Multivariate nonparametrical methods based on spatial signs and ranks: The R package SpatialNP

Seija Sirkiä
University of Jyväskylä

Sara Taskinen
University of Jyväskylä

Jaakko Nevalainen
National Public Health Institute and University of Tampere

Hannu Oja
University of Tampere

June 26, 2007

Abstract

Classical multivariate statistical inference methods are often based on the sample mean vector and covariance matrix. They are then optimal under the assumption of multivariate normality but loose in efficiency in the case of heavy tailed distribution. In this paper nonparametric and robust competitors based on the spatial signs and ranks are discussed and the **R** statistical software package to implement the procedures is documented. The location tests and estimates corresponding to the different score functions (sign, rank, signed rank) are reviewed in the one sample, several samples and multivariate regression cases. Also the tests for sphericity and independence of the random vectors are discussed. The inner standardization of the test statistics is then needed for the affine invariance/equivariance of the methods and it produces the corresponding scatter (or shape) matrix estimate. The main features of the **R** package **SpatialNP** is described and its use illustrated with several examples.

1 Introduction

Classical multivariate statistical inference methods (Hotelling's T^2, multivariate analysis of variance, multivariate regression, inference on the correlation structure) are based on the regular sample mean vector and covariance matrix. The standard multivariate techniques are optimal under the assumption of multivariate normality but unfortunately poor in efficiency for heavy tailed distributions and highly sensitive to outlying observations. In the paper nonparametric and robust competitors to the standard multivariate inference methods for high dimensional based on the spatial signs and ranks are discussed and the **R** statistical software package which implements the procedures is described.

The univariate concepts of sign and rank are based on the ordering of the data. In the multivariate case there are no natural orderings of the data points. An approach utilizing objective or criterion functions is therefore often used to extend these concepts to the multivariate case. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)'$ be an $n \times p$ data matrix with n observations and p variables. The multivariate spatial sign $\mathbf{u}_i$, multivariate spatial (centered) rank $\mathbf{r}_i$, and multivariate spatial signed-rank $\mathbf{q}_i$, $i = 1, ..., n$, may be implicitly defined using the three L_1 criterion functions with Euclidean norm

$$\begin{aligned} \text{ave}\{||\mathbf{y}_i||\} &= \text{ave}\{\mathbf{u}_i'\mathbf{y}_i\}, \\ \frac{1}{2}\text{ave}\{||\mathbf{y}_i - \mathbf{y}_j||\} &= \text{ave}\{\mathbf{r}_i'\mathbf{y}_i\}, \quad \text{and} \\ \frac{1}{4}\text{ave}\{||\mathbf{y}_i - \mathbf{y}_j|| + ||\mathbf{y}_i + \mathbf{y}_j||\} &= \text{ave}\{\mathbf{q}_i'\mathbf{y}_i\}. \end{aligned}$$

See Hettmansperger and Aubuchon (1988). Note also that the sign, centered rank, and signed-rank may be seen as *scores* $\mathbf{T}(\mathbf{y})$ corresponding to the three objective functions. The $\mathbf{T}(\mathbf{y}_i) = \mathbf{y}_i$, $i = 1, \ldots, n$, are the scores corresponding to the regular L_2 criterion $\text{ave}\{||\mathbf{y}_i||^2\} = \text{ave}\{\mathbf{y}_i'\mathbf{y}_i\}$.

Consider next these objective functions if applied to the residuals in the linear regression model. The first objective function, the *mean deviation* of the residuals, is the basis for the so called least absolute deviation (LAD) methods; it yields different median-type estimates and sign tests in the one-sample, two-sample, c-sample and finally general linear model settings. The second objective function is the *mean difference* of the residuals. The second and third objective functions generate Hodges-Lehmann type estimates and rank tests for different location problems. It is well known that in the

univariate normal case the asymptotic efficiency of the sign (rank) based method with respect to the optimal L_2 method is 0.637 (0.955). For heavy tailed univariate distributions, t_3 and t_{10}, the efficiencies are 1.621 (1.900) and 0.757 (1.054), respectively.

Möttönen and Oja (1995), Choi and Marden (1997), Marden (1999a) and Oja and Randles (2004) reviewed the theory of multivariate spatial sign and rank tests and the related estimates based on the above L_1 objective functions. Möttönen et al. (1997) calculated the asymptotic efficiencies $e_1(p, \nu)$ and $e_2(p, \nu)$ of the multivariate spatial sign and rank methods, respectively, in the p-variate t_ν distribution case. In the 3-variate case, for example, the asymptotic efficiencies are

$$\begin{aligned} &e_1(3,3) = 2.162, \quad e_1(3,10) = 1.009, \quad e_1(3,\infty) = 0.849, \\ &e_2(3,3) = 1.994, \quad e_2(3,10) = 1.081, \quad e_2(3,\infty) = 0.973 \end{aligned}$$

and in the 10-variate case one has even higher efficiencies

$$\begin{aligned} &e_1(10,3) = 2.422, \quad e_1(10,10) = 1.131, \quad e_1(10,\infty) = 0.951, \\ &e_2(10,3) = 2.093, \quad e_2(10,10) = 1.103, \quad e_2(10,\infty) = 0.989. \end{aligned}$$

This is, however, only one possible approach to multivariate analogues to common univariate nonparametric tests (sign test, rank test) and estimates (median, Hodges-Lehmann estimate). Randles (1989) developed an affine invariant sign test based on *interdirections*. Interdirections measure the angular distance between two observation vectors relative to the rest of the data. Randles (1989) was followed by a series of papers introducing nonparametric sign and rank interdirection tests. These tests are typically asymptotically equivalent with spatial sign and rank tests. The tests and estimates are, unfortunately, computationally heavy.

