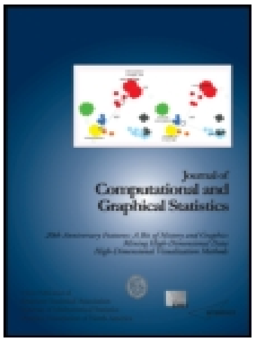

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Adjusted Empirical Likelihood and its Properties

Jiahua Chen, Asokan Mulayath Variyath & Bovas Abraham

To cite this article: Jiahua Chen, Asokan Mulayath Variyath & Bovas Abraham (2008) Adjusted Empirical Likelihood and its Properties, Journal of Computational and Graphical Statistics, 17:2, 426-443, DOI: 10.1198/106186008X321068

To link to this article: https://doi.org/10.1198/106186008X321068

Supplemental materials for this article are available online

Adjusted Empirical Likelihood and its Properties

Jiahua CHEN, Asokan Mulayath VARIYATH, and Bovas ABRAHAM

Computing a profile empirical likelihood function, which involves constrained maximization, is a key step in applications of empirical likelihood. However, in some situations, the required numerical problem has no solution. In this case, the convention is to assign a zero value to the profile empirical likelihood. This strategy has at least two limitations. First, it is numerically difficult to determine that there is no solution; second, no information is provided on the relative plausibility of the parameter values where the likelihood is set to zero. In this article, we propose a novel adjustment to the empirical likelihood that retains all the optimality properties, and guarantees a sensible value of the likelihood at any parameter value. Coupled with this adjustment, we introduce an iterative algorithm that is guaranteed to converge. Our simulation indicates that the adjusted empirical likelihood is much faster to compute than the profile empirical likelihood. The confidence regions constructed via the adjusted empirical likelihood are found to have coverage probabilities closer to the nominal levels without employing complex procedures such as Bartlett correction or bootstrap calibration. The method is also shown to be effective in solving several practical problems associated with the empirical likelihood.

Key Words: Algorithm; Confidence region; Constrained maximization; Coverage probability; Variable selection.

1. INTRODUCTION

Since the pioneering work by Owen (1988), Qin and Lawless (1994), and others, empirical likelihood (EL) methodology has become a powerful and widely applicable tool for nonparametric and semiparametric statistical inference. In this approach, the parameters are usually defined as functionals of the population distribution. The profile EL function exhibits many properties of the parametric likelihood functions without the associated restrictive model assumptions.

Jiahua Chen is Professor in the Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z2 Canada (E-mail: *jhchen@stat.ubc.ca*). Asokan Mulayath Variyath is Research Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *variyath@stat.tamu.edu*). Bovas Abraham is Professor, Department of Statistics and Actuarial Science, Waterloo, ON N2L 3G1, Canada (E-mail: *babraham@uwaterloo.ca*).

Journal of Computational and Graphical Statistics, Volume 17, Number 2, Pages 426–443
DOI: 10.1198/106186008X321068

Assume that 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 estimating 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 the Poisson distribution family and θ is its mean, a semiparametric model assumption takes F as a distribution having finite first and second moments with equal mean and variance. Hence, the parameter in this semiparametric model is specified by the 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 EL 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 the observations. However, this definition is applicable even when there are tied observations (Owen 2001). Without any further information on the distribution F, the EL 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 componentwise. In general, it is more convenient to work with the logarithm of the empirical likelihood

$$l_n(F) = \sum_{i=1}^{n} \log(p_i). \tag{1.1}$$

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

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

$$l_{\mathrm{EL}}(\theta) = \sup\{l_n(F) : \; p_i \geq 0, \; i = 1, \ldots, n; \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} p_i g(y_i, \theta) = 0\}. \tag{1.2}$$

