Austrian Journal of Statistics
April 2017, Volume 46, 13–22.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.667

The Spatial Sign Covariance Matrix and Its Application for Robust Correlation Estimation

Alexander Dürre
TU Dortmund

Roland Fried
TU Dortmund

Daniel Vogel
University of Aberdeen

Abstract

We summarize properties of the spatial sign covariance matrix and especially consider the relationship between its eigenvalues and those of the shape matrix of an elliptical distribution. The explicit relationship known in the bivariate case was used to construct the spatial sign correlation coefficient, which is a non-parametric and robust estimator for the correlation coefficient within the elliptical model. We consider a multivariate generalization, which we call the multivariate spatial sign correlation matrix. A small simulation study indicates that the new estimator is very efficient under various elliptical distributions if the dimension is large. We furthermore derive its influence function under certain conditions which indicates that the multivariate spatial sign correlation becomes more sensitive to outliers as the dimension increases.

Keywords: elliptical distribution, influence function, eigenvalues, fixed-point algorithm.

1. Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote a sample of independent p dimensional random variables from a distribution F and $s : \mathbb{R}^p \to \mathbb{R}^p$ with $s(\boldsymbol{x}) = \boldsymbol{x}/|\boldsymbol{x}|$ for $\boldsymbol{x} \neq 0$ and $s(0) = 0$ the spatial sign, then

$$S_n(\boldsymbol{t}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{X}_i - \boldsymbol{t}_n) s(\boldsymbol{X}_i - \boldsymbol{t}_n)^T$$

denotes the empirical spatial sign covariance matrix (SSCM) with location $\boldsymbol{t}_n$. The canonical choice for the location estimator $\boldsymbol{t}_n$ is the spatial median

$$\boldsymbol{\mu}_n = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} ||\boldsymbol{X}_i - \boldsymbol{\mu}||.$$

Besides its nice robustness properties like an asymptotic breakdown-point of 1/2, the spatial median has (under regularity conditions, see Kemperman 1987) the advantageous feature that it centres the spatial signs, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{X}_i - \boldsymbol{\mu}_n) = 0,$$

so that $S_n(\boldsymbol{\mu}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is indeed the empirical covariance matrix of the spatial signs of the data. If $\boldsymbol{t}_n$ is (strongly) consistent for a location $\boldsymbol{t} \in \mathbb{R}$, it was shown in Dürre, Vogel, and Tyler (2014) that under mild conditions on F the empirical SSCM is a (strongly) consistent estimator for its population counterpart $S(\boldsymbol{X}) = \mathsf{E}(s(\boldsymbol{X} - \boldsymbol{t})s(\boldsymbol{X} - \boldsymbol{t})^T)$, for $\boldsymbol{X} \sim F$. Results about $S(\boldsymbol{X})$ have been derived for continuous elliptical distributions F, i.e. if F possesses a density of the form

$$f(\boldsymbol{x}) = \det(V)^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))$$

for a location $\boldsymbol{\mu} \in \mathbb{R}^p$, a symmetric and positive definite shape matrix $V \in \mathbb{R}^{p \times p}$ and a function $g : [0, \infty) \to [0, \infty)$, which is often called the elliptical generator. Prominent members of the elliptical family are the multivariate normal distribution and elliptical t-distributions (e.g. Bilodeau and Brenner 1999, p. 208). If second moments exist, then $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{X} \sim F$, and V a multiple of its covariance matrix. The shape matrix V is unique only up to a multiplicative constant. In the following, we consider the trace-normalized shape matrix $V_0 = V/\mathrm{tr}(V)$, which is convenient since $S(\boldsymbol{X})$ also has trace 1. If F is elliptical, then $S(\boldsymbol{X})$ and V share the same eigenvectors and the respective eigenvalues have the same ordering. For this reason, the SSCM has been proposed for robust principal component analysis (e.g. Locantore, Marron, Simpson, Tripoli, Zhang, and Cohen 1999; Marden 1999). In the present article, we study the eigenvalues of the SSCM.

In the following we discuss properties of the SSCM and extend it to correlation estimation. In Section 2 we summarize results about the eigenvalues of the SSCM and illustrate by means of two examples how the eigenvalues of the SSCM are connected with those of the shape matrix. In Section 3 a new estimator for the correlation matrix based on the SSCM is introduced, which we call the multivariate spatial sign correlation matrix. We describe a fixed-point algorithm to calculate the estimator numerically. Furthermore we investigate the efficiency of the spatial sign correlation matrix in a small simulation under different elliptical distributions and derive its influence function under specific assumptions.

