


**RESEARCH ARTICLE**

# Principal balances of compositional data for regression and classification using partial least squares

V. Nesrstová<sup>1,2</sup>  | I. Wilms<sup>3</sup> | J. Palarea-Albaladejo<sup>4</sup>  | P. Filzmoser<sup>5</sup> |  
J. A. Martín-Fernández<sup>4</sup> | D. Friedecký<sup>6,7</sup> | K. Hron<sup>1</sup>

<sup>1</sup>Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czech Republic

<sup>2</sup>Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic

<sup>3</sup>Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

<sup>4</sup>Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

<sup>5</sup>Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

<sup>6</sup>Laboratory for Inherited Metabolic Disorders, Department of Clinical Biochemistry, University Hospital Olomouc, Olomouc, Czech Republic

<sup>7</sup>Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czech Republic

**Correspondence**

V. Nesrstová, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, Olomouc, Czech Republic.  
Email: [viktorie.nesrstova@gmail.cz](mailto:viktorie.nesrstova@gmail.cz)

**Funding information**

Austrian Science Foundation, Grant/Award Number: I 5799-N; Czech Science Foundation, Grant/Award Numbers: 22-15684L, 19-07155S; Internal Grant Agency of the Palacký University Olomouc, Grant/Award Numbers: IGA\_PrF\_2022\_008, IGA\_PrF\_2023\_009; Spanish Ministry of Science and Innovation and ERDF, Grant/Award Number: PID2021-123833OB-I00; Dutch Research Council (NWO)

**Abstract**

High-dimensional compositional data are commonplace in the modern omics sciences, among others. Analysis of compositional data requires the proper choice of a log-ratio coordinate representation, since their relative nature is not compatible with the direct use of standard statistical methods. Principal balances, a particular class of orthonormal log-ratio coordinates, are well suited to this context as they are constructed so that the first few coordinates capture most of the compositional variability of data set. Focusing on regression and classification problems in high dimensions, we propose a novel partial least squares (PLS) procedure to construct principal balances that maximize the explained variability of the response variable and notably ease interpretability when compared to the ordinary PLS formulation. The proposed PLS principal balance approach can be understood as a generalized version of common log-contrast models since, instead of just one, multiple orthonormal log-contrasts are estimated simultaneously. We demonstrate the performance of the proposed method using both simulated and empirical data sets.

**KEYWORDS**

balance coordinates, compositional data, high-dimensional data, metabolomic data, PLS regression and classification

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of Chemometrics* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

Compositional data (CoDa) occur in plenty of research fields, such as geochemistry,<sup>1</sup> metabolomics,<sup>2</sup> microbiome data,<sup>3</sup> time use data,<sup>4</sup> or ecology.<sup>5</sup> Let us consider a regression (or classification) task where  $\mathbf{y} = (y_1, \dots, y_n)^\top$  is a vector of  $n$  observations of a continuous (or binary) response variable and  $\mathbf{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq D}$  is an associated matrix of a  $D$ -part compositional predictor. Analyzing CoDa requires careful consideration since such data do not carry relevant information in their absolute values, but rather in the ratios between the parts that constitute the composition. Accordingly, CoDa are scale invariant, which means that the multiplication of a compositional vector by any positive constant keeps the ratios unaltered. Moreover, in this work, we address the case of high-dimensional compositions consisting of a large number  $D$  of parts.

In the growing literature regarding omics sciences, most data are actually of relative (hence compositional) nature,<sup>6</sup> and the development of methods for high-dimensional compositions is of increasing interest. The challenge concerns not only compositions consisting of many parts but also the fact that the number of samples  $n$  is usually substantially smaller than the number  $D$  of parts due to the nature of the technology and omics sciences in general (as, for example, hundreds of proteins or metabolites are examined). Such settings with more variables than observations hamper the use of the most popular regression and classification methods, including least squares (LS) regression and linear or quadratic discriminant analysis (LDA/QDA) models, where some pre-selection or dimension reduction of variables needs to be performed in advance.

For regression analysis with high-dimensional compositions, log-contrast models<sup>7</sup> have gained increasing popularity; see, for instance, Bates and Tibshirani,<sup>8</sup> Susin et al,<sup>9</sup> and Gordon-Rodriguez.<sup>10</sup> Log-contrasts are a building block in CoDa analysis through the log-ratio methodology.<sup>11</sup> Given a  $D$ -part composition, a log-contrast is a loglinear combination

$$\sum_{j=1}^D a_j \ln x_j, \text{ with } \sum_{j=1}^D a_j = 0, a_j \in \mathbb{R}.$$

Any log-ratio coordinate representation of CoDa consists of  $D - 1$  log-contrasts, with such number corresponding to the actual dimensionality of compositions. A typical feature of log-contrast models in Aitchison and Bacon-Shone<sup>7</sup> or Rivera-Pinto et al<sup>12</sup> is that only one of the possible  $D - 1$  log-contrasts corresponding to the dimensionality of  $D$ -part composition is estimated as predictor variable. Note that it is crucial to have compositional predictors expressed in the form of log-contrasts, because only then scale invariance is fulfilled. Representing the compositional predictors by just one log-contrast is in principle appealing, particularly when aiming to assess their relevance in explaining the response variable (with the possible caveat that the zero-sum constraint of regression coefficients must be met being considered). However, using only one log-contrast might be unnecessarily restrictive as other log-contrasts can be of interest as predictor variables, and this is something we investigate here. Moreover, if (and only if) such log-contrasts are orthogonal, then the usual interpretation of regression coefficients (in terms of assessing their association with the response variable) applies.<sup>13</sup>

However, estimating the set of all possible  $D - 1$  predictor log-contrasts could be computationally exhaustive or even unfeasible with high-dimensional compositions, definitely so when using existing log-contrast models. But, in fact, this would not be needed. Just having a few log-contrasts capturing most of the information contained in the original explanatory composition, while they appropriately relate to the response variable, would be necessary. We here introduce a model that sufficiently explains a response variable  $\mathbf{y}$  using the matrix  $\mathbf{X}$  of compositional predictors while performing dimension reduction through partial least squares (PLS) regression/classification.<sup>14</sup> Prior to PLS modeling, it is necessary to express CoDa in a proper log-ratio coordinate system. For example, a common choice is to use clr coefficients<sup>15</sup> and subsequently perform PLS analysis on them, although alternative coordinate representations can be considered.<sup>2,16</sup> Note that clr coefficients have a direct link to log-contrasts.<sup>17</sup>

Even though PLS is a convenient method to model relationships between response and explanatory variables, when dealing with compositions, the interpretation of log-contrasts might become challenging in high dimensions, and thus, simplification would be welcome. To this end, we propose to use a special class of log-ratio coordinates, so-called

balance coordinates,<sup>18</sup> which are interpretable in terms of contrasts between subgroups of compositional parts summarized by their geometric means. Note also that balances aggregate all pairwise log-ratios between parts with positive and negative sign in the respective log-contrast.<sup>19</sup> Furthermore, in order to achieve orthonormality between balance coordinates, we tailor the original principal balances (PB) approach of Pawlowsky-Glahn et al<sup>20</sup> and Martín-Fernández et al.<sup>17</sup> The original PB method was developed to enhance interpretability in dimension reduction of CoDa in the style of principal component analysis (we will denote this PCA-PB). We adapt it here to involve a response variable within a regression or classification problem by replacing PCA by PLS. Accordingly, up to  $D - 1$  PBs with decreasing explanatory power are obtained that can be either directly interpreted or used for further statistical analysis. In the following, we will refer to this new proposal as *PLS-PB*.

This manuscript is structured as follows. In Section 2, the basics of CoDa are presented together with the description of the proposed PLS-PB procedure. In Section 3, a simulation study is conducted to investigate the ability of PLS-PB to reflect and simplify the structure of PLS loadings as well as its prediction performance. The proposed method is then applied to two empirical data sets in Section 4. Section 5 concludes with some final remarks and future outlook.

## 2 | COMPOSITIONAL DATA AND PLS PRINCIPAL BALANCES

When analyzing CoDa, their relative nature needs to be appropriately accounted for. The sample space of CoDa is formed by equivalence classes of proportional vectors.<sup>21</sup> In practice, CoDa are typically (equivalently) represented in the form of percentages, proportions, or parts per million (ppm), that is, as data with a constant sum constraint and living on a simplex. Given the scale invariance property of compositions<sup>11</sup> and the geometric structure of their sample space (the so-called Aitchison geometry), a well-principled way to conduct analysis of CoDa is to express them in the form of log-ratios of parts, or more generically as their log-contrasts, and then proceed to further statistical processing on these.<sup>22,23</sup>

Log-contrast models have been recently used for regression or classification analysis with high-dimensional compositional covariates.<sup>3,8,10</sup> The general form of these models is given by

$$\mathbf{y} = \ln(\mathbf{X}) \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where  $\mathbf{y} = (y_1, \dots, y_n)^\top$  is a univariate response variable,  $\mathbf{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq D}$  is the matrix of observed compositional predictors (with  $\ln(\mathbf{X})$  denoting the elementwise logarithm of  $\mathbf{X}$ ),  $\boldsymbol{\beta}$  represents the regression coefficients with  $\sum_{j=1}^D \beta_j = 0$ , and  $\boldsymbol{\varepsilon}$  stands for the ordinary random error term. It is common practice to interpret the regression coefficients directly in terms of the original parts like in a standard multiple regression model. Nevertheless, from a compositional perspective, just one log-contrast is estimated and the common interpretation of regression coefficients can only be used with due caution. For this reason, considering more orthogonal (or orthonormal) log-contrasts is advantageous in our view.

The model in Equation (1) can, however, be immediately generalized to a setting where more orthonormal log-contrasts are estimated simultaneously. In order to reduce the dimension, it would be interesting to rank them according to decreasing relevance to explain or predict the response variable. Moreover, simplifying the interpretation of these log-contrasts would be beneficial. This is achieved by the novel PLS-PB method introduced in this work as showed in the following.

### 2.1 | Log-ratio representations of compositional data

As said, CoDa are commonly expressed in the form of log-ratios of parts (log-contrasts) for statistical analysis. Thus, clr coefficients<sup>11</sup> are often used. For a  $D$ -part composition  $\mathbf{x} = (x_1, \dots, x_D)^\top$ , its representation in the clr coefficients is defined as

$$\text{clr}(\mathbf{x}) = (\text{clr}_1(\mathbf{x}), \text{clr}_2(\mathbf{x}), \dots, \text{clr}_D(\mathbf{x})) = \left( \ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right), \quad (2)$$

where  $g(\mathbf{x})$  denotes the geometric mean of the parts of the composition  $\mathbf{x}$ . Clr coefficients and log-contrasts are closely linked together as it holds that<sup>17</sup>

$$\text{clr}_j(\mathbf{x}) = \ln \frac{x_j}{g(\mathbf{x})} = \sum_{k=1}^D a_k \ln x_k, \quad \sum_{k=1}^D a_k = 0, \quad \text{for } j = 1, \dots, D, \quad (3)$$

where  $a_k = -\frac{1}{D}$  for  $k \neq j$  and  $a_j = 1 - \frac{1}{D}$ . The clr coefficients impose a zero-sum constraint  $\sum_{k=1}^D \text{clr}_k(\mathbf{x}) = 0$  and thus lead to a singular covariance matrix, which is undesirable for some statistical methods including LS regression or LDA/QDA. However, their construction and interpretation are convenient in other cases, including PLS regression. Note that, by taking Equation (3) into account, the regression model in Equation (1) could have been developed directly in clr coefficients (this will be done separately later). The reason is methodological: while in Equation (1) the zero-sum constraint must be imposed,<sup>3</sup> by using clr coefficients in the first instance, it is automatically incorporated before proceeding with standard estimation. In fact, clr coefficients always add up to zero by definition, given the geometric mean placed in the denominator, and any LS-based estimator of a linear model (including PLS regression) preserves this constraint. This relates to the possibility of expressing the explanatory part of a linear regression model as the inner product of the vector of coefficients and the vector of predictor variables (we refer the interested reader to van den Boogaart et al<sup>24</sup> for further details). However, care must be taken when interpreting results in terms of this log-ratio coordinate representation as the clr coefficients should not be simply identified with the original compositional parts, because instead they aggregate all pairwise logratios with specific parts.<sup>23</sup>

