An Information Criterion for Empirical Likelihood with General Estimating Equations

Eric D. Kolaczyk
Department of Statistics
The University of Chicago
5734 University Avenue, Chicago, IL 60637

SUMMARY

We show that a completely analogous version of Akaike's information criterion (AIC) may be derived in the context of empirical likelihood for general estimating equations. The log-likelihood ratio statistic in AIC is replaced by the empirical log-likelihood ratio. The resulting empirical information criterion (EIC) statistic is shown to be an asymptotically unbiased estimate of risk, where the risk is proportional to the expected Kullback-Leibler distance between discrete, empirical distributions. The results of a series of simulations are presented in which the behavior of EIC, as an estimator, was studied for a small sample size. An example with real data is also presented. Finally, a result is mentioned which suggests the possibility of interesting relationships between experimental design and empirical likelihood distributions.

Some key words: Akaike's information criterion; empirical likelihood; estimating equations; experimental design; Kullback-Leibler distance; variance misspecification.

1 INTRODUCTION.

1.1 Akaike's information criterion.

Akaike's information criterion (AIC) statistic was originally derived as an unbiased estimate of the risk for approximating a larger model by a smaller submodel, where the parameters are estimated by the method of maximum likelihood and loss is defined with respect to Kullback-Leibler distance.

Specifically, assume $X_1, \ldots, X_n$ are independent and identically distributed random variables with respect to a density function $f(x|\theta)$, for some true parameter value $\theta =$

$\theta^* \in \mathbb{R}^L$. The Kullback-Leibler distance function

$$K(\theta^*, \theta) = \int \log \left\{ \frac{f(x|\theta^*)}{f(x|\theta)} \right\} f(x|\theta^*) dx \tag{1}$$

can be used as a measure of the loss entailed by modeling θ^* by θ. Define $\theta_k^* \in \mathbb{R}^k$ to be the projection of θ^* onto $\mathbb{R}^k \subset \mathbb{R}^L$, with respect to the metric defined by the Fisher information. Let $\hat{\theta}_k$ denote the maximum likelihood estimate of θ_k^*, and let $l(\hat{\theta}_k)$ be the corresponding log-likelihood ratio statistic. Akaike (1973) showed that under appropriate regularity conditions

$$AIC(k) = -2l(\hat{\theta}_k) + 2k - L \tag{2}$$

is asymptotically an unbiased estimator of

$$2nE_{\theta^*}\{K(\theta^*, \hat{\theta}_k)\} \ ,$$

the risk of modeling θ^* with $\hat{\theta}_k$.

AIC may be thought of as a likelihood-based variant of Mallows' C_p, in that both statistics evaluate models by estimating a risk measure which can be shown to incorporate both bias and variance terms. Although a number of different versions of AIC are in use, as well as extensions and modifications, we will be concerned in this paper only with AIC as defined in (2).

1.2 Empirical likelihood.

Owen (1988) introduced a nonparametric approach to likelihood methods in which an 'empirical' log-likelihood ratio statistic is maximized to produce a maximum empirical likelihood estimate $\tilde{\theta}$ of a parameter of interest, θ^*. The maximization is done with respect to certain constraints, most importantly those determined by a set of estimating equations for θ^*.

Empirical likelihood has been found to possess many of the properties of parametric likelihood in the i.i.d. setting. In particular, limiting chi-square distributions for likelihood ratio statistics (Owen (1988, 1990), Qin & Lawless (1994)), the ability to correct these statistics and their signed roots (DiCiccio, Hall, and Romano (1989), DiCiccio and Romano (1990)), and asymptotic normality of the maximum empirical likelihood estimates (Qin & Lawless (1994)). Other studies of asymptotic properties of empirical likelihood may be found in DiCiccio, Hall, and Romano (1991), DiCiccio and Romano (1990), and Hall (1990). Furthermore, an extension of the empirical likelihood methodology to linear models can be found in Owen (1991), while Kolaczyk (1994) and Owen (1992) consider extensions to generalized linear and projection pursuit models, respectively.

1.3 An information criterion for empirical likelihood ?

Given the many characteristics shared by empirical and parametric likelihood methods, it is natural to consider an empirical analogue of the AIC statistic, in which the parametric log-likelihood ratio statistic is replaced by the empirical version. In this paper, we show that the resulting *empirical information criterion* (EIC) statistic is asymptotically an unbiased estimate of the expected Kullback-Leibler distance between two discrete, empirical distributions.

The EIC statistic serves as a twofold contribution. First, it demonstrates that Akaike's original idea may be extended far beyond its initial parametric context. Second, it provides a well-defined tool, balancing bias and variance, whereby models defined with respect to quite general sets of estimating equations may be compared. The rest of this paper is organized as follows. In Section 2 the appropriate context is laid out and the EIC statistic derived. A corollary of the main result is also mentioned. Section 3 contains the results of a simulation designed to study the behavior of EIC in a simple setting with small samples. In passing, an interesting note is made with respect to empirical likelihood and experimental design. In Section 4, an example with real data is presented. Finally, a brief discussion appears in Section 5. Proofs connected with results in Sections 2 and 3 may be found in the appendix.

2 THE EMPIRICAL INFORMATION CRITERION.

2.1 Background and notation.

The context within which we will work is as follows. Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed, d-variate random variables with common distribution F. We are interested in a parameter $\theta \in \mathbb{R}^L$ associated with F. Information relating θ and F is summarized by a set of estimating functions $g(x,\theta) = \{g_1(x,\theta), \ldots, g_L(x,\theta)\}$ which are unbiased i.e.,

$$E_{\theta^*}\{g(X,\theta)\} = 0$$

when $\theta = \theta^*$.

We assume that θ has a nested structure and are interested in the risk inherent in replacing the full model θ by a submodel in $\mathbb{R}^k$, for $k < L$. To this end, define the metric

$$||\theta||_V^2 \equiv \theta' V^{-1} \theta$$

with respect to the matrix

$$V^{-1} \equiv E_{\theta^*}(\frac{\partial g'}{\partial \theta})\{E_{\theta^*}(gg')\}^{-1} E_{\theta^*}(\frac{\partial g}{\partial \theta}) \ ,$$

and let the submodel θ_k be defined as

$$\theta_k = \arg \min_{\theta' \in \Omega_k} ||\theta' - \theta||_V^2 \ .$$

In words, θ_k is the projection of θ onto $\mathbb{R}^k$ with respect to the metric $||\cdot||_V^2$.

