Journal of Multivariate Analysis 124 (2014) 199–213

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Model determination and estimation for the growth curve model via group SCAD penalty

Jianhua Hu [a,*], Xin Xin [a,b], Jinhong You [a]

[a] School of Statistics and Management, and Key Laboratory of Mathematical Economics, Shanghai University of Finance and Economics, Shanghai 200433, PR China

[b] School of Mathematics and Information Sciences, Henan University, Kaifeng, Henan 475000, PR China

ARTICLE INFO

Article history:
Received 11 March 2013
Available online 11 November 2013

AMS 2000 subject classifications:
primary 62H12
secondary 62F12
62H07

Keywords:
Degree of polynomial profile form
Growth curve model
Least squares estimation
Oracle property
SCAD penalty
Three-level variable selection

ABSTRACT

The growth curve model is a useful tool for studying the growth problems, repeated measurements and longitudinal data. A key point using the growth curve model to fit data is determining the degree of polynomial profile form, choosing suitable explanatory variables, shrinking some regression coefficients to zero and estimating nonzero regression coefficients. In this paper, we propose a three-level variable selection approach based on weighed least squares with group SCAD penalty to handle the aforementioned problems. Considering the rows and columns of regression coefficient matrix as groups with overlap to control the polynomial order and variables, respectively, our proposed procedure enables us to simultaneously determine the degree of polynomial profile, identify the significant explanatory variables and estimate the nonzero regression coefficients. With appropriate selection of the tuning parameters, we establish the oracle property of the procedure and the consistency of the proposed estimation. We investigate the finite sample performances of our procedure in simulation studies whose results are very supportive, and also analyze a real data set to illustrate the usefulness of our procedure.

1. Introduction

A growth curve model was first summarized by Potthoff and Roy [21], and studied subsequently by many authors including [9,12,14,16,17,19,22,25]. It is already shown to be very useful, particularly for studying growth problems on short time series, repeated observations often measured over multiple time points on a particular characteristic to investigate the temporal pattern of change on the characteristic [4] and longitudinal data especially with serial correlation [15] in a variety of scientific disciplines, such as medical research, biology, economics, education, forestry and so on. The interested reader can refer to [18,20] for a more detailed discussion and illustration of the usefulness of the growth curve model.

The basic idea of the growth curve model is to introduce some known functions, usually polynomial functions, so as to capture patterns of change for time-dependent measurements. In data analysis, the degree of polynomial profile is often unknown. This allows the possibility of selecting an underfitted (or overfitted) model, leading to biased (or inefficient) estimators and predictions. Meanwhile, if there are too many explanatory variables which are not important, we need to choose suitable explanatory variables for efficiency and accuracy. Therefore, a key point using the growth curve model to fit the data is simultaneously determining the degree of polynomial profile form, selecting suitable explanatory variables, choosing zero regression coefficients and estimating the nonzero regression coefficients.

Usually, a growth curve model can be written as

$$\mathbf{Y} = X\Theta Z^{\tau} + \mathcal{E} \tag{1.1}$$

* Corresponding author.
E-mail address: frank.jianhuahu@gmail.com (J. Hu).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.11.001

where $\mathbf{Y}$ is the observation matrix of the response consisting of p repeated measurements taken on n individuals, X is the treatment design matrix with order $n\times m$, Z is the profile matrix with order $p\times q$, and Θ is the unknown regression coefficient matrix with order $m \times q$. Assume that observations on individuals are independent, so that the rows of the random error matrix $\mathcal{E}$ are independent and identically distributed by a distribution with mean zero and covariance matrix Σ.

The treatment design matrix X discussed in this paper is assumed to comprise of intercept items and explanatory variables. The explanatory variables themselves may be continuous or categorical. Data sets with continuous explanatory variables often appear in real statistical problems. For example, in the problem considering children's reading recognition, cognitive stimulation and emotional support for children at home will be possible explanatory variables which may be continuous.

Several authors have investigated the problem of determining the degree of polynomial profile form in the model (1.1). Fujikoshi and Rao [8] considered the one of selecting the covariables in the growth curve model, but did not consider the problem of selecting the degree of freedom. Satoh et al. [23] treated determination of the degree of a polynomial growth curve as a problem of variable selection. They proposed the corresponding C_p and AIC for the situation where the matrix X is a design matrix across individuals on m groups. Here it is not necessary to select the columns of X though the criteria in [23] is also used to determinate the degree of polynomial profile and select the explanatory variables simultaneously.

