Applied Mathematics and Computation 245 (2014) 343–356

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Robust variable selection in semiparametric mean-covariance regression for longitudinal data analysis

Chaohui Guo *, Hu Yang, Jing Lv

[a] College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China

ARTICLE INFO

Keywords:
B-spline
Generalized estimating equations
Longitudinal data
Modified Cholesky decomposition
Partial linear models
Robustness

ABSTRACT

This paper considers robust semiparametric smooth-threshold generalized estimating equations for the analysis of longitudinal data based on the modified Cholesky decomposition and B-spline approximations. The proposed method can automatically eliminate inactive predictors by setting the corresponding parameters to be zero, and simultaneously estimate the mean regression coefficients, generalized autoregressive coefficients and innovation variances. In order to overcome the outliers in either the response or/and the covariate domain, we use a bounded score function and leverage-based weights to achieve better robustness. Moreover, the proposed estimators have desired large sample properties including consistency and oracle property. Finally, Monte Carlo simulation studies are conducted to investigate the robustness and efficiency of the proposed method under different contaminations.

1. Introduction

Longitudinal data sets are common in medical and public health studies, the significant feature of it is repeated measures over a certain period of time. Liang and Zeger [12] proposed generalized estimating equations (GEEs) which are taken as a milestone in the development of methodology for longitudinal data analysis. Partial linear models (PLMs) include a nonparametric function and linear part, which are more flexible and more interpretable than linear models and nonparametric models respectively. Wang et al. [24] considered marginal generalized semiparametric partially linear models for clustered data and proposed profile-type estimating equation, which differs from a standard marginal generalized estimating equation model [12] mainly through introducing the nonparametric component. However, these methods mentioned above are not robust, since its estimates are sensitive to potential outliers and influential observations. Hence, to seek a more robust method against outliers is a very important issue in longitudinal studies. In recent years, many authors studied the influence of outliers to the estimate and had developed many robust methods. An incomplete list of recent works on the robust GEE method include [4,6,9,18,19] and so on.

As far as we know, a better estimate for the covariance matrix will result in a better estimate for the mean parameter. But all methods above mainly payed attention to the estimate of mean parameters while regarded the covariance parameters as nuisance parameters. In fact, the covariance parameters may be not nuisance parameters and have substantive significance for its own interest, see Carroll [1] and Zhang and Li [29] for reference. Recently, motivated by the modified Cholesky decomposition, Ye and Pan [28] proposed joint mean and covariance regression models by using generalized estimating equations. The advantages of this decomposition are that it makes covariance matrices to be positive definite and the parameters in it have well-founded statistical concepts. Due to these advantages, Cholesky decomposition has received considerable

* Corresponding author.
E-mail address: guochaohui2010@126.com (C. Guo).

http://dx.doi.org/10.1016/j.amc.2014.07.086
0096-3003/

attention in the literature. Here we only list a few. See Leng et al. [10] and Mao et al. [13] constructed PLMs for the mean and the covariance structure for longitudinal data, which are more flexible than that of Ye and Pan [28]. However, the above approaches based on GEEs are also highly sensitive to outliers in the sample. Recently, Zheng et al. [31] established three robust estimating equations to hinder the effect of outliers in both mean and covariance estimation by borrowing the idea of [6]. However, as far as we know, there is little discussion on robust estimation on the semiparametric mean-covariance model (SMCM). In this paper, we consider the SMCM and decompose the inverse of covariance matrix by the modified Cholesky decomposition. The entries in this decomposition are autoregressive parameters and log innovation variances. See [10,28,31,32] for references. Zhang and Leng [30] proposed a new regression model to decompose covariance structures by using a novel Cholesky factor with the entries being a moving average and log innovation variances. These decompositions are based on linear time series analysis. Thus, the application of semiparametric mean-covariance regression with longitudinal data may be extended to the realm of nonlinear time series analysis, since the derivation of methods for nonlinear time series analysis acts as one of the crowning achievements that emerged from the theory of deterministic dynamical systems. Kodba et al. [8] and Perc [15] applied nonlinear time series analysis methods to analyse the chaotic behaviour of a very simple periodically driven resistor–inductor diode circuit and dynamics of human gait respectively and pointed out that the nonlinear time series analysis methods are superior to mathematical modelling, because of they can introduce basic concepts directly from the experimental data.

