Multiway Regularized Generalized Canonical Correlation Analysis

Arthur Tenenhaus, Laurent Le Brusquet, Gisela Lechuga

Multiway Regularized Generalized Canonical Correlation Analysis

Arthur Tenenhaus [1,2], Laurent Le Brusquet [1], Gisela Lechuga [1].

[1] *Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506) CNRS - CentraleSupelec - Université Paris-Sud - 3, rue Joliot Curie 91192, Gif-sur-Yvette , prenom.nom@centralesupelec.fr*

[2] *Bioinformatics/Biostatistics Platform IHU-A-ICM, Brain and Spine Institute 47-83, Bd de l'hôpital; Paris, France*

Résumé. L'Analyse Canonique Généralisée Régularisée (RGCCA) permet l'étude des relations entre différents blocs de données. Dans ce papier, une version multivoie de RGCCA (MGCCA) est proposée. MGCCA cherche à décrire et comprendre les relations entre tenseurs.

Mots-clés. Analyse Canonique Généralisée, Données Multivoie

Abstract. Regularized Generalized Canonical Correlation Analysis (RGCCA) is currently geared for the analysis two-way data matrix. In this paper, multiway RGCCA (MGCCA) extends RGCCA to the multiway data configuration. More specifically, MGCCA aims at studying the complex relationships between a set of three-way data table.

Keywords. Generalized Canonical Correlation, Multiway data

1 Introduction

On the one hand, multiblock analysis concerns the analysis of data structured in blocks of variables. In this framework, a column partition $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_l, \ldots, \mathbf{X}_L]$ is considered. Each $I \times J_l$ data matrix $\mathbf{X}_l$ is called a block and represents a set of J_l variables observed on I individuals. The number and nature of the variables usually differ from one block to another but the individuals are the same across blocks. The main aim is to investigate the relationships between blocks. It appears that Regularized Generalized Canonical Correlation Analysis (RGCCA) [Tenenhaus and Tenenhaus, 2011], is a general framework for multiblock data analysis.

On the other hand, suppose that measurements are avalaible from, for instance, I individuals on J variables measured at K occasions. In the litterature, such data is called three-way data because per occasion measurements are available for the same group of individuals on the same set of variables. These three-way data can be collected in an $I \times J \times K$ three-way array $\underline{\mathbf{X}}$, with frontal slices $\mathbf{X}_k$ containing the $I \times J$ data matrices

for each of the K occasions.

Several examples of either three-way data or multiblock data can be found in a variety of domains including chemometrics, psychometry, bioinformatics to name but a few. Nowadays, it frequently occurs to encounter data combining multiway and multiblock structures. For instance, neuroimaging is increasingly recognised as an intermediate phenotype to understand the complex path between genetics and behavioural or clinical phenotypes. In this context, the goal is primarily to identify a set of genetic biomarkers that explains some neuroimaging variabilities which implies some modifications of the behavioural. Often, the neuroimaging and the behavioral variabilities are evaluated througth time. Thus, it is crucial to perform multiple experiments (e.g. SNPs, functional MRI across time , behavioural data across time) on a single set of patients and the joint analysis of dataset that gathers two-way and three-way data becomes more and more crucial. RGCCA is currently geared for the analysis two-way data matrices. In this paper, multiway RGCCA (MGCCA) extends RGCCA to the multiway data configuration. More specifically, MGCCA aims at studying the complex relationships between a set of multi-way data table.

2 Regularized Generalized Canonical Correlation analysis

Let us consider L data blocks $\mathbf{X}_1, \ldots, \mathbf{X}_l, \ldots, \mathbf{X}_L$. Each block $\mathbf{X}_l$ is of dimension $I \times J_l$. We also associate to each matrix $\mathbf{X}_l$ a symmetric definite positive matrix $\mathbf{M}_l$ of dimensions $J_l \times J_l$. In addition, a design matrix $\mathbf{C} = (c_{jk})$ is defined with $c_{jk} = 1$ if blocks $\mathbf{X}_j$ and $\mathbf{X}_k$ are connected, and $= 0$ otherwise. RGCCA for multi-block data analysis is defined as the following optimization problem.

