frontiers | Frontiers in Bioinformatics

TYPE Methods
PUBLISHED 10 August 2023
DOI 10.3389/fbinf.2023.1211819

OPEN ACCESS

EDITED BY
Martin Hemberg,
Brigham and Women's Hospital and Harvard Medical School, United States

REVIEWED BY
Virginie Uhlmann,
European Bioinformatics Institute (EMBL-EBI), United Kingdom
Nikolaos Patikas,
Brigham and Women's Hospital and Harvard Medical School, United States

*CORRESPONDENCE
Carine Legrand,
carine.legrand@inserm.fr
Khanh Dao Duc,
kdd@math.ubc.ca

†These authors have contributed equally to this work

RECEIVED 03 May 2023
ACCEPTED 26 July 2023
PUBLISHED 10 August 2023

CITATION
Li W, Mirone J, Prasad A, Miolane N, Legrand C and Dao Duc K (2023), Orthogonal outlier detection and dimension estimation for improved MDS embedding of biological datasets. *Front. Bioinform.* 3:1211819. doi: 10.3389/fbinf.2023.1211819

Orthogonal outlier detection and dimension estimation for improved MDS embedding of biological datasets

Wanxin Li[1], Jules Mirone[2,3], Ashok Prasad[4], Nina Miolane[5], Carine Legrand[6*†] and Khanh Dao Duc[1,2*†]

[1]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, [2]Department of Mathematics, University of British Columbia, Vancouver, BC, Canada, [3]Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France, [4]Department of Chemical and Biological Engineering, School of Biomedical Engineering, Colorado State University, Fort Collins, CO, United States, [5]Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, United States, [6]Université Paris Cité, Génomes, biologie cellulaire et thérapeutique U944, INSERM, CNRS, Paris, France

Conventional dimensionality reduction methods like Multidimensional Scaling (MDS) are sensitive to the presence of orthogonal outliers, leading to significant defects in the embedding. We introduce a robust MDS method, called *DeCOr-MDS* (Detection and Correction of Orthogonal outliers using MDS), based on the geometry and statistics of simplices formed by data points, that allows to detect orthogonal outliers and subsequently reduce dimensionality. We validate our methods using synthetic datasets, and further show how it can be applied to a variety of large real biological datasets, including cancer image cell data, human microbiome project data and single cell RNA sequencing data, to address the task of data cleaning and visualization.

KEYWORDS

orthogonal outliers, outlier detection, outlier correction, multidimensional scaling, shape data, microbiome data, scRNA seq

1 Introduction

Multidimensional scaling (MDS) is a commonly used and fast method of data exploration and dimension reduction, with the unique capacity to take non-euclidean dissimilarities as its input. However, sensitivity to outliers is a major drawback (Harmeling et al., 2005; Blouvshtein and Cohen-Or, 2019). As arbitrary removal of outliers is undesirable, a possible alternative is to detect outliers and accommodate their influence on the MDS embedding, thus leveraging the information contained in outlying points.

Outlier detection has been widely used in biological data. Sheih and Yeung proposed a method using principal component analysis (PCA) and robust estimation of Mahalanobis distances to detect outlier samples in microarray data (Shieh and Hung, 2009). Chen *et al.* reported the use of two PCA methods to uncover outlier samples in multiple simulated and real RNA-seq data (Oh et al., 2008). Outlier influence can be mitigated depending on the specific type of outlier. In-plane outliers and bad leverage points can be harnessed using ℓ_1-norm (Spence and Lewandowsky, 1989; Cayton and Dasgupta, 2006; Forero and Giannakis, 2012), correntropy or M-estimators (Mandanas and Kotropoulos, 2017). Outliers which violate the triangular inequality can be detected and corrected based on their pairwise distances (Blouvshtein and Cohen-Or, 2019). Orthogonal outliers are another particular

case, where outliers have an important component, orthogonal to the hyperspace where most data is located. These outliers often do not violate the triangular inequality, and thus require an alternative approach.

Although MDS is known to be sensitive to such orthogonal outliers (Song et al., 2007; Legrand, 2017), none of the existing methods addresses this issue, to the best of our knowledge. We present here a robust MDS method, called *DeCOr-MDS*, Detection and Correction of Orthogonal outliers using MDS. *DeCOr-MDS* takes advantage of geometrical characteristics of the data to reduce the influence of orthogonal outliers, and estimate the dimension of the dataset. Our paper is organized as follows. We first describe the procedure and its implementation in detail. We then validate our method on synthetic data to confirm the accuracy and characterize the importance of different parts of our procedure. We further run the method on different experimental datasets from single cell images, microbiome sequencing data, and scRNA-seq data. Our experiments show that *DeCOr-MDS* can detect artefacts in cell shape data, improve the visualization of clusters in microbiome data, and be used as a step for quality control for scRNA-seq data, illustrating how it can be broadly applied to interpret and improve the performance of MDS on biological datasets. Finally, we discuss the advantages and limitations of our method and future directions.

2 Materials and methods

2.1 Background: height and volume of n-simplices

We recall some geometric properties of simplices, which our method is based on. For a set of n points $(x_1, \ldots, x_n)$, the associated n-simplex is the polytope of vertices $(x_1, \ldots, x_n)$ (a 3-simplex is a triangle, a 4-simplex is a tetrahedron and so on). The height $h(V_n, x)$ of a point x belonging to a n-simplex V_n can be obtained as (Sommerville, 1929)

$$h(V_n, x) = n\frac{V_n}{V_{n-1}}, \tag{1}$$

where V_n is the volume of the n-simplex, and V_{n-1} is the volume of the $(n-1)$-simplex obtained by removing the point x. V_n and V_{n-1} can be computed using the pairwise distances only, with the Cayley-Menger formula (Sommerville, 1929):

$$V_n = \sqrt{\frac{|det\,(CM_n)|}{2^n \cdot (n!)^2}}, \tag{2}$$

where $det(CM_n)$ is the determinant of the Cayley-Menger matrix CM_n, that contains the pairwise distances $d_{i,j} = \|x_i - x_j\|$, as

$$CM_n = \begin{bmatrix} 0 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 0 & d^2_{1,2} & \ldots & d^2_{1,n} & d^2_{1,n+1} \\ 1 & d^2_{2,1} & 0 & \ldots & d^2_{2,n} & d^2_{2,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & d^2_{n,1} & d^2_{n,2} & \ldots & 0 & d^2_{n,n+1} \\ 1 & d^2_{n+1,1} & d^2_{n+1,2} & \ldots & d^2_{n+1,n} & 0 \end{bmatrix}. \tag{3}$$

