Introducing group-sparsity and orthogonality constraints in RGCCA

Vincent Guillemot, Arnaud Gloaguen, Arthur Tenenhaus, Cathy Philippe, Hervé Abdi

Introducing group-sparsity and orthogonality constraints in RGCCA

Vincent Guillemot[1], Arnaud Gloaguen[2], Arthur Tenenhaus[2], Cathy Philippe[3] and Hervé Abdi[4]

[1] *Hub de Bioinformatique et Biostatistique, Institut Pasteur, Paris, FR*
[2] *Laboratoire des Signaux et Systèmes, CentraleSupelec, Gif-Sur-Yvette, FR*
[3] *Neurospin, CEA, Gif-Sur-Yvette, FR*
[4] *The University of Texas at Dallas, Richardson, TX, USA*

Résumé. RGCCA est une méthode flexible et rapide qui—généralisant de nombreuses méthodes existantes—permet l'analyse de données structurées en plusieurs blocs hétérogènes. Nous présentons l'ajout dans RGCCA de deux nouvelles contraintes : une contrainte de parcimonie de groupes et une contrainte d'orthogonalité sur les poids de RGCCA. Ces deux contraintes ont pour but d'augmenter l'interprétabilité de l'analyse de données de grande dimension qui possèdent une structure de groupe. Nous appliquons cette nouvelle méthode—abrégée en gSGCCA—à l'analyse de données de gliome malin pédiatrique structurées en trois blocs. Nous montrons sur ces données le gain en interprétabilité apporté par les contraintes de parcimonie et d'orthogonalité.

Mots-clés. RGCCA, parcimonie, parcimonie de groupe, structure

Abstract. RGCCA—a fast and flexible method—generalizes many other well-known methods in order to analyze data-sets comprising multiple blocks of variables. Here we extend RGCCA by adding two new constraints to the RCCCA optimization problem: 1) group sparsity and 2) orthogonality of the block weight vectors. These two constraints facilitate the interpretability of the results when analyzing high dimensional data with a group structure. We illustrate this new method—called gSGCCA—with the analysis of pediatric high-grade glioma data: a set comprising three data blocks. This analysis shows that these new constraints greatly improve the interpretability of the statistical analysis.

Keywords. RGCCA, sparsity, group-sparsity, structure

1 Introduction

Regularized Generalized Canonical Correlation Analysis (RGCCA) [8, 9, 3] is a recent multiblock component method that generalizes traditional component-based two table methods—such as partial least square correlation, redundancy analysis, and canonical correlation—in order to analyze data sets comprising multiple blocks of data. Just like with other component methods, RGCCA results are often difficult to interpret when there

are (too?) many variables; to mitigate this problem, RGCCA has been extended to become Sparse General Canonical Correlation Analysis (SGCCA) [7]: a version of RGCCA that incorporate an ℓ_1-norm based constraint in order to generate sparse block weight vectors. This sparsification constraint improves the interpretation of the results (because it selects important variables) but at a cost: the block weight vectors are not orthogonal—a pattern that often makes the results difficult to interpret. This trade-off between sparsity and orthogonality is not specific to RGGCA: it affects all component based-methods, especially those based on the singular value decomposition (SVD) and its extensions (e.g., the generalized SVD, GSVD). Recently, however, we found that this trade-off could be eliminated 1) for the SVD: the constrained SVD (CSVD) [4], combines orthogonality and sparsity constraints to the plain SVD, and 2) for the GSVD (including block constraints on observations and variables): as implemented in sparse Multiple Correspondence Analysis (sMCA) [5].

Here, we propose to extend the approach used for the GSVD to create gSGCCA: the version of RGCCA that includes 1) a group sparsity constraint (and it associated group sparse projection), and 2) an orthogonality constraint on the block weight vectors. To do so, we applied the same group projection as in sparse MCA, combined with an orthogonality projection with projections onto convex sets (POCS) [1]. We illustrate gSGCCA with the analysis of the pediatric glioma data used in [7].

