

ESS Update

Author: Art B. Owen

Title: Empirical Likelihood

Keywords: Bootstrap, Jackknife, Likelihood Ratio,
Nonparametric, Wilks's Theorem

Empirical likelihood is a technique for forming hypothesis tests and confidence sets based on nonparametric likelihood ratios. Many properties of parametric likelihood ratio functions have nonparametric parallels. The main one is that there is a nonparametric version of Wilks's famous result wherein an asymptotic chisquare distribution holds for the log likelihood ratio. The nonparametric version has the advantage of holding under very weak conditions. In particular, the data do not have to follow a parametric distribution for the chisquare result to hold for nonparametric likelihood ratios.

Empirical likelihood ratio inferences are of comparable accuracy to those based on the delta method, the jackknife and the simpler bootstrap methods. Each method has its advantages. Here we list a few of empirical likelihood's advantages: When constructing confidence regions for two or more parameters of interest, empirical likelihood gives a data determined shape for the confidence region. When constraints are known to hold among the parameters of interest, they can be imposed numerically. A Bartlett correction applies to empirical likelihood, providing a simple way to increase the accuracy of inferences. As in some, but not all of the other methods, all points in an empirical likelihood confidence region obey the usual range restrictions: variances are nonnegative, probabilities are in $[0, 1]$ and correlations are in $[-1, 1]$. Sometimes empirical likelihood methods require less computation than the alternatives, sometimes more.

At first nonparametric likelihood ratios may seem paradoxical, but they are defined below by taking very literally the notion that likelihood is the probability of observing the actual data values at hand.

1 Nonparametric Maximum Likelihood

The empirical distribution function is well known as a nonparametric maximum likelihood estimate (NPMLE). If X_1, \dots, X_n are i.i.d. in R^d from some distribution F_0 , the empirical distribution function is

$$F_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

where δ_X is a distribution taking the value X with probability one. The probability that F_n attaches to a set A , denoted $F_n(A)$, is $1/n$ times the number of sample observations belonging to A . In particular, $F_n(\{x\})$ is $1/n$ times the number of observations in the sample that are equal to x , and for $d = 1$ the usual empirical cumulative distribution function is equal to $F_n((-\infty, x])$, as a function of x .

We write $F\{x\}$ as a short form for $F(\{x\})$. The likelihood that F_n maximizes is $L(F) = \prod_{i=1}^n F\{X_i\}$. It is easy to see that F_n is the NPMLE. Let Y_1, \dots, Y_k be the distinct values observed among the X_i with Y_i appearing $n_i \geq 1$ times. Let F be any distribution on R^d and let $p_i = F\{Y_i\}$. Now using $\log(1+z) \leq z$ and $\sum_i p_i \leq 1$ we find

$$\begin{aligned} \log \left(\frac{L(F)}{L(F_n)} \right) &= \sum_{i=1}^k n_i \log \left(\frac{np_i}{n_i} \right) \\ &\leq \sum_{i=1}^k n_i \left(\frac{np_i}{n_i} - 1 \right) \\ &\leq 0 \end{aligned}$$

and so $L(F) \leq L(F_n)$.

That F_n is the NPMLE was apparently first noticed by Kiefer and Wolfowitz (1956). The NPMLE concept has since been put to good use in defining analogues of the empirical distribution function for problems where data are indirectly sampled or incompletely observed. The best known examples are for right censored data (Kaplan and Meier, 1958),

for more general censoring and truncation (Peto 1973, Turnbull 1974, 1976) and for biased sampling (Vardi (1982, 1985), Gill, Vardi and Wellner 1988).

Statisticians are usually interested in F through some statistical functional $T(F)$. For example, T might be the mean, or median, or a coefficient in a logistic regression. Then, if \hat{F} is an NPMLE of F , such as F_n in the i.i.d. case, it is natural to consider $\hat{T} = T(\hat{F})$ to be the NPMLE of $T(F)$. Here $T(\hat{F})$ is the estimator, $T(F)$ is the parameter and $T_0 = T(F_0)$ is the true value of the parameter.

