Journal of Multivariate Analysis 133 (2015) 216–231

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Tensor sliced inverse regression

Shanshan Ding [a,*], R. Dennis Cook [b]

[a] Department of Applied Economics and Statistics, University of Delaware, 225 Townsend Hall 531 S College Ave, Newark, DE 19711, USA

[b] School of Statistics, University of Minnesota, 313 Ford Hall 224 Church St SE, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 6 November 2013
Available online 28 September 2014

AMS subject classifications:
62B05
62F12
62H12
62H30
62H35

Keywords:
Sufficient dimension reduction
Sliced inverse regression
Central subspace
Central dimension folding subspace
Tensor data
Tensor decomposition

ABSTRACT

Sliced inverse regression (SIR) is a widely used non-parametric method for supervised dimension reduction. Conventional SIR mainly tackles simple data structure but is inappropriate for data with array (tensor)-valued predictors. Such data are commonly encountered in modern biomedical imaging and social network areas. For these complex data, dimension reduction is generally demanding to extract useful information from abundant measurements. In this article, we propose higher-order sufficient dimension reduction mainly by extending SIR to general tensor-valued predictors and refer to it as tensor SIR. Tensor SIR is constructed based on tensor decompositions to reduce a tensor-valued predictor's multiple dimensions simultaneously. The proposed method provides fast and efficient estimation. It circumvents high-dimensional covariance matrix inversion that researchers often suffer when dealing with such data. We further investigate its asymptotic properties and show its advantages by simulation studies and a real data application.

1. Introduction

Sliced inverse regression was proposed by Li [16]. It is a major supervised dimension reduction technique in non-parametric regression problems. It assumes that the response variable $Y \in \mathbb{R}^1$ depends on the predictor $X \in \mathbb{R}^p$ only through $K (K < p)$ unknown linear combinations of the predictor. Let $B = (\beta_1, \beta_2, \ldots, \beta_K) \in \mathbb{R}^{p\times K}$. This relationship can be described as $Y \perp\!\!\!\perp X|B^T X$, where '$\perp\!\!\!\perp$' stands for independence. To build SIR into the sufficient dimension reduction (SDR) framework, $B^T X$ is called a sufficient reduction of X [1,2]. The matrix B itself is not identifiable since it can be replaced by any non-singular transformation of its columns. However, the linear space spanned by the columns of B is identifiable, denoted as $\mathcal{S}_B$, or Span(B). As a consequence of this structure one can reduce the dimension of the predictor X by replacing it with its projection $\mathrm{P}_{\mathcal{S}_B} X$ onto the subspace $\mathcal{S}_B$, without loss of information on the conditional distribution of $Y|X$; that is,

$$Y \perp\!\!\!\perp X|\mathrm{P}_{\mathcal{S}_B} X. \tag{1}$$

When K is the smallest column rank of B such that (1) holds, the subspace $\mathcal{S}_B$ is called the central dimension reduction subspace (CS), denoted as $\mathcal{S}_{Y|X}$. The goal of SIR is to estimate $\mathcal{S}_{Y|X}$. We provide a brief review of the SIR procedure in Section 2.1.

Conventional SIR is simple and useful for dimension reduction of a vector-valued predictor $X \in \mathbb{R}^p$. However, it is inefficient to tackle problems with more general tensor-valued predictors, such as an m-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\times\cdots\times p_m}$. This type of data is commonly encountered in applications. For instance, EEG (electroencephalography) signals in biomedical

* Corresponding author.
E-mail addresses: sding@udel.edu (S. Ding), dennis@stat.umn.edu (R.D. Cook).

http://dx.doi.org/10.1016/j.jmva.2014.08.015
0047-259X/

engineering, gene expression in bioinformatics and images in pattern recognition are usually formed as two-mode tensors. Video sequences, spatial data and data in social networks often contain three- or multi-mode tensor predictors. Such data are often referred to as multivariate relational data because the tensor-valued predictors represent intrinsic spatial, repeated measured, or other correlated structure among variables. In the EEG data, for example, the brain signals of each subject forms a 256×64 matrix (two-mode tensor) with its rows and columns representing time and location information respectively. Due to the curse of dimensionality, SDR is desirable for such complex data. However, vectorizing these higher-order predictors could and typically does lose important information about the data structure and decrease estimation accuracy.

SDR for tensor-valued predictors has received increasing attention in recent literature. Pioneering work was done by Li et al. [17], where the authors proposed the idea of dimension folding and developed a class of moment-based dimension folding methods, including dimension folding SIR, to reduce a tensor predictor's multiple dimensions simultaneously. Their methods apply to many moment-based dimension reduction approaches but, as will be shown in later sections, are not very efficient, in operation, for dealing with higher-order tensor predictors. Other works include longitudinal SIR studied by Pfeiffer et al. [21] and dimension folding PCA and PFC developed by Ding and Cook [7]. These two studies focused only on two-mode tensor predictors, $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$.

