Computational Statistics and Data Analysis 103 (2016) 193–205

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Ensemble sufficient dimension folding methods for analyzing matrix-valued data

Yuan Xue [a], Xiangrong Yin [b,*], Xiaolin Jiang [c]

[a] School of Statistics, University of International Business and Economics, Beijing, China
[b] Department of Statistics, University of Kentucky, Lexington, KY, USA
[c] School of Banking and Finance, University of International Business and Economics, Beijing, China

ARTICLE INFO

Article history:
Received 17 August 2015
Received in revised form 4 May 2016
Accepted 5 May 2016
Available online 13 May 2016

Keywords:
Central dimension folding subspace
Central mean dimension folding subspace
Folded-MAVE ensemble
Folded-OPG ensemble
Folded-SR ensemble

ABSTRACT

The construction of novel sufficient dimension folding methods for analyzing matrix-valued data is considered. For a matrix-valued predictor, traditional dimension reduction methods fail to preserve the matrix structure. However, dimension folding methods can preserve the data structure and improve estimation accuracy. Folded-outer product of gradient (folded-OPG) ensemble estimator and two refined estimators, folded-minimum average variance estimation (folded-MAVE) ensemble and folded-sliced regression (folded-SR) ensemble are proposed to recover central dimension folding subspace (CDFS). Due to ensemble idea, estimation accuracies are improved for finite samples by repeatedly using the data. A modified cross validation method is used to determine the structural dimensions of CDFS. Simulated examples demonstrate the performance of folded ensemble methods by comparing with existing inverse dimension folding methods. The efficacy of folded-MAVE ensemble method is also evaluated by comparing with inverse dimension folding methods for analyzing the Standard & Poor's 500 stock data set.

1. Introduction

Ding and Cook (2014) investigated the relationship between the monthly Dow Jones industrial average index change rate and 19 daily stock price change rates over the 30 Dow Jones companies from January 2001 to December 2010. The response is a univariate continuous variable, however, the predictor for each observation is a 19×30 matrix. Traditional sufficient dimension reduction approaches can be applied to such matrix-valued data by vectorizing the data into a vector. As pointed out by Li et al. (2010, hereafter LKA) vectorizing a matrix-/array-valued predictor may lose sufficient information, data structure and related interpretation. LKA (2010) introduced a concept of central dimension folding subspace (CDFS) and proposed three inverse estimation methods, folded-sliced inverse regression (folded-SIR), folded-sliced average variance estimation (folded-SAVE) and folded-directional regression (folded-DR), to reduce the dimensions of a matrix-/array-valued predictor X as much as possible while preserving its intrinsic structure and the regression relation with a univariate response Y. If a predictor is vector-valued, sufficient dimension folding reduces to sufficient dimension reduction (Cook, 1994, 1996, 1998). Treating the predictor X as matrix-/array-valued instead of vectorizing it can preserve the original data structure, important aspects of interpretation, and reduce the number of parameters to enhance estimation accuracy (LKA, 2010). Along the same line, Pfeiffer et al. (2011) proposed an alternative inverse dimension folding method to folded-SIR and named

* Corresponding author.
E-mail address: yinxiangrong@uky.edu (X. Yin).

http://dx.doi.org/10.1016/j.csda.2016.05.001
0167-9473/

it longitudinal version of sliced inverse regression (LSIR). Ding and Cook (2014) proposed two inverse approaches, dimension folding principal components analysis (DF-PCA) and dimension folding principal fitted components (DF-PFC), to estimate CDFS.

Let vec(·) be an operation that stacks a matrix into a vector column by column. Inverse dimension folding approaches require certain conditions such as the linearity condition and the constant variance condition on the marginal distribution of vec(X). However, these conditions are not always satisfied, which can make the estimate not stable. Xue and Yin (2014) considered forward sufficient dimension folding methods for the conditional mean of Y given X, which can attain a more accurate estimate when conditional mean is of interest. Xue and Yin (2014) proposed a concept of central mean dimension folding subspace (CMDFS) and its two local estimation methods: folded-OPG and folded-MAVE in analogy to the development of their counterparts outer product of gradient (OPG) and minimum average variance estimation (MAVE) for a vector-valued predictor (Xia et al., 2002). A more general situation is considered by Xue and Yin (2015) which includes CDFS and CMDFS as two special cases. Xue and Yin (2015) performed sufficient dimension folding in reference to a general functional of the conditional distribution of Y given X. By applying sufficient dimension folding in reference to a specific functional, the relation between Y and X reflected in that functional is preserved.