2. Eigenvalues of the SSCM

Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of V_0 and $\delta_1 \geq \ldots \geq \delta_p \geq 0$ those of $S(\boldsymbol{X})$. Explicit formulae that relate the δ_i to the λ_i are only known for $p = 2$ (see Vogel, Köllmann, and Fried 2008; Croux, Dehon, and Yadine 2010), namely

$$\delta_i = \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}, \quad i = 1, 2. \tag{1}$$

Assuming $\lambda_2 > 0$, we have $\delta_1/\delta_2 = \sqrt{\lambda_1/\lambda_2} \leq \lambda_1/\lambda_2$, thus the eigenvalues of the SSCM are closer together than those of the corresponding shape matrix. It is shown in Dürre, Tyler, and Vogel (2016) that this holds true for arbitrary $p > 2$,

$$\lambda_i/\lambda_j \geq \delta_i/\delta_j \quad \text{for } 1 \leq i < j \leq p \tag{2}$$

as long as $\lambda_j > 0$. There is no explicit map between the eigenvalues known for $p > 2$. Dürre *et al.* (2016) give a representation of δ_i as one-dimensional integral, which permits fast and accurate numerical evaluations for arbitrary p,

$$\delta_i = \frac{\lambda_i}{2} \int_0^\infty \frac{1}{(1 + \lambda_i x) \prod_{j=1}^p (1 + \lambda_j x)^{\frac{1}{2}}} dx, \quad i = 1, \ldots, p. \tag{3}$$

We use this formula, which is implemented in R (R Core Team 2016) in the package sscor (Dürre and Vogel 2016b), to get an impression how the eigenvalues of $S(\boldsymbol{X})$ look like in comparison to those of V_0. We first look at equidistantly spaced eigenvalues

$$\lambda_i = \frac{2(p + 1 - i)}{p(p + 1)}, \quad i = 1, \ldots, p,$$

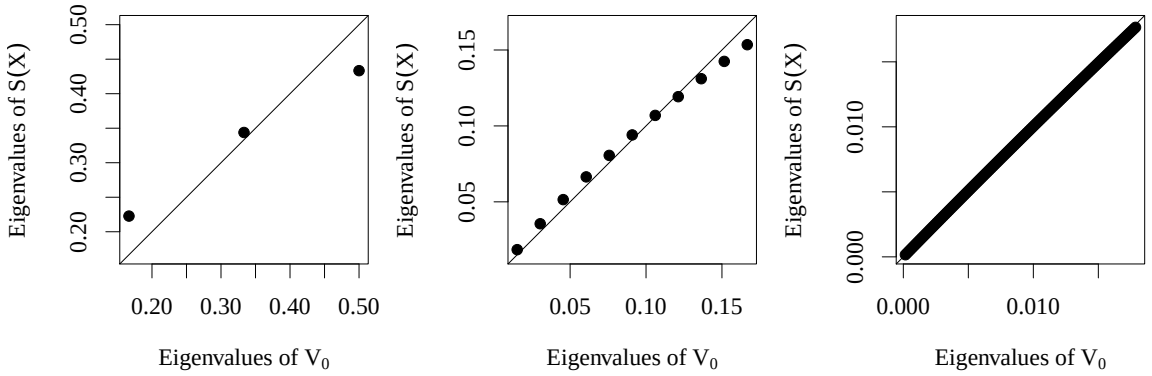

Figure 1: Eigenvalues of the SSCM w.r.t. the corresponding eigenvalues of the shape matrix in the equidistant setting $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

for different $p = 3,\ 11,\ 101$. The magnitude of the eigenvalues necessarily decreases as p increases, since $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \delta_i = 1$ per definition of V_0 and $S(\boldsymbol{X})$. As one can see in Figure 1, the eigenvalues of $S(\boldsymbol{X})$ and V_0 approach each other for increasing p. In fact the maximal absolute difference for $p = 101$ is roughly $2 \cdot 10^{-4}$. In the second scenario, we take $p - 1$ equidistantly spaced eigenvalues and one eigenvalue 5 times larger than the rest, i.e.,

$$\lambda_i = \begin{cases} \frac{5(p-1)}{p((p+1)/2+5)-5} & i = 1, \\ \frac{p-i}{p((p+1)/2+5)-5} & i = 1, \ldots, p-1. \end{cases}$$

