Statistical Methodology 9 (2012) 353–363

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

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

Constrained estimation and some useful results in several multivariate models

Sévérien Nkurunziza

University of Windsor, Department of Mathematics and Statistics, Sunset Avenue, Windsor, Ontario, N9B 3P4, Canada

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form
19 August 2011
Accepted 14 September 2011

Keywords:
Asymptotic distributional risk
Constrained estimation
Multivariate regression
QMLE
Shrinkage estimator

ABSTRACT

In this paper, we are interested in an estimation problem concerning the regression coefficient parameter matrices of M independent multivariate multiple linear models. More specifically, we consider the case where the M parameter matrices are suspected of satisfying some restrictions. Given such uncertainty, we study a class of shrinkage estimators which give an improvement over the performance of the quasi-maximum likelihood estimator (QMLE). To this end, we derive a theorem which is useful in establishing the asymptotic distributional risk function of a class of shrinkage estimators of the regression coefficient parameter matrices.

1. Introduction

In statistical modeling, it is common to model simultaneous influence of several covariates on a response variable. In particular, the statistical literature shows that the multiple regression technique is an extremely powerful methodology. Further, in order to model the influence of the same set of explanatory variables on several correlated responses, the multivariate multiple regression model (MMRM) seems to be the most appropriate methodology. Indeed, there are diverse situations where the correlation of the response variables is quite high. For example, systolic blood and diastolic blood pressures are positively correlated as is the number of cavities in the upper jaw and the lower jaw. In addition, classification and discrimination analysis are other common procedures used when several correlated responses occur. Some of the recent advances in MMRM are applied to artificial intelligence and machine learning theory, as described, for example, in [7]. Further, [5] suggested a shrinkage multivariate least squares estimator through canonical analysis to utilize the relationship of response variables.

E-mail address: severien@uwindsor.ca.

1572-3127/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2011.09.002

In this paper, we consider M independent $n \times m$-random matrices $\boldsymbol{Y}^{(i)}$, $\boldsymbol{U}^{(i)}$, for $i = 1, 2, \ldots, M$ which satisfy multivariate multiple linear model

$$\boldsymbol{Y}^{(i)} = \boldsymbol{X}^{(i)}\boldsymbol{\theta}^{(i)} + \boldsymbol{U}^{(i)} \tag{1.1}$$

where $\boldsymbol{X}^{(i)}$ is $n \times k$ random matrix, and $\boldsymbol{\theta}^{(i)}$ is a $k \times m$ parameter matrix. For each $i = 1, 2, \ldots, M$, we assume that the random matrices $\boldsymbol{Y}^{(i)}$ and $\boldsymbol{X}^{(i)}$ are observed and the matrix $\boldsymbol{U}^{(i)}$ is the unobserved noise. For concreteness, we consider the following motivating examples.

Motivating example: The motivating example is based on the data set found at http://lib.stat.cmu.edu/datasets/Plasma_Retinol, and consists of 315 observations. Three groups (i.e. $M = 3$) are considered. Namely, we have 122 who use vitamins fairly often, 82 use vitamins but not often, and 111 who never use vitamins. Thus, $n_1 = 122$, $n_2 = 82$ and $n_3 = 111$. In this data set, we also have $k = 12$ explanatory variables: $X_1 \equiv$ *AGE* (years), $X_2 \equiv$ *SEX* (1 = Male, 2 = Female), $X_3 \equiv$*SMOKSTAT*$_1$: Smoking status (1 = Never), $X_4 \equiv$ *SMOKSTAT*$_2$: Smoking status (1 = Former), $X_5 \equiv$ *QUETELET* (weight/(height 2)), $X_6 \equiv$ *CALORIES*: Number of calories consumed per day, $X_7 \equiv$ FAT: Grams of fat consumed per day, $X_8 \equiv$ *FIBER*: Grams of fiber consumed per day, $X_9 \equiv$ *ALCOHOL*: Number of alcoholic drinks consumed per week, $X_{10} \equiv$ *CHOLESTEROL*: Cholesterol consumed (mg/day), $X_{11} \equiv$ *BETADIET*: Dietary beta-carotene consumed (mcg per day), $X_{12} \equiv$ *RETDIET*: Dietary retinol consumed (mcg/day). Here, $m = 2$ and, we have $\boldsymbol{Y}^{(1)}$ which is a 122×2 matrix, $\boldsymbol{Y}^{(2)}$ which is a 82×2 matrix, and $\boldsymbol{Y}^{(3)}$ which is a 111×2 matrix. Further, the first column of $\boldsymbol{Y}^{(i)}$, $i = 1, 2, 3$ consists of the quantity (in ng/ml) of Plasma beta-carotene (*BETAPLASMA*), and the second column is the quantity (in ng/ml) of Plasma Retinol (*RETPLASMA*).

