Journal of Statistical Software

February 2020, Volume 92, Issue 8. *doi: 10.18637/jss.v092.i08*

Computing the Oja Median in R: The Package OjaNP

Daniel Fischer
Natural Resources Institute Finland & University of Tampere

Karl Mosler
University of Cologne

Jyrki Möttönen
University of Helsinki

Klaus Nordhausen
Vienna University of Technology

Oleksii Pokotylo
University of Cologne

Daniel Vogel
University of Aberdeen

Abstract

The Oja median is one of several extensions of the univariate median to the multivariate case. It has many desirable properties, but is computationally demanding. In this paper, we first review the properties of the Oja median and compare it to other multivariate medians. Then, we discuss four algorithms to compute the Oja median, which are implemented in our R package **OjaNP**. Besides these algorithms, the package contains also functions to compute Oja signs, Oja signed ranks, Oja ranks, and the related scatter concepts. To illustrate their use, the corresponding multivariate one- and C-sample location tests are implemented.

Keywords: Oja median, Oja signs, Oja signed ranks, Oja ranks, R, C++.

1. Introduction

The univariate median is a popular location estimator. It is, however, not straightforward to generalize it to the multivariate case since no generalization is known that retains all properties of the univariate estimator, and therefore different generalizations emphasize different properties of the univariate median. So besides the Oja median described here, there are several other multivariate median concepts. Hayford (1902) suggested the first generalization by simply using the vector of the marginal medians. Other popular multivariate medians are Tukey's median (Tukey 1975) and the spatial median (also known as L_1 median). The spatial median was initially defined as a bivariate median (Weber 1909, 1929) and subsequently

extended to the general multivariate case. These and more multidimensional medians are surveyed in Small (1990) and Oja (2013). While the vector of marginal medians is quite easy to compute, the other multivariate medians are more computationally expensive. Particularly the Oja median (Oja 1983) has, despite its compelling statistical properties, not been used very often in practice so far, since it is difficult to compute. The main topic of this paper is to describe the R (R Core Team 2019) package **OjaNP** (Fischer, Mosler, Möttönen, Nordhausen, Pokotylo, and Vogel 2020), which provides several algorithms for the computation of the Oja median in R and is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=OjaNP`.

The outline of this paper is as follows. In Section 2.1, we review and compare some multivariate medians and show which properties of the univariate median is generalized by which multivariate median. Our main focus is on the Oja median, whose basic properties are discussed in Section 2.2, followed by an introduction to Oja signs and ranks (Section 2.3), Oja signed ranks (Section 2.4) and Oja sign and rank covariance matrices (Section 2.5). To demonstrate the application of the Oja median and its sign and rank concepts, Section 2.6 discusses one- and C-sample tests of location.

In Section 3, we focus on the different algorithms provided by the package **OjaNP** to calculate the Oja median. Four different algorithms are available: two exact algorithms and two approximate algorithms based on different designs.

Section 4 shows how to use the package **OjaNP** in order to calculate the Oja median and related statistics. We provide simple examples, and additional benchmarks are calculated to analyze the performance of the implementations.

A concept frequently encountered in this paper is affine equivariance. We use it in the sense of full-rank affine equivariance, which is common in robust statistics. Given the k-dimensional sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ we let $\mathbb{X} = (\mathbf{x}_1 \ldots \mathbf{x}_n)^\top$ be the data matrix of dimension $n \times k$, containing the data points as rows. Subsequently the data sample is identified with $\mathbb{X}$. For a given affine-linear transformation $T : \mathbb{R}^k \to \mathbb{R}^k$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{b} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathbb{R}^{k \times k}$ non-singular, the data matrix $\mathbb{Y}$ of the transformed data $T(\mathbf{x}_1), \ldots, T(\mathbf{x}_n)$ is given by

$$\mathbb{Y} = T(\mathbb{X}) = \mathbb{X}\mathbf{A}^\top + \mathbf{1}\mathbf{b}^\top,$$

where $\mathbf{1}$ denotes the $n \times 1$ vector consisting of ones. We call an $\mathbb{R}^k$-valued location statistic $\boldsymbol{\mu}(\mathbb{X})$ affine equivariant, if

$$\boldsymbol{\mu}(T(\mathbb{X})) = T(\boldsymbol{\mu}(\mathbb{X})) \quad \text{for all } T \text{ as above.} \tag{1}$$

This applies analogously to set-valued location statistics $\boldsymbol{\mu}$, such as median sets. For a matrix-valued scatter statistic $\mathbf{S}$ taking on values in $\mathbb{R}^{k \times k}$, affine equivariance is commonly understood as

$$\mathbf{S}(T(\mathbb{X})) = \mathbf{A}\mathbf{S}(\mathbb{X})\mathbf{A}^\top.$$

For a more detailed introduction to affine equivariance, see Oja (2010).

2. Oja median and related concepts

2.1. Oja median and other multivariate medians

We start by introducing the univariate median for distributions. Given a distribution function F, let

$$\begin{aligned} F^{-1}\left(\frac{1}{2}-\right) &= \inf\left\{x \in \mathbb{R} : F(x) \geq \frac{1}{2}\right\} \quad \text{and} \\ F^{-1}\left(\frac{1}{2}+\right) &= \sup\left\{x \in \mathbb{R} : F(x) \leq \frac{1}{2}\right\}. \end{aligned}$$

Then the median (set) of F is given by the interval

$$\text{Med}(F) = \left[F^{-1}\left(\frac{1}{2}-\right), \quad F^{-1}\left(\frac{1}{2}+\right)\right]. \tag{2}$$

Any point of the interval divides the distribution in two halves of equal probability weight and can represent the median. In case a unique selection is needed, we use the gravity center of the median set as a (single-point) median and denote it by the lower case symbol,

$$\text{med}(F) = \frac{F^{-1}\left(\frac{1}{2}+\right) + F^{-1}\left(\frac{1}{2}-\right)}{2}. \tag{3}$$

For a given sample $\mathbb{X} = (x_1, \ldots, x_n)$, the median $\text{med}(\mathbb{X})$ is obtained as

$$\text{med}(\mathbb{X}) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd}, \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even}, \end{cases}$$

with $x_{(i)}$ being the ith order statistic. The latter definition is a special case of (3), obtained by taking F to be the empirical distribution of $\mathbb{X}$, which gives equal probability mass to each of the points $x_1, \ldots, x_n$. Note that $\text{med}(\mathbb{X})$ is an affine equivariant location statistic.

Now let $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_k)$ be a k-dimensional data set, where $\mathbb{X}_i$ denotes the ith column of $\mathbb{X}$ and corresponds hence to the ith variable. The many existing notions of a k-variate median for such data have in common that they reduce to the univariate median for $k = 1$. Multivariate medians are generally non-unique, and we select, as above, the gravity center of the median set to obtain a unique representation.

To the best of our knowledge the first generalization of the univariate median to the multivariate case is the vector of marginal medians **mmed** described in Hayford (1902).

Definition 1. *The vector of marginal medians* **mmed** *of the sample* $\mathbb{X}$ *is defined as*

$$\mathbf{mmed}(\mathbb{X}) = (\text{med}(\mathbb{X}_1), \ldots, \text{med}(\mathbb{X}_p))^\top.$$

The vector of marginal medians is easily computed but not affine equivariant.

Example 1. *A simple rotation of two-dimensional data visualizes the problem of not affine equivariant transformations. In the left part of Figure 1 a two-dimensional data set is plotted and rotated around the center point +. At the same time the marginal median* **mmed** *is*

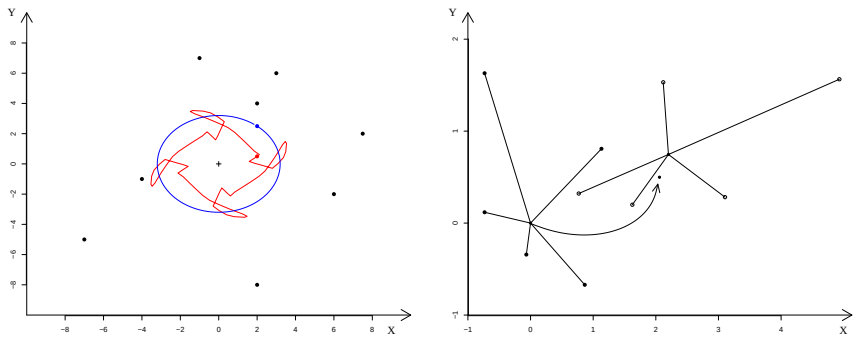

Figure 1: Transformation of different multivariate medians.

continuously plotted in red. An affine equivariant median would follow the rotation on a circle, meaning that it would not be affected by the degree of the rotation of the original data. The blue circle represents the Oja median and hence, illustrates the behavior of an affine equivariant median.

In the right part of Figure 1 we visualize the behavior of a rotation invariant estimator that is not scale invariant. We use the spatial median **smed** *as an example for a not scale invariant statistic, see Definition 3 below. For five example points the median is calculated and indicated as the center point, connected with lines to the data points. Then, the scale of those five points was transformed into the star shape on the right side of this plot. The location of the spatial median of the transformed data is indicated at the top of curved arrow. An affine equivariant median, however, would still be equivalent to the transformed median of the original data. The Oja median (and any other affine equivariant statistic) fulfills this criterion.*

The second generalization of the univariate median reviewed here is based on the fact that for a given sample $\mathbb{X}$ and its median $\mathrm{med}(x)$ the equation

$$\sum_{i=1}^{n} \mathbb{1}_{(-\infty,\mathrm{med}(x)]}(x_i) = \sum_{i=1}^{n} \mathbb{1}_{[\mathrm{med}(x),\infty)}(x_i)$$

holds, where $\mathbb{1}_A(x_i) = 1$ if $x_i \in A$ and $= 0$ otherwise. This means, there are as many observations smaller than $\mathrm{med}(x)$ as are bigger, as it was pointed out, e.g., in Hotelling (1929). Given a k-dimensional data set, we consider halfspaces in every direction $\mathbf{v} \in S^{k-1} = \{\mathbf{x} \mid \|\mathbf{x}\| = 1\}$. Let $H_\mathbf{v}$ denote the "minimal" halfspace with normal vector $\mathbf{v}$ that contains at least half of the data points, i.e., for any other halfspace $\tilde{H}_\mathbf{v}$ with these properties $H_\mathbf{v} \subset \tilde{H}_\mathbf{v}$ holds. The intersection of all $H_\mathbf{v}, \mathbf{v} \in S^{k-1}$, forms the Tukey median $\mathbf{Tmed}(\mathbb{X})$. If the data are in general position, each such halfspace is bordered by a hyperplane through exactly k data points, and the Tukey median is in general no singleton. (A set of k-variate data is in general position if at most k of them lie on the same hyperplane.) The unique Tukey median $\mathbf{tmed}(\mathbb{X})$ is defined as the gravity point of this median set; see Tukey (1975) and Donoho and Gasko (1992). It is affine equivariant (Chen 1995) and can be introduced as the maximizer of a depth function as follows.

Definition 2. *Let $\mathbb{X}$ be a k-dimensional sample as above. For any $\mathbf{x} \in \mathbb{R}^k$,*

$$\mathbf{tdep}_\mathbb{X}(\mathbf{x}) = \frac{1}{n} \min_{\|\mathbf{v}\|=1} \#\{i : \mathbf{v}^\top \mathbf{x}_i \geq \mathbf{v}^\top \mathbf{x}\} \tag{4}$$

is called the Tukey depth or location depth of $\mathbf{x}$ *w.r.t.* $\mathbb{X}$*. (Here* $\#\{S\}$ *denotes the cardinality of a set* S*.) The Tukey median* $\mathbf{Tmed}(\mathbb{X})$ *of the sample* $\mathbb{X}$ *is defined as the maximizer of the Tukey depth,*

$$\mathbf{Tmed}(\mathbb{X}) = \underset{\mathbf{x}\in\mathbb{R}^k}{\operatorname{argmax}}\{\mathbf{tdep}_{\mathbb{X}}(\mathbf{x})\}\,. \tag{5}$$

Another way to generalize the univariate median is to transfer its minimizing feature into higher dimensions. Consider again a univariate sample $\mathbb{X} = (x_1, \ldots, x_n)$ in $\mathbb{R}$. When minimizing the sum $\sum_{i=1}^n |x_i - x|$ over $x \in \mathbb{R}$ we obtain

$$\operatorname{med}(\mathbb{X}) = \underset{x\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - x|\,. \tag{6}$$

This minimizing feature of the univariate median is interpretable in two ways: It is the sum of absolute deviations, but it can also be viewed as the sum of one-dimensional simplices. The first interpretation will lead us to the spatial median, the second to the Oja median.

The spatial median is presumably the most popular multivariate median and almost as old as the marginal median. Weber (1909) first described the spatial median and used it to solve an economic problem: He was looking for the best place of a distribution center in the sense, that the sum of distances between outposts and the distribution center becomes minimal. The k-dimensional extension of this approach is the spatial median.

Definition 3. *The spatial median* **smed** *of a* k*-variate data sample* $\mathbb{X}$ *is defined as*

$$\mathbf{smed}(\mathbb{X}) = \underset{\mathbf{x}\in\mathbb{R}^k}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}\|_2 \right\}. \tag{7}$$

Here $\|\cdot\|_2$ denotes the Euclidean norm. The spatial median is sometimes also referred to as the L_1 median since it minimizes the L_1 norm of the n-variate vector of distances $\|\mathbf{x}_i - \mathbf{x}\|_2$. The spatial median is not affine equivariant as is visualized in the right part of Figure 1. Given an affine transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ which transforms the data points from the left star-shaped figure into the right star-shaped figure, the center of each star is the spatial median of the data points. However, due to the lack of affine equivariance, the transformed spatial median $T(\mathbf{smed}(\mathbb{X}))$ (marked with the arrow) does not coincide with the spatial median $\mathbf{smed}(T(\mathbb{X}))$ of the transformed data set. The spatial median is rotation invariant, but not scale invariant.

