The Annals of Statistics
2003, Vol. 31, No. 5, 1439–1459

A SCATTER MATRIX ESTIMATE BASED ON THE ZONOTOPE[1]

BY GLEB A. KOSHEVOY, JYRKI MÖTTÖNEN AND HANNU OJA

Russian Academy of Sciences, University of Oulu and University of Jyväskylä

We introduce a new scatter matrix functional which is a multivariate affine equivariant extension of the mean deviation $E(|x - \mathrm{Med}(x)|)$. The estimate is constructed using the data vectors (centered with the multivariate Oja median) and their angular distances. The angular distance is based on Randles interdirections. The new estimate is called the zonoid covariance matrix (the ZCM), as it is the regular covariance matrix of the centers of the facets of the zonotope based on the data set. There is a kind of symmetry between the zonoid covariance matrix and the affine equivariant sign covariance matrix; interchanging the roles of data vectors and hyperplanes yields the sign covariance matrix as the zonoid covariance matrix. (It turns out that the symmetry relies on the zonoid of the distribution and its projection body which is also a zonoid.) The influence function and limiting distribution of the new scatter estimate, the ZCM, are derived to consider the robustness and efficiency properties of the estimate. Finite-sample efficiencies are studied in a small simulation study. The influence function of the ZCM is unbounded (linear in the radius of the contamination vector) but less influential in the tails than that of the regular covariance matrix (quadratic in the radius). The estimate is highly efficient in the multivariate normal case and performs better than the regular covariance matrix for heavy-tailed distributions.

1. Introduction. Throughout the paper we assume that $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a random sample from a symmetric k-variate distribution with cumulative distribution function F having finite second moments. We wish to estimate the unknown symmetry center $\boldsymbol{\mu}$ and the unknown covariance matrix Σ. By symmetry we mean that $\mathbf{x}_i - \boldsymbol{\mu}$ and $\boldsymbol{\mu} - \mathbf{x}_i$ have the same distribution. A *location vector* $T(F)$ is an affine equivariant functional, and for symmetric distributions its value is the symmetry center. A $k \times k$ matrix valued functional $C = C(F)$ is a *scatter matrix* if it is symmetric, positive definite and affine equivariant.

Consider a standardized k-variate random variable $\mathbf{z}$ with mean vector $\mathbf{0}$ and covariance matrix I_k. Assume that the distribution of $\mathbf{z}$ with c.d.f. F_0 is *reflection and permutation invariant*, that is, $G\mathbf{z} \sim \mathbf{z}$, for all $k \times k$ reflection and permutation matrices G. (A reflection matrix is a diagonal matrix with diagonal elements $+1$ or -1; a permutation matrix is obtained by permuting the rows or columns of

Received January 2002; revised December 2002.

[1]Supported by the Academy of Finland.

AMS 2000 subject classification. 62H12.

Key words and phrases. Interdirection, parallelotope, multivariate mean deviation, multivariate median, multivariate signs, sign covariance matrix, scatter matrix, sign covariance matrix, zonoid, zonoid covariance matrix, zonotope.

the identity matrix.) Note that the permutation invariance property means that the components of $\mathbf{z}$ are exchangeable. The margins of $\mathbf{z}$ are identically distributed, symmetric around zero and uncorrelated. The standardized variable $\mathbf{z}$ generates the corresponding *location-scale model* as a totality of distributions of

$$\mathbf{x} = \Sigma^{1/2}\mathbf{z} + \boldsymbol{\mu}$$

for every positive definite $k \times k$ matrix Σ and k-vector $\boldsymbol{\mu}$. An elliptical model is an important special case. Then if F is the c.d.f. of $\mathbf{x}$, $T(F) = \boldsymbol{\mu}$ and $C(F) = \kappa^2\Sigma$ for all location vectors T and all scatter matrices C. As κ^2 depends on the functional C and the distribution F_0, a correction factor κ^{-2} is needed for the Fisher consistency of the estimate of Σ and for the comparisons between different scatter matrix estimates at a specific model.

Visuri, Koivunen and Oja (2000), Ollila, Hettmansperger and Oja (2002), Ollila, Oja and Croux (2002) and Visuri, Ollila, Koivunen, Möttönen and Oja (2003) introduced and investigated scatter matrices based on the Oja sign and rank vectors [Oja (1999)]. The affine equivariant sign vectors were built using hyperplanes (going through the origin and $k-1$ data points) and the normals to these hyperplanes. The sign covariance matrix is then the regular covariance matrix calculated from the multivariate signs of the centered observations. For the scatter matrix estimate based on the sign covariance matrix and its statistical properties, see Ollila, Oja and Croux (2002). In this paper we construct in a symmetric way, interchanging the roles of observation vectors and normals, a new scatter matrix estimate, which may be seen as the sign covariance matrix of the normals. The new estimate appears to be a multivariate affine invariant matrix valued extension of the mean deviation and it can be constructed using the data vectors (centered with the multivariate Oja median) and their angular distances. The angular distance is based on Randles' (1989) interdirections.

Koshevoy and Mosler (1997a, b, 1998) and Mosler (2002) proposed the use of zonoids and lift zonoids, k- and $(k+1)$-variate convex sets $Z(F)$ and $\mathrm{LZ}(F)$, respectively, to describe and investigate the properties of a multivariate distribution F. The volume of the zonoid $Z(F)$ is a global measure of scatter (an extension of the mean deviation) and, in the elliptic case, the shape of the zonoid is determined by the covariance structure of the distribution. The zonotope $Z(X)$, also a k-variate convex body based on data set X, is a natural estimate of zonoid $Z(F)$. It turns out that our new scatter matrix estimate can be constructed using the centers of the facets of a data based zonotope. Therefore the new estimate is named the zonoid covariance matrix. The relations between the new scatter matrix estimate and the sign covariance matrix rely on the zonoid of the distribution and its projection body (also a zonoid).

Our plan is as follows. In Section 2 we explain a kind of duality between observations and hyperplanes going through the origin and $k-1$ observations and introduce our affine equivariant scatter matrices, the SCM and ZCM, based on

Randles' (1989) angular distances. Section 3 discusses the statistical properties (influence function, limiting distribution, limiting efficiency and finite-sample efficiency) of the new scatter matrix estimate, the ZCM. In Section 4 it is shown that the new estimate may be constructed using the centers of the facets of the zonotope based on the data set. Some tools to investigate k-variate convex compacts as well as formal definitions of the concepts of the zonotope and zonoid are given. We close the paper with some final comments in Section 5.

2. Location and scatter estimates.

2.1. *Hyperplanes and interdirections.* Let $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a k-variate data set and consider hyperplanes going through the origin and $k-1$ observation points. To shorten the notation, write $I = (i_1, \ldots, i_{k-1})$ with $1 \le i_1 < \cdots < i_{k-1} \le n$ for an ordered set of indices. The new index I then refers to a $k-1$ subset of observations with indices listed in I, or to hyperplane

$$H(I) = \{\mathbf{y} \in \mathbb{R}^k : \det(\mathbf{y}_{i_1} \cdots \mathbf{y}_{i_{k-1}} \mathbf{y}) = 0\}.$$

Also, define vector $\mathbf{e}(I)$ implicitly by

$$\det(\mathbf{y}_{i_1} \cdots \mathbf{y}_{i_{k-1}} \mathbf{y}) = \mathbf{e}^T(I)\mathbf{y};$$

that is, $\mathbf{e}(I)$ is the vector of cofactors corresponding to the last column of the matrix $(\mathbf{y}_{i_1} \cdots \mathbf{y}_{i_{k-1}} \mathbf{y})$. Note that the vector $\mathbf{e}(I)$ is orthogonal to hyperplane $H(I)$ and its length is the volume of a $(k-1)$-variate parallelotope determined by vectors (segments) with indices in I. (For the definition of the parallelotope, see Section 4.) The $\mathbf{e}(I)$ are henceforth called *normals*.

The normals $\mathbf{e}(I)$ are affine equivariant in the sense that if the $\mathbf{e}^*(I)$ are constructed from the data set

$$A \cdot Y = \{A\mathbf{y}_1, \ldots, A\mathbf{y}_n\}$$

with a full rank $k \times k$ matrix A, then $\mathbf{e}^*(I) = A^*\mathbf{e}(I)$ with $A^* = \operatorname{abs}(\det(A))(A^{-1})^T$. See Ollila, Oja and Croux (2002). The normals $\mathbf{e}(I)$ are random vectors which also carry information about the covariance structure:

LEMMA 1. *If Y is a random sample from a distribution F with zero mean vector and covariance matrix Σ, then the normals $\mathbf{e}(I)$ are random vectors with*

$$E_F(\mathbf{e}(I)) = \mathbf{0} \quad \textit{and} \quad E_F(\mathbf{e}(I)\mathbf{e}^T(I)) = (k-1)!\det(\Sigma)\Sigma^{-1}.$$

An affine invariant measure of angular closeness between vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ may be constructed using hyperplanes $H(I)$ or normals $\mathbf{e}(I)$ as follows.

