International Journal of
***O*perations *R*esearch**

International Journal of Operations Research Vol. 15, No. 3, 89-100 (2018)

ℓ_1-norm Based Major Component Detection and Analysis for Asymmetric Radial Data

Qi An, Shu-Cherng Fang, Shan Jiang*, John E. Lavery

Department of Industrial and Systems Engineering, North Carolina State University, USA

Received June 2018; Revised August 2018; Accepted August 2018

Abstract: ℓ_1-norm based major component detection and analysis (ℓ_1 MCDA) is a state-of-the-art tool based exclusively on ℓ_1-norm to identify the major components of a multivariate data set with irregularly positioned "spokes" and "clutters". In this paper, we develop an algorithmic framework of ℓ_1 MCDA for treating radial data clouds without the assumption of symmetry. This two-phase algorithm first locates the central point of the data by a pre-selection procedure to screen out candidate points with sufficient data points in the vicinity followed by solving an ℓ_1-norm discrete minimization problem. It then calculates the major directions and median radii in those directions via a two-level median fitting process. Extensive computational experiments have been conducted on n-dimensional data sets of various configurations randomly generated from light-tailed and heavy-tailed distributions with possibly artificial outliers to support the accuracy and robustness of the proposed method.
Keyword — Principal component analysis, ℓ_1-norm, multivariate statistics

1. INTRODUCTION

Radial data sets are multivariate data consisting of clusters of data points with different values but equal or similar ratios between variables (Massart et al., 2001). In the graphical representation of a radial data set, we often observe clusters of data points radiating from a center that are positioned with many irregular "spokes". A "spoke" usually refers to a group of data points that show a linear structure and extend much further out in a certain direction from a particular point. The simplest radial data set appears as a "V-shape" where a "V-symmetry" is visually perceivable. The spokes may come from a heavy-tailed or contaminated distributions, resulting in a radial data set with high levels of noise or "clutters" (patterned outliers). It is important to note that not always a symmetry can be observed when recognizing a radial structure in the data, *i.e.*, the spokes may not be symmetrically positioned.

Data of this kind can be originated from the chemical analysis of ceramic samples, ancient proteinaceous pictorial ligands from mural paintings on cultural heritages, or the dilution of pollution from the emission point along a river course (Aruga, 2003). Other rather common sources that give data of radial nature include biomedical research, web search, geospatial activities, and airborne optical imaging (Luo et al., 2013).

It is meaningful to identify the spokes and clutters as part of knowledge mining in a multidimensional data set. In the context of urban terrain data, the radiating spokes correspond to the roads or fences with an intersection, while the clutters represent obstructions between the sensing mechanisms (Luo et al., 2013). Detection of roofs, walls, grounds is required for the full 3D building reconstruction. A data analysis method is in need to extract accurate spokes out of a data cloud with high noise and many outliers so some budget friendly techniques, like airborne imaging technique in urban planning, can be applied. Another motivation for finding existing major components of a point cloud is from recognizing data patterns for further compressing, for example, in cancer research that analyzes a patient's molecular profile to diagnose diseases (Rodriguez et al., 1997).

The asymmetric nature may pose a significant challenge for detecting the spokes and clutters in radial data, due to the major difficulty in identifying the center of the radial structure. Other problems also occur when multiple components are present in the radial data. In many cases, major components needs to be recovered without knowing the exact number of spokes. But if the spokes are very close to each other, they tend to interfere with the recovering of one another. Furthermore, high dimensionality of data can cause a great deal of computational burden for numerical implementations. Our proposed data analysis method intends to get around these difficulties.

Principal component analysis (PCA) is the most widely used statistical tool for identifying the major direction and spreads of a data set, particularly for statistically distributed data in mutually orthogonal directions (Jolliffe, 2002).

*Corresponding author's e-mail: sjiang8@ncsu.edu

http://doi.org/10.6886/IJOR.201809_15(3).0001

Standard PCA is based on the ℓ_2-norm and second order statistics, assuming no or extremely few outliers for a light-tailed distribution. But it is the usual case that a data cloud may follow a heavy-tailed distribution, or contain a significant number of outliers if not obtained under benign laboratory conditions, which often puts the validity of standard PCA in jeopardy. Ideally, an effective multivariate PCA procedure is needed that deals with the outliers and heavy tails present in a data set. A number of approaches to robustifying PCA have been explored in the literature (Candès et al., 2011; Choulakian, 2006; Croux et al., 2013).

Some robust PCAs that partially involve the ℓ_1-norm have been proposed in a growing tendency towards exploiting ℓ_1-norm features, as in signal and image processing (Gribonval and Nielsen, 2006), compressive sensing (Chartrand, 2007), shape-preserving geometric modeling (Jin et al., 2010; Nie et al., 2017; Wang et al., 2014). Croux and Ruiz-Gazen (2005) derived a robust PCA method that adopts the influence function of the projection-pursuit based estimator for principal components. Brooks et al. (2013) proposed a PCA procedure based on an efficient calculation of the optimal solution of the ℓ_1-norm best-fit hyperplane problem. It was implemented on the dataset sampled from Laplace distribution, a heavy-tailed distribution. The use of ℓ_1-norm in PCA methods prevails because ℓ_1-norm is a robust measure that can prevent the PCAs from easily being disturbed by the outliers. But the ℓ_1-norm is partially applied in most of these robust PCAs assuming the sparsity of principal components or errors, which leaves these reformulated PCAs inapplicable when the principal components and errors are not sparse. Besides, it still requires the orthogonal assumption of principal components but renders untreated the data sets with multiple irregularly positioned spokes.

Recently, Tian et al. (2013) designed an *ℓ_1-norm based major component detection and analysis* (ℓ_1 MCDA) method specifically for 2D data with two or more irregularly positioned spokes. Radial data set is one situation that can be addressed by ℓ_1 MCDA. This reformulation of the PCA approach is based not just in part but exclusively on ℓ_1-norm, in relation to the L_1 spline technique for shape preservation of highly irregular data. It does not required orthogonality of the major components to be imposed. Deng et al. (2014) further extended ℓ_1 MCDA to recovering major components for higher dimensional data of a similar structure. They numerically validate the effectiveness of ℓ_1-norm for dealing with data following a heavy-tailed distribution or containing a significant number of outliers.

Existing ℓ_1 MCDA explicitly requires that the data come from a symmetric statistical distribution (or from several distributions irregularly superimposed). It could not be directly applicable for an asymmetric radial data cloud, partly due to the difficulty in estimating the central point in the data set. For a symmetric data set, the central point is estimated by either an ℓ_2-norm based multidimensional average in standard PCA methods or an ℓ_1-norm based coordinate-wise median in the ℓ_1 MCDA method. But these estimates may not be justifiable for a radial data cloud displaying not symmetrically positioned spokes.

