Signed Peripherality Functions for Multivariate Statistical Inference

Subhabrata Majumdar, Snigdhansu Chatterjee

Abstract: We introduce a notion of multivariate ranking based on the composition of the multivariate sign function and data-dependent weights that are increasing or decreasing functions of a point-to-distribution distance measure in multivariate space, which we call the peripherality function. Using this weighted sign vector, we propose estimation and testing methods for the mean vector of an elliptical distribution that improve upon corresponding spatial sign-based procedures. When bounded and monotonically decreasing transformations on data depth functions are used as weights, the rank vectors can be used as proxies of the original data to estimate different components of the covariance matrix in the data-generating elliptical distribution: its eigenvectors, eigenvalues and the covariance matrix itself. Wimplement our proposed methods to perform robust Sufficient Dimension Reduction (Adragni and Cook, 2009) and functional outlier detection. Throughout the paper, several simulation studies and data examples outline the robustness and efficiency of our methods.

Keywords: Multivariate ranking, data depth, high-dimensional location testing, principal component analysis, sufficient dimension reduction, robust PCA, outlier detection

1 Introduction

Consider a point $\mathbf{x}$ from some $\mathcal{X} \subseteq \mathbb{R}^p$, and the spatial sign function $\mathbf{S} : \mathcal{X} \times \mathbb{R}^p \mapsto \mathbb{R}^p$, defined as

$$\mathbf{S}(\mathbf{x}; \boldsymbol{\mu}_x) = \frac{\mathbf{x} - \boldsymbol{\mu}_x}{\|\mathbf{x} - \boldsymbol{\mu}_x\|} \mathbb{I}(\mathbf{x} \neq \boldsymbol{\mu}_x),$$

where $\boldsymbol{\mu}_x \in \mathbb{R}^p$ is a fixed location parameter, $\mathbb{I}(\cdot)$ is the 0/1 indicator function and $\|.\|$ is the euclidean norm. This is a generalization of the real-valued case of the indicator for whether a point $x \in \mathbb{R}$ is to the right, left or at a scalar $\mu_x \in \mathbb{R}$, and was first introduced by Möttönen and Oja (1995). Given a data matrix $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, a popular approach of using spatial signs in the literature has been to take spatial signs of the data points, i.e. $\mathbf{S}(\mathbf{x}_i)$, and use the matrix of signs as a proxy of $\mathbb{X}$ to formulate robust versions of location and scale problems (Locantore et al., 1999; Oja, 2010; Wang et al., 2015).

The spatial sign maps all data points to the surface of a p-dimensional unit sphere (see Figure 1). While this transformation has its advantages (i.e. robustness), e.g. the eigenvectors remain the same while limiting the influence of outliers when the points are drawn from an elliptical distribution (Taskinen et al., 2012), valuable information is lost in the form of magnitudes of sample points. As a result, spatial sign-based procedures suffer from low efficiency. For example, eigenvector estimates obtained from the covariance matrix of spatial signs, which is called the Sign Covariance Matrix (SCM), are asymptotically inadmissible (Magyar and Tyler, 2014)- in the sense that there is an estimator (namely, Tyler's M-estimate of scatter (Tyler, 1987)) that has uniformly lower asymptotic risk than the SCM.

To utilize the magnitude information of sample points, at the same time preserving the robustness property, we propose a general class weighted spatial sign functions. To this end, we first define what we call a *peripherality function*.

Definition 1.1. For a set of probability measures $\mathcal{M}$ taking values in $\mathbb{R}^p$, a *peripherality function* $P : \mathbb{R}^p \times \mathcal{M} \mapsto \mathbb{R}^+ \cup \{0\}$, is a function that satisfies the following condition:

(P) For every probability measure $\mathbb{F} \in \mathcal{M}$, there exists a constant $\boldsymbol{\mu}_F \in \mathbb{R}^p$ such that for any $t \in [0, 1]$ and $\mathbf{x} \in \mathbb{R}^p$,

$$P(\boldsymbol{\mu}_F; \mathbb{F}) \leq P(\boldsymbol{\mu}_F + t(\mathbf{x} - \boldsymbol{\mu}_F); \mathbb{F}).$$

That is, for every fixed $\mathbb{F}$, the peripherality function achieves a minimum at μ_F, and is non-decreasing in every direction away from $\boldsymbol{\mu}_F$.

Peripherality functions quantify the relative position of $\mathbf{x}$ with respect to $\boldsymbol{\mu}_F$, and condition (P) formalizes a general notion of centrality. A whole class of such functions can be defined from *Data Depth* functions (Zuo and Serfling, 2000), which give the center-outward ranking of points in a multivariate dataset. Data depth has seen extensive use in the past decade, ranging from robust nonparametric classification (Dutta and Ghosh, 2012; Ghosh and Chaudhuri, 2005; Jornsten, 2004; Sguera et al., 2014) to parametric estimation of means (Zuo et al., 2004), covariance matrices (Zuo and Cui, 2005) and sparse

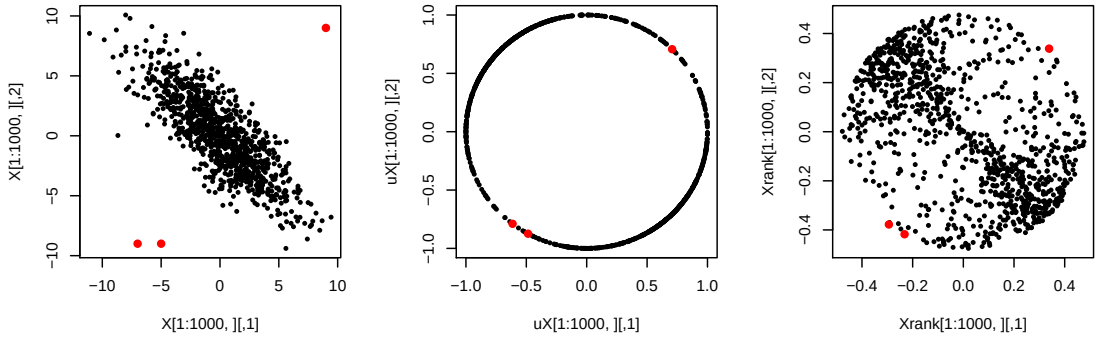


Figure 1: (Left) 1000 points randomly drawn from $\mathcal{N}_2\left((0,0)^T, \left(\begin{smallmatrix} 5 & -4 \\ -4 & 5 \end{smallmatrix}\right)\right)$, (center) their spatial signs, and (right) signed peripherality functions based on halfspace depth.

multivariate regression coefficients (Dutta and Genton, 2017; Majumdar and Chatterjee, 2018). Peripherality functions can be defined as some inverse ranking based on data depth, and the concept of *outlyingness* associated with data depth is essentially same as what we use in this paper. However, traditional depth functions require a wide array of theoretical conditions (Zuo and Serfling, 2000), and a sizeable portion of methods we discuss do not require the use of depth functions *per se* to be implemented.

