Journal of Statistical Planning and Inference 140 (2010) 971–981

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Empirical likelihood based variable selection

Asokan Mulayath Variyath [a,*], Jiahua Chen [b], Bovas Abraham [c]

[a] Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada A1C 5S7
[b] Department of Statistics, University of British Columbia, Vancouver, Canada V6T 1Z2
[c] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 25 July 2008
Received in revised form
23 September 2009
Accepted 28 September 2009
Available online 4 October 2009

Keywords:
Empirical likelihood
Variable selection

ABSTRACT

Information criteria form an important class of model/variable selection methods in statistical analysis. Parametric likelihood is a crucial part of these methods. In some applications such as the generalized linear models, the models are only specified by a set of estimating functions. To overcome the non-availability of well defined likelihood function, the information criteria under empirical likelihood are introduced. Under this setup, we successfully solve the existence problem of the profile empirical likelihood due to the over constraint in variable selection problems. The asymptotic properties of the new method are investigated. The new method is shown to be consistent at selecting the variables under mild conditions. Simulation studies find that the proposed method has comparable performance to the parametric information criteria when a suitable parametric model is available, and is superior when the parametric model assumption is violated. A real data set is also used to illustrate the usefulness of the new method.

1. Introduction

In the beginning of a scientific investigation, often a large number of covariates, $\mathbf{x}$, are suspected to have influence on the response variable, y, of interest. Because not all covariates have appreciable influence, analysis based on a regression model including all covariates considered initially is not advised. A variable selection procedure is commonly used to identify a minimum number of covariates needed to properly model the regression relationship.

A large number of variable selection procedures have been developed in the literature. There are at least four broad groups of them such as: sequential approach, prediction error approach, information theoretic approach and penalized likelihood approach. The sequential approaches such as forward selection, backward elimination and stepwise regression methods consider covariates one at a time. They either add a covariate or remove a covariate into a regression model that is judged most appropriate at the moment. Such procedures are known to be greedy. We regard cross-validation as a prediction error approach. It evaluates a model by comparing its prediction to the response of a left-out sample. The resulting model may have lower prediction error to a future observation. Information theoretic approach selects models that best approximate the true model based on Kullback–Leibler information or maximize Bayes posterior probability (Akaike, 1973; Schwarz, 1978). Mathematically, information criteria are parametric likelihood based. A new class methods based on penalized likelihood (Tibshirani, 1996; Fan and Li, 2001) are lauded for their computational efficiency and stability.

* Corresponding author.
E-mail address: variyath@mun.ca (A.M. Variyath).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.09.025

In this paper, we develop an information theoretic approach to variable selection problem where a parametric likelihood is not available. A particular example is the generalized linear models (GLM) (Nelder and Wedderburn, 1972) where the conditional distribution of y given covariates $\mathbf{x}$ is assumed to be a member of an exponential family. The exponential family assumption can be relaxed through quasi-likelihood (Wedderburn, 1974). More generally, we can specify the model through a set of estimating functions (Godambe, 1960) leaving out distributional assumptions. Instead of parametric likelihood, we use non-parametric empirical likelihood (Owen, 1988, 2001) in the information theoretic approach. The empirical likelihood methods have many similar properties as its parametric counterparts. It is natural to advance this technique to the area of variable selection. We investigate the use of empirical likelihood based Akaike information criterion (AIC) and Bayesian information criterion (BIC).

We first adopt the adjusted empirical likelihood (Chen et al., 2008; Variyath, 2006) to ensure that the empirical likelihood is well defined under all submodels being considered. We show that the empirical and the parametric likelihoods based AIC and BIC have the same first order asymptotic properties. Simulation studies show that when a parametric likelihood is available, the parametric and empirical likelihood based AIC and BIC have similar performances. The empirical likelihood based AIC and BIC are superior when the parametric model is mis-specified. We also present an iterative numerical algorithm with guaranteed numerical convergence.

The rest of the paper is organized as follows. In Section 2, we present the empirical likelihood based variable selection method. Theoretical properties are presented in Section 3 and computational issues are discussed in Section 4. Some simulation studies are given in Section 5. In Section 6, we illustrate the proposed method with a real data set and some concluding remarks are given in Section 7.

2. Empirical likelihood based information criteria

We first briefly discuss the concept of empirical likelihood, adjusted empirical likelihood and their connections to the estimating functions. The information criteria based on adjusted empirical likelihood method are then proposed.

2.1. Empirical likelihood and estimating functions

Given 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 empirical log-likelihood function of F is given by

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

where $p_i = F(\{y_i\}) = \Pr(Y_i = y_i)$. Inferences on any functional of F, for instance the population mean $\theta = (\theta_1, \ldots, \theta_d)$, can be made through profile empirical likelihood. The profile empirical log-likelihood of θ is defined as

$$l_{EL}(\theta) = \sup\left\{ l_n(F) : p_i > 0, i = 1, \ldots, n; \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(y_i - \theta) = 0 \right\}.$$

The maximization problem is very simple. Given θ, $l_n(F)$ under these restrictions is maximized when

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

for $i = 1, 2, \ldots, n$ where $\hat{\lambda}^{\tau}$ denote transpose of $\hat{\lambda}$ with the Lagrange multiplier $\hat{\lambda} = \hat{\lambda}(\theta)$ being the solution of

$$\sum_{i=1}^{n} \frac{y_i - \theta}{1 + \lambda^{\tau}(y_i - \theta)} = 0.$$

Hence, we may write

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

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

$$W(\theta) = -2 \sum_{i=1}^{n} \log(n\hat{p}_i) = 2 \sum_{i=1}^{n} \log\{1 + \hat{\lambda}^{\tau}(y_i - \theta)\}.$$

Owen (1990) shows that, when θ_0 is the true population mean, $W(\theta_0) \to \chi^2_d$ in distribution as $n \to \infty$ similar to a result for the parametric likelihood (Wilks, 1938). This fact is very useful for hypothesis test on θ and the construction of confidence regions.

