Received: 22 August 2022 | Revised: 11 October 2022 | Accepted: 14 November 2022

DOI: 10.1002/cem.3459

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

lrSVD: An efficient imputation algorithm for incomplete high-throughput compositional data

Javier Palarea-Albaladejo[1] | **Josep Antoni Martín-Fernández**[1] | **Anne Ruiz-Gazen**[2] | **Christine Thomas-Agnan**[2]

[1]Department of Computer Sciences, Applied Mathematics and Statistics, University of Girona, Girona, Spain

[2]Toulouse School of Economics, Toulouse, France

Correspondence
Javier Palarea-Albaladejo, Department of Computer Sciences, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain.
Email: javier.palarea@udg.edu

Funding information
French National Research Agency (ANR), Grant/Award Number: ANR-17-EURE-0010; Spanish Ministry of Science and Innovation, Grant/Award Numbers: MCIN/AEI/10.13039/501100011033, PID2021-123833OB-I00

Abstract

Compositional methods have been successfully integrated into the chemometric toolkit to analyse and model different types of data generated by modern high-throughput technologies. Within this compositional framework, the focus is put on the relative information conveyed in the data by using log-ratio coordinate representations. However, log-ratios cannot be computed when the data matrix is not complete. A new computationally efficient data imputation algorithm based on compositional principles and aimed at high-throughput continuous-valued compositions is introduced that relies on a constrained low-rank matrix approximation of the data. Simulation and real metabolomics data are used to demonstrate its performance and ability to deal with different forms of incomplete data: zeros, nondetects, missing values or a combination of them. The computer routines lrSVD and lrSVDplus are implemented in the R package zCompositions to facilitate its use by practitioners.

KEYWORDS

zeros, missing data, compositional data, singular value decomposition, log- ratio analysis

1 | INTRODUCTION

High-throughput technologies allow to quantify the relative abundances of fragments of genetic material, for example, through 16s rRNA and next generation sequencing, or to estimate the mixture of chemical signals in a biological sample, for example, via mass spectrometry and nuclear magnetic resonance (NMR) techniques. In recent literature there is a growing interest in the application and extension of the compositional data (CoDa) methodology for high-throughput sequencing and omics data.[1–5] On top of the particularities of CoDa as multivariate data carrying relative information, other distinguishing features of high-throughput CoDa include high dimensionality and sparsity. Furthermore, an important practical aspect to consider is the computational burden of their chemometric processing.

CoDa analysis assumes that the relevant information is contained in the ratios between the variables or constituting parts.[6,7] Given a CoDa matrix $\mathbf{T}_{n\times D}$ ($n \times D$; rows $\times$ columns), it is assumed that each observation $\mathbf{t}$ (a row of $\mathbf{T}$) is a member of an equivalence class. That is, the information contained in $\mathbf{t}$ is the same as in any other composition $k \cdot C(\mathbf{t})$ for any real scalar $k > 0$, where $C(\cdot)$ is the closure operation defined by $C(\mathbf{t}) = (t_1/\sum t_j,...,t_D/\sum t_j)$. This property, known

as scale invariance, implies that CoDa live in a quotient space[8] governed by the so-called Aitchison geometry.[7] In practice, CoDa are often expressed as elements of the D-part unit simplex $S^D = \{\mathbf{p} \in \mathbb{R}^D : p_j > 0; \sum p_j = 1; j = 1,...,D\}$, which is just a representative of that quotient space where the data are expressed in proportions (e.g., used to represent the relative abundance of species of metabolites in a sample or the relative abundance of genes as Kraken proportions). Although raw data are usually normalised to enable comparability between samples, the scale invariance property implies that compositional results will be generally independent of that.[9] A composition $\mathbf{x}$ can be mapped onto the ordinary real space by a vector of orthonormal log-ratio (olr, a.k.a. ilr commonly) coordinates $\mathbf{t}^* = \text{olr}(\mathbf{t}) = (t_1^*,...,t_{D-1}^*)$.[7] An example of olr-coordinates is given by

$$t_j^* = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{t_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} t_k}}, \; j = 1,...,D-1. \tag{1}$$

It is worth noting that orthogonality of the olr-coordinates refers to the fact that they are obtained from a log-ratio orthogonal basis, which does not imply that they are uncorrelated in a statistical sense. Importantly, a composition $\mathbf{t}$ and any other member of its equivalence class all have the same log-ratio coordinates. Formally, operations such as inner product, distance and norm in S^D can then be equivalently defined in coordinates. For example, the Aitchison distance d_a between two compositions $\mathbf{t}_1$ and $\mathbf{t}_2$ can be calculated as the Euclidean distance d_e between their corresponding vectors of olr-coordinates: $d_a(\mathbf{t}_1,\mathbf{t}_2) = d_e(\text{olr}(\mathbf{t}_1),\text{olr}(\mathbf{t}_2)) = d_e(\mathbf{t}_1^*,\mathbf{t}_2^*)$. Analogous definitions exist for the norm and scalar product, as well as for probability laws such as the log-ratio normal probability distribution. Moreover, as the Frobenius norm of an ordinary matrix is equivalent to the sum of the Euclidean norm of its rows,[10] the Frobenius norm for a CoDa matrix can be defined in terms of the Euclidean norms of the rows expressed in olr-coordinates.

Although it is not exclusive of the approach, a practical issue with working with log-ratios is that they cannot be computed when the data matrix is incomplete, typically because of the presence of zeros or missing values. For example, zeros in discrete-valued compositions derived from counting processes such as sequencing analysis. These count zeros are usually associated to limited sampling, e.g. fixed sequencing depth for rare taxa categories in microbiome studies. Bayesian multiplicative imputation[11] is a popular specialised method to deal with this case.[12,13] Moreover, continuous-valued compositions, such as those that can be obtained by spectrometry techniques in proteomics or metabolomics, often present what is called generically rounded zeros, that is, data not observed because their actual value is below a certain rounding-off error, detection limit (DL), signal-to-noise or other similar threshold (thus typically recorded as zero or marked as *nondetect* or *less-than* value). The existence of DLs is common in experimental studies, where some actual values fall beyond the sensitivity of the measuring device to distinguish them from background noise.[14]

Although rounded zeros have received most attention in regular CoDa analysis,[15] its treatment has been hardly studied for high-throughput CoDa where the number n of samples typically exceeds the number D of compositional parts. Thus, focusing on the case of wide and continuous-valued compositions, a new imputation procedure called lrSVD algorithm is introduced here for this case that allows to handle rounded zeros, general missing values and even a possible combination of them. The method is based on a constrained low-rank matrix approximation of the data, and the details are provided in Section 2. Its relative performance with regular and high-dimensional data sets is discussed in Section 3 by means of a simulation study. This includes an assessment of the computational cost, which is particularly relevant in high dimensions. Section 4 demonstrates the use of the proposed method and further assesses different aspects of its performance through an illustrative case study involving metabolomics data. Section 5 concludes with some final remarks.

