Statistica Sinica Preprint No: SS-2020-0254	
Title	A Robust Consistent Information Criterion for Model Selection Based on Empirical Likelihood
Manuscript ID	SS-2020-0254
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0254
Complete List of Authors	Chixiang Chen, Ming Wang, Rongling Wu and Runze Li
Corresponding Author	Ming Wang
E-mail	mwang@phs.psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

A Robust Consistent Information Criterion for Model Selection Based on Empirical Likelihood

Chixiang Chen[1], Ming Wang[2], Rongling Wu[2], Runze Li[3]

[1]*Department of Biostatistics, Epidemiology and Informatics, the University of Pennsylvania, Philadelphia, PA 19104, USA*

[2]*Division of Biostatistics and Bioinformatics, Department of Public Health Science, Pennsylvania State College of Medicine, Hershey, PA 17033, USA*

[3]*Department of Statistics and the Methodology Center, Pennsylvania State University, University Park, PA 16802, USA*

Abstract:

Conventional likelihood-based information criteria for model selection rely on the assumed distribution of the data. However, for complex data, specifying this underlying distribution turns out to be challenging, and existing criteria may be limited and not sufficiently general to handle various model-selection problems. Here, we propose a robust and consistent model-selection criterion based on the empirical likelihood function, which is data driven. In particular, this framework adopts plug-in estimators that can be achieved by solving external estimating equations not limited to the empirical likelihood. This avoids potential computational-convergence issues and allows for versatile applications, such as generalized linear models, generalized estimating equations, and penalized regressions. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

proposed criterion is derived initially from the asymptotic expansion of the marginal likelihood under a variable-selection framework, but more importantly, the consistent model-selection property is established in a general context. Extensive simulation studies confirm that the proposed model-selection criterion outperforms traditional criteria. Finally, an application to the Atherosclerosis Risk in Communities Study illustrates the practical value of the proposed framework.

Key words and phrases: Consistency, Empirical likelihood, Model selection.

1. Introduction

Model selection is a common problem in various disciplines, including variable selection in the mean structure, correlation structure selection for longitudinal data analysis, and tuning-parameter selection in penalized regression, among others. Currently, commonly used approaches for model selection rely on several likelihood-based information criteria, such as the Akaike information criterion (AIC) (Akaike, 1997), Bayesian information criterion (BIC) (Schwarz, 1978), and generalized information criteria (GIC) (Konishi and Kitagawa, 1996). However, these information criteria depend critically on the parametric distribution assumption, and have limited applications in model selection problems that are more complicated than variable selection (Chen and Lazar, 2012). More importantly, a distribution misspecification has a negative impact on the selection performance and is inevitably encountered in practice. For instance, in some survey studies, variables such as Beck's depression index or caffeine/alcohol

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

use can be highly skewed or over-dispersed because of sampling bias. Thus, it is usually difficult to identify a well-defined distribution. However, these complex data play critical roles in capturing the fundamental principles underlying natural, social, and engineering processes. As a result, approaches that are more advanced and rigorous must be adopted for valid inference.

To avoid the distribution specification, but still borrow the likelihood properties, a data-driven approach based on the empirical likelihood (EL) has been developed (Owen, 1988; Qin and Lawless, 1994), and is used widely for data analysis and statistical inference (Owen, 2001). However, few studies have examined EL-based information criteria for model selection. Kolaczyk (1995) proposed the empirical information criterion (EIC) based on the Kullback–Leibler distance between discrete empirical distributions, but it suffered from a severe lack of convergence. To alleviate this computational issue, Variyath et al. (2010) advocated an empirical AIC and an empirical BIC based on the adjusted EL by incorporating an extra parameter (Chen et al., 2008). However, only variable selection in the mean structure was considered, and they still required EL estimators. Several other EL-based criteria have been proposed for particular situations, but without theoretical justification (Tang and Leng, 2010; Chen and Lazar, 2012; Chang et al., 2018; Chen et al., 2019). To the best of our knowledge, few studies have investigated EL-based information criteria that are broadly ap-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

plicable to general model selection.

Here, we consider a general model selection context with a collection of candidate models $M_1, M_2, \ldots, M_k$, with the true model included. Under the Bayesian paradigm with a noninformative prior, the main focus is the marginal likelihood, which is given as

$$P(\mathbf{D}|M) = \int P(\mathbf{D}|\boldsymbol{\gamma}, M)P(\boldsymbol{\gamma}|M)d\boldsymbol{\gamma} \propto \int P(\mathbf{D}|\boldsymbol{\gamma})d\boldsymbol{\gamma}, \qquad (1.1)$$

where $\mathbf{D}$ denotes the full data, $\boldsymbol{\gamma}$ are the parameters in the candidate model M, and $P(\mathbf{D}|\boldsymbol{\gamma})$ is the likelihood function $L(\boldsymbol{\gamma}|\mathbf{D})$. When $L(\boldsymbol{\gamma}|\mathbf{D})$ is fully specified, the well-known criterion $\text{BIC} = -2\log L(\hat{\boldsymbol{\gamma}}|\mathbf{D}) + p\log n$ has been derived by selecting the model corresponding to the largest marginal probability in (1.1). However, there are some restrictions: (i) the distribution must be prespecified for the likelihood L; and (ii) the estimator $\hat{\boldsymbol{\gamma}}$ must be the maximum likelihood estimator. Herein, we present a robust EL-based consistent information criterion (ELCIC) that targets model selection under general contexts, and is not limited to variable selection. In particular, the robustness exhibits the distribution-free property and flexibility in wide application to general model selection problems. Consistency is defined as capturing the true model with probability tending to one, a separate research topic that has attracted considerable attention (Chen and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Chen, 2008; Variyath et al., 2010; Kim and Jeon, 2016). The likelihood part of the ELCIC is purely data-driven based on the EL. We relax the procedure for parameter estimation by using plug-in estimators to calculate this criterion, under which the consistency property still holds under mild conditions.

The rest of this paper is organized as follows. In Section 2, to demonstrate the formulation of our proposed criterion, we provide its theoretical derivation under the variable selection framework by expanding asymptotically the marginal probability (1.1). We then investigate the consistency property of our proposed criterion under more general model selection settings with mild conditions, the most important aspect herein. In Section 3, we consider three specific cases for illustration and evaluate the finite-sample performance using simulation studies. In Section 4, we apply our proposal to a real-data example. Finally, in Section 5, we discuss several promising extensions as future work.

