Journal of Geochemical Exploration 253 (2023) 107299

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Cell-wise robust covariance estimation for compositions, with application to geochemical data

Christopher Rieser [a], Kamila Fačevicová [b], Peter Filzmoser [a,*]

[a] TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstrase 8-10, Vienna, Austria
[b] Palacký University Olomouc, Department of Mathematical Analysis and Applications of Mathematics, 17. listopadu 12, Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Cell-wise outliers
Covariance matrix
Geochemistry
Log-ratio analysis

ABSTRACT

Cell-wise outliers are outliers in single entries of a compositional data matrix, and they can lead to a certain bias in the statistical analysis. Traditional row-wise robust methods downweight outlying observations for the estimation, independent of how many or which cells of an observation are contaminated. Cell-wise robustness still makes use of the information contained in non-contaminated cells. Here, cell-wise robustness is used for the estimation of the variation and the covariance matrix. For higher dimensional data also a regularized estimator is introduced. The advantages of the cell-wise robust estimators are demonstrated in simulation experiments and in a geochemistry application in the context of clustering and principal component analysis.

1. Introduction

The field of compositional data analysis (CoDA) has received a lot of attention after the publication of the book (Aitchison, 1986) of John Aitchison in 1986. The approach followed there is often called log-ratio methodology, as the building blocks are pairwise log-ratios between the variables (compositional parts). If D denotes the number of parts, then there are $D(D-1)/2$ (relevant) pairwise log-ratios for the analysis, thus each part is involved in $D-1$ different log-ratios. These pairwise log-ratios are usually not used separately as new variables within a multivariate statistical analysis, but they are appropriately aggregated, e.g. to form a coordinate system with $D-1$ coordinates (Filzmoser et al., 2018). This, however, creates difficulties if an observation has outliers only in single parts, since their effect in the coordinate representation will very much depend on the type of aggregation, as well as on the parts which are contaminated.

Outliers in single variables of traditional non-compositional data are called cell-wise outliers. If an observation contains cell-wise outliers, the remaining non-contaminated cells usually still contain valuable information for the analysis. This can be very different for compositional data. Contamination in single parts of a composition can be caused by atypical values in this part, or it can result from an improper imputation of zeros or missing values. When forming pairwise log-ratios with contaminated cells, also the resulting log-ratio pairs can be outlying, and the same is true for other types of aggregation with contaminated parts.

Robust multivariate statistical methods commonly downweight complete observations if they are deviating, independent of how many cells (variables) of such observations are contaminated. Since observations are usually arranged in the rows of a data matrix, this procedure is often called row-wise robustness (Maronna et al., 2019). Especially for a growing number of variables, this would lead to a severe loss of valuable information if only few cells of an observation are deviating. Therefore, methods have been developed to identify and downweight deviating cells rather than only outlying rows, and this is called cell-wise robustness (Rousseeuw and Van Den Bossche, 2018).

Row-wise robustness is well established in the context of CoDA (Filzmoser et al., 2018). Since many multivariate methods are based on an estimated covariance matrix, a robust version can be obtained by expressing the composition in coordinates, and performing robust covariance estimation with any of the proposed methods (Maronna et al., 2019, p 195–224), for example with the widely used MCD (minimum covariance determinant) estimator (Rousseeuw, 1985; Rousseeuw and Van Driessen, 1999). Since a complete row is downweighted (or not), it does not matter which parts are contaminated, and which form of coordinate representation is used.

It is not straightforward to implement cell-wise robustness for the CoDA case, especially after the compositions are expressed in coordinates. Depending on the parts which are contaminated, as well as on the coordinate representation used, it might be hardly possible to trace back which compositional parts caused cell-wise outliers, detected in the

* Corresponding author.
E-mail address: peter.filzmoser@tuwien.ac.at (P. Filzmoser).

https://doi.org/10.1016/j.gexplo.2023.107299
Received 15 May 2023; Received in revised form 19 July 2023; Accepted 11 August 2023
Available online 24 August 2023

space of the coordinates. It is much more straightforward to investigate pairwise log-ratios in order to identify those parts which caused the outlyingness. Several attempts in this direction have already been done, e.g. in Štefelová et al. (2021) and Walach et al. (2020) in the context of regression and outlier detection.

In this paper we focus on cell-wise robustness, but limited to the problem of robust covariance estimation. Nevertheless, a robustly estimated covariance matrix opens the door to many multivariate methods, and in the application part we will focus on cell-wise robust principal components. The main idea is to estimate the elements of the variation matrix robustly, and these elements contain the estimated variances of pairwise log-ratios (Aitchison, 1986). Afterwards, we make use of the well-known relationship between the variation matrix and the compositional covariance matrix (Aitchison, 1986).

The paper is organized as follows. Section 2 recalls some of the CoDA concepts which are needed to introduce the cell-wise robust covariance estimators in Section 3. Simulation studies in Section 4 reveal robustness properties of the new proposals, and a comparison of the methods based on geochemical data is provided in Section 5. The final Section 6 concludes.

2. Some CoDA concepts

Denote by $\mathbb{R}^D_+$ the space of D dimensional strictly positive real values. In the classical approach of CoDA the D-part simplex is of central importance: it is defined as the space $\mathscr{S}^D = \left\{\boldsymbol{x} = (x_1, \ldots, x_D)^T \in \mathbb{R}^D_+ \,|\Sigma_{i=1}^D x_i = 1\right\}$. Considering also a composition $\boldsymbol{y} = (y_1, \ldots, y_D)^T$, the D-part simplex is equipped with the Aitchison inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathscr{A}} := \frac{1}{2D} \sum_{i,j=1}^{D} ln\left(\frac{x_i}{x_j}\right) ln\left(\frac{y_i}{x_j}\right)$$

and the so-called perturbation $\boldsymbol{x} \oplus \boldsymbol{y} := \frac{1}{\Sigma_{j=1}^D x_j y_j}(x_1 y_1, \cdots, x_D y_D)^T$ and powering $\alpha \odot \boldsymbol{x} := \frac{1}{\Sigma_{j=1}^D x_j^\alpha}\left(x_1^\alpha, \cdots, x_D^\alpha\right)^T$ operation, for any $\alpha \in \mathbb{R}$, see Pawlowsky-Glahn and Egozcue (2001) and Billheimer et al. (2001). With these definitions it can be shown that $\left(\mathscr{S}^D, \langle \cdot, \cdot \rangle_{\mathscr{A}}, \oplus, \odot\right)$ is a Hilbert space with neutral element $\frac{1}{D}(1, \ldots, 1)^T \in \mathbb{R}^D_+$ and norm $\|\boldsymbol{x}\|_{\mathscr{A}} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathscr{A}}}$, see Pawlowsky-Glahn et al. (2015), p.23–30. Of central importance in CoDA is the clr (centered log-ratio)-map

$$\mathrm{clr} : \mathscr{S}^D \to \mathbb{R}^D, \quad \mathrm{clr}(\boldsymbol{x}) := \left(ln\left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_j}}\right), \ldots, ln\left(\frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_j}}\right)\right)^T. \tag{1}$$

It is well known that the clr-map is distance preserving on $\mathscr{S}^D$, see Pawlowsky-Glahn et al. (2015), p 35. Additionally, the clr-map possesses the following properties,

$$\mathrm{clr}(\boldsymbol{x} \oplus \boldsymbol{y}) = \mathrm{clr}(\boldsymbol{x}) + \mathrm{clr}(\boldsymbol{y}) \tag{2}$$

$$\mathrm{clr}(\alpha \odot \boldsymbol{x}) = \alpha \mathrm{clr}(\boldsymbol{x}) \tag{3}$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathscr{A}} = \langle \mathrm{clr}(\boldsymbol{x}), \mathrm{clr}(\boldsymbol{y}) \rangle_2 \tag{4}$$

with $\langle \cdot, \cdot \rangle_2$ denoting the standard inner product in $\mathbb{R}^D$, and $\alpha \in \mathbb{R}$ (Pawlowsky-Glahn et al., 2015, p 35). A drawback of the clr-map is that it is not bijective onto $\mathbb{R}^D$ and its components are constrained to sum to zero. The clr-map can also be written as a linear map of the coordinate-wise log-transformed variables of $\boldsymbol{x}$, i.e. $\mathrm{clr}(\boldsymbol{x}) = \mathbf{L}_{\mathscr{A}} ln(\boldsymbol{x})$, where $\mathbf{L}_{\mathscr{A}}$ is the so-called centering matrix

$$\mathbf{L}_{\mathscr{A}} := \frac{1}{D}\begin{pmatrix} D-1 & -1 & & -1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & & -1 & D-1 \end{pmatrix}$$

which will play a role in the next section. Alternatively, the so-called ilr (isometric log-ratio)-map (Egozcue et al., 2003) is often considered as it preserves properties (2)–(4) and is bijective. It is defined as

$$\mathrm{ilr}_{\mathbf{V}} : \mathscr{S}^D \to \mathbb{R}^{D-1}, \quad \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x}) := \mathbf{V}^T \mathrm{clr}(\boldsymbol{x}) \ , \tag{5}$$

where $\mathbf{V} \in \mathbb{R}^{D \times (D-1)}$ is any matrix with orthogonal columns that span the $D-1$ dimensional subspace $\left\{z \in \mathbb{R}^D | \Sigma_{j=1}^D z_j = 0\right\} \subset \mathbb{R}^D$. The centering matrix has the following important properties that also connect it to any admissible $\mathbf{V}$ as just described,

$$\mathbf{L}_{\mathscr{A}} = \mathbf{L}_{\mathscr{A}}^T, \quad \mathbf{L}_{\mathscr{A}} \mathbf{1} = \mathbf{0}, \quad \mathbf{L}_{\mathscr{A}} = \mathbf{V}\mathbf{V}^T,$$

where $\mathbf{1}$ is a vector of ones.

