A criterion-based PLS approach to multi-block and multi-group data analysis

Michel Tenenhaus

Abstract Each On the one hand, multi-block data analysis concerns the analysis of several sets of variables (blocks) observed on the same set of individuals. On the other hand, multi-group data analysis concerns the analysis of one set of variables observed on a set of individuals taking into account a group-structure at the level of the individuals. Two types of partition of an individuals×variables data matrix X are then defined. In the multi-block framework, the column partition $X=[X_1\ldots X_s]$ is considered. In this case, each block $X_j$ is an $n \times p_j$ data matrix and represents a set of $p_j$ variables observed on a set of $n$ individuals. The number and the nature of the variables differ from one block to another but the individuals must be the same across blocks. In the multi-group framework, the row partition $X=[X_1\ldots X_s\ldots X_I]$ is considered. In this framework, the same set of variables is observed on different groups of observations. Each matrix $X_i$ is an $n_i \times p$ data matrix, is called a group in this paper and represents a set of $p$ variables observed on a set of $n_i$ individuals. The number of observations of each block could differ from one block to another. Many methods exist for multi-block and multi-group data analysis. Regularized Generalized Canonical Correlation Analysis (RGCCA) has been proposed in Tenenhaus & Tenenhaus (2011) and appeared to include an amazing large number of criterion-based multi-block data analysis methods as particular cases. In this paper, we intend to extend RGCCA so that it can also be a unifying tool for multi-group data analysis. Only first dimension components will be discussed in this paper. Components related to other dimensions can be obtained by following the same procedures on deflated blocks or groups with respect to the previous dimension components.

Key-word: Multi-block data analysis, multi-group data analysis, regularized generalized canonical correlation analysis

1 Multi-block data analysis (for column partition)

With each $n \times p_j$ block $X_j$, it is useful to associate a $p_j \times p_j$ positive definite symmetric matrix $M_j$. This matrix $M_j$ is used to define a metric in the individual space $\mathbb{R}^{p_j}$. Moreover, a design matrix $C = \{c_{jk}\}$ is defined with $c_{jk} = 1$ if blocks $j$ and $k$ are related, and $= 0$ otherwise. The following optimization problem is defined:

---

1 Michel Tenenhaus

HEC Paris, 1 rue de la Libération, 78351 Jouy-en-Josas, e-mail: tenenhaus@hec.fr
Maximize \( \sum_{j=1}^{J} c_j g(\text{cov}(X_j, a_j, X_j a_j)) \)
subject to the constraints \( a_j^T M_j a_j = 1, j = 1, \ldots, J \) \( \) (1)

where \( g \) stands for the identity, the absolute value or the square function. Regularized
generalized canonical correlation analysis is a special case of problem (1) with \( M_j = \tau_j \mathbf{1} + (1-\tau_j)(1/n_j)X_j^T X_j \), the shrinkage constant \( \tau_j \) varying between 0 and 1. A monotone
convergent PLS algorithm similar to the PLS algorithm for RGCCA can be proposed to solve
optimization problem (1). Many methods are special cases of problem (1). They are
described in Tenenhaus & Tenenhaus (2011).

2 Multi-group data analysis (for row partition)

With each \( n_i \times p \) group \( X_i \), it is useful to associate a \( p \times p \) positive definite symmetric matrix
\( M_i \). This matrix \( M_i \) is used to defining a metric specific to group \( i \) in the individual space
\( \mathbb{R}^p \). A design matrix \( C = \{c_i \} \) and a function \( g \) are defined in the same way as for problem
(1).
The following optimization problem is defined:
Maximize \( \sum_{i=1}^{I} c_i g((1/n_i)X_i^T X_i a_i, (1/n_i)X_i^T X_i a_i) \)
subject to the constraints \( a_i^T M_i a_i = 1, i = 1, \ldots, I \) \( \) (2)
A monotone convergent PLS algorithm, similar to the PLS algorithm for RGCCA, is
available for solving optimization problem (2). We now describe some optimization
problems that are special cases of optimization problem (2).

