Journal of Statistical Software

May 2006, Volume 16, Issue 5. *http://www.jstatsoft.org/*

SAS/IML Macros for a Multivariate Analysis of Variance Based on Spatial Signs

Jaakko Nevalainen
University of Tampere

Hannu Oja
University of Tampere

Abstract

Recently, new nonparametric multivariate extensions of the univariate sign methods have been proposed. Randles (2000) introduced an affine invariant multivariate sign test for the multivariate location problem. Later on, Hettmansperger and Randles (2002) considered an affine equivariant multivariate median corresponding to this test. The new methods have promising efficiency and robustness properties. In this paper, we review these developments and compare them with the classical multivariate analysis of variance model. A new SAS/IML tool for performing a spatial sign based multivariate analysis of variance is introduced.

Keywords: affine invariance/equivariance, spatial sign, multivariate analysis of variance, multivariate sign test, multivariate median, SAS.

1. Introduction

Classical statistical techniques for multivariate location problems such as Hotelling's T^2 tests, multivariate analysis of variance (MANOVA) and multivariate multiple regression analysis rely on the assumption that the data were from a multivariate normal distribution. The inference methods are then based on the assumption of multivariate normality, the sample mean vector and the sample covariance matrix. However, these methods are extremely sensitive to outlying observations and they are inefficient for heavy tailed noise distributions.

Möttönen and Oja (1995) reviewed multivariate sign and rank tests and the corresponding estimates based on the L_1-type objective function. The tests and estimates were rotation invariant and equivariant, but not affine invariant/equivariant. Recently, new nonparametric multivariate extensions of the univariate sign methods have been proposed. Randles (2000) developed an affine invariant one-sample multivariate sign test. Hettmansperger and Randles (2002) considered an affine equivariant multivariate median corresponding to this test. Their approach combines the simultaneous use of the spatial median (Brown 1983), Tyler's M-estimate of scatter (Tyler 1987) and the transformation-retransformation technique (Chakraborty, Chaudhuri, and Oja 1998). Chakraborty *et al.* (1998) used a similar idea as Hettmansperger and Randles (2002), but not Tyler's scatter matrix. Like Randles' test, the Hettmansperger and Randles (2002) estimate is fairly easy to compute.

In this paper, we will first recall the classical MANOVA model. In Section 3 we review the multivariate spatial sign tests and estimators analoguous to their classical alternatives. In addition, we outline some ideas how to approximate the precision of the estimates. Section 4 introduces new SAS macros written in Interactive Matrix Language (IML) for performing the analysis. As far as the authors are aware, these procedures are not currently available in standard software packages. Finally, the use of the SAS/IML tools is illustrated by an example. The complete SAS/IML code is available at `http://www.jstatsoft.org/v16/i05/`.

In the following sections we will assume that there are c independent random samples of p-dimensional observations. Let

$$\mathbf{X} = (\mathbf{x}_{11} \cdots \mathbf{x}_{1n_1} \; \mathbf{x}_{21} \cdots \mathbf{x}_{2n_2} \; \cdots \; \mathbf{x}_{c1} \cdots \mathbf{x}_{cn_c})$$

denote the $p \times N$ data matrix, where $\mathbf{x}_{ij} = (x_{ij1} \; x_{ij2} \cdots x_{ijp})^\top$ represents the jth observation of the ith sample. Further write $N = n_1 + \cdots + n_c$ for the total number of observations. In practice, the data matrix is often given as a transpose of $\mathbf{X}$. We are interested in drawing conclusions on the parameter set $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}$, where $\boldsymbol{\mu}_i$ denotes the center of symmetry of the ith sample, and $\boldsymbol{\Sigma}$ the covariance (or scatter) matrix (assumed to be common for all the samples). Alternatively, one may wish to parametrize the model by $\boldsymbol{\mu}_1, \boldsymbol{\Delta}_{12}, \ldots, \boldsymbol{\Delta}_{1c}, \boldsymbol{\Sigma}$, where $\boldsymbol{\Delta}_{1i} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_1$ represents the difference between sample i and the first sample used as a reference (e.g. placebo). In general, we wish to estimate both sets of parameters, and construct the associated location tests.

Let $\mathbf{B}$ denote a nonsingular $p \times p$ matrix and $\mathbf{b}$ a $p \times 1$ vector. A location estimate $\hat{\boldsymbol{\mu}}_i(\mathbf{X})$ and a scatter matrix estimate $\hat{\boldsymbol{\Sigma}}(\mathbf{X})$ are affine equivariant if

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i\left(\mathbf{B}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top\right) &= \mathbf{B}\hat{\boldsymbol{\mu}}_i(\mathbf{X}) + \mathbf{b} \text{ and} \\ \hat{\boldsymbol{\Sigma}}\left(\mathbf{B}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top\right) &= \mathbf{B}\hat{\boldsymbol{\Sigma}}(\mathbf{X})\mathbf{B}^\top. \end{aligned}$$

A test statistic $\mathbf{T}(\mathbf{X})$ is affine invariant if

$$\mathbf{T}\left(\mathbf{B}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top\right) = \mathbf{T}(\mathbf{X}).$$

These definitions simply mean that a rescaling, a rotation or a shift of the data should results into corresponding transformation in the estimates, but the value of the test statistic should remain unchanged.

2. Classical MANOVA

When more than one attribute is measured per observational unit and the observational units arise from independent populations, the design is typically analyzed by multivariate analysis of variance techniques. Classical MANOVA assumption is that the outcome vectors $\mathbf{x}_{ij}$ $(p \times 1)$ are generated from the model

$$\mathbf{x}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij},$$

where $\boldsymbol{\mu}_i = (\mu_{i1}\,\mu_{i2}\,\cdots\,\mu_{ip})^\top$ is the location center for the ith sample (population), and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}\,\varepsilon_{ij2}\,\cdots\,\varepsilon_{ijp})^\top$ are independent and identically distributed random errors from a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. In a one-sample case, the classical test for the location problem is Hotelling's T^2 test.