DEFINITION 1. The measure of angular closeness of observation vectors $\mathbf{y}_i$ and $\mathbf{y}_j$, denoted by δ_{ij}, is given by

$$\delta_{ij} = \operatorname{ave}_I\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_j)\}.$$

The definition is natural as $\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_i)$ tells whether $\mathbf{y}_i$ is "above" or "below" hyperplane $H(I)$. Then $\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_j)$ is $+1$ when $\mathbf{y}_i$ and $\mathbf{y}_j$ are on the same side of the hyperplane $H(I)$ and -1 otherwise. The value of δ_{ij} is thus between -1 and $+1$. The closer the vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ are, the smaller the number of separating hyperplanes $H(I)$ and the larger is δ_{ij}. Note that $(1-\delta_{ij})/2$ is the well-known measure of angular distance between $\mathbf{y}_i$ and $\mathbf{y}_j$, the observed proportion of *interdirections* (hyperplanes separating $\mathbf{y}_i$ and $\mathbf{y}_j$). This concept of interdirection was first introduced by Randles (1989) and has been used to construct multivariate sign and rank and signed-rank tests. See Randles (1989), Peters and Randles (1990) and Hallin and Paindaveine (2002), for example.

Hyperplanes $H(I)$ going through the origin are uniquely defined by the directions of normals $\mathbf{e}(I)$. How should we then measure the angular closeness of normals $\mathbf{e}(I)$ and $\mathbf{e}(J)$? This can be done by exchanging the roles of observations and hyperplanes: Now let the $\mathbf{e}(I)$ be vectors in $\mathbb{R}^k$ and let the $\mathbf{y}_i$ define the hyperplanes $\{\mathbf{y}\in\mathbb{R}^k : \mathbf{y}_i^T\mathbf{y}=0\}$ going through the origin. Then an affine equivariant angular closeness of vectors $\mathbf{e}(I)$ and $\mathbf{e}(J)$ is defined in a symmetric way as follows.

DEFINITION 2. The measure of angular closeness of normals $\mathbf{e}(I)$ and $\mathbf{e}(J)$, denoted by $\Delta(I,J)$, is given by

$$\Delta(I,J)=\operatorname{ave}_i\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{y}_i)\operatorname{sign}(\mathbf{e}^T(J)\mathbf{y}_i)\}.$$

Finally note that, as $\mathbf{e}(I)$ and $-\mathbf{e}(I)$ yield the same hyperplane, the resulting measure of closeness of hyperplanes $H(I)$ and $H(J)$ is defined as $\operatorname{abs}(\Delta(I,J))$.

Recall that if $Y=\{\mathbf{y}_1,\dots,\mathbf{y}_n\}$ is a random sample from a distribution with symmetry center zero and covariance matrix Σ, then

$$\operatorname{ave}_i\{\mathbf{y}_i\mathbf{y}_i^T\}\quad\text{and}\quad\operatorname{ave}_I\{\mathbf{e}(I)\mathbf{e}^T(I)\}$$

estimate Σ and $(k-1)!\det(\Sigma)\Sigma^{-1}$, respectively. This suggests that also

$$\operatorname{ave}_{i,j}\{\delta_{ij}\mathbf{y}_i\mathbf{y}_j^T\}\quad\text{and}\quad\operatorname{ave}_{I,J}\{\Delta(I,J)\mathbf{e}(I)\mathbf{e}^T(J)\}$$

may be used to construct reasonable, affine equivariant estimates of the covariance matrix and of its inverse, respectively. In fact, the latter is the affine equivariant sign covariance matrix introduced by Visuri, Koivunen and Oja (2000) as is seen in Lemma 2 of the next section.

2.2. *Estimates based on signs.* The *multivariate Oja* (1983) *median* is defined as follows. Let $X=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ be a random sample from a k-variate symmetric distribution. To estimate the unknown symmetry center, shift the observations by a candidate $\boldsymbol{\mu}$. Write

$$Y=\{\mathbf{y}_1,\dots,\mathbf{y}_n\}=\{\mathbf{x}_1-\boldsymbol{\mu},\dots,\mathbf{x}_n-\boldsymbol{\mu}\}$$

for this data set of residuals. The affine equivariant Oja median, say $\hat{\boldsymbol{\mu}}$, then minimizes the objective function

$$D(\boldsymbol{\mu}) = \text{ave}_{i,I}\{\text{abs}(\mathbf{e}^T(I)\mathbf{y}_i)\} = \text{ave}_i\{\mathbf{S}_i^T\mathbf{y}_i\},$$

where

$$\mathbf{S}_i = \text{ave}_I\{\text{sign}(\mathbf{e}^T(I)\mathbf{y}_i)\mathbf{e}(I)\}, \qquad i = 1, \ldots, n,$$

are affine equivariant *multivariate sign vectors*. At $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$, the signs $\mathbf{S}_i$ are centered, that is, $\sum \mathbf{S}_i = \mathbf{0}$. See Oja (1999) and references therein.

Next construct the affine equivariant sign covariance matrix based on the centered signs $\mathbf{S}_1, \ldots, \mathbf{S}_n$ [Visuri, Koivunen and Oja (2000) and Ollila, Oja and Croux (2002)].

DEFINITION 3. The sign covariance matrix (SCM) is

$$\text{SCM} = \text{SCM}(X) = \text{ave}_i\{\mathbf{S}_i\mathbf{S}_i^T\}.$$

The SCM is affine equivariant in the sense that if SCM* is calculated from the data set $A \cdot X + \mathbf{b} = \{A\mathbf{x}_1 + \mathbf{b}, \ldots, A\mathbf{x}_n + \mathbf{b}\}$, then

$$\text{SCM}^* = \det(A^2)(A^{-1})^T\text{SCM}(A^{-1}).$$

Using normals and angular distances between normals, one immediately gets the following lemma.

LEMMA 2. *The sign covariance matrix satisfies*

$$\text{SCM} = \text{ave}_{I,J}\{\Delta(I, J)\mathbf{e}(I)\mathbf{e}^T(J)\}.$$

The inverse of the SCM may be used to estimate the regular covariance and correlation matrix. For the influence function, limiting distribution, efficiency and applications of the SCM see Ollila, Hettmansperger and Oja (2002), Ollila, Oja and Hettmansperger (2002) and Ollila, Oja and Croux (2002). In the next section we exchange the roles of observations and normals.

2.3. *Estimates based on signs of normals.* Again let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample from a k-variate distribution F symmetric around $\boldsymbol{\mu}$. To estimate the unknown symmetry center, again shift the observations by a candidate $\boldsymbol{\mu}$ and consider the shifted data set $Y = \{\mathbf{x}_1 - \boldsymbol{\mu}, \ldots, \mathbf{x}_n - \boldsymbol{\mu}\}$. Exchanging the roles of vectors and normals, the objective function of the Oja median may now be symmetrically written as

$$D(\boldsymbol{\mu}) = \text{ave}_{i,I}\{\text{abs}(\mathbf{e}^T(I)\mathbf{y}_i)\} = \text{ave}_I\{\mathbf{h}^T(I)\mathbf{e}(I)\},$$

where

$$\mathbf{h}(I) = \text{ave}_i\{\text{sign}(\mathbf{e}^T(I)\mathbf{y}_i)\mathbf{y}_i\}, \qquad i = 1, \ldots, n,$$

