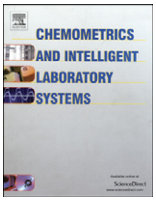
Chemometrics and Intelligent Laboratory Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



A variable selection method for simultaneous component based data integration

Zhengguo Gu*, Katrijn Van Deun

Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 12 January 2016
Received in revised form 13 May 2016
Accepted 29 July 2016
Available online xxxx

Keywords:
Data integration
Sparse simultaneous component analysis
Soft thresholding
Common/distinctive process

ABSTRACT

The integration of multiblock high throughput data from multiple sources is one of the major challenges in several disciplines including metabolomics, computational biology, genomics, and clinical psychology. A main challenge in this line of research is to obtain interpretable results 1) that give an insight into the common and distinctive sources of variations associated to the multiple and heterogeneous data blocks and 2) that facilitate the identification of relevant variables. We present a novel variable selection method for performing data integration, providing easily interpretable results, and recovering underlying data structure such as common and distinctive components. The flexibility and applicability of this method are showcased via numerical simulations and an application to metabolomics data.

1. Introduction

Big data from multiple sources are frequently seen in metabolomics, biology, genomics, and clinical psychology. To obtain a more complete picture about the object of interest (e.g. genes and persons), datasets from several sources are often integrated. Hence, joint analysis of such multi-source data has become increasingly popular in recent years. For example, Acar et al.[1] proposed a structure-revealing data fusion model for analyzing heterogeneous datasets with missing values and revealing shared and unshared components, and used the model to investigate measures obtained from mixtures analyzed by liquid chromatography-mass spectrometry and nuclear magnetic resonance.

A promising class of methods for analyzing data blocks are simultaneous component methods that have received attention in biomedical research [2], genomics [3,4], bioinformatics [5–8], cancer research [9], and psychology [10] in recent years. However, one drawback of simultaneous component integration methods is their low interpretability [11]: The simultaneous component representations are based on the contributions of all the variables, but it may well be the case that only a few variables play an important role in the integrated data, whereas the rest can be safely treated as trivial variables. Hence, a data-integration method that generates sparse component loadings can be very helpful. To this end, Van Deun et al. proposed a sparse simultaneous component method [11], a flexible framework that incorporates regularization penalties to induce sparse results, thereby greatly improving the interpretability. Similar ideas have been proposed in [1] and [2].

A challenge, however, to the sparse simultaneous component method is how to reveal the structure of integrated data. Specifically, researchers may want to know whether each data block has its unique data structure or whether there exists a common structure shared by several (or all) data blocks. Revealing the data structure is important because it helps identify which (biological, psychological etc.) processes govern all the sources and which govern a particular data source. To give an example, research on obesity and its genetic and environmental causes may involve joint analysis on survey, dietary, biomarker and genetic data. Finding common processes that are jointly governed by genes and environmental factors may be of great importance, but such subtle common processes are very difficult to identify because the major dominating variations in the integrated data are most likely to be the biological information with the general biological functions playing a role therein (see. e.g. [12]). In other situations, common processes are easy to identify, but distinctive ones are not. For example, personality research in cross-cultural psychology often sees cases where the dominating variations in data are those universal personality types shared cross all cultures studied, whereas it is rather difficult (though of substantial interest) to identify unique variations that belong to a certain culture (see e.g. [13]). Hence, statistical methods that

[*] Corresponding author at: Warandelaan 2, Tilburg, 5037 AB, The Netherlands.
E-mail addresses: z.gu@uvt.nl (Z. Gu), K.VanDeun@uvt.nl (K. Van Deun).

http://dx.doi.org/10.1016/j.chemolab.2016.07.013
0169-7439/© 2016 Elsevier B.V. All rights reserved.

can identify common and distinctive processes are greatly in need, yet existing methods such as the sparse simultaneous component method [11] do not provide adequate solutions.