Our weighted sign vector, which we call the *signed peripherality* function, is now defined as the product of the spatial sign and transformed peripherality functions:

$$\mathbf{R}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbb{F}) = \mathbf{S}(\mathbf{x}, \boldsymbol{\mu}_x).\kappa_P(P(\mathbf{x}, \mathbb{F})), \tag{1.1}$$

where $\kappa_P : \mathbb{R} \to \mathbb{R}$ is a monotone function. Its purpose is to either downweight or upweight the spatial sign of a point based on its peripherality with respect to a data cloud. As seen in the third panel of Figure 1, $\mathbf{R}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbb{F})$ maps points *inside* a ball of finite radius, thus limiting the influence of outlying points *and* preserving the magnitude information of data points. In this paper, we demonstrate certain interesting applications of this multivariate rank transformation in statistical inference. Notice that for the trivial choice $P(\mathbf{x}; \mathbb{F}) = \|\mathbf{x} - \boldsymbol{\mu}_F\|$ and $\boldsymbol{\mu}_x = \boldsymbol{\mu}_F, \kappa_P(x) = x$, we get $\mathbf{R}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbb{F}) = \mathbf{x}$. However, in this paper we illustrate how using the rank-transformed data in place of the original data points or their spatial signs leads to interesting robustness results for other choices of the peripherality functions.

The notion of multivariate ranks goes back to Puri and Sen (1971), where they take the vector consisting of marginal univariate ranks as multivariate rank vector. Subsequent definitions of multivariate ranks were proposed by Möttönen and Oja (1995), Hallin and Paindaveine (2002) and Chernozhukov et al. (2017). Compared to them, the advantages of our signed peripherality function are two-fold: interpretability and flexibility. Firstly, as evident in Figure 1, our rank transformation preserves the general shape of the data. In fact, if the map $\|\mathbf{x}\| \to P(\mathbf{x}, \mathbb{F})$ is one-one (given a fixed $\mathbb{F}$), then it is possible to get back $\mathbf{x}$ from $\mathbf{R}(\mathbf{x}; \boldsymbol{\mu}_x, \mathbb{F})$. It also provides an intuitive extension to any spatial sign-based methodology. Secondly, there is considerable scope of generalization within the present

framework. Given any general inner product space, the definition of signed peripherality remains the same when euclidean norm is replaced by the norm induced by the inner product. Even within $\mathbb{R}^p$, it is possible to generalize the ranking function in (1.1) by adding separate transformations on the sign and peripherality functions to obtain alternate representations of the data in lower or higher dimensional spaces. However, we limit our discussion in this paper to the original formulation in (1.1).

We assume $\mathbb{F}$ to be an elliptical distribution. Following Fang et al. (1990), elliptical distributions can be formally defined here using their characteristic function:

Definition 1.2. A p-dimensional random vector $\mathbf{X}$ is said to elliptically distributed if and only if there exist a vector $\boldsymbol{\mu} \in \mathbb{R}^p$, a positive semi-definite matrix $\Omega \equiv \Sigma^{-1} \in \mathbb{R}^{p\times p}$ and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $\mathbf{t} \mapsto \phi_{\mathbf{X}-\boldsymbol{\mu}}(\mathbf{t})$ of $\mathbf{X} - \boldsymbol{\mu}$ corresponds to $\mathbf{t} \mapsto \phi(\mathbf{t}^T\Sigma\mathbf{t}), \mathbf{t} \in \mathbb{R}^p$.

The density function of an elliptically distributed random variable takes the form:

$$h(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = |\Omega|^{1/2} g((\mathbf{x} - \boldsymbol{\mu})^T \Omega(\mathbf{x} - \boldsymbol{\mu})),$$

where g is a non-negative scalar-valued density function that is continuous and strictly increasing, and is called the *density generator* of the elliptical distribution. For ease of notation, we denote such a distribution by $\mathcal{E}(\boldsymbol{\mu}, \Sigma, g)$.

Organization of paper. We start by elaborating on the use of weighted spatial signs in the problem of estimating or testing for the location parameter $\boldsymbol{\mu}$ of $\mathbb{F}$ in Section 2. Specifically, we use the signed peripherality vectors in (1.1) to propose more efficient versions of the spatial median (Brown, 1983; Chaudhuri, 1996) and the high-dimensional testing procedure of Wang et al. (2015). In Section 3, we take inverse depth functions (Majumdar and Chatterjee, 2018) as our weights, and lay out a framework to estimate the different components of the covariance matrix Σ- namely, its eigenvectors (i.e. principal component analysis (PCA)), eigenvalues and Σ itself. The two subsequent sections are devoted to the applications of the robust PCA method in Sections 2 and 3. In Section 4 we use the robust estimates of $\boldsymbol{\mu}$ and eigenvectors of Σ in a supervised learning problem, while in Section 5 we propose a robust functional PCA based on our methodology. We conclude with a discussion in 6. The appendix contains auxiliary theoretical results and proofs.

Notation. We denote vectors $\mathbf{a} \in \mathbb{R}^A$ by bold small letters, matrices $\mathbb{A} \in \mathbb{R}^{A\times A}$ by blackboard bold or greek capital letters, vector-valued random variables $\mathbf{A}$ by bold capital letters and their elements by small letters. For the vector $\mathbf{a}$, we denote euclidean norm and ℓ_1-norm by $\|a\|$ and $\|a\|_1$, respectively. We denote expectations by $\mathbb{E}(\cdot)$, the variance of a scalar random variable or covariance of a vector-valued random variable by $\mathbb{V}(\cdot)$, and the covariance of two scalar random variables by $\mathbb{C}(\cdot)$. Asymptotic variances (or covariance matrices) and covariances are denoted by $A\mathbb{V}(\cdot)$ and $A\mathbb{C}(\cdot)$, respectively. Sample variances are denoted by $\mathbb{S}(\cdot)$, i.e. $\mathbb{S}(\mathbf{a}) = \mathbf{a}^T\mathbf{a}/a - (\sum_{i=1}^A a_i/a)^2$ and $\mathbb{S}(\mathbb{B}) = \mathbb{B}^T\mathbb{B}/B - \bar{\mathbf{b}}\bar{\mathbf{b}}^T$ for $\mathbb{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_B)^T \in \mathbb{R}^{B\times B_1}$, $\bar{\mathbf{b}} = \sum_{i=1}^B \mathbf{b}_i/B$.