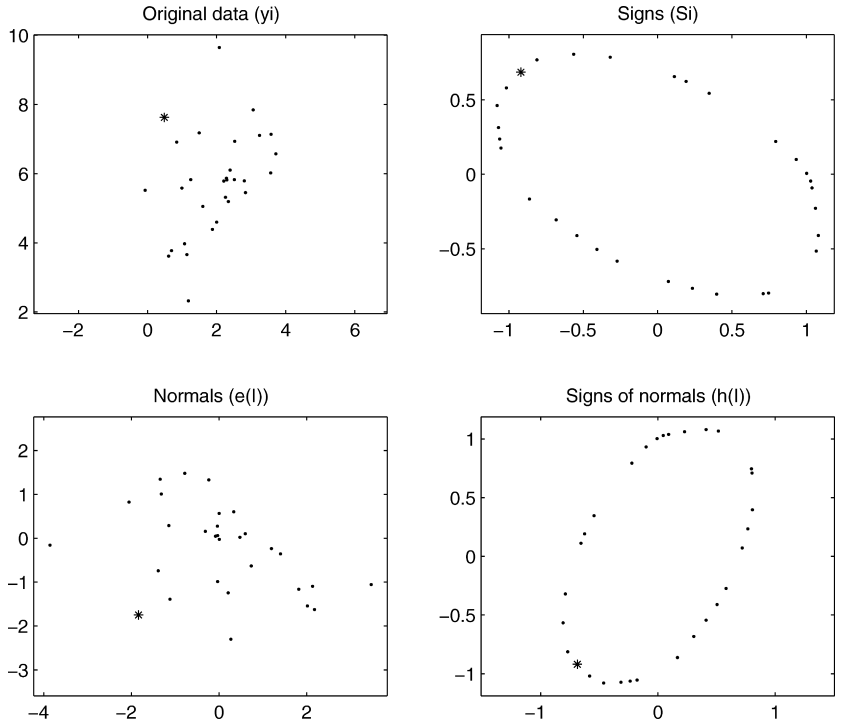

FIG. 1. *The original bivariate data vectors* $\mathbf{y}_i$, *their signs* $\mathbf{S}_i$, *normals to hyperplanes* $\mathbf{e}(I)$ *and signs of normals* $\mathbf{h}(I)$ *in the bivariate case. One of the observations* (*normals*) *is denoted by a star to illustrate the transformations*.

are *signs of normals*. Note that the signs are now based on the hyperplanes with normals $\mathbf{y}_i$. See Figure 1 for an illustration of the $\mathbf{y}_i$, $\mathbf{S}_i$, $\mathbf{e}(I)$ and $\mathbf{h}(I)$ for a small bivariate data set.

We next consider the scatter matrix estimate, which is now calculated from the signs of normals, the $\mathbf{h}(I)$. The data are again centered using the Oja median, which is a natural location estimate here also. See Section 4. We call it the *zonoid covariance matrix* (ZCM) as the $\mathbf{h}(I)$ yield all the centers of the facets of the zonotope based on Y. This will be explained in detail in Section 4.

DEFINITION 4. The zonoid covariance matrix (ZCM) based on data set X is

$$\mathrm{ZCM} = \mathrm{ZCM}(X) = \mathrm{ave}_I\{\mathbf{h}(I)\mathbf{h}^T(I)\}.$$

It is easy to see that $\mathrm{ZCM}(X)$ is a scatter matrix; that is, it is affine equivariant (in the usual sense),

$$\mathrm{ZCM}(A \cdot X + \mathbf{b}) = A\,\mathrm{ZCM}(X)\,A^T.$$

The zonoid covariance matrix can be also defined in terms of the original observation vectors and their interdirections as follows.

LEMMA 3. *The zonoid covariance matrix is*

$$\mathrm{ZCM}(X) = \mathrm{ave}_{i,j}\{\delta_{ij}\mathbf{y}_i\mathbf{y}_j^T\}.$$

Due to the centering of the data, the ZCM is invariant in shifts of the observations. Therefore, in the following derivations it is not a restriction to assume that X is a random sample from a distribution symmetric around the origin. We next show that, as the sign covariance matrix [see Ollila, Oja and Croux (2002)], the scatter matrix $\mathrm{ZCM}(X)$ also is asymptotically equivalent to a U-statistic. The general theory for the U-statistics can then be used to prove the limiting multinormality of the scatter matrix estimate (see the Appendix for the proof).

LEMMA 4. *Assume that $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a random sample from a distribution symmetric around zero. For any $K = \{i_1, \ldots, i_{k+1}\} \subset \{1, \ldots, n\}$ with $|K| = k+1$, write*

$$\begin{aligned} g(K) &= g(i_1, \ldots, i_{k+1}) \\ &= \frac{1}{k(k+1)} \sum_{I \cup \{i\} \cup \{j\} = K} \{\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T\}. \end{aligned}$$

Consider the U-statistic with symmetric kernel g,

$$U_n = \binom{n}{k+1}^{-1} \sum_K g(K).$$

Then under general assumptions (see the Appendix)

$$\sqrt{n}(U_n - \mathrm{ZCM}(X)) \xrightarrow{P} 0.$$

The *population zonoid covariance matrix* of F symmetric around the origin is naturally defined as the expectation of the kernel of the U-statistic, that is,

$$\mathrm{ZCM}(F) = E_F\{\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T\}$$

with distinct I, $\{i\}$ and $\{j\}$ (and a random sample X). Naturally $\mathrm{ZCM}(F)$ also is affine equivariant.

Consider next F_0 spherical around the origin and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from F_0. Write $r_i = \|\mathbf{x}_i\|$ and $\mathbf{u}_i = \|\mathbf{x}_i\|^{-1}\mathbf{x}_i$, $i = 1, \ldots, n$. Then r_i and $\mathbf{u}_i$ are independent, and the $\mathbf{u}_i$ as well as the $\mathbf{u}(I) = \|\mathbf{e}(I)\|^{-1}\mathbf{e}(I)$ are uniformly distributed on the unit sphere. As

$$E_{F_0}[\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\mathbf{x}_i \mid \mathbf{e}(I)] = E_{F_0}[|x_{i1}|] \cdot \mathbf{u}(I),$$

$$E_{F_0}[|x_{i1}|] = E_{F_0}[r_i] \cdot E_{F_0}[|u_{i1}|] = E_{F_0}[r_i]\frac{E[\chi_k]}{E[\chi_1]}$$

and $E_{F_0}[\mathbf{u}(I)\mathbf{u}^T(I)] = [1/k]I_k$, it follows that

$$\mathrm{ZCM}(F_0) = \frac{c_k^2 E_{F_0}^2(r_i)}{k} I_k$$

with

$$c_k = \frac{\Gamma(k/2)}{\Gamma((k+1)/2)\sqrt{\pi}}.$$

Note that the diagonal elements are proportional to the squared marginal mean deviations; therefore the zonoid covariance matrix can be thought of as a multivariate affine equivariant matrix valued extension of the mean deviation. Note that the regular covariance matrix is

$$\Sigma(F_0) = \frac{E_{F_0}(r_i^2)}{k} I_k.$$

A correction factor $E_{F_0}(r_i^2)/(c_k^2 E_{F_0}^2(r_i))$ is then needed for $\mathrm{ZCM}(F_0)$ to guarantee the Fisher consistency to $\Sigma(F_0)$.

3. Statistical properties of the zonoid covariance matrix in the elliptic case.

3.1. *Influence functions and efficiency in the elliptic case.* The influence function (IF) of a functional T at F measures the effect of an infinitesimal contamination located at a single point $\mathbf{z}$. We thus consider the contaminated distribution

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\mathbf{z}},$$

where $\Delta_{\mathbf{z}}$ is the cumulative distribution function of a distribution with probability mass 1 at $\mathbf{z}$. The influence function is defined as

$$\mathrm{IF}(\mathbf{z}, T, F) = \lim_{\varepsilon\downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{\partial}{\partial\varepsilon} T(F_\varepsilon)\Big|_{\varepsilon=0}.$$

The IF is a tool to describe robustness properties of an estimator, but it can also be used to compute asymptotic variance [cf. Hampel, Ronchetti, Rousseeuw and Stahel (1986) for more information on influence functions]. The influence function, limiting distribution and limiting efficiency of the Oja median have been investigated in several papers; we refer to Arcones, Chen and Giné (1994), Oja (1999) and Ollila, Hettmansperger and Oja (2002) and references therein.

Consider now the influence functions of the zonoid covariance matrix $\mathrm{ZCM}(F)$ for spherical and elliptical distributions F. In the spherical case, we have the following result. See the Appendix for the proof.

THEOREM 1. *For spherical* F_0,

$$\mathrm{IF}(\mathbf{z}; \mathrm{ZCM}, F_0) = c_k^2\big[2E_{F_0}(r)r - E_{F_0}^2(r)\big]\mathbf{u}\mathbf{u}^T - \frac{c_k^2 E_{F_0}^2(r)}{k} I_k,$$

where $r = \|\mathbf{z}\|$ *and* $\mathbf{u} = \|\mathbf{z}\|^{-1}\mathbf{z}$.