Oja (1983) introduced a multivariate median based on volumes of simplices. A k-dimensional simplex is the convex hull of $(k+1)$ spanning points in general location. Let

$$\mathbf{x}_1 = (x_{1,1}, \ldots, x_{1,k})^\top, \mathbf{x}_2 = (x_{2,1}, \ldots, x_{2,k})^\top, \ldots, \mathbf{x}_{k+1} = (x_{k+1,1}, \ldots, x_{k+1,k})^\top$$

be $k+1$ points in general location from $\mathbb{R}^k$. The volume $V(\mathbf{x}_1, \ldots, \mathbf{x}_{k+1})$ of the simplex spanned by the points $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$ is then given by

$$V(\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}) = \operatorname{abs}\left(\frac{1}{k!} \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{2,1} & \cdots & x_{k+1,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{k+1,2} \\ \vdots & \vdots & & \vdots \\ x_{1,k} & x_{2,k} & \cdots & x_{k+1,k} \end{pmatrix} \right), \tag{8}$$

see, e.g., Stein (1966). Let F be a distribution on $\mathbb{R}^k$ having finite first moment. The Oja median $\mathbf{Omed}(F)$ of F is defined as follows: for i.i.d. random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_k$, each distributed with F, define the Oja median set as

$$\mathbf{Omed}(F) = \underset{\boldsymbol{\mu} \in \mathbb{R}^k}{\operatorname{argmin}} \, \mathsf{E}(V(\mathbf{X}_1, \ldots, \mathbf{X}_k, \boldsymbol{\mu})).$$

For data we have the following definition. Let

$$P_{n,k} \;=\; \{\, p = (i_1, \ldots, i_k) \,|\, 1 \le i_1 < \cdots < i_k \le n\} \tag{9}$$

be the set of all ordered k-tuples out of $\{1, \ldots, n\}$, $1 \le k \le n$.

Definition 4. *The Oja median* **Omed** *of a k-dimensional sample $\mathbb{X} = (\mathbf{x}_1 \ldots \mathbf{x}_n)^\top$ of size $n > k$ is defined as*

$$\mathbf{Omed}(\mathbb{X}) = \underset{\mathbf{x} \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \sum_{(i_1, \ldots, i_k) \in P_{n,k}} V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}, \mathbf{x}) \right\}. \tag{10}$$

This means that the Oja median of a k-dimensional sample is any point $\mathbf{x} \in \mathbb{R}^k$ for which the sum of simplex volumes over all combinations of possible k data points is minimal. Note that $\mathbf{Omed}(\mathbb{X})$ equals $\mathbf{Omed}(F_{\mathbb{X}})$, where $F_{\mathbb{X}}$ is the empirical distribution on $\mathbf{x}_1 \ldots \mathbf{x}_n$. A unique version of the Oja median, denoted $\mathbf{omed}(\mathbb{X})$, is obtained by selecting the point of gravity of $\mathbf{Omed}(\mathbb{X})$.

2.2. Properties of the Oja median

As other multivariate extensions of the median, the Oja median is not unique. In the bivariate case, we have the following result.

Theorem 1. *If n is even and the data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$ are in general position, then the Oja median is unique.*

For details see Niinimaa (1995). The author also identifies a necessary and sufficient condition for the bivariate Oja median to be unique if n is odd. There appears to be no result for higher dimensions, but Oja (1999) conjectures that the Oja median is unique for even (odd) sample sizes if the dimension k is even (odd). Figure 2 visualizes the Oja median in several small-sample data situations.

Theorem 2. *The Oja median of a sample is a convex set.*

See Oja and Niinimaa (1985) for details. Theorem 2 is illustrated in Figure 3. Contour lines of the objective function (10) are plotted for two small data clouds. The objective function is convex for, both, even and odd sample sizes.

Theorem 3. *The Oja median is affine equivariant.*

Hence the Oja median is a proper location statistic in the sense of (1), i.e., $\mathbf{omed}(T(\mathbb{X})) = T(\mathbf{omed}(\mathbb{X}))$. The next result can be found, e.g., in Arcones, Chen, and Gine (1994), Shen (2008).

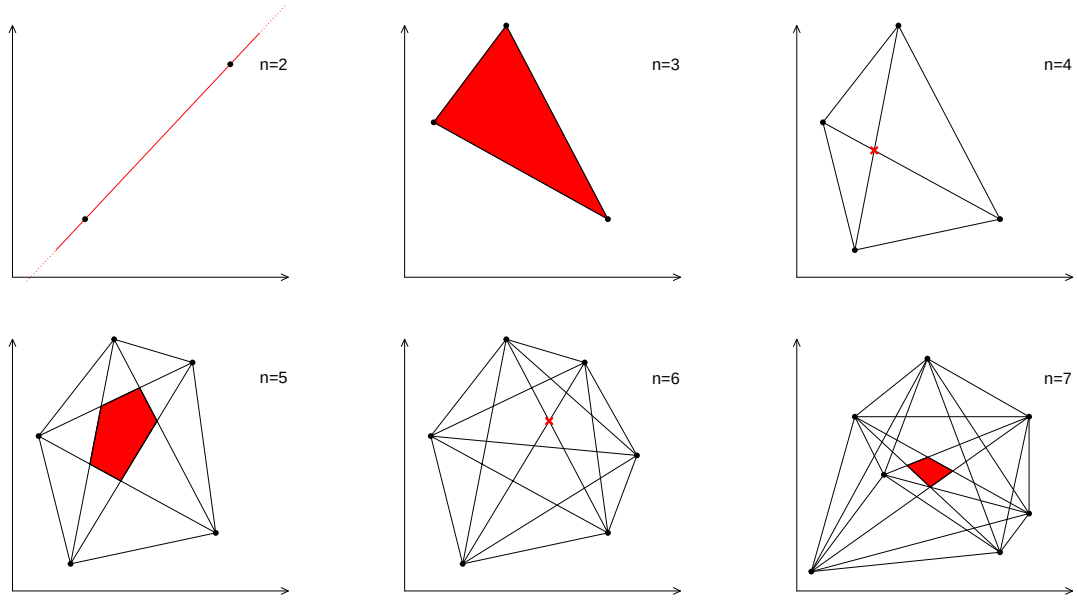

Figure 2: Example plots for the bivariate case.

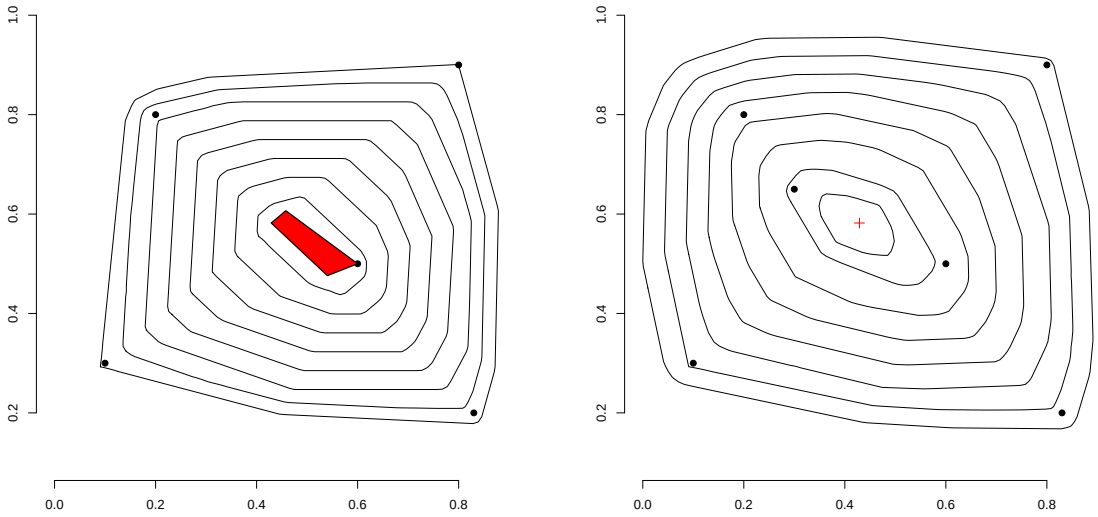

Figure 3: Contour plot for the bivariate case.

Theorem 4. *Under mild regularity conditions (including the existence of first moments) for an i.i.d. sample* $\mathbb{X} = (\mathbf{x}_1 \dots \mathbf{x}_n)^\top$ *from the distribution* F*, we have*

$$\sqrt{n}(\mathbf{omed}(\mathbb{X}) - \mathbf{omed}(F)) \to_d N_k(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^\top),$$

where $\mathbf{B}$ *is the Oja sign covariance matrix (OSCM) defined later in Section 2.5 at* F *and* $\mathbf{A}$ *is the expected covariance matrix between the Oja signs (***osgn***, see Section 2.3) and the optimal location score of* F*.*

Median	Affine equivariance	Breakdown point
Marginal median	No	$1/2$
Tukey median	Yes	$1/(k+1)$
Spatial median	No	$1/2$
Oja median	Yes	0
	Influence function	Asymptotic distribution
Marginal medians	Niinimaa and Oja (1995)	Babu and Rao (1988)
Tukey median	Romanazzi (2001)	Bai and He (1999)
Spatial median	Niinimaa and Oja (1995)	Möttönen, Nordhausen, and Oja (2010)
Oja median	Niinimaa and Oja (1995)	Shen (2008)

Table 1: Main properties of different multivariate medians.

Concerning robustness the Oja median has the following properties.

Theorem 5. *Let F have a finite first moment. Then the Oja median has a bounded influence function.*

The influence function is for example given in Niinimaa and Oja (1995). The Oja median has an asymptotic breakdown point of 0. Niinimaa, Oja, and Tableman (1990) show the following:

Theorem 6. *The finite-sample breakdown-point of the bivariate Oja median is $2/(n+2)$.*

Table 1 summarizes the main properties of the Oja median and other common multivariate medians discussed here. For another recent discussion of the different medians see also Oja (2013).

2.3. Oja signs and ranks

Closely related to the median is the concept of signs and ranks. In this section we introduce the multivariate Oja sign and Oja rank and relate them to other multivariate signs and ranks, corresponding to the marginal and the spatial median. Again, to motivate the subsequent derivations we take a brief look at the univariate case. Suppose we have a univariate data set $\mathbb{X} = (x_1 \dots x_n)^\top$, $n \in \mathbb{N}$. We call

$$\operatorname{sgn}_{\mathbb{X}}(x) = \operatorname{sgn}(x - \operatorname{med}(\mathbb{X})), \quad x \in \mathbb{R}, \tag{11}$$

the sign of x w.r.t. the data sample $\mathbb{X}$, where sgn is the univariate sign function ($\operatorname{sgn}(x) = \frac{x}{|x|}$ if $x \neq 0$ and zero otherwise), and $\operatorname{med}(\mathbb{X})$ is the univariate median of the sample $\mathbb{X}$. There are several possibilities of suitably assigning ranks to the data points. By

$$\operatorname{rnk}_{\mathbb{X}}(x) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(x - x_i), \quad x \in \mathbb{R}, \tag{12}$$

we define normalized central ranks, which may take on $2n+1$ possible values ranging from -1 to 1. In the following we call $\operatorname{rnk}_{\mathbb{X}}(x)$ simply the rank of x w.r.t. $\mathbb{X}$.

The median appropriately centers the data, i.e., the signs of the data points, centered by the median, sum up to zero:

$$\sum_{i=1}^{n} \mathrm{sgn}_{\mathbb{X}}(x_i) = \sum_{i=1}^{n} \mathrm{sgn}(x_i - \mathrm{med}(\mathbb{X})) = 0. \tag{13}$$

The sample mean does the same for the data points themselves. In other words we may say that the median has central rank zero,

$$\mathrm{rnk}_{\mathbb{X}}(\mathrm{med}(\mathbb{X})) = -\frac{1}{n} \sum_{i=1}^{n} \mathrm{sgn}(x_i - \mathrm{med}(\mathbb{X})) = 0\,, \tag{14}$$

and, in this respect, is the most central point. Identities (13) and (14) (which are different formulations of the fact that half of the data lies above and below the median) provide the essential link between signs and ranks and the median and are a motivating principle behind the multivariate sign and rank functions we will introduce next. Note that a k-variate sign function should be a vector that can point in any direction of the k-dimensional space. The same holds for a multivariate rank function based on signs.