As an example, suppose that $T_2(F)$ is the sampling variance of $T_1(F_n)$ based on an i.i.d. sample from F . Then $T_2(F_n)$, the bootstrap estimate of variance, can be considered an NPMLE, as can other simple bootstrap quantities.

2 Empirical Likelihood

In parametric theory, the maximizer of the likelihood function is used to estimate the parameter of interest, and for nonparametric likelihoods the same method provides NPMLE's. Confidence regions and hypothesis tests can be based on contours of parametric likelihood ratio functions, usually with a calibration from asymptotic theory. Similarly, a nonparametric likelihood ratio function can be defined and used to produce confidence regions and tests, again with an asymptotically justified calibration.

The first use of nonparametric likelihood ratios appears to be due to Thomas and Grunkemeier (1975). The statistic they considered was the survivor function, $S(x) = \Pr(X \geq x)$, the probability of surviving to at least time x . The Kaplan-Meier estimator provides the NPMLE $\hat{S}(x)$. The usual method for setting confidence intervals is based on Greenwood's formula for estimating the variance of $\hat{S}(x)$. A central limit theorem justifies taking as a 95% confidence interval $\hat{S}(x)$ plus or minus 1.96 estimated standard deviations of $\hat{S}(x)$. Such confidence intervals don't necessarily lie between 0 and 1, even though $S(x)$ is a probability.

Thomas and Grunkemeier define a likelihood ratio function $R(S) = L(S)/L(\hat{S})$ and give a heuristic proof that $-2 \log \max_{S(x)=S_0(x)} R(S)$ has an asymptotic chisquared distribution on 1 degree of freedom. Then for a 95% confidence interval they take $\{s \mid -2 \log R(s) \leq 3.84\}$, because $\Pr(\chi_{(1)}^2 \leq 3.84) = 0.95$. This interval is contained inside the interval $[0, 1]$ and in their simulation study it performed better than the interval based on Greenwood's formula.

In general for X_1, \dots, X_n i.i.d. in R^p , and F a distribution on R^p , define $R(F) = L(F)/L(F_n)$. Now for a statistical functional $T(F)$, it is natural to wonder how generally sets like $C = \{T(F) \mid R(F) \geq r\}$ can be used as confidence regions for $T(F_0)$, or equivalently whether hypothesis tests of $T(F_0) = t$ can be constructed by rejecting if and only if $t \notin C$. This likelihood ratio function was termed the empirical likelihood ratio function because the empirical distribution function F_n appears in the denominator. Note that any point in C is of the form $T(F)$ for some distribution F and hence all points of the confidence set satisfy range restrictions that T must satisfy.

The central result is for the mean, $T(F) = \int x dF(x)$, and it extends easily to more general statistics as described below. It is clear however that some care needs to be taken in formulating nonparametric likelihood confidence regions or else $C = R^d$ whenever $r < 1$. To see this, for $0 < \epsilon \leq 1$ and $x \in R^d$, define $F_{\epsilon, x} = (1 - \epsilon)F_n + \epsilon\delta_x$, a distribution with mean $(1 - \epsilon)\bar{X} + \epsilon x$. If $r < 1$, choosing ϵ small enough makes $R(F_{\epsilon, x}) > r$ and then letting x vary through R^d the mean $(1 - \epsilon)\bar{X} + \epsilon x$ sweeps out all of R^d .

If however, the X_i are known to belong to a bounded set B then a confidence set such as $\{\int x dF(x) \mid R(F) \geq r, F(B) = 1\}$ does not suffer from the problem described above. Choosing B can be difficult, even when the X_i are known to be bounded. Fortunately, under mild conditions, one can simply use as bounded sets B_n , the convex hull of X_1, \dots, X_n . B_n is the smallest polyhedron containing all of the X_i , and for $d = 1$ it is simply the interval $[X_{(1)}, X_{(n)}]$.