2.2 Orthogonal outlier detection and dimensionality estimation

We now consider a dataset $\mathbf{X}$ of size $N \times d$, where N is the sample size and d the dimension of the data. We associate with $\mathbf{X}$ a matrix $\mathbf{D}$ of size $N \times N$, which represents all the pairwise distances between observations of $\mathbf{X}$. We also assume that the data points can be mapped into a vector space with *regular observations* that form a *main* subspace of unknown dimension $d^\star$ with some small noise, and additional *orthogonal outliers* of relatively large orthogonal distance to the main subspace (Figure 1A). Our proposed method aims to infer from $\mathbf{D}$ the dimension of the main data subspace $d^\star$, using the geometric properties of simplices with respect to their number of vertices: Consider a $(n+2)$-simplex containing a data point x_i and its associated height, that can be computed using Eq. 1 in Section 2.1. When $n < d^\star$ and for S large enough, the distribution of heights obtained from different simplices containing x_i remains similar, whether x_i is an orthogonal outlier or a regular observation (see Figure 1B). In contrast, when $n \geq d^\star$, the median of these heights approximately yields the distance of x_i to the main subspace (Figure 1C). This distance should be significantly larger when x_i is an orthogonal outlier, compared with regular points, for which these distances are tantamount to the noise.

To estimate $d^\star$ and for a given dimension n tested, we thus randomly sample, for every x_i in $\mathbf{X}$, $S(n+2)$-simplices containing x_i, and compute the median of the heights h_i^n associated with these S simplices. Upon considering, as a function of the dimension n tested, the distribution of median heights $(\tilde{h}_1^n, \ldots, \tilde{h}_N^n)$ (with $1 \leq i \leq N$), we then identify $d^\star$ as the dimension at which this function presents a sharp transition towards a highly peaked distribution at zero. To do so, we compute $\tilde{h}_n$, as the mean of $(h_1^n, \ldots, h_N^n)$, and estimate $d^\star$ as

$$\bar{n} = \underset{n}{\text{argmax}}\, \frac{\tilde{h}_{n-1}}{\tilde{h}_n}. \tag{4}$$

Furthermore, we detect orthogonal outliers using the distribution obtained in $\bar{n}$, as the points for which $h_i^{\bar{n}}$ largely stands out from $\tilde{h}_{\bar{n}}$. To do so, we compute $\sigma_{\bar{n}}$ the standard deviation observed for the distribution $(h_1^{\bar{n}}, \ldots, h_N^{\bar{n}})$, and obtain the set of orthogonal outliers $\mathbf{O}$ as

$$\mathbf{O} = \left\{ i \mid h_i^{\bar{n}} > \tilde{h}_{\bar{n}} + c \times \sigma_{\bar{n}} \right\}, \tag{5}$$

where $c > 0$ is a parameter set to achieve a reasonable trade-off between outlier detection and false detection of noisy observations. Our implementation uses $c = 3$ by default (following the three σ rule Pukelsheim (1994), and which corresponds to $\sim 99.9\%$ of a Gaussian distribution being conserved), value which was also used in our experiments. In case users possess prior information or want to control the fraction of detected outliers, the value of c may be modified, with increasing c making the detection stricter. Also note that our method introduces another parameter S, as it samples S simplices to calculate the median of the corresponding heights. Therefore, S should be large enough so the resulting sample median well approximates the global median. Assuming the heights being sampled from a continuous distribution, this can be guaranteed as the sample median is asymptotically normal, with mean equal to the true median and the standard deviation proportional to $\frac{1}{\sqrt{S}}$ (Rider, 1960).

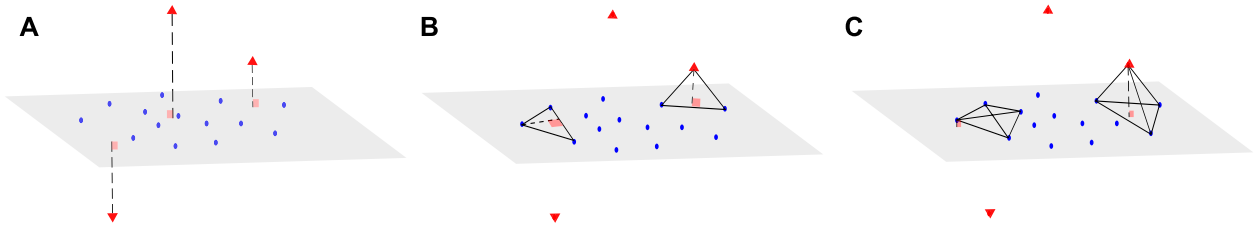

FIGURE 1
Example of a dataset with orthogonal outliers and n-simplices. **(A)** Representation of a dataset with regular data points (blue) belonging to a main subspace of dimension 2 with some noise, and orthogonal outliers (red triangle symbols) in the third dimension. **(B)** View of two instances of 3-simplices (triangles), one with only regular points (left) and the other one containing one outlier (right). The height drawn from the outlier is close to the height of the regular triangle. **(C)** Upon adding other regular points to obtain tetrahedrons (4-simplices), the height drawn from the outlier (right) becomes significantly larger than the height drawn from the same point (left) as in **(B)**.

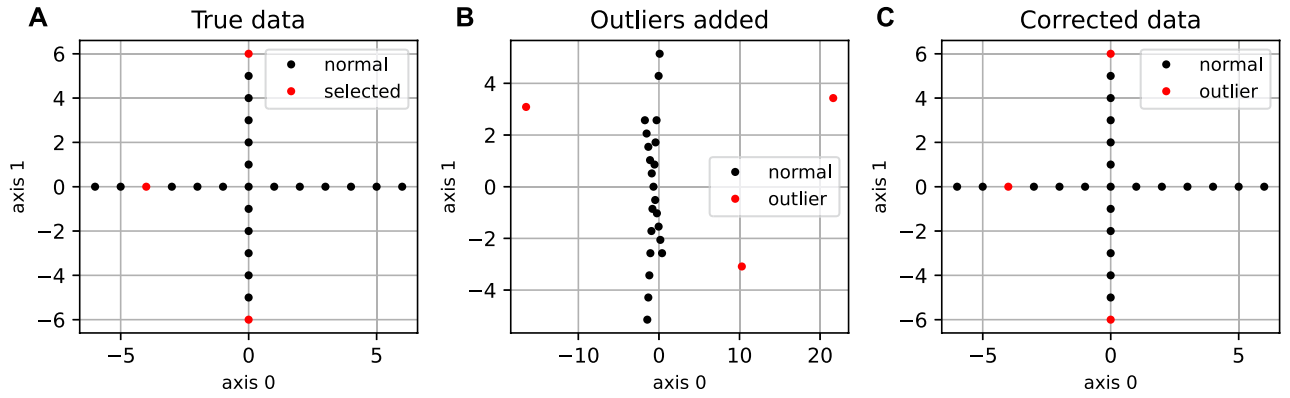

FIGURE 2
Application of DeCOr-MDS on a cross dataset. **(A)** Original cross dataset. The points selected to be orthogonal outliers are highlighted in red. **(B)** MDS embedding of the original data with an outlying component added to the selected points. **(C)** MDS embedding after preprocessing using DeCOr-MDS. Note that after correction, we recover the original cross structure.