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_L}{\text{maximize}} \sum_{j,k=1:j\neq k}^{L} c_{jk}\ \mathrm{g}\left(\mathrm{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)\right)\ \text{s.t. } \mathbf{w}_l^\top \mathbf{M}_l \mathbf{w}_l = 1, l = 1, \ldots, L \qquad (1)$$

with $g(x)$ a convex function of the scalar x. Typical choices of $g(x)$ are the identity, the absolute value or the square function. The vector $\mathbf{w}_l$ is called "vector of weights", the vector $\mathbf{X}_l\mathbf{w}_l$ is called "block component". RGCCA is recovered through the optimization problem (1) with $\mathbf{M}_l = \tau_l \mathbf{I}_{J_l} + (1-\tau_l)\mathbf{X}_l^\top \mathbf{X}_l$ and the shrinkage constant τ_l varying between 0 and 1.

By setting $\mathbf{a}_l = \mathbf{M}_l^{1/2}\mathbf{w}_l$ and $\mathbf{P}_l = \mathbf{X}_l\mathbf{M}_l^{-1/2}$, the optimization problem (1) becomes:

$$\underset{\mathbf{a}_1,\ldots,\mathbf{a}_L}{\text{maximize}} \sum_{j,k=1:j\neq k}^{L} c_{jk}\ \mathrm{g}\left(\mathrm{cov}(\mathbf{P}_j\mathbf{a}_j, \mathbf{P}_k\mathbf{a}_k)\right)\ \text{s.t. } \mathbf{a}_l^\top \mathbf{a}_l = 1, l = l, \ldots, L \qquad (2)$$

A monotone convergent algorithm (i.e. the bounded criteria to be maximized increases at each step of the procedure) is proposed in [Tenenhaus and Tenenhaus, 2011] and in [Tenenhaus and Tenenhaus, 2014] for the resolution of optimization problem (2).

3 Multiway RGCCA

Let us consider L up to order 3 tensors $\underline{\mathbf{X}}_1, \ldots, \underline{\mathbf{X}}_L$. Each tensor $\underline{\mathbf{X}}_l$ is of dimension $I \times J_l \times K_l$. The shared dimension across tensors is the row-mode. Let us consider their corresponding matricized versions $\mathbf{X}_1, \ldots, \mathbf{X}_L$. Each matrix $\mathbf{X}_l = [\mathbf{X}_{l1}, \ldots, \mathbf{X}_{l,K_l}]$ is of dimension $I \times J_l K_l$ and represents all the frontal slices of the tensor next to each other. We also associate to each $\underline{\mathbf{X}}_l$ a symmetric definite positive matrix $\mathbf{M}_l$ of dimensions $J_l K_l \times J_l K_l$ which usually has a block structure. A design matrix $\mathbf{C} = (c_{jk})$ is also defined with $c_{jk} = 1$ if $\underline{\mathbf{X}}_j$ and $\underline{\mathbf{X}}_k$ are connected, and $= 0$ otherwise. Taking into account the network of connections between the three-way data, our objective is to study the relationships between them. Using the same transforms as the ones used previously, $\mathbf{a}_l = \mathbf{M}_l^{1/2}\mathbf{a}_l$ and $\mathbf{P}_l = [\mathbf{P}_{l1}, \ldots, \mathbf{P}_{l,K_l}] = [\mathbf{X}_{l1}, \ldots, \mathbf{X}_{l,K_l}]\mathbf{M}_l^{-1/2}$, multi-way RGCCA (MGCCA) is defined as the following optimization problem:

$$\begin{cases} \underset{\mathbf{a}_1,\ldots,\mathbf{a}_L}{\text{maximize}}\ f(\mathbf{a}_1, ..., \mathbf{a}_L) = \displaystyle\sum_{j,k=1:j\neq k}^{L} c_{jk}\ \mathrm{g}\left(\langle \mathbf{P}_j\mathbf{a}_j, \mathbf{P}_k\mathbf{a}_k\rangle\right) \\ \\ \text{s.t. } \mathbf{a}_l^\top \mathbf{a}_l = 1 \text{ and } \mathbf{a}_l = \mathbf{c}_l \otimes \mathbf{b}_l, l = 1, \ldots, L \end{cases} \tag{3}$$

The sole difference between RGCCA and MGCCA relies on the kronecker constraints (structural constraints) that are applied on the outer weight vectors. These structural constraints are usual in the multi-way literature (see for instance [Bro, 1996]).

