Chemometrics and Intelligent Laboratory Systems 106 (2011) 173–182

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Identification of significant factors by an extension of ANOVA–PCA based on multi-block analysis

D. Jouan-Rimbaud Bouveresse [a,b,*], R. Climaco Pinto [b,c], L.M. Schmidtke [d], N. Locquet [a,b], D.N. Rutledge [a,b]

[a] INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005, Paris, France
[b] AgroParisTech, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris, France
[c] Computational Life Science Cluster (CliC), KBC, UmeåUniversity, S-90187, Umeå, Sweden
[d] National Wine and Grape Industry Center, School of Agriculture and Wine Sciences, Charles Sturt University, Wagga Wagga, NSW 2650, Australia

ARTICLE INFO

Article history:
Received 14 December 2009
Received in revised form 10 May 2010
Accepted 13 May 2010
Available online 25 May 2010

Keywords:
Multi-block analysis
Common Component and Specific Weights Analysis
ComDim
ANOVA–PCA
F-test

ABSTRACT

A modification of the ANOVA–PCA method, proposed by Harrington et al. to identify significant factors and interactions in an experimental design, is presented in this article. The modified method uses the idea of multiple table analysis, and looks for the common dimensions underlying the different data tables, or data blocks, generated by the "ANOVA-step" of the ANOVA–PCA method, in order to identify the significant factors. In this paper, the "Common Component and Specific Weights Analysis" method is used to analyse the calculated multi-block data set. This new method, called AComDim, was compared to the standard ANOVA–PCA method, by analysing four real data sets. Parameters computed during the AComDim procedure enable the computation of *F*-values to check whether the variability of each original data block is significantly greater than that of the noise.

1. Introduction

Several multi-block analysis procedures exist for the simultaneous study of multiple sets of matrices with different variables describing the same samples (for example, see [1–4]). These methods may be useful in chemometrics to combine information about the same set of samples contained in signals acquired using different techniques (IR spectroscopy; Raman spectroscopy; physico–chemical analyses; etc.). One such multi-block technique is "Common Component and Specific Weights Analysis"—CCSWA [5].

The objective of multi-block analysis methods is to describe p data blocks observed for the same n samples (i.e. a set of p data matrices ($\mathbf{X}_i$, $i=1$ to p) each with n rows, but not necessarily the same number of variables). The method consists in determining a common space for all p data blocks, with each matrix having a specific contribution ("salience") to the definition of each dimension of this common space. This is done by finding the directions describing common distributions of the samples in the spaces defined by the different data blocks (hence the name *Common Component*, abbreviated CC or *Common Dimension*, abbreviated CD). *Salience* indicates the importance of each block in the construction of the common dimension, and a "percentage of variability extracted" by each dimension can be computed. The particular implementation of CCSWA used in this work, "ComDim", was developed and coded in Matlab [6] by D. Bertrand [7].

The work presented in this article shows that an interesting extension of ComDim is to use it in the analysis of sets of blocks calculated from a single initial data matrix. AComDim, presented here, is one such application, based on replacing the many separate PCAs performed in the ANOVA–PCA method [8], also abbreviated APCA, by a single analysis using ComDim. In this case, the various "Factor matrices" and "Interaction matrices" calculated from the initial data matrix are all analysed simultaneously, resulting in a series of "Common Components" along which the samples are distributed, each associated with a vector of "saliences" reflecting the importance of the contribution of each data block to the corresponding "Common Component".

After a brief presentation of both the ComDim and the APCA methods, this article will present several real case studies, showing the interest of this new method, particularly in comparison to the standard APCA method.

2. Theory

2.1. Notation

Matrices will be denoted by bold uppercase letters (e.g., $\mathbf{X}$), column vectors will be denoted by bold lowercase letters (e.g., $\mathbf{u}$), and row vectors by bold lowercase letters followed by the uppercase

* Corresponding author. INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005, Paris, France. Tel.: +33 1 44 08 16 39.
E-mail address: delphine.bouveresse@agroparistech.fr
(D. Jouan-Rimbaud Bouveresse).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.05.005

symbol T (e.g., $\mathbf{u}^T$), standing for "transposed". Scalars will be indicated by a letter in italics (e.g., N or n).

2.2. ComDim [5,9,10]

This procedure iteratively calculates, for each successive common dimension, a series of score vectors (coordinates of the n samples along the direction defined by that common dimension). Each block has a specific weight, called "salience", associated with each dimension of the common space. Significant differences in the values of the specific weights for a given dimension reflect the fact that the dimension contains information which was present in some blocks but not others. The main idea of the Common Dimension procedure, ComDim, is to calculate a weighted sum of the *sample* variance–covariance matrix (and not the *variable* variance–covariance matrix, as is usually the case in multivariate analysis) of each block, and then extract its first normed Principal Component as the first "Common Dimension" or "Common Component". The algorithm then iteratively calculates the weight of each block to the calculated CC. Finally, the percentage of variability extracted by the CC can be calculated. After the computation of the first CC, each original block matrix is deflated, and the procedure repeated for the calculation of the second Common Component, and so forth. Therefore, each Common Component is the first PC of a weighted sum of deflated matrices.

In order to present ComDim from an algorithmic point of view, one assumes a set of n samples is described by p sets of different variables. Hence, p matrices $\mathbf{X}_i$ of sizes $n \times k_i$ ($i = 1$ to p) are available, for which one wants to determine the Common Components.