2. Methodology

2.1 Empirical Likelihood

Inspired by the empirical distribution, Owen (1988) introduced an EL approach for constructing likelihood-based confidence intervals. The full data are denoted by $\mathbf{D} = \{\mathbf{D}_i\}_{i=1}^n$ with $\mathbf{D}_i = (\mathbf{X}_i^{\mathrm{T}}, \mathbf{Y}_i^{\mathrm{T}})^{\mathrm{T}}$, which are assumed to be independent and identically distributed (i.i.d.) when a regular regression is considered. Given

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

some estimating equations $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ satisfying $E\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}) = \mathbf{0}$, for $i = 1, \ldots, n$, the empirical likelihood ratio (ELR) is defined by

$$R^F = \sup_{\boldsymbol{\gamma}, p_1, \ldots, p_n} \left\{ \prod_{i=1}^{n} np_i; p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{g}\big(\mathbf{D}_i, \boldsymbol{\gamma}\big) = \mathbf{0} \right\}. \tag{2.1}$$

Unlike with the traditional likelihood, the observations here use point-mass probabilities. Thus, the information from the data is borrowed automatically and efficiently from the constraints in (2.1) (Qin and Lawless, 1994), which is a desired property, and shows its considerable potential for model selection. Given the estimator denoted by $\hat{\boldsymbol{\gamma}}$, discussed further in subsequent sections, the negative logarithm of the ELR is calculated easily using Lagrange multipliers (Owen, 2001); that is,

$$l = -\log R^F(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \log\{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}})\}, \tag{2.2}$$

where the parameter estimate $\hat{\boldsymbol{\lambda}}$ is obtained using the Newton–Raphson method to solve

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}\big(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}\big)}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}\big(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}\big)} = \mathbf{0}. \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

2.2 Derivation of the ELCIC under Variable-Selection Framework

Before introducing our proposed model selection criterion and discussing its consistency, we obtain some insights into its formulation by considering variable selection in a regression framework. Here, $\boldsymbol{\gamma}$ is the parameter vector in the mean structure only. To implement EL-based selection, we should specify a full set of estimating equations (Kolaczyk, 1995; Variyath et al., 2010; Chen and Lazar, 2012). For any $i = 1, 2, \ldots, n$, we define

$$\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{g}_1(\mathbf{D}_i, \tilde{\boldsymbol{\gamma}}) \\ \mathbf{g}_2(\mathbf{D}_i, \tilde{\boldsymbol{\gamma}}) \end{pmatrix}, \tag{2.4}$$

where $\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ so that its dimension matches that of the prespecified full covariate matrix $\mathbf{X}_i$, and $\mathbf{g}_1(\mathbf{D}_i, \tilde{\boldsymbol{\gamma}})$ and $\mathbf{g}_2(\mathbf{D}_i, \tilde{\boldsymbol{\gamma}})$ correspond to the estimating equations for the parameters with and without involvement, respectively, in a candidate model. In variable selection, we denote the cardinality of $\tilde{\boldsymbol{\gamma}}$ as L, and that of $\boldsymbol{\gamma}$ from a candidate model as p, with $0 < p \leq L < \infty$. Note that (2.4) is constructed only to implement the variable selection.

As discussed previously, computational issues will be encountered when maximizing (2.1) to obtain the EL-based estimator, denoted by $\hat{\boldsymbol{\gamma}}_{EL}$. To ensure the existence of solutions, the convex hull of the estimating equations should

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

contain zero (Qin and Lawless, 1994; Chen et al., 2008), which is not guaranteed in practice. To overcome these computational issues and make the criterion more versatile, we consider plug-in estimators instead. In the variable selection framework, these are obtained by solving some external estimating equations $\sum_{i=1}^{n} \mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}) = \mathbf{0}$ in (2.4) to bypass the complex and unstable estimation procedure from (2.1). Given the plug-in estimators, denoted as $\hat{\boldsymbol{\gamma}}_{EE}$, and the corresponding Lagrange-multiplier estimator $\hat{\boldsymbol{\lambda}}_{EE}$, the negative logarithm of the EL ratio in (2.2) can be achieved. Instead of using the likelihood under a pre-specified distribution to maximize the marginal likelihood (1.1), we employ the ELR for $L(\boldsymbol{\gamma}|\mathbf{D})$, given as

$$L(\boldsymbol{\gamma}|\mathbf{D}) = R^F(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})\}^{-1}. \tag{2.5}$$

Next, we provide two conditions to facilitate the asymptotic expansion of $P(\mathbf{D}|M)$. Note that we use $\|\cdot\|$ to denote the Euclidean norm, and $|\cdot|$ to denote the determinant of a matrix. Furthermore, to simplify the notation, we drop the subscript for $\mathbf{D}$ based on the i.i.d. property, and evaluate all expectations at the true parameter values under the correctly specified models.

Condition 1. *(Regularities) Given the correctly specified model with the true parameter* $\boldsymbol{\gamma}_0$ *satisfying* $E\{\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma}_0)\} = \mathbf{0}$*,* $E\{\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D}, \boldsymbol{\gamma}_0)\}$ *is positive*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

definite and $\{\partial^2 \mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})\}/(\partial \boldsymbol{\gamma}^{\mathrm{T}} \partial \boldsymbol{\gamma})$ *is continuous in the neighborhood of* $\boldsymbol{\gamma}_0$. *Furthermore, we assume that* $\|\{\partial \mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})\}/(\partial \boldsymbol{\gamma}^{\mathrm{T}})\|$, $\|\{\partial^2 \mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})\}/(\partial \boldsymbol{\gamma}^{\mathrm{T}} \partial \boldsymbol{\gamma})\|$, *and* $\|\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})\|^3$ *are bounded by some integrable function around* $\boldsymbol{\gamma}_0$.