There are multiple ways to characterize the dependence structure between different compositional parts of $\boldsymbol{x}$ in CoDA depending on the transformation used. Given two different random vectors $\boldsymbol{u}, \boldsymbol{v}$, then the covariance matrix is given as $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{v}) := \mathbb{E}((\boldsymbol{u} - \mathbb{E}(\boldsymbol{u}))(\boldsymbol{v} - \mathbb{E}(\boldsymbol{v}))^T)$, where $\mathbb{E}$ denotes the expectation. Then we write in the following $\Sigma^{clr} := \mathrm{Cov}(\mathrm{clr}(\boldsymbol{x}), \mathrm{clr}(\boldsymbol{x}))$ for the covariance matrix of $\mathrm{clr}(\boldsymbol{x})$, $\Sigma^{\ln} := \mathrm{Cov}(ln(\boldsymbol{x}), ln(\boldsymbol{x}))$ for the covariance matrix of $ln(\boldsymbol{x})$ and $\Sigma^{\mathrm{ilr}_{\mathbf{V}}} := \mathrm{Cov}(\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x}), \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x}))$ for the covariance matrix of $\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x})$, for any fixed $\mathbf{V}$. Additionally, we write $\mathbf{T} \in \mathbb{R}^{D \times D}$ for the variation matrix which is defined entry-wise for any $i, j = 1, \ldots, D$ as $T_{ij} = \mathrm{Cov}\left(ln\left(\frac{x_i}{x_j}\right), ln\left(\frac{x_i}{x_j}\right)\right) = \mathrm{Var}\left(ln\left(\frac{x_i}{x_j}\right)\right)$. Thus, the entries of the variation matrix are the variances of all pairwise log-ratios. Values near zero represent proportionality of the respective parts and consequently also their close relationship.

As noted in Aitchison (1986), there are the following relations between the covariance matrix Σ^{clr} and the variation matrix $\mathbf{T}$,

$$\mathbf{T} = \mathbf{1}\boldsymbol{\sigma}^{2T} + \boldsymbol{\sigma}^2 \mathbf{1}^T - 2\boldsymbol{\Sigma}^{\mathrm{clr}}, \tag{6}$$

$$\Sigma^{\mathrm{clr}} = -\frac{1}{2}\mathbf{L}_{\mathscr{A}} \mathbf{T} \mathbf{L}_{\mathscr{A}}, \tag{7}$$

where $\boldsymbol{\sigma}^2 = diag\left(\Sigma^{\mathrm{clr}}\right)$ denotes the vector of the diagonal elements of Σ^{clr}. These relationships will be very helpful for cell-wise robustness, as they link associations between pairs of compositional parts with covariances.

3. Cell-wise robust covariance estimation

Denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathscr{S}^D$ a sample of N independent observations which are arranged as rows in the data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. The classical sample covariance matrix estimate of Σ^{clr} is given as $\widehat{\Sigma}^{\mathrm{clr}} = \frac{1}{N-1}\Sigma_{n=1}^N \left(\mathrm{clr}(\boldsymbol{x}_n) - \widehat{\boldsymbol{\mu}}^{\mathrm{clr}}\right)\left(\mathrm{clr}(\boldsymbol{x}_n) - \widehat{\boldsymbol{\mu}}^{clr}\right)^T$, where $\widehat{\boldsymbol{\mu}}^{\mathrm{clr}} := \frac{1}{N}\Sigma_{n=1}^N \mathrm{clr}(\boldsymbol{x}_n)$. It is a well-known fact that any contaminated data point $\boldsymbol{x}_n$ can have a huge effect on the classical sample covariance estimate of Σ^{clr} (Maronna et al., 2019, p 196–198). The same holds true if only a single entry of $\boldsymbol{x}_n$, say the k-th entry, is contaminated. A part of a composition is understood as contaminated, when it has an unusually high (or low) absolute value and which therefore distorts the whole relative structure of the respective composition. As we can write any l-th entry of $\mathrm{clr}(\boldsymbol{x}_n)$ also in terms of pairwise log-ratios

$$\mathrm{clr}(\boldsymbol{x}_n)_l = ln\left(\frac{x_{nl}}{\sqrt[D]{\prod_{j=1}^D x_{nj}}}\right) = \frac{1}{D}\left(ln\frac{x_{nl}}{x_{n1}} + \ldots + ln\frac{x_{nl}}{x_{nk}} + \ldots + ln\frac{x_{nl}}{x_{nD}}\right),$$

we can see that a contamination in x_{nk} propagates to $\mathrm{clr}(\boldsymbol{x}_n)_l$, for any

$l \in \{1, \ldots, D\}$, which in consequence also can have an impact on $\widehat{\boldsymbol{\mu}}^{\text{clr}}$, see also Mert et al. (2016). Something similar also holds for the ilr-map. Therefore, if a row of $\mathbf{X}$ contains only a single cell that displays outlying behavior, taking the clr- or ilr-map of this row can lead to a whole outlying vector. The goal is to find a covariance estimator which protects against outlying cells in compositional data.

The main idea to achieve cell-wise robustness is thus to restrict the influence of cell-wise outliers at the log-ratio level. The relationship (7) is of particular interest as it permits to estimate Σ^{clr} in terms of the variation matrix, which consists solely of the variances at the log-ratio level. To construct a robust estimator of Σ^{clr}, or equivalently $\Sigma^{\text{ilr}_{\mathbf{V}}}$, we suggest to replace each entry of the variation matrix $\mathbf{T}$, $t_{ij} = \text{Var}\left(ln\left(\frac{x_i}{x_j}\right)\right)$, by a robust counterpart. This simplifies things not only because we focus on pairwise log-ratios, but also because robust variance estimation is a simpler task compared to robust covariance estimation.

Several robust estimators of the variance have been proposed in the literature. A highly robust and efficient estimator of scale (standard deviation) is the Qn estimator (Rousseeuw and Croux, 1993), which is essentially defined as the first quartile of the absolute pairwise differences. To be more precise, for a vector $z = (z_1, \ldots, z_N)^T$, the Qn estimator of scale is defined as

$$\rho(z) = \rho_{n=1,\ldots,N}(z_n) = 2.219 \cdot \left\{ |z_i - z_j| ; i < j \right\}_{(k)}, \tag{8}$$

where $k = \binom{h}{2} \approx \binom{N}{2} \Big/ 4$ and $h = \lfloor N/2 \rfloor + 1$, and (k) is the k-th value of the sorted absolute pairwise differences (in ascending order) (Rousseeuw and Croux, 1993). A further advantage of this estimator is that it does not require a (robust) location estimation. These properties make the Qn estimator preferable over other robust scale estimators, such as the more well-known median absolute deviation (MAD).

3.1. The cell-wise robust Qn covariance estimator

We will use the Qn estimator for the elements of the variation matrix to obtain $t_{ij}^{\rho} = \rho_{n=1,\ldots,N}\left(ln\left(\frac{x_{ni}}{x_{nj}}\right)\right)^2$. We denote the resulting robust variation matrix by $\widehat{\mathbf{T}}_{\rho} := t_{ij}^{\rho}$. To construct a valid estimate of Σ^{clr} we then transform $\widehat{\mathbf{T}}_{\rho}$ according to Eq. (7). This estimate should however fulfill the following three properties that Σ^{clr} has.

- Σ^{clr} is a positive semi-definite matrix.
- Σ^{clr} has an eigenvector $\mathbf{1}$ to the zero eigenvalue, i.e. $\Sigma^{\text{clr}}\mathbf{1} = \mathbf{0}$.
- Σ^{clr} is invariant under multiplication with $\mathbf{L}_{\mathscr{A}}$, thus $\Sigma^{\text{clr}} = \mathbf{L}_{\mathscr{A}}\Sigma^{\text{clr}}\mathbf{L}_{\mathscr{A}}$.