The empirical likelihood method can also be used to analyze data from models defined by estimating functions (Owen, 1991; Kolaczyk, 1994; Qin and Lawless, 1994). Consider a random sample from a population whose parameter of interest $\theta = (\theta_1, \ldots, \theta_d)$ is defined by vector-valued estimating functions $g(y; \theta)$ of size $P \times 1$ such that $E\{g(y; \theta)\} = 0$. Based on n iid

observations, $y_1, \ldots, y_n$, the profile log-likelihood ratio function of θ is defined as

$$W(\theta) = \inf\left\{-2\sum_{i=1}^{n}\log(np_i) : p_i > 0, i = 1, \ldots, n; \sum_{i=1}^{n} p_i = 1; \sum_{i=1}^{n} p_i g(y_i;\theta) = 0\right\} = 2\sum_{i=1}^{n}\log\{1 + \hat{\lambda}^{\tau} g(y_i,\theta)\}$$

with $\hat{\lambda} = \hat{\lambda}(\theta)$ being the solution of

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

Under some regularity conditions, we also have $W(\theta_0) \to \chi^2_P$ in distribution as $n \to \infty$ assuming $E\{g(y;\theta_0)g^{\tau}(y;\theta_0)\}$ is full rank. Again this is very similar to conclusions in parametric likelihood.

2.2. Adjusted empirical likelihood

The definition of $W(\theta)$ relies on finding positive p_i's such that $\sum_i p_i g(y_i;\theta) = 0$ for each θ. The solution exists if and only if the convex hull of the $g(y_i;\theta), i = 1, 2, \ldots, n$ contains zero as an inner point. When the model is correct, the solution exists with probability tending 1 as the sample size $n \to \infty$ for θ in a neighborhood of θ_0. However, for finite n and at some θ value, the equation often does not have a solution in p_i. A convention is to set the empirical log-likelihood ratio statistic to ∞ when it happens (Owen, 2001). As pointed out by Chen et al. (2008), such a convention can be problematic in many applications, including the variable selection problem to be discussed. To avoid this problem, we introduce the adjusted empirical likelihood.

Let $g_i = g(y_i;\theta)$, $\bar{g}_n = n^{-1}\sum_{i=1}^{n} g_i$ and define $g_{n+1} = -a_n\bar{g}_n$ for some positive constant a_n. We adjust the profile empirical log-likelihood ratio function to

$$W^*(\theta) = \inf\left\{-2\sum_{i=1}^{n+1}\log[(n+1)p_i] : p_i > 0, i = 1, \ldots, n+1; \sum_{i=1}^{n+1} p_i = 1; \sum_{i=1}^{n+1} p_i g_i = 0\right\},$$
$$= 2\sum_{i=1}^{n+1}\log\{1 + \hat{\lambda}^{\tau}(\theta) g_i\}$$

with $\hat{\lambda} = \hat{\lambda}(\theta)$ being the solution of

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

Since 0 always lies on the line connecting $\bar{g}_n$ and g_{n+1}, the adjusted empirical log-likelihood ratio function is well defined after adding a pseudo-value g_{n+1} to the data set. The adjustment is particularly useful so that a numerical program does not crash simply because some undesirable θ value is assessed. For a wide range of a_n, $W(\theta)$ and $W^*(\theta)$ have the same first order asymptotic properties (see Chen et al., 2008).

2.3. Information criteria

A full GLM assumes that the mean of y relates to $\mathbf{x}^{\tau}\beta$ through a known link function with β being an unknown regression coefficient vector of size P. Let s be a subset of $\{1, 2, \ldots, P\}$, and $\mathbf{x}[s]$ and $\beta[s]$ be subvectors of $\mathbf{x}$ and β containing entries in positions specified by s. A sub-GLM assumes that the mean of y relates to $\mathbf{x}^{\tau}[s]\beta[s]$ through the same link function. That is, only covariates in positions specified by s have significant influence. Under exponential family distribution assumption, a log-likelihood function $l(\beta[s])$ of $\beta[s]$ is available. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) for a submodel s are given by

$$AIC(s) = \inf\{-2l(\beta[s]) + 2k : \beta[s]\} \tag{2}$$

and

$$BIC(s) = \inf\{-2l(\beta[s]) + k\log(n) : \beta[s]\}, \tag{3}$$

where k is the cardinality of s. A submodel s that minimizes $AIC(s)$ or $BIC(s)$ will be selected as the most appropriate model in characterizing the relationship between $\mathbf{x}$ and y. Akaike (1973, 1974) motivated AIC as selecting a submodel that best approximates the true model in terms of the estimated Kullback–Leibler distance (Kullback and Leibler, 1951). Schwarz (1978) derived BIC as the optimal choice in terms of having the highest posterior probability.

When a parametric likelihood is not available, the above approaches must be revised. For over-dispersed Poisson count data, Lebreton et al. (1992) proposed the use of quasi-likelihood. Let $l(\beta[s])$ be the log-likelihood based on the Poisson model assumption. They modified the AIC and BIC for over-dispersed count data as

$$QAIC(s) = \inf\{-2l(\beta[s])/\hat{c} + 2k : \beta[s]\},$$
$$QBIC(s) = \inf\{-2l(\beta[s])/\hat{c} + k\log(n) : \beta[s]\} \tag{4}$$

$\hat{c}$ = estimated VIF.

with $\hat{c}$ being an estimated variance inflation factor. The factor c is not the same as the over-dispersion parameter. For brevity, we do not give details of $\hat{c}$ here. The QAIC or QBIC are then used for variable selection. Their method, however, does not apply to situations where the model is defined through estimation functions.