Condition 2. *(Efficiency) For the estimating equations* $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *defined in (2.4), and given the correctly specified model with the true parameter* $\tilde{\boldsymbol{\gamma}}_0 = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$,

$$E\Big\{\frac{\partial \mathbf{g}_1(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big\} = -E\big\{\mathbf{g}_1(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)\mathbf{g}_1^{\mathrm{T}}(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)\big\},$$
$$E\Big\{\frac{\partial \mathbf{g}_2(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big\} = -E\big\{\mathbf{g}_1(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)\mathbf{g}_2^{\mathrm{T}}(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)\big\}.$$

Condition 1 includes several regular moment conditions to ensure a valid EL-based inference (Qin and Lawless, 1994). To simplify the formula for our proposed criterion, Condition 2 imposes some constraints on the estimating equations $\mathbf{g}_1$ and $\mathbf{g}_2$ in (2.4), which are related to the estimator efficiency. Note that Condition 2 contains a family of estimating equations that lead to asymptotically efficient estimators. For instance, if the score function is used, then Condition 2 is definitely satisfied by the property of the Fisher information under regular conditions (Pierce, 1982); if g comprises generalized estimating equations (GEEs) with a correctly specified correlation structure (Liang and Zeger, 1986), then Condition 2 holds as well. However, Condition 2 is not required

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

in the proof of model selection consistency, which is discussed further in next section.

Theorem 1. *Under Conditions 1 and 2, given $\hat{\boldsymbol{\gamma}}_{EE}$ obtained from the estimating equations $\mathbf{g}_1$ in (2.4) and the rank of $E[\{\partial \mathbf{g}(\mathbf{D}, \boldsymbol{\gamma}_0)\}/(\partial \boldsymbol{\gamma}^{\mathrm{T}})]$ being p, the same as the dimensionality of $\hat{\gamma}_{EE}$, and by applying the Laplace approximation and setting a noninformative prior to $\boldsymbol{\gamma}$, we have*

$$-2 \log P(\mathbf{D}|M) = -2 \log R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}) + p \log n + \tilde{C} + o_p(1),$$

where $\tilde{C} = \log(\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}) - p \log(2\pi) - 2 \log(\tilde{A})$, with $\boldsymbol{\Sigma}_{11} = E\big\{\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma}_0) \mathbf{g}^{\mathrm{T}}(\mathbf{D}, \boldsymbol{\gamma}_0)\big\}$; $\boldsymbol{\Sigma}_{12} = E\big[\partial \mathbf{g}(\mathbf{D}, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma}^{\mathrm{T}}\big]$, $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^{\mathrm{T}}$; $\tilde{A} = \int \exp\{(1/2) \boldsymbol{\delta}_1^{\mathrm{T}} (n \boldsymbol{\Sigma}_{11}) \boldsymbol{\delta}_1\} \rho_{\boldsymbol{\delta}_1}(\boldsymbol{\delta}_1) d\boldsymbol{\delta}_1$; and $\rho_{\boldsymbol{\delta}_1}(\cdot)$ is some prior function of the random variable $\boldsymbol{\delta}_1$, defined in the Supplementary Material.

Based on Theorem 1, our proposed ELCIC is defined as

$$\text{ELCIC} = -2 \log R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}) + p \log n. \tag{2.6}$$

Note that the proposed ELCIC is free of prior specification, a desired property for a well-defined model selection criterion, with finite values guaranteed, regardless of the fact that $\tilde{A}$ in Theorem 1 might be infinite for some prior functions.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Theorem 1 is derived under the framework of variable selection based on the Laplace approximation, with some extra constraints on the estimating equations, as specified in Condition 2. However, we show in Section 2.3 that the ELCIC is consistent (i.e., it captures the true model with probability approaching one), and so those conditions can be relaxed. Note that this consistency property holds under any general estimating equations, not just those defined in (2.4). In Section 2.3, we explore rigorously the consistency of our ELCIC.

2.3 Model-Selection Consistency of the ELCIC

In this section, we focus on the consistency of our proposal for general model selection, not limited to variable selection. Under this general context, let us redefine $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$ as some full estimating equations in (2.1) satisfying Condition 1. Here, the p-by-1 parameter vector $\boldsymbol{\gamma}$ includes all the parameters in the estimating equations $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$, such as those in the mean structure, the coefficients in the correlation matrix, or any other nuisance parameters. Note that we do not assume that $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$ correspond to the estimating equations for all the parameters in $\boldsymbol{\gamma}$. We further relax the assumption that the plug-in estimators $\hat{\boldsymbol{\gamma}}_{EE}$ can be derived from other external estimating equations, not necessarily specific to (2.4), but with some mild conditions satisfied (i.e., Condition 5). In Section 3.1, we discuss more concrete examples of specifying the function $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Theorem 2. *Under Condition 1 and the candidate model M specified correctly with the true value of $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}_0$, define $\mathbf{Q}_n = (1/n)\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) + \boldsymbol{\Sigma}_{12}(\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0)$ with plug-in estimators satisfying $\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0 = O_p(\mathbf{n}^{-1/2})$. Then, the negative logarithm of the likelihood is*

$$l = \frac{1}{2}(n^{1/2}\mathbf{Q}_n^{\mathrm{T}})\boldsymbol{\Sigma}_{11}^{-1}(n^{1/2}\mathbf{Q}_n) + o_p(1). \tag{2.7}$$

Theorem 2 provides insight into the order of l when the model is specified correctly, which is a crucial component in the derivation of the consistency of our proposed criterion. As a by-product of Theorem 2, Corollary 1 provides the asymptotic distribution of $2l$ in the variable selection framework.

Corollary 1. *Given the same conditions as those in Theorem 2, $\mathbf{g}$ and $\mathbf{g}_1$ defined in (2.4) in the case of variable selection, and $\tilde{\boldsymbol{\gamma}}_0 = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\boldsymbol{\gamma}}_{EE}$ satisfies*

$$\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0 = -\left\{E\Big(\frac{\partial \mathbf{g}_1(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_1(\mathbf{D}_i, \tilde{\boldsymbol{\gamma}}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}). \tag{2.8}$$

Then, we have that $2l$ converges in distribution to $\sum_{j=1}^{\tilde{L}}\Lambda_j\chi_1^2$, where $\Lambda_1, \ldots, \Lambda_{\tilde{L}}$ are nonzero eigenvalues of the matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{\Sigma}_*^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_{11}^{1/2}$, with $\tilde{L} = rank(\boldsymbol{\Omega})$ and $\boldsymbol{\Sigma}_* = \mathbf{I}_{L\times L} - \Big(\boldsymbol{\Sigma}_{12}\{E(\partial \mathbf{g}_1(\mathbf{D}, \tilde{\boldsymbol{\gamma}}_0)/\partial \boldsymbol{\gamma}^{\mathrm{T}})\}^{-1}, \mathbf{0}_{L\times(L-p)}\Big)$.