An obvious extension of (11) to the multivariate setting is its componentwise application, leading to the marginal sign function. We call

$$\mathbf{msgn}_{\mathbb{X}}(\mathbf{x}) = \mathbf{msgn}(\mathbf{x} - \mathbf{mmed}(\mathbb{X}))$$

the marginal sign of $\mathbf{x} \in \mathbb{R}^k$ w.r.t. the k-variate data sample $\mathbb{X} = (\mathbf{x}_1 \dots \mathbf{x}_n)^\top$, where $\mathbf{x} = (x_1 \dots x_k)^\top$,

$$\mathbf{msgn}(\mathbf{x}) = (\mathrm{sgn}(x_1) \dots \mathrm{sgn}(x_k))^\top,$$

and $\mathbf{mmed}(\mathbb{X})$ is the marginal median of $\mathbb{X}$. An equally straightforward generalization is the spatial sign of $\mathbf{x}$ w.r.t. $\mathbb{X}$:

$$\mathbf{ssgn}_{\mathbb{X}}(\mathbf{x}) = \mathbf{ssgn}(\mathbf{x} - \mathbf{smed}(\mathbb{X})), \quad \mathbf{x} \in \mathbb{R}^k,$$

where

$$\mathbf{ssgn}(\mathbf{x}) = \begin{cases} \frac{1}{\|\mathbf{x}\|}\,\mathbf{x} & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0}, \end{cases}$$

and $\mathbf{smed}(\mathbb{X})$ is the spatial median of $\mathbb{X}$. The corresponding rank functions, the marginal rank $\mathbf{mrnk}_{\mathbb{X}}$ and the spatial rank $\mathbf{srnk}_{\mathbb{X}}$, are obtained by replacing sgn in (12) by $\mathbf{msgn}$ and $\mathbf{ssgn}$, respectively. The Oja sign is defined as follows. For $0 \leq k \leq n$, let $N_{n,k} = \binom{n}{k}$ and $P_{n,k}$ as in (9). We call

$$\begin{aligned} \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}; \mathbf{m}) &= \frac{1}{N_{n,k-1}} \sum_{(i_1,\dots,i_{k-1}) \in P_{n,k-1}} \nabla_{\mathbf{x}} \left| \det(\mathbf{x}_{i_1} - \mathbf{m}, \dots, \mathbf{x}_{i_{k-1}} - \mathbf{m}, \mathbf{x} - \mathbf{m}) \right| \\ &= \frac{1}{N_{n,k-1}} \sum_{(i_1,\dots,i_{k-1}) \in P_{n,k-1}} \nabla_{\mathbf{x}} \left| \det \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ \mathbf{m} & \mathbf{x}_{i_1} & \dots & \mathbf{x}_{i_{k-1}} & \mathbf{x} \end{pmatrix} \right| \end{aligned} \tag{15}$$

the Oja sign of the point $\mathbf{x} \in \mathbb{R}^k$ w.r.t. the data sample $\mathbb{X}$ and the center location $\mathbf{m}$ and

$$\mathbf{osgn}_{\mathbb{X}}(\mathbf{x}) = \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}; \mathbf{omed}(\mathbb{X})) \tag{16}$$

simply the Oja sign of $\mathbf{x}$ w.r.t. $\mathbb{X}$. Furthermore

$$\begin{aligned}\mathbf{ornk}_{\mathbb{X}}(\mathbf{x}) &= \frac{1}{n-k+1}\sum_{i=1}^{n}\mathbf{osgn}_{\mathbb{X}}(\mathbf{x};\mathbf{x}_i)\\ &= \frac{1}{N_{n,k}}\sum_{(i_1,\ldots,i_k)\in P_{n,k}}\nabla_{\mathbf{x}}\left|\det\begin{pmatrix}1 & \ldots & 1 & 1\\ \mathbf{x}_{i_1} & \ldots & \mathbf{x}_{i_k} & \mathbf{x}\end{pmatrix}\right| \qquad (17)\end{aligned}$$

is the Oja rank of $\mathbf{x}$ w.r.t. $\mathbb{X}$. The notation $\nabla_{\mathbf{x}}$ means the gradient (the derivative as a column vector) w.r.t. $\mathbf{x}$. We define the derivative of $|\cdot|$ to be zero at the origin, i.e., $\frac{d}{dx}|x| = \mathrm{sgn}(x)$, $x \in \mathbb{R}$.

We note a qualitative difference in the definition of the Oja sign to the marginal and the spatial sign. The Oja sign function $\mathbf{osgn}_{\mathbb{X}}$ depends on the sample $\mathbb{X}$ not only through the centering point $\mathbf{omed}(\mathbb{X})$, but is a function of the whole sample. By introducing the parameter $\mathbf{m}$ in definition (15) we allow the data to be centered by a different central location than the Oja median. This may be useful under some circumstances, for example to speed up computation. The function `ojaSign` offers this option.

We also note that for $k = 1$ expressions (16) and (17) reduce to

$$\mathbf{osgn}_{\mathbb{X}}(x) = \frac{d}{dx}\,|\det(x - \mathrm{med}(\mathbb{X}))\,| = \mathrm{sgn}_{\mathbb{X}}(x)$$

and

$$\mathbf{ornk}_{\mathbb{X}}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dx}\left|\det\begin{pmatrix}1 & 1\\ x_i & x\end{pmatrix}\right| = \frac{1}{n}\sum_{i=1}^{n}\mathrm{sgn}(x - x_i) = \mathrm{rnk}_{\mathbb{X}}(x),$$

hence $\mathbf{osgn}_{\mathbb{X}}$ and $\mathbf{ornk}_{\mathbb{X}}$ are proper generalizations of $\mathrm{sgn}_{\mathbb{X}}$ and $\mathrm{rnk}_{\mathbb{X}}$, respectively, in the sense that for $k = 1$ they coincide with their univariate counterparts.

In dimensions 2 and 3, Oja signs and ranks allow a descriptive geometric interpretation: Recall that the k-dimensional volume of the simplex spanned by $k+1$ points $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$ in $\mathbb{R}^k$ is given by $V(\mathbf{x}_1, \ldots, \mathbf{x}_{k+1})$, cf. (8). Hence

$$\det\begin{pmatrix}1 & \ldots & 1 & 1\\ \mathbf{x}_1 & \ldots & \mathbf{x}_k & \mathbf{x}\end{pmatrix} = 0$$

characterizes all points $\mathbf{x} \in \mathbb{R}^k$ that lie on the $k-1$ dimensional hyperplane spanned by the k points $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Then for $k = 3$, $\mathbf{osgn}_{\mathbb{X}}(\mathbf{x})$ is the average of the $N_{n,k-1} = \binom{n}{k-1}$ k-dimensional vectors $\mathbf{v}_p$, $p = (i_1, \ldots, i_{k-1}) \in P_{n,k-1}$, where $\mathbf{v}_p$

- is perpendicular to the plane going through the k points $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k-1}}$ and $\mathbf{omed}(\mathbb{X})$,
- has length equal to $(k-1)!$ times the area of the triangle that is bordered by the k points and,
- points from the plane to the point $\mathbf{x}$.

If $\mathbf{x}$ itself lies on the plane or if the k points do not uniquely determine a plane, the vector $\mathbf{v}_p$ is zero. Likewise, $\mathbf{ornk}_{\mathbb{X}}(\mathbf{x})$ is the average of $N_{n,k} = \binom{n}{k}$ vectors $\mathbf{v}_p, p = (i_1, \ldots, i_k) \in P_{n,k}$, each perpendicular to the plane spanned by $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$. This holds analogously for $k = 2$ with

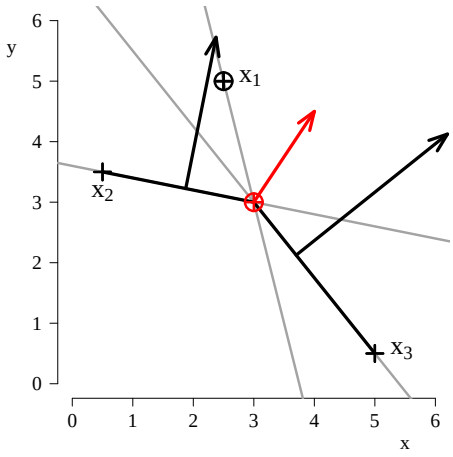

Figure 4: Three points in $\mathbb{R}^2$: The Oja sign (in red) of point $\mathbf{x}_1$ is the average of three vectors: the two black vectors and the null vector.

plane, *triangle*, and *area* being replaced by *straight line*, *line segment*, and *length*, respectively. The construction of the Oja sign is visualized in Figure 4 at the very simple example of three points in $\mathbb{R}^2$: The red $\oplus$ is an interior point (let it be denoted by $\mathbf{m}_0$) of the Oja median set. The latter is the triangle bordered by the three data points. Then $\mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_1; \mathbf{m}_0)$, the vector in red, is formed as the average of the two black vectors and the null vector (the latter being interpreted as a vector pointing perpendicularly from the straight line $\mathbf{x}_1$–$\mathbf{m}_0$ to $\mathbf{x}_1$). In this example, $\mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_1; \mathbf{m}_0)$ is invariant w.r.t. the specific choice of $\mathbf{m}_0$ as long as $\mathbf{m}_0$ is below both lines $\mathbf{x}_1$–$\mathbf{x}_2$ and $\mathbf{x}_1$–$\mathbf{x}_3$.

As in the univariate case, cf. (13) and (14), the multivariate signs of the data points sum up to zero, i.e.,

$$\sum_{i=1}^{n} \mathbf{sgn}_{\mathbb{X}}(x_i) = \mathbf{0},$$

where $\mathbf{sgn}_{\mathbb{X}}$ may be any of $\mathbf{msgn}_{\mathbb{X}}$, $\mathbf{ssgn}_{\mathbb{X}}$ or $\mathbf{osgn}_{\mathbb{X}}$. Equivalently,

$$\mathbf{rnk}_{\mathbb{X}}(\mathbf{med}(\mathbb{X})) = \mathbf{0}, \tag{18}$$

where $\mathbf{rnk}_{\mathbb{X}}$ may be any of $\mathbf{mrnk}_{\mathbb{X}}$, $\mathbf{srnk}_{\mathbb{X}}$ or $\mathbf{ornk}_{\mathbb{X}}$, and $\mathbf{med}(\mathbb{X})$ the respective multivariate median. This is generally only true if the appropriate median is chosen as centering point, for Oja signs in particular, the data has to be centered by the Oja median. Some further care must be given to these statements. If the median set is no singleton, they may be false for median points on its border. This also happens in the univariate case, when, for an even number n of observations, one chooses the median to be $x_{(\frac{n}{2})}$ or $x_{(\frac{n}{2}+1)}$. The function `ojaMedian` is likely to return points at the border of the median set. In (7) and (10), respectively, the spatial median and the Oja median were introduced as minimizers of objective functions, both generalizing the univariate case (6). Spatial and Oja ranks may be introduced as derivatives of these objective functions. Thus (18) is in concordance with the optimality property of the medians.

The negative rank function $-\mathbf{ornk}_{\mathbb{X}}(\mathbf{x})$ defines a hyperplane through $\mathbf{x}$, on the positive side of which the Oja median is found. This property is used by the exact bounded algorithm (Section 3.2) to reduce the search region for the median.

Similar to the multivariate medians, the multivariate signs differ by their equivariance and invariance properties. All multivariate signs are translation invariant, i.e., for $\mathbb{Y} = \mathbb{X} + \mathbf{1}_n\mathbf{b}^\top$, $\mathbf{b} \in \mathbb{R}^k$, we have

$$\mathbf{sgn}_{\mathbb{Y}}(\mathbf{x} + \mathbf{b}) = \mathbf{sgn}_{\mathbb{X}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^k,$$

where, again, **sgn** is any of **msgn**, **ssgn**, **osgn**. Marginal signs are furthermore invariant w.r.t. monotonously increasing, componentwise transformations, in particular

$$\mathbf{msgn}_{\mathbb{X}\mathbf{D}^\top}(\mathbf{D}\mathbf{x}) = \mathbf{msgn}_{\mathbb{X}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^k,$$

for $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_k)$ with $d_i > 0$, $i = 1, \ldots, k$. Spatial signs on the other hand are equivariant under orthogonal transformations, i.e.,

$$\mathbf{ssgn}_{\mathbb{X}\mathbf{U}^\top}(\mathbf{U}\mathbf{x}) = \mathbf{U}\,\mathbf{ssgn}_{\mathbb{X}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^k,$$

for any orthogonal $\mathbf{U} \in \mathbb{R}^{k\times k}$. Oja signs even obey a form of affine equivariance, an inverse proportional affine equivariance:

$$\mathbf{osgn}_{\mathbb{X}\mathbf{A}^\top}(\mathbf{A}\mathbf{x}) = \det(\mathbf{A})(\mathbf{A}^{-1})^\top\mathbf{osgn}_{\mathbb{X}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^k, \tag{19}$$

for any non-singular $\mathbf{A} \in \mathbb{R}^{k\times k}$, and they appear in the literature under the name affine equivariant signs. The respective multivariate rank functions have analogous equivariance and invariance properties. More details on Oja signs and ranks and their applications can be found, e.g., in Oja (1999). Hettmansperger and McKean (2011) and Nordhausen and Oja (2018) give an overview of multivariate sign and rank methods in general, Puri and Sen (1971) treat methods based on marginal signs and ranks and Oja (2010) describes spatial sign and rank methods.

2.4. Oja signed ranks

As in the Section 2.3 we consider first the univariate case. Suppose we have a univariate data set $\mathbb{X} = (x_1, \ldots, x_n)$. The signed rank of x w.r.t. $\mathbb{X}$ can be defined by

$$\mathrm{rnk}^+_{\mathbb{X}}(x) \quad = \quad \frac{1}{2n}\sum_{i=1}^{n}\sum_{a\in A}\mathrm{sgn}(x - ax_i), \tag{20}$$

where $A = \{-1, 1\}$. Note that the signed rank of the jth observation can be written as

$$\mathrm{rnk}^+_{\mathbb{X}}(x_j) \quad = \quad \frac{1}{2n}\{2\ \mathrm{rnk}(|x_j|) - 1\}\,\mathrm{sgn}(x_j), \tag{21}$$

where $|x_j|$ is ranked among absolute values $|x_1|, \ldots, |x_n|$. The Wilcoxon signed rank statistic $\sum_{i=1}^n \mathrm{rnk}(|x_i|)\,\mathrm{sgn}(x_i)$ is thus asymptotically equivalent with $n\sum_{i=1}^n \mathrm{rnk}^+_{\mathbb{X}}(x_i)$.