Lemma 1 *Hotelling's T^2 statistic for testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is*

$$T^2 = N\,\bar{\mathbf{x}}^\top \mathbf{S}^{-1}\bar{\mathbf{x}},$$

where $\bar{\mathbf{x}}$ is the sample mean vector and $\mathbf{S}$ is the sample covariance matrix. Furthermore,

$$\frac{N-p}{(N-1)p}\,T^2 \text{ has an } F_{p,N-p} \text{ distribution.}$$

In a multisample case, the interest is to test the null hypothesis of no difference in location between the samples

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_c$$

assuming a common covariance matrix $\boldsymbol{\Sigma}$. Under the null hypothesis, the maximum likelihood estimator of a joint $\boldsymbol{\mu}$ is the sample mean vector over the combined sample, and the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is the pooled sample covariance matrix. For hypothesis testing, we may use the two-sample Hotelling's T^2 statistic, or in a more general c-sample case, the Hotelling's trace statistic:

Lemma 2 *Hotelling's trace statistic for testing $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_c$ is*

$$T^2 = (N-c)\,\mathrm{Tr}\left(\mathbf{B}\mathbf{W}^{-1}\right),$$

where $\mathbf{B}$ is the between-samples sums of squares matrix and $\mathbf{W}$ the within-samples sums of squares matrix. Under the null hypothesis, the test statistic is asymptotically $\chi^2_{p(c-1)}$ distributed.

Write

$$\mathbf{z}_{ij} = \left(\frac{1}{N-c}\mathbf{W}\right)^{-1/2}(\mathbf{x}_{ij} - \bar{\mathbf{x}})$$

for standardized observations with the sample mean vector zero and the sample covariance matrix $\mathbf{I}_p$, and

$$\bar{\mathbf{z}}_i = \left(\frac{1}{N-c}\mathbf{W}\right)^{-1/2}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$

for their sample means. We can write

$$\begin{aligned}(N-c)\,\mathrm{Tr}\left(\mathbf{B}\mathbf{W}^{-1}\right) &= (N-c)\sum_{i=1}^{c} n_i \mathrm{Tr}\left((\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top \mathbf{W}^{-1}\right) \\ &= (N-c)\sum_{i=1}^{c} n_i \mathrm{Tr}\left(\mathbf{W}^{-1/2}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top \mathbf{W}^{-1/2}\right) \\ &= \sum_{i=1}^{c} n_i \mathrm{Tr}\left(\bar{\mathbf{z}}_i\bar{\mathbf{z}}_i^\top\right) = \sum_{i=1}^{c} n_i \|\bar{\mathbf{z}}_i\|^2. \qquad (1)\end{aligned}$$

Note that the limiting distribution is still $\chi^2_{p(c-1)}$ if the covariance matrix estimate $(N-c)^{-1}\mathbf{W}$ is replaced by the regular pooled sample covariance matrix $\mathbf{S}$. We will observe similarities between the trace statistic and a multivariate spatial sign test statistic later on.

3. Spatial sign MANOVA

In this section we present sign based competitors to the Hotelling's T^2 statistic and to the sample mean vector. Estimation and test constructions are based on the spatial signs of suitably standardized outcome vectors.

Multivariate extension of the sign concept, the spatial sign vector, is defined as

$$\mathbf{S}(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|^{-1}\mathbf{x} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0} \end{cases}$$

where $\|\mathbf{x}\| = (\mathbf{x}^\top \mathbf{x})^{1/2}$ is the Euclidean length of vector $\mathbf{x}$. Spatial signs are clearly rotation equivariant but not affine equivariant.

Let $\mathbf{V}$ denote the scatter matrix defined by Tyler (1987), which is the solution to

$$\mathsf{E}\left(\frac{\mathbf{V}^{1/2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{V}^{1/2}}{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right) = \frac{1}{p}\mathbf{I}_p.$$

Tyler's scatter matrix is affine equivariant, but unique only up to a multiplication by a constant; we will choose the symmetric version with $\mathrm{Tr}(\mathbf{V}) = p$. For a sign based analysis, it suffices to standardize by $\mathbf{z}_{ij} = \hat{\mathbf{V}}^{-1/2}(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_i)$. A location estimate is needed as well, and its selection will be discussed in the subsequent sections. The standardization is an analogue to the Mahalanobis transformation in the classical multivariate analysis, but instead of standardizing the sample variance-covariance matrix of the original data, this standardization produces a standardized variance-covariance matrix for the spatial sign vectors. For standardized data, the sign vectors then tend to lie uniformly on the unit sphere (see Figures 1, 2 and 3). Denote the direction vectors by $\mathbf{u}_{ij} = \mathbf{S}(\mathbf{z}_{ij})$ and the radius by $r_{ij} = \|\mathbf{z}_{ij}\|$.

Again assume that the outcome vectors are generated from

$$\mathbf{x}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij},$$

where the residuals can be decomposed as $\boldsymbol{\varepsilon}_{ij} = \boldsymbol{\Sigma}^{1/2} r_{ij}\mathbf{u}_{ij}$. Moving roughly from strong to minimal conditions, different model assumptions of the underlying distribution in terms of the direction vector $\mathbf{U}_{ij}$ and the radius $R_{ij} \geq 0$ can be listed as follows (Randles 2000):

1. Multivariate normal
   - $\mathbf{U}_{ij}$ is uniformly distributed on a p-dimensional unit sphere,
   - $R_{ij}^2 \sim \chi^2_p$, and
   - $\mathbf{U}_{ij}$ and R_{ij} are independent.
2. Elliptical symmetry
   - $\mathbf{U}_{ij}$ is uniformly distributed on a p-dimensional unit sphere and