2 | THE zlrSVD ALGORITHM

Nonparametric compositional imputation methods like multiplicative simple replacement[16] are still usable in the high-throughput context. But multivariate alternatives that account for the codependence structure such as the log-ratio EM algorithm[17] are limited to the $n > D$ case. The proposed lrSVD algorithm relies on a version of the well-known singular value decomposition (SVD) to produce a low-rank matrix approximation that is applicable to both regular ($n > D$) and wide ($n < D$) incomplete CoDa sets. It results from adapting a regularised iterative SVD algorithm for general missing data[18] to the compositional case and combining this with a method to consider lower and upper bounds for the imputed

data matrix cells.[19] Although primarily motivated by the rounded zeros (nondetection) problem, our lrSVD algorithm is designed to deal consistently with censored and general missing data, or in fact with both simultaneously. Thus, we will often refer generically to nonobserved data or incomplete data matrix in the following.

The original problem of finding an optimal low-rank approximation of a data matrix $\mathbf{T}$ is formulated as

$$\begin{array}{ll} \min & ||\mathbf{T}-\mathbf{R}||_F^2 \\ \text{s.t.} & \text{rank}(\mathbf{R}) \leq r, \end{array} \tag{2}$$

where $||\cdot||_F$ denotes the Frobenius norm. With complete data, a unique solution is guaranteed by selecting the first r terms of the SVD of $\mathbf{T}$.[20] However, Equation (2) does not have an explicit solution when $\mathbf{T}$ is incomplete. For this case, Josse and Husson[18] introduce an iterative algorithm where, after an initial imputation, nonobserved values in $\mathbf{R}$ are sequentially updated according to a regularised version of the SVD procedure until convergence. This approach is generalised to the case where information about lower and upper bounds for the nonobserved values is available. A box constraint $\mathbf{L} \leq \mathbf{R} \leq \mathbf{U}$ is considered, where the inequalities are cell-wise (e.g., $R_{ij} \leq U_{ij}$ for all i,j). However, it is not generally guaranteed that a matrix $\mathbf{R}$ will be found fulfilling the rank constraint and the box constraint simultaneously. For this reason, an additional intermediate matrix $\mathbf{M}_{n\times D}$ satisfying the box constraint is introduced. Then, an optimal couple $(\mathbf{M},\mathbf{R})$ is searched so that it optimises a convex linear combination of the Frobenius distance of $\mathbf{M}$ to the initial matrix $\mathbf{T}$ and the Frobenius distance of $\mathbf{M}$ to the matrix $\mathbf{R}$ of (up to) rank r. Namely, given two matrices $\mathbf{L}$ and $\mathbf{U}$, and the incomplete matrix $\mathbf{T}$ satisfying the box constraints $L_{ij} \leq T_{ij} \leq U_{ij}$ for all (i,j) corresponding to observed values, the approximation of $\mathbf{T}$ is the matrix $\mathbf{M}$ that solves the optimisation problem

$$\begin{array}{ll} \min & (1-\beta)\cdot||\mathbf{M}-\mathbf{R}||_F^2+\beta\cdot||\mathbf{M}-\mathbf{T}||_{F_{OBS}}^2 \\ \text{s.t.} & \text{rank}(\mathbf{R}) \leq r, \\ & \mathbf{L} \leq \mathbf{M} \leq \mathbf{U}, \end{array} \tag{3}$$

where $0<\beta<1$ and $||\cdot||_{F_{OBS}}$ means that only the observed values in $\mathbf{T}$ are involved. Problem (3) and the metric concepts involved are here translated into the geometry of the simplex for consistence with the compositional nature of the data. Thus, the Frobenius norm is defined in terms of the Aitchison distance d_a, the rank of $\mathbf{R}$ is understood as the rank of its expression in log-ratio coordinates $\mathbf{R}^* = \text{olr}(\mathbf{R})$, defined row-wise as in Equation (1), and the SVD reduction is applied to the matrix $\mathbf{M}^* = \text{olr}(\mathbf{M})$. In addition, for the particular case of rounded zeros, the matrix of lower bounds is $\mathbf{L} = \mathbf{0}$ (i.e., a matrix of zeros) and the matrix of upper bounds is given by $\mathbf{DL}$ (containing thresholds that can be different for each column and/or row). For the observed values the upper bound is set to the maximum of the corresponding column, which ensures that they satisfy the constraints.

The procedure devised is outlined in the following points and sketched in pseudocode in the Algorithm 1 box (for the case of zeros without loss of generality).

1. Initial imputation $\mathbf{M}^{(0)}$: rounded zeros in $\mathbf{T}$ are firstly replaced by multiplicative simple replacement; using $0.65\cdot\mathbf{DL}$ to obtain $\mathbf{M}^{(0)}$.
2. SVD reduction $\mathbf{R}^{*(1)}$: $\mathbf{M}^{(0)}$ is expressed in olr-coordinates $\mathbf{M}^{*(0)}$ and SVD is applied to $\mathbf{M}^{*(0)} = \mathbf{U}^{(0)}\cdot\boldsymbol{\Lambda}^{(0)^{1/2}}\cdot\mathbf{V}^{(0)^t}$ to obtain $\mathbf{R}^{*(1)} = \mathbf{U}_r^{(0)}\cdot\boldsymbol{\Lambda}_r^{(0)^{1/2}}\cdot\mathbf{V}_r^{(0)^t}$, a reduced rank-$r$ approximation of $\mathbf{M}^{*(0)}$ given a preset value r. The eigenvalues in $\boldsymbol{\Lambda}^{(0)}$ are denoted by $\lambda_1^{(0)} \geq \lambda_2^{(0)} \geq \ldots \geq \lambda_{D-1}^{(0)} \geq 0$.
3. Intermediate matrix $\mathbf{M}^{(1)}$ calculation: matrix $\mathbf{M}^{(0)}$ is updated to obtain $\mathbf{M}^{(1)}$ as follows. First, the matrix $\mathbf{R}^{*(1)}$ is regularised using $\mathbf{R}^{*(1)} - \mathbf{P}^{(1)}$, where $\mathbf{P}^{(1)} = \sigma^{(0)^2}\cdot\mathbf{U}_r^{(0)}\cdot\boldsymbol{\Lambda}_r^{(0)^{-1/2}}\cdot\mathbf{V}_r^{(0)^t}$ is the penalisation matrix, with

$$\sigma^{(0)^2} = \frac{n\cdot(D-1)}{\min\{n-1,D-1\}}\cdot\frac{\sum_{k=r+1}^{D-1}\lambda_k^{(0)}}{(n-r-1)(D-r-1)}$$

being the reduction factor for regularised methods.[18] Second, the raw regularised matrix $\mathbf{R}^{(1)}$ is obtained by inverse mapping olr^{-1}. The nonobserved cells in $\mathbf{M}^{(0)}$ are then updated using the values in $\mathbf{R}^{(1)}$. Moreover, the observed values in $\mathbf{M}^{(0)}$ are updated using the expression $\mathbf{R}^{(1)^{(1-\beta)}}\cdot\mathbf{T}^{\beta}$. Finally, if any, the values of $\mathbf{M}^{(1)}$ larger than $\mathbf{DL}$ are replaced by $\mathbf{DL}$.