This paper is to extend the ℓ_1 MCDA method for handling asymmetric radial data clouds. We proposed the ℓ_1-norm based central point analysis (ℓ_1 CPA) in a preliminary paper (An et al., 2018) for identifying a central point in an asymmetric radial data set. In this paper, we continue the investigation and introduce an algorithmic framework to locate an appropriate central point for the data set and recover its major directions and spreads. The first step of central point estimation accounts for preselecting a set of central point candidates and identifying the exact central point among the candidates by solving an ℓ_1-norm constrained discrete optimization model. The second step involves a fundamental reformulation of PCA framework, designed to accommodate handling a data set from an asymmetric heavy-tailed distribution. In contrast to the early ℓ_1 MCDA, this scheme avoids translating the data points into angular coordinates. Otherwise the applicability of ℓ_1 MCDA is subject to the appropriateness of higher-dimensional angular coordinate definition. In the rest of this paper, Section 2 explicitly proposes the extended ℓ_1 MCDA method for asymmetric radial data set. Section 3 provides extensive numerical experiments in support of the effectiveness of the proposed ℓ_1 MCDA method for handling various asymmetric radial data clouds possibly with heavy tails and patterned artificial outliers. Conclusion is given in Section 4.

2. EXTENDED ℓ_1 MCDA METHOD FOR ASYMMETRIC RADIAL DATA CLOUDS

In this section, we revisit our preliminary result on estimating a central point for asymmetric radial data and then present a complete algorithmic framework of ℓ_1 MCDA for detecting the major directions and spreads of the spokes. Throughout the paper, consider a radial data cloud $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$ in the n-dimensional real space with $\bar{\boldsymbol{x}}^m \in \mathbb{R}^n$, $m = 1, \ldots, M$. The ℓ_1 MCDA is divided into two steps:

1. Calculate the central point of data and subtract it out of the data.
2. Calculate the major directions of the shifted point cloud and the spreads of the point cloud along these major directions.

2.1 Calculation of the central point

The central point $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ is typically defined as minimizing the sum of its distances to all points in the given data cloud, *i.e.*,

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{m=1}^{M} d_p(\boldsymbol{x}, \bar{\boldsymbol{x}}^m), \tag{1}$$

where $d_p(\cdot,\cdot)$ is an ℓ_p-norm distance function for some $p \geq 1$. In standard PCA, the central point is obtained by using the ℓ_2-norm based distance function $d_2(\cdot,\cdot)$, which leads to the multidimensional average of the data set. Some literature on Robust PCAs (Crou et al., 2013; Fritz et al., 2012) use the distance function $d_2(\cdot,\cdot)$ for calculating a central point termed "L_1-median". These central point estimates might sometimes be flawed, especially for data sets containing heavy tails or outliers. The reason could be that ℓ_2-norm may considerably exaggerate the influence of heavy tails or outliers, or, that some heavy-tailed distributions even do not have a meaningful average.

In view of this, existing ℓ_1 MCDA uses the ℓ_1-norm based distance function $d_1(\cdot,\cdot)$ to compute coordinate-wise median as the central point estimate. One advantage of adopting the ℓ_1-norm is that it always exists for any data set, even for those assuming heavy-tailed distributions. Considering that ℓ_1-norm is more resistant to outliers than ℓ_2-norm, its robustness is another advantage. The central point estimate induced by ℓ_2-norm would adjust for the errors caused by a few outliers at the expense of deviating from the normal data points. But still, the multidimensional median may not be guaranteed to be appropriate for estimating a central point in an asymmetric data set. For example, a data cloud representing a corner, the central point should be the vertex of the "V". Yet a multidimensional median might be inside the cornered formed by the two meeting edges of "V". It might make sense to calculate multidimensional average or the coordinate-wise median for a symmetric data set, since the expected central point estimate is exactly where the symmetry happens. But for an asymmetric data set, like a V-shaped one, whether there exists a specific point around which the spokes are symmetrically positioned becomes questionable. In this situation, both the multidimensional average and the coordinate-wise median may result in a point far from the ideal center. It is necessary to propose an appropriate central point estimation method for principal component analysis of an asymmetric data set.

Instead, we characterize a central point with two desirable characteristics: (i) the total distance from a "central point" to all the data points is expected to be minimal; (ii) a "central point" should be close to where the spokes intersect and extend from. The first characteristic has long been recognized for conventional PCA methods. Since the ℓ_1-norm based distance function comes suitable for handling heavy tails and outliers, we keep using it. The second characteristic is specifically introduced for data with multiple irregularly positioned spokes, for example, a V-shape radial data set. A "central point" representing the corner is expected to include sufficiently many data points of the cloud in a relatively small vicinity, or, in other words, the central point should be able to include a good size (high percentage) of data points of the cloud in a minimal vicinity (small neighborhood).

To accommodate the two distinct characteristics of a desired central point, it takes two procedures to estimate a central point for an asymmetric radial data cloud. Procedure I is to prescreen possible candidates fitting characteristic (ii) by locating points that may include a given percentage of data points of the cloud in a smallest neighborhood. A set of candidates may be obtained by varying the given percentage at different quantiles. Procedure II is to choose among the set of candidates the one fitting characteristic (i), *i.e.*, the one with the smallest ℓ_1-norm based total distance to all the data points.

Step I: Preselect candidates of central point

The first step is to prescreen a set of central point candidates that should include sufficiently many data points in a minimal vicinity. In the process of generating such candidates, a range of predetermined quantiles are used to control the percentage of data points that are expected to be included in a small neighborhood. For simple implementation, we prescreen the candidate points in a coordinate-wise manner. More specifically, along each dimension we find the coordinates that require the smallest neighborhood in terms of the ℓ_1-norm based distance to include a given percentage of data points. Gathering all possible combinations of coordinates leads to a finite set of candidates for the central point. The corresponding algorithm for identifying candidates for the central point of the data cloud $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$ is as follows:

Algorithm A

Input: Parameters n, M, J; Data set $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$ with $\bar{\boldsymbol{x}}^m \in \mathbb{R}^n$, $m = 1,\ldots,M$; Quantile set $\{p_j\%\}_{j=1}^J$ with $p_j \in (0,100)$, $j = 1,\ldots,J$.

Initialize: Candidate point sets $C = \emptyset$. Coordinate index $i = 1$.

