ISSN 1973-2201 *STATISTICA, anno LXXXIV, n. 4, 2024*
DOI: https://doi.org/10.60923/issn.1973-2201/20506

SURPRISING GEOMETRICAL PROPERTIES OF HIGH-DIMENSION LOW-SAMPLE SIZE DATA WITH DEVASTATING CONSEQUENCES FOR DATA ANALYSIS

Ana Maria Pires [1]
Department of Mathematics and CEMAT, IST, Universidade de Lisboa, Lisboa, Portugal
João António Branco
Department of Mathematics and CEMAT, IST, Universidade de Lisboa, Lisboa, Portugal

SUMMARY

The advent of modern technology, permitting the measurement of thousands of variables simultaneously, has given rise to floods of data characterized by many large or even huge datasets. This new paradigm presents extraordinary challenges to data analysis and the question arises: how can conventional data analysis methods, devised for moderate or small datasets, cope with the complexities of modern data? The case of high-dimension low-sample size data is particularly revealing of some of the drawbacks. We look at the case where the number of variables measured in an object is at least the number of observed objects and conclude that (under the further assumptions that the data are observations from continuous random variables and that linear combinations of the variables are meaningful operations) this configuration leads to geometrical and mathematical oddities and is an insurmountable barrier for the direct application of traditional methodologies. If scientists are going to base their conclusions on high-dimension low-sample size data, ignoring fundamental mathematical results arrived at in this paper and blindly use software to analyze data, the results of their analyses may not be trustful, and the findings of their experiments may never be validated. That is why new methods together with the wise use of traditional approaches are essential to progress safely through the present reality.

Keywords: Curse of dimensionality; High-dimension low-sample size data; Mahalanobis distance; Multivariate outliers; Nearest-neighbors; Projection-pursuit.

1. INTRODUCTION

When n "objects" (patients, subjects, cells, samples, etc) are measured on p distinct, possibly correlated, variables (or features) we say we have a multivariate p-dimensional

[1] Corresponding Author. E-mail: ana.maria.n.pires@gmail.com

dataset. Examples can be found in all areas of science as well as in many current human activities. Statistical methods to deal with this kind of datasets were developed all along the twentieth century mostly under the assumption that the number of variables is much smaller than the number of observations (see e.g., Johnson and Wichern, 2007). As notorious exceptions we can mention the suite of partial least squares methods developed by H. Wold and S. Wold, which have been routinely used in chemometrics for more than 40 years (Sjöström *et al.*, 1983; Frank and Friedman, 1993; Wold *et al.*, 2001; Castro-Reigía *et al.*, 2024). However, the automatic acquisition of data, due to the extraordinary development of new technologies observed in the last decades, has changed this paradigm, and nowadays it is quite common to have datasets with a number of variables much larger than the number of observations, in many subject areas other than chemometrics. A variety of examples of such high-dimension low-sample size datasets can be found in genomics (in a recent example, Shen *et al.*, 2024, base their conclusions on a dataset with $n = 108$ and $p = 135239$), astronomy, climate or finance, to name just a few subjects. How are we dealing with this new scenario? A few references (namely, Clarke *et al.*, 2008; Johnstone and Titterington, 2009; Bickel *et al.*, 2018; Chakrabarti and Sen, 2019; Zhang *et al.*, 2023) acknowledge the difficulties encountered and recognize our ignorance about the basic properties of high-dimension low-sample size data spaces. Despite this ignorance, a continuous stream of new methods claiming to be able to deal with this kind of data has flooded the scientific literature. As examples we can cite methods described in Lee and Cook (2010); Pires and Branco (2010); Cai *et al.* (2015); Liu *et al.* (2017); Sarkar *et al.* (2020); Cavalheiro *et al.* (2024), or methods implemented in the `R` package `rrcovHD` (R Core Team, 2024; Todorov, 2024).

In this paper we present some new mathematical results about the geometry of high-dimension low-sample size datasets, which reveal interesting fatal features that have been vastly ignored so far. The consequences of these findings are determinant for the correct approach to analyze high-dimension low-sample size data. As a matter of fact, it will become clear that it is nonsense to freely use many of the traditional data analysis methods directly to study that type of data. Of particular concern are, according to our results, methods that are based on distances between two points, and also methods that search for orthogonal projections optimizing given criteria. The later are often referred to as "projection-pursuit" (PP) methods. A long standing reference on PP has been Huber (1985), who claims *"The most exciting feature of PP is that it is one of the very few multivariate methods able to bypass the "curse of dimensionality" caused by the fact that high-dimensional space is mostly empty"*. The fact that the great expectations about PP have never been fulfilled may be explained by one of our main results, also corroborated, as we discuss later, by results of Bickel *et al.* (2018), who say in a cautionary note *"…with a limited amount of high-dimensional data, the results of projection pursuit and related ICA methods should be interpreted with great care."*

This paper is organized as follows. In Section 2 we establish the background, describe the problem and present the main findings, together with a few simple examples. The consequences of the results and possible solutions are discussed in Section 2. Mathematical details and proofs are given in Appendices A and B. `R` code is in Appendix C.

2. Geometric aspects of multivariate data

2.1. When p is small

In this case n is usually much larger than p ($p \ll n$). If, for instance, $p = 2$ the data can be represented by n points in a bivariate scatter plot, where the two axes represent the two variables. The main features of the data (relationship/correlation between the variables, grouping of objects, outliers, and others) can be visualized on this scatter plot. When $p = 3$ things are not so easy. We may produce a 3-d scatter plot, but this scatter plot is ultimately represented in a two dimensional space (computer screen/sheet of paper) and what we really observe is a projection of the 3-d point cloud onto a 2-d subspace. This can be informally described as a picture of the point cloud taken from a given position and angle. To understand a 3-d dataset we must take many pictures from varying positions and angles, which means that visualization of the data is, no doubt, more difficult when $p = 3$ than when $p = 2$. What about when $p > 3$? Conceptually we could always apply a similar procedure, that is, project the p-dimensional data points onto 2-d subspaces. It is easy to understand that, as the number of dimensions increases, we would need an increasingly large number of pictures to get just a glimpse of the data. This is yet a manifestation of the "curse of dimensionality" (Bellman, 1957).

At this point we must recognize the need and importance of multivariate statistical methods. These methods are designed to analyze datasets with n observations on p variables organized in a data matrix, $X = \{x_{ij}\}_{i=1,\ldots,n;j=1,\ldots,p}$, with n rows and p columns, where x_{ij} denotes the value of the jth variable for the ith object. We restrict our attention to numerical variables and we assume throughout that the n points are in "general position", which means that there are no redundancies in the data, like, for instance, two exact replicas of an observation (see definition in Appendix A).

Multivariate statistical methods (considered in a broad sense, i.e., including machine learning and related topics) can extract and quantify relevant properties of the data, and also produce, in many cases, two dimensional graphical representations to complement the analysis. Most multivariate statistical methods use matrix algebra techniques. For instance, principal components (Hotelling, 1933), which are successive uncorrelated directions with maximal variance, are defined by the eigenvectors of the covariance matrix, S, or of the correlation matrix, R, whereas the variances along the principal components can be obtained by computing the corresponding eigenvalues. A very important tool in multivariate analysis, which is related to a number of methods, including principal components, is the Mahalanobis distance (Mahalanobis, 1936).

When comparing or analyzing data on a single variable it is often informative to compute the distance of each observation to the center of the data (usually identified by the sample mean) or the distance between two observations, taking into consideration the intrinsic variability of the data, usually measured by the standard deviation. Recall that for a dataset consisting of n observations on one variable, $(x_1, \ldots, x_n)$, the z-scores are $z_i = (x_i - \overline{x})/s$, where $\overline{x}$ and s are, respectively, the arithmetic mean and standard deviation of the n observations. The z-scores have been used for instance to detect out-

liers, though they have a number of important limitations (Barnett and Lewis, 1994). The mean of $(z_1,\ldots,z_n)$ is zero while both its standard deviation and variance are equal to 1. Also, $|z_i| = \{(x_i - \overline{x})^2/s^2\}^{1/2}$ is both the standardized distance between x_i and $\overline{x}$ and the Euclidean distance between z_i and $\overline{z} \equiv 0$, whereas the Euclidean distance between z_i and z_j, $|z_i - z_j|$, is equal to $\{(x_i - x_j)^2/s^2\}^{1/2}$, the standardized distance between x_i and x_j.

Consider now a multivariate $n \times p$ dataset and imagine a standardizing transformation with similar properties as the z-scores transformation. Computing the z-scores separately for each variable does not achieve the standardization of a multivariate dataset unless all the variables are uncorrelated (that is, in case both S and R are diagonal matrices). The multivariate transformation which standardizes a multivariate dataset, called the Mahalanobis transformation, takes into account the variances and covariances between all the variables and is defined in matrix notation by $z_i = S^{-1/2}(x_i - \overline{x})$, where x_i is a vector containing the p measurements of the ith object, $\overline{x}$ is the mean vector, containing the p means of the individual variables and $S^{-1/2}$ is a square root of the inverse of S. This transformation has the same form of the z-scores transformation, but uses matrices in place of single values. It is easy to verify that the standardized data matrix, Z, formed with the z_i vectors, is such that all the variables have zero mean and unit variance, and all pairwise covariances are zero (in other words, the mean vector of Z is the null vector, and both its covariance and correlation matrices are the identity matrix, I). The Euclidean distance between a standardized observation, z_i, and the null vector is $\left(\sum_{j=1}^{p} z_{ij}^2\right)^{1/2}$, which can be written in matrix notation as

$$\left(z_i^T z_i\right)^{1/2} = \left\{(x_i - \overline{x})^T S^{-1} (x_i - \overline{x})\right\}^{1/2},$$

and defines the Mahalanobis distance between the observation x_i and the mean of the n observations, $d_{x_i,\overline{x}}$. The Euclidean distance between two standardized observations, z_i and z_j, is

$$\left\{\left(z_i - z_j\right)^T \left(z_i - z_j\right)\right\}^{1/2} = \left\{\left(x_i - x_j\right)^T S^{-1} \left(x_i - x_j\right)\right\}^{1/2},$$

and defines the Mahalanobis distance between x_i and x_j, d_{x_i,x_j}. Mahalanobis distance is widely used to define statistical depth functions, such as Mahalanobis Depth, which play a significant role in multivariate analysis. Like the z-scores, Mahalanobis distances to the mean have been used to detect outliers in multivariate datasets (Gnanadesikan and Kettenring, 1972), a relevant task in "anomaly detection" problems. Mahalanobis distances between two objects are useful, for instance, in clustering applications, or for nearest-neighbor algorithms.

2.2. *When p is larger than n*

Before we discuss the very complicated issue of the visualization of this kind of data, let us consider an apparently naive question: can we use Mahalanobis distances when the number of variables is larger than the number of observations? A first quick answer would be: no, because when $p \geq n$ the covariance matrix can not be inverted (S is singular) and, therefore, Mahalanobis distances are not defined. However, it is still possible to define Mahalanobis distances by reasoning as follows. Three points (in general position) in a 3-d space ($p = 3$, $n = 3$) define a plane, in other words, they lie on a certain 2-d subspace. If we adopt a coordinate system in that subspace, Mahalanobis distances can be computed, because the new covariance matrix will not be singular. A similar argument can be applied to higher-dimensional spaces. A dataset with n observations in p variables, with $p \geq n$, can be represented, without any loss of information, in a new set of $q = n - 1$ variables, for which Mahalanobis distances can be computed. Moreover, such a set of variables is easy to find. Consider, for instance, the $n - 1$ principal components of X corresponding to the non-null eigenvalues of S. These new variables are linear combinations of the original variables and it is easy to move from one system to the other. Therefore, we conclude that we can still standardize datasets with $p \geq n$, by applying the Mahalanobis transformation in the smallest subspace where the data is fully contained and the covariance matrix is invertible, that is, using $q = n - 1$ new variables which are linear combinations of the original p variables. Mahalanobis distances defined in this way are equivalent to generalized Mahalanobis distances (Mardia, 1977), for which the non-existing inverse is replaced by the Moore-Penrose pseudoinverse, and, for mathematical convenience, we are going to use this form in the definitions.