4. Convergence check: Steps 2 and 3 are repeated until $||\mathbf{M}^{*(k)} - \mathbf{M}^{*(k-1)}||_F^2 \leq \delta$, with δ being the prefixed convergence criterion value (e.g., $\delta = 10^{-6}$).
5. Imputed matrix $\mathbf{M}$ calculation: inverse mapping olr^{-1} is applied to the rows of $\mathbf{M}^{*(k)}$ to obtain $\mathbf{M}$. This is done by preserving the originally observed cells in $\mathbf{T}$ and just imputing the nonobserved cells by their corresponding inverse olr-coordinates. Note that the rows in $\mathbf{M}$ are by construction elements of the unit simplex S^D. Hence, if the original data are closed to another constant, the imputed data matrix is rescaled accordingly. Otherwise, if $\mathbf{T}$ is not closed, the imputed cells are adjusted to provide a compositionally equivalent data set in the original units.[15] In any case, the relative structure of the data is preserved, and the basic properties of scale invariance and subcompositional coherence are guaranteed for the imputed data matrix.

Algorithm 1 lrSVD algorithm pseudocode

Input: Incomplete data matrix $\mathbf{T}$, matrices of lower $\mathbf{L}$ and upper $\mathbf{U}$ limits

1: Initial imputation $\mathbf{M}^{(0)}$: multiplicative simple replacement (zeros) or observed geometric mean (missing)

2: $\mathbf{M}^{(0)}$ expressed in olr-coordinates $\rightarrow \mathbf{M}^{*(0)}$

3: **while** $||\mathbf{M}^{*(k)} - \mathbf{M}^{*(k-1)}||_F^2 \leq \delta$ **do**

4: $\quad \text{SVD}(\mathbf{M}^{*(k)}) \rightarrow \mathbf{R}^{*(k+1)}$

5: $\quad$ Update $\mathbf{M}^{(k)} = \begin{cases} \text{olr}^{-1}(\text{regularised } \mathbf{R}^{*(k+1)}) & \text{if non-observed cell} \\ \mathbf{R}^{(k)^{(1-\beta)}} \cdot \mathbf{T}^{\beta} & \text{if observed cell} \end{cases}$

6: **end while**

Output: Final imputed matrix $\mathbf{M} = \begin{cases} \text{olr}^{-1}(\mathbf{M}^*) & \text{if non-observed cell} \\ \mathbf{T} & \text{if observed cell} \end{cases}$ and scaled to original units

As the procedure is formulated in olr-coordinates (Equation 1), it is important to note that the solution to (3) (understood in a simplicial sense) is the same regardless of the system of olr-coordinates used. It might be argued that other log-ratio representations could be used, such as additive log-ratio (alr) or centred log-ratio (clr) representations. However, alr-coordinates are a nonisometric mapping between the simplex and the real space, which would deform the norms involved in the optimisation problem.[7] Moreover, even though the clr representation is isometric, the fact that the geometric mean of all parts of the composition is embedded in this mapping favours the propagation of numerical errors.[21] To minimise error propagation using olr-coordinates in the lrSVD algorithm, the parts of the composition are prearranged in decreasing order according to number of nonobserved cells, so that the part placed on the first position includes the largest amount of zeros. Note that what we call olr-coordinate throughout this work can be exactly exchanged for the more widely used term ilr-coordinate (with the *i* standing for *isometric*). We just consider that using olr better honours the most outstanding feature of such log-ratio representation and helps to differentiate it from the clr representation which is also isometric. Finally, the procedure is easily extended for general missing values, and for censored and missing data simultaneously, by setting the upper bound for missing cells to the observed maximum of the corresponding column. As to the initial imputation step, missing values are replaced here by the geometric mean of the observed values in the column.

2.1 | Parameter settings

As shown above, the lrSVD algorithm involves three basic choices with regard to parameter settings: the initial imputation (Step 1), the dimension of the low-rank approximation r (Step 2) and the weighting parameter β (Step 3). Note that the two latter aspects have been discussed in detail in the referenced literature and the overall conclusions are equally applicable here.

Regarding the sensitivity to the initial imputation, our tests revealed that the results are in fact very robust as illustrated in Section 4. Although we could consider several starting points,[19] we concluded that the potential benefit is not worth the computational cost of such an extra layer of complexity.

As to the parameter r, this corresponds to the number of latent variables or components used for the low-dimensional representation of the original data set as for standard principal component analysis. Note that this must be lower than $D-1$, that is, lower than the actual dimension of the sample space, and for wide data where $n<D$ a rank reduction only takes place if $r<n-1$. For complete data, a number of criteria have been proposed to assist in making this choice and none has really proved its overall superiority. Previous work concludes that the particularities of the data set at hand largely affect their performance.[22] Moreover, methods allowing for missing values are to our knowledge very rare. One is implemented in the R package *missMDA* and involves to sequentially removing each observed value of the given data matrix, computing its prediction from the low-rank model for a collection of values of r and then select the one minimising the mean squared error between observed and predicted values.[23] Our attempt to emulate this procedure while respecting the properties of CoDa did not produce satisfactory results unfortunately. Hence, our practical recommendation is for the users to conduct their own exploration and sensitivity analysis with the data at hand, considering the purpose of the study and the statistical methods to be used. Some further insight about this is provided in Section 4.

With regard to β, this is defined in $(0,1)$ and determines the weight given to each term of the objective function in Equation (3). With a large β the second term is penalised more, so that $\mathbf{M}$ is forced to be closer to $\mathbf{T}$ and further away from satisfying the low-rank constraint. Contrarily, when the penalty to the first term is larger, that is, for small β values, the algorithm makes $\mathbf{M}$ to be further away from $\mathbf{T}$ while enforcing more strictly the low-rank constraint. From our own investigation using the reference data in this study, the distributions of the imputed values were overall very similar for a range of values β within the interval $(0,1)$. The lrSVD algorithm showed no convergence problems for $\beta=0$; however, we found some issues as β approached 1 and larger number of iterations were required for convergence. Based on these analyses and previous work,[19] we recommend using $\beta\approx 0.5$ and only use the elements of $\mathbf{M}$ corresponding to nonobserved values in $\mathbf{T}$ to perform the imputation, that is, leaving the observed values in $\mathbf{T}$ unchanged in $\mathbf{M}$. This takes jointly into account the observed values and the low-rank approximation at every step of the iterative procedure and leads to faster convergence.