For the mean, if $F(B_n) = 1$, there is no loss in further assuming that F puts probability one on the sample: i.e. $F(\{X_1, \dots, X_n\}) = 1$. If $F(B_n) < 1$, then either F places probability one on the observed sample X_1, \dots, X_n , or there is another distribution \tilde{F} with $\tilde{F}(\{X_1, \dots, X_n\}) = 1$, $\int x d\tilde{F} = \int x dF$ and $L(\tilde{F}) > L(F)$.

Owen (1990) proves the following:

Empirical Likelihood Theorem (ELT). *Let X, X_1, X_2, \dots be i.i.d. random vectors in R^d , with $E(X) = \mu_0$, and $\text{var}(X)$ finite and of rank $q > 0$. For positive $r < 1$ let*

$$C_{r,n} = \left\{ \int x dF(x) \mid R(F) \geq r, F(\{X_1, \dots, X_n\}) = 1 \right\}.$$

Then $C_{r,n}$ is a convex set and

$$\lim_{n \rightarrow \infty} P(\mu_0 \in C_{r,n}) = P(\chi_{(q)}^2 \leq -2 \log r).$$

Moreover if $E(\|X\|^4) < \infty$ then

$$|P(\mu_0 \in C_{r,n}) - P(\chi_{(q)}^2 \leq -2 \log r)| = O(n^{-1/2}).$$

It is not necessary for F to be a bounded distribution; the ELT only requires it to have a finite nonzero variance. This ensures that the sets B_n do not grow too rapidly with n .

The empirical likelihood ratio function $R(F)$ for F satisfying $F(\{X_1, \dots, X_n\}) = 1$ is simply a multinomial likelihood ratio function on the distinct observed values Y_j . When F_0 is a discrete distribution taking values in a finite set, the ELT above follows from Wilks's theorem applied to the multinomial distribution.

When F_0 is a continuous distribution, the ELT is surprising because the number of parameters, p_i in the previous section, is equal to the number of data points n . (There are

$n - 1$ free parameters because $\sum_i p_i = 1$.) Parametric MLE's are not necessarily consistent under these conditions (Neyman and Scott, 1948). Also, in the finite discrete case it is eventually true that $R(F_0) > 0$, but for the continuous case, $R(F_0) = 0$ no matter how large n is.

The degrees of freedom in the chisquare is d , unless the distribution of X is completely restricted to a hyperplane of dimension $q < d$.

The error in the probability approximation has the same rate as the one for parametric likelihood ratios. Under some further conditions to justify Edgeworth expansions, the error is in fact $O(n^{-1})$. This is due to a cancellation phenomenon. When $d = 1$, one sided confidence intervals have a coverage error $O(n^{-1/2})$, but when forming central confidence intervals, the lead terms in the coverage error cancel leaving an error of $O(n^{-1})$. This phenomenon also holds for parametric likelihood ratios. These issues are discussed in the survey paper of Hall and La Scala (1990).

3 Extensions

3.1 Smooth Functions of Means

Versions of the ELT exist for more general statistics than the mean. One large class of statistics is obtained by taking a smooth function of means $T(F) = g(\int Z(x)dF(x))$. For example the correlation between U and V can be written as a smooth function of the expected value of $(U, V, U^2, V^2, UV)'$ and this vector can be written as $Z(X)$ where $X = (U, V)'$. Similarly, the coefficients in a linear regression can be written through smooth functions of means. For F close to F_0 , the linear Taylor approximation $T(F) \doteq T_L(F) = T(F_0) + J'_0 \int (Z(x) - \mu_0)dF(x)$ holds where J'_0 is the Jacobian matrix of partial derivatives of components of g with respect to components of $E(Z)$ evaluated at μ_0 . The approximation is usually close enough that the set $\{T(F) \mid R(F) \geq r, F(\{X_1, \dots, X_n\}) = 1\}$ approximates

$\{T_L(F) \mid R(F) \geq r, F(\{X_1, \dots, X_n\}) = 1\}$ and the ELT applies to the coverage probability of the latter set. A proof along these lines is given in Owen (1990), and Owen (1988) has a proof for one dimensional Frechet differentiable statistical functionals. DiCiccio, Hall and Romano (1991) show that the coverage error in the ELT for smooth functions of means is $O(n^{-1})$ under mild conditions.

