Multiway Generalized Canonical Correlation Analysis

Arnaud Gloaguen, Cathy Philippe, Vincent Frouin, Laurent Le Brusquet, Arthur Tenenhaus

Chimiométrie XIX - 2018

Abstract/Résumé

Multiway Generalized Canonical Correlation Analysis

Arnaud Gloaguen[1,2] Cathy Philippe[2] Vincent Frouin[2] Laurent Le Brusquet[1] Arthur Tenenhaus[1,3]

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

[2] NeuroSpin/UNATI – CEA, Université Paris-Saclay – France, prenom.nom@cea.fr

[3] Plateforme de Bioinformatique/Biostatistiques IHU-A-ICM, Institut du Cerveau et de la Moelle épinière - Paris

Keywords: Multiway data analysis, PLS, Canonical Correlation Analysis, RGCCA.

Regularized generalized canonical correlation analysis (RGCCA) is a general framework that encompasses several important multivariate analysis methods. In this paper, we extend RGCCA to the case where data blocks have a tensor structure.

1 Introduction

We consider L data matrices $\mathbf{X}_1, \dots, \mathbf{X}_l, \dots, \mathbf{X}_L$. 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 the nature of the variables may differ from one block to another, but the individuals must be the same across blocks. We assume that all variables are centered. The objective of RGCCA is to find block components $\mathbf{y}_l = \mathbf{X}_l \mathbf{w}_l, l = 1, \dots, L$ (where the block-weight vector $\mathbf{w}_l$ is a column-vector with J_l elements) summarizing the relevant information between and within the blocks. The second generation RGCCA, detailed in [1], subsumes fifty years of multiblock component methods and is defined as the following optimization problem:

$$\max_{\mathbf{w}_1,\dots,\mathbf{w}_L} \sum_{k,l=1}^{L} c_{kl}\, g\left(I^{-1}\mathbf{w}_k^\top \mathbf{X}_k^\top \mathbf{X}_l \mathbf{w}_l\right) \quad \text{s.t.} \quad \mathbf{w}_l^\top \mathbf{M}_l \mathbf{w}_l = 1 \quad, l = 1, \dots, L \qquad (1)$$

With g a convex function, $\mathbf{M}_l$ a positive definite matrix and $\mathbf{C} = (c_{kl})$ a symmetric $L \times L$ matrix of nonnegative elements describing the network of connections between blocks that the user wants to take into account. Usually, $c_{kl} = 1$ for two connected blocks and 0 otherwise.

2 Multiway Generalized Canonical Correlation Analysis

In this paper, we adopt the standardized notations and terminology proposed by [2] for multiway data analysis. Let us consider L tensors $\underline{\mathbf{X}}_1, \dots, \underline{\mathbf{X}}_l, \dots, \underline{\mathbf{X}}_L$. Each tensor $\underline{\mathbf{X}}_l$ is of dimension $I \times J_l \times K_l$ and represents a set of $\boldsymbol{J}_l$ variables observed over K_l modalities on I individuals. The number of frontal and lateral slices and the nature of the variables can differ from one tensor to another, but the individuals must be the same across tensors. Let $\mathbf{X}_{\boldsymbol{l}}$ be the matricized version of $\underline{\mathbf{X}}_l$. Each matrix $\mathbf{X}_l$ is of dimension $I \times J_l K_l$ and represents all the frontal slices of $\underline{\mathbf{X}}_l$ next to each other. The major drawback of the optimization problem (1) applied to matricized three-way data $\underline{\mathbf{X}}_1, \dots, \underline{\mathbf{X}}_L$ is that the data structure is not preserved and leads potentially (i) to very large $J_l K_l$ block-weight vector to estimate and (ii) to an estimation procedure that ignores the original three-way structure of the data. These aspects can yield a lack of relevant interpretations and additional structural constraints are required. From that perspective, we propose Multiway Generalized Canonical Correlation Analysis (MGCCA) that specifically address the peculiar three-way structure of the data within the RGCCA optimization process.