2 Method

Group sparse GCCA (gSGCCA) is defined as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{\operatorname{argmax}} f\left(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J\right) &= \sum_{\substack{j,k=1 \\ j\neq k}}^{J} c_{jk}\, g\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j,\mathbf{X}_k\mathbf{a}_k\right)\right) \\ \text{subject to } \begin{cases} \|\mathbf{a}_j\|_2 = 1 \\ \|\mathbf{a}_j\|_{\mathcal{G}_j} \leq s_j \quad , \ \forall j = 1,\ldots,J. \\ \mathbf{a}_j \perp \mathbf{A}_j \end{cases} \end{aligned} \tag{1}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_J$ are J centered blocks of data, the function g is defined as any continuously differentiable convex function, and the design matrix $\mathbf{C} = \{c_{jk}\}$ is a symmetric $J \times J$ matrix of non-negative elements describing the network of connections between blocks that are to be taken into account. Moreover, $\mathbf{a}_1,\ldots,\mathbf{a}_J$ are block weight vectors (i.e.. the weights applied to each block to obtain the block components), $\mathbf{A}_1,\ldots,\mathbf{A}_J$ are the previously estimated weight vectors combined into matrices, $\mathcal{G}_1,\ldots,\mathcal{G}_J$ are the groups of variables for a block, $s_1,\ldots,s_J$ are positive scalars controlling the group sparsity constraint for the block weight vectors, and the group norm is defined as: $\|\mathbf{x}\|_{\mathcal{G}} = \sum_{g=1}^{G} \|\mathbf{x}_{\iota_g}\|_2$, where $\mathbf{x}_{\iota_g}$ is the subvector of $\mathbf{x}$ that contains only the elements of group $\mathcal{G}_j$. The $\ell_{1,2}$-ball

associated with this norm is noted $\mathscr{B}_{1,2}(\cdot)$. In the next section, we present a monotone convergent algorithm for solving optimization Problem (1).

2.1 The gSGCCA algorithm

The maximization of function f over the parameter vectors $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_L)$, is implemented using cyclic Block Coordinate Ascent (BCA [2]); a procedure that updates in turn, each of the parameter vectors while keeping the others fixed. Specifically, let $\nabla_j f(\mathbf{a})$ be the partial gradient of $f(\mathbf{a})$ with respect to $\mathbf{a}_j$. We want to find an update $\hat{\mathbf{a}}_j \in \Omega_j = \{\|\mathbf{a}_j\|_2 = 1, \text{ and } \|\mathbf{a}_j\|_{\mathcal{G}_j} \leq s_j, \text{ and } \mathbf{a}_j \perp \mathbf{A}_j\}$ such that $f(\mathbf{a}) \leq f(\mathbf{a}_1, ..., \mathbf{a}_{j-1}, \hat{\mathbf{a}}_j, \mathbf{a}_{j+1}, ..., \mathbf{a}_J)$. Because f is a continuously differentiable multi-convex function and because a convex function lies above its linear approximation at $\mathbf{a}_j$ for any $\tilde{\mathbf{a}}_j \in \Omega_j$, the following inequality holds:

$$f(\mathbf{a}_1, ..., \mathbf{a}_{j-1}, \tilde{\mathbf{a}}_j, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_J) \geq f(\mathbf{a}) + \nabla_j f(\mathbf{a})^\top (\tilde{\mathbf{a}}_j - \mathbf{a}_j) := \ell_j(\tilde{\mathbf{a}}_j, \mathbf{a}). \tag{2}$$

On the right-hand side of (2), only the term $\nabla_j f(\mathbf{a})^\top \tilde{\mathbf{a}}_j$ is relevant to $\tilde{\mathbf{a}}_j$ and, so, the solution maximizing the minorizing function $\ell_j(\tilde{\mathbf{a}}_j, \mathbf{a})$ over $\tilde{\mathbf{a}}_j \in \Omega_j$ is obtained by considering:

$$\hat{\mathbf{a}}_j = \underset{\tilde{\mathbf{a}}_j \in \Omega_j}{\operatorname{argmax}} \ \nabla_j f(\mathbf{a})^\top \tilde{\mathbf{a}}_j = \underset{\tilde{\mathbf{a}}_j \in \Omega_j}{\operatorname{argmin}} \ \|\nabla_j f(\mathbf{a}) - \tilde{\mathbf{a}}_j\|_2^2. \tag{3}$$

This last equality follows from $\|\mathbf{a}_j\|_2 = 1$ as $\mathbf{a}_j \in \Omega_j$. This core optimization problem is a projection onto the intersection between the ball defined by the groups, the ℓ_2-ball, and the space orthogonal to the already estimated block weight vectors, assembled in $\mathbf{A}_j$. This projection on $\mathscr{B}_{1,2}(s_j) \cap \mathscr{B}_2(1) \cap \mathbf{A}_j^\perp$ is performed using POCS with two components: the projection onto the intersection of the group ball and the ℓ_2-ball, and the projection onto the orthogonal spaces defined by the already estimated loading vectors combined in the matrix $\mathbf{A}_j$. The complete gSGCCA algorithm is presented in Algorithm 1.