It is reasonable to assume that the 3 samples are independent as they correspond to 3 different groups. Assumption $(\mathcal{A}_5)$ given in Section 2 is the mathematical interpretation of this realistic assumption. However, for each sample, the response variables are likely correlated, which justifies the need for multivariate regression. In other words, within the ith group ($i = 1, 2, 3$), the two columns of $\boldsymbol{Y}^{(i)}$ are likely correlated.

Remark 1.1. Another application context of the proposed method, related to the above motivating example, consists of the scenario where there are M geographical areas for which, for example, the same number of observations were randomly selected. Also, the level of using vitamins can be analyzed by creating two categorical variables: VITUSE$_1$ = 1 if vitamins are used often, 0 otherwise; and VITUSE$_2$ = 1 if vitamins are used but not often, 0 otherwise. Once again, in this context, it is reasonable to assume that the M samples are independent as they correspond to M different geographical areas. However, within the ith geographical area, ($i = 1, 2, \ldots, M$), the two columns of $\boldsymbol{Y}^{(i)}$ are more likely correlated.

In this paper, we are interested in estimating the parameter matrices $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, \ldots, M$ when matrices $\boldsymbol{\theta}^{(i)}$ are suspected to lie in a certain candidate subspace of dimension $q < m$, as for example,

$$\boldsymbol{L}_1\boldsymbol{\theta}^{(1)}\boldsymbol{L}_2 = \boldsymbol{L}_1\boldsymbol{\theta}^{(2)}\boldsymbol{L}_2 = \cdots = \boldsymbol{L}_1\boldsymbol{\theta}^{(M)}\boldsymbol{L}_2, \tag{1.2}$$

where $\boldsymbol{L}_1$ is a given $q \times k$-matrix of full rank with $k < q$ and $\boldsymbol{L}_2$ is a known $m \times r$ full rank matrix with $r \leqslant m$. In the sequel, for the sake of simplicity, we consider the case where $n_1 = n_2 = \cdots = n_M = n$.

It should be noticed that the subspace candidate in (1.2) extends that of [7, p. 168]. Thus, the restriction in (1.2) is useful, for example, in model assessment and variable selection, and in profile analysis. Also, it is useful in economical modeling where, for instance, different groups of countries decide to unify their economic policies, as is the case for the European Union countries. Indeed, since the economic policies of the united countries are supposed to be harmonized, it is reasonable to suspect homogeneity of their economic indicators. In this context, the parameter matrices $\boldsymbol{\theta}^{(i)}$, $i = 1, 2, \ldots, M$ would satisfy some restrictions as in (1.2).

To simplify the notation, let $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)'}, \boldsymbol{\theta}^{(2)'}, \ldots, \boldsymbol{\theta}^{(M)'}\right)'$, and let $\boldsymbol{A} \otimes \boldsymbol{B}$ stand for the Kronecker product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Then, the restriction in (1.2) can be rewritten as $(\boldsymbol{L}_3 \otimes \boldsymbol{L}_1)\,\boldsymbol{\theta}\boldsymbol{L}_2 = \boldsymbol{0}$,