The influence function in general elliptic cases then easily follows by the affine equivariance property. Ollila, Oja and Croux (2002) derived the influence function of the scatter matrix estimate based on the SCM. If the correction factors are used, the same influence functions are obtained in these two cases and the scatter matrix estimates are asymptotically equivalent.

As $\mathrm{ZCM}(X)$ is asymptotically equivalent to a U-statistic, the limiting multinormality follows. (The assumption on finite second moments is needed here.) The limiting variances and covariances of the elements of the ZCM can be derived using the influence function presentation above. In the following, the mean of a random $k \times k$ matrix D is a $k \times k$ matrix $E(D)$ with elements $(E(D))_{ij} = E(D_{ij})$, $i, j = 1, \ldots, k$, and the covariance matrix of D is structured as a $k^2 \times k^2$ matrix $\mathrm{Cov}(D) = E(D \otimes D^T) - E(D) \otimes (E(D))^T$. $\mathrm{Cov}(D)$ then consists of k^2 $k \times k$ blocks with $\mathrm{Cov}(D_{i_1 j_1}, D_{i_2 j_2})$ for the element (j_2, i_2) of the block (i_1, j_1).

THEOREM 2. *Let* $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *be a random sample from a distribution* F_0 *spherically symmetric around* $\mathbf{0}$. *The limiting distribution of* $\sqrt{n}(\mathrm{ZCM}(X) - \mathrm{ZCM}(F_0))$ *is multivariate normal with zero mean matrix and covariance matrix*

$$c_k^4 E_{F_0}^2(r)\big[4E_{F_0}(r^2) - 3E_{F_0}^2(r)\big]E_{F_0}\big[\mathbf{u}\mathbf{u}^T \otimes \mathbf{u}\mathbf{u}^T\big] - \frac{c_k^4 E_{F_0}^4(r)}{k^2} I_{k^2},$$

where $r = \|\mathbf{z}\|$ *and* $\mathbf{u} = \|\mathbf{z}\|^{-1}\mathbf{z}$ *with* $\mathbf{z} \sim F_0$.

Again, see the Appendix for the proof.

In the spherical case, two quantities τ_1^2 and τ_2^2, namely the limiting variances of the on-diagonal and off-diagonal elements of the scatter matrix, fully characterize the limiting distribution and therefore also the limiting efficiency properties of the scatter matrix. See Croux and Haesbroeck (2000). If the correction factors are used, the ZCM and the scatter matrix estimate based on the SCM are asymptotically equivalent with the same quantities τ_1^2 and τ_2^2 and the same limiting efficiencies. In the bivariate case, the estimates coincide. In the multinormal case, the asymptotic relative efficiencies (w.r.t. the regular sample covariance matrix estimate) of the on-diagonal elements (ratios of τ_1^2) are

$$0.935, 0.960, 0.981 \text{ and } 0.994 \qquad \text{for dimensions } k = 2, 3, 5 \text{ and } 10,$$

and the same figures for off-diagonal elements (ratios of τ_2^2) are

$$0.956, 0.973, 0.987 \text{ and } 0.996 \qquad \text{for dimensions } k = 2, 3, 5 \text{ and } 10.$$

The efficiencies go to 1 as the dimension $k \to \infty$. For heavy-tailed distributions, the ZCM and the scatter matrix based on the SCM perform better than the regular covariance matrix. See Ollila, Oja and Croux (2002) for the efficiencies in the multivariate t distribution case.

In the general elliptic case, the limiting distribution of a scatter matrix is determined by the covariance matrix (of the background distribution) Σ and scalars τ_1^2 and τ_2^2, the limiting variances of the on-diagonal and off-diagonal elements in the corresponding spherical case. In most applications such as principal component analysis, canonical correlation analysis and multivariate regression analysis, the asymptotic relative efficiencies of the estimates based on different scatter matrices are simply ratios of τ_1^2 or ratios of τ_2^2 and independent of Σ. In principal component analysis, for example, the limiting variance–covariance matrices of eigenvectors and standardized eigenvalues are proportional to τ_2^2. This is used in the next section, where we consider the finite-sample efficiencies and compare them to the asymptotic relative efficiencies.

3.2. *Finite-sample efficiencies.* Ollila, Oja and Croux (2002) derived the asymptotic relative efficiencies of the sign covariance matrix in the multivariate t distribution case. Since the zonoid covariance matrix ZCM and sign covariance matrix SCM (with correction factors) have the same asymptotic efficiencies, we can now concentrate on the small sample properties of ZCM.

The finite-sample efficiencies of the zonoid covariance matrix are estimated as follows. We generated $m = 10{,}000$ samples of sizes $n = 20, 50, 100, 300$ from k-variate elliptical t distributions with $\nu = 5, 6, 8, 15, \infty$ degrees of freedom and covariance matrix $\Sigma = \operatorname{diag}(1, \ldots, k)$. The choice $\nu = \infty$ then refers to the k-variate normal distribution. Next the eigenvector and standardized eigenvalue estimates were constructed using both the zonoid covariance matrix and the regular sample covariance matrix. (The standardized eigenvalues are the regular eigenvalues divided by their geometrical means.) No correction factors are needed in these estimation problems and, as stated before, the asymptotic relative efficiencies are obtained as ratios of τ_2^2.

We consider the finite-sample efficiencies of the first eigenvector and first standardized eigenvalue estimates. The estimated mean squared error (MSE) of the first eigenvector estimate is given by

$$\operatorname{MSE}(\hat{\mathbf{v}}_1) = \frac{1}{m} \sum_{j=1}^{m} \bigl(\arccos\{|\mathbf{v}_1^T \hat{\mathbf{v}}_1^{(j)}|\}\bigr)^2,$$

where m is the number of simulated samples, $\hat{\mathbf{v}}_1^{(j)}$ is the estimate for the first eigenvector computed from the jth sample and $\mathbf{v}_1 = (0, \ldots, 0, 1)^T$ is the true

first eigenvector of Σ. The estimated MSE for the logarithm of the first standardized eigenvalue is

$$\mathrm{MSE}(\log \hat{\lambda}_1^*) = \frac{1}{m} \sum_{j=1}^{m} (\log(\hat{\lambda}_1^*)^{(j)} - \log \lambda_1^*)^2,$$

where $(\hat{\lambda}_1^*)^{(j)}$ is the estimate for the first standardized eigenvalue from the jth sample and $\lambda_1^* = k/(k!^{1/k})$ is the true first standardized eigenvalue of Σ. The estimated relative efficiencies are then the ratios of the estimated mean squared errors for the competing two estimates. The estimated efficiencies are as shown in Tables 1 and 2. The case $\{k = 4, n = 300\}$ was left out because of extremely long simulation times.

As $n \to \infty$, the ratios of the MSE converge in both considered cases to the ratios of τ_2^2, that is, to the limiting efficiencies of the off-diagonal elements in the spherical case. [See Croux, Ollila and Oja (2002) for this.] The efficiencies are very high in the multivariate normal case. For heavy-tailed distributions, the ZCM outperforms the regular covariance matrix. Note, however, that for small sample sizes, the finite-sample efficiencies tend to be much lower than the limiting efficiencies; the regular sample covariance matrix is much better than what one can expect from the asymptotic figures. The efficiencies of the estimates based on the ZCM and SCM are naturally identical in the bivariate case and also quite similar in all other considered cases.

TABLE 1

Simulated finite-sample efficiencies of the eigenvector estimates of the ZCM *relative to eigenvector estimates based on the sample covariance matrix. Samples were generated from a k-variate t distribution with ν degrees of freedom and $\Sigma = \mathrm{diag}(1, \ldots, k)$*

		Degrees of freedom (ν)				
k	***n***	**5**	**6**	**8**	**15**	∞
2	20	1.068	1.062	1.021	0.991	0.952
	50	1.289	1.216	1.141	1.014	0.942
	100	1.502	1.341	1.170	1.024	0.956
	300	1.679	1.428	1.174	1.025	0.947
	∞	2.000	1.447	1.184	1.031	0.956
3	20	1.034	1.024	1.022	1.001	0.995
	50	1.118	1.124	1.064	1.020	0.976
	100	1.280	1.221	1.140	1.040	0.976
	300	1.708	1.428	1.210	1.042	0.971
	∞	1.960	1.429	1.179	1.038	0.973
4	20	1.024	1.021	1.008	1.004	0.985
	50	1.098	1.079	1.052	1.016	0.987
	100	1.193	1.170	1.106	1.021	0.987
	∞	1.929	1.413	1.173	1.040	0.982