3 Application on glioma data

We applied gSGCCA to the glioma data previously analyzed with SGCCA ([7, 6]). This data-set comprises three blocks of variables: 1) gene expression data (GE), 2) comparative genomic hybridization data (CGH) and, 3) the location of the tumor in the brain. Here, we focused on the analysis of only six groups of genes highly associated with different types of brain tumors and with brain tumor development. The six groups are defined similarly for both the GE and CGH blocks. For this analysis, we used three different versions of RGCCA: 1) a classical three-block RGCCA with a complete design (i.e., all blocks are inter-connected); 2) a structured version of RGCCA where each group of genes is a block (here the design is complete within each type of data, GE or CGH, and all the

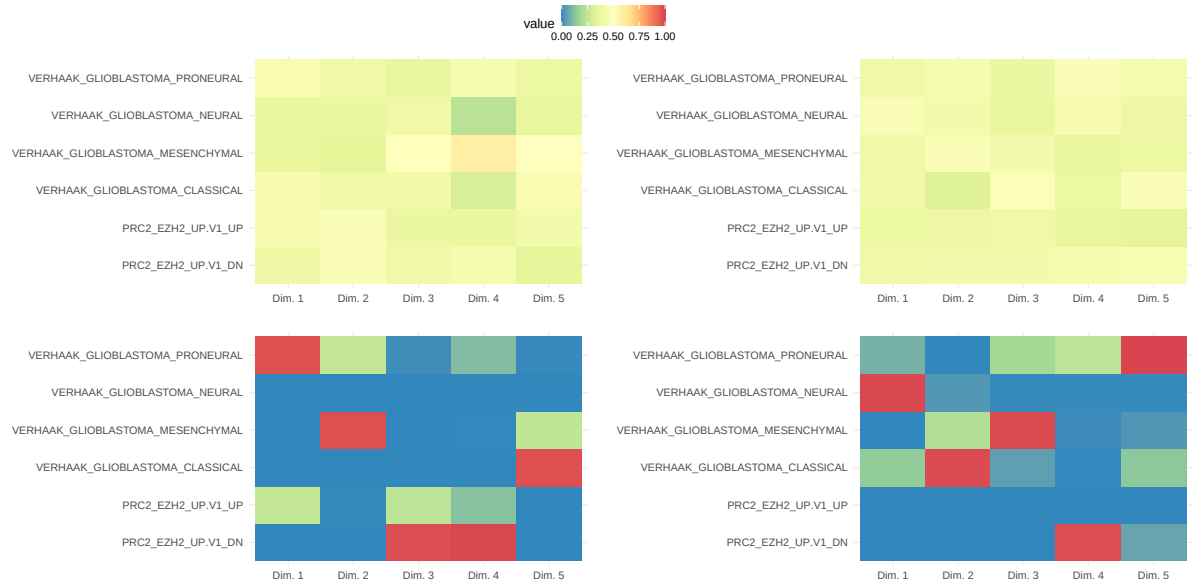

Figure 1: Comparison of the block weight vectors, summarized by groups, without sparsity constraints (upper graphs) or with a maximum sparsity constraint (lower graphs), for the GE block (on the left) and the CGH block (on the right).