This models the case where the dependence is mainly driven by one principal component. As

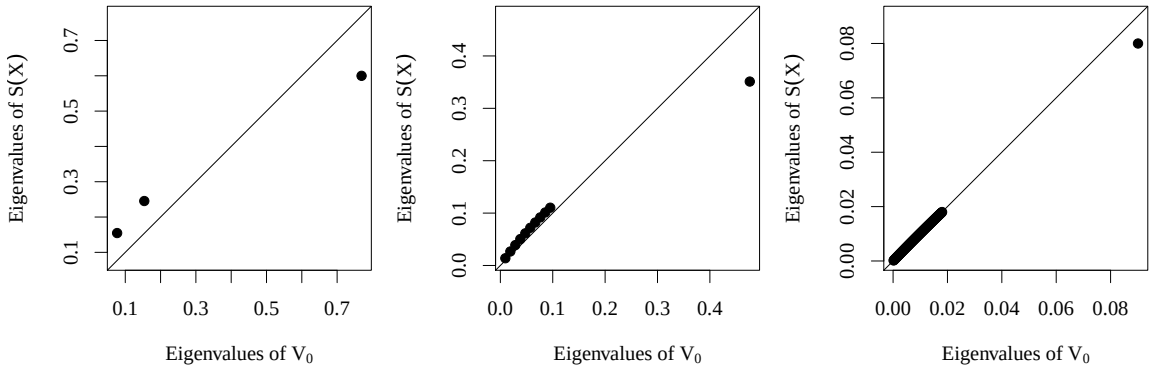

Figure 2: Eigenvalues of the SSCM wrt the corresponding eigenvalues of shape matrix in the setting of one large eigenvalue for $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

one can see in Figure 2, the distance between the two largest eigenvalues is smaller for $S(\boldsymbol{X})$ than for V_0. This is not surprising in the light of (2). Thus in general, the eigenvalues of the SSCM are less separated than those of V_0, which is one reason why the use of the SSCM for robust principal component analysis has been questioned (e.g. Bali, Boente, Tyler, and Wang 2011; Magyar and Tyler 2014). However, the differences appear to be generally small in higher dimensions.

3. Estimation of the correlation matrix

In the bivariate case, a robust estimator for the correlation coefficient based on the SSCM

can be obtained by inverting (1). Let

$$\rho_n = \hat{v}_{12}/\sqrt{\hat{v}_{11}\hat{v}_{22}} \quad \text{where} \quad V_n = (\hat{v}_{ij})_{i,j=1,2} = S_n^2,$$

and S_n is the bivariate SSCM. We call this estimator the spatial sign correlation coefficient. For more information, see [6]. Under mild regularity assumptions this estimator is consistent under elliptical distributions and asymptotically normal with variance

$$\mathrm{ASV}(\rho_n) = (1-\rho^2)^2 + \frac{1}{2}(a + a^{-1})(1-\rho^2)^{3/2}, \tag{4}$$

where $a = \sqrt{v_{11}/v_{22}}$ is the ratio of the marginal scales and $\rho = v_{12}/\sqrt{v_{11}v_{22}}$ is the generalized correlation coefficient, which coincides with the usual moment correlation coefficient if second moments exists. Equation (4) indicates, that for fixed ρ, the variance of ρ_n is minimal for $a = 1$, but can get arbitrarily large if a tends to infinity or 0.

Therefore a two-step procedure has been proposed, the *two-stage spatial sign correlation* $\rho_{\sigma,n}$, which first margin-wise standardizes the data by a robust scale estimator, e.g., the median absolute deviation (MAD), and then computes the spatial sign correlation of the standardized data. Under mild conditions (see Dürre and Vogel 2016a), this two-step procedure yields an asymptotic variance of

$$\mathrm{ASV}(\rho_{\sigma,n}) = (1-\rho^2)^2 + (1-\rho^2)^{3/2}, \tag{5}$$

which equals that of ρ_n for the most favourable case of $a = 1$. Since (5) only depends on the parameter ρ, the two-stage spatial sign correlation coefficient is very suitable to construct robust and non-parametric confidence intervals for the correlation coefficient under ellipticity. It turns out that these intervals are quite accurate even for rather small sample sizes of $n = 10$ and in fact more accurate than those based on the sample moment correlation coefficient (Dürre and Vogel 2016a).

