shao, 1997). In a simplistic description, both criteria choose a model specified by a subset of covariates by their penalized likelihood values. In general, a parametric distributional assumption for ϵ is required.

To avoid the parametric assumption, Kolaczyk (1994) and Variyath (2006) discussed the use of empirical likelihood based information criteria. Assume that we have a set of independent observations as given earlier. Let s be a subset of indices of covariate variables. We use $x_i[s]$ and $\beta[s]$ to denote the corresponding subset of covariates and regression coefficients. For each given s, the profile empirical likelihood is given by

$$\begin{aligned} l_{EL}(\beta_0, \beta[s], \sigma^2) &= \sup \left\{ \sum_{i=1}^{n} \log p_i : \ p_i > 0; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i (y_i - \beta_0 - x_i^\tau[s]\beta[s]) = 0; \right. \\ & \left. \sum_{i=1}^{n} p_i x_i^\tau (y_i - \beta_0 - x_i^\tau[s]\beta[s]) = 0. \right\}. \end{aligned}$$

Note that the number of constraints remains a constant with respect to s in the definition of the profile likelihood. This is very important in order to differentiate the plausibility of sub-models formed by including only a subset of covariates.

The profile empirical likelihood ratio function $W(\beta_0, \beta[s], \sigma^2)$ is similarly defined. For convenience, we will omit the entries of β_0 and σ^2 in this notation. The empirical likelihood based Akaike information criterion

$$EAIC(s) = 2 \inf\{W(\beta[s]) : \beta_0, \beta[s], \sigma^2\} + 2k$$

with k being the number of covariates in s. We choose the sub-model with corresponding s minimizes $EAIC(s)$. Similarly, we define

$$EBIC(s) = 2 \inf\{W(\beta[s]) : \beta_0, \beta[s], \sigma^2\} + k \log n.$$

The empirical version of BIC chooses the sub-model s that minimizes $BIC(s)$.

We refer to Kolaczyk (1994) and Variyath (2006) for specific discussion on the properties of EAIC and EBIC for variable selection. Before these methods can be used, one must solve a technical problem: make sure that both EAIC and EBIC are well defined.

Consider the most extreme case when s is empty, but the dimension of x, m, is relatively large combined with not so large n. In this case, the constraints implies that we must find

a value of β_0 such that

$$\begin{aligned} \sum_{i=1}^{n} p_i &= 1; \\ \sum_{i=1}^{n} p_i(y_i - \beta_0) &= 0; \\ \sum_{i=1}^{n} p_i x_i^{\tau}(y_i - \beta_0) &= 0 \end{aligned}$$

have solutions in p_i's. In the special case when $y_i = x_i$ and $m = 1$, it implies that $\beta_0^2 = (\sum p_i x_i)^2 = \sum p_i x_i^2$. This is possible only if p_i degenerates which implies $EAIC(s) = EBIC(s) = \infty$. With the adjusted empirical likelihood, the solution exists for any choice of β_0. Thus, sensible and informative values of EAIC(s) and BAIC(s) are always defined after the adjustment and can be easily computed.

We now consider the cancer study example given by Stamey et. al (1989) for variable selection problem. This study examines the correlation between the levels of prostate specific antigen(PSA) and 8 clinical measurements of 97 men who were yet to receive a radical prostatectomy. These clinical measurements considered here for the multiple linear regression of $\log(PSA)$ are logarithm of cancer volume (lcavol), logarithm of weight (lweight), age, logarithm of benign prostate hyperplasia amount(lbph), seminal vesicle invasion (svi), logarithm of capsular penetration (lcp), Gleason score (gleason), and percentage of Gleason score 4 or 5(pgg45). The aim of this analysis is to predict the logorithm of PSA based on the 8 covariates. First ordinary least square estimates were obtained. We apply variable selection methods AIC, BIC, EAIC and EBIC to identify the appropriate model. Defining the sub-models by setting the regression coefficients to zero for all covariates not in the sub-model, we face computational issue due the non-existence of solution to the equation $\sum_{i=1}^{n} p_i g(y_i, x_i, \beta) = 0$ for a given value of β. The adjusted empirical likelihood works smoothly in this example. Covariates lcavol, lweight and svi were identified to form an appropriate model by EAIC, BIC and EBIC whereas model identified by AIC includes an