3 | PERFORMANCE ASSESSMENT BY SIMULATION

3.1 | Distortion measures

The performance of the proposed method was assessed in terms of the distortion it introduces in comparison with a reference complete data set $\mathbf{X}$. In line with previous reports,[11,24] we considered the following two measures:

- Average difference in covariance structure (ADCS)

$$\mathrm{ADCS}=\frac{1}{D-1}||\mathbf{S}_{\mathbf{X}}^{*}-\mathbf{S}_{\mathbf{M}}^{*}||_{F}, \tag{4}$$

  where $||\cdot||_F$ denotes the Frobenius matrix norm as applied to the difference between the sample covariance matrix of the original and imputed data sets expressed in olr-coordinates ($\mathbf{S}_{\mathbf{X}}^{*}$ and $\mathbf{S}_{\mathbf{M}}^{*}$, respectively). Recall that the Frobenius norm is invariant to the olr representation.

- Compositional error deviation (CED)

$$\mathrm{CED}=\frac{1/n_K\sum_{k\in K}d_a(\mathbf{x}_k,\mathbf{m}_k)}{\max\limits_{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}_{-K}}d_a(\mathbf{x}_i,\mathbf{x}_j)}, \tag{5}$$

  where d_a denotes the Aitchison distance between compositions,[25] n_K is the number of compositions $\mathbf{x}_k$ (data matrix rows) that contain at least one zero (K denotes the index set referring to such compositions), and $\mathbf{m}_k$ is the

corresponding imputed composition. The denominator is the maximum Aitchison distance between compositions in the subset formed by fully observed compositions ($\mathbf{X}_{-K}$).

The ADCS measures the effect of the imputation on the covariance structure, with no distortion corresponding with $ADCS = 0$. The CED is a scaled average of the differences between original and imputed samples. It is normalised by the maximum observed difference between compositions to account for the overall spread in the data.

3.2 | Comparative performance

A simulation study was set up to assess the behaviour and performance of the proposal in different scenarios and in relation to alternative imputation methods. This was based on the ADCS and CED measures presented in Equations (4) and (5) above, as well as on computing time, that is, time to produce the imputed data matrix. Without loss of generality, the focus was on the case of zeros or nondetects. To facilitate comparison, these methods were applied using their default parameter settings.

Complete CoDa sets $\mathbf{X}_{n\times D}$ were simulated from a normal distribution on the unit simplex,[26] considering a vector of means equal to zero and covariance matrix $\boldsymbol{\Sigma}$ in the associated space of olr-coordinates. Based on the design implemented in a previous study,[27] variances equal to 1 and a two-block correlation structure for the matrix of olr-coordinates $\mathbf{X}^*_{n\times(D-1)}$ were set. These parameters determined the shape of the cloud of compositions in the simplex. The subgroups consisted of 2/3 and 1/3 of the olr-coordinates, with the correlations within each group being either homogeneously weak $\rho = 0.2$ or strong $\rho = 0.8$ and the correlations between variables in the two different groups being $\rho = 0$. This determined simulated data matrices with a Rank 2 covariance matrix representing kind of two extreme situations.

The simulated data were mapped into the simplex by inverse olr mapping $\text{olr}^{-1}(\mathbf{X}^*)$. Following a previous study,[24] incomplete matrices $\mathbf{T}$ were generated by artificially imposing zeros in every second column for all values below a certain quantile Q_p, with p determining the proportion of zeros in the column. A collection of 40 scenarios resulted from combining values for the correlations $\rho = \{0.2,0.8\}$, numbers of compositional parts $D = \{6,30,60,120,240\}$ and proportion of zeros per column $p = \{0.1,0.2,0.3,0.4\}$. Note that the proposed method is aimed for wide data sets having in mind the context of proteomics or metabolomics studies where typically the number of variables is up to the hundreds and the occurrence of nonobserved values is up to moderate. Given this fact, what was observed in similar studies in the referenced literature, and the need for workable computing times for the simulation experiment, our comparative performance study was limited to the case of up to 240 compositional parts and up to 40% nonobserved values in every second column. Different numbers of observations n were also initially considered. However, the results were very much comparable, and thus, only results corresponding to $n = 50$ are shown here. Hence, both regular and wide data set cases in combination with variation in the previous parameters were considered.

The zeros were subsequently replaced to produce imputed CoDa matrices $\mathbf{M}$ using the following methods:

- Multiplicative simple replacement (`multRepl` function in *zCompositions* package) as reference nonparametric method, applied using the default imputation by 65% of the DL followed by multiplicative adjustment.
- Log-ratio EM algorithm (`lrEM` function in *zCompositions* package) as a multivariate model-based counterpart, applied using maximum likelihood estimation and starting log-ratio covariance matrix based on a preliminary `multRepl` step. This latter avoided some issues across the simulations derived from the possibility of randomly generating incomplete matrices not including enough fully observed cases to allow for ordinary initial covariance matrix estimate.
- PLS regression-based method as implemented in the `impRZilr` function of the *robCompositions* package (setting method to PLS) as a competing alternative for rounded zero imputation in wide data.
- Proposed lrSVD algorithm as implemented in the `lrSVD` function of the *zCompositions* package.

A total of 500 data sets were simulated for each scenario and ADCS, CED and computing time were summarised by the average across simulation runs. Imposing zeros in half the variables tends to generate a large number of different zero patterns, including cases where all samples contain at least one zero. It also induces a somewhat exaggerated concentrations of zeros in single variables, particularly for the highest values of p. This scenario is specially challenging for methods operating on a multivariate and pattern-wise basis, as relatively little information might be available in some cases to fit the relationships between variables. Figures 1 and 2 summarise the results based on ADCS, CED and computing time for the weak and strong two-block correlation structures (`impRZilr` using PLS regression is denoted