1. Sequence the data points $\bar{\boldsymbol{x}}^1,\ldots,\bar{\boldsymbol{x}}^M$ in an ascending/descending order based on the value of the i-th coordinate $\bar{x}_i^m$, $m = 1,\ldots,M$. Create coordinate set $C_i = \emptyset$

2. For each $j = 1,\ldots,J$, let $N_j = \lceil p_j\%M \rceil$ denote the number of data points to be included in a neighborhood, where $\lceil\cdot\rceil$ is the ceiling function. Calculate r_m=max$\{\bar{x}_i^m - \bar{x}_i^{m-\triangle N_j},\ \bar{x}_i^{m+\triangle N_j} - \bar{x}_i^m\}$

for each integer $m \in (\triangle N_j, M - \triangle N_j]$, where $\triangle N_j = \lceil \frac{N_j}{2} \rceil$. Find all indices m^* such that $m^* = \arg \min_{\triangle N_j < m \leq M - \triangle N_j} r_m$, and let $C_i \to C_i \cup \bar{x}_i^{m^*}$.

3. Update $i \to i + 1$; If $i \leq n$, return to Step 1; otherwise, terminate the algorithm.

Output: The candidate point set $C = C_1 \times \ldots \times C_n = \{\boldsymbol{x} \in \mathbb{R}^n : x_i \in C_i,\ i = 1, \ldots, n\}$.

Note that in item 2, along the i-th dimension, the coordinate $\bar{x}_i^m$ requires a neighborhood with the minimal radius r_m to include at least $\frac{p_j}{2}\%$ of coordinates $\{\bar{x}_i^m\}_{m=1}^M$ on its left and at least $\frac{p_j}{2}\%$ on its right. The coordinate $\bar{x}_i^{m^*}$ characterized by requiring a neighborhood of the least radius is then calculated and added to the set C_i. With all predetermined percentages $p_j\%$, $j = 1, \ldots, J$, being traversed, the set C_i is chosen as a subset of the i-th coordinate values of all data points and comprises a class of possible i-th coordinate values for a central point candidate.

Step II: Determine central point

In the second phase, we select among all candidate points the one that minimizes the total ℓ_1-norm distance to all data points of the cloud $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$. Specifically, we need to solve an optimization problem that minimizes the measure $\sum_{m=1}^M d_1(\boldsymbol{x}, \bar{\boldsymbol{x}}^m)$ under the restriction of $\boldsymbol{x} \in C$, where $d_1(\boldsymbol{x}, \bar{\boldsymbol{x}}^m) = \| \boldsymbol{x} - \bar{\boldsymbol{x}}^m \|_1$ is the ℓ_1-norm distance between $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}^m$ for $m = 1, \ldots, M$. For $C = C_1 \times \ldots \times C_n$, the problem can be explicitly written as the following ℓ_1-norm constrained discrete optimization program (P):

$$\begin{aligned} \min \quad & \sum_{m=1}^M r_m \\ \text{s.t.} \quad & \|\boldsymbol{x} - \bar{\boldsymbol{x}}^m\|_1 \leq r_m,\ m = 1, \ldots, M, \\ & x_i \in C_i,\ i = 1, \ldots, n, \\ & r_m \in \mathbb{R}_+,\ m = 1, \ldots, M. \end{aligned}$$

Denote $C_i = \{c_j^i \mid j = 1, ..., |C_i|\}$ for $i = 1, \ldots, n$. As C_i gets bigger in size, directly solving an exact solution to (P) comes with an unbearable computational complexity. We then reformulated (P) as a 0-1 mixed integer linear programming (MILP) program, which can be subsequently treated using commercial MILP solvers (*e.g.*, CPLEX, Gurobi) for fast computation. State-of-the-art linearization method are considered here.

Problem (P) can be reformulated into (MILP) program as below.

$$\min \quad \sum_{m=1}^M r_m \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^n t_i^m \leq r_m,\ m = 1, \ldots, M, \tag{3}$$

$$x_i - \bar{x}_i^m \leq t_i^m,\ m = 1, \ldots, M,\ i = 1, \ldots, n, \tag{4}$$

$$-x_i + \bar{x}_i^m \leq t_i^m,\ m = 1, \ldots, M,\ i = 1, \ldots, n, \tag{5}$$

$$\sum_{j=1}^{|C_i|} c_j^i u_j^i = x_i,\ i = 1, \ldots, n, \tag{6}$$

$$\sum_{j=1}^{|C_i|} u_j^i = 1,\ i = 1, \ldots, n, \tag{7}$$

$$\sum_{j \in G_l^i} u_j^i = \lambda_l^i,\ i = 1, \ldots, n,\ l = 1, \ldots, \lceil \log_2 |C_i| \rceil, \tag{8}$$

$$x_i \in \mathbb{R},\ r_m,\ t_i^m \in \mathbb{R}_+,\ i = 1, \ldots, n,\ m = 1, \ldots, M, \tag{9}$$

$$u_j^i \geq 0,\ \lambda_l^i \in \{0, 1\},\ i = 1, \ldots, n,\ j = 1, \ldots, |C_i|,\ l = 1, \ldots, \lceil \log_2 |C_i| \rceil. \tag{10}$$

For purpose of illustrating complexity of the program (MILP), assume in each coordinate set C_i there exists one unique coordinate value that correspond to each percentage $p_j\%$, $j = 1, \ldots, J$, *i.e.*, $|C_i| = J$ for $i = 1, \ldots, n$, without loss of generality. The reformulation model (MILP) requires n continuous variables of x_i, M nonnegative continuous variables of r_m, nM nonnegative continuous variables of t_i^m, nJ $[0, 1]$-bounded continuous variables of u_j^i, $n\lceil \log_2 J \rceil$ binary variables of λ_l^i, $M + 2nM$ linear inequality constraints, and $2n + n\lceil \log_2 J \rceil$ linear equality constraints.

Solving for an optimal solution $\boldsymbol{x}^* \in \mathbb{R}^n$ of program (MILP), we obtain a desired central point chosen from the candidate set C for the asymmetric radial data cloud $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$. We then subtract the central point out from any of the

data points. This way, the whole data set is repositioned to be centering at the origin of the space. To avoid notation abuse, we still denote the modified data cloud as $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$.

An et al. (2018) performed extensive numerical studies to support the effectiveness of ℓ_1-norm based central point analysis (ℓ_1 CPA) by comparing against the multidimensional average and the coordinate-wise median. They showed that ℓ_1 CPA is suitable in providing a correct central point for an asymmetric radial data set demonstrating great robustness.

Having estimated a central point $\bar{\boldsymbol{x}}_0$, it is then subtracted from each data point in the data sample $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$. This way, the data sample is repositioned to be centering around the origin.

