Journal of Statistical Software
July 2011, Volume 43, Issue 5. http://www.jstatsoft.org/

Multivariate L_1 Methods: The Package MNM

Klaus Nordhausen
University of Tampere

Hannu Oja
University of Tampere

Abstract

In the paper we present an R package **MNM** dedicated to multivariate data analysis based on the L_1 norm. The analysis proceeds very much as does a traditional multivariate analysis. The regular L_2 norm is just replaced by different L_1 norms, observation vectors are replaced by their (standardized and centered) spatial signs, spatial ranks, and spatial signed-ranks, and so on. The procedures are fairly efficient and robust, and no moment assumptions are needed for asymptotic approximations. The background theory is briefly explained in the multivariate linear regression model case, and the use of the package is illustrated with several examples using the R package **MNM**.

Keywords: least absolute deviation, mean deviation, mean difference, multivariate linear regression, R, shape matrix, spatial sign, spatial signed-rank, spatial rank, transformation-retransformation method.

1. Introduction

Classical multivariate statistical inference methods (Hotelling's T^2, multivariate analysis of variance, multivariate regression, tests for independence, canonical correlation analysis, principal component analysis, and so on) are based on the use of the L_2 norm. These standard moment-based multivariate techniques are optimal under the multivariate normality of the residuals but poor in their efficiency for heavy-tailed distributions. They are also highly sensitive to outlying observations. In this paper we present an R package **MNM** – available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=MNM` – which uses different L_1 norms and the corresponding scores (spatial signs, spatial signed-ranks, and spatial ranks) in the analysis of multivariate data. The theory of the multivariate L_1 methodology is explained in details in Oja (2010).

We briefly explain the approach in the multivariate multiple linear regression model setting. This is necessary for the correct use of the arguments of the functions (`score`, `stand`, etc.) in **MNM**. Let $(\mathbf{X}, \mathbf{Y})$ be the $n \times (q + p)$ data matrix where $\mathbf{X}$ is the matrix of q explaining

variables and $\mathbf{Y}$ the matrix of p-variate response variable. We assume that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E},$$

where $\boldsymbol{\beta}$ is an $q \times p$ matrix of regression coefficients, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^\top$ is a random sample of p-variate residuals "centered" at the origin. In this paper, L_1 objective functions are used to find an estimate for the unknown $\boldsymbol{\beta}$. We also consider the partitioned model

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{E}$$

where $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) is a $n \times q_1$ (resp. $n \times q_2$) matrix. The null hypothesis $H_0: \boldsymbol{\beta}_2 = 0$ can then be tested using the score functions corresponding to L_1 norms. Moreover, if we are also interested in the scatter matrix estimation or testing, we may use the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\Omega}^\top,$$

where the residuals in $\mathbf{E}$ are now "centered and standardized" in a certain way. The matrix $\boldsymbol{\Sigma} = \Omega\Omega^\top$ is the scatter (or shape) matrix of the residuals in the regression model. Note that the classical one-sample and several-sample location problems and the one-sample scatter problem are simple but important special cases here. See Chapters 6 to 9 and Chapter 11 in Oja (2010). In the book, also the problem of testing independence between the subvectors (Chapter 10) and the analysis of data from a randomized block design (Chapter 12) are considered. Therefore in **MNM**, following the presentation in the book, there are own functions for these cases as well.

The tests and estimates for the multivariate location problem based on multivariate spatial signs, signed-ranks, and ranks have been widely discussed in the literature. See, for example, Möttönen and Oja (1995), Choi and Marden (1997), Marden (1999), and Oja and Randles (2004). The scatter matrix estimates by Tyler (1987) and Dümbgen (1998) are often used for robust standardization of the data. The location tests and estimates are robust and they have good efficiency properties even in the multivariate normal model (Möttönen *et al.* 1997). The work in the area is collected together in Oja (2010).

We wish to mention that the procedures based on spatial signs and ranks, however, offer only one possible multivariate extension of nonparametric tests (sign test, rank test) and corresponding estimates (median, Hodges-Lehmann estimate). Randles (1989) followed by a series of papers, for example, develop multivariate nonparametric tests based on so-called *interdirections*. These tests are typically asymptotically equivalent to spatial sign and rank tests described here but, unfortunately, computationally heavy. The multivariate inference methods based on marginal signs and ranks are described in detail in the monograph by Puri and Sen (1971). The R package **ICSNP** (Nordhausen *et al.* 2010) provides some tools for the tests based on marginal signs and ranks, including affine invariant modifications of the tests (see for example Nordhausen *et al.* 2008). Still another extension which is based on the affine equivariant signs and ranks is described in Oja (1999). For some implementations of this approach, see the R package **OjaNP** (Fischer *et al.* 2010). There exists a scattered collection of functions for univariate sign and rank methods; no general R package is available so far. However base R contains many tests and estimates as does the package **exactRankTests** (Hothorn and Hornik 2011) and its successor **coin** (Hothorn *et al.* 2006, 2008). For univariate regressions based on signs and ranks see among others the packages **Rfit** (Kloke 2010) and **quantreg** (Koenker 2011). Some aspects of using R for univariate analysis based on signs and

ranks are for example covered in Hettmansperger and McKean (2010), Terpstra and McKean (2005) and Wilcox (2010) to name a few.

The goal of the paper is to explain the use of the L_1 methods based on the spatial signs and ranks in the analysis of multivariate data and illustrate how the analysis can be implemented using the R package **MNM**. The structure of the paper is as follows. In the next Section 2 three multivariate L_1 objective functions and the corresponding score functions are discussed. The use of these score functions in the general multivariate linear regression case (with inner and outer standardization) is explained in Section 3. For other cases the reader is referred to Oja (2010). In Section 4 the main functions of the R package **MNM** are described. In Section 5 the use of the functions is illustrated with several examples. A summary is given in Section 6.

2. Multivariate L_1 objective functions and score functions

2.1. L_1 objective functions

For estimation and testing, write $\mathbf{e}_i = \mathbf{e}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i$, $i = 1, \ldots, n$. The regular least-squares (LS) estimate minimizes the L_2 criterion function $D_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n\{||\mathbf{e}_i||^2\} = \frac{1}{n}\sum_{i=1}^n\{\mathbf{e}_i^\top \mathbf{e}_i\}$. In this paper we consider the L_1 type criterion functions

$$
\begin{aligned}
D_{1n}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n\{\|\mathbf{e}_i\|\},\\
D_{2n}(\boldsymbol{\beta}) &= \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n\{\|\mathbf{e}_i - \mathbf{e}_j\|\}, \quad \text{and}\\
D_{3n}(\boldsymbol{\beta}) &= \frac{1}{4n^2}\sum_{i=1}^n\sum_{j=1}^n\{\|\mathbf{e}_i - \mathbf{e}_j\| + \|\mathbf{e}_i + \mathbf{e}_j\|\}.
\end{aligned}
$$

Multivariate spatial sign and spatial rank methods are based on the L_1 objective functions D_{1n}, D_{2n}, and D_{3n} and the corresponding score functions. The first objective function $\frac{1}{n}\sum_{i=1}^n\{\|\mathbf{e}_i\|\}$ is the *mean deviation* of the residuals from the origin, and it is the basis for the so called least absolute deviation (LAD) methods. It yields different median-type estimates and spatial sign tests in the one-sample, several-sample and finally general linear model settings. The second objective function $(1/(2n^2))\sum_{i=1}^n\sum_{j=1}^n\{\|\mathbf{e}_i - \mathbf{e}_j\|\}$ is the *mean difference* of the residuals which in fact measures how close together the residuals are. The second and third objective functions generate Hodges-Lehmann type estimates and rank tests for different location problems.

2.2. Spatial sign, spatial rank, and spatial signed-rank

Let

$$
\begin{aligned}
\mathbf{U}(\mathbf{e}) &= \|\mathbf{e}\|^{-1}\mathbf{e}, \quad \text{if } \mathbf{e} \neq \mathbf{0}\\
&= \mathbf{0}, \qquad\quad \text{if } \mathbf{e} = \mathbf{0}.
\end{aligned}
$$

The multivariate spatial sign $\mathbf{U}_i$, multivariate spatial signed-rank $\mathbf{Q}_i$, and multivariate spatial (centered) rank $\mathbf{R}_i$ of the residual $\mathbf{e}_i$, $i = 1, \ldots, n$, are defined as

$$\begin{aligned} \mathbf{U}_i &= \mathbf{U}(\mathbf{e}_i), \\ \mathbf{R}_i &= \frac{1}{n}\sum_{j=1}^{n}\{\mathbf{U}(\mathbf{e}_i - \mathbf{e}_j)\}, \quad \text{and,} \\ \mathbf{Q}_i &= \frac{1}{2n}\sum_{j=1}^{n}\{\mathbf{U}(\mathbf{e}_i - \mathbf{e}_j) + \mathbf{U}(\mathbf{e}_i + \mathbf{e}_j)\}. \end{aligned}$$

In the univariate case, one gets just regular sign, (centered) rank, and signed-rank. The three objective functions D_{1n}, D_{2n}, and D_{3n} then satisfy

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\{||\mathbf{e}_i||\} &= \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{U}_i^\top \mathbf{e}_i\}, \\ \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\|\mathbf{e}_i - \mathbf{e}_j\|\} &= \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{R}_i^\top \mathbf{e}_i\}, \quad \text{and} \\ \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\|\mathbf{e}_i - \mathbf{e}_j\| + \|\mathbf{e}_i + \mathbf{e}_j\|\} &= \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{Q}_i^\top \mathbf{e}_i\}. \end{aligned}$$

2.3. Multivariate score functions in MNM

The general strategy in the analysis of the multivariate data is first to replace the residuals $\mathbf{e}_i$ by some scores $\mathbf{T}_i = \mathbf{T}(\mathbf{e}_i)$ or, in more complex designs, the estimated residuals $\hat{\mathbf{e}}_i$ by centered and/or standardized scores $\hat{\mathbf{T}}_i = \mathbf{T}(\hat{\mathbf{e}}_i)$, $i = 1, \ldots, n$. (The estimated residuals are $\hat{\mathbf{e}}_i = \mathbf{y}_i - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i$ or sometimes $\hat{\mathbf{e}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i)$ where $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are estimated under a full model or under a restricted model depending on the problem at hand.) The statistical tests are then based on the new data matrix

$$\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_n)^\top \qquad \text{or} \qquad \hat{\mathbf{T}} = (\hat{\mathbf{T}}_1, \ldots, \hat{\mathbf{T}}_n)^\top.$$