Two general set-ups constitute the internal mechanism for the maximization of the optimization problem (3). In the first, the function is to be maximized over different parameter vectors (i.e. $\mathbf{a}_1, \ldots, \mathbf{a}_J$), and this is approached by updating each of the parameter vectors in turn, keeping the others fixed. If each update improves the function value, the function will be optimized gradually over the complete set of parameter vectors. This principle is called block relaxation by [De Leeuw, 1994]. The second set-up relies on iterative majorization [Hunter and Lange, 2004].

Actually, at each block relaxation substep the core optimization that has to be considered in MGCCA is:

$$\mathbf{a}_l^{s+1} = \underset{\mathbf{a},\|\mathbf{a}\|=1}{\text{argmax}}\ \langle \mathbf{P}_l\mathbf{a}, \mathbf{z}_l^s\rangle \ \text{ subject to } \ \mathbf{a} = \mathbf{c} \otimes \mathbf{b}. \tag{4}$$

where $\mathbf{z}_l^s$ is the so-called inner component defined by

$$\mathbf{z}_l^s = \sum_{k<l} c_{lk} \mathrm{g}' \left(\langle \mathbf{P}_l \mathbf{a}_l^s, \mathbf{P}_k \mathbf{a}_k^{s+1} \rangle \right) \mathbf{P}_k \mathbf{a}_k^{s+1} + \sum_{k>l} c_{lk} \mathrm{g}' \left(\langle \mathbf{P}_l \mathbf{a}_l^s, \mathbf{P}_k \mathbf{a}_k^s \rangle \right) \mathbf{P}_k \mathbf{a}_k^s \tag{5}$$

The optimization problem (4) boils down to find a set of weight vectors $\mathbf{c}$ and $\mathbf{b}$ that produce a block component $\mathbf{y}_l = \mathbf{P}_l \mathbf{a}_l$ with maximal scalar product with $\mathbf{z}_l$. This problem is equivalent to:

$$\begin{aligned}
\mathbf{a}_l^{s+1} &= \underset{\mathbf{b},\mathbf{c};\|\mathbf{c}\otimes\mathbf{b}\|=1}{\operatorname{argmax}} (\mathbf{z}_l^s)^\top \mathbf{P}_l (\mathbf{c} \otimes \mathbf{b}) = \underset{\mathbf{b},\mathbf{c};\|\mathbf{c}\otimes\mathbf{b}\|=1}{\operatorname{argmax}} (\mathbf{z}_l^s)^\top \left[\mathbf{P}_l \left(\mathbf{c} \otimes \mathbf{I}_{J_l} \right) \right] \mathbf{b} \\
&= \underset{\mathbf{b},\mathbf{c};\|\mathbf{c}\otimes\mathbf{b}\|=1}{\operatorname{argmax}} (\mathbf{z}_l^s)^\top \left[\sum_{k=1}^{K} c_k \mathbf{P}_{lk} \right] \mathbf{b} = \underset{\mathbf{b},\mathbf{c};\|\mathbf{c}\otimes\mathbf{b}\|=1}{\operatorname{argmax}} \left[\sum_{k=1}^{K} c_k (\mathbf{z}_l^s)^\top \mathbf{P}_{lk} \right] \mathbf{b} \\
&= \underset{\mathbf{b},\mathbf{c};\|\mathbf{c}\otimes\mathbf{b}\|=1}{\operatorname{argmax}} \mathbf{c}^\top \mathbf{Q}_l \mathbf{b}
\end{aligned} \tag{6}$$

where $\mathbf{Q}_l$ is a $K_l \times J_l$ matrix defined by $\mathbf{Q}_l = [\mathbf{P}_{l1}^\top \mathbf{z}_l^s, \ldots, \mathbf{P}_{lK_l}^\top \mathbf{z}_l^s]^\top$.