Estimates of the true parameter θ^*, or a more parsimonious submodel θ_k^*, may be obtained through the use of empirical likelihood methods. For fixed θ, the empirical log-likelihood ratio is defined as

$$l_{EL}(\theta) = \sup \left\{ \sum_{i=1}^{n} \log(np_i) \right\} \ , \tag{3}$$

where $p_i = dF(x_i) = pr(X = x_i)$ are probability atoms (or weights) on the sample, $\{x_1, \ldots, x_n\}$, subject to the constraints

$$0 \le p_i \le 1 \qquad \sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i g(x_i, \theta) = 0 \ .$$

The optimal weights, $\{p_i(\theta)\}_{i=1}^n$, may be found using the method of Lagrange multipliers. Specifically, we differentiate the function

$$H = \sum_{i=1}^{n} \log(np_i) + \lambda(1 - \sum_{i=1}^{n} p_i) - n\tau' \sum_{i=1}^{n} p_i g(x_i, \theta) \tag{4}$$

with respect to the p_i's to obtain

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \lambda - n\tau' g(x_i, \theta) = 0 \ ,$$

where $\lambda \in \mathbb{R}$ and $\tau \in \mathbb{R}^L$ are Lagrange multipliers. The identity $\sum p_i \frac{\partial H}{\partial p_i} = 0$ implies that $\lambda = n$, and hence we may solve for $p_i(\theta)$ as

$$p_i(\theta) = \frac{1}{n} \frac{1}{1 + \tau' g(x_i, \theta)} \ . \tag{5}$$

Substitution into Equation (3) then yields

$$l_{EL}(\theta) = -\sum_{i=1}^{n} \log[n\{\, 1 + \tau' g(x_i, \theta) \,\}] \ .$$

The maximization in $\{p_i(\theta)\}$ in (3) is actually converted to a minimization in τ, and can be implemented using any of a number of numerical optimization algorithms. See Owen (1990) and Qin and Lawless (1994) for details.

Maximizing $l_{EL}(\theta)$ with respect to θ yields the maximum empirical likelihood estimate $\tilde{\theta}$ of θ^*. If the dimensions of θ and g are equal, as we have assumed here, $p_i(\tilde{\theta}) = 1/n$ for all i, and $l_{EL}(\tilde{\theta}) = 0$. In order to obtain the constrained maximum empirical likelihood estimate, $\tilde{\theta}_k$, of θ_k^*, the same maximization as above is performed, but now subject to the additional constraints that those components of θ *not* in $\mathbb{R}^k$ be held fixed at zero.

By construction, the set $\{p_i(\theta)\}_{i=1}^n$ is a proper probability distribution on the sample $\{x_1, \ldots, x_n\}$. Hence, a natural analogue of the loss function in (1) within the present context of empirical likelihood is given by

$$K_{EL}(\theta^*, \theta) = \sum_{i=1}^{n} \log \left\{ \frac{p_i(\theta^*)}{p_i(\theta)} \right\} p_i(\theta^*) \ .$$

In words, the loss entailed in modeling θ^* by θ is defined to be the Kullback-Leibler distance between the two *discrete* empirical distributions corresponding to these two parameter values.

Note that the loss $K_{EL}(\theta^*, \tilde{\theta}_k)$, like $K(\theta^*, \hat{\theta}_k)$, is a random variable. Therefore, similar to the AIC methodology, we will opt instead to estimate the *risk* associated with modeling θ^* by the fitted submodel $\tilde{\theta}_k$, that is, $E_{\theta*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\}$. However, in contrast to $K(\theta^*, \cdot)$, which is a constant for fixed θ, the randomness of $K_{EL}(\theta^*, \tilde{\theta}_k)$ results not only from that of $\tilde{\theta}_k$ but also from that of the p_i's themselves. Hence $K_{EL}(\theta^*, \cdot)$ is a random variable even for a fixed value of θ.

2.2 Main result.

In this subsection we derive the empirical information criterion (EIC) statistic as an asymptotically unbiased estimate of the risk $2nE_{\theta*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\}$. The main result may be stated as follows.

Theorem 1 . *Assume the regularity conditions of Lemmas 1 and 2 below. Then the statistic*

$$EIC(k) = -2l_{EL}(\tilde{\theta}_k) + 2k - L$$

is asymptotically an unbiased estimate of

$$2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\} \ ,$$

the risk of modeling θ^* by the fitted submodel $\tilde{\theta}_k$. We call the EIC statistic the Empirical Information Criterion.

The derivation of this result, perhaps not surprisingly, is quite similar in form to that of AIC. As a result, the framework common to both is sketched here, while those points requiring work unique to EIC have been condensed into Lemmas 1 and 2. Proofs for these lemmas may be found in the appendix.

Proof of Theorem 1.

Begin by noting that, for a fixed sample, a Taylor series expansion of $K_{EL}(\theta^*, \theta)$ in θ about θ^* yields the expression

$$K_{EL}(\theta^*, \theta) \approx \frac{1}{2}(\theta - \theta^*)' \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} K_{EL}(\theta^*, \theta) \right\}_{|\theta^*} (\theta - \theta^*) \ . \tag{6}$$

The following result will be useful in expressing K_{EL} in a simpler form.

Lemma 1 . *Assume that $E_{\theta^*}\{g(X, \theta^*)g'(X, \theta^*)\}$ is positive definite, $\partial g(x, \theta)/\partial\theta$ is continuous in a neighborhood about θ^*, $||\partial g(x, \theta)/\partial\theta||$ and $||g(x, \theta)||$ are both bounded by some integrable function $G(x)$ in this neighborhood, and $E_{\theta^*}\{\partial g(X, \theta^*)/\partial\theta\}$ is of full rank L. Then for θ in the ball $||\theta - \theta^*|| \leq n^{-1/3}$,*

$$K_{EL}(\theta^*, \theta) = \frac{1}{2}(\theta - \theta^*)' V^{-1} (\theta - \theta^*) \, + \, o_p(1) \ .$$

Compare the result of Lemma 1 with the analogous result for the regular Kullback-Leibler distance in (1) ,

$$K(\theta^*, \theta) = \frac{1}{2}(\theta - \theta^*)' J (\theta - \theta^*) + o(1) \qquad \text{a.s.}$$

The change from $o(1)$ to $o_p(1)$ in going from K to K_{EL} reflects the reliance of the latter on the sample.