2.3 Correcting the dimensionality estimation for a large outlier fraction

The method presented in the previous section assumes that at dimension $d^\star$, the median height calculated for each point reflects the distance to the main subspace. This assumption is valid when the fraction of orthogonal outliers is small enough, so that the sampled n-simplex likely contains regular observations only, aside from the evaluated point. However, if the number of outliers gets large enough so that a significant fraction of n-simplices also contains outliers, then the calculated heights would yield the distance between x_i and an outlier-containing hyperplane, whose dimension is larger than a hyperplane containing only regular observations. The apparent dimensionality of the main subspace would thus increase and generates a positive bias on the estimate of $d^\star$.

Specifically, if $\mathbf{X}$ contains a fraction of p outliers, and if we consider $o_{n,p,N}$ the number of outliers drawn after uniformly sampling $n + 1$ points (to test the dimension n), then $o_{n,p,N}$ follows a hypergeometric law, with parameters $n + 1$, the fraction of outliers $p = N_o/N$, and N. Thus, the expected number of outliers drawn from a sampled simplex is $(n + 1) \times p$. After estimating $\bar{n}$ (from Section 2.2), and finding a proportion of outliers $\bar{p} = |\mathbf{O}|/N$ using Eq. 5, we hence correct $\bar{n}$ *a posteriori* by substracting the estimated bias δ, as the integer part of the expectation of $o_{n,p,N}$, so the debiased dimensionality estimate $n^\star$ is

$$n^\star = \bar{n} - \left\lfloor (\bar{n} + 1) \times p \right\rfloor. \tag{6}$$

2.4 Outlier distance correction

Upon identifying the main subspace containing regular points, our procedure finally corrects the pairwise distances that contain outliers in the matrix $\mathbf{D}$, in order to apply a MDS that projects the outliers in the main subspace. In the case where the original coordinates cannot be used (e.g., as a result of some transformation or if the distance is non Euclidean), we perform the two following steps: 1) We first apply a MDS on $\mathbf{D}$ to place the points in a euclidean space of dimension d, as a new matrix of coordinates $\tilde{X}$. 2) We run a PCA on the full coordinates of the estimated set of regular data points (i.e., $\tilde{X}\backslash O$), and project the outliers along the first $\bar{n}^\star$ principal components of the PCA, since

these components are sufficient to generate the main subspace. Using the projected outliers, we accordingly update the pairwise distances in **D** to obtain the corrected distance matrix **D***. Note that in the case where **D** derives from a euclidean distance between the original coordinates, we can skip step 1), and directly run step 2) on the full coordinates of the estimated set of regular data points.

2.5 Overall procedure and implementation

The overall procedure, called DeCOr-MDS, is described in Algorithm 1. The values for the parameters S and c were set by default and in our experiments to $S = 100$ and $c = 3$. We also provide an implementation in Python 3.8.10 available on this github repository: https://github.com/wxli0/DeCOr-MDS.

Input D the pairwise distance matrix of the dataset of size $N \times d$, E_{dim} the set of dimensions ($\leq d$) to be tested, c and S user-specified constants
Output $\bar{n}$* the relevant dimension of the dataset, O the list of orthogonal outliers, and D* the matrix of corrected pairwise distances
for n in E_{dim} **do**
  for i in [1, N] **do**
    for j in [1, S] **do**
      Sample a $(n+2)$-simplex $V_{i,j}$ containing x_i
      Compute the height (using D and Eq. 1) $h_i^{(j,n)} \coloneqq h_i^j(V_{i,j}, x_i)$
    end for
    $h_i^n \coloneqq median(h_i^{(1,n)}, h_i^{(2,n)}, \ldots, h_i^{(S,n)})$
  end for
  $\tilde{h}_n \coloneqq mean(h_1^n, h_2^n, \ldots, h_N^n)$
  $\sigma(n) \coloneqq std(h_1^n, h_2^n, \ldots, h_N^n)$
end for
$\bar{n} \coloneqq \arg\max \frac{\tilde{h}_{n-1}}{\tilde{h}_n}$
$O \coloneqq \{i \mid h_i^{\bar{n}} > \tilde{h}_{\bar{n}} + c \times \sigma_{\bar{n}}\}$
$p \coloneqq |O|/N$
$\bar{n}^* = \bar{n} - \lfloor (\bar{n}+1) \times p \rfloor$
(Skip if using original coordinates) Apply a MDS on D to create an euclidean space of dimension d, resulting $\tilde{X}$
Apply a PCA on $\tilde{X} \backslash O$ to get the main subspace of dimensionality $\bar{n}$
for outlier i in O **do**
  Project x_i on the main subspace, and correct the coordinates of x_i in $\tilde{X}$
end for
Recompute the pairwise distance matrix D* from $\tilde{X}$.
return $\bar{n}$*, O and D*

Algorithm 1. DeCOr-MDS.

The complexity of the algorithm can be briefly evaluated as follows.

1. Given one n-simplex, the volume computation has a complexity of $\mathcal{O}(n^3)$. Since we compute the height for S simplices and repeat the process for all E_{dim} dimensions, the total complexity of this step amounts to $\mathcal{O}(SNn^3E_{\text{dim}})$,
2. The complexity of PCA over the regular data points is $\mathcal{O}(Nd \times \min(N, d) + d^3)$,
3. The complexity of MDS over the pairwise distance matrix D is $\mathcal{O}(N^3)$,
4. The computation of the corrected distances matrix is $\mathcal{O}(dN^2)$.

Note that with the tested dimensions being smaller than the data dimension (n, $E_{\text{dim}} < d$), and the number of simplices being significantly smaller than the total number of data points ($S \ll N$), the burden of evaluating the simplices (step 1) and correcting outliers (step 4) is in practice less than the cost of the PCA (step 2) and MDS (step 3).

2.6 Datasets

2.6.1 Synthetic datasets

The "cross" dataset (Spence and Lewandowsky, 1989), which is a two-dimensional dataset representing a simple cross structure (Figure 2) was generated with $N = 25$ points, and $d^* = 2$. We introduced orthogonal outliers by randomly sampling three points and by adding a third coordinate of random amplitude to them. Other synthetic datasets were generated by sampling Gaussian-distributed coordinates in the main subspace, and adding some small noise in the whole space with variance between 0.0001 and 0.0003. A fraction p of the points was considered to define the orthogonal outliers, with coordinates modified by randomly increasing the coordinate(s) orthogonal to the plane; the amount increased is drawn from a uniform distribution between −30 and 30 or -100 to 100. These datasets were generated for a main subspace of dimension 2, 10 and 40, with $p = 0.05$ and $N = 200$ for dimension $d^* = 2$, $p = 0.05$ and $N = 1000$ for dimension $d^* = 10$, and p varying between 0.02 and 0.1 for $d^* = 40$, and $N = 1,000$. For all the synthetic datasets, the pairwise distance matrix was calculated using the Euclidean distance.