It is well known that the traditional variable selections such as C_p, AIC and Bayesian information criterion (BIC) do not shrink the regression coefficients to zero. This results in the accuracy of the regression coefficient estimators in the final model being somewhat difficult to understand.

To overcome the limitation, we consider an alternative method, to the traditional best subset variable selections, that can simultaneously estimate regression coefficients and shrink some regression coefficients to zero, thereby, removing them from the final model. For linear regression models, Fan and Li [6] novelly proposed a family of variable selection procedures by the smoothly clipped absolute deviation penalty (SCAD). They showed that the proposed method outperforms the best subset variable selection in terms of computational cost and stability. An attractive feature of it is that with a proper choice of regularization parameters, the resulting estimator possesses an oracle property, namely, the true zero regression coefficients are automatically estimated as zero, and the remaining coefficients are estimated as well as if the correct submodel were known in advance.

To simultaneously choose the degree of polynomial profile and explanatory variables, we shall propose a three-level variable selection procedure based on group SCAD penalty and weighed least squares so that we can simultaneously determine the order of polynomial profile form, identify the significant explanatory variables, shrink some regression coefficients to zero and estimate nonzero regression coefficients.

Our contributions are to divide the rows and columns of regression parametric matrix in the model (1.1) into column and row groups with a special overlap, discover the equivalent relationships between row group and variable selection and between column group and determine of the order of polynomial profile form, then use a smoothly clipped absolute deviation penalty function to penalize row groups, column groups and regression coefficients, and finally minimize the penalized (weighted) least squares with three-level SCAD penalty to lead to a penalized (weighted) least squares estimator of the regression parametric matrix. It will be shown that the oracle property of the procedures and the consistency of the estimation can be achieved under an appropriate selection of the tuning parameters. Our results indicate that our proposed procedure outperforms the best subset selection and the method that SCAD is directly applied to the vector version of model (1.1).

The rest of this paper is as follows. A least squares estimation via a group SCAD penalty with a special overlap is proposed in Section 2. The estimation of the error covariance matrix is provided in Section 3. Based on the estimated covariance, we proposed a three-level variable selection procedure via weighted least squares estimation with a group SCAD penalty in Section 4. The selection of the tuning parameters is discussed in Section 5. Some simulation studies are conducted in Section 6. One practical problem, which can be characterized by the model (1.1), is illustrated in Section 7. Finally, the brief concluding remarks are stated in Section 8. The proof of the main results are collected in the Appendix.

2. Least squares estimation with group SCAD penalty

For convenience of notation, we write Θ into $\Theta = (\Theta_{ij})_{2\times 2}$ with Θ_{11} a $m_0 \times q_0$ matrix. Here $q_0(\leq q)$ is the true degree of polynomial profile form and $m_0(\leq m)$ is the true number of explanatory variables. Let $\boldsymbol{\theta}_{\cdot j}$ denote the jth $(1 \leq j \leq q)$ column of matrix Θ and $\boldsymbol{\theta}_{i\cdot}$ denote the ith $(1 \leq i \leq m)$ row of matrix Θ. There is an $m_0q_0 \times m_0q_0$ elementary transformation matrix $L_0 = (L_{10}, L_{20})$ such that $\mathrm{vec}(\Theta_{11}) = L_{10}\boldsymbol{\beta}_{10} + L_{20}\boldsymbol{\beta}_{20}$, where $\boldsymbol{\beta}_{10}$ is a d_0-dimensional vector with each element nonzero and $\boldsymbol{\beta}_{20} = \mathbf{0}$. The following is an example with

$$\Theta_{11} = \begin{pmatrix} \theta_{11} & 0 \\ 0 & \theta_{22} \\ \theta_{31} & \theta_{32} \end{pmatrix}, \quad L_{10} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad L_{20} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{\beta}_{10} = \begin{pmatrix} \theta_{11} \\ \theta_{22} \\ \theta_{31} \\ \theta_{32} \end{pmatrix}, \quad \boldsymbol{\beta}_{20} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Here the *vec operator* transforms a matrix into a vector by *stacking the rows of the matrix one under another*.