Using Lemma 1 and the definition of $|| \cdot ||_V^2$, and taking $\theta = \tilde{\theta}_k$, we may write

$$\begin{aligned} 2nK_{EL}(\theta^*, \tilde{\theta}_k) &\approx n(\tilde{\theta}_k - \theta^*)' V^{-1} (\tilde{\theta}_k - \theta^*) \\ &\equiv n||\tilde{\theta}_k - \theta^*||_V^2 \\ &= n||\tilde{\theta}_k - \theta_k^*||_V^2 + n||\theta_k^* - \theta^*||_V^2 \ . \end{aligned} \tag{7}$$

Hence $2nK_{EL}(\theta^*, \tilde{\theta}_k)$ may be decomposed into the sum of two terms measuring variance and bias, respectively. As the variance term $n||\tilde{\theta}_k - \theta_k^*||_V^2$ has an approximate $\chi^2_{(k)}$ distribution (see Lemma 2 below), taking expectations of both sides in (7) yields

$$2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\} \approx n||\theta_k^* - \theta^*||_V^2 + k \ .$$

The problem is therefore reduced to one of estimating $n||\theta_k^* - \theta^*||_V^2$.

To this end, expand $-2l_{EL}(\theta)$ about $\tilde{\theta}$ and $\tilde{\theta}_k$, yielding

$$-2l_{EL}(\theta) = n(\theta - \tilde{\theta})' \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} l_{EL}(\theta) \right\}_{|_{\tilde{\theta}}} (\theta - \tilde{\theta}) \tag{8}$$

and

$$-2l_{EL}(\theta) = -2l_{EL}(\tilde{\theta}_k) + n(\theta - \tilde{\theta}_k)' \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} l_{EL}(\theta) \right\}_{|_{\tilde{\theta}_k}} (\theta - \tilde{\theta}_k) \ . \tag{9}$$

Substituting $\theta = \theta_k^*$, setting (8) equal to (9), and replacing the matrices of second partial derivatives by V^{-1} (justification similar to that of Lemma 1 above) it follows that

$$\begin{aligned} -2l_{EL}(\tilde{\theta}_k) &\approx n||\theta_k^* - \tilde{\theta}||_V^2 - n||\theta_k^* - \tilde{\theta}_k||_V^2 \\ &= n||\theta_k^* - \theta^*||_V^2 + \left[n||\theta^* - \tilde{\theta}||_V^2 - n||\theta_k^* - \tilde{\theta}_k||_V^2 \right] \\ &\quad -2n\langle \theta^* - \tilde{\theta}, \theta_k^* - \theta^* \rangle_V \ . \end{aligned} \tag{10}$$

Here $\langle \cdot, \cdot \rangle_V$ denotes the inner product with respect to the matrix V^{-1}. The relationship among the three components of (10) may be summarized as follows.

Lemma 2 . *In addition to the conditions in Lemma 1, assume that $\partial^2 g(x,\theta)/\partial\theta\partial\theta'$ is continuous in a neighborhood of θ^* and $||\partial^2 g(x,\theta)/\partial\theta\partial\theta'||$ is bounded by some integrable function $G(x)$ in that neighborhood. Then the quantities $n||\theta^* - \tilde{\theta}||_V^2 - n||\theta_k^* - \tilde{\theta}_k||_V^2$ and $n||\tilde{\theta}_k - \theta_k^*||_V^2$ are asymptotically independent and distributed as $\chi^2_{(L-k)}$ and $\chi^2_{(k)}$ random variables, respectively. Furthermore, the contribution of $2n\langle \theta^* - \tilde{\theta}, \theta_k^* - \theta^* \rangle_V$ is negligible with respect to that of the other components.*

From Lemma 2 it follows that

$$-2E_{\theta^*}\{l_{EL}(\tilde{\theta}_k)\} \approx n||\theta_k^* - \theta^*|| + L - k \ \ .$$

Adjusting by the appropriate constants, we arrive at our main result. $\square$

2.3 Comments on Theorem 1.

Theorem 1 holds for a very general set of estimating functions, essentially those with mild moment and smoothness conditions on g, $\partial g/\partial\theta$, and $\partial^2 g/\partial\theta\partial\theta'$. In the special case where the estimating functions g are the maximum likelihood estimating functions, the following corollary holds.

Corollary 1 . *Suppose $x_1, \ldots, x_n$ are independent and identically distributed observations from a density $f(x|\theta)$, with $\theta = \theta^*$, and the estimating functions are defined as*

$$g(x,\theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) \ \ .$$

Then $EIC(k)$ is asymptotically an unbiased estimate of

$$2nE_{\theta^*}\{K(\theta^*, \tilde{\theta}_k)\} \ \ .$$

Proof of Corollary 1.

By definition of g, the second Bartlett identity holds, which in the present notation may be written as

$$-E_{\theta^*}\{\frac{\partial}{\partial\theta}g(X,\theta^*)\} = E_{\theta^*}\{g(X,\theta^*)g'(X,\theta^*)\}.$$

This implies that $V^{-1} = J$, where J is the Fisher information matrix. Noting that

$$2nK(\theta^*,\theta) \approx n||\theta - \theta^*||_J^2 \ ,$$

the result follows. □

The reader should be aware that although $\tilde{\theta} \equiv \hat{\theta}$ in the scenario of Corollary 1, generally it is *not* expected that $\tilde{\theta}_k = \hat{\theta}_k$. Hence, the quantities being estimated by $EIC(k)$ and $AIC(k)$,

$$2nE_{\theta^*}\{K(\theta^*,\tilde{\theta}_k)\} \qquad \text{and} \qquad 2nE_{\theta^*}\{K(\theta^*,\hat{\theta}_k)\} \ ,$$

respectively, are typically not the same.

To better understand the cause of this difference, suppose that we have $X_i = (Z_i, Y_i)$ random variable pairs, $Z_i' \in \mathbb{R}^L$, $Y_i \in \mathbb{R}$, such that $Y_i|Z_i \sim \text{Normal}(Z_i\theta^*, 1)$. The maximum likelihood approach to estimating θ^* results in an estimate $\hat{\theta}$ based on the Normal equations

$$Z(Z'\theta - Y) = 0 \ , \tag{11}$$

where the ith column of the $L \times n$ matrix Z is Z_i, and $Y = (Y_1, \ldots, Y_n)'$. In the parametric likelihood approach, the submodel θ_k^* is estimated by simply *dropping* the last $L - k$ regressors in the matrix Z before solving (11). On the other hand, empirical likelihood keeps all regressors in the model, whether estimating θ^* or θ_k^*. In the latter case, the optimization of $-2l_{EL}(\theta)$ is done over θ with the last $L - k$ components set to zero. This approach imposes the conditions that the errors are uncorrelated with the

Z_i's and have mean zero. Simply dropping regressors will not achieve these conditions. See Owen (1992, Section 8) for a more detailed illustration.

On a final note, we emphasize the fact that the framework for EIC contains no parametric assumptions regarding the distribution of the data. Only unbiasedness of the estimating functions is required. In the case where a family of distributions for the data may be assumed safely, a parametric approach and AIC are probably preferable to empirical likelihood and EIC. However, in the case where the family has been correctly assumed, but, for example, the variance is misspecified, the results regarding empirical likelihood and EIC as an asymptotically unbiased estimator still hold; on the other hand, it is uncertain what AIC is estimating. These issues are explored in more detail in the next section.