Each matrix is first column-centered (to obtain $\mathbf{X}_{ic}$), and then normalised (division by its Frobenius norm), to obtain the scaled matrix $\mathbf{X}_{is}$. Although in certain cases, it can decrease the signal-to-noise ratio, the normalisation of the data matrices needs to be done to ensure that all data blocks have similar orders of magnitudes, so that no table predominates over the others, which would reduce the influence of the matrices with low orders of magnitude. Since the p original data blocks are column-centered and normed, p also corresponds to the total variance of the data at the beginning of the procedure. A parameter *unexpl* is set equal to p:

$$unexpl = p \tag{1}$$

For each $\mathbf{X}_{is}$, a matrix sample variance–covariance $\mathbf{W}_i$ of dimensions $n \times n$ is computed as:

$$\mathbf{W}_i = \mathbf{X}_{is} * \mathbf{X}_{is}^T \tag{2}$$

The Common Components are computed in an iterative fashion. At each iteration, the weighted sum of the p $\mathbf{W}_i$ matrices is computed, resulting in a global $\mathbf{W_G}$ matrix. In the first iteration, all the weights, λ_i, are set to 1.

$\mathbf{W_G} = 0$;
for i = 1 to p

$\lambda_i = 1$;

$$\mathbf{W_G} = \mathbf{W_G} + \lambda_i * \mathbf{W}_i \tag{3}$$

end

$\mathbf{W_G}$ is then decomposed by singular value decomposition (SVD), yielding $\mathbf{U_W}$ (matrix of row-singular vectors), $\mathbf{S_W}$ (diagonal matrix with the singular values sorted in decreasing order), and $\mathbf{V_W}$ (matrix of column-singular vectors):

$$\mathbf{W_G} = \mathbf{U_W} * \mathbf{S_W} * \mathbf{V_W}^T \tag{4}$$

The first column of $\mathbf{U_W}$ (i.e., the normed score vector of $\mathbf{W_G}$ associated with the largest singular value) is chosen as the first estimation of the "Common Component score" of $\mathbf{W_G}$, denoted as $\boldsymbol{q}$. A new estimation of λ_i is calculated using $\boldsymbol{q}$ and $\mathbf{W_G}$, and an *unfit* value is then determined as a function of the updated λ_i values:

$unfit = 0$;
$\boldsymbol{q} = \mathbf{U_W}(:,1)$;
for $i = 1$ to p

$$\lambda_i = \boldsymbol{q}^T * \mathbf{W}_i * \boldsymbol{q} \tag{5}$$

$$\mathbf{Aux} = \mathbf{W}_i - \lambda_i \times \boldsymbol{q} * \boldsymbol{q}^T \tag{6}$$

$$unfit = \mathrm{sum}(\mathrm{sum}(\mathbf{Aux}.\ *\mathbf{Aux})) \tag{7}$$

end

where the symbol '.' before the product symbol means that each element of **Aux** is multiplied by itself.

Aux is a "residuals" matrix of the variability unaccounted for by the Common Components calculated up to that point. The *unfit* value is the variance of all the blocks unexplained by all those CCs.

The calculation of $\mathbf{W_G}$ (from the updated λ_i values—Eq. (3)), then $\boldsymbol{q}$ (after SVD of the updated $\mathbf{W_G}$—Eq. (4)), and then λ_i is iterated (as in Eqs. (5)–(7)) until convergence of *unfit*. The final $\boldsymbol{q}$ vector is the first Common Component (its elements being equivalent to normalized scores). The final λ_i value indicates the weight of the original $\mathbf{X_i}$ in the Common Component ("salience"), and reflects the dispersion of the samples along that dimension, and so can be seen as a measure of variance.

A percentage of variance contained in the Common Components is given by:

$$expl = 100 \times (unexpl - unfit^{\circ}) / p \tag{8}$$

unexpl is then updated as:

$$unexpl = unfit \tag{9}$$

Each $\mathbf{X_{is}}$ data matrix is then "deflated" (Eqs. (10) and (11)) using the normalized scores vectors:

$$\mathbf{Aux} = \mathbf{I} - \boldsymbol{q} * \boldsymbol{q}^T \tag{10}$$

$$\mathbf{X}_{is} = \mathbf{Aux} * \mathbf{X}_{is} = \left(\mathbf{I} - \boldsymbol{q} * \boldsymbol{q}^T\right) * \mathbf{X}_{is} \tag{11}$$

(where $\mathbf{I}$ is the $n \times n$ Identity matrix).

New estimations of $\mathbf{W}_i$ are computed from these "deflated" $\mathbf{X}_{is}$ matrices, and the following Common Components are computed as before.

2.3. APCA [8]

Analysis of variance—principal component analysis (APCA) was introduced in 2005 by Harrington et al. [8] for the detection of biomarkers in high dimensional proteomic data sets. Since then, it has been applied in several other situations [11–13]. The aim of the method is to determine whether known characteristics of the samples (or "factors", in the Experimental Design terminology) produce a variation in the data which is significantly larger than the variations due to noise. A clear explanation of the method is given in [13]. To briefly summarise the method here, one assumes that f factors, each described by l_f levels, are known for matrix $\mathbf{X}$. First, $\mathbf{X}$ is column-centered, yielding $\mathbf{X_c}$. A "Factor 1 matrix" $\mathbf{M_1}$ is created from $\mathbf{X_c}$, by replacing each row by the average vector of all rows whose level for