An alternative way to map CoDa from their original sample space into the real space is through orthonormal log-ratio (olr) coordinates, also known as isometric log-ratio (ilr) coordinates, which are derived from the Euclidean vector space structure of the Aitchison geometry and overcome the singularity issue, hence being more generally applicable in statistical analysis.<sup>23,25,26</sup>

Note that balances are one concrete instance of olr coordinates that are constructed by means of a sequential binary partition (SBP) of the parts of a composition.<sup>18</sup> This procedure sequentially splits parts into two non-overlapping groups. Thus, at the  $k$ th partition,  $k = 1, \dots, D - 1$ , the balance  $b_k$  between two subgroups is given by

$$b_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{\sqrt[r_k]{x_{n_1} \cdot \dots \cdot x_{n_{r_k}}}}{\sqrt[s_k]{x_{d_1} \cdot \dots \cdot x_{d_{s_k}}}},$$

where  $r_k$  denotes the number of parts in the first group (numerator of the log-ratio) and  $s_k$  denotes the number of parts in the second group (in the denominator of the log-ratio), with  $n_1, \dots, n_{r_k}$  and  $d_1, \dots, d_{s_k}$  being the indices of the parts of the first and second group, respectively. From a  $D$ -part composition, the number of balances derived in the SBP is  $D - 1$ , which corresponds to the actual dimensionality of the composition. The interpretation of balances is straightforward: they represent the relative dominance of the average of one group of parts with respect to the average of the other group. They are orthonormal log-contrasts by construction, which means that the respective vectors of log-contrast coefficients are re-scaled to have unit norm and are mutually orthogonal:

$$b_k = \sum_{j=1}^D a_{kj} \ln x_j, \quad \text{with } a_{kj} = \begin{cases} \sqrt{\frac{s_j}{(r_j + s_j)r_j}} & \text{if } j \in \{n_1, \dots, n_{r_k}\} \\ -\sqrt{\frac{r_j}{(r_j + s_j)s_j}} & \text{if } j \in \{d_1, \dots, d_{s_k}\} \\ 0 & \text{otherwise.} \end{cases}$$

Recently, the use of balances has been questioned by some.<sup>1,27,28</sup> Nonetheless, their ability to represent the original information in terms of contrasts or comparisons between two groups of parts remains appealing when compared to using general log-contrasts, particularly so in high-dimensional settings.

However, as the number  $D$  of parts of a composition increases, it becomes more challenging to build a SBP and obtain a collection of balances that are interpretable. It is then desirable to construct just a few balances which capture the majority of the information. In an unsupervised learning settings, principal balances have been proposed for this aim.<sup>20</sup> In Martín-Fernández et al.,<sup>17</sup> PBs are formally defined as follows.

**Definition 2.1.** Given a composition  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$ , principal balances are log-contrasts  $\sum_{j=1}^D a_{kj} \cdot \ln x_j$ ,  $k = 1, \dots, D-1$ , such that  $\mathbf{a}_k = (a_{k1}, \dots, a_{kD})^\top$  are constant vectors which maximize the variances  $\text{var}\left[\sum_{j=1}^D a_{kj} \cdot \ln x_j\right]$  and:

- for  $k = 1, \dots, D-1$  the coefficients  $a_{kj}$  take one of the three values  $(-c_1, 0, c_2)$ ,  $c_1$  and  $c_2$  being some strictly positive numbers,
- for  $k = 1, \dots, D-1$ , it holds that  $\sum_{j=1}^D a_{kj} = 0$  and  $\sum_{j=1}^D a_{kj}^2 = 1$ ,
- for  $k = 2, 3, \dots, D-1$ ,  $\mathbf{a}_k$  is orthogonal to the previous  $\mathbf{a}_{k-1}, \mathbf{a}_{k-2}, \dots, \mathbf{a}_1$ , that is  $\sum_{j=1}^D a_{kj} \cdot a_{(k-l)j} = 0$ ,  $l = 1, 2, \dots, k-1$ .

PBs facilitate dimension reduction using PCA in the clr space, but the interpretation of the resulting principal components is simplified through their expression in terms of balances. Accordingly, the first PB maximizes the sample variance, and each subsequent balance then maximizes the remaining variance in the data, while satisfying the orthonormality constraint. Up to  $D-1$  PBs can be derived, although in practice much fewer are typically needed to capture the main modes of variability. In the following subsection, we embed such dimension reduction by PB coordinates into a regression setting using a PLS formulation.

## 2.2 | PLS regression and classification

PLS is a multivariate method that is used to model a linear relationship between a response variable and a set of (not necessarily compositional) explanatory variables. The linear relationship is however not modeled directly, but via the construction of latent variables (PLS components). Values of new latent variables are called scores, and coefficients that determine the influence of each variable on the score are called loadings.<sup>29</sup>

More precisely, PLS regression aims to estimate the regression coefficient vector  $\mathbf{b} = (b_1, \dots, b_D)$  in the linear regression model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (4)$$

where  $\mathbf{e}$  stands for a random error term. Both the response variable and covariates are centered prior to the analysis, so no intercept is included in Equation (4). As we deal with CoDa, the matrix  $\mathbf{X}$  in Equation (4) is expressed in the form of clr coefficients, and this clr matrix is denoted  $\text{clr}(\mathbf{X})$  in the following, such that the model in Equation (4) can be written as

$$\mathbf{y} = \text{clr}(\mathbf{X})\mathbf{b} + \mathbf{e}, \quad (5)$$

While the estimation of  $\mathbf{b}$  in terms of the original variables (here clr coefficients as used for compositional PCA) is the final goal, the regression fit itself and prediction are performed on the PLS components, which are linear combinations of clr variables in the  $n \times D$  matrix  $\text{clr}(\mathbf{X})$ . Because of Equation (3), these latent PLS components are just the log-contrasts we are searching for.

Namely, the matrix  $\text{clr}(\mathbf{X})$  is decomposed as

$$\text{clr}(\mathbf{X}) = \mathbf{TP}^\top + \mathbf{E}_X,$$

where  $\mathbf{T}$  is a score matrix,  $\mathbf{P}$  is a loading matrix, and  $\mathbf{E}_X$  is an error matrix.<sup>29</sup> Both matrices  $\mathbf{T}, \mathbf{P}$  have  $k$  columns,  $k \leq \min(D, n)$ , indicating the number of PLS components. The goal of PLS is then to maximize the covariance between the scores (coordinates corresponding to the latent variables) and  $\mathbf{y}$ , under the constraint of uncorrelated scores (the most usual case) or orthogonal loading vectors (representing weights given to the original variables in the construction of the PLS components). Let  $\mathbf{p}$  denote a (column) loading vector of matrix  $\mathbf{P}$ . Then it holds for a score vector  $\mathbf{t}$  that  $\mathbf{t} = \text{clr}(\mathbf{X})\mathbf{p}$ , and the maximization problem can then be written as

$$\max_{\mathbf{p}} \text{cov}(\text{clr}(\mathbf{X})\mathbf{p}, \mathbf{y}), \text{ subject to } \|\text{clr}(\mathbf{X})\mathbf{p}\| = 1, \quad (6)$$

where  $\mathbf{p}$  is considered to be a weighting vector. The constraint of unit length ensures that the maximization problem is unique.<sup>29</sup> Such maximization problem results in the first score vector  $\mathbf{t}$ . The subsequent score vectors are obtained in the same way, with the condition that they must be orthogonal to the previous ones. The score vector  $\mathbf{t} = \text{clr}(\mathbf{X})\mathbf{p}$  contains  $n$  observations of a log-contrast because the sum of elements of each loading vector  $\mathbf{p}$  equals to zero in the clr coefficient representation of the composition acting as covariate. The resulting log-contrasts can then be used in the regression model

$$\mathbf{y} = (\mathbf{TP}^\top)\mathbf{b} + \mathbf{e}_T = \mathbf{T}\mathbf{v} + \mathbf{e}_T, \quad (7)$$

with  $\mathbf{v} = \mathbf{P}^\top \mathbf{b}$ , for prediction purposes as well as to determine the number of log-contrasts that are sufficient for a good prediction of  $\mathbf{y}$ . There exist several algorithms to find a solution to this optimization problem. We resort to the well-known SIMPLS algorithm.

Similarly, PLS can be used for classification purposes. This method is then commonly called PLS discriminant analysis (PLS-DA). Here, we will focus on binary response variables, typically using codes 0 for observations that do not belong to a certain group and 1 for those that do belong.

Beyond representing a generalization of log-contrast models, a step further with the proposed formulation is to simplify it in the form of balances and then rely on PBs instead of PLS loadings. Thus, we can investigate which groups of parts contribute in positive or negative sense to their values based on the position of parts either in the numerator (positive) or denominator (negative) of the coordinate and their corresponding log-contrast coefficients. In doing so, a small price is paid in terms of prediction ability of the resulting regression model, but a benefit in interpretability of log-contrasts as latent variables is generally obtained, as will be further discussed in Section 3.

## 2.3 | Algorithmic implementation of PLS principal balances

PLS-PB can be straightforwardly constructed by adapting the PCA-PB approach from.<sup>17</sup> Specifically, we take the constrained PCs algorithm introduced in that work as reference. This algorithm builds PBs based on loadings obtained from PCA. Our proposal is to do it based on the loadings from PLS. Calculations were conducted on the R system for statistical computing,<sup>30</sup> using the package `pls`<sup>31</sup> for PLS estimation and the package `compositions`<sup>32</sup> for CoDa operations and manipulations.

The core of the modification is as follows: Instead of maximizing explained variance in accordance with PCA, we maximize the covariance between the response variable and the new established balance (through the respective balance coefficients) while keeping orthonormality of the new coordinate system. Thus, balances and balance coefficients play the role of score vectors and loadings, respectively. In the relationship  $\mathbf{t} = \text{clr}(\mathbf{X})\mathbf{p}$ , balance coefficients are in the place of vector  $\mathbf{p}$ . Accordingly, the first PLS loading vector is used to derive the first PB, and the other balances are then obtained by maximizing the absolute value of the covariance of subsequently derived balances with the response variable. Algorithm 1 summarizes this procedure for obtaining the PLS-PB.

