Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 14

5-1-2017

Multivariate Rank Outlyingness and Correlation Effects

Olusola Samuel Makinde
Department of Statistics, Federal University of Technology, P.M.B. 704, Akure, Nigeria, osmakinde@futa.edu.ng

Recommended Citation
Makinde, O. S. (2017). Multivariate rank outlyingness and correlation effects. Journal of Modern Applied Statistical Methods, 16(1), 246-260. doi: 10.22237/jmasm/1493597580

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 246-260.
doi: 10.22237/jmasm/1493597580

ISSN 1538 − 9472

Multivariate Rank Outlyingness and Correlation Effects

Olusola Samuel Makinde
Federal University of Technology
Akure, Nigeria

The effect of correlation on multivariate rank outlyingness, a result of deviation of multivariate rank functions from property of spherical symmetry, is examined. Possible affine invariant versions of this multivariate rank are surveyed, and outlyingness of affine invariant and non-invariant spatial rank functions under general affine transformation are compared.

Keywords: rank function, outlyingness function, symmetry, correlation

Introduction

Ordering of data and the search for the units lying far from the centroid is closely related to searching for outliers in the data cloud. In a univariate setting, this ordering is a linear ranking from smallest to largest. Given sample points $X_1, X_2, \ldots, X_n$, we can order them by their rank values. Ordering of univariate objects based on rank does not depend heavily on the underlying distribution of the data, nor involve estimation of parameters of probability distributions. Similarly in a multivariate setting, we can order multivariate sample points $X_1, X_2, \ldots, X_n$ by their rank function.

An appealing way of working with probability distributions in $\mathbb{R}^d$, especially in nonparametric inference, is through "descriptive measures" that characterize features of particular interest (Serfling, 2004, p. 260). One attractive approach is to base the measures on outlyingness of multivariate rank. In the last couple of decades, notions of multivariate signs and ranks have become a useful tool in analyzing multivariate data, as it does not depend heavily on distributional assumptions, and characterizes the central and extreme observations quite effectively (Makinde & Chakraborty, 2015). Use of multivariate rank for ordering

Olusola Samuel Makinde is a Lecturer in the Department of Statistics. Email at osmakinde@futa.edu.ng.

of data preserves the direction of the data. Möttönen & Oja (1995), Möttönen, Oja & Tienari (1997) used the notion of spatial ranks to construct multivariate tests of location.

A related notion to multivariate ranks is the data depth. Data depth measures depth or centrality of a *d*-dimensional observation with respect to a multivariate data cloud or underlying multivariate distribution. Depth functions in literature include Mahalanobis depth, half-space depth, simplicial depth, likelihood depth, and projection depth, among others. Liu, Parelius & Singh (1999) proposed various ideas on analyzing multivariate data using data depths. We refer readers to Liu, Parelius & Singh (1999) for detailed discussion on depth functions. Statistical approaches based on most of these depth functions suffer computational complexities of the depth functions.

The spatial rank and its outlyingness can be applied in classification and clustering (Makinde, 2015). It has been applied in construction of geometric quantile (Chaudhuri, 1996; Serfling, 2004). It is well known that multivariate rank is not invariant under arbitrary affine transformations, so it may be affected by deviation of population distribution from spherical symmetry. Effect of this deviation on spatial rank outlyingness will be investigated. Based on this, we shall introduce a way of constructing affine invariant multivariate rank outlyingness.

Spatial Rank

Signs and ranks are commonly used in statistical methodology to develop methods or procedures that are independent of distribution assumptions. Use of rank for computing statistical quantities gives robust estimators (e.g. estimator for location) as they are not affected by the presence of outlying values in the data. For the univariate data, *sign* of $x \in \mathbb{R}$ can be defined as

$$sign(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

or equivalently,

$$sign(x) = \begin{cases} \dfrac{x}{|x|}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

Univariate centred rank of x with respect to data points $X_1, X_2, \ldots, X_n$ from distribution F can be defined as

$$rank(x) = \frac{1}{n}\sum_{i=1}^{n} sign(x - X_i).$$

Following are some of the basic properties of $rank(x)$,

1. $|\, rank(x)\,| \leq 1$.
2. $|\, rank(x)\,| = 0$ implies x is the median and $|\, rank(x)\,| = 1$ implies x is an extreme point.
3. $E(|\, rank(x)\,|) = 2F(x) - 1$