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

To solve the numerical problem related to $l_{\mathrm{EL}}(\theta)$, a prerequisite is that the convex hull of $\{g(y_i,\theta), i = 1, 2, \ldots, n\}$ must have the zero vector as an interior point. In the simple example of population mean, $l_{\mathrm{EL}}(\theta)$ is well defined for all θ in the convex hull of $\{y_i, i = 1, 2, \ldots, n\}$. In applications that we shall discuss in more detail, the dimension m of g can be much larger than the dimension q of the parameter θ. It can be difficult to determine the region Θ over which $l_{\mathrm{EL}}(\theta)$ is well defined. In fact, it is possible that Θ is empty. A convention is to define $l_{\mathrm{EL}}(\theta) = -\infty$ for $\theta \notin \Theta$. However, this strategy has two drawbacks: (1) It is often difficult to specify the region Θ which is data specific, and (2) for any pair of $\theta_1, \theta_2 \notin \Theta$, the values $l_{\mathrm{EL}}(\theta_1)$ and $l_{\mathrm{EL}}(\theta_2)$ provide no information on their relative plausibility, even if one is 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 $l_{\mathrm{EL}}(\theta)$. Even finding a proper initial value can be difficult in this case.

In this article, we make a novel adjustment to the commonly used EL. With this adjustment, the profile-adjusted empirical likelihood (AEL) is well defined for all parameter values. Hence, it is not necessary to determine whether $\theta \in \Theta$ and consequently finding the maximum point of the adjusted $l_{\mathrm{EL}}(\theta)$ becomes much simpler. Further, the new solution provides a good initial value for θ when Θ is nonempty in the original problem. We also show that the asymptotic properties of the EL are preserved. The adjusted EL confidence regions based on the chi-square limiting distribution are always 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 by Tsao (2004). As will be seen in our simulation later, the improved coverage probability is achieved without resorting to complex techniques such as Bartlett-correction or bootstrap calibration. In fact, the algorithm for the AEL converges more quickly than the algorithm for the EL. Finally, a numerical algorithm for computing the profile AEL is given with guaranteed convergence.

The rest of the article is organized as follows. In Section 2, we give additional details of EL 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 empirical log-likelihood ratio function. We demonstrate the usefulness of the new method by simulation and by examples. We draw several conclusions in Section 6, and two proofs are given in the 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 that defines the parameter θ of the population distribution F through $E\{g(Y, \theta)\} = 0$. The empirical log-likelihood $l_n(F)$ is defined by (1.1) and the profile empirical log-likelihood $l_{\mathrm{EL}}(\theta)$ is defined as in (1.2). Let Θ be the set of parameter values such that for each $\theta \in \Theta$, the solution to

$$\sum_{i=1}^{n} p_i g(y_i, \theta) = 0 \tag{2.1}$$

exists for the p_i values. For each $\theta \in \Theta$, $l_n(F)$ under constraint (2.1) is maximized when

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

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 the profile empirical log-likelihood function,

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

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

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

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

$$\{\theta : -2W(\theta) \leq \chi^2_m(1 - \alpha)\}, \tag{2.2}$$

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 regions constructed via the normal approximation, the EL regions have data-driven shape, are range respecting, and often have better coverage properties (Owen 1990; Chen et al. 2003).

The results for the population mean are more general. Similar conclusions are true for linear models, generalized linear models, models defined by estimating equations, and many others (Owen 1991; Kolaczyk 1994; Qin and 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

$$\sum_{i=1}^{n} \frac{g(y_i, \theta)}{1 + \lambda' g(y_i, \theta)} = 0 \tag{3.1}$$

for the Lagrange multiplier λ. We look for a solution λ that satisfies $1 + \lambda' g(y_i, \theta) > 0$ for all $i = 1, 2, \ldots, n$. A necessary and sufficient condition for its existence is that the zero vector is an interior 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 interior point with probability 1 as $n \to \infty$. When θ is not close to θ_0, or when n is small, there is a good chance that the solution to

(3.1) does not exist. This can be a serious limitation in some applications as shown in the examples in the next section. In this section, we propose the following AEL.