Note that the second point is a direct consequence of the third point, and the third point directly follows from the idempotency of $\mathbf{L}_{\mathscr{A}}$ and the representation of the clr-map through the latter. Therefore, to construct a positive semi-definite estimator we propose to take the nearest positive semidefinite matrix to $-\frac{1}{2}\mathbf{L}_{\mathscr{A}}\widehat{\mathbf{T}}_{\rho}\mathbf{L}_{\mathscr{A}}$ and adjust the median quantile. All together we perform the following steps:

1. Estimate the elements t_{ij}^{ρ} of $\widehat{\mathbf{T}}_{\rho}$ by $t_{ij}^{\rho} = \rho_{n=1,\ldots,N}\left(ln\left(\frac{x_{ni}}{x_{nj}}\right)\right)^2$, where $i \neq j$ (otherwise, the entries are zero).
2. Project $\widehat{\mathbf{T}}_{\rho}$ on the appropriate subspace by $\widehat{\Sigma}_0^{\text{clr}} := -\frac{1}{2}\mathbf{L}_{\mathscr{A}}\widehat{\mathbf{T}}_{\rho}\mathbf{L}_{\mathscr{A}}$.
3. Find the nearest positive semi-definite matrix (p.s.d) $\widehat{\Sigma}_1^{\text{clr}}$ by solving $min_{\widehat{\Sigma}_1^{\text{clr}} \text{isp.s.d}} \left\| \widehat{\Sigma}_0^{\text{clr}} - \widehat{\Sigma}_1^{\text{clr}} \right\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. Note that also other measures could be considered that take into account the symmetry of the matrix.
4. The final estimator is $\widehat{\Sigma}^{\text{clr}} = c\widehat{\Sigma}_1^{\text{clr}}$ where c is chosen such that the median of the squared Mahalanobis distances is equal to the quantile 0.5 of the χ^2 distribution with degrees of freedom equal to the rank of $\widehat{\Sigma}_1^{\text{clr}}$. This alignment allows to use the quantile 0.975 of this χ^2 distribution as outlier cutoff value for the squared Mahalanobis distances (Maronna et al., 2019, p 206).

For these steps to lead to an admissible estimate of $\widehat{\Sigma}^{\text{clr}}$ we need to check that the three properties as described above hold. Clearly $\widehat{\Sigma}_1^{\text{clr}}$ is positive semi-definite by definition and so we only need to check the latter two properties. As $\mathbf{L}_{\mathscr{A}}\mathbf{1} = \mathbf{0}$ holds, we can follow directly $\widehat{\Sigma}_0^{\text{clr}}\mathbf{1} = \mathbf{0}$. Step 3. does not affect the eigenvectors, as the solution to the nearest positive semi-definite matrix problem is equal to a zero thresholding of the eigenvalues, see Higham (1988). Therefore, $\widehat{\Sigma}_1^{\text{clr}}\mathbf{1} = \mathbf{0}$ holds, and consequently the same is true for $\widehat{\Sigma}^{\text{clr}}$. The third property directly follows from the second one by $\mathbf{L}_{\mathscr{A}}\widehat{\Sigma}^{\text{clr}}\mathbf{L}_{\mathscr{A}} = \left(\mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^T\right)\widehat{\Sigma}^{\text{clr}}\left(\mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^T\right) = \widehat{\Sigma}^{\text{clr}} - \frac{1}{D}\widehat{\Sigma}^{\text{clr}}\mathbf{1}\mathbf{1}^T - \frac{1}{D}\mathbf{1}\mathbf{1}^T\widehat{\Sigma}^{\text{clr}} + \frac{1}{D^2}\mathbf{1}\mathbf{1}^T\widehat{\Sigma}^{\text{clr}}\mathbf{1}\mathbf{1}^T = \widehat{\Sigma}^{\text{clr}}$.

In the following, we will refer to $\widehat{\Sigma}^{\text{clr}}$ as the cell-wise robust Qn estimator, abbreviated simply as "Qn".

3.2. The cell-wise robust regularized Qn covariance estimator

In the case where $N \leq D$, it is known in the non-compositional data setting, that non-robust as well as robust covariance estimates display an increased estimation error, and that estimates can be approved upon by adding a shrinkage term, often in form of a unit matrix, see Ledoit and Wolf (2004); Chen et al. (2011) and Hubert et al. (2018). Note, however, that also other shrinkage matrices can be used. Inspired by these ideas we also propose a shrinkage version. We first perform steps 1. to 3. as described above obtaining $\widehat{\Sigma}_1^{\text{clr}}$. Instead of continuing to step 4., we now define $\widehat{\Sigma}_2^{\text{clr}} := (1-\alpha)\widehat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{L}_{\mathscr{A}}$, where α is a fixed chosen constant, see below. Lastly, we obtain the final estimator as $\widehat{\Sigma}^{\text{clr}} := c\widehat{\Sigma}_2^{\text{clr}}$, where c is chosen such that the median of the squared Mahalanobis distances is equal to the quantile 0.5 of the χ^2 distribution with degrees of freedom equal to the rank of $\widehat{\Sigma}_2^{\text{clr}}$. Note that equivalently we could also compute instead step 1. to 2., replace $\widehat{\Sigma}_0^{\text{clr}}$ by a shrunk version $(1-\alpha)\widehat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{I}$ and then continue with step 3. to 4., with $(1-\alpha)\widehat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{I}$ instead of $\widehat{\Sigma}_1^{\text{clr}}$. The motivation for adding $\mathbf{L}_{\mathscr{A}}$ instead of the unit matrix is that the latter does not fulfill the properties of an admissible covariance estimator. $\mathbf{L}_{\mathscr{A}}$ however trivially does. One can also argue that $\mathbf{L}_{\mathscr{A}}$ can be considered the compositional pendant to the unit matrix. This can easily be seen from the fact that if a compositional variable $\mathbf{x}$ is standard Gaussian distributed, $\text{ilr}_{\mathbf{V}}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then the covariance of $ln(\mathbf{x})$ is given by $\mathbf{V}^T\mathbf{V} = \mathbf{L}_{\mathscr{A}}$ as $\text{ilr}_{\mathbf{V}}(\mathbf{x})^T\text{ilr}_{\mathbf{V}}(\mathbf{x}) = ln(\mathbf{x})\mathbf{V}\mathbf{V}^T ln(\mathbf{x})$ and $\mathbf{V}\mathbf{V}^T = (\mathbf{V}\mathbf{V}^T)^+$, where $+$ is the Moore-Penrose generalized inverse. This shrinkage step introduces an additional tuning parameter α. In the non-compositional literature, optimal choices of α are well-known. For an unbiased estimate $\widehat{\Sigma}$ of Σ, the optimal choice of α such that the expected Frobenius distance between Σ and $(1-\alpha)\widehat{\Sigma} + \alpha\mathbf{\Gamma}$, i.e. $\mathbb{E}\left(\|\Sigma - ((1-\alpha)\widehat{\Sigma} + \alpha\mathbf{\Gamma})\|_F^2\right)$, is minimized, with $\mathbf{\Gamma}$ being a shrinkage target matrix, is given by

$$\alpha = \frac{\sum_{i,j=1}^{D}\left(\text{Var}\left(\widehat{\Sigma}_{ij}\right) - \text{Cov}\left(\Gamma_{ij}, \widehat{\Sigma}_{ij}\right)\right)}{\sum_{i,j=1}^{D}\mathbb{E}\left(\left(\Gamma_{ij} - \widehat{\Sigma}_{ij}\right)^2\right)},$$

see Schäfer and Strimmer (2005), where subscripts ij denote the corresponding matrix entries. Taking $\mathbf{\Gamma} = \mathbf{L}_{\mathscr{A}}$ we can drop $\text{Cov}\left(\Gamma_{ij}, \widehat{\Sigma}_{ij}\right)$ due to

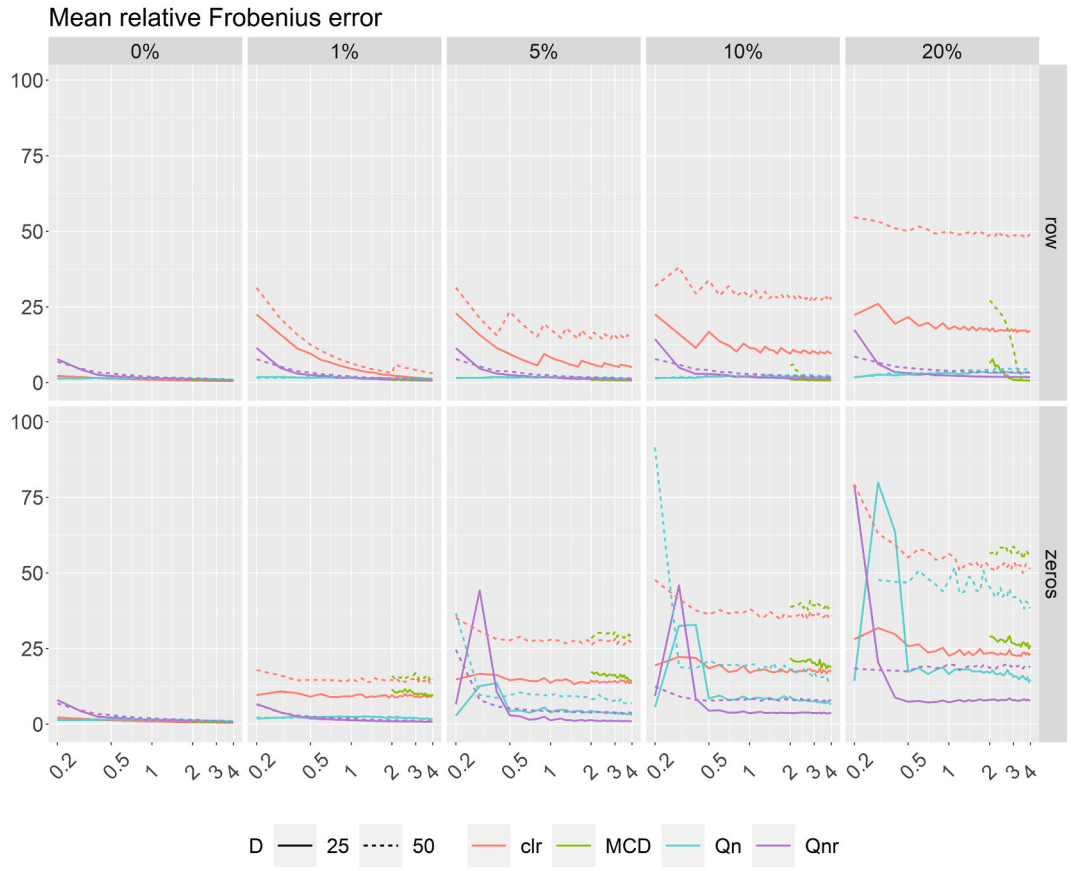

Fig. 1. Results of the simulation study in terms of the mean relative Frobenius error. The classical estimate based on the clr representation (clr) and row-wise robust estimate (MCD) are compared with two newly proposed cell-wise robust estimates (Qn and Qnr). The column blocks correspond to different levels of corruption and the horizontal axis is given by the value of $k = N/D$, with D being number of compositional parts.