**Algorithm 1** CONSTRUCTION OF PLS-PB**Initiation:** center response variable  $\mathbf{y}$ , compute clr coefficients of composition in  $\mathbf{X}$  and center them**PB:**1. First PB:  $\mathbf{pb}_1$  (based on first PLS loading vector  $\mathbf{p}_1$ ):i. PLS regression of centered  $\mathbf{y}$  on centered  $\text{clr}(\mathbf{X})$ ii. First loading vector:  $\mathbf{p}_1$ iii. Signs of values in  $\mathbf{p}_1$ :  $\mathbf{s}_{\text{sign}} = \text{sign}(\mathbf{p}_1)$ iv. Using  $\mathbf{p}_1$ , derive  $D - 1$  candidate sign of balances  $\mathbf{s}_1, \dots, \mathbf{s}_{D-1}$  with codes  $\{-1, 0, +1\}$ :

- for  $i = 1, \dots, D$ :

$$s_{i1} = \begin{cases} +1 & \text{if } p_i = \max_{\{1 \leq i \leq D\}} \mathbf{p}_1 \\ -1 & \text{if } p_i = \max_{\{1 \leq i \leq D\}} (-\mathbf{p}_1) \\ 0 & \text{otherwise.} \end{cases}$$

- $\mathbf{s}_j, j = 2, \dots, D-1$ : copy codes  $\{+1, -1\}$  in  $\mathbf{s}_{j-1}$ , add  $+1$  or  $-1$  using  $\mathbf{s}_{\text{sign}}$  and the remaining components of  $\mathbf{p}_1$  (excluding the values chosen in the previous step). For  $i = 1, \dots, D$ :

$$s_{ij} = \begin{cases} s_{i(j-1)} & \text{if } s_{i(j-1)} \neq 0 \\ \text{sign}(\mathbf{p}_{1i}) & \text{if } |p_{1i}| = \max_{\{k: s_{k(j-1)}=0\}} \{|p_{1k}|\} \\ 0 & \text{otherwise.} \end{cases}$$

- result: matrix of signs  $\mathbf{S}_{D \times (D-1)}$ ; sign of balances in columns

v. Create  $\mathbf{B}_{D \times (D-1)} = (\mathbf{b}_1, \dots, \mathbf{b}_{(D-1)})$  matrix of balance coefficients: for  $i^{\text{th}}$  row and  $j^{\text{th}}$  column,  $i = 1, \dots, D, j = 1, \dots, D - 1$ 

$$b_{ij} = \begin{cases} \sqrt{\frac{s_j}{(r_j + s_j)r_j}} & \text{if } s_{ij} = 1, \\ -\sqrt{\frac{r_j}{(r_j + s_j)s_j}} & \text{if } s_{ij} = -1, \\ 0 & \text{otherwise.} \end{cases}$$

where  $r_j$  is number of  $+1$  values in column  $j$  and  $s_j$  is number of  $-1$  values in column  $j$ vi. First PB:  $\mathbf{pb}_1 = \max_{\{1 \leq j \leq (D-1)\}} |\text{cov}(\ln(\mathbf{X})\mathbf{b}_j, \mathbf{y})|$ 2. Derive the other balances based on  $\mathbf{pb}_1$ :i. If  $\mathbf{pb}_1$  contains 0 value(s): repeat the procedure for  $\mathbf{pb}_1$  (steps 1.[i-vi.]) using only variables assigned with 0 in  $\mathbf{pb}_1$ 

ii. Further partition:

- Down the numerator: repeat steps 1.[i-vi.] using variables in the numerator of  $\mathbf{pb}_1$
- Down the denominator: repeat steps 1.[i-vi.] using variables in the denominator of  $\mathbf{pb}_1$

**Final step:** Sort PB:  $\mathbf{pb}_1, \mathbf{pb}_2, \dots, \mathbf{pb}_{(D-1)}$  such that

$$|\text{cov}(\mathbf{pb}_{(1)}, \mathbf{y})| > |\text{cov}(\mathbf{pb}_{(2)}, \mathbf{y})| > \dots > |\text{cov}(\mathbf{pb}_{(D-1)}, \mathbf{y})|$$

The codes with functions together with simulations are available in GitHub (<https://github.com/NestrstovaV/PLS-PBs.git>). Empirical data sets cannot be publicly available due to ownership rights, but are available upon reasonable request.

### 3 | NUMERICAL ASSESSMENT

In this section, we first introduce three illustrative examples to visually demonstrate that PLS-PBs help to arrive at a simplified structure of PLS loadings (Section 3.1). In line with an application to metabolomics in Section 4, here we refer to (bio)markers (explanatory variables), that is, biological measurements or signals most associated to some health or biological outcome/status of interest (response variable).<sup>2</sup> We therefore focus on the identification of meaningful biomarkers as the main purpose of the data analysis. Subsequently, using the same data generating processes as in Section 3.1, we set up a more thorough simulation study to formally compare the predictive performance of the proposed PLS-PB with the original PCA-PB in Martín-Fernández et al.<sup>17</sup> (Section 3.2).

#### 3.1 | Artificial settings for comparison: PLS-PB against PLS loadings

To assess the behavior of PLS-PB across various settings, we consider three artificial cases inspired by the study in Štefelová et al.<sup>2</sup> The first case contains just one block of markers among a given collection of signals defining the explanatory composition. In the other two examples, we extend such setting to include several groups of markers. In all cases, we consider  $n = 250$  samples and  $D = 100$  signals in the composition. The number of PLS-PB is then 99 ( $D - 1$ ). Following on Štefelová et al.,<sup>2</sup> compositions were simulated using so-called pivot coordinates, an instance of olr coordinates<sup>33</sup> (see Appendix A for more details). The resulting covariance matrices are visualized in Figure 1.

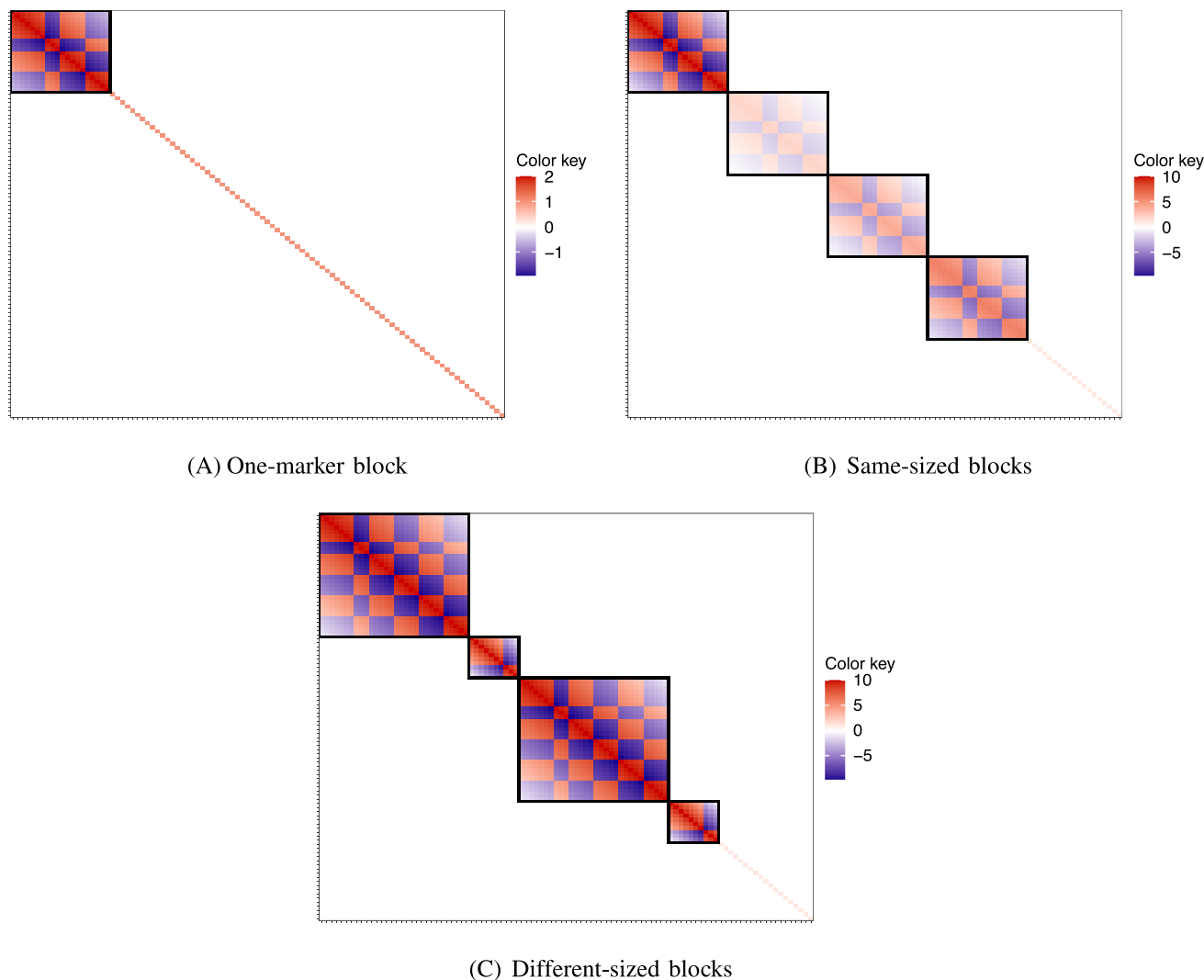
The three cases aim to reflect realistic metabolomic scenarios. Metabolomic data in general, by its very nature, contain a number of correlations, primarily because of the very close first-order relationship through the metabolic pathways,<sup>34–36</sup> where individual metabolites are converted. In some selected cases, correlations can be very strong ( $>0.75$ , e.g., for fatty acids<sup>37</sup>). This phenomenon is particularly applicable in a younger branch of metabolomics, called lipidomics, where individual lipids within their lipid classes show systematic correlations and we can get into a situation where even most lipids in the data matrix can be correlated.<sup>38,39</sup> This all is reflected in the regression and correlation structure provided in the simulation study, which aims to provide a reasonable theoretical model covering scenarios, which are general enough. Specifically, we consider (i) a first case where all markers are in one block with groups of markers that line up sequentially, with the covariances within each group being generated so that more distant markers are less correlated, and (ii) cases 2 and 3, where multiple (four) blocks of correlated markers are allowed.

Figure 1A displays the covariance matrix in the first case. The generated compositions then contain one block of 20 markers, and the remaining 80 signals are regarded irrelevant (i.e., as the covariance matrix is  $(D - 1) \times (D - 1)$ , the remaining 79 rows/columns correspond to irrelevant signals). Within the block of markers, there are four groups consisting respectively of 7, 3, 5, and 5 markers. The markers line up sequentially, and they are, in turn, positively or negatively associated within each group. All these 20 markers relate with the response; that is, the corresponding model coefficients are not zero, and its sign is determined by the covariance matrix (the remaining 79 coefficients are set to zero).

Figure 1B displays the covariance matrix for the case including four blocks of markers (each block consisting of 20 markers). There are thus in total 80 markers and 20 irrelevant signals (i.e., the last 19 rows/columns in Figure 1B), which can be considered as noise. The elements in each marker block are generated following a decreasing sequence such that the elements further away from the diagonal have smaller values. This approach ensures that the scattering of pivot coordinates in each block differs. The effect of the first block of pivot coordinates is the strongest, and the second block should produce the “weakest” markers (in the sense of pivot coordinates; covariances between variables in the second block are the smallest). Because of the way pivot coordinates are constructed, it can be said that the relevance of markers (in terms of the original parts) decreases in each subsequent block. Finally, all markers in blocks 1–3 are related to the response variable, but those in the fourth block are not and, hence, should not be considered as genuine markers (along with the irrelevant signals).

Finally, in the third case, four blocks of markers are again considered, but here they have different sizes as shown in Figure 1C. The first and third blocks consist of 30 markers, while the second and fourth consist of 10 markers each. As the blocks differ in size, the structure of larger blocks consists of respectively 7, 5, 5, 3, 5, and 5 groups of markers, with the sign of the covariances alternating across groups. The entries in the covariance matrix are in the same range in each block, with the diagonal elements being identical. Similarly to the second case, the last 19 rows/columns in Figure 1C correspond to irrelevant signals, and the fourth block is not related to the response variable.



