Journal of Multivariate Analysis 145 (2016) 178–189

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

The Dual Central Subspaces in dimension reduction

Ross Iaci [a,*], Xiangrong Yin [b], Lixing Zhu [c]

[a] Department of Mathematics, The College of William and Mary, Williamsburg, VA, 23185, United States
[b] Department of Statistics, University of Kentucky, Lexington, KY, 40536, United States
[c] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

ARTICLE INFO

Article history:
Received 4 May 2015
Available online 21 December 2015

AMS subject classifications:
62H20
62H12
62G07
62B10

Keywords:
Canonical Correlation Analysis
Dimension reduction
Dual Central Subspaces
Multivariate analysis
Visualization

ABSTRACT

Existing dimension reduction methods in multivariate analysis have focused on reducing sets of random vectors into equivalently sized dimensions, while methods in regression settings have focused mainly on decreasing the dimension of the predictor variables. However, for problems involving a multivariate response, reducing the dimension of the response vector is also desirable and important. In this paper, we develop a new concept, termed the Dual Central Subspaces (DCS), to produce a method for simultaneously reducing the dimensions of two sets of random vectors, irrespective of the labels predictor and response. Different from previous methods based on extensions of Canonical Correlation Analysis (CCA), the recovery of this subspace provides a new research direction for multivariate sufficient dimension reduction. A particular model-free approach is detailed theoretically and the performance investigated through simulation and a real data analysis.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Methods for dimension reduction in multivariate association studies for two sets of random vectors generally focus on reducing the dimensions of both sets of variables, where the role of predictor and response is unimportant, while multivariate regression centers on the dimension reduction of the vector labeled the predictor variables.

A popular method pioneered by Hotelling [10] for the pairwise extraction of the significant relationships that exist between two random vectors is Canonical Correlation Analysis (CCA). Kettenring [14,15] investigated five measures, extending Hotelling's theory [10] to multiple sets, while Van der Burg and De Leeuw [31] developed a method termed nonlinear canonical correlation analysis. More recently, many methods advancing this area of research have been proposed, see for example [34,39,13,12] and references therein. Importantly, all of these methods require that the number of coefficient vectors that provide the dimension reduction be equal. While this restriction simplifies the problem, if the number of coefficient vectors that recover the true associations between the random vectors are not equal then this could result in a critical loss of information. Therefore, methods that allow the number of coefficient vectors to be different and thus, provide a sufficient dimension reduction, are crucial in multivariate analysis.

To this end, we introduce the Dual Central Subspaces (DCS), and subsequently provide a new method to estimate these subspaces, which provides a simultaneous sufficient dimension reduction of two multivariate random vectors. That is, our approach provides a dimension reduction of both vectors without requiring the dimensions of the reduction to be equal. To identify the DCS, we consider a higher-order information measure based on the Kullback–Leibler (KL) divergence, rather

* Corresponding author.
E-mail addresses: riaci@wm.edu (R. Iaci), yinxiangrong@gmail.com (X. Yin), lzhu@hkbu.hk (L. Zhu).

http://dx.doi.org/10.1016/j.jmva.2015.12.003
0047-259X/© 2015 Elsevier Inc. All rights reserved.

than extending traditional methods for estimating the Central Subspaces (CS) that recover information from lower moments, such as SIR and SAVE. The KL index was introduced in [13] to provide a measure of overall association between random vectors, the main focus of their paper; a more detailed discussion of the differences between the use of the index in both papers is provided below. An advantage, and motivation, for using this information based measure is that it is able to detect both linear and nonlinear relationships that exist between random vectors, which enables a more complete recovery of the DCS while treating both vectors equivalently. Moreover, in using this method no distributional assumptions, except for the existence of the joint density, are required and the estimation of the DCS becomes an optimization problem. The method is directly applicable for random vectors labeled as predictor and response and thus, also provides a powerful tool for dimension reduction in a multivariate regression setting.