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 be

$$\begin{aligned} W^*(\theta) \quad &= \quad \sup\Bigg\{\sum_{i=1}^{n+1}\log[(n+1)p_i] : \ p_i \ge 0, \ i = 1,\dots,n+1; \\ &\qquad \sum_{i=1}^{n+1} p_i = 1; \ \sum_{i=1}^{n+1} p_i g_i = 0\Bigg\}. \end{aligned} \tag{3.2}$$

Since the convex hull of $\{g_i, i = 1, 2, \dots, n, n+1\}$ for any given θ contains 0, $W^*(\theta)$ is well defined for all θ.

It can be 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 $W(\theta) = -\infty$ whereas $W^*(\theta)$ 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 maximizing $W(\theta)$.

The value of a_n should be chosen according to the specific nature of the 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 mild: it is equivalent to adding a few artificial but comparable observations to the set of n original observations. Nevertheless, we recommend the use of a_n with smaller magnitude.

When θ is far from θ_0, some g_i values can be much larger than the rest of the 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 semiparametric 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 low around θ compared to in the neighborhood of θ_0. When the model is incorrect, which occurs in the model selection example in the next section, there will be no hypothetical θ_0 to rely on. Our strategy is to replace $\bar{g}_n$ by the median or trimmed mean of the g_i. The particular form of $\bar{g}_n$ can be chosen by the user.

In most applications, the focus is on θ in a small neighborhood of θ_0 so that $\bar{g}_n$ is of moderate size. Our general recommendation is to have $a_n = \max(1, \log(n)/2)$ coupled with a trimmed version of $\bar{g}_n$ when appropriate. Investigation of the optimal choice of a_n in various practically important situations is still under way. When the sample size n increases, our estimate of θ_0 will get into its $n^{-1/2}$ range, hence this a_n will make $a_n\bar{g}_n = o_p(1)$. The effect of the 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 the confidence regions.

We now give a simple example to illustrate the convex hull problem and the adjustment. We generate 50 observations from an independent bivariate standard normal distribution.

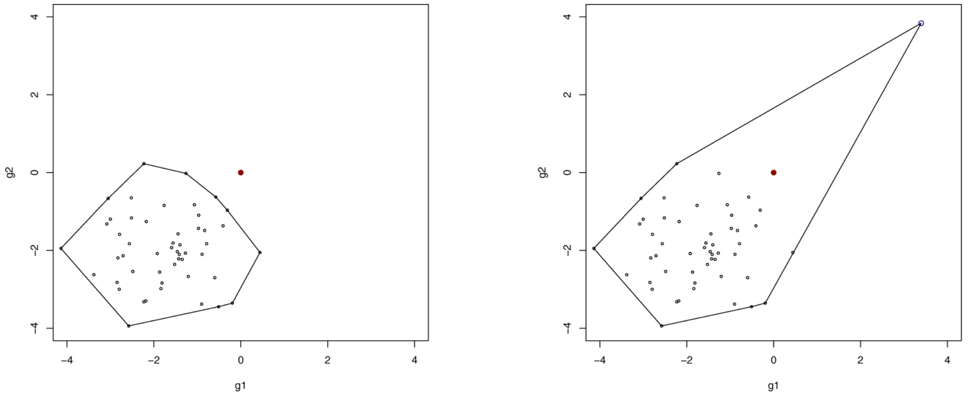

Figure 1. Convex hull (left) and adjusted convex hull with $a_n = \log(n)/2$ (right). The bold dot is (0,0).

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 interior point as in Figure 1 (right). Since $(2, 2)$ is well outside of the convex hull, the adjustment in this case appears to be substantial. However, the empirical log-likelihood is merely adjusted from $-\infty$ to some large negative finite number, and this does not really have a large impact. At the same time, one will not notice anything substantial in the adjustment of EL at $(\mu_1, \mu_2) = (0, 0)$.

4. ASYMPTOTIC PROPERTIES

The most important result in the context of EL is the limiting distribution of the empirical log-likelihood ratio function at θ_0. We show that the AEL has the same asymptotic properties as the unadjusted EL.

