A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia

Imene Garali, Isaac M. Adanyeguh,* Farid Ichou,* Vincent Perlbarg, Alexandre Seyer, Benoit Colsch, Ivan Moszer, Vincent Guillemot, Alexandra Durr, Fanny Mochel* and Arthur Tenenhaus*

*These authors contributed equally to this work.
Corresponding author: Arthur Tenenhaus, Laboratoire des Signaux et Systèmes at CentraleSupelec Gif-sur-Yvette, France. Tel.: +33 (0)169851422; E-mail: arthur.tenenhaus@centralesupelec.fr

Abstract

The growing number of modalities (e.g. multi-omics, imaging and clinical data) characterizing a given disease provides physicians and statisticians with complementary facets reflecting the disease process but emphasizes the need for novel statistical methods of data analysis able to unify these views. Such data sets are indeed intrinsically structured in blocks, where each block represents a set of variables observed on a group of individuals. Therefore, classical statistical tools cannot be applied without altering their organization, with the risk of information loss. Regularized generalized canonical correlation analysis (RGCCA) and its sparse generalized canonical correlation analysis (SGCCA) counterpart are component-based methods for exploratory analyses of data sets structured in blocks of variables. Rather than operating sequentially on parts of the measurements, the RGCCA/SGCCA-based integrative analysis method aims at summarizing the relevant information between and within the blocks. It processes a priori information defining which blocks are supposed to be linked to one another, thus reflecting hypotheses about the biology underlying the data blocks. It also requires the setting of extra parameters that need to be carefully adjusted.

Imene GARALI obtained her PhD degree from Aix-Marseille University in 2015 in image processing. Since 2016, she is a postdoctoral researcher at the Bioinformatics and Biostatistics Core Facility of the Brain and Spine Institute, La Pitié-Salpêtrière Hospital, Paris, France.

Isaac Adanyeguh is PhD student at the Pierre and Marie Curie University whose research focuses on biomarker identification and disease modelling in polyglutamine disorders using multimodal neuroimaging approaches.

Farid Ichou is currently working as metabolomic core manager and researcher at ICANalytics department, institute of cardiometabolism and nutrition, Paris, France. His research interests include study of gut-derived metabolite, discovery approach and translational research in preclinical and clinical studies for cardiometabolic and neurodegenerative diseases.

Vincent Perlbarg is a research engineer at the Bioinformatics and Biostatistics Core Facility of the Brain and Spine Institute, La Pitié-Salpêtrière Hospital, in charge of biomarkers identification in neuroimaging.

Alexandre Seyer is currently researcher at the SpectMet platform of the MedDay Pharmaceuticals company, Paris, France. His research focus on the identification of new therapeutic targets for nervous system disorders through metabolomic and lipidomic analysis in preclinical and clinical studies.

Benoit Colsch is currently research scientist in the LEMM Laboratory at CEA-Saclay, France. His research interests are based on the development of qualitative and quantitative methods using LC-MS in lipidomics and metabolomics projects in the field of neurosciences.

Ivan Moszer is currently managing the Bioinformatics and Biostatistics Core Facility of the Brain and Spine Institute, La Pitié-Salpêtrière Hospital (Paris, France), which provides scientists and clinicians with expert support in omics data processing and high-dimensional heterogeneous data analysis.

Vincent Guillemot is currently a research engineer at Institut Pasteur, in the Statistical Genetics group, and in the Bioinformatics / Biostatistics Core Facility.

Alexandra Durr is full professor and consultant in genetics at the Pitié-Salpêtrière University Hospital in Paris, developing translational neurogenetics at the Brain and Spine Institute.

Fanny Mochel is an associate professor of genetics at the University Pierre and Marie Curie (UPMC) and the Pitié-Salpêtrière university hospital. She runs the French reference center for Neurometabolic diseases in adults. Her research is focused on the characterization and treatment of brain energy deficiencies using both in vitro (metabolomics) and in vivo (nuclear magnetic resonance spectroscopy) approaches.

Arthur Tenenhaus is currently associate professor in the L2S Laboratory at CentraleSupélec, France, and researcher at the Bioinformatics and Biostatistics Core Facility of the Brain and Spine Institute, La Pitié-Salpêtrière Hospital.

Here, we provide practical guidelines for the use of RGCCA/SGCCA. We also illustrate the flexibility and usefulness of RGCCA/SGCCA on a unique cohort of patients with four genetic subtypes of spinocerebellar ataxia, in which we obtained multiple data sets from brain volumetry and magnetic resonance spectroscopy, and metabolomic and lipidomic analyses. As a first step toward the extraction of multimodal biomarkers, and through the reduction to a few meaningful components and the visualization of relevant variables, we identified possible markers of disease progression.

Introduction

The growing number of modalities (e.g. multi-omics, imaging and clinical data) characterizing a given disease provides physicians and statisticians with complementary facets of the disease process. However, novel statistical methods of data analysis are needed to unify these views. This data set is indeed intrinsically structured in blocks, where each block represents a set of variables observed on a group of individuals. The number and the nature of the variables can differ from one block to another. Therefore, classical statistical tools cannot be applied without altering the structure of the multiblock data set. The integration and visualization of these multivariate data sets is also challenging, explaining the need of dedicated modeling algorithms able to cope with the inherent properties of these structured data sets.

In this article, we present the principles of regularized generalized canonical correlation analysis (RGCCA) [1, 2], and its sparse generalized canonical correlation analysis (SGCCA) counterpart [3], a component-based framework for the integrative exploration of multimodal and high-dimensional data sets. We apply it to an original multiblock data set generated from a unique, considering the rarity of these diseases, cohort of patients with spinocerebellar ataxia (SCA) and controls. We show how the obtained results are useful, as SGCCA allows both the extraction of biomarkers and the reduction of the multiblock data sets into a few meaningful components that can be easily described as a set of graphical representations. The main objectives of the article are thus to provide users with practical guidelines for the application of RGCCA and SGCCA, and to illustrate their versatility and relevance on the SCA data set.

This article is organized as follows. In 'Multiblock component methods' section, the RGCCA and SGCCA optimization problems are briefly presented, and a synthetic overview of methods, which are special cases of RGCCA/SGCCA, is given. In 'Practical guidelines for using RGCCA and SGCCA' section, practical guidelines defining how to use RGCCA/SGCCA are provided. 'Case study: the SCA data set' section illustrates on a real and challenging multiblock data set, the usefulness of RGCCA/SGCCA for data integration.

Multiblock component methods

The following section describes a general framework for multiblock component methods, RGCCA and variations, that was previously published [1–3] and assessed [4–6]. For the sake of comprehension of the use of these methods, their theoretical bases will be briefly described in the next subsections. In short, RGCCA is a rich technique that encompasses several important multivariate analysis methods (see Table 1 for the overview). The objective of RGCCA is to find, for each block, a weighted composite of variables (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 biology underlying the data blocks. In addition, RGCCA integrates a variable selection procedure, called SGCCA, allowing the identification of the most relevant features. Finally, as a component-based method, RGCCA/SGCCA can provide users with graphical representations to visualize the sources of variability within blocks and the amount of correlation between blocks.

Regularized generalized canonical correlation analysis

We consider J data matrices $\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_J$. Each $n \times p_j$ data matrix $\mathbf{X}_j = \left[\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jp_j} \right]$ is called a block and represents a set of p_j variables observed on n individuals. The number and 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 multiblock component methods is to find block components $\mathbf{y}_j = \mathbf{X}_j \mathbf{w}_j, \ j = 1, \ldots, J$ (where $\mathbf{w}_j$ is a column vector with p_j elements) summarizing the relevant information between and within the blocks. The second-generation RGCCA [2] subsumes 50 years of multiblock component methods (see [2] for a complete review). It provides important improvements to the initial version of RGCCA [1] and is defined as the following optimization problem:
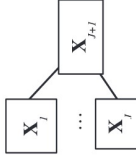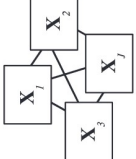
$$\max_{\mathbf{w}_1, \ldots, \ \mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk} g(\text{cov}(\mathbf{X}_j \mathbf{w}_j, \ \mathbf{X}_k \mathbf{w}_k)) \tag{1}$$
$$\text{s.t. } (1 - \tau_j)\text{var}(\mathbf{X}_j \mathbf{w}_j) + \tau_j \|\mathbf{w}_j\|_2^2 = 1, \ j = 1, \ \ldots, J$$
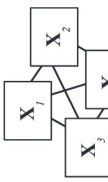
where:
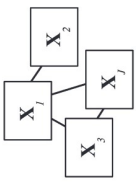
- The scheme function g is any continuous convex function and allows to consider different optimization criteria. Typical choices of g are the identity (leading to maximizing the sum of covariances between block components), the absolute value (yielding maximization of the sum of the absolute values 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 the τ_j to 0 will force the block components to unit variance ($\text{var}(\mathbf{X}_j \mathbf{w}_j) = 1$), in which case the covariance criterion boils down to the correlation. The correlation criterion is better

