RGCCA for the analysis of Microbiome data

Nasrine Bourokba
L'Oréal R&I Singapour – NBOUROKBA@rd.loreal.com

Cécile Clavaud
L'Oréal R&I, Aulnay-sous-Bois, France – CCLAVAUD@rd.loreal.com

Stéphanie Nouveau
L'Oréal R&I, Aulnay-sous-Bois, France – SNOUVEAU@rd.loreal.com

Philippe Bastien
L'Oréal R&I, Aulnay-sous-Bois, France – PBASTIEN@rd.loreal.com

Arthur Tenenhaus
Laboratoire des Signaux et Systèmes (UMR CNRS 8506), CentraleSupelec, Gif-sur-Yvette, France,
arthur.tenenhaus@centralesupelec.fr

Abstract

We follow the practical guidelines described in [1] on how to use RGCCA for the analysis of microbiome data. We illustrate the flexibility and usefulness of RGCCA on a dataset of 200 volunteers from two Chinese cities, in which we obtained metabolomics data. Through the reduction to a few meaningful components and the visualization of relevant variables, we identified possible relevant metabolites.

Keywords: Regularized Generalized Canonical Correlation Analysis, Microbiome data.

1. Introduction

Microbiome data is intrinsically structured in blocks of variables. Indeed, a set of Operational Taxonomic Units (OTUs), at a certain level of taxonomy, can be grouped into specific genera and define, when observed on a set of individuals, a data matrix called block thereafter. Considering the number of genus, a microbiome dataset can then be viewed as multiblock data set. Dedicated modelling algorithms able to cope with the inherent properties of these multiblock datasets are therefore mandatory for harnessing their complexity and provide relevant information.

In this article, we present the principles of Regularized Generalized Canonical Correlation Analysis (RGCCA) [2, 3, 4], a component-based framework for the integrative exploration of multiblock and high-dimensional datasets. We apply it to an original microbiome dataset generated Pr. P. Lee (City University of Hong Kong) and show how the obtained results are useful as RGCCA allows both the identification of relevant variables and the reduction of the multiblock datasets into a few meaningful components that can be easily described as a set of graphical representations.

This paper is organized as follows. In section 2, the RGCCA optimization problems are briefly presented. Section 3 illustrates on a real and challenging microbiome dataset the usefulness of RGCCA.

2. Regularized Generalized Canonical Correlation Analysis

The following section describes a general framework for multiblock component methods, RGCCA and variations, that was previously published [2, 3, 4]. For the sake of comprehension of the use of this method, their theoretical bases will be briefly described in the next subsection.

Let us consider J data matrices $\mathbf{X}_1, \dots, \mathbf{X}_j, \dots, \mathbf{X}_J$. Each $n \times p_j$ data matrix $\mathbf{X}_j = \left[\mathbf{x}_{j1}, \dots, \mathbf{x}_{jp_j}\right]$ is called a block and represents a set of p_j variables observed on n 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, for each block, a weighted composite of variables $\mathbf{y}_j = \mathbf{X}_j\mathbf{w}_j, j = 1, \dots, J$ (where $\mathbf{w}_j$ is a column-vector with p_j elements), called block component, summarizing the relevant information between and within the blocks. The block components are obtained such that (i) block components explain well their own block and/or (ii) block components that are assumed to be connected are highly correlated. Indeed, RGCCA can process a priori information defining which blocks are supposed to be linked to one another, thus reflecting hypotheses about the structural connection between blocks. The second generation RGCCA [4] subsumes fifty years of multiblock component methods. It provides important improvements to the initial version of RGCCA [2] and is defined as the following optimization problem:

$$(1) \qquad \max_{\mathbf{w}_1, \dots, \mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}(\mathbf{X}_j\mathbf{w}_j,\ \mathbf{X}_k\mathbf{w}_k)\right) \qquad \mathrm{s.t.}\ (1-\tau_j)\mathrm{var}(\mathbf{X}_j\mathbf{w}_j) + \tau_j\|\mathbf{w}_j\|_2^2 = 1\ ,\ j = 1,\ \dots, J$$