TABLE 2

Simulated finite-sample efficiencies of the standardized eigenvalue estimates of the ZCM *relative to standardized eigenvalue estimates based on the sample covariance matrix. Samples were generated from a k-variate t distribution with ν degrees of freedom and $\Sigma = \mathrm{diag}(1, \ldots, k)$*

		Degrees of freedom (ν)				
k	n	**5**	**6**	**8**	**15**	∞
2	20	1.157	1.110	1.051	0.992	0.947
	50	1.259	1.187	1.084	1.003	0.955
	100	1.373	1.242	1.119	1.025	0.953
	300	1.540	1.323	1.162	1.031	0.953
	∞	2.000	1.447	1.184	1.031	0.956
3	20	1.192	1.142	1.072	1.010	0.960
	50	1.398	1.255	1.121	1.025	0.974
	100	1.483	1.303	1.133	1.021	0.970
	300	1.584	1.343	1.168	1.033	0.975
	∞	1.960	1.429	1.179	1.038	0.973
4	20	1.174	1.120	1.061	1.009	0.965
	50	1.431	1.273	1.146	1.027	0.969
	100	1.633	1.370	1.174	1.041	0.979
	∞	1.929	1.413	1.173	1.040	0.982

4. Zonotopes, zonoids and scatter matrices.

4.1. *Data based zonotope $Z(Y)$ and its volume.* In this section we discuss the concepts of the zonoid and zonotope and show how the ZCM and SCM are related to these. Some new notation and definitions are needed first. For k-variate sets $K, K_1, K_2 \subset \mathbb{R}^k$, write

$$c \cdot K = \{c\mathbf{k} : \mathbf{k} \in K\} \quad \text{and} \quad K_1 + K_2 = \{\mathbf{k}_1 + \mathbf{k}_2 : \mathbf{k}_1 \in K_1, \mathbf{k}_2 \in K_2\}.$$

The set $K_1 + K_2$ is called the *Minkowski sum* of sets K_1 and K_2.

Let $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a random sample from a k-variate distribution with c.d.f. F symmetric around the origin. In the following constructions we use data based line segments

$$[-\mathbf{y}_i, \mathbf{y}_i] = \{\alpha\mathbf{y}_i - (1-\alpha)\mathbf{y}_i : \alpha \in [0, 1]\} \subset \mathbb{R}^k.$$

By definition, *zonotopes* are finite Minkowski sums of line segments. If the number of segments is $r \le k$, an r-variate *parallelotope* is obtained. Then k data based segments yield parallelotopes

$$[-\mathbf{y}_{i_1}, \mathbf{y}_{i_1}] + \cdots + [-\mathbf{y}_{i_k}, \mathbf{y}_{i_k}]$$

and

$$Z(Y) = \frac{1}{n}\sum_{i=1}^{n}\{[-\mathbf{y}_i, \mathbf{y}_i]\}$$

is the zonotope based on the complete data set. (A multiplicative factor $1/n$ is used to make it converge to a limit as sample size $n \to \infty$.)

Note that $Z(Y)$ is symmetrically located around the origin. This construction is natural as we assume that the observations come from a symmetrical distribution. It is *affine equivariant* in the sense that if

$$A \cdot Y = \{A\mathbf{y}_1, \ldots, A\mathbf{y}_n\}$$

with a positive definite $k \times k$ matrix A is a transformed data set, then the zonotope of the transformed data is

$$Z(A \cdot Y) = A \cdot Z(Y).$$

The affine equivariance property then means that zonotope $Z(Y)$ carries information on the shape and geometry of the multivariate data cloud Y. See Figures 2 and 3 for illustrations of $Z(Y)$ in the two-variate and three-variate cases.

The volume of the k-variate parallelotope is

$$\operatorname{vol}\big([-\mathbf{y}_{i_1}, \mathbf{y}_{i_1}] + \cdots + [-\mathbf{y}_{i_k}, \mathbf{y}_{i_k}]\big) = 2^k \operatorname{abs}\big(\det(\mathbf{y}_{i_1} \ldots \mathbf{y}_{i_k})\big)$$

and the volume of zonotope $Z(Y)$ is given by

$$\operatorname{vol}(Z(Y)) = \operatorname{vol}\left(\frac{1}{n}\sum_{i=1}^{n}[-\mathbf{y}_i, \mathbf{y}_i]\right) = \frac{2^k}{n^k}\sum\big\{\operatorname{abs}\big(\det(\mathbf{y}_{i_1} \ldots \mathbf{y}_{i_k})\big)\big\},$$

where the last sum is over all k-tuples $1 \le i_1 < \cdots < i_k \le n$. Note that $\operatorname{vol}(Z(Y))$ is a scalar valued multivariate extension of the *mean deviation*. Recall that if $Y = \{\mathbf{x}_1 - \boldsymbol{\mu}, \ldots, \mathbf{x}_n - \boldsymbol{\mu}\}$, the Oja median $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(X)$ is the choice to minimize $\operatorname{vol}(Z(Y))$ and therefore a natural location estimate here.

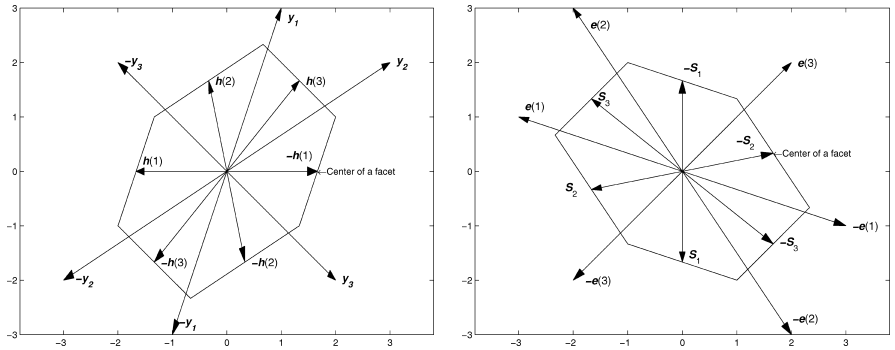

FIG. 2. *A small two-variate data set* $Y = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$ *and the zonotope* $Z(Y)$ *with centers of facets* $\pm\mathbf{h}(1), \pm\mathbf{h}(2), \pm\mathbf{h}(3)$. *The set of normals* $E = \{\mathbf{e}(1), \mathbf{e}(2), \mathbf{e}(3)\}$ *and the zonotope* $Z(E)$ *with centers of facets* $\pm\mathbf{S}_1, \pm\mathbf{S}_2, \pm\mathbf{S}_3$.

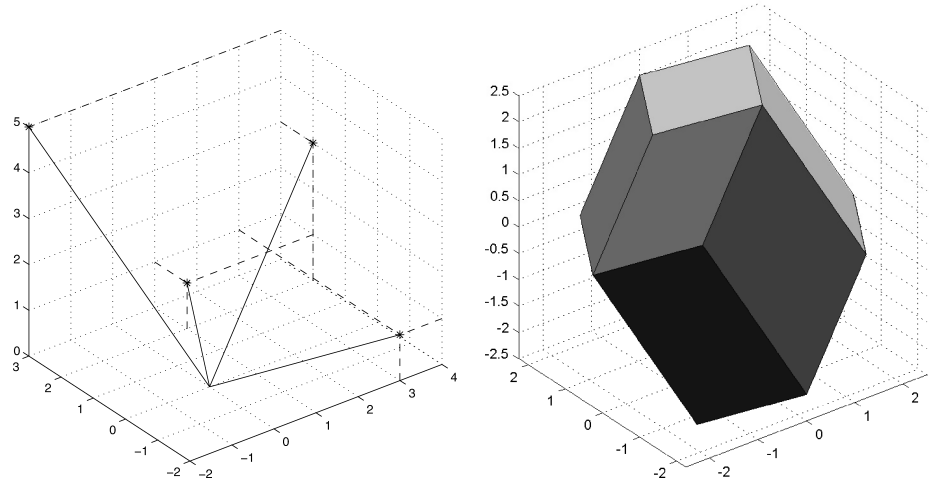

FIG. 3. *A small three-variate data set Y* (a) *and the related zonotope* $Z(Y)$ (b).