2.6.2 Cell shape dataset

The cell shape dataset contains mouse osteosarcoma 2D imaged cells (Alizadeh et al., 2019), that were processed into a 100×2 vector of coordinates that define the cell shape contour, used as a test dataset in the Python package Geomstats (Miolane et al., 2020) (for more details, see also (Miolane et al., 2021) and the associated Github link). We more specifically considered the subset of "DUNN" cells (that denotes a specific lineage) from the control group (no treatment on the cells), which yields 207 cells in total. The pairwise distance matrix of all cell shapes was obtained from the same reference (Miolane et al. (2020; 2021)) using the so-called Square Root Velocity metric that derives from the L_2 distance between velocities of the curves (Srivastava et al., 2010).

2.6.3 HMP dataset

The Human Microbiome Project (HMP) (Turnbaugh et al., 2007) dataset represents the microbiome measured across thousands of human subjects. The human microbiome corresponds to the set of microorganisms associated to the human body, including the gut flora, or the skin microbiota. The data used here corresponds to the HMP1 phase of clinical production. The hypervariable region v13 of ribosomal RNA was sequenced for each sample, which allowed to identify and count each specific microorganism, called phylotype. The processing and classification were performed by the HMP using

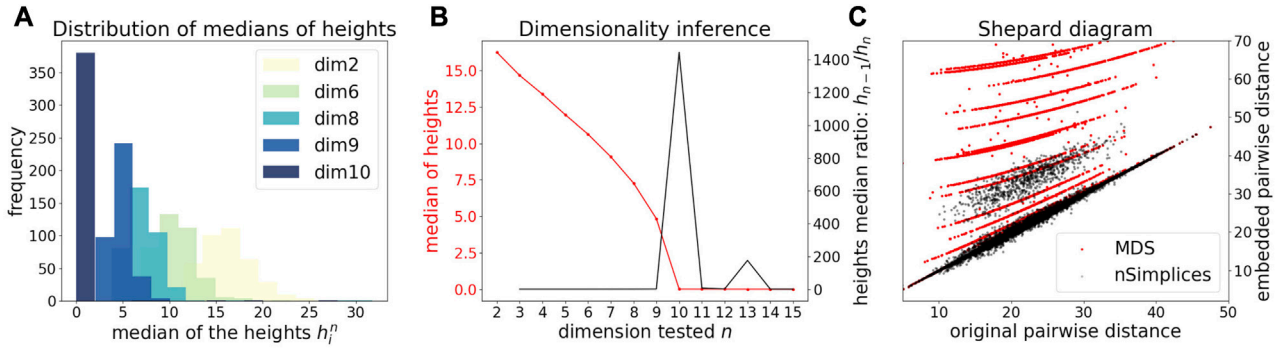

FIGURE 3
Application of DeCOr-MDS on a synthetic dataset with a main subspace of dimension 10. **(A)** Distribution of median heights per data point x_i as a function of the tested dimension n. **(B)** Dimensionality inference based on the ratio of median heights (red curve, see also Eq. 1), with the optimal ratio (black curve) found for the true dimension 10. **(C)** Shepard diagram comparing the pairwise distances between regular points and outliers that are projected to the main subspace (true δ_{ij}), with the same distances obtained after directly running MDS on the original pairwise distance matrix (red dots), or after correcting these distances using our procedure (black dots).

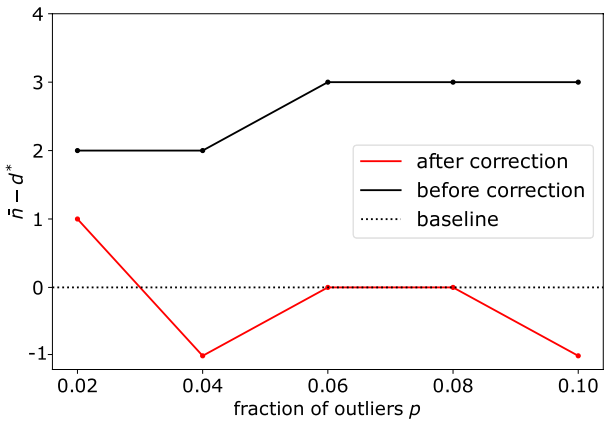

FIGURE 4
Application of DeCOr-MDS on a dataset with a main subspace of dimension 40: Dimension correction effect versus the fraction of outliers. The vertical axis represents the remaining bias between the inferred and actual dimensions, before and after bias correction. After correction, the differences between the estimated dimensions and the true dimension are always closer to 0 regardless of the fraction of outliers.

MOTHUR, and made available as low quality counts (https://www.hmpdacc.org/hmp/HMMCP/) (Turnbaugh et al., 2007). We downloaded this dataset, and subsequently, counts were filtered and normalized as previously described (Legrand, 2017). For our analysis, we also restricted our dataset to samples collected in nose and throat. Samples and phylogenies with less than 10 strictly positive counts were filtered out (Legrand, 2017), resulting in an $n \times p$-matrix where $n = 270$ samples and $p = 425$ phylotypes. Next, the data distribution was identified with an exponential distribution, by fitting its rate parameter. Normalization was then achieved by replacing the abundances (counts) with the corresponding quantiles. Lastly, the matrix of pairwise distances was obtained using the Euclidean distance.

2.6.4 scRNA-seq dataset

The scRNA-seq dataset contains single-cell transcriptomic profiles from mouse pancreatic cells (raw count data accession number: GSE84133), which were first processed using standard quality control methods from McCarthy et al. (2017). From the gene count matrix, which originally contained 1,886 cells with 13,357 genes, we focused on the cells from Mouse 2, yielding 1,063 cells with 13,357 genes. We further lognormalized the data (Luecken and Theis, 2019) and selected highly variable genes using *scanpy* package (Wolf et al., 2018). This procedure resulted in a normalized gene count matrix of 1,603 cells with 2,601 genes. To obtain a matrix of pairwise distances, we used the Euclidean distance.