$\mathbf{\Gamma}$ having constant expectation. With this choice of $\mathbf{\Gamma}$ and taking $\widehat{\Sigma} = \widehat{\Sigma}_1^{\text{clr}}$ (here, for simplicity, $\widehat{\Sigma}_1^{\text{clr}}$ denotes the estimator in terms of random variables), the expectation in the denominator becomes $\Sigma_{i,j=1}^{D}\mathbb{E}\left(\left(\Gamma_{ij} - \left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ij}\right)^2\right) = \Sigma_{i\neq j}\mathbb{E}\left(\left(-\frac{1}{D} - \left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ij}\right)^2\right) + \Sigma_{i=1}^{D}\mathbb{E}\left(\left(\left(1-\frac{1}{D}\right) - \left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ii}\right)^2\right)$, leading to

$$\alpha = \frac{\Sigma_{i,j=1}^{D}\text{Var}\left(\left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ij}\right)}{\Sigma\Sigma_{i\neq j}\mathbb{E}\left(\left(\left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ij} + \frac{1}{D}\right)^2\right) + \Sigma_{i=1}^{D}\mathbb{E}\left(\left(\left(\widehat{\Sigma}_1^{\text{clr}}\right)_{ii} + \frac{1}{D} - 1\right)^2\right)}.$$

Recently, also in the non-robust compositional setting, such shrinkage estimators have been proposed (Jin et al., 2022). As an estimate of α requires knowledge of the variance of terms involving $\widehat{\Sigma}_1^{\text{clr}}$, and as these might be error-prone due to cell-wise outliers in the data, we propose to replace the variance and expectation by robust counterparts, and to estimate α by

$$\alpha = \frac{\Sigma_{i,j=1}^{D}\left(\rho_{k=1,\dots,K}\left(\left(\widehat{\Sigma}_1^{\text{clr},k}\right)_{ij}\right)\right)^2}{\Sigma_{i\neq j}\text{med}_{k=1,\dots,K}\left(\left(\left(\widehat{\Sigma}_1^{\text{clr},k}\right)_{ij} + \frac{1}{D}\right)^2\right) + \Sigma_{i=1}^{D}\text{med}_{k=1,\dots,K}\left(\left(\left(\widehat{\Sigma}_1^{\text{clr},k}\right)_{ii} + \frac{1}{D} - 1\right)^2\right)},$$

where med denotes the median, and $\widehat{\Sigma}_1^{\text{clr},k}$ is the k-th estimate, for $k \in \{1,\dots,K\}$, and K is the number of bootstrap samples of size N (with replacement), see Wilcox (2010) and Efron and Tibshirani (1994), p 87–108. More precisely, we sample with replacement N times full observations, not cell-wise, from the observed samples $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ and calculate $\widehat{\Sigma}_1^{\text{clr}}$ by the proposed algorithm, which we denote instead by $\widehat{\Sigma}_1^{\text{clr},k}$, where k denotes the number of bootstrap samples we draw; in our computations we set $K = 500$.

We will refer to the resulting estimator as the cell-wise robust regularized Qn estimator, abbreviated simply as "Qnr".

4. Simulation study

In this section we conduct a simulation study to compare the performance of the proposed estimators in comparison with other existing non-robust and row-wise robust versions. For different dimensions $D \in \{25, 50\}$ and different numbers of samples $N = Dk$ with $k \in \{0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 3, 4\}$, we simulate a random covariance matrix Σ^{clr} and a mean vector $\boldsymbol{\mu}^{\text{clr}} \in \mathbb{R}^D$ from a standard Gaussian distribution. More precisely, we generate a $D-1$ dimensional correlation matrix Σ as in Agostinelli et al. (2015), with code adapted from the R cellWise package (Raymaekers, 2022), and turn it into an admissible covariance matrix in clr coordinates by $\Sigma^{\text{clr}} = \mathbf{V}\Sigma\mathbf{V}^T$, where $\mathbf{V}$ is the same matrix as for an ilr coordinate function. Note that our setup differs in that we include an additional step to turn these into admissible covariance matrices in clr coordinates. Following this, we generate an $N \times D$ data matrix $\mathbf{X}$ such that the log-transformation of each row $\mathrm{x}_n \in \mathscr{S}^D$ is singular normally distributed $ln(\boldsymbol{x}_n) \sim \mathcal{N}\left(\boldsymbol{\mu}^{\text{clr}}, \Sigma^{\text{clr}}\right)$. To corrupt the data, two different setups are used, which we denote in the following by "row-wise" and "zero" corruption.

- Row-wise corruption: A given percentage of observations is randomly selected and replaced by observations that are sampled as

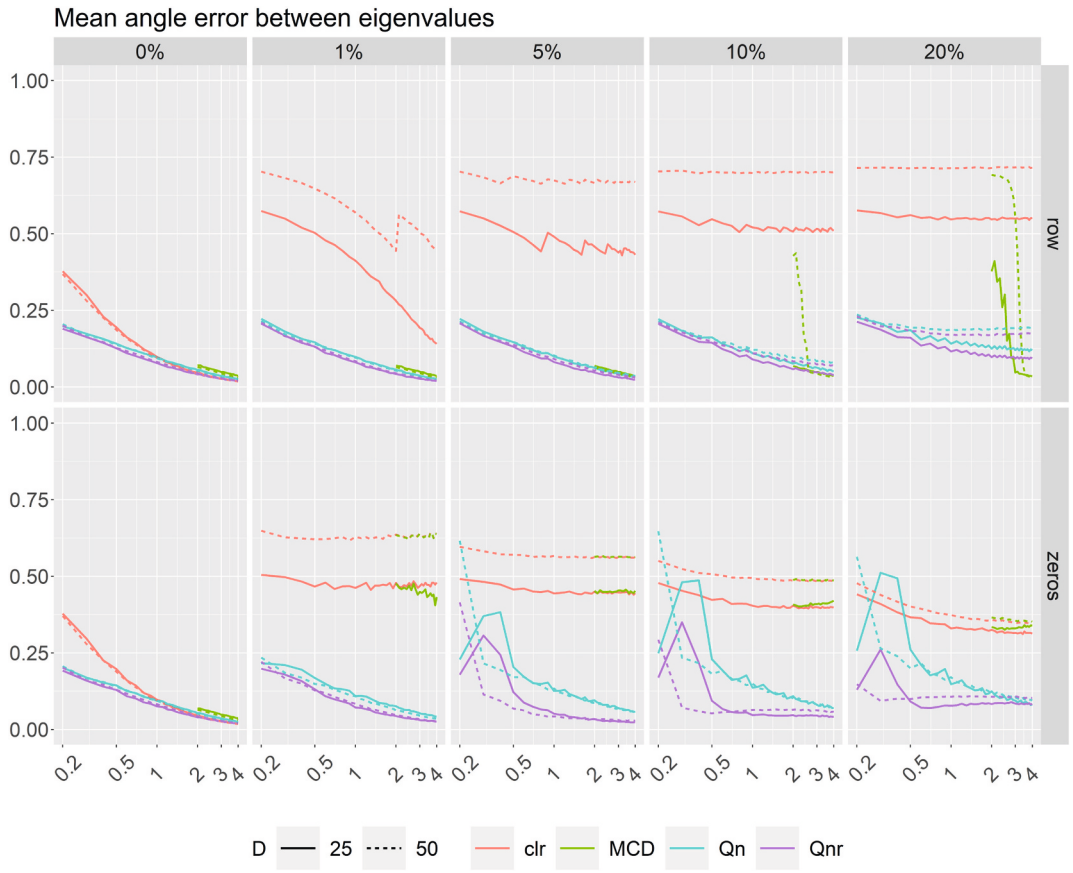

Fig. 2. Results of the simulation study in terms of the mean angle error between eigenvalues. The classical estimate based on the clr representation (clr) and row-wise robust estimate (MCD) are compared with two newly proposed cell-wise robust estimates (Qn and Qnr). The column blocks correspond to different levels of corruption and the horizontal axis is given by the value of $k = N/D$, with D being number of compositional parts.

above, but from another randomly generated covariance matrix, and a different mean sampled from a standard Gaussian distribution.