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 be 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 profile empirical log-likelihood ratio function defined by (3.2) and $a_n = o_p(n^{2/3})$. As $n \to \infty$, we have*

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

in distribution.

Proof: See Appendix.

If the rank of 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 foregoing 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 regions using AEL are no longer confined within the convex hull of the observed values. Thus, the new method has the potential to effectively improve the under-coverage problem due to small sample size or high dimension.

We are naturally interested in knowing the asymptotic behavior of $W^*(\theta)$ and $W(\theta)$ when $\theta \neq \theta_0$. Interestingly, this problem has not received much attention in the literature.

Theorem 2. *Assume that the conditions in Theorem 1 hold and that for some $\theta \neq \theta_0$*

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

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

The proof of Theorem 2 is also given in the Appendix. Based on Theorems 1 and 2, it is easily seen that the nonparametric 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 and Lawless (1994) showed that there exists a local maximum of $W(\theta)$ in an $n^{-1/3}$ neighborhood of θ_0. The same is true for $W^*(\theta)$. Because the proofs of this and other results do not contain new techniques, we present only one result here.

Theorem 3. *Assume that the conditions in Theorem 1 hold, and that $\partial^2 g(y,\theta)/(\partial\theta\partial\theta')$ is continuous in the neighborhood of θ_0. Also assume that $\|g(y,\theta)\|^3$ and $\|\partial^2 g(y,\theta)/(\partial\theta\partial\theta')\|$ are bounded by some integral function $G(y)$ in the neighborhood 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

$$\Sigma = \{E(\partial g/\partial\theta)' \, (Egg')^{-1} \, E(\partial g/\partial\theta)\}^{-1}.$$

5. NUMERICAL ALGORITHM, SIMULATIONS, AND EXAMPLES

Since the constraints in the AEL are always satisfied, the numerical computation of the profile AEL can be done with any existing algorithm. In particular, we recommend the modified Newton–Raphson algorithm proposed by Chen et al. (2002). To maximize $W^*(\theta)$ with respect to θ, we use the simplex method introduced by Nelder and Mead (1965) for the sake of its stability. Most optimization software includes a built-in function for the simplex method.

For given θ and 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 $g_{n+1} = -a_n \bar{g}_n$. The sample mean $\bar{g}_n$ may be replaced by a trimmed mean or any other robust substitute.

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

3. Compute the first and second partial derivatives of

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

with respect to λ evaluated at λ^k. Let these 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 $1 + (\lambda^k + \delta)^{'} g_i \le 0$ 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. (2002) for the proof of the algorithmic convergence. The foregoing operations result in the value of $W^*(\theta)$. We then use the simplex method to optimize this function. A set of good initial values of θ will be helpful in this step. In general, one may use a rough estimate of θ such as via the method of moments. We demonstrate the usefulness of AEL in the following examples.

5.1 CONFIDENCE REGION

Constructing confidence regions for the population means received primary attention in the pioneering article by Owen (1988). The empirical likelihood confidence regions have data-driven shape, are range respecting, and are Bartlett-correctable (DiCiccio et al. 1991). When the sample size is not large, one may replace chi-square distribution calibration by bootstrap calibration to improve the accuracy of the coverage probability. As discussed by Tsao (2004), however, confidence regions computed via the unadjusted EL are by definition confined to be 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 regions are no longer confined to be within the convex hull of the data. It is hence better suited to solve the under-coverage problem.

For illustration, we applied the new method to construct confidence intervals for the examples discussed by DiCiccio et al. (1991). We used their settings. The coverage probabilities are based on 5,000 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

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	0.9222	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.

of g_{n+1}. We reproduce the Bartlett-correction results from DiCiccio et al. (1991) for comparison. The coverage probabilities of intervals based on unadjusted EL, Bartlett-corrected, and AEL are given in Table 1 for nominal levels of 80, 90, 95, and 99%. It is clear that the coverage probabilities of the AEL are closer to the target values for the sample sizes and population distributions considered. The results are particularly impressive since the AEL does not involve 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 5,000 times with $a_n = \log(n)/2$ for bivariate distributions: (1) two independent standard normal distributions, (2) two independent $\chi^2_{(1)}$ distributions, and (3) two independent $t_{(5)}$ distributions. The coverage probabilities are reported in Table 2. It is clear that the AEL compares favorably with the usual unadjusted EL. The coverage probabilities are substantially improved with the AEL method. Our simulations also revealed that the AEL is computationally simpler: the time taken was only about one-fifth to one-seventh of the time for the unadjusted EL.