3.1. The multivariate spatial sign correlation matrix

One can construct an estimator of the correlation matrix R by filling the off-diagonal positions of the matrix estimate with the bivariate spatial sign correlation coefficients of all pairs of variables. This was proposed in Dürre, Vogel, and Fried (2015). Equation (3) allows an alternative approach: First standardize the data marginally by a robust scale estimator and compute the SSCM of the transformed data. Then apply a singular value decomposition

$$S_n(\boldsymbol{t}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \hat{U}\hat{\Delta}\hat{U}^T,$$

where $\hat{\Delta}$ contains the ordered eigenvalues $\hat{\delta}_1 \geq \ldots \geq \hat{\delta}_p$. One obtains estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ by inverting (3). Although theoretical results are yet to be established, we found in our simulations that the following fix point algorithm

$$\begin{aligned}
\hat{\lambda}_i^{(0)} &= \hat{\delta}_i, && i = 1, \ldots, p, \\
\tilde{\lambda}_i^{(k+1)} &= 2\hat{\delta}_i \left(\int_0^\infty \frac{1}{(1+\hat{\lambda}_i^{(k)}x)\prod_{j=1}^p (1+\hat{\lambda}_j^{(k)}x)^{\frac{1}{2}}} dx, \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots \\
\hat{\lambda}_i^{(k+1)} &= \tilde{\lambda}_i^{(k+1)} \left(\sum_{j=1}^p \tilde{\lambda_j}^{(k+1)} \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots
\end{aligned}$$

works reliably and converges fast, converging usually within 5 iterations if p is large. Let $\hat{\Lambda}$ denote the diagonal matrix containing $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, then $\hat{V} = \hat{U}\hat{\Lambda}\hat{U}^T$ is a suitable estimator for the shape of the standardized data and $\hat{R}$ with $\hat{\rho}_{ij} = \hat{v}_{ij}/\sqrt{\hat{v}_{ii}\hat{v}_{jj}}$ an estimator for the correlation matrix, which we call the *multivariate spatial sign correlation matrix*. As opposed

to the pairwise approach, the multivariate spatial sign correlation matrix is positive semi-definite by construction.

3.2. Simulation under elliptical distributions

By a small simulation study we want to obtain an impression of the efficiency of the multivariate spatial sign correlation matrix. We compare the variances of the moment correlation, the pairwise as well as the multivariate spatial sign correlation under several elliptical distributions: normal, Laplace and t distributions with 5 and 10 degrees of freedom. The latter three generate heavier tails than the normal distribution. The Laplace distribution is obtained by the elliptical generator $g(x) = c_p \exp(-\sqrt{|x|}/2)$, where c_p is the appropriate integration constant depending on p (e.g. Bilodeau and Brenner 1999, p. 209).

First we take the identity matrix as shape matrix and compare the variances of an off-diagonal element of the matrix estimates for different dimensions $p = 2,\ 3,\ 5,\ 10,\ 50$ and sample sizes $n = 100,\ 1000$. We use the R packages mvtnorm (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2015) and MNM (Nordhausen and Oja 2011) for the data generation. The results based on 10000 runs are summarized in Table 1.

Table 1: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for spherical normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	sscor pairwise	1.9	1.9	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_{10}	cor	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	1.3
	sscor pairwise	2.0	1.9	1.9	2.0	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.3	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_5	cor	2.0	2.1	2.1	2.1	2.1	2.6	2.6	2.6	2.6	2.6
	sscor pairwise	2.0	2.0	1.9	2.0	1.9	2.1	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.4	1.2	1.1	2.1	1.7	1.4	1.2	1.0
L	cor	1.6	1.5	1.3	1.2	1.1	1.6	1.5	1.3	1.2	1.1
	sscor pairwise	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.1	2.0	1.7	1.4	1.2	1.1

Except for the moment correlation at the t_5 distribution, the results for $n = 100$ and $n = 1000$ are very similar. Note that the variance of the moment correlation decreases at the Laplace distribution as the dimension p increases, but not so for the other distributions considered. The lower dimensional marginals of the Laplace distribution are, contrary to the normal and the t-distributions, not within the same distributional class (see Kano 1994), and the kurtosis of the one-dimensional marginals of the Laplace distribution in fact decreases as p increases.