- Zero corruption: We randomly select a certain percentage of the columns of the data matrix $\mathbf{X}$ and set for each selected column all entries below the quantile 0.25 to zero. Finally, in accordance with the common CoDA practice, we replace the zeros by 0.5 times the smallest non-zero entry of the corresponding values of the column.

Thus, with row-wise corruption we would expect that the new estimators show similar performance as a row-wise robust estimator. The setting with zero corruption mimics a typical scenario observed with geochemical data, where a proportion of samples falls below a detection limit for specific variables. Here we should see the benefits of a cell-wise robust method.

Each simulation is repeated 500 times, and we will report averages over these replications. To evaluate the performance of the methods we use mean values of two different error measures between an estimated Σ^{clr} and the true $\widehat{\Sigma}^{\text{clr}}$. These measures are

- Relative Frobenius error: $\frac{\left\|\widehat{\Sigma}^{\text{clr}} - \Sigma^{\text{clr}}\right\|_F}{\left\|\Sigma^{\text{clr}}\right\|_F}$
- Angle error between eigenvalues: $1 - \frac{\widehat{\mathbf{a}}^T \mathbf{a}}{\sqrt{\widehat{\mathbf{a}}^T \widehat{\mathbf{a}}}\sqrt{\mathbf{a}^T \mathbf{a}}}$,

where $\widehat{\mathbf{a}}$ resp. $\mathbf{a}$ are the vectors of eigenvalues of $\widehat{\Sigma}^{\text{clr}}$ resp. Σ^{clr}. Big values of the relative Frobenius error indicate difficulties in the estimation of the elements of the covariance matrix. The angle error between the eigenvalues is in the interval $[0, 1]$, and a big value means that the shape of the covariance matrix is not appropriately estimated. From this point of view, the two measures complement each other to a certain extent.

Figs. 1 and 2 compare the considered estimators in terms of mean relative Frobenius error and mean angle error between eigenvalues, respectively. The top panels refer to row-wise corruption, and the bottom panels to zero corruption. In the panels from left to right we increase the amount of contamination. The horizontal axes show the ratio $k = N/D$, and thus in the left part we have a higher-dimensional low-sample size situation, in the right part "classical" tall data matrices with $N > D$. Note that some repetitions (less than 1 %) of the simulation resulted in extremely high values of the relative Frobenius error for the Qn estimator, which can particularly happen in the setting with zeros, leading to degenerate solutions for robust scale estimation. These failures were removed from the following considerations. Based on the respective plots, the performance of the estimators differs between scenarios with and without contamination. When no contamination is present (left column) and $D \leq N$, the classical estimate, denoted in the plots as clr, performs well. Moreover, the mean values of both errors considered are similar to errors obtained for the Qn and Qnr estimators. Slightly worse but still comparable values of the mean angle error are obtained for the MCD estimate. Even in the case of no contamination, but for $D > N$, we can already see an advantage of Qn and Qnr over clr; in this case, the MCD does not give a solution since it is ill-defined for such a situation.

When contamination increases, the classical estimate is naturally distorted by the presence of outliers. For both corruption types, this effect gets more serious with increasing number of variables D and, on the contrary, is quite stable with respect to the N to D ratio. The performance of the other considered estimators depends on the corruption type. The use of the MCD estimator is appropriate for the case of row-wise contamination. The upper panel of Fig. 2 clearly shows that

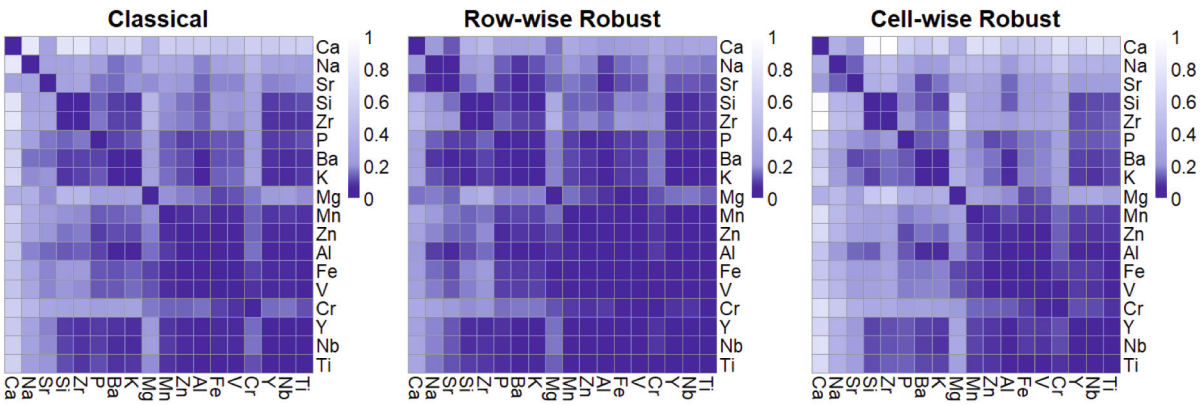

Fig. 3. Variation matrices scaled by the maximum of the entries over all three methods.

particularly with higher percentage of outlying rows, the MCD estimator results in the lowest mean angle error between eigenvalues (if N is sufficiently large). On the other hand, for a lower percentage of the outlying rows, both the Qn estimator and its regularized version Qnr give comparable results as MCD, but these two estimators are also able to cope with settings where D is close or even bigger than N.

The lower panels of Figs. 1 and 2 mimic the scenario of cell-wise outliers, where downweighting of the whole observations may yield too much information loss. The MCD estimator therefore results in a very high error, which is in term of the mean relative Frobenius distance even worse than the one of the classical estimate. The Qn estimator copes with the cell-wise outliers significantly better. This estimator brings valuable results for both high- and low-dimensional settings and its performance enhances with increasing number of observations. The Qn estimate can be refined with its regularized version Qnr, which, particularly under the high-dimensional setting, results in a very low error rate. For all levels of corruption considered and the N to D ratio above 0.5, the performance of the Qnr estimator is comparable to the situation where there is no corruption in the data.

5. Application to a dataset from geochemistry

The performance of the proposed covariance estimators will be demonstrated on a dataset which originates from the large-scale geochemical mapping project GEMAS. This dataset, which is freely available in the R package robCompositions (Templ et al., 2011), consists of the concentration of chemical elements, measured at 2108 locations in most countries of Europe, see Reimann et al. (2014) for details. Here we consider elements measured by XRF (as total element concentrations), and they should provide a close link to the lithology. Elements with more than 3 % of values below the detection limit were excluded, and thus we considered the 18 elements Al, Ba, Ca, Cr, Fe, K, Mg, Mn, Na, Nb, P, Si, Sr, Ti, V, Y, Zn, and Zr. Our goal here is to compare different versions of estimating the variation matrix and the corresponding covariance matrix, and their effect on Q-mode clustering and on principal component analysis (PCA). All computations were done in the software environment R (R Core Team, 2023).

It is unknown if this dataset contains outliers (row-wise or cell-wise) or not. However, according to the results from the simulations we might not have to expect a huge difference between the row-wise and cell-wise robust covariance estimations.

5.1. Estimates of covariance and variation matrices

As a first step, we compare the estimated covariance and variation matrices:

- *Classical:* Σ^{clr} is estimated by the sample covariance matrix of the clr-transformed data. The estimated variation matrix is obtained by the relationship given in Eq. (6).
- *Row-wise robust:* The data are first expressed in (any) ilr coordinates. Then the MCD estimator is used to obtain a row-wise robust covariance estimate, which is then transformed to the clr representation. The row-wise robust variation matrix is again obtained through Eq. (6).
- *Cell-wise robust:* The cell-wise robust variation and covariance matrices are computed as explained in Section 3.

Note that here we do not compare with a robust covariance based on shrinkage, as introduced in Section 3, because the estimated value of α is essentially zero, implying that no shrinkage would be done, and the number of observations is big enough in comparison to the number of measured elements.

The resulting variation and covariance matrices are visualized as

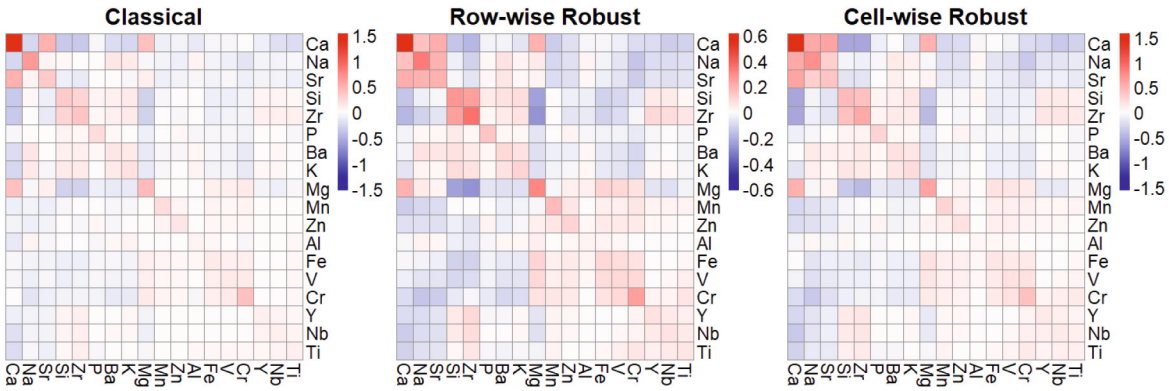

Fig. 4. Clr covariance matrices estimated by the different methods.