When we perform the computations just described we arrive at the following surprising result (details and proofs are given in Appendix B).

THEOREM 1. *Let* $D(x, y; S^*) = \left\{(x-y)^T S^* (x-y)\right\}^{1/2}$, *where* x *and* y *are* p*-dimensional vectors and* S^* *is a symmetric positive semi-definite* $p \times p$*-matrix.*

For every dataset with n *points,* $x_1, \ldots, x_n$, *and* $p \geq n - 1$ *variables we have that:*

(i) $d_{x_i, \overline{x}} = D(x_i, \overline{x}; S^*) = (n-1)\, n^{-1/2}$, *for every* $i = 1, \ldots, n$, *and*

(ii) $d_{x_i, x_j} = D\left(x_i, x_j; S^*\right) = \{2(n-1)\}^{1/2}$, *for every* $i \neq j = 1, \ldots, n$,

with $S^* \equiv S^{-1}$, *if the covariance matrix of* $(x_1, \ldots, x_n)$, S, *is invertible, or* $S^* \equiv S^-$, *the Moore-Penrose pseudoinverse of* S, *otherwise.*

In other words, whatever the points $x_1, \ldots, x_n$, *as long as* $p \geq n - 1$, *the standardized data always form a regular pattern in which the distance from every point to the center is a constant and the distance between any two points is another constant, both constants depending only on* n.

Despite Mahalanobis distances having been known and used for almost 90 years, we have not been able to find these simple results in the literature.

Incidentally, we are also able to show, following almost the same proof as in Theorem 1, that for every dataset, irrespective of the number of observations or variables, we have that (see proof in Appendix B)

(iii) $d_{x_i,\bar{x}} \leq (n-1)\, n^{-1/2}$, for every $i = 1,\ldots,n$, and

(iv) $d_{x_i,x_j} \leq \{2(n-1)\}^{1/2}$ for every $i \neq j = 1,\ldots,n$.

A proof of (iii) was published in Trenkler and Puntanen (2005) and also in Gath and Hayes (2006).

Under the conditions of Theorem 1, *(ii)* implies that every observation is the nearest-neighbor of every other observation. This is in line with asymptotic properties described in Hinneburg *et al.* (2000).

The results given in Theorem 1 can also be connected to asymptotic results in Hall *et al.* (2005) and Ahn *et al.* (2007).

The present choice of the covariance matrix ($S^* \equiv S^-$) is natural and has the power of revealing singular properties of the space, with the consequences shown along the text. There may be other alternatives in the literature leading to different results and consequences. However, that does not mean that the properties revealed by Theorem 1 and their consequences are going to disappear.

Figure 1 illustrates Theorem 1 for datasets with 3 points and a number of variables larger or equal than 2. This is the only case we can represent directly in a 2-*d* plot (the plane containing the three points), and we can see that every dataset with 3 points in $p \geq 2$ variables corresponds, when standardized and apart from an arbitrary rotation, to the "same" equilateral triangle (the general position assumption excludes cases where the 3 points lie on a straight line). Similarly, for $n = 4$ points in $p \geq 3$ variables, every dataset corresponds, when standardized, to the "same" regular tetrahedron (again apart from an arbitrary rotation). For $n \geq 5$ we have to imagine a regular geometric object in a $p \geq n-1 \geq 4$ dimensional space which is the appropriate member in the sequence: equilateral triangle, regular tetrahedron, ... These objects are called regular simplices (triangle $\equiv$ 2-simplex; tetrahedron $\equiv$ 3-simplex; pentachoron $\equiv$ 4-simplex; ...; $(n-1)$-simplex; ...). By the above result, we can envisage the standardized version of an $n \times p$ data matrix, with $p \geq n-1$, as the n vertices of a regular $(n-1)$-simplex, such that the length of every edge is $d_{x_i,x_j} = \{2(n-1)\}^{1/2}$, and every vertex is located at a distance of $d_{x_i,\bar{x}} = (n-1)\, n^{-1/2}$ from the center of the simplex. This is a very simple regular structure with all the points at the boundary of its convex hull (the simplex) while the interior of the convex hull is completely empty. This last property is also shared by the original data, because they can be seen as a non-singular linear transformation of the standardized data: $x_i = S^{1/2} z_i + \bar{x}$.

We have thus shown that the point cloud of a multivariate dataset with $p \geq n-1$ is like an empty shell which, when standardized, looks the same whatever the data. This

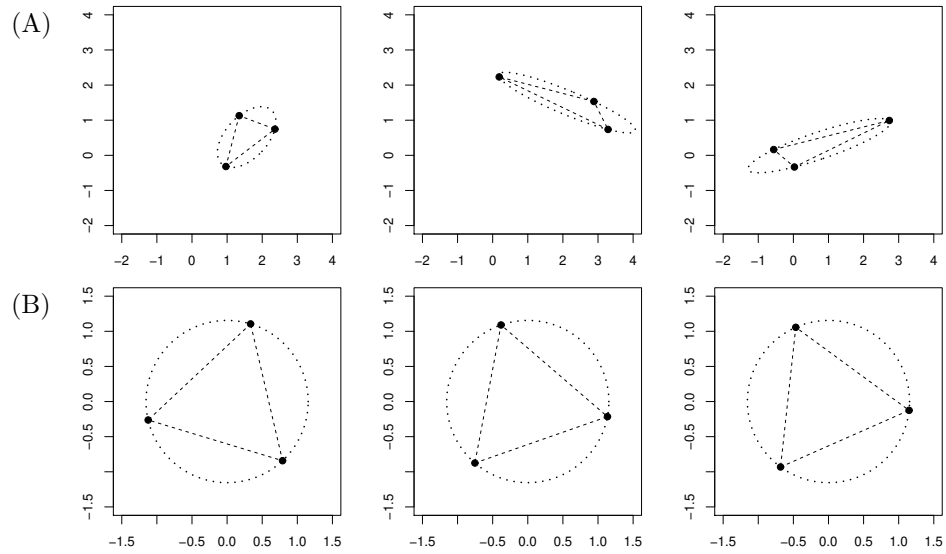

Figure 1 – Three datasets with 3 random points in 2 dimensions. (A) Original observations, x_1, x_2 and x_3. (B) Standardized observations, $z_i = S^{-1/2}(x_i - \overline{x})$, $i = 1, 2, 3$.

reality has strong implications. If we insist on considering the multivariate relation between all the variables (i.e., if we accept that computing Mahalanobis distances and linear projections onto a subspace are meaningful operations), it is not possible, unless external information is provided, to: (i) separate outliers from non outlying observations, (ii) detect any kind of deviations from symmetric structures, (iii) distinguish between linear and non linear structures, (iv) identify any type of clustering, or (v) trust nearest-neighbor algorithms.

Let us now consider projections of these data structures, that is, imagine traveling "around" the point cloud and taking a large number of interesting pictures. The next result shows that these pictures can show virtually anything we want to see and have, therefore, to be used very carefully when extracting conclusions about the data (again, details and proofs are given in Appendix B).

THEOREM 2. *For every data matrix with n points and $p \geq n-1$ variables, X, and every non-singular two dimensional arrangement of n points, Y, it is always possible to find (explicitly) an orthogonal projection of X which is "similar" to Y.*

The result given in Theorem 2 can be connected to some results in the mathematical literature (Baryshnikov and Vitale, 1994; Eastwood and Penrose, 2000). Remarkably, the "piling effect" (Ahn and Marron, 2010), characterized by the existence of directions

such that projections of data onto those directions take only two distinct values, is just a special case of Theorem 2.

Figure 2 illustrates Theorem 2 using two well known datasets: the colon cancer microarray data (Alon *et al.*, 1999) and the Olivetti Research Laboratory (ORL) face database (Samaria and Harter, 1994). Note that in Theorem 2 there is a one to one correspondence between the points in X and in Y, $x_1 \to y_1, \ldots, x_n \to y_n$, which means that we can obtain, by permutation of the Y labels, similar visual arrangements with different local structure, such as nearest-neighbors. In other words, closeness of the points can be changed under arbitrary projections.

The implications of Theorem 2 for data analysis are the same as of Theorem 1: if there is no external information, there is little we can conclude about the multivariate structure of an high-dimension low-sample size dataset. For instance, does a projection where all the observations but one are projected onto the same point, provide evidence that the observation projected onto the isolated point is an outlier? It can not be, because we know, from Theorem 2, that we can find such a projection for every observation of every high-dimensional low-sample size dataset, irrespective of the presence of any outlying observation.

Theorem 3 below clarifies the claims made in the previous paragraph, bringing further theoretical insights into the subject of outlier detection for high-dimension low-sample size datasets (as before, details and proofs are given in Appendix B).

THEOREM 3. *Let X be a data matrix with n points, $x_1, \ldots, x_n$, on p variables, and consider the outlyingness measure defined for every $i \in \{1, \ldots, n\}$ as:*

$$\text{out}(x_i, X; m, s) = \sup_{\alpha \in \mathbb{R}^p, \|\alpha\|=1} \frac{|\alpha^T x_i - m(\alpha^T x_1, \ldots, \alpha^T x_n)|}{s(\alpha^T x_1, \ldots, \alpha^T x_n)}, \tag{1}$$

where m is a location measure and s is a scale measure. The outlyingness values depend on (m, s) and on the relation between n and p, as follows:

(m, s)	$p < n-1$	$p \geq n-1$
(mean, st.dev.)	$d_{x_i, \bar{x}}$ *(function of X and i)*	$d_{x_i, \bar{x}} \left(\equiv (n-1)/\sqrt{n}\right)$
(median, mad)	*not explicit (function of X and i)*	$+\infty$ *(for all X and i)*

The outlyingness measure defined by Eq. (1) when (m, s)=(median, mad) is known in the literature as the "Stahel-Donoho outlyingness measure". It was proposed independently by Stahel (1981) and Donoho (1982), as a robust "analogue" of the Mahalanobis distance (from an observation to the center of the data).

The fact that the Stahel-Donoho outlyingness measure is equal to $+\infty$, for all X and i, whenever $p \geq n-1$, shows how wild the high-dimension low-sample size space is (All observations of all datasets are outliers!).

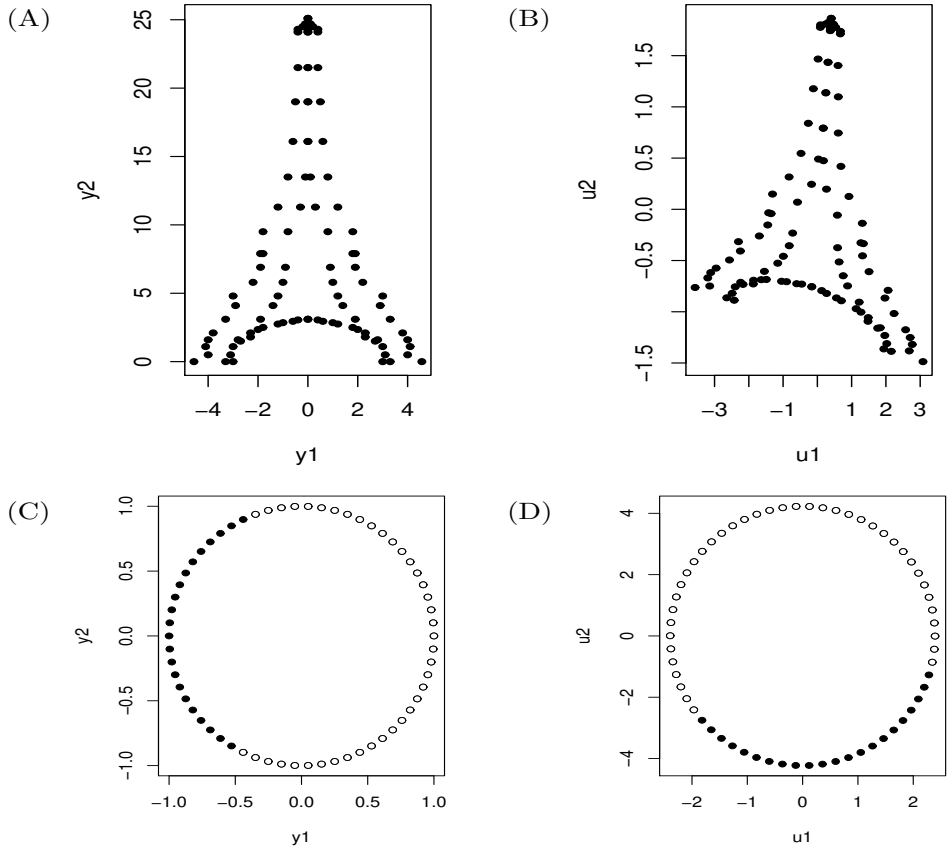

Figure 2 – Illustration of Theorem 2. (A) and (C) Arbitrary configurations of points in 2 dimensions. (B) An orthogonal projection of high-dimension low-sample size data which is similar to the configuration in (A), where the data comes from the ORL face database with $n = 400$ observations (pictures) on $p = 10304$ variables (92×112 pixels). (D) An orthogonal projection of high-dimension low-sample size data which is similar to the configuration in (C), where the data comes from the colon cancer microarray dataset with $n = 62$ observations (arrays) on $p = 2000$ variables (genes); the black circles correspond to the 22 normal tissue samples and the white circles correspond to the 40 tumor tissue samples (the arbitrary configuration in (C) specifies separation of the two groups).

