Accepted Manuscript

Automatic variable selection for longitudinal generalized linear models

Gaorong Li, Heng Lian, Sanying Feng, Lixing Zhu

PII: S0167-9473(12)00449-5
DOI: 10.1016/j.csda.2012.12.015
Reference: COMSTA 5470

To appear in: *Computational Statistics and Data Analysis*

Received date: 3 October 2011
Revised date: 2 November 2012
Accepted date: 22 December 2012

Please cite this article as: Li, G., Lian, H., Feng, S., Zhu, L., Automatic variable selection for longitudinal generalized linear models. *Computational Statistics and Data Analysis* (2012), doi:10.1016/j.csda.2012.12.015

Automatic Variable Selection for Longitudinal Generalized Linear Models

Gaorong Li[*,a], Heng Lian[b], Sanying Feng[a], Lixing Zhu[c]

[a] *College of Applied Sciences, Beijing University of Technology, Beijing 100124, P. R. China*
[b] *Division of Mathematical Sciences, SPMS, Nanyang Technological University, Singapore*
[c] *Department of Mathematics, Hong Kong Baptist University, Hong Kong, P. R. China*

Abstract

We consider the problem of variable selection for the generalized linear models (GLMs) with longitudinal data. An automatic variable selection procedure is developed using smooth-threshold generalized estimating equations (SGEE). The proposed procedure automatically eliminates inactive predictors by setting the corresponding parameters to be zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE. The proposed method shares some of the desired features of existing variable selection methods: the resulting estimator enjoys the oracle property; the proposed procedure avoids the convex optimization problem and is flexible and easy to implement. Moreover, we propose a penalized weighted deviance criterion for a data-driven choice of the tuning parameters. Simulation studies are carried out to assess the performance of SGEE, and a real dataset is analyzed for further illustration.

Key words: Generalized linear model; Longitudinal data; Automatic variable selection; Generalized estimating equations; Oracle property

1. Introduction

Generalized linear models (GLMs, McCullagh and Nelder, 1989) extend the framework of linear models, by allowing for non-Gaussian data and nonlinear link functions. They have become a favored tool for modelling clustered and longitudinal data, in particular, for repeated or correlated non-Gaussian data, such as binomial or Poisson type response that is commonly encountered in longitudinal studies. The generalized estimating equations (GEE) method was introduced in a seminal paper of Liang and Zeger (1986) as a useful extension of GLMs to correlated data, and has also become a very popular estimation method.

In the present paper, we consider the marginal longitudinal GLMs. Suppose that $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{im_i})^T$ is the multivariate response for the ith subject, and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im_i})^T$ is the $m_i \times p$ matrix of the covariates for the ith subject ($i = 1, \ldots, n$). Observations from different subjects are independent; but those from the same subjects are correlated.

[*] Corresponding author. *E-mail address:* ligaorong@gmail.com (Gaorong Li).
E-mail addresses: HengLian@ntu.edu.sg (Heng Lian), fsy5801@sina.com (Sanying Feng), lzhu@hkbu.edu.hk (Lixing Zhu)