2.2 Calculation of the major directions and median length

The "major directions" are the directions along which the spokes spread. The "median length" measures how far the spoke spreads along each direction. Assume $f_\theta(r)$ is the probability density function of the distance between a data point lying in the direction θ and the origin. The median of $f_\theta(r)$ is defined to be the median length of the spoke in the direction θ. Or, the value of median length $r^*(\theta)$ is such that $\int_0^{r^*(\theta)} f_\theta(r)dr = 0.5$. A direction $\hat{\theta}$ is a major direction if $r^*(\theta)$ is a local maximum in the angular space θ.

Our goal is to estimate the major directions and the median length for each spoke in an asymmetric radial data set which might contain heavy tails or outliers. To deemphasize the impact of outliers, we follow the current framework that reformulates PCA approach based exclusively on the ℓ_1-norm.

It is estimated by a two-level median of sample points over a small angular neighborhood. The algorithm for calculating the major directions and median radii in those directions is as follows:

Algorithm B

Input: Data set $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$ with $\bar{\boldsymbol{x}}^m \in \mathbb{R}^n$, $m = 1, \ldots, M$.

1. For each $\bar{\boldsymbol{x}}^m = (\bar{x}_1^m, \ldots, \bar{x}_n^m)$, $m = 1, \ldots, M$, calculate its length

$$\bar{r}^m = \sqrt{(\bar{x}_1^m)^2 + \cdots + (\bar{x}_n^m)^2}, \tag{11}$$

and corresponding direction vector $\bar{\theta}^m = (\bar{\theta}_1^m, \bar{\theta}_2^m, \ldots, \bar{\theta}_n^m)$ with

$$\bar{\theta}_i^m = \frac{\bar{x}_i^m}{\bar{r}^m},\ i = 1, \ldots, n. \tag{12}$$

2. Randomly choose a $\bar{m} \in \{1, \ldots, M\}$ with the direction $\bar{\theta}^{\bar{m}}$ to start. Create a visited set $Q = \{\bar{m}\}$.

3. Apply the k-nearest-neighbors (k-NN) method (Friedman et al., 1977) with an appropriate $p_0 \in (0, 1)$ to find a neighbor $\mathcal{N}^{\bar{m}}$ with $k = p_0 * M$ elements (directions) that are nearest to $\bar{\theta}^{\bar{m}}$ in a given angular measure (ℓ_1-norm or ℓ_2-norm).

4. For each direction $\bar{\theta}^{\bar{m}_i} \in \mathcal{N}^{\bar{m}}$, $i = 1, \ldots, k$, apply the k-NN method to determine a neighbor $\mathcal{N}^{\bar{m}_i}$ with k ($p_0 * M$) elements (directions) that are nearest to $\bar{\theta}^{\bar{m}_i}$. For each direction $\bar{\theta}^{\bar{m}_{ij}} \in \mathcal{N}^{\bar{m}_i}$, $j = 1, \ldots, k$, apply the k-NN method to determine a neighbor $\mathcal{N}^{\bar{m}_{ij}}$ with k ($p_0 * M$) elements (directions) that are nearest to $\bar{\theta}^{\bar{m}_{ij}}$. Calculate the median $\tilde{r}^{\bar{m}_{ij}}$ of the lengths $\{\bar{r}^m \mid \bar{\theta}^m \in \mathcal{N}^{\bar{m}_{ij}}\}$ and then the median $\hat{r}^{\bar{m}_i}$ of the lengths $\{\tilde{r}^{\bar{m}_{ij}} \mid \bar{\theta}^{\bar{m}_{ij}} \in \mathcal{N}^{\bar{m}_i}\}$.

5. Determine the maximum of the median lengths $\hat{r}^{\bar{m}_i}$, $i = 1, \ldots, k$. Let $i^* = \arg\max_{i=1,\ldots,k}\{\hat{r}^{\bar{m}_i}\}$. If the maximum is achieved at $\bar{m}_{i^*} \in Q$, go to Step 6. Otherwise, update $Q \to Q \cup \{\bar{m}_{i^*}\}$, $\bar{m} = \bar{m}_{i^*}$, and return to Step 3.

6. Apply the k-NN method to determine a neighbor $\mathcal{N}$ with $\lceil \frac{k}{2} \rceil$ elements (directions) that are nearest to the local maximal direction identified in Step 5.

Output: The median of $\bar{\theta}^m$ in $\mathcal{N}$ is the estimated major direction of the data cloud. The median of the lengths $\{\bar{r}^m \mid \bar{\theta}^m \in \mathcal{N}\}$ is the estimated median length for that major direction.

Note that a well-known fact indicates there is no explicit formula to determine the optimum value of k (in our case, $p_0 * M$) for applying the k-NN method. Trail and error is needed to find the appropriate value of p_0 in practice.

The above procedure will terminate since, in the worst case, it traverses all directions $\bar{\theta}^m$ in a finite data set. This procedure outputs one major direction of a spoke for an input data set. To find all major directions, one can choose different starting point for multiple implementations of this algorithm. One way is to randomly find a direction that is orthogonal to the major directions obtained so far and selects the data point nearest to this direction as the new starting point.

Algorithm B calculates two-level medians in order to develop a scalable procedure for high dimensional data sets. There are other approaches, like fitting quadratic surfaces in the ℓ_1-norm to the data in the neighborhoods [16]. The neighbors required in Algorithm B are calculated by the k-nearest-neighbors (k-NN) method. One can try other methods for finding neighbors. The weakness of quadratic fitting approach is the fact that it may not be linearly scalable as the dimension increases, when both the size of the local neighborhood and the computing time increase quadratically.

3. NUMERICAL EXPERIMENTS

In this section, we perform numerical experiments to examine the effectiveness of our proposed ℓ_1 MCDA. We first test ℓ_1 MCDA on various types of radial data clouds with or without heavy tails and outliers, and then investigate the ability of ℓ_1 MCDA to recover major spokes from asymmetric radial data sets in high dimensional space.

All computational experiments are conducted using MATLAB R2016a on a PC equipped with the Intel Core i7-2600 CPU, 8 GB RAM and Windows 7 (64 bit) operating system. Mixed integer programming problems are solved by *Gurobi* (7.0.1) using the MIP Solver.

We use different types of asymmetric radial data clouds to examine the effectiveness of ℓ_1 MCDA. The asymmetric radial data cloud may contain one spoke or multiple irregularly positioned spokes. The asymmetry feature presents a significant challenge for the state-of-the art robust PCA methods. The data clouds are sampled from Gaussian distribution (light-tailed) or Student t distribution with 1 degree of freedom (heavy-tailed). Clustered outliers may be included in the asymmetric radial data clouds to test the robustness of ℓ_1 MCDA. Overall, we randomly generate the following different types of asymmetric radial data clouds.