Simultaneous component methods are a promising tool for the analysis of multiblock and multiset large data (in the number of variables). Yet, to improve interpretability both variable selection and accounting for common and distinctive sources of variations are needed. In this paper, we introduce a novel variable selection method for simultaneous component data integration. We will show that this method is suitable for revealing complex multi-group structures where, for example, there is a (sparse) common component shared by all the datasets, and several distinctive components that belong to particular datasets. This paper is organized as follows. After briefly introducing sparse simultaneous component methods, we present our variable selection method, which is followed by a series of numerical simulations. Afterwards, we showcase an application of the proposed method to metabolomics data. Discussions and conclusions are offered in the end.

2. Methods

In this section, we first briefly review the model for sparse simultaneous component analysis, the objective function of which is not suitable for identifying common and distinctive processes. Afterwards, we present our novel variable selection method that can be easily adjusted for finding common and distinctive processes. In the last part of this section, we introduce a resampling-based stability selection method to be incorporated into our variable selection method.

2.1. Sparse simultaneous component data integration

As an extension of principal component analysis (PCA), simultaneous component analysis (SCA) is applied to situations where multiple data blocks with the same persons or the same variables are to be integrated [11]. Let $\mathbf{X}_k$ denote a $I_k \times J_k$ matrix representing the kth data block ($k = 1, \cdots, K$) with information (e.g. scores) of I_k persons on J_k variables. Principal component analysis then decomposes $\mathbf{X}_k$ into

$$\mathbf{X}_k = \mathbf{T}_k \mathbf{P}_k^T + \mathbf{E}_k, \tag{1}$$

where $\mathbf{T}_k$ with size $I_k \times R$ denotes the component scores for R components, where $\mathbf{P}_k$ with size $J_k \times R$ is referred to as the component loadings, and where $\mathbf{E}_k$ denotes the residuals [11,14]. (The superscript ^{T} denotes the transpose of a matrix.) Furthermore, certain constraints, such as $\mathbf{T}_k^T \mathbf{T}_k = \mathbf{I}$ and a principal axis orientation, are also needed to identify the solution.

PCA can be interpreted as a least squares minimization problem

$$\left(\hat{\mathbf{T}}_k, \hat{\mathbf{P}}_k\right) = \arg\min_{\mathbf{T}_k, \mathbf{P}_k} \left\| \mathbf{X}_k - \mathbf{T}_k \mathbf{P}_k^T \right\|_2^2 \tag{2}$$

with respect to $\mathbf{T}_k$ and $\mathbf{P}_k$. To improve the interpretability of the results of PCA, the Lasso penalty $\|\mathbf{P}_k\|_1 = \sum_{j_k, r} |p_{j_k r}|$[15] is imposed on $\mathbf{P}_k$ to induce sparse loadings (for a detailed discussion, see [11,16]), and thus the model becomes

$$\left(\hat{\mathbf{T}}_k, \hat{\mathbf{P}}_k\right) = \arg\min_{\mathbf{T}_k, \mathbf{P}_k} \left\| \mathbf{X}_k - \mathbf{T}_k \mathbf{P}_k^T \right\|_2^2 + \lambda_L \parallel \mathbf{P}_k \|_1, \; (\lambda_L \geq 0), \tag{3}$$

where λ_L is the tuning parameter for the Lasso penalty. This penalty has the property to shrink the loadings, some (or many for large λ_L's) exactly to zero.

Building upon the sparse PCA and simultaneous component methods (for a review, see [7]), Van Deun et al.[11] presented a flexible framework for integrating multiblock data where a few penalties were introduced, thereby greatly improving the interpretability of the integrated data. The sparse simultaneous component method solves the following penalized least squares minimization problem

$$\begin{aligned}\left(\hat{\mathbf{T}}, \hat{\mathbf{P}}_k\right) &= \arg\min_{\mathbf{T}, \mathbf{P}_k} \left\| \mathbf{X}_C - \mathbf{T}\mathbf{P}_C^T \right\|_2^2 + \lambda_L \parallel \mathbf{P}_C \|_1 \\ &\quad + \sum_{k,r} \left(\lambda_G \sqrt{J_k} \parallel \mathbf{P}_k \|_2 + \lambda_E \parallel \mathbf{P}_k \|_{1,2} \right)\end{aligned} \tag{4}$$