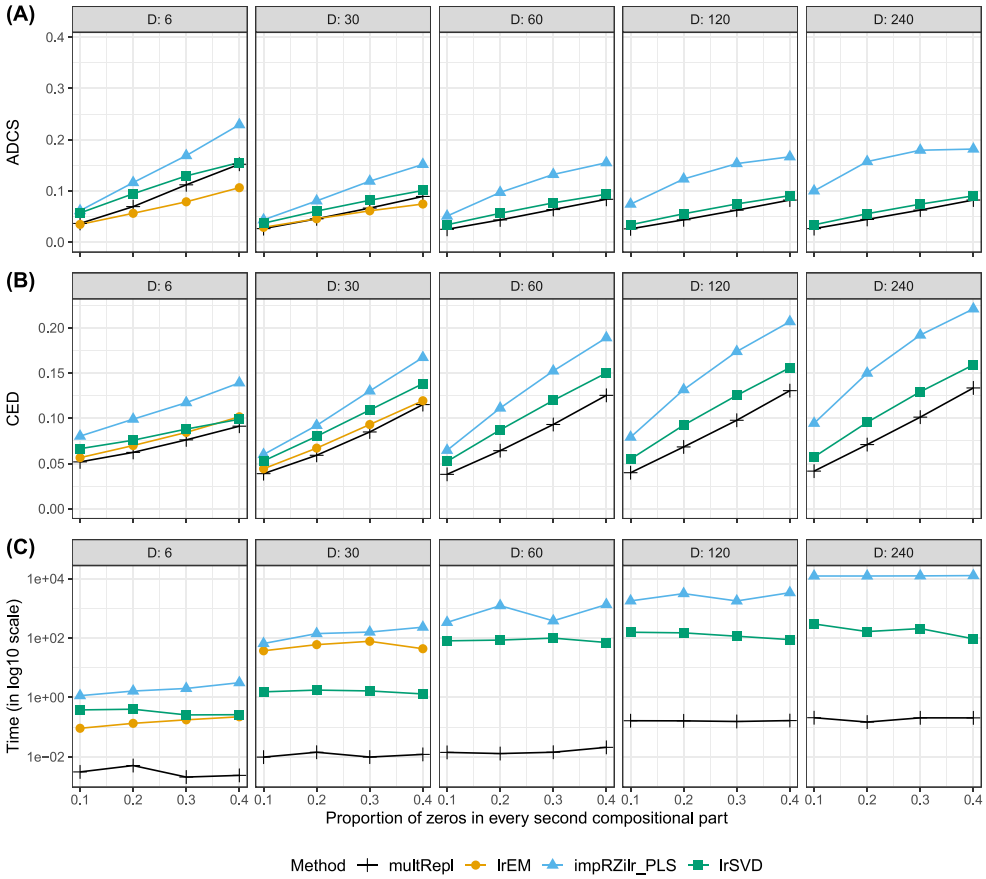

FIGURE 1 Comparative performance based on simulated data with a two-block correlation structure (weak correlation case: $\rho = 0.2$). (A) Distortion in covariance structure (ADCS), (B) distortion in compositional samples (CED) and (C) computing time (in $\log_{10}$-scaled seconds)

impRZilr_PLS in the graphs). Note that it was needed to slightly edit the original computer implementation of `impRZilr` to be able to fix the number of PLS components and facilitate comparison with lrSVD (the number of latent components in `impRZilr` and the low-rank approximation in `lrSVD` were both set to $r = 2$).

When the associations between variables are generally low ($\rho = 0.2$; Figure 1), making use of the codependence structure to inform imputation is not expected to be an advantage. In our study, the simulation results indicated that this particularly affected `impRZilr`, which provided the worst results for both distortion measures. The lrEM algorithm, even though it also relies on the codependence structure, outperformed the other methods as p increases in the regular multivariate setting ($n = 50, D = 6$) and provided results comparable to `multRepl` and `lrSVD` for $D = 30$. The lrEM algorithm could not be used; however, once $D > n$, and it is also expected to become unstable as D approaches n. Generally, imputation by `multRepl` or `lrSVD` showed a comparable performance and caused the lowest distortion in this case, particularly in terms of the covariance structure (ADCS). The increases in both ADCS and CED were mostly linear with p for all methods. As to computation cost, this increased dramatically with D for `impRZilr`, regardless of the value of p, showing marked differences in the most high-dimensional scenario ($D = 240$), whereas it remained stable and relatively low for the others (note that computing time is represented in seconds in $\log_{10}$ scale to facilitate visualisation).

When the associations are generally high ($\rho = 0.8$; Figure 2), then the codependence structure is expected to be a relevant source of information. Thus, `multRepl` which ignores this was the most affected and provided the worst results overall, with the difference increasing with p, except in relation to computing time, which was the lowest across all scenarios. For $D = 6$, `lrEM` provided some lower distortion than the others; for $D = 30$, `impRZilr` and `lrEM` produced very comparable results; and `impRZilr` generally showed a better performance than the others from that scenario on. The proposed lrSVD algorithm remained close to `impRZilr` with increasing D and p, particularly in terms of CED,

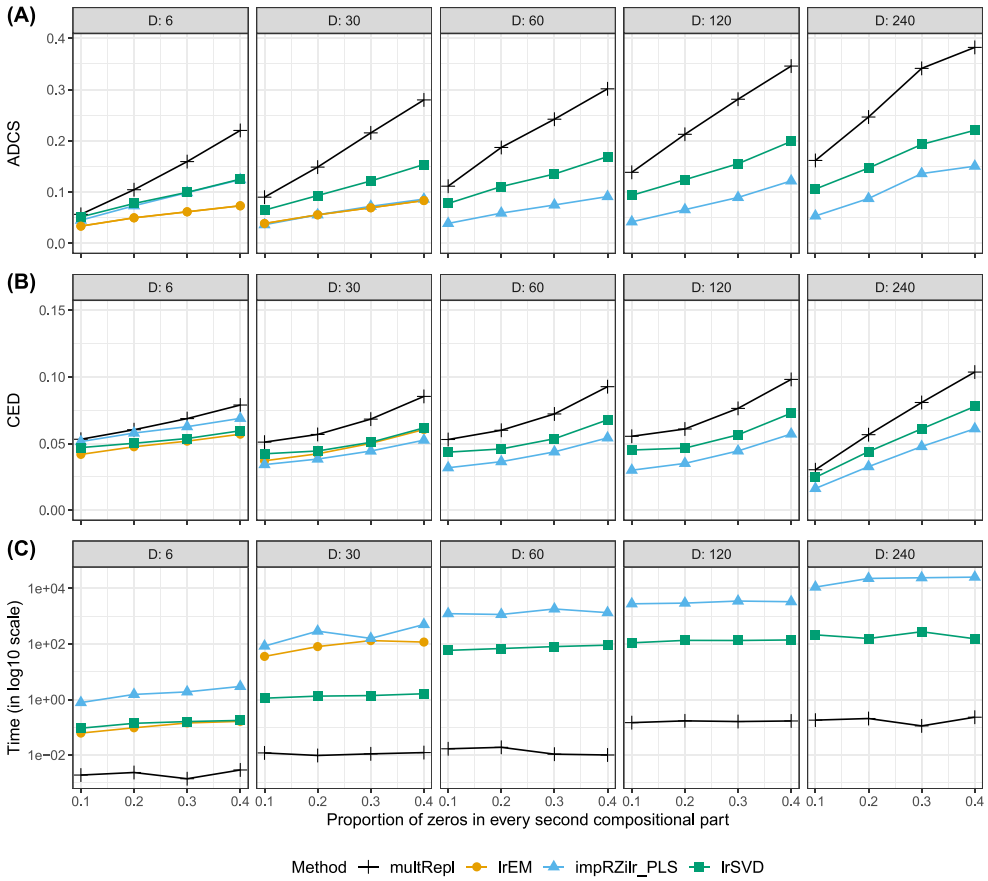

FIGURE 2 Comparative performance based on simulated data with a two-block correlation structure (strong correlation case: $\rho = 0.8$). (A) Distortion in covariance structure (ADCS), (B) distortion in compositional samples (CED) and (C) computing time (in $\log_{10}$-scaled seconds)

and had markedly lower computational cost that remained fairly stable with increasing dimensions (note that computing time is represented in seconds in $\log_{10}$ scale to facilitate visualisation).