4.2. *Support function and facets of zonotope* $Z(Y)$. Next note that any closed convex set $K \subset \mathbb{R}^k$ is uniquely defined by its *support function*

$$\mathbf{p} \to \psi(K, \mathbf{p}) = \sup\{\mathbf{p}^T\mathbf{y} : \mathbf{y} \in K\}, \qquad \mathbf{p} \in \mathbb{R}^k,$$

or by the restriction of the support function to the sphere. As

$$\psi([-\mathbf{y}, \mathbf{y}], \mathbf{p}) = \operatorname{sign}(\mathbf{p}^T\mathbf{y})\mathbf{p}^T\mathbf{y},$$

the support function of $Z(Y)$ is given by

$$\psi\big(Z(Y), \mathbf{p}\big) = \frac{1}{n}\sum_{i=1}^{n}\{\operatorname{sign}(\mathbf{p}^T\mathbf{y}_i)\mathbf{p}^T\mathbf{y}_i\}.$$

If $\mathbf{p}$ is a unit vector, that is, $\mathbf{p}^T\mathbf{p} = 1$, then $\psi(Z(Y), \mathbf{p}) = \operatorname{ave}_i|\mathbf{p}^T\mathbf{y}_i|$ is the marginal mean deviation in the direction given by $\mathbf{p}$. The sets

$$\{\mathbf{y} \in \mathbb{R}^k : \mathbf{p}^T\mathbf{y} = \psi(K, \mathbf{p})\}, \qquad \mathbf{p} \in \mathbb{R}^k$$

are called the *faces* of K. Moreover, the $(k-1)$-dimensional faces are called the *facets* of K.

Using these concepts, it is easy to see (step by step) that the following hold:

1. The support function of $Z(Y) = \operatorname{ave}_i\{[-\mathbf{y}_i, \mathbf{y}_i]\}$ at $\mathbf{e}(I)$ equals
$$\psi\big(Z(Y), \mathbf{e}(I)\big) = \mathbf{h}^T(I)\mathbf{e}(I).$$
2. $Z(Y)$ has $2\binom{n}{k-1}$ facets; the facets are translates of $(k-1)$-variate parallelotopes
$$\frac{1}{n}\{[-\mathbf{y}_{i_1}, \mathbf{y}_{i_1}] + \cdots + [-\mathbf{y}_{i_{k-1}}, \mathbf{y}_{i_{k-1}}]\}.$$

The $\mathbf{h}(I)$ and $-\mathbf{h}(I)$ are the centers of these facets.

3. The zonoid covariance matrix ZCM is the covariance matrix calculated from the centers of facets of the zonoid $Z(Y)$.

Symmetrically consider also the zonotope

$$Z(E) = \text{ave}_I\{[-\mathbf{e}(I), \mathbf{e}(I)]\}.$$

Then similarly the following hold:

1. The support function of $Z(E)$ at $\mathbf{y}_i$ equals
$$\psi\big(Z(E), \mathbf{y}_i\big) = \mathbf{S}_i^T \mathbf{y}_i.$$
2. The zonotope $Z(E)$ has a huge number of facets. The $\mathbf{S}_i$ and $-\mathbf{S}_i$ are the centers of some of the facets of $Z(E)$.
3. The sign covariance matrix SCM is the covariance matrix calculated from selected centers of facets of the zonoid $Z(E)$.

The symmetry of the data vectors, their signs, normals and the signs of normals is also seen in that $\text{vol}(Z(Y))$ is proportional to

$$\text{ave}_{i,I}\big\{\text{sign}\big(\mathbf{e}^T(I)\mathbf{y}_i\big)\mathbf{e}^T(I)\mathbf{y}_i\big\} = \text{ave}_I\big\{\mathbf{h}^T(I)\mathbf{e}(I)\big\} = \text{ave}_i\big\{\mathbf{S}_i^T\mathbf{y}_i\big\}.$$

See Figure 2 for an illustration of this symmetry.

4.3. *Zonoid of the distribution F, $Z(F)$.* Expected value of the random convex set $K \in \mathbb{R}^k$, denoted by $E(K)$, may be defined through support functions; the support function of set $E(K)$ is

$$\psi\big(E(K), \mathbf{p}\big) = E\big(\psi(K, \mathbf{p})\big), \qquad \mathbf{p} \in \mathbb{R}^k.$$

The theoretical counterpart of the zonotope is then the *associated zonoid of F*, denoted by $Z(F)$, which is

$$Z(F) = E_F\{[-\mathbf{y}, \mathbf{y}]\},$$

where the c.d.f. of $\mathbf{y}$ is F. We emphasize that our definition of the zonoid (of the reflected distribution around the origin) is natural for symmetrical distributions; for an alternative definition and uses of the zonoid, see Mosler (2002) and references therein.

Zonoid $Z(F)$ is symmetrically located around $\mathbf{0}$. The multivariate (scalar valued) mean deviation,

$$\text{vol}(Z(F)) = 2^k E_F\big\{\text{abs}\big(\det(\mathbf{y}_1, \ldots, \mathbf{y}_k)\big)\big\},$$

is the expected volume of the random parallelotope. See Koshevoy and Mosler (1998) and Mosler (2002).

Let the distribution F_0 of $\mathbf{z}$ be spherically symmetric around the origin with covariance matrix I_k; thus the radius $r = \|\mathbf{z}\|$ and direction vector $\mathbf{u} = \|\mathbf{z}\|^{-1}\mathbf{z}$ are independent and $\mathbf{u}$ is uniformly distributed on the periphery of a unit sphere. The

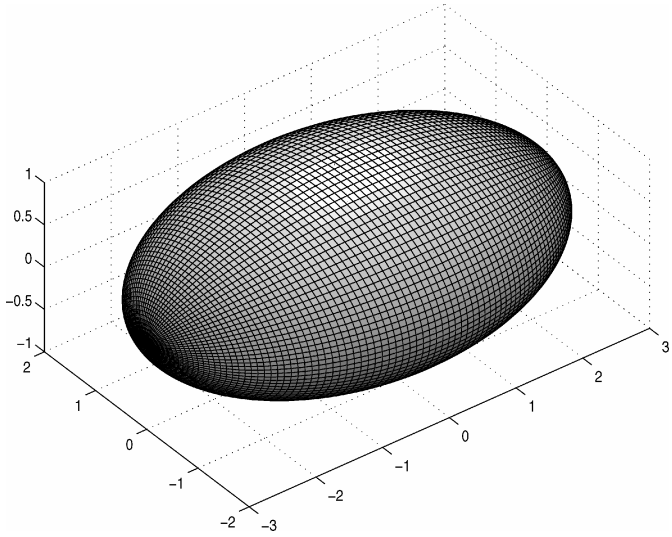

FIG. 4. *A zonoid $Z(F)$ for an elliptic distribution in the three-variate case.*

zonoid $Z(F_0)$ is then a sphere with radius $c_k E_{F_0}(r)$, which is also the marginal mean deviation. The zonoid covariance matrix $\mathrm{ZCM}(F_0) = [c_k^2 E_{F_0}^2(r)/k]I_k$ is the covariance matrix of $c_k E_{F_0}(r)\mathbf{u}$ which is uniformly distributed on the boundary of the zonoid $Z(F_0)$. See Section 2.3.

Consider next elliptic $\mathbf{y} = \Sigma^{1/2}\mathbf{z}$ with c.d.f. F. By the affine equivariance property, $Z(F) = \Sigma^{1/2} \cdot Z(F_0)$ and therefore

$$Z(F) = \{\mathbf{y} \in \mathbb{R}^k : \mathbf{y}^T \Sigma^{-1} \mathbf{y} \le c_k^2 E_{F_0}^2(r)\}$$

is an ellipsoid with shape determined by Σ. See Figure 4 for an illustration of a zonoid $Z(F)$ for an elliptic distribution in the three-variate case. Again, $\mathrm{ZCM}(F)$ is the regular covariance matrix of random variable $\Sigma^{1/2} c_k E_{F_0}(r)\mathbf{u}$ with a distribution concentrated on the boundary of $Z(F)$.

Finally, consider the zonoid $E_F\{[-\mathbf{e}(I), \mathbf{e}(I)]\}$ which appears to be the projection body of $Z(F)$. For the definition of the projection body and for its properties, see, for example, Gardner [(1995), Chapter 4]. The projection body of the ellipsoid $Z(F)$ is the ellipsoid

$$\{\mathbf{y} \in \mathbb{R}^k : \mathbf{y}^T \Sigma \mathbf{y} \le \{\det(\Sigma)\}^2 c_{F_0}^2\},$$

where now

$$c_{F_0} = \frac{\Gamma^k(k/2) E_{F_0}^{k-1}(r)}{\sqrt{\pi}\Gamma^{k-1}((k+1)/2)}.$$

Compare Ollila, Oja and Croux (2002). In the elliptic case, the shape of the projection body is thus given by Σ^{-1}. The sign covariance matrix, $\mathrm{SCM}(F)$, is the

regular covariance matrix of random variable $\det(\Sigma)\Sigma^{-1/2}c_{F_0}\mathbf{u}$ with a distribution concentrated on the boundary of the projection body of $Z(F)$.