To provide a more concrete comparison between the confidence regions constructed with the unadjusted and the adjusted EL, we considered a dataset given by Owen (2001, p. 31). The original source of this dataset is Iles (1993). It consists of four types of prey

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.

(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 numbers 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 scatterplot of the original data points. It can be seen that the confidence regions based on the AEL retain the data-driven shape and contain the confidence regions based on the unadjusted EL.

We finally remark that there is a potential for further gain in accuracy of the confidence regions by choosing a proper size of a_n. We have some preliminary results on using the Bartlett correction technique to guide our choice of a_n, but it is outperformed by $\log(n)/2$. Since the theory may not be of interest to many and is yet to be developed fully, it is left as a topic for future research.

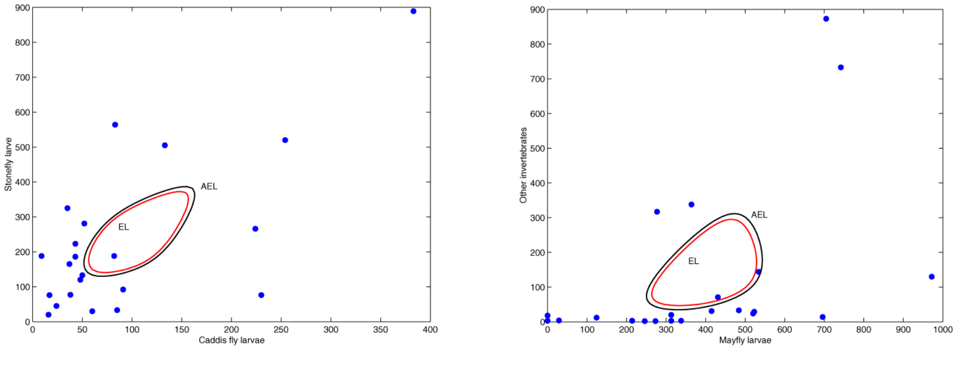

Figure 2. Comparison of confidence regions based on EL and AEL.

5.2 Estimation of a Covariance Matrix With Known Zero Entries

Chaudhuri et al. (2007) discussed the problem of estimating a covariance matrix of a random vector with many known zero entries. Such restrictions appear in many applications, see Grzebyk et al. (2004). We refer the readers to Chaudhuri et al. (2007) for motivation and more references.

Suppose we have a random vector $Y = (Y_1, Y_2, Y_3, Y_4)'$ 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 the normality assumption, both the mean $\mu = E\{Y\}$ and the covariance matrix can be estimated by the maximum likelihood estimator. However, maximizing the likelihood function under the constraint of some covariances being zero is not a simple problem. Chaudhuri et al. (2007) presented many numerical algorithms for this problem. One method is to ignore the normality assumption and to use EL 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. Chaudhuri et al. (2007) proposed computing the profile EL at a given feasible μ and Σ:

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

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

When the sample size is small ($n = 10$), Chaudhuri et al. (2007) experienced problems with the EL procedure due to an inability to find feasible starting values. With the AEL, however, this problem disappears immediately. According to our own simulation, the minimizer of $W^*(\mu, \Sigma)$ (with $a_n = n^{-1}$) provides excellent initial values for maximizing $l_{\mathrm{EL}}(\mu, \Sigma)$. We used $a_n = n^{-1}$ instead of $a_n = \max\{1, \log n/2\}$ to better approximate the solution of $l_{\mathrm{EL}}(\mu, \Sigma)$. Interestingly, the minimizer of $W^*(\mu, \Sigma)$ itself serves as an efficient estimator when $a_n = n^{-1}$. The best choice of a_n in terms of efficiency is beyond the scope of this article. The following simulation study demonstrates that the AEL provides good initial values and more efficient estimation.