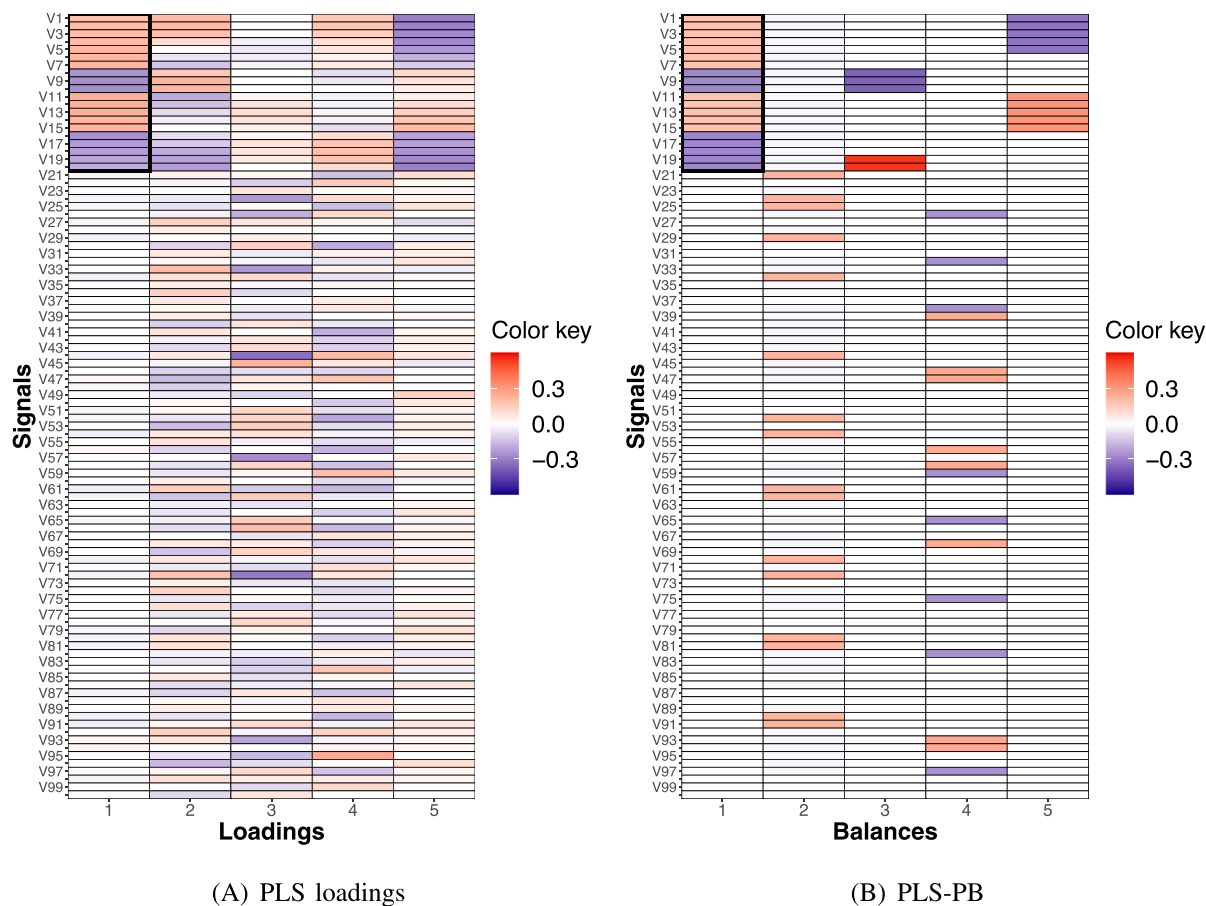


**FIGURE 1** Visual representation of covariance matrices used in the examples. Colors represent the covariance between pairs of pivot coordinates. (A) shows a single block of 20 meaningful markers, (B) four same-sized blocks of 20 markers each, and (C) four blocks of markers containing either 30 or 10 markers. Marker blocks are highlighted by a black frame.

We are now ready to compare PLS-PB to PLS loadings. For this, it is important to note that the structure of PLS-PB must not necessarily reflect the structure of the corresponding PLS loadings. This is due to the fact that orthonormality is required when constructing PLS-PB, while this is not the case for PLS loadings. These examples provide a first insight into the performance of PLS-PB to correctly identify markers in a collection of signals.

Figure 2 displays the comparison of PLS loadings and PLS-PB for the first case with one marker block. As we are typically interested in the first few PBs most strongly related to the response, we focus our discussion on the first five PLS loadings and PLS-PBs. These are displayed in the Figure 2A,B, respectively. Both PLS-PB and PLS loadings succeed in separating out the relevant markers (i.e., first columns in Figure 2) from signals corresponding to just random noise. However, the structure of PLS-PBs is more parsimonious, in the sense that it leads to a sparser solution including more zeros. The first PB (i.e., column 1 in Figure 2B) captures precisely the information contained in the first loading vector, highlighting all the markers (i.e., first 20 colored rows in Figure 2B). The third and fifth balances then capture several differences between markers in the block. As to the other balances, note that the PLS-PB solution does return false positives (e.g., second and fourth balances where noise signals are included).

Figure 3 displays the resulting PLS loadings and PLS-PB for case 2 with same-sized blocks of markers. The PLS-PBs in Figure 3B display a fairly neat structure. The first PB reproduces the information in the first loading vector as it identifies markers in the first block (those having the highest covariances). The second and third balances represent the third marker block (signals from V41 to V60) and distinguish groups of markers with either positive (balance 3) or



**FIGURE 2** Comparison of the first five PLS loadings (left) and PLS-PB (right) for the first example with one block of markers (case 1). Signal variables generated are arranged by rows. Colors represent values of loading vectors (left) and coefficients of PLS-PB (right). Signals in a numerator of a balance get a positive value, in a denominator a negative value. Signals not included in a balance get 0. Block of markers (first 20 signals) is highlighted into a black frame.

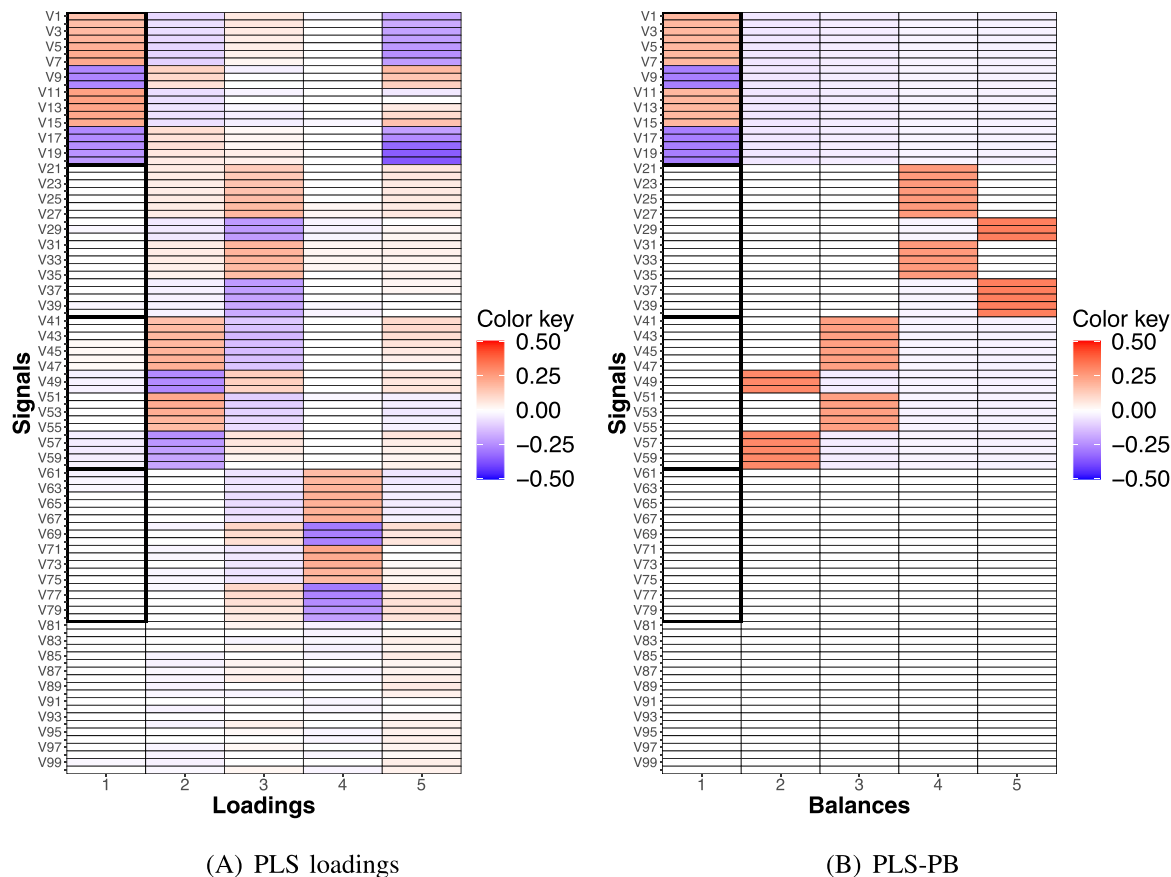
negative (balance 2) covariances. A similar pattern is observed for the second block (signals from V21 to V40) in the fourth and fifth balances. Note that all PBs correctly exclude the signals corresponding to random noise (i.e., the last 20); PLS loadings also give (close to) zero weight to these irrelevant signals. In contrast, the fourth block of predictors (not related to the response variable) is correctly flagged as irrelevant by PLS-PB (i.e., white cells in all balances) whereas PLS loadings still highlight them, particularly in the fourth column of Figure 3A.

Finally, Figure 4 displays the results for the case of varying block sizes (case 3). The largest blocks (i.e., the first and the third ones; signals V1-V30 and V41-V70) are correctly highlighted by the first PB. Moreover, the second PB identifies almost all markers from these blocks with negative covariances (only V69 and V70 are not included). The third PB captures markers in the second block (positive covariance; signals V31-V37). Similarly to the second PB, the fourth PB highlights the majority of markers from the largest blocks that have a positive covariance and, thus, the corresponding relationship with the response variable. As in case 2, PBs correctly excluded noise signals (the last 20), as well as the fourth block of predictors, which were all unrelated to the response variable.

In summary, these examples illustrate the potential of PB as a convenient counterpart to PLS loadings, as they deliver a simplified structure and the orthonormality constraint enables to, for example, perform interpretable regression analysis.<sup>1</sup>

### 3.2 | Simulation-based assessment

We investigate the predictive ability of PLS-PB in comparison to PLS loadings, although it is important to note that this is not the main focus of the method proposed in this work. Moreover, a natural alternative to PLS-PB in a regression