- One-spoke radial Gaussian-based distribution with or without artificial outliers;
- One-spoke radial Student t-based distribution with or without artificial outliers;
- Two-spoke (V-shaped) radial Gaussian-based distribution with or without artificial outliers;
- Two-spoke (V-shaped) radial Student t-based distribution with or without artificial outliers;
- Four-spoke radial Gaussian-based distribution with or without artificial outliers;
- Four-spoke radial Student t-based distribution with or without artificial outliers.

To generate these types of data clouds, we start with a symmetric data sample $\{\bar{\boldsymbol{x}}^m\}_{m=1}^M$ with $\bar{\boldsymbol{x}}^m \in \mathbb{R}^n$, $m = 1, \ldots, M$, from multivariate Gaussian and Student t distributions which can be obtained using the MATLAB *mvnrnd* and *mvtrnd* modules, respectively. The symmetric data sample is designed with a mean vector $(0, \cdots, 0)^T$ and a covariance/correlation matrix

$$\theta(n) = \begin{bmatrix} 1 & b & \ldots & b \\ b & 1 & \ldots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \ldots & 1 \end{bmatrix}_{n\times n},$$

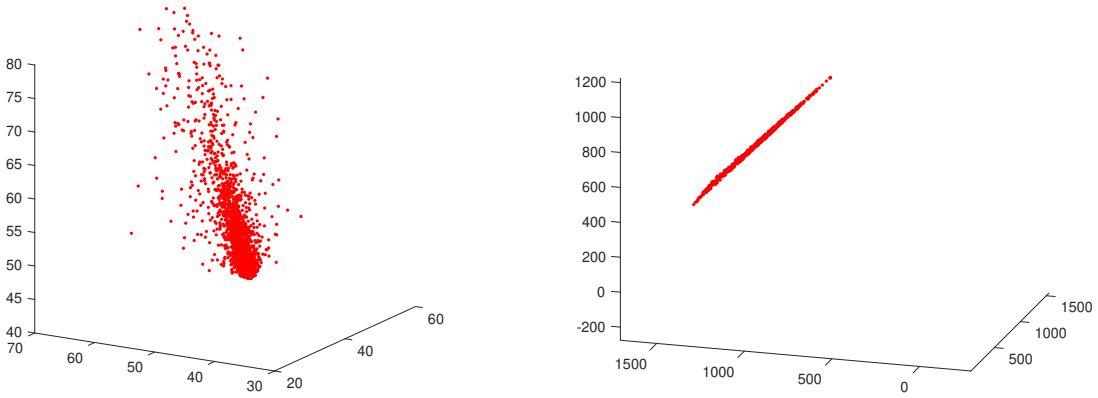
where $0 < b < 1$ is a given parameter defined as $b = \frac{K^2-1}{n-1+K^2}$. The constant $K > 0$ is the ratio of the longest major direction's length to that of each other major direction in the symmetric data, which is also the ratio of the maximum eigenvalue of $\theta(n)$ to the minimal eigenvalue of $\theta(n)$. In our numerical experiment, we set $K = 10$. The generated data set is symmetric around the origin with two spokes in the opposite directions of $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})^T$ and $(-\frac{1}{\sqrt{n}}, \cdots, -\frac{1}{\sqrt{n}})^T$, respectively. Next, we generate different types of asymmetric radial data sets in the following manner.

- One-spoke radial data set is one radial data spoke emanating from the origin along the direction of $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})^T$, created by taking $(|\bar{x}_1^m|, |\bar{x}_2^m|, \ldots, |\bar{x}_n^m|)$ for each $\bar{\boldsymbol{x}}^m = (\bar{x}_1^m, \bar{x}_2^m, \ldots, \bar{x}_n^m)$ in the symmetric data sample. Figure 1(a) displays an example of an one-spoke radial data cloud with 8000 3D data points generated from Student t distribution. Since a Studnet-t-distribution based data cloud contains a percentage of heavy tails, there are many points outside the plot range.
- V-shaped data set is one obtained through overlying two separate one-spoke radial data clouds. Each of the two spokes is originally one radial data spoke along the major direction $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})^T$, then rotated to another randomly generated major direction and multiplied by a factor, 1.0 and 1.5, respectively. Figure 2(a) shows an example of a 3D V-shaped data cloud with 8000 Student t-distributed data points in each spoke.
- Four-spoke asymmetric radial data set emanates from the origin and spreads out along four irregularly positioned major directions. It is created by overlying four one-spoke radial data samples rotated to different randomly generated directions. The four spokes are then multiplied by a factor, 1.0, 1.5, 0.6, 3.0, respectively. Note that the four spokes are not orthogonal to each other. Figure 3(a) shows an example of a 3D radial data cloud with four Student-t based spokes each containing 8000 points.

In all of the above asymmetric radial sets, each spoke radiates from the origin so the data center is always the origin. To create a general asymmetric radial data set, we add an arbitrary $\boldsymbol{x}_0$ to each data point in any given sample. The position x_0 then becomes the theoretical center for that data sample.

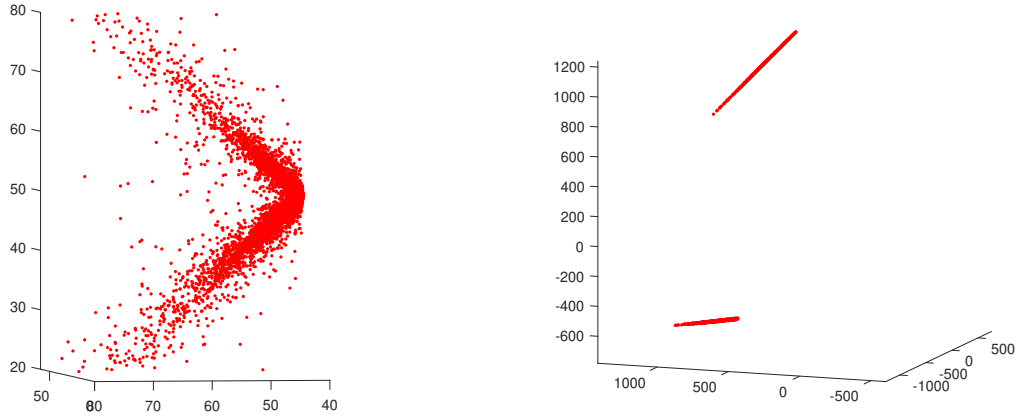
We generate random data sets from both Gaussian and Student t distributions. Considering the data set generated from Student t distribution contains a number of heavy tails while one generated from Gaussian distribution does not, these two distributions represented a significant contrast of challenge for ℓ_1 MCDA .