The package **MNM** uses the score functions

$$\begin{aligned} \mathbf{T}(\mathbf{e}) &= \mathbf{e} \quad \text{(identity score)}, \\ &= \mathbf{U}(\mathbf{e}) \quad \text{(spatial sign)}, \\ &= \mathbf{R}(\mathbf{e}) = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{U}(\mathbf{e} - \mathbf{e}_i)\} \quad \text{(spatial rank), and} \\ &= \mathbf{Q}(\mathbf{e}) = \frac{1}{2n}\sum_{i=1}^{n}\{\mathbf{U}(\mathbf{e} - \mathbf{e}_i) + \mathbf{U}(\mathbf{e} + \mathbf{e}_i)\} \quad \text{(spatial signed-rank)}. \end{aligned}$$

The spatial sign score $\mathbf{U}_i = \mathbf{U}(\mathbf{e}_i)$, the spatial rank score $\mathbf{R}_i = \mathbf{R}(\mathbf{e}_i)$, and the spatial signed-rank score $\mathbf{Q}_i = \mathbf{Q}(\mathbf{e}_i)$ thus correspond to the three L_1 criterion functions as explained in the previous section. Inner centering and/or standardization are used in **MNM** to attain the affine invariance property of the tests and the affine equivariance property of the estimates.

An estimate for the scatter (or shape) matrix of the residuals is then obtained as a side product.

2.4. Important matrices

For theoretical studies we often need to know the matrices

$$\mathbf{A} = E\{\mathbf{T}(\mathbf{e}_i)\mathbf{L}(\mathbf{e}_i)^\top\} \quad \text{and} \quad \mathbf{B} = E\left\{\mathbf{T}(\mathbf{e}_i)\mathbf{T}(\mathbf{e}_i)^\top\right\},$$

where $\mathbf{L}(\mathbf{e}) = -\nabla \log f(\mathbf{e})$ is the optimal location score for the density of the residuals f. Then one can show, see e.g., Möttönen *et al.* (1997) and Chapter 8 in Oja (2010), that, with distinct i, j, and k,

- for the identity score,

$$\mathbf{A} = E\left[\mathbf{e}_i\mathbf{e}_i^\top\right] \quad \text{and} \quad \mathbf{B} = E\left[\mathbf{e}_i\mathbf{e}_i^\top\right],$$

- for the spatial sign score,

$$\mathbf{A} = E\left[\frac{1}{\|\mathbf{e}_i\|}\left(\mathbf{I}_p - \frac{\mathbf{e}_i\mathbf{e}_i^\top}{\|\mathbf{e}_i\|^2}\right)\right] \quad \text{and} \quad \mathbf{B} = E\left[\frac{\mathbf{e}_i\mathbf{e}_i^\top}{\|\mathbf{e}_i\|^2}\right],$$

- for the spatial rank score,

$$\mathbf{A} = E\left[\frac{1}{\|\mathbf{e}_i - \mathbf{e}_j\|}\left(\mathbf{I}_p - \frac{(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^\top}{\|\mathbf{e}_i - \mathbf{e}_j\|^2}\right)\right] \quad \text{and} \quad \mathbf{B} = E\left[\frac{(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_k)^\top}{\|\mathbf{e}_i - \mathbf{e}_j\| \cdot \|\mathbf{e}_i - \mathbf{e}_k\|}\right].$$

- for the spatial signed-rank score,

$$\mathbf{A} = E\left[\frac{1}{\|\mathbf{e}_i + \mathbf{e}_j\|}\left(\mathbf{I}_p - \frac{(\mathbf{e}_i + \mathbf{e}_j)(\mathbf{e}_i + \mathbf{e}_j)^\top}{\|\mathbf{e}_i + \mathbf{e}_j\|^2}\right)\right]$$

and

$$\mathbf{B} = \frac{1}{4}E\left[\left(\frac{\mathbf{e}_i - \mathbf{e}_j}{\|\mathbf{e}_i - \mathbf{e}_j\|} - \frac{\mathbf{e}_i + \mathbf{e}_j}{\|\mathbf{e}_i + \mathbf{e}_j\|}\right)\left(\frac{\mathbf{e}_i - \mathbf{e}_k}{\|\mathbf{e}_i - \mathbf{e}_k\|} - \frac{\mathbf{e}_i + \mathbf{e}_k}{\|\mathbf{e}_i + \mathbf{e}_k\|}\right)^\top\right].$$

Of course, assumptions are needed for the existence of the matrix $\mathbf{A}$. See Section 3.1 for these assumptions. Note that natural estimates of $\mathbf{A}$ and $\mathbf{B}$ are obtained by replacing, in the above formulae, the expected values by the averages and the residual $\mathbf{e}_i$ by estimated residuals $\hat{\mathbf{e}}_i$, $i = 1, \ldots, n$. In the following, the theory is presented using a general score function $\mathbf{T}(\mathbf{e})$.

3. Multivariate linear regression model

3.1. Model and assumptions

We consider the data matrix $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X}$ is a $n \times q$ matrix of explaining variables (fixed) and $\mathbf{Y}$ is a $n \times p$ matrix of response variables. The multivariate linear regression model is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\Omega}^\top,$$

where $\boldsymbol{\beta}$ is a $q \times p$ matrix of unknown regression coefficients, $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$ is a scatter matrix, and $\mathbf{E}$ is an $n \times p$ matrix of unobserved centered and standardized residuals. The following assumptions are needed for asymptotic approximations, that is, for the limiting distributions of the test statistics and the estimates. In practical data analysis with the package **MNM**, the validity of the asymptotic p values and the covering probabilities of the confidence ellipsoids thus depends on whether the assumptions hold. In some cases, if the assumptions are not true, one can still apply permutation versions of the tests or use bootstrapping techniques to estimate the accuracy of the estimates.

Design assumptions: *The $n \times q$ design matrix (sequence) $\mathbf{X}$ satisfies*

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \mathbf{D} \quad \textit{and} \quad \frac{\max_{1\le i\le n}\{\|\mathbf{C}\mathbf{x}_i\|^2\}}{\sum_{i=1}^n\{\|\mathbf{C}\mathbf{x}_i\|^2\}} \to 0, \qquad \textit{as } n \to \infty,$$

for some positive definite $q \times q$ matrix $\mathbf{D}$ and for all $p \times q$ matrices $\mathbf{C}$.

Distributional assumptions: *The rows of $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ are i.i.d. from*

(i) a distribution with $E(\mathbf{e}_i) = \mathbf{0}$ and $E(\mathbf{e}_i\mathbf{e}_i^\top) = \mathbf{I}_p$ (identity score), or

(ii) a continuous distribution with bounded density (spatial sign, rank, and signed-rank scores), standardized so that $E(\mathbf{T}(\mathbf{e}_i)) = \mathbf{0}$ and $E(\mathbf{T}(\mathbf{e}_i)\mathbf{T}(\mathbf{e}_i)^\top) \propto \mathbf{I}_p$.

It is important to note that, in our approach, the parameters are fixed so that the transformed residuals (not the original ones) are standardized. Note also that no moment assumptions are needed for the spatial sign, signed-rank, or rank methods.

3.2. Testing problem I: Inner and outer standardization

We wish first to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$. (Of course, the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ may be tested just by replacing $\mathbf{y}_i$ by $\mathbf{y}_i - \boldsymbol{\beta}_0^\top\mathbf{x}_i$.) Write $\mathbf{T}_i = \mathbf{T}(\mathbf{y}_i)$ and $\mathbf{T}_i(\boldsymbol{\beta}) = \mathbf{T}(\mathbf{y}_i - \boldsymbol{\beta}^\top\mathbf{x}_i)$, $i = 1, \ldots, n$, and

$$\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_n)^\top \quad \text{and} \quad \mathbf{T}(\boldsymbol{\beta}) = (\mathbf{T}_1(\boldsymbol{\beta}), \ldots, \mathbf{T}_n(\boldsymbol{\beta}))^\top.$$

Then, under our assumptions and under the null hypothesis, $n^{-1/2}\text{vec}\,(\mathbf{T}^\top\mathbf{X}) \to_d N_{pq}(\mathbf{0}, \mathbf{D} \otimes \mathbf{B})$ and the test statistic

$$Q^2 = Q^2(\mathbf{X}, \mathbf{Y}) = n \cdot tr\,(\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{T}) \ \to_d \chi^2_{pq}.$$

where $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $\mathbf{P}_\mathbf{T} = \mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top$ are $n \times n$ projection matrices. Note that Q^2 depends on $\mathbf{Y}$ through $\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1/2}$. The $n \times p$ matrix $\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1/2}$ gives *outer standardized scores* as $[\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1/2}]^\top[\mathbf{T}(\mathbf{T}^\top\mathbf{T})^{-1/2}] = \mathbf{I}_p$.

For the spatial sign and rank score, the test with outer standardized scores is not necessarily affine invariant, however. Affine invariance means here that $Q^2(\mathbf{X}\mathbf{V}, \mathbf{Y}\mathbf{W}) = Q^2(\mathbf{X}, \mathbf{Y})$, for all nonsingular $q\times q$ and $p\times p$ matrices $\mathbf{V}$ and $\mathbf{W}$, respectively. An affine invariant modification of the test statistic is obtained using *inner standardization* of the scores as follows.

1. Find $\mathbf{S}^{-1/2}$ such that if $\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}\mathbf{y}_i)$ then $\hat{\mathbf{T}}^\top\hat{\mathbf{T}} \propto \mathbf{I}_p$.
   This is called *inner standardization.*

2. The invariant test statistic is then $Q^2 = n \cdot tr\left(\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\hat{\mathbf{T}}}\right)$.

