Journal of Multivariate Analysis 166 (2018) 241–265

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Angle-based joint and individual variation explained

Qing Feng, Meilei Jiang *, Jan Hannig, J.S. Marron

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 7 April 2017
Available online 24 March 2018

AMS subject classifications:
62H99

Keywords:
Data integration
Heterogeneity
Perturbation theory
Principal angle
Singular value decomposition

ABSTRACT

Integrative analysis of disparate data blocks measured on a common set of experimental subjects is a major challenge in modern data analysis. This data structure naturally motivates the simultaneous exploration of the joint and individual variation within each data block resulting in new insights. For instance, there is a strong desire to integrate the multiple genomic data sets in The Cancer Genome Atlas to characterize the common and also the unique aspects of cancer genetics and cell biology for each source. In this paper we introduce Angle-Based Joint and Individual Variation Explained capturing both joint and individual variation within each data block. This is a major improvement over earlier approaches to this challenge in terms of a new conceptual understanding, much better adaption to data heterogeneity and a fast linear algebra computation. Important mathematical contributions are the use of score subspaces as the principal descriptors of variation structure and the use of perturbation theory as the guide for variation segmentation. This leads to an exploratory data analysis method which is insensitive to the heterogeneity among data blocks and does not require separate normalization. An application to cancer data reveals different behaviors of each type of signal in characterizing tumor subtypes. An application to a mortality data set reveals interesting historical lessons. Software and data are available at GitHub https://github.com/MeileiJiang/AJIVE_Project.

1. Introduction

A major challenge in modern data analysis is data integration, combining diverse information from disparate data sets measured on a common set of experimental subjects. Simultaneous variation decomposition has been useful in many practical applications. For example, Kühnle [14], Lock and Dunson [19], and Mo et al. [24] performed integrative clustering on multiple sources to reveal novel and consistent cancer subtypes based on understanding of joint and individual variation. The Cancer Genome Atlas (TCGA) [25] provides a prototypical example for this problem. TCGA contains disparate data types generated from high-throughput technologies. Integration of these is fundamental for studying cancer on a molecular level. Other types of application include analysis of multi-source metabolomic data [15], extraction of commuting patterns in railway networks [10], recognition of brain-computer interface [49], etc.

A unified and insightful understanding of the set of data blocks is expected from simultaneously exploring the joint variation representing the inter-block associations and the individual variation specific to each block. Lock et al. [20] formulated this challenge into a matrix decomposition problem. Each data block is decomposed into three matrices modeling different types of variation, including a low-rank approximation of the joint variation across the blocks, low-rank approximations of the individual variation for each data block, and residual noise. Definitions and constraints were proposed

* Corresponding author.
E-mail addresses: qing.feng1014@gmail.com (Q. Feng), jiangm@live.unc.edu (M. Jiang), jan.hannig@unc.edu (J. Hannig), marron@unc.edu (J.S. Marron).

https://doi.org/10.1016/j.jmva.2018.03.008
0047-259X/© 2018 Elsevier Inc. All rights reserved.

for the joint and individual variation together with a method named JIVE; see https://genome.unc.edu/jive/ and O'Connell and Lock [27] for Matlab and R implementations of JIVE, respectively.

JIVE was a promising framework for studying multiple data matrices. However, Lock et al. [20] algorithm and its implementation was iterative (thus slow) and performed rank selection based on a permutation test. It had no guarantee of achieving a solution that satisfied the definitions of JIVE, especially in the case of some correlation between individual components. The example in Fig. B.16 in Appendix B shows that this can be a serious issue. An important related algorithm named COBE was developed by Zhou et al. [50]. COBE considers a JIVE-type decomposition as a quadratic optimization problem with restrictions to ensure identifiability. While COBE removed many of the shortcomings of the original JIVE, it was still iterative and often required longer computation time than the Lock et al. [20] algorithm. Neither Zhou et al. [50] nor Lock et al. [20] provided any theoretical basis for selection of a thresholding parameter used for separation of the joint and individual components.