3 Results

3.1 Using n-simplices for orthogonal outlier detection and dimensionality reduction

We propose a robust method to reduce and infer the dimensionality of a dataset from its pairwise distance matrix, by detecting and correcting orthogonal outliers. The method, called *DeCOr-MDS*, can be divided into three sub-procedures detailed in Sections 2.2–2.4, with the overall algorithm provided in Section 2.5. The first procedure detects orthogonal outliers and estimates the subspace dimension using the statistics of simplices that are sampled from the data, using Eqs 4, 5. The second procedure corrects for potential bias in estimated dimension when the fraction of outlier is large. The third procedure corrects the pairwise distance of the original data, by replacing the distance to orthogonal outliers by that to their estimated projection on the main subspace. In the next sections, we report the results obtained upon running the procedure on synthetic and various biological datasets, that demonstrate the performance and accuracy of the method. For all these experiments, we also reported the runtime in Supplementary Table S1, showing how the method can be used in practice with reasonable time on experimental datasets (less than 10 min in our workstation, with x86_64 CPU, 132 GB RAM and 447 GB disk storage).

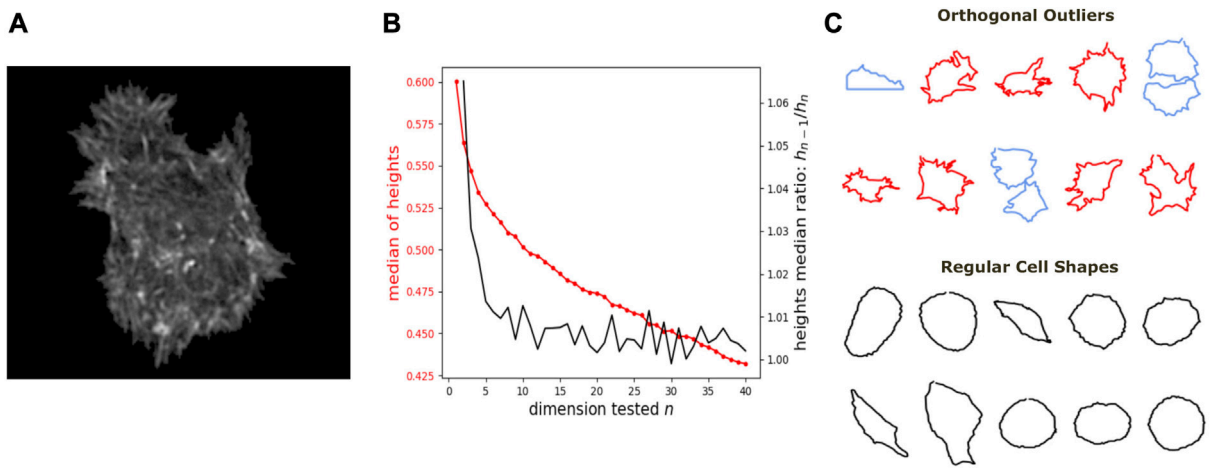

FIGURE 5
Application of DeCOr-MDS on a cell shapes dataset. **(A)** An example of osteocarcoma cell image obtained from fluorescence microscopy. We process and extract the cell contour in our analysis. **(B)** Dimensionality inference of the dataset obtained from 207 cell shapes using DeCOr-MDS. We estimate the dimension of the main subspace $n^* = 2$. The red line shows the median of heights and the black curve shows the heights median ratios between h_{n-1} and h_n. **(C)** Examples of cell shapes, including regular cells (in black), and orthogonal outliers detected. Among these outliers, we highlight cell shapes that are likely to be invalid due to segmentation errors (in blue), with the other outliers shown in red.

3.2 Performance on synthetic datasets

We first illustrate and evaluate the performance of the method on synthetic datasets, (for a detailed description of the datasets and their generation, see the Methods Section 2.6). On a simple dataset of points forming a 2D cross embedded in 3D (Figure 2A), we observed that the MDS is sensitive to the presence of orthogonal outliers and distorts the cross when reducing the data in 2D (Figure 2B). In contrast, our procedure recovers the original geometry of the uncontaminated dataset, with the outliers being correctly projected (Figure 2C). The same results were obtained when sampling regular points from a 2D plane (Supplementary Figure S1). We further tested higher dimensions, and illustrate in Figure 3A how the distribution of heights becomes concentrated around 0, when testing for the true dimension ($d^* = 10$), as suggested in the Methods Section 2.2. As a result, our method allows to infer the main subspace dimension from Eq. 4, as shown in Figure 3B. In addition, the procedure accurately corrects the pairwise distances to orthogonal points with the distances to their projections on the main subspace, as shown in Figure 3C.

When the dimension of the subspace and fraction of outliers get significantly large, we also illustrate the importance of the correction step (see Methods Section 2.3), due to the sampling of simplices that contain several outliers. Upon using synthetic datasets with $d^* = 40$ and varying the number (fraction) of outliers from 20 (2%) to 100 (10%), we observe this bias appearing before correction, with d^* being overestimated by 2 or 3 dimensions (Figure 4). Using the debiased estimate n^* from Eq. 6 successfully reduced the bias, with an error ≤ 1 for all the parameters tested.

3.3 Application to cell shape data

We further show how *DeCOr-MDS* can be broadly applied to biological data, ranging from images to high throughput sequencing. We first studied a dataset of single cell images, from osteosarcoma cells (see Figure 5A), which were processed to extract from their contour a 100 × 2 array of *xy* coordinates representing a discretization of a closed curve (see Dataset Section 2.6). We obtained a pairwise distance matrix on this set of curves by using the so-called *Square Root Velocity* (SRV) metric, which defines a Euclidean distance on the space of velocities that derive from a regular parameterization of the curve (Srivastava et al., 2010; Miolane et al., 2020). Using *DeCOr-MDS*, we found a main subspace of dimension 2 (Figure 5B), with 14 (7%) outliers detected among the 207 cells of this dataset. The comparison between the resulting embedding and that obtained from a simple MDS is shown in Supplementary Figure S2, and reveals that outliers, when uncorrected, affect the embedding coordinates, while our correction mitigates it. By examining in more details the regular and inferred outlier cells (Figure 5C, with all cell shapes shown in Supplementary Figure S3), we found regular observations to approximately describe elliptic shapes, which is in agreement with the dimension found, since ellipses are defined by 2 parameters. One can also visually interpret the orthogonal outliers detected as being more irregular, with the presence of more spikes and small protusions. Interestingly, the procedure also identified as outliers some images containing errors, due to bad cropping or segmentation (with 2 cells shown instead of one), which should thus be removed of the dataset for downstream analysis.