Equation (5) yields an asymptotic variance of 2 for the pairwise spatial sign correlation matrix elements regardless of the specific elliptical generator. This can also be observed in the simulation results. The moment correlation is twice as efficient under normality, but it has a higher variance at heavy tailed distributions. For uncorrelated t_5 distributed random variables, the spatial sign correlation outperforms the moment correlation. Looking at the multivariate spatial sign correlation, we see a strong increase of efficiency for larger p. For $p = 50$ the variance is comparable to that of the moment correlation. Since the asymptotic variance of the SSCM does not depend on the elliptical generator, this is expected to apply also for the

multivariate spatial sign correlation, and this claim is confirmed by the simulations. The multivariate spatial sign correlation is more efficient than the moment correlation even under slightly heavier tails for moderately large p.

Following a referee's suggestion, we simulate also from other shape matrices, e.g., the equi-correlation matrix

$$V = \begin{pmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.5 \\ 0.5 & \dots & 0.5 & 1 \end{pmatrix}.$$

The results can be found in Table 2. Except for the general smaller asymptotic variances we get the same picture. The asymptotic variance of the multivariate spatial sign correlation matrix is shrinking with growing dimension and approaches that of the sample correlation under normality, albeit more slowly than in the uncorrelated case.

Table 2: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for equi-correlated normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	1.0	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_{10}	cor	0.8	0.7	0.7	0.8	0.8	0.8	0.7	0.8	0.7	0.8
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_5	cor	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.5	1.5	1.5
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	0.9	0.8	0.7	1.2	1.0	0.9	0.8	0.8
L	cor	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	sscor pairwise	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.9	0.7	1.2	1.0	0.9	0.8	0.7

An increase of efficiency for larger p is not uncommon for robust scatter estimators. It can be observed amongst others for M-estimators, the Tyler shape matrix, the MCD, and S-estimators (see e.g. Croux and Haesbroeck 1999; Taskinen, Croux, Kankainen, Ollila, and Oja 2006). All of these are affine equivariant estimators, requiring $n > p$. This restriction is not necessary for the spatial sign correlation matrix.

3.3. Sensitivity to outliers

One may expect that the efficiency gain for large p is at the expense of robustness. We therefore investigate the influence function of one off-diagonal element of the multivariate spatial sign correlation. The influence function is based on the concept that estimators are functionals working on distributions. In this setting the specific estimate based on a given dataset equals the functional evaluated at the corresponding empirical distribution. Denote $\check{\rho}$ the functional representation of the multivariate spatial sign correlation with matrix-elements $\check{\rho}_{i,j},\ 1 \le i < j \le p$. Then the influence function $IF(\boldsymbol{x}, \check{\rho}_{i,j}, F)$ is defined as

$$IF(\boldsymbol{x}, \check{\rho}_{i,j}, F) = \lim_{\epsilon \to 0} \frac{\check{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) - \check{\rho}_{i,j}(F)}{\epsilon}$$