Table 1. Special cases of RGCCA in a situation of J blocks

Method	Scheme function $g(x)$	Shrinkage constants (τ_j; $j = 1,\dots,J$)	Design matrix (C)
PCA [7]	x	$\tau_1 = 1$	$X_1 \!-\! X_1$ $\quad C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
Canonical Correlation Analysis (CCA) [8]	x	$\tau_1 = 0$ and $\tau_2 = 0$	$X_1 \!-\! X_2$ $\quad C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
Interbattery Factor Analysis [9] (or PLS [10])	x	$\tau_1 = 1$ and $\tau_2 = 1$	
Redundancy analysis of X_1 with respect to X_2 (RR) [11]	x	$\tau_1 = 1$ and $\tau_2 = 0$	
GCCA [12]	x^2	$\tau_j = 0$, $j = 1,\dots, J+1$	$X_1 \cdots X_J \to X_{J+1}$ $\quad C = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}$
GCCA [13]	x^2	$\tau_j = 0$, $j = 1\dots,J_1$, $J+1$ $\tau_j = 1$, $j = J_1+1,\dots, J$	
HPCA [14]	x^4	$\tau_j = 1$, $j = 1,\dots, J$ $\tau_{J+1} = 0$	
MCOA [15], CPCA [16], CPCA-W [17] and MFA [18]	x^2	$\tau_j = 1$, $j = 1,\dots, J$ $\tau_{J+1} = 0$	$c_{j,\,J+1} = 1$, $j = 1,\dots,J$ and 0 otherwise
SUM of CORrelations method (SUMCOR) [19]	x	$\tau_j = 0$, $j = 1,\dots, J$	X_1, X_2, X_3, X_J $\quad C = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & 1 \\ 1 & \cdots & 1 & 1 \end{pmatrix}$
Sum of SQuared CORrelations method (SSQCOR) [20]	x^2	$\tau_j = 0$, $j = 1,\dots, J$	
Sum of ABsolute value CORrelations method (SABCOR) [21]	$|x|$	$\tau_j = 0$, $j = 1,\dots, J$	
SUM of COVariances method (SUMCOV-1). SUMCOV-1 is the 'one component per block' version of MAXBET [22]	x	$\tau_j = 1$, $j = 1,\dots, J$	
Sum of SQuared COVariances method (SSQCOV). SSQCOV-1 is the 'one component per block' version of MAXBET B [23]	x^2	$\tau_j = 1$, $j = 1,\dots, J$	
Sum of ABsolute value COVariances method (SABCOV) [1, 24]	$|x|$	$\tau_j = 1$, $j = 1,\dots, J$	

(continued)

Table 1. Continued

Method	Scheme function $g(x)$	Shrinkage constants ($\tau_j, j = 1,\ldots,J$)	Design matrix (C)
SUMCOV-2. SUMCOV-2 is the 'one component per block' version of MAXDIFF [22]	x	$\tau_j = 1, j = 1,\ldots,J$	$$\mathbf{C} = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & & \vdots \\ \vdots & & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}$$
SSQCOV-2. SSQCOV-2 is the 'one component per block' version of MAXDIFF B [23]	x^2	$\tau_j = 1, j = 1,\ldots,J$	
PLS path modeling—Mode B [25]	$|x|$	$\tau_j = 0$	$\mathbf{C} = (c_{jk})$, $c_{jk} = 1$ for two connected block and $c_{jk} = 0$ otherwise

Note: $\mathbf{X}_{J+1} = \mathbf{X}_1, \ldots, \mathbf{X}_J$ is called superblock and is defined as the concatenation of the J blocks. $(J+1)$th block defined as a superblock, the concatenation of the J blocks.

in explaining the correlated structure across data sets, thus discarding the variance within each individual data set. 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 $\mathbf{w}_j^\top \left(\tau_j \mathbf{I} + (1 - \tau_j)(1/n)\mathbf{X}_j^\top \mathbf{X}_j \right)\mathbf{w}_j = 1$ in Equation (1). The choices $\tau_j = 1$, $\tau_j = 0$ and $0 < \tau_j < 1$ are, respectively, referred as Modes A, B and Ridge.

From optimization problem in Equation (1), the term 'generalized' in the acronym of RGCCA embraces at least three notions. The first one relates to the generalization of two-block methods—including Canonical Correlation Analysis [8], Interbattery Factor Analysis [9] and Redundancy Analysis [10]—to three or more sets of variables. The second one relates to the ability of taking into account some hypotheses on between-block connections: the user decides which blocks are connected and which ones are not. The third one relies on the choices of the shrinkage parameters allowing to capture both correlation or covariance-based criteria.

Variable selection in RGCCA: SGCCA

The quality and interpretability of the RGCCA block components $\mathbf{y}_j = \mathbf{X}_j\mathbf{w}_j$, $j = 1,\ldots,J$ are likely affected by the usefulness and relevance of the variables of each block. Accordingly, it is an important issue to identify within each block a subset of significant variables that are active in the relationships between blocks. SGCCA extends RGCCA to address this issue of variable selection. Specifically, RGCCA with all τ_j, $j = 1, \ldots, J$ equal to 1 is combined with an L1 penalty that gives rise to SGCCA [3]. The SGCCA optimization problem is defined as follows:

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk} g\Big(\text{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)\Big) \quad \text{s.t.} \begin{cases} \|\mathbf{w}_j\|_2 = 1 \\ \|\mathbf{w}_j\|_1 \leq s_j \end{cases}, \; j = 1, \ldots, J,$$

(2)

where s_j is a user-defined positive constant that determines the amount of sparsity for $\mathbf{w}_j$, $j = 1, \ldots,J$. The smaller the s_j, the larger the degree of sparsity for $\mathbf{w}_j$.

Higher stage block components

It is possible to obtain more than one block component per block for RGCCA and SGCCA. Higher stage block components can be obtained using a deflation strategy [1]. This strategy forces all the block components within a block to be uncorrelated. This deflation procedure can be iterated in a 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 [2].

Implementation

The function `rgcca()` of the RGCCA package [26] implements a monotonically convergent algorithm for the optimization problem in Equation (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 Equation (1). Two numerically equivalent approaches for solving the RGCCA optimization problem are available. A primal formulation described in [1] requires the handling of matrices of dimension $p_j \times p_j$. A dual formulation described in [27] requires the handling of matrices of

dimension $n \times n$. Therefore, the primal formulation of the RGCCA algorithm will be used when $n \geq p_j$, and the dual form will be preferred when $n < p_j$. The `rgcca()` function of the RGCCA package implements these two formulations and selects automatically the best one. The SGCCA algorithm is similar to the RGCCA algorithm and keeps the same convergence properties. The algorithm associated with the optimization problem in Equation (2) is available through the function `sgcca()` of the RGCCA package.

Moreover, multiblock data faces two types of missing data structure: (i) if an observation i has missing values on a whole block j and (ii) if an observation i has some missing values on a block j (but not all). For these two situations, it is possible to exploit the algorithmic solution proposed for partial least squares (PLS) regression path modeling to deal with missing data ([28], p. 171). Work is in progress to implement this missing data solution within the RGCCA package.

Special cases of RGCCA

Many different multiblock methods were published for 50 years. The choice of the 'best' multiblock method must be in line with the nature of the data set and the objective of the analysis. The introduction of the design matrix $\mathbf{C}$, the shrinkage parameters τ_js' and the scheme function g makes RGCCA highly versatile. A practical guideline for appropriately specifying these extra parameters is proposed in the next two sections. From a statistical data analysis perspective, RGCCA subsumes a remarkably large number of well-known methods as particular cases—including principal component analysis (PCA) [7], generalized Canonical Correlation Analysis (GCCA) [12], PLS regression [10], consensus PCA (CPCA) [16], hierarchical PCA (HPCA) [14], multiple co-inertia analysis (MCOA) [15], etc. For an exhaustive list of methods, see[2]. All the methods cited above (and many others) are recovered with RGCCA by appropriately defining the triplet ($\mathbf{C}$, τ_j, g). Table 1 gives the correspondence between the triplet ($\mathbf{C}$, τ_j, g) and the associated methods. SGCCA offers a sparse counterpart to all the covariance-based methods of RGCCA. RGCCA/SGCCA provides a framework for exploratory data analysis of multiblock data sets that has immediate practical consequences for a unified statistical analysis and implementation strategy. It is noteworthy that a complete review on dimension reduction approaches for simultaneous exploratory analyses of multiple data sets, and especially multi-omics data sets, has been recently published [5]. In that review, RGCCA/SGCCA is discussed and appears to occupy a key position as many of the single-block, two-block and multiblock component methods—referred as PCA (Principal Component Anlaysis), sPCA (sparse Principal Component Anlaysis), CCA (Canonical Correlation Analysis), RDA (Redundancy analysis), rCCA (Regularized canonical correlation), sCCA (sparse Regularized canonical correlation), PLS (Partial Least Squares), sPLS (sparse Partial Least Squares), sPLSDA (sparse Partial Least Squares - dicriminant analysis), cPCA (consensus PCA), CIA (Co-Inertia Analysis), multiple factor analysis (MFA), MCIA (Multiple Co-Inertia Analysis) and GCCA (Generalized Canonical Correlation Analysis)—are special cases of RGCCA/SGCCA.