3.4 Application to HMP data

As another example of application to biological data, we next considered a dataset from the Human Microbiome Project (HMP). The Human Microbiome Project aims at describing and studying the microbial contribution to the human body. In particular, genes contributed by microbes in the gut are of primary importance in health and disease (Turnbaugh et al., 2007). The resulting data is an

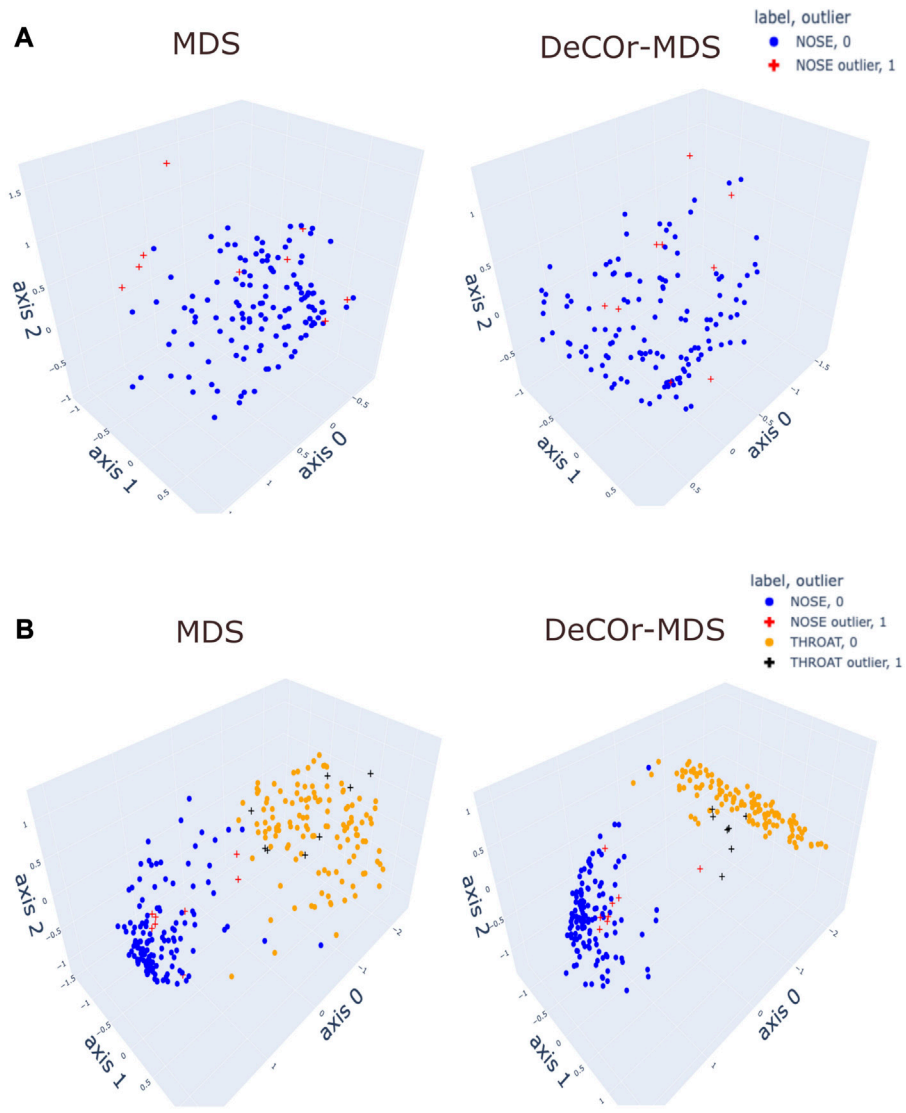

FIGURE 6
Application of DeCOr-MDS on HMP dataset. **(A)**: Structure restituted on 3 axes using *MDS* (left) and our procedure (right) using data from the nose site. The points marked with cross represent orthogonal outliers detected by *DeCOr-MDS*, which are also put closer to regular points after correction. **(B)** Same comparison as in **(A)** using data from nose and throat. The two clusters formed by nose and throat have a better separation using DeCOr-MDS.

array which typically contains the abundance of different elements of the microbiome (typically 10^2–10^3), denoted phylotypes, measured in different human subjects. To analyze such high dimensional datasets, dimensionality reduction methods including MDS (often denoted Principal Coordinates Analysis PCoA), are typically applied and used to visualize the data (Brooks et al., 2018; Trevelline and Kohl, 2022; Zhou et al., 2022).

To assess our method incrementally, we restricted first the analysis to a representative specific site (nose), yielding a 136 × 425 array that was further normalized to generate Euclidean pairwise distance matrices (see Material and Methods Section 2.6 for more details). Upon running *DeCOr-MDS*, we estimated the main dimension to be 3, with 9 (6.62%) orthogonal outliers detected, as shown in Figure 6A. This is also supported by another study that the estimated dimension of HMP dataset is 2 or 3 (Tomassi et al., 2021). We also computed the average distance between these orthogonal outliers and the barycenter of regular points in the reduced subspace, and obtained a decrease from 1.21 when using *MDS* to 0.91 when using *DeCOr-MDS*. This decrease suggests that orthogonal outliers get corrected and projected closer to the regular points, to improve the visualization of the data in the reduced subspace, like in our experiments with the synthetic datasets (Figure 2 and Supplementary Figure S2). In Figure 6B, we next aggregated data points from another site (throat) to study how the method performs in this case, yielding a 270 × 425 array that was further normalized to generate Euclidean pairwise distance matrices. As augmenting the dataset brings a separate cluster of data points, the dimension of the main dataset was then estimated to be 2, with 13 (5%) orthogonal outliers detected, as shown in Figure 6B. The average distance

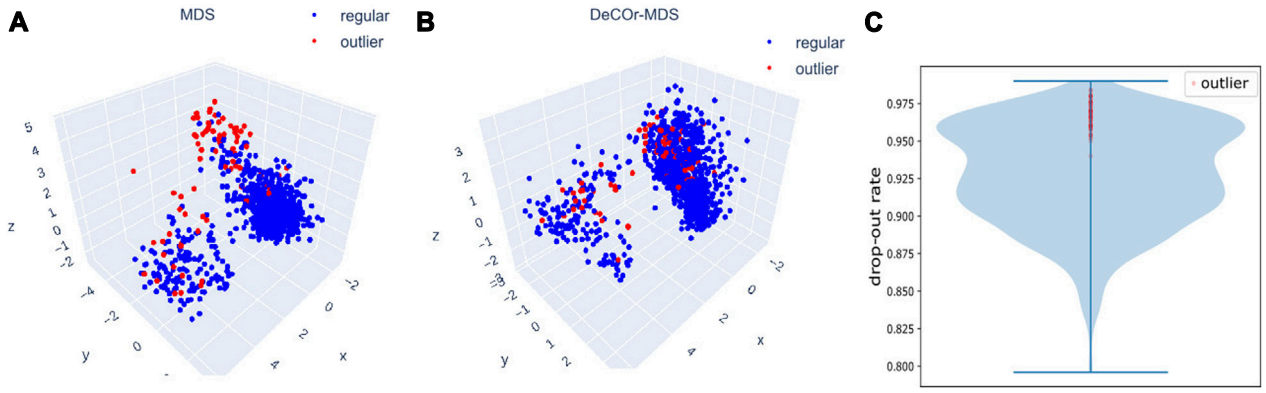

FIGURE 7
Application of *DeCOr-MDS* on the scRNA-seq dataset. Structure restituted on 3 axes using *MDS* **(A)**, and using *DeCOr-MDS* **(B)**. The red points in **(A)** and **(B)** are outliers detected by *DeCOr-MDS*. The blue points are regular points. **(C)** Violin plot for the drop-out rates for the scRNA-seq dataset, for the top 500 highly expressed genes. Drop-out rates for outliers detected by *DeCOr-MDS* are shown in red.

between the projected outliers and the barycenter of projected regular points are approximately the same when using *MDS* (1.46) as when using *DeCOr-MDS* (1.45) for nose, and are also approximately the same when using *MDS* (1.75) to when using *DeCOr-MDS* (1.74) for throat. This decrease also suggests that orthogonal outliers get corrected and projected closer to the regular points.

