Journal of Multivariate Analysis 135 (2015) 89–105

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Spatial sign correlation

Alexander Dürre *, Daniel Vogel, Roland Fried

Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany

ARTICLE INFO

Article history:
Received 1 April 2014
Available online 16 December 2014

AMS 2000 subject classifications:
62H12
62G05
62G35

Keywords:
Elliptical distribution
Gaussian rank correlation
Gnanadesikan–Kettenring estimator
Kendall's tau
Spatial median
Spatial sign covariance matrix
Spearman's rho

ABSTRACT

A robust correlation estimator based on the spatial sign covariance matrix (SSCM) is proposed. We derive its asymptotic distribution and influence function at elliptical distributions. Finite sample and robustness properties are studied and compared to other robust correlation estimators by means of numerical simulations.

1. Introduction

The research presented in this article is motivated by the problem of robust and high-dimensional correlation estimation. By *robust* we mean insusceptible to outliers and erroneous observations, that is, we examine alternatives to the commonly used, but highly non-robust Pearson correlation. Over the last few decades, many robust multivariate scatter estimators, and consequently robust correlation matrices, have been proposed, see, e.g., [27] for a review. Much attention has been paid to affine equivariant estimators. If we denote by $\mathbb{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ the $n \times p$ data matrix containing the p-dimensional observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as rows, then the data set $\mathbb{Y}_n = \mathbb{X}_n A^T + \mathbf{1}_n \boldsymbol{b}^T$ is obtained by applying the affine linear transformation $\boldsymbol{x} \mapsto A\boldsymbol{x} + \boldsymbol{b}$ to each data point. An affine equivariant scatter estimator $\hat{S}_n$ satisfies $\hat{S}_n(\mathbb{Y}_n) = A\hat{S}_n(\mathbb{X}_n)A^T$ for any $\boldsymbol{b} \in \mathbb{R}^p$ and any full rank square matrix A, i.e. it behaves as the covariance matrix under linear transformations of the data.

The second attribute *high-dimensional* means two things: being fast to compute, also in high-dimensions, and being defined also for sparse, high-dimensional data, i.e. in the $p > n$ situation. Both properties basically prohibit robust, affine equivariant estimators: they are usually hard to compute in high dimensions, and they are not defined in the $p > n$ setting—or coincide with a multiple of the sample covariance matrix [47] and are thus not robust. In fact, both requirements suggest the use of pairwise correlation estimators. In a pairwise correlation estimate $\hat{P}_n \in \mathbb{R}^{p\times p}$ each entry $\hat{\rho}_{i,j}$ is computed only from the ith and the jth coordinate of the data, implying that the computing time increases quadratically with p.

In the present article, we are thus concerned with correlation estimation in the bivariate setting. We propose a new, highly robust correlation estimator that is fast to compute, making it appealing as a building block for pairwise correlation matrix estimation in high dimensions. The new proposal is based on the spatial sign covariance matrix. The *spatial sign* of a

* Corresponding author.
E-mail address: alexander.duerre@udo.edu (A. Dürre).

http://dx.doi.org/10.1016/j.jmva.2014.12.002
0047-259X/

multivariate observation is its projection (after a suitable centering) onto the p-dimensional unit sphere. Spatial signs play an important role in robust multivariate data analysis. Since every observation is basically shrunk to length 1, the impact of any contamination is bounded. Spatial signs have been used, e.g., for robust tests of multivariate location (e.g. [31]), tests of independence [43], testing for sphericity [41] or canonical correlation analysis [42]. Using spatial signs as score function in estimation leads to the spatial median as a multivariate location estimator or, in the regression setting, to the least absolute deviation (LAD) regression. For a recent overview of spatial sign methods see [34].

The *spatial sign covariance matrix* (SSCM) is simply the covariance matrix of the spatial signs of the (suitably centered) observations. It is known that, within symmetric data models, the SSCM consistently estimates the eigenvectors of the covariance matrix, but not the eigenvalues. In fact, the connection between the eigenvalues of the population SSCM and the covariance matrix is an open problem. We solve this problem for the special case of two-dimensional elliptical distributions. This enables us to robustly estimate a two-dimensional covariance matrix (up to scale) based on the SSCM and hence devise a correlation estimator, which we call *spatial sign correlation*. We further derive the asymptotic distributions and influence functions of the SSCM and the spatial sign correlation. The main advantage of the new estimator is its simplicity. It is very fast to compute, it requires neither an iterative algorithm nor any ranking or sorting of the data. It is furthermore distribution-free within the elliptical model, it behaves equally well for very heavy-tailed and very peaked distributions, which is true for hardly any other robust scatter estimator.[1]

The paper has two parts: In part 1, consisting of Sections 2–4, we develop the spatial sign correlation estimator and derive its asymptotics. Section 2 deals with the asymptotics of the spatial sign covariance matrix in the two-dimensional case, and in Section 3, the asymptotics of the spatial sign correlation are studied. Since this estimator can become very inefficient when the marginal variances strongly differ, we explore in Section 4 a two-stage version of the estimator, where the correlation estimation is preceded by a component-wise standardization of the data. Being aware that the spatial sign correlation is one out of many that were introduced for similar purposes, the second part, consisting of Sections 5 and 6, gathers together analytic results about a variety of alternatives and compares them in an elaborate simulation study to provide some guidance within the ever increasing number of robust correlation estimates. All proofs are deferred to the Appendix.

We close this section by introducing some recurrent terms and notation. In order to study the properties of the new estimator analytically we will assume the data to stem from the elliptical model. A continuous distribution F on $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

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

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix V. We call $\boldsymbol{\mu}$ the *location* or *symmetry center* and V the *shape matrix*, since it describes the shape of the elliptical contour lines of the density. The class of all continuous elliptical distributions F on $\mathbb{R}^p$ having these parameters is denoted by $\mathscr{E}_p(\boldsymbol{\mu}, V)$. The shape matrix V is unique only up to scale, that is, $\mathscr{E}_p(\boldsymbol{\mu}, V) = \mathscr{E}_p(\boldsymbol{\mu}, cV)$ for any $c > 0$. For scale-free functions of V, such as correlations, which we consider here, this ambiguity is irrelevant. A common view on the *shape* of an elliptical distribution is to treat it as an equivalence class of positive definite random matrices being proportional to each other. We adopt this notion here: in the results of this exposition, V can be any representative from its equivalence class. If second moments exist, one can always take the covariance matrix, or any suitably scaled multiple of it. However, the results are more general, the existence of second, or even first, moments is not required. Throughout the paper we let

$$V = U\Lambda U^T \tag{2}$$

denote an eigenvalue decomposition of V, where U is an orthogonal matrix containing the eigenvectors of V as columns and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ is such that $0 < \lambda_p \leq \cdots \leq \lambda_1$. We use $\|\cdot\|$ to denote the L_2 norm of a vector.

2. The spatial sign covariance matrix

We define the spatial sign covariance matrix of a multivariate distribution and derive its connection to the shape matrix V in case of a two-dimensional elliptical distribution. For $\boldsymbol{x} \in \mathbb{R}^p$ define the *spatial sign* $\boldsymbol{s}(\boldsymbol{x})$ of $\boldsymbol{x}$ as $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$ if $\boldsymbol{x} \neq \boldsymbol{0}$ and $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{0}$ otherwise. Let $\boldsymbol{X}$ be a p-dimensional random vector ($p \geq 2$) having distribution F. We call $\boldsymbol{\mu}(F) = \boldsymbol{\mu}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathbb{E}(\|\boldsymbol{X}-\boldsymbol{\mu}\| - \|\boldsymbol{X}\|)$ the *spatial median* and, following the terminology of Visuri et al. [50],

$$S(F) = S(\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{s}(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{s}(\boldsymbol{X}-\boldsymbol{\mu})^T\right) \tag{3}$$

the *spatial sign covariance matrix (SSCM)* of F (or $\boldsymbol{X}$). If there is no unique minimizing point of $\mathbb{E}(\|\boldsymbol{X}-\boldsymbol{\mu}\| - \|\boldsymbol{X}\|)$, then $\boldsymbol{\mu}(F)$ is the barycenter of the minimizing set. This may only happen if F is concentrated on a line. For results on existence and uniqueness of the spatial median see [16,19,29] or [20]. If the first moments of F are finite, then the spatial median allows the more descriptive characterization as $\arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathbb{E}\|\boldsymbol{X}-\boldsymbol{\mu}\|$. Let $\mathbb{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ be a data sample of size n, where the $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are i.i.d., each with distribution F. Define

$$\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}) = \operatorname*{ave}_{i=1,\ldots,n} \boldsymbol{s}(\boldsymbol{X}_i - \boldsymbol{t})\boldsymbol{s}(\boldsymbol{X}_i - \boldsymbol{t})^T \tag{4}$$

[1] For these statements to be true, the SSCM has to based on an appropriate location estimator.