Journal of Multivariate Analysis 139 (2015) 347–359

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A two-step estimation method for grouped data with connections to the extended growth curve model and partial least squares regression

Ying Li [a,*], Peter Udén [a], Dietrich von Rosen [a,b]

[a] Swedish University of Agricultural Sciences, Uppsala, Sweden
[b] Linköping University, Linköping, Sweden

ARTICLE INFO

Article history:
Received 17 June 2014
Available online 17 April 2015

AMS subject classifications:
62H12
62J07
62H99

Keywords:
Extended growth curve model
Grouped data
Krylov space
PLS
Two-step method

ABSTRACT

In this article, the two-step method for prediction, which was proposed by Li et al. (2012), is extended for modelling grouped data, which besides having near-collinear explanatory variables, also having different mean structure, i.e. the mean structure of some part of the data is more complex than other parts. In the first step, inspired by partial least squares regression (PLS), the information for explanatory variables is summarized by a multilinear model with Krylov structured design matrices, which for different groups have different size. The multilinear model is similar to the classical growth curve model except that the design matrices are unknown and are functions of the dispersion matrix. Under such a multilinear model, natural estimators for mean and dispersion matrices are proposed. In the second step, the response is predicted through a conditional predictor where the estimators obtained in the first step are utilized.

1. Introduction

One common problem in statistical analysis is to build relationships among a set of explanatory variables $\boldsymbol{x}$ and a response variable y. To apply ordinary least squares (OLS) theory is one of the classical choices. However, the OLS predictor does not work well in prediction of new observations when the explanatory variables are near-collinear. It may be due to the unrealistic large $\boldsymbol{\Sigma}$ and the unstable estimator of $\boldsymbol{\Sigma}^{-1}$ in the OLS coefficient, i.e. $\boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}$ is the variance of $\boldsymbol{x}$ and $\boldsymbol{\omega}$ is the covariance between $\boldsymbol{x}$ and y. Therefore, several methods have over the years been proposed which intend to construct new estimators using an approximation of $\boldsymbol{\Sigma}^{-1}$. For example, ridge regression uses $(\boldsymbol{\Sigma} + k\boldsymbol{I})^{-1}$ to approximate $\boldsymbol{\Sigma}^{-1}$ where k is a parameter needed to be estimated and $\boldsymbol{I}$ is the identity matrix. Principal component regression (PCR) uses the largest eigenvalue and the corresponding eigenvectors of $\boldsymbol{\Sigma}$ to approximate $\boldsymbol{\Sigma}^{-1}$. Partial least squares regression (PLS) can be considered as approximating $\boldsymbol{\Sigma}^{-1}$ by a polynomial sequence of $\boldsymbol{\Sigma}$, which generates a Krylov space, i.e. the linear space generated by $(\boldsymbol{\Sigma\omega} : \boldsymbol{\Sigma}^2\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^i\boldsymbol{\omega})$.

In Li and von Rosen [7]'s paper, a two-step method for prediction was developed. The main purpose of the paper was to link PLS with classical multivariate linear models theory. In the first step, which could be considered as a dimension reduction step, a multivariate linear model was applied to summarize the information in the explanatory variables. Inspired by PLS, a Krylov structured matrix was used as design matrix in the linear model. In the second step, the prediction step, a conditional approach was applied. Concerning dimension reduction, it is worth mentioning that Cook and coworkers (see Cook et al. [2])

* Corresponding author.
E-mail address: Ying.Li@slu.se (Y. Li).

http://dx.doi.org/10.1016/j.jmva.2015.03.011
0047-259X/

develop envelope models, a general methodology for model reduction in prediction problems, which has a close connection to PLS, i.e. a proper version of the developed method is identical to the population version of PLS (see Cook et al. [1]).

Compared with envelope models, the two-step method in this article is more specific in the use of the Krylov design matrix. The two-step method in a special case is identical to the population version of PLS. In comparison with PLS, the two-step method is non-algorithmic and gives higher flexibility to model different types of structures, for example, different groups.