4 | ILLUSTRATION USING METABOLOMICS DATA

In this section, we demonstrate the practical application and relative performance of the lrSVD algorithm using as reference a biomolecular data set in the context of predictive modelling of livestock greenhouse gas emissions. The data consist of high-throughput spectral profiles, representative of metabolite signals, acquired by NMR spectrometry analysis of rumen fluid of 67 samples from livestock fed with a low forage diet.[28] The raw samples went through a number of ordinary preprocessing stages, including phase and baseline correction, binning to integrate the area under the signal peaks and normalisation by referencing all the integrals to a same integral (corresponding to methyl of propionate), which resulted in $D = 127$ nonzero NMR integrals per sample. Methane originated from food digestion in ruminants is a main contributor to greenhouse gas emissions, and it was measured (CH_4 in grams per kilogram of dry matter intake) along with the ruminal metabolite composition from each animal. These data also serve here to further assess the performance of the method in high dimensions, investigating the effect of the number of latent components chosen, the sensitivity to the starting point, as well as illustrating its use for the simultaneous imputation of zeros and missing data.

A total of 500 data sets were simulated using the compositional centre and variation matrix from the original complete NMR data as reference. Seventy-five signals were considered minor (short spectral peaks) and potentially able to include zeros associated to amounts below the DL of the measurement device. Thus, for each simulated data set, zeros were imposed in a random selection of between 20% and 60% minor signals by using the Q_p quantile as threshold, with

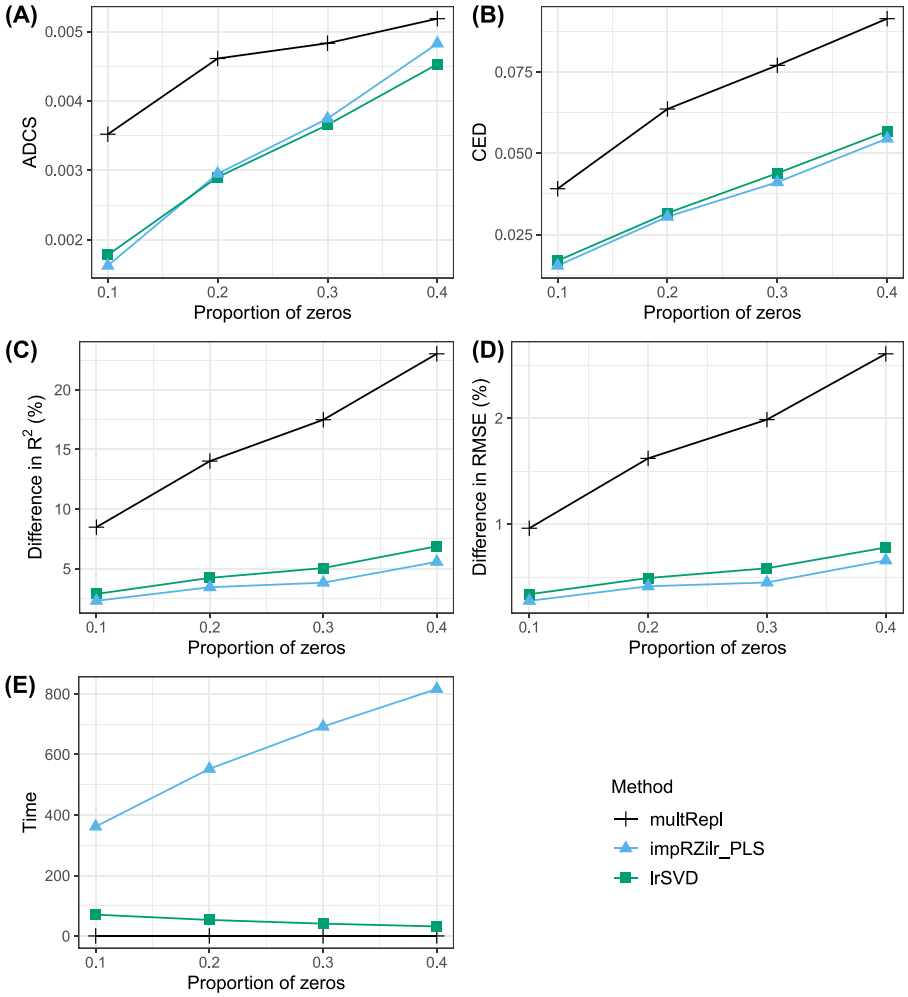

FIGURE 3 Comparative performance based on NMR high-throughput data (67×127). Distortion (A) in covariance structure (ADCS) and (B) in compositional samples (CED). Percentage relative difference (C) in R^2 and (D) in root mean squared error of prediction (RMSE). (E) Computing time (in seconds)

p from 0.1 to 0.4 by 0.1. Figure 3 summarises the overall performance of `multRepl`, `impRZilr` and the proposed lrSVD algorithm based on ADCS, CED and computing time as used above, as well as percentage relative difference between complete and imputed data in the R^2 coefficient and root mean square error of prediction (RMSE) from a PLS regression model fitted to methane yield (expressed in log scale) on the ruminal NMR signals. The number of components for lrSVD, `impRZilr` and the final PLS regression model was set to $r = 2$ based on results from a preliminary principal component analysis of the complete data.

The results suggest that PLS- and lrSVD-based imputation had similar performance for all increasing proportions of zeros, with multiplicative simple replacement introducing somewhat higher distortion in terms of ADCS and CED (Figure 3A,B). As to the PLS regression fit, the former methods had a lower impact on the ordinary goodness of fit statistics relative to the complete data (Figure 3C,D), with the `impRZilr` routine performing slightly better than lrSVD but showing markedly poorer results in terms of computing time.

4.1 | Effect of the number of components chosen

The simulations generated from the NMR data set were also used to assess the effect on the distortion measures, ADCS and CED, of different values for the parameter r. Recall that this determines the number of components or dimensions used for the low-rank approximation of the data in the lrSVD algorithm. A range between 1 and 20 was considered for

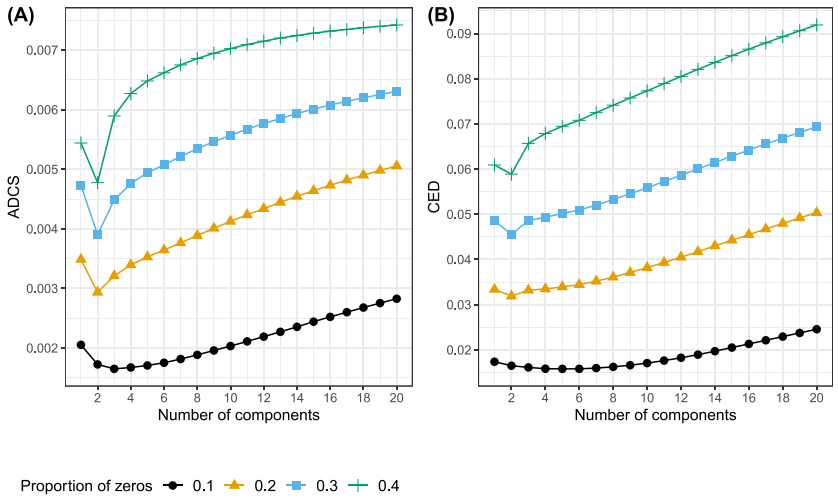

FIGURE 4 Effect of the number of components chosen in the lrSVD algorithm. Distortion (A) in covariance structure (ADCS) and (B) in compositional samples (CED)