Condition 3. *(Regularity for Misspecified Model) Let $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$. Then, for any*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

$\boldsymbol{\gamma}$ in the neighborhood of $\boldsymbol{\gamma}_*$, we have that $E\|\mathbf{g}(\mathbf{D},\boldsymbol{\gamma})\|^{2+\delta} < \infty$, with some $\delta > 0$.

Condition 4. *(Identifiability) For any $\boldsymbol{\gamma}$ in the neighborhood of $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$, we have the condition that $\|E\mathbf{g}(\mathbf{D},\boldsymbol{\gamma})\| > 0$.*

Condition 3 extends the regularities in Condition 1 when the candidate model M is misspecified. We require the $(2+\delta)$th moment of the estimating equations to be finite. Condition 4 is the identifiability assumption, which further implies that the model is identifiable if only the correctly specified model satisfies $E\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0) = \mathbf{0}$. This is also the key to model selection.

Theorem 3. *Under Conditions 1, 3, and 4, for any $\boldsymbol{\gamma}$ in the neighborhood of $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$, we have $n^{1-c}\|\bar{\mathbf{g}}_n\|^2 \log(n) l^{-1} = O_p(1)$, where $\bar{\mathbf{g}}_n = (1/n)\sum_{i=1}^n \mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})$, for $\frac{1}{2} < c < 1$.*

Theorem 3 ensures that if the candidate model is misspecified, then the negative log likelihood l tends to infinity with order of at least $\log n$. Together with Theorems 2 and 3 and the following Condition 5, we are ready to present our main result.

Condition 5. *(Well-behaved Estimator) Denote $\mathbf{V} = \boldsymbol{Cov}\big((1/n)\sum_{i=1}^n \mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0) + \boldsymbol{\Sigma}_{12}(\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0)\big)$, with $\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0 = O_p(\mathbf{n}^{-1/2})$. Then, given the correctly specified candidate model, we have that $tr(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{V}) < \infty$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Condition 5 is a mild condition requiring well-behaved plug-in estimators. On the one hand, if this plug-in estimator $\hat{\boldsymbol{\gamma}}_{EE}$ is from Corollary 1, then $tr(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{V}) < \infty$ in Condition 5 is equivalent to asking for finite eigenvalues of $\boldsymbol{\Omega}$ defined in Theorem 2. On the other hand, if $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$ are the estimating equations for $\boldsymbol{\gamma}$ and the estimator $\hat{\boldsymbol{\gamma}}_{EL}$ is obtained from maximizing the ELR in (2.1), then we have that $tr(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{V}) = L - p$, which satisfies Condition 5.

Theorem 4. *Under Conditions 1, 3–5 and given the true model denoted by M_0, we have $P\big[\min\{\text{ELCIC}(M) : M \neq M_0\} > \text{ELCIC}(M_0)\big] \to 1$ as $n \to \infty$.*

The proof strategy is based on standard large-sample theory, but important findings from Theorem 4 reveal the underlying merits and implications of the ELCIC for general model selection. First, the theorem holds under very mild conditions, in particular, involving neither Condition 2 nor the Laplace approximation. Moreover, the proof does not rely on a very specific form of the full estimating equations, not necessarily limited to the estimating equations $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ in (2.4). This implies that the ELCIC has potential for various model selection problems, and not just variable selection. Second, the plug-in estimators are subject to few restrictions other than Condition 5. In practice, common estimation procedures could be applied, such as the least squares, score functions, GEEs, or loss functions, thereby making the ELCIC more flexible and versatile. Note that the parameters $\boldsymbol{\gamma}$ are not limited to the primary parameters of interest

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

in the candidate model. They can also include other nuisance parameters, such as the correlation coefficients in the GEE method for longitudinal data, and the parameters in logistic regressions for observing probabilities in the method of inverse probability weights. Therefore, the consistency property in Theorem 4 allows the ELCIC to deal with a broad range of model selection problems, as exemplified by the three case studies presented in Section 3.

3. Case Studies and Numerical Results

3.1 Full Estimating Equations

We discuss how to specify the full estimating equations $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ for the ELCIC in three cases. To avoid confusion, we denote the parameter vector in the mean structure as $\boldsymbol{\beta}$, and the overall parameter vector in the estimating equations as $\boldsymbol{\gamma}$.

Case 1: Generalized Linear Models (GLMs). Nelder and Wedderburn (1972) introduced the GLM concept to unify the theories for different models in categorical analysis. In this case, the full estimating equations $\mathbf{g}$ in (2.1) can be defined simply as the score functions, that is

$$\mathbf{g}(\mathbf{D}_i, \boldsymbol{\beta}) = \mathbf{X}_i(Y_i - \mu_i(\tilde{\boldsymbol{\beta}})), \tag{3.1}$$

where $\mu_i(\tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ is the conditional expectation of Y_i modeled by

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

$f(\mathbf{X}_i^T\tilde{\boldsymbol{\beta}})$, with some prespecified canonical link function f. Given that (3.1) is valid only when the mean structure is specified correctly, without requiring the second moment, the ELCIC under the full estimating equations (3.1) can handle the variance structure being misspecified, such as over-dispersion, which is often encountered in count data analysis. Only variable selection is of research interest in this case.

Case 2: Generalized Estimating Equations (GEE). We now extend our focus to model selection for longitudinal data. Liang and Zeger (1986) introduced the marginal model to conduct statistical inference without specifying the joint distribution of longitudinal data. Note that a correctly specified mean structure is always key for estimation consistency, and in the GEE approach, the efficiency is improved by identifying the correct "working" correlation structure. In this case, we specify the full estimating equations properly so that the ELCIC can select the marginal mean and the correlation structures simultaneously. In contrast, the main existing criteria, such as the quasi-likelihood criterion (QIC) (Pan, 2001), cannot handle joint selection.