The inference methods based on marginal signs and ranks are described in Puri and Sen (1971) but they are not affine invariant/equivariant. In a series of papers, Hallin and Paindaveine constructed *optimal signed-rank tests* for the location and scatter problems in the elliptical model; see the seminal papers by Hallin and Paindaveine (2002, 2006) and Hallin et al. (2006). The location tests were based on the spatial signs and ranks of the Euclidean lengths of the standardized observations. For yet another different approach, see Oja (1999) and the references therein.

The paper is organized as follows. In the Section 2 the theory is recalled: First the score functions (sign, rank, signed rank) corresponding to the three

objective functions are introduced. Their covariance matrices play a special role. Then the tests corresponding to the different choices of the score functions are listed in the one sample, several samples and multivariate regression cases. Also the tests for sphericity and independence of the random vectors are discussed. The inner standardization of the test statistics needed for the affine invariance/equivariance of the methods is described. The related location and scatter estimates are discussed as well. The main features of the **R** package **SpatialNP** are then briefly described in Section 3. In Section 4 the use of the package is illustrated with several practical examples.

2 Multivariate spatial signs and ranks

2.1 Spatial signs and ranks

Let

$$\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)'$$

be an $n \times p$ data matrix with n observations and p variables. The data based *spatial sign, spatial rank and spatial signed-rank functions* $\mathbf{U}(\mathbf{y})$, $\mathbf{R}(\mathbf{y}) = \mathbf{R}(\mathbf{y}; \mathbf{Y})$ and $\mathbf{Q}(\mathbf{y}) = \mathbf{Q}(\mathbf{y}; \mathbf{Y})$ are defined as

$$\begin{aligned} \mathbf{U}(\mathbf{y}) &= \begin{cases} \|\mathbf{y}\|^{-1}\mathbf{y}, & \mathbf{y} \neq \mathbf{0} \\ \mathbf{0}, & \mathbf{y} = \mathbf{0} \end{cases}, \\ \mathbf{R}(\mathbf{y}; \mathbf{Y}) &= \text{ave}\{\mathbf{U}(\mathbf{y} - \mathbf{y}_i)\} \quad \text{and} \\ \mathbf{Q}(\mathbf{y}; \mathbf{Y}) &= \frac{1}{2}\left[\mathbf{R}(\mathbf{y}; \mathbf{Y}) + \mathbf{R}(\mathbf{y}; -\mathbf{Y})\right]. \end{aligned}$$

Clearly the spatial sign function $\mathbf{U}(\mathbf{y})$ and signed-rank function $\mathbf{Q}(\mathbf{y}; \mathbf{Y})$ are odd, that is, $\mathbf{U}(-\mathbf{y}) = -\mathbf{U}(\mathbf{y})$ and $\mathbf{Q}(-\mathbf{y}; \mathbf{Y}) = -\mathbf{Q}(\mathbf{y}; \mathbf{Y})$.

Definition 1 *The observed spatial signs are* $\mathbf{u}_i = \mathbf{U}(\mathbf{y}_i), \quad i = 1, \ldots, n$. *We write also* $\mathbf{u}_{ij} = \mathbf{U}(\mathbf{y}_i - \mathbf{y}_j), \quad i, j = 1, \ldots, n$. *As in the univariate case, the observed central spatial ranks are averages of signs of pairwise differences*

$$\mathbf{r}_i = \mathbf{R}(\mathbf{y}_i; \mathbf{Y}) = ave_j\{\mathbf{U}(\mathbf{y}_i - \mathbf{y}_j)\}, \quad i = 1, \ldots, n.$$

Finally, the observed spatial signed-ranks are given as

$$\mathbf{q}_i = \mathbf{Q}(\mathbf{y}_i; \mathbf{Y}) = \frac{1}{2} ave_j\{\mathbf{U}(\mathbf{y}_i - \mathbf{y}_j) + \mathbf{U}(\mathbf{y}_i + \mathbf{y}_j)\}, \quad i = 1, ..., n.$$

The spatial sign $\mathbf{u}_i$ is just a direction vector of length one (lying on the unit p-sphere $\mathcal{S}_p$) whenever $\mathbf{y}_i \neq \mathbf{0}$. The centered ranks $\mathbf{r}_i$ and signed-ranks $\mathbf{q}_i$ lie in the unit p-ball $\mathcal{B}_p$. The direction of $\mathbf{r}_i$ ($\mathbf{q}_i$) roughly tells the direction of $\mathbf{y}_i$ from the center of the data cloud (from the origin), and its length tells how far away this point is from the center (from the origin). The spatial signs, ranks and signed-ranks are only orthogonally equivariant, not affine equivariant. The covariation of the marginals of sign and rank vectors will be described by their covariance matrices as follows.

