Computational Statistics and Data Analysis 92 (2015) 68–83

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and inference on central mean subspace for multivariate response data

Liping Zhu [a], Wei Zhong [b,*]

[a] Institute of Statistics and Big Data, Renmin University of China, 59 Zhongguancun Street, Haidian District, Beijing, 100872, China

[b] Wang Yanan Institute for Studies in Economics (WISE), Department of Statistics, School of Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, 422 Siming South Road, Xiamen, 361005, China

ARTICLE INFO

Article history:
Received 8 November 2013
Received in revised form 21 May 2015
Accepted 25 May 2015
Available online 5 June 2015

Keywords:
Central mean subspace
Multivariate response
Profile least squares
Semiparametric efficiency
Sufficient dimension reduction

ABSTRACT

In this paper, we introduce the notion of the central mean subspace when the response is multivariate, and propose a profile least squares approach to perform estimation and inference. Unlike existing methods in the sufficient dimension reduction literature, the profile least squares method does not require any distributional assumptions on the covariates, and facilitates statistical inference on the central mean subspace. We demonstrate theoretically and empirically that the properly weighted profile least squares approach is more efficient than its unweighted counterpart. We further confirm the promising finite-sample performance of our proposal through comprehensive simulations and an application to an etiologic study on essential hypertension conducted in P. R. China.

1. Introduction

Since the seminal work of Li (1991), sufficient dimension reduction has emerged as an important statistical technique in high dimensional data analysis. The goal of sufficient dimension reduction is to replace the original high dimensional covariates with a few linear combinations while retaining full information of regression. In the regression context we are often interested in the mean function of a univariate response Y given $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$, denoted $E(Y \mid \mathbf{x})$. How to estimate $E(Y \mid \mathbf{x})$ precisely and efficiently when $\mathbf{x}$ is high dimensional has long been a challenging issue. To combine both the flexibility of nonparametric modeling and the interpretability of parametric modeling, Cook and Li (2002) assumed that, there exists a $p \times d_a$ matrix $\boldsymbol{\alpha}$ such that $E(Y \mid \mathbf{x}) = E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})$. This assumption implies that the p-vector $\mathbf{x}$ can be replaced with d_a linear combinations $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$, and such a replacement will not lose information of the mean function. In practice d_a is usually small, say $d_a = 1, 2$ or 3, thus we can estimate $E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})$ precisely and efficiently as long as a consistent estimator of $\boldsymbol{\alpha}$ is available. Observing that $\boldsymbol{\alpha}$ is not identifiable, Cook and Li (2002) defined the central mean subspace, denoted $\mathcal{S}_{E(Y|\mathbf{x})}$, as the smallest column space of $\boldsymbol{\alpha}$. In other words, $\mathcal{S}_{E(Y|\mathbf{x})} = \mathrm{span}(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}$ with the smallest column dimension and satisfies $E(Y \mid \mathbf{x}) = E(Y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})$.

Many approaches have been developed to recover $\mathcal{S}_{E(Y|\mathbf{x})}$ in the area of sufficient dimension reduction. In the particular case with $d_a = 1$, for example, Li and Duan (1989) observed that the ordinary least squares estimator is simple yet useful in estimating $\mathcal{S}_{E(Y|\mathbf{x})}$ when $\mathbf{x}$ follows an elliptical distribution. Cook and Li (2002) further proved that, $\mathrm{span}\left\{\{\mathrm{var}(\mathbf{x})\}^{-1}\mathrm{cov}(\mathbf{x}, Y)\right\} \subseteq \mathcal{S}_{E(Y|\mathbf{x})}$ when $\mathbf{x}$ satisfies the linearity condition that $E(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is a linear function of $\mathbf{x}$. Powell et al. (1989) and Härdle and Stoker (1989) introduced an average derivative estimation method. Ichimura (1993)

* Corresponding author.
E-mail address: wzhong@xmu.edu.cn (W. Zhong).