We finally mention (the proof will be left to a forthcoming paper) that, under general assumptions, the following hold: (a) the zonoid covariance matrix $\mathrm{ZCM}(F)$ is the regular covariance matrix of a distribution concentrated on the boundary of the zonoid $Z(F)$; (b) the sign covariance matrix $\mathrm{SCM}(F)$ is the regular covariance matrix of a distribution concentrated on the boundary of the zonoid which is the projection body of $Z(F)$.

5. Final comments. In this paper, we introduced a new, highly efficient (under normality) and fairly robust scatter matrix estimate ZCM based on the observed zonotope. The estimate is an affine equivariant multivariate extension of the mean deviation and a natural competitor of the regular covariance matrix. The estimate is closely related to the affine equivariant sign covariance matrix and can be constructed using the data vectors and their angular distances. For high dimensions and large sample sizes, the computation of the ZCM is heavy; the computational load is about the same as in the affine equivariant sign covariance matrix case.

In the location-scale model, a correction factor is needed to compare different scatter matrix estimates. The functional

$$V(F) = \left[\frac{k}{\mathrm{Tr}(C(F))}\right] C(F)$$

with standardized eigenvalues, $\mathrm{Tr}(V) = k$, is the *shape matrix* related to the scatter matrix $C(F)$. In the location-scale model, the shape matrices are directly comparable without any modifications. Note that in several applications, such as principal component analysis (PCA), canonical correlation analysis (CCA) or multivariate multiple regression, the test and estimation procedures may be based on the shape matrix only. See Ollila, Oja and Croux (2002) and Ollila, Hettmansperger and Oja (2002) for applications.

As found in Section 4, the scatter matrix estimate may thus be constructed using the zonotope based on the centered data set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. The *lift zonotope* based on k-variate uncentered data set X is the $(k+1)$-variate convex body

$$\mathrm{LZ}(X) = \frac{1}{n}\sum_i \{[\mathbf{0}, (1, \mathbf{x}_i^T)^T]\}.$$

If X is a random sample from F, the corresponding population lift zonoid is

$$\mathrm{LZ}(F) = E_F\{[\mathbf{0}, (1, \mathbf{x}_i^T)^T]\}.$$

Unlike the zonoid, the lift zonoid fully characterizes the distribution F. In our forthcoming paper, the scatter matrix estimate based on the lift zonotope is constructed; this approach is again symmetrically related to the approach based on affine equivariant multivariate ranks. See Oja (1999) and Visuri, Ollila, Koivunen, Möttönen and Oja (2003).

6. Proofs of the results.

6.1. *Auxiliary notation and results.* For observations centered with $\boldsymbol{\mu}$, write

$$\det\bigl(\mathbf{x}_{i_1}-\boldsymbol{\mu},\ldots,\mathbf{x}_{i_{k-1}}-\boldsymbol{\mu},\mathbf{x}-\boldsymbol{\mu}\bigr)=\bigl(\mathbf{e}(I)+G(I)\boldsymbol{\mu}\bigr)^T(\mathbf{x}-\boldsymbol{\mu}),$$

where

$$\begin{aligned}(G(I))_{ij}=\bigl[&\det\bigl(-\mathbf{1}_j\ \mathbf{x}_{i_2}\cdots\mathbf{x}_{i_{k-1}}\ \mathbf{1}_i\bigr)+\det\bigl(\mathbf{x}_{i_1}\ -\mathbf{1}_j\cdots\mathbf{x}_{i_{k-1}}\ \mathbf{1}_i\bigr)+\cdots\\&+\det\bigl(\mathbf{x}_{i_1}\ \mathbf{x}_{i_2}\cdots-\mathbf{1}_j\ \mathbf{1}_i\bigr)\bigr]\end{aligned}$$

and $\mathbf{1}_j$ is a k-vector with the jth element 1 and the other elements 0. Then write accordingly

$$G(I)=G_1(I)+\cdots+G_{k-1}(I),$$

where $G_j(I)$ does not depend on $\mathbf{x}_{i_j}$.

Next consider fixed k-vector $\mathbf{e}$ and fixed $k\times k$ matrix G. Then

$$\begin{aligned}&E\bigl\{\operatorname{sign}\bigl((\mathbf{e}+G\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\bigr)(\mathbf{x}-\boldsymbol{\mu})\bigr\}\\&\quad=\nabla_{\mathbf{e}}E\bigl\{\operatorname{abs}\bigl(\mathbf{e}^T\mathbf{x}-\mathbf{e}^T\boldsymbol{\mu}+\boldsymbol{\mu}^TG^T\mathbf{x}-\boldsymbol{\mu}^TG^T\boldsymbol{\mu}\bigr)\bigr\}\end{aligned}$$

and, applying $\nabla_{\boldsymbol{\mu}}$ to both sides, at $\boldsymbol{\mu}=\mathbf{0}$, one obtains

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}}E\bigl\{\operatorname{sign}\bigl((\mathbf{e}+G\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\bigr)(\mathbf{x}-\boldsymbol{\mu})\bigr\}\\&\quad=\nabla_{\mathbf{e}}E\bigl\{\operatorname{sign}\bigl(\mathbf{e}^T\mathbf{x}\bigr)\bigl(-\mathbf{e}+G^T\mathbf{x}\bigr)\bigr\}\\&\quad=G^TD(\mathbf{e}),\end{aligned}$$

where $D(\mathbf{e})$ is even in $\mathbf{e}$.

Now we are ready to prove the following lemma.

LEMMA 5. *Assume that $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ is a random sample from a distribution symmetric around the origin. Let $I=(i_1,\ldots,i_{k-1})$ and $i\notin I$. At $\boldsymbol{\mu}=0$,*

$$\nabla_{\boldsymbol{\mu}}E\bigl\{\operatorname{sign}\bigl(\bigl(\mathbf{e}(I)+G(I)\boldsymbol{\mu}\bigr)^T(\mathbf{x}_i-\boldsymbol{\mu})\bigr)(\mathbf{x}_i-\boldsymbol{\mu})h_l(I)\bigr\}=0,\qquad l=1,\ldots,k.$$

PROOF. The proof follows from

$$E\bigl\{G^T(I)D(\mathbf{e}(I))h_l(I)\bigr\}=\sum_{j=1}^{k-1}E\bigl\{G_j^T(I)D_j(\mathbf{e}(I))h_l(I)\bigr\}=0$$

as $G_j(I)$ does not depend on $\mathbf{x}_{i_j}$, $D_j(\mathbf{e}(I))$ is even in $\mathbf{x}_{i_j}$ and $h(I)$ is odd in $\mathbf{x}_{i_j}$. □

PROOF OF LEMMA 1. Suppose first that $E_F(\mathbf{y}_i) = 0$ and $\mathrm{Cov}_F(\mathbf{y}_i) = I_k$. Then, for example,

$$e_1(I) = \sum \{\pm y_{i_1 j_2} \cdot \cdots \cdot y_{i_{k-1} j_k}\},$$

where the sum goes over all $(k-1)!$ permutations $(j_2, \ldots, j_k)$ of $(2, \ldots, k)$. As $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_{k-1}}$ are independent, it follows that $E_F(e_1(I)) = 0$, $E_F(e_1^2(I)) = (k-1)!$ and $E_F(e_1(I)e_2(I)) = 0$. Thus $E_F(\mathbf{e}(I)) = \mathbf{0}$ and $\mathrm{Cov}_F(\mathbf{e}(I)) = (k-1)!I_k$. The result then follows from the affine equivariance. □

PROOF OF LEMMA 4. We write $\mathrm{ZCM}(X, \boldsymbol{\mu})$ for the zonoid covariance matrix based on the data set centered by $\boldsymbol{\mu}$. If the centering is by the Oja median, the regular zonoid covariance matrix is obtained. We assume here that the Oja median is $\sqrt{n}$-consistent with $\boldsymbol{\mu}$. [This is true under general assumptions; see Arcones, Chen and Giné (1994).] As mentioned in the Introduction, we also assume the second moments are finite. The proof has two parts. First we show that $\mathrm{ZCM}(X, \mathbf{0})$ and U_n are asymptotically equivalent, that is, $\sqrt{n}(\mathrm{ZCM}(X, \mathbf{0}) - U_n) \to_P \mathbf{0}$. Second, one has to show that $\sqrt{n}(\mathrm{ZCM}(X, \mathbf{0}) - \mathrm{ZCM}(X, \hat{\boldsymbol{\mu}})) \to_P \mathbf{0}$, which means that the estimate can be replaced by the true value in asymptotical considerations.