\( \text{PCA with group-structure} \)

All columns of each group \( X_i \) are supposed to be standardized (mean zero and unit variance).

We consider a situation where all groups are related: \( c_i = 1 \) for \( i \neq \ell \) and \( 0 \) otherwise. We
consider the following metric for each group: \( M_i = (1/n_i)X_i^T X_i \). We consider the case where the function \( g \) is the identity. Then problem (2) for this specific situation can be written as:
Maximize \( \sum_{i=1}^{I} \text{cos}(1/n_i)X_i^T X_i a_i, (1/n_i)X_i^T X_i a_i) \| (1/n_i)X_i^T X_i a_i \| \| (1/n_i)X_i^T X_i a_i \|
subject to the constraints \( \text{Var}(X_i a_i) = 1, i = 1, \ldots, I \) \( \) (3)

The elements of the vector \( (1/n_i)X_i^T X_i a_i \) represent correlations between the various columns of
the group \( X_i \) and the group component \( X_i a_i \). The vector \( a_i \) is called “vector of weights”
and the vector \( (1/n_i)X_i^T X_i a_i \) is called “vector of loadings”. The maximization of \( \| (1/n_i)X_i^T X_i a_i \| \) subject to the constraint \( \text{Var}(X_i a_i) = 1 \) yields the first standardized principal
component of \( X_i \) for which weights and loadings are proportional. Therefore, problem (3)
leads to a compromise between separate one dimension PCA’s of the various groups \( X_i \)’s and
vectors of loadings with small angles.

\( \text{Global-PCA with group structure} \)

In addition to the context of problem (3), we consider a super-group \( X_{1:\ell} \) identical to the
global data matrix \( X \). We consider a situation where all groups 1 to \( I \) are only related to the
super-group \( X_{i+1} : c_{\ell i} = 1 \) for \( i = 1, \ldots, I \) and \( \ell = I + 1 \) and 0 otherwise. We consider a function \( g \) equal to the square function.

Then optimization problem (2) for this specific situation can be written as:

Maximize

\[
\sum_{i=1}^{I} \cos^2\left(\frac{1}{\sigma_i} \mathbf{X}_i \mathbf{a}_i, \frac{1}{n_i} \mathbf{X}_i^T \mathbf{X}_i, \mathbf{a}_{i+1}\right) \times \left\| \frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\|^2
\]

subject to the constraints \( \text{Var}(\mathbf{X}_i, \mathbf{a}_i) = 1, \ i = 1, \ldots, I + 1 \). (4)

Optimization problem (4) leads to a compromise between the various one dimension PCA’s of the \( \mathbf{X}_i \)'s, one dimension PCA of \( \mathbf{X}_{i+1} = \mathbf{X} \) and vectors of loadings with small angles. The various compromise one dimension group PCA’s will tend to be close to the one dimension compromise global PCA Qannari et al. (2011).

Other optimization problems for multi-group data analysis have been proposed in Qannari et al. (2011). Two of them are special cases of problem (2). Their first optimization problem is similar to

Maximize

\[
\sum_{i=1}^{I} \cos^2\left(\frac{1}{\sigma_i} \mathbf{X}_i \mathbf{a}_i, \frac{1}{n_i} \mathbf{X}_i^T \mathbf{X}_i, \mathbf{a}_{i+1}\right) \times \left\| \frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\|^2
\]

subject to the constraints \( \text{Var}(\mathbf{X}_i, \mathbf{a}_i) = 1, \ i = 1, \ldots, I + 1 \). (5)

Optimization problem (5) will yield a solution for which the angles between vectors of loadings for each group and vector of loadings for the super-group are as small as possible. However, it does not take into account the level of these loadings. Their second optimization problem is similar to

Maximize

\[
\sum_{i=1}^{I} \cos^2\left(\frac{1}{\sigma_i} \mathbf{X}_i \mathbf{a}_i, \frac{1}{n_i} \mathbf{X}_i^T \mathbf{X}_i, \mathbf{a}_{i+1}\right) \times \left\| \frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\|^2
\]

subject to the constraints: \( \text{Var}(\mathbf{X}_i, \mathbf{a}_i) = 1, \ i = 1, \ldots, I \) and \( \left\| \frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\|^2 = 1 \). (6)