Because the Stahel-Donoho outlyingness measure can not be computed explicitly, several numerical algorithms have been proposed and implemented in software. But, in view of Theorem 3, the results of such algorithms can only be arbitrary if applied to datasets with $p \geq n-1$.

The Stahel-Donoho outlyingness measure has often been used in the development of robust multivariate statistical methods. A big issue is that some of those methods have been applied to, and are even recommended, when $p \geq n-1$ (e.g., in Filzmoser *et al.*, 2020, one can read: *"the Stahel-Donoho estimator can handle data sets with more variables than observations"*. In the light of Theorem 3, how can this be?).

It is clear now that both Theorem 1 and Theorem 2 imply that, for $p \geq n-1$, we can not distinguish outliers from non-outliers, which in turn implies that it is not possible to build robust methods. This conclusion, how odd as it may sound, is in line with a result from Tyler (2010), who shows that it is not possible to build robust affine equivariant estimators of multivariate location and scatter.

We now know that the $p \geq n-1$ data space is quite odd. It is then natural to ask whether those singularities appear suddenly when p reaches $n-1$ or whether the properties of the space start changing gradually as p approaches $n-1$ from below. The upper bound of the Mahalanobis distance, together with the fact that the expected value of $d^2_{x_i,\bar{x}}$ is $(n-1)p/n$ for $p \leq n-1$, whatever the distribution of the data (Mardia, 1977), show that a transition must start far before p reaches $n-1$ (this may also be related to Corollary 1.1 in Donoho and Tanner, 2005, which states a rigorous result about such transitions for samples from a multivariate normal distribution). Figure 3 illustrates this aspect. The plots show orthogonal projections of the colon cancer microarray dataset, selected to be as similar as possible, in the sense of least squares, to configuration (C) of Figure 2, but using only a subset with p randomly chosen variables. For the plots on the top row, with $p \geq n-1$ ($= 61$), Theorem 2 still applies, and it is possible to find orthogonal projections onto a two dimensional space which replicate a given configuration. On the contrary, for the remaining plots on Figure 3, all with $p < n-1$, it is no longer possible to replicate any configuration. However, certain aspects of the given structure remain visible, even in the case $p = 10$, where the separation of the two groups is already noticeable. This example shows that it may be dangerous to interpret motifs seen in projections if the ratio n/p is not large enough, even though it may be greater than 1.

Taking into account the results presented one should investigate many recent scientific discoveries, in various subject areas, relying on the analysis of high-dimension low-sample size datasets. This can be easily done by (i) generating a number of artificial datasets similar to the dataset under study, for instance, with the same number of variables and observations, and with equal means and variances/covariances/correlations, (ii) studying those artificial datasets in the same manner as the original data, (iii) comparing the results. One example of the application of this strategy is presented next, but others could be given.

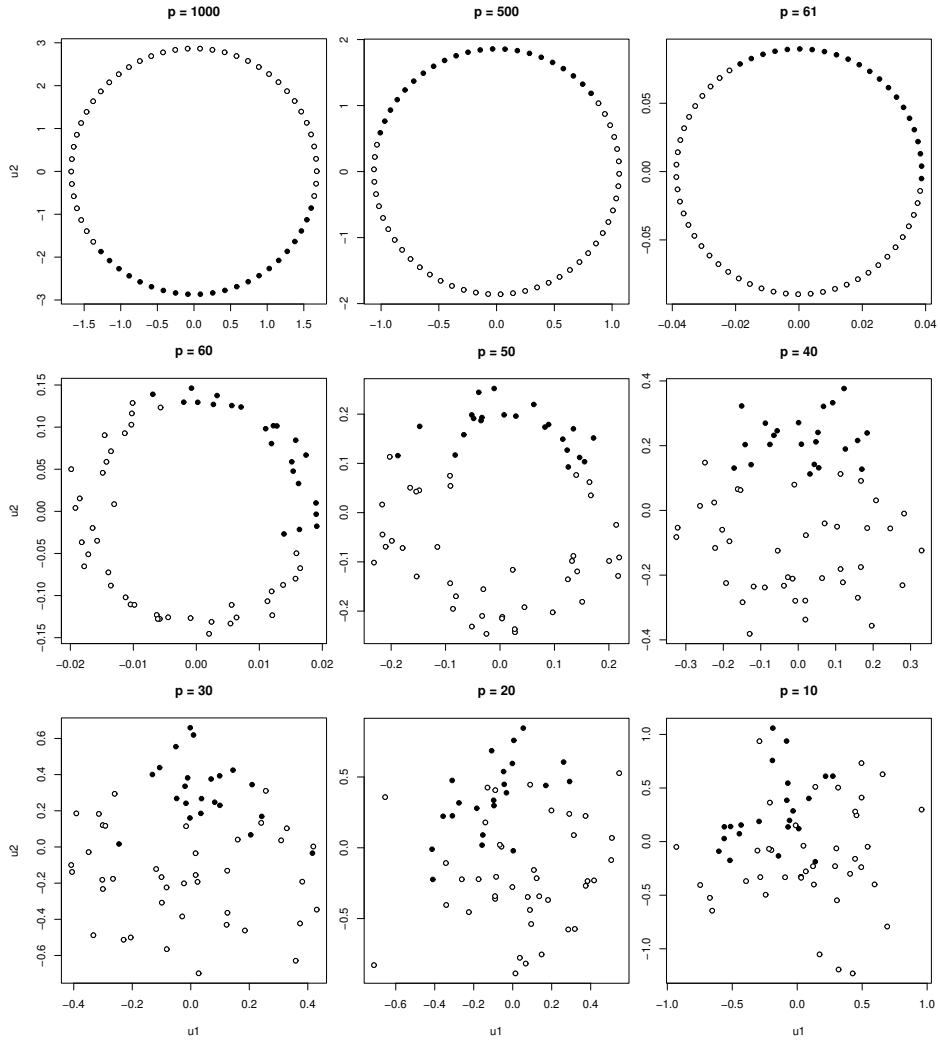

Figure 3 – Nine orthogonal projections, selected to be as similar as possible to configuration (C) of Figure 2, obtained from subsets of the colon cancer microarray data (*p* randomly chosen variables).

2.3. Example

In this example we consider again the colon cancer microarray dataset and assess the performance of several outlier detecting methods advertised as appropriate for high-dimension including high-dimension low-sample size datasets. There is an extensive literature about the outliers of this dataset (according to Shieh and Hung, 2009, and the references therein).

The methods selected for this study are: the ROBPCA of Hubert *et al.* (2005), the projection-pursuit based methods of Croux and Ruiz-Gazen (2005) and Croux *et al.* (2007), denoted respectively, PPPROJ and PPGRID, three methods proposed in Filzmoser *et al.* (2008), denominated by the authors as PCOUT, SIGN1 and SIGN2, the method proposed in Shieh and Hung (2009), referred to as SH, and, finally, the classical baseline method which uses Mahalanobis distances in the space of the first k principal components (CLA). All these methods are described in Filzmoser and Todorov (2013) and implemented in the `R` package `rrcovHD` (R Core Team, 2024; Todorov, 2024). All the methods selected require parameters, whose values have been fixed to make the methods comparable, in terms of effective number of dimensions used (k) and target level (α, proportion of false positives for normal data, approximately 0.05). There are two further methods in `rrcovHD`, the old method of Locantore *et al.* (1999), which was not selected due to issues related with the computation of the target level, and the sparse PCA method proposed by Croux *et al.* (2013) which was not selected because of being extremely slow. More details about this example, including the `R` code and outputs used to obtain the results shown here, are provided in Appendix C. The number of outliers pointed out by each method in each group are shown in Table 1 (in the lines identified by Case = Data). We see that all the methods find some outliers, but the specific observations pointed out vary with the method, as can be confirmed in Appendix C.

TABLE 1

The number of outliers detected by each method in each group of the colon cancer microarray dataset (Data). The median of the number of outliers detected by each method in each group over 500 simulations of randomly generated data from normal multivariate distributions, with parameters similar to the empirical parameters of the real data (Sim0).

Case	n	CLA	ROBPCA	PCOUT	SIGN1	SIGN2	SH	PPGRID	PPPROJ
Data	22	2	5	4	8	4	3	2	3
Sim0	22	0	6	4	9	2	2	1	2
Data	40	3	10	6	21	8	6	3	7
Sim0	40	1	12	6	18	3	2	2	2

To show that the flagged outliers may well be illusions, created by the high-dimension low-sample size, we simulated 500 datasets with multivariate normal distribution but similar to the original colon cancer dataset (that is, with two groups, same number of variables and observations, same means and covariances per group). We then applied the eight outlier detecting methods to each dataset and registered the number of outliers

detected. The medians of the numbers of detections observed are also shown in Table 1 (in the lines identified by Case = Sim0). These figures are far too large than would be anticipated and are not too different, depending on the method, from the number of outliers "detected" in the real data. A deeper analysis of the results is in order.

As mentioned above, the probability of declaring an observation from a normal distribution as an outlier, is pre-specified at $\alpha = 0.05$ for all the methods. Thus, the number of outliers detected in a sample of n observations from a normal distribution follows a binomial(n, α) distribution, at least approximately (due to several distributional approximations involved when translating the target level into cut-off values). Therefore, the sample containing the number of outliers found in each of the simulated datasets is expected to behave approximately as a sample with 500 observations from a binomial(n, 0.05). From now on we refer only to the part of the simulation with $n = 40$, as the conclusions for $n = 22$ were similar.

The plots in the left column of Figure 4 show the observed frequency distributions of the number of outliers detected with all the variables ($p = 2000$) included. The '+' symbol in the plots indicate the expected frequencies under the binomial(40,0.05) distribution. All the observed frequencies are different from the expected frequencies, and, while part of the differences may be related to the approximations mentioned in the previous paragraph, we have to conclude that some detection methods are not working as they should, namely ROBPCA, PCOUT and SIGN1. As a consequence we can not trust the results provided by these methods when applied to the real dataset.

We are convinced that the responsibility for the failure of some of the outlier detection methods can be ascribed, at least partially, to the high-dimension low-sample size nature of the data. To support this claim we repeated the simulation with normal data for a much smaller number of variables. The illustration described in Figure 3 shows that $p = 10$ may still be too large when $n = 40$, so we have selected randomly $p = 5$ variables from the original 2000. The plots with the number of outliers detected are shown on the right column in Figure 4.