3 SIMULATIONS.

In practice, information criterion statistics typically are used for choosing an "optimal" model amongst many similar choices. However, as has been pointed out by numerous authors (e.g. Mallows (1973)), the (asymptotic) unbiasedness of the statistic as an estimate of risk, and its success at choosing an "optimal" model are not necessarily one and the same issue. Accordingly, in an effort to complement the theoretical results of the previous section, we have restricted our attention here simply to studying the behavior of EIC as an *estimator* in small samples, and have chosen to leave issues of *model selection* for later work. For this purpose, a series of simulations were conducted.

The model and design used in the simulations were kept simple, and are similar to those of Atkinson (1980). Specifically, for a sample of size $n = 45$, an equal number of data are observed at each of the three design points $z = -1, 0$, and 1, according to the model

$$y = 1.0 + 0.5z + 0.1z^2 + \epsilon \ , \tag{12}$$

where $\epsilon \sim$ Normal$(\,0, \sigma^2(z)\,)$. The coefficients were chosen so that the true model is quadratic but a linear submodel may be entertained realistically. Two choices of variance functions were used, with the intention of studying the effects of variance misspecification on EIC and AIC: $\sigma^2(z) \equiv \sigma^2$ and $\sigma^2(z) = e^{2c|z|}\sigma^2$.

3.1 Correctly specified variance.

Suppose that $\sigma^2(z) \equiv \sigma^2$ is constant, but unknown. In our simulations, σ was taken to be 1. The full model in (12) has $L = 3$ parameters, and we consider the linear submodel with $k = 2$ parameters. The natural choice for estimating θ^* and θ_2^* in both the parametric and empirical likelihood contexts is via the Normal equations, as discussed in Section 2.3. The associated Kullback-Leibler distance is

$$K(\theta^*, \theta) = \frac{1}{3} \sum_{z \in \{-1,0,1\}} \frac{\{(1, z, z^2)(\theta^* - \theta)\}^2}{2\sigma^2} \ ,$$

and the AIC statistic is

$$AIC(k) = \frac{RSS_k - RSS_L}{\sigma^2} + 2k - L \ , \tag{13}$$

where RSS_j is the residual sum of squares from the regression with j non-zero parameters. Since σ^2 is unknown, it needs to be replace by an estimate in (13). We chose $\hat{\sigma}^2 = \frac{1}{n-1}\sum(y_i - \hat{y}_i)^2$, yielding

$$\widehat{AIC}(k) = (n - L)\frac{RSS_k}{RSS_L} + 2k - n \ ,$$

which is actually just Mallows' C_p statistic.

A total of $10,000$ trials were conducted. In each trial, a sample of size $n = 45$ was generated as specified above. The parameter estimates $\hat{\theta}_2$ and $\tilde{\theta}_2$ were calculated from the data, as were the corresponding AIC and EIC statistics. The loss functions $K(\theta^*, \hat{\theta}_2)$, $K(\theta^*, \tilde{\theta}_2)$, and $K_{EL}(\theta^*, \tilde{\theta}_2)$ were calculated as well. Because our estimating equations are the maximum likelihood estimating equations, note that EIC is an asymptotically unbiased estimator of both $2nE_{\theta^*}\{K(\theta^*, \tilde{\theta}_k)\}$ and $2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\}$.

Figure 1 shows smooth estimates of the densities of $\log(AIC)$ and $\log(EIC)$. Although the two statistics are not estimating exactly the same quantity, it is interesting to note that their densities are almost identical. That of $\log(EIC)$ has a slightly longer tail. The biases of AIC and EIC in estimating the risks $2nE_{\theta^*}\{K(\theta^*, \hat{\theta}_2)\}$ and $2nE_{\theta^*}\{K(\theta^*, \tilde{\theta}_2)\}$, respectively, were roughly equivalent at 0.059 and 0.052. The bias of EIC in estimating the risk $2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\}$ was negative and of an order of magnitude larger than either of these, -0.774. The standard deviation of AIC, 1.725, was slightly smaller than that of EIC, 1.904. Combining the above biases and variances yields mean squared errors of 2.977, for AIC, and 3.626 and 4.223, for EIC with respect to K and K_{EL}.

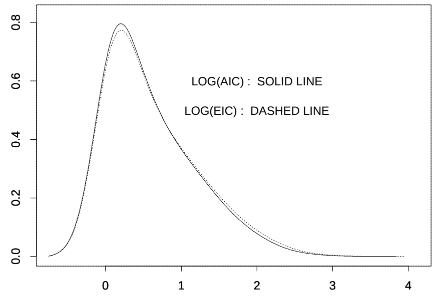

Figure 1: Smooth estimates of the densities of $\log(AIC)$ (solid line) and $\log(EIC)$ (dashed line), based on $10,000$ trials, for the submodel θ_2^*.

The relative performances of AIC and EIC, as indicated by these mean squared errors, are perhaps not surprising. Parametric maximum likelihood methods, and hence AIC, are most appropriate in the present context. Of greater interest is the difference in the biases with which EIC estimates its two corresponding risks. Figure 2 gives some insight into this difference. A scatterplot of $\log\{2nK_{EL}(\theta^*, \tilde{\theta}_2)\}$ versus $\log\{2nK(\theta^*, \tilde{\theta}_2)\}$ is shown for the 10,000 trials. The position of the point cloud with respect to the line $y = x$ suggests that for small sample sizes K_{EL} is on average slightly larger than K, although perhaps more so at larger values than at small. There is also evidence of increased variability in the former over the latter. This difference in variability may be seen more clearly by consulting the smooth estimates of the marginal densities included at the top and righthand sides of the scatterplot. The difference was also anticipated in Section 2, where it was mentioned that the errors in approximating $2nK(\theta^*, \theta)$ and $2nK_{EL}(\theta^*, \theta)$ by quadratic forms were $o(1)$ and $o_p(1)$, respectively.

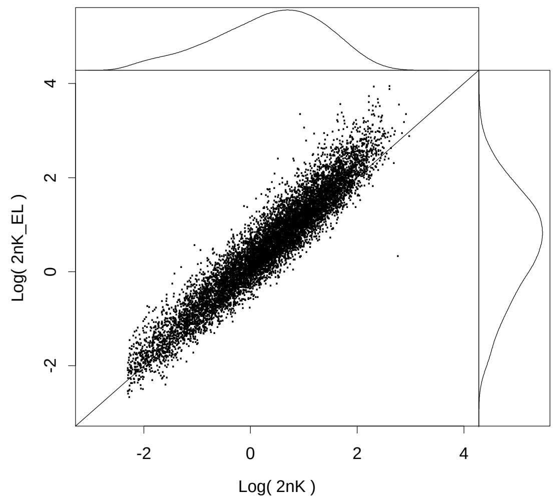

Figure 2: Scatterplot of $\log\{2nK_{EL}(\theta^*, \tilde{\theta}_2)\}$ versus $\log\{2nK(\theta^*, \tilde{\theta}_2)\}$, for 10,000 trials. The line $y = x$ is included for reference. Smooth estimates of the marginal densities appear on the top and righthand sides.