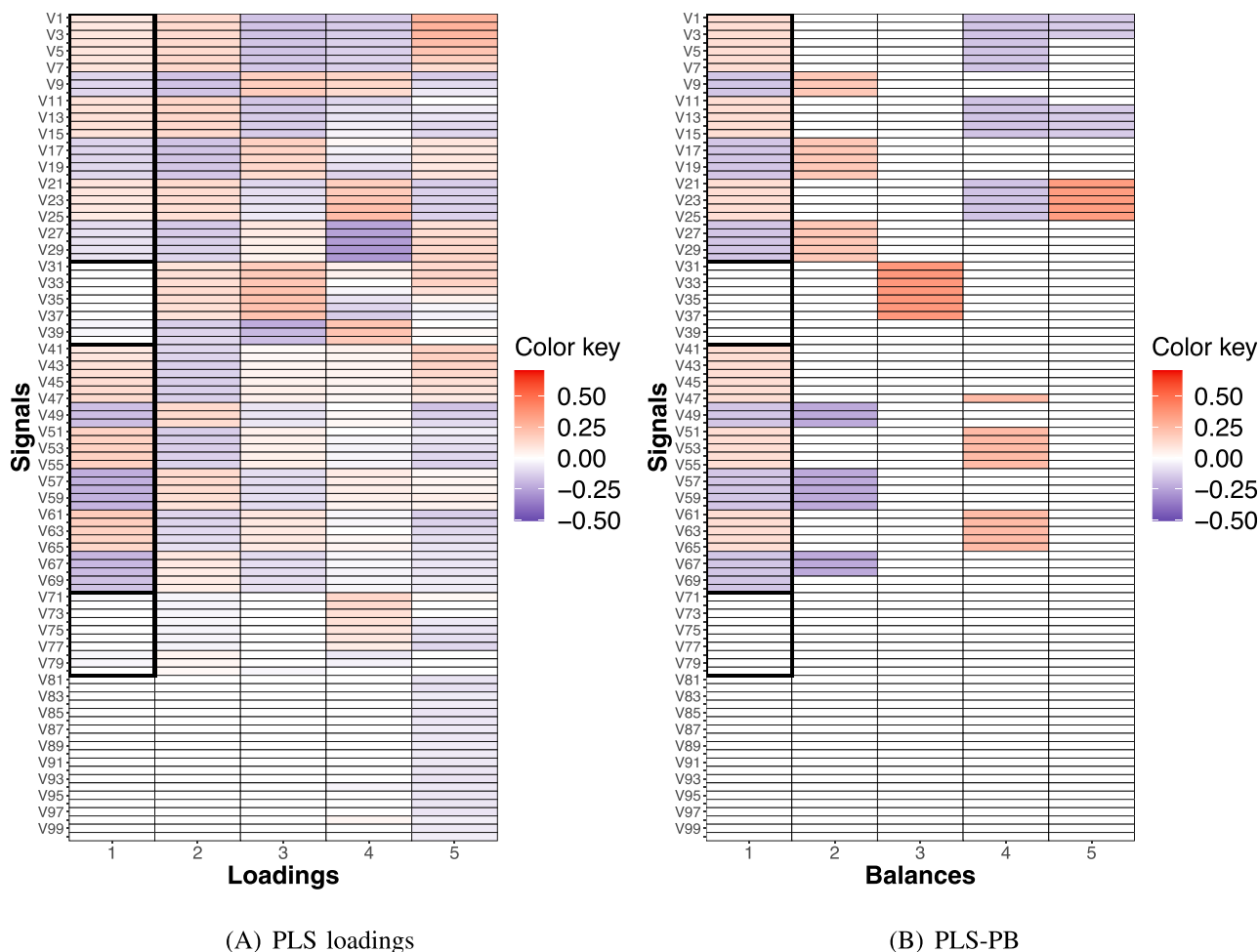
**FIGURE 3** Comparison of the first five PLS loadings (left) and PLS-PB (right) for the same-size blocks of markers setting (case 2). Signal variables generated are arranged by rows. Colors represent values of loading vectors (left) and coefficients of PLS-PB (right). Signals in the numerator of a balance have a positive value (negative value if in the denominator). Signals not included in a balance are given value 0. Blocks of markers (signals V1–V80) are highlighted into a black frame. The last block of predictors (i.e., signals V61–V80) is not related to the response variable.

setting are PCA-based PBs as mentioned before. While PLS-PB correspond to PLS regression with a simplified loading structure, PCA-PB should follow the behavior of the well-known principal component regression (PCR<sup>29</sup>), where the number of explanatory variables in a regression model is reduced using PCA. Both PLS and PCA regression are popular tools, for example, in chemometrics and molecular biology applications to cope with high-dimensionality and/or multicollinearity issues. We therefore devise a simulation study to compare the prediction performance of PLS PB, PCA PB, and PLS loadings. It is known that PLS regression generally leads to better prediction performance than PCR when just a few latent components are involved.

We consider three scenarios for the simulation study based on the three cases introduced in the previous section, that is: only one block of markers, several blocks of markers of the same size, and several blocks of markers of different sizes. The computed PBs (either PLS or PCA) are used to fit linear regression models like Equation (7), including only one PB up to all possible PBs, that is, 99. The *root mean squared error of prediction*,

$$\text{RMSEP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

is used as prediction performance measure, where  $y_i$  are the actual values and  $\hat{y}_i$  the corresponding predicted values. This was estimated by five-fold cross-validation (CV) to provide a more realistic assessment, and it was evaluated over 100 simulation runs in each case.

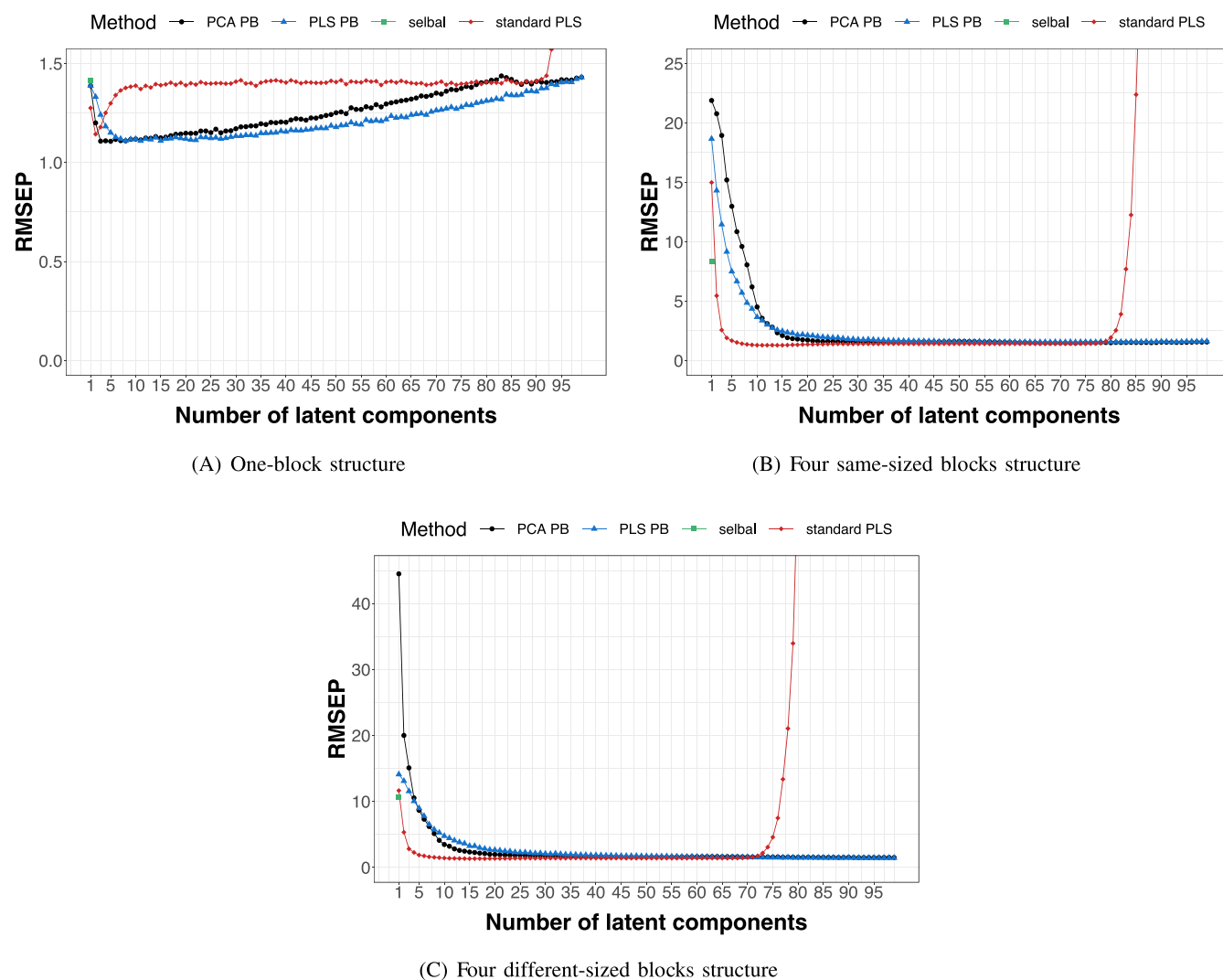


**FIGURE 4** Comparison of the first five PLS loadings (left) and PLS-PB (right) for the different-size blocks of markers setting (case 3). Signal variables generated are arranged by rows. Colors represent values of loading vectors (left) and coefficients of PLS-PB (right). Signals in the numerator of a balance have a positive value (negative value if in the denominator). Signals not included in a balance are given value 0. Blocks of markers (signals V1–V80) are highlighted into a black frame. The last block of predictors (i.e., signals V71–V80) is not related to the response variable.

Figure 5 shows the results for all three simulation scenarios, comparing cross-validated RMSEP of PLS-PB against PCA-PB and PLS loadings. The number of PBs used (PLS loadings in case of standard PLS), ranging from 1 to 99 ( $D - 1$ ) on the horizontal axis, can be understood as an index of model complexity. Each point represents the average value of the RMSEP over the 100 simulation runs.

First, we focus on the performance of PLS-PB against PCA-PB. When considering just a few PBs in the model, both perform similarly in the one-block setting, with their RMSEP being around 1.388 for both PLS-PB and PCA-PB for one balance (Figure 5A). In contrast, PLS-PB outperforms PCA-PB in the multiple marker settings (Figure 5B,C), with improvements of 14.7% and 68.3% in RMSEP for the first balance in cases 2 and 3, respectively. This is particularly relevant since, in practice, having just a few balances is preferred to facilitate interpretation of the results. Note that the value of the RMSEP coincides for both PLS and PCA-PB when the maximum number of PBs is used because both systems of PBs are orthogonal rotations of each other, as it is the case with any other olr coordinate representation.

Looking now at differences between PLS-PB and standard PLS, it is not surprising that the latter performs best for the lowest numbers of latent components. Our aim is to construct interpretable PBs that explain most of the variation in the response and, in doing so, remain competitive in terms of predictive performance in relation to ordinary PLS. Hence, although the PB-based approach shows a weaker prediction performance than standard PLS, it compensates this in terms of interpretability by providing a sparser solution (as demonstrated in Section 3.1). Moreover, note that for models with numerous latent components, the performance of standard PLS worsens dramatically, as reflected by the



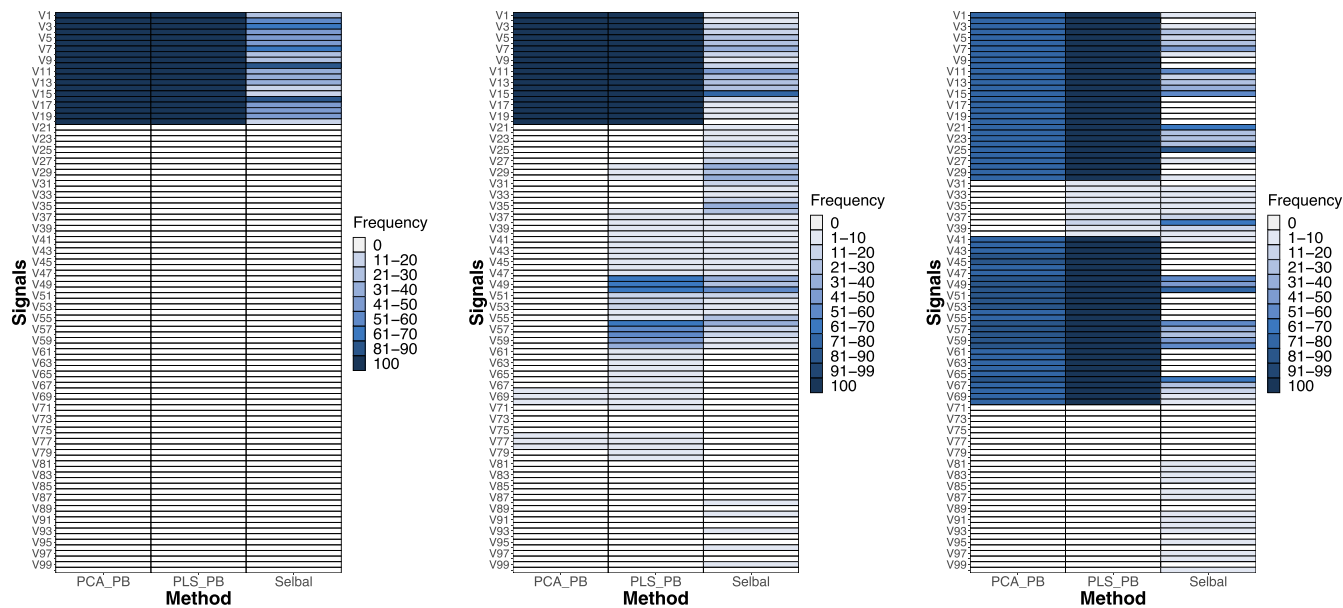
**FIGURE 5** Cross-validated root mean squared error of prediction using PB, PLS-PB (blue), PCA-PB (black), and ordinary PLS loadings (red) for each simulation scenario according to number of latent components used (either PB or ordinary loadings)

**TABLE 1** Comparison of mean RMSEP of the first PCA-PB and PLS-PB with mean RMSEP of selbal balance.

Example	PCA-PB	PLS-PB	selbal
1. One-block structure	1.388	1.388	1.413
2. Four same-sized blocks structure	21.89	18.663	8.307
3. Four different-sized blocks structure	44.486	14.109	10.605

large values of the RMSEP. This might be caused by numerical instability resulting from applying the SIMPLS algorithm as it provides non-orthogonal components causing degeneration when increasing the number of components.