The scheme function g is any continuous convex function and allows to consider different optimization criteria. Typical choices of g are the identity (leading to maximizing the sum of covariances between block components), the absolute value (yielding maximization of the sum of the absolute values of the covariances) or the square function (thereby maximizing the sum of squared covariances). The design matrix $\mathbf{C} = \{c_{jk}\}$ is a symmetric $J \times J$ matrix of nonnegative elements describing the network of connections between blocks that the user wants to take into account. Usually, $c_{jk} = 1$ for two connected blocks and 0 otherwise. The τ_j are called shrinkage parameters ranging from 0 to 1. Setting τ_j to 0 will force the block components to unit variance ($\mathrm{var}(\mathbf{X}_j\mathbf{w}_j) = 1$), in which case the covariance criterion boils down to the correlation. The correlation criterion is better in explaining the correlated structure across datasets, thus discarding the variance within each individual dataset. Setting τ_j to 1 will normalize the block weight vectors ($\mathbf{w}_j^\top\mathbf{w}_j = 1$) , which applies the covariance criterion. A value between 0 and 1 will lead to a compromise between the two first options and correspond to the following constraint $\mathbf{w}_j^\top\left((1-\tau_j)n^{-1}\mathbf{X}_j^\top\mathbf{X}_j + \tau_j\mathbf{I}\right)\mathbf{w}_j = 1$ in (1). We mention that depending on the choices of the triplet $(\mathbf{C}, \tau_j, g)$, RGCCA recovers several important multivariate analysis methods (see [4] for a complete overview).

Optimization problem (1) focuses on the construction of the first block-components. It is possible to obtain more than one block-component per block. Higher stage block components can be obtained using a deflation strategy (see [4] for details). This strategy forces all the block components within a block to be uncorrelated. This deflation procedure can be iterated in a very flexible way. It is not necessary to keep all the blocks in the procedure at all stages: the number of components summarizing a block can vary from one block to another.

Finally, as a component-based method, RGCCA can provide users with graphical representations to visualize the sources of variability within blocks and the amount of correlation between blocks.

The function `rgcca()` of the RGCCA package [5] implements a monotonically convergent algorithm for the optimization problem (1) – i.e. the bounded criterion to be maximized increases at each step of the iterative procedure –, which hits at convergence a stationary point of (1).

In this paper, we applied RGCCA to a microbiome dataset by following the eight-step guideline described in [1]: (i) construction of the multiblock dataset, (ii) preprocessing, (iii) definition of the between-block connections, (iv) determination of the shrinkage parameters, (v) choice of the scheme function, (vi) determination of the number of components per block, (vii) visualization of the results and (viii) assessment of the reliability of parameter estimates using bootstrap confidence intervals.

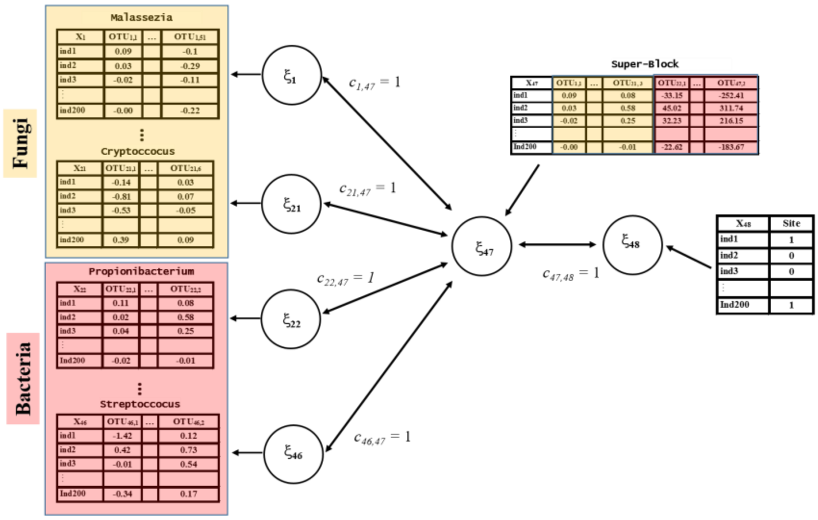
3. Application of RGCCA to a microbiome dataset