If $T(F)$ is the variance of X under sampling from F , empirical likelihood confidence regions have asymptotically correct coverage so long as F satisfies $E(X^4) < \infty$ and, to avoid trivialities, $T(F) > 0$. By contrast, normal theory likelihood ratio confidence intervals do not have correct coverage, even as $n \rightarrow \infty$, unless $E((X - E(X))^4) = 3E((X - E(X))^2)^2$ holds. Thus, for normal theory confidence intervals to be asymptotically correct the kurtosis of the sampling distribution must be zero, whereas empirical likelihood regions require finite kurtosis.

3.2 Estimating Equations

Another generalization is to parameters θ defined implicitly through estimating equations $\int g(x, \theta) dF(x) = 0$ and usually estimated by statistics $\hat{\theta}$ solving $0 = \int g(x, \theta) dF_n(x) = (1/n) \sum_{i=1}^n g(X_i, \hat{\theta})$. Here $X_i \in R^d$, $\theta \in R^q$ and $g(x, \theta) \in R^p$. Usually $p = q$ so that there are as many estimating equations as there are unknown parameters. These estimators are also called M -estimators because they generalize the method of maximum likelihood where for X_i i.i.d. with density $f(x, \theta)$ one takes $g(x, \theta)$ to be $\partial \log f(x, \theta) / \partial \theta$.

Owen (1991) gives an ELT for M -estimates under very weak conditions, that don't even require $\hat{\theta} = T(F_n)$ to be a good estimate of $\theta = T(F_0)$. Qin and Lawless (1994) prove a sharper ELT under stronger conditions that ensure that $T(F_n)$ is a good estimator. Zhang (1996) gives conditions for one dimensional M -estimates to have coverage error $O(1/n)$.

Consider the linear regression model $E(Y_i | X_i = x) = Z(x)\beta$. Here $Z(x)$ is a row vector such as $(1, x, x^2)$ and β is a column vector of parameters. The usual estimating equations

for this model, based on i.i.d. observation pairs (X_i, Y_i) are $E(Z(X)'(Y - Z(X)\beta)) = 0$, and the sample coefficients $\hat{\beta}$ solve the normal equations $\sum_{i=1}^n Z(X_i)'(Y_i - Z(X_i)\hat{\beta}) = 0$. Then $-2 \log \mathcal{R}(\beta_0)$ has an asymptotic $\chi_{(p)}^2$ distribution under very mild conditions described in Owen (1991). The main condition is that the random variables $Z(X_i)'(Y_i - Z(X_i))$ have a finite variance matrix of rank $p > 0$. There is no need to assume that the variance of Y_i given $X_i = x$ is independent of x .

3.3 Other Extensions

Empirical likelihood inferences are also available for regression on deterministic predictors x_i , fixed for example by design, under mild conditions described in Owen (1991). There the random variables $Z(x_i)'(Y_i - Z(x_i))$ are not i.i.d., but a modified ELT applies to them. Kolaczyk (1994) considers generalized linear models and Owen (1992) considers estimating equations derived from certain projection pursuit models.

Chen and Hall (1993) consider empirical likelihood for quantiles, and discuss Bartlett correctability and smoothing for this problem.

Hall and Owen (1993) provide an ELT for kernel density estimates. This can be used to form pointwise confidence intervals and simultaneous confidence bands for the density function. Owen (1995) constructs empirical likelihood based confidence bands for the univariate empirical cumulative distribution function.

Mykland (1995) defines a dual likelihood for data from a martingale. The dual likelihood ratio function reduces to the empirical likelihood ratio function for independent observations. Martingale methods allow one to consider problems arising in time series and in point process models.

Qin and Lawless (1994) study the case where the number of estimating equations exceeds the number of parameters. This can be thought of as knowing a number of prior constraints that the parameters must satisfy. An example is univariate regression constrained to pass

through the origin with one estimating equation to stipulate mean zero residuals, and another for residuals uncorrelated with the predictor. They define a maximum empirical likelihood estimate (MELE) and show that the MELE combines the estimating equations in an asymptotically efficient way. They prove an ELT for this case and find a diagnostic for whether the postulated constraints truly hold.