Definition 2 *Let* $\mathbf{Y}$ *be a data matrix. Then the spatial sign covariance matrix* $SCov(\mathbf{Y})$*, and the symmetrized spatial sign covariance matrix* $SSCov(\mathbf{Y})$ *are*

$$\begin{aligned} SCov(\mathbf{Y}) &= ave\left\{\mathbf{u}_i\mathbf{u}_i'\right\} \quad and \\ SSCov(\mathbf{Y}) &= ave\left\{\mathbf{u}_{ij}\mathbf{u}_{ij}'\right\}. \end{aligned}$$

We also define

Definition 3 *Let* $\mathbf{Y}$ *be a data matrix. Then the spatial rank covariance matrix* $RCov(\mathbf{Y})$*, and the spatial signed-rank covariance matrix* $SRCov(\mathbf{Y})$ *are*

$$\begin{aligned} RCov(\mathbf{Y}) &= ave\left\{\mathbf{r}_i\mathbf{r}_i'\right\} \quad and \\ SRCov(\mathbf{Y}) &= ave\left\{\mathbf{q}_i\mathbf{q}_i'\right\}. \end{aligned}$$

The matrices $SCov(\mathbf{Y})$, $SSCov(\mathbf{Y})$, $RCov(\mathbf{Y})$ and $SRCov(\mathbf{Y})$ are not genuine scatter matrices as they are not affine equivariant. They are equivariant under orthogonal transformations only. See also that the $SSCov$ and $RCov$ are shift or location invariant. Finally note that the sign covariance matrix and the symmetrized sign covariance matrix are standardized in the sense that $tr(SCov(\mathbf{Y})) = tr(SSCov(\mathbf{Y})) = 1$. For the use of spatial sign and rank covariance matrices, see Marden (1999b) and Visuri et al. (2000).

2.2 One sample location and scatter

Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)'$ be a random sample from a symmetrical distribution satisfying $-(\mathbf{y}_i - \mu) \sim (\mathbf{y}_i - \mu)$ for unknown symmetry center μ. For the

scatter problem we often need a stronger assumption that the $\mathbf{y}_i$ has an elliptically symmetric distribution with density function

$$f(\mathbf{y}) = \det(\mathbf{\Sigma}^{-1/2}) g\left(\|\mathbf{\Sigma}^{-1/2}(\mathbf{y}-\mu)\|\right)$$

with symmetry center μ, scatter matrix $\mathbf{\Sigma}$, and unspecified g. We wish to test the null hypothesis $H_0 : \ \mu = \mathbf{0}$ and to estimate the unknown μ. A general idea to construct tests and estimates is to use an odd vector valued *score function* $\mathbf{T}(\mathbf{y})$ yielding individual scores $\mathbf{t}_i = \mathbf{T}(\mathbf{y}_i)$, $i = 1, \ldots, n$. The test statistic is simply

$$\text{ave}\left\{\mathbf{T}(\mathbf{y}_i)\right\}$$

Then, under the null hypothesis, $\sqrt{n}\ \text{ave}\left\{\mathbf{T}(\mathbf{y}_i)\right\} \to_d N_p(\mathbf{0}, \mathbf{B})$ where $\mathbf{B} = E\left\{\mathbf{T}(\mathbf{y}_i)\mathbf{T}(\mathbf{y}_i)'\right\}$. (The assumption on the existence of $\mathbf{B}$ is needed.) A natural estimate of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \text{ave}\left\{\mathbf{T}(\mathbf{y}_i)\mathbf{T}(\mathbf{y}_i)'\right\}.$$

There are two ways to standardize the test statistic.

- One can use an *outer standardization:* Under general assumptions and under the null hypothesis,

$$Q^2 = n\ ||\hat{\mathbf{B}}^{-1/2}\text{ave}\left\{\mathbf{T}(\mathbf{y}_i)\right\}||^2 \to_d\ \chi^2_p$$

- Sometimes it is possible to use an *inner standardization*: Then one first finds a $p \times p$ transformation matrix H such that, for $\mathbf{z}_i = \mathbf{H}\mathbf{y}_i$, $i = 1, \ldots, n$,

$$p \cdot \text{ave}\left\{T(\mathbf{z}_i)T(\mathbf{z}_i)'\right\} = \text{ave}\left\{T(\mathbf{z}_i)'T(\mathbf{z}_i)\right\}\mathbf{I}_p$$

  The test statistic using inner standardization is then

$$Q^2 = np \cdot \frac{\|\text{ave}\left\{\mathbf{T}(\mathbf{z}_i)\right\}\|^2}{\text{ave}\left\{\|\mathbf{T}(\mathbf{z}_i)\|^2\right\}}$$

  also with the limiting χ^2_p null distribution.

The approximate p-value may thus be based on the limiting chi squared distribution. For small sample sizes, an alternative way to construct the p-value is to use the *sign-change argument*. Let $\mathbf{J}$ be a $n \times n$ diagonal matrix

with diagonal elements ± 1. Then the p-value of a conditionally distribution-free sign-change test statistic is

$$E_{\mathbf{J}}\left[I\left(Q^2(\mathbf{JY}) \geq Q^2(\mathbf{Y})\right)\right]$$

where $\mathbf{J}$ has a uniform distribution over its all 2^n possible values.

The matrix

$$\mathbf{C} = (\mathbf{H}'\mathbf{H})^{-1}$$

is an affine equivariant scatter (or shape) matrix corresponding to score function $\mathbf{T}(\mathbf{y})$. The companion location estimate $\hat{\mu}$ is determined by estimating equations

$$\sum_{i=1}^{n} T(\mathbf{y}_i - \hat{\mu}) = \mathbf{0}.$$

1. **Hotelling's T^2 and the sample mean:** Classical Hotelling's T^2 test is obtained with score function $\mathbf{T}(\mathbf{y}) = \mathbf{y}$ corresponding to the L_2 criterion. Now $\mathbf{B} = \mathbf{C} = Cov(\mathbf{Y})$ is the regular *sample covariance matrix* and both the outer and inner standardizations yield the same well-known Hotelling's one sample test statistic

   $$Q^2 = n\bar{\mathbf{y}}'\mathbf{B}^{-1}\bar{\mathbf{y}}.$$

   The test statistic is affine invariant in the sense that

   $$Q^2(\mathbf{YH}') = Q^2(\mathbf{Y}), \quad \text{for all } \mathbf{H}.$$

   The companion estimate is the *sample mean vector*.