We simulated datasets $Y = (Y_1, Y_2, Y_3, Y_4)'$ from a 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 1,000 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 AEL with $a_n = n^{-1}$ where the existence of the solutions is guaranteed. The maximum AEL mean estimator was then used to provide an initial value for the unadjusted EL. This time, we failed to find a solution in only 17.4% of the samples. Thus, the AEL is useful as an initial-value locator even if one insists on using the original EL.

Further, we are interested in the efficiency of the maximum AEL estimator itself. We computed the total bias and root mean square error (RMSE) of the maximum unadjusted and adjusted EL estimates according to the following definitions:

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

and

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

where $\hat{\sigma}_{ij}^{(k)}$ is the estimate of σ_{ij} in the kth simulation dataset and M is the total number of simulations. For this comparison, we set $n = 20$ and the number of simulations $M = 10{,}000$. For the unadjusted EL, we did not find solutions in 2.1% of the simulations. Hence, total bias and root mean square error were computed based on the remaining 97.9% of the 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 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 AEL 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 small as $n = 10$ may be unusual. 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 be used both to search for feasible starting values and to directly provide efficient estimates of unknown parameters in this and similar applications. The best choice of a_n to achieve the highest efficiency is beyond the scope of this article.

5.3 VARIABLE SELECTION IN REGRESSION ANALYSIS

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

$$y_i = \beta_0 + x_i^{'}\beta + \epsilon_i$$

$i = 1, 2, \ldots, n$ 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 meaningfully different from 0. Thus, a variable selection step is often applied to reduce the complexity of the model and hence reduce the variability in the 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). 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 Mulayath (2006) discussed the 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 EL is given by

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

Note that the number of constraints remains a constant with respect to s in the definition of the profile likelihood. This is important for differentiating the plausibility of submodels formed by a subset of covariates.

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

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

with k being the number of covariates in s. We choose the submodel corresponding to the s that minimizes $\mathrm{EAIC}(s)$. Similarly, we define

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

This empirical version of BIC chooses the submodel s that minimizes $\mathrm{EBIC}(s)$.

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

Consider the most extreme case where s is empty, but the dimension of x (i.e., m) is relatively large while n is not so large. In this case, the constraints imply that we must find a value of β_0 such that

$$\sum_{i=1}^{n} p_i = 1;$$
$$\sum_{i=1}^{n} p_i(y_i - \beta_0) = 0;$$
$$\sum_{i=1}^{n} p_i x_i^{'}(y_i - \beta_0) = 0$$

have solutions in the p_i values. In the special case where $y_i = x_i$ and $m = 1$, this 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 $\text{EAIC}(s) = \text{EBIC}(s) = \infty$. With the AEL, the solution exists for any choice of β_0. Thus, sensible and informative values of $\text{EAIC}(s)$ and $\text{EBIC}(s)$ are always defined after the adjustment and can be easily computed. We choose $a_n = \log(n)/2$ because it worked well for constructing confidence regions. Yet the key in this application is to guarantee that the optimization problem has a solution for any candidate model of interest. Hence, the size of a_n is not of particular importance.

We now consider the cancer study example given by Stamey et. al (1989) for the variable selection problem. This study examines the correlation between the levels of prostate specific antigen (PSA) and eight clinical measurements of 97 men who were yet to receive a radical prostatectomy. The 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 logarithm of PSA based on the eight covariates. We apply variable selection methods AIC, BIC, EAIC, and EBIC to identify the appropriate model. Defining the submodels by setting the regression coefficients to zero for all covariates not in the submodel, we face computational issues due to the nonexistence of a solution to the equation $\sum_{i=1}^{n} p_i g(y_i, x_i, \beta) = 0$ for given β. The AEL works well in this example. Covariates lcavol, lweight, and svi were identified to form an appropriate model by EAIC, BIC, and EBIC whereas the model identified by AIC includes an additional variable lbph. These models, the 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 pick up the most significant covariates. The AIC picks a slightly larger model than EAIC does.