The signed rank can be extended to the multivariate case, provided we have a concept of multivariate sign. We get spatial signed rank $\mathbf{srnk}^+_{\mathbb{X}}(\mathbf{x})$ if we replace sgn in (20) by **ssgn**. The Oja signed rank can be defined in the following way. Let $\mathbb{A}$ be the set of 2^k possible vectors $(\pm 1, \ldots, \pm 1)$, i.e., $\mathbb{A} = \{\mathbf{a} = (a_1, \ldots, a_k) : a_1 = \pm 1, \ldots, a_k = \pm 1\}$. Then

$$\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{x}) \quad = \quad \frac{1}{2^k N_{n,k}}\sum_{P_{n,k}}\sum_{\mathbf{a}\in\mathbb{A}}\nabla_{\mathbf{x}}\left|\det\begin{pmatrix} 1 & \cdots & 1 & 1 \\ a_1\mathbf{x}_{i_1} & \cdots & a_k\mathbf{x}_{i_k} & \mathbf{x}\end{pmatrix}\right| \tag{22}$$

is the Oja signed rank of $\mathbf{x}$ w.r.t. $\mathbb{X}$. It is easy to see that the Oja signed rank coincides with the univariate signed rank when $k = 1$. Furthermore, it can be shown that

- $\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{x})$ is odd: $\mathbf{ornk}^+_{\mathbb{X}}(-\mathbf{x}) = -\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{x})$.
- $\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{x})$ points (approximately) in the direction of $\mathbf{x}$.
- $\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{0}) = \mathbf{0}$.
- $\mathbf{ornk}^+_{\mathbb{X}}(\mathbf{x})$ is bounded, piecewise constant and increases in magnitude as $\mathbf{x}$ moves away from $\mathbf{0}$.

2.5. Oja sign and rank matrices

Multivariate signs can be useful for obtaining information about the spread of the data and dependencies among the variables. We call

$$\mathrm{OSCM}(\mathbb{X}; \mathbf{m}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_i; \mathbf{m}) \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_i; \mathbf{m})^\top$$

the Oja sign covariance matrix of $\mathbb{X}$ w.r.t. the central location $\mathbf{m} \in \mathbb{R}^k$,

$$\mathrm{OSCM}(\mathbb{X}) = \mathrm{OSCM}(\mathbb{X}; \mathbf{omed}(\mathbb{X})) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_i)\, \mathbf{osgn}_{\mathbb{X}}(\mathbf{x}_i)^\top$$

the Oja sign covariance matrix of $\mathbb{X}$ and

$$\mathrm{ORCM}(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{ornk}_{\mathbb{X}}(\mathbf{x}_i)\, \mathbf{ornk}_{\mathbb{X}}(\mathbf{x}_i)^\top$$

the Oja rank covariance matrix of $\mathbb{X}$. In an analogous way we define marginal and spatial sign and rank matrices MSCM, SSCM and MRCM, SRCM by replacing $\mathbf{osgn}_{\mathbb{X}}$ by $\mathbf{msgn}_{\mathbb{X}}$, $\mathbf{ssgn}_{\mathbb{X}}$ and $\mathbf{ornk}_{\mathbb{X}}$ by $\mathbf{mrnk}_{\mathbb{X}}$ and $\mathbf{srnk}$, respectively. A comparative study of all sign and rank matrices presented here can be found in Visuri, Koivunen, and Oja (2000). Figure 5 shows a small two-dimensional data sample, together with the different multivariate medians and the corresponding signs and sign covariance matrices visualized as ellipses. The ellipses have radius $\sqrt{\chi^2_{2,0.9}}$, i.e., the positive definite matrix S is depicted by the ellipse $\mathbf{x}^\top S^{-1} \mathbf{x} = \chi^2_{2,0.9}$. The affine equivariance (19) for Oja signs and ranks translates into a similar equivariance for the corresponding covariance matrices,

$$\mathrm{OSCM}(\mathbb{X}\mathbf{A}^\top + \mathbf{1}_n \mathbf{b}^\top) = \det(\mathbf{A})^2 (\mathbf{A}^{-1})^\top \mathrm{OSCM}(\mathbb{X}) \mathbf{A}^{-1}$$

and

$$\mathrm{ORCM}(\mathbb{X}\mathbf{A}^\top + \mathbf{1}_n \mathbf{b}^\top) = \det(\mathbf{A})^2 (\mathbf{A}^{-1})^\top \mathrm{ORCM}(\mathbb{X}) \mathbf{A}^{-1},$$

for $\mathbf{b} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathbb{R}^{k \times k}$ with full rank. This has, roughly speaking, the consequence that OSCM and ORCM estimate the inverse of the covariance matrix up to scale. More specifically, Ollila, Oja, and Croux (2003) show that $\mathrm{OSCM}(\mathbb{X}; \mathbf{m})$ converges to a multiple of the inverse of the covariance matrix at the $\sqrt{n}$ rate, if $\mathbf{m}$ is a $\sqrt{n}$-convergent location estimator and the data stem from a linear transformation of a reflection and permutation invariant distribution having finite second moments. The statistical theory of the Oja rank covariance matrix ORCM is treated in Visuri, Ollila, Koivunen, Möttönen, and Oja (2003) and Ollila, Croux, and Oja (2004).

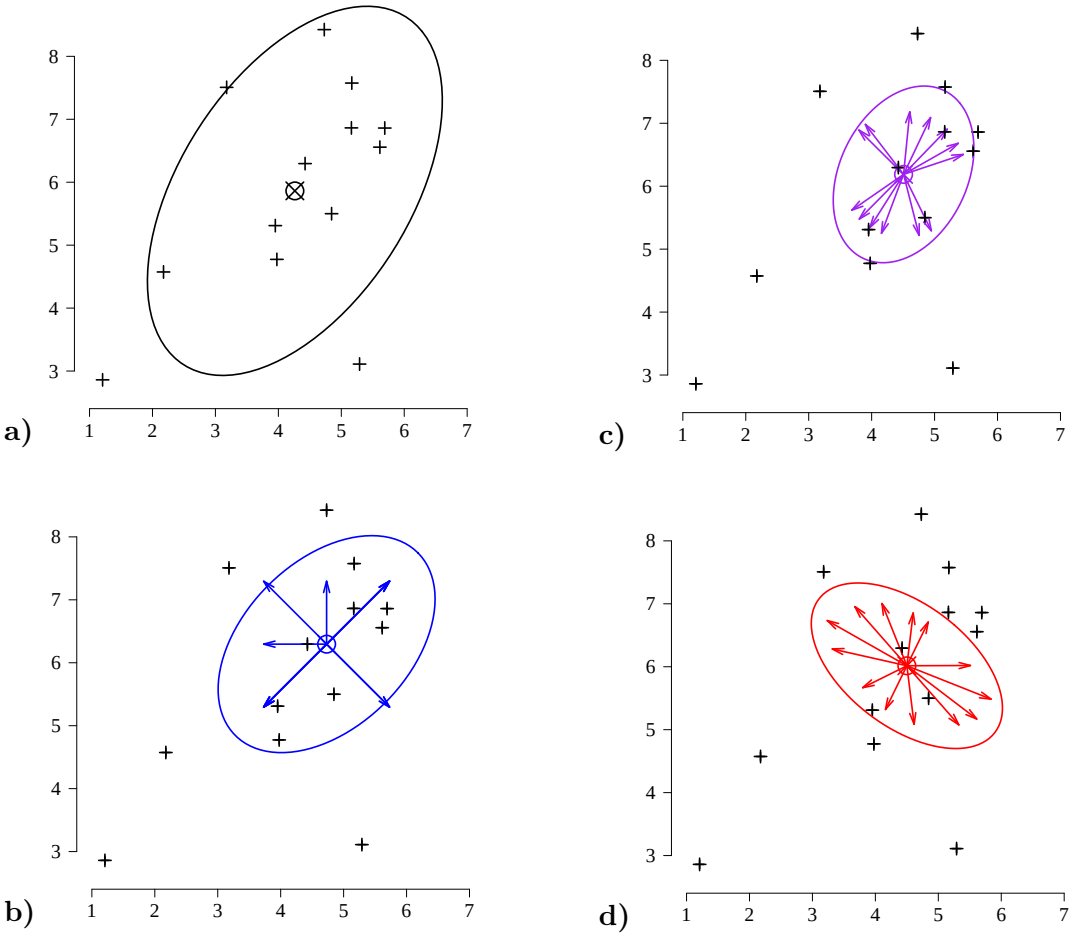

Figure 5: Data sample together with **a)** mean and sample covariance matrix, **b)** marginal median, marginal signs and MSCM, **c)** spatial mean, spatial signs and SSCM and **d)** Oja median, Oja signs and OSCM.

A further consequence of this remarkable property is that Oja sign and rank matrices provide easily-obtained, positive definite, consistent estimates for the covariance matrix up to scale, which is a significant advantage over the other sign and rank covariance matrices. Multivariate data analysis is primarily aimed at analyzing the dependencies and interactions between variables. For multivariate methods such as correlation, canonical correlation analysis, principal component analysis, or factor analysis, it is fully sufficient to know the covariance matrix only up to scale. In particular, the OSCM and the ORCM directly estimate a multiple of the concentration matrix or precision matrix, i.e., the inverse covariance matrix, which plays an important role in graphical models. The application of OSCM in this context is examined in Vogel, Köllmann, and Fried (2008) and Vogel and Fried (2008). Due to their (inverse proportional) affine equivariance, the theory developed in Vogel and Fried (2011) also applies to the OSCM and the ORCM.

Furthermore Ollila *et al.* (2003) and Ollila *et al.* (2004) also derive the limiting distributions of OSCM and ORCM in the elliptical model. Contrary to the other non-parametric covariance matrices based on marginal and spatials signs and ranks, the OSCM and the ORCM are

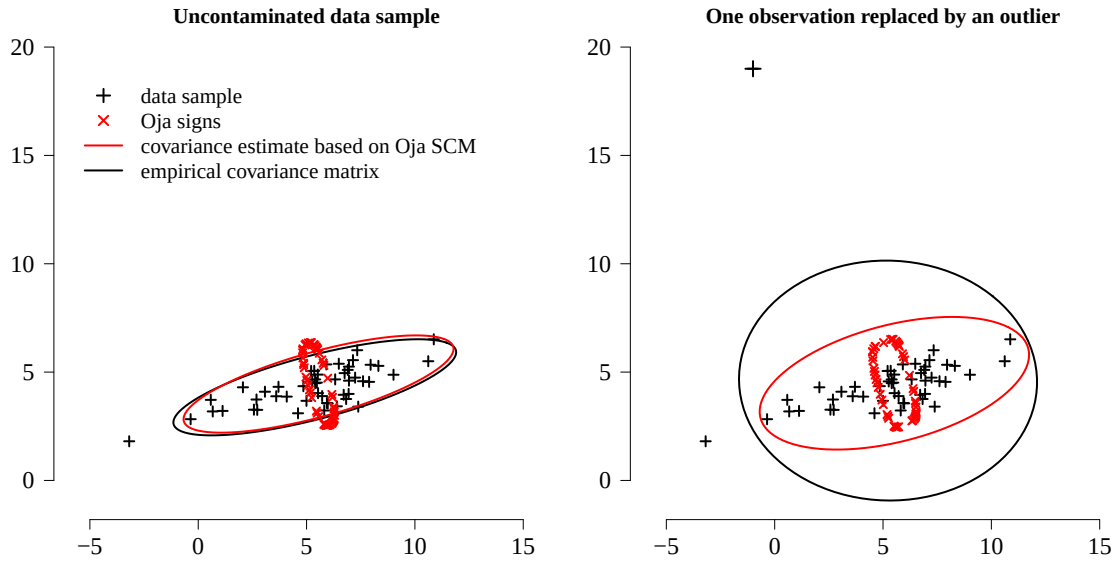

Figure 6: The effect of an outlier on the empirical covariance matrix (black) and on the OSCM-based covariance estimate (red). In the right-hand plot, one observation has been replaced by an outlier at $(-1, 19)$.

not invariant w.r.t. the elliptical generator within the elliptical model. Their asymptotic distribution depends on the tail behavior of the population distribution. But, and this is also in contrast to marginal and spatial non-parametrics, Oja sign and rank matrices are very efficient at the normal model in low dimensions. Their performance almost equals that of the empirical covariance matrix, the normal maximum likelihood estimator. They outperform the empirical covariance matrix ECM at heavier tailed distributions, but are generally less efficient at light tails. They maintain the good efficiency at small sample sizes n and small dimensions k, which is not true for many robust scatter estimators.