Lastly, we compare the proposed PLS-PB approach to the selbal algorithm introduced in Rivera-Pinto et al<sup>12</sup> to identify an optimal balance between parts in high-dimensional compositions for regression and classification problems in a microbiome analysis context. The RMSEP of the latter is depicted in Figure 5 as a single green point, given that the selbal algorithm only selects a single balance. We can thus compare its performance to PLS-PB based on the first PB. Together with the depicted RMSEP value in Figure 5, the resulting mean values of the RMSEP for the three simulation scenarios is shown in Table 1. While selbal exhibits better performance in terms of prediction (cases 2 and 3), its ability to reflect the actual structure of markers in the data is notably poorer as discussed in the following.



(A) One-block structure

(B) Four same-sized blocks structure

(C) Four different-sized blocks structure

**FIGURE 6** Simulation: ability of the first balance to capture the data structure in the three simulation scenarios using the PCA-PB, PLS-PB, and selbal algorithm.

For each balance-based algorithm (PCA-PB, PLS-PB, and selbal), we evaluate its ability to correctly identify the biomarkers. For each simulation run, the algorithms are applied to determine the first PB (PCA-PB and PLS-PB) and the optimal single balance (selbal). Then, for each signal, we record whether it is included in such balances or not. The heatmaps in Figure 6 show the number of times a signal appears in the selected balance for each method, across the simulation runs. The selbal algorithm is unable to correctly distinguish marker from noise signals, especially in cases 2 and 3 (see respectively Figure 6B,C). In contrast, PLS-PB and PLS-PCA perform rather similarly across all cases, correctly identifying markers more often than selbal. They highlight most of the markers in cases 2 and 3, and they correctly classify noisy signals as irrelevant in more instances. In cases 2 and 3, both PLS-PB and PCA-PB struggle to identify important markers in the weakest/smallest marker block. The most noticeable difference between PLS-PB and PCA-PB is observed for the third marker block in case 2 (Figure 6B) and the two largest marker blocks in case 3 (Figure 6C), where PLS-PB more often captures the correct markers.

## 4 | APPLICATIONS

We here demonstrate the application of the PLS-PB approach on two empirical data sets. First, we consider a regression problem and then a classification task by simply accommodating the binary response into the PLS model.

### 4.1 | NMR data set

We use the data set from Štefelová et al.<sup>2</sup> consisting of high-throughput spectral profiles obtained by nuclear magnetic resonance (NMR). The data set involves a 127-part compositional predictor (metabolite signals, also called integrals) measured in  $n = 211$  rumen fluid samples from cattle. This was collected along with individual measurements of animal methane yield ( $\text{CH}_4$  in grams per kilogram of dry matter intake), which plays the role of a continuous response variable that we aim to model in terms of the metabolite composition.

We apply the proposed PLS-PB method and analyze its ability to predict the response. Considering a varying number of PBs, we aim to detect an optimum that combines sensible prediction accuracy with preferably a small number of PB. Moreover, we investigate whether the PBs reflect the structure of PLS loadings while simplifying the interpretation.

### 4.1.1 | Optimal number of principal balances

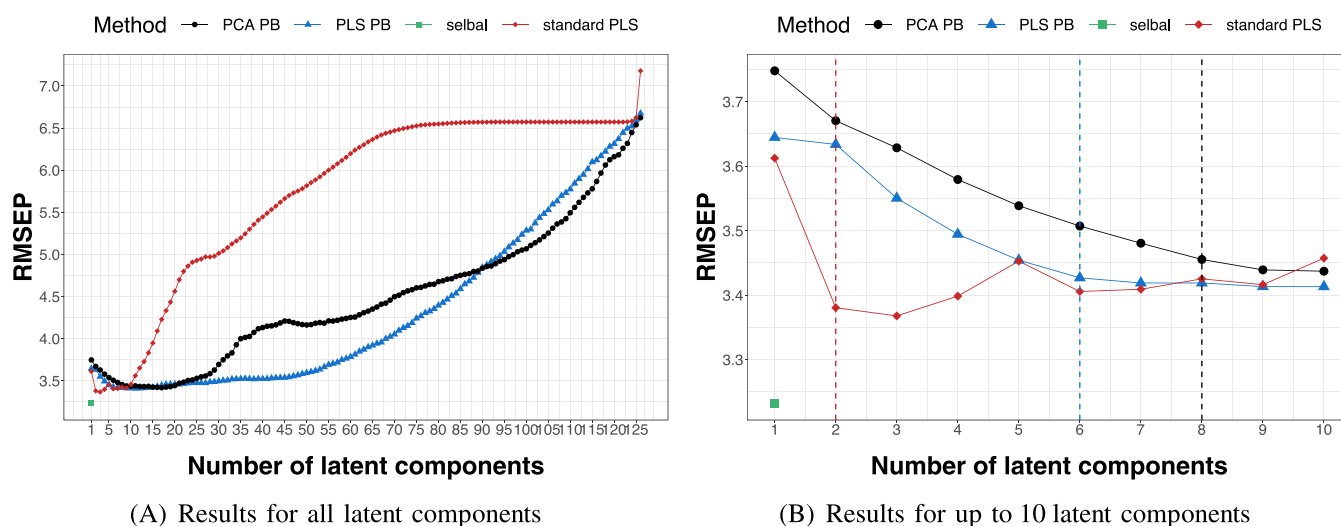
Similar to the simulation study (Section 3.2), we compute PLS and PCA-PBs as well as standard PLS and compare prediction performance for models including several PBs (ranging from the first PB or PLS loading up to all  $D - 1 = 126$  of them). Again, RMSEP is the measure used for comparison, and its values are estimates based on five-fold CV and averaged over 100 runs.

For each number of latent components (either PB or ordinary loadings), the mean and standard deviation of the RMSEP values were computed across the 100 runs. Then, given the lowest mean RMSEP, the model using the fewest number of balances within one standard error from such a minimum is chosen (one standard error rule; see Friedman et al.<sup>40</sup>). Accordingly, the most parsimonious model among those of best prediction performance is selected.

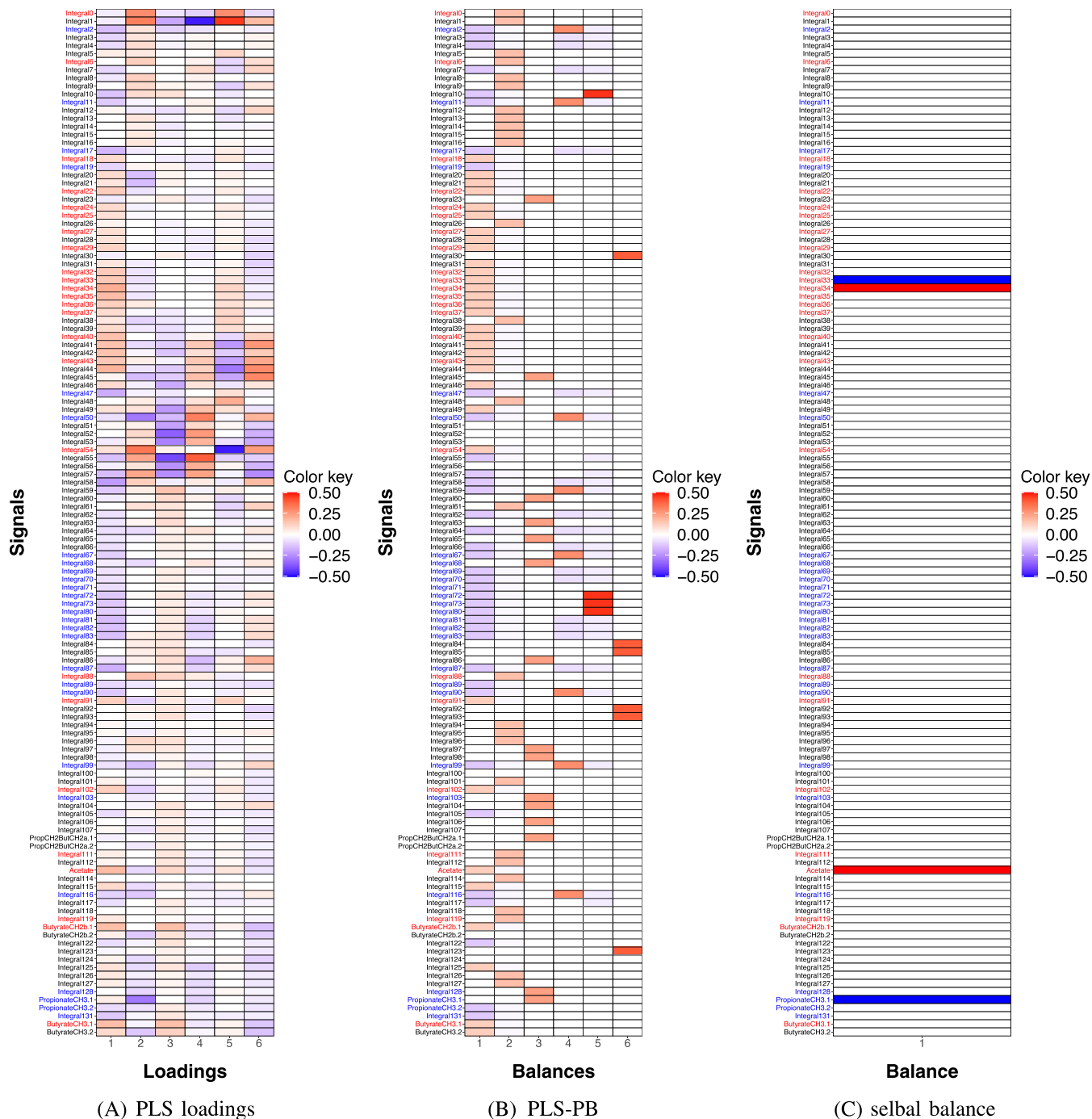
The results based on PLS-PB (blue) and PCA-PB (black), together with the results of standard PLS (red) and the selbal algorithm (green), are shown in Figure 7A. We can observe that PLS-PB outperforms PCA-PB for most part of the range of PBs. Unlike in the previous simulation study, the effect of a possible overfitting issue can be observed here as the RMSEP increases with the number of PBs. Also, as expected, the RMSEP values coincide (disregarding the minimal numerical difference) again for both approaches at the maximum number of PBs. Similarly to the simulation study in Section 3.2, standard PLS outperforms PLS-PB for the lowest numbers of latent components. However, weaker prediction performance of PLS-PB is compensated by the interpretability, which can be later seen in Figure 8. Moreover, it can be clearly seen in Figure 7A that standard PLS performs worse with increasing the number of loadings considered.