Qin (1994) considers problems in which parametric likelihoods on one sample may be combined with empirical likelihoods on another. Qin (1995) considers problems in which parametric likelihoods apply over part of the data range and nonparametric likelihoods are used on the rest of the range.

4 Computation

A convenient approach for computation is through the profile empirical likelihood ratio function

$$\mathcal{R}(\theta) = \sup \left\{ R(F) \mid T(F) = \theta, F(\{X_1, \dots, X_n\}) = 1 \right\}.$$

Confidence regions are of the form $\{\theta \mid \mathcal{R}(\theta) \geq r\}$ and hypothesis tests of $H_0 : \theta = \theta_0$ reject if and only if $-2 \log \mathcal{R}(\theta_0)$ is smaller than a critical value from the appropriate chi-square distribution. In the case of the mean,

$$\mathcal{R}(\mu) = \max \left\{ \prod_{i=1}^k \left(\frac{np_i}{n_i} \right)^{n_i} \mid 0 \leq p_i, \sum_{i=1}^k p_i = 1 \right\}.$$

It is a nuisance to have to keep track of ties among the observations, and Owen (1988) shows that an observation based likelihood is equivalent, not just for the mean, but in general. Let observation i carry “weight” $w_i \geq 0$, and to the distinct value Y_j , attach the probability $p_j = \sum_{i|X_i=Y_j} w_i$ for $j = 1, \dots, k$.

In terms of weights the likelihood ratio is $\prod_{i=1}^n nw_i$ and the computation becomes to

maximize $\prod_{i=1}^n w_i$, or equivalently $\sum_{i=1}^n \log w_i$ subject to $w_i \geq 0$, $\sum_i w_i = 1$ and $\sum_i w_i X_i = \mu$. If $\mu \in B_n$ then an argument based on Lagrange multipliers shows that

$$w_i = w_i(\lambda) = \frac{1}{n} \frac{1}{1 + \lambda'(X_i - \mu)}$$

where $\lambda \in R^d$, the Lagrange multiplier, satisfies

$$0 = \sum_{i=1}^n \frac{X_i - \mu}{1 + \lambda'(X_i - \mu)}.$$

This value of λ may be found by minimizing $\sum_{i=1}^n \log w_i(\lambda)$ over λ . That maximizing the likelihood ratio over w_i is equivalent to minimizing it over λ is an example of convex duality. The global solution over $\lambda \in R^d$ may be found by many different algorithms, some of which are listed in Owen (1990). If $\mu \notin B_n$, then it is impossible to reweight the observations to have mean μ , and hence $\mathcal{R}(\mu) = 0$. Owen (1990) describes how to modify convex optimization methods so that they find $\mathcal{R}(\mu)$ when $\mu \in B_n$ and otherwise indicate that $\mu \notin B_n$. A function in the S language (Becker, Chambers and Wilks, 1988) for calculating $\mathcal{R}(\mu)$ is available on the internet at <http://playfair.stanford.edu/reports/owen/e1.S>.

Forcing $F(\{X_1, \dots, X_n\}) = 1$ reduces the problem from infinite dimensions to n . Convex duality further reduces the dimension to d . Simpler computations for the case $d = 1$ are described in Owen (1988).

If $\theta \in R^m$ is defined through estimating equations $E(g(X, \theta)) = 0$, then a test of the simple hypothesis $H_0 : \theta = \theta_0$ can be carried out by computing the random variables $Z_i = g(X_i, \theta_0)$ and then testing whether their common mean may be zero. Composite hypotheses which specify only some components of θ or only that some function $c(\theta) = 0$ require maximization over the other components of the minimum over λ described above.

Algorithms for complicated cases than the mean are discussed in Hall and La Scala

(1990), Owen (1990, 1992) and Wood, Do and Broom (1994).