The gain in using the OSCM or the ORCM instead of the sample covariance matrix lies in their higher robustness, which comes at practically no loss in efficiency (but unfortunately at a large increase in computing time). However, similar to the Oja median, the Oja sign and rank matrices do not qualify as globally robust estimators in the sense of having a breakdown point near 1/2. They require first moments; their influence functions are unbounded but grow linearly (instead of quadratic as in the case of the sample covariance matrix), and the asymptotic breakdown point is zero. Very few misplaced observations suffice to let the bias of the estimators become arbitrarily large, but the bias is significantly smaller than that of the sample covariance matrix. This is visualized in Figure 6: The left-hand plot shows a data sample of size $n = 50$ drawn from a two-dimensional normal distribution. The black ellipse represents the empirical covariance matrix, the red ellipse the OSCM-based covariance estimate. In the right-hand plot, one observation has been replaced by an outlier at $(-1, 19)$, and the ellipses represent the corresponding estimates of the contaminated data set.

2.6. The one- and C-sample location tests based on Oja signs and ranks

The notions of multivariate location and spread together with the corresponding concepts of sign and rank allow the general derivation of multivariate inference methods. The general idea

is there to view the signs, signed ranks and ranks as scores, which replace the observations in the classical multivariate procedures. In principle, robust counterparts of any multivariate method can be derived this way. We demonstrate here the multivariate one-sample and C-sample location tests. For that purpose, denote for a sample point $\mathbf{x}_i$ the corresponding score $\mathbf{s}(\mathbf{x}_i; \mathbf{m})$, where $\mathbf{m}$ is an optional location w.r.t. to which the score is computed.

The one-sample tests

Assume $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a sample of size n from a k-variate symmetric distribution F with symmetry center $\boldsymbol{\mu}$. We are interested in testing the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

Denote $\bar{\mathbf{s}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \boldsymbol{\mu}_0)$ as the average of the score values under the null hypothesis and $\boldsymbol{\Sigma}_\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \boldsymbol{\mu}_0)\mathbf{s}(\mathbf{x}_i, \boldsymbol{\mu}_0)^\top$. The test statistic is then

$$Q = n\,\bar{\mathbf{s}}^\top \boldsymbol{\Sigma}_\mathbf{s}^{-1} \bar{\mathbf{s}}\,.$$

Using Oja signs or Oja signed ranks as scores, this yields a straightforward extension of Hotelling's classical one-sample T^2-test. The test is invariant under affine transformations and asymptotically distribution-free and has a limiting χ^2_k distribution. Test decisions can also be based on permutation principles by randomly changing the signs of the scores. The tests are described in detail in Hettmansperger, Nyblom, and Oja (1994) and Hettmansperger, Möttönen, and Oja (1997). Similar tests can also be naturally constructed using marginal or spatial signs and signed ranks. See Puri and Sen (1971) and Oja (2010) for details.

The C-sample tests

Let $\mathbb{X}_c = (\mathbf{x}_{c,1}, \ldots, \mathbf{x}_{c,n_c})$, $c = 1, \ldots, C$, correspond to k-variate samples coming from $C \geq 2$ groups having distributions F_c that differ only in location parameters $\boldsymbol{\mu}_c$, $c = 1, \ldots, C$. The null hypothesis is $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_C$, i.e., the C groups have the same location.

Denote $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_C)$ as the combined sample and $n = \sum_{i=1}^{C} n_i$. Then $\bar{\mathbf{s}}_c = \frac{1}{n_c}\sum_{j=1}^{n_c} \mathbf{s}(\mathbf{x}_{c,j})$, $c = 1, \ldots, C$, is the average score value of group c computed w.r.t. to the location of the combined sample. Similarly, $\boldsymbol{\Sigma}_\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i)\mathbf{s}(\mathbf{x}_i)^\top$ is computed for the combined groups.

The test statistic is obtained as

$$Q = \sum_{c=1}^{C} n_c \bar{\mathbf{s}}_c^\top \boldsymbol{\Sigma}_\mathbf{s}^{-1} \bar{\mathbf{s}}_c\,.$$

When Oja signs and Oja ranks are used as scores, C-sample tests for multivariate location are obtained that are asymptotically distribution-free and affine invariant. The limiting distribution of the test statistic is $\chi^2_{k(C-1)}$, but p values can be obtained by permuting observations between the groups. The two tests are described in Hettmansperger and Oja (1994); Hettmansperger, Möttönnen, and Oja (1998). Similar tests based on other concepts of signs and ranks are described also in Puri and Sen (1971) and Oja (2010).

3. Description of the algorithms

The package **OjaNP** contains four different algorithms to calculate the Oja median. Two exact algorithms and two approximate algorithms. The first exact algorithm was developed

in Ronkainen, Oja, and Orponen (2003) as well as one of the approximate algorithms. The second exact algorithm (Mosler and Pokotylo 2015) is based on the first one: It accelerates the computation considerably by introducing bounds to the region of search. The numerical calculation is a non-trivial problem which consumes enormous calculation resources and hence, the exact algorithms are limited to small data situations only and, as a consequence, approximate algorithms are needed. These offer parameters to regulate the speed vs. accuracy trade-off, and the user has to decide from case to case which algorithm to choose with which tuning parameters. In Section 4 we will give an overview over the several options and their effect onto the calculation precision and time. Before that, we are going to describe the four algorithms.

3.1. Exact algorithm

Ronkainen *et al.* (2003) implemented the ideas from Niinimaa, Oja, and Nyblom (1992) and generalized them into higher dimensions based on the result described in Hettmansperger, Möttönen, and Oja (1999), whereby the vertices of the Oja median set are always located on intersections of hyperplanes that are spanned by data points.

Ronkainen *et al.* (2003) constructed a Las Vegas algorithm as follows. (This is a simplified version; a more detailed description can be found in the original paper.)

1. Let $\mathcal{H}$ be the set of all $(k-1)$-dimensional hyperplanes spanned by the points in $\mathbb{X}$.
2. Take the data point $\mathbf{x}_c \in \mathbb{X}$ closest to the mean as an initial candidate point.
3. Sample $k-1$ hyperplanes out of $\mathcal{H}$ such that the candidate point is on their intersection L.
4. Calculate the Oja depth of each intersection point between L and the hyperplanes in $\mathcal{H}$.
5. Take the point $\mathbf{x}_c^*$ with the highest Oja depth as next candidate point for the Oja median.
6. Repeat steps 3 to 5 until no improvement in the objective function is possible (or latest after n repetitions).
7. The result for the exact Oja median is the last candidate point $\mathbf{x}_c^*$.

Ronkainen *et al.* (2003) focused on computational stability rather than efficiency. In case a candidate point is a data point, there are $k-1$ possible intersection hyperplanes L. Instead of only following the one determined by the gradient of the objective function, the algorithm tries all possible ones.

This algorithm finds just one of the vertices of the median set. While searching for the median, the algorithm may pass through several vertices of the median set, although it is not guaranteed that it visits all of them. The reason is that on step 5 only the first of possibly two points having highest Oja depth is taken as $\mathbf{x}_c^*$. However, in case of a non-unique median, there exist two such points lying on an edge of the median set. To deliver all vertices of the median set, the algorithm can be modified as follows: It has to store both points as vertices and, in addition, check all lines passing through them.

The algorithm is implemented in C++ and was initially published as stand-alone software on the personal webpage of Tommi Ronkainen but cannot be accessed anymore in that form. The implementation was modified such that it can be used directly from R.

3.2. Exact bounded algorithm

Based on the exact algorithm of Ronkainen *et al.* (2003), Mosler and Pokotylo (2015) developed a faster exact algorithm. This algorithm uses the centered rank functions to build bounded regions which contain the median. The negative rank function $-\mathbf{ornk}_{\mathbb{X}}(\mathbf{x})$ is a vector that points in a direction of ascent of the depth function. It defines a hyperplane through $\mathbf{x}$, on the positive side of which the Oja median is found. The halfspaces defined by the negative rank function are used to build a bounded region that contains the median. In this algorithm, these halfspaces are selected in an iterative way and the further search is restricted to their intersection. The hyperplanes bordering such a search region will be called bounding hyperplanes or simply bounds.

The steps of the algorithm are as follows. (A more detailed description can be found in the original paper):

1. Let $\mathcal{H}$ be the set of all hyperplanes spanned by the points in $\mathbb{X}$.

2. Create the initial rectangular bounded region $\mathbf{B}$, limited by hyperplanes that are perpendicular to the coordinate axes and go through the maximal and minimal coordinates of the data points on these axes.

3. Iteratively reduce the bounded region $\mathbf{B}$ by adding hyperplanes that go through a properly chosen central point of the region and have their normal vectors equal to the corresponding negative rank function. Specifically, the mean value of the bounds' intersection points is selected as a central point in our implementation. The bounds of $\mathbf{B}$ are cut off by newly added hyperplanes.

4. The region is reduced until the desired final volume of the bounded region is reached.

5. Add the bounds from $\mathbf{B}$ to $\mathcal{H}$.

   - At the first iteration: Take a random initial line L on the border of $\mathbf{B}$.
   - At further iterations: Sample $k-1$ hyperplanes out of $\mathcal{H}$ in that way, that the candidate point is on their intersection line L.

6. Calculate the Oja depth on each intersection point between L and hyperplanes in $\mathcal{H}$ that lies in the bounded region.

7. Take the point $\mathbf{x}_c^*$ with the highest Oja depth as next candidate point for the Oja median.

8. Repeat steps 6 to 8 until no improvement in the objective function is possible (or latest after n repetitions).

9. The result for the exact Oja median is the last candidate point $\mathbf{x}_c^*$.

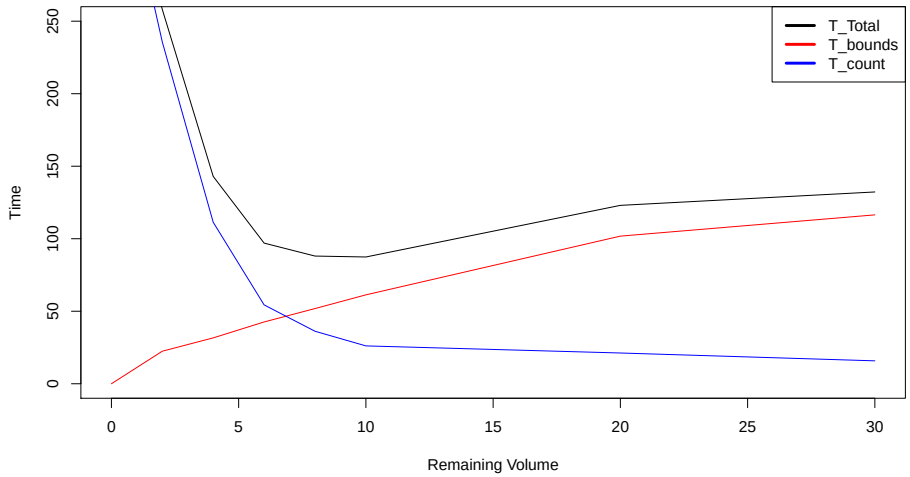

Figure 7: Dependence of calculation time on the size of the bounded region; time needed for bounding (T_{bounds}), for minimizing (T_{count}), and total time (T_{total}). For comparison: total time of the first exact algorithm is about 340 seconds.

The bounded regions reduce the complexity of the searching procedure by reducing the number of hyperplanes that cross the searching lines as well as the number of their intersections to be considered in the minimization procedure.

The algorithm is driven by the desired final volume of the bounded region, which is the volume of the minimal rectangle containing the region and having edges parallel to the coordinate axes. As this parameter is reduced, the time needed to build the bounded regions (T_{bounds}) increases, while the minimization time (T_{count}) decreases along with the number of hyperplanes and their intersections; see Figure 7. The total computing time (T_{total}) decreases rapidly with the volume, but then slowly grows again. Beyond some point the procedure becomes less efficient. For comparison, the original algorithm needs a total time T_{total} of ca. 340 seconds in this example. It appears that the fastest computation is obtained if bounds are imposed until the volume of the bounded region ranges around 10^{-8} of the original volume. Note that the bounds may cut off some of the vertices of the median set. Moreover, if the central point of the bounded region lies in the median set on step 3, its negative rank function is zero, and this point is directly returned as a median, as on Figure 9.

Due to limitations in computing memory and long calculation time, the exact algorithm and its bounded version are only able to calculate the Oja median for small data sets in low dimensions. For example, the calculation of the median in a data set of size 100×5 needs 12 GB RAM. Therefore, approximate algorithms are needed.

Obviously, the bounded algorithm can be stopped at any iteration and some mean value of the last bounded region be taken as an approximation of the Oja median. However, unlike the approximate algorithms presented below, this approach requires the calculation of all hyperplanes. Due to its high computational requirements it is less suited as an approximate algorithm for big data sets in high dimensions.

For calculation speed reasons, also this algorithm is implemented in C++.

3.3. Grid-based algorithm

The third algorithm, which calculates an approximation to the Oja median, was also proposed in Ronkainen *et al.* (2003). Technically it is a Monte Carlo algorithm. The algorithm lays a uniform grid over the data set. At each grid point a test is performed whether the point is a possible candidate. The amount of candidate points is reduced as long as only one grid point is left. This point is afterwards the center for a smaller but denser grid, where again each grid point is tested. The algorithm stops when the distance between two grid points gets smaller than a predefined parameter. A second tuning parameter is the significance level of the point tests. The steps of the algorithm are:

1. Create a grid $\mathcal{G}$ with equidistant knot distance h, covering the whole data set.
2. Choose randomly a set of hyperplanes, build the test statistic and test each of the grid knots in $\mathcal{G}$ whether it is an Oja median.
3. Remove those grid knots which have been tested not to be an Oja median.
4. If there is more than one knot left, sample additional hyperplanes and repeat the test for the remaining knots.
5. Repeat these steps until only one grid knot is left over. If the last test removes all remaining ones, take the last set.
6. Build a new grid around the last remaining old grid knot with equidistant knot distance $h/2$.
7. Repeat all these steps until the grid distance reaches a predefined threshold.
8. The last point is taken as the Oja median.