In the next section, we provide some guidelines to choose the triplet ($\mathbf{C}$, τ_j, g) according to the objectives of the user and the nature of the data.

Practical guidelines for using RGCCA and SGCCA

There are eight steps, discussed hereafter, that need to be applied: (i) construction of the multiblock data set, (ii) preprocessing, (iii) definition of the between-block connections, (iv) determination of the shrinkage or sparsity 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.

Construction of the multiblock data set

The variables that compose each block have to be defined carefully: not only according to their nature (e.g. one block that contains all the voxels of an image, one block for all the metabolites, etc.) but also according to external information. Nowadays, a huge amount of external information is available and can be used to define each block. For example, a block that contains all the metabolites can be divided into several data blocks; hence, metabolites belonging to one pathway are gathered within the same block. A block that contains all the voxels of an image can be grouped by regions: voxels belonging to one specific region are then gathered within the same block. This grouping strategy makes more interpretable blocks and facilitates the interpretation of the RGCCA/SGCCA model. RGCCA/SGCCA can be viewed as a 'divide and conquer' strategy that allows incorporating prior information when defining the blocks.

Preprocessing

In general, and especially for the covariance-based criterion, the data might be preprocessed to ensure comparability between variables and blocks. To make variables comparable, standardization is applied (zero mean and unit variance). To make blocks comparable, a strategy is to divide each block by the square root of its number of variables [16]. This two-step procedure leads to $\text{trace}\left(\mathbf{X}_j^\top \mathbf{X}_j\right) = n$ for each block (i.e. the sum of the eigenvalues of the covariance matrix of $\mathbf{X}_j$ is equal to 1 whatever the block). Such a preprocessing will be implicitly used throughout this article.

Another way to make blocks more comparable is to divide each block $\mathbf{X}_j$ by the square root of the first eigenvalue of $\frac{1}{n}\mathbf{X}_j^\top \mathbf{X}_j$. This is exactly the normalization procedure used for MFA [18]. The rationale of this normalization is the same as in PCA where variables are standardized to have the same influence in the analysis; here, it can be seen as an extension to blocks of variables where the first singular value plays the role of the standard deviation. This second strategy may be preferred in a situation where the numbers of uninformative and noisy variables are unbalanced between blocks. General guidelines for centering and scaling in component analysis are available in [29]. Several normalization strategies used in the context of simultaneous component analysis are discussed in [30].

Definition of the design matrix C

The between-block connections are encoded through the design matrix $\mathbf{C} = (c_{jk})$; usually $c_{jk} = 1$ for two connected blocks and 0 otherwise. The customization of the design matrix can be defined according to biological assumptions reflecting the biology underlying the data blocks. For instance, multi-omics data (transcriptomics, metabolomics, etc...) and other modalities such as neuroimaging, electrophysiological data and scores of disease severity are routinely acquired to study the complexity of the neurodegenerative cascades. It can be roughly considered that the path between omics data and behavioral data is mediated by neuroimaging data (i.e. no direct relationship between

omics and behavioral data is imposed). The prior information on the between-block connections can be injected in the design matrix.

Furthermore, from a statistical viewpoint, the design matrix is a flexible way to reach one of the methods listed in Table 1.

Determination of the shrinkage parameters $0 < \tau_j < 1$, $j=1$, ..., J and the sparsity parameter s_j, $j=1$, ..., J

The RGCCA model introduces some extra parameters, particularly a shrinkage parameter. The shrinkage parameters $0 < \tau_j < 1$, $j = 1$, ..., J interpolate smoothly between maximizing the covariance (all $\tau_j = 1$) and maximizing the correlation (all $\tau_j = 0$). More precisely, we can define the choice of the shrinkage parameters by providing interpretations on the properties of the resulting block components:

- $\tau_j = 1$ yields the maximization of a covariance-based criterion. It is recommended when the user wants a stable component (large variance) while simultaneously taking into account the correlations between blocks. The user must, however, be aware that variance dominates over correlation.
- $\tau_j = 0$ yields the maximization of a correlation-based criterion. It is recommended when the user wants to maximize correlations between connected components. This option can yield unstable solutions in case of multicollinearity and cannot be used when a data block is rank deficient (e.g. $n < p_j$).
- $0 < \tau_j < 1$ is a good compromise between variance and correlation: the block components are simultaneously stable and as well correlated as possible with their connected block components. This setting can be used when the data block is rank deficient. Ledoit and Wolf [31] consider $\mathbf{M}_j = \tau_j \mathbf{I} + (1 - \tau_j)(1/n)\mathbf{X}_j^\top \mathbf{X}_j$ as a shrinkage estimate of the true covariance matrix Σ_{jj} related to block j. In case of multicollinearity within blocks or when the number of observations is smaller than the number of variables ($P_j \gg n$), the sample covariance matrix $(1/n)\mathbf{X}_j^\top \mathbf{X}_j$ is a poor estimation of the true covariance matrix. The usual strategy for finding a better estimation is to consider the class of linear combinations of the identity matrix and the sample covariance matrix, $\{\tau_j \mathbf{I} + (1 - \tau_j)(1/n)\mathbf{X}_j^\top \mathbf{X}_j\}$ [32]. Shrinkage parameters between 0 and 1 allow stepping closer to the correlation criterion, even in the case of high multicollinearity or when the number of individuals is smaller than the number of variables. For each block, the determination of the shrinkage parameter is made fully automatic by using one of the various formulas that have been proposed for finding an optimal shrinkage parameter [32]. Depending on the context, the shrinkage parameters should also be determined based on V-fold cross-validation.

Barker and Rayens [33] PLS for discrimination offer a good opportunity to illustrate the impact of the shrinkage parameters. They consider a block $\mathbf{X}$ of explanatory variables and a block $\mathbf{Y}$ of dummy variables describing a categorical variable. They are looking for a block component $\mathbf{Xa}$ (with $\mathbf{a}$ normalized) and a standardized component $\mathbf{Yb}$ maximizing the following criterion:

$$\max_{\mathbf{a}, \mathbf{b}} \quad \text{cor}^2(\mathbf{Xa}, \mathbf{Yb}) \times \text{var}(\mathbf{Xa}).$$

The rationale of the Barker and Rayens's criterion is based on the following idea: we are not looking for a block component $\mathbf{Yb}$ that explains its own block well (as $\mathbf{Y}$ is a group coding matrix) but

one that correlates with $\mathbf{Xa}$, hence removing from the covariance criterion $(\text{cov}^2(\mathbf{Xa}, \mathbf{Yb}) = \text{cor}^2(\mathbf{Xa}, \mathbf{Yb}) \times \text{var}(\mathbf{Xa}) \times \text{var}(\mathbf{Xb}))$, the $\text{var}(\mathbf{Yb})$ part. Using the RGCCA formalism, the Barker and Rayens's optimization problem is recovered as follows:

$$\max_{\mathbf{a}, \mathbf{b}} \quad \text{cov}^2(\mathbf{Xa}, \mathbf{Yb}) \quad \text{s.t.} \quad \|\mathbf{a}\|_2 = 1 \text{ and } \text{var}(\mathbf{Yb}) = 1,$$

that is for $\tau_X = 1$ and $\tau_Y = 0$.

The choice to set $\tau_X = 1$ and more generally to set τ_j equals to 1 in the optimization problem in Equation (1) is to some extent surprising. Indeed, it yields a sample covariance matrix equal to the identity for each block. It corresponds to the highest level of regularization that can be applied to RGCCA. The level of regularization can be relaxed by decreasing the value of τ_j. However, in high-dimensional settings, the highest level of regularization has proven to be necessary or even insufficient [34, 35]. Additional penalties that promote sparsity are often required. The sparsity parameter s_j, $j = 1$, ..., J is usually set based on cross-validation procedures (see next section for an illustration). Alternatively, values of s_j, $j = 1$, ..., J can simply be chosen to result in desired amounts of sparsity.

Choice of the scheme function g

It is possible to choose any continuous convex function. In the literature, classical scheme functions are $g(x) = x$ (horst scheme), $g(x) = |x|$ (centroid scheme), $g(x) = x^2$ (factorial scheme) or, more generally, for any even integer m, $g(x) = x^m$ (m-scheme). The horst scheme penalizes structural negative correlation between block components, while both the centroid scheme and the m-scheme enable two components to be negatively correlated. 'How the results of RGCCA/SGCCA depend on the values of m?' The answer to this question is related to the notion of fairness. According to [22], a fair model is a model where all blocks contribute equally to the solution in opposition to a model dominated by only a few of the J sets. If fairness is a major objective, the user must choose m=1. m>1 is preferable if the user wants to discriminate between blocks [2]. In practice, m is equal to 1, 2 or 4. The higher the value of m, the more the method acts as block selector [2].

Determination of the number of block components

Cross-validation is usually used to determine the number of block components to retain. Depend on the context (supervised or unsupervised), two types of cross-validation can be considered in the framework of RGCCA/SGCCA:

a. When the analysis is oriented toward the prediction of a specific phenotype, then the number of components per block can be selected based on the cross-validated prediction accuracy.
b. When no external information is available, then the number of components per block can be estimated as follows. For each block j, some percent of the elements of $\mathbf{X}_j$ is removed at random from the data matrix. The RGCCA block components are estimated from this partially observed data set. For each block, the missing values are imputed using the reconstruction formula. The number of components that results in the lowest sum of squared errors of the missing values is retained.

Besides, the average variance explained (AVE) by a block component y_j can also inform on the number of component to retain. The AVE of X_j, denoted by $AVE(X_j)$, is defined as:

$$AVE(X_j) = \frac{1}{p_j} \sum_{h=1}^{p_j} cor^2(x_{jh}, y_j).$$

$AVE(X_j)$ varies between 0 and 1 and reflects the proportion of variance captured by y_j. The number of block components to retain for X_j can be determined using the 'elbow' criterion or alternatively, the number of components that explains a predefined percentage of the total variance of X_j.

Visualization of the results

As a component-based method, RGCCA/SGCCA provides the users with graphical representations, including factor plot, correlation circle and biplot. These graphical displays allow visualizing the sources of variability within blocks, the relationships between variables within and between blocks and the amount of correlation between blocks.

Assessment of the reliability of parameter estimates

It is possible to use a bootstrap resampling method [36, 37] to assess the reliability of parameter estimates obtained using RGCCA/SGCCA. B bootstrap samples of the same size as the original data are repeatedly sampled with replacement from the original data. RGCCA/SGCCA is then applied to each bootstrap sample to obtain the RGCCA/SGCCA 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. As 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.

For SGCCA, the percentage of times a specific variable had a non-null weight across bootstrap sample can be derived. In addition, the stability of the selected variables can be measured according to the Fleiss'κ score [38] that estimates the agreement among the B bootstrap samples. The Fleiss'κ score is always ≤ 1, and the higher the value of κ is, the more stable the methods are with respect to sampling. This resampling procedure, intuitive and pragmatic, is classically used in the PLS community. We may note that alternative resampling-based strategy for variable and stability selection could be considered [39]. This alternative approach has been tested for simultaneous component analysis [40].

Case study: the SCA data set

Description of the SCA data set

Neurodegenerative disorders have become the leading cause of disability in Western societies, e.g. SCAs that are autosomal dominant diseases responsible for severe movement disorders. Heterogeneous and high-dimensional sources of information such as omics data (transcriptomics, metabolomics, etc.) and other modalities such as neuroimaging and/or electrophysiological data are routinely acquired to study such complex diseases. Disease mathematical models are thus critically needed to identify biomarkers that are relevant to disease mechanisms and can be used in therapeutic trials. As gene-based therapeutic approaches are being developed in SCA [41], it becomes increasingly important to identify readouts for trials with sufficient effect sizes. Clinical scores are useful, but insufficient, and a single biomarker is likely to fail reflecting the complexity of the neurodegenerative cascades leading to the onset and progression of SCA. An integrated multimodal biomarkers approach is therefore needed to (i) better understand disease pathophysiology and (ii) generate composite scores with greater effect sizes than isolated biomarkers.

SCA belongs to the group of polyglutamine repeat disorders and is characterized by a predominant atrophy of two brain regions: the cerebellum and the pons. More than 40 genetically different SCAs have been defined. The most common—SCA1, SCA2 and SCA3, which together affect about half of the families with a history of SCA—are caused by abnormal CAG repeat expansions, encoding elongated polyglutamine tracts within the proteins associated with each type [42]. Progressive cerebellar ataxia is the prominent symptom of all SCAs. In SCA7, patients present with additional non-neurological signs commonly seen in patients with mitochondrial dysfunction such as pigmentary retinopathy and cardiomyopathy. Depending on the SCA genotype, CAG repeat length explains about 50–70% of the variability in age at onset, i.e. individuals with longer repeats tend to have an earlier onset [43].

The volume of the pons has been shown to be the most sensitive to change in patients with SCA [44], including at the presymptomatic phase of the disease in individuals carrying abnormal CAG repeats but who have not yet developed symptoms [45]. Accordingly, the pons volume is likely to closely reflect disease progression and can also be studied longitudinally in controls, unlike motor scales evaluating cerebellar dysfunctions. Therefore, following previous work that we conducted on metabolic dysfunction in polyglutamine repeat disorders [46–48], we chose to perform multiblock analyses to discover relevant associations between the pons volume and various metabolic modalities—calorimetry, metabolomics and lipidomics on plasma, and metabolic imaging by magnetic resonance

Table 2. Characteristics of the SCA cohort

	Controls	SCA1	SCA2	SCA3	SCA7	*P-value*
Number of subjects	35	18	14	22	13	
Sex (M/F)	17/18	9/9	8/6	10/12	7/6	
BMI (kg/m^2)	25±4	24±6	27±5	24±5	23±3	0.104
Age at examination (years)	48±13	45±15	46±12	50±11	46±14	0.735
Age at disease onset (years)	–	41±12	35±11	42±11	38±13	0.397
SARA score (/40)	0.8±1	10±6	14±7	14±7	10±8	
Disease CAG repeats	–	48±7	40±3	70±6	43±5	

Table 3. Description of the SCA multiblock data set

Block $\mathbf{X}_j$	Number of variables for $\mathbf{X}_j$	Modalities
$\mathbf{X}_1$: Arginine_Proline	$p_1 = 14$	Metabolic
$\mathbf{X}_2$: BCAA_Threonine	$p_2 = 9$	pathways
$\mathbf{X}_3$: Carnitine_Lysine	$p_3 = 7$	including
$\mathbf{X}_4$: CE_sterols_bile_acids	$p_4 = 47$	754 metabolites
$\mathbf{X}_5$: Essential_fatty_acids	$p_5 = 15$	
$\mathbf{X}_6$: Fatty_acids_Ketone_bodies	$p_6 = 23$	
$\mathbf{X}_7$: GABA_Glutamine_Histidine	$p_7 = 18$	
$\mathbf{X}_8$: Glucose_Alanine_Pyruvate	$p_8 = 4$	
$\mathbf{X}_9$: Glycerides	$p_9 = 177$	
$\mathbf{X}_{10}$: Glycine_Serine	$p_{10} = 16$	
$\mathbf{X}_{11}$: Krebs_cycle	$p_{11} = 5$	
$\mathbf{X}_{12}$: Phenylalanine_Tyrosine	$p_{12} = 12$	
$\mathbf{X}_{13}$: Phospholipids	$p_{13} = 292$	
$\mathbf{X}_{14}$: Purines	$p_{14} = 14$	
$\mathbf{X}_{15}$: Pyrimidines	$p_{15} = 9$	
$\mathbf{X}_{16}$: Sphingolipids	$p_{16} = 56$	
$\mathbf{X}_{17}$: Tryptophan	$p_{17} = 23$	
$\mathbf{X}_{18}$: Urea_cycle	$p_{18} = 5$	
$\mathbf{X}_{19}$: Various	$p_{19} = 8$	
$\mathbf{X}_{20}$: MRS	$p_{20} = 19$	MRS of the cerebellum
$\mathbf{X}_{21}$: CAL	$p_{21} = 3$	Calorimetry information
$\mathbf{X}_{22}$: Superblock	$p_{22} = 776$	$\mathbf{X}_{22} = [\mathbf{X}_1, \ldots, \mathbf{X}_{21}]$
$\mathbf{X}_{23}$: Pons	$p_{23} = 1$	The volume of the pons

spectroscopy (MRS)—in patients with SCA compared with controls, to gain insight into the pathophysiology of SCA. Our ultimate goal—outside the scope of these analyses—is to study prospectively these biomarkers in longitudinal studies and generate composite scores with greater effect sizes than the pons volume alone.