To achieve our goal and for simplicity, we assume a balanced design with T observations for each subject. For subject i, the marginal mean is denoted by $\boldsymbol{\mu}_i$ and the variance–covariance matrix by $\mathbf{V}_i$. The over-dispersion parameter is denoted by ϕ (assumed known, but can also be estimated consistently) and the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

correlation-coefficient vector by $\boldsymbol{\rho}^c = (\rho_1^c, \ldots, \rho_{T-1}^c)^{\mathrm{T}}$. Here, the superscript c indicates the type of correlation structure. For instance, under a stationary structure, we have $\boldsymbol{\rho}^{STA} = (\rho_1^{STA}, \ldots, \rho_{T-1}^{STA})^{\mathrm{T}}$. Thus, the full estimating equations in (2.1) are defined as

$$\mathbf{g}\Big(\mathbf{D}_i, \boldsymbol{\beta}, \boldsymbol{\rho}^c\Big) = \begin{pmatrix} \mathbf{H}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \Big(\mathbf{Y}_i - \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\Big) \\ \mathbf{U}_i(\tilde{\boldsymbol{\beta}}) - \boldsymbol{h}(\boldsymbol{\rho}^c)\phi \end{pmatrix}, \tag{3.2}$$

where $\tilde{\boldsymbol{\beta}}$ is defined as $(\boldsymbol{\beta}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{H}_i$ denotes the first derivative of $\boldsymbol{\mu}_i$ with respect to $\tilde{\boldsymbol{\beta}}$, and $\mathbf{U}_i(\tilde{\boldsymbol{\beta}}) = \big(U_{i1}(\tilde{\boldsymbol{\beta}}), U_{i2}(\tilde{\boldsymbol{\beta}}), \ldots, U_{i(T-1)}(\tilde{\boldsymbol{\beta}})\big)^{\mathrm{T}}$, with

$$U_{im}(\tilde{\boldsymbol{\beta}}) = \sum_{j=1}^{T-m} e_{ij}(\tilde{\boldsymbol{\beta}}) e_{i,j+m}(\tilde{\boldsymbol{\beta}}), \text{ for } m = 1, \ldots, T-1. \tag{3.3}$$

In addition, e_{ij} represents the standardized residual term $(y_{ij} - \mu_{ij})/\sqrt{\nu_{ij}}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, T$. Finally, $\boldsymbol{h}(\boldsymbol{\rho}^c)$ is defined as $\big(\rho_1^c\big(T{-}1{-}p/n\big), \ldots, \rho_{T-1}^c(1-p/n)\big)^{\mathrm{T}}$.

Note that $\tilde{\boldsymbol{\beta}}$ is proposed to achieve variable selection in a marginal mean structure, and a stationary correlation structure is used based on (3.3) to select the correlation structure. Thus, the ELCIC can select marginal mean and correlation structures simultaneously, because the expectation of the full estimating equations (3.2) is zero only when both structures are specified correctly. Note

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

too that Chen and Lazar (2012) proposed the ELBIC to select a "working" correlation structure alone, under which the full estimating equations comprise a subset of our proposed ones in (3.2). Accordingly, the ELCIC unifies the selection procedure by allowing for both marginal mean and correlation structures. The theorems in the Supplementary Material provide a theoretical justification for this criterion, something that is currently lacking in the literature to date.

Case 3: Penalized Generalized Estimating Equations (PGEE). Penalized regression is among the most popular research topics of the past two decades (Tibshirani, 1996; Fan and Li, 2001). It uses penalties to shrink the effect of unnecessary features toward zero by identifying a proper tuning parameter. In this case, we focus mainly on selecting the tuning parameter, for which there are two common approaches, namely cross-validation (CV) and some BIC-type methods (Chen and Chen, 2008). As is well known, CV leads to a high rate of false positives, whereas BIC-type methods are less time consuming and tend to have lower rates of false positives. However, BIC-type criteria cannot be applied to semiparametric or nonparametric contexts. Here, we consider the PGEE proposed by Wang et al. (2012), for which the BIC is no longer suitable, but the ELCIC can be easily embedded.

The PGEE is a combination of the GEE and the first derivative of the smoothly clipped absolute deviation (SCAD) penalty, thereby facilitating spar-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

sity in the marginal mean structure. Wang et al. (2012) investigated the selection consistency and asymptotic normality with a diverging number of covariates. Here, we consider only cases with a fixed number of covariates, and use the full estimating equations $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ in (3.2) to implement model selection. Note again that the consistency of the ELCIC does not require that the estimating equations for $\hat{\boldsymbol{\beta}}_{EE}$ be contained in $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$, which justifies theoretically the application of the ELCIC to cases of penalized regression. Moreover, based on the selection consistency and asymptotic normality in Wang et al. (2012), there exists a tuning parameter that identifies true zeros correctly with probability tending to one, and that makes the nonzero part of the estimators converge to the true one with order $O_p(n^{-\frac{1}{2}})$ under fixed p. Therefore, we can apply the ELCIC to locate this "optimal" tuning parameter. Furthermore, using the same rationale as that in Case 2, the full estimating equations $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ in (3.2) facilitate joint selection of marginal mean and "working" correlation structures, something that is unfeasible based on CV.

As well as the three traditional cases presented above, the ELCIC can deal with complicated scenarios to which the existing criteria might apply either poorly or not at all, such as variable selection for the augmented inverse probability weighting (AIPW) method, which is commonly used for missing-data analysis (Robins et al., 1994), with extensive work in longitudinal data, sur-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

vival analysis, and causal inference (Bang and Robins, 2005; Seaman and Copas, 2009; Scharfstein et al., 1999; Long et al., 1997). In the Supplementary Material, we discuss such cases in detail while evaluating our proposal numerically.

3.2 Numerical Results

Here, we report on simulation studies conducted under the three cases in Section 3.1, and in each case, we compare the ELCIC with popular existing criteria to show that the former is robust. Given the limited space available here, the Supplementary Material provides additional simulation studies for variable selection under the AIPW framework.