For further details, especially about the testing procedure, we refer to the original paper Ronkainen *et al.* (2003). It may happen that there is continually more than one point left on step 5. In this case sampling of the additional hyperplanes may not help and the algorithm hangs. We restrict the algorithm to 5000 iterations on step 5, after which the grid point with the best test statistic is passed to step 6 and the number of iterations is reduced to 100. We repeat until the grid threshold is reached and return the grid point with the best test statistic as an Oja median approximation.

The grid-based algorithm is part of the initially stand-alone tool that contains also the implementation of the first exact algorithm. It is as well written in C++ to speed up the computational burden.

3.4. Evolutionary algorithm

The fourth algorithm to calculate the Oja median is an evolutionary algorithm, which is based on mutations of the latest candidate points. It was developed by the Department of Computer Science, Efficient Algorithms and Complexity Theory at the TU Dortmund, but has not been published before. The algorithm works as follows:

1. Set the level of initial mutation variance σ_0^2.

2. Take 10 randomly chosen observations from $\mathbb{X}$.

3. Evaluate the objective function for all these points and take the minimum as starting candidate point η.

4. Choose $k+1$ random numbers $x_1, \ldots, x_k, l$ from a $N(0, \sigma_0^2)$ distribution and calculate the k-variate mutation vector

$$\nu = \frac{|l|}{\sqrt{x_1^2 + \cdots + x_k^2}} (x_1, \ldots, x_k)^\top .$$

   The mutation vector ν has a normally distributed length with given variance σ_0^2 and uniformly distributed direction.

5. Calculate m mutation points $\eta_i = \eta + \nu_i$ for $i = 1, \ldots, m$ from the last candidate point η.

6. Calculate the ratio r how often the objective function is bigger at the mutations than at η.

7. If $r > 0.2$ then $\sigma_0^2 \cdot \kappa$ else $\sigma_0^2 \cdot \frac{1}{\kappa}$ for some $\kappa > 1$.

8. Choose as a new candidate point the mutation with the smallest objective function value.

9. Repeat steps 4 to 8 until the variance for the next mutation drops under a predefined value s.

10. If the algorithm has not terminated after n_t steps, stop the calculation.

Step 7 controls the dynamic of the mutation. If at more than 20% of the mutations the objective function has a smaller value than at the last candidate point, the algorithm increases the variability of the mutation; hence the search area is enlarged. Step 10 ensures that the algorithm terminates in any case.

As the other implementations also this algorithm is implemented in C++.

3.5. Other algorithms for the Oja median in R

There are other implementations of algorithms for the Oja median available in R, but they are mostly restricted to two dimensions.

The function `med` in the package **depth** (Genest, Masse, and Plante 2019) uses the Fortran code of Niinimaa *et al.* (1992) and is restricted to the bivariate case.

Another method to compute the Oja median was suggested by Roger Koenker on R-help on 2003-08-16 (see https://hypatia.math.ethz.ch/pipermail/r-help/2003-August/037702.html using the **quantreg** package; Koenker 2019):

```
R> oja.median <- function(x) {
+    n <- dim(x)[1]
+    A <- matrix(rep(1:n, n), n)
```

```
+    i <- A[col(A) < row(A)]
+    j <- A[n + 1. - col(A) > row(A)]
+    xx <- cbind(x[i,  ], x[j,  ])
+    y <- xx[, 1] * xx[, 4] - xx[, 2] * xx[, 3]
+    z1 <- (xx[, 4] - xx[, 2])
+    z2 <-  -(xx[, 3] - xx[, 1])
+    return(quantreg::rq(y ~ cbind(z1, z2) - 1)$coef)
+  }
```

4. The R package OjaNP

The main purpose of the **OjaNP** package is to provide users with the possibility to compute the Oja median. The package includes, however, also other useful functions. The main functions of the package are visualized in Figure 8.

Most of the function names are self-explanatory. For details about them we refer to the corresponding help pages. In the following we will first explain the function `ojaMedian` and its options in detail. Then we demonstrate the use of some of the functions with a small but illustrative data set, which is also contained in the package.

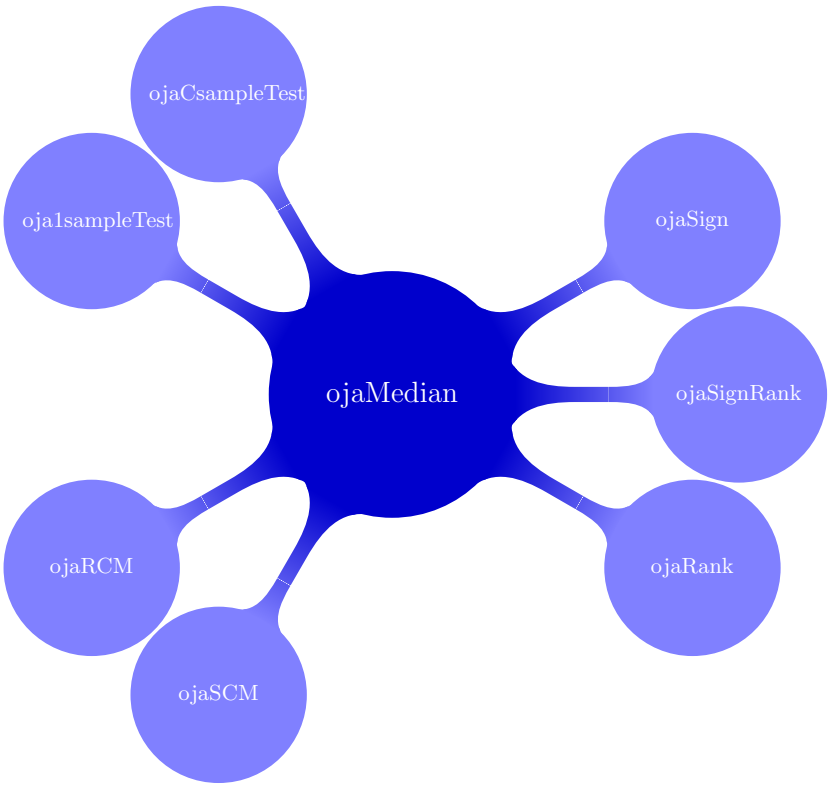

Figure 8: The main functions in package **OjaNP**.

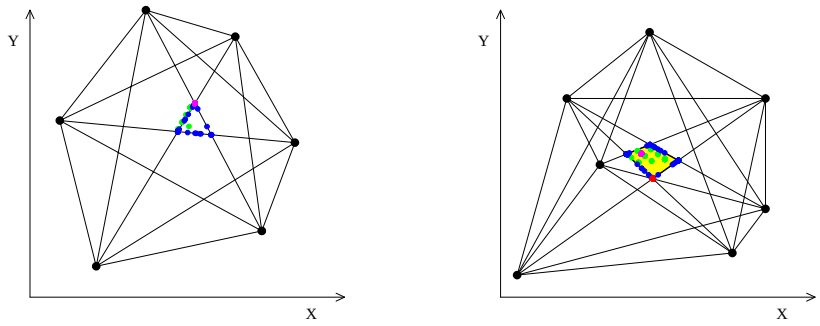

Figure 9: Example plot for all four algorithms: exact algorithms (red), evolutionary algorithm (blue), grid algorithm (green). The convex median set is marked with yellow.

4.1. The computation of the Oja median in OjaNP

The main function of **OjaNP** is

```
ojaMedian(X, alg = "evolutionary", sp = 1, na.action = na.fail,
  control = ojaMedianControl(...), ...)
```

The user can choose via the `alg` option between four algorithms to calculate the Oja median. Furthermore, we have an option to calculate the Oja median repeatedly and average these results in order to receive less varying results. The amount of repetitions can be controlled with the `sp` parameter.

In what follows we are going to explain the different parameters which control the flow of the different algorithms in detail and give insights how to choose the parameters in a given data situation. The default algorithm of the `ojaMedian` function is the evolutionary algorithm.

The evolutionary algorithm relaxes the affine equivariance property of the Oja median. In order to restore it we first perform a scatter matrix transformation to obtain an invariant coordinate system (implemented in **ICS**; Nordhausen, Oja, and Tyler 2008), apply the algorithm to the transformed data and re-transform afterwards. That way we restore the affine equivariance for this implementation.

Figure 9 shows the exemplary outcome of the four implemented algorithms in simple data situations for the unique (left) and non-unique (right) case. We have chosen simple data situations with 6 and 7 data points, run the different algorithms 500 times and plotted the outcome into the figure.

In the left part of Figure 9 we have a data set that has a unique Oja median. This one is correctly determined by the exact implementations (red), whereas both approximate algorithms have a systematic behavior which does not differ strongly from the non-unique case in the right side of the graphic. The evolutionary algorithm (blue) determines the Oja median always along lines of intersection with the result of a bordered area. The right-hand side of Figure 9 exhibits data that have a non-unique Oja median. Here the two exact algorithms find a vertex (red) of the median set, while the evolutionary algorithm (blue) yields any point of the border of the median set, that is the area with the lowest Oja depth (yellow). The grid algorithm, however, terminates in this case usually within the convex median set.

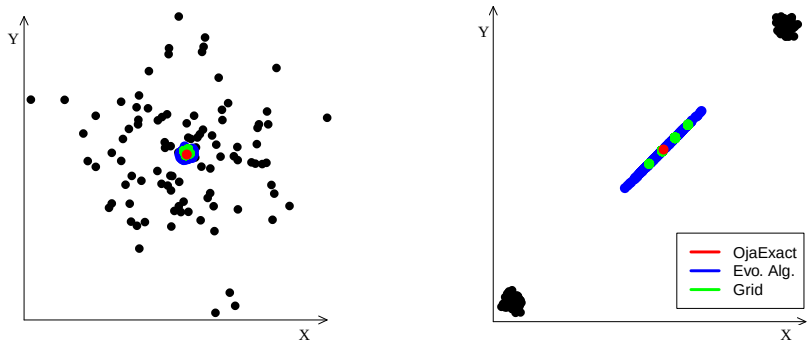

Figure 10: Example plot for all four algorithms: exact algorithms (red), evolutionary algorithm (blue), grid algorithm (green).

In a next step we are going to analyze the outcome of the algorithms in more complex data situations. The first typical data situation is a multivariate normal distributed data cloud and we calculate the Oja median with the exact (red), the grid (green) and the evolutionary algorithm (blue). As we can see the evolutionary algorithm has the biggest variation within the results, but we have to keep in mind that the default setting of our function is tuned to calculate fast results. In the next paragraph we will discuss how to get more precise results in trade-off for calculation time. For most applications the faster, less accurate settings appear to be preferable.

The second data set of interest consists of two data clouds, which are far away from each other. The motivation for that is to see whether the algorithms take this into account or if they get stuck in one cloud. As you can see in the right part of the figure, that all included algorithms are capable of calculating the Oja median to be in the center between both data clouds.

Let us now compare the runtime of the two approximate algorithms depending on the dimension and size of the data in the bivariate case. We do not perform a runtime analysis for the exact algorithms. Their performance depends strongly on the computer used (particularly memory), much more so than the approximate ones.

On an average computer, bivariate problems up to 1,000 observations are solvable in acceptable time, but for higher dimensions and sizes this decision has to be made from case to case. For example in the seven-dimensional case with tens of observations it takes minutes for the exact algorithm to find the solution. The biggest limitation of the exact algorithm is the memory required to store all hyperplanes. Even if we would allow infinite calculation time, the algorithm would still not be able to calculate the exact Oja median in more complex data situations because it cannot pre-calculate and access the total amount of hyperplanes, and hence we are facing a corresponding address space problem. This also applies to the exact bounded algorithm. Compared with the original exact algorithm, the bounded one finds the solution approximately two to five times faster.

In order to analyze the runtime for different dimensions we simulated 10,000 multivariate normal distributed random numbers (with $\mu = \mathbf{0}$, $\mathbf{\Sigma} = \mathbf{I}$). The approximate grid algorithm (solid green line) is only able to calculate the Oja median up to 5 dimensions in acceptable time for this amount of data; this is why we did not take higher dimension situations into

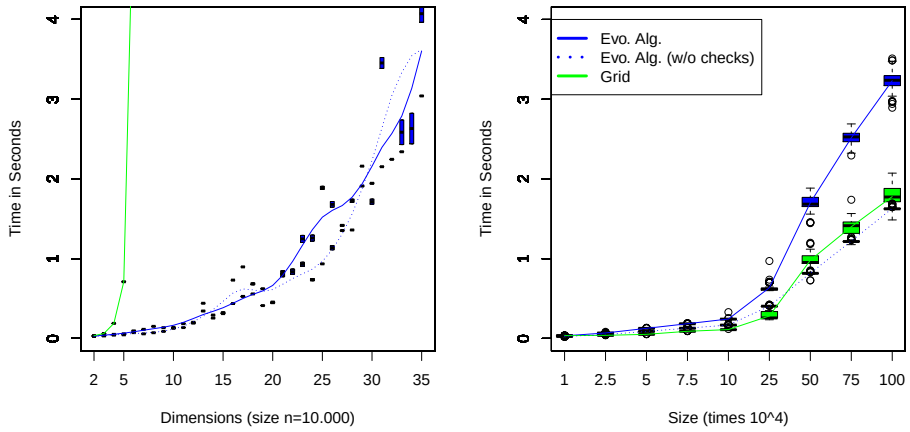

Figure 11: Runtimes for the approximate algorithms: the approximate grid algorithm (solid green line), the evolutionary algorithm (solid blue line), the raw method without transformations or validation checks (dotted blue line).

consideration. The evolutionary algorithm (solid blue) can calculate in the same time the Oja median of a 35-dimensional data set, and even higher dimensional problems are solvable. Since the ICS step (and there especially the data validation checks) takes a lot of computational time we implemented also a raw command to access the algorithms directly without transformations or validation checks (dotted blue line). In the analysis of different dimensions the raw algorithm does not bring huge advantages, but in the runtime analysis concerning the size of the data in the bivariate case (right-hand side of Figure 11) we can detect a huge advantage for more than 10^5 observation. The raw algorithm is even able to calculate the Oja median for sample size $5 \cdot 10^7$. Hence, our advice in time-critical situations with high sample size is to use the raw method. If the affine equivariant property is still required, we advise to perform beforehand ICS separately without performing the included data validation steps.