3.2 Misspecified variance.

As mentioned in Section 2.3, it is of some interest to assess the effects of variance misspecification on the behavior of EIC and AIC. In particular, although it would be reasonable to expect that EIC may suffer a loss of efficiency in the presence of variance misspecification, it should still retain its unbiasedness as an estimator. On the other hand, having chosen a particular family of distributions, with an assumed variance function, it is not clear that AIC should remain an asymptotically, unbiased estimator of the associated risk.

We explored these matters by repeating the above simulation, for $n = 45$, with $\sigma^2(z) = e^{2c|z|}\sigma^2$. This choice of variance function results in a variance of σ^2 at $z = 0$, but increased (or decreased) variance $a^2\sigma^2$ at $z = -1$ and 1, where $a \equiv e^c$. Three values of c were chosen so as to yield variance misspecification factors of $a = 1.25, 1.50$, and 2.00. A total of $10,000$ trials were conducted for each value of a, as in the case of correctly specified variance (which we denote by $a = 1.00$).

Figure 3(a) shows the mean squared errors as a function of a. The mean squared error of EIC as an estimator of $2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_2)\}$ remains fairly constant, while its mean squared error as an estimator of $2nE_{\theta^*}\{K(\theta^*, \tilde{\theta}_2)\}$ increases noticeably by $a = 1.50$ and dramatically by $a = 2.00$. The mean squared error of AIC increases in a similar, though perhaps even slightly more drastic, fashion. Figures 3(b) and 3(c) decompose the mean squared error into bias and variance components. Clearly the bias in the estimators is the driving factor in the inflation of the mean squared errors. The variance, surprisingly, appears to decrease slightly.

These results suggest that the behavior of EIC, as an approximately unbiased estimator of $2nE_{\theta^*}\{K_{EL}(\theta^*, \tilde{\theta}_k)\}$ in finite samples, is effected little by mild variance misspecification. Furthermore, the variation of the estimator does not suffer. On the other

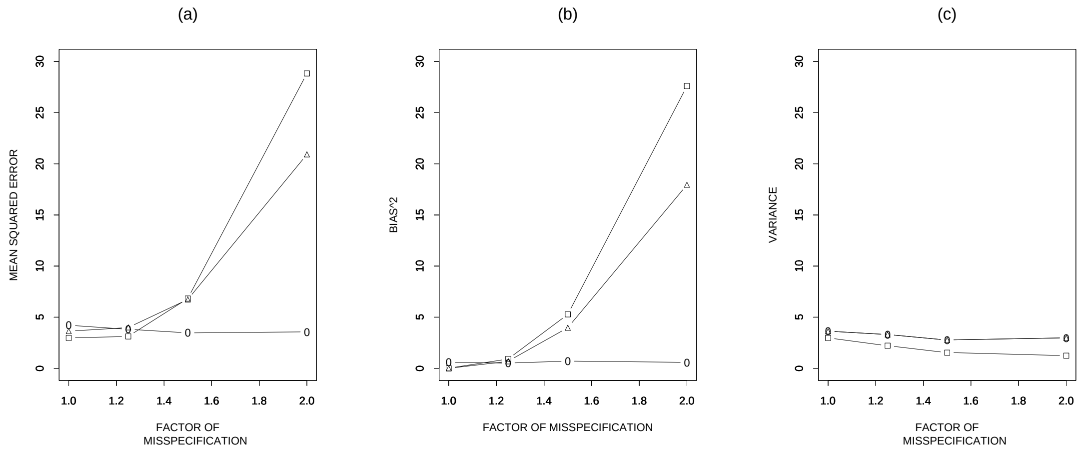

Figure 3: Figure (a) shows the mean squared error for AIC (□), and EIC with respect to K (△) and K_{EL} (O). Figures (b) and (c) decompose the mean squared error into bias (squared) and variance components, respectively.

hand, AIC suffers a fairly severe increase in bias, although its variance also remains roughly constant. Additional studies, with more complex models, varying sample sizes, and a finer sampling of the range of misspecification would be helpful in confirming these conclusions. On a side note, we mention that a similar study of the effects of variance misspecification on parameter estimation and coverage of confidence regions for empirical and parametric likelihood methods also would be of interest. See Kolaczyk (1994, pg. 205) for comments in this direction.

3.3 Comments on computing.

The simulation of data and calculation of all statistics were done using the S-Plus software package. The empirical weights, $\{p_i(\theta^*)\}$ and $\{p_i(\tilde{\theta}_2)\}$, were computed using Owen's Nel2.0 software. In general, the existence of a unique solution to the empirical likelihood optimization is only guaranteed when 0 is inside the convex hull of the data i.e., $ch\{g(x_1, \theta), \ldots, g(x_n, \theta)\}$. For datasets in which 0 is near or outside the convex hull, convergence to a valid solution is difficult or impossible, respectively. See Owen (1990, pg. 106) for more details. In our simulations, there were 2, 4, 1, and 4 trials in which convergence to a proper solution (i.e. probability distribution) was not reached, for $a = 1.00, 1.25, 1.50$, and 2.00, respectively. Eight of these cases involved the true parameter, θ^*, while the other three involved the fitted parameter, $\tilde{\theta}_2$. The reader should note that the calculation of $\{p_i(\theta^*)\}$ was done for the purpose of simulation only, and has no bearing on the usage of EIC in practice, where lack of knowledge of θ^* clearly precludes such calculations. For all trials in which convergence failed, the values were dropped from the study and the summaries were computed based on the remaining trials.

3.4 An aside regarding experimental design.

Although the design matrix used in the above simulations is quite simple, it has an interesting numerical property with regard to empirical likelihood distributions.