The solution of problem (6) amounts to performing a one dimension PCA of \( \mathbf{X} \), taking into account that each group \( \mathbf{X}_i \) is column-centered. Optimization problem (6) leads to a compromise between one dimension PCA’s of the \( \mathbf{X}_i \)'s, one dimension PCA of \( \mathbf{X}_{i+1} = \mathbf{X} \) and small angles between the associated vectors of loadings. In other words, the compromise one dimension group PCA’s will tend to be close to the one dimension global PCA.

2.1 Regularization

In the resolution of optimization problems (3), using a PLS algorithm, the constraint \( \text{Var}(\mathbf{X}_i, \mathbf{a}_i) = 1 \) imply the inversion of \( \mathbf{X}_i^T \mathbf{X}_i \). This implies more individuals than variables for group \( i \) and no severe multicollinearity problem. These conditions are not always satisfied. Therefore we propose a more stable version of optimization problems (3) by considering the following optimization problem:

Maximize

\[
\sum_{i=1}^{I} \cos\left(\frac{1}{\sigma_i} (\frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i, (\frac{1}{n_i} \mathbf{X}_i^T \mathbf{X}_i) \mathbf{a}_{i+1}\right) \times \left\| (\frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\| \left\| (\frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\|
\]

subject to the constraints \( \left\| \mathbf{a}_i \right\| = 1, \ i = 1, \ldots, I \). (7)

The maximization of \( \left\| (\frac{1}{n_i} \mathbf{X}_i \mathbf{a}_i \right\| \) subject to the constraint \( \left\| \mathbf{a}_i \right\| = 1 \) yields the first principal component of \( \mathbf{X}_i \). Therefore, problem (7) leads also to a compromise between one dimension
PCA’s of the various groups $X_i$’s and vectors of loadings with small angles. Groups with important first principal components weight more in problem (7) than in problem (3). It is also possible to consider a compromise between problems (3) and (7) by introducing a shrinkage constant $\tau_i$ varying between 0 and 1. This gives the following optimization problem:

Maximize $\sum_{i,j=1}^I \cos((1/n_i)X'_iX_j a_i, (1/n_j)X'_jX_i a_i) \times \|X'_iX_j a_i\| \times \|X'_jX_i a_i\|$

subject to the constraints $\tau_i \|a_i\|^2 + (1 - \tau_i) \text{Var}(X_i a_i) = 1, \, i = 1,...,I$.  

(8)

The same constraint modifications can be proposed for optimization problems (4), (5) and (6).

3 A unified approach to multi-block and multi-group data analysis

Optimization problems (1) and (2) can in fact be written as the following unique optimization problem:

Maximize $\sum_{i,j=1}^I c_{ij} g(\langle Q_i a_i, Q_j a_j \rangle)$

subject to the constraints $a_i^TM_i a_i = 1, i = 1,...,I$  

(9)

where the dimensions of the various matrices $Q_i$ are $m \times p_i$ and where the matrices $M_i$ are symmetric definite positive and of dimensions $p_i \times p_i$. In optimization problem (1) $Q_i = X_i$, where the columns of the matrix $X_i$ are obtained by centering the columns of block $X_i$, and where $m$ is equal to the number $n$ of individuals. In optimization problem (2), $Q_i = (1/n_i) X_i^T X_i$, where the columns of the matrix $X_i$ are obtained by standardizing the columns of group $X_i$, and where $m$ and $p_i$ are both equal to the number $p$ of variables. By setting $b_i = M_i^{1/2} a_i$ and $P_i = Q_i M_i^{-1/2} P_i$, optimization problem (9) becomes

Maximize $\sum_{i,j=1}^I c_{ij} g(\langle P_i b_i, P_j b_j \rangle)$

subject to the constraints $b_i^T b_j = 1, i = 1,...,I$  

(10)

A monotone convergent PLS algorithm similar to the PLS algorithm for RGCCA can be developed for optimization problem (10).

References