2 Weighted sign vectors in location problem

In this section, we focus on the general situation of estimating or testing for the location parameter $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_F$ using the weighted sign vectors in (1.1). We impose the following very general conditions on the weights $w(\mathbf{x}, \mathbb{F}) := \kappa_P(P(\mathbf{x}, \mathbb{F}))$:

(A1) The weights are affine invariant, i.e.

$$w(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{A}\mathbb{F} + \mathbf{b}) = w(\mathbf{x}, \mathbb{F})$$

for any $\mathbf{b} \in \mathbb{R}^p$ and full rank matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$.
(A2) Given a fixed $\mathbb{F}$, the weights are square-integrable:

$$\int [w(\mathbf{y}, \mathbb{F})]^2 d\mathbb{G}(\mathbf{y}) < \infty; \mathbb{G} \in \mathcal{M}.$$

Because of the affine invariance, it is possible to simplify the weight as a function of the norm of $\mathbf{z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$: $w(\mathbf{x}, \mathbb{F}) = f(r)$, with $r = \|\mathbf{z}\|$.

A straightforward use of weighted signs in a location problem is to construct an outlier-robust alternative to the Hotelling's T^2 test using their sample mean vector and covariance matrix. Formally, this means testing for $H_0 : \boldsymbol{\mu} = \mathbf{0}_p$ vs. $H_1 : \boldsymbol{\mu} \neq \mathbf{0}_p$ based on the test statistic:

$$T_{n,w} = n\bar{\mathbf{x}}_w^T (\mathbb{S}(\mathbb{X}_w))^{-1} \bar{\mathbf{x}}_w,$$

with $\bar{\mathbf{x}}_w = \sum_{i=1}^n \mathbf{x}_{w,i}/n, \mathbf{x}_{w,i} = w(\mathbf{x}_i)\mathbf{S}(\mathbf{x}_i)$ for $i = 1, 2, ..., n$, and $\mathbb{X}_w = (\mathbf{x}_{w,1}, \ldots, \mathbf{x}_{w,n})^T$. However, the following holds true for this weighted sign test:

Proposition 2.1. *Consider n random variables $\mathbf{Z}_1, ..., \mathbf{Z}_n$ distributed independently and identically as $\mathcal{E}(\boldsymbol{\mu}, k\mathbf{I}_p, g); k \in \mathbb{R}$, and the class of hypothesis tests defined above. Then, given any $\alpha \in (0, 1)$, local power at $\boldsymbol{\mu} \neq \mathbf{0}_p$ for the level-α test based on $T_{n,w}$ is maximum when $w(\mathbf{Z}_1) = c$, a constant independent of $\mathbf{Z}_1$.*

This essentially means that power-wise the (unweighted) spatial sign test (Oja, 2010) is optimal in the given class of hypothesis tests when the data comes from a spherically symmetric distribution. Our simulations show that this empirically holds for non-spherical but elliptic distributions as well.

2.1 The weighted spatial median

To explore usage of weighted spatial signs in the location problem that improve upon the state-of-the-art, we concentrate on the following optimization problem:

$$\boldsymbol{\mu}_w = \underset{\boldsymbol{\mu}_0 \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}(w(\mathbf{X})|\mathbf{X} - \boldsymbol{\mu}_0|). \tag{2.1}$$

This can be seen as a generalization of the Fermat-Weber location problem (which has the spatial median (Brown, 1983; Chaudhuri, 1996) as the solution) using data-dependent weights. Using affine equivariant weights in (2.1) ensures that the weights are independent

	t_3	t_5	t_{10}	t_{20}	Normal
$p = 5$	1.28	1.20	1.16	1.14	1.13
$p = 10$	1.15	1.10	1.07	1.07	1.06
$p = 20$	1.09	1.05	1.04	1.03	1.03
$p = 50$	1.05	1.02	1.01	1.01	1.01

Table 1: Table of $ARE(\boldsymbol{\mu}_w; \boldsymbol{\mu}_s)$ for different spherical distributions

of $\boldsymbol{\mu}_0$, allowing the optimization problem to have a unique solution. We call this solution the *weighted spatial median* of $\mathbb{F}$, and denote it by $\boldsymbol{\mu}_w$. In a sample setup it is estimated by iteratively solving the equation $\sum_{i=1}^{n} w(\mathbf{X}_i)\mathbf{S}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_w)/n = \mathbf{0}_p$.

The sample weighted spatial median $\hat{\boldsymbol{\mu}}_w$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\mu}_w$, with the following result gives its asymptotic distribution:

Theorem 2.2. *Let $\mathbb{A}_w, \mathbb{B}_w$ be two matrices, dependent on the weight function w such that*

$$\mathbb{A}_w = \mathbb{E}\left[\frac{w(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|}\left(1 - \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right)\right]; \quad \mathbb{B}_w = \mathbb{E}\left[\frac{(w(\boldsymbol{\epsilon}))^2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right],$$

where $\boldsymbol{\epsilon} \sim \mathcal{E}(\mathbf{0}, \Sigma, g)$. Then

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_w - \boldsymbol{\mu}_w) \rightsquigarrow \mathcal{N}_p(\mathbf{0}, \mathbb{A}_w^{-1}\mathbb{B}_w\mathbb{A}_w^{-1}). \tag{2.2}$$

Setting $w(\boldsymbol{\epsilon}) = 1$ above yields the asymptotic covariance matrix for the spatial median. Following this, the asymptotic relative efficiency (ARE) of $\boldsymbol{\mu}_w$ corresponding to some non-uniform weight function with respect to the spatial median, say $\boldsymbol{\mu}_s$ will be:

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \left[\frac{\det(\mathbb{A}^{-1}\mathbb{B}\mathbb{A}^{-1})}{\det(\mathbb{A}_w^{-1}\mathbb{B}_w\mathbb{A}_w^{-1})}\right]^{1/p}$$

with $\mathbb{A} = \mathbb{E}[1/\|\boldsymbol{\epsilon}\|(\mathbb{I}_p - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2)]$ and $\mathbb{B} = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2]$. This is further simplified under spherical symmetry:

Corollary 2.3. *For the spherical distribution $\mathcal{E}(\boldsymbol{\mu}, k\mathbf{I}_p, g); k \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^p$, we have*

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \frac{\left[\mathbb{E}\left(\frac{f(r)}{r}\right)\right]^2}{\mathbb{E}f^2(r)\left[\mathbb{E}\left(\frac{1}{r}\right)\right]^2}.$$

Table 1 summarizes the AREs for several families of elliptic distributions, numerically calculated using 10,000 random samples, and taking $f(r) = 1/(1 + r)$. It is evident from the table that the weighted spatial median outperforms its unweighted counterpart for all data dimensions and distribution families. While the performance is much better for small values of p, weighting the signs seems to have less and less effect as p grows larger. Assuming a first-order autoregressive (AR(1)) covariance structure, i.e. $\sigma_{ij} = \rho^{|i-j|}$ with $\rho \in (0, 1)$, results in largely similar ARE values as those obtained in table 1 that assume $\Sigma = \mathbf{I}_p$.