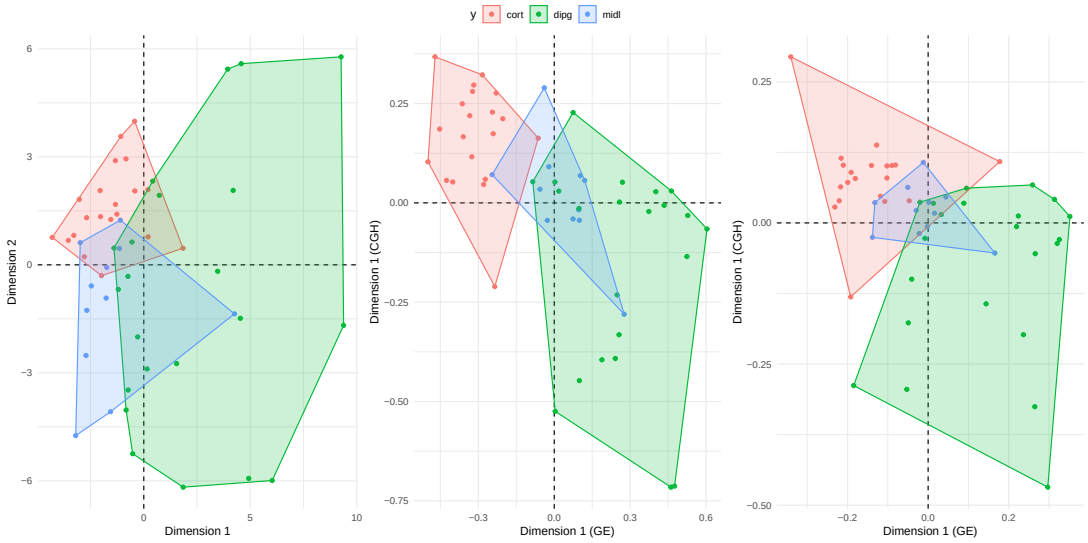

Figure 2: Factor scores of three different versions of RGCCA. Left: the two first dimensions of RGCCA applied to a 13 block dataset, 6 blocks for the GE functional groups, 6 blocks for the CGH functional groups and 1 block for the response, followed by a principal component analysis of the resulting GE and CGH block components. Middle and right: the first dimension of CGH (y-axis) as a function of the first dimension of GE (x-axis). Middle: RGCCA with no sparsity. Right: gSGCCA with maximum sparsity.

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_J$, $\mathcal{G}_1, \ldots, \mathcal{G}_J$, ε, R, $s_{1,\ell}, \ldots, s_{J,\ell}$.
Initialization: $\forall j = 1, \ldots, J, \mathbf{A}_j \leftarrow [\,]$;
Result: The estimated weight vectors combined into matrices $\mathbf{A}_1, \ldots, \mathbf{A}_J$
for $\ell = 1, \ldots, R$ **do**
  Initialize $\mathbf{a}_j^0$ for all j;
  $s \leftarrow 0$;
  while $\left\|\mathbf{a}_j^{(s+1)} - \mathbf{a}_j^{(s)}\right\| \geq \varepsilon$, $\forall j = 1, \ldots, J$ **do**
    for $j = 1, \ldots, J$ **do**
      Compute the inner component:

$$\nabla_j^s f \leftarrow \frac{1}{n}\mathbf{X}_j^t \left[\sum_{k=1}^{j-1} c_{jk} g'\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j^s, '\mathbf{X}_k\mathbf{a}_k^{s+1}\right)\right)\mathbf{X}_k\mathbf{a}_k^{s+1} \;+ \sum_{k=j+1}^{J} c_{jk} g'\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j^s, \mathbf{X}_k\mathbf{a}_k^s\right)\right)\mathbf{X}_k\mathbf{a}_k^s\right]$$

      Compute the outer weight:

$$\mathbf{a}_j^{s+1} \leftarrow \operatorname{proj}\left(\nabla_j^s f, \mathscr{B}_{1,2}(s_{j,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{A}_j^{\perp}\right)$$

    end
    $s \leftarrow s + 1$;
  end
  $\forall j = 1, \ldots, J, \mathbf{A}_j \leftarrow \left[\mathbf{A}_j, \mathbf{a}_j^{(s+1)}\right]$;
end

Algorithm 1: General algorithm of gSGCCA implementing group-sparsity and orthogonality of the block weight vectors.

GE and CGH blocks are connected to the location block), and 3) gSGCCA with maximum sparsity and a complete design (like option 1).

The block weight vectors are shown in Figure 1 for Versions 1 and 3. Each functional group is represented by its norm. This figure shows that incorporating a group-sparsity constraint in RGCCA greatly improves the interpretability of the block weight vectors because with gSGCCA only a handful of groups were selected for each dimension of GE and CGH.

The block components are shown on Figure 2, which shows that the improvement in interpretability for the loadings—observed on the block weight vectors—comes at the cost of a diminished class separation observed on the observations (this effect occurs because sparsifying the loadings automatically reduces the variance of the observations factor scores).

4 Conclusion and perspectives

We present in this paper a new method—called gSGCCA—that adds group-sparsity and orthogonality constraints to RGCCA. The application of gSGCCA to a medical example illustrates that, compared to the original RGCCA, gSGCCA provides results easier to interpret.

Future work will focus on developing a user friendly framework for the selection of the sparsity parameters to achieve some optimum trade-off between sparsity and prediction performance. We will also work on including metrics in gsGCCA to generalize its application to a wider range of data types.

References

[1] P.L. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, 1993.

[2] J. De Leeuw. Block-relaxation algorithms in statistics. In Hans-Hermann Bock, Wolfgang Lenski, and Michael M. Richter, editors, *Information Systems and Data Analysis*, pages 308–324, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg.

[3] I. Garali et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings in Bioinformatics*, 19:1356–1369, 2018.

[4] V. Guillemot et al. A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE*, 14:e0211463, 2019.

[5] V. Guillemot et al. Sparse Multiple Correspondence Analysis. In *52èmes Journées de Statistique*, Nice, France, 2020.

[6] S. Puget et al. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. *PloS one*, 7:e30313, 2012.

[7] A. Tenenhaus et al. Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*, 15:569–83, 2014.

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

[9] M. Tenenhaus, A. Tenenhaus, and P.J.F. Groenen. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, 82:737–777, 2017.