The evolutionary algorithm has many tuning parameters, some of which control its accuracy. As we have seen in Figure 10, the default settings for these tuning parameters are preset to deliver fast results. In trade-off for higher computational time, the user can adjust the settings to obtain a more precise algorithm. The key parameters are `useAllSubsets`, `nSubsetsUsed`, and `sigmaLog10Dec`. The latter is the main abort criterion of the algorithm. It forces the algorithm to stop if the logarithmized initial variance differs more then the value of `sigmaLog10Dec` from the actual logarithmized variance. In other words, when the variance of the mutation vector is getting small enough, the algorithm stops.

The settings for `useAllSubsets` and `nSubsetsUsed` control how many spanned hyperplanes should be taken into account during the calculation of the Oja median. Since the total amount of possible hyperplanes could be huge (it is $\binom{n}{k}$ for n observations in the k-variate case), the flag for `useAllSubsets` should be used carefully. It is more advisable to control this with the argument `nSubsetsUsed`. Raise this value together with `sigmaLog10Dec` for more precise values, lower them for faster results.

The dynamics of the evolutionary algorithm are controlled via the parameters `sigmaInit`, `sigmaAda` and `adaFactor`. All these take control over the variance adjustments of the mutation vector. The parameter `sigmaInit` sets the initial variance of the mutation, the settings for `sigmaAda` control after how many mutation steps the mutation variance is adjusted and

`adaFactor` defines how the variance is adjusted. In most cases the default settings work nicely.

To conclude this section, we would like to mention that there are many R packages available to compute various medians: The **depth** package (Genest *et al.* 2019) contains Tukey's median, Liu's median, spatial, marginal, and also the Oja median. The authors use the early Fortran implementation by Niinimaa *et al.* (1992) that is only able to calculate the bivariate Oja median and that is much slower than the here presented implementations. Other packages containing different algorithms to compute the spatial median are, e.g., **ICSNP** (Nordhausen, Sirkiä, Oja, and Tyler 2018b), **MNM** (Nordhausen, Möttönen, and Oja 2018a; Nordhausen and Oja 2011) and **pcaPP** (Filzmoser, Fritz, and Kalcher 2018). Some of these packages also offer functions for multivariate signs and ranks and methods based upon them. Consequently, the aforementioned packages provide other multivariate medians that follow different generalizations, as described in Section 2.

4.2. Short demonstation of the package's main function

In this section we will demonstrate the main functions of the package using the `biochem` data set. This data set consists of the amounts of two chemical components in the brain measured at 22 mice. Ten of the mice belong to a control group and twelve received a drug.

This is a very basic example just to demonstrate the basic use of the main functions. For details about the functions see also their corresponding help pages.

We first load the package and the data and create data objects for easier handling as well as fixing the random seed for reproducibility.

```
R> library("OjaNP")
R> data("biochem", package = "OjaNP")
R> set.seed(1)
R> X <- as.matrix(biochem[, 1:2])
R> GROUP <- biochem$group
R> GRlabel <- as.numeric(GROUP)
```

Next we compute the bivariate Oja median of the two components using the default evolutionary algorithm.

```
R> OMev <- ojaMedian(X)
R> OMev

comp.1 comp.2
 1.150  0.425
```

As this toy data set is quite small, it is no problem to compute here also the exact Oja median using either of the algorithms provided.

```
R> OMex <- ojaMedian(X, alg = "exact")
R> OMex

   comp.1    comp.2
1.1515385 0.4269231
```

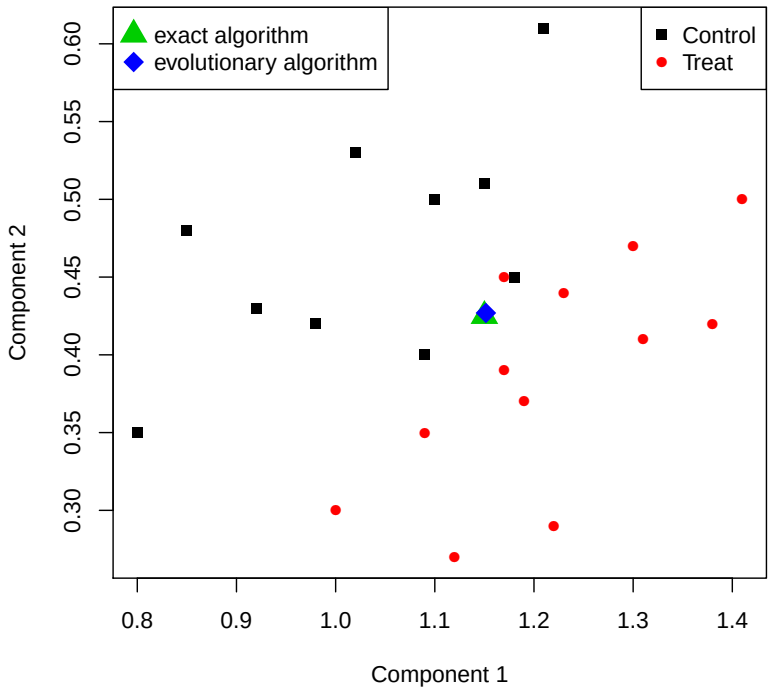

Figure 12: Visualization of the `biochem` data and of the Oja median when computed using the exact algorithm and the evolutionary algorithm.

```
R> OMbo <- ojaMedian(X, alg = "bounded_exact")
R> OMbo

   comp.1    comp.2
1.1515385 0.4269231
```

As can be seen, the difference between the exact and the approximate estimate is rather small, which is also visualized in Figure 12, produced by the following code:

```
R> plot(X[, 1], X[, 2], col = GRlabel, pch = GRlabel + 14,
+    xlab = "Component 1", ylab = "Component 2")
R> points(OMev[1], OMev[2], cex = 2, pch = 17, col = 3)
R> points(OMex[1], OMex[2], cex = 2, pch = 18, col = 4)
R> legend("topright", legend = levels(GROUP), col = 1:2, pch = 15:16)
R> legend("topleft", legend = c("exact algorithm", "evolutionary algorithm"),
+    col = 3:4, pch = 17:18, pt.cex = 2)
```

Next we look at the Oja signs of the data.

```
R> head(ojaSign(X))

             [,1]         [,2]
[1,]  0.011223776  0.114446387
[2,] -0.063951049  0.019644522
```

```
[3,] -0.056363636 -0.036818182
[4,] -0.060769231  0.064644522
[5,] -0.062587413 -0.005810023
[6,] -0.017272727  0.119545455
```

These signs are computed with respect to the Oja median. But the `ojaSign` function has several options to compute them also with respect to some other location. For example the vector of marginal medians can be specified as

```
R> head(ojaSign(X, center = "compMedian"))
```

```
              [,1]          [,2]
[1,]  0.010681818  0.1145454545
[2,] -0.063409091  0.0195454545
[3,] -0.059772727 -0.0240909091
[4,] -0.058863636  0.0781818182
[5,] -0.063409091  0.0004545455
[6,] -0.015681818  0.1200000000
```

The Oja signs covariance matrix can be similarly obtained as

```
R> ojaSCM(X)
```

```
               comp.1        comp.2
comp.1  0.0021343943 -0.0003989038
comp.2 -0.0003989038  0.0075945852
```

Next we test whether the Oja median of the control group corresponds to the value `c(1, 0.5)`.

```
R> oja1sampleTest(X[1:10, ], mu = c(1, 0.5))
```

```
        OJA 1 SAMPLE SIGN TEST

data:  X[1:10, ]
Q.S = 3.3745, df = 2, p-value = 0.185
alternative hypothesis: true location is not equal to c(1,0.5)
```

The test decision is here based on the limiting distribution. The sample size is rather small in this example, and one may prefer to use permutation p values. Using the `method` argument of the function, p values can be computed by permutation.

```
R> oja1sampleTest(X[1:10, ], mu = c(1, 0.5), method = "permutation")
```

```
        OJA 1 SAMPLE SIGN TEST

data:  X[1:10, ]
Q.S = 3.3745, replications = 1000, p-value = 0.203
alternative hypothesis: true location is not equal to c(1,0.5)
```

To demonstrate the C-sample location test, we use the rank test to test whether the two groups differ and want to base the decision on permutation principles.

```
R> ojaCsampleTest(X ~ GROUP, scores = "rank", method = "permutation")

        OJA C SAMPLE RANK TEST

data:  X by GROUP
Q.R = 15.17, permutations = 1000, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0,0)
```

4.3. A more complex data example

For a more complex example we choose the data set `LASERI` that is contained in the **ICSNP** package. The data set contains for 223 individuals the cardiovascular responses to a passive head-up tilt. For this purpose several haemodynamic variables are measured before, during and after the tilt. One question of interest here, e.g., is if the haemodynamic system 5 minutes after the tilt has reached already pre-tilt levels.

For demonstration purpose we consider the three variables HR (heart rate), CO (cardiac output) and SVRI (systemic vascular resistance index). The testing problem is then if the 3-variate difference between pre-tilt and post-tilt values has location zero. The differences of interest are saved in the data set as variables `HRT1T4`, `COT1T4` and `SVRIT1T4`. For more details about the data set see its help page.

After preparing the data we first compute the Oja median of the differences using three different algorithms and time the methods.

```
R> library("OjaNP")
R> data("LASERI", package = "ICSNP")
R> set.seed(1)
R> X <- as.matrix(subset(LASERI, select = c(HRT1T4, COT1T4, SVRIT1T4)))
R> system.time(OMev <- ojaMedian(X))

   user  system elapsed
  0.038   0.000   0.043

R> OMev

      HRT1T4       COT1T4     SVRIT1T4
   3.4775306    0.4529893 -208.2378473

R> system.time(OMex <- ojaMedian(X, alg = "exact"))

   user  system elapsed
 82.405   0.176  82.588

R> OMex
```

```
     HRT1T4       COT1T4     SVRIT1T4
  3.4179008    0.4152541 -198.9544360
```

```
R> system.time(OMbo <- ojaMedian(X, alg = "bounded_exact"))
```

```
    user  system elapsed
  16.391   0.000  16.394
```

```
R> OMbo
```

```
     HRT1T4       COT1T4     SVRIT1T4
  3.4179008    0.4152541 -198.9544360
```

Clearly the default evolutionary algorithm is the fastest (0.04 seconds) but the results differ from the exact algorithms where as expected the bounded algorithm is much faster than the other exact algorithm (16.4 vs. 82.4 seconds).

The data together with the three estimates is visualized in Figure 13 which is however better used online interactively (https://plot.ly/~fischuu/9). We used the **plotly** (Sievert, Parmer, Hocking, Chamberlain, Ram, Corvellec, and Despouy 2019) package for the visualization.

```
R> library("plotly")
R> plotThis <- rbind(X, OMev, OMex)
R> useColors <- factor(c(SEX, "EV", "EX"),
+    labels = c("Female", "Male", "Evo", "Exact"))
R> plot_ly(as.data.frame(plotThis), x = ~ HRT1T4, y = ~ COT1T4,
+    z = ~ SVRIT1T4, color = useColors,
+    colors = c("#BF382A", "#0C4B8E", "green", "yellow"),
+    type = "scatter3d", mode = "markers", marker = list(size = 2)) %>%
+      layout(scene = list(xaxis = list(title = "HRT"),
+        yaxis = list(title = "COT"), zaxis = list(title = "SVRIT")))
```

Next we test for each sex separately if the Oja median of the differences is significantly different from zero using Oja ranks.

```
R> SEX <- LASERI$Sex
R> oja1sampleTest(X[SEX == "Male", ], scores = "rank")
```

```
PLEASE NOTE: You have requested to compute 1414808
hyperplanes in R^3. This may take a while.

OJA 1 SAMPLE SIGNED RANK TEST

data:  X[SEX == "Male", ]
Q.R = 73.11, df = 3, p-value = 8.882e-16
alternative hypothesis: true location is not equal to c(0,0,0)
```

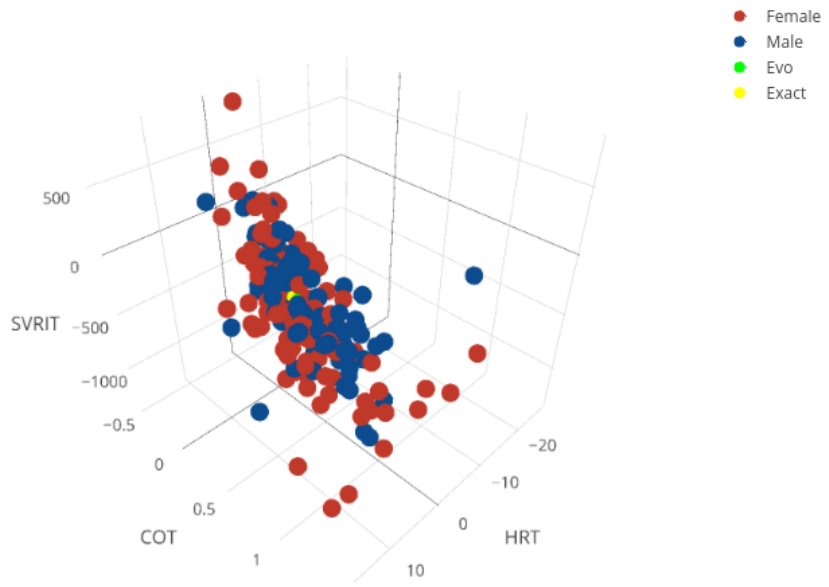

Figure 13: Visualization of the `LASERI` data, together with the Oja median calculated with the evolutionary (green) and the exact (yellow) algorithm. An interactive version of the figure can be found here: https://plot.ly/~fischuu/9.