2.2 A high-dimensional test of location

It is possible to take an alternative approach to the location testing problem by using the covariance-type U-statistic $C_{n,w} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{X}_{w,i}^T \mathbf{X}_{w,j}$. This class of test statistics are especially attractive since they are readily generalized to cover high-dimensional situations, i.e. when $p > n$. The Chen and Qin (CQ) high-dimensional test of location for multivariate normal $\mathbf{X}_i$ (Chen and Qin, 2010) is a special case of this test that uses the statistic $C_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{X}_i^T \mathbf{X}_j$. The testing procedure of Wang et al. (2015) (from here on referred to as WPL test) shows that one can improve upon the power of the CQ test for non-gaussian elliptical distributions by using spatial signs $\mathbf{S}(\mathbf{X}_i)$ in place of the actual variables.

Under mild regularity conditions in the lines of Wang et al. (2015), the following results hold for our generalized test statistic $C_{n,w}$ under H_0 as $n, p \to \infty$:

$$\frac{C_{n,w}}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(\mathbb{B}_w^2)}} \rightsquigarrow N(0,1), \tag{2.3}$$

and under contiguous alternatives $H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$,

$$\frac{C_{n,w} - \frac{n(n-1)}{2}\boldsymbol{\mu}_0^T \mathbb{A}_w^2 \boldsymbol{\mu}_0 (1 + o(1))}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(\mathbb{B}_w^2)}} \rightsquigarrow N(0,1). \tag{2.4}$$

We provide the details behind deriving these two results in the supplementary material, which extend the results of Wang et al. (2015) in a weightd sign setup using modified regularity conditions.

Following this, the ARE of this test statistic with respect to its unweighted version, i.e. the WPL statistic, is expressed as:

$$ARE(C_{n,w}, \mathrm{WPL}; \boldsymbol{\mu}_0) = \frac{\boldsymbol{\mu}_0^T \mathbb{A}_w^2 \boldsymbol{\mu}_0}{\boldsymbol{\mu}_0^T \mathbb{A}^2 \boldsymbol{\mu}_0} \sqrt{\frac{\mathrm{Tr}(\mathbb{B}^2)}{\mathrm{Tr}(\mathbb{B}_w^2)}} (1 + o(1)),$$

when $\Sigma = k\mathbb{I}_p$, this again simplifies to $\mathbb{E}^2(f(r)/r)/[\mathbb{E}f^2(r).\mathbb{E}^2(1/r)]$. Thus the ARE values are exactly same as those in Table 1, indicating that for large data dimensions the WPL test and that based on $C_{n,w}$ are almost equivalent.

However, in a practical high-dimensional setup sample sizes are often small. Thus, comparing the the two tests with respect to their *finite sample* efficiencies instead gives a better idea of their practical utility. We do so in table 2, which lists empirical powers calculated from 1000 replications of each setup under an AR(1) covariance structure (with $\rho = 0.8$). While under $H_0 : \boldsymbol{\mu} = \mathbf{0}_p$ all tests have similar performance, $C_{n,w}$ beats the other two under deviations from the null.

3 Inverse depth and estimates of scatter

We focus on the estimation of scatter functionals of $\mathbb{F}$ using peripherality-weighted sign vectors in this section. To this end, in addition to conditions (P) and (A1), we assume the following conditions on the peripherality functions:

$\boldsymbol{\mu} = \text{rep}(.15, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.051	0.376	0.418
500	50	0.060	0.832	0.866
1000	20	0.044	0.541	0.584
1000	50	0.039	0.973	0.987
$\boldsymbol{\mu} = \text{rep}(0, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.049	0.061	0.063
500	50	0.039	0.061	0.064
1000	20	0.042	0.060	0.063
1000	50	0.043	0.050	0.050

Table 2: Table of empirical powers of level-0.05 tests for the Chen and Qin (CQ), WPL and $C_{n,w}$ statistics

(B1) The following holds for any $\mathbf{x} \in \mathbb{R}^p$ and $t \in (0, 1)$:

$$P(\mathbf{x}; \mathbb{F}) \leq P(\boldsymbol{\mu}_F + t(\mathbf{x} - \boldsymbol{\mu}_F), \mathbb{F}),$$

(B2) There exists a positive constant $M(P, \mathbb{F})$ such that $P(\mathbf{x}; \mathbb{F}) \to M(P, \mathbb{F})$ as $\|\mathbf{x}\| \to \infty$.

Condition (A2) in Section 2 is now replaced by the stronger condition (B2). Note that all the current conditions are essentially the opposite of those imposed on a traditional depth function (Zuo and Serfling, 2000). Given a depth function, the peripherality functions in this section can be defined as a bounded monotonically decreasing function of it. Consequently, we use the term *inverse depth function* (Majumdar and Chatterjee, 2018) in place of peripherality functions in this section, denoting them by $D^-(\mathbf{x}, \mathbb{F})$, or equivalently $D^-_{\mathbf{X}}(\mathbf{x})$. For ease of exposition, we stick to the following inverse transformation: $D^-_{\mathbf{X}}(\mathbf{x}) = \sup_{\mathbf{y}} D^-_{\mathbf{X}}(\mathbf{y}) - D^-_{\mathbf{X}}(\mathbf{x})$ from now on. The theoretical results go through with minor changes for other choices of inverse depths.

Given a choice of inverse depth function $D^-_{\mathbf{X}}(\mathbf{x})$, we transform the original random variable as: $\tilde{\mathbf{X}} = D^-_{\mathbf{X}}(\mathbf{x})\mathbf{S}(\mathbf{x} - \boldsymbol{\mu})$. We use the transformed data matrix $\tilde{\mathbb{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)^T$ as proxy of the original data $\mathbb{X}$ for improved estimation of the components of the population covariance matrix Σ in this section. To this end, in Section 3.1 we give some results characterizing the population covariance matrix of these vectors, i.e. $\mathbb{V}\tilde{\mathbf{X}}$, and its affine equivariant counterpart. Section 3.2 gives asymptotic properties of their finite sample versions. In Section 3.3, we use eigenvectors of these estimates for robust PCA, deriving their influence functions, as well as asymptotic and finite sample efficiencies. Finally in Section 3.4, we calculate robust estimates of eigenvalues and propose a plug-in robust estimate of Σ.