These properties suggest that $rank(x)$ is not only a useful descriptive statistics, it also characterizes the distribution. Now, we want to define sign and rank functions in a multivariate set up following Chakraborty (2001). Suppose $\boldsymbol{x} \in \mathbb{R}^d$, then the l_p sign of $\boldsymbol{x}$ is

$$sign_p(\boldsymbol{x}) = \begin{cases} \dfrac{\partial}{\partial \boldsymbol{x}}\|\boldsymbol{x}\|_p = \dfrac{\upsilon(\boldsymbol{x})}{\|\boldsymbol{x}\|_p^{\,p-1}}, & \boldsymbol{x} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{x} = \boldsymbol{0} \end{cases}$$

where

$$\|\boldsymbol{x}\|_p = \left(x_1^p + x_2^p + \cdots + x_d^p\right)^{\frac{1}{p}} \text{ and}$$
$$\upsilon(\boldsymbol{x}) = \left(sign(x_1)|x_1|^{p-1}, \ldots, sign(x_d)|x_d|^{p-1}\right).$$

The l_p rank of $\boldsymbol{x} \in \mathbb{R}^d$ with respect to data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is defined as

$$rank_p(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} sign_p(\boldsymbol{x} - \boldsymbol{X}_i).$$

when p = 1, $sign(\boldsymbol{x}) = \left(sign(x_1), sign(x_2), \ldots, sign(x_d)\right)^T$, the vector of co-ordinatewise signs and for p = 2,

$$sign_2(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$$

where $\|.\|_2$ is the Euclidean norm defined as $\|\boldsymbol{y}\|_2 = \left(y_1^2 + y_2^2 + \cdots + y_d^2\right)^{\frac{1}{2}}$. $sign_2(\boldsymbol{x})$ is called the spatial sign vector.

Suppose $\boldsymbol{X}$ is a d-dimensional random vector having a distribution F, which is assumed to be absolutely continuous with respect to the Lebesgue measure $\mathbb{R}^d$. The spatial rank function (Möttönen & Oja, 1995) of any point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to F is defined as

$$rank_F(\boldsymbol{x}) = E_F\left(\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|}\right). \tag{1}$$

Here $\|.\|$ is the usual Euclidean norm. It follows immediately from the definition that $rank_F(\boldsymbol{x}) = \boldsymbol{0}$ implies that $\boldsymbol{x}$ is the spatial median of the multivariate distribution F. Koltchinskii (1997) established that this spatial rank function is a one-to-one function of the distribution function F and hence it characterizes the distribution. Moreover the direction of the vector $rank_F(\boldsymbol{x})$ suggests the direction in which $\boldsymbol{x}$ is extreme compared to the distribution. Using this idea, Serfling (2004) introduced $\|rank_F(\boldsymbol{x})\|$ as a measure of outlyingness and defined several descriptive measures. Smaller values of $\|rank_F(\boldsymbol{x})\|$ implies that $\boldsymbol{x}$ is more central to the distribution and larger values of $\|rank_F(\boldsymbol{x})\|$ indicates that $\boldsymbol{x}$ is more extreme. If $\|rank_F(\boldsymbol{x})\| = 0$, then $\boldsymbol{x}$ is the spatial median.

Spatial rank helps determine the geometric position of points in $\mathbb{R}^d$ with respect to the data cloud, and hence can be viewed as a descriptive statistic (Guha, 2012). Suppose F is spherically symmetric and characterized by location parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, $\|rank_F(\boldsymbol{x})\|$ increases as $\|\boldsymbol{x} - \boldsymbol{\theta}\|$ increases. This result is stated formally in Theorem 1 below:

Theorem 1. If $\boldsymbol{x}$ has spherically symmetric distribution F with $\boldsymbol{\theta}$ as the centre of symmetry, then for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$rank_F(\boldsymbol{x}) = q(\|\boldsymbol{x}-\theta\|)\frac{\boldsymbol{x}-\theta}{\|\boldsymbol{x}-\theta\|}$$

for some increasing, non-negative function q.