Table 3: Variable Selection Problem - Prostate Cancer Data

Variables	OLS	AIC	EAIC/BIC/EBIC
Intercept	0.6694 (1.2963)	0.1456(0.5975)	-0.2681(0.5435)
lvavol	0.5870(0.0879)	0.54860.0741)	0.5516(0.0747)
lweight	0.4545(0.1700)	0.3909(0.1660)	0.5085(0.1502)
age	-0.0196(0.0112)	-	-
lbph	0.1071(0.0584)	0.0901(0.0562)	-
svi	0.7662(0.2443)	0.7117(0.2100)	0.6662(0.2098)
lcp	-0.1055(0.0910)	-	-
gleason	0.0451(0.1575)	-	-
pgg45	0.0045(0.0044)	-	-

additional variable lbph. These models, estimates of the corresponding regression parameters, the estimates of standard errors (in brackets), and the ordinary least square (OLS) estimates based on the full model are given in Table 3. We can see that all the methods picked up the most significant covariates. The AIC picked a slightly larger model where as EAIC does not.

6. Conclusions

In this paper, we suggest a method to overcome the difficulty posed by the non-existence of solutions while computing the profile empirical likelihood. We demonstrate that the proposed adjusted empirical likelihood is well defined for all parameter values. The new method substantially enhances the applicability of empirical likelihood. We show that the resulting adjusted empirical likelihood retains the optimal asymptotic properties. Further, the confidence regions constructed by the new method have closer to nominal coverage proba-

bilities in the examples considered. The algorithm associated with the adjusted empirical likelihood also converges much faster, reducing the computational burden. The usefulness of this new method is illustrated via a number of application examples.

Appendix

Because the proofs of some similar results are well known and can be easily found in Owen (2001), our proofs here will be very brief and somewhat simplistic.

Proof of Theorem 1.:

Let the eigenvalues of $\mathrm{Var}\{g(Y,\theta_0)\}$ be $\sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_m^2$. Without loss of generality, we assume $\sigma_1^2 = 1$. Let λ be the solution to

$$\sum_{i=1}^{n+1} \frac{g_i}{1+\lambda^\tau g_i} = 0. \tag{7}$$

We first show that $\lambda = O_p(n^{-1/2})$. For brevity, we claim that $\lambda = o_p(1)$ which is easy to verify. Our task is to refine this assessment.

Let $g^* = \max_{1\le i\le n} \|g_i\|$. The moment assumption implies,

$$g^* = o_p(n^{1/2}) \quad \text{and} \quad \bar{g}_n = O_p(n^{-1/2}).$$

Let $\rho = \|\lambda\|$ and $\hat{\lambda} = \lambda/\rho$. Multiplying $n^{-1}\theta^\tau$ to (7), we get,

$$\begin{aligned}
0 &= \frac{\hat{\lambda}^\tau}{n} \sum_{i=1}^{n+1} \frac{g_i}{(1+\lambda^\tau g_i)} \\
&= \frac{\hat{\lambda}^\tau}{n} \sum_{i=1}^{n+1} g_i - \rho \sum_{i=1}^{n+1} \frac{(\hat{\lambda}^\tau g_i)^2}{(1+\rho\hat{\lambda}^\tau g_i)} \\
&\le \hat{\lambda}^\tau \bar{g}_n (1 - a_n/n) - \frac{\rho}{n(1+\rho g*)} \sum_{i=1}^{n} (\hat{\lambda}^\tau g_i)^2 \\
&= \hat{\lambda}^\tau \bar{g}_n - \frac{\rho}{n(1+\rho g*)} \sum_{i=1}^{n} (\hat{\lambda}^\tau g_i)^2 + O_p(n^{-3/2} a_n).
\end{aligned} \tag{8}$$

The inequality above is valid because the $(n+1)$th term in of the second summation is non-negative. Consequently, the variance assumption on $Var\{g(Y,\theta_0)\}$ implies that

$$n^{-1} \sum_{i=1}^{n} (\hat{\lambda}^\tau g_i)^2 \ge (1-\epsilon)\sigma_1^2 = 1 - \epsilon$$

in probability for some $1 > \epsilon > 0$. Therefore, as long as $a_n = o_p(n)$, (8) implies

$$\frac{\rho}{(1+\rho g*)} \leq \hat{\lambda}^\tau \bar{g}_n \times (1-\epsilon)^{-1} = O_p(n^{-1/2})$$

which further implies $\rho = O_p(n^{-1/2})$ and hence $\lambda = O_p(n^{-1/2})$.

Next, denote $\hat{V}_n = n^{-1}\sum_{i=1}^n g_i^\tau g_i$, we find

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n+1} \frac{g_i}{1+\lambda^\tau g_i} \\ &= \bar{g}_n - \lambda^\tau \hat{V}_n + o_p(n^{-1/2}). \end{aligned}$$

Hence, when $n \to \infty$, $\lambda = \hat{V}_n^{-1}\bar{g}_n + o_p(n^{-1/2})$.