5 Bartlett Correction

The limiting distribution of $\log \mathcal{R}(\theta_0)$ for a smooth function of means is typically $\chi_{(p)}^2$ where p is the number of components in θ . More careful analysis shows that typically $E(\log \mathcal{R}(\theta_0)) = p(1 + a/n) + O(n^{-2})$ for some value a , which may depend on θ_0 . A Bartlett corrected $100(1 - \alpha)\%$ confidence region has the form $C_{r,n,a} = \{T(F) \mid R(F) \geq r/(1 + a/n), F(\{X_1, \dots, X_n\}) = 1\}$ where $\Pr(\chi_{(p)}^2 \leq -2 \log(r)) = 1 - \alpha$. It is natural to expect some improvement from Bartlett correction, but in fact a surprisingly large improvement obtains and $\Pr(\theta_0 \in C_{r,n,a}) = 1 - \alpha + O(n^{-2})$.

This order of improvement is surprising because Bartlett correction is based only on discrepancies in the mean of $-2 \log \mathcal{R}(\theta_0)$ and takes no explicit account of the variance, skewness or other higher moments. Yet it results in a very small order of magnitude for the coverage errors.

In practice, learning the right value $a(\theta)$ may be harder than learning θ_0 itself, but it turns out that one can substitute a sample estimate $\hat{a} = a(\hat{\theta})$ and get $\Pr(\theta_0 \in C_{r,n,\hat{a}}) = 1 - \alpha + O(n^{-2})$ as well.

These results were found by DiCiccio, Hall and Romano (1991) and are surveyed in Hall and La Scala (1990). Zhang (1996) gives conditions under which empirical likelihood is Bartlett correctable for one dimensional M -estimates. Bartlett correction was originally established for parametric likelihood ratios. As yet no Bartlett corrections are known for the bootstrap or for nonparametric likelihoods other than empirical likelihood. Lazar and Mykland (1995) show that Bartlett correction of empirical likelihood does not work when forming a confidence region for only a subset of the parameters of interest.

6 Bootstraps and Least Favorable Families

Empirical likelihood developed from the work of Thomas and Grunkemeier (1975) mentioned above. Owen (1990) surveys other related work in the survival analysis literature and in the bootstrap literature.

Two of the bootstrap papers deserve special mention. A variant of Efron's (1981) non-parametric tilting bootstrap uses the same family of multinomial distributions as empirical likelihood. Efron prefers another family based on Kullback-Liebler distance because one can then apply exponential tilting to the bootstrap samples.

The empirical likelihood ratio function is nearly the same as the posterior distribution for Rubin's (1981) Bayesian Bootstrap with a non-informative prior distribution. Rubin resamples by generating random observation weights w_i from a Dirichlet distribution.

The best explanation for why empirical likelihood works is due to DiCiccio and Romano (1990) using Stein's concept of least favorable families. By maximizing over w_i for each fixed value of $T(F)$, one reduces the problem to a parametric subfamily, defined through $w_i = w_i(\lambda)$ in the case of the mean. This family has the same dimension as the statistic $T(F)$. Thus using empirical likelihood is like working with a data determined parametric subfamily of the simplex $w_i \geq 0$ with $\sum_i w_i = 1$.

The true distribution F_0 can be embedded into a parametric family of distributions in many ways. The problem of estimating $T(F_0)$ can be much easier in some parametric families than in others. Moreover, data from F_0 can never be used to distinguish between two different families containing F_0 . If one were to pick a family through F_0 for convenience, without having specific prior knowledge, then it would be best if that family did not make the problem artificially easy. A least favorable family is one in which estimating $T(F_0)$ is just as hard as it is nonparametrically. The parametric family used by empirical likelihood is asymptotically least favorable.

DiCiccio and Romano (1990) describe several data based approximations of the least favorable family, and show that inference in any of these families based on either likelihood ratios or resampling is asymptotically justified.