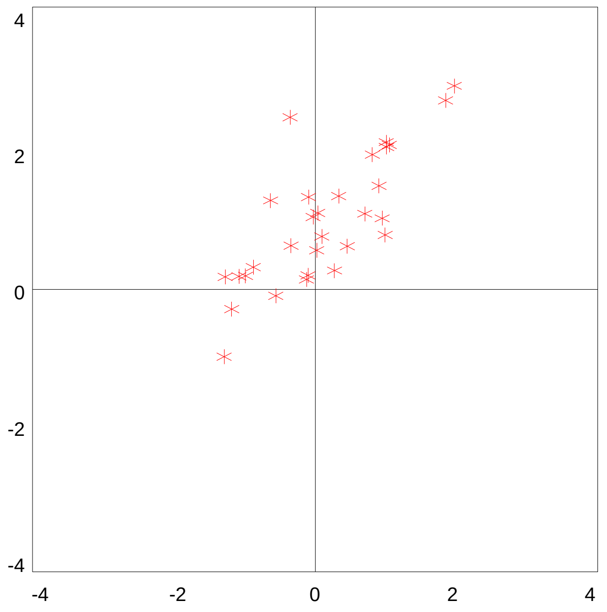

Figure 1: Observations $\mathbf{x}_j$ from a bivariate normal distribution

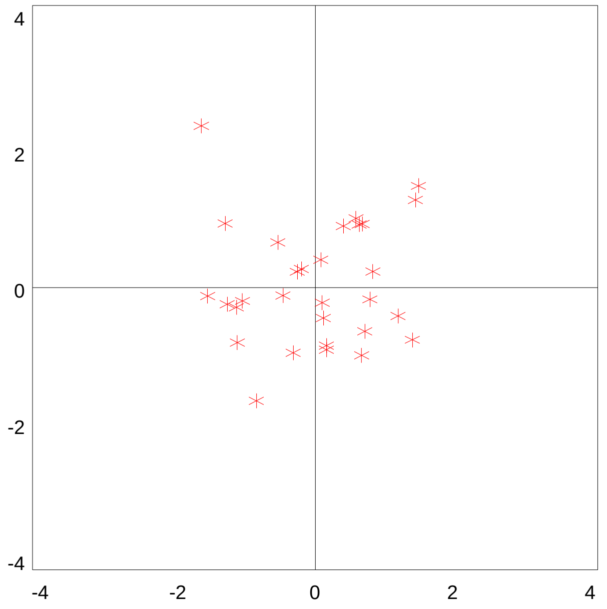

Figure 2: Standardized observations $\mathbf{z}_j$

- $\mathbf{U}_{ij}$ and R_{ij} are independent.

3. Elliptical directions
   - $\mathbf{U}_{ij}$ is uniformly distributed on a p-dimensional unit sphere.
4. Symmetry
   - $R_{ij}\mathbf{U}_{ij}$ has the same distribution as $-R_{ij}\mathbf{U}_{ij}$.
5. Directional symmetry
   - $\mathbf{U}_{ij}$ has the same distribution as $-\mathbf{U}_{ij}$.

The families are not subsets of each other: for a hierarchy between these symmetry assumptions see Randles (2000). Multivariate spatial sign methods are typically distribution-free in the family of elliptical directions. If the underlying distribution is skewed, the location parameter $\boldsymbol{\mu}$ is the population median vector rather than the mean vector (symmetry center in models 1, 2 and 4). Similarly, $\boldsymbol{\Sigma}$ is the covariance matrix in the multivariate normal model, and proportional to the covariance matrix (if it exists) in the elliptical symmetry model.

3.1. One-sample case

Consider testing the null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against the alternative hypothesis $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ (without loss of generality). For standardized signs $\mathbf{u}_j = \mathbf{S}\left(\hat{\mathbf{V}}^{-1/2}(\mathbf{x}_j - \mathbf{0})\right)$, seek an estimate of Tyler's scatter matrix as the solution to the implicit equation

$$\text{ave}\left\{\mathbf{u}_j\mathbf{u}_j^\top\right\} = \frac{1}{p}\mathbf{I}_p.$$

Obviously, the estimate is not influenced by r_j. Hence, a distribution-free test in the family of elliptical directions is given by

Lemma 3 *Under the null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$, the limiting distribution of the multivariate spatial sign test statistic*

$$Q^2 = Np\|\text{ave}\{\mathbf{u}_j\}\|^2$$

is χ^2 with p degrees of freedom.

The development was given by Randles (2000). The test statistic Q^2 is affine invariant.

For small samples, Randles (2000) proposes the use of a sign change test. For the family of directionally symmetric distributions, it leads into a conditionally distribution-free test. Let $\mathbf{U}$ denote a $p \times N$ matrix with $\mathbf{u}_j$ as the jth column. Furthermore, let $\mathbf{S}_1, \ldots, \mathbf{S}_M$, denote independent random $N \times N$ diagonal sign change matrices with 2^N equiprobable values of $\text{diag}(\pm 1, \ldots, \pm 1)$. Since $\hat{\mathbf{V}}$ is sign change invariant, the p-value can be estimated by

$$\hat{p} = \frac{\#\{Q^2(\mathbf{US}_m) \geq Q^2(\mathbf{U})\}}{M},$$

that is, by the proportion of cases where $Q^2(\mathbf{US}_m) \geq Q^2(\mathbf{U})$, $m = 1, ..., M$.