Consider the model specified by $E\{g(y;\mathbf{x}^\tau\beta)\}=0$ with g being a vector valued function of dimension P. A submodel is specified by $E\{g(y;\mathbf{x}^\tau[s]\beta[s])\}=0$. Note that g does not depend on s. For a given s, let $g_i=g(y_i;\mathbf{x}_i^\tau[s]\beta[s])$, $\bar{g}_n=n^{-1}\sum_{i=1}^n g_i$ and $g_{n+1}=-a_n\bar{g}_n$ for some positive constant a_n as before. The adjusted empirical log-likelihood ratio becomes

$$W^*(\beta[s]) = \inf\left\{-2\sum_{i=1}^{n+1}\log[(n+1)p_i] : p_i>0, i=1,\ldots,n+1; \sum_{i=1}^{n+1}p_i=1; \sum_{i=1}^{n+1}p_ig_i=0\right\}$$
$$= 2\sum_{i=1}^{n+1}\log\{1+\hat{\lambda}^\tau g_i\} \tag{5}$$

with $\hat{\lambda}=\hat{\lambda}(\beta[s])$ being the solution of

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

We define the adjusted profile empirical log-likelihood ratio as

$$W^*(s)=\inf\{W^*(\beta[s]) : \beta[s]\}. \tag{6}$$

The empirical likelihood versions of AIC and BIC are then defined as

$$EAIC(s)=W^*(s)+2k,$$
$$EBIC(s)=W^*(s)+k\log(n). \tag{7}$$

3. Asymptotic properties

It is well known that under some mild conditions the parametric BIC is consistent for variable selection while the parametric AIC is not. In this section, we prove that under similar conditions and when P is constant, the empirical likelihood based BIC is consistent but the empirical likelihood based AIC is not. In situations where P also goes to infinity as $n\to\infty$, the BIC may not be consistent as pointed out in Chen and Chen (2008). However, we do not consider this case in this paper. We leave all proofs in the appendix.

Theorem 1. *Suppose* $(y_i,\mathbf{x}_i), i=1,2,\ldots,n$ *is a set of independent and identically distributed random vectors. Let* $g_i(\beta)=g(y_i,\mathbf{x}_i^\tau\beta)$ *be the estimating function for* $\beta\in R^P$ *such that for each i,* $E\{g_i(\beta_0)\}=0$. *Let* $A=E\{g_i(\beta_0)g_i^\tau(\beta_0)\}$ *and* $B=E\{(\partial g/\partial\beta)|_{\beta=\beta_0}\}$ *where* $g=g(y;\mathbf{x}^\tau\beta)$. *We also assume that*:

(i) *A is positive definite and B has rank P.*
(ii) *The second derivatives of each component of g, say* $g[k]$, $\partial^2 g[k]/\partial\beta^2$, *a* $P\times P$ *matrix with the (ij)th entry* $\partial^2 g[k]/(\partial\beta_i\partial\beta_j)$, *is continuous in* β *in a neighborhood of* β_0.
(iii) *there exists some function* $G(y,\mathbf{x})$ *with finite expectation such that*

$$\left|\frac{\partial g}{\partial\beta}\right|<G(y,\mathbf{x}); \quad |g(y,\mathbf{x}^\tau\beta)|^3<G(\mathbf{x},y),$$

where the inequality is interpreted component-wise for the matrix and vector on the left-hand side.

Under above assumptions, there exists a sequence of adjusted empirical likelihood estimates $\hat{\beta}$ *of* β *such that*

$$\sqrt{n}(\hat{\beta}-\beta_0)\to N(0,\{B^\tau A^{-1}B\}^{-1}), \sqrt{n}(\hat{\lambda}-0)\to N(0,U),$$

where

$$U=A^{-1}-A^{-1}B\{B^\tau A^{-1}B\}^{-1}B^\tau A^{-1}.$$

We point out that Condition (i) simply means neither the estimating function g nor the parameter vector β degenerate. The moment condition (iii) ensures the applicability of central limit theorem to the leading term in the expanded likelihood, while the condition (ii) makes the expansion possible.

Recall that we denote a submodel by a size k subset s of $\{1,2,\ldots,P\}$. When a submodel s is a true model, it implies $\beta_0[\bar{s}]=0$. That is, components of β_0 not in s are zero. The following theorem shows that when $\beta_0[\bar{s}]=0$ is assumed in the analysis and the assumption is true, then adjusted empirical log-likelihood ratio statistic has chisquared limiting distribution with k fewer degrees of freedom, just like the usual empirical likelihood.

Theorem 2. *Assume the same conditions as in Theorem* 1 *and that* $\beta_0[\overline{s}] = 0$ *for a submodel s of size k. Then when* $a_n = o_p(n^{1/2})$, *we have* $W^*(s) \to \chi^2_{(P-k)}$ *in distribution as* $n \to \infty$, *where* $W^*(s)$ *is defined in* (6).

When the null hypothesis of $\beta[\overline{s}] = 0$ is not true, the likelihood ratio go to ∞ as $n \to \infty$. We state the following theorem in terms of the adjusted empirical likelihood which also applies to the usual empirical likelihood.

Theorem 3. *Assume the conditions of Theorem* 1 *and* $a_n = o_p(n^{1/2})$. *Then for any* $\beta \neq \beta_0$ *such that* $E\{g(y, \mathbf{x}^\tau\beta)\} \neq 0$, *the adjusted empirical likelihood function as defined in* (5), $W^*(\beta) \to \infty$ *in probability as* $n \to \infty$.

Assume there exists a subset s_0 *of* $\{1, 2, \ldots, P\}$ *such that for any other subset s,* $E\{g(y, \mathbf{x}^\tau[s]\beta[s])\} = 0$ *for some* β *if any only if s contains* s_0. *Then, EBIC is consistent and EAIC is not consistent.*

Note that the assumption, $E\{g(y, \mathbf{x}^\tau[s]\beta[s])\} = 0$ if and only if s contains s_0, means that s_0 is identifiable. In addition, our proof is valid when P does not change with n.