References

- Becker, R.A., Chambers, J.M. & Wilks, A.R. (1988). *The New S Language*. Wadsworth Brooks/Cole, Pacific Grove CA.
- Chen, S.X. (1993). On the Accuracy of Empirical Likelihood Confidence Regions for Linear Regression Model. *Ann. Inst. Statist. Math.* **45**, 621–637.
- Chen, S.X., & Hall, P.G. (1993). Smoothed Empirical Likelihood Ratios for Quantiles. *Ann. Statist.* **21**, 1166–1181.
- DiCiccio, T.J., Hall, P.J. & Romano, J. (1991). Empirical Likelihood is Bartlett-Correctable. *Ann. Statist.* **19**, 1053–1061.
- DiCiccio, T.J. & Romano, J. (1990). Nonparametric Confidence Limits by Resampling Methods and Least Favorable Families. *I.S.I. Review* **58**, 59–76.
- Efron, B. (1981). Nonparametric Standard Errors and Confidence Intervals (with Discussion). *Canadian Journal of Statistics* **9**, 139–172.
- Gill, R.D., Vardi, Y. & Wellner, J.A. (1988). Large sample theory of empirical distributions in biased sampling. *Ann. Statist.* **16**, 1069–1112.
- Hall, P. & La Scala, B. (1990). Methodology and Algorithms of Empirical Likelihood. *I.S.I. Review* **58**, 109–127.
- Hall, P. & Owen, A.B. (1993). Empirical Likelihood Confidence Bands in Density Estimation. *Jour. Comp. Graph. Stat.* **2**, 273–289.
- Kaplan, E.L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *J. Am. Statist. Assoc.* **53**, 457–481.
- Kolaczyk, E. (1994). Empirical Likelihood for Generalized Linear Models. *Statist. Sinica* **4**, 199–218.

- Kiefer, J. & Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Ann. Math. Statist.* **27**, 887–906.
- Lazar, N., & Mykland, P. A. (1995). “Empirical Likelihood in the Presence of Nuisance Parameters”. Technical report no. 400, Dept. of Statistics, University of Chicago.
- Mykland, P. A. (1995). Dual Likelihood. *Ann. Statist.* **23**, 396–421.
- Neyman, J. & Scott, E.L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica* **16**, 1–16.
- Owen, A.B. (1988). Empirical Likelihood Ratio Confidence Intervals For a Single Functional. *Biometrika* **75**, 2, 237–249.
- Owen, A.B. (1990). Empirical Likelihood Ratio Confidence Regions. *Ann. Statist.* **18**, 90–120.
- Owen, A.B. (1991). Empirical Likelihood for Linear Models. *Ann. Statist.* **19**, 1725–1747.
- Owen, A.B. (1992). “Empirical Likelihood and Generalized Projection Pursuit”. Dept. of Statistics Tech. Rep. 393, Stanford University, Stanford CA.
- Owen, A.B. (1995). Nonparametric Likelihood Confidence Bands for a Distribution Function. *J.A.S.A.* **90**, 516–521.
- Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Stat.* **22**, 86–91.
- Qin, J. (1994). Semi-Empirical Likelihood Ratio Confidence Intervals for the Difference of two Sample Means. *Ann. Inst. Statist. Math.* **46**, 117–126.
- Qin, J. (1995). “Semi-Empirical Likelihood to Detect Changes of Distribution Function”. Manuscript.
- Qin, J. & Lawless, J.F. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist.* **22**, 300–325.
- Rubin, D.B. (1981). The Bayesian Bootstrap. *Ann. Statist* **9**, 130–134.
- Thomas, D.R. & Grunkemeier, G.L. (1975). Confidence Interval Estimation of Survival Probabilities for Censored Data. *J. Am. Statist. Assoc.* **70**, 865–871.
- Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly

censored data. *J.A.S.A.* **69**, 169–173.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J.R.S.S.-B* **38**, 290–295.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–620.

Vardi, Y. (1985). Empirical Distributions in Selection Bias Models. *Ann. Statist.* **13**, 178–203.

Wood, A.T.A., Do, K.-A. & Broom, B.M. (1994). “Sequential Linearization of Empirical Likelihood Constraints with Application to U-Statistics”. Centre for Mathematics and its Applications, Research Report No. SRR 033-94.

Zhang, B. (1996). On the Accuracy of Empirical Likelihood Confidence Intervals for M-Functionals. *Jour. Nonpar. Statist.* To appear.