3.1 Depth Covariance Matrix

Consider the spectral decomposition of Σ: $\Sigma = \Gamma\Lambda\Gamma^T$, Γ being orthogonal and Λ diagonal with positive diagonal elements $\lambda_1 \geq \ldots \geq \lambda_p$. Also normalize the original random

variable as $\mathbf{Z} = \Gamma^T \Lambda^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. In this setup, we can represent the transformed random variable as

$$\begin{aligned}
\tilde{\mathbf{X}} &= D^-_{\mathbf{X}}(\mathbf{X})\mathbf{S}(\mathbf{X} - \boldsymbol{\mu}) \\
&= D^-_{\Gamma\Lambda^{1/2}\mathbf{Z}+\boldsymbol{\mu}}(\Gamma\Lambda^{1/2}\mathbf{Z} + \boldsymbol{\mu}).\mathbf{S}(\Gamma\Lambda^{1/2}\mathbf{Z}) \\
&= \Gamma D^-_{\mathbf{Z}}(\mathbf{Z})\mathbf{S}(\Lambda^{1/2}\mathbf{Z}) \\
&= \Gamma\Lambda^{1/2} D^-_{\mathbf{Z}}(\mathbf{Z})\mathbf{S}(\mathbf{Z})\frac{\|\mathbf{Z}\|}{\|\Lambda^{1/2}\mathbf{Z}\|}. \qquad (3.1)
\end{aligned}$$

$D^-_{\mathbf{Z}}(\cdot)$ is an even function in its argument because of affine invariance, as is $\|\mathbf{z}\|/\|\Lambda^{1/2}\mathbf{z}\|$. Since the sign function $\mathbf{S}(\cdot)$ is odd in its argument, it follows that $\mathbb{E}\tilde{\mathbf{X}} = \mathbf{0}$. Consequently, we obtain an expression for $\tilde{\Sigma} := \mathbb{V}\tilde{\mathbf{X}}$, which we call the *Depth Covariance Matrix* (DCM).

Theorem 3.1. *Let the random variable $\mathbf{X} \in \mathbb{R}^p$ follow an elliptical distribution with center $\boldsymbol{\mu}$ and covariance matrix $\Sigma = \Gamma\Lambda\Gamma^T$, its spectral decomposition. Then, given a depth function $D_{\mathbf{X}}(.)$ the covariance matrix of the transformed random variable $\tilde{\mathbf{X}}$ is*

$$\tilde{\Sigma} = \Gamma\Lambda_D\Gamma^T, \quad \textit{with} \quad \Lambda_D = \mathbb{E}_{\mathbf{Z}}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^2 \frac{\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}}\right], \qquad (3.2)$$

where $\mathbf{Z} = (Z_1, ..., Z_p)^T \sim N(\mathbf{0}, \mathbb{I}_p)$, so that Λ_D a diagonal matrix with diagonal entries

$$\lambda_{D,i} = \mathbb{E}_{\mathbf{Z}}\left[\frac{(D^-_{\mathbf{Z}}(\mathbf{z}))^2\lambda_i z_i^2}{\sum_{j=1}^p \lambda_j z_j^2}\right].$$

The matrix of eigenvectors Γ remains unchanged in the transformation $\mathbf{X} \to \tilde{\mathbf{X}}$. Thus the multivariate rank vectors can be used for robust principal component analysis (which we discuss shortly). However, as seen in (3.2), the diagonal entries of Λ_D do not change if a scale change is done on all entries of Λ, meaning the Λ_D matrices corresponding to $\mathbb{F}$ and $c\mathbb{F}$ for some $c \neq 0$ are same. Because of this reason, the DCM is not equivariant under affine transformations.

We follow the general framework of M-estimation with data-dependent weights (Huber, 1981) to construct an affine equivariant counterpart of the DCM. Specifically, we implicitly define the Affine-equivariant Depth Covariance Matrix (ADCM) as

$$\Sigma_{Dw} = \frac{1}{\mathbb{V}\tilde{Z}_1}\mathbb{E}_{\mathbf{X}}\left[\frac{(D^-_{\mathbf{X}}(\mathbf{x}))^2(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T}{(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}_{Dw}(\mathbf{x} - \boldsymbol{\mu})}\right]. \qquad (3.3)$$

Its affine equivariance follows from the fact that the weights $(D^-_{\mathbf{X}}(\mathbf{x}))^2$ depend only on the standardized quantities $\mathbf{z} \sim \mathcal{E}(\mathbf{0}, k\mathbf{I}_p, g)$. We solve (3.3) iteratively by obtaining a sequence of positive definite matrices $\Sigma^{(k)}_{Dw}$ until convergence:

$$\Sigma^{(k+1)}_{Dw} = \frac{1}{\mathbb{V}\tilde{Z}_1}\mathbb{E}_{\mathbf{X}}\left[\frac{(D^-_{\mathbf{X}}(\mathbf{x}))^2(\Sigma^{(k)}_{Dw})^{1/2}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T(\Sigma^{(k)}_{Da})^{1/2}}{(\mathbf{x} - \boldsymbol{\mu})^T(\Sigma^{(k)}_{Dw})^{-1}(\mathbf{x} - \boldsymbol{\mu})}\right].$$

To ensure existence and uniqueness of this estimator, we consider the class of scatter estimators Σ_M that are obtained as solutions of the following equation:

$$\mathbb{E}_{\mathbf{Z}_M}\left[u(\|\mathbf{z}_M\|)\frac{\mathbf{z}_M\mathbf{z}_M^T}{\|\mathbf{z}_M\|^2} - v(\|\mathbf{z}_M\|)\mathbb{I}_p\right] = 0 \tag{3.4}$$

with $\mathbf{z}_M = \Sigma_M^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$. Under the following assumptions on the scalar valued functions u and v, the above equation produces a unique solution (Huber, 1981):

(C1) The function $u(r)/r^2$ is monotone decreasing, and $u(r) > 0$ for $r > 0$;
(C2) The function $v(r)$ is monotone decreasing, and $v(r) > 0$ for $r > 0$;
(C3) Both $u(r)$ and $v(r)$ are bounded and continuous;
(C4) $u(0)/v(0) < p$;
(C5) For any hyperplane in the sample space $\mathcal{X}$, (i) $P(H) = \mathbb{E}_{\mathbf{X}}\{\mathbb{I}\mathbf{x} \in H\} < 1 - pv(\infty)/u(\infty)$ and (ii) $P(H) \leq 1/p$.

Putting things into context, in our case we have $v(r) = \mathbb{V}\tilde{Z}_1, u(r) = (D^-_{\mathbf{X}}(\mathbf{x}))^2 = (D^-_{\mathbf{Z}}(\mathbf{z}))^2$. Since $v(r)$ is a constant, (C2) and (C3) are trivially satisfied. As for u, most well-known depth functions can be expressed as simple functions of the norm of the standardized random variable. For example, for Halfspace Depth (HSD) we have $D_{\mathbf{Z}}(\mathbf{z}) = (1-G(\|\mathbf{z}\|)$, for Mahalanobis Depth (MhD) we have $D_{\mathbf{Z}}(\mathbf{z}) = (1+\|\mathbf{z}\|^2)^{-1}$ and for Projection Depth (PD) we have $D_{\mathbf{Z}}(\mathbf{z}) = (1+\|\mathbf{z}\|//MAD(G))^{-1}$, where G is the cumulative distribution function of g, and MAD is median absolute deviation. Combined with the choice of inverse transformation, we get

$$u_{HSD}(r) = G^2(r); \quad u_{MhD}(r) = \frac{r^4}{(1+r^2)^2}; \quad u_{PD}(r) = \frac{r^2}{(1+r/MAD(G))^2}.$$