From equality (6), we deduce that $\mathbf{c}$ and $\mathbf{b}$, solution of the optimization problem (4), are the first left and right singular vectors of the matrix $\mathbf{Q}_l$. It appears that $\mathbf{c}$ and $\mathbf{b}$ are constrained to be normalized, satisfying automatically the unit norm constraint on $\mathbf{a}_l$. Note that a similar optimization is found for Multilinear Partial Least Squares [Bro, 1996].

The entire MGCCA algorithm is described in Algorithm 1.

Data: $\mathbf{X}_j$s, τ_js, g, ε
Result: $\mathbf{a}_j$s
Initialization: choose random unit norm $\mathbf{a}_l^0$ for each $l = 1, \ldots, L$;
$s = 0$;
while $f(\mathbf{a}_1^{s+1}, \ldots, \mathbf{a}_J^{s+1}) - f(\mathbf{a}_1^s, \ldots, \mathbf{a}_J^s) < \varepsilon$ **do**
  for $l = 1, 2, \ldots, L$ **do**
    - $\mathbf{z}_l^s = \sum_{k<l} c_{lk} \mathrm{g}' \left(\langle \mathbf{P}_l \mathbf{a}_l^s, \mathbf{P}_k \mathbf{a}_k^{s+1} \rangle \right) \mathbf{P}_k \mathbf{a}_k^{s+1} + \sum_{k>l} c_{lk} \mathrm{g}' \left(\langle \mathbf{P}_l \mathbf{a}_l^s, \mathbf{P}_k \mathbf{a}_k^s \rangle \right) \mathbf{P}_k \mathbf{a}_k^s$
    - $\mathbf{a}_l^{s+1} = \mathbf{c}_l^{s+1} \otimes \mathbf{b}_l^{s+1}$

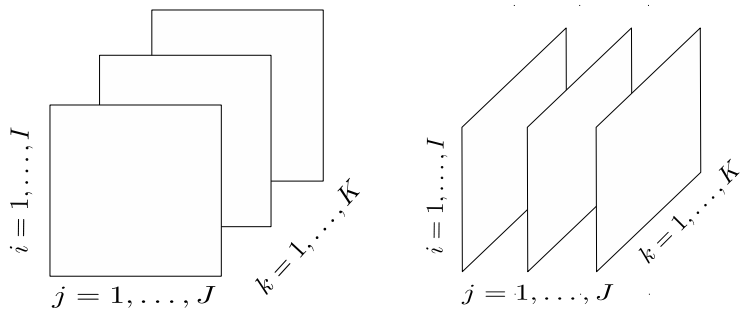
    where $\mathbf{c}_l^{s+1}$ and $\mathbf{b}_l^{s+1}$ are obtained as the first left and right singular vectors of the matrix $\mathbf{Q}_l = [\mathbf{P}_{l1}^\top \mathbf{z}_l^s, \ldots, \mathbf{P}_{lK_l}^\top \mathbf{z}_l^s]^\top$ of dimension $K_l \times J_l$.
  end
  $s = s + 1$;
end

Algorithm 1: RGCCA algorithm for three-way data analysis

4 Alternative formulation and interpretations

This section proposes alternative expressions of the matricized matrices that appear in the MGCCA algorithm. This reformulation allows complementary views/interpretations.