This is proved in Guha (2012). Following Theorem 1, smaller rank outlyingness indicates more central observation and larger rank outlyingness indicates extreme observation. The following results hold for rank outlyingness:

Fact: Let $\|rank_F(\boldsymbol{x})\|$ denote the measure of outlyingness of $rank_F(\boldsymbol{x})$. Then

1. $\|rank_F(\boldsymbol{x}+\boldsymbol{\theta})\| = \|rank_F(\boldsymbol{x})\|$ for a constant vector $\boldsymbol{\theta}$.
2. $\|rank_F(\boldsymbol{Ax})\| = \|rank_F(\boldsymbol{x})\|$ for an orthogonal matrix $\boldsymbol{A}$.

The first expression above implies that rank outlyingness is invariant under location shift or translation while the second indicates that rank outlyingness is invariant under orthogonal scale transformation. In practice, the rank functions $rank_F$ will hardly be known completely and we need to estimate them from the training sample. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample from a population having distribution F. We define the empirical rank function as

$$rank_{F_n}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x}-\boldsymbol{X}_i}{\|\boldsymbol{x}-\boldsymbol{X}_i\|}$$

Theorem 2. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent and identically distributed d-dimensional random vectors having distribution function F, which is absolutely continuous, then as $n \to \infty$,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\|rank_{F_n}(\boldsymbol{x})\| - \|rank_F(\boldsymbol{x})\|\right| \to 0.$$

The proof follows from Koltchinskii's (1997) work on the convergence of the empirical version of spatial rank to its population analogue.

Chaudhuri (1996) defined spatial quantiles as vectors in $\mathbb{R}^d$ that are indexed by a vector $\boldsymbol{u}$ in d-dimensional unit ball. Define an open ball

$B^d = \{\boldsymbol{u} \mid \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\| < 1\}$. For any $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{t} \in \mathbb{R}^d$, also define $\Phi(\boldsymbol{u}, \boldsymbol{t}) = \|\boldsymbol{t}\| + \langle \boldsymbol{u}, \boldsymbol{t} \rangle$, where $\langle .,. \rangle$ denotes the usual Euclidean inner product. Spatial quantile corresponding to $\boldsymbol{u}$ and based on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is defined as

$$\hat{Q}_n(\boldsymbol{u}) = \arg\min_{Q \in \mathbb{R}^d} \sum_{i=1}^{n} \Phi(\boldsymbol{u}, \boldsymbol{X}_i - Q).$$

It follows from Theorem 1.1.2 of Chaudhuri (1996) that

$$\sum_{i=1}^{n} \frac{Q_n(\boldsymbol{u}) - \boldsymbol{X}_i}{\|Q_n(\boldsymbol{u}) - \boldsymbol{X}_i\|} + n\boldsymbol{u} = 0$$

if $Q_n(\boldsymbol{u}) \neq \boldsymbol{X}_i$ for all $1 \leq i \leq n$. This implies

$$\boldsymbol{u} = \frac{1}{n} \sum_{i=1}^{n} \frac{Q_n(\boldsymbol{u}) - \boldsymbol{X}_i}{\|Q_n(\boldsymbol{u}) - \boldsymbol{X}_i\|}. \tag{2}$$

Serfling (2004) defined $rank_{F_n}(\boldsymbol{x})$ as the inverse function of the spatial quantile function, $\hat{Q}_n(\boldsymbol{u})$. Mathematically, we can write (2) as $\boldsymbol{u} = rank_{F_n}(\hat{Q}_n(\boldsymbol{u})) = rank_{F_n}(\boldsymbol{x})$ and so $\hat{Q}_n(\boldsymbol{u}) = \boldsymbol{x}$ implies $rank_{F_n}(\boldsymbol{x}) = \boldsymbol{u}$. It follows that $rank_{F_n}(\boldsymbol{x})$ is the inverse function of the multivariate geometric quantile function $Q_n(\boldsymbol{u})$ in the sense that $rank_{F_n}(\boldsymbol{x}) = \boldsymbol{u}$ implies that $Q_n(\boldsymbol{u}) = \boldsymbol{x}$ and vice-versa.

