Statistica Sinica **33** (2023), 1-22
doi:https://doi.org/10.5705/ss.202020.0197

MODEL SELECTION OF GENERALIZED ESTIMATING EQUATION WITH DIVERGENT MODEL SIZE

Shicheng Wu[1], Xin Gao[1] and Raymond J. Carroll[2]

[1] *York University and* [2] *Texas A&M University*

Abstract: We consider the problem of model selection for a high-dimensional generalized estimating equation (GEE) in a marginal regression analysis for clustered or longitudinal data. Because the GEE method only makes assumptions about the first two moments, the full likelihood is not specified. Therefore, the likelihood-based model selection criteria cannot be applied directly. This paper introduces a generalized model selection criterion based on a quadratic form of the residuals. Using the large deviation result of the quadratic forms, we choose appropriate penalty terms on the model complexity. Lastly, we establish the model selection consistency of the proposed criterion for a divergent number of covariates.

Key words and phrases: Generalized estimation equation, generalized information criterion, large deviation, model selection consistency.

1. Introduction

With big data, model selection is essential to determine a subset of useful covariates. We consider the problem of model selection on generalized estimating equations (GEE) for clustered or longitudinal data. Because the full likelihood of multivariate clustered data is often difficult to specify, Liang and Zeger (1986) extended the generalized linear models (McCullough and Nelder (1989)) to include correlated data, thus proposing the GEE. The GEE estimate is consistent, even when the working correlation matrix is misspecified. Li (1997) investigated the consistency of the GEE using a minimax approach. Xie and Yang (2003) established a more comprehensive large-sample theory for the GEE, including consistency and asymptotic normality. Balan and Schiopu-Kratina (2005) provide a rigorous study on the GEE under a pseudo-likelihood framework. These works all assume that the number of covariates p is fixed, and that the number of clusters n goes to infinity. Recently, a great amount of work has been devoted to high-dimensional data analysis; see Donoho (2000), Fan and Li (2001), Fan and Lv (2008), and Lv and Fan (2009) for a comprehensive review.

Corresponding author: Xin Gao, Department of Mathematics and Statistics York University, Toronto, ON M3J 1P3, Canada. E-mail: xingao@mathstat.yorku.ca.