The human microbiome contains a vast array of microbes (e.g. bacteria and fungi) that are essential to health and provide important metabolic capabilities. Several statistical analysis tools have been proposed to examine differences between microbial communities and to identify the key OTUs that are responsible of the differences. None of them use the intrinsic multiblock structure of microbiome data. Indeed, it appears that microbiome data are intrinsically structured in block since each OTUs belongs to one specific genus. Therefore, we apply RGCCA on a cohort of 200 volunteers from two Chinese cities. The main objective of this analysis was to identify OTUs within each genus that (i) explain their own block and (ii) important for the discrimination between the two cities.

(ii) Construction of the multiblock dataset. The variables that compose each block have to be defined carefully. For microbiome data, each block represents a set of OTUs belonging to a specific genus. Overall, 25 blocks (resp. 21) associated with bacteria (resp. fungi) are considered. This grouping strategy makes blocks more interpretable and facilitates the interpretation of the RGCCA model. Figure 1 gives details about the structure used during the RGCCA modeling process.

(i) Data processing and normalization. The sparse nature of microbiome data makes preprocessing and normalization steps crucial. After removing samples with less than 10 OTU counts, we removed OTUs with proportional counts across all samples below 0.1%. We then applied the Cumulative Sum Scaling normalization [6] on the log transformed counts using the metagenomeSeq package [7]. In addition, after a centering step, to make blocks comparable, we divide each block by the square root of its number of variables.

(iii) Definition of the design matrix C. In the search of OTUs discriminating the two Chinese cities – we applied RGCCA to identify variables from the 46 blocks (i.e number of genus considered in the analysis) associated with the binary variable indicating the site of each volunteer. The between-block connections associated with this objective of analysis are presented in Figure 1. We chose a consensus PCA structure oriented toward the explanation of the site by imposing an additional connection between the superblock and the site variable.



X_1	$OTU_{1,1}$	…	$OTU_{1,51}$
ind1	0.09		-0.1
ind2	0.03		-0.29
ind3	-0.02		-0.11
⋮			
ind200	-0.00		-0.22

X_{21}	$OTU_{21,1}$	…	$OTU_{21,6}$
ind1	-0.14		0.03
ind2	-0.81		0.07
ind3	-0.53		-0.05
⋮			
ind200	0.39		0.09

X_{22}	$OTU_{22,1}$	…	$OTU_{22,2}$
ind1	0.11		0.08
ind2	0.02		0.58
ind3	0.04		0.25
⋮			
Ind200	-0.02		-0.01

X_{46}	$OTU_{46,1}$	…	$OTU_{46,2}$
ind1	-1.42		0.12
ind2	0.42		0.73
ind3	-0.01		0.54
⋮			
Ind200	-0.34		0.17

X_{47}	$OTU_{1,1}$	…	$OTU_{21,6}$	$OTU_{22,1}$	…	$OTU_{47,2}$
ind1	0.09		0.08	-33.15		-252.41
ind2	0.03		0.58	45.02		311.74
ind3	-0.02		0.25	32.23		216.15
⋮						
Ind200	-0.00		-0.01	-22.62		-183.67

X_{48}	Site
ind1	1
ind2	0
ind3	0
⋮	
Ind200	1

Figure 1. Path diagram describing the between-block connections encoded by the design matrix **C**.