2. **Spatial sign test and the spatial median:** The spatial sign test is obtained with score function $\mathbf{T}(\mathbf{y}) = \mathbf{U}(\mathbf{y})$. Then $\mathbf{B} = SCov(\mathbf{Y})$ is the spatial sign covariance matrix, and $\mathbf{C}$ is the celebrated *Tyler's shape matrix* (Tyler, 1987). Q^2 with outer standardization is invariant under orthogonal transformations only, but inner standardization gives affine invariance. It is remarkable that Q^2 with inner standardization is strictly distribution-free in the elliptic model. The invariant test was first proposed by Randles (2000). The companion location estimate is the spatial median (Gower, 1974; Brown, 1983). An inner standardization with respect to location and shape simultaneously is

given by a $p \times p$ matrix $\mathbf{H}$ and p-vector $\mathbf{h}$ such that, for $\mathbf{z}_i = \mathbf{H}(\mathbf{y}_i - \mathbf{h})$, $i = 1, \ldots, n$,

$$\text{ave}\{\mathbf{U}(\mathbf{z}_i)\} = \mathbf{0} \quad \text{and} \quad \text{ave}\{\mathbf{U}(\mathbf{z}_i)\mathbf{U}(\mathbf{z}_i)'\} = \frac{1}{p}\mathbf{I}_p.$$

This procudes the *Hettmansperger-Randles estimate* (Hettmansperger and Randles, 2002), simultaneous estimate of location $\mathbf{h}$ and shape $\mathbf{C} = (\mathbf{H}'\mathbf{H})^{-1}$. The location estimate is the *transformation-retransformation (TR) spatial median* using Tyler's scatter matrix. The Dümbgen shape estimate (Dümbgen, 1998), a symmetrized version of Tyler's shape matrix, is given by transformation matrix $\mathbf{H}$ satisfying

$$\text{ave}\{\mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)\mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)'\} = \frac{1}{p}\mathbf{I}_p.$$

Note that no location estimate is needed here.

3. **Spatial signed-rank test and the spatial Hodges-Lehmann estimate:** The spatial signed-rank test is obtained with the score function $\mathbf{T}(\mathbf{y}) = \mathbf{Q}(\mathbf{y})$. Now $B = SRCov(\mathbf{Y})$ is the spatial signed-rank covariance matrix, and $\mathbf{C}$ is the corresponding affine equivariant shape matrix. See Sirkiä et al. (2007). The inner standardization gives affine invariance. The companion estimate is the *spatial Hodges-Lehmann (HL) estimate*.

Note that in the inner standardization we find a transformation matrix $\mathbf{H}$ such that

$$SCov(\mathbf{YH}'), \quad SSCov(\mathbf{YH}'), \quad RCov(\mathbf{YH}') \quad \text{or} \quad SRCov(\mathbf{YH}'),$$

is proportional to the identity matrix, respectively. In the elliptic model, the scatter (or shape) matrix $\mathbf{C} = (\mathbf{H}'\mathbf{H})^{-1}$ (properly standardized) then estimates the population quantity which is proportional to the regular covariance matrix. See Sirkiä et al. (2007) for a discussion of these matrices and their use in testing for the sphericity of the distribution.

2.3 Several samples location

Let now data matrix

$$\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_c')'$$

consist of c independent random samples

$$\mathbf{Y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, ..., \mathbf{y}_{in_i})', \quad i = 1, \ldots, c,$$

from p-variate distributions with cdf's $F_1, F_2, \ldots, F_c$. Write also $n = n_1 + \ldots + n_c$. We wish to test the null hypothesis $H_0 : F_1 = F_2 = \ldots = F_c$ saying that all observations come from the same population. The location tests are constructed assuming that $F_i(\mathbf{y}) = F(\mathbf{y} - \mu_i)$, $i = 1, \ldots, c$. Again, we wish to use a general location score function $\mathbf{T}(\mathbf{y})$ in test construction.

In the approach based on the inner standardization, first a $p \times p$ matrix $\mathbf{H}$ and p-vector $\mathbf{h}$ are found such that, for $\mathbf{z}_{ij} = \mathbf{H}(\mathbf{y}_{ij} - \mathbf{h})$, $i = 1, ..., c;\ j = 1, ..., n_i$,

$$\begin{aligned} \text{ave}\left\{\mathbf{T}(\mathbf{z}_{ij})\right\} &= \mathbf{0} \text{ and} \\ p \cdot \text{ave}\left\{\mathbf{T}(\mathbf{z}_{ij})\mathbf{T}(\mathbf{z}_{ij})'\right\} &= \text{ave}\left\{\|\mathbf{T}(\mathbf{z}_{ij})\|^2\right\}\mathbf{I}_p \end{aligned}$$

The several-samples location test statistic is then

$$Q^2 = p \cdot \frac{\sum n_i \|\text{ave}_j \mathbf{T}(\mathbf{z}_{ij})\|^2}{\text{ave}\|\mathbf{T}(\mathbf{z}_{ij})\|^2}$$

Under general assumptions, the limiting distribution of the test statistic Q^2 is a chi squared distribution with $p(c-1)$ degrees of freedom.

The p-value can also be calculated for the conditionally distribution-free *permutation test* version. Let $\mathbf{P}$ be a $n \times n$ permutation matrix (obtained from an identity matrix by permuting rows or columns). The p-value of the permutation test statistic is then

$$E_{\mathbf{P}}\left[I\left(Q^2(\mathbf{PY}) \geq Q^2(\mathbf{Y})\right)\right]$$

where $\mathbf{P}$ has a uniform distribution over all possible $n!$ permutations.

Possible choices are again

1. **Hotelling's T^2 and MANOVA:** Classical MANOVA test is obtained with score function $\mathbf{T}(\mathbf{y}) = \mathbf{y}$ corresponding to the L_2 criterion.

2. **MANOVA based on spatial signs:** MANOVA based on the spatial signs is obtained with score function $\mathbf{T}(\mathbf{y}) = \mathbf{U}(\mathbf{y})$. This is an extension of *Mood's test* to the multivariate case.

3. **MANOVA based on spatial ranks:** This approach uses the spatial rank function $\mathbf{R}(\mathbf{y})$ as a score function. Note that the spatial ranks are automatically centered and no shift estimate is needed. This extends *Wilcoxon-Mann-Whitney* and *Kruskal-Wallis tests*.

2.4 Multivariate regression

Let now data matrix

$$(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & ... & \mathbf{x}_n \\ \mathbf{y}_1 & \mathbf{y}_2 & ... & \mathbf{y}_n \end{pmatrix}'$$

consist of n independent observations from a $(q+p)$-dimensional distribution ($\mathbf{x}_i$ is a q-vector and $\mathbf{y}_i$ is a p-vector, $i = 1, ..., n$). In the multivariate multiple regression model it is commonly assumed that, for fixed $\mathbf{X}$, the response matrix $\mathbf{Y}$ is generated by

$$\mathbf{Y} \ = \ (\mathbf{1}_n, \mathbf{X}) \ \beta \ + \ \mathbf{Z}$$

where β is a $(q+1) \times p$ matrix of regression coefficients and $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)'$ is an $n \times p$ matrix consisting of independent symmetric residual vectors. Again, we wish to use a general location score function $\mathbf{T}(\mathbf{y})$ in test construction.

Let

$$\mathbf{Z}(\beta) \ = \ \mathbf{Y} \ - \ (\mathbf{1}_n, \mathbf{X}) \ \beta$$

be the residual matrix corresponding to a choice β. Then the test statistic for testing $H_0 : \ \beta = \beta_0$ can be based on

$$\begin{pmatrix} \mathbf{1}_n' \\ \mathbf{X}' \end{pmatrix} \mathbf{T}(\mathbf{Z}(\beta_0)),$$

where the ith row of $\mathbf{T}(\mathbf{Z})$ is $\mathbf{T}(\mathbf{z}_i)$. The corresponding estimate $\hat{\beta}$ solves

$$\begin{pmatrix} \mathbf{1}_n' \\ \mathbf{X}' \end{pmatrix} \mathbf{T}(\mathbf{Z}(\hat{\beta})) = \mathbf{0}.$$

2.5 Testing for independence

As in the previous subsection, let again data matrix

$$(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \\ \mathbf{y}_1 & \mathbf{y}_2 & \dots & \mathbf{y}_n \end{pmatrix}'$$

consist of n independent observations from a $(q+p)$-dimensional distribution ($\mathbf{x}_i$ is a q-vector and $\mathbf{y}_i$ is a p-vector, $i = 1, \dots, n$). We now wish to test the hypothesis

$$H_0 : \quad \mathbf{x}_i \text{ and } \mathbf{y}_i \text{ are independent}$$

and use a general location score functions $\mathbf{T}(\mathbf{y})$ in test construction.

In the approach based on the inner standardization, we first find (as before) affine transformations

$$\mathbf{X} \rightarrow \mathbf{X}^* \text{ and } \mathbf{Y} \rightarrow \mathbf{Y}^*$$

such that

$$\begin{aligned} \text{ave}\left\{\mathbf{T}(\mathbf{x}_i^*)\right\} &= \mathbf{0} \text{ and} \\ q \cdot \text{ave}\left\{\mathbf{T}(\mathbf{x}_i^*)\mathbf{T}(\mathbf{x}_i^*)'\right\} &= \text{ave}\left\{\|\mathbf{T}(\mathbf{x}_i^*)\|^2\right\}\mathbf{I}_q \end{aligned}$$

and

$$\begin{aligned} \text{ave}\left\{\mathbf{T}(\mathbf{y}_i^*)\right\} &= \mathbf{0} \text{ and} \\ p \cdot \text{ave}\left\{\mathbf{T}(\mathbf{y}_i^*)\mathbf{T}(\mathbf{y}_i^*)'\right\} &= \text{ave}\left\{\|\mathbf{T}(\mathbf{y}_i^*)\|^2\right\}\mathbf{I}_p. \end{aligned}$$

The affine invariant test statistic is then

$$Q^2 = \frac{npq \cdot \|\text{ave}\left\{\mathbf{T}(\mathbf{x}_i^*)\mathbf{T}(\mathbf{y}_i^*)'\right\}\|^2}{\text{ave}\left\{\|\mathbf{T}(\mathbf{x}_i^*)\|^2\right\} \cdot \text{ave}\left\{\|\mathbf{T}(\mathbf{x}_i^*)\|^2\right\}}$$

The limiting null distribution is a chi squared distribution with pq degrees of freedom. A p-value can also be calculated for the conditionally distribution-free *permutation test* version:

$$E_{\mathbf{P}}\left[I\left(Q^2(\mathbf{X}, \mathbf{PY}) \geq Q^2(\mathbf{X}, \mathbf{Y})\right)\right]$$

where $\mathbf{P}$ has a uniform distribution over all possible $n!$ permutations

Our scores functions yield

1. **Classical Wilks (Wilks, 1935) test for independence:** This is obtained with score function $\mathbf{T}(\mathbf{y}) = \mathbf{y}$ and is "optimal" in the multivariate normal case.