The symmetric matrix $\mathbf{S}$ satisfying $\mathbf{S}^{-1/2}\mathbf{S}(\mathbf{S}^{-1/2})^\top = \mathbf{I}_p$ is then the corresponding scatter (or shape) matrix estimate for $\mathbf{\Sigma} = \mathbf{\Omega}\mathbf{\Omega}^\top$. For the spatial sign score, $\mathbf{S}$ is Tyler's shape matrix, and $\mathbf{S}^{-1/2}$ is Tyler's transformation. See Tyler (1987).

3.3. Estimation problem: With and without inner standardization

Next we wish to estimate the unknown $q \times p$ matrix $\boldsymbol{\beta}$. The estimate $\hat{\boldsymbol{\beta}}$ based on score function $\mathbf{T}$ solves

$$\mathbf{T}(\hat{\boldsymbol{\beta}})^\top \mathbf{X} = \mathbf{0}.$$

Then, under general assumptions,

$$\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\to_d\; N_{qp}\left(\mathbf{0}, \mathbf{D}^{-1} \otimes (\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})\right)$$

where $\mathbf{A}$ and $\mathbf{B}$ were given in Section 2.4. For the practical estimation of the covariance matrix of $\mathrm{vec}(\hat{\boldsymbol{\beta}})$, estimates of $\mathbf{A}$ and $\mathbf{B}$ are easily available as described in Section 2.4.

Different scores then yield the following estimates.

- Identity score: The regular LS estimate which minimizes $\sum_{i=1}^n \|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i\|^2$.
- Spatial sign score: The multivariate least absolute deviation (LAD) estimate which minimizes $\sum_{i=1}^n \|\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i\|$.
- Spatial rank score: The multivariate mean difference (MD) estimate which minimizes $\sum_{i=1}^n \sum_{j=1}^n \|(\mathbf{y}_i - \mathbf{y}_j) - \boldsymbol{\beta}^\top(\mathbf{x}_i - \mathbf{x}_j)\|$.

The spatial signed-rank score is used only in the one-sample location case, and it gives the multivariate Hodges-Lehmann location estimate. See Chapter 7 in Oja (2010). Note that the spatial rank score does not yield an estimate for the intercept vector. The spatial signed-rank score applied to the estimated residuals can then be used for the estimation of the intercept parameter.

The regular LS estimate is fully regression equivariant. For the concept of regression equivariance, see e.g., Ollila *et al.* (2002). The LAD and MD estimates can be made affine equivariant using *inner standardization* as follows. Find a transformation matrix $\mathbf{S}^{-1/2}$ and $\hat{\boldsymbol{\beta}}$ such that if $\hat{\mathbf{e}}_i = \mathbf{S}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i)$ and $\hat{\mathbf{T}}_i = \mathbf{T}(\hat{\mathbf{e}}_i)$, $i = 1, \ldots, n$, then simultaneously

$$\hat{\mathbf{T}}^\top \mathbf{X} = \mathbf{0} \;\text{ and }\; \hat{\mathbf{T}}^\top \hat{\mathbf{T}} \propto \mathbf{I}_p.$$

Then $\hat{\boldsymbol{\beta}}$ is affine equivariant and $\mathbf{S}$ is the scatter/shape estimate of $\mathbf{\Sigma}$ based on the score function $\mathbf{T}$.

In the package **MNM**, an equivariant LAD estimate, for example, is calculated using a fixed point algorithm as follows. First the residuals, second the regression coefficient matrix, and finally the residual scatter matrix are updated using repeatedly the following three steps.

1. $\mathbf{e}_i \leftarrow \mathbf{S}^{-1/2}(\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i), \quad i = 1, \ldots, n$
2. $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \left[\sum_{i=1}^n \{\|\mathbf{e}_i\|^{-1}\mathbf{x}_i\mathbf{x}_i^\top\}\right]^{-1} \sum_{i=1}^n \{\mathbf{x}_i \mathbf{U}(\mathbf{e}_i)^\top\}\mathbf{S}^{1/2}$
3. $\mathbf{S} \leftarrow \frac{p}{n}\, \mathbf{S}^{1/2} \left[\sum_{i=1}^n \{\mathbf{U}(\mathbf{e}_i)\mathbf{U}(\mathbf{e}_i)^\top\}\right] \mathbf{S}^{1/2}.$

If inner standardization is not used then one just repeats steps 1 and 2 with $\mathbf{S} = \mathbf{I}_p$. In the one-sample location case one then gets (i) the spatial median (without inner standardization), or (ii) the Hettmansperger-Randles estimate (with inner standardization). The Hettmansperger-Randles estimate combines the Tyler's transformation and the spatial median. See Hettmansperger and Randles (2002). In **MNM**, similar algorithms are used for the calculation of the value of the MD estimate (with and without inner standardization) as well.

3.4. Testing problem II: Inner and outer standardization

Consider now the partitioned model

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

where $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) is a $n \times q_1$ (resp. $n \times q_2$) matrix. We wish to test the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. (i) If the *outer standardization* is used, one finds scores $\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_{i1})$, $i = 1, \ldots, n$, such that $\hat{\mathbf{T}}^\top \mathbf{X}_1 = \mathbf{0}$. (ii) In the *inner standardization*, the standardized scores $\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_{i1}))$, $i = 1, \ldots, n$, satisfy both $\hat{\mathbf{T}}^\top \mathbf{X}_1 = \mathbf{0}$ and $\hat{\mathbf{T}}^\top \hat{\mathbf{T}} \propto \mathbf{I}_p$. The *score test statistic* is now

$$Q^2 = n \cdot tr\left(\mathbf{P}_{\hat{\mathbf{X}}_2}\mathbf{P}_{\hat{\mathbf{T}}}\right)$$

where $\hat{\mathbf{X}}_2 = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$. With the inner standardization, the test is fully invariant. Under the null hypothesis, the test statistic has an approximate χ^2 distribution with $q_2 p$ degrees of freedom. The *Wald-type test statistic* which uses $\mathrm{vec}(\hat{\boldsymbol{\beta}}_2)$ and its estimated covariance matrix in the full model is asymptotically equivalent with the score test statistic.

3.5. Inference for shape

Let $\mathbf{K}_{p,p}$ be the commutation matrix, that is, a $p^2 \times p^2$ block matrix with (i, j)-block being equal to a $p \times p$ matrix that has one at entry (j, i) and zero elsewhere, and $\mathbf{J}_{p,p}$ for $\mathrm{vec}(\mathbf{I}_p)\mathrm{vec}(\mathbf{I}_p)^\top$. Matrix

$$\mathbf{C}_{p,p} = \frac{1}{2}(\mathbf{I}_{p^2} + \mathbf{K}_{p,p}) - \frac{1}{p}\mathbf{J}_{p,p}$$

projects a vectorized matrix $\mathrm{vec}(\mathbf{A})$ to the space of symmetrical and centered vectorized matrices. The tests and estimates for the shape parameter are based on the squared norm of such a projection,

$$Q^2(\mathbf{A}) = \|\mathbf{C}_{p,p}\mathrm{vec}(\mathbf{A})\|^2,$$

which is proportional to the variance of the eigenvalues of a symmetrized version of $\mathbf{A}$. For symmetrical positive definite $p \times p$ matrices $\mathbf{A}$, it then holds that $Q^2(\mathbf{A}) = 0$ if and only if $\mathbf{A} \propto \mathbf{I}_p$.

For simplicity, assume that $\boldsymbol{\beta} = \mathbf{0}$ and we wish to estimate unknown $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$ and test the null hypothesis $H_0 : \boldsymbol{\Sigma} \propto \mathbf{I}_p$. Matrix $\boldsymbol{\Sigma}$ is then also defined by the condition

$$E\left(\mathbf{T}(\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i)\mathbf{T}(\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i)^\top\right) \propto \mathbf{I}_p.$$

The algorithm for the estimate $\mathbf{S} = \mathbf{S}(\mathbf{Y})$ of $\boldsymbol{\Sigma}$ (up to a multiplying constant) with respect to the origin then uses the steps (i) $\hat{\mathbf{T}}_i \leftarrow \mathbf{T}(\mathbf{S}^{-1/2}\mathbf{y}_i)$, $i = 1, \ldots, n$, and (ii) $\mathbf{S} \leftarrow [p/tr(\hat{\mathbf{T}}^\top\hat{\mathbf{T}})]\mathbf{S}^{1/2}\hat{\mathbf{T}}^\top\hat{\mathbf{T}}\mathbf{S}^{1/2}$.

The test statistic for testing $H_0: \mathbf{\Sigma} \propto \mathbf{I}_p$ (hypothesis of sphericity) is simply

$$Q^2\left(n^{-1}\mathbf{T}^\top\mathbf{T}\right) = \left\|\mathbf{C}_{p,p}\text{vec}\left(n^{-1}\mathbf{T}^\top\mathbf{T}\right)\right\|^2.$$

Under the null hypothesis,

$$(n/\tau)Q^2 \ \rightarrow_d \ \chi^2_{(p+2)(p-1)/2}$$

where τ is sometimes unknown and has to be estimated. Recall that in our approach $n^{-1}\mathbf{T}^\top\mathbf{T}$ is the regular covariance matrix, spatial sign covariance matrix, or rank covariance matrix depending on which score function is chosen.

4. R package MNM (Multivariate Nonparametrical Methods)

4.1. General features

The package provides multivariate tests and estimates and other procedures based on the (i) identity score, (ii) spatial sign score, and (iii) spatial rank score. Most functions in the package have an argument `score` which can be set to `"identity"`, `"sign"` or `"rank"` (or sometimes also to `"symmsign"` for symmetrized signs). Let $\mathbf{T}(\mathbf{e})$ be the chosen score function. In the procedures, the user can also choose between inner and outer standardization. In the outer standardization the model is

$$\mathbf{y}_i = \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{e}_i, \quad \text{where } E(\mathbf{T}(\mathbf{e}_i)) = \mathbf{0}$$

and in the inner standardization one assumes that

$$\mathbf{y}_i = \boldsymbol{\beta}^\top\mathbf{x}_i + \mathbf{\Omega}\mathbf{e}_i, \quad \text{where } E(\mathbf{T}(\mathbf{e}_i)) = \mathbf{0} \text{ and } E(\mathbf{T}(\mathbf{e}_i)\mathbf{T}(\mathbf{e}_i)^\top) \propto \mathbf{I}_p.$$

The argument `stand` should then be `"outer"` or `"inner"` for outer or inner standardization, respectively. The default values are `score = "identity"` and `stand = "outer"`. For different standardizations, see also Oja (2010).