Case 1. The main goal in this case is to determine how variable selection is affected by a distribution misspecification. We narrow our focus to the Poisson regression, a special GLM case. The true mean structure is

$$\log(\mu_i) = \beta_0 + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \qquad \text{for} \quad i = 1, \ldots, n,$$

where $\beta_0 = 0.5$ and $\boldsymbol{\beta} = (0.5, 0.5, 0)^{\mathrm{T}}$. Furthermore, $\mathbf{X}_i$ is a covariate vector from a three-dimensional multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \mathbf{V})$, where the variance–covariance matrix $\mathbf{V}$ is an AR1 matrix with unit variance and a correlation coefficient of 0.5. To account for the variance in the Poisson distri-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

bution being misspecified, we apply a negative binomial as the true distribution with $k = 2$ or 8 failures. However, we use the AIC, GIC, and BIC for variable selection under the assumption of the Poisson distribution, and apply the ELCIC under the full estimating equations $\mathbf{g}$ specified in (3.1). The correct specification of the Poisson distribution is also considered as a benchmark. We generate 500 Monte Carlo data sets with sample size $n = 100, 200$, or 400, and we report the selection rates for each candidate model for comparison. Table 1 shows that if the variance structure is specified correctly, then the ELCIC is comparable to the BIC, but performs slightly less well, which is understandable, given that that BIC incorporates all the likelihood information needed for the data. However, if the variance structure is misspecified, then the situation is reversed, and the ELCIC is much more robust than the AIC or BIC with an increasing sample size n. The ELCIC offers more advantages when the data have higher over-dispersion, and unlike the ELCIC; the consistency property does not hold for the AIC or GIC. Moreover, although the GIC somehow relaxes the distribution assumption and is more robust than the AIC, it is more sensitive to the distribution misspecification than is the ELCIC, even under a relatively large sample size.

Case 2. We apply the ELCIC to the GEE framework, and compare it with the popular QIC. Suppose that the true underlying correlation structure is exchangeable (EXC) with correlation coefficient $\rho = 0.5$. We assume count out-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

comes with the true marginal mean defined as

$$\log(\mu_{ij}) = \beta_0 + x_{i1}\beta_1 + x_{ij2}\beta_2 \qquad \text{for} \quad i = 1, \ldots, n,\ j = 1, \ldots, T,$$

where x_{i1} is the subject-level (cluster-level) covariate generated from a uniform distribution $\mathrm{U}[0, 1]$, and $x_{ij2} = j - 1$ is a time-dependent covariate. A redundant covariate X_{ij3} is generated from a standard normal distribution $\mathrm{N}(0, 1)$. The number of observations (i.e., cluster size) is $T = 3$, and the true parameters are $\boldsymbol{\beta} = (-1, 1, 0.5)^{\mathrm{T}}$. As discussed previously, the ELCIC with the full estimating equations (3.2) can select the marginal mean and correlation structures simultaneously. However, the QIC is insufficiently powerful to implement joint selection, Therefore, we instead use the correlation information criterion (CIC) (Hin and Wang, 2009) to identify the correlation structure under the full marginal mean structure and then implement the QIC to select the variables in the marginal mean under the selected correlation structure. We also compare the performance with that of QIC/b, which is the QIC, but with the BIC penalty. We generate 500 Monte Carlo data sets with sample size $n = 100$ or 300 and observation times $T = 3$ or 5, and summarize the selection rates for each combination of marginal mean and correlation structures in Table 2. As can be seen, the two-stage selection procedures based on the CIC and QIC (QIC/b) are less

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

powerful, particularly for $T = 3$, given that the second stage of the variable selection relies heavily on the first stage for the correlation structure selection. In contrast, the ELCIC maintains a much higher selection rate across the various scenarios, thereby exhibiting more robustness for model selection in the framework of longitudinal data. See Section 5 for a detailed discussion. To show the flexibility of the ELCIC in terms of handling nuisance parameters, we also implement the variable selection by using the first part of the estimating equations in (3.2) as $\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$, thereby regarding the correlation coefficients as nuisance parameters. As shown by the results in the Supplementary Material, we observe a higher selection rate than that of the QIC.

Case 3. We simulate data from the model $\mathbf{Y}_i = \mathbf{X}_i^T\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, for $i = 1, \ldots, n$, with $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0, 0, 0, 0)^{\mathrm{T}}$. Note that $\mathbf{X}_i$ is a $T \times 7$ matrix from a multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \mathbf{V})$, where $T = 3$ and the variance-covariance matrix $\mathbf{V}$ is an AR1 matrix with unit variance and a correlation coefficient of 0.5. The random errors $\{\boldsymbol{\epsilon}_i\}$ are generated from a multivariate normal distribution with zero mean and an exchangeable covariance matrix with $\sigma^2 = 1$ and $\rho = 0.5$. We generate 500 Monte Carlo data sets with sample size $n = 100$ or 200, and record the following evaluation measures: consistency, false-positive rate, overall variable selection rate, overall variable over-selection rate, correlation structure selection rate, joint selection rate, and variable over-selection rate

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

with the true correlation structure selected. Both ELCIC_1 and ELCIC_2 use the full estimating equations (3.2), where ELCIC_1 is calculated under the true correlation structure $\boldsymbol{\rho}^{EXC}$ to compare with CV for variable selection, and ELCIC_2 is used to jointly select the marginal mean and correlation structures. In Table 3, ELCIC_1 generally gives substantially lower false-positive rates and higher variable selection rates than those of the CV-based method, and ELCIC_2 performs the simultaneous selection satisfactorily when CV is not applicable.

4. Real-Data Example

We apply our method to the Atherosclerosis Risk in Communities Study (ARIC), designed originally to investigate the causes and clinical outcomes of atherosclerosis and trends in the rates of hospitalized myocardial infarction and coronary heart diseases. Our outcome is platelet count, which has been studied in the literature and shown to be an essential factor in coronary heart diseases (Renaud and De Lorgeril, 1992). The objective of this application is to investigate the temporal pattern of platelet count, and to identify potential risk factors among various baseline variables, such as age (year), gender (female or male), diabetes (1=yes; 0=no), smoker (1=yes; 0=no), body mass index (kg), total cholesterol (mmol/L), total triglycerides (mmol/L), and the time-dependent visit variable (coded as 0,1,2,3). We select Washington County to identify a total of 1,463

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

white patients at approximately three-year intervals (1987–1989, 1990–1992, 1993–1995, and 1996–1998) who were diagnosed with hypertension at the first examination. Note that there were 441 dropouts during the follow-up. To illustrate the application of our proposal, we assume missing completely at random, for simplicity; for more-sophisticated manipulation of missing data, see Chen et al. (2019). We apply the full estimating equations (3.2) with the GEE in Case 2 to construct the ELCIC and facilitate the joint selection of the marginal mean and correlation structures. The results are summarized in Table 4 in the Supplementary Material, and indicate that both the QIC and the ELCIC recommend the marginal mean , including time, gender, age, diabetes, and cholesterol, with the AR1 correlation structure as the optimal model. Note that the variables selected by the ELCIC match the significant ones when we fit the full model (Model 1). The same marginal mean model is identified by the PGEE procedure from Case 3 using the same covariate pool.