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)
lcavol	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)	–	–

6. CONCLUSIONS

In this article, we suggested a method to overcome the difficulty posed by the nonexistence of solutions while computing the profile empirical likelihood. We demonstrated that the proposed AEL is well defined for all parameter values. The new method substantially enhances the applicability of EL. We showed that the resulting AEL retains the asymptotic optimality properties. Further, the confidence regions constructed by the new method have closer to nominal coverage probabilities in the examples considered. The algorithm associated with the AEL also converges much faster, reducing the computational burden. The usefulness of this new method was illustrated via a number of examples.

A. APPENDIX

Because the proofs of some similar results are well known and can be easily found in Owen (2001), our proofs here will be 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 that $\sigma_1^2 = 1$. Let λ be the solution to

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

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 that

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

Let $\rho = \|\lambda\|$ and $\hat{\lambda} = \lambda/\rho$. Multiplying $n^{-1}\lambda'$ to both sides of (A.1), we get

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

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

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

in probability for some $1 > \epsilon > 0$. Therefore, as long as $a_n = o_p(n)$, (A.2) implies that

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

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

Next, denoting $\hat{V}_n = n^{-1} \sum_{i=1}^{n} g_i' g_i$, we find that

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

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

Finally, we expand W^* as follows.

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

Substituting the expansion of λ, we get that

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

which converges to a 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 foregoing proof still goes through with no changes. Even if a_n has larger order such that $a_n = o_p(n)$, the 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 $n \to \infty$, $\|\bar{g}_n'\bar{g}_n\| \to \delta^2 > 0$ in probability. Note that $g_i - \bar{g}_n$ has mean zero, and satisfies 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

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

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

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

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

The proof of $-2n^{-1/3} W(\theta) \to \infty$ is similar. □

Remark: First, the sample mean $\bar{g}_n$ can again be replaced by the sample median or trimmed means without invalidating the foregoing 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.

ACKNOWLEDGMENTS

This research is support by the Natural Science and Engineering Research Council of Canada and by the MITACS. The authors thank the Editor, Associate Editor and referees for helpful comments.

[Received November 2006. Revised September 2007.]

REFERENCES

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–282.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007), "Estimation of a Covariance Matrix with Zeros," *Biometrika*, 94, 199–216.

Chen, J., Chen, S., and 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, 230–237.

DiCiccio, T., Hall, P., and Romano, J. (1991), "Empirical Likelihood is Bartlett-Correctable," *The Annals of Statistics*, 19, 1053–1061.

Grzebyk, M., Wild, P., and Chouanière, D. (2004), "On Identification of Multi-factor Models with Correlated Residuals," *Biometrika*, 91, 141–151.

Iles, T. C. (1993), "Multiple Regression," in *Biological Data Analysis: A Practical Approach*, ed. J. C. Fry, Oxford: Oxford University Press, pp. 127–172.

Kolaczyk, E. D. (1994), "Empirical Likelihood for Generalized Linear Models," *Statistica Sinica*, 4, 199–218.

Mulayath Variyath, A. (2006), "Variable Selection in Generalized Linear Models by Empirical Likelihood," Ph.D. Thesis, Department of Statistics and Actuarial Science, University of Waterloo. Waterloo, Ontario, Canada.

Nelder, J. A., and Mead, R. (1965), "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308–313.

Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Interval for a Single Functional," *Biometrika*, 75, 237–249.

——— (1990), "Empirical Likelihood Confidence Regions," *The Annals of Statistics*, 18, 90–120.

——— (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19, 1725–1747.

——— (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC.

Qin, J., and 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., and 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.

Wilks, S. S. (1938), "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annuals of Mathematical Statistics*, 9, 60–62.