For most testing functions, the p values can be based on limiting distributions of the test statistic (`method = "approximation"`). The test is then asymptotically distribution-free. Another possibility is to use conditionally distribution-free test versions (`method = "permutation"` or `method = "signchange"`) which is based on permutation or sign-change arguments, respectively. See again Oja (2010) for more details. The approximations based on the limiting distributions may not be good for small sample sizes. (As far as we know, there are no simulation studies to consider this problem.) Therefore permutation tests are very much recommended for the p value calculation with small sample sizes. In general a comparison of asymptotic and permutation based p values is a good strategy when analyzing data, Also bootstrapping could be used to estimate the accuracy of the estimates but is not yet available in **MNM**.

Many formulas of the tests and estimates mentioned above contain matrix inverses. In the implementation of the methods we avoid computing the explicit inverse when possible. However, if the same inverse matrix is used repeatedly like for example in the p value calculation for the permutation tests, we choose to compute it. (Note that, for the permutation tests,

we need to compute the inverse only once.) Since most of the matrices to be inverted are symmetric, the inverses are found via Cholesky decomposition when appropriate.

Note that the functions which use ranks, signed-ranks, and symmetrized signs may sometimes be slow and memory consuming since they operate with pairwise differences and pairwise sums of observation vectors.

We will now describe the main functions of the package in more detail.

4.2. One-sample location problem

The one-sample location estimates with their estimated covariance matrices are given by the function `mv.1sample.est`. One then gets the regular mean vector, the spatial median, or the multivariate Hodges-Lehmann estimate with outer or inner standardization depending on the values of the options for `score` and `stand`. Choices `score = "sign"` and `stand = "outer"` give the spatial median, and choices `score = "sign"` and `stand = "inner"` the Hettmansperger-Randles estimate, for example. In the one-sample case, the option `"rank"` refers to signed-ranks.

The function `mv.1sample.est` returns a list of class `mvloc` with a location estimate as a component `location` and its estimated covariance matrix as a component `vcov`. For objects in this class we provide `print`, `summary` and `plot` methods. For the comparisons of different location estimates, the function `plot` produces a simultaneous scatter plot matrix for up to three different location estimates with their estimated confidence ellipsoids.

The function `mv.1sample.test` can be used to test the null hypothesis that the observations come from a distribution symmetric around the origin. The null value can be respecified with the argument `mu`. Depending again on the values of the arguments `score` one gets either the Hotellings T^2-test, or the spatial sign test, or the spatial signed-ranked test. The choice `stand = "inner"` makes the latter two tests affine invariant. Note that the Hotelling's T^2-test version implemented here slightly differs from the regular version given in most textbooks (implemented as `HotellingsT2` in the package **ICSNP**, Nordhausen *et al.* 2010, for example); the covariance matrix for the test statistic is here computed with respect to the null value `mu`.

Main references for the tests and estimates in the one-sample location case are the papers by Chaudhuri (1992), Möttönen and Oja (1995), Randles (2000), Vardi and Zhang (2000), Hettmansperger and Randles (2002), and Oja and Randles (2004). See also Chapters 5 to 8 in Oja (2010).

4.3. One-sample shape problem

The function `mv.shape.est` needs arguments `score`, `estimate`, and `location`. One estimates $\mathbf{\Sigma}$ satisfying either

$$\Sigma^{-1/2} E(\mathbf{T}(\mathbf{y}_i - \boldsymbol{\mu})\mathbf{T}(\mathbf{y}_i - \boldsymbol{\mu})^\top)(\Sigma^{-1/2})^\top = \mathbf{I}_p$$

or

$$E(\mathbf{T}(\mathbf{\Sigma}^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}))\mathbf{T}(\mathbf{\Sigma}^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}))^\top) \propto \mathbf{I}_p$$

depending on whether `estimate = "outer"` or `estimate = "inner"`. The observations are centered by natural companion location estimates $\hat{\boldsymbol{\mu}}$ if not otherwise stated by the argument `location`. For example, the choices `score = "sign"` and `estimate = "outer"` give the spatial sign covariances matrix, and `score = "rank"` and `estimate = "outer"` the spatial rank covariances matrix. Tyler's shape matrix is obtained with the choices `score =`

"sign" and estimate = "inner", and Dümbgen's shape matrix with score = "symmsign" and estimate = "inner". All the shape matrix estimates are rescaled to have trace p. If score = "identity" one gets just the regular covariance matrix.

The function plotShape can be used for a graphical comparison of different, at most three, shape matrices. The shape matrix estimates (standardized to have determinant one) are illustrated with ellipsoids plotted in a scatter matrix. Note that a center for the shape matrices needs to be specified too.

The function mv.shape.test can be used to test the null hypothesis that the observations are coming from a spherical distribution. Null hypothesis then implies that the population shape matrices (with any scores) are proportional to the identity matrix. The scores "identity", "sign", and "symmsign" are available. The location center is estimated if location = "estimate"; it is also possible to choose location = "origin". Naturally, any null hypothesis for a shape matrix can be tested by first transforming the data to be spherical under the null hypothesis.

In the elliptic model, all shape matrices are proportional. This means that their eigenvectors are the same and their eigenvalues are proportional. The function mvPCA can then be used for principal component analysis (PCA). As the shape matrices are scaled to have trace p, the eigenvalues are only proportional to the true variances of the principal components. Function mvPCA returns a list of class mvPCA with methods print, summary, predict, and plot. The use of mvPCA is made as similar as possible to the use of traditional functions princomp and prcomp.

Main references for the one-sample shape estimation and testing are Tyler (1987), Dümbgen (1998), Croux *et al.* (2002), Sirkiä *et al.* (2009), and Chapter 9 in Oja (2010).

4.4. Testing for independence of the subvectors

We assume that the data matrix is decomposed as $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ and we wish to test the null hypothesis that $n \times p_1$ data matrix $\mathbf{Y}_1$ and $n \times p_2$ data matrix $\mathbf{Y}_2$ are independent. The observation vectors are again replaced by inner centered and standardized scores, $(\mathbf{Y}_1, \mathbf{Y}_2) \to (\hat{\mathbf{T}}_1, \hat{\mathbf{T}}_2)$ such that

$$\hat{\mathbf{T}}_1^\top \mathbf{1}_n = \mathbf{0}, \quad \hat{\mathbf{T}}_2^\top \mathbf{1}_n = \mathbf{0}, \quad \hat{\mathbf{T}}_1^\top \hat{\mathbf{T}}_1 \propto \mathbf{I}_{p_1}, \quad \text{and} \quad \hat{\mathbf{T}}_2^\top \hat{\mathbf{T}}_2 \propto \mathbf{I}_{p_2}.$$

The test is then based on the canonical correlations between $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$. Hence the function mv.ind.test always uses inner centering and standardization, and the tests are affine invariant.

The main reference here is Taskinen *et al.* (2005). See also Chapter 10 in Oja (2010).

4.5. Several-samples location problem

The function mv.Csample.test can be used to test the null hypothesis that all c random samples come from the same p-variate distribution. The sample membership should be given by a factor variable (argument g) with at least two levels. The regular MANOVA (slightly modified) is obtained if score = "identity". Multivariate extensions of Mood's test and Kruskal-Wallis test are obtained with choices score = "sign" and score = "rank", respectively. Affine invariance of the tests is again attained if stand = "inner".

The function `mv.2sample.est` provides estimates of the location difference with its estimated covariance matrix in the two-sample case. The sample memberships are again given using argument `g` which now must be a factor with exactly two levels. The output is an object of class `mvloc` with methods `print`, `summary` and `plot`. The estimates available are then the difference of the sample means, the difference of the spatial medians, and the two-samples spatial Hodges-Lehmann estimate. Equivariant estimates are obtained with the choice `stand = "inner"`.

The main references are Möttönen and Oja (1995), Oja and Randles (2004), and Chapter 11 in Oja (2010).

4.6. Randomized blocks

The blocked design for the comparison of the effects of c treatments is the generalization of the paired-sample design. In the randomized block design the c subjects in each block are randomly assigned to all c treatments, $c \geq 2$. For an analysis of multivariate data arising from a randomized complete block design the functions `mv.2way.est` and `mv.2way.test` are available. The block membership is given by argument `block`, and the treatment by argument `treatment`. Both factors then have at least two levels.

The function `mv.2way.test` tests the null hypothesis of no treatment differences. The regular balanced two-way MANOVA is obtained with the choice `score = "identity"`, and a multivariate extension of the Friedman test with `score = "rank"`. Function `mv.2way.est` gives a list of class `mvcloc` with its own `print` and `summary` methods. All pairwise estimates of the location differences with their estimated covariance matrices are provided. These individual results are again of class `mvloc` and hence can be plotted using `plot`.

The main references here are Möttönen *et al.* (2003) and Chapter 12 in Oja (2010).

4.7. Multivariate linear regression

A formula object that specifies the model is the main argument in the regression function `mv.l1lm`. The left side in the formula must be a numeric matrix with at least two columns. Working with `mv.l1lm` is similar to working with other regression functions in R. Note that the results from `lm` and from `mv.l1lm` with the identity score function, however, differ slightly due to different divisors in the formula for the covariance matrix. For the regular L_2 regression the function `lm` is computationally more efficient and has more options than `mv.l1lm`.

The general algorithm in Section 3.3 is in fact a Weiszfeld algorithm modified to the multivariate linear regression case. The original Weiszfeld algorithm is for the one sample location case and may have problems if the residuals become excessively small (or zero). Vardi and Zhang (2000) developed a modified version to deal with zero residuals but there is no extension of their approach to the general linear regression case. In our modified Weisfeld algorithm we use the modified L_1 norm

$$||\mathbf{e}||_\epsilon = \begin{cases} ||\mathbf{e}|| & ||\mathbf{e}|| > \epsilon \\ \epsilon & ||\mathbf{e}|| \leq \epsilon \end{cases},$$

which gives a continuous modified spatial sign function

$$\mathbf{U}_\epsilon(\mathbf{e}) = \begin{cases} ||\mathbf{e}||^{-1}\mathbf{e} & ||\mathbf{e}|| > \epsilon \\ \epsilon^{-1}\mathbf{e} & ||\mathbf{e}|| \leq \epsilon \end{cases}.$$

These smoothed versions as approximations of the L_1 norm and the spatial sign function have been used before in the proofs for the asymptotic properties of the tests and estimates, see Möttönen *et al.* (1997). In our experience, the smoothed versions work well in the algorithms and yield reliable results. In the function `mv.l1lm` ϵ is called `eps.S` and has by default a value of 1e-06.

The function `mv.l1lm` returns an object of class `mvl1lm`. If `score = "rank"` then the estimate of the intercept parameter is the Hodges-Lehmann estimate of the residuals and must be computed separately. The returned object is then also made different for `score = "rank"`. Function `mv.l1lm` with `score = "rank"` can not be used in the one-sample location problem. The returned objects from `mv.l1lm` can be treated with methods `print`, `summary`, `coef`, `vcov`, `fitted`, `residuals`, and `predict` in a regular way. Method `plot` provides a joint scatter plot matrix for the fitted values and the residuals from the estimated model.

The method `anova` for the objects of class `mvl1lm` works as follows. If only `object` (and no `object2`) is provided `anova` returns the results from the test for the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ (testing problem I in Section 3.2). In this case the `test` argument is ignored. In testing problem II (Section 3.4) both arguments, `object` and `object2`, are used. Argument `object` is a fit from a full unrestricted model (with explaining variables in $\mathbf{X}_1$ and $\mathbf{X}_2$) and `object2` is the output for a restricted model (with explaining variables in $\mathbf{X}_1$ only). The test can be based either on the score test statistic (default, `test = "Score"`) or on the Wald-type test statistic (`test = "Wald"`). Naturally the fits in `object` and in `object2` must be based on the same data set, same score function (`identity`/`sign`/`rank`), and same way of standardization (`outer`/`inner`).

Note that the one-sample and c-sample location problems are special cases of the multivariate regression problem. The results from `mv.l1lm` and from specialized functions for one-sample and several-sample cases may differ slightly, however, as the covariance matrices of the estimates may be calculated in a different way and for estimation the stabler algorithm of Vardi and Zhang (2000) is used. In general we recommend the use of the specialized functions if available.

The main references are Bai *et al.* (1990), Arcones (1998), Chakraborty (2003), and Zhou (2010). The theory is explained also in Chapter 13 of Oja (2010).

Besides the functions mentioned above the package offers also some other auxiliary functions like `affines.trans`, `pairs2`, `rmvpowerexp` or `runifsphere`. For details, see the help pages. The plan is that in the future the package will include functions for canonical correlation analysis and the analysis of clustered data.

5. Examples for multivariate analysis using MNM

In this section we illustrate the use of **MNM** for different problems and designs discussed earlier. For the output the option `options(digits = 4)` in R 2.13.0 (R Development Core Team 2011) is used. We also use the packages **MNM** 1.0-0, **mvtnorm** 0.9-99 (Genz *et al.* 2011), **robustbase** 0.7-3 (Rousseeuw *et al.* 2011; Todorov and Filzmoser 2009) and **DAAG** 1.0-6 (Maindonald and Braun 2011).

In all examples, random seeds are provided for reproducibility of the results. In a few cases the output was slightly modified to fit into the text.

5.1. One-sample location problem

Outer vs. inner standardization

We first use **MNM** to illustrate the comparison of estimates using outer and inner standardization. The estimates to be compared are the sample mean vector, the regular spatial median, and the spatial median with inner standardization (affine equivariant Hettmansperger-Randles estimate). For the comparison, we generate 300 observations from a $N_3(\mathbf{0}, diag(1, 1, 100))$ distribution.

The data are then generated as follows.

```
R> library("MNM")
R> set.seed(1234)
R> X <- rmvnorm(300, c(0, 0, 0), diag(c(1, 1, 100)))
R> names(X) <- c("x_1", "x_2", "x_3")
```

The three estimates are computed with the following function calls.

```
R> Est.X1 <- mv.1sample.est(X)
R> Est.X2 <- mv.1sample.est(X, score = "s", stand = "o")
R> Est.X3 <- mv.1sample.est(X, score = "s", stand = "i")
```

The best way to have a first look at the estimation results is to use the `summary` function. For the third estimate, for example, we get the following summary.

```
R> summary(Est.X3)
```

```
The equivariant spatial median of X is:
[1] -0.0159 -0.0172 -0.3695

And has the covariance matrix:
       [,1]    [,2]    [,3]
[1,] 0.0040  0.0000  0.0035
[2,] 0.0000  0.0041 -0.0020
[3,] 0.0035 -0.0020  0.3185
```

The three location estimates now are

```
R> rbind(Est.X1$location, Est.X2$location, Est.X3$location)
```

```
          [,1]     [,2]    [,3]
[1,]  0.013028 -0.05559 -0.2352
[2,] -0.005535 -0.04049 -0.5057
[3,] -0.015895 -0.01717 -0.3695
```

For a visual comparison of the estimates and their 95% confidence we can write

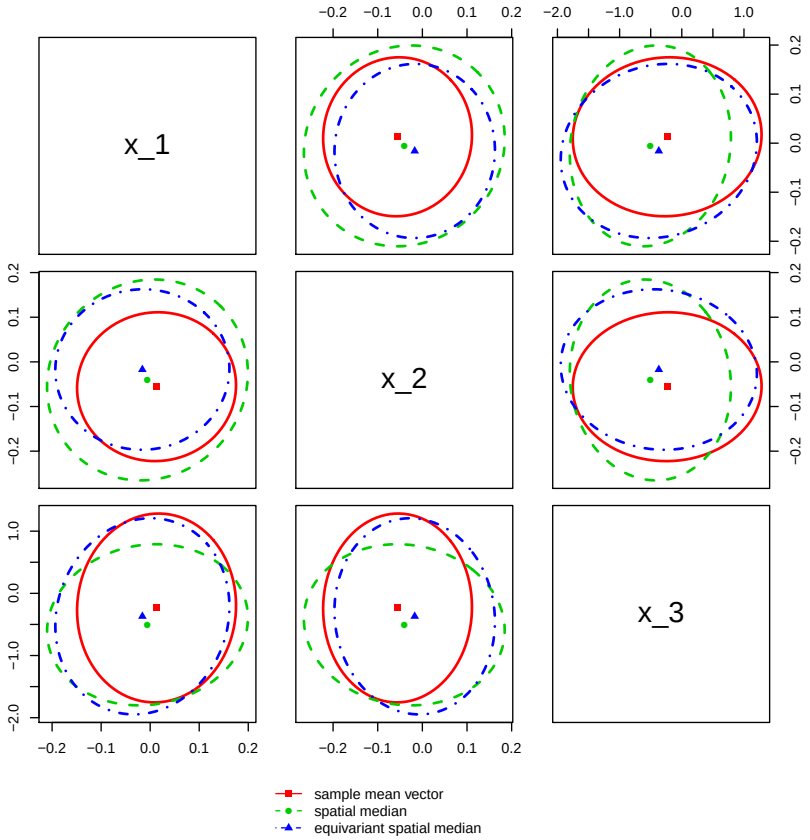

Figure 1: Comparison of the mean vector, the regular spatial median, and the affine equivariant spatial median (Hettmansperger-Randles estimate) for a random sample coming from a multivariate normal distribution.

```
R> plot(Est.X1, Est.X2, Est.X3, lty.ell = c(1, 2, 4), pch.ell = 15:17,
+     lwd.ell = c(2, 2, 2), alim = "e", labels = names(X))
```

Figure 1 then shows, as expected, that the mean vector has the smallest confidence ellipsoid for the data coming from a multivariate normal distribution. The affine equivariant spatial median (Hettmansperger-Randles estimate) seems better than the regular spatial median. The regular spatial median seems efficient in the direction of the largest scale (third component) but has otherwise a poor efficiency. In general we recommend the use of inner standardization if the scales of the marginal variables differ a lot.

Comparison of the estimates for a heavy-tailed distribution

We next compare location estimates that are based on identity, sign, and signed-rank scores and use inner standardization. We compare the behavior of the estimates using a random sample from a spherical power exponential distribution with shape parameter $\beta = 0.4$ (The power exponential distribution is an elliptical distribution that has light or heavy tails depending on the value of the shape parameter β. Special cases are, for example, a multivariate

normal distribution ($\beta = 1$) or a multivariate Laplace distribution ($\beta = 0.5$). The limiting case as $\beta \to \infty$ is a multivariate generalization of the uniform distribution in a sphere. See Gómez *et al.* (1998) for details).

The dataset is generated as follows.

```
R> set.seed(4321)
R> Y <- rmvpowerexp(150, c(0, 0, 0), Beta = 0.4)
R> names(Y) <- c("y_1", "y_2", "y_3")
```

The three estimates are obtained using the following function calls.

```
R> Est.Y1 <- mv.1sample.est(Y)
R> Est.Y2 <- mv.1sample.est(Y, score = "r", stand = "i")
R> Est.Y3 <- mv.1sample.est(Y, score = "s", stand = "i")
```

The observed values of the estimates are

```
R> rbind(Est.Y1$location, Est.Y2$location, Est.Y3$location)
```

```
       [,1]     [,2]   [,3]
[1,] 0.3110 -0.08433 0.1616
[2,] 0.2408 -0.10970 0.1301
[3,] 0.1317 -0.11474 0.1353
```

A visual comparison of the estimates and their confidence ellipsoids is obtained as before.

```
R> plot(Est.Y1, Est.Y2, Est.Y3, lty.ell = c(1, 2, 4), pch.ell = 15:17,
+    lwd.ell = c(2, 2, 2), alim = "e", labels = names(Y))
```

Figure 2 shows that the sample mean vector is poor in its efficiency for heavy tailed distribution. As expected, the affine equivariant spatial median has the smallest confidence ellipsoid in this case.

5.2. One-sample shape problem

In this section we show how the shape matrices can be estimated with the package **MNM**. We use the dataset `salinity` with 28 observations and four variables. The dataset is available in the package **robustbase**. We consider the first three variables only with two clearly visible outliers.

First we load the data and extract the variables of interest.

```
R> library("robustbase")
R> data("salinity")
R> sal.X <- salinity[, 1:3]
```

Then we compute three different scatter/shape matrices, the regular covariance matrix, Tyler's shape matrix and Dümbgen's shape matrix as follows.

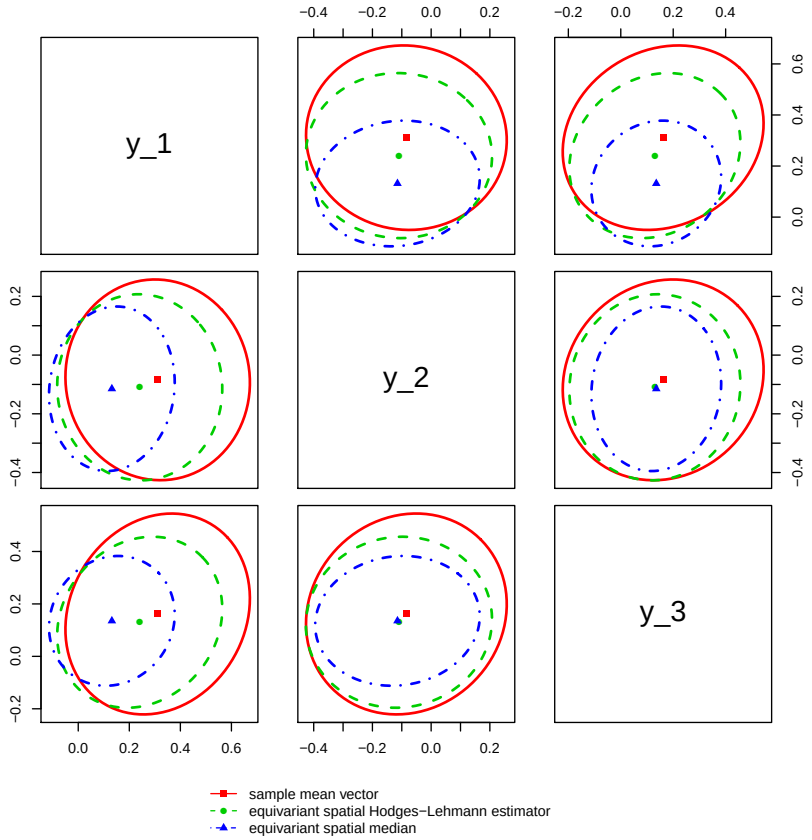

Figure 2: Comparison of the mean vector, the affine equivariant Hodges-Lehmann estimate, and the affine equivariant spatial median (Hettmansperger-Randles estimate) for a random sample coming from a heavy-tailed power-exponential distribution.

```
R> covSal <- mv.shape.est(sal.X)
R> tylerSal <- mv.shape.est(sal.X, score = "si", estimate = "i")
R> dumbgenSal <- mv.shape.est(sal.X, score = "sy", estimate = "i")
```

These three matrices are not directly comparable, however, since they are not scaled in the same way. For a visual comparisons, we again plot ellipsoids based on shape matrices and centered using suitable location estimates (the mean vector and the affine equivariant spatial median). The affine equivariant spatial median (Hettmansperger-Randles estimate) is obtained with

```
R> HR.median <- mv.1sample.est(sal.X, score = "s", stand = "i")$location
```

We then combine the shape matrices and location centers and give the combinations the names as follows.

```
R> EST1 <- list(location = colMeans(sal.X), scatter = covSal,
+    est.name = "regular cov")
```

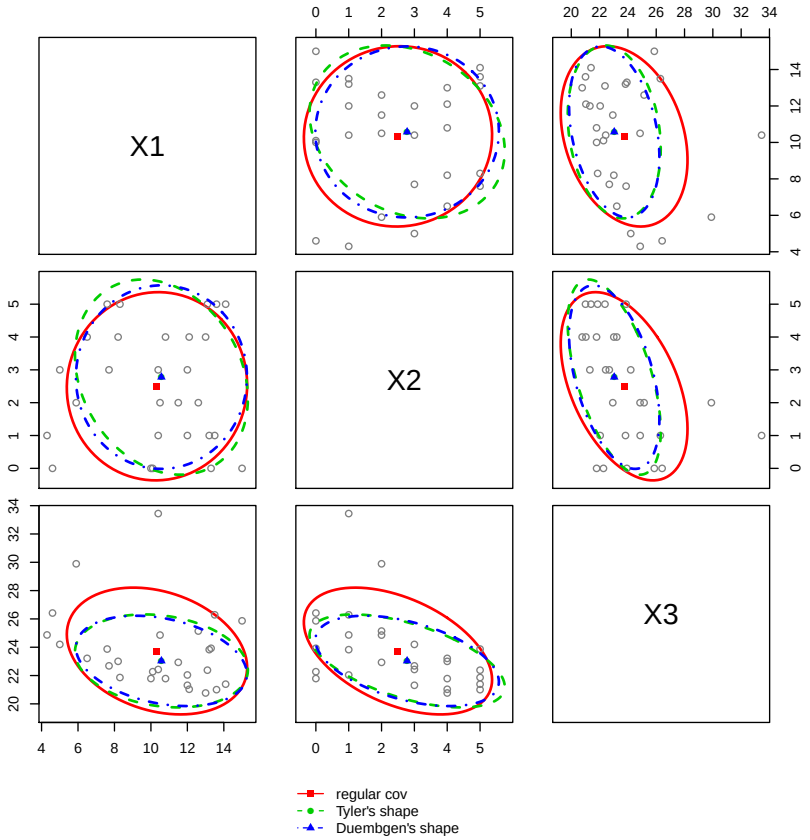

Figure 3: Comparison of the regular covariance matrix, Tyler's shape matrix, and Dümbgen's shape matrix for the salinity dataset.

```
R> EST2 <- list(location = HR.median, scatter = tylerSal,
+    est.name = "Tyler's shape")
R> EST3 <- list(location = HR.median, scatter = dumbgenSal,
+    est.name = "Duembgen's shape")
```

For the comparison of different approaches, the function `plotShape` then plots estimated 50% tolerance ellipsoids based on different combinations with the following function call.

```
R> plotShape(EST1, EST2, EST3, X = sal.X, lty.ell = c(1, 2, 4),
+    pch.ell = 15:17, lwd.ell = c(2, 2, 2))
```

As we can see in Figure 3, Tyler's shape matrix and Dümbgen's shape matrix give for this data similar ellipsoids. The ellipsoid based on the covariance matrix and mean vector is attracted by the two outliers.

At first sight, the figure seems to suggest that the first two variables `X1` and `X2` come from a spherical distribution. This may not be true, however, as in the plot the variables are rescaled in a different way. We next call the test functions for testing for sphericity using the identity score, the sign score, and the symmetrized sign score.

```
R> mv.shape.test(sal.X[, 1:2])

        Mauchly test for sphericity

data:  sal.X[, 1:2]
L = 0.0186, df = 2, p-value = 0.01864

R>  mv.shape.test(sal.X[, 1:2], score = "si")

        Test for sphericity based on UCOV

data:  sal.X[, 1:2]
Q2 = 2.326, df = 2, p-value = 0.3126

R> mv.shape.test(sal.X[, 1:2], score = "sy")

        Test for sphericity based on TCOV

data:  sal.X[, 1:2]
Q2 = 13.15, df = 2, p-value = 0.001397
```

The tests based on identity scores and symmetrized signs scores reject the null hypothesis. In the figure the ellipsoid based on the sign score seems different from the others.

5.3. Two-sample location problem

In the two-sample location problem and in the multivariate regression we use the Australian athletes dataset available as `data("ais")` in the package **DAAG**. We are mainly interested in the differences between male and female athletes when the response variables are the hematocrit percentage (variable `hc`) and the hemaglobin concentration (variable `hg`).

The data can be loaded as follows.

```
R> library("DAAG")
R> data("ais")
```

A scatter plot for a visual comparison of males and females is given by the following call.

```
R> with(ais, pairs(cbind(hc, hg), col = sex))
```

Figure 4 shows a clear difference in location. For this two-sample location problem we can use the function `mv.Csample.test`. If we use the sign score and inner standardization we get

```
R> with(ais, mv.Csample.test(cbind(hc, hg), sex, score = "s", stand = "i"))

        Equivariant several samples location test using spatial signs

data:  cbind(hc, hg) by sex
Q.2 = 113.6, df = 2, p-value < 2.2e-16
alternative hypothesis: true location difference between some groups is not
                        equal to c(0,0)
```

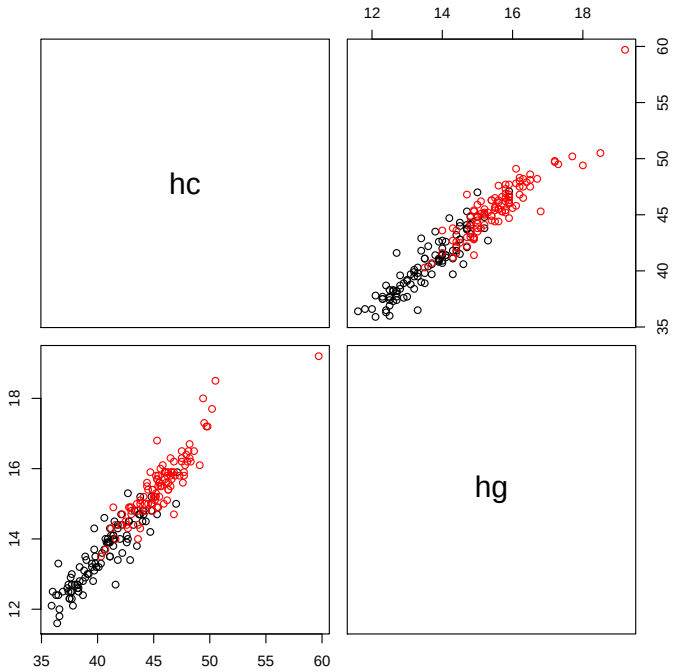

Figure 4: Hematocrit percentage (variable `hc`) and hemaglobin concentration (variable `hg`) of the athletes. The males are marked with red color.

The test clearly shows that there is a difference between the genders. Also the sample size of $n = 202$ leaves no doubt about the accuracy of the χ^2 approximation. For small sample sizes one should rather use the permutation version of the test which gives in this case the following result.

```
R> with(ais, mv.Csample.test(cbind(hc, hg), sex, score = "s", stand = "i",
+    method = "p"))

        Equivariant several samples location test using spatial signs

data:  cbind(hc, hg) by sex
Q.2 = 113.6, replications = 1000, p-value < 2.2e-16
alternative hypothesis: true location difference between some groups is not
                        equal to c(0,0)
```

and as expected no disagreement here. To get an estimate of the difference we use next the function `mv.2sample.est`:

```
R> summary(with(ais, mv.2sample.est(cbind(hc, hg), sex, score = "s",
+    stand = "i")))

The difference between equivariant spatial medians of cbind(hc, hg) by sex is:
[1] -5.314 -2.012
```

```
And has the covariance matrix:
       [,1]   [,2]
[1,] 0.1330 0.0507
[2,] 0.0507 0.0233
```

5.4. Multivariate linear regression

As mentioned earlier, the two-sample location case is a special case of the multivariate linear model. The results for the two-sample location problem can therefore be obtained using the function `mv.l1lm` as well.

The estimate of the location difference between males and females with the sign score and inner standardization can be obtained by fitting the following model.

```
R> model.sex <- mv.l1lm(cbind(hc, hg) ~ sex, data = ais, score = "s",
+    stand = "i")
R> summary(model.sex)

Multivariate regression using spatial sign scores and inner standardization

Call:
mv.l1lm(formula = cbind(hc, hg) ~ sex, scores = "s", stand = "i", data = ais)

Testing that all coefficients = 0:
Q.2 = 197.7 with 4 df, p.value < 2.2e-16

Results by response:

Response hc :
            Estimate Std. Error
(Intercept)    40.27      0.259
sexm            5.31      0.365

Response hg :
            Estimate Std. Error
(Intercept)    13.51      0.108
sexm            2.01      0.153
```

The estimate for the location difference is as obtained when using `mv.2sample.est`. For the p value for testing the null hypothesis of no difference, we fit the model with the intercept term only and compare the resulting fit to the fit coming from the previous model.

```
R> model.int <- mv.l1lm(cbind(hc, hg) ~ 1, data = ais, score = "s",
+    stand = "i")
R> anova(model.sex, model.int)
```

```
Comparisons between multivariate linear models

Full model:       mv.l1lm(formula = cbind(hc, hg) ~ sex, scores = "s",
                          stand = "i", data = ais)
Restricted model: mv.l1lm(formula = cbind(hc, hg) ~ 1, scores = "s",
                          stand = "i", data = ais)

Score type test that coefficients not in the restricted model are 0:
Q.2 = 113.6 with 2 df, p.value < 2.2e-16
```

The variables red blood cell count (`rcc`), body mass index (`bmi`) and the percentage of body fat (`pcBfat`) are good explaining factors for our response variables `hc` and `hg`. They are also variables with location differences between males and females. See Figure 5 produced by the following call.

```
R> with(ais, pairs(cbind(hc, hg, rcc, bmi, pcBfat), col = sex))
```

To see whether the differences between males and females are due to differences in `rcc`, `bmi`, and `pcBfat`, we first fit the full model (now with the rank score and inner standardization).

```
R> model.full <- mv.l1lm(cbind(hc, hg) ~ rcc  + bmi + pcBfat  + sex,
+    data = ais, score = "r", stand = "i")
R> summary(model.full)

Multivariate regression using spatial rank scores and inner standardization

Call:
mv.l1lm(formula = cbind(hc, hg) ~ rcc + bmi + pcBfat + sex, scores = "r",
                  stand = "i", data = ais)

Inner HL-estimator for the residuals (intercept):
              hc   hg
(Intercept) 11.0 3.27

Testing that all coefficients = 0:
Q.2 = 124.8 with 8 df, p.value < 2.2e-16

Results by response:

Response hc :
       Estimate Std. Error
rcc      6.4467     0.2747
bmi      0.1132     0.0436
pcBfat  -0.0766     0.0263
sexm     0.2585     0.3598

Response hg :
```

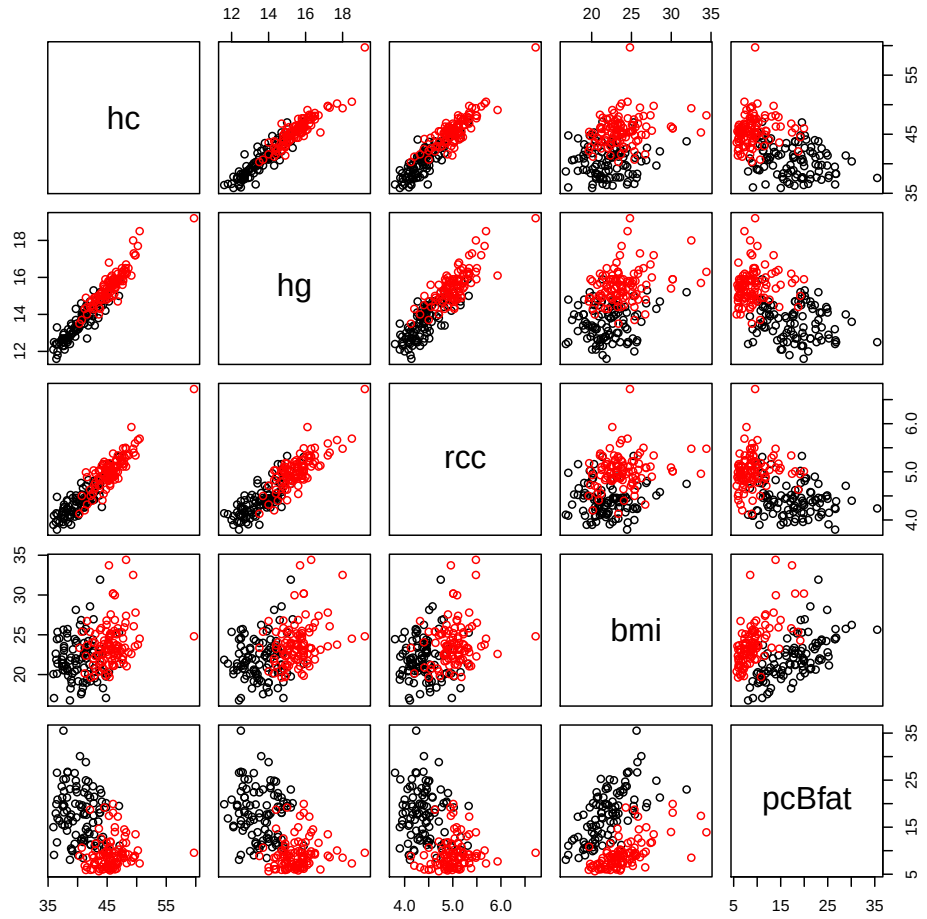

Figure 5: Hematocrit percentage (variable `hc`) and hemaglobin concentration (variable `hg`), red blood cell count (`rcc`), body mass index (`bmi`), and the percentage of body fat (`pcBfat`) of the athletes data. The males are marked with red color.

```
        Estimate Std. Error
rcc       2.0882     0.1229
bmi       0.0778     0.0195
pcBfat   -0.0334     0.0118
sexm      0.2353     0.1609
```

As mentioned earlier the output for the rank score differs slightly from that of the other scores - the intercept parameter is here reported separately and not in the table of other coefficients.

To test the null hypothesis that there is no difference between males and females, we find the fit from a restricted model (without gender) and then use `anova` to compare the full model and restricted model.

```
R> model.res <- mv.l1lm(cbind(hc,hg) ~ rcc  + bmi + pcBfat,
+    data = ais, score = "r", stand = "i")
```

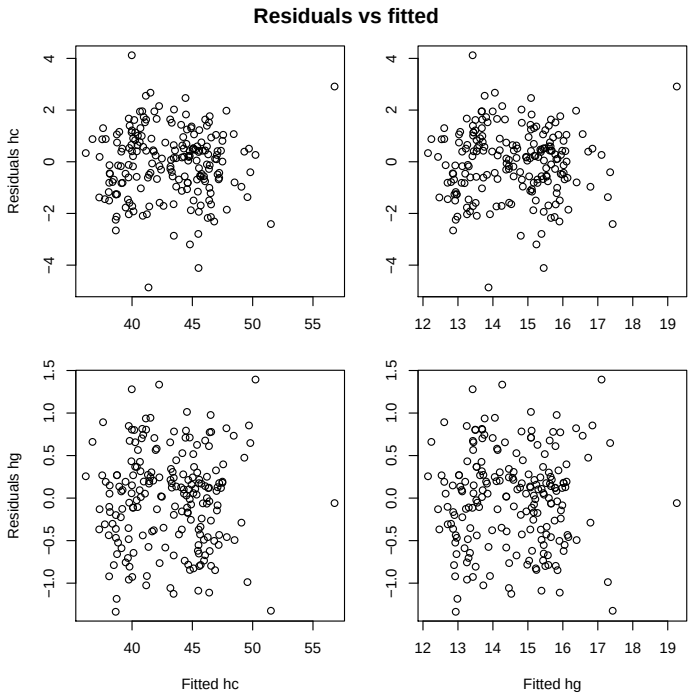

Figure 6: A plot of residuals coming from the estimated model `model.res`.

```
R> anova(model.full, model.res)

Comparisons between multivariate linear models

Full model:       mv.l1lm(formula = cbind(hc, hg) ~ rcc + bmi + pcBfat + sex,
                          scores = "r", stand = "i", data = ais)
Restricted model: mv.l1lm(formula = cbind(hc, hg) ~ rcc + bmi + pcBfat,
                          scores = "r", stand = "i", data = ais)

Score type test that coefficients not in the restricted model are 0:
Q.2 = 2.496 with 2 df, p.value = 0.2871
```

Hence the data provides no evidence for a difference between males and females. For model checking one could still check the residuals for any hidden structures as follows.

```
R> plot(model.res)
```

Figure 6 then suggests that the fit `model.rest` is satisfactory.

6. Summary

The package **MNM** provides functions for most standard inference problems in multivariate analysis. In most cases, the user can choose between the scores `identity`, `sign` or `rank`.

The identity score is optimal in a multivariate normal model but the performance becomes poor for data with heavy-tailed distributions or with outliers. In those cases `sign` and `rank` scores are better choices. A problem with `rank` score is that the procedures are based on pairwise differences and/or pairwise sums of the observations. The computation is then slow if the dimension is large and memory consuming when the number of observations is huge. The usefulness of `rank` scores depends heavily on the user's hardware. The use of `sign` score is much less demanding and have also other properties that make it attractive in high dimensions.

Methods based on `sign` and `rank` scores have usually an option for their outer or inner standardization. Using outer standardization means that the methods are not affine invariant/equivariant under linear transformations but can be used when the marginal variables are similarly scaled. For coordinate-free tests and estimates, inner standardization is needed.

Test functions in **MNM** often give the user the possibility to compute p values based on (i) the limiting distribution of the test statistic or based on (ii) permutation and sign-change arguments. For small sample sizes, a good practise is to compute both and then decide which one to choose and report. Spatial sign and rank methods have been applied in Behseta and Chenouri (2011) or Tahvanainen *et al.* (2009), for example.

The package provides estimation and testing procedures for independent and identically distributed observations only. Extension for clustered data are planned to be added to **MNM** as well. For the theory, see Nevalainen *et al.* (2010). Long term plans include also to write some parts of the code in C or C++ and develop new estimation algorithms for the multivariate regression problem.

Acknowledgments

The authors are grateful for the comments of the associate editor and the two anonymous referees. The work of Klaus Nordhausen and Hannu Oja was supported by grants from the Academy of Finland.

References

Arcones MA (1998). "Asymptotic Theory for M-Estimators over a Convex Kernel." *Econometric Theory*, **14**, 387–422.

Bai ZD, Chen R, Miao BQ, Rao CR (1990). "Asymptotic Theory of Least Distances Estimate in Multivariate Linear Models." *Statistics*, **4**, 503–519.

Behseta S, Chenouri S (2011). "Comparison of Two Populations of Curves with an Application in Neuronal Data Analysis." *Statistics in Medicine*, **30**, 1441–1454.

Chakraborty B (2003). "On Multivariate Quantile Regression." *Journal of Statistical Planning and Inference*, **110**, 109–132.

Chaudhuri P (1992). "Multivariate Location Estimation Using Extension of R-Estimates through U-Statistics Type Approach." *The Annals of Statistics*, **20**, 897–916.

Choi K, Marden J (1997). "An Approach to Multivariate Rank Tests in Multivariate Analysis of Variance." *Journal of the American Statistical Society*, **92**, 1581–1590.

Croux C, Ollila E, Oja H (2002). "Sign and Rank Covariance Matrices: Statistical Properties and Application to Principal Component Analysis." In Y Dodge (ed.), *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pp. 257–270. Birkhäuser, Basel.

Dümbgen L (1998). "On Tyler's M-Functional of Scatter in High Dimension." *Annals of the Institute of Statistal Mathematics*, **50**, 471–491.

Fischer D, Möttönen J, Nordhausen K, Vogel D (2010). ***OjaNP**: Multivariate Methods Based on the Oja Median and Related Concepts*. R package version 0.9-4, URL http://CRAN.R-project.org/package=OjaNP.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2011). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 0.9-99, URL http://CRAN.R-project.org/package=mvtnorm.

Gómez E, Gómez-Villegas MA, Marín JM (1998). "A Multivariate Generalization of the Power Exponential Family of Distributions." *Communications in Statististics – Theory and Methods*, **27**, 589–600.

Hettmansperger TP, McKean JW (2010). *Robust Nonparametric Statistical Methods*. 2nd edition. CRC Press, Boca Raton.

Hettmansperger TP, Randles RH (2002). "A Practical Affine Equivariant Multivariate Median." *Biometrica*, **89**, 851–860.

Hothorn T, Hornik K (2011). ***exactRankTests**: Exact Distributions for Rank and Permutation Tests*. R package version 0.8-20, URL http://CRAN.R-project.org/package=exactRankTests.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**(8), 1–23. URL http://www.jstatsoft.org/v28/i08/.

Kloke J (2010). ***Rfit**: Rank Estimation for Linear Models*. R package version 0.09, URL http://CRAN.R-project.org/package=Rfit.

Koenker R (2011). ***quantreg**: Quantile Regression*. R package version 4.67, URL http://CRAN.R-project.org/package=quantreg.

Maindonald JH, Braun WJ (2011). ***DAAG**: Data Analysis and Graphics Data and Functions*. R package version 1.06, URL http://CRAN.R-project.org/package=DAAG.

Marden J (1999). "Multivariate Rank Tests." In S Ghosh (ed.), *Design of Experiments, and Survey Sampling*, pp. 401–432. M. Dekker, New York.

Möttönen J, Hüsler J, Oja H (2003). "Multivariate Nonparametric Tests in Randomized Complete Block Design." *Journal of Multivariate Analysis*, **85**, 106–129.

Möttönen J, Oja H (1995). "Multivariate Spatial Sign and Rank Methods." *Journal of Nonparametric Statistics*, **5**, 201–213.

Möttönen J, Oja H, Tienari J (1997). "On the Efficiency of Multivariate Spatial Sign and Rank Tests." *The Annals of Statistics*, **25**, 542–552.

Nevalainen J, Larocque D, Oja H, Pörsti I (2010). "Nonparametric Analysis of Clustered Multivariate Data." *Journal of the American Statistical Association*, **105**, 864–872.

Nordhausen K, Oja H, Tyler DE (2008). "Tools for Exploring Multivariate Data: The Package **ICS**." *Journal of Statistical Software*, **28**(6), 1–31. URL http://www.jstatsoft.org/v28/i06/.

Nordhausen K, Sirkiä S, Oja H, Tyler DE (2010). ***ICSNP**: Tools for Multivariate Nonparametrics*. R package version 1.0-7, URL http://CRAN.R-project.org/package=ICSNP.

Oja H (1999). "Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: A Review." *Scandinavian Journal of Statistics*, **26**, 319–343.

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

Oja H, Randles RH (2004). "Multivariate Nonparametric Tests." *Statistical Science*, **19**, 598–605.

Ollila E, Hettmansperger TP, Oja H (2002). "Estimates of Regression Coefficients Based on Sign Covariance Matrix." *Journal of the Royal Statistical Society B*, **64**, 447–466.

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

Randles RH (1989). "A Distribution-Free Multivariate Sign Test Based on Interdirections." *Journal of the American Statistical Association*, **84**, 1045–1050.

Randles RH (2000). "A Simpler, Affine Equivariant Multivariate, Distribution-Free Sign Test." *Journal of the American Statistical Association*, **95**, 1263–1268.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Maechler M (2011). ***robustbase**: Basic Robust Statistics*. R package version 0.7-3, URL http://CRAN.R-project.org/package=robustbase.

Sirkiä S, Taskinen S, Oja H, Tyler D (2009). "Tests and Estimates of Shape Based on Spatial Signs and Ranks." *Journal of Nonparametric Statistics*, **21**, 155–176.

Tahvanainen A, Leskinen M, Koskela J, Ilveskoski E, Nordhausen K, Oja H, Kähönen M, Kööbi T, Mustonen J, Pörsti I (2009). "Ageing and Cardiovascular Responses to Head-Up Tilt in Healthy Subjects." *Atherosclerosis*, **207**, 445–451.

Taskinen S, Oja H, Randles RH (2005). "Multivariate Nonparametric Tests of Independence." *Journal of the American Statistical Association*, **100**, 916–925.

Terpstra JT, McKean JW (2005). "Rank-Based Analysis of Linear Models Using R." *Journal of Statistical Software*, **14**(7), 1–26. URL http://www.jstatsoft.org/v15/i07/.

Todorov V, Filzmoser P (2009). "An Object-Oriented Framework for Robust Multivariate Analysis." *Journal of Statistical Software*, **32**(3), 1–47. URL http://www.jstatsoft.org/v32/i03/.

Tyler DE (1987). "A Distribution-Free M-Estimator of Multivariate Scatter." *The Annals of Statistics*, **15**, 234–251.

Vardi Y, Zhang CH (2000). "The Multivariate L_1-Median and Associated Data Depth." In *Proceedings of the National Academy of Sciences of the United States of America*, volume 97, pp. 1423–1426.

Wilcox RR (2010). *Introduction to Robust Estimation and Hypothesis Testing*. 2nd edition. Elsevier Academic Press, Burlington.

Zhou W (2010). "A Multivariate Wilcoxon Regression Estimate." *Journal of Nonparametric Statistics*, **22**, 859–877.

Affiliation:

Klaus Nordhausen
School of Health Sciences
University of Tampere
33014 University of Tampere, Finland
E-mail: klaus.nordhausen@uta.fi
URL: http://www.uta.fi/~klaus.nordhausen/

Hannu Oja
School of Health Sciences
University of Tampere
33014 University of Tampere, Finland
E-mail: hannu.oja@uta.fi
URL: http://www.uta.fi/~hannu.oja/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 43, Issue 5 *Submitted:* 2010-02-05
July 2011 *Accepted:* 2011-06-17