illustration (out of up to $n-1=66$ potentially, because $n=67<D-1=126$). Figure 4 shows that, as expected, the distortion increases with the proportion of zeros for all values of r. The curves generally exhibit a dip around the actual number of components $r=2$ of the original complete data; however, it is more pronounced for higher proportions of zeros. This suggests that the decision about r becomes more relevant as the proportion of zeros increases. Note that the pattern is similar for both ADCS and CED; however, the relative impact on the covariance structure appears to be higher (ADCS; Figure 4A). After that point, the distortion steadily increases with the number of components, suggesting that overfitting (i.e., choosing an excessive number of components) may have more detrimental effects than falling short. In any case, for a given proportion of zeros, some departure from the *ideal* specification seems not to have dramatic consequences on the distortion introduced by the imputation procedure.

4.2 | Sensitivity to starting point

The lrSVD algorithm requires an initial imputation step to start the iterative process. Focusing on the rounded zeros problem, multiplicative simple replacement is used for this as detailed in Section 2, setting the initial imputed value around 2/3 ($\approx 65\%$) of the given DL as it is ordinarily recommended. Such criterion is based on previous studies,[16] and it is also justified by its approximation to the expected value of a triangular probability distribution between zero and the DL. We conduct here a sensitivity analysis to investigate the potential effect that changing this setup might have in the final lrSVD imputation. For this, zeros were forced in every second column by using the Q_p quantile as threshold, with $p=0.1$.[24] Given the input data matrix $\mathbf{T}$ with rounded zeros, we generated 100 uniformly random values in the interval $[0.05,0.95]$ that were used instead of the default 0.65 fraction to produce the initial imputation.

Figure 5A,B shows the results based on ADCS and CED, respectively. It can be observed that the results were very stable for both distortion measures when using values in a relatively broad and symmetric vicinity of the ordinary 0.65 fraction (indicated by a vertical dashed line, corresponding values ADCS = 0.021 and CED = 0.147 indicated with horizontal dashed lines). This suggests a negligible effect on the final lrSVD imputation across a sensible range of possible starting imputations. Moreover, these results also support the use of 0.65 as default (central) fraction in agreement with previous studies. Finally, these graphs show that, after the stable section around the default fraction is surpassed, distortion increases exponentially as the zero limit is approached.

4.3 | Simultaneous imputation of zeros and missing values

As a distinctive feature, the ability of the proposed lrSVD algorithm to deal simultaneously with zeros and missing data is illustrated in this section. We created a version of the original NMR data including 15% of zeros in each of a

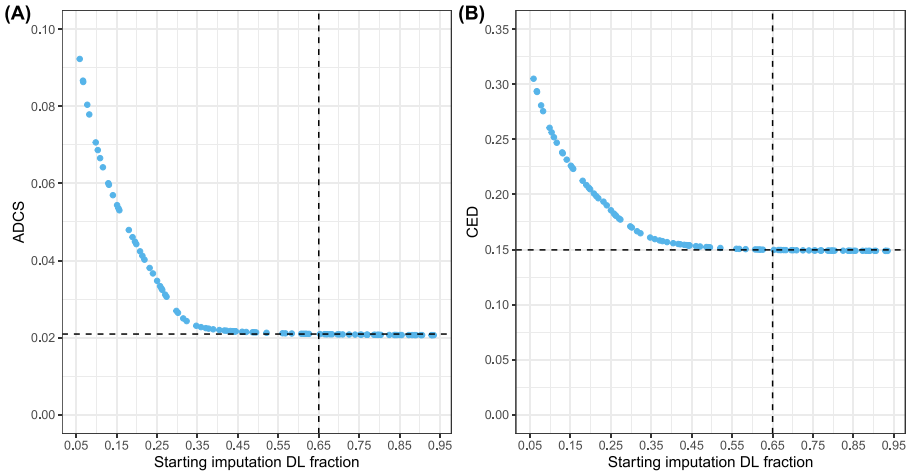

FIGURE 5 Analysis of sensitivity to starting multiplicative imputation based on the NMR data set using (A) ADCS and (B) CED distortion measures for 100 random fractions of the detection limit in [0.05,0.95]. The dashed lines indicate reference results using default $0.65 \cdot \mathbf{DL}$ imputation

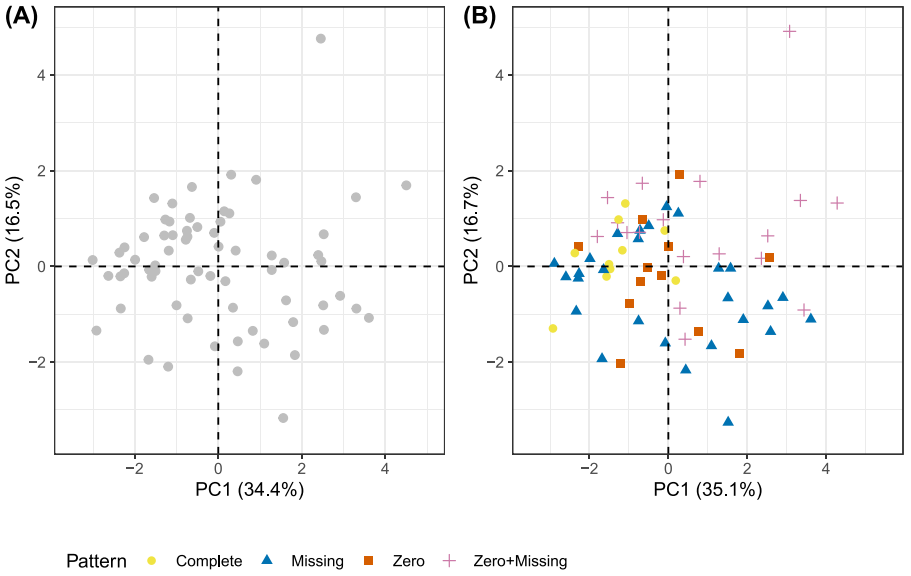

FIGURE 6 Imputation of zeros and missing values in the NMR data set using the lrSVD algorithm. Compositional PCA scores plot of (A) original fully observed data set and (B) imputed data set

random selection of 20% of the minor signals (generated using their $Q_{0.15}$ quantiles as detailed above) and 10% of missing values in each of a random selection of 20% of the major signals (missing completely at random). Original and imputed data sets are visualised in Figure 6 using the first two principal components (PC1 and PC2) from a compositional principal component analysis (PCA), that is, PCA applied on the data expressed in clr coefficients (note that only PCA scores displaying the samples and not loadings are shown to facilitate visual assessment). Different colours and symbols denote the patterns of samples being complete (fully observed) or including either zeros, missing values or both. It can be observed that the overall configuration of points and the fractions of total variance explained by the first two PCs as derived from the original (Figure 6A; 50.9% variance explained) and imputed (Figure 6B; 51.8% variance explained) data sets are highly comparable, and little distortion is introduced for any of the types of partly observed samples.