We conclude that, except for the PCOUT method, all the observed frequencies have moved closer to the expected frequencies and in the case of ROBPCA and SIGN1 there is in fact a fantastic recovery. In this low dimensional situation all the methods but PCOUT are in agreement and doing more or less close to what they are supposed to do.

As we wondered what could possibly justify the observed behaviour of the eight methods we looked into the algorithmic details of each method and can then add the following explanations for the results.

PCOUT: according to Filzmoser *et al.* (2008), PCOUT was designed to work in high dimensional cases. This explains why it does not work very well in low dimensions. However, on the high-dimensional situation, the method uses robust versions of the Mahalanobis distance, which, on the light of Theorem 3 do not exist and, therefore, explains the problems encountered.

ROBPCA and SIGN1: when $p \geq n$ both methods work internally with the data rotated to its proper subspace (of dimension $n-1$) and then look for projections showing outliers. For that, they rely on a sampling algorithm to compute approximations of the

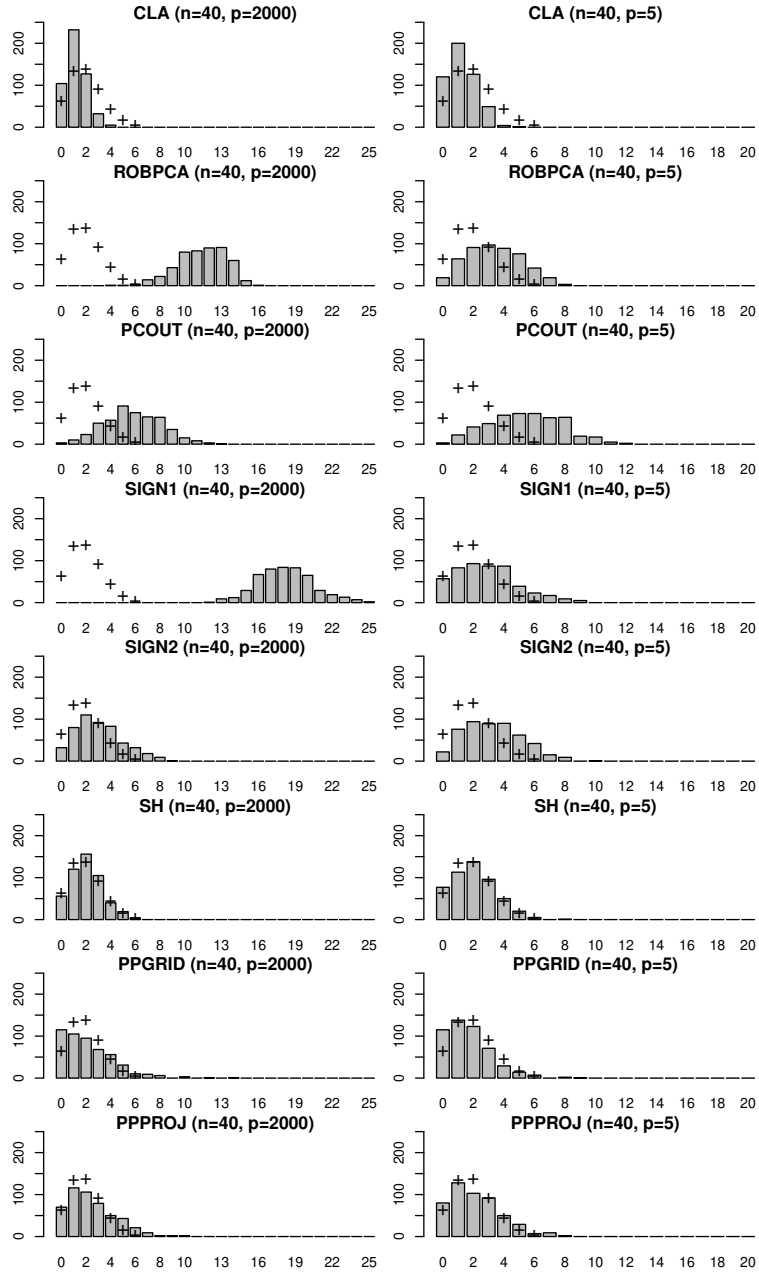

Figure 4 – Results of the detection of outliers by various methods (see text for details) in 500 simulated multivariate normal data sets with $n = 40$ observations and $p = 2000$ variables (left) or $p = 5$ variables (right). The '+' signs represent expected frequencies under a binomial(40, 0.05) distribution.

Stahel–Donoho outlyingness measure. As Theorem 3 shows this is completely useless for $p \geq n-1$. When the number of variables is small ($p \ll n$) there are no such issues and both methods work as expected.

SIGN2 and SH: SIGN2 is similar to SIGN1, and SH is similar to ROBPCA. However, in both cases a dimension reduction by principal components is the first step of the procedure, which is then followed by outlier detection in a reduced space (the number of principal components to retain is chosen by fixing a proportion of explained variance, in the case of SIGN1, or by an automatic selection method based on the scree plot, in the case of SH). By bringing the high-dimension low-sample size situation to the usual $p \ll n$ the difficulties described above are avoided, and, therefore, the methods work as expected no matter the dimensionality. The same argument explains the reasonable and similar behaviour of CLA, PPGRID and PPPROJ.

To have a better understanding of the problem we must get an idea of how the methods compare in terms of power. So we repeated the simulation, inserting one shift outlier in each dataset, by adding a fixed constant $\delta = 3$ to all the variables of every first observations. We then computed estimates of the power and level for the detection of this single outlier. Results are shown in Tables 2 and 3, on the lines identified by Case = Sim1.

TABLE 2

Level: proportion (in %) of false positives for each method, in the group with $n = 40$, over 500 simulations of randomly generated data from normal multivariate distributions, with parameters similar to the empirical parameters of the real data (Sim0: without outliers; Sim1: with one generated outlier).

Case	p	CLA	ROBPCA	PCOUT	SIGN1	SIGN2	SH	PPGRID	PPPROJ
Sim0	2000	3.01	28.47	14.54	45.48	7.61	5.13	5.52	6.12
Sim1	2000	2.98	26.24	12.96	44.95	7.11	4.48	6.15	5.97
Sim0	5	3.10	8.41	13.52	7.23	8.20	5.07	4.25	5.38
Sim1	5	2.58	7.51	13.12	6.66	7.70	4.61	3.93	5.18

TABLE 3

Power: proportion (in %) of true positives for each method, in the group with $n = 40$, over 500 simulations of randomly generated data from normal multivariate distributions, with parameters similar to the empirical parameters of the real data (Sim1: with one generated outlier).

Case	p	CLA	ROBPCA	PCOUT	SIGN1	SIGN2	SH	PPGRID	PPPROJ
Sim1	2000	72.8	98.8	96.2	99.2	60.0	90.8	57.6	84.6
Sim1	5	75.6	80.6	85.2	87.8	85.4	78.4	82.8	81.2

The results in Table 2, for the empirical level when there are no outliers (Sim0), are

in accordance with the proportions of outliers detected in each case that can be inferred from the plots in Figure 4. When the simulated data contains one outlier (Sim1), we observe approximately the same behaviour. From Table 3 we can conclude that all methods have some reasonable power in detecting a single outlier, at least under this experimental setup. However, the results for ROBPCA, PCOUT and SIGN1, can not be considered superior to the other methods, as the level is not respected. For SIGN2 and PPGRID a degradation is observed when the number of variables is increased from 5 to 2000. SH and PPPROJ emerge as the most powerful methods. Due to obvious limitations of the simulation study we do not claim that its conclusions can be generalized to other datasets.

This example clearly shows that knowledge of the geometric properties of high-dimension low-sample size data, in particular of the theoretical results given in this paper, is of crucial importance for the correct development and application of statistical methods, when dealing with this type of data.

3. Conclusions

New technologies are great in providing floods of data full of information and potential knowledge, but the extraction of such information can often prove very difficult. High-dimension low-sample size data spaces are typical in revealing such difficulties that mathematicians label "curse of dimensionality". In the present work we focus on spaces where the number of variables (p) is at least the number of observations (n), a case that occurs, in statistical practice, mostly within the high-dimension low-sample size data framework. We look at the behaviour of the Mahalanobis distance and the idea of projection, so essentials in the analysis of multivariate data.

When $p \geq n-1$ we prove that, the Mahalanobis distance becomes degenerated and its known distance properties are lost. Under the same restriction it is confirmed that the idea of projection loses interest and becomes mostly useless. These conclusions imply that many procedures for analyzing data will not work in these spaces.

Although we have seen users insisting in using artifacts to analyze this kind of data, it seems that they are not aware of the geometric complexities of spaces under these conditions. Disclosing what happens to the Mahalanobis distance and to the usual system of projections, we gain insight into the geometric properties of those spaces and hopefully contribute to: (i) prevent unconscious applications of inadequate methods and (ii) help to devise new methods where those geometric properties have to be taken into account.

Having said this, it should be understood that we are not suggesting that traditional statistical methods are not worthwhile and their use should be abandoned when analyzing high-dimension low-sample size data. On the contrary, we consider that this formidable source of statistical knowledge should not be forgotten but used properly. To allow traditional methods to operate, a first step towards reducing the dimensionality of the data should be given. For dimensionality reduction to be achieved one can

think of a number of techniques including a panoply of variable selection and regularization procedures.

As shown in the example regarding the outlier detection methods, a strategy that uses a reduced number of principal components (empirical suggestion: number of components $\leq n/10$) may prevent the most dramatic consequences.

A common characteristic of the aforementioned suggestions is the lack of affine in/equivariance (as mentioned after Theorem 3). We can also consider other procedures that are not affine in/equivariant, like, for instance, the L_1 distances or privileged directions, as done in Hennig (2020).

A piece of mathematical statistics work that comes in support of our own findings is in Bickel *et al.* (2018). This relevant research follows an approach different from our own and considers the asymptotic properties of projection-pursuit methods in an high-dimensional environment ($p \to \infty$). The conclusions include a result similar to our result on projections: *"...given enough (high-dimensional) data one may find in it whatever structure one wants to look for."*. More specifically, it is not difficult to show that Corollary 1 of Theorem 2, on page 9153 of Bickel *et al.* (2018), is an asymptotic version of our Theorem 2 (as $p \to \infty$ and $n \to \infty$, with $p/n \to \gamma > 1$).

During our research work we were somehow surprised that we were unable to find any of the main results in the literature. That has changed. Currently, as of the date of submission, it is possible to find references using Theorem 1 (e.g., Hennig, 2020; Provost *et al.*, 2023) and Theorem 2 (e.g., Radojicic *et al.*, 2021; Loperfido, 2023), based on an earlier version of this work, meanwhile published as research report (Pires and Branco, 2019). This is another indication that the material here discussed is both original and relevant.

Acknowledgements

This work received financial support from Portuguese National Funds through Fundação para a Ciência e a Tecnologia.

Appendix

A. Definitions and notation

- Data matrix $(n \times p)$: $X = \begin{pmatrix} x_1^T \\ \cdots \\ x_n^T \end{pmatrix} = (x_1 \cdots x_n)^T$.

- Points in general position: the p-dimensional observations in the $(n \times p)$ data matrix X are said to be in general position if X has maximal rank, that is, if $\text{rank}(X) = \min(p, n)$. This property holds with probability one for continuous variables.

- Mean of X $(n \times 1)$: $\overline{x} = \sum_{i=1}^{n} x_i/n$.

- Centered data matrix ($n \times p$): $X_c = (x_1 - \overline{x} \cdots x_n - \overline{x})^T = \left(I_n - v_1 v_1^T/n\right) X$ (v_x denotes a vector with all its elements equal to x).

- Covariance matrix of X ($p \times p$):

$$S = \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^T/(n-1) = X_c^T X_c/(n-1).$$

  If X is in general position then $\text{rank}(S) = \text{rank}(X_c) = \min(p, n-1)$.

- Standardized data matrix ($n \times p$): $Z = X_c S^{-1/2}$, where $S^{-1/2}$ is any square root of S^{-1} (which means that $S^{-1/2} S^{-1/2} = S^{-1}$).