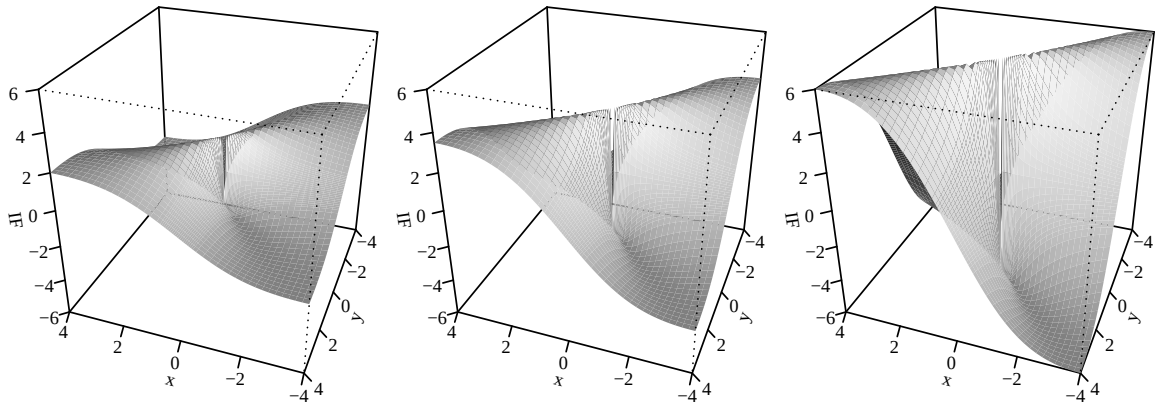

Figure 3: Partial influence functions of the off-diagonal element of multivariate spatial sign correlation $\check{\rho}_{12}$ for $\boldsymbol{x} = (x, y, 0, \ldots, 0)$ under spherical distribution for $p = 2$ (left), $p = 5$ (centre) and $p = 10$ (right).

where $\Delta_{\boldsymbol{x}}$ denotes the Dirac measure putting its mass at $\boldsymbol{x}$. For further explanations and details about the influence function, see Huber and Ronchetti (2009).
Since we do not have an explicit representation for the estimated eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, it seems to be challenging to calculate the influence function for arbitrary F and $\boldsymbol{x}$. Nevertheless, we can get results if we restrict ourselves to the case where F is elliptical with shape $V = I_p$ and $\boldsymbol{x}$ lies in a special hyperplane of $\mathbb{R}^p$. Furthermore we look at the case where the proportions of the marginal scales are known, respectively the data is not standardized prior to the computation of the SSCM. The following proposition, the proof of which can be found in the appendix, states the influence function in the outlined situation.

Proposition 1. *Let F be elliptical with shape $V = I_p$ and $\boldsymbol{\mu} = 0$. If we let $\check{\rho}_{i,j}$ denote the functional representation of the off-diagonal element of the multivariate spatial sign correlation without pre-standardization and let $\boldsymbol{x} = (x, y, 0, \ldots, 0)^T$ with $x, y \in \mathbb{R}$, then*

$$IF(\boldsymbol{x}, \check{\rho}_{1,2}, F) = (p+2)\frac{xy}{x^2+y^2}. \tag{6}$$

For $p = 2$, Proposition 1 is a special case of Proposition 4 in Dürre *et al.* (2015) which gives the influence function for arbitrary V. Although Proposition 1 is restricted to the situation where there is only contamination in the first two components, it provides evidence that the sensitivity of the multivariate spatial sign correlation increases with increasing dimension. One can see in Figure 3 respectively formula (6) that the influence functions are proportional to each other and that $|IF(\boldsymbol{x}, \check{\rho}_{1,2}, F)|$ increases linearly in p for fixed $\boldsymbol{x} = (x, y, 0 \ldots, 0)$. This result indicates that the multivariate spatial sign correlation is more effected by outliers if p is large.

4. Conclusion

We have discussed properties of the spatial sign covariance matrix, in particular those concerning its eigenvalues under elliptical distributions. We expand on the eigenvalue representation as one-dimensional integrals given in Dürre *et al.* (2016). First we use it to investigate the function mapping the eigenvalues of the shape matrix onto the ones of the spatial sign covariance. The eigenvalues of the spatial sign covariance matrix are closer together than the ones of the shape matrix on a logarithmic scale, see Dürre *et al.* (2016). Two examples suggest

that this behaviour diminishes as the dimension increases. One may suspect that the map between the eigenvalues of the spatial sign covariance matrix and the shape matrix converges towards the identity modulo a multiplicative constant as the dimension tends to infinity.
Our second application of the integral representation is the construction of the multivariate spatial sign correlation matrix. By a fixed-point algorithm one can invert the map between the eigenvalues of the shape and the spatial sign covariance matrix and, based on this, estimate the correlation matrix of an elliptical distributed random vector. We found the fixed-point algorithm to work reliably and fast for various shape matrices and dimensions. Simulations show that the resulting estimator is highly efficient in larger dimensions. Its asymptotic variance appears to approach that of the sample correlation under normality as the dimension is growing. Asymptotics confirming the simulation results are of great interest. The calculated partial influence function indicates that the efficiency gain of the spatial sign correlation matrix is at the cost of robustness. So the estimator does not seem to be very robust in the case of very high dimensions, but is nevertheless very efficient under heavy-tailed distributions.