Each type of data cloud may contain a percentage of outliers. We generate artificial outliers from a uniform distribution on a simplex in $\mathbb{R}^n$ in a way that each spoke corresponds to one simplex. The n vertices of a simplex are randomly chosen at the intersection of the ball $x_1^2 + x_2^2 + \cdots + x_n^2 = 1500^2$ with a hyperplane $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 1000$. The normal vector $(\alpha_1, \alpha_2, \ldots, \alpha_n)^T \in \mathbb{R}^n$ of the simplex is one that has 45° angle with the major direction of the spoke. In Figures 1(b), 2(b), 3(b), we report examples of Student t-based one-spoke, V-shaped, and four-spoke radial data sets with 10% additional artificial outliers, respectively.



(a) (b)

Fig. 1:One-spoke radial data set and outliers



(a) (b)

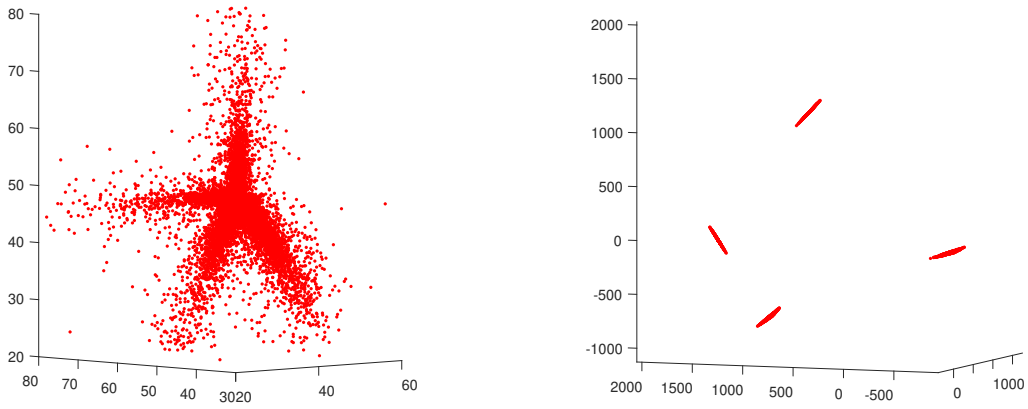
Fig. 2:V-shape radial data set and outliers

For each type of data cloud, we implement ℓ_1 MCDA on 30 randomly generated data asymmetric radial sets, without and with 10% additional artificial outliers, respectively.

The theoretical values of the major directions are predetermined and the theoretical value of the median lengths is calculated from numerical integration. To measure the accuracy of our estimation, we calculate the average for the absolute difference between the output major directions and the theoretical major directions (*i.e.*, "av. abs_err of angle" in all Tables) as well as the average for the relative difference between the output median length and the theoretical median length (*i.e.*, "av. rel_err of length" in all Tables).

In this setting, standard PCA, any of Croux and Ruiz-Gzen's projection-pursuit and Ke and Kanada's ℓ_1 factorization outputs one single major direction estimate not yielding any meaningful information about the real major directions for an asymmetric radial data set. But ℓ_1 MCDA is successful in locating multiple major directions and estimate their spreads with accuracy comparable to the theoretical values.

Tables 1-2 summarize the computational results for Gaussian-based and Student t-based one-spoke, V-shaped, and four-spoke radial data sets, resulting from implementation of ℓ_1 MCDA with local neighborhood of 2%, 1%, 0.5%



(a) (b)

Fig. 3:Four-spoke radial data set with random spokes and outliers

of total points. Tables 1-2 both indicate that, as the number of spokes increase, the overall accuracy of ℓ_1 MCDA relatively decreases but no so much that ℓ_1 MCDA performs effectively in estimating the major directions and spreads.

This decrease in accuracy is due to the fact that if the multiple randomly generalized spokes are not located separately enough, they might interfere with each other. Comparison of Tables 1-2 indicates that when dealing with data containing heavy tails, ℓ_1 MCDA, despite becoming less accurate, is still able to provide proper estimates.

Table 1: Results of ℓ_1 MCDA on Gaussian-based radial data sets

# of spokes	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
1	0.0365	0.0681	0.0206	0.0246	0.0152	0.0191
2	0.0581	0.1196	0.0254	0.0648	0.0157	0.0380
4	0.0618	0.1512	0.0272	0.0950	0.0169	0.0605

Table 2: Results of ℓ_1 MCDA on Student t-based radial data sets

# of spokes	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
1	0.0784	0.1758	0.0438	0.1301	0.0269	0.0738
2	0.1021	0.3658	0.0546	0.1511	0.0370	0.0819
4	0.1652	0.3833	0.0688	0.1684	0.0468	0.1065

We can also observe that ℓ_1 MCDA obtains better estimates by using a relatively large neighbor in our case. The choice of neighbor size is one important implementation detail. In our numerical experiments, ℓ_1 MCDA obtains good accuracy by using a small neighbor size (up to 2% of the size of given data) for the test cases. Estimates of major directions and spreads obtained by ℓ_1 MCDA become more accurate as the neighbor size increases.

Tables 3-4 present the computational results for Gaussian-based and Student t-based one-spoke, V-shaped, and four-spoke radial data sets with 10% outliers. Though the existence of outliers slightly compromises the overall accuracy, ℓ_1 MCDA delivers estimates of major directions and spreads for data containing outliers with high quality. This supports the robustness of ℓ_1 MCDA. In this case, a too small neighbor size may not be desirable for obtaining accurate estimates. Overall, ℓ_1 MCDA provides an effective tool to deal with asymmetric radial data set that standard and robust PCAs cannot handle.