4. Numerical consideration

Numerically, for each s and $\beta[s]$, we first use the modified Newton–Raphson algorithm of Chen et al. (2002) for computing Lagrange multiplier to obtain $W^*(\beta[s])$. We then use the simplex method of Nelder and Mead (1965) to maximize $W^*(\beta[s])$ with respect to $\beta[s]$. Due to the use of adjusted empirical likelihood, $W^*(\beta[s])$ is always well defined and the modified Newton–Raphson algorithm converges fast. The simplex method is numerically stable which is particularly useful in this application. Most of the optimization softwares include built-in functions for this method.

Our numerical computation for $W^*(s)$ goes as follows.

1. Initiation: (a) use any standard software package to obtain the estimate of β and $\beta[s]$ under the full model and candidate submodel, respectively; (b) let $\beta^0[s]$ be the average of the two estimates over the components in s; (c) set $\lambda^0 = 0$, $c = 0, \gamma^c = 1$, and $\varepsilon = 10^{-8}$.
2. At the c th iteration: (a) compute $g_i = g_i(\beta^c[s])$, $i = 1, 2, \ldots, n$ and $g_{n+1} = -a_n\overline{g}_n$. In our simulation, some trimmed mean is used in place of $\overline{g}_n$ for robustness; (b) compute the first and second derivatives, R_λ and $R_{\lambda\lambda}$, of $R = \sum_{i=1}^{n+1} \log\{1 + \lambda^c g_i\}$ with respect to λ; (c) compute $\Delta(\lambda^c) = -R_{\lambda\lambda}^{-1}R_\lambda$.
3. Check convergence: if $\|\Delta(\lambda^c)\| < \varepsilon$ go to Step 6; otherwise continue.
4. Check feasibility: compute $\delta^c = \gamma^c\Delta(\lambda^c)$. If $1 + (\lambda^c - \delta^c)^\tau g(\beta) \leq 0$ for some i or $R(\lambda^c - \delta^c) < R(\lambda^c)$, let $\gamma^c = \gamma^c/2$ and repeat the current Step 4; otherwise, continue to the next step.
5. Updating parameters: (a) $\lambda^{c+1} = \lambda^c - \delta^c, c = c + 1$; (b) $\gamma^{c+1} = (c+1)^{-1/2}$. Go back to Step 2 for looping.
6. Report the outcome as
$$W^*(\beta[s]) = \sum_{i=1}^{n+1} \log\{1 + \lambda^c g_i\}.$$
7. Minimize $W^*(\beta[s])$ with respect to $\beta[s]$ using some existing software package. Report the outcome as $W^*(s)$.

After $W^*(s)$ is evaluated for all s, we compute *EAIC*(s) and/or *EBIC*(s) as per Eq. (7) and select the model with minimum EAIC or EBIC value.

In some applications, the number of all possible submodels can be so large that it becomes infeasible and/or insensible to evaluate them all. To speed up our simulation analysis we adopted a few simple strategies to reduce the number of submodels to be evaluated. For each data set generated, we first fit the full model and identify covariates whose p-values are smaller than, say 10^{-5}, for testing its significance. We then only consider submodels containing at least one of these important covariates. In addition, a submodel can have such a large EAIC or EBIC that including more covariates will not change its fate. All submodels containing this submodel will be subsequently left out from further evaluation.

5. Simulation studies

When the data are known to be from some parametric model, both parametric and non-parametric methods are applicable. We compared the EAIC, EBIC with the parametric AIC and BIC based on data generated from linear and logistic regression models. Simulation results, which are not provided here to save space, indicate that performances are comparable. Hence, the new methods are viable alternatives even if the parametric model assumption is plausible.

Next, we evaluate the new method when parametric model assumptions are violated. We compare the EAIC and EBIC to AIC and BIC under an assumed Poisson regression model. We also compare them to QAIC and QBIC which are specifically designed to deal with over-dispersed data. The over-dispersed Poisson regression model is specified through a conditional density given by

$$f(y|u; \mu) = \frac{(u\mu)^y \exp\{-u\mu\}}{y!}$$

with u being a random variable such that $E(u) = 1$ and $\text{var}(u) = \omega$. Marginally, we have

$$E(y) = \mu;\ \text{var}(y) = \mu(1 + \mu\omega).$$

In particular, the distribution of u is chosen to be gamma with parameters $(\omega, 1/\omega)$ with ω being the over-dispersion parameter.

We consider a four-covariates generalized linear model such that

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

with $\beta = (0.5, 0.5, 0.6, 0, 0)$. The model with the first two covariates is the true model. We consider the following two scenarios of over-dispersion.

Case I: *Constant over-dispersion.* Covariates $x = (x_1, x_2, x_3, x_4)$ were generated from the multivariate normal distribution with mean 0 and covariance structure $\text{cov}(x_i, x_j) = (0.5)^{|i-j|}$. We choose four levels of over-dispersion $\omega = 0, 1/8, 1/6, 1/4$. A large value of ω indicates severe over-dispersion.

Case II: *Over-dispersion depends on covariates.* Covariates $x = (x_1, x_2, x_3, x_4)$ were generated from the multivariate normal distribution with mean (0, 0, 3, 4) and covariance structure $\text{cov}(x_i, x_j) = (0.5)^{|i-j|}$. The over-dispersion parameter ω for individual response variable y is made a function of (x_1, x_2, x_3, x_4) through the relationship $\omega = (ax_3 + bx_4)^{-1}$. Three pairs of constants (a, b), (10,000, 10,000), (1, 1) and (0.5, 0.5) were used in the simulation.

In each simulation run, we generated $n = 100$ observations for the response y from a conditional distribution specified earlier. For each model set up, we analyzed 1000 simulated data sets. We recorded the percentages of times when the selected model (1) is the true model (TM); (2) contains the true model with at most one additional covariate (TM + 1); (3) contains the true model with at most two additional covariates (TM + 2).