A novel solution, *Angle-based Joint and Individual Variation Explained (AJIVE)*, is proposed here for addressing this matrix decomposition problem. It provides an efficient *angle-based algorithm* ensuring an identifiable decomposition and also an insightful new interpretation of extracted variation structure. The key insight is the use of row spaces, i.e., a focus on scores, as the principal descriptor of the joint and individual variation, assuming columns are the n data objects, e.g., vectors of measurements on patients. This focuses the methodology on variation patterns across data objects, which gives straightforward definitions of the components and thus provides identifiability. These variation patterns are captured by the *score subspaces* of $\mathbb{R}^n$. Segmentation of joint and individual variation is based on studying the relationship between these score subspaces and using perturbation theory to quantify noise effects [36].

The main idea of AJIVE is illustrated in the flowchart of Fig. 1. AJIVE works in three steps. First we find a low-rank approximation of each data block (shown as the far left color blocks in the flowchart) using SVD. This is depicted (using blocks with colored dashed line boundaries) on the left side of Fig. 1 with the black arrows signifying thresholded SVD. Next, in the middle of the figure, SVD of the concatenated bases of row spaces from the first step (the gray blocks with colored boundaries) gives a joint row space (the gray box next to the circle), using a mathematically rigorous threshold derived using perturbation theory in Section 2.3. This SVD is a natural extension of Principal Angle Analysis, which is also closely related to the multi-block extension of Canonical Correlation Analysis [26] as well as to the flag means of the row spaces [5]; see Section 4.2 for details. Finally, the joint and individual space approximations are found using projection of the joint row space and its orthogonal complements on the data blocks as shown as colored boundary gray squares on the right with the three joint components at the top and the individual components at the bottom.

Using score subspaces to describe variation contained in a matrix not only empowers the interpretation of analysis but also improves understanding of the problem and the efficiency of the algorithm. An identifiable decomposition can now be obtained with all definitions and constraints satisfied even in situations when individual spaces are somewhat correlated. Moreover, the need to select a tuning parameter used to distinguish joint and individual variation is eliminated based on theoretical justification using perturbation theory. A consequence is an algorithm which uses a fast built-in singular value decomposition to replace lengthy iterative algorithms. For the example in Section 1.1, implemented in Matlab, the computational time of AJIVE (10.8 s) is about 11 times faster than the old JIVE (121 s) and 39 times faster than COBE (422 s). The computational advantages of AJIVE get even more pronounced on data sets with higher dimensionality and more complex heterogeneity such as the TCGA data analyzed in Section 3.1. For a very successful application of AJIVE on integrating fMRI imaging and behavioral data, see Yu et al. [48].

Other methods that aim to study joint variation patterns and/or individual variation patterns have also been developed. Westerhuis et al. [42] discuss two types of methods. One main type extends traditional Principal Component Analysis (PCA), including Consensus PCA and Hierarchical PCA first introduced by Wold et al. [45,46]. An overview of extended PCA methods is discussed in Smilde et al. [35]. Abdi et al. [1] discuss a multiple block extension of PCA called multiple factor analysis. This type of method computes the block scores, block loadings, global loadings and global scores.

The other main type of method is extensions of Partial Least Squares (PLS) [44] or Canonical Correlation Analysis (CCA) [9] that seek associated patterns between the two data blocks by maximizing covariance/correlation. For example, Wold et al. [46] introduced multi-block PLS and hierarchical PLS (HPLS) and Trygg and Wold [37] proposed *O2-PLS* to better reconstruct joint signals by removing structured individual variation. A multi-block extension can be found in Löfstedt et al. [21].

Yang and Michailidis [47] provide a very nice integrative joint and individual component analysis based on non-negative matrix factorization. Ray et al. [31] do integrative analysis using factorial models in the Bayesian setting. Schouteden et al. [33,34] propose a method called DISCO-SCA that is a low-rank approximation with rotation to sparsity of the concatenated data matrices.

A connection between extended PCA and extended PLS methods is discussed in Hanafi et al. [6]. Both types of methods provide an integrative analysis by taking the inter-block associations into account. These papers recommend use of normalization to address potential scale heterogeneity, including normalizing by the Frobenius norm, or the largest singular value of each data block, etc. However, there are no consistent criteria for normalization and some of these methods have convergence problems. An important point is that none of these approaches provide simultaneous decomposition highlighting joint and individual modes of variation with the goal of contrasting these to reveal new insights.