2. **Extension of quadrant test by Blomqvist (1950):** Test of independence using the marginal (standardized) spatial sign vectors; $\mathbf{T}(\mathbf{y}) = \mathbf{U}(\mathbf{y})$.

3. **Extension of Spearman's rho (Spearman, 1904):** This approach uses the marginal standardized spatial ranks; $\mathbf{T}(\mathbf{y}) = \mathbf{R}(\mathbf{y})$.

3 R-package SpatialNP

The **R**-package **SpatialNP** contains implementations of most of the methods described above. It depends, directly or indirectly, on **R** version 2.5.0 and packages **ICSNP**, **ICS**, **mvtnorm** and **survey**. Some of the methods described in this paper are in fact implemented in package **ICSNP** but for a part of these a wrapper function is provided in **SpatialNP**. The examples given in the next section use the wrappers when available. The classical methods are mostly implemented in **R** base packages except for Hotelling's T^2 which is implemented in **ICSNP**. A short introduction of the functions follows.

Functions `spatial.symmsign`, `spatial.rank` and `spatial.signrank` compute the spatial symmetrized signs, ranks and signed ranks, respectively. Spatial signs are implemented in package **ICSNP** as function `spatial.sign`. It is possible to compute the scores without any standardization or with inner standardization as explained above, or even with respect to a predefined shape. In cases of the functions `spatial.sign` and `spatial.signrank` also a location vector is involved. The default location used when no other vector is given is the vector of column means.

The covariance matrices defined in Definitions 2 and 3 can be computed using functions `SCov`, `SSCov`, `RCov` and `SRCov`. Of these, `SCov` and `SRCov` again require a location vector and as with the score functions the vector of column means serves as default.

Spatial median and its affine equivariant counterpart, Hettmansperger-Randles estimate, are implemented in package **ICSNP** as `spatial.median` and `HR.Mest`. The corresponding affine equivariant location estimate using signed-rank scores is implemented in package **SpatialNP** as function `ae.hl.estimate`. Optionally, it is also possible to compute the non affine equivariant version, similar to the spatial median, using this function. The wrapper `spatial.location` covers all four cases with a `score` argument to choose between the estimates.

As mentioned above, the classical Hotelling's T^2 test is implemented in package **ICSNP**; the name of the function is `HotellingsT2`. Function `sr.loc.test` in **SpatialNP** covers both sign and (signed) rank based versions of the location test. The choice of the score is made via argument `score`. Both functions handle one as well as several sample cases (note that in the one sample case when ranks are chosen as scores the test is in fact based on signed ranks). At the moment the conditionally distribution free

version of the test is only provided for the sign based test.

Multivariate regression estimates based on signs and ranks can be found using function `sr.regression`. The choice of score to be used is again made via argument `score`. Testing for hypothesis concerning the regression coefficients is so far unimplemented.

Function `sr.indep.test` performs the independence test. As before, the argument `score` is used to choose between different scores. Both the asymptotic and conditionally distribution free p-values are provided.

Of the sphericity tests the ones based on signs and symmetrized signs are implemented as function `sr.sphere.test`. Also here the argument `score` controls the choice of score.

Further, there are functions for computing the inner standardization matrices. The ones corresponding to signs and symmetrised signs, Tyler's and Dümbgen's matrices, respectively, are implemented in package **ICSNP**. The names of these functions are `tyler.shape` and `duembgen.shape`. The ones based on ranks and signed ranks are called `rank.shape` and `signrank.shape` and are in package **SpatialNP**. A wrapper function `spatial.shape` is provided for a unified access to all four shape matrixces. Naturally, `tyler.shape` and `signrankmat` require a location vector for the computation, with vector of column means again as default. The algorithms are iterative and all four functions allow for providing the starting point of the iteration, as well as computing the so-called k-step versions of the matrices.

Alternatively, one can compute the simultaneous estimates of the location and shape, `HR.Mest` and `ae.hl.estimate` (or `spatial.location` for both) mentioned earlier; the resulting estimates of shape are returned as attributes to the location estimate.

A utility function `to.shape` is also provided for scaling a matrix to a shape matrix.

4 Examples

These examples involve three data sets that are provided in **R** packages. The `pulmonary` data set in **ICSNP** consist of changes in three pulmonary measurements of twelve workers after six hours of exposure to cotton dust. The famous `iris` data of sepal and petal width and length measurements of three iris subspecies is in package **datasets**. The `frets` data set in package **boot** consists of two measurements of the heads of the oldest and second

adult brothers from 25 families. The datasets are loadable by a command like

```
data(pulmonary)
```

once the corresponding packages are loaded.

4.1 Spatial signs and ranks

First consider the bivariate data formed by the first two variables of the `frets` data, the width and length of the head of the older brother. Top left panel of Figure 1 shows the original data. One of the observations is marked with a black dot to show how it is transformed. The signs, with respect to the mean vector and regular covariance matrix, of this data are computed by the calls

```
frets.2<-frets[,c(1,2)]
spatial.sign(frets.2,center=colMeans(frets.2),shape=cov(frets.2))
```

and are shown in the top right panel, together with the unit circle. The bottom left panel shows the unstandardized and the bottom right panel the standardized spatial ranks which are computed as

```
spatial.rank(frets.2,shape=FALSE)
spatial.rank(frets.2,shape=TRUE)
```

respectively. Note how the inner standardization makes the ranks appear uniformly distributed inside the unit circle.