In this article, we consider constructing new methods to estimate CDFS based on the ensemble idea introduced by Yin and Li (2011) for a vector-valued predictor. Yin and Li (2011) considered a general function family $\mathfrak{F}$ of Y, and introduced a probability measure on $\mathfrak{F}$. Functions $f_1, \ldots, f_m$ are randomly sampled from $\mathfrak{F}$ according to that probability. They assembled the central mean subspaces (CMS) of $f_l(Y)$, $l = 1, \ldots, m$, given a vector-valued predictor to recover the central subspace (CS). The ensemble idea can be applied to estimate CDFS via a set of CMDFS's as well. Folded-MAVE has three appealing advantages in estimating CMDFS: (1) it can exhaustively estimate CMDFS; (2) there is no strong assumption assumed on the distribution of X and (3) the estimation procedure of folded-MAVE can be broken down into iterative quadratic optimizations and each step has an explicit solution. However, folded-MAVE method may not recover directions outside CMDFS. Xue and Yin (2014) pointed out that CMDFS is not invariant under one-to-one transformation of the response variable. Based on the ensemble idea, we can repeatedly use data to obtain a better estimate of CDFS, which is achieved by estimating CMDFS for a sufficient large class of transformations of the response variable. We construct folded-OPG ensemble, folded-MAVE ensemble and folded-sliced regression (folded-SR) ensemble methods to estimate CDFS. The first two ensemble methods are similar to those developed by Yin and Li (2011), and the folded-SR ensemble method is developed in parallel with its counterpart sliced regression (SR; Wang and Xia, 2008) for a vector-valued predictor.

The rest of this article is organized as follows. In Section 2, we study the ensemble theory on characterizing CDFS for a matrix-valued predictor. In Section 3, we introduce algorithms of folded-OPG ensemble, folded-MAVE ensemble and folded-SR ensemble and propose a cross validation criterion to estimate the structural dimensions of CDFS. In Section 4, we comment on the consistency and convergence rate of folded-MAVE ensemble. Simulation studies and a real data analysis are included in Sections 5 and 6, respectively, followed by a discussion on characterizing CDFS for an array- or higher dimensional valued predictor in Section 7. A short discussion is in Section 8.

2. Characterizing CDFS

In this section, we develop sufficient dimension folding methods to estimate CDFS based on the ensemble idea (Yin and Li, 2011) for a matrix-valued predicator. Definitions and theorems in this section are straightforward extensions from Yin and Li (2011). Let the predictor variable X be a $p \times q$ random matrix with its support Ω_X and Y be an s-dimensional random vector with its support Ω_Y. Let $\mathfrak{F}$ be a family of functions $f : \Omega_Y \to \mathbb{F} \subseteq \mathbb{R}$, which means that a transformation f can project a vector-valued Y to the scalar field $\mathbb{F}$. Let $P_{\mathcal{S}}$ denote the orthogonal projection onto subspace $\mathcal{S}$. Xue and Yin (2014) denoted CMDFS for the conditional mean $E[f(Y)|X]$ as $\mathcal{S}_{E[f(Y)|\circ X\circ]}$ which is equal to $\mathcal{S}_{E[f(Y)|X\circ]} \otimes \mathcal{S}_{E[f(Y)|\circ X]}$. The subspaces $\mathcal{S}_{E[f(Y)|\circ X]}$ and $\mathcal{S}_{E[f(Y)|X\circ]}$ are the intersections of all the respective subspaces $\mathcal{S}_L$ and $\mathcal{S}_R$ of $\mathbb{R}^p$ and $\mathbb{R}^q$ such that

$$E[f(Y)|X] = E[f(Y)|P_{\mathcal{S}_L} X P_{\mathcal{S}_R}]. \tag{1}$$

LKA (2010) denoted CDFS of Y versus X as $\mathcal{S}_{Y|\circ X\circ} = \mathcal{S}_{Y|X\circ} \otimes \mathcal{S}_{Y|\circ X}$ where $\mathcal{S}_{Y|\circ X}$ and $\mathcal{S}_{Y|X\circ}$ are the intersections of all the respective subspaces $\mathcal{S}_L$ and $\mathcal{S}_R$ of $\mathbb{R}^p$ and $\mathbb{R}^q$ such that

$$Y \perp\!\!\!\perp X | P_{\mathcal{S}_L} X P_{\mathcal{S}_R}. \tag{2}$$

Definition 1. Let $\mathfrak{F}$ be a family of measurable $\mathbb{F}$-valued functions defined on Ω_Y. If

$$\text{span}\{\mathcal{S}_{E[f(Y)|\circ X\circ]} : f \in \mathfrak{F}\} = \mathcal{S}_{Y|\circ X\circ},$$

the family $\mathfrak{F}$ is said to characterize the CDFS.

Our goal is to identify the family $\mathfrak{F}$, so that the dimension folding subspaces for the conditional means $E[f(Y)|X]$, when combined for a set of $f(Y)$, can recover the dimension folding subspace for Y versus X. Let F_Y denote the distribution of Y and $L_1(F_Y)$ be the class of functions $f(Y)$ such that $E|f(Y)| < \infty$, with the norm $E|f(Y)|$. Let $L_2(F_Y)$ be the class of functions $f(Y)$ with finite variances, with the inner product $\langle f_1, f_2\rangle = E[f_1(Y)f_2(Y)]$. And let $\mathfrak{B}$ be the family of measurable indicator functions of Y, which means $\mathfrak{B} = \{I_{\mathcal{B}} : \mathcal{B}$ is a Borel set in $\Omega_Y\}$. We have the following theorem: