Received: 14 November 2017 | Revised: 28 September 2018 | Accepted: 2 October 2018

DOI: 10.1002/bimj.201700248

RESEARCH PAPER

Biometrical Journal

Multiset sparse redundancy analysis for high-dimensional omics data

Attila Csala | Michel H. Hof | Aeilko H. Zwinderman

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands

Correspondence
Attila Csala, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, 1105 AZ, The Netherlands.
Email: a@csala.me

Abstract

Redundancy Analysis (RDA) is a well-known method used to describe the directional relationship between related data sets. Recently, we proposed sparse Redundancy Analysis (sRDA) for high-dimensional genomic data analysis to find explanatory variables that explain the most variance of the response variables. As more and more biomolecular data become available from different biological levels, such as genotypic and phenotypic data from different omics domains, a natural research direction is to apply an integrated analysis approach in order to explore the underlying biological mechanism of certain phenotypes of the given organism. We show that the multiset sparse Redundancy Analysis (multi-sRDA) framework is a prominent candidate for high-dimensional omics data analysis since it accounts for the directional information transfer between omics sets, and, through its sparse solutions, the interpretability of the result is improved. In this paper, we also describe a software implementation for multi-sRDA, based on the Partial Least Squares Path Modeling algorithm. We test our method through simulation and real omics data analysis with data sets of 364,134 methylation markers, 18,424 gene expression markers, and 47 cytokine markers measured on 37 patients with Marfan syndrome.

KEYWORDS

high-dimensional data, multivariate statistics, omics data, redundancy analysis

1 | INTRODUCTION

In recent years, technological advancement has enabled large amounts of biomolecular data generation from different biological levels, such as from the genome, epigenome, transcriptome, proteome, and metabolome. Omics data from multiple biological levels are often measured on the same patients that share a certain phenotype, and the conceptual information flow between these different biological levels is well defined by the central dogma of molecular biology (Crick, 1970). Integromics is an emergent research area with the interest of simultaneously analyzing multiple molecular and clinical data sources from various omics domains, in order to reveal complex biological pathways that are involved in certain phenotypes (Buescher & Driggers, 2016). In this paper, we describe a statistical tool that models the conceptual information flow between multiple omics data sets and extract genotypic data from those data sets in order to model biological pathways of the phenotype of interest. We refer to this method as the multiset sparse Redundancy Analysis (multi-sRDA).

Redundancy Analysis (RDA) is a well-known multivariate method that models the information flow between two data sets by maximizing the redundancy index between explanatory and response variables, thus RDA measures the effect of the explanatory

data set on the response data set (Johansson, 1981). RDA describes the relationships between data sets by finding a linear combination of the explanatory variables that explain the most variance of the response variables. Thus, if one wishes to perform an integrated analysis to explain the variability of phenotypes with multiple sets of genotypic data, multiset Redundancy Analysis (multi-RDA) is a prominent candidate.

Although RDA's conceptual model was already described by van den Wollenberg (1977), and has been widely used in fields such as ecology (Oksanen et al., 2007) and psychology (Israels, 1984), it has received little attention in high-dimensional genetic and genomic data analyses (Huang, Chaudhary, & Garmire, 2017). A well-known characteristic of omics data is its high dimensionality ($p \gg n$, where p is the number of variables and n is the number of observations). In this setting, RDA's software implementations are not applicable, since they rely on using standard multivariate least squares regression steps where high-dimensional covariance matrices are noninvertible (Oksanen et al., 2007).

Recently, we showed how penalization methods can be introduced for RDA to overcome the high dimensionality problem. By applying the Elastic Net (ENet) penalization to the multivariate regression equation in a Partial Least Squares (PLS) framework, we showed that sparse Redundancy Analysis (sRDA) is able to deal with high-dimensional genetic and genomic data and select the important explanatory variables with high precision (Csala, Voorbraak, Zwinderman, & Hof, 2017). In this paper, we show how sRDA can be extended into a multiset multivariate method so that the effect of multiple explanatory variables, from multiple data sets, on a set of outcome variables, measured on one data set, can be assessed simultaneously. Multi-sRDA is a particularly attractive method for omics data analysis since it is able to account for the directional information flow between data sets from different omics levels. In addition, Multi-sRDA provides easily interpretable sparse solutions through selecting the important explanatory variables that explain the most variation in their response data set, while excluding the unimportant variables from the final model that explain no to little of the variance of their response set.

In this paper, we extend our sRDA approach to multiple sets and we demonstrate multi-sRDA's capacity for high-dimensional omics data by conducting simulation studies and analyzing real biomedical data from three different omics domains, measured on the same 37 patients, all diagnosed with the same disease phenotype, namely, Marfan Syndrome. For the simulation studies, we generate multiple high-dimensional data sets to demonstrate multi-sRDA's ability to find the important explanatory variables in a high dimensional, multiple data set setting. For the real data analysis, we build a conceptual model that incorporates 364,134 methylation markers, 18,424 gene expression markers, and 47 serum cytokine markers. In this setting, multi-sRDA models the information transfer from the epigenome through the transcriptome to the proteome. Through this real data analysis, we show how multi-sRDA can be used to analyze multiple high-dimensional omics sets in order to find the combination of those foremost biomolecular markers from various biological levels that explain the most variance of the phenotypic variables, while modeling for the conceptual model of the central dogma of molecular biology.

This paper is organized as follows. Section 2 describes sRDA and multi-sRDA, with an algorithmic implementation of multi-sRDA in the PLS framework. Section 3 describes the simulation studies that are used to assess multi-sRDA's ability of finding the important explanatory variables, and Section 4 describes an application of multi-sRDA on real biomedical data. Section 5 discusses the findings of the real data analysis and concludes the paper.

2 | DIRECTIONAL MULTISET MULTIVARIATE ANALYSES FOR HIGH-DIMENSIONAL DATA

2.1 | Sparse redundancy analysis

We start with a description of sRDA for two data sets; for explanatory data set $\boldsymbol{X}$ and for response data set $\boldsymbol{Y}$. Consider data on n individuals distributed over two data sets, $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{X}$ is an $n \times p$-dimensional matrix containing p explanatory variables (i.e., $\boldsymbol{X} \in \mathbb{R}^{n \times p}$), and $\boldsymbol{Y}$ is a $n \times q$-dimensional matrix with q response variables (i.e., $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$). Let $\boldsymbol{\xi}$ be the $n \times 1$-dimensional latent variable of $\boldsymbol{X}$ (i.e., $\boldsymbol{\xi} = \boldsymbol{X}\boldsymbol{\alpha}$, where $\boldsymbol{\xi} \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{\alpha} \in \mathbb{R}^{p \times 1}$) and $\boldsymbol{\eta}$ the $n \times 1$-dimensional latent variable of $\boldsymbol{Y}$ (i.e., $\boldsymbol{\eta} = \boldsymbol{Y}\boldsymbol{\beta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{\beta} \in \mathbb{R}^{q \times 1}$). The column vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the associated weights of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, respectively.

RDA estimates the amount of variance of a given set of response variables in terms of a given set of explanatory variables. The objective of RDA is to define a linear combination of $\boldsymbol{X}$ (denoted as $\boldsymbol{\xi} = \boldsymbol{X}\boldsymbol{\alpha}$) that maximizes the sum of the squared correlations between $\boldsymbol{\xi}$ and $\boldsymbol{Y}$, that is,

$$\sum_{j=1}^{q} \mathrm{Cor}(\boldsymbol{\xi}, \boldsymbol{y}_j)^2. \tag{1}$$

The weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be estimated in an iterative multiple least-squared regression framework (Fornell, Barclay, & Rhee, 1988). Estimating $\boldsymbol{\beta}$ involves q univariate least-squared regressions

$$\beta_j = (\xi'\xi)^{-1}\xi'\boldsymbol{y}_j,$$

where $\boldsymbol{y}_j$ is the j-th column from the matrix $\boldsymbol{Y}$. Estimating $\boldsymbol{\alpha}$ involves a multivariate least squares regression, for which the standard closed-form solution can be used if the number of explanatory variables is smaller than the number of observations (i.e., $p < n$);

$$\boldsymbol{\alpha} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}. \tag{2}$$

Since in omics data the number of explanatory variables is typically much larger than the number of individuals (i.e., $p \gg n$), the closed-form solution for the multivariate regression step leads to multicollinearity issues. In this setting, the covariance matrix in Equation (2) is noninvertible. To solve this problem, we recently proposed using the ENet penalization, which involves both the LASSO and the Ridge regularization. ENet is applied to the multivariate regression step in the RDA framework (Csala et al., 2017), which turns Equation (2) into

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\text{argmin}}\boldsymbol{\alpha}' \left(\frac{\boldsymbol{X}'\boldsymbol{X} + \lambda_2\mathbf{I}}{1+\lambda_2} \right) \boldsymbol{\alpha} - 2\boldsymbol{\eta}'\boldsymbol{X}\boldsymbol{\alpha} + \lambda_1|\boldsymbol{\alpha}|_1, \tag{3}$$

where $|\boldsymbol{\alpha}|_1$ denotes the 1-norm form of vector $\boldsymbol{\alpha}$, λ_1 is the LASSO penalty parameter, and λ_2 is the Ridge penalty parameter.

Thus sRDA can be applied to high-dimensional data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ to maximize the correlation described in Equation (1) by the following steps:

sparse Redundancy Analysis (sRDA) Algorithm
Given explanatory data set $\boldsymbol{X}$ and response data set $\boldsymbol{Y}$

(i) Preliminary steps
- Center and scale $\boldsymbol{X}$ and $\boldsymbol{Y}$
- Set $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ to arbitrary vectors $[1, 1, \ldots, 1]'$ with length p and q, respectively
- Define convergence criterion $CRT = 1$ and a small positive tolerance $\gamma = 10^{-6}$

(ii) Iterative alternating regression
While $CRT \geq \gamma$

$\boldsymbol{\eta} = \boldsymbol{Y}\boldsymbol{\beta}^{(0)}$

$\boldsymbol{\alpha}^{(1)} = \underset{\boldsymbol{\alpha}^{(0)}}{\text{argmin}}\, \boldsymbol{\alpha}^{(0)\prime} \left(\frac{X'X+\lambda_2\mathbf{I}}{1+\lambda_2} \right) \boldsymbol{\alpha}^{(0)} - 2\boldsymbol{\eta}'\boldsymbol{X}\boldsymbol{\alpha}^{(0)} + \lambda_1|\boldsymbol{\alpha}^{(0)}|_1$

$\xi = \boldsymbol{X}\boldsymbol{\alpha}^{(1)}$

$\boldsymbol{\beta}^{(1)\prime} = [\xi'\xi]^{-1}[\xi'\boldsymbol{Y}]$

$\boldsymbol{\eta} = \boldsymbol{Y}\boldsymbol{\beta}^{(1)}$

$CRT = \sum(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}^{(0)})^2$

$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(1)}$

(iii) Upon convergence, return $\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}$

More detail about the sRDA can be found in Csala et al. (2017).

2.2 | Multiset sparse redundancy analysis

Now we describe multi-sRDA for the case when the true explanatory variables are distributed over multiple data sets. Suppose that the information transfer between the data sets is well-defined and it can be modeled with explanatory and response data set pairs. The objective function of multi-sRDA is to maximize the correlation given in Equation (1) for every pairs of explanatory and response data set, that is,

$$\sum_{i=1}^{k}\sum_{j=1}^{q} \text{Cor}(\boldsymbol{\xi}_i, \boldsymbol{y}_j)^2, \tag{4}$$

where $\boldsymbol{\xi}_i$ is the latent vector of the k-th explanatory data set and $\boldsymbol{y}_j$ is the j-th column of response data set $\boldsymbol{Y}$. More precisely, suppose there are three data sets $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{Y}$, assuming a directional information flow from $\boldsymbol{X}_1$ to $\boldsymbol{X}_2$ and information flows from both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ toward $\boldsymbol{Y}$, thus $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are explanatory for $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ is also an explanatory for $\boldsymbol{X}_2$ (see Figure 1). Our

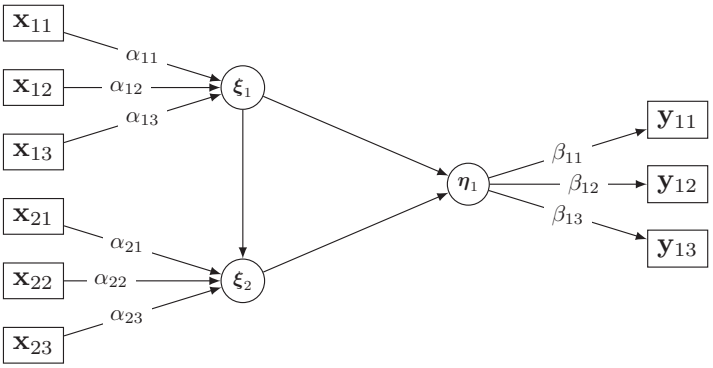

FIGURE 1 Multiset sRDA based on PLS path modeling framework, where $\mathbf{x}_{11}$, $\mathbf{x}_{12}$, and $\mathbf{x}_{13}$ denote the variables of matrix $\boldsymbol{X}_1$, and α_{11}, α_{12}, and α_{13} are elements in vector $\boldsymbol{\alpha}_1$, denoting the individual regression weights of the explanatory variables from matrix $\boldsymbol{X}_1$ on their latent variable $\boldsymbol{\xi}_1$ *Notes*: Latent variables $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are explanatory for response latent variable $\boldsymbol{\eta}_1$, and $\boldsymbol{\xi}_1$ is also an explanatory for $\boldsymbol{\xi}_2$

objective is to find the combination of variables in data sets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that explain the most variance in $\boldsymbol{Y}$ while accounting for the fact that $\boldsymbol{X}_1$ also explains variance in $\boldsymbol{X}_2$ (i.e., $\boldsymbol{X}_1$ has an explanatory-response relationship with $\boldsymbol{X}_2$). The resulting $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ weights for data sets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively, indicate the strength of the contribution of the explanatory variables. Due to the LASSO penalty, the sparse solution provides the set of variables that has the highest contribution among all variables. The $\boldsymbol{\beta}_1$ weights describe the strength of the correlation between the response variables from data set $\boldsymbol{Y}$ with the linear combination of their important explanatory variables. Our implementation of multi-sRDA is based on Wold's Partial Least Squares path modeling algorithm (PLS-PM) (Esposito Vinzi & Russolillo, 2013; Sanchez, 2013; Wold, 1975).

First, we present the algorithm for the case of three data sets (see Figure 1). Afterward, we generalize the algorithm to an arbitrary number of data sets.

Multiset sparse Redundancy Analysis algorithm for three data sets
Given data sets $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{Y}$

(i) *Preliminary steps*

- Center and scale $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{Y}$
- Set $\boldsymbol{\alpha}_1^{(0)}$, $\boldsymbol{\alpha}_2^{(0)}$, and $\boldsymbol{\beta}^{(0)}$ to arbitrary vectors $[1, 1, \ldots, 1]'$ with length p_1, p_2, and q, respectively
- Define convergence criterion $CRT_1 = 1$, $CRT_2 = 1$ and a small positive tolerance $\gamma = 10^{-6}$

(ii) *Iterative alternating regression*
While $CRT_1 \geq \gamma$ and $CRT_2 \geq \gamma$

Step (a) *Estimate initial latent variables*

$\boldsymbol{\xi}_1 \propto \boldsymbol{X}_1\boldsymbol{\alpha}_1^{(0)}$; where $\propto$ indicates that $\boldsymbol{\xi}_1$ is normalized to unit variance

$\boldsymbol{\xi}_2 \propto \boldsymbol{X}_2\boldsymbol{\alpha}_2^{(0)}$

$\boldsymbol{\eta} \propto \boldsymbol{Y}\boldsymbol{\beta}^{(0)}$

Step (b) *Model the relationship between data sets by calculating the correlation\regression weights between their latent variables (i.e., use regression weights for explanatory-response data set pairs and correlation weights otherwise, see the general case for the precise description)*

$$\boldsymbol{\Theta} = \begin{bmatrix} 0 & \theta_{12} & \theta_{13} \\ Cor(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) & 0 & \theta_{23} \\ Cor(\boldsymbol{\xi}_1, \boldsymbol{\eta}) & Cor(\boldsymbol{\xi}_2, \boldsymbol{\eta}) & 0 \end{bmatrix};$$

where θ_{13} are the weights from the regression model $[\boldsymbol{\xi}_1'\boldsymbol{\xi}_1]^{-1}\boldsymbol{\xi}_1'\boldsymbol{\xi}_2$, and θ_{13} and θ_{23} are the weights from the multiple regression model $[(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)'(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)]^{-1}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)'\boldsymbol{\eta}$.

Step (c) *Reestimate the latent variables*

$\boldsymbol{W} = [\boldsymbol{\xi}_1', \boldsymbol{\xi}_2', \boldsymbol{\eta}]\boldsymbol{\Theta}$

Step (d) *Estimate the new $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ weights and calculate the latent variables*

$$\boldsymbol{\alpha}_1^{(1)} = \underset{\boldsymbol{\alpha}_1}{\operatorname{argmin}}\, \boldsymbol{\alpha}_1' \left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1 + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\alpha}_1 - 2\boldsymbol{w}_1'\boldsymbol{X}_1\boldsymbol{\alpha}_1 + \lambda_1 |\boldsymbol{\alpha}_1|_1; \text{ where } \boldsymbol{w}_1 = \boldsymbol{\xi}_1\boldsymbol{\theta}_1$$

$\boldsymbol{\xi}_1 \propto \boldsymbol{X}_1\boldsymbol{\alpha}_1^{(1)}$

$$\boldsymbol{\alpha}_2^{(1)} = \underset{\boldsymbol{\alpha}_2}{\operatorname{argmin}}\, \boldsymbol{\alpha}_2' \left(\frac{\boldsymbol{X}_2'\boldsymbol{X}_2 + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\alpha}_2 - 2\boldsymbol{w}_2'\boldsymbol{X}_2\boldsymbol{\alpha}_2 + \lambda_1 |\boldsymbol{\alpha}_2|_1; \text{ where } \boldsymbol{w}_2 = \boldsymbol{\xi}_2\boldsymbol{\theta}_2$$

$$\boldsymbol{\xi}_2 \propto \boldsymbol{X}_2\boldsymbol{\alpha}_2^{(1)}$$

$$\boldsymbol{\beta}^{(1)'} = \left[\boldsymbol{w}_3'\boldsymbol{w}_3\right]^{-1}\left[\boldsymbol{w}_3'\boldsymbol{Y}\right]; \text{ where } \boldsymbol{w}_3 = \boldsymbol{\eta}\boldsymbol{\theta}_3$$

$$\boldsymbol{\eta} \propto \boldsymbol{Y}\boldsymbol{\beta}^{(1)}$$

Step (e) *Evaluate the convergence criteria and discard the old $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ weights*

$$CRT_1 = \sum \left(\boldsymbol{\alpha}_1^{(1)} - \boldsymbol{\alpha}_1^{(0)}\right)^2$$

$$CRT_2 = \sum \left(\boldsymbol{\alpha}_2^{(1)} - \boldsymbol{\alpha}_2^{(0)}\right)^2$$

$$\boldsymbol{\alpha}_1^{(0)} = \boldsymbol{\alpha}_1^{(1)},\ \boldsymbol{\alpha}_2^{(0)} = \boldsymbol{\alpha}_2^{(1)} \text{ and } \boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(1)}$$

(iii) *Upon convergence, return $\boldsymbol{\alpha}_1^{(0)}, \boldsymbol{\alpha}_2^{(0)}$, and $\boldsymbol{\beta}^{(0)}$*

The previous example can be generalized to K explanatory and a single response data set. In order to obtain the optimal $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}$ weights for latent variables $\boldsymbol{\xi}_k$ and $\boldsymbol{\eta}$ with arbitrary $\boldsymbol{\alpha}_k^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ weights are estimated (Step (ii)(a) in *multi-sRDA Algorithm*). $\boldsymbol{Z}$ is the matrix of all $K+1$ latent variables (i.e., $\boldsymbol{Z} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K, \boldsymbol{\eta}]$), and the link between the response and explanatory latent variables are modeled by the following multiple regression equations:

$$\boldsymbol{\zeta}_j = \theta_{j0} + \sum_{j'} \theta_{jj'}\boldsymbol{\zeta}_{j'\to j} + \boldsymbol{\epsilon}_j, \tag{5}$$

where $\boldsymbol{\zeta}_j$ denotes a response latent variable, $\boldsymbol{\zeta}_{j'\to j}$ denotes a latent variable that is explanatory for $\boldsymbol{\zeta}_j$, and $\boldsymbol{\epsilon}_j$ denotes the residual vector of $\boldsymbol{\zeta}_j$ (Esposito Vinzi & Russolillo, 2013; Sanchez, 2013). The association between the latent variables are expressed by $\boldsymbol{\Theta}$ (i.e., $\theta_{jj'}$ is the coefficient denoting the effect of $\boldsymbol{\zeta}_{j'\to j}$ on $\boldsymbol{\zeta}_j$)

$$E\left(\boldsymbol{\zeta}_j | \boldsymbol{Z}_{j'\to j}'\right) = \sum_{j'} \theta_{jj'}\boldsymbol{\zeta}_{j'\to j}. \tag{6}$$

Coefficient $\theta_{jj'}$ is obtained from the multiple multivariate regression of $\boldsymbol{\zeta}_j$ on its latent explanatory variables $\boldsymbol{Z}_{j'\to j}$, and the latent variables $\boldsymbol{\zeta}_j$ are updated with the effect of their explanatories (Step (ii)(b)).

Once matrix $\boldsymbol{\Theta}$ is obtained, the latent variables are reestimated and stored in matrix $\boldsymbol{W}$ (i.e., $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\Theta}$, Step (ii)(c)). Then latent explanatory variable $\boldsymbol{\xi}_k$ is recalculated with $\boldsymbol{\alpha}_k^{(1)}$ weights that are obtained from the penalized multivariate regression of $\boldsymbol{w}_k$ on data set $\boldsymbol{X}_k$. The latent response variable $\boldsymbol{\eta}$ is calculated with the updated $\boldsymbol{\beta}^{(1)}$ weights which are obtained by the univariate regression of $\boldsymbol{Y}$ on $\boldsymbol{w}_{(K+1)}$ (Step (ii)(d)).

This process is repeated until convergence and convergence is reached if the summed squared differences of $\boldsymbol{\alpha}_k^{(1)}$ and $\boldsymbol{\alpha}_k^{(0)}$ are smaller than γ, a predefined small positive tolerance (e.g., $\gamma = 10^{-6}$, Step (ii)(e)).

For an arbitrary number of explanatory data sets and one response data set, the correlation described in Equation (4) can be maximized by the following steps:

Multiset sparse Redundancy Analysis general Algorithm

Given $K+1$ data sets, from which K number are explanatory data sets $\boldsymbol{X}$, that is, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ (indexed by k), and a response data set $\boldsymbol{Y}$

(i) *Preliminary steps*

- Center and scale $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ and $\boldsymbol{Y}$
- Set $\boldsymbol{\alpha}_k^{(0)}$ (i.e., $\boldsymbol{\alpha}_1^{(0)}, \ldots, \boldsymbol{\alpha}_k^{(0)}, \ldots, \boldsymbol{\alpha}_K^{(0)}$) and $\boldsymbol{\beta}^{(0)}$ to arbitrary vectors $[1, 1, \ldots, 1]'$ with length p_k, and q, respectively
- Define convergence criterion $CRT = 1$ and a small positive tolerance $\gamma = 10^{-6}$

(ii) *Iterative alternating regression*

While $CRT \geq \gamma$

Step (a) *Estimate initial latent variables*

$$\boldsymbol{\xi}_k \propto \boldsymbol{X}_k\boldsymbol{\alpha}_k^{(0)}; \text{ where } k \text{ is the index from 1 to } K \text{ and } \propto \text{ indicates that } \boldsymbol{\xi}_k \text{ is normalized to unit variance}$$

$$\boldsymbol{\eta} \propto \boldsymbol{Y}\boldsymbol{\beta}^{(0)}$$

Define matrix $\boldsymbol{Z} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K, \boldsymbol{\eta}]$ and define column j of $\boldsymbol{Z}$ as $\boldsymbol{\zeta}_j$

Step (b) *Model the relationship between data sets by calculating the correlation\regression weights between their latent variables*

Define $\boldsymbol{\Theta} \in \mathbb{R}^{j \times j}$ as a correlation matrix where $\theta_{j,j'}$ is:

If $j' < j$ **then**

$$\theta_{j,j'} = \begin{cases} \theta_{jj'} & \text{if } \boldsymbol{\zeta}_{j'} \text{ explains } \boldsymbol{\zeta}_j \\ cor(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_{j'}) & \text{else} \end{cases}$$

else

Define $\boldsymbol{Z}_{j' \to j}$ as the submatrix of $\boldsymbol{Z}$ formed by all columns $\boldsymbol{\zeta}_{j'}$ that are explanatory for $\boldsymbol{\zeta}_j$

$$\theta_{j,j'} = \begin{cases} [\boldsymbol{Z}'_{j' \to j} \boldsymbol{Z}_{j' \to j}]^{-1} \boldsymbol{Z}'_{j' \to j} \boldsymbol{\zeta}_j & \text{if } \boldsymbol{\zeta}_{j'} \text{ explains } \boldsymbol{\zeta}_j \\ cor(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_{j'}) & \text{else} \end{cases}$$

end if

Step (c) *Reestimate the latent variables*

$\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\Theta}$

Step (d) *Estimate the new* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *weights and calculate the latent variables*

$\boldsymbol{\alpha}_k^{(1)} = \underset{\boldsymbol{\alpha}_k}{\operatorname{argmin}} \, \boldsymbol{\alpha}_k' \left(\frac{\boldsymbol{X}_k' \boldsymbol{X}_k + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\alpha}_k - 2\boldsymbol{w}_k' \boldsymbol{X}_k \boldsymbol{\alpha}_k + \lambda_1 |\boldsymbol{\alpha}_k|_1$

$\boldsymbol{\xi}_k \propto \boldsymbol{X}_k \boldsymbol{\alpha}_k^{(1)}$

$\boldsymbol{\beta}^{(1)'} = \left[\boldsymbol{w}'_{(K+1)} \boldsymbol{w}_{(K+1)} \right]^{-1} \left[\boldsymbol{w}'_{(K+1)} \boldsymbol{Y} \right]$

$\boldsymbol{\eta} \propto \boldsymbol{Y} \boldsymbol{\beta}^{(1)}$

Step (e) *Evaluate the convergence criteria and discard the old* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *weights*

$CRT = \sum \left(\boldsymbol{\alpha}_k^{(1)} - \boldsymbol{\alpha}_k^{(0)} \right)^2$

$\boldsymbol{\alpha}_k^{(0)} = \boldsymbol{\alpha}_k^{(1)}$ and $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(1)}$

(iii) Upon convergence, return $\left(\boldsymbol{\alpha}_1^{(0)}, \ldots, \boldsymbol{\alpha}_k^{(0)}, \ldots, \boldsymbol{\alpha}_K^{(0)} \right), \boldsymbol{\beta}^{(0)}$

Note that the only assumption of this model is that there is a linear relationship between the explanatory and response variables that can be modeled through latent variables (Equations 5 and 6). Also, PLS-PM does not require any assumptions regarding the sample size or the distribution of the explanatory and response variables, therefore it is considered to be an exploratory approach rather than a confirmatory one (Vinzi, Trinchera, & Amato, 2010).

In the next section, we describe how to determine the best penalization parameters for the penalized multivariate equation step (i.e., selecting λ_1 and λ_2 when computing $\boldsymbol{\alpha}_k^{(1)}$).

2.3 | Selecting the best parameter for the penalization variables

As the main goal of the analysis is to maximize the correlation given in Equation (1) for every explanatory and response pairs of data sets, we define the best penalization parameters as the values for λ_1 and λ_2 that leads to the maximization of the correlation described in Equation (4). To determine the best penalization parameters for the penalized multivariate equation (i.e., selecting λ_1 and λ_2 when computing $\boldsymbol{\alpha}_k^{(1)}$), we define a search space of values for both variables and assessing all possible combinations of these through a standard 10-fold cross validation (CV) procedure. This procedure is repeated for every explanatory data set and the selected best penalization parameters are the ones that maximize the sum of squared correlations between response variables in data set $\boldsymbol{Y}$ with their latent explanatories $\boldsymbol{W}_{\to \eta}$.

Finding the best penalization parameters becomes quickly computationally expensive with extending the search space for λ_1 and λ_2. For example, extending the search space from two possible values for both λ_1 and λ_2 to two values for λ_1 and three values

for λ_2 introduces 80 extra multi-sRDA runs in a 4 data set setting, since the total number of runs are defined by: search space $(\lambda_1) \times$ search space $(\lambda_2) \times k \times 10$. This high computational burden can be mitigated by introducing Univariate Soft Thresholding (UST) penalization, which can be seen as a special version of ENet, where $\lambda_2 \to \infty$. With UST, the covariance matrix from the multivariate equation in Equation (3) is ignored, which implies that with UST we ignore the correlation between the variables within the explanatory data set $\boldsymbol{X}$.

2.4 | Assessing the statistical significance of the proposed model and quantifying multi-sRDA's ability of including the relevant variables in the final model

For assessing the statistical significance of the proposed model, we conduct a permutation study as follows. The null distribution of the sum of squared correlations between the latent explanatory variables and the response variables are approximated by removing the correlation between explanatory and response data sets by permuting the rows of the matrices, thus the observations are shuffled keeping the internal correlation structure intact while removing the correlation between the involved matrices. We fit the model on the permuted data sets to obtain the $\boldsymbol{\alpha}$ weights. The estimated $\boldsymbol{\alpha}$ weights are used to determine the sum of squared correlations, and the null distribution of the sum of squared correlations is obtained by repeating the permutation test many times.

In addition, a resampling method is used to approximate the confidence interval of the optimum estimate of sum of absolute correlations. We use bootstrapping by taking a sample of observations from the original data, with replacement. We do this 100 times and fit a model on each bootstrap sample and report the mean and selected quantiles of the resulting distribution.

In order to assess multi-*RDA*'s ability to select the relevant (i.e., truly associated) variables from each data set, we calculate the true-positive rate (TPR) and true-negative rate (TNR) during the simulation studies. That is, for TPR we study the proportion of the number of truly associated variables identified by the algorithm (i.e., those that the model assigned with nonzero $\boldsymbol{\alpha}$ weights) with the total number of truly associated variables that were simulated during the data generation. For TNR, we assess the proportion of the number of truly nonassociated variables excluded from the final model (i.e., those truly nonassociated variables that the model assigned with nonzero regression weights) to the number of truly nonassociated variables.

It is important to note that the TPR and TNR measures are hampered by the restricted computational resources. That is, finding the best penalization parameters is increasingly computationally expensive as the search space for the possible parameters extends and it is possible that the optimal value is not included in the search space at all. For example, the model might not include all the associated variables due to a restricted search space for λ_1; during the CV study one might restrict the search space for five possible values for λ_1 and these all might result in a model with fewer nonzero weights than the total number of the truly associated variables, that is some of the truly associated variables will not be identified. In order to account for this bias, we provide a measure which is less restrained by the search spaces of the penalization parameters, and is less affected when λ_1 is smaller than the number of the real associated variables. We call this the truly associated variable inclusion (TAVI) measure. Thus TAVI gives a better indication about how many times some of the truly associated variables are included in the model. TAVI is then defined as the proportion of truly associated variables identified by the algorithm, to either the total number of truly associated variables or the value of the selected λ_1 parameter, whichever is smaller, that is

$$TAVI = \frac{\sum_{i=1}^{p^{relevant}} I\left(\boldsymbol{\alpha}_{p_{irrelevant}+i} \neq 0\right)}{\min\left(p^{relevant}, \lambda_1\right)}.$$

3 | SIMULATION STUDIES

We assessed multi-sRDA's ability of selecting the relevant variables from the explanatory data sets and the response data sets. That is, we quantified the methods ability of assigning nonzero regression weights to the variables from the explanatory data sets that explain the most variation in the response data set and indicate which response variables has the highest variance. To do this, we designed four simulation studies. The data were created in a similar fashion for all simulation studies, therefore we first describe the general approach of data generation and then we describe the particular parameters for the different studies.

3.1 | Data generation

In general, we created $K = 3$ explanatory data sets, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_3$ and one response data set, $\boldsymbol{Y}$. For all explanatory data sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_3$, we generated $p^{irrelevant}$ irrelevant variables that were not associated with their latent variables. The irrelevant

variables were sampled from the standard normal distribution with mean 0 and a standard deviation of 1. We generated $p^{relevant}$ relevant variables that were associated with their latent variables, and the strength of the association was indicated with the regression weights $\boldsymbol{\alpha}^{relevant}$. The strength of the associations between data sets were modeled by θ regression weights, that is, $\boldsymbol{\xi}_3 = \theta_2\boldsymbol{\xi}_2 + \theta_1\boldsymbol{\xi}_1$.

More precisely:

Generate $K+1$ latent variables:

(i) $\boldsymbol{\Xi} \in \mathbb{R}^{n\times(K+1)}$

(ii) $\boldsymbol{\xi}_1 \in \mathbb{R}^n$ distributed $\mathcal{N}(0,1)$

(iii) For $i = 2, \ldots, K+1$:

- **(a)** $m = 0$
- **(b)** $s2 = 0$
- **(c)** For $c = i-1, \ldots, 1$:
  - i. $m = m + \theta_c\boldsymbol{\xi}_c$
  - ii. $s2 = s2 + \theta_c^2$
- **(d)** $\boldsymbol{\xi}_\mathbf{i} \sim \mathcal{N}(m, \sqrt{1-s2})$

Generate the k-th data sets (where k is the running index from 1 to $K+1$):

(iv) $\mathbf{X}_k \in \mathbb{R}^{n\times\left(p^{irrelevant}+p^{relevant}\right)}$

(v) For $i = 1, \ldots, p^{irrelevant}$:

$\mathbf{x}_i \in \mathbb{R}^n$ distributed $\mathcal{N}(0,1)$

(vi) For $c = \left(p^{irrelevant}+1\right), \ldots, \left(p^{irrelevant}+p^{relevant}\right)$:

$\boldsymbol{x}_i \in \mathbb{R}^n$ distributed $\mathcal{N}\left(\boldsymbol{\alpha}^{relevant}\boldsymbol{\xi}_k, \sqrt{1-(\boldsymbol{\alpha}^{relevant})^2}\right)$

Steps (iv)–(vi) are repeated $K+1$ times to obtain $K+1$ data sets and the last data set obtained was treated as the response data set (i.e., $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ were explanatory data sets and $\boldsymbol{Y} = \boldsymbol{X}_{K+1}$ was the response data set).

3.2 | Data simulation

First, we designed three different simulation studies with different sample sizes; small ($n = 100$), medium ($n = 250$), and large ($n = 500$) sample size. Otherwise, all three simulation studies shared the same parameters, namely, the number of data sets was set to 4 ($K = 4$), the number of irrelevant variables per data set were $p^{irrelevant}_{1,\ldots,4}$ = (1,000, 500, 200, 10), the number of relevant variables were $p^{relevant}_{1,\ldots,4}$ = (10, 8, 8, 2), the strength of the association between the latent variables were set to θ = (0.8, 0.7, 0.4), and the regression weights for the relevant variables in the four different data sets were $\boldsymbol{\alpha}^{relevant}_{1,\ldots,4}$ = ([0.5, 0.5, 0.5, 0.3, 0.3, 0.2, 0.2, 0.2, 0.2, 0.2], [0.5, 0.5, 0.5, 0.4, 0.4, 0.3, 0.2, 0.2], [0.5, 0.5, 0.5, 0.3, 0.3, 0.2, 0.2, 0.2], [0.5, 0.1]).

In order to assess *multi-RDA*'s ability to select the associated variables from each data set, we calculated the TPR, the TNR, and the TAVI rate over 1,000 simulations. The TPR measures are reported for all three explanatory data sets (i.e., TPR_{X_1}, TPR_{X_2}, and TPR_{X_3}). The results are presented in Table 1. TPR values ranged from 0.72 to 0.91 and it showed a tendency to increase by increasing sample size. The measures for TAVI are reported in Table 1 for all three data sets (i.e., $TAVI_{X_1}$, $TAVI_{X_2}$, and $TAVI_{X_3}$). The values for TAVI ranged from 0.87 to 1.

The TNR measures are reported for the three different explanatory data sets (i.e., TNR_{X_1}, TNR_{X_2}, and TNR_{X_3}) and the results are reported in Table 1. The TNR values ranged from 0.985 to 0.998 and were not much affected by the varying sample size. We observed and reported earlier that ENet was slightly superior to UST in precision of the relevant explanatory variables selection in such high-dimensional setting, but ENet imposed huge computational costs compared to UST (Csala et al., 2017). Therefore, given the size of our data, we used UST penalization for all simulation studies.

In addition, we designed a simulation study that resembles the size of the omics data used for the real data analysis in Section 4. This study had the following parameters; a sample size of 37 ($n = 37$), the number of data sets was 3 ($K = 3$), the number of irrelevant variables per data set were $p^{irrelevant}_{1,\ldots,3}$ = (360,000; 18,000; 47), the number of relevant variables were $p^{relevant}_{1,\ldots,3}$ = (100; 80; 8), the strength of the association between the latent variables were set to θ = (0.8; 0.4), and the regression weights for the relevant variables $\boldsymbol{\alpha}^{relevant}_{1,\ldots,3}$ in the three data sets varied between 0.2 and 0.5. The TPR, the TNR, and the TAVI rate were

TABLE 1 Results of simulation study

	$n = 100$	$n = 250$	$n = 500$
TPR_{X_1}	0.72	0.76	0.75
TPR_{X_2}	0.76	0.88	0.90
TPR_{X_3}	0.79	0.81	0.79
$TAVI_{X_1}$	0.89	0.99	0.99
$TAVI_{X_2}$	0.87	0.99	0.99
$TAVI_{X_3}$	0.92	0.99	1.00
TNR_{X_1}	0.99	0.99	0.99
TNR_{X_2}	0.99	0.99	0.99
TNR_{X_3}	0.99	0.99	0.99

Notes: True positive rates (TPR), the rate of truly associated variable inclusion (TAVI), and true negative rates (TNR) computed from 1000 replicates.

calculated over 100 simulations. TPR resulted in 0.66 for the first explanatory data set and in 0.79 for the second explanatory data set. TAVI resulted in 0.68 for the first explanatory data set and in 0.81 for the second explanatory data set and we observed the value 0.99 for TNR for both explanatory data sets.

4 | HIGH-DIMENSIONAL OMICS DATA ANALYSIS

We applied multi-sRDA to high-dimensional omics data in order to explain variability in proteome variables explained by biomarkers from the epigenome and transcriptome. Our data sets included 364,134 methylation markers measured in blood leukocytes by the Illumina Infinium HumanMethylation450 BeadChip, 18,424 gene expression markers obtained from skin biopsy with Affymetrix Human Exon 1.0ST Arrays, and 47 cytokine markers measured in blood plasma (Radonic et al., 2012). The data included in the analysis were measured on 37 patients with Marfan syndrome who participated in the Dutch Compare Trial (Groenink et al., 2013). The conceptual model we built regarded the methylation data as explanatory for both the gene expression measurements and the cytokine markers, and the gene expression measurements were regarded as explanatory for the cytokine markers. Given the size of the data, we used UST penalization for our model, with which one multi-sRDA run took 364 s. In order to find the best λ_1 parameters for UST, we used 10-fold CV, which resulted in selecting 150 nonzero explanatory variables from the methylation markers, with the sum of absolute correlations criterion of 6,842.32, and selecting 15 nonzero explanatory variables from the gene expression markers, with the sum of absolute correlation criterion of 8.61 (see Figure 2). We show the sum of absolute correlations instead of the sum of squared correlations for ease of interpretation.

In order to assess the statistical significance of our findings, we conducted a permutation study with the real data. We performed the permutation study 100 times and observed that the sum of absolute correlations obtained with the best penalization values were significantly different from the null distribution of the sum of absolute correlation for the methylation markers (with p-value of 9.447×10^{-11}), but not for the gene expression markers (with p-value of 0.085) (see Figure 3). We used bootstrapping to estimate the confidence interval of the sum of absolute correlation. We obtained a 95% CI [3872.59, 9147.52] for the sum of absolute correlation for the methylation markers and a 95% CI [8.65, 16.34] for the sum of absolute correlation for the gene expression markers.

The names and weights of the 150 methylation markers and the 15 genes that were included in our final model can be found in Table 3 and 2 in Section 5. Our method also provided the genes and cytokines with the highest correlation coefficients in the response data set, thus those gene expression values whose variability was most affected by the combination of the 150 selected methylation markers and those cytokine markers whose variability was most affected by the combination of the 15 gene expression variables and the 150 selected methylation markers. These variables can be found in Table 4 with their corresponding $\boldsymbol{\beta}$ weights, where the $\boldsymbol{\beta}$ weights are the individual correlation coefficients with the combination of their explanatory variables.

The final model is represented in Figure 4. The methylation markers are positioned in the upper right corner in the figure, with their corresponding latent variable $\boldsymbol{\xi}_1$, and only 15 markers with the highest α weights are represented out of the 150. The gene expression markers are plotted in the bottom with their latent variable $\boldsymbol{\xi}_2$ and the top 40 markers that have the highest $\boldsymbol{\beta}$ weights are plotted. The cytokine markers are represented upper left on the plot with their latent variable $\boldsymbol{\eta}$, and the top 15 cytokine markers that have the highest $\boldsymbol{\beta}$ weights are represented. The observed correlation was 0.807 between $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, 0.763 between $\boldsymbol{\xi}_1$ and $\boldsymbol{\eta}$, and 0.611 $\boldsymbol{\xi}_2$ and $\boldsymbol{\eta}$.

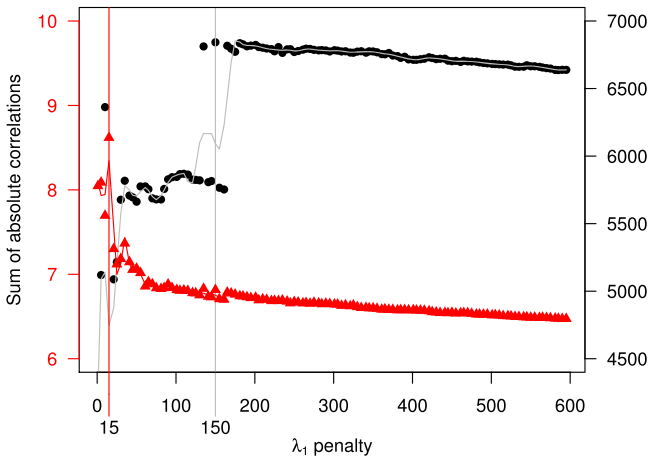

FIGURE 2 Representation of the 10-fold CV results for finding the best penalization parameters for the number of nonzero $\boldsymbol{\alpha}$ weights (i.e., the λ_1 penalty) *Notes*: On the *x*-axis, the number of nonzeros selected are given and on the *y*-axis the values for the sum of absolute correlations are represented. On the left in red, these values are associated with the gene expression markers and on the right in black with the methylation markers. A total of 150 methylation markers were selected for the final model that maximize the sum of squared correlations with the cytokine markers (weights $\boldsymbol{\alpha}$ are given in Table 4), resulted in the sum of absolute correlations criterion of 6,842.32. The results from the CVs are represented with black dots, with scale on the right side of the figure. Similarly, 15 gene expression markers are selected (with $\boldsymbol{\alpha}$ weights given in Table 3) that maximizes the sum of squared correlation with the cytokine markers, with the sum of absolute correlation criterion of 8.61, and the CV results are represented with red triangles, with scale of the left side of the figure

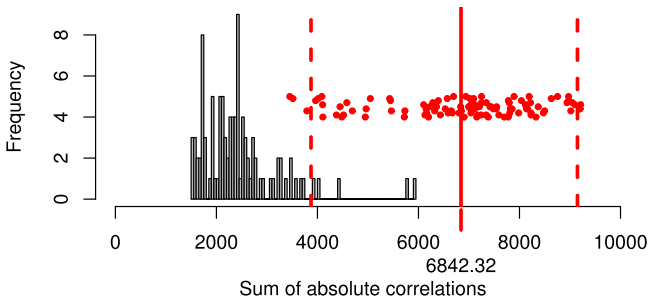

FIGURE 3 The null distribution of the sum of absolute correlations of the methylation markers (in gray) plotted with the sum of absolute correlation (6842.32) observed with the best optimal penalization parameters (p-value 9.447×10^{-11}) *Notes:* The red dashed lines represent the 95% confidence interval of the distribution obtained through bootstrapping and the red dots represent the sum of absolute correlation values obtained for the 100 bootstrap samples

We used an online overrepresentation analysis tool (available at `https://reactome.org`) to test whether the variables included in the final model can be associated with any known biological pathways. Several pathways were identified, including "*Cytokine Signaling in Immune system*" pathway, with genes involved TCEB1, PPP2CB, HIST1H3A, KPNA2, BIRC2 (with p-value 4.32×10^{-2}), "*Cellular responses to stress*" pathway, with genes involved HIGD1A, TCEB1, AQP8, HSPH1, HIST1H3A (with p-value 4.41×10^{-2}), and the "*Regulation of HSF1-mediated heat shock response*" pathway with the gene involved HSPH1 (with p-value 4.96×10^{-3}).

5 | DISCUSSION

In the present paper, we have shown how sRDA can be extended into a multiset multivariate method (multi-sRDA) so that the effect of multiple explanatory variables, from multiple data sets, on multiple outcome variables, measured on one data set, can be assessed simultaneously, while modeling the sequential information transfer between the involved data sets. We showed through simulation studies and through genomewide high-dimensional omics data analysis that our proposed multi-sRDA model is able to deal with data sets containing hundreds of thousands of variables and is able to indicate those explanatory genotypic variables that explain the most variation in the response phenotypic variables, with high precision. To quantify the precision of

TABLE 2 Gene expression variables that were selected for the final model with their corresponding $\boldsymbol{\alpha}$ weights

Gene Name	α Weight
AASDH	0.0746
AZIN1	0.0720
C12orf5	0.0742
C16orf61	0.0740
C20orf15	0.0781
HSPH1	0.0734
KIAA1841	0.0736
KPNA2	0.0782
KPNA2_A	0.0745
MINPP1	0.0743
NDUFAF4	0.0772
PTGR1	0.0763
SGMS2	0.0727
SNX16	0.0796
TAC4	0.0751

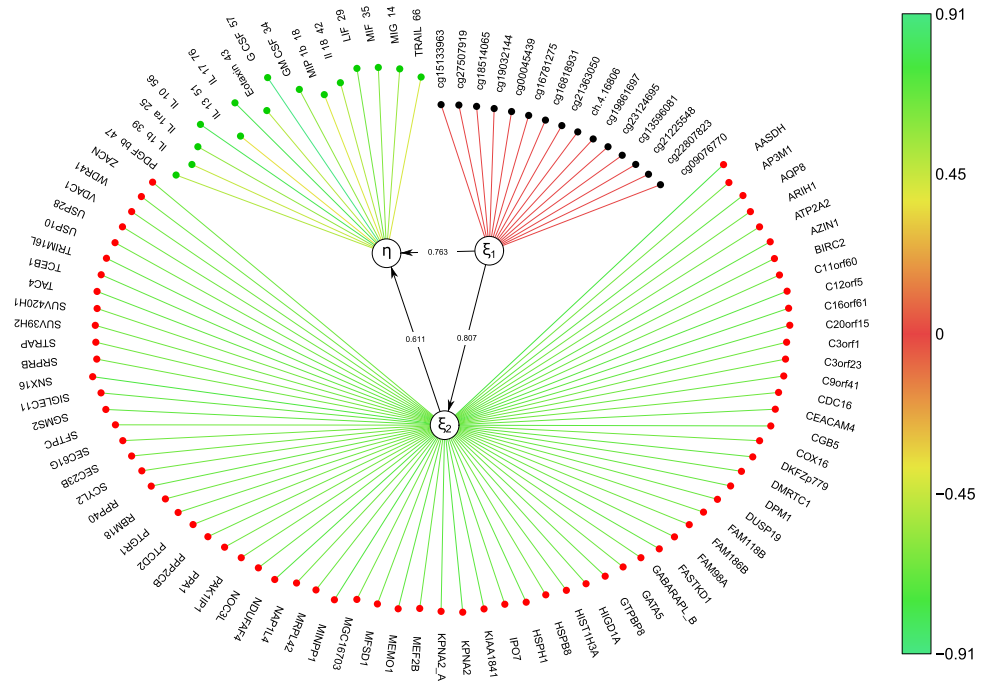

FIGURE 4 Plot of the model obtained by applying multi-sRDA to the Marfan data set *Notes:* Upper right are the $\boldsymbol{\alpha}$ weights of the methylation markers, from which only 15 represented on the plot with their corresponding latent variable $\boldsymbol{\xi}_1$. Bottom are the $\boldsymbol{\beta}$ weights of the gene expression markers, from which the highest 40 correlated are plotted and $\boldsymbol{\xi}_2$ represents the gene expression markers' latent variable. Upper left are the $\boldsymbol{\beta}$ weights of the cytokine markers, from which the highest 15 correlated are plotted and η represents the latent variable of the cytokine markers. The correlation between $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ is 0.807, between $\boldsymbol{\xi}_1$ and $\boldsymbol{\eta}$ is 0.763, and between $\boldsymbol{\xi}_2$ and $\boldsymbol{\eta}$ is 0.611

multi-sRDA, we ran simulation studies and used two true-positive measures and a negative rate measure. Our simulation studies indicate that multi-sRDA is able to find important explanatory variables with high precision (TPR and TAVI range from 0.66 to 1, depending on sample size and the number of variables) and that the unimportant variables are almost always excluded from the final model (TNR ranges from 0.985 to 0.998). As we described, TAVI is less restrained by the size of the search spaces of the penalization parameters, and is less affected in the case when only a subset of the real associated variables are included in the model, thus when the number of nonzero $\boldsymbol{\alpha}$ weights is smaller than the number of the real associated variables, which

TABLE 3 Methylation sites that were selected for the final model with their corresponding α weights

Methylation Site	α Weights
cg21535222	0.0078
cg20314620	0.0077
cg27559870	0.0078
cg00571339	0.0079
cg00719323	0.0081
cg15133963	0.0084
cg17459215	0.0079
cg09451427	0.0077
cg10550308	0.0077
cg15999104	0.0077
cg26309929	0.0082
cg27507919	0.0084
cg01101647	0.0078
cg03415545	0.0078
cg07784166	0.0078
cg23008238	0.0079
cg19176453	0.0082
cg01287342	0.0077
cg12217600	0.0082
cg18862888	0.0077
cg07209141	0.0077
cg16993220	0.0078
cg25188298	0.0079
cg01230784	0.0077
cg00716025	0.0077
cg18514065	0.0082
cg13191049	0.0078
cg12391328	0.0078
cg21487894	0.0079
cg26217813	0.0080
cg03751734	0.0081
cg20110347	0.0079
cg02021210	0.0078
cg02394421	0.0080
cg03541057	0.0078
cg04993276	0.0078
cg05309877	0.0081
cg05650632	0.0079
cg07091551	0.0079
cg15568225	0.0078
cg16514995	0.0078
cg26352401	0.0081
cg09994323	0.0080
cg14932133	0.0078
cg17673205	0.0077

(Continues)

TABLE 3 (Continued)

Methylation Site	α Weights
cg19032144	0.0083
cg20362287	0.0077
cg20419724	0.0078
cg25058482	0.0081
cg25325588	0.0077
ch.2.24120	0.0077
ch.2.15408	0.0078
ch.2.33765	0.0077
cg05997021	0.0078
cg06606548	0.0078
cg11271830	0.0077
cg12904680	0.0078
cg15146004	0.0078
cg23817981	0.0077
cg25562664	0.0077
ch.3.57315	0.0078
cg00045439	0.0086
cg00669856	0.0081
cg02945007	0.0080
cg05493561	0.0078
cg11818376	0.0077
cg16781275	0.0082
cg16818931	0.0083
cg21363050	0.0085
cg22904577	0.0079
cg27096981	0.0078
ch.4.16806	0.0084
ch.4.32075	0.0077
ch.4.33836	0.0079
cg06017028	0.0079
cg16537383	0.0079
cg17231906	0.0078
cg24194998	0.0081
cg25564121	0.0080
cg26891849	0.0077
ch.5.91394	0.0078
cg00949008	0.0080
cg01802635	0.0080
cg04673565	0.0080
cg06442073	0.0079
cg10684686	0.0078
cg12882103	0.0078
cg17774634	0.0078
cg18715868	0.0079
cg19004138	0.0079
cg03413884	0.0077

(Continues)

TABLE 3 (Continued)

Methylation Site	α Weights
cg05896120	0.0077
cg13791269	0.0078
cg15394787	0.0077
cg19861697	0.0084
cg20752420	0.0082
cg20956366	0.0078
cg23124695	0.0084
cg27552912	0.0079
ch.7.16370	0.0077
cg00940812	0.0081
cg13562276	0.0078
cg16717480	0.0079
cg17813074	0.0078
cg24704908	0.0078
cg26669044	0.0078
cg13596081	0.0084
cg21225548	0.0083
cg22807823	0.0084
cg10715637	0.0078
cg23656443	0.0082
cg24903527	0.0082
cg01991180	0.0081
cg03569616	0.0081
cg09076770	0.0082
cg11031737	0.0077
cg14706107	0.0082
cg02076642	0.0082
cg05666055	0.0078
cg14263244	0.0079
cg24883413	0.0077
cg01837275	0.0081
cg02632314	0.0081
cg04570827	0.0077
cg06549607	0.0078
cg08542640	0.0079
cg14210817	0.0080
cg22417789	0.0077
cg27163659	0.0077
cg05100432	0.0080
cg10208821	0.0077
cg13015872	0.0078
cg20279895	0.0079
cg23114881	0.0080
cg24478145	0.0081
cg20560091	0.0080
cg27625481	0.0079

(Continues)

TABLE 3 (Continued)

Methylation Site	α Weights
cg01289020	0.0081
cg02215945	0.0081
cg05160228	0.0079
cg07691874	0.0079
cg19803976	0.0080
cg05297461	0.0080
cg07688933	0.0079
cg18244544	0.0078
cg18708075	0.0078
cg03933490	0.0078
cg13888886	0.0081
cg18166947	0.0077
ch.22.9096	0.0077

we observed several times during the simulation studies. Measuring TAVI showed that in fact multi-sRDA found a subset of the important explanatory variables almost every time for all the involved data sets in the medium and large sample size setting (range from 0.99 to 1).

Note that the simulation studies are limited due to the fact that the generated model is quite simple since the latent variable is not a combination of the explanatory variables, but instead the explanatory variables are directly sampled from a one-dimensional distribution which might not represent well the structure of the omics data used for the real data analysis. Also, there is a somewhat unexpected behavior observed in the optimization curve of the λ_1 parameters for the methylation markers during the real data analysis (see Figure 2). Our best explanation is that around the value 140 for the λ_1 penalty, the model starts to generate a linear combination of $\boldsymbol{\alpha}$ weights that correlate much stronger with the response variables that leads to the substantial increase in the sum of absolute correlation. Nevertheless, the overrepresentation analysis of the resulting multi-sRDA model provided reasonable pathways, that is, "*Cytokine Signaling in Immune system*" pathway given one of the data sets were the cytokine markers and "*Cellular responses to stress*" makes sense too from a biomedical perspective since the characteristic of Marfan syndrome is the hampered strength and elasticity of the connective tissue.

We also applied regularized canonical correlation analysis (rCCA) (Waaijenborg & Zwinderman, 2009) to the same omics data sets that we used in our real data analysis. Canonical correlation analysis is a multivariate statistical method similar to RDA, but instead of maximizing the correlation between the explanatory latent variable and the response variables, it searches for latent variable pairs that have the maximum covariance with each other. Well-known drawbacks are that CCA's results are not readily interpretable (i.e., the $\boldsymbol{\beta}$ weights cannot be interpreted as correlation coefficients like in RDA) and that CCA cannot model the presumed sequential information transfer between data sets. Nevertheless, for comparison, we applied rCCA on the three omics data sets we described in Section 4. By treating all data set as response, thus running the analysis only with multivariate regression steps to calculate the weight vectors, multi-sRDA's framework was used for rCCA. We observed the value of 6118.28 for the sum of absolute correlation between the outcome cytokine variables and the latent variable of the methylation markers and the value of 7.47 for the sum of absolute correlation between the outcome cytokine variables and the latent variable of the methylation markers. These values are substantially lower than we found with our sRDA method, which makes sense given CCA's different objective function. Thus if we can assume directionality between data sets, it is better to use our new method. Recently, minimal Bayes Information Criterion (BIC) was proposed for selecting the best optimal λ_1 for sparse canonical correlation analysis (Wilms & Croux, 2015) which is an attractive alternative to replace CV in order to mitigate computational burden, although CV outperformed BIC in the average number of iterations needed for algorithmic convergence in what the authors call "Ultra high-dimensional" scenario, with $n = 50$ and $p = 1,000$.

Although multi-sRDA is based on the PLS framework, it should not be confused with penalized multiblock PLS models (Kawaguchi & Yamashita, 2017). Penalized multiblock PLS, like CCA, does not model the sequential information flow between data sets. When there is a well-defined information transfer assumed between data sets, multi-sRDA has the advantage of accounting for this information flow and extracting the biologically relevant variables from the data sets, while penalized multiblock PLS extract the variables that has the highest correlation with each other, disregarding the explanatory-response relationships between data sets (Karaman et al., 2015).

TABLE 4 The 40 highest β weights from the resulting model for the gene expression and cytokine markers

Gene Expression Values		Cytokine Markers	
Gene Name	**β Weight**	**Cytokine Marker**	**β Weight**
AASDH	0.761	Eotaxin 43	0.508
AQP8	0.699	FGF basic 44	0.462
AZIN1	0.703	G CSF 57	0.314
BIRC2	0.675	GM CSF 34	0.585
C11orf60	0.673	GRO a 61	0.266
C12orf5	0.717	HGF 62	0.184
C16orf61	0.704	IFN a2 20	0.249
C20orf15	0.702	IFN g 21	0.097
C9orf41	0.678	IL 10 56	0.211
CEACAM4	0.686	Il 12 p40 28	0.277
CGB5	0.710	IL 12 p70 75	0.798
DKFZp779	0.694	IL 13 51	0.333
DMRTC1	0.678	IL 15 73	0.367
DPM1	0.684	Il 16 27	0.312
DUSP19	0.672	IL 17 76	0.793
FAM98A	0.674	Il 18 42	0.495
FASTKD1	0.692	Il 1a 63	0.273
GTPBP8	0.679	IL 1b 39	0.907
HIGD1A	0.678	IL 1ra 25	0.589
HIST1H3A	0.678	IL 2 38	0.345
HSPH1	0.702	Il 2Ra 13	0.208
IPO7	0.679	Il 3 64	0.407
KIAA1841	0.709	IL 4 52	0.146
KPNA2	0.734	IL 5 33	0.196
KPNA2_A	0.714	IL 6 19	0.145
MEF2B	0.680	IL 7 74	0.094
MEMO1	0.707	IL 9 77	0.167
MFSD1	0.674	IP 10 48	0.286
MINPP1	0.737	LIF 29	0.509
MRPL42	0.672	M CSF 67	0.310
NDUFAF4	0.734	MCP 3 26	0.174
NOC3L	0.681	MIF 35	0.319
PTGR1	0.672	MIG 14	0.657
RPP40	0.688	MIP 1a 55	0.364
SGMS2	0.698	MIP 1b 18	0.366
SIGLEC11	0.693	PDGF bb 47	0.618
SNX16	0.745	SCF 65	0.559
TAC4	0.693	TNF b 30	0.244
TCEB1	0.691	TRAIL 66	0.418
TRIM16L	0.691	VEGF 45	0.316

The proposed multi-sRDA model is easily extendable for multidimensional latent variable extraction. After the first latent variables $\boldsymbol{\xi}_k$ are obtained for data sets $\boldsymbol{X}_k$, $\boldsymbol{\xi}_k$ explained effects are subtracted from original data sets $\boldsymbol{X}_k$ in order to obtain residual data sets $\boldsymbol{X}_k^{res}$, on which the analysis can be repeated to obtain the following latent variables. Since the multidimensional latent variables are orthogonal to each other, each of them explains a different portion of variance of their response data set. By definition, the first latent variable has the highest absolute sum of squared correlation with its response data set, and all the following variables explain a smaller or equal portion of variance of the response data set. The number of latent variable to be extracted from a data set can be determined using the permutation test we proposed in Section 2; one might stop extracting further latent variables when the permutation test indicates nonsignificant results.

However, obtaining multiple latent variables introduces further computational burdens, which, as we already mentioned before, practically restricted the present model to the *UTS* penalization scheme. When the data sets consist hundreds of thousands variables, estimating the optimal penalization parameters are costly due to the CV procedure and extracting l number of latent variables would increase the computational costs l-fold. Currently, we are investigating how to implement multidimensional latent variable extraction so that it is also applicable in real omics data settings.

In this paper, we extended our sRDA approach to multiple explanatory sets and we demonstrated multi-sRDA's applicability for high-dimensional omics data through conducting simulation studies and analyzing real biomedical data from three different omics domains. We conclude that multi-sRDA can be used to analyze multiple high-dimensional omics sets in order to explore the combination of those foremost biomolecular markers from various biological levels that explain the most variance of the phenotypic variables, while modeling the conceptual model of the central dogma of molecular biology.

ACKNOWLEDGMENT

The authors would like to thank two anonymous referees and the associate editor for their comments and insights that significantly improved this paper. The first author also would like to thank for the support and help of Christopher, Grace, Jesper, Jim, Leon, Martha, Nastya, Stephanie, and William.

CONFLICTS OF INTEREST

The authors have declared no conflict of interest.

ORCID

Attila Csala https://orcid.org/0000-0003-4969-5555

REFERENCES

Buescher, J. M., & Driggers, E. M. (2016). Integration of omics: More than the sum of its parts. *Cancer & Metabolism*, *4*(1), 4.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563.

Csala, A., Voorbraak, F. P. J. M., Zwinderman, A. H., & Hof, M. H. (2017). Sparse redundancy analysis of high-dimensional genetic and genomic data. *Bioinformatics*, *33*(20), 3228–3234.

Esposito Vinzi, V., & Russolillo, G. (2013). Partial least squares algorithms and methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(1), 1–19.

Fornell, C., Barclay, D. W., & Rhee, B.-D. (1988). A model and simple iterative algorithm for redundancy analysis. *Multivariate Behavioral Research*, *23*(3), 349–360.

Groenink, M., Den Hartog, A. W., Franken, R., Radonic, T., De Waard, V., Timmermans, J., … Mulder, B. J. M. (2013). Losartan reduces aortic dilatation rate in adults with Marfan syndrome: A randomized controlled trial. *European Heart Journal*, *34*(45), 3491–3500.

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, *8*, 84.

Israels, A. Z. (1984). Redundancy analysis for qualitative variables. *Psychometrika*, *49*(3), 331–346.

Johansson, J. K. (1981). An extension of Wollenberg's redundancy analysis. *Psychometrika*, *46*(1), 93–103.

Karaman, Ä., Nørskov, N. P., Yde, C. C., Hedemann, M. S., Bach Knudsen, K. E., & Kohler, A. (2015). Sparse multi-block PLSR for biomarker discovery when integrating data from LC–MS and NMR metabolomics. *Metabolomics*, *11*(2), 367–379.

Kawaguchi, A., & Yamashita, F. (2017). Supervised multiblock sparse multivariable analysis with application to multimodal brain imaging genetics. *Biostatistics*, *18*(4), 651–665.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. (2007). The vegan package. *Community Ecology Package*, *10*, 631–637.

Radonic, T., de Witte, P., Groenink, M., de Waard, V., Lutter, R., van Eijk, M., … Zwinderman, A. H. (2012). Inflammation aggravates disease severity in Marfan syndrome patients. *PLoS ONE*, *7*(3), 1–9.

Sanchez, G. (2013). *PLS path modeling with R*. Berkeley: Trowchez Editions.

van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, *42*(2), 207–219.

Vinzi, V. E., Trinchera, L., & Amato, S. (2010). PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement. In *Handbook of partial least squares* (pp. 47–82). Berlin: Springer.

Waaijenborg, S., & Zwinderman, A. H. (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics*, *10*(1), 315.

Wilms, I., & Croux, C. (2015). Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, *57*(5), 834–851.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In *Quantitative Sociology* (pp. 307–357). New York: Elsevier.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Csala A, Hof MH, Zwinderman AH. Multiset sparse redundancy analysis for high-dimensional omics data. *Biometrical Journal*. 2019;61:406–423. https://doi.org/10.1002/bimj.201700248