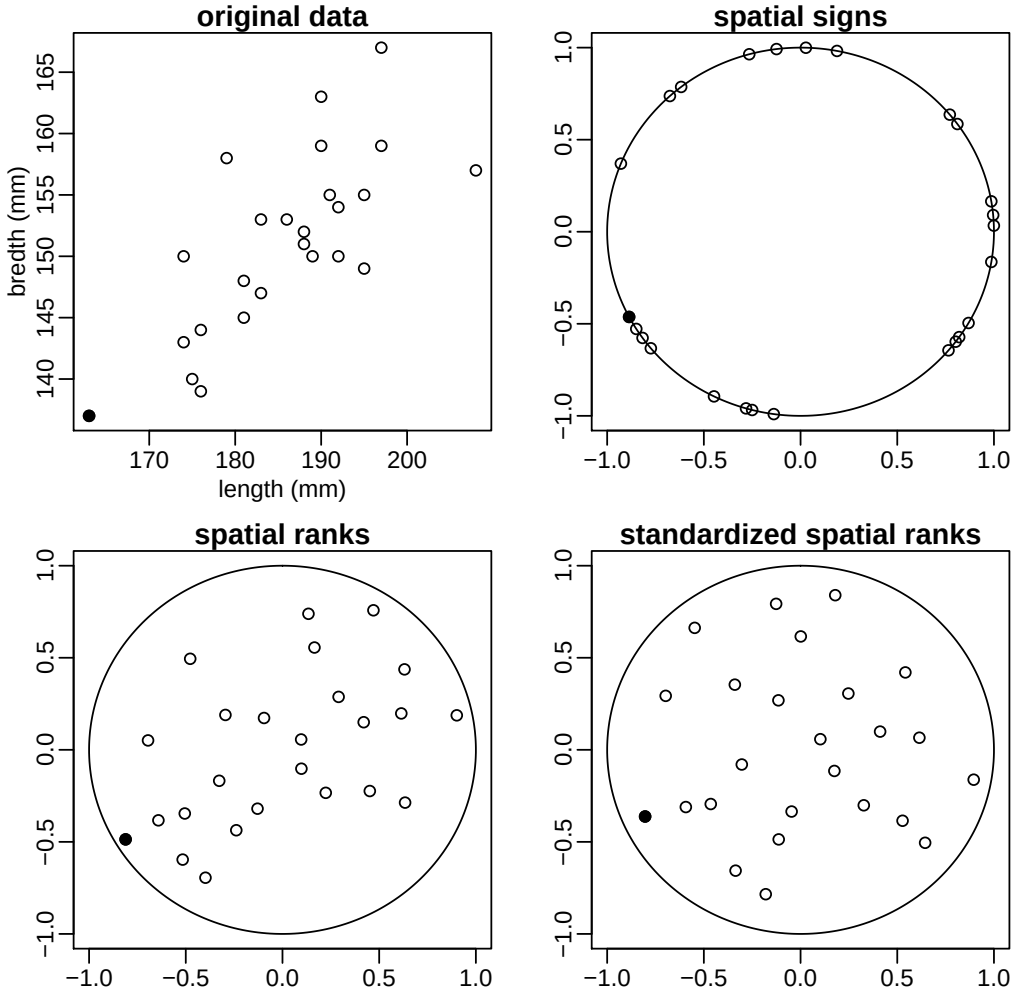
4.2 One sample location and shape

For the one sample location problem consider the `pulmonary` data. Test based on spatial signs for the null hypothesis that the exposure to cotton dust has no effect on pulmonary functions, i.e. that the observations are centered on the origin is done as

```
> sr.loc.test(pulmonary)

        One sample location test using spatial signs
```

Figure 1: Example plot of spatial signs and ranks



```
data:  pulmonary
Q.2 = 7.3771, df = 3, p-value = 0.0608
alternative hypothesis: true location is not equal to c(0,0,0)
```

because the sign based test for one sample location being equal to the origin is the default one. The signed rank based estimate of the location is

```
> spatial.location(pulmonary,score="signrank")
[1] -0.1392316 -0.1576346  2.8323453
attr(,"shape")
           [,1]       [,2]       [,3]
[1,]  0.1884525  0.2473997  -4.455964
[2,]  0.2473997  0.3749895  -2.783621
[3,] -4.4559639 -2.7836213 398.318583
```

which also gives the affine equivariant signed rank shape matrix as an attribute. For example the non affine equivariant (scaled to a shape matrix for easier comparison) and affine equivariant rank based matrices are found by

```
> to.shape(RCov(pulmonary))
           [,1]      [,2]       [,3]
[1,]  0.2137115 0.3492100 -0.3751607
[2,]  0.3492100 0.6614510  0.1739982
[3,] -0.3751607 0.1739982 58.9920789
> spatial.shape(pulmonary,score="rank")
           [,1]       [,2]       [,3]
[1,]  0.1855779  0.2453626  -4.179161
[2,]  0.2453626  0.3736209  -2.545533
[3,] -4.1791606 -2.5455330 384.048080
```

respectively.