Patients and controls

The SCA study (NCT 01470729) was approved by the local ethical committee (AOM10094, CPP Ile de France VI, Ref: 105–10) and performed in a unique cohort of patients with SCA—SCA1 ($n = 18$), SCA2 ($n = 14$), SCA3 ($n = 22$) and SCA7 ($n = 13$). Healthy controls ($n = 35$) with similar sex ratio, age and body mass index (BMI) than patients were also recruited. All participants signed informed consent to be included in the study. Their demographic characteristics are summarized in Table 2. The scale for the assessment and rating of ataxia (SARA, score up to 40) was used to evaluate the severity of the disease [49]. The four SCA subtypes were comparable in terms of duration of disease and SARA scores.

Application of SGCCA to the SCA data set

We collected standard clinical and brain volumetric metrics in our cohort of SCA patients and healthy controls, and then jointly analyzed modalities (or blocks) reflecting metabolic regulations using calorimetry, metabolomics and lipidomics on plasma, and metabolic imaging by MRS. A full description of the methods used for the acquisition of each modality is available as Supplemental Materials. The main objective of this integrative analysis was to identify variables within each block that (i) well explain their own block and (ii) influence the relationships

between 'connected' blocks. The example of the SCA data set was well suited to illustrate the versatility and relevance of SGCCA, as the number of variables within each block made it difficult to identify the most important variables, so that a variable selection procedure was needed. In this section, we intend to instantiate the eight-step guideline described in 'Practical guidelines for using RGCCA and SGCCA' section. Moreover, some additional advices to set up the extra parameters according to the nature of the data and the scientific objectives are given. We then illustrate how relationships between the most relevant variables can be displayed and the results interpreted by visualizing the observations and variables in a common space.

Construction of the multiblock data set

The SCA data set was organized into 23 blocks. A detailed description of each block, including the number of variables per block, is reported in Table 3. Annotated metabolites were classified into metabolite sets mapping various biochemical pathways. Nineteen sets were proposed including 754 metabolites classified by the confidence level of annotation and detected using our metabolomic and lipidomic methods. These blocks $\mathbf{X}_1, \ldots, \mathbf{X}_{19}$ were defined based on biological knowledge about metabolic pathways from KEGG (Kyoto Encyclopedia of Genes and Genomes), HMDB (Human Metabolome Database) and literature [50–56]. Blocks $\mathbf{X}_{20}$ and $\mathbf{X}_{21}$ contained information on brain MRS, denoted MRS, and calorimetry information, denoted CAL. In the framework of CPCA and HPCA methods, a superblock defined as the concatenation of all the blocks is also used. In the SCA data set, the superblock was defined as $\mathbf{X}_{22} = [\mathbf{X}_1, \ldots, \mathbf{X}_{21}]$, and the corresponding global components were derived. The space spanned by the global components was viewed as a compromise space that integrated all the modalities. This global space was useful for visualization and eased the interpretation of the results. Finally, $\mathbf{X}_{23}$ contained the volume of the pons.

Preprocessing

Adjustments for confounding factors (age, gender and BMI) were carried out by residualization (before preprocessing) for each variable of the SCA data set. Residualization consists in regressing each block by age, gender and BMI. To make blocks more comparable, the residual variables were standardized (zero mean and unit variance) and then divided by $\sqrt{p_j}$ within each block.

Definition of the design matrix C

In the search of biomarkers associated with the four subtypes of SCA—SCA1, SCA2, SCA3 and SCA7—we applied SGCCA to identify variables from the 21 blocks associated with the pons volume. The between-block connections associated with this objective of analysis are presented in Figure 1. We chose a CPCA structure oriented toward the explanation of the volume of the pons by imposing an additional connection between the superblock and the pons. The 'divide and conquer' strategy, by incorporating prior knowledge in the definition of the blocks, yielded valuable improvements and more interpretable results.

Choice of the scheme function g

In this case, it was not expected that all the blocks, especially the metabolic pathways, contributed equivalently to the process. The block selector behavior of SGCCA was favored by using the scheme function $g(x) = x^4$.
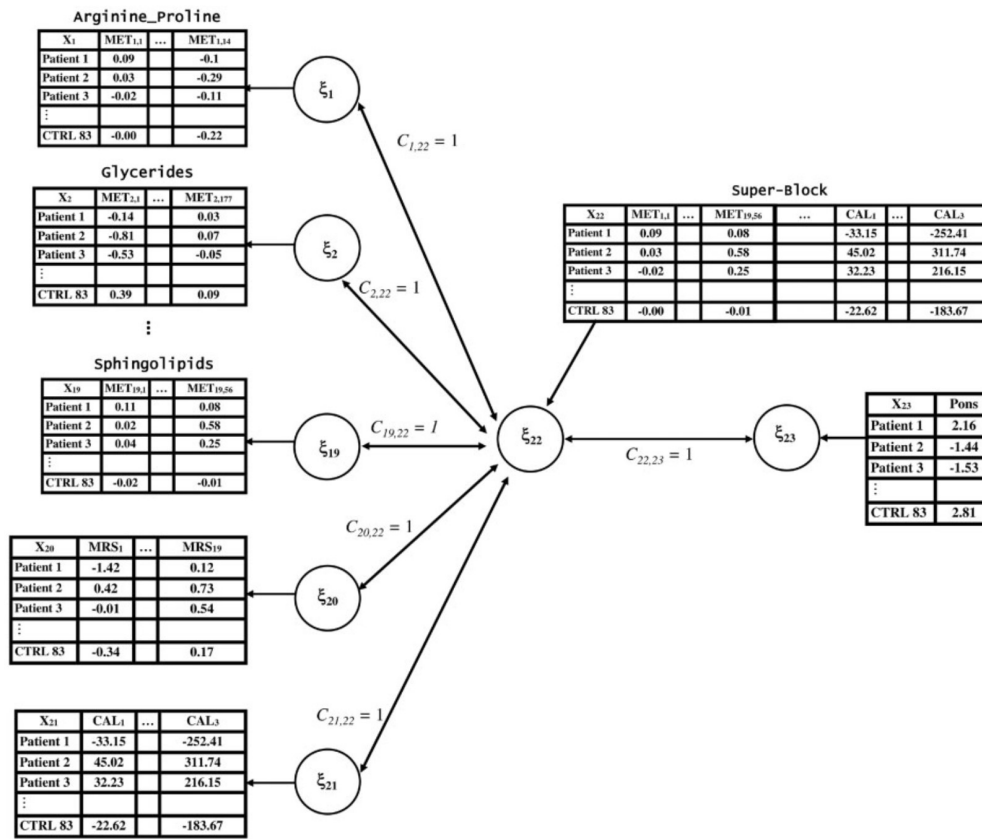
Figure 1. Between-block connections. $\mathbf{X}_1,\ldots, \mathbf{X}_{21}$ are connected to the superblock $\mathbf{X}_{22}$, and $\mathbf{X}_{22}$ is connected to the volume of the pons $\mathbf{X}_{23}$. These between-block connections are encoded through the design matrix C: $c_{j,22} = 1$, $j = 1, \ldots 21$, $c_{22,23} = 1$ and $c_{jk} = 0$ otherwise.

Determination of the sparsity parameter and the number of block components

SGCCA requires determining the sparsity parameters. For each block $\mathbf{X}_j$, the sparsity parameter s_j, $j = 1, \ldots, J$ was set using a leave-one-out cross-validation procedure. The value of the parameter s_j was chosen in a range defined by the following formula $1 + \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\} \times \sqrt{p_j}$, which allowed us to impose the same degree of sparsity for all the blocks. To select the optimal value, linear models predicting the volume of the pons with respect to the block components were performed, and the optimal parameter was selected with respect to the mean squared error of these models. The optimal values were equal to $s_j = 1 + 0.2 \times \sqrt{p_j}$, $j = 1, \ldots, J$.

Moreover, using a deflation strategy, four components per block were built. We denote by $\mathbf{y}_j^{(h)}$ (respectively, $\mathbf{w}_j^{(h)}$) the *h*th block component (respectively, *h*th block weight vector) associated with $\mathbf{X}_j$.

Visualization

As the fourth global component was the most discriminant between patients and controls, the graphical display of the individuals obtained by crossing the global components $\mathbf{y}_{22}^{(1)}$ and $\mathbf{y}_{22}^{(4)}$ and marked with their status (SCA1, SCA2, SCA3, SCA7 and controls) is shown in Figure 2. It is noteworthy that, despite some overlap, the first global component exhibited a separation among some SCA groups, especially patients with SCA7 who were mainly grouped at the bottom. Moreover, the fourth global component captured the discriminative information between

patients and healthy controls as controls concentrated on the right and patients on the left.