In this paper, we propose a higher-order SDR approach by extending SIR to general m-mode tensor-valued predictors; we refer to it as tensor SIR. The proposed method makes more efficient use of the tensor structure and leads to $\sqrt{n}$ consistent and asymptotically normal estimator of the sufficient reduction subspace. We further compare tensor SIR with the aforementioned methods in the two-mode tensor case. Tensor SIR outperforms dimension folding SIR by (i) circumventing high-dimensional covariance matrix inversion; (ii) alleviating computational cost and improving estimation accuracy; and (iii) having easy interpretation and good theoretical properties. In comparison to longitudinal SIR, tensor SIR places fewer restrictions on the covariance structure of $\mathrm{vec}(\mathcal{X})$. It provides the maximum likelihood estimation of the sufficient reduction when $\mathbf{X}|Y$ is matrix-normally distributed and $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ has a Kronecker structure.

The rest of this paper is organized as follows. Section 2 introduces tensor SIR for two-mode tensor predictors, called two-tensor SIR. Section 3 is devoted to the development of tensor SIR for more general m-mode tensor predictors. We develop the asymptotic properties for the proposed methods in Section 4. Section 5 establishes connections between tensor SIR and other high-order SDR methods. Sections 6 and 7 contain simulation results and data analyses. Discussion is given in Section 8.

2. Two-tensor SIR

Without loss of generality, we assume that the predictors discussed in this paper have mean zero. Let $P_B = B(B^TB)^\dagger B^T$ be the projection onto $\mathrm{Span}(B)$, and $P^T_{B(A)} = AB(B^TAB)^\dagger B^T$ be the projection onto $\mathrm{Span}(B)$ relative to A, where $B \in \mathbb{R}^{p\times d}$ and $A \in \mathbb{R}^{p\times p}$ $(A > 0)$ are two matrices, and $\dagger$ is the Moore–Penrose inverse. Before introducing tensor SIR, we provide a brief review for the conventional SIR.

2.1. A review of SIR

In the classical setting, $X \in \mathbb{R}^p$ is a predictor vector and $Y \in \mathbb{R}^1$ is a response variable. SIR serves to reduce the predictor's dimension by finding the CS $\mathcal{S}_{Y|X}$ so that the projected predictor $P_{\mathcal{S}_{Y|X}}X$ retains the full information on $Y|X$. Let $\mathcal{S}_{Y|X} = \mathrm{Span}(\eta)$, where $\eta \in \mathbb{R}^{p\times d}$ $(d \le p)$. Let Σ and $\hat{\Sigma}$ be the covariance and sample covariance matrices of X. Under the linearity condition (Condition 3.1 in [16]), $\mathrm{E}(X|\eta^TX)$ is a linear function of η^TX. That is, $\mathrm{E}(X|\eta^TX) = A\eta^TX$, where A has an explicit expression $A = \Sigma\eta(\eta^T\Sigma\eta)^\dagger$ (Proposition 4.2, [2]). Therefore,

$$\mathrm{E}(X|Y) = \mathrm{E}[\mathrm{E}(X|\eta^TX, Y)|Y] = \mathrm{E}[\mathrm{E}(X|\eta^TX)|Y] = P^T_{\eta(\Sigma)}\mathrm{E}(X|Y), \tag{2}$$

which indicates $\mathrm{E}(X|Y) \in \mathrm{Span}(\Sigma\eta)$. Correspondingly, $\Sigma^{-1}\mathrm{Span}\{\mathrm{cov}[\mathrm{E}(X|Y)]\} \subseteq \mathcal{S}_{Y|X}$. Conventional SIR estimates $\mathcal{S}_{Y|X}$ by the sample estimate $\hat{\Sigma}^{-\frac{1}{2}}$ times the leading d eigenvectors of $\widehat{\mathrm{cov}}[\hat{\Sigma}^{-\frac{1}{2}}\hat{\mathrm{E}}(X|Y)]$. To allow relatively easy estimation of the inverse mean $\mathrm{E}(X|Y)$, the response Y is replaced with a discrete version by partitioning the range of Y into certain slices. One estimates $\mathrm{E}(X|Y)$ by the intraslice mean.

2.2. Two-tensor SIR

To introduce the idea of tensor SIR, we first consider a simple case when the predictor $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ is two-mode tensor-valued (matrix-valued) and the response Y is univariate. We propose an SDR procedure called two-tensor SIR. It is a special case of tensor SIR dealing with matrix-valued predictors.

The sufficient dimension reduction for $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ is defined as follows. Let '$\otimes$' stand for the Kronecker product.

Definition 1 (*Li et al. [17]*)**.** Let $B_1 \in \mathbb{R}^{p_1\times d_1}$ $(d_1 \le p_1)$ and $B_2 \in \mathbb{R}^{p_2\times d_2}$ $(d_2 \le p_2)$ be two semi-orthogonal matrices that satisfy

$$Y \perp\!\!\!\perp \mathbf{X}|B_1^T\mathbf{X}B_2. \tag{3}$$