- Mahalanobis distance between an observation, x_i, and $\overline{x}$:

$$d_{x_i,\overline{x}} = \left\{(x_i - \overline{x})^T S^{-1} (x_i - \overline{x})\right\}^{1/2}.$$

- Mahalanobis distance between two observations, x_i and x_j:

$$d_{x_i,x_j} = \left\{(x_i - x_j)^T S^{-1} (x_i - x_j)\right\}^{1/2}.$$

- Matrix D ($n \times n$):

$$D = X_c S^{-1} X_c^T = (n-1) X_c (X_c^T X_c)^{-1} X_c^T = Z Z^T, \tag{2}$$

  where $d_{ij} = (x_i - \overline{x})^T S^{-1} (x_j - \overline{x})$. Note that $d^2_{x_i,\overline{x}} = d_{ii}$ and $d^2_{x_i,x_j} = d_{ii} + d_{jj} - 2 d_{ij}$.

- Augmented data matrix ($p \times (n+1)$): $X_d = \left[\ v_1 \mid X_c\ \right]$, which verifies

$$X_d (X_d^T X_d)^{-1} X_d^T = H = \frac{D}{n-1} + \frac{v_1 v_1^T}{n}, \tag{3}$$

  where H is the well known "hat matrix" (Hoaglin and Welsch, 1978).

B. Proofs and details

Proof (of Theorem 1). For $p > n-1$, S is singular and S^{-1}, which does not exist, is replaced by the Moore-Penrose pseudoinverse (S^-). This is equivalent to reducing the number of variables to $n-1$ (*e.g.*, the standardized non-trivial principal components, which means discarding only the non-informative components, *i.e.*, those with zero variance/eigenvalue). Therefore, we only have to prove the $p = n-1$ case.

In that case, X_c has dimension $n \times (n-1)$ and rank $n-1$, whereas X_d has dimension $n \times n$ and rank n, so it can be inverted, resulting in

$$H = X_d(X_d^T X_d)^{-1} X_d^T = X_d X_d^{-1}(X_d^T)^{-1} X_d^T = I = \frac{D}{n-1} + \frac{v_1 v_1^T}{n},$$

which is equivalent to

$$D = (n-1)\left(I - \frac{v_1 v_1^T}{n}\right). \tag{4}$$

That is, $d_{ii} = (n-1)(1-1/n) = (n-1)^2/n$, for all i, which proves (i). On the other hand, from Eq. (4), for any $i \neq j$, $d_{ij} = -(n-1)/n$, therefore,

$$d^2_{x_i,x_j} = d_{ii} + d_{jj} - 2d_{ij} = 2\frac{(n-1)^2}{n} + 2\frac{n-1}{n} = 2(n-1),$$

which proves (ii).

Equation (3) writes $h_{ij} = d_{ij}/(n-1) + 1/n$, and applying the well known result (Hoaglin and Welsch, 1978), $0 \leq h_{ii} \leq 1$, the conclusion that $d_{ii} \leq (n-1)^2/n$ follows immediately, thus proving (iii). This is a short alternative proof of Theorem 2.1 in Gath and Hayes (2006).

To prove (iv), let $T = I - H$ and check that T is symmetric and idempotent. Consider the ith and jth row (or column) vectors of T, t_i and t_j, with $i \neq j$. Then $t_i^T t_j = -h_{ij}$ and $t_i^T t_i = 1 - h_{ii}$ (we need the assumption that $h_{ii} < 1$, for all i, which is not restrictive, because the cases where $h_{ii} = 1$ are considered in (ii)). Let β be the angle between t_i and t_j, then

$$\cos\beta = \frac{t_i^T t_j}{\left(t_i^T t_i\right)^{1/2}\left(t_j^T t_j\right)^{1/2}} = -\frac{h_{ij}}{(1-h_{ii})^{1/2}\left(1-h_{jj}\right)^{1/2}},$$

implying that $-(1-h_{ii})^{1/2}\left(1-h_{jj}\right)^{1/2} \leq h_{ij} \leq (1-h_{ii})^{1/2}\left(1-h_{jj}\right)^{1/2}$. Considering this result and the inequality $\left\{(1-h_{ii})^{1/2} - \left(1-h_{jj}\right)^{1/2}\right\}^2 \geq 0$, it follows that

$$0 \leq 1 - h_{ii} + 1 - h_{jj} - 2(1-h_{ii})^{1/2}\left(1-h_{jj}\right)^{1/2} \leq 1 - h_{ii} + 1 - h_{jj} + 2h_{ij},$$

which is equivalent to $h_{ii} + h_{jj} - 2h_{ij} \leq 2$. Finally, simple manipulations of the definitions yield

$$d^2_{x_i,x_j} = d_{ii} + d_{jj} - 2d_{ij} = (n-1)(h_{ii} + h_{jj} - 2h_{ij}) \leq 2(n-1).$$

□

PROOF (OF THEOREM 2). We need to prove that, given X and Y, there is an orthogonal transformation, characterized by the $p \times 2$ matrix Q, whose columns are orthonormal vectors, such that $XQ = Y^*$, where Y^* is an affine transformation of Y (that is, $Y^* = YA + v_1 b^T$, for some non-singular 2×2 matrix A and some two dimensional vector b).

Without loss of generality we assume that Y $(n \times 2)$ is such that $\overline{y} = v_0$ and $S_Y = Y^T Y/(n-1) = I_2$. As in the proof of Theorem 1 we need to consider only the case $p = n-1$, for which the covariance matrix of X, S, is invertible. Let $Z = X_c S^{-1/2}$, be the standardized data matrix, and consider the $(n-1) \times 2$ matrix U defined by $U = Z^T Y/(n-1)$. We show next that (i) the columns of U are orthonormal vectors, and (ii) $Y = ZU$.

(i) $U^T U = Y^T Z Z^T Y/(n-1)^2 = Y^T D Y/(n-1)^2 = Y^T \left(I_n - v_1 v_1^T/n\right) Y/(n-1) = Y_c^T Y_c/(n-1) = I_2$, which follows from Equation (4) and by the assumption that Y is standardized.

(ii) $ZU = ZZ^T Y/(n-1) = DY/(n-1) = \left(I_n - v_1 v_1^T/n\right) Y = Y_c = Y$, for the same reasons.

The next step is to rewrite (ii) in terms of X or, equivalently, in terms of X_c, $Y = ZU = X_c S^{-1/2} U$. Replacing, in the last equality, the matrix $S^{-1/2}U$ by its singular value decomposition, $S^{-1/2}U = V_1 L V_2$, where the columns of V_1 $(p \times 2)$ and V_2 (2×2) are orthonormal and L is a 2×2 diagonal matrix, leads to $Y = X_c V_1 L V_2$, which is equivalent to $Y V_2^T L^{-1} = X_c V_1$. Therefore, and concluding the proof, $Q = V_1$, $A = V_2^T L^{-1}$ and $b = V_1^T \overline{x}$. □

The two-dimensional projections are usually the most interesting to consider, but Theorem 2 can be easily generalized to projections on a subspace of arbitrary dimension, $k = 1, \ldots, n-2$. The case $k = 1$ includes the "piling effect" as a special case, and also the curious case of projections where all but one point, which can be any of the n points, coincide (see Theorem 3). Putting together all the steps in the proof of Theorem 2, it is possible to conclude that the solution for $k = 1$ is given by $Q = S^{-1} X_c^T Y / \| S^{-1} X_c^T Y \|$, where Y does not have to be standardized.

Note that the equality $H = I_n$ for $p \geq n-1$ is the is the key to prove both Theorem 1 and Theorem 2.

The plots in Figure 2, produced as described in the proof of Theorem 2 above, illustrate the kind of "similarity" between Y and Y^* that can be obtained when $p \geq n-1$, no matter the data, and that can be described as perfect "similarity". On the contrary, in the cases considered in Figure 3 where $p < n-1$, perfect "similarity" is no longer possible. To produce plots as "similar" as possible to Y we use a least squares criterion and look for the transformation, $XQ = \hat{Y}^*$, minimizing $\| Y^* - \hat{Y}^* \|$, concluding that U must be replaced by $U = \left(Z^T Z\right)^{-1} Z^T Y/(n-1)$ (all the other formulas are unchanged).

PROOF (OF THEOREM 3). Assuming a fixed, given, i. For $(m, s) = (\text{mean}, \text{st.dev.})$ and $p \leq n-1$ trivial properties of the mean and standard deviation allow us to rewrite

the expression of the function to be maximized in Equation (1), as

$$\frac{|\alpha^T x_i - \text{mean}(\alpha^T x_1, \ldots, \alpha^T x_n)|}{\text{st.dev.}(\alpha^T x_1, \ldots, \alpha^T x_n)} = \frac{\left|\alpha^T (x_i - \overline{x})\right|}{\sqrt{\alpha^T S \alpha}}.$$

A direct application of the Maximization Lemma in page 80 of Johnson and Wichern (2007) to the square of the second member, leads to the conclusion that the maximum is $\sqrt{(x_i - \overline{x})^T S^{-1} (x_i - \overline{x})} = d_{x_i, \overline{x}}$ and that it is attained at $\alpha = c\, S^{-1}(x_i - \overline{x})$ for any constant $c \neq 0$. Choosing $c = 1/||S^{-1}(x_i - \overline{x})||$, makes $||\alpha|| = 1$ and gives the desired result.

As argued before, in the proofs of Theorems 1 and 2, the results for $p > n-1$ are obtained from the results for $p = n-1$, simply by replacing S^{-1}, which does not exist, with S^{-} (the Moore-Penrose pseudoinverse). This concludes the proof for the (mean, st.dev.) pair.

Note that, for the projected data we have, for $j = 1, \ldots, n$,

$$\alpha^T x_j = \alpha^T \left(x_j - \overline{x}\right) + \alpha^T \overline{x} = \frac{(x_i - \overline{x})^T S^{-1} (x_j - \overline{x})}{||S^{-1}(x_i - \overline{x})||} + \alpha^T \overline{x} = \frac{d_{ij}}{||S^{-1}(x_i - \overline{x})||} + \alpha^T \overline{x},$$

with d_{ij} defined in Equation (2). Apart from d_{ij}, the other quantities involved, $\alpha^T \overline{x}$ and $||S^{-1}(x_i - \overline{x})||$, do not depend on j. For $p < n-1$, d_{ij} depends on X and will in general take n distinct values (one for each j). For $p = n-1$, however, d_{ij} takes only two distinct values: $d_{ij} = (n-1)^2/n$, if $j = i$, and $d_{ij} = -(n-1)/n$, if $j \neq i$, as shown in the proof of Theorem 1, a result that is important for the rest of the proof.

For (m, s) =(median, mad) the function to be maximized in Equation (1) is

$$\frac{|\alpha^T x_i - \text{median}(\alpha^T x_1, \ldots, \alpha^T x_n)|}{\text{mad}(\alpha^T x_1, \ldots, \alpha^T x_n)}.$$

It is not possible to simplify this expression because the median of a linear combination of vectors is not guaranteed to be the linear combination of the medians of the vectors. For $p < n-1$ it is possible to prove that the expression is a continuous but non-differentiable function of α. A numerical algorithm can be used to obtain approximate solutions, that will, in general, depend on X and lead to a projection with n distinct values.