Figure 3 shows the variables projected on the compromise space. The sparsity-inducing penalty of SGCCA made the interpretation of the variable space easier. Indeed, only the variables associated with non-null elements in the block weight vector $\mathbf{w}_j^{(1)}$ and $\mathbf{w}_j^{(4)}$, $j = 1, \ldots, J$ (i.e. the ones that contribute to the construction of the first and fourth dimensions) were projected on the compromise space. A variable that is highly expressed for a category of individuals will be projected with a high weight (far from the origin) in the direction of that category. Likewise, the most discriminant variables between patients and controls appeared to be metabolites measured by MRS in the vermis such as total creatine, a marker of energy metabolism and myoinositol (myoIns), a putative glial marker (Figure 3). Interestingly, we previously identified these variables as significantly different between patients and controls [52]. We also showed that these metabolites were associated with SARA scores, which reflect higher disease severity [52]. Moreover, the separation among patients with SCA, and especially patients with SCA7, seemed to be driven by certain lipid species detected in plasma by lipidomic analyses such as sphingolipids and phospholipids (Figure 3).

Figure 3 allows visualizing relationships between variables belonging to the different blocks. This figure suggests relationships between blocks that can be confirmed by a block clustering. As the fourth dimension was the most informative axis for the explanation of the pons, we considered the variables that contribute to the construction of $\mathbf{y}_j^{(4)}$, $j = 1, \ldots, 21$. Let $\mathbf{X}_j^*$ be the block that contains the variables associated with non-null elements in the
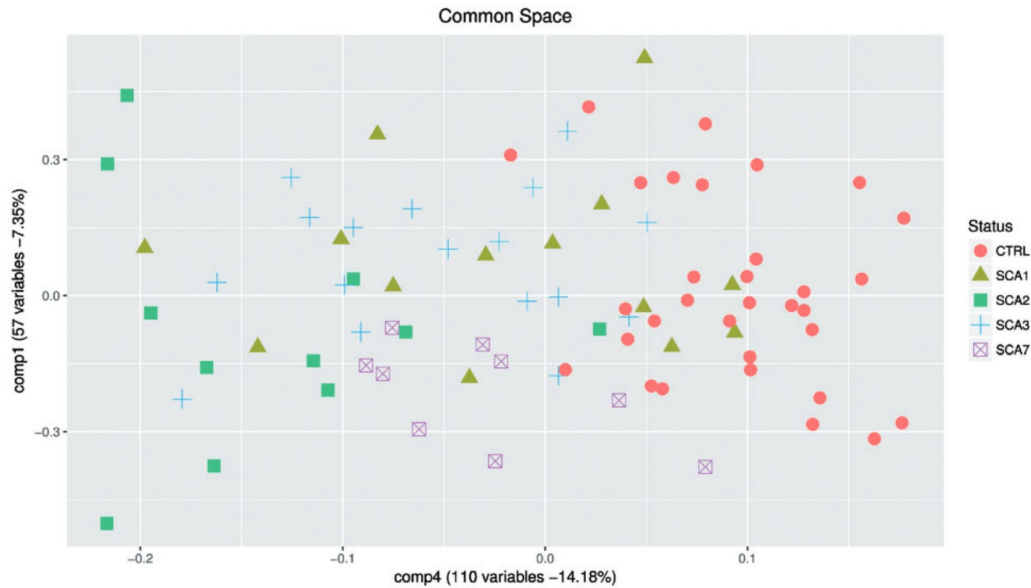
Figure 2. Sample space associated with the dimensions 1 and 4 of the superblock. Individuals are marked according to the status (CTRL, SCA1, SCA2, SCA3 and SCA7). CTRL: healthy controls.

block weight vector $w_j^{(4)}$. This subsection presents the block clustering of $\mathbf{X}_1^*, \ldots, \mathbf{X}_{21}^*, \mathbf{X}_{23}$ based on the McKeon's measure [57]. The McKeon's measure quantifies the homogeneity of a set of block components and is defined by the following equation:

$$r_1(\mathbf{X}_1\mathbf{w}_1, \ldots, \mathbf{X}_J\mathbf{w}_J) = \frac{\frac{1}{J(J-1)/2}\sum_{j<k}\mathrm{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)}{\frac{1}{J}\sum_j \mathrm{var}(\mathbf{X}_j\mathbf{w}_j)} \quad (3)$$

Equation (3) allows evaluating the homogeneity of the solution of any multiblock component methods. The computation of the McKeon's measure was carried out using RGCCA (full between-block connections, $\tau_j = 1$ for all blocks, and $g(x) = x$ for a fair analysis). Figure 4 represents the resulting block clustering of $\mathbf{X}_1^*, \ldots, \mathbf{X}_{21}^*, \mathbf{X}_{23}$. Blocks that were the most closely related (e.g. GABA–glutamine–histidine and Krebs cycle) contained variables that were partially redundant, as they belonged to more than one pathway and could thus serve as an internal validation. As previously discussed, the volume of the pons, the most distinctive feature in this model between patients and controls, clustered with the vermis MRS profile. Lipid species, including sphingolipids and phospholipids, also clustered with one another providing further validation to our model.

Assessment of the reliability of parameter estimates
To assess the reliability of parameter estimates obtained using SGCCA, 1000 bootstrap samples were derived. SGCCA was then applied to each bootstrap sample (with the sparsity parameters determined at the previous step) to obtain estimates $\mathbf{w}_j^b$, where j denotes the block index and b the bootstrap sample index. The percentage of times a specific variable had a non-null weight was therefore derived. Figure 5 reports those percentages for the blocks that contributed mostly to the construction of the first global component.

Among the lipid species that tended to separate patients with SCA7 from other patients, certain ceramides (Cer) were especially represented in the sphingolipids group, as well as

certain phosphatidylcholines (PCs) in the phospholipids group (Figure 5A and B, respectively). To our knowledge, to date, there has been no metabolomic or lipidomic studies conducted in SCA patients. However, one lipidomic analysis was performed in the cerebellum of a preclinical model of SCA2 and found significant changes in some sphingolipids and cholesterol by-products [58]. Likewise, although conclusions cannot be made without further biological validation, it is noteworthy that both lipid classes, Cer and PC, are highly expressed in the retina [59]. Furthermore, defects in their synthesis are associated with pigmentosa retinopathy [60, 61], which is a distinctive feature of SCA7 compared with SCA1, SCA2 and SCA3.