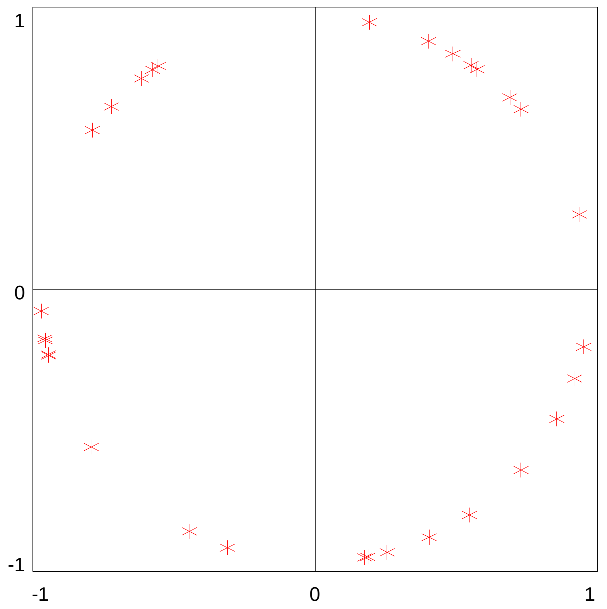

Figure 3: Direction vectors $\mathbf{u}_j$

Möttönen, Oja, and Tienari (1997) studied the limiting efficiency of multivariate sign tests for multivariate t-distributions. They show that the efficiency relative to Hotelling's test is 0.785 even in a bivariate normal case (∞ degrees of freedom). In four dimensions, they obtained relative efficiencies of 0.884, 1.051 and 2.250 for ∞, 10 and 4 degrees of freedom, respectively. In dimension 10, the same figures were 0.951, 1.131 and 2.422. See also Randles (1989) for the family of elliptically symmetric power family distributions.

Hettmansperger and Randles (2002) introduced the simultaneous estimation of location and scatter in the one-sample case. They computed a multivariate location estimate and a scatter matrix estimate to satisfy

$$\text{ave}\{\mathbf{u}_j\} = \mathbf{0} \text{ and } \text{ave}\{\mathbf{u}_j\mathbf{u}_j^\top\} = \frac{1}{p}\mathbf{I}_p \tag{2}$$

for standardized signs $\mathbf{u}_j = \mathbf{S}\left(\hat{\mathbf{V}}^{-1/2}(\mathbf{x}_j - \hat{\boldsymbol{\mu}})\right)$. The solutions to the equations are the transformation-retransformation spatial median and Tyler's scatter matrix, respectively. Standardization by the resulting location and scatter estimates distributes direction vectors uniformly into a unit sphere centered at $\mathbf{0}$ (Figure 3). Another important property of the estimates is that they are affine equivariant. The property is reached by the above utilization of the transformation-retransformation procedure (Chakraborty *et al.* 1998).

3.2. Several samples case

Next consider c independent random samples with cumulative distribution functions $F(\mathbf{x} - \boldsymbol{\mu}_1), F(\mathbf{x} - \boldsymbol{\mu}_2), \ldots, F(\mathbf{x} - \boldsymbol{\mu}_c)$, i.e. it is assumed that the underlying distributions have a joint scatter matrix and differ only in location. Our interest is to test the null hypothesis of no treatment difference

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_c$$

or, equivalently,

$$H_0 : \mathbf{\Delta}_{12} = \cdots = \mathbf{\Delta}_{1c} = \mathbf{0}.$$

Furthermore, we wish to estimate the centers of symmetry $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c$ for each sample, and the treatment differences $\mathbf{\Delta}_{12}, \ldots, \mathbf{\Delta}_{1c}$ with respect to a reference location $\boldsymbol{\mu}_1$.

Start by constructing the standardized sign vectors $\mathbf{u}_{ij} = \mathbf{S}\left(\hat{\mathbf{V}}^{-1/2}(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}})\right)$, where $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$ are the null case estimates (obtained as in the one-sample estimation case). Then, if $\hat{\mathbf{V}}$ is a $\sqrt{N}$-consistent estimate and $\hat{\boldsymbol{\mu}}$ the corresponding transformation-retransformation spatial median, we have for elliptical F that

Lemma 4 *Under the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_c$, the multisample multivariate spatial sign test statistic*

$$Q^2 = p \sum_{i=1}^{c} n_i \|\text{ave}_j\{\mathbf{u}_{ij}\}\|^2 \tag{3}$$

has a limiting χ^2 distribution with $p\,(c-1)$ degrees of freedom.

("ave$_j$" means the average taken over j.) A conditionally distribution-free test can be obtained by permuting (Oja and Randles 2004): Let $\mathbf{P}_1, \ldots, \mathbf{P}_M$ denote random $N \times N$ permutation matrices with $N!$ equiprobable values obtained by permuting the rows of an identity matrix $(N \times N)$. As $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$ are permutation invariant, p-value can be estimated as

$$\hat{p} = \frac{\#\{Q^2(\mathbf{UP}_m) \geq Q^2(\mathbf{U})\}}{M}$$

where $\mathbf{U}$ is the data set consisting of standardized signs.

The test statistic resembles the Hotelling's trace test statistic (1) in a classical MANOVA setting. But (3) is based on the directions only. For limiting efficiencies, see Randles (1989) and Möttönen *et al.* (1997).

Figure 4 displays an illustration of a bivariate non-null case. The direction vectors of the two samples are concentrated on different parts of the unit circle.

Estimation is extended to a c-sample case as follows. Choose $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c$ and $\hat{\mathbf{V}}$ so that they satisfy

$$\text{ave}\{\mathbf{u}_{1j}\} = \cdots = \text{ave}\{\mathbf{u}_{cj}\} = \mathbf{0} \text{ and } \text{ave}\left\{\mathbf{u}_{ij}\mathbf{u}_{ij}^\top\right\} = \frac{1}{p}\mathbf{I}_p$$

for standardized signs $\mathbf{u}_{ij} = \mathbf{S}\left(\hat{\mathbf{V}}^{-1/2}(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_i)\right)$. The resulting estimates are the sample transformation-retransformation spatial medians utilizing a joint Tyler's scatter matrix. Due to the affine equivariance property, the differences $\mathbf{\Delta}_{12}, \ldots, \mathbf{\Delta}_{1c}$ can be constructed as the differences of the transformation-retransformation spatial medians.

3.3. Estimation of accuracy

The following asymptotic result gives a way to approximate the precision of the estimates.

Lemma 5 *In the elliptically symmetric case*

$$\sqrt{N}\,(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \longrightarrow_D N_p\left(\mathbf{0}, \frac{p}{(p-1)^2}\,\left[\mathsf{E}\left[r^{-1}\right]\right]^{-2}\mathbf{V}\right).$$

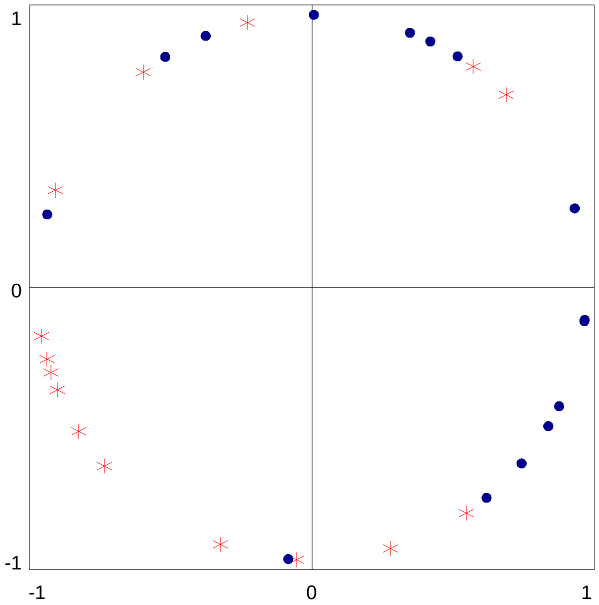

Figure 4: Direction vectors $\mathbf{u}_{ij}$ from two bivariate normal distributions

Therefore, an estimate of the covariance matrix is achieved by

$$\widehat{\mathsf{COV}}(\hat{\boldsymbol{\mu}}) = \frac{p}{N(p-1)^2}\left[\text{ave}\{r_j^{-1}\}\right]^{-2}\hat{\mathbf{V}}.$$

See for example Ollila, Oja, and Croux (2003b), Ollila, Hettmansperger, and Oja (2003a), and Hettmansperger and Randles (2002). In the case of several samples, an estimate $\widehat{\mathsf{COV}}(\hat{\boldsymbol{\mu}}_i)$ is obtained by replacing N by n_i in Lemma 5. The covariance matrix estimate of $\hat{\boldsymbol{\Delta}}_{1j}$ is easily obtained as $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c$ are asymptotically independent.

Another possibility to estimate precision is to use distribution-free methods such as bootstrapping and delete-1 jackknife. These methods are quite attractive since they require no assumption of the underlying distribution or, assuming that some basic prerequisites are fulfilled, a large sample size.

To get a bootstrap covariance matrix estimate, generate bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$ by sampling (with replacement) from the observed sample $\mathbf{X}$, keeping sample size fixed. Then compute the desired estimate from each bootstrap sample.

Lemma 6 *A bootstrap estimator of the covariance matrix of $\hat{\boldsymbol{\mu}}$ is*

$$\widehat{\mathsf{COV}}(\hat{\boldsymbol{\mu}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\mu}}_b^* - \text{ave}\{\hat{\boldsymbol{\mu}}^*\}\right)\left(\hat{\boldsymbol{\mu}}_b^* - \text{ave}\{\hat{\boldsymbol{\mu}}^*\}\right)^\top$$

where $\hat{\boldsymbol{\mu}}_b^* = \hat{\boldsymbol{\mu}}(\mathbf{X}_b^*)$ is the location estimate from the bth bootstrap sample.

In case of more than one sample, we wish to make use of the model assumption of a common scatter matrix $\mathbf{V}$. After standardization by the estimates $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{V}}$, $\mathbf{z}_{ij}$ vectors are approximately ”independent and identically distributed”. The idea is to sample (with replacement) from the data set (as if it were one sample)

$$\mathbf{Z} = \{\mathbf{z}_{11}, \ldots, \mathbf{z}_{1n_1}, \cdots, \mathbf{z}_{c1}, \ldots, \mathbf{z}_{cn_c}\}.$$

Then transform each $\mathbf{z}_{ij}^*$ back to obtain $\mathbf{x}_{ij}^* = \hat{\mathbf{V}}^{1/2}\mathbf{z}_{ij}^* + \hat{\boldsymbol{\mu}}_i$. These vectors then constitute the bootstrap sample $\mathbf{X}^* = \left(\mathbf{x}_{ij}^*\right)$. Then we can proceed as usual.

Note that some healthy caution is needed with bootstrapping. As pointed out by Stromberg (1997), there is in fact a "high" probability of generating a single bootstrap sample with an unusually large proportion of outlying observations. This proportion might even exceed the breakdown point of the estimator. Thus, even for robust methods, bootstrap estimation may sometimes fail in the presence of outliers. Yet another problem could arise when the sample size is small compared to the dimension of the data.

To overcome possible limitations of bootstrapping, an alternative approach is a delete-1 jackknife estimator.

Lemma 7 *The delete-1 jackknife estimator of covariance matrix of $\hat{\boldsymbol{\mu}}$ in the one-sample case is*

$$\widehat{\mathsf{COV}}(\hat{\boldsymbol{\mu}}) = \frac{N-1}{N} \sum_{i=1}^{N} \left(\hat{\boldsymbol{\mu}}^{(i)} - \hat{\boldsymbol{\mu}}\right) \left(\hat{\boldsymbol{\mu}}^{(i)} - \hat{\boldsymbol{\mu}}\right)^\top$$

where $\hat{\boldsymbol{\mu}}^{(i)}$ is the location estimate from a sample without the ith observation.

We have not used jackknife methods in a case of several samples.

Delete-1 jackknife does not always work well, for example, in conjunction with a nonsmooth estimator such as the vector of marginal medians (Shao and Wu 1989). However, delete-1 jackknife appears to perform nicely with the transformation-retransformation spatial median.

4. SAS/IML modules

The programs are organized as macros, which consist of frequently used modules and the master code. This section outlines the functionality of the modules, so that an advanced user can modify and make further use of them. The SAS/IML programs (`sgnmanova_1.sas` and `sgnmanova_c.sas`) are available at http://www.jstatsoft.org/v16/i05/.

4.1. Modules for estimation of location and scatter

Modules `estimate_1` (one-sample case) and `estimate_c` (c-sample case) perform the estimation procedure. The estimation algorithm uses the steps

1. Compute the direction vectors $\mathbf{u}_{ij}$ by the current estimate values.
2. Update $\hat{\mathbf{V}}$.
3. Update $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c$.
4. Return to 1 and continue until convergence.

Vector of the componentwise medians and a $p \times p$ identity matrix are used as starting values. Iteration steps are given by

$$\begin{aligned}\hat{\mathbf{V}} &\leftarrow p\,\hat{\mathbf{V}}^{1/2}\,\text{ave}_{ij}\left\{\mathbf{u}_{ij}\mathbf{u}_{ij}^\top\right\}\,\hat{\mathbf{V}}^{1/2} \text{ and}\\ \hat{\boldsymbol{\mu}}_i &\leftarrow \hat{\boldsymbol{\mu}}_i + \left[\text{ave}_j\left\{r_{ij}^{-1}\right\}\right]^{-1}\hat{\mathbf{V}}^{1/2}\text{ave}_j\{\mathbf{u}_{ij}\}.\end{aligned}$$

(Hettmansperger and Randles 2002; Vardi and Zhang 2001; Oja and Randles 2004). The symmetric transformation matrix $\hat{\mathbf{V}}^{-1/2}$ is found via the spectral decomposition of the matrix $\hat{\mathbf{V}}$.

We have also implemented a protection against landing iteration on a data point (Vardi and Zhang 2001). Their modification ensures that iteration moves on even then. The need for such protection is rare, but it does have practical value in bootstrapping, since—for some bootstrap samples—the iteration might encounter a large mass of data on a single point.

`estimate_1` Input for the module are the data matrix and the desired level of precision. The module returns a $(p+1) \times p$ matrix where the first row is the location estimate and the remaining rows consist of the scatter matrix estimate.

`estimate_c` Input for the module are the data matrix, the desired level of precision and the number of samples. The module returns a $(p+c) \times p$ matrix where the first c rows are the location estimates and the remaining rows consist of the scatter matrix estimate.

4.2. Modules for hypothesis testing

Modules for testing the null hypothesis are named as `test_1` (one-sample case) and `test_c` (multisample case).

`test_1` Input for the module are the data matrix, the desired level of precision and the number of sign change permutations. Module estimates the scatter matrix under H_0 (fixed location), and returns a 1×4 vector with value of the test statistic, a p-value based on the limiting distribution, a p-value based on a sign change permutation distribution and its standard error (from a binomial distribution) as elements.

`test_c` Input for the module are the data matrix, the desired level of precision and the number of permutations. Calls the `estimate_1` module. The module returns a 1×4 vector with value of the test statistic, a p-value based on the limiting distribution, a p-value based on a permutation distribution and its standard error (from binomial distribution) as elements.

Small number of permutations guarantees a reasonable computation time.

4.3. Modules for estimation of accuracy

Module `asymptotic` estimates the covariance matrix of $\hat{\boldsymbol{\mu}}$ based on the limiting distribution. Similarly, module `bootstrap` estimates the covariance matrix by bootstrapping, and module `jackknife` estimates it by the delete-1 jacknife. Note that `jackknife` module is available only for the one-sample case.

`asymptotic` Input for the module consist of the data matrix, the estimated parameters values and the desired level of precision. In a multisample case the number of samples has to be given as well. The module returns a $p \times p$ covariance matrix estimate in a one-sample case, and a $cp \times (p+1)$ matrix in a multisample case, where the first column identifies the rows which contain the covariance matrix estimate of $\hat{\boldsymbol{\mu}}_i$.

`bootstrap` Input for the module consist of the data matrix, the desired level of precision and the number of bootstrap samples. In a multisample case the number of samples has to be given as well. Calls the respective `estimate` modules. The module returns a $p \times p$ covariance matrix estimate in a one-sample case, and a $cp \times (p+1)$ matrix in a multisample case, where the first column identifies the rows which contain the covariance matrix of $\hat{\boldsymbol{\mu}}_i$.

`jackknife` Input for the module consist of the data matrix, location estimate and the desired level of precision. Calls the `estimate_1` module. The module returns a $p \times p$ covariance matrix estimate.

It is a good idea to start with a small number of bootstrap samples.

5. Examples

5.1. Multivariate normal distribution

We simulated a two-sample case ($n_1 = n_2 = 50$)

$$\mathbf{x}_{1j} \sim N_3(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ and } \mathbf{x}_{2j} \sim N_3(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 1 \\ & 3 & 1 \\ & & 3 \end{pmatrix}.$$

One-sample analysis. We start by analysing the first sample data as a one-sample problem. The null hypothesis of interest is $H_0 : \boldsymbol{\mu}_1 = \mathbf{0}$. The SAS statements

```
%INCLUDE ’<full path>\sgnmanova_1.sas’;
sgnmanova_1(y3onesam, eps=1E-9, nperm=1000, nboot=500);
```

produce the output

```
                                      Q2
Value of the test statistic:   0.6305067

                                    P_AS
p-value (large sample appr.):  0.8894144

                              P_PERM      SE_P
p-value (sign change test):    0.898 0.0095706 ( 1000 permutations)
```

The null hypothesis is thus not rejected; p-values based on the limiting χ^2_3-distribution and on a permutation distribution were 0.889 and 0.898, respectively. The estimate of $\boldsymbol{\mu}_1$ and covariance matrix estimates of $\hat{\boldsymbol{\mu}}_1$ are obtained from the output as well (the output is reduced to fit it on the page):

```
   MU
-0.054617 0.0340526 0.0930009

   COV_AS
0.0236718 0.0211768 0.0165879
0.0211768 0.0594587 0.0227896
0.0165879 0.0227896 0.049833

   COV_BS
0.025025  0.0262765 0.0181631
0.0262765 0.0674611 0.0308509
0.0181631 0.0308509 0.0511991

   COV_JK
0.0238118 0.0238157 0.0178145
0.0238157 0.061081  0.0289153
0.0178145 0.0289153 0.0480224
```

The subindices ”AS”, ”BS” and ”JK” refer to the approximation method by asymptotics, bootstrapping and jackknife, respectively (see Section 3.3). The estimates are very similar. For comparison, the sample mean vector is $\bar{\mathbf{x}}_1 = (-0.02\ 0.09\ 0.06)^\top$ and the estimated covariance matrix of the sample mean is

$$\begin{pmatrix} 0.020 & 0.020 & 0.014 \\ & 0.054 & 0.018 \\ & & 0.048 \end{pmatrix}.$$

The mean is slightly more accurate in the normal case. But, if just one observation of the data set is contaminated (by adding, say, 10 to all its components), the covariance matrix estimates are:

$$\begin{aligned} \widehat{\mathsf{COV}}_{\mathrm{AS}}(\hat{\boldsymbol{\mu}}_1) &= \begin{pmatrix} 0.027 & 0.024 & 0.019 \\ & 0.062 & 0.025 \\ & & 0.048 \end{pmatrix}, \ \widehat{\mathsf{COV}}_{\mathrm{BS}}(\hat{\boldsymbol{\mu}}_1) = \begin{pmatrix} 0.028 & 0.027 & 0.021 \\ & 0.068 & 0.031 \\ & & 0.051 \end{pmatrix}, \\ \widehat{\mathsf{COV}}_{\mathrm{JK}}(\hat{\boldsymbol{\mu}}_1) &= \begin{pmatrix} 0.026 & 0.024 & 0.019 \\ & 0.062 & 0.030 \\ & & 0.044 \end{pmatrix} \text{ and } \widehat{\mathsf{COV}}(\bar{\mathbf{x}}_1) = \begin{pmatrix} 0.059 & 0.059 & 0.042 \\ & 0.093 & 0.046 \\ & & 0.065 \end{pmatrix}.. \end{aligned}$$

The covariance matrix of $\hat{\boldsymbol{\mu}}_1$ is almost unaffected, but the covariance matrix of the sample mean nearly doubles in size. This reflects the robustness of the spatial median against outliers. Despite of a single outlier, bootstrapping worked well. We will return to the robustness studies in the two-sample case.

Two-sample analysis. Now we move on to the sample comparisons. The interest is to test for differences in location, and to estimate the location, shift and scatter. Analysis for the two-sample data set was performed by the SAS statements

MANOVA	Parameter	Estimate	Standard error
Spatial sign	$\boldsymbol{\mu}_1$	$(-0.05\ 0.03\ 0.09)^\top$	$(0.15\ 0.28\ 0.24)^\top$
	$\boldsymbol{\mu}_2$	$(0.71\ 0.55\ 0.93)^\top$	$(0.15\ 0.28\ 0.24)^\top$
	$\boldsymbol{\Delta}_{12}$	$(0.77\ 0.52\ 0.85)^\top$	$(0.21\ 0.39\ 0.34)^\top$
Classical	$\boldsymbol{\mu}_1$	$(-0.02\ 0.09\ 0.06)^\top$	$(0.14\ 0.26\ 0.22)^\top$
	$\boldsymbol{\mu}_2$	$(0.71\ 0.55\ 0.91)^\top$	$(0.14\ 0.26\ 0.22)^\top$
	$\boldsymbol{\Delta}_{12}$	$(0.73\ 0.46\ 0.86)^\top$	$(0.20\ 0.37\ 0.32)^\top$

Table 1: Estimates for location and shift. Standard errors are based on large sample approximations.

```
%INCLUDE '<full path>\sgnmanova_c.sas';
sgnmanova_c(y3, 2, 1E-9, 1000, 500);
```

Resulting location and shift estimates are shown in Table 1. The sample covariance matrix and Tyler's scatter matrix (used to transform the data), both standardized to $\mathrm{Tr}(\cdot) = 3$, are very much alike:

$$\frac{3}{\mathrm{Tr}(\mathbf{S})}\mathbf{S} = \begin{pmatrix} 0.42 & 0.51 & 0.38 \\ & 1.48 & 0.42 \\ & & 1.10 \end{pmatrix} \text{ and } \hat{\mathbf{V}} = \begin{pmatrix} 0.42 & 0.47 & 0.43 \\ & 1.48 & 0.50 \\ & & 1.10 \end{pmatrix}$$

Test results are presented in Table 2. To demonstrate the robustness of the multivariate spatial sign test we contaminated the elements of a single observation in the first sample by adding a positive constant to all its elements. The effect on the multivariate spatial test is small, but Hotelling's trace test fails completely for large contamination values.

Contamination	Hotelling's	Multivariate spatial sign test	
factor	trace	χ^2_3	1000 permutations
none	0.002	0.006	0.003 (0.002)
1	0.003	0.008	0.007 (0.003)
10	0.155	0.013	0.011 (0.003)
100	0.726	0.014	0.012 (0.003)

Table 2: p-values for testing $H_0 : \boldsymbol{\Delta}_{12} = \mathbf{0}$. The standard error of the p-value estimate is given in parentheses.