It is easy to verify that the above choices of u satisfy (C1) and (C3). To check (C4) and (C5), notice that $\mathbf{Z}$ has a spherically symmetric distribution, so that its euclidean norm and spatial sign are independent. Since $D_{\mathbf{Z}}(\mathbf{z})$ depends only on $\|\mathbf{z}\|$, we now have

$$\mathbb{V}\tilde{Z}_1 = \mathbb{V}\left(D^-_{\mathbf{Z}}(\mathbf{Z})\frac{Z_1}{\|\mathbf{Z}\|}\right) = \mathbb{V}(D^-_{\mathbf{Z}}(\mathbf{Z}))\mathbb{V}(S_1(\mathbf{Z})) = \frac{1}{p}\mathbb{V}(D^-_{\mathbf{Z}}(\mathbf{Z})),$$

as $\mathbb{V}(\mathbf{S}(\mathbf{Z})) = \mathbb{V}((S_1(\mathbf{Z}), S_2(\mathbf{Z}), ..., S_p(\mathbf{Z}))^T) = \mathbb{I}_p/p$. Now for MhD and PD $u(\infty) = 1, u(0) = 0$, so (C4) and (C5) are immediate. To achieve this for HSD, we only need to replace $u_{HSD}(r)$ with $u_{HSD*}(r) = G^2(r) - 1/4$.

3.2 Calculating the sample DCM and ADCM

Let us denote $\mathbb{S}(\mathbf{x};\boldsymbol{\mu}) = \mathbf{S}(\mathbf{x}-\boldsymbol{\mu})\mathbf{S}(\mathbf{x}-\boldsymbol{\mu})^T$. Then, given the depth function and *known* location center $\boldsymbol{\mu}$, one can show that the vectorized form of $\sqrt{n}$-times the sample DCM, i.e. $n^{-1/2}\sum_{i=1}^n (D^-_{\mathbf{X}}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})$ has an asymptotic multivariate normal distribution with mean $\sqrt{n}.\,\mathrm{vec}(\mathbb{E}[((D^-_{\mathbf{X}}(\mathbf{X}))^2\mathbb{S}(\mathbf{x};\boldsymbol{\mu})])$ and some covariance matrix by straightforward application of the central limit theorem (CLT). But in practice the depth at any point $\mathbf{x}$ is estimated by the sample depth, i.e. depth calculated using empirical distribution function

in place of $\mathbb{F}$. We also need to replace the known location parameter $\boldsymbol{\mu}$ by some estimator $\hat{\boldsymbol{\mu}}_n$.

Denote this sample depth by $D^n_{\mathbf{X}}(\mathbf{x})$, and the inverse depth by $D^{n-}_{\mathbf{X}}(\mathbf{x})$. Here we make the following assumption regarding the sample depths:

(D1) *Uniform convergence*: $\sup_{\mathbf{x}\in\mathbb{R}^p} |D^n_{\mathbf{X}}(\mathbf{x}) - D_{\mathbf{X}}(\mathbf{x})| \to 0$ as $n \to \infty$.

The above holds under very mild conditions for several depth functions, e.g. projection depth (Zuo, 2003) and simplicial depth (Dümbgen, 1992). For $\hat{\boldsymbol{\mu}}_n$ we use a robust estimate of location, examples of which include the spatial median (Brown, 1983; Haldane, 1948), Oja median (Oja, 1983) and projection median (Zuo, 2003).

To plug in the above estimates to the sample DCM and go through with the CLT, we need the following result.

Lemma 3.2. *Consider a random variable $\mathbf{X} \in \mathbb{R}^p$ having a continuous and symmetric distribution with location center $\boldsymbol{\mu}$ such that $\mathbb{E}\|\mathbf{x} - \boldsymbol{\mu}\|^{-3/2} < \infty$. For a size-$n$ random sample from this distribution, suppose $\hat{\boldsymbol{\mu}}_n$ is an estimator of $\boldsymbol{\mu}$ so that $\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} = O_P(n^{-1/2})$. Then, given assumption (D1) we have*

$$\frac{1}{n}\sum_{i=1}^{n}(D^{n-}_{\mathbf{X}}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) = \frac{1}{n}\sum_{i=1}^{n}(D^{-}_{\mathbf{X}}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu}) + \mathbb{R}_n,$$

where $\|\mathbb{R}_n\|_\infty = o_P(n^{-1/2})$.

We are now in a position to state the result for consistency of the sample DCM:

Theorem 3.3. *Consider a size-n random sample from the distribution in Lemma 3.2. Then, given a depth function $D_{\mathbf{X}}(.)$ and an estimate of center $\hat{\boldsymbol{\mu}}_n$ so that $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = O_P(1)$,*

$$\text{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D^{n-}_{\mathbf{X}}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right\} - \mathbb{E}[\text{vec}\left\{(D^{-}_{\mathbf{X}}(\mathbf{x}))^2\mathbb{S}(\mathbf{x};\boldsymbol{\mu})\right\}] \rightsquigarrow \mathcal{N}_{p^2}(\mathbf{0}, \mathbb{V}_D(\mathbb{F})) + \mathbf{R}_n,$$

with $\mathbb{V}_D(\mathbb{F}) = \mathbb{V}[\text{vec}\{(D^{-}_{\mathbf{X}}(\mathbf{x}))^2\mathbb{S}(\mathbf{x};\boldsymbol{\mu})\}]$ and $\|\mathbf{R}_n\|_\infty = o_P(n^{-1/2})$.

When $\mathbb{F}$ is elliptical, an elaborate form of the covariance matrix $\mathbb{V}_D(\mathbb{F})$ explicitly specifying each of its elements (more directly those of its Γ^T-rotated version) can be obtained (see Appendix A). This form is useful when deriving limiting distributions of eigenvectors and eigenvalues of the sample DCM.

For ADCM, we simultaneously estimate its sample version and the mean vector $\boldsymbol{\mu}$ by solving for the location and scatter functionals $(\boldsymbol{\mu}_{Dw}, \Sigma_{Dw})$:

$$\mathbb{E}\left[\frac{\Sigma^{-1/2}_{Dw}(\mathbf{x} - \boldsymbol{\mu}_{Dw})}{\|\Sigma^{-1/2}_{Dw}(\mathbf{x} - \boldsymbol{\mu}_{Dw})\|}\right] = \mathbf{0}_p, \tag{3.5}$$

$$\frac{1}{\mathbb{V}\tilde{Z}_1}\mathbb{E}\left[\frac{(D^{-}_{\mathbf{X}}(\mathbf{x}))^2\Sigma^{-1/2}_{Dw}(\mathbf{x} - \boldsymbol{\mu}_{Dw})(\mathbf{x} - \boldsymbol{\mu}_{Dw})^T\Sigma^{-1/2}_{Dw}}{(\mathbf{x} - \boldsymbol{\mu}_{Dw})^T\Sigma^{-1}_{Dw}(\mathbf{x} - \boldsymbol{\mu}_{Dw})}\right] = \mathbb{I}_p. \tag{3.6}$$

In the M-estimation framework of Huber (1981), for any fixed Σ_M there exists a unique and fixed solution of the general location problem $\mathbb{E}_{\mathbf{Z}_M}(w(\|\mathbf{z}_M\|\mathbf{z}_M) = \mathbf{0}_p$ under the following condition:

(D2) The function $w(r)r$ is monotone increasing for $r > 0$.