.

```
R> oja1sampleTest(X[SEX == "Female", ], scores = "rank")

PLEASE NOTE: You have requested to compute 2246720
hyperplanes in R^3. This may take a while.

OJA 1 SAMPLE SIGNED RANK TEST

data:  X[SEX == "Female", ]
Q.R = 79.553, replications = 1000, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(0,0,0)
```

Therefore neither women nor men return in the time frame considered here to their pre-tilt levels.

To conclude this example we test then still if these differences differ between women and men using a two sample Oja sign test where the p value is computed using permutation arguments.

```
R> ojaCsampleTest(X ~ SEX, scores = "sign", method = "permutation")

        OJA C SAMPLE SIGN TEST

data:  X by SEX
Q.S = 12.982, permutations = 1000, p-value = 0.007
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

This indicates with a p value of 0.007 that also this null hypothesis needs to be rejected.

5. Conclusions

There are many different multivariate medians. In this paper we explained how the different medians extend different properties of the univariate median to the multivariate case. The Oja median has very convincing statistical properties, but is also among the computationally more challenging ones. We described the R package **OjaNP**, which provides four different algorithms for its computation. Along with the concept of the Oja median comes the notion of Oja signs and ranks and multivariate scatter estimators based upon them. The package provides also functions for these useful tools, which can then be used for robust multivariate inferential procedures. As examples, we described and implemented the one-sample and the C-sample location test based on Oja signs and ranks and showed their practical use in simple and complex data situations.

Acknowledgments

The work of Klaus Nordhausen was supported by the Academy of Finland (grant 268703). Oleksii Pokotylo is supported by the Cologne Graduate School of Management, Economics and Social Sciences. The work of Daniel Vogel was supported by the DFG collaborate research grant SFB 823. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Arcones MA, Chen Z, Gine E (1994). "Estimators Related to U-Processes with Applications to Multivariate Medians: Asymptotic Normality." *The Annals of Statistics*, **22**(3), 1460–1477. doi:10.1214/aos/1176325637.

Babu GJ, Rao CR (1988). "Joint Asymptotic Distribution of Marginal Quantiles and Quantile Functions in Samples from a Multivariate Population." *Journal of Multivariate Analysis*, **27**(1), 15–23. doi:10.1016/0047-259x(88)90112-1.

Bai ZD, He X (1999). "Asymptotic Distributions of the Maximal Depth Estimators for Regression and Multivariate Location." *The Annals of Statistics*, **27**(5), 1616–1637. doi:10.1214/aos/1017939144.

Chen Z (1995). "Robustness of the Half-Space Median." *Journal of Statistical Planning and Inference*, **46**(2), 175–181. doi:10.1016/0378-3758(94)00105-5.

Donoho D, Gasko M (1992). "Breakdown Properties of Location Estimates Based on Half-Space Depth and Projected Outlyingness." *The Annals of Statistics*, **20**(4), 1803–1827. doi:10.1214/aos/1176348890.

Filzmoser P, Fritz H, Kalcher K (2018). ***pcaPP***: *Robust PCA by Projection Pursuit*. R package version 1.9-73, URL https://CRAN.R-project.org/package=pcaPP.

Fischer D, Mosler K, Möttönen J, Nordhausen K, Pokotylo O, Vogel D (2020). ***OjaNP***: *Multivariate Methods Based on the Oja Median and Related Concepts*. R package version 1.0-0, URL https://CRAN.R-project.org/package=OjaNP.

Genest M, Masse JC, Plante JF (2019). ***depth**: Depth Functions Tools for Multivariate Analysis*. R package version 2.1-1.1, URL https://CRAN.R-project.org/package=depth.

Hayford J (1902). "What Is the Center of an Area or the Center of a Population." *Journal of the American Statistical Association*, **8**(58), 47–58. doi:10.2307/2276137.

Hettmansperger TP, McKean JW (2011). *Robust Nonparametric Statistical Methods*. 2nd edition. CRC Press, Boca Raton.

Hettmansperger TP, Möttönen J, Oja H (1997). "Affine-Invariant Multivariate One-Sample Signed-Rank Tests." *Journal of the American Statistical Association*, **92**(440), 1591–1600. doi:10.1080/01621459.1997.10473681.

Hettmansperger TP, Möttönen J, Oja H (1999). "The Geometry of the Affine Invariant Multivariate Sign and Rank Methods." *Journal of Nonparametric Statistics*, **11**(1–3), 271–285. doi:10.1080/10485259908832784.

Hettmansperger TP, Möttönnen J, Oja H (1998). "Affine Invariant Multivariate Rank Tests for Several Samples." *Statistica Sinica*, **8**, 785–800.

Hettmansperger TP, Nyblom J, Oja H (1994). "Affine Invariant Multivariate One-Sample Sign Tests." *Journal of the Royal Statistical Society B*, **56**(1), 221–234. doi:10.1111/j.2517-6161.1994.tb01973.x.

Hettmansperger TP, Oja H (1994). "Affine Invariant Multivariate Multisample Sign Tests." *Journal of the Royal Statistical Society B*, **56**(1), 235–249. doi:10.1111/j.2517-6161.1994.tb01974.x.

Hotelling H (1929). "Stability in Competition." *The Economic Journal*, **39**(153), 41–57. doi:10.2307/2224214.

Koenker R (2019). ***quantreg**: Quantile Regression*. R package version 5.54, URL http://CRAN.R-project.org/package=quantreg.

Mosler K, Pokotylo O (2015). "Computation of the Oja Median by Bounded Search." In K Nordhausen, S Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods*, pp. 185–203. Springer-Verlag.

Möttönen J, Nordhausen K, Oja H (2010). "Asymptotic Theory of the Spatial Median." In J Antoch, M Hŭsková, PK Sen (eds.), *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurecková*, volume 7, pp. 182–193.

Niinimaa A (1995). "Bivariate Generalizations of the Median." In EM Tiit, T Kollo, H Niemi (eds.), *Multivariate Statistics and Matrices in Statistics*, pp. 163–180. VSP BV, Zeist.

Niinimaa A, Oja H (1995). "On the Influence Functions of Certain Bivariate Medians." *Journal of the Royal Statistical Society B*, **57**(3), 565–574. doi:10.1111/j.2517-6161.1995.tb02048.x.

Niinimaa A, Oja H, Nyblom J (1992). "Algorithm AS 277: The Oja Bivariate Median." *Journal of the Royal Statistical Society C*, **41**(3), 611–633. doi:10.2307/2348099.

Niinimaa A, Oja H, Tableman M (1990). "The Finite-Sample Breakdown Point of the Oja Bivariate Median and of the Corresponding Half-Samples Version." *Statistics & Probability Letters*, **10**(4), 325–328. doi:10.1016/0167-7152(90)90050-h.

Nordhausen K, Möttönen J, Oja H (2018a). ***MNM**: Multivariate Nonparametric Methods. An Approach Based on Spatial Signs and Ranks*. R package version 1.0-3, URL https://CRAN.R-project.org/package=MNM.

Nordhausen K, Oja H (2011). "Multivariate L_1 Methods: The Package **MNM**." *Journal of Statistical Software*, **43**(5), 1–28. doi:10.18637/jss.v043.i05.

Nordhausen K, Oja H (2018). "Robust Nonparametric Inference." *Annual Review of Statistics and Its Application*, **5**(1), 473–500. doi:10.1146/annurev-statistics-031017-100247.

Nordhausen K, Oja H, Tyler DE (2008). "Tools for Exploring Multivariate Data: The Package **ICS**." *Journal of Statistical Software*, **28**(6), 1–31. doi:10.18637/jss.v028.i06.

Nordhausen K, Sirkiä S, Oja H, Tyler DE (2018b). ***ICSNP**: Tools for Multivariate Nonparametrics*. R package version 1.1-1, URL https://CRAN.R-project.org/package=ICSNP.

Oja H (1983). "Descriptive Statistics for Multivariate Distributions." *Statistics & Probability Letters*, **1**(6), 327–332. doi:10.1016/0167-7152(83)90054-8.

Oja H (1999). "Affine Invariant Multivariate Sign and Rank Tests." *Scandinavian Journal of Statistics*, **26**(3), 319–343. doi:10.1111/1467-9469.00152.

Oja H (2010). *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks*. Springer-Verlag, New York.

Oja H (2013). "Multivariate Median." In C Becker, R Fried, S Kuhnt (eds.), *Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather*, pp. 3–16. Springer-Verlag, Berlin.

Oja H, Niinimaa A (1985). "Asymptotic Properties of the Generalized Median in the Case of Multivariate Normality." *Journal of the Royal Statistical Society B*, **47**(2), 372–377. doi:10.1111/j.2517-6161.1985.tb01366.x.

Ollila E, Croux C, Oja H (2004). "Influence Function and Asymptotic Efficiency of the Affine Equivariant Rank Covariance Matrix." *Statistica Sinica*, **14**(1), 297–316.

Ollila E, Oja H, Croux C (2003). "The Affine Equivariant Sign Covariance Matrix: Asymptotic Behavior and Efficiencies." *Journal of Multivariate Analysis*, **87**(2), 328–355. doi:10.1016/s0047-259x(03)00045-9.

Puri ML, Sen PK (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Romanazzi M (2001). "Influence Function of Halfspace Depth." *Journal of Multivariate Analysis*, **77**(1), 138–161. doi:10.1006/jmva.2000.1929.

Ronkainen T, Oja H, Orponen P (2003). "Computation of the Multivariate Oja Median." In R Dutter, P Filzmoser, U Gather, PJ Rousseeuw (eds.), *Developments in Robust Statistics: Proceedings of the International Conference on Robust Statistics (ICORS'01, Stift Vorau, Austria, July 2001)*, pp. 344–359. Springer-Verlag, Berlin Heidelberg.

Shen G (2008). "Asymptotics of Oja Median Estimate." *Statistics & Probability Letters*, **78**(14), 2137–2141. doi:10.1016/j.spl.2008.02.004.

Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, Despouy P (2019). ***plotly**: Create Interactive Web Graphics via **plotly.js***. R package version 4.9.1, URL https://CRAN.R-project.org/package=plotly.

Small CG (1990). "A Survey of Multidimensional Medians." *International Statistical Review*, **58**(3), 263–277. doi:10.2307/1403809.

Stein P (1966). "A Note on the Volume of a Simplex." *The American Mathematical Monthly*, **73**(3), 299–301. doi:10.2307/2315353.

Tukey JW (1975). "Mathematics and the Picturing of Data." In *Proceedings of the International Congress of Mathematicians*, volume 2, pp. 523–531. Vancouver.

Visuri S, Koivunen V, Oja H (2000). "Sign and Rank Covariance Matrices." *Journal of Statistical Planning and Inference*, **91**(2), 557–575. doi:10.1016/s0378-3758(00)00199-3.

Visuri S, Ollila E, Koivunen V, Möttönen J, Oja H (2003). "Affine Equivariant Multivariate Rank Methods." *Journal of Statistical Planning and Inference*, **114**(1–2), 161–185. doi:10.1016/s0378-3758(02)00469-x.

Vogel D, Fried R (2008). "Estimating Partial Correlations Using the Oja Sign Covariance Matrix." In P Brito (ed.), *COMPSTAT 2008 – Proceedings in Computational Statistics*, volume II, pp. 721–729. Physica-Verlag, Heidelberg.

Vogel D, Fried R (2011). "Elliptical Graphical Modelling." *Biometrika*, **98**(4), 935–951. doi:10.1093/biomet/asr037.

Vogel D, Köllmann C, Fried R (2008). "Partial Correlation Estimates Based on Signs." In J Heikkonen (ed.), *Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering*. TICSP Series # 43.

Weber A (1909). *Über den Standort der Industrien*. Mohr, Tübingen.

Weber A (1929). *Theory of the Location of Industries*. The University of Chicago Press, Chicago.

Affiliation:

Daniel Fischer
Natural Resources Institute Finland (Luke)
Production Systems
Jokioinen, Finland
and
School of Health Sciences
University of Tampere
Tampere, Finland
E-mail: daniel.fischer@luke.fi

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
February 2020, Volume 92, Issue 8 *Submitted:* 2016-05-06
doi:10.18637/jss.v092.i08 *Accepted:* 2018-09-06