Next, we test the performance of ℓ_1 MCDA for handling asymmetric radial data set in a high dimensional space. We generate nD radial data sets with four fixed spokes of varying dimension $n = 3, ..., 10$, by superimposing four one-spoke radial samples that have been rotated to the directions $(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})^T$, $(1, 0, 0, ..., 0)^T$, $(0, 1, 0, ..., 0)^T$, and $(0, 0, 1, ..., 0)^T$, respectively. One-spoke samples, either Gaussian-based or Student t-based, are rotated in ℓ_2-norm after multiplying by a factor (1.0, 1.5, 0.6, 3.0, respectively). To account for the robustness testing, we conduct

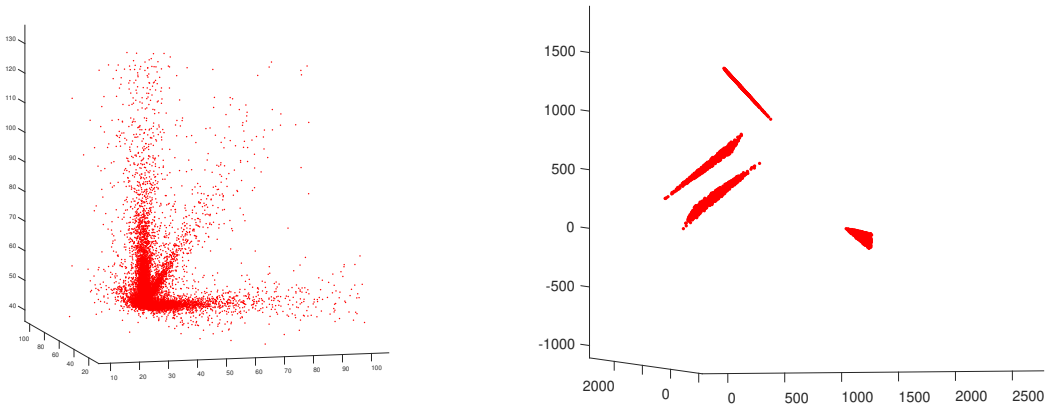
experiments on the nD four-spoke radial data sample possibly containing 10% artificial outliers. Figures 4(a) and 4(b) show an example of the 3D Student t-based four-spoke radial data without and with outliers, respectively.

Table 3: Results of ℓ_1 MCDA on Gaussian-based radial data sets (with 10% outliers)

# of spokes	k-NN=40(0.5%) av. abs_err of angle	k-NN=40(0.5%) av. rel_err of length	k-NN=80(1%) av. abs_err of angle	k-NN=80(1%) av. rel_err of length	k-NN=160(2%) av. abs_err of angle	k-NN=160(2%) av. rel_err of length
1	0.0874	0.1114	0.0650	0.0513	0.0491	0.0305
2	0.1199	0.1826	0.0837	0.0713	0.0721	0.0386
4	0.1301	0.3147	0.0939	0.1246	0.0958	0.0854

Table 4: Results of ℓ_1 MCDA on Student t-based radial data sets (with 10% outliers)

# of spokes	k-NN=40(0.5%) av. abs_err of angle	k-NN=40(0.5%) av. rel_err of length	k-NN=80(1%) av. abs_err of angle	k-NN=80(1%) av. rel_err of length	k-NN=160(2%) av. abs_err of angle	k-NN=160(2%) av. rel_err of length
1	0.1522	0.1819	0.1288	0.1403	0.0979	0.0966
2	0.1732	0.3950	0.1342	0.1638	0.1070	0.1014
4	0.1869	0.4729	0.1456	0.1879	0.1129	0.1115



(a) (b)

Fig. 4:Four-spoke radial data set with fixed spokes and outliers

For each n, we implement ℓ_1 MCDA on 30 randomly generated data sets. Tables 5-6 provides the computational results for Gaussian-based and Student-t based four-spoke radial data sets with local neighborhood of 2%, 1%, 0.5% of total points. Tables 7-8 gives the result from implementation of ℓ_1 MCDA on data sets with 10% outliers. Tables 5-8 indicate that ℓ_1 MCDA is a suitable tool to estimate major spokes and spreads for an asymmetric radial data set in a high dimensional space. As the dimension of data space increases, the performance of ℓ_1 MCDA decreases in terms of accuracy. Overall, our result confirm that ℓ_1 MCDA works to deliver major direction and spread estimates in high dimensional spaces.

A closer look at Tables 5-6 reveals that the performance of ℓ_1 MCDA suffers from the existence of heavy tails slightly in estimating major direction, but mostly in estimating spread. We observe that ℓ_1 MCDA obtains more accurate estimates by using a relatively small neighbor down to 0.5% in most cases for Gaussian-based data, or using a relatively large neighbor up to 2% for Student t-based data.

Tables 7-8 indicate though existence of outliers might compromise the accuracy, ℓ_1 MCDA can estimate major directions and spreads for data containing a percentage of outliers if choosing an appropriate neighbor size. This attests to the robustness of ℓ_1 MCDA in a high dimensional space. In our case, the performance of ℓ_1 MCDA increases in accuracy as we choose a relatively small neighbor.

Table 5: Results of ℓ_1 MCDA on four-spoke Gaussian-based high-dimensional radial data sets

n	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
4	0.0290	0.0462	0.0204	0.0323	0.0147	0.0314
5	0.0293	0.0384	0.0213	0.0438	0.0159	0.0693
6	0.0281	0.0571	0.0234	0.0872	0.0190	0.1236
7	0.0322	0.0938	0.0261	0.1282	0.0215	0.1638
8	0.0349	0.1265	0.0281	0.1730	0.0233	0.2059
9	0.0364	0.1676	0.0306	0.1946	0.0262	0.2344
10	0.0410	0.1993	0.0342	0.2370	0.0259	0.2656

Table 6: Results of ℓ_1 MCDA on four-spoke Student t-based high-dimensional radial data sets

n	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
4	0.0566	0.1686	0.0374	0.1537	0.0308	0.0571
5	0.0535	0.2512	0.0422	0.1903	0.0310	0.1236
6	0.0597	0.3047	0.0458	0.2297	0.0373	0.1891
7	0.0562	0.3758	0.0477	0.3133	0.0396	0.2196
8	0.0595	0.4134	0.0513	0.3774	0.0399	0.2621
9	0.0637	0.4993	0.0484	0.4334	0.0403	0.3248
10	0.0631	0.5456	0.0524	0.4915	0.0406	0.3655

Table 7: Results of ℓ_1 MCDA on four-spoke Gaussian-based high-dimensional radial data sets (with 10% outliers)

n	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
4	0.0625	0.0512	0.0564	0.0745	0.0764	0.0925
5	0.0718	0.0561	0.0773	0.0625	0.1054	0.0883
6	0.0883	0.0779	0.0920	0.1043	0.1133	0.1356
7	0.1052	0.1095	0.1093	0.2052	0.1385	0.2289
8	0.1273	0.1423	0.1329	0.2296	0.1543	0.2525
9	0.1354	0.2036	0.1463	0.2583	0.1638	0.2723
10	0.1532	0.2826	0.1693	0.3125	0.1853	0.3329