We see from Tables 1 and 2 that when there is no over-dispersion, empirical likelihood methods have similar performance with parametric methods. When there is over-dispersion, the percentage of times the true model is selected by EAIC and EBIC is higher than AIC and BIC. Note that var(y) increases with the over-dispersion parameter ω, and this is accompanied by the decrease in the percentage of times the true model is selected. In comparison, the rate of decrease is faster for AIC than for EAIC, and faster for BIC than for EBIC. For example in case I when $\omega = 1/6$, the true model is selected 53% of the times by AIC and 65% of times by EAIC. The corresponding figures are 80% and 90% for BIC and EBIC. Simulation results show that the QAIC and QBIC have similar performances to EAIC and EBIC. This is encouraging for the empirical likelihood because it does not model the over-dispersion.

Table 1
Comparison of variable selection methods based %time the model selected in Poisson regression model: case I.

ω	Model	AIC	QAIC	EAIC	BIC	QBIC	EBIC
0	TM	72.9	72.0	70.9	93.7	93.1	92.4
	TM + 1	96.6	96.4	96.2	99.5	99.6	99.5
	TM + 2	100.0	100.0	100.0	99.7	99.8	99.7
1/8	TM	58.2	66.5	66	82.9	88.3	88.8
	TM + 1	93.0	94.4	95.0	96.9	97.1	97.4
	TM + 2	99.1	99.8	99.8	98.9	98.3	98.0
1/6	TM	52.8	63.6	64.6	80.0	88.4	90.0
	TM + 1	92.0	95.5	95.3	98.8	99.3	99.2
	TM + 2	100.0	100.0	99.8	100.0	99.7	99.7
1/4	TM	45.7	60.6	63.9	76.2	89.7	87.9
	TM + 1	87.4	93.8	94.7	97.2	99.1	97.7
	TM + 2	100.0	99.9	99.8	99.9	99.6	98.6

Table 2
Comparison of variable selection methods based %time the model selected in Poisson regression model: case II.

(a, b)	Model	AIC	QAIC	EAIC	BIC	QBIC	EBIC
(10,000, 10,000)	TM	58.7	58.6	58.0	69.8	69.6	67.9
	TM + 1	97.9	97.6	96.7	99.7	99.7	99.5
	TM + 2	100.0	100.0	100.0	100.0	100.0	100.0
(1, 1)	TM	45.0	49.5	49.3	57.0	60.8	60.9
	TM + 1	93.4	95.8	95.9	94.3	98.0	97.9
	TM + 2	99.9	99.8	99.8	99.7	99.2	98.8
(0.5, 0.5)	TM	30.1	38.4	40.9	43.5	50.4	52.0
	TM + 1	85.5	91.7	93.8	95.0	96.4	96.4
	TM + 2	99.9	99.6	99.8	99.4	98.4	97.6

Table 3
Comparison of variable selection methods using NCZC and NIZC in Poisson regression model: case I.

ω	Average	AIC	QAIC	EAIC	BIC	QBIC	EBIC
0	NCZC	1.695	1.684	1.671	1.937	1.930	1.925
	NIZC	0	0	0	0.001	0.002	0.003
1/8	NCZC	1.512	1.609	1.614	1.802	1.858	1.902
	NIZC	0	0.001	0.003	0.004	0.004	0.034
1/6	NCZC	1.448	1.591	1.603	1.789	1.884	1.899
	NIZC	0.004	0.007	0.005	0.014	0.030	0.033
1/4	NCZC	1.331	1.546	1.588	1.736	1.896	1.882
	NIZC	0.005	0.009	0.009	0.016	0.044	0.056

Table 4
Comparison of variable selection methods using NCZC and NIZC in Poisson regression model: case II.

(a, b)	Average	AIC	QAIC	EAIC	BIC	QBIC	EBIC
(10,000, 10,000)	NCZC	1.566	1.562	1.547	1.695	1.693	1.674
	NIZC	0.269	0.269	0.269	0.272	0.271	0.283
(1, 1)	NCZC	1.386	1.457	1.456	1.519	1.584	1.612
	NIZC	0.309	0.318	0.328	0.333	0.339	0.355
(0.5, 0.5)	NCZC	1.158	1.306	1.350	1.393	1.496	1.518
	NIZC	0.403	0.425	0.440	0.439	0.471	0.488

We also computed the average number of correct zero coefficients (NCZC) and incorrect zero coefficients (NIZC) selected by these methods over 1000 repetitions under all set ups. The results are given in Tables 3 and 4. From Table 3, the average numbers of correct zero coefficients in general decrease when the over-dispersion becomes more severe. In comparison, however, the rates of decreases for EAIC and EBIC are slower than other methods. At the same time, the average numbers of incorrect zero coefficients increase, and the rates are higher for EAIC and EBIC. This is expected since when over-dispersion increases, variability in $\hat{\beta}$ also increases leading to a decrease in the significance level of $\hat{\beta}$. Hence, the contribution of some of the covariates may be reduced. The trends are similar in Table 4 for case II. In conclusion, when a model is mis-specified, the EAIC and EBIC perform better than the AIC and BIC.

6. Example

We consider the data set 'doctor visits' which is discussed in detail in Cameron and Trivedi (1998). The original source of this data set is the Australian Health Survey 1977–78. We used this data set to apply variable selection methods with Poisson regression.

The response of interest in this survey is the health of adults. This is measured in terms of the number of doctor visits (Y) made by an adult in the past two weeks. The data set contains information on the number of doctor visits and several measures of health service utilization and socioeconomic parameters on 5190 adults. Cameron et al. (1988) analyzed this data in the study of an economic model of joint determination of health service utilization and health insurance choices. Cameron and Trivedi (1986) also studied this data set in a different context. The main objective of our analysis is to model the relationship between the response and the covariates and to identify the simplest model that gives a good description of the data-generating mechanism. A short description of the response variable and the covariates is given in Table 5.