In recent years, many penalization or shrinkage based variable selection methods have been developed to select significant variables among all the candidate variables, e.g., [3,33–35], etc. All the methods mentioned above only consider the independent data. But variable selection is also a fundamentally important issue in longitudinal studies, which could greatly enhance the prediction performance of the fitted model and select significant variables. In recent years, penalty function had been adopted to select active variables in the longitudinal data analysis. For example, Fan et al. [4] developed penalized robust estimating equations for longitudinal linear regression models with the SCAD penalty. Wang et al. [23] considered the SCAD-penalized GEE for analyzing longitudinal data with high-dimensional covariates. Zheng et al. [32] considered robust variable selection method in joint mean and covariance models through using three penalized robust generalized estimating equations with three SCAD penalties. All these procedures are based on penalty functions to select variables, which are singular at zero. Consequently, these variable selection procedures need to solve the convex optimization which lead to a computational burden. Borrowing the idea of Ueki [22], Li et al. [11] developed the smooth-threshold generalized estimating equations (SGEEs) for longitudinal generalized linear models which can efficiently avoid the convex optimization problem. These facts motivate us to develop robust smooth-threshold generalized estimating equations for the SMCM with longitudinal data. This paper has made the following contributions: (i) we establish consistency and asymptotic normality of the mean regression coefficients, generalized autoregressive coefficients and innovation variances, and obtain the optimal convergent rate for estimating the nonparametric functions. (ii) The proposed method can alleviate the effect of outliers in either the response or/and the covariate domain by using the bounded Huber's score function and Mallows weights. (iii) The proposed method can automatically eliminate inactive predictors by setting the corresponding parameters to be zero and estimate nonzero coefficients through semiparametric smooth-threshold generalized estimating equations.

The rest of the article is organized as follows. Section 2 introduces the model and estimation method. Theoretical properties of the proposed estimators are also given in this Section. Section 3 describes the semiparametric smooth-threshold generalized estimating equations and oracle property. In Section 4, an efficient algorithm is proposed to implement the procedures. Moreover, we discuss how to select the tuning parameters so that the corresponding estimators are robust and sparse. Simulation studies are carried out in Section 5 to investigate the performance of the proposed estimators under two types of contaminations. Some concluding remarks are given in Section 6. All the proofs are relegated to the Appendix.

2. Robust generalized estimating equations in joint mean and covariance semiparametric models

2.1. The joint mean and covariance semiparametric models

For a vector or a matrix $\boldsymbol{A}$, we define $\boldsymbol{A}$ as random vector or matrix of population, $\boldsymbol{A}_i$ as the ith corresponding sample and $\boldsymbol{A}_{0i}$ as the true value of $\boldsymbol{A}_i$ throughout our paper. In this paper, we consider an experiment with m subjects and n_i observations over time for the ith subject, where $n=\sum_{i=1}^{m} n_i$ is the total of observations. Suppose that $\boldsymbol{y}_i=\left(y_{i1},\ldots,y_{in_i}\right)^T$ is the response for the ith subject $(i=1,\ldots,m)$ at time points $\boldsymbol{t}_i=(t_{i1},\ldots,t_{in_i})$. We assume that t_{ij} is the time or any time-dependent covariate and $\{t_{ij}\}\in[0,1]$. We further specify a marginal model through defining the first two moments of the response y_{ij}, i.e., $E(y_{ij}|\boldsymbol{x}_{ij},t_{ij})=\mu_{0ij}, V(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{t}_i)=\boldsymbol{\Sigma}_{0i}$, where $\boldsymbol{x}_{ij}$ is a p-dimension covariate vector and $\boldsymbol{x}_i=(\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i})^T$ is a corresponding covariate matrix for the ith subject. On account of the modified Cholesky decomposition, there exists a lower triangle matrix $\boldsymbol{\Phi}$ with 1's on the main diagonal satisfying $\boldsymbol{\Phi}_i\boldsymbol{\Sigma}_{0i}\boldsymbol{\Phi}_i^T=\boldsymbol{D}_{0i}$, where $\boldsymbol{D}_{0i}$ is a diagonal matrix with positive entries and the lower-diagonal entries of $\boldsymbol{\Phi}_i$ are defined as the negatives of the autoregressive coefficients ϕ_{ijk} of

$$y_{ij}-\mu_{ij}=\sum_{k=1}^{j-1}\phi_{ijk}(y_{ik}-\mu_{ik})+\varepsilon_{ij}. \tag{1}$$

The diagonal entries of $\boldsymbol{D}_{0i}$ are taken as the innovation variances with entries $\sigma_{0ij}^2=\mathrm{var}(\varepsilon_{ij})$, where $\varepsilon_{ij}=y_{ij}-\hat{y}_{ij}$. The modified Cholesky decomposition makes $\boldsymbol{\Sigma}_{0i}$ to be positive definite, and the parameters ϕ and $\log(\sigma_{ij}^2)$ to be unconstrained. Based on