Naturally, the same phenomenon is reflected in the corresponding estimates. For a contamination factor of 100,

$$\begin{aligned} \hat{\boldsymbol{\mu}}_1 &= (\ -0.02 \quad 0.07 \quad 0.17\ )^\top, \text{ and} \\ \bar{\mathbf{x}}_1 &= (\ 1.98 \quad 2.09 \quad 2.06\ )^\top. \end{aligned}$$

The Hettmansperger-Randles estimate is still close to the true value, but the sample mean vector is totally destroyed.

5.2. Multivariate Cauchy distribution

In this section we study the behavior of the estimates for a heavy-tailed error distribution. We simulated a data set ($N = 50$) from a multivariate Cauchy distribution using the model

$$\mathbf{x}_j = \frac{\mathbf{y}_j}{z_j},$$

where $\mathbf{y}_j \sim N_3(\mathbf{0}, \mathbf{I}_3)$, $z_j^2 \sim \chi_1^2$, and $\mathbf{y}_j$ and z_j are independent. Then $\mathbf{x}_j$ has a spherical multivariate Cauchy distribution. The distribution does not possess finite moments, and it has very heavy tails.

An analysis by `sgnmanova_1` macro gives

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \begin{pmatrix} 0.01 & -0.06 & -0.19 \end{pmatrix}^\top \\ \hat{\mathbf{V}} &= \begin{pmatrix} 0.902 & 0.107 & 0.166 \\ & 1.124 & -0.081 \\ & & 0.974 \end{pmatrix} \end{aligned}$$

i.e. natural estimates of the center of symmetry and the spatial sign covariance. Different covariance matrix estimates of the location estimate are

$$\begin{aligned} \widehat{\mathsf{COV}}_{\mathrm{AS}}(\hat{\boldsymbol{\mu}}) &= \begin{pmatrix} 0.034 & 0.004 & 0.006 \\ & 0.043 & -0.003 \\ & & 0.037 \end{pmatrix}, \\ \widehat{\mathsf{COV}}_{\mathrm{BS}}(\hat{\boldsymbol{\mu}}) &= \begin{pmatrix} 0.048 & 0.011 & 0.012 \\ & 0.050 & -0.002 \\ & & 0.046 \end{pmatrix}, \text{ and} \\ \widehat{\mathsf{COV}}_{\mathrm{JK}}(\hat{\boldsymbol{\mu}}) &= \begin{pmatrix} 0.039 & 0.011 & 0.007 \\ & 0.050 & -0.001 \\ & & 0.035 \end{pmatrix}, \end{aligned}$$

giving results mainly in the same direction. Due to the extreme values generated by the underlying Cauchy distribution, bootstrapping appears to slightly overestimate the elements of the variance-covariance matrix.

Due to the lack of finite moments of the noise distribution, a classical analysis is not helpful at all:

$$\begin{aligned} \bar{\mathbf{x}} &= \begin{pmatrix} 0.36 & -1.72 & 0.04 \end{pmatrix}^\top \\ \mathbf{S} &= \begin{pmatrix} 17.720 & -38.555 & -2.000 \\ & 142.475 & 3.274 \\ & & 6.156 \end{pmatrix} \end{aligned}$$

By coincidence, the p-values were close to each other: $p = 0.749$ and $p = 0.776$ for the spatial sign test and the Hotelling's T^2 test, respectively.

6. Concluding remarks

Hettmansperger and Randles (2002) recognized that the conditions for the existence and the uniqueness of simultaneous solutions to the estimating equations have not been established. In authors' experience, however, the algorithm appears always to converge.

Lopuhaä and Rousseeuw (1991) showed that the spatial median has a 50% breakdown point. The breakdown point of Tyler's scatter matrix is positive, and generally within the interval $[\,1/(p+1), 1/p\,]$. Both the location estimator and the scatter estimator have bounded influence functions. Given these robustness qualities, the minimal model assumptions and the good efficiency properties, multivariate spatial sign methods are attractive alternatives to the classical procedures particularly for skewed or heavy-tailed distributions, or in the presence of outliers.

Acknowledgements

Authors wish to thank the Academy of Finland and the School of Statistical Information, Inference and Data Analysis (SIIDA) for supporting this research.

References

Brown BM (1983). "Statistical Uses of the Spatial Median." *Journal of the Royal Statistical Society B*, **45**, 25–30.

Chakraborty B, Chaudhuri P, Oja H (1998). "Operating Transformation Retransformation on Spatial Median and Angle Test." *Statistica Sinica*, **8**, 767–784.

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

Lopuhaä HP, Rousseeuw PJ (1991). "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices." *The Annals of Statistics*, **19**, 229–248.

Möttönen J, Oja H (1995). "Multivariate Spatial Sign and Rank Methods." *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.

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

Ollila E, Hettmansperger TP, Oja H (2003a). "Affine Equivariant Multivariate Sign Methods." Under revision.

Ollila E, Oja H, Croux C (2003b). "The Affine Equivariant Sign Covariance Matrix: Asymptotic Behavior and Efficiencies,." *Journal of Multivariate Analysis*, **87**, 328–355.

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-invariant, Multivariate, Distribution-free Sign Test." *Journal of the American Statistical Association*, **95**, 1263–1268.

Shao J, Wu CFJ (1989). "A General Theory for Jackknife Variance Estimation." *The Annals of Statistics*, **17**, 1176–1197.

Stromberg AJ (1997). "Robust Covariance Estimates Based on Resampling,." *Journal of Statistical Planning and Inference*, **57**, 321–334.

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

Vardi Y, Zhang CH (2001). "A Modified Weiszfeld Algorithm for the Fermat-Weber Location Problem." *Mathematical Programming A*, **90**, 559–566.

Affiliation:

Jaakko Nevalainen
Department of Mathematics, Statistics and Philosophy
33014 University of Tampere, Finland
E-mail: jaakko.nevalainen@uta.fi
URL: http://www.uta.fi/~jaakko.nevalainen/

Hannu Oja
Tampere School of Public Health
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 16, Issue 5 *Submitted:* 2004-11-23
May 2006 *Accepted:* 2006-05-17