For $p = n-1$, we only need to consider the orthogonal projection on the direction obtained for the (mean, st.dev.) pair, $\alpha_0 = S^{-1}(x_i - \overline{x})/||S^{-1}(x_i - \overline{x})||$. As shown above, $\alpha_0^T x_i = a$, and $\alpha_0^T x_j = b$, for $j \neq i$, where a and b are real numbers with $a \neq b$. We thus have

$$\text{median}(\alpha_0^T x_1, \ldots, \alpha_0^T x_n) = b \quad \text{and} \quad \text{mad}(\alpha_0^T x_1, \ldots, \alpha_0^T x_n) = 0,$$

as well as

$$\lim_{\alpha\to\alpha_0,\,\alpha\in\mathbb{R}^p,\,\|\alpha\|=1} |\alpha^T x_i - \text{median}(\alpha^T x_1,\ldots,\alpha^T x_n)| = |a-b| > 0 \tag{5}$$

and

$$\lim_{\alpha\to\alpha_0,\,\alpha\in\mathbb{R}^p,\,\|\alpha\|=1} \text{mad}(\alpha^T x_1,\ldots,\alpha^T x_n) = 0, \tag{6}$$

which implies

$$\lim_{\alpha\to\alpha_0,\,\alpha\in\mathbb{R}^p,\,\|\alpha\|=1} \frac{|\alpha^T x_i - \text{median}(\alpha^T x_1,\ldots,\alpha^T x_n)|}{\text{mad}(\alpha^T x_1,\ldots,\alpha^T x_n)} = +\infty. \tag{7}$$

□

It can be shown that, $\alpha_0 = S^{-1}(x_i - \overline{x})/\|S^{-1}(x_i - \overline{x})\|$, the projection direction in the proof above, coincides with the solution given in the first paragraph after the proof of Theorem 2, $Q = S^{-1}X_c^T y/\|S^{-1}X_c^T y\|$.

Figure 5 illustrates Theorem 3 and its proof. We use two artificial datasets with $p = 2$, one with $n = 3$ ($p = n-1$) and another one with $n = 30$ ($p < n-1$). At the top we show the scatter plots of the datasets. The remaining 4 plots show the results of the numerical computation, by grid search, of the outlyingnesses, for (m, s) =(mean, std.dev.) and (m, s) =(median, mad), at the points highlighted in the scatter plots with a solid circle. The grid uses 1000 points of the polar angle, θ, defining the bidimensional directions as $\alpha^T = (\cos\theta, \sin\theta)$, $\theta \in [0, \pi)$. Table 4 gives the theoretical results, if known, as well as the numerical results.

TABLE 4
Theoretical and numerical results.

		$n = 3$		$n = 30$	
(m, s)	result	theoretical	numerical	theoretical	numerical
mean	out_i	$2/\sqrt{3}$	1.1547	4.2796	4.2796
st.dev.	$\hat{\theta}_i$	$\pi/4$	0.7862	2.2646	2.2642
median	out_i	$+\infty$	428.6	—	12.195
mad	$\hat{\theta}_i$	$\pi/4$	0.7862	—	2.1636

Notation: $\text{out}_i = \max_\theta g(i,\theta)$ denotes the outlyingness and $\hat{\theta}_i = \text{argmax}_\theta g(i,\theta)$ identifies the direction at which the maximum is attained.

Note that for the case (m, s) =(median, mad) there may be other directions α_0 verifying Equations (5), (6) and (7). Indeed, for $n \geq 5$ there are infinitely many such directions. This is because, in order to verify the Equations we only need that the projection of the ith point is at a, and that the projections of at least $\lfloor n/2 \rfloor + 1$ any other points are at b. The projections of the remaining points can be anywhere.

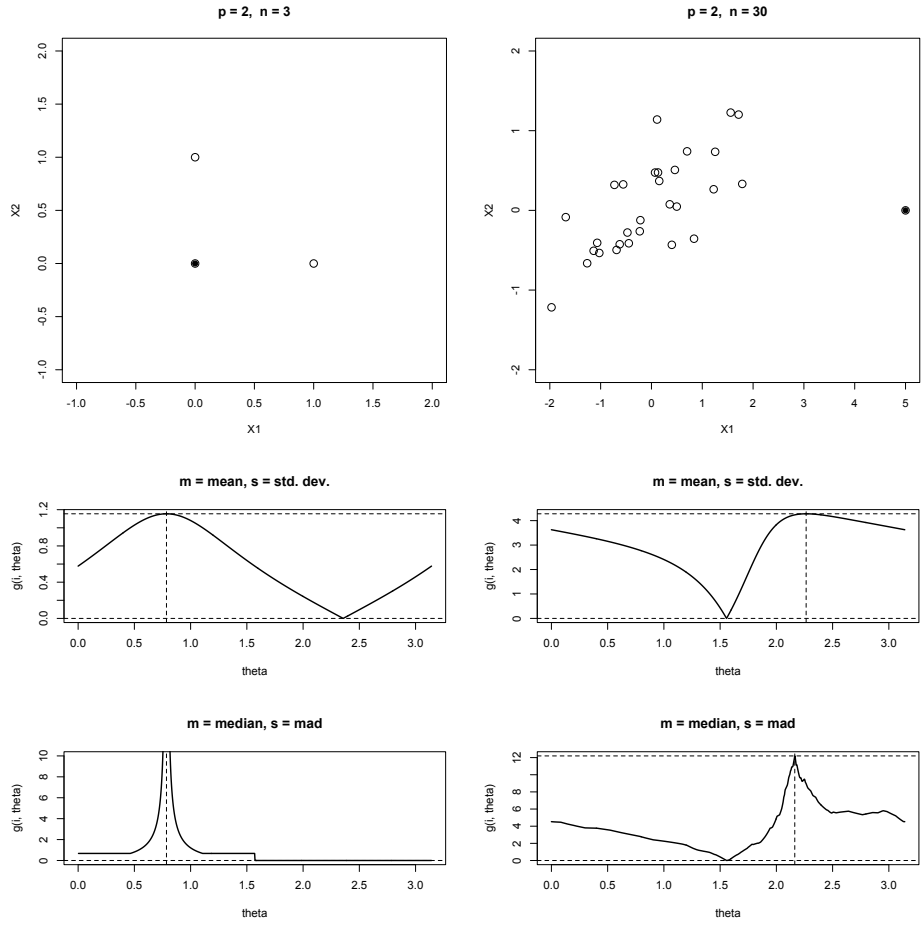

Figure 5 – Two artificial data sets in dimension two and the plots of the functions (denoted $g(i,\theta)$) that have to be maximized in order to compute the outlyingnesses, for $(m,s) = (\text{mean}, \text{std.dev.})$ and $(m,s) = (\text{median}, \text{mad})$, at the points highlighted in the scatter plots with a solid circle. The horizontal axes represent the polar angle, θ, defining the bidimensional directions as $\alpha^T = (\cos\theta, \sin\theta)$, $\theta \in [0, \pi)$.

C. Data and R code for the example in Section 2.3

The data used in this example, described in Alon *et al.* (1999), were downloaded from the Princeton University Gene Expression Project database.

The data were then imported into R (R Core Team, 2024) where all the computations were performed. The only preprocessing applied was a $\log_2$ transform of the full data matrix (following Shieh and Hung, 2009).

The data matrix has 62 rows (samples or cases) and 2000 columns (genes). The 62 cases are divided into two groups, the normal tissue samples (22, labeled N...) and the tumor samples (40, labeled T...). The purpose of the example is to assess the performance of some outlier detecting methods advertised as appropriate for high-dimension low-sample size datasets. The methods considered are identified by the abbreviated names (CLA, ROBPCA, PCOUT, SIGN1, SIGN2, SH, PPGRID, PPPROJ, see the main text for references), and are all available in the R package `rrcovHD` (Todorov, 2024).

Each method was applied to each group separately, and the number of outliers detected by each method is shown in Table 1. To complement that information we provide here the R code used and the full results obtained, including the labels of the outlying observations:

```
######################################## reading and preprocessing the data
colon_gene_expr = read.table("colon.cancer.txt", header=F)
#each line represents a gene or variable and each column represents a tissue or observation
colon_gene_expr = as.matrix(colon_gene_expr)
colon_gene_expr = t(log(colon_gene_expr, 2))

colon_tissue_id = c(-1,1,-2,2,-3,3,-4,4,-5,5,-6,6,-7,7,-8,8,-9,9,-10,10,-11,11,-12,12,-13,-14,-15,-16,-17,-18,-19,-20,-21,
-22,-23,-24,-25,-26,27,-27,-28,28,29,-29,-30,-31,-32,32,-33,33,34,-34,-35,35,36,-36,-37,-38,-39,39,-40,40)
# the numbers correspond to patients, a positive sign to a normal tissue, and a negative sign to a tumor tissue.
colon_grouping = colon_tissue_id * 0
colon_grouping[colon_tissue_id < 0] = 2
colon_grouping[colon_tissue_id > 0] = 1
######################################## packages
library(rrcovHD)
######################################## function
find_hd_outliers = function(X, k, crit.pca.distances, qcrit, explvar, method){
   n = nrow(X)
   if (method == "CLA"){
      result = (1:n)[!PcaClassic(X, k=k, crit.pca.distances=crit.pca.distances)@flag]
   }
   if (method == "ROBPCA"){
      result = (1:n)[!PcaHubert(X, k=k, crit.pca.distances=crit.pca.distances)@flag]
   }
   if (method == "PCOUT"){
      result = (1:n)[!OutlierPCOut(X, explvar=explvar)@flag]
   }
   if (method == "SIGN1"){
      result = (1:n)[!OutlierSign1(X, qcrit=qcrit)@flag]
   }
   if (method == "SIGN2"){
      result = (1:n)[!OutlierSign2(X, explvar=explvar, qcrit=qcrit)@flag]
   }
   if (method == "SH"){
      result = (1:n)[!OutlierPCDist(X, explvar=explvar)@flag]
   }
   if (method == "PPGRID"){
      result = (1:n)[!PcaGrid(X, k=k, crit.pca.distances=crit.pca.distances)@flag]
   }
   if (method == "PPPROJ"){
      result = (1:n)[!PcaProj(X, k=k, crit.pca.distances=crit.pca.distances)@flag]
   }
   result
}
######################################## get results for the original colon dataset
glabel = c("N", "T")
list_methods = c("CLA", "ROBPCA", "PCOUT", "SIGN1", "SIGN2", "SH", "PPGRID", "PPPROJ")

for (im in list_methods){
```

```
    cat("\nNumber of outliers by method ", im, ":  ", sep="")
    res_id = c()
    for (ig in 1:2){
       result = find_hd_outliers(colon_gene_expr[colon_grouping == ig, ], k=0, crit.pca.distances=0.975, qcrit=0.95, explvar=0.8, method=im)
       aux_id = colon_tissue_id[colon_grouping == ig][result] * (-1) ** (ig + 1)
       for (ij in aux_id){
           res_id = c(res_id, paste(glabel[ig], ij, sep=''))
       }
       cat(length(result), " (", glabel[ig], ") ", sep="")
    }
    cat("\nTissue ids: ", res_id, "\n")
}

Number of outliers by method CLA:  2 (N) 3 (T)
Tissue ids:  N33 N34 T5 T37 T39

Number of outliers by method ROBPCA:  5 (N) 10 (T)
Tissue ids:  N8 N9 N12 N34 N36 T2 T4 T5 T6 T9 T12 T19 T25 T34 T37

Number of outliers by method PCOUT:  4 (N) 6 (T)
Tissue ids:  N3 N8 N9 N12 T5 T10 T30 T33 T36 T37

Number of outliers by method SIGN1:  8 (N) 21 (T)
Tissue ids:  N2 N3 N8 N9 N10 N12 N29 N34 T2 T5 T6 T9 T10 T12 T17 T18 T19 T20 T21 T25 T26 T28 T29 T30 T31 T32 T34 T37 T38

Number of outliers by method SIGN2:  4 (N) 8 (T)
Tissue ids:  N8 N9 N12 N34 T2 T5 T6 T9 T12 T29 T32 T37

Number of outliers by method SH:  3 (N) 6 (T)
Tissue ids:  N8 N12 N34 T2 T5 T9 T12 T25 T37

Number of outliers by method PPGRID:  2 (N) 3 (T)
Tissue ids:  N5 N32 T8 T12 T37

Number of outliers by method PPPROJ:  3 (N) 7 (T)
Tissue ids:  N5 N12 N32 T2 T5 T6 T9 T12 T30 T37

############################ Software versions:
> packageVersion('rrcovHD')
[1] '0.3.1'
> packageVersion('rrcov')
[1] '1.7.6'
> packageVersion('MASS')
[1] '7.3.60.2'
> getRversion()
[1] '4.4.0'
```

