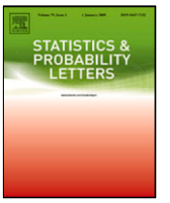
Statistics and Probability Letters 203 (2023) 109920

Contents lists available at ScienceDirect

Statistics and Probability Letters

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



On the use of robust estimators of multivariate location for heterogeneous data

Raúl Pérez-Fernández

Department of Statistics and O.R. and Mathematics Didactics, University of Oviedo, Calle Federico García Lorca 18, 33007 Oviedo, Spain

ARTICLE INFO

Article history:
Received 18 March 2021
Accepted 9 August 2023
Available online 12 August 2023

Keywords:
Estimation of multivariate location
Robustness
Orthogonal equivariance
Componentwise increasingness
Heterogeneous data

ABSTRACT

The properties of orthogonal equivariance and componentwise increasingness are often discussed for estimators of multivariate location. The former property is linked to a coordinate-free nature of the data (as in spatial data), whereas the latter property is linked to an ordered nature of the data (typically formed by several univariate variables). Since both properties do not get along well with each other, if robustness in the presence of outliers is to be pursued, one must choose between them. Admittedly, most of the literature on multivariate estimation of location, in which affine equivariance is pursued, typically abandons the latter property. Unfortunately, in the more and more common presence of heterogeneous data (as in spatio-temporal data), both orthogonal equivariance (therefore, affine equivariance) and componentwise increasingness might bring very disappointing results. For this very reason, both properties should be forfeited and only required for some of the components.

1. Introduction

Statisticians are often interested in the estimation of different parameters describing a certain population. It is therefore unsurprising that estimators of location have become a prominent subject of study in statistics. Actually, the analysis of estimators of location currently reaches far beyond the most probabilistic and sampling-oriented approach to statistics and has become a core topic in fields as varied as aggregation theory (where estimators of location are referred to as idempotent aggregation functions, averaging functions or means) (Grabisch et al., 2011), descriptive statistics (where they are referred to as central tendencies, measures of central tendency or averages) (Wilcox and Keselman, 2003), cluster analysis (where they are referred to as cluster centers) (MacQueen, 1967) and operations research (where they are referred to as optimal facility locations) (Owen and Daskin, 1998).

The field of multivariate statistics is concerned with the simultaneous analysis of more than one variable. If the variables are assumed to be independent, the estimation of the multivariate location can be performed for each variable individually. However, if there is no guarantee of independence, the estimation of the multivariate location becomes more involved, resulting in a vast literature. Some prominent examples of estimators of multivariate location are the (weighted) centroid (Gagolewski et al., 2020), the Euclidean center (Sylvester, 1857) and different extensions of the median to higher dimensions (Small, 1990) such as the componentwise median (Gagolewski, 2017), the medoid (Kaufman and Rousseeuw, 1990), the spatial median (Vardi and Zhang, 2000; Weber, 1909; Weiszfeld, 1937), the orthomedian (Grübel, 1996), Tukey's halfspace median (Tukey, 1975), Oja's simplex median (Oja, 1983), the convex hull peeling median (Eddy, 1982) and the simplicial depth median (Liu, 1990).

E-mail address: perezfernandez@uniovi.es.

https://doi.org/10.1016/j.spl.2023.109920

The field of robust statistics focuses on reducing the influence of outliers (Rousseew and Hubert, 2011) in several tasks such as parameter estimation (Croux and Dehon, 2014; Huber, 1964), regression (Rousseeuw, 1984), principal component analysis (Hubert et al., 2005) and cluster analysis (Cuesta-Albertos et al., 1997; Gagolewski et al., 2016; Gallegos and Ritter, 2005). In particular, we are here interested in robust statistics for the estimation of multivariate location, which is oftentimes performed jointly with the estimation of scatter. Some prominent estimators of multivariate location and scatter are: M-estimators (Maronna, 1976), the Stahel–Donoho estimator (Donoho, 1982; Stahel, 1981), the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984), the Minimum Volume Ellipsoid (MVE) estimator (Rousseeuw, 1985) and S-estimators (Lopuhaä, 1989; Rousseeuw and Leroy, 1987).

The property of affine equivariance (Rousseeuw and Hubert, 2013) (sometimes referred to as affine invariance (Maronna, 1976)) has been largely venerated as a necessary property for estimators of multivariate location since it assures that the estimator behaves well after applying an affine transformation to the space (e.g., a translation, scaling, rotation or reflection). At least, the weaker property of orthogonal equivariance, which assures that the estimator behaves well after applying an orthogonal transformation to the space (e.g., a rotation or reflection), is typically required. Unfortunately, orthogonal equivariance does not get along well with componentwise increasingness — as first pointed out in Gagolewski (2015, 2017) and further explored in Gagolewski et al. (2020), Pérez-Fernández et al. (2019) – thus restricting estimators of multivariate location that satisfy both properties to the family of weighted centroids (i.e., componentwise extensions of a single weighted arithmetic mean). Since weighted centroids are known to be non-robust, it thus becomes necessary to choose between orthogonal equivariance and componentwise increasingness. In this paper, it is discussed that this choice must be made depending on the characteristics of the data. More precisely, in the presence of heterogeneous data, both properties should be forfeited and only required for some of the components.

The remainder of the paper is structured as follows. In Section 2, some preliminaries on the estimation of multivariate location are provided and, in particular, the properties of orthogonal equivariance and componentwise increasingness are presented. Section 3 presents some real-life examples of use of the estimation of multivariate location for heterogeneous data. Section 4 discusses some guidelines for the estimation of multivariate location in the context of heterogeneous data, for which none of orthogonal equivariance or componentwise increasingness should be required. We end with some concluding remarks in Section 5.

2. Two basic types of multivariate data

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n points in $\mathbb{R}^d$. We denote by $\mathbf{x}(j)$ the jth component of a point $\mathbf{x} \in \mathbb{R}^d$, and, given $D \subseteq \{1, \ldots, d\}$, we denote by $\mathbf{x}(D)$ the components of $\mathbf{x}$ corresponding to the indices in D. A function $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called idempotent if it is such that $T(\mathbf{x}, \ldots, \mathbf{x}) = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^d$. Throughout this paper, we deal with idempotent functions of the type $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ and refer to them as estimators of multivariate location. The point $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is referred to as the estimate of multivariate location based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

2.1. Estimation of multivariate location for lists of univariate variables

A common type of multivariate data just consists of many univariate variables (e.g., height and weight of different individuals). A natural property for this type of data is that of componentwise increasingness (or componentwise monotonicity, as in Gagolewski et al., 2020; Pérez-Fernández et al., 2019), which requires that, if all the data points are increased on a certain component, then the estimate should also increase on that component.

Definition 1. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called *componentwisely increasing* if, for any $j \in \{1, \ldots, d\}$, it holds that $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)(j) \leq T(\mathbf{y}_1, \ldots, \mathbf{y}_n)(j)$, for any $(\mathbf{x}_1, \ldots, \mathbf{x}_n), (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in (\mathbb{R}^d)^n$ such that $\mathbf{x}_i(j) \leq \mathbf{y}_i(j)$ for any $i \in \{1, \ldots, n\}$.

It is to be noted that, as discussed in Gagolewski (2017), componentwisely increasing estimators of multivariate location necessarily are componentwise extensions of increasing estimators of univariate location.

Proposition 1. *(Gagolewski, 2017) An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is componentwisely increasing if and only if there exist d increasing estimators of univariate location $T_1, \ldots, T_d : \mathbb{R}^n \to \mathbb{R}$ such that $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)(j) = T_j(\mathbf{x}_1(j), \ldots, \mathbf{x}_n(j))$, for any $j \in \{1, \ldots, d\}$.*

The most typical estimator of multivariate location that is componentwisely increasing is the centroid, which is also referred to as the vector of means and is obtained by computing the (arithmetic) mean in each of the components. The family of weighted centroids generalizes the notion of centroid by incorporating a weight for each of the points, thus computing the same weighted arithmetic mean in each of the components (Gagolewski et al., 2020). Unfortunately, the weighted centroids (and, thus, the centroid) are not robust in the presence of outliers. A highly-robust estimator of multivariate location that is componentwisely increasing is the componentwise median, which is obtained by computing the median in each of the components.

2.2. Estimation of multivariate location for coordinate-independent variables

The choice of coordinates for some specific types of multivariate data is arbitrary, and the chosen coordinate system could have been naturally subjected to an orthogonal transformation. An example of such type of multivariate data is that of spatial data, where the location of objects is usually expressed in terms of a somehow arbitrary coordinate system. A natural property when dealing with spatial data is that of orthogonal equivariance, which requires that the estimation of multivariate location should behave well even if the data points are subjected to any orthogonal transformation (e.g. rotation or reflection). We recall that an orthogonal transformation is characterized by an orthogonal matrix, i.e., a matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$ such that $\mathbf{O}^T = \mathbf{O}^{-1}$.

Definition 2. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called *orthogonal equivariant* if, for any $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^d)^n$ and any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$, it holds that $T(\mathbf{O}\mathbf{x}_1, \ldots, \mathbf{O}\mathbf{x}_n) = \mathbf{O}\,T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

A typical estimator of multivariate location that is orthogonal equivariant is the spatial median (Vardi and Zhang, 2000; Weber, 1909; Weiszfeld, 1937), which is obtained by computing the point that minimizes the sum of Euclidean distances to all the points. Another example of estimator of multivariate location that is orthogonal equivariant is the orthomedian (Grübel, 1996), which is obtained by orthogonalizing the componentwise median. Interestingly, this orthogonalization process turns the componentwise median orthogonal equivariant, at the cost of losing the componentwise increasingness.

2.3. A tension between orthogonal equivariance and componentwise increasingness in the estimation of multivariate location

Probably due to the somehow standard assumption of the points being drawn from an elliptical distribution, the property of orthogonal equivariance is considered a "must" in the literature on estimators of multivariate location and, quite surprisingly, the property of componentwise increasingness is oftentimes forgotten. A potential explanation could be the inherent tension between these two properties: the unique estimators of multivariate location satisfying both properties are the weighted centroids.

Theorem 1. *(Gagolewski et al., 2020) An (idempotent) estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is orthogonal equivariant and componentwisely increasing if and only if it is a weighted centroid.*

This implies that there exists no robust, orthogonal equivariant and componentwisely increasing estimator of multivariate location, and probably explains why most works on robust estimation of multivariate location only focus on the property of orthogonal equivariance. Actually, the literature on robust multivariate location usually considers orthogonal equivariance not to be sufficient and aims at an even stronger property: affine equivariance (see, e.g., Lopuhaä and Rousseeuw, 1991).

Definition 3. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called *affine equivariant* if, for any $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^d)^n$, any invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and any $\mathbf{t} \in \mathbb{R}^d$, it holds that $T(\mathbf{A}\mathbf{x}_1 + \mathbf{t}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{t}) = \mathbf{A}\,T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{t}$.

Note that the spatial median and the orthomedian are examples of estimators of multivariate location that are orthogonal equivariant but are not affine equivariant. An example of affinitization process based on the transformation into a data-driven coordinate system and a subsequent retransformation into the original coordinate system is due to Chakraborty and Chaudhuri (1996). Although originally proposed as a process for turning the componentwise median affine equivariant (at the cost of losing the componentwise increasingness), this procedure has been studied for making other estimators of multivariate location affine equivariant (see, e.g., Chakraborty et al. (1998) for a study on the affinitization of the spatial median).

3. Some examples of estimation of multivariate location for heterogeneous data

3.1. Spatio-temporal data

Unfortunately, real-life data might be quite complex and be a combination of both aforementioned types of multivariate data. For instance, consider the spatio-temporal data arising in Kong et al. (2016) for earthquake early warning. The smartphones of the users of the app *MyShake* are used as seismic sensors and, in case of an earthquake, the location and time of each trigger detection is recorded. In particular, each of the data points is of the form $\mathbf{x} \in \mathbb{R}^3$, where $\mathbf{x}(1)$ and $\mathbf{x}(2)$ represent the spatial location[1] of the triggered sensor (horizontal and vertical position, respectively, both measured in meters) and $\mathbf{x}(3)$ represents its trigger detection time (in seconds).

[1] Interestingly, it was not until *The International Prime Meridian Conference* was held in Washington in October 1884 that the geographic coordinate system based on latitude and longitude was considered a standard, and still nowadays one can find many alternative geographic coordinate systems. Here, as this coordinate system is angle-based and the estimation of location for directional data (Fisher, 1985; Mardia, 1975) lies out of the scope of this paper, we will transform the latitude and longitude to another coordinate system based on a projection to a map, for instance, the Universal Transverse Mercator (UTM) coordinate system. Note that this simplification is not very problematic for locations distributed over a small area of the Earth.

In order to estimate the multivariate location, quoting from (Kong et al., 2016), "the origin time is set to the earliest trigger time, and the centroid of [all the trigger detections] is used as the epicenter". We thus distinguish two components in the data. On the one hand, we have the spatial component, which is bivariate in nature and for which an orthogonal equivariant (even affine equivariant) estimator of multivariate location is considered: the centroid. On the other hand, we have the temporal component, which is univariate in nature and for which an increasing estimator of univariate location is considered: the minimum. Thus, the estimation of multivariate location is given by the centroid (in the first two components) and the minimum (in the third component). Formally, the considered estimator of multivariate location $T_{\mathrm{ST}} : (\mathbb{R}^3)^n \to \mathbb{R}^3$ is defined by

$$T_{\mathrm{ST}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(1),\ \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(2),\ \min_{i=1}^{n} \mathbf{x}_i(3) \right).$$

Note that this estimator is not robust. If one aims at substituting this estimator by a classic robust estimator of multivariate location, neither orthogonal equivariant estimators, nor componentwisely increasing estimators should be used. On the one hand, if the former ones are used, a miscalibration in the geolocation of the sensor could potentially affect the estimation of location for the temporal component . On the other hand, if the latter ones are used, the choice of coordinate system for the geolocation of the sensor could potentially affect the result of the estimation of location — something unacceptable due to the coordinate-free nature of the geolocation .

3.2. Spatio-bathymetric data

In Bost et al. (2009), the complete at-sea movements of twelve macaroni penguins (*Eudyptes chrysolophus*) were monitored from April to October 2006. In particular, each of the data points is of the form $\mathbf{x} \in \mathbb{R}^3$, where $\mathbf{x}(1)$ and $\mathbf{x}(2)$ represent the spatial location (in meters) and $\mathbf{x}(3)$ represents depth (in meters) at which a macaroni penguin was located. Similarly to the case of the spatio-temporal data, for identifying the foraging areas, neither orthogonal equivariant robust estimators of multivariate location, nor componentwisely increasing robust estimators of multivariate location can be used. Again, the estimator of multivariate location should be orthogonal equivariant with respect to the spatial component but increasing with respect to the depth.

4. Estimation of multivariate location for heterogeneous data

4.1. Appropriate estimation of multivariate location for heterogeneous data

It thus becomes clear that, in some cases, we should acknowledge that the variables composing the data might be different in nature, and that neither orthogonal equivariance nor componentwise increasingness should be required. For the purpose of multivariate location estimation, we may aim at identifying the (sets of) variables for which it is natural to require orthogonal equivariance and the variables for which it is natural to require increasingness. Therefore, we aim at partitioning $\{1, \ldots, d\}$ into m disjoint subsets $D_1, \ldots, D_m$, where each D_ℓ represents the indices of a set of variables that are independent in nature of the choice of coordinate system and where said coordinate system could be subjected to any orthogonal transformation. For the sake of simplicity, we assume that the indices in each D_ℓ are consecutive and refer to the set $\mathcal{D} = \{D_1, \ldots, D_m\}$ as a partition of the variables. Both componentwise increasingness and orthogonal equivariance might be defined just for the restriction to one such set of variables.

Definition 4. Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables and $\ell \in \{1, \ldots, m\}$. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called componentwise increasing on D_ℓ if, for any $j \in D_\ell$, it holds that $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)(j) \leq T(\mathbf{y}_1, \ldots, \mathbf{y}_n)(j)$, for any $(\mathbf{x}_1, \ldots, \mathbf{x}_n), (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in (\mathbb{R}^d)^n$ such that $\mathbf{x}_i(j) \leq \mathbf{y}_i(j)$ for any $i \in \{1, \ldots, n\}$.

Definition 5. Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables and $\ell \in \{1, \ldots, m\}$. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called orthogonal equivariant on D_ℓ if, for any $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^d)^n$ and any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$ such that $\mathbf{O}_{jj} = 1$ for any $j \notin D_\ell$, it holds that $T(\mathbf{O}\mathbf{x}_1, \ldots, \mathbf{O}\mathbf{x}_n) = \mathbf{O}\,T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

It then seems intuitive to decide on the basis of the data whether orthogonal equivariance or componentwise increasingness should be pursued for each group of variables. Since being componentwisely increasing on D_ℓ is equivalent to being increasing on each of the variables associated with the indices in D_ℓ, it is assumed that all D_ℓ for which componentwise increasingness is required are such that $|D_\ell| = 1$.

Definition 6. Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables. An estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is appropriate for $\mathcal{D}$ if there exist m estimators of location $\{T_\ell : (\mathbb{R}^{|D_\ell|})^n \to \mathbb{R}^{|D_\ell|}\}_{\ell=1}^m$ such that

$$T(\mathbf{x}_1, \ldots, \mathbf{x}_n) = (T_1(\mathbf{x}_1(D_1), \ldots, \mathbf{x}_n(D_1)), \ldots, T_m(\mathbf{x}_1(D_m), \ldots, \mathbf{x}_n(D_m))),$$

where, for any $\ell \in \{1, \ldots, m\}$, T is orthogonal equivariant on D_ℓ (and, thus, T_ℓ is orthogonal equivariant) if $|D_\ell| > 1$ and componentwisely increasing on D_ℓ (and, thus, T_ℓ is componentwisely increasing) if $|D_\ell| = 1$.

On the one hand, in case we are dealing with coordinate-independent data, m should equal 1, thus T being orthogonal equivariant. On the other hand, in case we are dealing with a list of univariate variables, m should equal d, thus T being componentwisely increasing.

Any estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ that is appropriate for a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables, can easily be proven to be componentwisely increasing if and only if all T_ℓ are componentwisely increasing. Unfortunately, there is no guarantee of orthogonal equivariance even in case all T_ℓ are orthogonal equivariant. Interestingly, T is translation equivariant if and only if all T_ℓ are translation equivariant. It is recalled that an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is called *translation equivariant* if, for any $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{a} \in \mathbb{R}^d$, it holds that $T(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a}) = T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}$.

Proposition 2. *Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables. Any estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ that is appropriate for $\mathcal{D}$:*

(i) is componentwisely increasing if and only if all T_ℓ are componentwisely increasing;
(ii) does not need to be orthogonal equivariant, even in case all T_ℓ are orthogonal equivariant;
(iii) is translation equivariant if and only if all T_ℓ are translation equivariant.

4.2. Robust estimation of multivariate location for heterogeneous data

A popular measure of robustness is the (finite sample) breakdown point (Donoho and Huber, 1983; Gather and Davies, 2005; Hampel, 1971), which measures the smallest proportion of contaminated observations that may cause an estimator of location to take arbitrarily large values.

Definition 7. Consider $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^d)^n$. The set $\mathcal{Z}_q(\mathbf{X})$ of q-corrupted (by replacement) lists given $\mathbf{X}$ is defined as the set of all lists $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \in (\mathbb{R}^d)^n$ such that $|\{i \in \{1, \ldots, n\} \mid \mathbf{x}_i \neq \mathbf{z}_i\}| \leq q$. The maximum bias caused by q-corruption at $\mathbf{X}$ for an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is defined as $b(T, \mathbf{X}, q) = \sup_{\mathbf{Z} \in \mathcal{Z}_q(\mathbf{X})} \|T(\mathbf{X}) - T(\mathbf{Z})\|$. The finite-sample breakdown point of an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ at $\mathbf{X}$ is defined as $\epsilon(T, \mathbf{X}) = \inf_{q \in \{1,\ldots,n\}, b(T,\mathbf{X},q) = +\infty} \frac{q}{n}$. The finite-sample breakdown point of an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is defined as $\epsilon(T) = \inf_{\mathbf{X} \in (\mathbb{R}^d)^n} \epsilon(T, \mathbf{X})$

The maximum breakdown point that can be attained by a translation equivariant estimator of location is 0.5. This maximum value is attained by three popular estimators of multivariate location: the componentwise median, the spatial median and the orthomedian. Interestingly, the componentwise median is componentwisely increasing, whereas the spatial median and the orthomedian are orthogonal equivariant.

As pointed out in previous sections, there exists no robust estimator of multivariate location that is, at the same time, orthogonal equivariant and componentwise increasing. This result still holds if componentwise increasingness is substituted by the weaker property of componentwise increasingness on at least one variable. Obviously, the result also holds if orthogonal equivariance is substituted by affine equivariance.

Theorem 2. *Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables. The breakdown point ϵ of an (idempotent and) orthogonal equivariant estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ that is componentwise increasing on D_ℓ for at least one $\ell \in \{1, \ldots, m\}$ is $\frac{1}{n}$.*

Therefore, if there exists at least one component for which componentwise increasingness is pursued, then orthogonal and affine equivariant should be abandoned in the context of robust estimation of multivariate location. Obviously, this does not mean that orthogonal and affine equivariance should be completely abandoned since these properties can still be required for the variables for which componentwise increasingness is not pursued.

In order to propose a robust estimator of multivariate location that is appropriate for a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables, it suffices that all $T_\ell : (\mathbb{R}^{|D_\ell|})^n \to \mathbb{R}^{|D_\ell|}$ are robust. In particular, it can be assured that the optimal breakdown point of 0.5 is attained if (and only if) all T_ℓ attain the optimal breakdown point of 0.5.

Proposition 3. *Consider a partition $\mathcal{D} = \{D_1, \ldots, D_m\}$ of the d variables. The breakdown point ϵ of an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ that is appropriate for $\mathcal{D}$ (with associated estimators of location $\{T_\ell : (\mathbb{R}^{|D_\ell|})^n \to \mathbb{R}^{|D_\ell|}\}_{\ell=1}^m$ with respective breakdown points $\{\epsilon_\ell\}_{\ell=1}^m$) satisfies that $\epsilon = \min_{\ell \in \{1,\ldots,m\}} \epsilon_\ell$.*

A possible option would then be to consider the classical median for those components such that $|D_\ell| = 1$ and the orthomedian for those components such that $|D_\ell| > 1$. The use of the orthomedian is encouraged over the spatial median because – although it is not componentwisely increasing – it satisfies the weaker property of (SC)-monotonicity described in Pérez-Fernández et al. (2019). In case affine equivariance is pursued, the orthomedian should be substituted by a robust affine equivariant estimator of multivariate location such as the transformation–retransformation median (Chakraborty and Chaudhuri, 1996), the Stahel–Donoho estimator (Stahel, 1981; Donoho, 1982) or the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984).

Coming back to the examples of the spatio-temporal data in Kong et al. (2016) and the spatio-bathymetric data in Bost et al. (2009), a natural robust estimator of multivariate location is given by the orthomedian for the spatial component and the median for the temporal/depth component. It is to be noted that, since the temporal component in the case of the spatio-temporal data in Kong et al. (2016) prioritizes early trigger detections, the use of an order statistic smaller than the median might be recommendable, even despite the loss in robustness.

5. Conclusions

The estimation of (univariate) location typically builds upon the property of increasingness. Surprisingly, the estimation of multivariate location has historically neglected this property and has built around the properties of orthogonal equivariance and the even stronger affine equivariance. This is probably explained by the incompatibility of these latter properties with componentwise increasingness when combined with the robustness in the presence of outliers. In this work, it is discussed that, in this era of big data in which countless variables from very different nature are often monitored, both (componentwise) increasingness and orthogonal equivariance might lead to a very undesirable behavior. Thus, the choice between (componentwise) increasingness and orthogonal equivariance must depend on the type of data one is dealing with. In the presence of heterogeneous data, both properties should be forfeited and only required for some of the components.

We end by concluding that, as principal component analysis and other techniques for feature selection and extraction require to subject the data to a carefully chosen affine transformation, if robustness is prioritized over componentwise increasingness, then the latter property must be abandoned. Other weaker monotonicity-related properties that are coordinate-free in nature, such as (SC)-monotonicity (as described in Pérez-Fernández et al. (2019)), should be then pursued.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The author is deeply indebted to Marek Gagolewski for illuminating suggestions on an initial version of the manuscript.

Funding

This research has been partially supported by the Spanish MINECO (TIN2017-87600-P).

Appendix. Proofs

Proof of Proposition 2. (i) Follows straightforwardly from Proposition 1. More specifically, $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is componentwisely increasing if and only if there exist d increasing estimators of univariate location $T_{0,1}, \ldots, T_{0,d} : \mathbb{R}^n \to \mathbb{R}$ such that $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)(j) = T_{0,j}(\mathbf{x}_1(j), \ldots, \mathbf{x}_n(j))$, for any $j \in \{1, \ldots, d\}$. Equivalently, $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is componentwisely increasing on all D_ℓ if and only if for all ℓ there exist d_ℓ increasing estimators of univariate location $T_{\ell,1}, \ldots, T_{\ell,d_\ell} : \mathbb{R}^n \to \mathbb{R}$ such that $T_\ell(\mathbf{x}_1, \ldots, \mathbf{x}_n)(j) = T_{\ell,j}(\mathbf{x}_1(j), \ldots, \mathbf{x}_n(j))$, for any $j \in \{1, \ldots, d_\ell\}$. The result follows from identifying each $T_{0,j}$ with its corresponding $T_{\ell,j'}$.

(ii) Due to the fact that the univariate median is orthogonal equivariant in one dimension, it suffices to consider the componentwise median (which is not orthogonal equivariant).

(iii) Consider any $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{a} \in \mathbb{R}^d$. If all T_ℓ are translation equivariant, then it holds that

$$\begin{aligned}
&T(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a}) \\
&= (T_1(\mathbf{x}_1(D_1) + \mathbf{a}(D_1), \ldots, \mathbf{x}_n(D_1) + \mathbf{a}(D_1)), \ldots, T_m(\mathbf{x}_1(D_m) + \mathbf{a}(D_m), \ldots, \mathbf{x}_n(D_m) + \mathbf{a}(D_m))) \\
&= (T_1(\mathbf{x}_1(D_1), \ldots, \mathbf{x}_n(D_1)) + \mathbf{a}(D_1), \ldots, T_m(\mathbf{x}_1(D_m), \ldots, \mathbf{x}_n(D_m)) + \mathbf{a}(D_m)) \\
&= \left(T_1(\mathbf{x}_1(D_1), \ldots, \mathbf{x}_n(D_1)), \ldots, T_m(\mathbf{x}_1(D_m), \ldots, \mathbf{x}_n(D_m))\right) + \mathbf{a} \\
&= T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}.
\end{aligned}$$

If T is translation equivariant, then for any $\ell \in \{1, \ldots, m\}$ it holds that

$$\begin{aligned}
&T_\ell(\mathbf{x}_1(D_\ell) + \mathbf{a}(D_\ell), \ldots, \mathbf{x}_n(D_\ell) + \mathbf{a}(D_\ell)) \\
&= T(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a})(D_\ell) \\
&= T(\mathbf{x}_1, \ldots, \mathbf{x}_n)(D_\ell) + \mathbf{a}(D_\ell) \\
&= T_\ell(\mathbf{x}_1(D_\ell), \ldots, \mathbf{x}_n(D_\ell)) + \mathbf{a}(D_\ell).
\end{aligned}$$

■

Proof of Theorem 2. If an estimator of multivariate location $T : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is componentwise increasing on D_ℓ for at least one $\ell \in \{1, \ldots, m\}$, then it is $\mathbf{e}_j$-SP-monotone for any $j \in D_\ell$ in the sense of Definition 4 in Pérez-Fernández et al. (2019). Proposition 14 and Theorem 28 in Pérez-Fernández et al. (2019) imply that T is a weighted centroid. The result then follows from the fact that the breakdown point of a weighted centroid is $\frac{1}{n}$. ■

Proof of Proposition 3. On the one hand, suppose that $\epsilon < \min_{\ell \in \{1,\ldots,m\}} \epsilon_\ell$. This implies that $\sup_{\mathbf{Z} \in \mathcal{Z}_{n\epsilon}(\mathbf{X})} \|T(\mathbf{X}) - T(\mathbf{Z})\| = +\infty$, and, for any $\ell \in \{1, \ldots, m\}$, $\sup_{\mathbf{Z} \in \mathcal{Z}_{n\epsilon}(\mathbf{X})} \|T_\ell(\mathbf{X}(D_\ell)) - T_\ell(\mathbf{Z}(D_\ell))\| < +\infty$. The contradiction follows from the fact that

$$\begin{aligned}\|T(\mathbf{X}) - T(\mathbf{Z})\|^2 &= \sum_{j=1}^{d} \big(T(\mathbf{X})(j) - T(\mathbf{Z})(j)\big)^2 = \sum_{j=1}^{d} \big(T_{\ell_j}(\mathbf{X}(D_{\ell_j}))(j) - T_{\ell_j}(\mathbf{Z}(D_{\ell_j}))(j)\big)^2 \\ &\leq \sum_{j=1}^{d} \|T_{\ell_j}(\mathbf{X}(D_{\ell_j})) - T_{\ell_j}(\mathbf{Z}(D_{\ell_j}))\|^2 < +\infty\,,\end{aligned}$$

where ℓ_j represents the subindex ℓ associated with the component j.

On the other hand, suppose that $\epsilon > \min_{\ell \in \{1,\ldots,m\}} \epsilon_\ell$. Consider ℓ_* such that $\epsilon_{\ell *} = \min_{\ell \in \{1,\ldots,m\}} \epsilon_\ell$. It holds that

$$\sup_{\mathbf{Z} \in \mathcal{Z}_{n\epsilon_{\ell *}}(\mathbf{X})} \|T(\mathbf{X}) - T(\mathbf{Z})\| < +\infty \text{ and } \sup_{\mathbf{Z} \in \mathcal{Z}_{n\epsilon_{\ell *}}(\mathbf{X})} \|T_{\ell_*}(\mathbf{X}(D_{\ell_*})) - T_{\ell_*}(\mathbf{Z}(D_{\ell_*}))\| = +\infty.$$

The contradiction follows from the fact that

$$\|T(\mathbf{X}) - T(\mathbf{Z})\|^2 = \sum_{\ell=1}^{m} \|T(\mathbf{X})(D_\ell) - T(\mathbf{Z})(D_\ell)\|^2 = \sum_{\ell=1}^{m} \|T_\ell(\mathbf{X}(D_\ell)) - T_\ell(\mathbf{Z}(D_\ell))\|^2\,. \blacksquare$$

References

Bost, C.A., Thiebot, J.B., Pinaud, D., Cherel, Y., Trathan, P.N., 2009. Where do penguins go during the inter-breeding period? Using geolocation to track the winter dispersion of the macaroni Penguin. Biol. Lett. 5, 473–476.

Chakraborty, B., Chaudhuri, P., 1996. On a transformation and re-transformation technique for constructing an affine equivariant multivariate median. Proc. Amer. Math. Soc. 124 (8), 2539–2547.

Chakraborty, B., Chaudhuri, P., Oja, H., 1998. Operating transformation retransformation on spatial median and angle test. Statist. Sinica 8, 767–784.

Croux, C., Dehon, C., 2014. Robust estimation of location and scale. In: Wiley StatsRef: Statistics Reference Online. Wiley, pp. 1–9.

Cuesta-Albertos, J.A., Gordaliza, A., Matrán, C., 1997. Trimmed k-means: An attempt to robustify quantizers. Ann. Statist. 25, 553–576.

Donoho, D.L., 1982. Breakdown properties of multivariate location estimators (Ph.D. thesis). Harvard University, Boston.

Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: A Festschrift for Erich L. Lehmann. Wadsworth, Belmont, USA, pp. 157–184.

Eddy, W.F., 1982. Convex hull peeling. In: Proceedings of the COMPSTAT Symposium. Toulouse, pp. 42–47.

Fisher, N.I., 1985. Spherical medians. J. R. Stat. Soc. Ser. B Stat. Methodol. 47, 342–348.

Gagolewski, M., 2015. Data Fusion: Theory, Methods, and Applications. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, p. 290.

Gagolewski, M., 2017. Penalty-based aggregation of multidimensional data. Fuzzy Sets and Systems 325, 4–20.

Gagolewski, M., Bartoszuk, M., Cena, A., 2016. Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. Inform. Sci. 363, 8–23.

Gagolewski, M., Pérez-Fernández, R., De Baets, B., 2020. An inherent difficulty in the aggregation of multidimensional data. IEEE Trans. Fuzzy Syst. 28 (3), 602–606.

Gallegos, M.T., Ritter, G., 2005. A robust method for cluster analysis. Ann. Statist. 33, 347–380.

Gather, U., Davies, P.L., 2005. Breakdown and groups. Ann. Statist. 33 (3), 977–1035.

Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E., 2011. Aggregation functions: Means. Inform. Sci. 181, 1–22.

Grübel, R., 1996. Orthogonalization of multivariate location estimators: The orthomedian. Ann. Statist. 24 (4), 1457–1473.

Hampel, F.R., 1971. A general qualitative definition of robustness. Ann. Math. Stat. 42, 1887–1896.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: A new approach to robust Principal Component Analysis. Technometrics 47 (1), 64–79.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Kong, Q., Allen, R.M., Schreier, L., Y.-W., K., 2016. MyShake: A smartphone seismic network for earthquake early warnking and beyond. Sci. Adv. 2, 1–9.

Liu, R.Y., 1990. On a notion of data depth based on random simplices. Ann. Statist. 18 (1), 405–414.

Lopuhaä, H.P., 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance. Ann. Statist. 17 (4), 1662–1683.

Lopuhaä, H.P., Rousseeuw, P.J., 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann. Statist. 19 (1), 229–248.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press, Berkeley, pp. 281–297.

Mardia, K.V., 1975. Statistics of directional data (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 37, 349–393.

Maronna, R.A., 1976. Robust M-estimators of multivariate location and scatter. Ann. Statist. 4, 51–67.

Oja, H., 1983. Descriptive statistics for multivariate distributions. Statist. Probab. Lett. 1, 327–332.

Owen, S.H., Daskin, M.S., 1998. Strategic facility location: A review. European J. Oper. Res. 111, 423–447.

Pérez-Fernández, R., De Baets, B., Gagolewski, M., 2019. A taxonomy of monotonicity properties for the aggregation of multidimensional data. Inf. Fusion 52, 322–334.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79 (388), 871–880.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications. Reidel Publishing Company, Dordrecht, pp. 283–297.
Rousseeuw, P., Hubert, M., 2013. High-breakdown estimators of multivariate location and scatter. In: Robustness and Complex Data Structures. Springer, pp. 49–66.
Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.
Rousseew, P.J., Hubert, M., 2011. Robust statistics for outlier detection. Wiley Interdisc. Rev. - Data Mining Knowl. Discov. 1, 73–79.
Small, C.G., 1990. A survey of multidimensional medians. Internat. Statist. Rev. 58 (3), 263–277.
Stahel, W.A., 1981. Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen (Ph.D. thesis). ETH, Zurich.
Sylvester, J.J., 1857. A question in the geometry of situation. Q. J. Math. 1, 79.
Tukey, J.W., 1975. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians. Vancouver, pp. 523–531.
Vardi, Y., Zhang, C.-H., 2000. The multivariate L1-median and associated data depth. Proc. Natl. Acad. Sci. 97 (4), 1423–1426.
Weber, A., 1909. Ueber den Standort der Industrien. Mohr Siebeck Verlag, Tübingen.
Weiszfeld, E., 1937. Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Math. J. First Ser. 43, 355–386.
Wilcox, R.R., Keselman, H.J., 2003. Modern robust data analysis methods: Measures of central tendency. Psychol. Methods 8, 254–274.