This is easy to verify for our choice of the weights: $w(\|\mathbf{z}_M\|) = D^-_{\mathbf{Z}_M}(\mathbf{z}_M)/\|\mathbf{z}_M\|$. Uniqueness of simultaneous fixed point solutions of (3.5) and (3.6) is guaranteed when $\mathbf{X}$ has a symmetric distribution (Huber, 1981), which we also satisfy since $\mathbf{X} \sim \mathbb{F}$, an elliptical distribution,

In practice, it is difficult to calculate the scale multiple $\mathbb{V}\tilde{Z}_1$ analytically for known depth functions and an arbitrary $\mathbb{F}$. We bypass this by instead obtaining a standardized version of the scatter functional: $\Sigma^*_{Dw} = \Sigma_{Dw}/\mathbb{V}\tilde{Z}_1$ (and $\boldsymbol{\mu}_{Dw}$) using the following iterative algorithm:

1. Start from some initial estimates $(\boldsymbol{\mu}^{(0)}_{Dw}, \Sigma^{(0)}_{Dw})$. Set $t = 0$;
2. Calculate the standardized observations $\mathbf{z}^{(t)}_i = (\Sigma^{(t)}_{Dw})^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}^{(t)}_{Dw})$;
3. Update the location estimate:
$$\boldsymbol{\mu}^{(t+1)}_{Dw} = \frac{\sum_{i=1}^n \tilde{\mathbf{x}}_i/\|\mathbf{z}^{(t)}_i\|}{\sum_{i=1}^n \|\mathbf{z}^{(t)}_i\|^{-1}};$$
4. Update the scatter estimate:
$$\Sigma^{*(t+1)}_{Dw} = \frac{1}{n}\sum_{i=1}^n \frac{(D^{n-}_{\mathbf{X}}(\mathbf{x}_i))^2(\mathbf{x}_i - \boldsymbol{\mu}^{(t+1)}_{Dw})(\mathbf{x}_i - \boldsymbol{\mu}^{(t+1)}_{Dw})^T}{\|\mathbf{z}^{(t)}_i\|^2}; \quad \Sigma^{*(t+1)}_{Dw} \leftarrow \frac{\Sigma^{*(t+1)}_{Dw}}{\det(\Sigma^{*(t+1)}_{Dw})^{1/p}}.$$
5. Continue until convergence.

3.3 Robust PCA using eigenvectors

We are mainly interested in using the DCM for robust principal component analysis (PCA). We assume that the eigenvalues of Σ are distinct: $\lambda_1 > \lambda_2 > ... > \lambda_p$, for ease of obtaining asymptotic distributions of eigenvectors of the DCM (Appendix B). Limiting distributions for the case of non-unique eigenvalues can be obtained following Magyar and Tyler (2014).

3.3.1 Influence functions

To demonstrate the robustness of eigenvector estimates of the DCM and ADCM, we derive their influence functions. Influence functions quantify how much influence a sample point, especially an infinitesimal contamination, has on any functional of a probability

distribution (Hampel et al., 1986). Given any probability distribution $\mathbb{H} \in \mathcal{M}$, the influence function of any point $\mathbf{x}_0 \in \mathcal{X}$ for some functional $T(\mathbb{H})$ on the distribution is defined as

$$IF(\mathbf{x}_0; T, \mathbb{H}) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(T(\mathbb{H}_\epsilon) - T(\mathbb{H}))$$

where $\mathbb{H}_\epsilon$ is $\mathbb{H}$ with an additional mass of ϵ at $\mathbf{x}_0$, i.e. $\mathbb{H}_\epsilon = (1-\epsilon)\mathbb{H} + \epsilon\Delta_{\mathbf{x}_0}$; $\Delta_{\mathbf{x}_0}$ being the distribution with point mass at $\mathbf{x}_0$. When $T(\mathbb{H}) = E_{\mathbb{H}} f$ for some $\mathbb{H}$-integrable function f, $IF(\mathbf{x}_0; T, \mathbb{H}) = f(\mathbf{x}_0) - T(\mathbb{H})$.

It now follows that for the DCM,

$$IF(\mathbf{x}_0; \tilde{\Sigma}, \mathbb{F}) = (D^-_{\mathbf{X}}(\mathbf{x}_0))^2 \mathbb{S}(\mathbf{x}_0; \boldsymbol{\mu}) - \tilde{\Sigma}.$$

Following Croux and Haesbroeck (2000), we get the influence function of the i^{th} eigenvector of $\tilde{\Sigma}$, say $\boldsymbol{\gamma}_D = (\gamma_{D,1}, ..., \gamma_{D,p})^T$ for $1 \le i \le p$:

$$\begin{aligned}
IF(\mathbf{x}_0; \boldsymbol{\gamma}_{D,i}, \mathbb{F}) &= \sum_{k=1;k\neq i}^{p} \frac{1}{\lambda_{D,i} - \lambda_{D,k}} \left\{ \boldsymbol{\gamma}_k^T IF(\mathbf{x}_0; \tilde{\Sigma}, \boldsymbol{\gamma}_i) \right\} \boldsymbol{\gamma}_k \\
&= \sum_{k=1;k\neq i}^{p} \frac{1}{\lambda_{D,i} - \lambda_{D,k}} \left\{ \boldsymbol{\gamma}_k^T (D^-_{\mathbf{X}}(\mathbf{x}_0))^2 \mathbb{S}(\mathbf{x}_0; \boldsymbol{\mu}) \boldsymbol{\gamma}_i - \lambda_{D,i} \boldsymbol{\gamma}_k^T \boldsymbol{\gamma}_i \right\} \boldsymbol{\gamma}_k \\
&= \sum_{k=1;k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k} z_{0i} z_{0k}}{\lambda_{D,i} - \lambda_{D,k}} . \frac{(D^-_{\mathbf{Z}}(\mathbf{z}_0))^2}{\mathbf{z}_0^T \Lambda \mathbf{z}_0} \boldsymbol{\gamma}_k \\
&= \sum_{k=1;k\neq i}^{p} \frac{(D^-_{\mathbf{Z}}(\mathbf{z}_0))^2 \mathbb{S}_{ik}(\Lambda^{1/2}\mathbf{z}_0; \mathbf{0})}{\lambda_{D,i} - \lambda_{D,k}} \boldsymbol{\gamma}_k, \qquad (3.7)
\end{aligned}$$

where $\Gamma^T \Lambda^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu}) = \mathbf{z}_0 = (z_{01}, ..., z_{0p})^T$. Clearly this influence function is bounded, indicating good robustness properties of the principal component estimates.