Effect of correlation on rank outlyingness

The distribution of a random variable $\boldsymbol{X}$ is said to be spherically symmetric about a parameter $\boldsymbol{\theta}$ if, for any orthogonal matrix $\boldsymbol{B}$,

$$\boldsymbol{X} - \boldsymbol{\theta} \overset{d}{=} \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\theta})$$

The density function of any spherically symmetric distribution of a random variable $\boldsymbol{X}$, if it exists, is of the form $f(\boldsymbol{x}) \propto g\left((\boldsymbol{x} - \boldsymbol{\theta})^T (\boldsymbol{x} - \boldsymbol{\theta})\right)$ for some nonnegative scalar function $g(.)$. Similarly, the distribution of a random vector $\boldsymbol{X}$ is said to be elliptically symmetric about $\boldsymbol{\theta}$ if there exists a $d \times d$ nonsingular matrix $\boldsymbol{A}$ such that $\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\theta})$ has a spherically symmetric distribution about $\mathbf{0}$. See Liu (1990), Liu & Singh (1993), Liu, Parelius & Singh (1999) and Serfling (2006) for further reading on multivariate symmetry. The deviation of rank outlyingness from the property of spherical symmetry implies that there exists correlation among variables in the population from which the sample is drawn.

Now, examine the effect of correlation among variables on rank outlyingness. Define $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{b}$ and $\boldsymbol{Y}_i = \boldsymbol{AX}_i + \boldsymbol{b}$ for nonsingular matrix $\boldsymbol{A}$ and constant vector $\boldsymbol{b}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{y}-\boldsymbol{Y}_i}{\left\|\boldsymbol{y}-\boldsymbol{Y}_i\right\|} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{X}_i\right)}{\left\|\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{X}_i\right)\right\|} \neq \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\boldsymbol{x}-\boldsymbol{X}_i\right)}{\left\|\boldsymbol{x}-\boldsymbol{X}_i\right\|}. \tag{3}$$

Table 1. Descriptive statistics of rank outlyingness of bivariate normal objects, bivariate Laplace objects and bivariate t objects with 3 degrees of freedom.

		$\delta = 0$				**$\delta = 2$**			
	Statistics	**$\rho = 0$**	**$\rho = 0.5$**	**$\rho = 0.75$**	**$\rho = 0.9$**	**$\rho = 0$**	**$\rho = 0.5$**	**$\rho = 0.75$**	**$\rho = 0.9$**
Bivariate normal distribution	*Minimum*	0.0378	0.0272	0.0156	0.0048	0.0799	0.0806	0.0809	0.0804
	25% quantile	0.4396	0.4136	0.4143	0.3739	0.4497	0.4430	0.4258	0.3870
	Median	0.6263	0.6405	0.6157	0.5794	0.6069	0.5900	0.5774	0.5503
	Mean	0.6021	0.5986	0.5873	0.5693	0.6053	0.6007	0.5900	0.5711
	75% quantile	0.7827	0.7852	0.7767	0.7665	0.7948	0.7724	0.7408	0.7524
	Maximum	0.9647	0.9649	0.9846	0.9941	0.9637	0.9678	0.9714	0.9900
Bivariate Laplace distribution	*Minimum*	0.0687	0.0673	0.0589	0.0607	0.0459	0.0588	0.0655	0.0732
	25% quantile	0.4346	0.4429	0.4114	0.3797	0.3688	0.3693	0.3749	0.3770
	Median	0.6133	0.6076	0.5717	0.5410	0.6244	0.6089	0.5749	0.5691
	Mean	0.5952	0.5894	0.5791	0.5649	0.5934	0.5868	0.5762	0.5618
	75% quantile	0.7611	0.7646	0.7821	0.7742	0.7986	0.7942	0.7853	0.7664
	Maximum	0.9693	0.9763	0.9800	0.9832	0.9819	0.9925	0.9955	0.9976
Bivariate *t* distribution with 3 d.f.	*Minimum*	0.1054	0.1129	0.1050	0.0871	0.0883	0.0865	0.0899	0.0698
	25% quantile	0.4076	0.4075	0.3900	0.3569	0.4260	0.4158	0.4098	0.3951
	Median	0.6188	0.5967	0.5705	0.5433	0.6054	0.6009	0.5817	0.5566
	Mean	0.5940	0.5849	0.5716	0.5546	0.5945	0.5885	0.5783	0.5630
	75% quantile	0.8034	0.7875	0.7682	0.7656	0.7734	0.7715	0.7890	0.7600
	Maximum	0.9833	0.9843	0.9927	0.9978	0.9948	0.9964	0.9986	0.9996