4.3 Several samples location

To test whether the head measurements of brothers differ by their location it is possible to use the call (disregarding the pairwise nature of the data)

```
> sr.loc.test(frets[,c(1,2)],frets[,c(3,4)])

        Several samples location test using spatial signs
```

```
data:  frets[, c(1, 2)] and frets[, c(3, 4)]
Q.2 = 0.5448, df = 2, p-value = 0.7615
alternative hypothesis: true common location is not equal to c(0,0)
```

The pairwise version of the test can be performed by taking first the differences of the data, as in

```
> sr.loc.test(frets[,c(1,2)]-frets[,c(3,4)],score="rank")

        One sample location test using spatial signed ranks

data:  frets[, 1:2] - frets[, 3:4]
Q.2 = 2.8236, df = 2, p-value = 0.2437
alternative hypothesis: true location is not equal to c(0,0)
```

In case of several samples (also alternatively in the two sample case) the subsamples are given by a factor. For example for the `iris` data,

```
> sr.loc.test(iris[,1:4],g=iris[,5])

        Several samples location test using spatial ranks

data:  iris[, 1:4] by iris[, 5]
Q.2 = 171.4739, df = 8, p-value < 2.2e-16
alternative hypothesis: true common location is not equal to c(0,0,0,0)
```

4.4 Multivariate regression

The sign based affine equivariant regression coefficients for the `frets` data, predicting the younger brother's head measures by those of the older brother are computed by (note that a data frame is not usable with `formula`)

```
> frets.y<-as.matrix(frets[,c(3,4)])
> frets.o<-as.matrix(frets[,c(1,2)])
> sr.regression(frets.y~frets.o)
                  [,1]       [,2]
(Intercept) 36.4708397 53.7866582
frets.ol1    0.3576193  0.2522857
frets.ob1    0.5421662  0.3215173
```

or, the non affine equivariant rank based coefficients without the intercept term

```
> sr.regression(frets.y~frets.o-1,score="rank",ae=FALSE)
                 [,1]      [,2]
(Intercept)        NA        NA
frets.ol1   0.4097044 0.2718346
frets.ob1   0.5513059 0.3465670
```

Note that based on ranks it is not possible to estimate coefficients of a model without an intercept term, because of the inherent centering of the ranks. The intercept, if it is required, is always estimated separately and if it is not required the result include a row of `NA` as the intercept term, as above.

4.5 Testing for independence

The test for the independence of the head measurements between brothers based on signs is performed by

```
> sr.indep.test(frets[,c(1,2)],frets[,c(3,4)])

        Multivariate independence test using spatial signs

data:  frets[, c(1, 2)] and frets[, c(3, 4)]
Q.2 = 13.7155, df = 4, p-value = 0.00826
alternative hypothesis: true measure of dependence is not equal to 0
```

which produces also two warning messages because observations too close to center of symmetry could not be used in internal shape estimation. Here the use of a factor is reasonable:

```
> sr.indep.test(frets,g=gl(2,2),score="symmsign")

        Multivariate independence test using spatial symmetrized signs

data:  frets by gl(2, 2)
Q.2 = 17.5997, df = 4, p-value = 0.001477
alternative hypothesis: true measure of dependence is not equal to 0
```

which tests for the independence using symmetrised signs. For the same reason as above there is also a warning message.

4.6 Testing for sphericity

Finally, testing for sphericity of the head measurements of brothers based on signs is done simply as

```
> sr.sphere.test(frets)

        Test of sphericity using spatial signs

data:  frets
Q.2 = 224.7055, df = 9, p-value < 2.2e-16
alternative hypothesis: true shape is not equal to diag(4)
```

Testing based on symmetrised signs yields a similar result:

```
> sr.sphere.test(frets,score="symmsign")

        Test of sphericity using spatial symmetrized signs

data:  frets
Q.2 = 114.4421, df = 9, p-value < 2.2e-16
alternative hypothesis: true shape is not equal to diag(4)
```

References

N Blomqvist. On a measure of dependence between two random variables. *Annals of Mathematical Statistics*, 21:593–600, 1950.

BM Brown. Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B*, 45:25–30, 1983.

K Choi and J Marden. An approach to multivariate rank tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 92: 1581–1590, 1997.

L Dümbgen. On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, 50:471–491, 1998.

JC Gower. The mediancentre. *Applied Statistics*, 23:466–470, 1974.

M Hallin and D Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Annals of Statistic*, 30: 1103–1133, 2002.

M Hallin and D Paindaveine. Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Annals of Statistics*, 34:2707–2756, 2006.

M Hallin, H Oja, and D Paindaveine. Semiparametrically efficient rank-based inference for shape. II. optimal R-estimation of shape. *Annals of Statistics*, 34:2757–2789, 2006.

TP Hettmansperger and JC Aubuchon. Comment on "rank-based robust analysis of linear models. I. Exposition and review" by David Draper. *Statistical Science*, 3:262–263, 1988.

TP Hettmansperger and RH Randles. A practical affine equivariant multivariate median. *Biometrika*, 89:851–860, 2002.

JI Marden. Multivariate rank tests. In S Ghosh, editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 401–432. CRC, 1999a.

JI Marden. Some robust estimates of principal components. *Statistics & Probability Letters*, 43:349–359, 1999b.

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

J Möttönen, H Oja, and J Tienari. On the efficiency of multivariate spatial methods. *Annals of Statistics*, 25:542–552, 1997.

H Oja. Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics*, 26:319–343, 1999.

H Oja and R Randles. Multivariate nonparametric tests. *Statistical Science*, 19:598–605, 2004.

ML Puri and PK Sen. *Nonparametric Methdos in Multivariate Analysis*. Wiley, New York, 1971.

R Randles. A simpler, affine-invariant, multivariate, distribution-free sign test. *Journal of the American Statistical Association*, 95:1263–1268, 2000.

R Randles. A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84:1045–1050, 1989.

S Sirkiä, S Taskinen, H Oja, and D Tyler. Tests and estimates of shape based on spatial signs and ranks. *submitted*, 2007.

C Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

DE Tyler. A distribution-free M-estimate of multivariate scatter. *Annals of Statistics*, 15:234–251, 1987.

S Visuri, V Koivunen, and H Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91:557–575, 2000.

SS Wilks. On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3:309–326, 1935.