The results of the Monte Carlo simulation described in the main text are presented below, together with the R script used to obtain them.

```
##########################################
### This script simulates the detection of outliers
### in multivariate normal data (no outliers)
### It requires libraries rrcovHD and MASS

library(MASS)

simulMdetection = function(Xdata, grouping, B, list_methods, k, crit.pca.distances, qcrit, explvar, seed){
    ### group principal components and means of the original dataset
    ### (to be replicated in the simulated data)
    pca.cd1 = prcomp(Xdata[grouping == 1, ])
    pca.cd2 = prcomp(Xdata[grouping == 2, ])
    mean.1 = colMeans(Xdata[grouping == 1, ])
    mean.2 = colMeans(Xdata[grouping == 2, ])
    n1 = sum(grouping == 1)
    n2 = sum(grouping == 2)
    p1 = sum(pca.cd1$sdev > 1e-12)
    p2 = sum(pca.cd2$sdev > 1e-12)
    number_out = array(0, c(B, 2, length(list_methods)))
    for (i in 1:B){
        auxX1 = mvrnorm(n1, mu=rep(0, p1), Sigma=diag(p1), empirical=T) %*% diag(pca.cd1$sdev[1:p1]) %*% t(pca.cd1$rot[, 1:p1])
        auxX2 = mvrnorm(n2, mu=rep(0, p2), Sigma=diag(p2), empirical=T) %*% diag(pca.cd2$sdev[1:p2]) %*% t(pca.cd2$rot[, 1:p2])
        auxX1 = t(t(auxX1) + mean.1)
        auxX2 = t(t(auxX2) + mean.2)
        for (im in 1:length(list_methods)){
            result1 = find_hd_outliers(auxX1, k=k, crit.pca.distances=crit.pca.distances, qcrit=qcrit, explvar=explvar, method=list_methods[im])
            result2 = find_hd_outliers(auxX2, k=k, crit.pca.distances=crit.pca.distances, qcrit=qcrit, explvar=explvar, method=list_methods[im])
            number_out[i, 1, im] = length(result1)
            number_out[i, 2, im] = length(result2)
            }
        print(i)
    }
```

```
for (im in 1:length(list_methods)){
        cat("\n", list_methods[im], ":", sep="")
        cat("\nn = ", n1, ":", sep="")
        print(table(number_out[, 1, im]))
        cat("\nn = ", n2, ":", sep="")
        print(table(number_out[, 2, im]))
     }
     for (im in 1:length(list_methods)){
         cat("\n", list_methods[im], ":", sep="")
         cat("\nn = ", n1, ": ", round(mean(number_out[, 1, im]) / n1, 4), sep="")
         cat("\nn = ", n2, ": ", round(mean(number_out[, 2, im]) / n2, 4), sep="")
     }
     number_out
}

### results for data in 2000 dimensions
res2000 = simulMdetection(colon_gene_expr, colon_grouping, B=500, list_methods=list_methods, k=0,
                          crit.pca.distances=0.975, qcrit=0.95, explvar=0.8, seed=200)

CLA:
n = 22:
  0   1   2
280 204  16
n = 40:
  0   1   2   3   4
104 232 127  32   5

ROBPCA:
n = 22:
  1   2   3   4   5   6   7   8
  1   1   7  50 130 221  82   8
n = 40:
 4  5  6  7  8  9 10 11 12 13 14 15 16
 1  1  2 14 22 43 80 83 90 91 60 12  1

PCOUT:
n = 22:
  0   1   2   3   4   5   6   7   8   9  10
 15  31  63 108 108  96  53  20   1   4   1
n = 40:
 0  1  2  3  4  5  6  7  8  9 10 11 12 13
 3 10 23 50 57 91 75 65 64 35 15  8  3  1

SIGN1:
n = 22:
 1  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
 1  4 10 38 60 84 89 87 55 40 17  9  1  3  1  1
n = 40:
12 13 14 15 16 17 18 19 20 21 22 23 24 25
 1  9 12 29 67 80 84 83 65 29 19 13  7  2

SIGN2:
n = 22:
  0   1   2   3   4   5   6   7
 52 107 153  93  58  26   8   3
n = 40:
  0   1   2   3   4   5   6   7   8   9
 32  80 110  92  83  43  32  18   9   1

SH:
n = 22:
  0   1   2   3   4   5   6   7
 51 111 141 108  49  33   6   1
n = 40:
  0   1   2   3   4   5   6
 56 120 156 105  40  19   4

PPGRID:
n = 22:
  0   1   2   3   4   5   6   7   8   9
157 145  94  48  31  17   5   1   1   1
n = 40:
  0   1   2   3   4   5   6   7   8  10  12  14
115 105  95  68  56  31  10   9   6   3   1   1

PPPROJ:
n = 22:
  0   1   2   3   4   5   6   7   8
 85 141 100  79  48  23   9  13   2
n = 40:
  0   1   2   3   4   5   6   7   8   9  10
 70 116 106  79  50  43  21   9   2   2   2

CLA:    n = 22: 0.0215  n = 40: 0.0301  ROBPCA:  n = 22: 0.258   n = 40: 0.2847
PCOUT:  n = 22: 0.1732  n = 40: 0.1454  SIGN1:   n = 22: 0.4171  n = 40: 0.4548
SIGN2:  n = 22: 0.1021  n = 40: 0.0761  SH:      n = 22: 0.1019  n = 40: 0.0513
```

```
PPGRID: n = 22: 0.0673  n = 40: 0.0552  PPPROJ:  n = 22: 0.0951  n = 40: 0.0612

### results for data in 5 dimensions (selected randomly)

set.seed(100)
xvar = sample(2000, 5)
print(xvar)
[1] 1738  503 1382 1648  985

res5 = simulMdetection(colon_gene_expr[, xvar], colon_grouping, B=500, list_methods=list_methods,
                       k=5, crit.pca.distances=0.95, qcrit=0.95, explvar=0.99999, seed=200)

CLA:
n = 22:
  0   1   2   3
344 134  21   1
n = 40:
  0   1   2   3   4   5
120 200 126  49   4   1

ROBPCA:
n = 22:
  0   1   2   3
 30  82 155 233
n = 40:
 0  1  2  3  4  5  6  7  8
19 64 91 97 89 76 42 19  3

PCOUT:
n = 22:
  0   1   2   3   4   5   6   7   8
 20  54  62 101 102  95  41  18   7
n = 40:
 0  1  2  3  4  5  6  7  8  9 10 11 12
 3 22 41 49 69 73 73 63 64 19 17  5  2

SIGN1:
n = 22:
  0   1   2   3   4   5   6   7   8   9
 68 112 111 105  57  28   9   7   1   2
n = 40:
 0  1  2  3  4  5  6  7  8  9
57 83 93 87 87 39 23 17  9  5

SIGN2:
n = 22:
  0   1   2   3   4   5   6   7
 49 106 125 103  69  26  16   6
n = 40:
 0  1  2  3  4  5  6  7  8 10
22 76 94 89 90 62 42 15  9  1

SH:
n = 22:
  0   1   2   3   4   5   6
 48 111 150 104  59  21   7
n = 40:
  0   1   2   3   4   5   6   8
 77 113 138  96  50  20   5   1

PPGRID:
n = 22:
  0   1   2   3   4   5   6
165 150 100  62  15   7   1
n = 40:
  0   1   2   3   4   5   6   8   9
115 138 123  71  29  14   7   2   1

PPPROJ:
n = 22:
  0   1   2   3   4   5   6   8
109 140 109  78  40  15   6   3
n = 40:
  0   1   2   3   4   5   6   7   8
 80 128 103  92  50  29   7   9   2


CLA:    n = 22: 0.0163  n = 40: 0.031   ROBPCA: n = 22: 0.0992  n = 40: 0.0841
PCOUT:  n = 22: 0.1629  n = 40: 0.1352  SIGN1:  n = 22: 0.1042  n = 40: 0.0723
SIGN2:  n = 22: 0.1099  n = 40: 0.082   SH:     n = 22: 0.1005  n = 40: 0.0507
PPGRID: n = 22: 0.0579  n = 40: 0.0425  PPPROJ: n = 22: 0.0806  n = 40: 0.0538



##########################################
```

```
### This script simulates the detection of outliers
### in multivariate normal data with 1 outlier
### It requires libraries rrcovHD and MASS

simulMdetection_out1 = function(Xdata, grouping, B, list_methods, k,
                                crit.pca.distances, qcrit, explvar, seed, dd){
    ### group principal components and means of the original dataset
    ### (to be replicated in the simulated data)
    pca.cd1 = prcomp(Xdata[grouping == 1, ])
    pca.cd2 = prcomp(Xdata[grouping == 2, ])
    mean.1 = colMeans(Xdata[grouping == 1, ])
    mean.2 = colMeans(Xdata[grouping == 2, ])
    n1 = sum(grouping == 1)
    n2 = sum(grouping == 2)
    p1 = sum(pca.cd1$sdev > 1e-12)
    p2 = sum(pca.cd2$sdev > 1e-12)
    number_out = array(0, c(B, 4, length(list_methods)))
    # out_correct = array(0, c(B, 2, length(list_methods)))
    for (i in 1:B){
        auxX1 = mvrnorm(n1, mu=rep(0, p1), Sigma=diag(p1), empirical=T) %*% diag(pca.cd1$sdev[1:p1]) %*% t(pca.cd1$rot[, 1:p1])
        auxX2 = mvrnorm(n2, mu=rep(0, p2), Sigma=diag(p2), empirical=T) %*% diag(pca.cd2$sdev[1:p2]) %*% t(pca.cd2$rot[, 1:p2])
        auxX1 = t(t(auxX1) + mean.1)
        auxX2 = t(t(auxX2) + mean.2)
        auxX1[1, ] = auxX1[1, ] + dd
        auxX2[1, ] = auxX2[1, ] + dd
        for (im in 1:length(list_methods)){
            result1 = find_hd_outliers(auxX1, k=k, crit.pca.distances=crit.pca.distances, qcrit=qcrit,
                                       explvar=explvar, method=list_methods[im])
            result2 = find_hd_outliers(auxX2, k=k, crit.pca.distances=crit.pca.distances, qcrit=qcrit,
                                       explvar=explvar, method=list_methods[im])
            number_out[i, 1, im] = length(result1)
            number_out[i, 2, im] = length(result2)
            if (length(result1) > 0){number_out[i, 3, im] = sum(result1 == 1)}
            if (length(result2) > 0){number_out[i, 4, im] = sum(result2 == 1)}
            }
        print(i)
    }
   for (im in 1:length(list_methods)){
        cat("\n", list_methods[im], ":", sep="")
        cat("\nn = ", n1, ":", sep="")
        print(table(number_out[, 1, im]))
        cat("\nn = ", n2, ":", sep="")
        print(table(number_out[, 2, im]))
     }
     for (im in 1:length(list_methods)){
         cat("\n", list_methods[im], ":", sep="")
         cat("\nn = ", n1, ": ", round(mean(number_out[, 1, im] - number_out[, 3, im]) / (n1 - 1), 4), sep="")
         cat("\nn = ", n2, ": ", round(mean(number_out[, 2, im] - number_out[, 4, im]) / (n2 - 1), 4), sep="")
         cat("\nn = ", n1, ": ", round(mean(number_out[, 3, im]), 4), sep="")
         cat("\nn = ", n2, ": ", round(mean(number_out[, 4, im]), 4), sep="")
     }
     number_out
}
res2000_out1 = simulMdetection_out1(colon_gene_expr, colon_grouping, B=500, list_methods=list_methods,
                                    k=0, crit.pca.distances=0.975, qcrit=0.95, explvar=0.8, seed=200, dd=3)

CLA:     n = 22: 0.0187  n = 40: 0.0298  n = 22: 0.726  n = 40: 0.728
ROBPCA:  n = 22: 0.1909  n = 40: 0.2624  n = 22: 0.992  n = 40: 0.988
PCOUT:   n = 22: 0.1405  n = 40: 0.1296  n = 22: 0.964  n = 40: 0.962
SIGN1:   n = 22: 0.3874  n = 40: 0.4495  n = 22: 1      n = 40: 0.992
SIGN2:   n = 22: 0.093   n = 40: 0.0711  n = 22: 0.976  n = 40: 0.6
SH:      n = 22: 0.081   n = 40: 0.0448  n = 22: 0.974  n = 40: 0.908
PPGRID:  n = 22: 0.0729  n = 40: 0.0615  n = 22: 0.812  n = 40: 0.576
PPPROJ:  n = 22: 0.0788  n = 40: 0.0597  n = 22: 0.924  n = 40: 0.846


res5_out1 = simulMdetection_out1(colon_gene_expr[, xvar], colon_grouping, B=500, list_methods=list_methods,
                                    k=5, crit.pca.distances=0.95, qcrit=0.95, explvar=0.99999 seed=200, dd=3)

CLA:     n = 22: 0.0106  n = 40: 0.0258  n = 22: 0.606  n = 40: 0.756
ROBPCA:  n = 22: 0.0713  n = 40: 0.0751  n = 22: 0.87   n = 40: 0.806
PCOUT:   n = 22: 0.1354  n = 40: 0.1312  n = 22: 0.858  n = 40: 0.852
SIGN1:   n = 22: 0.0855  n = 40: 0.0666  n = 22: 0.904  n = 40: 0.878
SIGN2:   n = 22: 0.0935  n = 40: 0.077   n = 22: 0.894  n = 40: 0.854
SH:      n = 22: 0.0756  n = 40: 0.0461  n = 22: 0.856  n = 40: 0.784
PPGRID:  n = 22: 0.0441  n = 40: 0.0393  n = 22: 0.868  n = 40: 0.828
PPPROJ:  n = 22: 0.0635  n = 40: 0.0518  n = 22: 0.842  n = 40: 0.812
```