(iv) Choice of the scheme function g. In this case, it was not expected that all the genus, contributed equivalently to the process. The block selector behavior of RGCCA was favored by using the scheme function $g(x) = x^4$ (see [4] for details).

(v) Determination of the shrinkage parameters and (vi) the number of block components. Due to the high number of variables, we set the shrinkage parameters to 1 for all blocks. It yields stable block-components (large variance) while simultaneously taking into account the correlations between connected blocks. Also, one component for the first order blocks and two components (using a deflation strategy) for the superblock were built. We mention that the shrinkage parameters and the number of components per block could have been determined by V-fold cross-validation with respect to the prediction of the site.

(vii) Visualization of the results. As a component-based method, RGCCA provides the users with graphical representations. This graphical displays allows visualizing the sources of variability within blocks, the relationships between variables within and between blocks and the amount of correlation between blocks. The space spanned by the global components is viewed as a compromise space that integrated all the modalities. This global space was useful for visualization and eased the interpretation of the results. The graphical display of the individuals obtained by crossing the two first global components and marked with their status is shown in Figure 2.

(viii) Assessment of the reliability of parameter estimates. It is possible to use a bootstrap resampling method to assess the reliability of parameter estimates obtained using RGCCA. *B* bootstrap samples of the same size as the original data is repeatedly sampled with replacement from the original data. RGCCA is then applied to each bootstrap sample to obtain the RGCCA estimates. For RGCCA, we calculate the mean and variance of the estimates across the bootstrap samples, from which we derived *t*-ratio and *p*-value (under the assumption that the parameter estimates exhibited asymptotic normality) to indicate how reliably parameters were estimated. Since several *p*-values are constructed simultaneously, Bonferroni or FDR corrections can be applied for controlling the Family-Wise Error Rate or the False Discovery Rate, respectively. Table 1 reports these confidence intervals and p-values for one specific block (Corynebacterium).

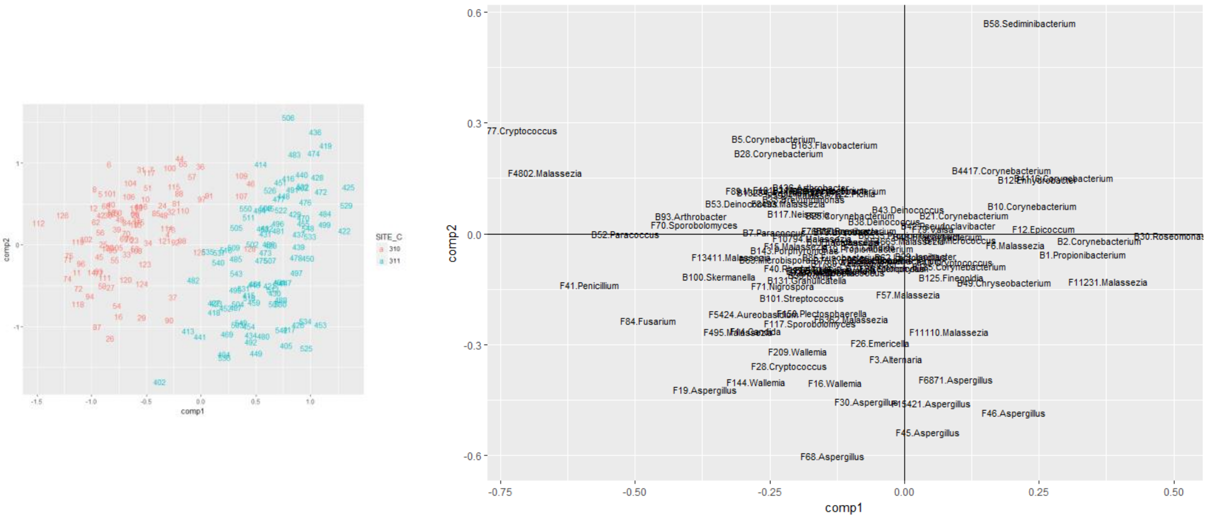


Figure 2. Sample space (left) and variable space (right) corresponding to the two first dimensions of the superblock. Individuals are colored according to the sites.