New notations are introduced in this section. Let $\mathbf{P}_l = [\mathbf{P}_{l1}, \ldots, \mathbf{P}_{lK_l}]$ and $\mathbf{P}_k = [\mathbf{P}_{k1}, \ldots, \mathbf{P}_{kK_k}]$, be two matricized matrices of dimension $I \times J_l K_l$ and $I \times J_k K_k$. Let $\mathbf{a}_l = \mathbf{c}_l \otimes \mathbf{b}_l$ and $\mathbf{a}_k = \mathbf{c}_k \otimes \mathbf{b}_k$ be the corresponding outer weight vectors. Let $\mathbf{P}^l_{..k} = \mathbf{P}_{kl}$ be the k^{th} frontal slice of $\underline{\mathbf{P}}_l$ of dimension $I \times J_l$ and $\mathbf{P}^l_{.j.}$ be the j^{th} lateral slice of $\underline{\mathbf{P}}_l$ of dimension $I \times K_l$. Figure 4 depicts the frontal and lateral slices that are handled within MGCCA.



(a) Frontal slices (b) Lateral slices

Figure 1: Frontal ($\mathbf{P}_{..k}$) and lateral ($\mathbf{P}_{.j.}$) slices of the tensor $\underline{\mathbf{X}}$ that are handled within MGCCA

$$\mathbf{P}_l\mathbf{a}_l = \mathbf{P}_l(\mathbf{I}_{K_l} \otimes \mathbf{b}_l)\mathbf{c}_l = \left(\sum_{j=1}^{J_l} b_{lj}\mathbf{P}^l_{.j.}\right)\mathbf{c}_l \tag{7}$$

$$= \mathbf{P}_l(\mathbf{c}_l \otimes \mathbf{I}_{J_l})\mathbf{b}_l = \left(\sum_{k=1}^{K_l} c_{lk}\mathbf{P}^l_{..k}\right)\mathbf{b}_l \tag{8}$$

From equations (7) and (8), we conclude that the outer component $\mathbf{P}_l\mathbf{a}_l$ can be expressed as a linear combination of the columns of either $\sum_{j=1}^{J_l} b_{lj}\mathbf{P}^l_{.j.}$ or $\sum_{k=1}^{K_l} c_{lk}\mathbf{P}^l_{..k}$. It yields that the dot product between components can be expressed as follows:

$$\langle \mathbf{P}_l\mathbf{a}_l, \mathbf{P}_k\mathbf{a}_k \rangle = (\mathbf{c}_l \otimes \mathbf{b}_l)^t\mathbf{P}_l^t\mathbf{P}_k(\mathbf{c}_k \otimes \mathbf{b}_k)$$

$$= \mathbf{b}_l^t\left[\sum_{h=1}^{K_l} c_{lh}\mathbf{P}^l_{..h}\right]^t\left[\sum_{h=1}^{K_k} c_{kh}\mathbf{P}^k_{..h}\right]\mathbf{b}_k \tag{9}$$

$$= \mathbf{c}_l^t\left[\sum_{j=1}^{J_l} b_{lj}\mathbf{P}^l_{.j.}\right]^t\left[\sum_{j=1}^{J_k} b_{kj}\mathbf{P}^k_{.j.}\right]\mathbf{c}_k \tag{10}$$

From expressions (7)-(10), it appears that the matricization operations can be avoided and replaced by weigthed means of either the lateral or the frontal slices of the tensors $\underline{\mathbf{X}}_l$ or $\underline{\mathbf{X}}_k$.

5 Conclusion

We have shown that a simple monotone convergent algorithm can be used for multi-block/multiway data. A lot of experiences have been gathered in multi-block data analysis. The field of multiblock/multiway data analysis has been less explored. Our proposal of using RGCCA for multiblock/multiway data analysis is new and opens a new field to be explored.

References

[Bro, 1996] Bro, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, 10:47–61.

[Tenenhaus and Tenenhaus, 2011] Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76:257–284.

[Tenenhaus and Tenenhaus, 2014] Tenenhaus, A. and Tenenhaus, M. (2014). Regularized Generalized Canonical Correlation Analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, 238:391–403.

[De Leeuw, 1994] De Leeuw, J. (1994). Block relaxation algorithms in statistics. *Information Systems and Data Analysis, Springer, Berlin, Bock, H.H. Lenski W. Richter M.M. (Eds.)*, 308–325.

[Hunter and Lange, 2004] Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.