5. Discussion

Because of complex features and data structures, existing approaches for analysis and inference based on specifying a distribution may not work well. In this work, we present a data-driven information criterion framework for model selection under different contexts. By further relaxing the estimation procedure, our

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

ELCIC overcomes the limitation of classic EL-based criteria. More importantly, it can be extended easily, depending on the model selection needs in practice, to situations in which no existing information criteria would theoretically fit well, such as Case 3 and the extra cases shown in the Supplementary Material. Further discussion about the conditions in the theorems and the theoretical limitations of the existing GIC are provided in the Supplementary Material.

Several extensions of the ELCIC are open to research, including its extension to (ultra) high-dimensional cases. Prompted by those who reviewed this paper, we have begun to investigate this framework by means of a literature search and empirical simulation studies, which show that the ELCIC outperforms alternative criteria; see the Supplementary Material for more details. Furthermore, robust model prediction criteria would be of independent research interest, and an informative degree of freedom instead of p could be used to borrow more information. However, this would increase the criterion complexity and case specificity and, hence, reduce its flexibility in practice. As shown in our work, the ELCIC has indicated its potential and robustness for exploring statistical issues related to model selection with highly complex and diverse data.

Supplementary Material

The online Supplementary Material provides (i) detailed proofs of Theorems 1–4 and Corollary 1, (ii) results from additional simulation studies, includ-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

ing cases for variable selection under the AIPW estimator framework and the ultrahigh-dimensional setup, (iii) discussions about the conditions in the theorems and the limitations of the existing GICs, and (iv) the results for the real-data application using the ARIC study.

Acknowledgments The authors thank the editor, associate editor, and referees for their valuable suggestions. Ming Wang was supported partially by Grants TR002014 and KL2 TR002015 from the National Center for Advancing Translational Sciences (NCATS) and by a grant from the Pennsylvania Department of Health using Tobacco CURE Funds. The contents of this paper are solely the responsibility of the authors, and do not represent the official views of the National Institute of Health and the Pennsylvania Department of Health.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716-723.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**,962-973.

Chang, J., Tang, C., and Wu, T.(2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *Ann. Stat.* **46**,3185-3216.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Chen, C., Shen, B., Zhang, L., Xue, Y., and Wang, M. (2019). Empirical-likelihood-based criteria for model selection on marginal analysis of longitudinal data with dropout missingness. *Biometrics* **75**,950-965.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**,759-771.

Chen, J. and Lazar, N. A. (2012). Selection of working correlation structure in generalized estimating equations via empirical likelihood. *J. Comput. Graph. Stat.* **21**,18-41.

Chen, J., Variyath, A. M., and Abraham, B.(2008). Adjusted empirical likelihood and its properties. *J. Comput. Graph. Stat.* **17**,426-443.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties.*J. Am. Stat. Assoc.* **96**, 1348-1360.

Hin, L. Y. and Wang, Y. G.(2009). Working-correlation-structure identification in generalized estimating equations. *Statist. Medic. 28*,642-658.

Kim, Y. and Jeon, J. J. (2016). Consistent model selection criteria for quadratically supported risks. *Ann. Stat.* **44**, 2467-2496.

Kolaczyk, E. D. (1995). An information criterion for empirical likelihood with general estimating equations.*Technical Report 417, Department of Statistics, The University of Chicago.*

Konishi, S. and Kitagawa, G.(1996). Generalised information criteria in model selection.*Biometrika* **83**, 875-890.

Liang, K. Y. and Zeger, S. L.(1986). Longitudinal data analysis using generalized linear models. *Biometrika*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

73, 13-22.

Long, Q., Zhang, X., and Johnson, B. A.(2011). Robust estimation of area under roc curve using auxiliary variables in the presence of missing biomarker values. *Biometrics* **67**, 559-567.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **135**, 370-384.

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

Owen, A. B. (2001).*Empirical likelihood*. CRC press.

Pan, W.(2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Stat.* **10**, 475-478.

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

Renaud, S. d. and De Lorgeril, M.(1992). Wine, alcohol, platelets, and the french paradox for coronaryheart disease. *The Lancet* **339**, 1523-1526.

Robins, J. M., Rotnitzky, A., and Zhao, L. P.(1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846-866.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models.*J. Am. Stat. Assoc.* **94**, 1096-1120.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.

Seaman, S. and Copas, A.(2009). Doubly robust generalized estimating equations for longitudinal data. *Statist. Medici.* **28**, 937-955.

Tang, C. Y. and Leng, C. (2010). Penalized high-dimensional empirical likelihood.*Biometrika* **97**, 905-920.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R Stat. Soc. Series B Stat. Methodol.* **58**, 267-288.

Variyath, A. M., Chen, J., and Abraham, B. (2010). Empirical likelihood based variable selection. *J. Stat. Plan. Inference.* **140**, 971-981.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**, 353-360.

Department of Biostatistics, Epidemiology and Informatics, the University of Pennsylvania

E-mail: chixiang.chen@pennmedicine.upenn.edu

Department of Public Health Science, Pennsylvania State College of Medicine

E-mail: muw22@psu.edu; Phone: 717-53105745; Fax: 717-531-5779

Department of Public Health Science, Pennsylvania State College of Medicine

E-mail: ronglingwu@phs.psu.edu

Department of Statistics and the Methodology Center, Pennsylvania State University

E-mail: rzli@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Table 1: Performance of ELCIC compared with the AIC and BIC for the scenarios under a Poisson distribution with potential over-dispersed outcomes. 500 Monte Carlo data are generated with sample size $n = 100, 200$. The model with $\{x_1, x_2\}$ is the true one. NB: Negative binomial with k as number of failures.