3.5 Application to scRNA-seq data

We further evaluated *DeCOr-MDS* on single cell RNA-seq (scRNA-seq) data. In general, analyzing scRNA seq data requires dimensionality reduction for visualization (including MDS-based methods Canzar et al. (2021); Senabouth et al. (2019)), and specific quality control procedures to mitigate various technical artifacts McCarthy et al. (2017); Luecken and Theis (2019). We applied our method as a potentially relevant tool for this purpose. We applied *DeCOr-MDS* first on a dataset containing the expression level of 1063 cells for 2,601 genes (detailed in Methods Section 2.6). We found the dimension of the main subspace to be 3, with 77 (7%) outliers detected. In Figures 7A,B, we compared the embeddings in 3D using *MDS* and *DeCOr-MDS*. Similarly to the previous experiments, the mean distance between the orthogonal outliers and barycenter of regular points in the reduced subspace decreases when using *DeCOr-MDS* (from 4.22 to 2.51), improving the visualization of regular points. In Figure 7C, we further examined the drop-out rates (indicating zero count for a given gene) of the cells among the top 500highly expressed genes, determined by the median of counts per gene. Among these highly expressed genes, we identified 97.4% of the detected outliers that have drop-out rates greater than 0.95, while this was the case for 27.4% of the regular cells. Upon performing a pairwise *t*-test on the total counts for the top 500 highly expressed genes from the outlier group and the regular cell group, we found that the total counts are significantly different between the two groups (p-value < 0.001). Therefore, our method led to detect some outliers associated with high drop-out counts for highly expressed genes, which were not captured by the standard processing and quality control methods used in the first place.

4 Discussion

We proposed *DeCOr-MDS*, a novel approach using geometric characteristics to detect dimension, and to correct orthogonal outliers in high dimensional space. That is, to the best of our knowledge, the first statistical tool that addresses the challenge of the presence of orthogonal outliers in high dimensional space. We validated the method using synthetic datasets and demonstrated its potential applications to analyze biological datasets, including cell shape data, count arrays from microbiome data and scRNA-seq data. The visualization and numerical comparison confirmed that *DeCOr-MDS* effectively detects dimensionalities in many instances, corrects orthogonal outliers, and demonstrates superior performance to classical dimension reduction methods.

The notion of simplices is used frequently with the aim of robustness, either to detect the coreness of data [data depth and multivariate median, Liu (1990); Aamari et al. (2021)], or to detect outlying features [detection of extreme directions, Meyer and Wintenberger (2021)]. Simplices can also be used to build a flexible network of points for informative visualization (McInnes et al., 2018). Outlier detection and accommodation have been addressed by a wide array of methods, which can be broadly divided into three categories: 1) robust metrics (Spence and Lewandowsky, 1989; Cayton and Dasgupta, 2006; Oh et al., 2008; Shieh and Hung, 2009; Forero and Giannakis, 2012), 2) robust estimation (Mandanas and Kotropoulos, 2017), or 3) exploiting the characteristics of outliers (Forero and Giannakis, 2012; Blouvshtein and Cohen-Or, 2019). Our method resorts to both (3) by using the geometry of data, and 1) by using the median as centrality estimator. Our method also aims at estimating dimension. A common approach to do so is the screeplot (or elbow) test in principal components analysis, where a notable drop in the proportion of variance (or distance) explained can be taken as a cutoff, and as the most relevant dimension. High-dimensional biological datasets challenge this strategy, because fine-scale

structure confounds in practice downstream analyses. Because of this, authors often use an arbitrary large set of 10, or sometimes 20 or 50 components (Astle and Balding, 2009; Barfield et al., 2014; Demmitt et al., 2017; Sakaue et al., 2020; Arciero et al., 2021; Deng et al., 2021). Power analyses based on simulations also provide a way to assess an adequate number of components (Barfield et al., 2014). In this work, we proposed an alternative approach, by exploiting the structure of the dataset to determine essential versus non-essential dimensions.

Limitations of DeCOr-MDS include the non-automated choice of the cutoff parameter c. This parameter sets the maximum tolerated number of standard deviations σ before a point is considered an outlier. A value for $c = 3$, which corresponds approximately to the 0.1% most extreme points in a Gaussian distribution, may be selected, for instance. Dimension detection is also imperfect for heterogeneous datasets where the distribution of regular points (e.g. with distant clusters may prevent the height criterion for outlier detection to be effective. In this case a possible solution would be to first perform a clustering analysis (for instance k-means) to assess if the distance between clusters is comparable with the distance between the outlier and the main subspace, and if that's the case separately perform our method on each cluster. There are various potential directions to improve the dimension detection in real datasets of high dimension. This may be achieved by studying the behaviour of the Cayley-Menger determinant, which is central in the procedure, in higher dimensions. One may also associate the height criterion with a distribution criterion (Legrand, 2017), which would be sensitive to clusters or other notable structure, as was apparent in the HMP dataset. Another beneficial improvement would be to reduce computing time, for instance by implementing a parallelized version or using a call to a compiled program. Finally, one could optimize the cutoff parameter c automatically, either through a hyperparameter search, or by using a data-driven procedure, during the exploration phase of the algorithm.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/x5796.

Author contributions

Conceptualization: CL and KDD, data curation: WL, JM, NM, AP, CL, and KDD, funding acquisition: CL and KDD, formal analysis: JM, CL, and KDD, investigation: WL and JM, methodology: CL and KDD, resources: AP, CL, and KDD, software: WL, JM, NM, CL, and KDD, supervision: CL and KDD, validation: WL, visualization: WL, JM, CL, and KDD, writing–original draft: WL, JM, CL, and KDD, writing–review and editing: WL, CL, and KDD. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by a NSERC Discovery grant (PG 22R3468) and a MITACS PIMS fellowship.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1211819/full#supplementary-material

References

Aamari, E., Arias-Castro, E., and Berenfeld, C. (2021). From graph centrality to data depth. https://arxiv.org/abs/2105.03122.