It can be observed in Figure 7A that the RMSEP for all three methods decreases rapidly at the beginning of the range as PBs (or ordinary loadings) are aggregated. Figure 7B zooms in on the results for the range of the first 10 latent components. While the smallest RMSEP for PLS-PB occurs for a model consisting of 11 PBs, a more parsimonious model using just six PBs provides comparable performance according to the one standard error rule (marked by a blue vertical dashed line in Figure 7B). For PCA-PB, the minimum occurs for a model containing the first 17 PBs, while the model with eight PBs lies within a one standard error of this minimum (marked by a black vertical dashed line in Figure 7B). For standard PLS, the minimum was reached for three loadings, and the optimal model determined by one standard error is the one with two loadings (red vertical dashed line in Figure 7B). The prediction performance was also assessed using the selbal algorithm. In this case, the resulting RMSEP is 3.227, much lower than for the other three methods. However, it was shown in Section 3.2 that selbal has a rather poorer ability to capture the structure of the data compared to PLS-PB, which can be seen in Figure 8C.

It is important to note that even if the PLS-PB approach does not outperform the PCA-PB approach, it is by construction expected to provide a more interpretable structure of PB, as these are tailored to maximize association with the response variable. The next section discusses the interpretative advantage of the PLS-PB approach.



**FIGURE 7** Prediction performance of PLS-PB (blue), PCA-PB (black), and standard PLS (red) methods on NMR data set for different choices of latent components used (either PB or ordinary loadings). Optimal number determined according to one standard error rule from minimum CV RMSEP is indicated by vertical dashed lines for each method. Results for the selbal algorithm were represented by a single dot (green)



**FIGURE 8** Comparison of the first six PLS loadings, six PLS-PB, and a selbal balance from the NMR data set. Red and blue labels are used to highlight markers identified in previous studies. Red (blue) colored text is used to highlight the names of markers having a positive (negative) relationship with the response variable.

#### 4.1.2 Comparison of PLS-PB to PLS loadings

We compare the PLS-PBs to PLS loadings to examine whether the former suitably reflect the latter (in terms of signs of coefficients) and, at the same time, facilitate interpretation.

In Figure 8, we display their values for the first six PBs, the optimal number determined in the previous section. The first PLS-PB reproduces the structure of signs of the coefficient values observed in the first PLS loading vector well, highlighting biologically meaningful markers identified in Štefelová et al.,<sup>2</sup> related to methane yield. For example, very distinct groups of identified markers are the group from Integral32 to Integral37 and a group from Integral67 to



Integral83 (with Integral68 being picked in the third balance). These markers are colored red and blue in Figure 8. Markers colored in red are those having a positive relationship with the response variable, whereas blue-colored markers are those that have a negative relationship with the response variable. The other PLS-PB, whose structure does not necessarily coincide with the structure of the respective PLS loadings due to the orthogonality constraint, capture some other patterns, related to both marker and non-marker variables. Moreover, PLS loadings also provide misleading information, as in the third loading there is a group of signals (Integral49 to Integral57) which is highlighted, but it is not biologically meaningful.<sup>2</sup> On the contrary, PLS-PB provides a neater and more parsimonious view.

## 4.2 | Metabolomic data set

We now consider a metabolomic data set consisting of  $n = 46$  observations and  $D = 209$  metabolites, thus representing the common high-dimensional setting with  $n < D$ . The response variable is in this case dichotomous and states cancerous ( $y_i = 1$ ) or healthy ( $y_i = 0$ ) tissues, having 23 patients suffering from lung cancer and other 23 being healthy.<sup>41</sup> The aim is to enable classification of tissues and, in particular, to reveal pathobiochemical changes of the disease. The samples were analyzed by a targeted metabolomic method based on HILIC liquid chromatography coupled with triple quadrupole mass spectrometry. This method allows to detect altogether 350 metabolites in different biofluids, tissues, and cells and covers main metabolic pathways.

We applied the PLS-PB and PLS-PCA methods and assessed their relative performance by five-fold CV over the range of possible numbers of PB as detailed previously. The RMSEP was replaced by the misclassification error, defined as

$$ME = \frac{1}{n} \sum_{i=1}^n I(\hat{y}_i \neq y_i), \quad (8)$$

where  $y_i$  denotes the group number of the  $i$ th object,  $\hat{y}_i$  is the estimated group number, and the index function  $I$  gives 1 if the group numbers are not the same and 0 otherwise.

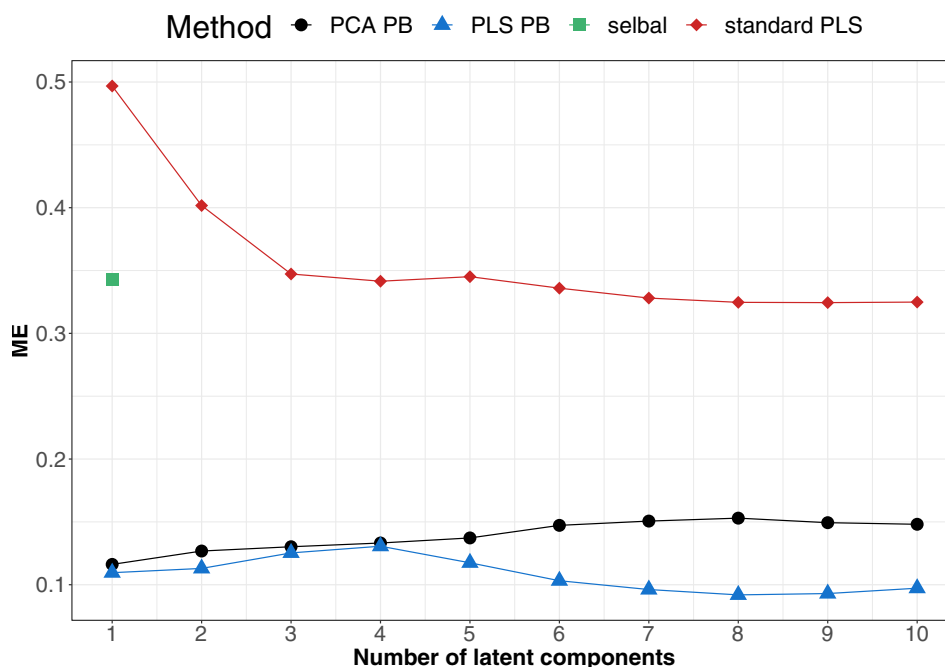


FIGURE 9 Prediction performance of PLS-PB (blue), PCA-PB (black), and standard PLS (red) methods on metabolomic data set across the first 10 latent components. Results for the selbal algorithm were represented by a single dot (green)

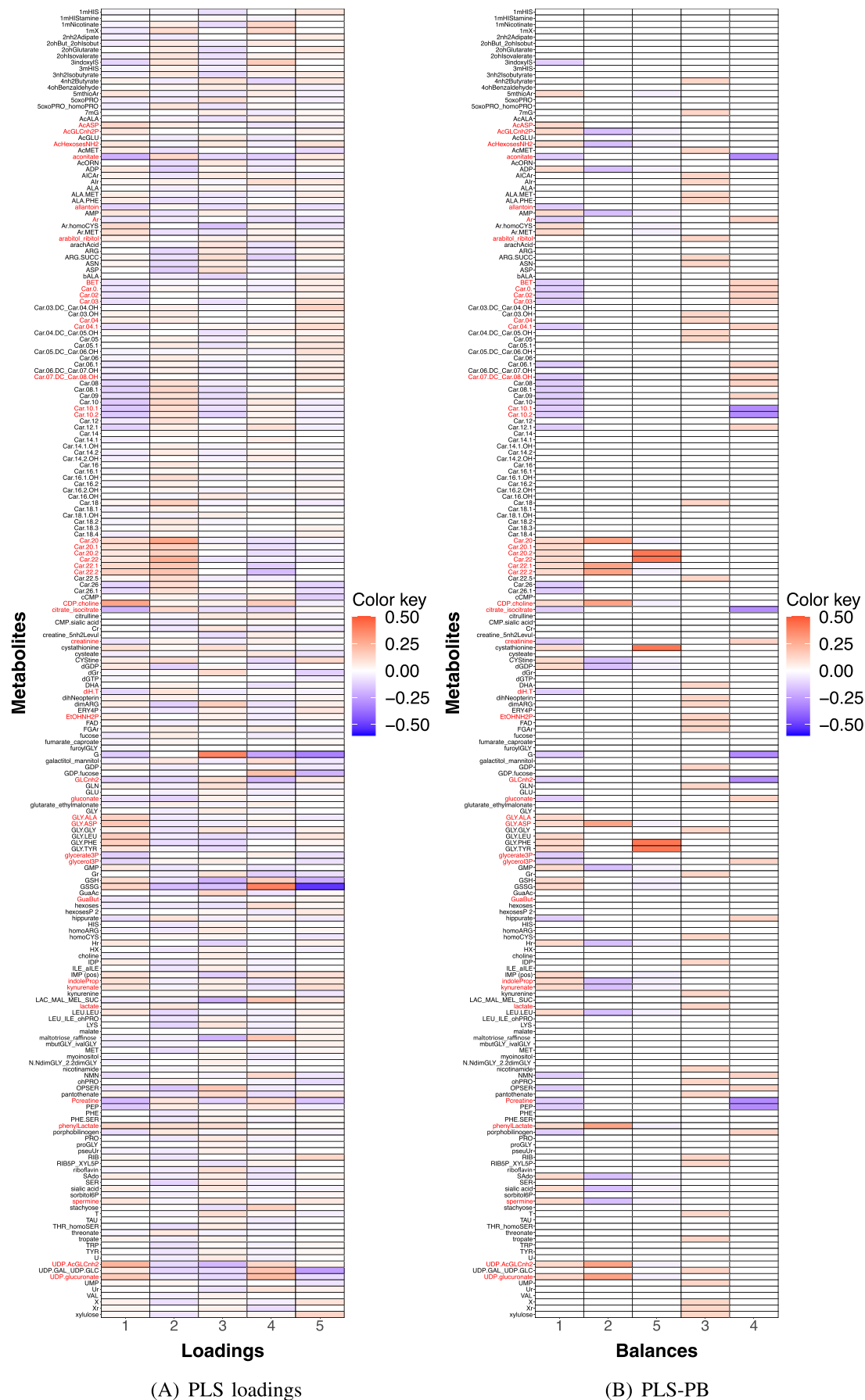


FIGURE 10 Legend on next page.

**FIGURE 10** Comparison of the first five PLS loadings and PLS-PB. Red colored text is used to highlight metabolites, which were marked as significant using  $p$ -values after Bonferroni correction.

Figure 9 displays the results for the first 10 latent components. Similar to the NMR data study, PLS-PB generally outperforms PCA-PB. Even though the numerical difference in ME is not dramatic, it can be seen that the ME of PLS-PB drops, whereas the ME of PCA-PB rather levels off. On the other hand, the performance of standard PLS is considerably worse than the other methods. Their performance was again compared to the result of selbal, for which the ME was 0.343. Selbal thus shows worse performance than PLS-PB and PCA-PB. Moreover, this latter does not recover the structure of markers very well.

Figure 10 displays PLS loadings and PLS-PB for the first five balances. Again, PLS-PBs appear to be less noisy than the PLS loadings counterparts. The latter puts an unnecessarily large emphasis on absolute differences. PLS-PBs, in contrast, are easier to navigate in the outcome matrix and show more agreement with univariate statistical analysis.<sup>41</sup> We can see general trends in decreasing short chain and medium chain (Car.0–Car.12) compared to increased very long chain (Car.20–Car.22) acyl carnitines, which are closely metabolically connected. Furthermore, selected groups of metabolites such as glycine dipeptides (GLY.ALA–GLY.TYR) and pyrimidine nucleotides (UDP.glucuronate, UDP.AcGlcNH<sub>2</sub>, and CDP.choline) show systematic trends. The PLS-PB approach in fact splits acylcarnitines into two separate groups, which could be then subject of future research.