Result : *Observe data* $y_i = \theta^{(0)} + \theta^{(1)} z_i + \theta^{(2)} z_i^2 + u_i$, $i = 1, \ldots, n$, *where the* z_i*'s take on values only in* $\{-1, 0, 1\}$. *Let the estimating functions* g *be defined by the Normal equations, and the* $\{u_i\}$ *be any set of numbers such that* $\sum g\{(z_i, y_i), \theta\} g'\{(z_i, y_i), \theta\}$ *is positive definite for* θ*'s of interest. Then for fixed* θ *and* $z = -1, 0,$ *or* 1,

$$\sum_{i: z_i = z} p_i(\theta) = \frac{\#\{i : z_i = z\}}{n} .$$

In other words, the total proportion of weight in $\{p_i(\theta)\}$ *given to a particular design point will be exactly that proportion of the sample represented by the point.*

Although such a result is true trivially in the case where $\theta = \tilde{\theta}$, the assertion here is that it holds true independent of θ and the assignment of the z_i's. The proof, which may be found in the appendix, relies on the fact that the constraint $\sum p_i g(z_i, \theta) = 0$ may be rewritten in terms of an equivalent set of three constraints, one for each of the three design points.

The above result suggests that there may be some interesting relationships between empirical likelihood and experimental design, the exploration of which the author will defer to a latter work. As an example, suppose a sample of size n (for n even) will be taken in which the z_i's are severely unbalanced in a $\frac{n-2}{2}/\frac{n-2}{2}/2$ split over the design points $\{-1, 0, 1\}$. Then the amount of weight given to (or the "say" allowed for) either of the two datum at $z = 1$ *a priori* can be no more than $\frac{2}{n}$, regardless of their influence in the regression.

4 AN EXAMPLE.

We consider the urinary crystal data of Andrews and Herzberg (1985), which allows for the study of EIC in a context where both the data are real and the possible models are of a more complex nature than in Section 3. A total of 79 urine samples were collected and examined for the presence or absence of calcium oxalate crystals, denoted by $Y = 1$ or 0, respectively. In addition, measurements were taken on six covariates: specific gravity, $z^{(1)}$; pH level, $z^{(2)}$; osmolarity, $z^{(3)}$; conductivity, $z^{(4)}$; urea concentration, $z^{(5)}$; and calcium concentration, $z^{(6)}$. The data were 'cleaned' and transformed (by logarithms) as in Cox and Snell (1989).

A logistic regression model of the form

$$\log \frac{\pi_i}{1 - \pi_i} = \theta^{(0)} + \theta^{(1)} z^{(1)} + \ldots + \theta^{(6)} z^{(6)} \quad ,$$

where $\pi_i = pr(Y_i = 1)$, is an obvious candidate for modeling the risk of urinary crystals as a function of $z^{(1)}, \ldots, z^{(6)}$. However, the covariate pairs $(z^{(1)}, z^{(3)})$, $(z^{(3)}, z^{(4)})$, and $(z^{(3)}, z^{(5)})$ all have fairly high correlations (0.8 or above), which suggests that a model with one of these covariates dropped might be an improvement.

Figure 4 shows the values of the EIC and AIC statistics when each one of the covariates $z^{(1)}, \ldots, z^{(5)}$ were dropped from the model. Calcium concentration, $z^{(6)}$, was kept in all submodels. EIC and AIC appear to be in agreement as to the relative risks of dropping $z^{(2)}$, $z^{(3)}$, and $z^{(4)}$ from the full linear logistic model. Also, the values of the two statistics for the full model, by definition, are both equal to $L = 7$. In the cases where either $z^{(1)}$ or $z^{(5)}$ are dropped, EIC and AIC are almost the reverse of each other in predicting the risks. Overall, both methods estimate a risk value of 7.0 or above for all models except the one in which $z^{(3)}$ is dropped, where the risk is estimated to be about 5.0. Considering the high correlation of this covariate with three of the other 5, this result is perhaps not surprising.

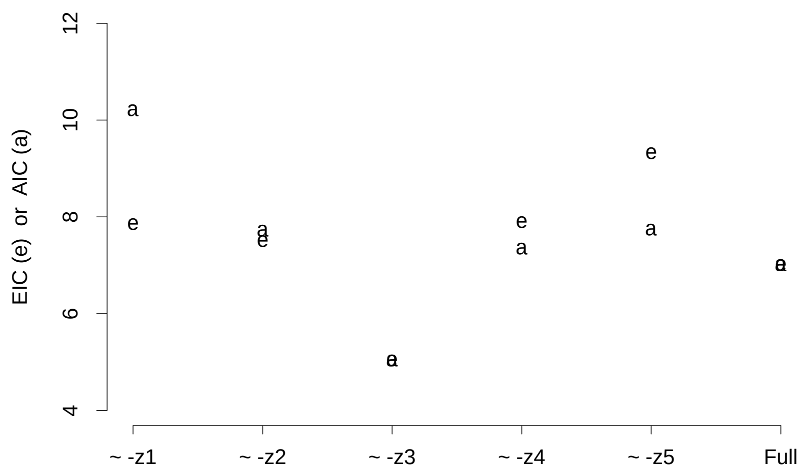

Figure 4: EIC (e) and AIC (a) values estimating risk for submodels of the full linear model, in which a single covariate is dropped.

In analyzing this data, Cox and Snell (1989) note that none of the cross-product terms seem to improve the model by their inclusion, although addition of a quadratic term in $z^{(5)}$ offers some improvement. Figure 5 shows EIC and AIC values for the same linear submodels considered in Figure 4, except that the full model is taken to be quadratic in $z^{(5)}$. Note that the overall relation of the AIC values to each other is exactly the same as in Figure 4, the difference just being an increase equal to roughly the change in log-likelihood in going from the full linear to the full quadratic model. We would not expect a similarly simple change in EIC values, however, since the linear submodels are fit using different full-model matrices with empirical likelihood (see Section 2.3).

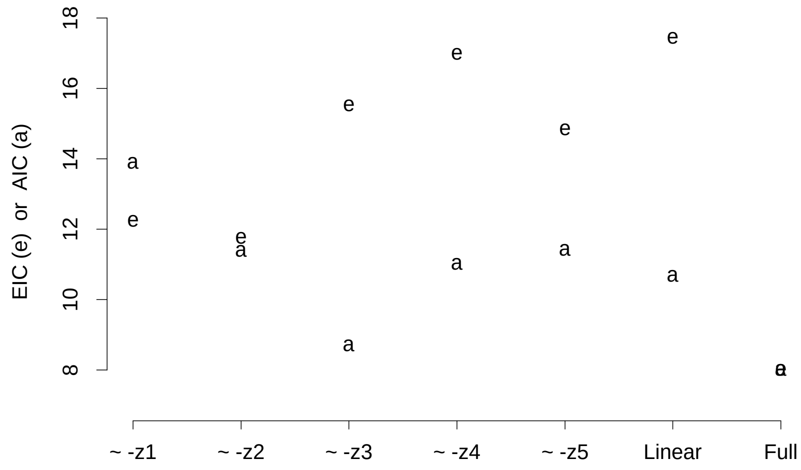

Figure 5: EIC (e) and AIC (a) values estimating risk for the same linear submodels as in Figure 4, but with the full model taken to be quadratic in $z^{(5)}$.