The MGCCA optimization problem is defined as the following optimization problem:

$$\max_{\mathbf{w}_1,\dots,\mathbf{w}_L} \sum_{k,l=1}^{L} c_{kl}\, g(I^{-1}\mathbf{w}_k^\top \mathbf{X}_k^\top \mathbf{X}_l \mathbf{w}_l) \quad \text{s.t.} \quad \mathbf{w}_l^\top \mathbf{M}_l \mathbf{w}_l = 1 \quad \text{and} \quad \mathbf{w}_l = \mathbf{w}_l^K \otimes \mathbf{w}_l^J\,, l = 1, \dots, L \tag{2}$$

The weight vectors $\mathbf{w}_l\,, l = 1, \dots, L$ are modeled as the Kronecker product between a weight vector $\mathbf{w}_l^K$ associated with the K_l frontal slices and a weight vector $\mathbf{w}_l^J$ associated with the J_l lateral slices: $\mathbf{w}_l = \mathbf{w}_l^K \otimes \mathbf{w}_l^J, l = 1, \dots, L$. This structural constraints yield a more parsimonious model ($J_l + K_l$ instead of $J_l \times K_l$ parameters to estimate), and allows to separately study the effects of the variables and the modalities. These Kronecker constraints are usual in the multi-way literature ([3], [4] and [5]). A fast algorithm that converges to a stationary point of (2) is implemented (proof of convergence published soon) and is available as part of the RGCCA package [6].

3 Analysis of Raman Spectroscopy data

This study aims at analyzing the efficiency of a moisturizer thanks to Raman spectroscopy. For this, on $I = 13$ volunteers, a Raman spectroscopy was performed on both of their arms over $K = 7$ different depths each of which lead to $J = 751$ wavelengths. However, only one of their arms had received the moisturizer. Thus, one arm plays the role of the control, the other one of the test. In addition, the acquisition was repeated at $L = 5$ time points. The resulting dataset is characterized by 5 three-way tensors of dimension $26 \times 751 \times 7$. In this context, MGCCA will provide time points, wavelengths and depths that help to discriminate between treated and non-treated arms. Figure 1.c shows the $\mathbf{w}_l^L$ and 1.d the $\mathbf{w}_l^K$coefficients obtain for $l = 1, \dots, 5$. It appears that at each time point, wavelengths that discriminate the most between treated and non-treated arms are the ones associated to the absorption band of water.

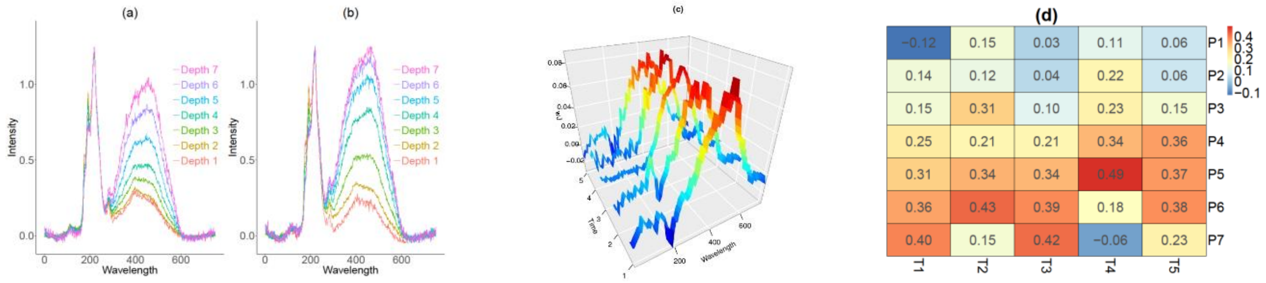

Figure 1 – For an individual, intensity of absorption depending on the wavelength and the depth of the skin, for a non-treated (a) and a treated (b) arm. (c) shows the $\mathbf{w}_l^J$coefficients (associated to the wavelengths) and (d) shows the $\mathbf{w}_l^K$coefficients (associated to the depths P1 to P7) for $l = 1, \dots, 5$ (associated to the time).

4 Conclusion

Work in progress includes to incorporate within the optimization problem (2) specific penalties on the weight vectors $\mathbf{w}_l^J$ and $\mathbf{w}_l^K$.

5 References

[1] Tenenhaus, M., Tenenhaus, A., & Groenen, P., Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. Psychometrika, 2017, 82 (3), pp.737-777

[2] Kiers, H. A. Towards a standardized notation and terminology in multiway analysis. Journal of chemom., 2000, 14(3), 105-122.

[3] Bro, R. (1996). Multiway calidration. Multilinear PLS. Journal of Chemometrics, 10, 47–61.

[4] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM review, 51(3), 455–500.

[5] Zhou, H., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association,108(502), 540–552.

[6] Tenenhaus, A., Guillemot, V. (2017). RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multi-Block Data. R package version 2.1. https://CRAN.Rproject.org/package=RGCCA