Ninety-eight percent of the responses (Y values) are 0, 1, or 2 and there are few large values with maximum being 9. The mean of the response variable is 0.302 and the standard deviation is 0.798. The raw data indicate that there is over-dispersion. Since we are utilizing the information on covariates to model the response, we need to verify that inclusion of covariates eliminates over-dispersion. Correlation analysis indicates that some covariates are highly correlated. As expected, X2 and X3 are highly correlated since one is the square of the other. The correlation of X7 with X2 and X3 is around 0.6, and of X7 with X4 and X5 is around 0.4. Other correlations are negligible. This indicates that the estimates of the regression coefficients of submodels may differ from those of the full model. The best submodel may not be identified by the significance levels of the regression estimates of the full model. We fit a Poisson regression model to the data. The estimates of regression coefficients are given in Table 6.

Table 5
Variable codes and descriptions of data set 'doctor visits'.

Variables	Description
Y-Dvisits	Number of doctor visits in past two weeks
X1-SEX	1 if female, 0 if male
X2-AGE	Age in years divided by 10
X3-AGESQ	AGE squared
X4-INCOME	Annual income in Australian dollars divided by 1000
X5-LEVYPLUS	1 if covered by private health insurance
X6-FREEPOOR	1 if covered by government because of low income
X7-FREEREPA	1 if covered free by government because of old age, disability pension, or veteran status
X8-ILLNESS	Number of illnesses in past two weeks
X9-ACTDAYS	Number of days of reduced activity in past two weeks due to illness or injury
X10-HSCORE	General health questionnaire score using Goldberg's method
X11-CHCOND1	1 if chronic condition(s) but not limited in activity, 0 otherwise
X12-CHCOND2	1 if chronic condition(s) and limited in activity, 0 otherwise

Table 6
Estimates of Poisson regression coefficients for the full model.

Variable	Coefficient (β)	Standard error (s.e.)	$z = \beta/\text{s.e.}$	$P[Z > z]$
Intercept	−2.2238	0.1898	−11.716	$<10e-11$
X1-SEX	0.1569	0.0561	2.795	0.0052
X2-AGE	1.0563	1.0008	1.055	0.2912
X3-AGESQ	−0.8487	1.0778	−0.787	0.4310
X4-INCOME	−0.2053	0.0884	−2.323	0.0202
X5-LEVYPLUS	0.1232	0.0716	1.720	0.0855
X6-FREEPOOR	−0.4400	0.1798	−2.447	0.0144
X7-FREEREPA	0.0798	0.0921	0.867	0.3861
X8-ILLNESS	0.1869	0.0183	10.227	$<10e-11$
X9-ACTDAYS	0.1268	0.0050	25.198	$<10e-11$
X10-HSCORE	0.0301	0.0101	2.979	0.0029
X11-CHCOND1	0.1141	0.0666	1.712	0.0869
X12-CHCOND2	0.1416	0.0831	1.698	0.0896

Table 7
Estimates of regression coefficients and standard errors for different models identified.

Variables	AIC	EAIC	QAIC	BIC	EBIC	QBIC
Intercept	−2.0891	−2.2049	−2.0520	−2.2444	−2.0486	−2.2226
	(0.1008)	(0.0691)	(0.0995)	(0.0679)	(0.0517)	(0.0542)
X1	0.1620	0.2003	0.1755	0.2056	0.2627	0.2100
	(0.0558)	(0.0542)	(0.0554)	(0.0542)	(0.0527)	(0.0542)
X2	0.3551	0.5168	0.4335	0.5694	–	0.5426
	(0.1432)	(0.1319)	(0.1371)	(0.1307)	–	(0.1302)
X4	−0.1998	–	−0.1711	–	–	–
	(0.0843)	–	(0.0819)	–	–	–
X5	0.0837	–	–	–	–	–
	(0.0535)	–	–	–	–	–
X6	−0.4696	−0.4375	−0.4963	–	–	–
	(0.1764)	(0.1731)	(0.1753)	–	–	–
X8	1.1861	0.1988	0.1960	0.1997	0.2303	0.2157
	(0.0183)	(0.0175)	(0.0176)	(0.0175)	(0.0165)	(0.0168)
X9	0.1266	0.1277	0.1278	0.1279	0.1363	0.1334
	(0.0050)	(0.0049)	(0.0049)	(0.0049)	(0.0045)	(0.0046)
X10	0.0311	0.0334	0.0324	0.0320	–	–
	(0.0101)	(0.0990)	(0.0099)	(0.0099)	–	–
X11	0.1211	–	–	–	–	–
	(0.0664)	–	–	–	–	–
X12	0.1589	–	–	–	–	–
	(0.0818)	–	–	–	–	–

From Table 6, we see that X8 and X9 are highly significant. This is expected since these covariates are clear indicators of illness. The covariate SEX is also significant indicating that females visit the doctors more frequently than males.

In view of the background information, a Poisson regression model might be appropriate. Since the raw data indicate over-dispersion, we also consider a negative binomial model with two types of variance functions $\text{var}(y) = \mu + \alpha\mu$ and $\text{var}(y) = \mu + \alpha\mu^2$, both deviate from the Poisson model. We applied the likelihood ratio test to the null hypothesis of $\alpha = 0$. The hypothesis is rejected implying significant over-dispersion. Apparently, Poisson model based AIC and BIC for model selection are not suitable. We therefore conclude that the empirical likelihood based AIC and BIC discussed in this paper are appropriate for variable selection for this data set. We also analyzed the data with QAIC and QBIC. The final model chosen by methods under consideration together with their corresponding parameter estimations are given in Table 7.

From Table 7 we see that all the variable selection methods identified the important covariates based on the full model analysis. The performances of QAIC and EAIC, and of QBIC and EBIC, are very similar. The AIC and BIC select larger models compared to EAIC and EBIC. Overall, EBIC selected the simplest model.