Table 8: Results of ℓ_1 MCDA on four-spoke Student t-based high-dimensional radial data sets (with 10% outliers)

n	k-NN=40(0.5%)		k-NN=80(1%)		k-NN=160(2%)	
	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length	av. abs_err of angle	av. rel_err of length
4	0.0902	0.2050	0.0912	0.2623	0.1121	0.3325
5	0.1083	0.2843	0.0958	0.3023	0.1293	0.3976
6	0.1042	0.2969	0.1234	0.3749	0.1343	0.5382
7	0.1395	0.3738	0.1495	0.4583	0.1564	0.6752
8	0.1639	0.4820	0.1894	0.6323	0.2034	0.7372
9	0.1776	0.5723	0.2004	0.7784	0.2432	0.8393
10	0.2488	0.7293	0.2595	0.8592	0.2854	0.9034

4. CONCLUSION

As continuation of principal component analysis (PCA), we have developed a complete algorithmic framework of ℓ_1 major component detection analysis (ℓ_1 MCDA) for treating multivariate data of radial structure with multiple asymmetrically positioned spokes. The extended ℓ_1 MCDA method consists of locating the central point from which each spoke diverges and calculating the major directions and median lengths for those directions. Our ℓ_1 MCDA method does not require the assumption of major components being orthogonal or sparse and features the exclusive use of ℓ_1 norm. It is also designed for allowing implementation in a high-dimensional space. In contrast to the early ℓ_1 MCDA, this ℓ_1 MCDA procedure avoids translating the data points into angular coordinates, otherwise the applicability of ℓ_1 MCDA is subject to the appropriateness of higher-dimensional angular coordination definition. Extensive numerical experiments have been conducted to show its remarkable capability of detecting major components for asymmetric ra-

dial data and its robustness. They also corroborate that ℓ_1 MCDA can provide as a practice tool in recovering complex spoke structures from large-scale data clouds in a high dimensional space.

The ℓ_1 MCDA provides the foundation for identification of data structure and further compression, with the guideline of exclusive use of ℓ_1 operations. It can serve as a robust tool in terrain modeling, geometric modeling, image analysis, information mining and general pattern recognition. For future research and development, one direction is to design a major component detection algorithm with better scalability or one that can be implemented on the parallel and distributed computers in a way that the applicability of ℓ_1 MCDA can be extended to data of high dimensions (10^4 to 10^6). It would also be interesting to ponder the question of how the ℓ_1 MCDA can be extended to deal with missing data.

Acknowledgement

This work has been supported by the US Army Research Office Grant #W911NF-15-1-0223.

References

1. An, Q., Fang, S.C., Nie T. and Jiang S. (2018) 'ℓ_1-norm based central point analysis for asymmetric radial data', *Annals of Data Science*, Vol. 5 No. 3, pp. 469-486.
2. Aruga R. (2003) 'The problem of multivariate classification of samples with radial (or V-shaped) chemical data', *Talanta*, Vol. 60 No. 5, pp. 937-944.
3. Brooks J., Dulá J. and Boone E. (2013) 'A pure ℓ_1-norm principal component analysis', *Computational Statistics & Data Analysis*, Vol. 61, pp. 83-98.
4. Candès, E.J., Li X., Ma Y. and Wright, J. (2011) 'Robust principal component analysis', *Journal of the ACM*, Vol. 58 No. 3, pp. 11.
5. Chartrand R. (2007) 'Exact reconstruction of sparse signals via nonconvex minimization', *IEEE Signal Process. Lett.*, Vol. 14 No. 10, pp. 707-710.
6. Choulakian V. (2006) 'L_1-norm projection pursuit principal component analysis', *Computational Statistics & Data Analysis*, Vol. 50 No. 6, pp. 1441-1451.
7. Croux C., Filzmoser P. and Fritz H. (2013) 'Robust sparse principal component analysis', *Technometrics*, Vol. 55 No. 2, pp. 202-214.
8. Croux C., Ruiz-Gazen A. (2005) 'High breakdown estimators for principal components: the projection-pursuit approach revisited', *Journal of Multivariate Analysis*, Vol. 95 No. 1, pp. 206-226.
9. Deng Z., Lavery J.E., Fang S.-C. and Luo J. (2014) 'ℓ_1 major component detection and analysis (ℓ_1 MCDA) in three and higher dimensional spaces', *Algorithms*, Vol. 7 No. 3, pp. 429-443.
10. Friedman J., Bentley J. and Finkel R. (1977) 'An algorithm for finding best matches in logarithmic expected time', *ACM Transactions on Mathematical Software (TOMS)*, Vol. 3 No. 3, pp. 209-226.
11. Fritz H., Filzmoser P. and Croux C. (2012) 'A comparison of algorithms for the multivariate L_1-median', *Computational Statistics*, Vol. 27 No. 3, pp. 393-410.
12. Gribonval R., Nielsen M. (2006) 'Sparse approximations in signal and image processing', *Signal Process*, Vol. 86 No. 3, pp. 415-416.
13. Jin Q., Lavery J.E. and Fang S.-C. (2010) 'Univariate cubic L^1 interpolating splines: Analytical results for linearity, convexity and oscillation on 5-point windows', *Algorithms*, Vol. 3 No. 3, pp. 276-293.
14. Jolliffe I. (2002) Principal Component Analysis, 2nd ed., Springer: New York City.
15. Luo J., Deng Z., Bulatov D., Lavery J.E. and Fang S.-C. (2013) 'Comparison of an ℓ_1-regression-based and a RANSAC-based planar segmentation procedure for urban terrain data with many outliers', *Image and Signal Processing for Remote Sensing XIX*, Vol. 8892, pp. 09.
16. Massart B., Guo Q., Questier F., Massart D.L., Boucon C., De Jong S. and Vandeginste B.G. (2001) 'Data structures and data transformations for clustering chemical data', *TrAC Trends in Analytical Chemistry*, Vol. 20 No. 1, pp. 35-41.
17. Nie T., Wang Z., Fang S.-C. and Lavery J.E. (2017) 'Convex shape preservation of cubic L^1 spline fits', *Annals of Data Science*, Vol. 4 No. 1, pp. 1-25.
18. Rodriguez R., Schuur E.R., Lim H.Y., Henderson G.A., Simons J.W. and Henderson D.R. (1997) 'Prostate attenuated replication competent adenovirus (ARCA) CN706: a selective cytotoxic for prostate-specific antigen-positive prostate cancer cells', *Cancer Research*, Vol. 57 No. 13, pp. 2559-2563.

19. Tian Y., Jin Q., Lavery J.E. and Fang S.-C. (2013) 'ℓ_1 major component detection and analysis (ℓ_1 MCDA): Foundations in two dimensions', *Algorithms*, Vol. 6 No. 1, pp. 12-28.
20. Wang Z., Lavery J.E. and Fang S.-C. (2014) 'Approximation of irregular geometric data by locally calculated univariate cubic L^1 spline fits', *Annals of Data Science*, Vol. 1 No. 1, pp.5-14.