Acknowledgements

Alexander Dürre and Roland Fried were supported in part by the Collaborative Research Grant 823 of the German Research Foundation. The authors are grateful to the referee for the constructive comments, which helped to improve the presentation of the article.

Appendix

Proof of Proposition 1: Let denote $\check{S}$ the functional representation of the spatial sign covariance matrix $\check{S}(F) = \mathsf{E}_F\left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{\boldsymbol{X}^T\boldsymbol{X}}\right)$ where $\boldsymbol{X}$ has distribution F. Since $\check{S}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) = (1-\epsilon)I_p + \epsilon\boldsymbol{x}\boldsymbol{x}^T$ is a block diagonal matrix, we get the following eigenvalue decomposition $\check{S}((1-\epsilon)F + \epsilon\boldsymbol{x}\boldsymbol{x}^T) = U_{xy}\Delta_\epsilon U_{xy}^T$ where

$$U_{x,y} = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ \frac{y}{\sqrt{x^2+y^2}} & \frac{-x}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{pmatrix} \quad \text{and} \quad \Delta_\epsilon = \begin{pmatrix} \frac{1+(p-1)\epsilon}{p} & 0 & 0 & \dots & 0 \\ 0 & \frac{1-\epsilon}{p} & 0 & \dots & 0 \\ 0 & 0 & \frac{1-\epsilon}{p} & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & \frac{1-\epsilon}{p} \end{pmatrix}.$$

We need to know how the perturbation of the eigenvalues of the SSCM translates into the eigenvalues of the shape matrix. The function $\Phi : \mathbb{R}^p \to \mathbb{R}^p$ which maps the eigenvalues of the shape to the eigenvalues of the SSCM is injective (see Proposition 1 in Dürre *et al.* 2016). Therefore the shape matrix related to Δ_ϵ contains only two distinct eigenvalues: λ_1 and $\lambda_2 = \ldots = \lambda_p$. We can simplify the situation even further since the eigenvalues are not uniquely defined and standardize them such that $\lambda_2 = \ldots, \lambda_p = 1$. On the other hand we have $\sum_{i=1}^p \delta_i = 1$ and therefore $\delta_i = \frac{1-\delta_1}{p-1}$, $i = 2, \ldots, p$. Consequently in this case the connection between the eigenvalues can be expressed by the one-dimensional function $f : [0,1] \to [0,\infty)$ which maps the first eigenvalue of Δ_ϵ to the first of the shape matrix.
Let $\gamma : \mathbb{R}^{p\times p} \to [-1,1]$ denote the function which computes the correlation coefficient between the first and second component given the shape matrix: $\gamma(A) = \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$ and denote further $k(\epsilon) = \frac{1+(p-1)\epsilon}{p}$, then straightforward calculations yields,

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{\check{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) - \check{\rho}_{i,j}(F)}{\epsilon} &= \lim_{\epsilon\to 0} \frac{\gamma\left(U_{xy} f\left(\frac{1+(p-1)\epsilon}{p}\right) U_{xy}^T\right)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \frac{(f[k(\epsilon)]-1)xy}{\sqrt{y^2 + f[k(\epsilon)]x^2}\sqrt{x^2 + f[k(\epsilon)]y^2}} =: \frac{\partial}{\partial\epsilon} h(f[k(\epsilon)])\Big|_{\epsilon=0}. \end{aligned}$$

By the chain rule we get:

$$\frac{\partial}{\partial \epsilon} h(f[k(\epsilon)])\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} h(x)\Big|_{x=1} \cdot \frac{\partial}{\partial y} f(y)\Big|_{y=1/p} \cdot \frac{\partial}{\partial \epsilon} k(\epsilon)\Big|_{\epsilon=0}.$$

Whereas differentiation of h and k is straightforward, we do not have an explicit representation of f. Since we only need its derivative, we can apply the inverse function theorem. Using (3) and Leibniz's rule we arrive at

$$\begin{aligned}\frac{\partial}{\partial x} f(x)\Big|_{x=1/p} &= \frac{1}{\frac{\partial}{\partial x} f^{-1}(x)|_{x=1}} \\ &= 1/\left(\frac{1}{2}\int_0^\infty \frac{1}{(1+z)^{\frac{p}{2}+1}} dz - \frac{3}{4}\int_0^\infty \frac{z}{(1+z)^{\frac{p}{2}+2}} dz\right) =: \frac{1}{A_1 + A_2}.\end{aligned}$$

