Statistica Sinica **32** (2022), 1-19
doi:https://doi.org/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] and Runze Li[3]

[1]*University of Maryland,* [2]*Pennsylvania State College of Medicine and* [3]*Pennsylvania State University*

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 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 (1994)), 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

Corresponding author: Ming Wang, Department of Public Health Science, Pennsylvania State College of Medicine, Hershey, PA 17033, USA. E-mail: muw22@psu.edu.