As illustration of the effect of correlation on rank outlyingness in (3), a small simulation study is presented. Consider a population to be bivariate elliptically symmetric with centre of symmetry $\boldsymbol{\mu} = (\delta \;\; 0)^T$ and scale matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Simulate a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, where sample size n is taken to be 100, and estimate the rank outlyingness function. For various values of ρ, Table 1 presents rank outlyingness for bivariate normally distributed sample, bivariate Laplace distributed sample and bivariate t distributed sample with 3 degrees of freedom.

The outlyingness function behaves anomalously for different values of $\rho \in [0,1)$ irrespective of class distribution. For each family of distribution, descriptive statistics are not in any specific order of ρ. The reason is that though the distribution of $\boldsymbol{X}_i$ is taking more ellipsoid form as ρ increases, the rank outlyingness is being computed with respect to sphere as spatial rank is non-invariant under affine transformation. To overcome the problem of affine non-invariance property of spatial rank, affine invariant versions of rank outlyingness are suggested next.

Affine Invariant Rank Function

Approach based on Cholesky decomposition of the covariance matrix

Spatial rank function can also be defined (Makinde & Chakraborty, 2015) as

$$rank_F^*(\boldsymbol{x}) = E_F\left(\frac{\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{X})}{\left\|\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{X})\right\|}\right)$$

where $\boldsymbol{V}$ is a $d \times d$ matrix such that $\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{c}\boldsymbol{\Sigma}$ for some constant $\boldsymbol{c}$. If the covariance of the distribution F exists, we can take $\boldsymbol{V}$ to be the Cholesky decomposition of the covariance matrix. For the empirical versions, one can estimate $\boldsymbol{\Sigma}$ by minimum covariance determinant (MCD) estimator of Rousseeuw (1984) and then $\boldsymbol{V}$ by its square root matrix. Note that, the Choleski decomposition of $\boldsymbol{\Sigma}$ (or, its estimate) may not produce an affine invariant rank function but the outlyingness function $\left\|rank_F^*(\boldsymbol{x})\right\|$ will be affine invariant (Makinde & Chakraborty, 2015).

Transformation and re-transformation approach

Chakraborty & Chaudhuri (1996) proposed transformation and re-transformation methodology for conversion of non-equivariant and non-invariant measures under affine transformation to affine equivariant and affine invariant versions respectively, using data driven coordinate system. and then used to construct an affine equivariant median. This technique was also used in Chakraborty & Chaudhuri (1998) to construct robust estimate of location; in Chakraborty, Chaudhuri & Oja (1998) to construct an affine equivariant median and angle test; in Chakraborty (2001) to construct an affine equivariant quantile and also in Dutta & Ghosh (2012); and in Makinde & Chakraborty (2015) to construct affine invariant classifier. The concept is to form an appropriate data driven coordinate system and express all the data points in terms of the new coordinate system. Then compute the spatial rank of the transformed data. Define

$$S_n = \left\{\alpha \mid \alpha \subset \{1, 2, \ldots, n\} \text{ and } |\alpha| = d+1\right\}$$

as the collection of all $d + 1$ subset of $\{1, 2, ..., n\}$. For a fixed $\alpha = \{i_0, i_1, \ldots, i_d\} \subset S_n$, we define $\boldsymbol{X}(\alpha)$ to be a $d \times d$ matrix whose columns are $X_{i_1} - X_{i_0}, X_{i_2} - X_{i_0}, \ldots, X_{i_d} - X_{i_0}$. That is, one of the $d + 1$ data points determines the origin and the lines joining that origin to the remaining d data point will form the coordinate system.