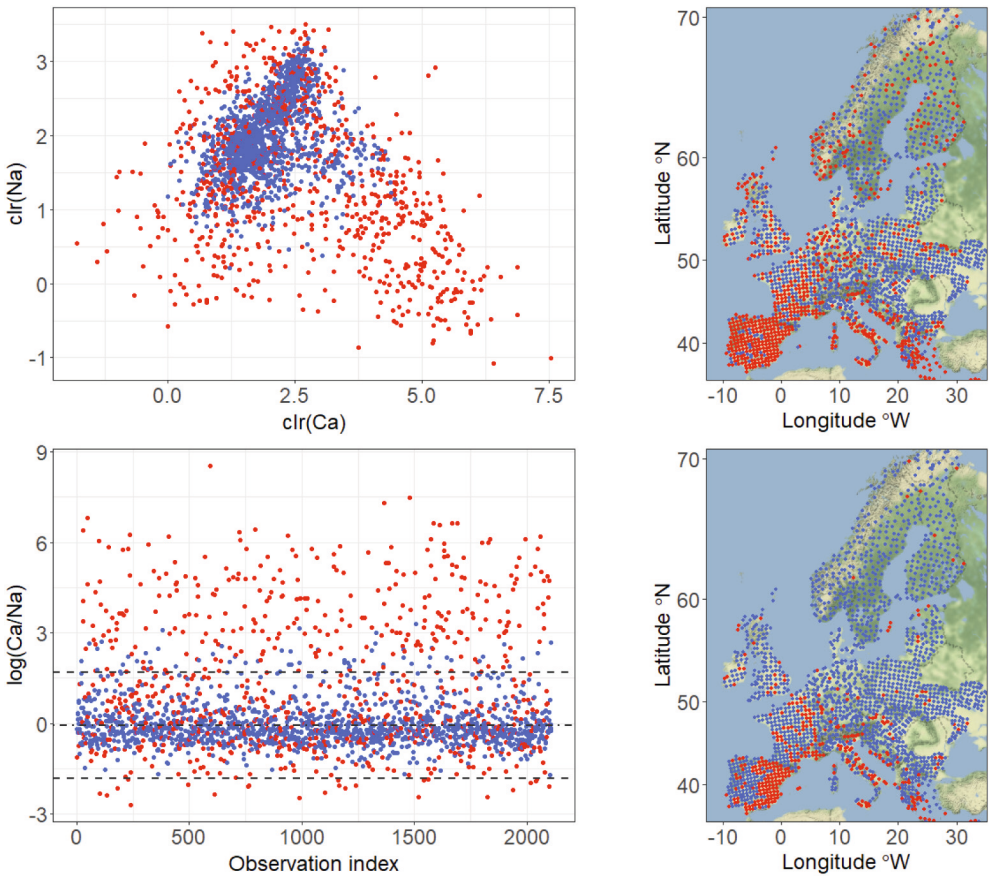

Fig. 5. Top left: clr-transformed Ca versus Na, with row-wise outliers in red; classical covariance is negative, robust one is positive. Top right: row-wise outliers in red in the map. Bottom left: log-ratio Ca/Na versus index, with extremes indicated by values outside median plus/minus 2 times Qn, with color as above. Bottom right: extremes detected according to value of the Ca to Na log-ratio in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

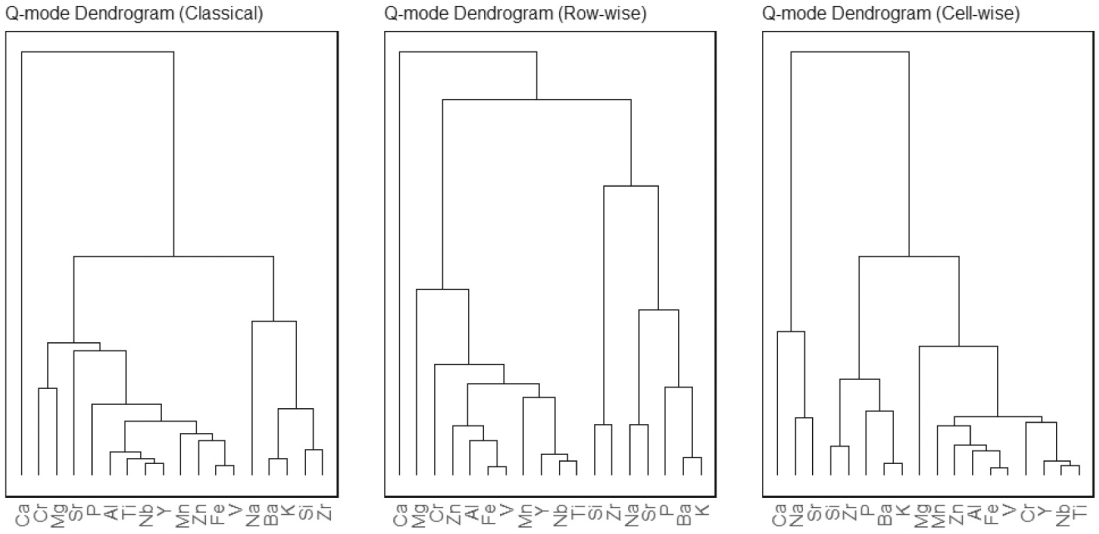

Fig. 6. Dendrograms resulting from the Q-mode clustering, based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimate of the variation matrix.

heatmaps in Figs. 3 and 4. Note that the elements have been reordered by the order given from the cell-wise robust dendrogram for the Q-mode clustering, see Section 5.2, which might simplify the comparison. The estimated variation matrices, visualized in Fig. 3 with a common normalized scale, reveal that Ca is to some extent different from most of the other elements, and to a lesser extent also Mg and Cr.

The most obvious difference in the covariances in Fig. 4 is the negative relationship between clr representations of Ca and Na in the classical version, while the robust estimates indicate a positive covariance. This phenomenon is explained visually in Fig. 5. The upper left panel shows the clr-transformed variables Ca and Na, and the red color refers to observations which are identified as outliers with the row-wise robust MCD estimator. Indeed, the two variables show a negative trend if the outliers are included, and a positive if they are downweighted. The upper right panel shows the color information at the sample locations in the map. As an example, a row-wise robust analysis would downweight

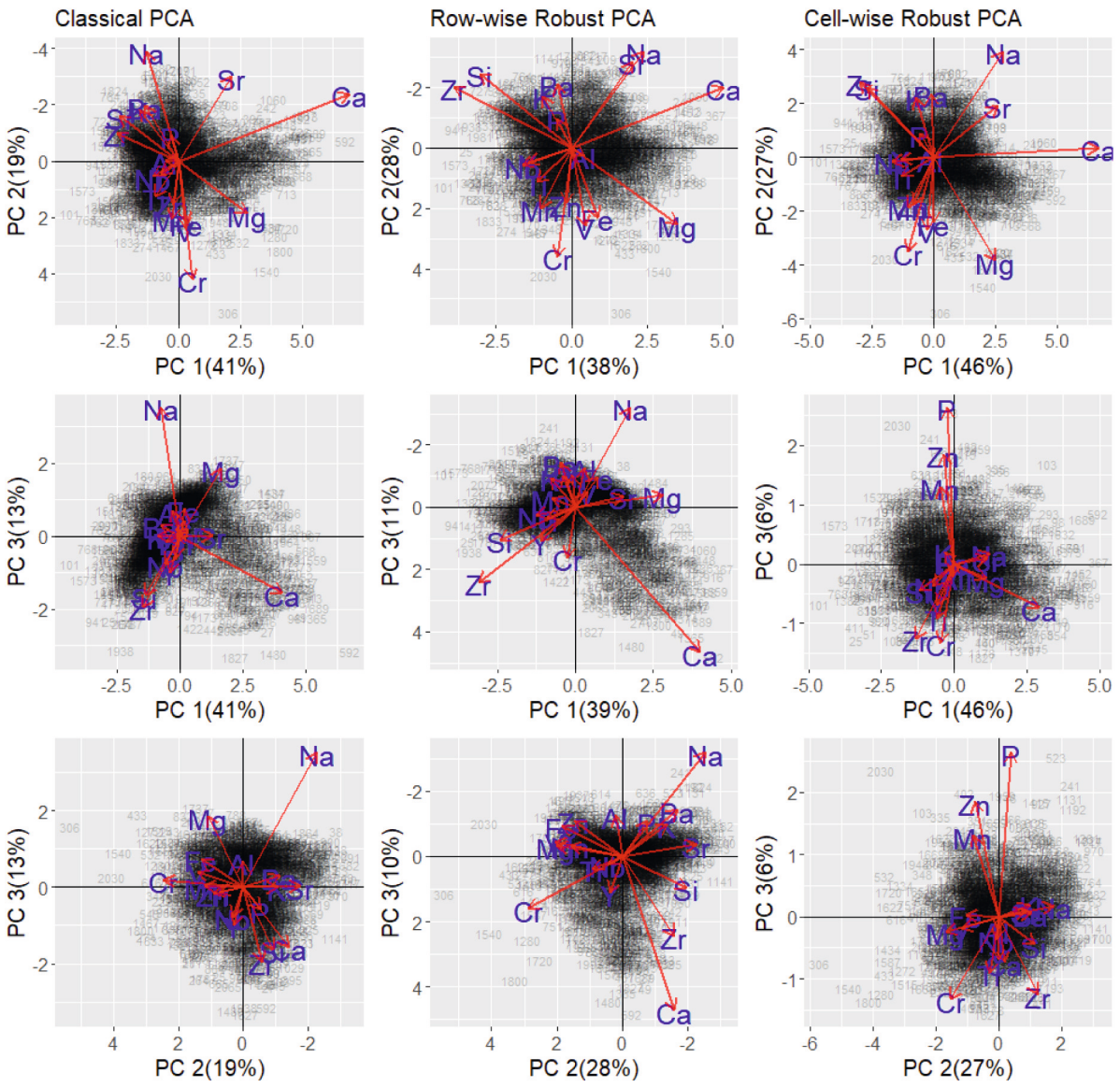

Fig. 7. Biplot of the PCA results based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix.