Discussion and conclusion

R/SGCCA stands as a unique, general and original way for analyzing high-dimensional multiblock data sets. It allows the selection of a few meaningful variables that underline the between-block connections encoded by the design matrix **C**. This design matrix is highly modular to fit any prior knowledge the user has on the links between blocks. The variable selection property results in models more easily interpretable than a model based on all the variables. Being able to select variables means that one can also study the stability of the variable selection process and possibly deduce patterns in the way the variables are selected under sampling. Moreover, the selection of a few meaningful variables from longitudinal studies will enable their combination into a composite score. Such a composite score can be used as a proxy for disease severity and acts as a basis for future therapeutic studies. Indeed, composite scores are likely to provide both a better reflection of the disease process pathology and a larger effect size than any biomarker alone. This is crucial for the assessment of experimental treatments in many neurodegenerative conditions, and especially in rare diseases like SCA where patient's recruitment is challenging. Finally, we showed that having blocks of heterogeneous sizes and nature is taken into account routinely by SGCCA.
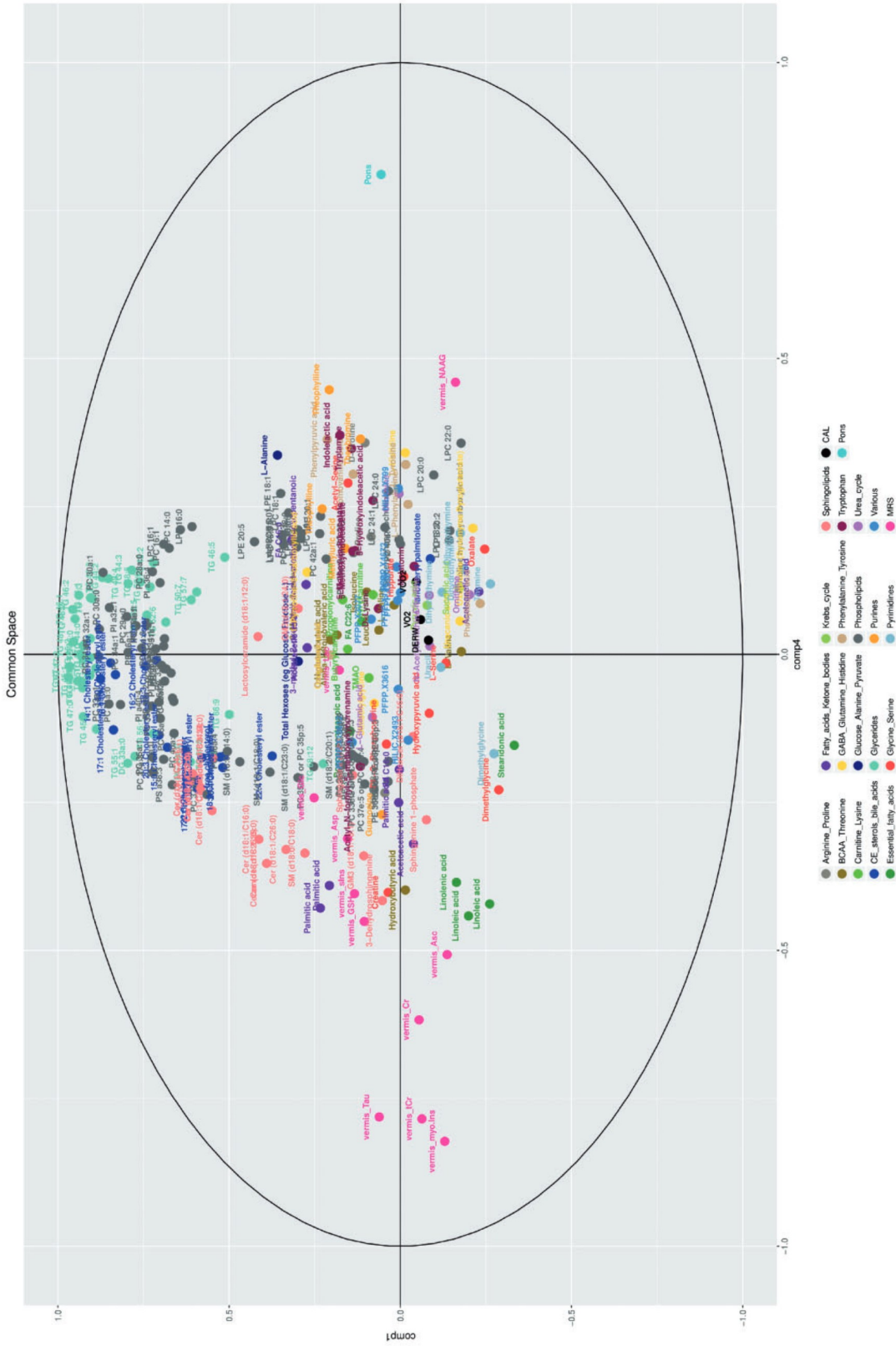
Figure 3. Variables space associated with the dimensions 1 and 4 of the superblock. Only the variables that contribute to the construction of components 1 and 4 are visualized in the variables space.
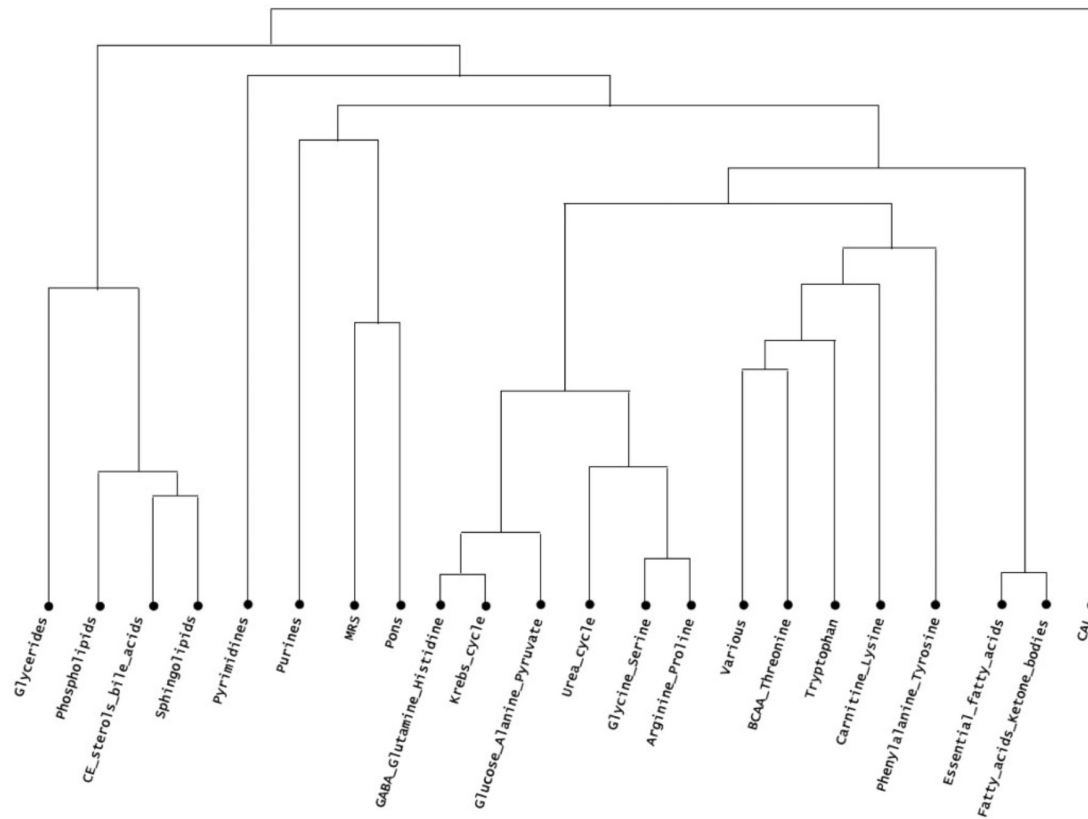
Figure 4. Dendrogram for block clustering based on McKeon measure derived using RGCCA with a full between-block connections, $\tau_j = 1$ for all blocks, and $g(x) = x$.
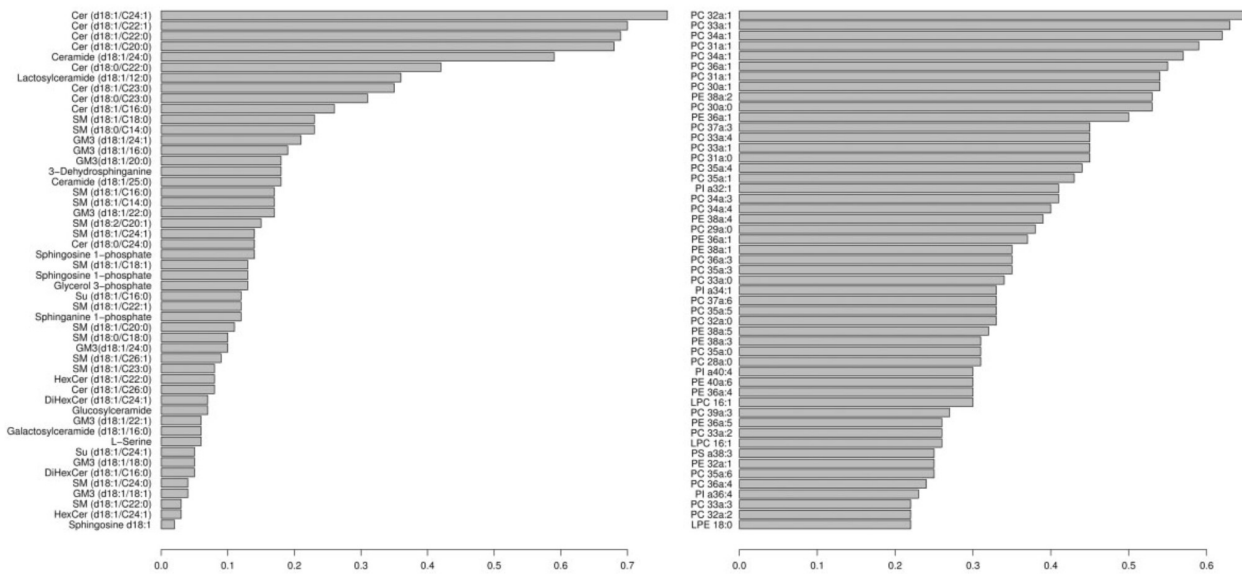

Figure 5. (A) Percentages of times the 'Sphingolipid' variables participate to the construction of the block component. (B) Percentages of times the 'Phospholipids' variables participate to the construction of the block component. We presented the most important variables, as the block contained 292 variables.

Likewise, the application of SGCCA on our SCA data set pinpoints at a possible role of some lipid species in the pathophysiology of SCA7 compared with other SCA, which merits further metabolic explorations.

Of note, RGCCA can also be applied to longitudinal data using the multiway formalism, which accounts for multiple measurements (either in time or in type of acquisition) of a given feature [62]. Multigroup structure (i.e. same sets of

variables observed on different groups of individuals) can also be tackled with RGCCA: the aim is to uncover similar relationships between variables across the various groups [63]. Beyond the example data set used in this study, this framework proves equally efficient to manage and interpret a large variety of biological data types, typically information produced by next-generation sequencing approaches (e.g. DNA-seq, RNA-seq, Methyl-seq, etc.) that are increasingly used to further investigate normal or pathological biological processes.