The first part follows easily from

$$\begin{aligned}
\mathrm{ave}_I\{\mathbf{h}(I)\mathbf{h}^T(I)\} &= \frac{1}{\binom{n}{k-1}n^2} \sum_I \sum_i \sum_j \{\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\mathrm{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T\} \\
&= \frac{\binom{n}{k+1}}{\binom{n}{k-1}n^2} U_n + \frac{\binom{n-1}{k-1}}{\binom{n}{k-1}n^2} \sum_i \mathbf{x}_i\mathbf{x}_i^T \\
&= U_n + o_P\left(\frac{1}{n}\right).
\end{aligned}$$

For fixed $\boldsymbol{\mu}$, the statistic $\mathrm{ZCM}(X, \mathbf{0}) - \mathrm{ZCM}(X, \boldsymbol{\mu})$ is similarly asymptotically equivalent to a U-statistic. Its expected value and variances and covariances are continuous in $\boldsymbol{\mu}$ and the variances are $O(1/n)$ uniformly in a neighborhood of the origin. This implies that the variance of $\sqrt{n}(\mathrm{ZCM}(X, \mathbf{0}) - \mathrm{ZCM}(X, \hat{\boldsymbol{\mu}}))$ goes to zero with n. Lemma 5 and the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\mu}}$ together imply that also

$$\sqrt{n}E\big[\mathrm{ZCM}(X, \mathbf{0}) - \mathrm{ZCM}(X, \hat{\boldsymbol{\mu}})\big] \to \mathbf{0}$$

and the result follows. □

PROOF OF THEOREM 1. The functional $\mathrm{ZCM}(F)$ is the expectation of the kernel $g(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k+1}})$ of U-statistic U_n given in Lemma 4. The influence function of the U-statistic at contaminated value $\mathbf{z}$ is then

$$(k+1)\big[E\big(g(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k+1}})|\mathbf{x}_{i_1} = \mathbf{z}\big) - E\big(g(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k+1}})\big)\big]$$

and therefore the conditional expectations $E(g(\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_k})|\mathbf{x}_{i_j})$ are needed. For observation $\mathbf{x}_i$, write $r_i = \|\mathbf{x}_i\|$ and $\mathbf{u}_i = \|\mathbf{x}_i\|^{-1}\mathbf{x}_i$, $i = 1,\dots,n$. For normal $I = (i_1,\dots,i_{k-1})$, write $\mathbf{u}(I) = \|\mathbf{e}(I)\|^{-1}\mathbf{e}(I)$.

For conditional expectations we first need

$$E_F\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T|\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_{k-1}}\} = c_k^2E^2(r)\mathbf{u}(I)\mathbf{u}^T(I),$$

which gives

$$E_F\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T|\mathbf{x}_{i_1}\} = \frac{c_k^2E^2(r)}{k-1}(I_k - \mathbf{u}_{i_1}\mathbf{u}_{i_1}^T).$$

Next note that

$$\begin{aligned}&E_F\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T|\mathbf{x}_i,\mathbf{x}_{i_1},\dots,\mathbf{x}_{i_{k-1}}\}\\ &\quad= c_kE(r)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\mathbf{x}_i\mathbf{u}^T(I)\end{aligned}$$

and consequently

$$E_F\{\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_i)\operatorname{sign}(\mathbf{e}^T(I)\mathbf{x}_j)\mathbf{x}_i\mathbf{x}_j^T|\mathbf{x}_i\} = c_k^2E(r)r_i\mathbf{u}_i\mathbf{u}_i^T.$$

The influence function of the functional $\mathrm{ZCM}(F)$ for $\mathbf{z} = r\mathbf{u}$ is then

$$c_k^2\left[(k-1)\frac{E^2(r)}{k-1}(I_k - \mathbf{u}\mathbf{u}^T) + 2rE(r)\mathbf{u}\mathbf{u}^T\right] - (k+1)\frac{c_k^2E^2(r)}{k}I_k$$

and the result follows. □

PROOF OF THEOREM 2. The assumptions are as in the proof of Lemma 4. All the (nonconstant) linear combinations of the elements of matrix $\mathrm{ZCM}(X)$ are asymptotically equivalent to scalar valued U-statistics with kernels $g_l(\mathbf{x}_i,\dots,\mathbf{x}_{k+1})$ such that $\mathrm{Var}(g_l(\mathbf{x}_i,\dots,\mathbf{x}_{k+1})|\mathbf{x}_1)$ are bounded and positive. All linear combinations are then asymptotically normal, which implies that $\sqrt{n}(\mathrm{ZCM}(X) - \mathrm{ZCM}(F_0))$ also is asymptotically multinormal. The limiting variances are then easily derived using the influence function. □

Acknowledgments. The authors thank two anonymous referees for very careful reading and for helpful suggestions which greatly improved the presentation.

REFERENCES

ARCONES, M. A., CHEN, Z. and GINÉ, E. (1994). Estimators related to U-processes with applications to multivariate medians: Asymptotic normality. *Ann. Statist.* **22** 1460–1477.

CROUX, C. and HAESBROECK, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika* **87** 603–618.

CROUX, C., OLLILA, E. and OJA, H. (2002). Sign and rank covariance matrices: Statistical properties and application to principal component analysis. In *Statistical Data Analysis Based on the L_1 Norm and Related Methods* (Y. Dodge, ed.) 257–270. Birkhäuser, Basel.

GARDNER, R. J. (1995). *Geometric Tomography.* Cambridge Univ. Press.

HALLIN, M. and PAINDAVEINE, D. (2002). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Ann. Statist.* **30** 1103–1133.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* Wiley, New York.

KOSHEVOY, G. and MOSLER, K. (1997a). Multivariate Gini indices. *J. Multivariate Anal.* **60** 252–276.

KOSHEVOY, G. and MOSLER, K. (1997b). Zonoid trimming for multivariate distributions. *Ann. Statist.* **25** 1998–2017.

KOSHEVOY, G. and MOSLER, K. (1998). Lift zonoids, random convex hulls and the variability of random vectors. *Bernoulli* **4** 377–399.

MOSLER, K. (2002). *Multivariate Dispersion*, *Central Regions*, *and Depth*: *The Lift Zonoid Approach*. *Lecture Notes in Statist.* **165**. Springer, New York.

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332.

OJA, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scand. J. Statist.* **26** 319–343.

OLLILA, E., HETTMANSPERGER, T. P. and OJA, H. (2002). Affine equivariant multivariate sign methods. Unpublished manuscript.

OLLILA, E., OJA, H. and CROUX, C. (2002). The affine equivariant sign covariance matrix: Asymptotic behavior and efficiency. Unpublished manuscript.

OLLILA, E., OJA, H. and HETTMANSPERGER, T. P. (2002). Estimates of regression coefficients based on the sign covariance matrix. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 447–466.

PETERS, D. and RANDLES, R. H. (1990). A multivariate signed-rank test for the one-sample location problem. *J. Amer. Statist. Assoc.* **85** 552–557.

RANDLES, R. H. (1989). A distribution-free multivariate sign test based on interdirections. *J. Amer. Statist. Assoc.* **84** 1045–1050.

VISURI, S., KOIVUNEN, V. and OJA, H. (2000). Sign and rank covariance matrices. *J. Statist. Plann. Inference* **91** 557–575.

VISURI, S., OLLILA, E., KOIVUNEN, V., MÖTTÖNEN, J. and OJA, H. (2003). Affine equivariant multivariate rank methods. *J. Statist. Plann. Inference* **114** 161–185.

G. A. KOSHEVOY
CENTRAL INSTITUTE OF ECONOMICS
AND MATHEMATICS
RUSSIAN ACADEMY OF SCIENCES
NAKHIMOVSKII 47
MOSCOW 117418
RUSSIA
E-MAIL: koshevoy@cemi.rssi.ru

J. MÖTTÖNEN
DEPARTMENT OF MATHEMATICAL SCIENCES/
STATISTICS
UNIVERSITY OF OULU
P.O. BOX 3000
FIN-90014 UNIVERSITY OF OULU
FINLAND
E-MAIL: jyrki.mottonen@oulu.fi

H. OJA
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF JYVÄSKYLÄ
P.O. BOX 35
FIN-40351 JYVÄSKYLÄ
FINLAND
E-MAIL: ojahannu@cc.jyu.fi