n	Distribution	Criteria	Candidate Models						
			x_1	x_2	x_3	$\boldsymbol{x_1, x_2}$	x_1, x_3	x_2, x_3	x_1, x_2, x_3
100	POISSON	AIC	0	0	0	0.852	0	0	0.148
		GIC	0	0	0	0.798	0	0	0.202
		BIC	0	0	0	0.980	0	0	0.020
		ELCIC	0	0	0	0.95	0	0	0.050
	NB $k = 8$	AIC	0	0	0	0.786	0.002	0	0.212
		GIC	0	0	0	0.792	0	0	0.206
		BIC	0	0	0	0.926	0.002	0	0.072
		ELCIC	0.002	0.002	0	0.940	0.002	0	0.054
	NB $k = 2$	AIC	0.002	0.002	0	0.592	0.006	0.002	0.396
		GIC	0.004	0.008	0	0.714	0.012	0.006	0.256
		BIC	0.004	0.008	0	0.774	0.012	0.002	0.200
		ELCIC	0.022	0.052	0	0.850	0.018	0.002	0.056
200	POISSON	AIC	0	0	0	0.830	0	0	0.170
		GIC	0	0	0	0.800	0	0	0.200
		BIC	0	0	0	0.978	0	0	0.002
		ELCIC	0	0	0	0.962	0	0	0.038
	NB $k = 8$	AIC	0	0	0	0.718	0	0	0.282
		GIC	0	0	0	0.776	0	0	0.224
		BIC	0	0	0	0.916	0	0	0.084
		ELCIC	0	0	0	0.952	0	0	0.048
	NB $k = 2$	AIC	0	0	0	0.562	0.002	0	0.436
		GIC	0	0	0	0.764	0.002	0	0.234
		BIC	0	0	0	0.814	0.002	0	0.184
		ELCIC	0.002	0	0	0.946	0.004	0	0.048
$n = 400$	POISSON	AIC	0	0	0	0.848	0	0	0.152
		GIC	0	0	0	0.810	0	0	0.190
		BIC	0	0	0	0.990	0	0	0.010
		ELCIC	0	0	0	0.990	0	0	0.010
	NB $k = 8$	AIC	0	0	0	0.746	0	0	0.254
		GIC	0	0	0	0.820	0	0	0.180
		BIC	0	0	0	0.938	0	0	0.062
		ELCIC	0	0	0	0.968	0	0.002	0.030
	NB $k = 2$	AIC	0	0	0	0.576	0	0	0.424
		GIC	0	0	0	0.772	0	0	0.228
		BIC	0	0	0	0.804	0	0	0.196
		ELCIC	0	0	0	0.980	0	0	0.020

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Table 2: Performance of ELCIC compared with QIC for the scenarios under longitudinal count data. 500 Monte Carlo data sets are generated with sample size $n = 100, 300$ and number of observations within-subject $T = 3, 5$. The model with $\{x_1, x_2\}$ and an exchangeable (EXC) correlation structure with the correlation coefficient $\rho = 0.5$ is the true model. AR1: auto-correlation 1; IND: independence; QIC/b: QIC with the BIC penalty.

Set-ups	Criteria		Candidate Models					
			x_1, x_2, x_3	$\boldsymbol{x_1, x_2}$	x_1, x_3	x_2, x_3	x_1	x_3
$n = 100$	ELCIC	EXC	0.040	0.844	0	0.002	0	0
$T = 3$		AR1	0.008	0.106	0	0	0	0
		IND	0	0	0	0	0	0
	QIC	EXC	0.090	0.494	0	0	0	0
		AR1	0.044	0.372	0	0	0	0
		IND	0	0	0	0	0	0
	QIC/b	EXC	0.012	0.570	0	0.002	0	0
		AR1	0.018	0.398	0	0	0	0
		IND	0	0	0	0	0	0
$n = 300$	ELCIC	EXC	0.026	0.958	0	0	0	0
$T = 3$		AR1	0	0.016	0	0	0	0
		IND	0	0	0	0	0	0
	QIC	EXC	0.078	0.574	0	0	0	0
		AR1	0.030	0.318	0	0	0	0
		IND	0	0	0	0	0	0
	QIC/b	EXC	0.012	0.640	0	0	0	0
		AR1	0.002	0.346	0	0	0	0
		IND	0	0	0	0	0	0
$n = 100$	ELCIC	EXC	0.052	0.946	0	0	0	0
$T = 5$		AR1	0	0.002	0	0	0	0
		IND	0	0	0	0	0	0
	QIC	EXC	0.102	0.834	0	0	0	0
		AR1	0.006	0.058	0	0	0	0
		IND	0	0	0	0	0	0
	QIC/b	EXC	0.016	0.920	0	0	0	0
		AR1	0	0.064	0	0	0	0
		IND	0	0	0	0	0	0
$n = 300$	ELCIC	EXC	0.02	0.98	0	0	0	0
$T = 5$		AR1	0	0	0	0	0	0
		IND	0	0	0	0	0	0
	QIC	EXC	0.098	0.894	0	0	0	0
		AR1	0.002	0.006	0	0	0	0
		IND	0	0	0	0	0	0
	QIC/b	EXC	0.008	0.984	0	0	0	0
		AR1	0	0.008	0	0	0	0
		IND	0	0	0	0	0	0

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0254

Table 3: Performance of ELCIC compared with the cross-validation (CV) method for the scenarios under longitudinal count data. 500 Monte Carlo data are generated with sample size $n = 100, 200$ and number of observations within-subject $T = 3$.

n	Criteria	MS	FP	OVS	OVOS	CS	JS	VOS
100	CV	0.017	1.710	0.225	0.775	–	–	–
	ELCIC_1	0.016	0.595	0.575	0.425	–	–	–
	ELCIC_2	0.016	0.610	0.565	0.435	0.91	0.52	0.39
200	CV	0.007	1.825	0.190	0.810	–	–	–
	ELCIC_1	0.007	0.405	0.655	0.345	–	–	–
	ELCIC_2	0.007	0.405	0.655	0.345	0.975	0.64	0.335

Note: MS: consistency $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$; FP: average number of falsely selecting nonzero variables; OVS: number of selecting the true mean structure/500; OVOS: number of over selecting the mean structure/500; CS: number of selecting the true correlation structure/500; JS: number of jointly selecting the true mean and the correlation structures/500; VOS: number of over selecting the mean structure under the true correlation structure selected/number of selecting the true correlation structure.