subject to

$$\mathbf{T}^T\mathbf{T} = \mathbf{I}; \lambda_L, \lambda_E, \lambda_G \geq 0.$$

The flexible framework Eq. (4) decomposes the concatenated data $\mathbf{X}_C$ consisting of K data blocks $\mathbf{X}_k$ (with respect to the same set of I persons) into component scores $\mathbf{T}$ and concatenated component loadings $\mathbf{P}_C$ consisting of K blocks of component loadings $\mathbf{P}_k$. Note that $\mathbf{T}$ is the same for each of the data blocks. Besides a Lasso penalty on the concatenated component loadings $\mathbf{P}_C$, the flexible framework also incorporates an Elitist Lasso penalty $\sum_{k,r} (\|\mathbf{P}_k\|_{1,2}) = \sum_{k,r} \left(\sum_{j_k} |p_{j_k r}| \right)^2$ [17,18] and a Group Lasso penalty $\sum_{k,r} \sqrt{J_k} \|\mathbf{P}_k\|_2 = \sum_{k,r} \sqrt{J_k \sum_{j_k} \left(p_{j_k r}^2 \right)}$ [19]. The Elitist Lasso penalty and the Group Lasso penalty become very useful when variable selection in $\mathbf{P}_C$ pertains to groups (i.e. data blocks): The Elitist Lasso penalty specializes in selecting variables within groups. For example, in regression analysis, the Elitist Lasso penalty retains the highest coefficients within each group. The Group Lasso penalty on the other hand performs variable selection on the group level; that is, all the variables in the data blocks with the highest sum of squared coefficients are selected.

2.2. A novel variable selection method for sparse simultaneous component based data integration

First note that it is necessary to rewrite the minimization problem (Eq. (4)) in its vectorization form as follows:

$$\begin{aligned}\left(\hat{\mathbf{T}}, \hat{\mathbf{P}}_k\right) &= \arg\min_{\mathbf{T}, \mathbf{P}_k} \left\| \mathbf{X}_C - \mathbf{T}\mathbf{P}_C^T \right\|_2^2 + \lambda_L \|\mathbf{P}_C\|_1 \\ &\quad + \sum_{k,r} \left(\lambda_G \sqrt{J_k} \|\mathbf{P}_k\|_2 + \lambda_E \|\mathbf{P}_k\|_{1,2} \right) \\ &= \arg\min_{\mathbf{T}, \mathbf{P}_k} \left\| \mathrm{Vec}(\mathbf{X}_C) - (\mathbf{I} \otimes \mathbf{T}) \mathrm{Vec}\left(\mathbf{P}_C^T\right) \right\|_2^2 + \lambda_L \left\| \mathrm{Vec}\left(\mathbf{P}_C^T\right) \right\|_1 \\ &\quad + \sum_{k,r} \left(\lambda_G \sqrt{J_k} \left\| \mathrm{Vec}\left(\mathbf{P}_k^T\right) \right\|_2 + \lambda_E \left\| \mathrm{Vec}\left(\mathbf{P}_k^T\right) \right\|_{1,2} \right),\end{aligned} \tag{5}$$

subject to

$$\mathbf{T}^T\mathbf{T} = \mathbf{I}; \lambda_L, \lambda_E, \lambda_G \geq 0.$$

This minimization problem will be solved by alternatingly updating $\mathbf{T}$ and $\mathbf{P}_k$. Because the problem of minimizing Eq. (5) given a fixed $\mathbf{P}_k$ is a problem with known solution (see [20]), we shall focus on explaining how to update $\mathbf{P}_k$.

The minimization problem (Eq. (5)), given $\mathbf{T}$ is fixed, is not a standard minimization problem due to the penalties, and therefore