It is common that data due to experimental conditions consist of different treatment groups, e.g. gender and season are factors which form groups. Treatment group effects have a variety of forms. For example, it may be expressed as having different mean levels among groups. In other cases, data from different groups may have different mean structures e.g. linear in one group and non-linear in another group. An example is the silage data set discussed in Li et al. [8] which is a data set with an inherent group effect. The study aims to determine the chemical compounds in silage by Fourier mid-infrared spectroscopy. Data was collected over several years from different experiments, which comprised a variety of previous ensiling studies with objective to investigate crop type, harvesting date, silage additives, etc. It is natural that the chemical compounds e.g. sugar content is high in one experiment, but relatively low in another experiment, due to different crop types. In addition, part of the data from special experiments has silage additive compounds, e.g. ammonia, which do not exist in other experiments. The chemical compounds from different groups (experiments) do not only have different levels, but also different structures. As a result, a non-standard model is needed to fit a part of the data compared to other parts. If there is no near-collinear structure in the data and enough independent observations, it is appropriate to apply the extended growth curve model, which will be explained later. If using classical PLS, it is likely that one has to model the groups separately, since the classical PLS is not designed for grouped data. In the literature, there are some other versions of PLS which are proposed to handle grouped data such as least square PLS [4], sequential and orthogonalized PLS [9]. In this paper, the two-step approach in Li and von Rosen [7] will be extended to deal with group structures.

The basic model and the two-step method are explained in Section 2. In Section 3, the extended two-step method is stated. The main result of the paper appears in Section 4, where the estimators of mean and variance parameters are derived. Finally, we remark that the main aim of the present paper is not to improve PLS. The focus is instead to combine PLS thinking with multivariate linear models through Krylov spaces which naturally appear when approximating $\boldsymbol{\Sigma}$.

2. Background

2.1. Basic model

Let $\boldsymbol{y}$ be a k-dimensional random vector and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$ be a $p \times k$-random matrix, jointly normally distributed with $E[\boldsymbol{X}] = \boldsymbol{\mu}_{xc} = (\boldsymbol{\mu}_{x1}, \boldsymbol{\mu}_{x2}, \ldots, \boldsymbol{\mu}_{xk})$ and $E[\boldsymbol{y}'] = \boldsymbol{\mu}_{yc} = (\mu_{y1}, \mu_{y2}, \ldots, \mu_{yk})$ and $D[\boldsymbol{X}] = \boldsymbol{I}_k \otimes \boldsymbol{\Sigma}$, $C[\boldsymbol{X}, \boldsymbol{y}] = \boldsymbol{I}_k \otimes \boldsymbol{\omega}$, i.e.

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{y}' \end{pmatrix} \sim N_{(p+1),k}\left(\begin{pmatrix} \boldsymbol{\mu}_{xc} \\ \boldsymbol{\mu}_{yc} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\omega} \\ \boldsymbol{\omega}' & \sigma_y^2 \end{pmatrix}, \boldsymbol{I}_k \right), \tag{2.1}$$

where $N_{q,r}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$ stands for the matrix normal distribution. The model in (2.1) usually is used for data with group effects, i.e. data from k different groups share a common covariance structure, but have different means. If $k = 1$, there is no group effect involved in the data and the model in (2.1) becomes

$$\begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} \sim N_{(p+1)}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\omega} \\ \boldsymbol{\omega}' & \sigma_y^2 \end{pmatrix} \right). \tag{2.2}$$

2.2. The two-step method

Based on the model in (2.2), the two-step method for prediction is formulated as follows: let $\boldsymbol{\omega}$ be known,

(i) suppose $\mathbf{x} = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Sigma})$ and γ is an unknown parameter,

(ii) predict via the conditional expectation: $\hat{y} = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}_x}) + \mu_y$.

Via Step (i), information from $\mathbf{x}$ is summarized i.e. $\widehat{\boldsymbol{\mu}_x}$ is obtained by deriving an estimator of γ. Step (ii) is the prediction step, i.e. $\hat{y}$ is predicted through $\widehat{\boldsymbol{\mu}_x}$ and $\widehat{\boldsymbol{\Sigma}}$. However, the major problem is to estimate $\boldsymbol{\Sigma}^{-1}$ in Step (ii). If predictors are collinear in the data, then the usual estimators, e.g. a moment estimator, will being singular or close to be singular leading to a poor predictor. Based on the Cayley–Hamilton theorem, $\boldsymbol{\Sigma}^{-1}$ can be presented in a polynomial form, i.e. $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} c_i \boldsymbol{\Sigma}^{i-1} \approx \sum_{i=1}^{a} c_i \boldsymbol{\Sigma}^{i-1}$, for some constants c_i and $a \leq p$. In theory, c_i is a function of $\boldsymbol{\Sigma}$. However, here we treat it as an unknown constant needed to be estimated. Then,

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\sum_{i=1}^{p} c_i \boldsymbol{\Sigma}^{i-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} \\ &\approx \sum_{i=1}^{a} \boldsymbol{\Sigma}^{i}\boldsymbol{\omega}(c_i\gamma) + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\boldsymbol{G}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.3}$$