5 | FINAL REMARKS

Incomplete CoDa matrices are a common practical issue that hinders the representation of the relative information by means of log-ratio coordinates. The proposed lrSVD algorithm builds on compositional principles and recent developments in the area based on constrained low-rank approximations of data matrices. Although equally applicable to regular multivariate data sets, the approach is particularly relevant for the case of wide data sets as generated in modern analytical chemistry and molecular biology. In this context, procedures having a low computational cost and requiring minimal supervision are particularly valuable to be embedded into chemometric pipelines. The method shows robust results across a range of scenarios from low to high dimensions, from weak to strong association structures, and for varying amounts of observed values at a relatively low computational cost. In addition, as a distinctive feature, it deals with censored, missing or both types of nonobserved values simultaneously.

The lrSVD algorithm has been integrated into the open source R package *zCompositions*, which provides a common framework for imputation of CoDa sets through straightforward computer routines. In particular, the function `lrSVD` in *zCompositions* deals with zeros (or nondetects) or, alternatively, with missing values by simply toggling the argument `imp.missing` of the function to `TRUE`. Moreover, the function `lrSVDplus` tackles the case of zeros and missing values simultaneously. More details and illustrative examples are described in the help documentation accompanying the package.

As a note for future work, some types of data analyses might benefit from a multiple imputation approach to incorporate the uncertainty derived from the own imputation process into variability estimates, for example, standard errors in regression analysis. Another possible extension would be the use of robustified SVD methods to manage the potential effect of aberrant, outlying values in the data set.

ACKNOWLEDGEMENTS

This work was supported by Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and ERDF A way of making Europe (Grant PID2021-123833OB-I00) (to JP-A and JAM-F) and the French National Research Agency (ANR) (Grant ANR-17-EURE-0010) (to AR-G and CT-A).

DATA AVAILABILITY STATEMENT

The data underlying this article were provided by Scotland's Rural College by permission. Data will be shared on reasonable request to the corresponding author with permission of Scotland's Rural College.

ORCID

Javier Palarea-Albaladejo https://orcid.org/0000-0003-0162-669X
Josep Antoni Martín-Fernández https://orcid.org/0000-0003-2366-1592
Anne Ruiz-Gazen https://orcid.org/0000-0001-8970-8061
Christine Thomas-Agnan https://orcid.org/0000-0002-6430-3110

REFERENCES

1. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microb*. 2017;8:2224.
2. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*. 2018;34(16):2870-2878. doi:10.1093/bioinformatics/bty175
3. Calle ML. Statistical analysis of metagenomics data. *Genom Inform*. 2019;17:e6. doi:10.5808/GI.2019.17.1.e6
4. Monti GS, Filzmoser P. Sparse least trimmed squares regression with compositional covariates for high-dimensional data. *Bioinformatics*. 2021;37(21):3805-3814. doi:10.1093/bioinformatics/btab572
5. 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 Min*. 2021;14:315-330.
6. Filzmoser P, Hron K, Templ M. *Applied compositional data analysis. With worked examples in R*, Springer Series in Statistics: Springer; 2018.
7. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Modeling and analysis of compositional data*: John Wiley & Sons; 2015.
8. Barceló-Vidal C, Martín-Fernández JA. The mathematics of compositional analysis. *Austrian J Stat*. 2016;45(4):57-71.
9. Walach J, Filzmoser P, Hron K. Data normalization and scaling: consequences for the analysis in omics sciences. *Data Analysis for Omics Sciences: Methods and Applications*: Elsevier; 2018.
10. Harville DA. *Matrix Algebra From a Statistician's Perspective*: Springer-Verlag; 2008.

11. Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat Model*. 2015;15(2):134-158.
12. Racedo S, Portnoy I, Vélez JI, San-Juan-Vergara H, Sanjuan M, Zurek E. A new pipeline for structural characterization and classification of RNA-Seq microbiome data. *BioData Min*. 2021;14(1):31. doi:10.1186/s13040-021-00266-7
13. Baruzzo G, Patuzzi I, Di Camillo B. Beware to ignore the rare: how imputing zero-values can improve the quality of 16s rrna gene studies results. *BMC Bioinform*. 2022;22(15):618. doi:10.1186/s12859-022-04587-0
14. Palarea-Albaladejo J, Martn-Fernndez JA, Olea RA. A bootstrap estimation scheme for chemical compositional data with nondetects. *J Chemom*. 2014;28(7):585-599.
15. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometr Intell Lab Syst*. 2015;143:85-96.
16. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing data in compositional data sets using nonparametric imputation. *Math Geol*. 2003;35(3):253-278. doi:10.1023/A:1023866030544,
17. Palarea-Albaladejo J, Martín-Fernández JA. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Comput Geosci*. 2008;34(8):902-917.
18. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016;70(1):1-31.
19. Tatsukawa M, Tanaka M. Box constrained low-rank matrix approximation with missing values. In: Proceedings of the 7th International Conference on Operations Research and Enterprise Systems—Volume 1: ICORES. SciTePress; 2018.
20. Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika*. 1936;1:211-218.
21. Mert MC, Filzmoser P, Hron K. Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations. *Math Geosci*. 2016;48:941-961.
22. Peres-Neto PR, Jackson DA, Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal*. 2005;49(4):974-997.
23. Josse J, Husson F. Handling missing values in exploratory multivariate data analysis methods. *J Soc Fr Stat*. 2012;153(2):79-99.
24. Templ M, Hron K, Filzmoser P, Gardlo A. Imputation of rounded zeros for high-dimensional compositional data. *Chemometr Intell Lab Syst*. 2016;155(C):183-190.
25. Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V. Logratio analysis and compositional distance. *Math Geol*. 2000;32(3):271-275. doi:10.1023/A:1007529726302
26. Mateu Figueras G, Pawlowsky-Glahn V, Egozcue JJ. The normal distribution in some constrained sample spaces. *SORT*. 2013;37(1):29-56.
27. Audigier V, Husson F, Josse J. Multiple imputation for continuous variables using a Bayesian principal component analysis. *J Stat Comput Simul*. 2016;86(11):2140-2156. doi:10.1080/00949655.2015.1104683
28. Bica R, Palarea-Albaladejo J, Kew W, Uhrin D, Pacheco D, Macrae A, Dewhurst RJ. Nuclear magnetic resonance to detect rumen metabolites associated with enteric methane emissions from beef cattle. *Sci Rep*. 2020;10(1):5578. doi:10.1038/s41598-020-62485-y

How to cite this article: Palarea-Albaladejo J, Antoni Martín-Fernández J, Ruiz-Gazen A, Thomas-Agnan C. lrSVD: An efficient imputation algorithm for incomplete high-throughput compositional data. *Journal of Chemometrics*. 2022;e3459. doi:10.1002/cem.3459