Key Points

- The RGCCA-based integrative procedure requires the setting of extra parameters that need to be carefully adjusted. We provide practical guidelines for the use of RGCCA/SGCCA.
- The flexibility and usefulness of RGCCA/SGCCA was illustrated on a unique cohort of patients with four genetic subtypes of SCA, in which we obtained multiple data sets from brain volumetry, MRS and metabolomic and lipidomic data sets.
- We show how to graph RGCCA output.

Supplementary Data

Supplementary data are available online at http://bib.oxford journals.org/.

Availability of data and materials

The data that support the findings are part of a larger ongoing study. They are thus not publicly available yet. However, the data presented in this article are available from the authors on reasonable request.

Ethics approval

Ethical approval for experiments involving patients was granted by the Comité de Protection des Personnes-Ile de France Paris VI (ID RCB: 2010-A01324-35).

References

1. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis. *Psychometrika* 2011;**76**:257–84.
2. Tenenhaus M, Tenenhaus A, Groenen PJF. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Accepted Psychometrika* 2017.
3. Tenenhaus A, Philippe C, Guillemot V, et al. Variable selection for generalized canonical correlation analysis.. *Biostatistics* 2014;**15**(3):569–83.
4. Günther OP, Shin H, Nq RT, et al. Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection study. *OMICS* 2014;**18**(11):682–95.
5. Meng C, Zeleznik OA, Thallinger GG, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**(4):628–41.
6. Meng C, Kuster B, Culhane AC, et al. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;**15**(1):1.
7. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933;**24**:417–41.
8. Hotelling H. Relation between two sets of variates. *Biometrika* 1936;**28**:321–77.
9. Tucker LR. An inter-battery method of factor analysis. *Psychometrika* 1958;**23**:111–36.
10. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. *Proc Conf Matrix Pencils* 1983;**973**:286–93.
11. Van den Wollenberg A. Redundancy analysis – an alternative to canonical correlation analysis. *Psychometrika* 1977;**42**:207–19.
12. Carroll J. A generalization of canonical correlation analysis to three or more sets of variables. *Proc 76th Conv Am Psych. Assoc* 1968;**3**:227–8.
13. Carroll J. Equations and tables for a generalization of canonical correlation analysis to three or more sets of variables. 1968, Unpublished Companion paper to Carroll.
14. Wold S, Kettaneh N, Tjessem K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J Chemom* 1996;**10**:463–82.
15. Chessel D, Hanafi M. Analyses de la co-inertie de K nuages de points. *Rev Stat Appl* 1996;**44**(2):35–60.
16. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemom* 1998;**12**:301–21.
17. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemom* 2003;**17**(6):323–37.
18. Escofier B, Pagès J. Multiple factor analysis, (AFMULT package). *Comput Stat Data Anal* 1994;**18**:121–40.
19. Horst P. Relations among m sets of variables. *Psychometrika* 1961;**26**:126–49.
20. Kettenring J. Canonical analysis of several sets of variables. *Biometrika* 1971;**58**:433–51.
21. Hanafi M. PLS Path modelling: computation of latent variables with the estimation mode B. *Comput Stat* 2007;**22**:275–92.
22. Van de Geer JP. Linear relations among k sets of variables. *Psychometrika* 1984;**49**:70–94.
23. Hanafi M, Kiers H. Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Comput Stat Data Anal* 2006;**51**:1491–508.
24. Kramer N. Analysis of high dimensional data with partial least squares and boosting. Doctoral dissertation, Technical University of Berlin, 2007.

25. Wold H. Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation: Part 2*. North-Holland, Amsterdam: K.G. Jöreskog and H. Wold, 1982, pp. 1–54.

26. Tenenhaus A and Guillemot, V. RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis forMulti-Block Data. R package version 2.12. 2017. https://cran.r-project.org/web/packages/RGCCA/.

27. Tenenhaus A, Philippe C, Frouin V. Kernel generalized canonical correlation analysis. *Comput Stat Data Anal* 2015;**90**:114–31.

28. Tenenhaus M, Vinzi VE, Chatelin YM, et al. PLS path modeling. *Comput Stat Data Anal* 2005;**48**(1):159–205.

29. Bro S, Smilde AK. Centering and scaling in component analysis. *J Chemom* 2003;**17**(1):16–33.

30. Van Deun K, Smilde AK, van der Werf MJ, et al. A structured overview of simultaneous component based data integration. *BMC Bioinformatics* 2009;**10**:246.

31. Ledoit O, Wolf M. A well conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal* 2004;**88**:365–411.

32. Schäfer J, Strimmer KA. Shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biolgy* 2005;**4**(1):32.

33. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom* 2003;**17**:166–73.

34. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 2004;**10**(6):989–1010.

35. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Ser B Stat Methodol* 2010;**72**(1):3–25.

36. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat* 1979;**7**:1–26.

37. Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987;**82**:171–85.

38. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**(5):378.

39. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B Stat Methodol* 2010;**72**(4):417–73.

40. Gu Z, Van Deun K. A variable selection method for simultaneous component based data integration. *Chemom Intell Lab Syst* 2016;**158**:187–99.

41. Keiser MS, Kordower JH, Gonzalez-Alegre P, et al. Broad distribution of ataxin 1 silencing in rhesus cerebella for spinocerebellar ataxia type 1 therapy. *Brain* 2015;**138**(Pt 12):3555–66.

42. Rüb U, Schöls L, Paulson H, et al. Clinical features, neurogenetics and neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. *Prog Neurobiol* 2013;**104**:38–66.

43. Durr A. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. *Lancet Neurol* 2010;**9**(9):885–94.

44. Klaes A, Reckziegel E, Franca MC, Jr et al. MR Imaging in Spinocerebellar Ataxias: a systematic review. *AJNR Am J Neuroradiol* 2016;**37**(8):1405–12.

45. Jacobi H, Reetz K, du Montcel ST, et al. Biological and clinical characteristics of individuals at risk for spinocerebellar ataxia types 1, 2, 3, and 6 in the longitudinal RISCA study: analysis of baseline data. *Lancet Neurol* 2013;**12**(7):650–8.

46. Mochel F, Charles P, Seguin F, et al. Early energy deficit in Huntington disease: identification of a plasma biomarker traceable during disease progression. *PLoS One* 2007;**2**(7):e647.

47. Mochel F, Haller RG. Energy deficit in Huntington disease: why it matters. *J Clin Invest* 2011;**121**(2):493–9.

48. Adanyeguh IM, Rinaldi D, Henry PG, et al. Triheptanoin improves brain energy metabolism in patients with Huntington disease. *Neurology* 2015;**84**(5):490–5.

49. Schmitz-Hubsch T, du Montcel ST, Baliko L, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology* 2006;**66**:1717–20.

50. Wishart DS, Tzur D, Knox C. HMDB: the Human Metabolome Database. *Nucleic Acids Res* 2007;**35**:521–6.

51. Wishart DS, Knox C, Guo AC. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 2009;**37**:603–10.

52. Wishart DS, Knox C, Guo AC. HMDB 3.0 | The Human Metabolome Database in 2013. *Nucleic Acids Res* 2013;**41**:801–7.

53. Kanehisa M, Sato Y, Kawashima M, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.

54. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**:457–62.

55. Lamari F, Mochel F, Sedel F, et al. Disorders of phospholipids, sphingolipids and fatty acids biosynthesis: toward a new category of inherited metabolic diseases. *J Inherit Metab Dis* 2013;**36**:411–25.

56. Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2008;**3636**(**Suppl 1**):D623–31.

57. McKeon JJ. Canonical analysis: some relation between canonical correlation, factor analysis, discriminant analysis, and scaling theory. *Psychom Monogr* 1966;**13**.

58. Lastres-Becker I, Brodesser S, Lütjohann D, et al. Insulin receptor and lipid metabolism pathology in ataxin-2 knock-out mice. *Hum Mol Genet* 2008;**17**:1465–81.

59. Martin RE, Elliott MH, Brush RS, et al. Detailed characterization of the lipid composition of detergent-resistant membranes from photoreceptor rod outer segment membranes. *Invest Ophthalmol Vis Sci* 2005;**46**(4):1147–54.

60. McMahon A, Butovich IA, Kedzierski W. Epidermal expression of an Elovl4 transgene rescues neonatal lethality of homozygous Stargardt disease-3 mice. *J Lipid Res* 2011;**52**(6):1128–38.

61. Lamari F, Mochel F, Saudubray JM. An overview of inborn errors of complex lipid biosynthesis and remodelling. *J Inherit Metab Dis* 2015;**38**(1):3–18.

62. Tenenhaus AL, Brusquet L, Lechuga G. Multiway regularized generalized canonical correlation analysis. In 47èmes Journée de Statistique de la SFdS (JdS 2015), Lille, France, 2015.

63. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res* 2014;**238**:391–403.