In fact, the story told by EIC has changed noticeably. In particular, based on the EIC values, it appears that the models may be separated into three groups. The full model has a much smaller risk than any of the submodels. The linear submodels with $z^{(1)}$ or $z^{(2)}$ dropped have similar risks (as in Figure 4), estimated to be about 50% higher than that of the full model. Finally, the linear submodels in which $z^{(3)}$, $z^{(4)}$, or $z^{(5)}$ are dropped, as well as the full linear model, all have estimated risks roughly twice that of the full model. Of these last four models, only the risk for that in which $z^{(5)}$ is dropped has remained similar in its relation to AIC as compared to Figure 4. The risks of the other three models have increased greatly in comparison.

This example suggests that the relationship between urinary crystals and the covariates in this dataset may be more intricate than indicated by a parametric analysis, but a more detailed study of the effects on empirical likelihood of the addition of covariates to the model matrix would be necessary before declaring this relationship 'real'.

5 DISCUSSION.

Empirical likelihood allows one to perform inference on parameters within quite general, nonparametric frameworks. The introduction of the EIC statistic provides users of empirical likelihood with a tool whereby the risks of similar, competing models may be estimated and compared. However, the reliability of using EIC to chose an "optimal" model is not discussed here. It is probably reasonable to imagine that the properties associated with such usage would be quite similar to those of AIC.

Our simulation results show that even in contexts where the statistician is fairly comfortable with the assumptions accompanying maximum likelihood approaches, EIC and empirical likelihood may still be quite valuable in the presence of variance misspecification. EIC was seen to retain its properties of approximate unbiasedness as a risk estimator in small samples, whether the variance was correctly specified or not. AIC appears to suffer noticeably as an estimator of risk in the presence of even mild misspecification. Similar studies exploring the effects of variance misspecification on inference would be of interest.

On the topic of simulations, we note that although the model used in Section 3 was fairly simple, it allowed for detailed study of the properties of both EIC and the loss function $K_{EL}(\theta^*, \theta)$. The model was also useful in illuminating the possibility of interesting relationships between empirical likelihood and experimental design. The author has found in conducting simulations involving EIC that some care must be given to the choice of sample size, parameter values, and noise level, in order to obtain convergence, for all submodels, in all but an acceptably small number of trials (see Kolaczyk (1992) for comments regarding an earlier simulation study). As might be expected, highly constrained submodels, severely unbalanced designs, and extreme samples can individually or together lead to a solution near the edge of or outside of the convex hull of the data, and hence problems with convergence. This suggests that current efforts at improving empirical likelihood optimization algorithms are indeed well spent, and perhaps even necessary for conducting larger, more extensive simulations.

Acknowledgements

This research was conducted in part using computing facilities supported by The University of Chicago Block Fund. The author would like to thank Colin Mallows for helpful discussions during the initial version of this work, as well as Jing Qin for encouraging the present extension.

Appendix

Proof of Lemma 1.

From the definition of $p_i(\theta)$ in (5)

$$\frac{\partial^2}{\partial\theta\partial\theta'} K_{EL}(\theta^*, \theta) = \frac{\partial^2}{\partial\theta\partial\theta'} \sum_{i=1}^{n} \log\{1 + \tau'(\theta)g(x_i, \theta)\} p_i(\theta^*) \ . \tag{14}$$

Restrict attention for the moment to the summation in the right hand side of (14), and consider $\theta \in \mathcal{B}_n$, where $\mathcal{B}_n \equiv \{\theta| \ ||\theta - \theta^*|| \leq n^{-1/3}\}$. When $E_{\theta^*}\{||g(X, \theta^*)||^3\} < \infty$, a Taylor series expansion allows for the expression

$$\begin{aligned} \sum_{i=1}^{n} \log\{1 + \tau'(\theta)g(x_i, \theta)\} p_i(\theta^*) &= \sum_{i=1}^{n} \left[\tau'(\theta)g(x_i, \theta) - \frac{\{\tau'(\theta)g(x_i, \theta)\}^2}{2} \right] p_i(\theta^*) \\ &\quad + o(n^{-2/3}) \qquad \text{a.s.} \end{aligned} \tag{15}$$

Similarly, a second Taylor series expansion yields

$$\begin{aligned} p_i(\theta) &= \frac{1}{n} \frac{1}{1 + \tau'(\theta)g(x_i, \theta)} \\ &= \frac{1}{n}\{1 - o_p(1)\} \\ &= \frac{1}{n} + o_p(n^{-1}) \ , \end{aligned} \tag{16}$$

for $\theta \in \mathcal{B}_n$.

Substituting (16) for $p_i(\theta^*)$ and borrowing the result that

$$t(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\right\} + o(n^{-1/3}) \quad \text{a.s.}$$

from Qin and Lawless (1994), we express (15) as

$$\begin{aligned} & \sum_{i=1}^{n} \log\{1+\tau'(\theta)g(x_i,\theta)\}p_i(\theta^*) \\ = \; & \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\right\}'\left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\right\} + o_p(n^{-2/3}) \quad .(17) \end{aligned}$$

The right hand side of (17) may be rewritten as

$$\begin{aligned} RHS \;=\; & \frac{1}{2}\left[E\{g(X,\theta*)\} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}g(x_i,\theta^*)un^{-1/3}\right]' \\ & \times [Eg(X,\theta^*)g'(X,\theta^*)]^{-1} \\ & \times \left[E\{g(X,\theta*)\} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}g(x_i,\theta^*)un^{-1/3}\right] + o_p(n^{-2/3}) \qquad (18) \end{aligned}$$

by expanding the summands in (17) about θ^* and appealing to the SLLN. Here u is a vector in the unit ball.

Differentiating (18) twice with respect to u results in

$$\left\{\frac{\partial^2}{\partial\theta\partial\theta'}K_{EL}(\theta^*,\theta)\right\}_{|\theta*} = V^{-1} + o_p(n^{-2/3}) \;, \qquad (19)$$

and Lemma 1 follows by substituting (19) into (6). □

Proof of Lemma 2.

By Theorem 1 of Qin and Lawless (1994), $n^{1/2}(\theta^* - \tilde{\theta})$ is asymptotically normal in distribution, with mean 0 and covariance matrix V.