References

J. Ahn, J. S. Marron (2010). *The maximal data piling direction for discrimination*. Biometrika, 97, no. 1, pp. 254–259.

J. Ahn, J. S. Marron, K. M. Muller, Y.-Y. Chi (2007). *The high-dimension, low-sample-size geometric representation holds under mild conditions*. Biometrika, 94, no. 3, pp. 760–766.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine (1999). *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proceedings of the National Academy of Sciences, 96, no. 12, pp. 6745–6750.

V. Barnett, T. Lewis (1994). *Outliers in Statistical Data. 3rd Edition*. John Wiley & Sons, Kluwer Academic Publishers, Boston/Dordrecht/London.

Y. M. Baryshnikov, R. A. Vitale (1994). *Regular simplices and Gaussian samples*. Discrete & Computational Geometry, 11, pp. 141–147.

R. E. Bellman (1957). *Dynamic Programming*. Princeton University Press, Princeton.

P. J. Bickel, G. Kur, B. Nadler (2018). *Projection pursuit in high dimensions*. Proceedings of the National Academy of Sciences, 115, no. 37, pp. 9151–9156.

T. T. Cai, T. Liang, H. H. Zhou (2015). *Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions*. Journal of Multivariate Analysis, 137, pp. 161–172.

D. Castro-Reigía, J. Ezenarro, M. Azkune, I. Ayesta, M. Ostra, J. M. Amigo, I. García, M. C. Ortiz (2024). *Yoghurt standardization using real-time NIR prediction of milk fat and protein content*. Journal of Food Composition and Analysis, 128, p. 106015.

L. P. Cavalheiro, S. Bernard, J. P. Barddal, L. Heutte (2024). *Random forest kernel for high-dimension low sample size classification*. Statistics and Computing, 34, no. 1, p. 9.

A. Chakrabarti, R. Sen (2019). *Some statistical problems with high dimensional financial data*. In F. Abergel, B. K. Chakrabarti, A. Chakraborti, N. Deo, K. Sharma (eds.), *New Perspectives and Challenges in Econophysics and Sociophysics*, Springer International Publishing, Cham, pp. 147–167.

R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, Y. Wang (2008). *The properties of high-dimensional data spaces: implications for exploring gene and protein expression data*. Nature Reviews Cancer, 8, no. 1, pp. 37––49.

C. CROUX, P. FILZMOSER, H. FRITZ (2013). *Robust sparse principal component analysis*. Technometrics, 55, no. 2, pp. 202–214.

C. CROUX, P. FILZMOSER, M. R. OLIVEIRA (2007). *Algorithms for projection–pursuit robust principal component analysis*. Chemometrics and Intelligent Laboratory Systems, 87, no. 2, pp. 218–225.

C. CROUX, A. RUIZ-GAZEN (2005). *High breakdown estimators for principal components: the projection-pursuit approach revisited*. Journal of Multivariate Analysis, 95, no. 1, pp. 206–226.

D. L. DONOHO (1982). *Breakdown properties of multivariate location estimators*. Ph.D. Qualifying Paper Harvard University.

D. L. DONOHO, J. TANNER (2005). *Neighborliness of randomly projected simplices in high dimensions*. Proceedings of the National Academy of Sciences, 102, no. 27, pp. 9452–9457.

M. EASTWOOD, R. PENROSE (2000). *Drawing with complex numbers*. The Mathematical Intelligencer, 22, pp. 8––13.

P. FILZMOSER, R. MARONNA, M. WERNER (2008). *Outlier identification in high dimensions*. Computational Statistics & Data Analysis, 52, no. 3, pp. 1694–1711.

P. FILZMOSER, S. SERNEELS, R. MARONNA, C. CROUX (2020). *Robust multivariate methods in chemometrics*. In S. BROWN, R. TAULER, B. WALCZAK (eds.), *Comprehensive Chemometrics (Second Edition)*, Elsevier, Oxford, pp. 393–430.

P. FILZMOSER, V. TODOROV (2013). *Robust tools for the imperfect world*. Information Sciences, 245, pp. 4–20.

I. E. FRANK, J. H. FRIEDMAN (1993). *A statistical view of some chemometrics regression tools*. Technometrics, 35, no. 2, pp. 109–135.

E. G. GATH, K. HAYES (2006). *Bounds for the largest mahalanobis distance*. Linear Algebra and its Applications, 419, no. 1, pp. 93–106.

R. GNANADESIKAN, J. R. KETTENRING (1972). *Robust estimates, residuals, and outlier detection with multiresponse data*. Biometrics, 28, no. 1, pp. 81–124.

P. HALL, J. S. MARRON, A. NEEMAN (2005). *Geometric representation of high dimension, low sample size data*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67, no. 3, pp. 427–444.

C. HENNIG (2020). *Minkowski distances and standardisation for clustering and classification on high-dimensional data*. In T. IMAIZUMI, A. NAKAYAMA, S. YOKOYAMA (eds.), *Advanced Studies in Behaviormetrics and Data Science: Essays in Honor of Akinori Okada*, Springer Singapore, Singapore, pp. 103–118.

A. HINNEBURG, C. C. AGGARWAL, D. A. KEIM (2000). *What is the nearest neighbor in high dimensional spaces?* In *VLDB'2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, Cairo, Egypt*, pp. 506—-515.

D. C. HOAGLIN, R. E. WELSCH (1978). *The hat matrix in regression and ANOVA*. The American Statistician, 32, no. 1, pp. 17–22.

H. HOTELLING (1933). *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24, no. 6, pp. 417—-441.

P. J. HUBER (1985). *Projection pursuit*. The Annals of Statistics, 13, no. 2, pp. 435–475.

M. HUBERT, P. J. ROUSSEEUW, K. VANDEN BRANDEN (2005). *ROBPCA: a new approach to robust principal component analysis*. Technometrics, 47, no. 1, pp. 64–79.

R. A. JOHNSON, D. W. WICHERN (2007). *Applied Multivariate Statistical Analysis. 6th edn.* Prentice Hall, New Jersey.

I. M. JOHNSTONE, D. M. TITTERINGTON (2009). *Statistical challenges of high-dimensional data*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367, no. 1906, pp. 4237–4253.

E. K. LEE, D. COOK (2010). *A projection pursuit index for large p small n data*. Statistics and Computing, 20, pp. 381—-392.

B. LIU, Y. WEI, Y. ZHANG, Q. YANG (2017). *Deep neural networks for high dimension, low sample size data*. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2287–2293.

N. LOCANTORE, J. S. MARRON, D. G. SIMPSON, N. TRIPOLI, J. T. ZHANG, K. L. COHEN (1999). *Robust principal component analysis for functional data*. Test, 8, pp. 1––73.

N. LOPERFIDO (2023). *Kurtosis removal for data pre-processing*. Advances in Data Analysis and Classification, 17, pp. 239––267.

P. C. MAHALANOBIS (1936). *On the generalised distance in statistics*. Proceedings of the National Institute of Sciences of India, 2, pp. 49–55.

K. V. MARDIA (1977). *Mahalanobis distances and angles*. In P. R. KRISHNAIAH (ed.), *Multivariate Analysis IV*, North-Holland, Amsterdam, pp. 495—-511.

A. M. PIRES, J. A. BRANCO (2010). *Projection-pursuit approach to robust linear discriminant analysis*. Journal of Multivariate Analysis, 101, no. 10, pp. 2464–2485.

A. M. PIRES, J. A. BRANCO (2019). *High dimensionality: The latest challenge to data analysis*. URL https://arxiv.org/abs/1902.04679.

M. L. PROVOST, R. BAPTISTA, J. D. ELDREDGE, Y. MARZOUK (2023). *An adaptive ensemble filter for heavy-tailed distributions: tuning-free inflation and localization*. URL https://arxiv.org/abs/2310.08741.

R CORE TEAM (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

U. RADOJICIC, K. NORDHAUSEN, J. VIRTA (2021). *Kurtosis-based projection pursuit for matrix-valued data*. URL https://arxiv.org/abs/2109.04167.

F. SAMARIA, A. HARTER (1994). *Parameterisation of a stochastic model for human face identification*. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pp. 138–142.

S. SARKAR, R. BISWAS, A. K. GHOSH (2020). *On some graph-based two-sample tests for high dimension, low sample size data*. Machine Learning, 109, pp. 279––306.

X. SHEN, C. WANG, X. ZHOU, W. ZHOU, D. HORNBURG, S. WU, M. P. SNYDER (2024). *Nonlinear dynamics of multi-omics profiles during human aging*. Nature Aging, 4, pp. 1619––1634.

A. D. SHIEH, Y. S. HUNG (2009). *Detecting outlier samples in microarray data*. Statistical Applications in Genetics and Molecular Biology, 8, no. 1.

M. SJÖSTRÖM, . S. WOLD, W. LINDBERG, J. A. PERSSON, H. MARTENS (1983). *A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables*. Analytica Chimica Acta, 150, pp. 61–70.

W. STAHEL (1981). *Breakdown of covariance estimators*. Research Report 31, Fachgrupp fur Statistik, E.T.H. Zurich.

V. TODOROV (2024). *R: rrcovHD: Robust Multivariate Methods for High Dimensional Data*. URL https://CRAN.R-project.org/package=rrcovHD. R package version 0.3-1.

G. TRENKLER, S. PUNTANEN (2005). *A multivariate version of samuelson's inequality*. Linear Algebra and its Applications, 410, pp. 143–149.

D. E. TYLER (2010). *A note on multivariate location and scatter statistics for sparse data sets*. Statistics & Probability Letters, 80, no. 17, pp. 1409–1413.

S. WOLD, M. SJÖSTRÖM, L. ERIKSSON (2001). *PLS-regression: a basic tool of chemometrics*. Chemometrics and Intelligent Laboratory Systems, 58, no. 2, pp. 109–130.

C. ZHANG, J. YE, X. WANG (2023). *A computational perspective on projection pursuit in high dimensions: Feasible or infeasible feature extraction*. International Statistical Review, 91, no. 1, pp. 140–161.