At last, we expand $\tilde{W}^*$ as follows.

$$\begin{aligned} -2\tilde{W}^*(\theta_0) &= 2\sum_{i=1}^{n+1} \log(1+\lambda^\tau g_i) \\ &= 2\sum_{i=1}^{n+1} \{\lambda^\tau g_i - (\lambda^\tau g_i)^2/2\} + o_p(1). \end{aligned}$$

Substituting the expansion of λ, we get,

$$-2\tilde{W}^*(\theta_0) = n\bar{g}_n^\tau \hat{V}_n^{-1}\bar{g}_n + o_p(1)$$

which converges to chi-square distribution with m degrees of freedom as $n \to \infty$.

Remark: when $\bar{g}_n$ is replaced by any other $O_p(n^{-1/2})$ random quantity in the definition of g_{n+1}, the above proof still goes through with no need of any changes. Even if a_n has larger order such that $a_n = o_p(n)$, the above proof still works. The assumption of $a_n = o_p(n^{2/3})$ makes the next proof simpler. Using a large a_n in most cases is not advisable.

Proof of Theorem 2:

Note again that $g_i = g(y_i, \theta)$, $i = 1, \ldots, n$ and similarly define $\bar{g}_n$ and g_{n+1}. By the law of large numbers, as $n \to \infty$, $\|\bar{g}_n^\tau \bar{g}_n\| \to \delta^2 > 0$ in probability. Note that $g_i - \bar{g}_n$ has mean zero, and satisfy all moment conditions to ensure that

$$\max\{\|g_i - \bar{g}\|\} = o_p(n^{1/2}).$$

Let $\tilde{\lambda} = n^{-2/3}\bar{g}M$ for some positive constant M. Hence, we have

$$\max\{|\tilde{\lambda}^\tau g_i|, i = 1, \ldots, n, n+1\} = o_p(1).$$

Thus, with probability going to one, $1 + \tilde{\lambda}^\tau g_i > 0$ for all $i = 1, \ldots, n, n+1$. Using the duality of the maximization problem, we find

$$\begin{aligned} \tilde{W}^*(\theta) &= -\sup_\lambda \{\sum_{i=1}^{n+1} \log(1 + \lambda^\tau g_i) \\ &\leq -\sum_{i=1}^{n+1} \log(1 + \tilde{\lambda}^\tau g_i) \\ &= -n^{1/3}\delta^2 M + o_p(1). \end{aligned}$$

Since M can be arbitrarily large, we have $-2n^{-1/3}W^*(\theta) \to \infty$ for any $\theta \neq \theta_0$.

Remark: First, the sample mean $\bar{g}_n$ can again be replaced by the sample median or trimmed means without invalidating the above proof. Second, the order of $W^*(\theta)$ tending to infinity is clearly higher than $n^{1/3}$. Since the result is useful mostly for obtaining asymptotic properties of other procedures, the exact order is not of great interest here.

References

Akaike, H. (1973). Information theory as a extension of the maximum likelihood principle. 267-282. Petrov, B. N. & Csaki, F. (eds.) *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest.

Chaudhuri, S., Drton, M. & Richardson, T. S. (2006). Estimation of a convariance matrix with zeros, Technical manuscript, Department of Statistics and Applied Probability, National University of Singapore, Singapore.

Chen, J., Chen, S. & Rao, J. N. K. (2003). Empirical likelihood confidence intervals for a population containing many zero values. *Canadian Journal of Statistics*, **31**, 53-67.

Chen, J., Sitter, R. R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89**,2 30-237.

DiCiccio, T., Hall, P. & Romano, J. (1991). Empirical likelihood is bartlett-correctable. *the Annals of Statistics*, **19**, 1053-1061.

Grzebyk, M., Wild, P. & Chouanière, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91**, 141-151.

Kolaczyk, E. D. (1994). Empirical likelihood for generalized linear models. *Statistica Sinica* **4**, 199-218.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minizatin. *Computer Journal* **7**, 308-313.

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

Owen, A. B. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, **18**, 90-120.

Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, **19**, 1725-1747.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman & Hall/CRC, New York.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300-325.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221-264.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, R. & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: Radical prostatectomy treated patients. *Journal of Urology* **16**, 1076–1083.

Tsao, M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *The Annals of Statistics*, **32**, 1215-1221.

Variyath, A. M. (2006). Variable selection in generalized linear models by empirical likelihood. Ph.D. proposal, University of Waterloo, Waterloo, Canada, 2006.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *The Annuals of Mathematical Statistics*, **9**, 60-62.