In the notation of the above authors, define

$$Q_{1n}(\theta,\tau) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\tau' g(x_i,\theta)}g(x,\theta)$$

and

$$Q_{2n}(\theta,\tau) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\tau' g(x_i,\theta)} \left\{\frac{\partial g(x_i,\theta)}{\partial \theta}\right\}' \tau \ .$$

Expand $Q_{1n}(\tilde{\theta},\tilde{\tau})$ and $Q_{2n}(\tilde{\theta},\tilde{\tau})$ about $(\theta^*,0)$ to obtain the authors' equations $(A.2)$ and $(A.3)$. *Also,* expand $Q_{1n}(\tilde{\theta}_k,\tilde{\tau}_k)$ and $Q_{2n}(\tilde{\theta}_k,\tilde{\tau}_k)$ similarly about $(\theta_k^*,0)$ to obtain the analogous expressions. Note that here the empirical likelihood optimization problem is interpreted as one in the joint parameter (θ,τ). Also note that while $\theta_k \in \mathbb{R}^k$, we still have $\tau_k \in \mathbb{R}^L$, since all L estimating functions are used in both the full model and submodels.

The result of these four expansions are the two approximations

$$V^{-1} n^{1/2}(\theta^* - \tilde{\theta}) \approx E\frac{\partial g'}{\partial \theta}\{-Egg'\}^{-1} n^{1/2} Q_{1n}(\theta^*,0) \tag{20}$$

and

$$V_k^{-1} n^{1/2}(\theta_k^* - \tilde{\theta}_k) \approx E\frac{\partial g'}{\partial \theta_k}\{-Egg'\}^{-1} n^{1/2} Q_{1n}(\theta_k^*,0) \ , \tag{21}$$

where V_k^{-1} is a $k \times L$ matrix containing the first k rows of V^{-1}.

Under the assumption that $\theta_k^* \in \mathcal{B}_n$, approximate equality in the first k rows of the righthand sides of (20) and (21) holds, for large n, implying that

$$V^{-1} n^{1/2}(\theta^* - \tilde{\theta}) \approx V^{-1} n^{1/2}(\theta_k^* - \tilde{\theta}_k) \ . \tag{22}$$

It follows from equation (22) that $\theta_k^* - \tilde{\theta}_k$ is approximately the projection (with respect to V^{-1}) of $\theta^* - \tilde{\theta}$. The asymptotic distributions and approximate independence of

$$n||\theta^* - \tilde{\theta}||_V^2 - n||\theta_k^* - \tilde{\theta}_k||_V^2 \quad \text{and} \quad n||\theta_k^* - \tilde{\theta}_k||_V^2$$

then follow from standard arguments concerning projections of multivariate normal random variables.

The negligibility of the contribution of $n\langle\theta^* - \tilde{\theta}, \theta_k^* - \theta^*\rangle_V$ to (10) follows by first noting that the standard deviation of its asymptotic distribution is $n^{1/2}||\theta_k^* - \theta^*||_V$. The arguments of inspection of Akaike (1973) may then be repeated here verbatim. □

Proof of Result.

Begin by writing

$$\begin{aligned}\sum_{i=1}^{n} p_i g\left\{(z_i, y_i), \theta\right\} &= \sum_{i=1}^{n} p_i (1, z_i, z_i^2)' \left\{y_i - (1, z_i, z_i^2)\theta\right\} \\ &= \sum_{i:z_i=-1} p_i (1,-1,1)' \left(y_i - \theta_0 + \theta_1 - \theta_2\right) \\ &\quad + \sum_{i:z_i=0} p_i (1,0,0)' \left(y_i - \theta_0\right) \\ &\quad + \sum_{i:z_i=1} p_i (1,1,1)' \left(y_i - \theta_0 - \theta_1 - \theta_2\right) \\ &= (0,0,0)' \ . \end{aligned} \tag{23}$$

Simple relations among these three equations may be used to show that the constraint in (23) is equivalent to the constraints

$$\sum_{i:z_i=-1} p_i \left(y_i - \theta_0 + \theta_1 - \theta_2\right) = \sum_{i:z_i=0} p_i \left(y_i - \theta_0\right) = \sum_{i:z_i=1} p_i \left(y_i - \theta_0 - \theta_1 - \theta_2\right) = 0 \ .$$

Hence the function H in (4) may be rewritten as

$$\begin{aligned} H &= \sum_{i=1}^{n} \log(np_i) + \lambda(1 - \sum_{i=1}^{n} p_i) - nt_1 \sum_{i:z_i=-1} p_i \left(y_i - \theta_0 + \theta_1 - \theta_2\right) \\ &\quad - nt_2 \sum_{i:z_i=0} p_i \left(y_i - \theta_0\right) - nt_3 \sum_{i:z_i=1} p_i \left(y_i - \theta_0 - \theta_1 - \theta_2\right) \ , \end{aligned}$$

for $\lambda, t_1, t_2, t_3 \in \mathbb{R}$. Differentiating with respect to p_i yields

$$\frac{\partial H}{\partial p_i} = \begin{cases} \frac{1}{p_i} - \lambda - nt_1 \left(y_i - \theta_0 + \theta_1 - \theta_2\right) \ , & \text{if } z_i = -1 \\ \frac{1}{p_i} - \lambda - nt_2 \left(y_i - \theta_0\right) \ , & \text{if } z_i = 0 \\ \frac{1}{p_i} - \lambda - nt_3 \left(y_i - \theta_0 - \theta_1 - \theta_2\right) \ , & \text{if } z_i = 1 \, . \end{cases}$$

The identity $\sum_{i=1}^{n} p_i \frac{\partial H}{\partial p_i} = 0$ implies that $\lambda = n$, as in Section 2.1. The identities

$$\sum_{i:z_i=-1} p_i \frac{\partial H}{\partial p_i} = \sum_{i:z_i=0} p_i \frac{\partial H}{\partial p_i} = \sum_{i:z_i=1} p_i \frac{\partial H}{\partial p_i} = 0$$

then yield our result. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory,* pp. 267-281. Academiai Kiado: Budapest.

Andrews, D.F. & Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York.

Atkinson, A.C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika* **67**, 413-418.

Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data.* Chapman and Hall, New York.

DiCiccio, T.J., Hall, P. & Romano, J. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika* **76**, 465-476.

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

DiCiccio, T.J. & Romano, J. (1989). On adjustments to the signed root of the empirical likelihood ratio statistic. *Biometrika* **76**, 447-456.

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

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *Ann. Statist.* **18**, 121-140.

Kolaczyk, E.D. (1992). Empiricising Akaike's Information Criterion. Technical Memorandum, AT&T Bell Laboratories, Murray Hill, New Jersey.

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

Mallows, C.L. (1973). Some Comments on C_p. *Technometrics* **15**, 661-676.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals 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. (1992). Empirical likelihood and generalized projection pursuit. Technical Report 393, Dept. of Statistics, Stanford University.

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