7. Conclusions

In this paper, we proposed an empirical likelihood based variable selection method. This method is particularly useful when there is no suitable parametric model for the data. We also designed a computational strategy for implementation of the new method. The method is shown to be consistent, and simulations indicate that the method has comparable performance with parametric methods even when parametric distributional assumptions are appropriate. The empirical likelihood based variable selection is superior to parametric likelihood based methods when there is mis-specification. Thus the new methods are robust to model mis-specifications.

Acknowledgement

The authors would like to thank the editor and the anonymous referee for their constructive suggestions, which improved the earlier version of this paper. The work of Dr. Variyath was partially supported by a grant from the Memorial University of Newfoundland and the work of Drs. Chen and Dr Abraham was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

Appendix

Proof of Theorem 1. Because the result and the proof is similar to Qin and Lawless (1994), we only provide a partial proof by assuming $\hat{\beta}$ is consistent for β. Under this assumption and due to the smoothness of the estimating function, $\hat{\beta}$ must be a stationary point of $W^*(\beta)$. Hence, $\hat{\beta}$ together with some $\hat{\lambda}$ depending on $\hat{\beta}$, solves

$$Q_{1,n+1}(\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{g_i(\beta)}{1+\lambda^\tau g_i(\beta)} = 0,$$

$$Q_{2,n+1}(\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{1}{1+\lambda^\tau g_i(\beta)}\left(\frac{\partial g_i}{\partial\beta}\right)^\tau \lambda = 0.$$

We now omit (β,λ) in the notation if a function is evaluated at $(\beta_0,0)$. Expanding $Q_{1,n+1}(\hat{\beta},\hat{\lambda})$ and $Q_{2,n+1}(\hat{\beta},\hat{\lambda})$ at $(\beta,\lambda) = (\beta_0,0)$ and ignoring the high order terms leads to

$$Q_{1,n+1}(\hat{\beta},\hat{\lambda}) = Q_{1,n+1} + \left\{\frac{\partial Q_{1,n+1}}{\partial\beta}\right\}(\hat{\beta}-\beta_0) + \left\{\frac{\partial Q_{1,n+1}}{\partial\lambda}\right\}\hat{\lambda} = 0, \tag{8}$$

$$Q_{2,n+1}(\hat{\beta},\hat{\lambda}) = Q_{2,n+1} + \left\{\frac{\partial Q_{2,n+1}}{\partial\beta}\right\}(\hat{\beta}-\beta_0) + \left\{\frac{\partial Q_{2,n+1}}{\partial\lambda}\right\}\hat{\lambda} = 0, \tag{9}$$

Let $A = E\{gg^\tau\}$ and $B = E\{\partial g/\partial\beta\}$. Note that at $(\beta_0,0)$, we have

$$\frac{\partial Q_{1,n+1}}{\partial\beta} = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{\partial g_i}{\partial\beta} \to B,$$

$$\frac{\partial Q_{1,n+1}}{\partial\lambda} = -\frac{1}{n+1}\sum_{i=1}^{n+1} g_i g_i^\tau \to -A,$$

$$\frac{\partial Q_{2,n+1}}{\partial\beta} = 0,$$

$$\frac{\partial Q_{2,n+1}}{\partial \lambda} = \frac{1}{n+1}\sum_{i=1}^{n+1}\left\{\frac{\partial g_i}{\partial \beta}\right\}^{\tau} \to B^{\tau}.$$

Replacing the above quantities by their limits, and keeping track of the orders, we find

$$\hat{\lambda} = -n^{-1}\{A^{-1} - A^{-1}B(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}\}Q_{1,n+1} + o_p(n^{-1/2}), \tag{10}$$

$$\hat{\beta} - \beta_0 = n^{-1}(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}Q_{1,n+1} + o_p(n^{-1/2}). \tag{11}$$

Note that the choice of $a_n = o_p(n^{1/2})$ together with the moment conditions on g makes $Q_{1,n+1} = \overline{g}_n(1 - a_n/n) = O_p(n^{-1/2})$. This fact also reveals that the above remainder term is $o_p(n^{-1/2})$.

Applying the central limit theorem to $Q_{1,n+1}$ and using Slustzky's theorem, we get the conclusions of Theorem 1. □

Proof of Theorem 2. Let $g_i = g_i(\beta[s]) = g(y_i, \mathbf{x}_i^{\tau}[s]\beta[s])$ and $\hat{\lambda}$ be the Lagrange multiplier corresponding to $\hat{\beta}[s]$, the maximum point of $W^*(\beta[s])$. With these notation, we may write

$$W^*(s) = 2\sum_{i=1}^{n+1}\log\{1 + \hat{\lambda}^{\tau}g_i(\hat{\beta}[s])\}.$$

At the same time, we have

$$\hat{\lambda}^{\tau}g_i(\hat{\beta}[s]) = \hat{\lambda}^{\tau}g_i + \hat{\lambda}^{\tau}\left\{\frac{\partial g_i}{\partial \beta[s]}\right\}^{\tau}(\hat{\beta}[s] - \beta_0[s]) + o_p(n^{-1}).$$

The expansions (10) and (11) for $\hat{\lambda}$ and $\hat{\beta}$ are still valid with new $B = E\{\partial g/\partial \beta[s]\}$ and the same matrix A. Hence, by simple Taylor's expansion,

$$\begin{aligned} W^*(s) &= 2\sum_{i=1}^{n+1}\{1 + \hat{\lambda}^{\tau}g_i(\hat{\beta}[s])\} \\ &= 2\hat{\lambda}^{\tau}\sum_{i=1}^{n+1}g_i + 2\hat{\lambda}^{\tau}\left\{\sum_{i=1}^{n+1}\frac{\partial g_i}{\partial \beta[s]}\right\}^{\tau}(\hat{\beta}[s] - \beta_0[s]) - \hat{\lambda}^{\tau}\left\{\sum_{i=1}^{n+1}g_i g_i^{\tau}\right\}\hat{\lambda} + o_p(1) \\ &= n^{-1}Q_{1,n+1}^{\tau}\{A^{-1} - A^{-1}B(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}\}Q_{1,n+1} + o_p(1). \end{aligned}$$