Since Li's sliced inverse regression [17] method, there have been many statistical studies that have focused on dimension reduction in a regression setting. For example, see Cook and Weisberg (SAVE, [8]), Li (PHD, [18]), Yin and Cook (Cov_k, [36]), Xia et al. (MAVE, [32]), the seminal papers of Ma and Zhu [21–23] and for a detailed review see [3] or [9]. All of these methods consider only a univariate response and thus, dimension reduction is performed only on the predictor variables. A few methods have been developed in a multivariate regression setting, but the dimension reduction is focused only on the predictors; see for example [6,35,20]. Methods for sufficient dimension reduction, especially with a multivariate response, for example [41,25], could also be considered to develop a method to identify the DCS, but prefer the flexibility of the information based procedure in this initial work. More recently, Cook et al. [5] developed an envelope model for multivariate linear regression that not only reduces the dimension of the predictors, but also the noninformative responses in order to obtain a more efficient estimator. While their method and those of others, such as Su and Cook [28–30], Cook et al. [4] and Cook and Su [7], have made significant advances in this area, the focus of these techniques is only on the regression mean function for a specified regression model. The proposed method of Li et al. [19] for achieving a dimension reduction in a multivariate response regression setting could be considered for developing a method for the identification of the DCS, however the linearity conditions and the exhaustive nature of recovering all the directions using this SIR based method are viewed to be somewhat restrictive. Importantly, procedures based on spectral decompositions, for example SIR and SAVE, and moment based methods in general, have been shown to perform poorly, even under strong conditions like normality, when nonlinear relationships exist between the responses and predictors. To investigate this in the context of estimating the DCS, and further motivate our use of the KL information based method, we use an alternating search procedure to estimate the DCS using the projective resampling SIR procedure of Li et al. [20] and compare the performance to our method in simulation.

The article is organized as follows. In Section 2.1 we introduce the concept of the DCS, discuss the theoretical properties, and its role in providing a new method for multivariate sufficient dimension reduction. Identification of the DCS and computational aspects of our approach are described in Section 2.2. Simulation studies are performed in Section 3 and, in Section 4, we revisit the Los Angeles County dataset that was initially investigated in [26] to gain further insight into the associations that exist between mortality and environmental conditions using our method. Proofs of the presented results and the projective resampling SIR study are provided in the Appendix.

2. Methodology

2.1. The Dual Central Subspaces

In this section, we define the Dual Central Subspaces (DCS) to reduce the dimensions of two sets of variables sufficiently and discuss the relevant properties. Even though contextually each vector may be regarded as the response or predictor, the labeling of the vectors as predictor and response is used only for the convenient exposition of the method. Importantly, this novel concept allows the size of the dimension for which the reduction occurs to vary for each random vector.

Let $\mathcal{S}$ denote a generic subspace, $\mathcal{S}(\mathbf{A}_r)$ represent the r-dimensional subspace in $\mathbf{R}^p$ spanned by the columns of a $p \times r$ full rank matrix $\mathbf{A}$ and finally, let $P_{\mathcal{S}}$ designate the projection onto $\mathcal{S}$ with respect to the usual inner product. Consider two sets of random variables, a $p \times 1$ vector $\mathbf{X}$ and a $q \times 1$ vector $\mathbf{Y}$, the Dimension Reduction Subspace (DRS) for reducing the dimension of $\mathbf{X}$ is defined as the subspace $\mathcal{S}$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}}\mathbf{X}. \tag{1}$$

Here, the notation means that $\mathbf{Y}$ is independent of $\mathbf{X}$ given $P_{\mathcal{S}}\mathbf{X}$, the projection of $\mathbf{X}$ onto the subspace $\mathcal{S}$. The Central Subspace (CS), denoted $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, is defined as the intersection of all DRSs, which importantly is also a DRS. Note that, when $q = 1$ and $\mathbf{Y}$ is considered the response, this is equivalent to the CS defined in [1–3].

In a multivariate dimension reduction CCA context, it is also necessary to reduce the dimension of $\mathbf{Y}$ sufficiently. To this end, we define the CS of $\mathbf{Y}$, denoted $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, by simply interchanging the roles of $\mathbf{X}$ and $\mathbf{Y}$ in the above definition. That is, we define the DRS for the dimension reduction of $\mathbf{Y}$ as the subspace $\mathcal{S}$ such that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | P_{\mathcal{S}}\mathbf{Y}$. Again, $P_{\mathcal{S}}$ is the usual projection onto the subspace $\mathcal{S}$ and the CS is defined as the intersection of all DRSs. Thus, similar to the role of the CS associated with $\mathbf{X}$, the CS of $\mathbf{Y}$ will also play an important role in dimension reduction, and with this subspace the information from $\mathbf{Y}$ can be preserved.

In the sense of reducing the dimensions of both $\mathbf{X}$ and $\mathbf{Y}$, the two sets of variables can be treated equally and thus, we term the subspaces, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathcal{S}_{\mathbf{X}|\mathbf{Y}}$, the Dual Central Subspaces (DCS). The dimensions of the respective Central Subspaces are

Download English Version:

https://daneshyari.com/en/article/1145197

Download Persian Version:

https://daneshyari.com/article/1145197

Daneshyari.com