For A_1 and A_2 we can apply formula 3.193-3 in Gradshteyn and Ryzhik (2000):

$$\int_0^\infty \frac{x^{\mu-1} dx}{(1+\beta x)^\nu}\, dx = B(\mu, \nu-\mu) \quad \text{for} \quad \nu > \mu > 0$$

where B denotes the beta function. Setting $\beta = 1$, $\mu = 1$ and $\nu = p/2+1$ for A_1 respectively $\mu = 2$ and $\nu = p/2+2$ for A_2 and using the relationship between beta and gamma function we arrive at $A_1 = \frac{1}{p}$ and $A_2 = \frac{3}{2p(p/2+1)}$. Straightforward term manipulations yield the stated formula (6). □

References

Bali JL, Boente G, Tyler DE, Wang JL (2011). "Robust Functional Principal Components: A Projection-pursuit Approach." *The Annals of Statistics*, **39**(6), 2852–2882.

Bilodeau M, Brenner D (1999). *Theory of Multivariate Statistics*. Springer Science & Business Media.

Croux C, Dehon C, Yadine A (2010). "The k-step Spatial Sign Covariance Matrix." *Advances in data analysis and classification*, **4**(2-3), 137–150.

Croux C, Haesbroeck G (1999). "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator." *Journal of Multivariate Analysis*, **71**(2), 161–190.

Dürre A, Tyler DE, Vogel D (2016). "On the Eigenvalues of the Spatial Sign Covariance Matrix in More than Two Dimensions." *Statistics & Probability Letters*, **111**, 80–85.

Dürre A, Vogel D (2016a). "Asymptotics of the Two-stage Spatial Sign Correlation." *Journal of Multivariate Analysis*, **144**, 54–67.

Dürre A, Vogel D (2016b). *sscor: Spatial Sign Correlation*. R package version 0.2, URL http://CRAN.R-project.org/package=sscor.

Dürre A, Vogel D, Fried R (2015). "Spatial Sign Correlation." *Journal of Multivariate Analysis*, **135**, 89–105.

Dürre A, Vogel D, Tyler DE (2014). "The Spatial Sign Covariance Matrix with Unknown Location." *Journal of Multivariate Analysis*, **130**, 107–117.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2015). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-3, URL http://CRAN.R-project.org/package=mvtnorm.

Gradshteyn I, Ryzhik I (2000). *Table of Integrals, Series, and Products. Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger.* 6th edition. Amsterdam: Elsevier/Academic Press.

Huber PJ, Ronchetti E (2009). *Robust Statistics.* Wiley.

Kano Y (1994). "Consistency Property of Elliptic Probability Density Functions." *Journal of Multivariate Analysis*, **51**(1), 139–147.

Kemperman JHB (1987). "The Median of a Finite Measure on a Banach Space." In Y Dodge (ed.), *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pp. 217–230. Amsterdam: North-Holland.

Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K (1999). "Robust Principal Component Analysis for Functional Data." *Test*, **8**(1), 1–28.

Magyar AF, Tyler DE (2014). "The Asymptotic Inadmissibility of the Spatial Sign Covariance Matrix for Elliptically Symmetric Distributions." *Biometrika*, **101**(3), 673–688.

Marden JI (1999). "Some Robust Estimates of Principal Components." *Statistics & Probability Letters*, **43**(4), 349–359.

Nordhausen K, Oja H (2011). "Multivariate L_1 Methods: The Package MNM." *Journal of Statistical Software*, **43**(5), 1–28. URL http://www.jstatsoft.org/v43/i05/.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Taskinen S, Croux C, Kankainen A, Ollila E, Oja H (2006). "Influence Functions and Efficiencies of the Canonical Correlation and Vector Estimates Based on Scatter and Shape Matrices." *Journal of Multivariate Analysis*, **97**(2), 359–384.

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

Affiliation:

Alexander Dürre
Department of Statistics
Technische Universität Dortmund
44221 Dortmund, Germany
E-mail: alexander.duerre@udo.edu
URL: https://www.statistik.tu-dortmund.de/duerre-en.html