Note that $Q_{1,n+1}$ is asymptotic normal with covariance matrix A. In addition,

$$\begin{aligned} &\{A^{-1} - A^{-1}B(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}\}A\{A^{-1} - A^{-1}B(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}\} \\ &\quad = \{A^{-1} - A^{-1}B(B^{\tau}A^{-1}B)^{-1}B^{\tau}A^{-1}\} \end{aligned}$$

which makes the quadratic form in $Q_{1,n+1}$ satisfy the idempotent condition as given in Serfling (1980, p. 128). Hence, we have $W^*(s) \to \chi^2_{(p-k)}$ in distribution as $n \to \infty$. □

Proof of Theorem 3. Denote again that $g_i(\beta) = g(y_i, \mathbf{x}_i^{\tau}\beta)$ for $i = 1, \ldots, n$ and similarly for $\overline{g}_n(\beta)$ and $g_{n+1}(\beta)$. Since $E\{g(y, \mathbf{x}^{\tau}\beta)\} \neq 0$, there must exist some $\delta > 0$ such that $\overline{g}_n^{\tau}(\beta)\overline{g}_n(\beta) \to \delta^2$ in probability by the law of large numbers. Note that $g_i(\beta) - \overline{g}_n(\beta)$ has mean zero and satisfies all moment conditions to ensure that

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

Let $\tilde{\lambda} = n^{-2/3}\overline{g}_n(\beta)\log(n)$, where $\log(n)$ is used as a quantity that goes to infinite at a slow rate. We notice that

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

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

$$W^*(\beta) = \sup_{\lambda}\left[2\sum_{i=1}^{n+1}\log\{1 + \lambda^{\tau}g_i(\beta)\}\right] \geq 2\sum_{i=1}^{n+1}\log\{1 + \tilde{\lambda}^{\tau}g_i(\beta)\} = 2\sum_{i=1}^{n}\log\{1 + \tilde{\lambda}^{\tau}g_i(\beta)\} + o_p(1) = 2n^{1/3}\delta^2\log(n) + o_p(1).$$

The last equality is from the fact that $\overline{g}_n^{\tau}(\beta)\overline{g}_n(\beta) \to \delta^2$. Hence, $W^*(\beta) \to \infty$. Similarly, the unadjusted empirical likelihood ratio function $W(\beta) \to \infty$.

We next consider the selection consistency.

Let us consider the EAIC first. Consider the situation when s_0 is empty. That is, none of the covariates are significant. Let $s = \{1\}$ which contains a single covariant. It is seen that $W^*(s) - W^*(s_0) \to \chi^2_1$ based on expansion in the proof of Theorem 2. Thus, $\lim_{n\to\infty}P\{W^*(s) - W^*(s_0) > 2\} > 0$. That is, by EAIC, the model s has non-diminishing probability to be selected over the true model s_0. Thus, EAIC is not consistent.

Next, we consider EBIC. Suppose s is a model which does not contain s_0. Then, $Eg(y, \mathbf{x}[s]^{\tau}\beta[s]) \neq 0$ for any $\beta[s]$. Therefore, we have $W^*(s) \geq 2n^{1/3}\delta^2 \log(n) + o_p(1)$. This order implies

$$P\{EBIC(s) < EBIC(s_0)\} \leq P\{W^*(s) < W^*(s_0) - P\log n\} \to 0$$

because P is a constant as $n \to \infty$. That is, EBIC will not select any model s that does not contain s_0.

Furthermore, if s contains s_0 and $k > 0$ additional insignificant variables, by Theorem 2, we have

$$W^*(s_0) - W^*(s) \to \chi^2_k.$$

Thus, as $n \to \infty$

$$P\{EBIC(s) < EBIC(s_0)\} = P(W^*(s_0) - W^*(s) > k\log n) \to 0$$

Thus, the model s will not be selected by EBIC as $n \to \infty$.

Because P is finite, there are only finite number of s competing against s_0. Each of them has o(1) probability being selection. EBIC is hence selection consistent. This completes the proof. □

References

Akaike, H., 1973. Information theory as a extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Cameron, A.C., Trivedi, P.K., 1986. Economic models based on count data: Comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1, 29–53.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, United Kingdom.

Cameron, A.C., Trivedi, P.K., Milne, F., Piggot, J., 1988. A micro econometric model of the demand for health care and health insurance in Australia. Rev. Econom. Stud. 55, 85–106.

Chen, J., Chen, Z., 2008. Extended Bayesian information criterion for model selection with large sample space. Biometrika 95, 759–771.

Chen, J., Sitter, R.R., Wu, C., 2002. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika 89, 230–237.

Chen, J., Variyath, A.M., Abraham, B., 2008. Adjusted empirical likelihood and its properties. J. Comput. Graph. Statist. 17, 426–443.

Fan, J., Li, R., 2001. Variable selection via non concave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Godambe, V.P., 1960. An optimal property of regular maximum likelihood estimation. Ann. Math. Statist. 31, 568–571.

Kolaczyk, E.D., 1994. Empirical likelihood for generalized linear models. Statist. Sinica 4, 199–218.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22, 79–86.

Lebreton, J.D., Burnham, K.P., Clobert, J., Anderson, D.R., 1992. Modeling survival and testing biological hypothesis using marked animals: a unified approach with case studies. Ecological Monograph 62, 67–118.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 135, 370–384.

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. Ann. Statist. 18, 90–120.

Owen, A.B., 1991. Empirical likelihood for linear models. Ann. Statist. 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. Ann. Statist. 22, 300–325.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288.

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

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models and Gauss–Newton Method. Biometrika 61, 439–447.

Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist. 9, 60–62.