Assuming that elements of α are naturally ordered and that $\boldsymbol{X}_i$'s are independent and identically distributed observations with common probability distribution, which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, $\boldsymbol{X}(\boldsymbol{\alpha})$ is invertible with probability one (Chakraborty, 2001). So, $\boldsymbol{X}(\boldsymbol{\alpha})$ is the transformation matrix and for each $i \notin \alpha$, the data set $\boldsymbol{X}_i$ is transformed into a new coordinate system, $\boldsymbol{Y}_i = \{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{X}_i$ and then compute the rank of $\boldsymbol{y} = \{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{x}$. $\boldsymbol{X}(\alpha)$ is chosen in such a way that the columns of $\Sigma^{-\frac{1}{2}}\boldsymbol{X}(\alpha)$ are as orthogonal as possible. Because population covariance matrix $\boldsymbol{\Sigma}$ is unknown in practice, compute its estimate from the data. The choice of α depends on the value of α that minimizes

$$\zeta(\alpha) = \frac{trace\left(\{\boldsymbol{X}(\alpha)\}^T \Sigma^{-1}\boldsymbol{X}(\alpha)\right)/d}{\det\left(\{\boldsymbol{X}(\alpha)\}^T \Sigma^{-1}\boldsymbol{X}(\alpha)\right)^{1/d}}$$

so that $\zeta(\alpha)$ becomes very close to 1. Obviously, once α is selected, the computation of affine invariant spatial rank is straightforward in any dimension.

The affine invariant spatial rank is defined as

$$rank_F(\boldsymbol{x}) = E\left(\frac{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}]}{\left\|\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}]\right\|}\right) \tag{4}$$

The sample version is defined as

$$rank_{F_n}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]}{\left\|\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]\right\|} \tag{5}$$

Suppose $\boldsymbol{X}_i$, $1 \le i \le n$ be samples on $\mathbb{R}^d$ from a distribution F, it is easy to show that the rank function (defined in (5) above) of a data point $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{b}$ is $rank_G(\boldsymbol{y}) = rank_F(\boldsymbol{x})$, where G is the distribution of $\boldsymbol{y}$. This is shown by the theorem below.

Theorem 3. Suppose $\boldsymbol{X}_i$, $1 \le i \le n$ is a sample on $\mathbb{R}^d$ having a distribution F. For any $\alpha \in S_n$, $rank_{F_n}(\boldsymbol{x})$ defined in (5) is affine invariant.

Hence, the transformed multivariate rank is invariant under affine transformation. Any statistic based on this transformed rank is affine invariant and can handle the problem associated with deviation from spherical symmetry. Gao (2003**)** defined another version of spatial depth based on rank outlyingness defined in (1) and can be made affine invariant by replacing outlyingness of the rank function in (1) by its affine invariant version.

Numerical Example

To illustrate these methodologies, an example based on ordering of iris data (Fisher, 1936) is presented and quantiles of outlyingness functions of the variants of multivariate rank for the three species of iris flower are compared. The species are iris setosa, iris versicolor and iris virginica.

Presented in Table 2 are the quantiles and mean of the outlyingness of affine invariant and non-affine invariant rank for three species of iris data. The data is available on package R. We denote outlyingness function of affine invariant multivariate rank based on Cholesky decomposition of the covariance matrix by CD approach, outlyingness function of affine invariant multivariate rank based on transformation and re-transformation approach by TR approach and outlyingness function of affine non-invariant multivariate rank defined in equation (1) by non-invariant.

Observe that quantiles of rank outlyingness based on Cholesky decomposition of the covariance matrix and one based on transformation and re-transformation approach are close to each but far away from corresponding quantiles of values of outlyingness based on affine non-invariant multivariate rank. The implication of this is that correlation among the four variables (sepal length, sepal width, petal length and petal width) of each observation in the data can affect the performance of any statistical method or test based on non-affine invariant rank outlyingness.

Table 2. Ordering of species of Iris data based on the outlyingness functions of affine invariant and non-affine invariant ranks.