Figure 2 shows the individuals and variables projected on the compromise space. Despite some overlap, the first global component exhibited a strong separation among the two sites. A variable that is highly expressed for a category of individuals will be projected with a high weight in the direction of that category. Likewise, the separation among volunteers from the two cities seemed to be driven by the OTUs located on the left and right of the variables map.

Table 1. Bootstrap confidence intervals and associated p-values for the OTUs, example of OTUs belonging to Corynebacterium genus.

Corynebacterium	Initial weights	Lower Bound	Upper Bound	p-value	adjusted p-value (Bonferonni)	adjusted p-value (FDR)
B2	0,5326	0,3329	0,6842	0	0	0
B5	-0,4475	-0,5535	-0,1924	0,0001	0,0006	0,0003
B28	-0,4049	-0,5315	-0,1736	0,0001	0,0014	0,0005
B10	0,4262	0,1483	0,5337	0,0005	0,0063	0,0016
B4116	0,2548	0,1096	0,4845	0,0019	0,0227	0,0045
B114	-0,1972	-0,467	-0,0529	0,0139	0,1664	0,0277
B185	0,1304	-0,0404	0,3691	0,1157	1	0,1984
B4417	0,1334	-0,0734	0,3431	0,2043	1	0,2743
B25	-0,1478	-0,3424	0,0806	0,2251	1	0,2743
B21	0,0816	-0,0791	0,3313	0,2286	1	0,2743
B75	-0,0509	-0,2692	0,1493	0,5741	1	0,6263
B298	0,0265	-0,1717	0,2505	0,7147	1	0,7147

4. Conclusion

RGCCA stands as a unique, general and original way for analyzing high-dimensional multiblock datasets. It allows the identification of a few meaningful variables that underline the between-block connections encoded by the design matrix **C**. RGCCA highlights important discriminative OTUs and provides interpretable graphical outputs to better understand the role of each genus to the discrimination.

Of note, the quality and interpretability of the RGCCA block components are likely affected by the usefulness and relevance of the variables of each block. Accordingly, RGCCA integrates a variable selection procedure, called SGCCA [8], allowing the selection of the most informative variables. The SGCCA algorithm is similar to the RGCCA algorithm and keeps the same convergence properties. The algorithm associated with SGCCA is available through the function `sgcca()` of the RGCCA package. Work in progress includes the application of SGCCA to the microbiome dataset presented in this paper.

References

[1] Garali I, Adanyeguh I, Ichou F, Perlbarg V, Seyer A, Colsch B, Moszer I, Guillemot V, Durr A, Mochel F, Tenenhaus A (2017) A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia, accepted for publication, Briefings in Bioinformatics.

[2] Tenenhaus A, Tenenhaus M (2011) Regularized generalized canonical correlation, vol. 76, pp. 257-284, Psychometrika.

[3] Tenenhaus A, Philippe C, Frouin V (2015) Kernel generalized canonical correlation analysis, Computational Statistics & Data Analysis, vol. 90, pp. 114-131.

[4] Tenenhaus M, Tenenhaus A, Groenen PJF, (2017) Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods, Accepted for publication, Psychometrika.

[5] Tenenhaus A and Guillemot V (2017) RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multi-Block Data. R package version 2.1.

[6] Paulson JN, Stine OC, Bravo HC, Pop M (2013) Differential abundance analysis for microbial arkergene surveys. Nature methods 10(12), 1200–1202.

[7] Paulson JN, Pop M, Bravo HC (2015) metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor package: 1.6.0

[8] Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. (2014) Variable selection for generalized canonical correlation analysis., Biostatistics, vol. 15, no. 3, pp. 569-583.