almost all samples from Spain. The bottom plots try to show the differences to a cell-wise treatment. The lower left panel shows the log-ratio Ca versus Na, for which the Qn estimator is used. The horizontal lines indicate median (dashed), and median ± 2 times the Qn estimator of this log-ratio (dot-dashed). Thus, values outside the solid lines are "extremes" of this log-ratio. The color information is the same as before, and thus we can see that row-wise outliers are not only found in the extremes but in the entire range. The map at the bottom right has as color information the extremes in red and "normal" values of the log-ratio in blue, and it clearly differs from the coloring in the map above. Here, Ca-rich sediments, e.g. in the eastern and southern part of Spain are highlighted and thus downweighted in a cell-wise robust analysis. The cell-wise robust method thus tries to incorporate information of the trend visible in the data majority, but acts pairwise, and thus incorporates also other relevant information of a particular log-ratio.

5.2. Q-mode clustering

Q-mode clustering generally performs a grouping of the variables. In the CoDA context, the basis for Q-mode clustering is the estimated variation matrix, see Fig. 3, and usually Ward's hierarchical clustering is used (Filzmoser et al., 2018, p 113). The results are presented as dendrograms in Fig. 6. As already seen in Fig. 3, Ca makes the most important difference in the three versions: while Ca joins as a single element at the very last stage for the classical and the row-wise version, it is in a cluster with Na and Sr in the cell-wise version. From a geochemical point of view, this cluster is meaningful, and we will discuss these and other relationships in more detail for the PCA results.

5.3. Principal component analysis

The estimated covariance matrix is the basis for computing the principal components. Accordingly, differences identified in Fig. 4 are also supposed to be visible in the different versions of PCA (classical, row-wise and cell-wise robust). The corresponding results are shown as biplots in Fig. 7, and the PCA scores are plotted in the maps in Figs. 8–10. The results of the row-wise robust version can also be compared to the results in Reimann et al. (2014), where almost the same elements have been used for row-wise robust PCA.

When comparing the different versions of PCA, we can immediately see a bigger difference from classical PCA to the robust counterparts, which obviously is based on the effect of outliers on the covariance

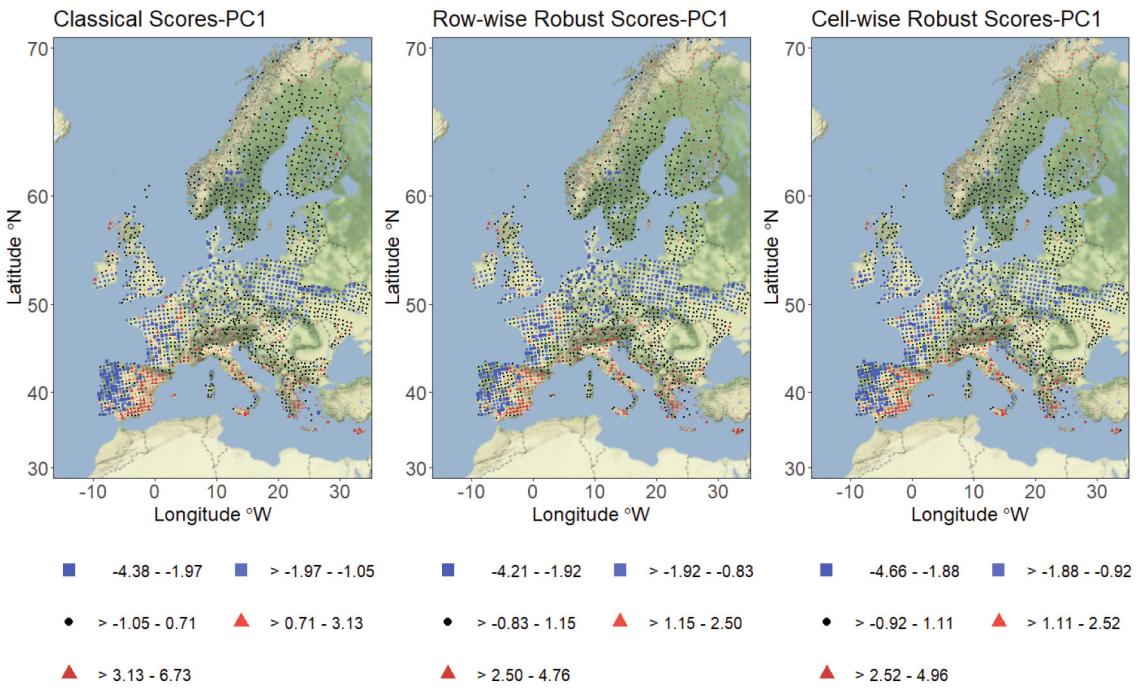

Fig. 8. Scores of PC1 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

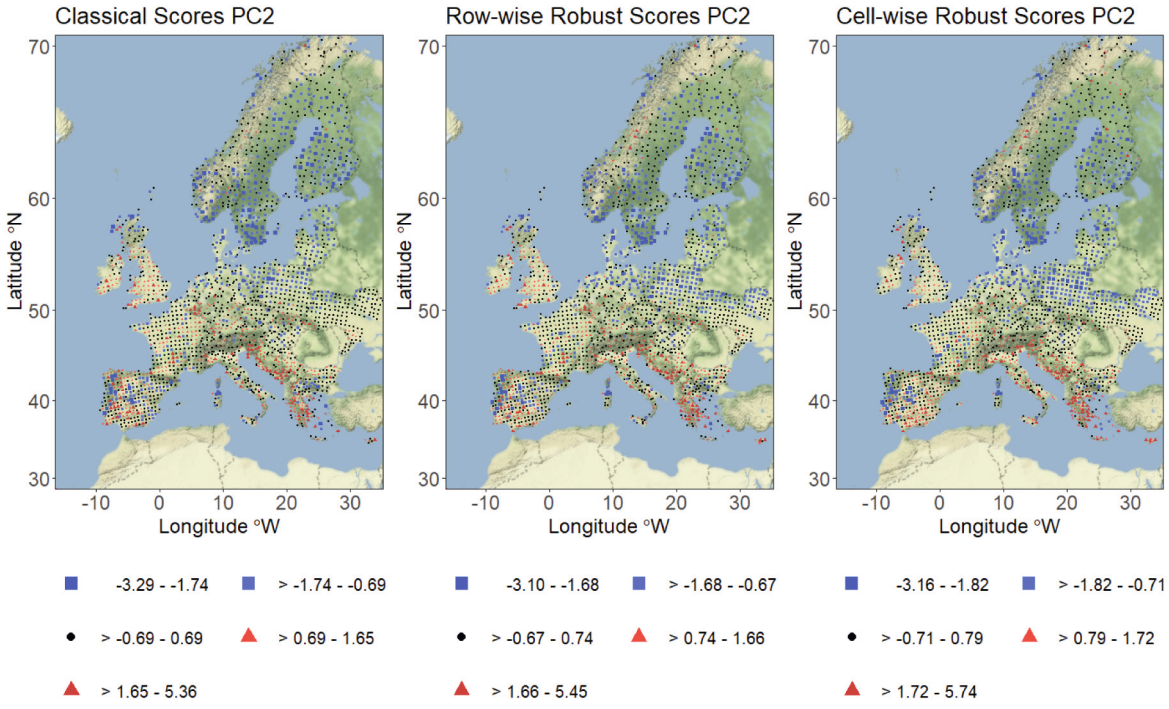

Fig. 9. Scores of PC2 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

estimation. Differences between row-wise and cell-wise robust PCA are not so clearly pronounced, but the element Ca plays a key role: in contrast to the row-wise robust version, the element Ca is practically not affecting PC2 and PC3 in the cell-wise robust version. Ca is related to calcareous rocks, which are very diverse as a parent material in the form of chalk, limestone, dolomite, or marble. Thus, having a component which is not involving Ca might be of advantage to better distinguish the parent material.