http://dx.doi.org/10.1016/j.csda.2015.05.006
0167-9473/

and Härdle et al. (1993) proposed a profile least squares approach. For general d_a not necessary to be 1, if $\mathbf{x}$ satisfies both the linearity condition and the constant variance condition that $\mathrm{var}(\mathbf{x} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ is a constant matrix, Li (1992) proved that $\mathrm{span}\left\{\{\mathrm{var}(\mathbf{x})\}^{-1} E\left\{(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^{\mathrm{T}}(Y-E(Y))\right\}\right\} \subseteq \mathcal{S}_{E(Y|\mathbf{x})}$. Nevertheless, requiring the covariates $\mathbf{x}$ satisfy the aforementioned distributional assumptions, such as the linearity and the constant variance conditions, limits the applicability of these methods. Xia et al. (2002) designed a minimum average variance estimation (MAVE) to recover $\mathcal{S}_{E(Y|\mathbf{x})}$, as long as the covariates are continuous. Recently, Ma and Zhu (2014) developed a semiparametric approach which recovers $\mathcal{S}_{E(Y|\mathbf{x})}$ through solving several estimating equations. The semiparametric approach completely removes the distributional assumptions on $\mathbf{x}$, and facilitates statistical inferences on $\mathcal{S}_{E(Y|\mathbf{x})}$. In addition, Ma and Zhu (2014) found that the efficient estimator of $\mathcal{S}_{E(Y|\mathbf{x})}$ is indeed not practical, though some locally efficient estimators of $\mathcal{S}_{E(Y|\mathbf{x})}$ are available. All the aforementioned methods can only be used when the response is univariate.

In this paper, we adapt the notion of central mean subspace of Cook and Li (2002) to multivariate response data. Write the response vector as $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$. We define the central mean subspace $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ as the smallest column space of $\boldsymbol{\beta}$ if it satisfies

$$E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \text{ or equivalently, } E(Y_k \mid \mathbf{x}) = E(Y_k \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \text{ for } k = 1, \ldots, q. \tag{1.1}$$

At the population level, we show that the properties of $\mathcal{S}_{E(Y|\mathbf{x})}$, particularly those stated in Proposition 1 of Cook and Li (2002), also applies to $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$. At the sample level, we design a profile least squares approach to estimate $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$. Our profile least squares approach can be regarded as an extension of the approach proposed by Ichimura (1993) and Härdle et al. (1993). In the particular homoscedastic scenario for univariate response data, our profile least squares method is also equivalent to the estimating equation approach of Ma and Zhu (2014). However, our approach is different from the existing researches in that ours is designed for multivariate response data and allows for multiple linear combinations. Our profile least squares approach also inherits the merits of the semiparametric approach of Ma and Zhu (2014). In particular, our approach does not require any distributional assumptions of $\mathbf{x}$, and statistical inference built upon profile least squares can be easily implemented. We further observe that, the profile least squares approach, if it is properly weighted by considering correlations among multivariate response variables, can improve the efficiency of the unweighted approach in estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$.

Estimation of the central mean subspace for multivariate response data has not yet received much attention in the past decades, largely due to the fact that

$$\mathcal{S}_{E(\mathbf{y}|\mathbf{x})} = \bigcup_{k=1}^{q} \mathcal{S}_{E(Y_k|\mathbf{x})}.$$

Thus, one may argue that estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ jointly amounts to estimating $\mathcal{S}_{E(Y_k|\mathbf{x})}$ marginally, for $k = 1, \ldots, q$. We will show that, estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ jointly will help to improve the estimation efficiency as long as the dimension of $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ is nonparametrically manageable. In particular, even if the response variables are all conditionally uncorrelated, estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ jointly is still more efficient than estimating $\mathcal{S}_{E(Y_k|\mathbf{x})}$ marginally. When the response variables are correlated, the efficiency can be further improved if we take into account the correlations among the response variables. The efficiency is an important issue from both the theoretical and the practical perspectives, particularly when statistical inference on the parameters is much concerned.

There are many sufficient dimension reduction approaches to identify and to estimate the central mean subspace when the response is univariate. To the best of our knowledge, however, very few existing works provide inferential results about the central mean subspace. Ma and Zhu (2014) is probably the first attempt on this topic. They show that the efficient estimator of the central mean subspace is indeed not practical, and their work is designed for univariate response. We consider the inference issue when the response is multivariate, and show that the efficiency in estimating the central mean subspace can be improved by estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ jointly and considering the correlations among the response variables.

The rest of this article is organized as follows. In Section 2, we first adapt the notation of the central mean subspace to multivariate response data, then derive the properties of $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ at the population level. At the sample level, we introduce a profile least squares approach to estimating $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$, and establish the asymptotic properties of the resultant estimators. We demonstrate the methodologies through simulations and an analysis of the essential hypertension data in Section 3. We conclude this paper with a brief discussion in Section 4. All proofs are given in the Appendix.

2. The methodology development

2.1. The central mean subspace for multivariate response data

Our interest is in estimating the mean function $E(\mathbf{y} \mid \mathbf{x})$, where $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$ and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$. Sufficient dimension reduction hinges on finding a $p \times d$ matrix $\boldsymbol{\beta}$ such that the d-vector $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ contains all the information about $\mathbf{y}$ that is available from the mean function. In other words, we hope to find $\boldsymbol{\beta}$ such that $E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. In our context, q and d are small while p is relatively large.