Iris Species	Approaches	*Minimum*	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Maximum*
	CD approach	0.2461	0.5500	0.6805	0.6447	0.7740	0.8827
Setosa	*TR approach*	0.2456	0.5383	0.6792	0.6437	0.7791	0.8820
	Non-invariant	0.1398	0.5138	0.6372	0.6160	0.7640	0.9436
	CD approach	0.2727	0.5520	0.6811	0.6506	0.7682	0.8755
Versicolor	*TR approach*	0.2710	0.5532	0.6862	0.6485	0.7603	0.8676
	Non-invariant	0.2992	0.4759	0.6406	0.6197	0.7317	0.9116
	CD approach	0.3879	0.5578	0.6724	0.6543	0.7382	0.8877
Virginica	*TR approach*	0.3513	0.5394	0.6722	0.6531	0.7596	0.9063
	Non-invariant	0.2808	0.4850	0.6447	0.6204	0.7439	0.9538

Observe that range of outlyingness of observations is noticeably bigger in affine non-invariant rank compare to the affine invariant rank. The minimum outlyingness value is least in affine non-invariant rank and may therefore mis-identify an observation as outlying. Hence, both affine invariant rank outlyingness functions perform quite well.

Conclusion

The effect of correlation on spatial rank outlyingness was considered and its possible applications. The spatial rank outlyingness based on the training sample does not depend on any distributional assumption and does not require any estimation of model parameters. These give a nonparametric flavor to any statistical technique based on multivariate rank. It is also computationally simple and can be applied to very high dimensional data as well. The rank outlyingness is not affine invariant and as a remedial measure we suggested a transformation of the data to a new coordinate system to make the rank outlyingness affine invariant.

The first idea of transformation is based on transformation retransformation approach proposed by Chakraborty (2001). This makes the spatial ranks affine invariant and hence the rank outlyingness becomes affine invariant. The other transformation considered is based on the square root of the scale matrix $\mathbf{\Sigma}$. It requires the estimation of $\mathbf{\Sigma}$ and may result in a non-robust rank outlyingness. Though the resulting spatial ranks are not affine invariant, rank outlyingness is affine invariant and usually computationally very simple if we use the sample covariance matrix as an estimate of $\mathbf{\Sigma}$. When variables of the data are independent of one another, then both affine invariant versions of rank outlyingness reduces to the usual rank outlyingness.

References

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics, 53*(2), 380–403. doi: 10.1023/a:1012478908041

Chakraborty, B. & Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. *Proceedings of the American Mathematical Society*, *124*(8), 2539–2546. doi: 10.1090/s0002-9939-96-03657-x

Chakraborty, B. & Chaudhuri, P. (1998). On an adaptive transformation and retransformation estimate of multivariate location. *Journal of the Royal Statistical Society: Series B*, *60*(1), 145–157. doi: 10.1111/1467-9868.00114

Chakraborty B., Chaudhuri, P. & Oja, H. (1998). Operating transformation and re-transformation on spatial median and angle test. *Statistica Sinica, 8*, 767–784.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of American Statistical Association, 91*(434), 862–872. doi: 10.1080/01621459.1996.10476954

Dutta, S. & Ghosh, A. K. (2012). *On classification based on L_p depth with an adaptive choice of p.* Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Gao, Y. (2003). Data depth based on spatial rank. *Statistics & Probability Letters, 65*(3), 217 – 225. doi: 10.1016/j.spl.2003.06.003

Guha, P. (2012). *On scale-scale curves for multivariate data based on rank regions*. PhD thesis, University of Birmingham, UK.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics, 25*(2), 435 – 477. doi: 10.1214/aos/1031833659

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics, 18*(1), 405–414. doi: 10.1214/aos/1176347507

Liu, R. Y. & Singh, K. (1993). A quality index based on multivariate data depth and multivariate rank tests. *Journal of the American Statistical Association, 88*(421), 252–260. doi: 10.1080/01621459.1993.10594317

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics, 27*(3), 783–858. doi: 10.1214/aos/1018031260

Makinde, O. S. (2015). *On some classification methods for high dimensional and functional data*. PhD Thesis, University of Birmingham

Makinde, O. S. & Chakraborty, B. (2015). On some classifiers based on distribution functions of multivariate ranks. In Nordhausen, K and Taskinen, S. (Eds). *Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja*. NY: Springer, 249–264. doi: 10.1007/978-3-319-22404-6_15

Möttönen, J. & Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics, 5*, 201–213.

Möttönen, J., Oja, H. & Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics, 25*(2), 542–552. doi: 10.1214/aos/1031833663

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880. doi: 10.1080/01621459.1984.10477105

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference, 123*(2), 259–278. doi: 10.1016/s0378-3758(03)00156-3

Serfling, R. (2006). Multivariate symmetry and asymmetry, In S. Kotz, N. Balakrishnan, C. B. Read & B. Vidakovic, Eds. *Encyclopedia of Statistical Sciences, Second Ed., 8*, 5338–5345. NY: Wiley. doi: 10.1002/0471667196.ess5011.pub2

Appendix

Proof of Theorem 3. For any $d \times d$ nonsingular matrix $\boldsymbol{A}$, let $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{X}_i + \boldsymbol{b}$. Since $\boldsymbol{X}(\alpha)=[\boldsymbol{X}_{i1} - \boldsymbol{X}_{i0}, \boldsymbol{X}_{i2} - \boldsymbol{X}_{i0}, \ldots, \boldsymbol{X}_{id} - \boldsymbol{X}_{i0}]$, we have

$$\begin{aligned}
\boldsymbol{Y}(\alpha) &= [\boldsymbol{Y}_{i1} - \boldsymbol{Y}_{i0}, \boldsymbol{Y}_{i2} - \boldsymbol{Y}_{i0}, \ldots, \boldsymbol{Y}_{id} - \boldsymbol{Y}_{i0}] \\
&= \left[\boldsymbol{A}\boldsymbol{X}_{i1} + \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b}), \boldsymbol{A}\boldsymbol{X}_{i2} + \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b}), \ldots, \boldsymbol{A}\boldsymbol{X}_{id} + \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{X}_{i0} + \boldsymbol{b})\right] \\
&= [\boldsymbol{A}\boldsymbol{X}_{i1} - \boldsymbol{A}\boldsymbol{X}_{i0}, \boldsymbol{A}\boldsymbol{X}_{i2} - \boldsymbol{A}\boldsymbol{X}_{i0}, \ldots, \boldsymbol{A}\boldsymbol{X}_{id} - \boldsymbol{A}\boldsymbol{X}_{i0}] \\
&= \boldsymbol{A}[\boldsymbol{X}_{i1} - \boldsymbol{X}_{i0}, \boldsymbol{X}_{i2} - \boldsymbol{X}_{i0}, \ldots, \boldsymbol{X}_{id} - \boldsymbol{X}_{i0}] \\
&= \boldsymbol{A}\boldsymbol{X}(\alpha)
\end{aligned}$$

The transformed multivariate rank of a data point $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{x} \in \mathbb{R}^d$ is

$$
\begin{aligned}
rank_{G_n}(\boldsymbol{y}) &= rank_{G_n}(\boldsymbol{AX}+\boldsymbol{b}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{Y}(\alpha)\}^{-1}[\boldsymbol{y}-\boldsymbol{Y}_i]}{\left\|\{\boldsymbol{Y}(\alpha)\}^{-1}[\boldsymbol{y}-\boldsymbol{Y}_i]\right\|} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{AX}(\alpha)\}^{-1}\left[\boldsymbol{Ax}+\boldsymbol{b}-[\boldsymbol{AX}_i+\boldsymbol{b}]\right]}{\left\|\{\boldsymbol{AX}(\alpha)\}^{-1}\left[\boldsymbol{Ax}+\boldsymbol{b}-[\boldsymbol{AX}_i+\boldsymbol{b}]\right]\right\|} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}[\boldsymbol{x}-\boldsymbol{X}_i]}{\{\boldsymbol{X}(\alpha)\}^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}[\boldsymbol{x}-\boldsymbol{X}_i]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]}{\{\boldsymbol{X}(\alpha)\}^{-1}[\boldsymbol{x}-\boldsymbol{X}_i]} = rank_{F_n}(\boldsymbol{x}).
\end{aligned}
$$