Focusing now on the results of the cell-wise robust PCA version, we can conclude:

- PC1 has positive loadings for Ca, Sr, Na, Mg, which characterize calcareous rocks (e.g. in eastern and southern Spain, south France, the Apennines in Italy), ophiolites (e.g. in Greece and Cyprus), greenstones and the crystalline less weathered rocks in Scandinavia. Negative contributions from Zr and Si relate to parent materials as

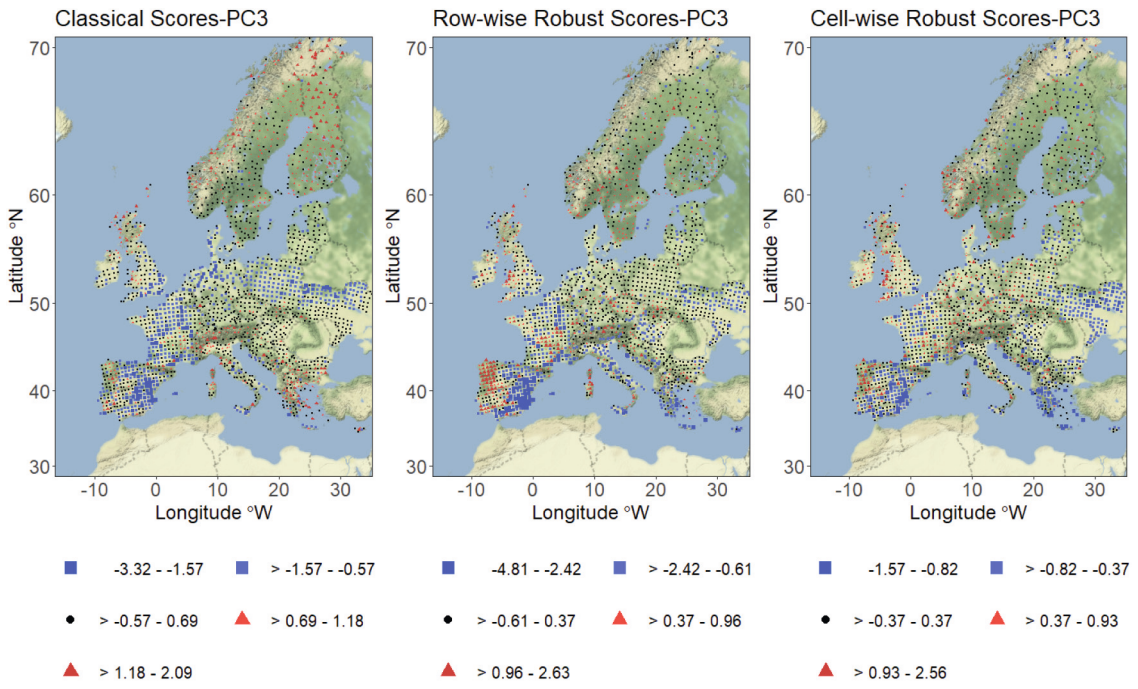

Fig. 10. Scores of PC3 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

coarse grained, sandy soil, and soil developed on granite, typically in central and southern Europe.

- PC2 has positive loadings for Na, Sr, Zr, Si, Ba, K, Sr, which are indicative of felsic lithologies. Negative loadings are for Mg, Cr, V, Fe, Zn, Mn, which indicate mafic bedrocks. Overall, this component shows a separation into northern and southern Europe, following the boundary of the last glaciation.
- PC3 is unique, with high loadings for P, Zn and Mn. This might be related to climatic conditions (north-western parts of Spain, western part of UK, south-west of Scandinavia), where wet and cold climate favor the build-up of organic material. Negative loadings are for Ca, Cr, Zr, Ti, indicative of ophiolite (e.g. Greece, Balkan States, Cyprus), but also limestone (southern/eastern Spain).

6. Summary and conclusions

In the field of robust statistics, research in cell-wise robustness, i.e., robustness with respect to outliers in single cells of a data matrix, received a lot of attention in the last years. This started with an investigation on error propagation of outliers in single cells (Alqallaf et al., 2009), where it has been demonstrated that the effect becomes worse especially if the number of variables increases. Even a small proportion of contaminated cells could then result in a high proportion of rows containing outliers (i.e. outlying cells), which could cause breakdown of traditional row-wise robust methods (Rousseeuw and Van Den Bossche, 2018; Agostinelli et al., 2015).

Cell-wise outliers play also an important role in CoDA, and depending on the specific log-ratio transformation they may also propagate to the entire row (observation) (Mert et al., 2016). The goal is to avoid this form of propagation, and to still make use of the non-contaminated cells (log-ratios) in a statistical analysis.

The focus here was on cell-wise robust covariance estimation by robustly estimating the elements of the variation matrix and by making use of the well-known relationship between both matrices. Robust estimation of the variation matrix is straightforward, by simply applying robust variance estimation to all elements of this matrix. What is not so straightforward is to obtain a covariance estimator which (a) fulfills the properties of a clr covariance matrix and (b) has reasonable robustness properties. We proposed two estimators which are appropriate in the sense of (a), both based on the highly robust and efficient Qn scale estimator. One proposed estimator has an additional regularization, which has advantages especially if the ratio of the number of observations to the number of variables gets small (smaller than 1). Simulations (Section 4) have shown that even without contamination, the estimators are comparable to the classical clr sample covariance estimator, or even preferable if the dimension gets larger than the sample size. In case of contamination, they are clearly preferable over the classical estimator, but in many cases also preferable over a row-wise robust estimator. Particularly if the dimension is close to or even bigger than the sample size, the regularized cell-wise version shows excellent properties, and it achieves good results in cases where the row-wise estimator even cannot be computed. The simulation settings with zero corruption might be of special interest to applications, where detection limit problems occur, and it is surprising how well the regularized Qn estimator performs in this situation.

A covariance estimate is the basis for many multivariate methods, such as PCA, and in CoDA also the variation matrix estimate is very useful, e.g. for Q-mode clustering. Both methods have been demonstrated at a geochemistry dataset, based on classical and robust versions of the estimates. Generally speaking, the principle of robustness is to fit the model to the data majority, and thus to downweight observations which deviate from this majority trend (Maronna et al., 2019, p 195–224). While in row-wise robustness the downweighting is done for complete observations, cell-wise methods here in the variation matrix context only downweight deviating values of particular pairwise log-ratios, and thus they incorporate all “useful” information of the observations from pairwise log-ratios.

One of the proposed estimators of the variation matrix was a shrinkage estimator. Note that the same type of estimator could also be defined in a non-robust manner. The resulting estimator could then serve as a classical counterpart in situations where the dimension exceeds the number of observations. This use-case is of increasing importance in areas where more and more variables are measured, e.g. in bioinformatics (Gloor et al., 2017). In our future research we will investigate this case in more detail, also in combination with sparsity constraints.

CRediT authorship contribution statement

The authors declare that this paper is their own work, and that nobody else is involved.

Declaration of competing interest

The authors declare that they do not have any conflict of interests.

Data availability

The data are already available.

Acknowledgements

KF was supported by the Czech Science Foundation (grant number 22-15684L), and CR and PF by the Austrian Science Fund (project number I 5799-N).

References

Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H., 2015. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test 24 (3), 441–461.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., 2009. Propagation of outliers in multivariate data. Ann. Stat. 37 (1), 311–331.

Billheimer, D., Guttorp, P., Fagan, W.F., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96 (456), 1205–1214.

Chen, Y., Wiesel, A., Hero, A.O., 2011. Robust shrinkage estimation of high-dimensional covariance matrices. IEEE Trans. Signal Process. 59 (9), 4097–4107.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35 (3), 279–300.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis: With Worked Examples in R. Springer.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome datasets are compositional: and this is not optional. Front. Microbiol. 8, 2224.

Higham, N.J., 1988. Computing a nearest symmetric positive semidefinite matrix. Linear Algebra Appl. 103, 103–118. https://doi.org/10.1016/0024-3795(88)90223-6. URL. https://www.sciencedirect.com/science/article/pii/0024379588902236.

Hubert, M., Debruyne, M., Rousseeuw, P., 2018. Minimum covariance determinant and extensions. Wiley Interdiscip. Rev.: Comput. Stat. 10 (3), e1421.

Jin, S., Notredame, C., Erb, I., 2022. Compositional Covariance Shrinkage and Regularised Partial Correlations. https://doi.org/10.48550/ARXIV.2212.00496. URL. https://arxiv.org/abs/2212.00496.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88 (2), 365–411.

Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M., 2019. Robust Statistics: Theory and Methods (With R). John Wiley & Sons.

Mert, C., Filzmoser, P., Hron, K., 2016. Error propagation in compositional data analysis: theoretical and practical considerations. Math. Geosci. 48 (8), 941–961.

Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15, 384–398.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.

Raymaekers, J., 2022. cellWise: Analyzing Data With Cellwise Outliers, r Package Version 2.5.0. URL. https://CRAN.R-project.org/package=cellWise.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014. Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set. Schweizerbarth.

Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, pp. 283–297.

Rousseeuw, P., Croux, C., 1993. Alternatives to the median absolute deviation. J. Am. Stat. Assoc. 88 (424), 1273–1283 arXiv: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1993.10476408 https://doi.org/10.1080/01621459.1993.10476408. URL. https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408.

Rousseeuw, P., Van Den Bossche, W., 2018. Detecting deviating data cells. Technometrics 60 (2), 135–145.

Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41 (3), 212–223.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4 (1).

Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., Hron, K., 2021. Robust regression with compositional covariates including cellwise outliers. ADAC 15 (4), 869–909.

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications, pp. 341–355.

Walach, J., Filzmoser, P., Kouřil, Š., Friedecký, D., Adam, T., 2020. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. J. Chemom. 34 (1).

Wilcox, R.R., 2010. Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. Springer.