## 5 | CONCLUSIONS

This manuscript introduces a new procedure to construct PBs within a log-ratio analysis framework for high-dimensional CoDa. We extend previous work in PBs by exploiting PLS as a dimension reduction tool that accounts for the relationship between a response variable of interest and a high-dimensional composition playing the role of predictor. The algorithm determines  $D - 1$  data-driven PLS-PBs that maximize their covariance with the response variable.

The proposal is applicable to both regression and classification problems and our numerical experiments firstly demonstrate that the resulting PLS-PBs provide a simplified structure of PLS loadings and outperform the original PCA-PB in terms of prediction performance. Secondly, when compared with the recently proposed selbal algorithm, which targets the same goal as PLS-PB, it is shown that although the selbal method may perform better in terms of prediction, it shows poorer ability to capture the data structure. Finally, PLS-PBs simplify the structure and enhance the interpretation of the results when compared with standard PLS. The method is further demonstrated on two empirical data sets regarding regression analysis with NMR spectral data and a classification task with metabolomic data. In both cases, the usefulness of the PLS-PB approach for variable selection and biomarker discovery is illustrated.

Building on the PLS-PB framework presented here, possibilities for further developments include its robustification to manage the potential influence of outlying samples in the results or the ability to deal with sparse data.

## ACKNOWLEDGMENTS

We thank the editor and referees for their constructive comments, which substantially improved the quality of the manuscript. JPA, JAMF, and KH gratefully acknowledge the support of the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and ERDF A way of making Europe (Grant PID2021-123833OB-I00); KH and PF were supported by the Czech Science Foundation, Project 22-15684L, and by the Austrian Science Foundation, Project I 5799-N, respectively; VN and KH were supported by IGA\_PrF\_2022\_008 Mathematical models and IGA\_PrF\_2023\_009 Mathematical models from the Internal Grant Agency of the Palacký University Olomouc and the Czech Science Foundation, Project 19-07155S. IW was supported by the Dutch Research Council (NWO) under grant number VI. Vidi.211.032.

## PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3518>.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

V. Nesrstová  <https://orcid.org/0000-0002-6137-7632>

J. Palarea-Albaladejo  <https://orcid.org/0000-0003-0162-669X>

## REFERENCES

- Hron K, Coenders G, Filzmoser P, Palarea-Albaladejo J, Faměra M, Grygar TM. Analysing pairwise logratios revisited. *Math Geosci*. 2021;53(7):1643-1666.
- Štefelová N, Palarea-Albaladejo J, Hron K. Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. *Stat Anal Data Mining*. 2021;14(4):315-330.
- Monti GS, Filzmoser P. Robust logistic zero-sum regression for microbiome compositional data. *Adv Data Anal Classif*. 2022;16(2):301-324.
- Dumuid D, Pedišić Z, Palarea-Albaladejo J, Martín-Fernández JA, Hron K, Olds T. Compositional data analysis in time-use epidemiology: what, why, how. *Int J Environ Res Public Health*. 2020;17:2220.
- Perujo N, Romani AM, Martín-Fernández JA. Microbial community-level physiological profiles: new analysis by a compositional data approach. *Ecol Indic*. 2020;117.
- Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol*. 2017;8:2224.
- Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika*. 1984;71:323-330.
- Bates S, Tibshirani R. Log-ratio lasso: scalable, sparse estimation for log-ratio models. *Biometrics*. 2019;75(2):613-624.
- Susin A, Wang Y, Lê Cao K-A, Calle ML. Variable selection in microbiome compositional data analysis. *NAR Genomics Bioinforma*. 2020;2(2):lqaa029.
- Gordon-Rodriguez E, Quinn TP, Cunningham JP. Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics*. 2022;38(1):157-163.
- Aitchison J. *The Statistical Analysis of Compositional Data*. Chapman & Hall; 1986. Reprinted 2003 with additional material by The Blackburn Press, London, UK.
- Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a new perspective for microbiome analysis. *MSystems*. 2018;3(4):e00053-18.
- Coenders G, Pawlowsky-Glahn V. On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT*. 2020;44(1):201-220.
- Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58(2):109-130.
- Gallo M. Discriminant partial least squares analysis on compositional data. *Stat Model*. 2010;10(1):41-56.
- Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T. PLS-DA for compositional data with application to metabolomics. *J Chemom*. 2015;29(1):21-28.
- Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Advances in principal balances for compositional data. *Math Geosci*. 2018;50(3):273-298.
- Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. *Math Geol*. 2005;37:795-828.
- Hron K. Advances in compositional data analysis. *Wiley StatsRef: Stat Ref Online*. 2018;1-15.
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana Delgado R. Principal balances. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI, eds. *4<sup>th</sup> International Workshop on Compositional Data Analysis (CoDaWork 2011)*. International Centre for Numerical Methods in Engineering (CIMNE); 2011:1-10.
- Barceló-Vidal C, Martín-Fernández JA. The mathematics of compositional analysis. *Austrian J Stat*. 2016;45(4):57-71.
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Modeling and Analysis of Compositional Data*. John Wiley & Sons; 2015.
- Filzmoser P, Hron K, Templ M. *Applied Compositional Data Analysis*. Springer; 2018.
- van den Boogaart KG, Filzmoser P, Hron K, Templ M, Tolosana-Delgado R. Classical and robust regression analysis with compositional data. *Math Geosci*. 2021;53:823-858.
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. *Math Geol*. 2003;35(3):279-300.
- Martín-Fernández JA. Comments on: compositional data: the sample space and its structure. *TEST*. 2019;28(3):653-657.
- Greenacre M. Variable selection in compositional data analysis using pairwise logratios. *Math Geosci*. 2019;51(5):649-682.
- Greenacre M, Grunsky E, Bacon-Shone J. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Comput Geosci*. 2021;148:104621.
- Varmuza K, Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press; 2009.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <https://www.R-project.org/>; 2020.
- Liland KH, Mevik B-H, Wehrens R, Hiemstra P. pls: partial least squares and principal component regression. R package version 2.7-3; 2020.
- van den Boogaart KG, Tolosana-Delgado R, Bren M. compositions: compositional data analysis. R package version 2.0-1; 2021.
- Fišerová E, Hron K. On interpretation of orthonormal coordinates for compositional data. *Math Geosci*. 2011;43(4):455-468.
- Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30.

35. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* 2019;28(11):1947-1951.
36. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* 2023;51(D1):D587-D592.
37. Tofte N, Suvitaival T, Trost K, et al. Metabolomic assessment reveals alteration in polyols and branched chain amino acids associated with present and future renal impairment in a discovery cohort of 637 persons with type 1 diabetes. *Front Endocrinol.* 2019;10:818.
38. Lam SM, Chua GH, Li X-J, Su B, Shui G. Biological relevance of fatty acyl heterogeneity to the neural membrane dynamics of rhesus macaques during normative aging. *Oncotarget.* 2016;7(35):55970.
39. Kvasnička A, Friedecký D, Tichá A, et al. Slide—novel approach to apocrine sweat sampling for lipid profiling in healthy individuals. *Int J Mol Sci.* 2021;22(15):8054.
40. Hastie T, Tibshirani R, Friedman J, et al. *The Elements of Statistical Learning*, Springer Series in Statistics. Vol 1. 2nd ed. Springer; 2009.
41. Cífková E, Brumarová R, Ovčáčiková M, et al. Lipidomic and metabolomic analysis reveals changes in biochemical pathways for non-small cell lung cancer tissues. *Biochim Biophys Acta (BBA) - Mol Cell Biol Lipids.* 2022;1867(2):159082.

**How to cite this article:** Nesrstová V, Wilms I, Palarea-Albaladejo J, et al. Principal balances of compositional data for regression and classification using partial least squares. *Journal of Chemometrics.* 2023;e3518. doi:10.1002/cem.3518

## APPENDIX A: ARTIFICIAL SETTINGS FOR COMPARISON: SIMULATION DESIGN

Compositions were simulated using so-called pivot coordinates, an instance of *olr* coordinates,<sup>33</sup> and assuming multivariate normality. In general, pivot coordinates  $\mathbf{z}^{(l)} = \text{ilr}(\mathbf{x}^{(l)})$  are defined as

$$\begin{aligned} z_j^{(l)} &= \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^{(l)}}{\sqrt{\prod_{k=j+1}^D x_k^{(l)}}} \\ &= \frac{1}{\sqrt{(D-j+1)(D-j)}} \left[ \ln \left( \frac{x_j^{(l)}}{x_{j+1}^{(l)}} + \dots + \frac{x_j^{(l)}}{x_D^{(l)}} \right) \right], l = 1, \dots, D, j = 1, \dots, D-1, \end{aligned}$$

where  $\mathbf{x}^{(l)} = (x_1^{(l)}, \dots, x_D^{(l)})^\top$  is a rearranged composition  $\mathbf{x}$  having the  $l$ -th part on the first position. It follows that via pivot coordinates, the relative information about the  $l$ -th part is captured by the first coordinate, which is advantageously used here for setting up our example. That is, it holds  $z_1^{(l)} = \sqrt{\frac{D}{D-1}} \text{clr}_l(\mathbf{x})$  for  $l = 1, \dots, D$ .

The steps to generate data (proposed in Štefelová et al.<sup>2</sup>) can be summarized as follows:

- First, generate pivot coordinates  $\mathbf{z}_i = (z_{i,1}, \dots, z_{i,D-1})^\top$  from a multivariate normal distribution  $N_{D-1}(\mathbf{0}, \boldsymbol{\Sigma})$ ,  $i = 1, \dots, n$ . The elements (with  $i, j = 1, \dots, D-1$ ) in the covariance matrix  $\boldsymbol{\Sigma}$  are shown in Figure 1. In case 1, there is one marker block consisting of 20 markers. The markers line up sequentially in groups of size 7, 3, 5, and 5, respectively. The groups alternate in sign with markers in the first and third groups (resp., two and four) being positively (resp., negatively) associated as indicated in red (resp. purple) color, see Figure 1A (all other covariances are equal to zero). Case 2 has four marker blocks consisting of 20 markers: All blocks have a structure similar to case 1, but they vary in the strength of the signal, as reflected by the shading in Figure 1B. In case 3, there are four marker blocks that vary in size: The first and third blocks consist of 30 markers while the second and fourth consist of only 10 markers. The covariance structure within each marker block is similar to the previous cases.
- Second, to obtain matrix  $\mathbf{X}$ , pivot coordinates need to be back-transformed:

$$\mathbf{x}_i = \text{ilr}^{-1}(\mathbf{z}_i) = (x_{i,1}, \dots, x_{i,D})^\top, i = 1, \dots, n.$$

- Finally, the response variable in our simulation design then results from

$$y_i = \beta_1 z_{i,1} + \beta_2 z_{i,2} + \dots + \beta_{2r-1} z_{i,2r-1} + \beta_{2r} z_{i,2r} + \varepsilon_i,$$

where  $\varepsilon_i \sim N(0,1)$ ,  $i = 1, \dots, n$ . In case 1, the coefficients corresponding to the first  $2r$  pivot coordinates that act as markers are  $\beta_j \sim U(0.1,1)$ ,  $j = 1, \dots, 2r$ , with their sign determined by the structure of the group within the marker block (i.e., seven coefficients having positive signs, the following three negative, etc.) In cases 2 and 3, the coefficients corresponding to markers in blocks 1–3 are obtained similarly to case 1. The coefficients corresponding to markers in the fourth marker block, in contrast, are set to zero, thus reflecting the presence of a block of correlated markers that does not relate to the response variable.