Alizadeh, E., Xu, W., Castle, J., Foss, J., and Prasad, A. (2019). Tismorph: A tool to quantify texture, irregularity and spreading of single cells. *PLoS One* 14, e0217346. doi:10.1371/journal.pone.0217346

Arciero, E., Dogra, S. A., Malawsky, D. S., Mezzavilla, M., Tsismentzoglou, T., Huang, Q. Q., et al. (2021). Fine-scale population structure and demographic history of british pakistanis. *Nat. Commun.* 12 (1), 7189. doi:10.1038/s41467-021-27394-2

Astle, W., and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* 24, 451–471. doi:10.1214/09-sts307

Barfield, R. T., Almli, L. M., Kilaru, V., Smith, A. K., Mercer, K. B., Duncan, R., et al. (2014). Accounting for population stratification in dna methylation studies. *Genet. Epidemiol.* 38, 231–241. doi:10.1002/gepi.21789

Blouvshtein, L., and Cohen-Or, C. (2019). Outlier detection for robust multi dimensional scaling. *IEEE Trans. Pattern Analysis Mach. Intell.* 41, 2273–2279. doi:10.1109/tpami.2018.2851513

Brooks, A. W., Priya, S., Blekhman, R., and Bordenstein, S. R. (2018). Gut microbiota diversity across ethnicities in the United States. *PLoS Biol.* 16, e2006842. doi:10.1371/journal.pbio.2006842

Canzar, S., Do, V. H., Jelić, S., Laue, S., Matijević, D., and Prusina, T. (2021). Metric multidimensional scaling for large single-cell data sets using neural networks. *bioRxiv*, 1–16.

Cayton, L., and Dasgupta, S. "Robust euclidean embedding," in Proceedings of the 23rd International Conference on Machine Learning. Editors W. Cohen and A. Moore, 169–176. June 2006, Pittsburgh, Pennsylvania, USA, doi:10.1145/1143844.1143866

Demmitt, B. A., Corley, R. P., Huibregtse, B. M., Keller, M. C., Hewitt, J. K., McQueen, M. B., et al. (2017). Genetic influences on the human oral microbiome. *BMC Genomics* 18, 1–15. doi:10.1186/s12864-017-4008-8

Deng, S., Caddell, D. F., Xu, G., Dahlen, L., Washington, L., Yang, J., et al. (2021). Genome wide association study reveals plant loci controlling heritability of the rhizosphere microbiome. *ISME J.* 15, 3181–3194. doi:10.1038/s41396-021-00993-z

Forero, P. A., and Giannakis, G. B. (2012). Sparsity-exploiting robust multidimensional scaling. *IEEE Trans. Signal Process.* 60, 4118–4134. doi:10.1109/tsp.2012.2197617

Harmeling, S., Dornhege, G., Tax, D., Meinecke, F., and Müller, K.-R. (2005). From outliers to prototypes: Ordering data. *Neurocomputing* 69, 1608–1618. doi:10.1016/j.neucom.2005.05.015

Legrand, C. (2017). Exploring and controlling for underlying structure in genome and microbiome case-control association studies, Ph.D. thesis. Heidelberg, Germany: University of Heidelberg.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statistics* 18, 405–414. doi:10.1214/aos/1176347507

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: A tutorial. *Mol. Syst. Biol.* 15, e8746. doi:10.15252/msb.20188746

Mandanas, F. D., and Kotropoulos, C. L. (2017). Robust multidimensional scaling using a maximum correntropy criterion. *IEEE Trans. Signal Process.* 65, 919–932. doi:10.1109/tsp.2016.2625265

McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics* 33, 1179–1186. doi:10.1093/bioinformatics/btw777

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426.

Meyer, N., and Wintenberger, O. (2021). Sparse regular variation. *Adv. Appl. Probab.* 53, 1115–1148. doi:10.1017/apr.2021.14

Miolane, N., Caorsi, M., Lupo, U., Guerard, M., Guigui, N., Mathe, J., et al. (2021). Iclr 2021 challenge for computational geometry & topology: Design and results. https://arxiv.org/abs/2108.09810.

Miolane, N., Guigui, N., Le Brigant, A., Mathe, J., Hou, B., Thanwerdas, Y., et al. (2020). Geomstats: A python package for riemannian geometry in machine learning. *J. Mach. Learn. Res.* 21, 1–9.

Oh, J. H., Gao, J., and Rosenblatt, K. "Biological data outlier detection based on kullback-leibler divergence," in Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine (IEEE), Philadelphia, PA, USA, November 2008, 249–254.

Pukelsheim, F. (1994). The three sigma rule. *Am. Statistician* 48, 88–91. doi:10.2307/2684253

Rider, P. R. (1960). Variance of the median of small samples from several special populations. *J. Am. Stat. Assoc.* 55, 148–150. doi:10.1080/01621459.1960.10482056

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Lai Too, C., et al. (2020). Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nat. Commun.* 11, 1569. doi:10.1038/s41467-020-15194-z

Senabouth, A., Lukowski, S. W., Hernandez, J. A., Andersen, S. B., Mei, X., Nguyen, Q. H., et al. (2019). ascend: R package for analysis of single-cell rna-seq data. *GigaScience* 8, giz087. doi:10.1093/gigascience/giz087

Shieh, A. D., and Hung, Y. S. (2009). Detecting outlier samples in microarray data. *Stat. Appl. Genet. Mol. Biol.* 8, 1–24. doi:10.2202/1544-6115.1426

Sommerville, D. M. Y. (1929). *An introduction to the geometry of n dimensions*. Mineola, New York, United States: Dover Publications.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional anomaly detection. *IEEE Trans. Knowl. Data Eng.* 19, 631–645. doi:10.1109/tkde.2007.1009

Spence, I., and Lewandowsky, S. (1989). Robust multidimensional scaling. *Psychometrika* 54, 501–513. doi:10.1007/bf02294632

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2010). Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Analysis Mach. Intell.* 33, 1415–1428. doi:10.1109/tpami.2010.184

Tomassi, D., Forzani, L., Duarte, S., and Pfeiffer, R. M. (2021). Sufficient dimension reduction for compositional data. *Biostatistics* 22, 687–705. doi:10.1093/biostatistics/kxz060

Trevelline, B. K., and Kohl, K. D. (2022). The gut microbiome influences host diet selection behavior. *Proc. Natl. Acad. Sci.* 119, e2117537119. doi:10.1073/pnas.2117537119

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi:10.1038/nature06244

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15–5. doi:10.1186/s13059-017-1382-0

Zhou, Q., Deng, J., Pan, X., Meng, D., Zhu, Y., Bai, Y., et al. (2022). Gut microbiome mediates the protective effects of exercise after myocardial infarction. *Microbiome* 10, 82–19. doi:10.1186/s40168-022-01271-6