In Figure 2 we consider first eigenvectors of the DCM and two other well-known robust estimates of scatter: the Sign Covariance Matrix (SCM) and Tyler's shape matrix (Tyler, 1987). We set $\mathbb{F} \equiv \mathcal{N}_2(\mathbf{0}, \text{diag}(2, 1))$ and plot norms of the eigenvector influence functions for different values of $\mathbf{x}_0$. Influence function for the i^{th} eigenvectors of these two matrices (say $\boldsymbol{\gamma}_{S,i}$ and $\boldsymbol{\gamma}_{T,i}$, respectively) are as follows:

$$IF(\mathbf{x}_0; \boldsymbol{\gamma}_{S,i}, \mathbb{F}) = \sum_{k=1;k\neq i}^{p} \frac{\mathbb{S}_{ik}(\Lambda^{1/2}\mathbf{z}_0, \mathbf{0})}{\lambda_{S,i} - \lambda_{S,k}} \boldsymbol{\gamma}_k; \quad \lambda_{S,i} = \mathbb{E}_{\mathbf{Z}} \left(\frac{\lambda_i z_i^2}{\sum_{j=1}^{p} \lambda_j z_j^2} \right), \qquad (3.8)$$

$$IF(\mathbf{x}_0; \boldsymbol{\gamma}_{T,i}, \mathbb{F}) = (p+2) \sum_{k=1;k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k}}{\lambda_i - \lambda_k} \mathbb{S}_{ik}(\mathbf{z}_0; \mathbf{0}) \boldsymbol{\gamma}_k. \qquad (3.9)$$

Their corresponding plots (panels (b) and (c) in Figure 2) demonstrate an 'inlier effect', i.e. points close to symmetry center and the center itself having high influence: which results in loss of efficiency. The influence function for eigenvector estimates of the sample covariance matrix is obtained by replacing $(p+2)$ by $\|\mathbf{z}_0\|^2$ in (3.9), hence is unbounded and makes

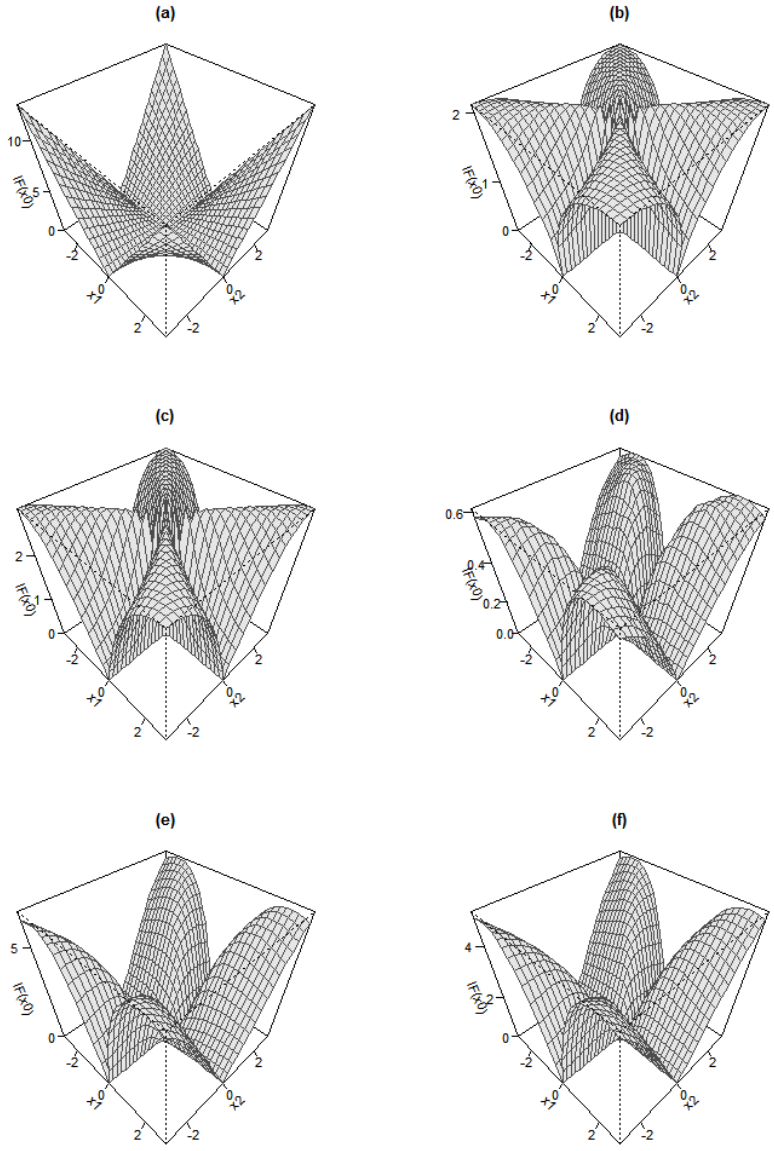


Figure 2: Plot of the norm of influence function for first eigenvector of (a) sample covariance matrix, (b) SCM, (c) Tyler's scatter matrix and DCMs for (d) Halfspace depth, (e) Mahalanobis depth, (f) Projection depth for a bivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}, \Sigma = \text{diag}(2, 1)$

the corresponding estimates non-robust. In comparison, all three DCMs considered have bounded influence functions *and* small values of the influence function at 'deep' points.

For the ADCM, we first notice that the influence function of any affine equivariant estimate of scatter can be expressed as

$$IF(\mathbf{x}_0, C, \mathbb{F}) = \alpha_C(\|\mathbf{z}_0\|)\mathbb{S}(\mathbf{z}_0; \mathbf{0}) - \beta_C(\|\mathbf{z}_0\|)C$$

for scalar valued functions α_C, β_C (Hampel et al., 1986). Following this, the influence function of an eigenvector $\boldsymbol{\gamma}_{C,i}$ of C is derived:

$$IF(\mathbf{x}_0, \boldsymbol{\gamma}_{C,i}, \mathbb{F}) = \alpha_C(\|\mathbf{z}_0\|) \sum_{k=1, k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k}}{\lambda_i - \lambda_k} . \mathbb{S}_{ik}(\mathbf{z}_0; \mathbf{0}) \boldsymbol{\gamma}_k.$$

When $C = \Sigma_M$, i.e. the solution to (3.3), Huber (1981) shows that

$$\alpha_C(\|\mathbf{z}_0\|) = \frac{p(p+2)u(\|\mathbf{z}_0\|)}{\mathbb{E}_{\mathbf{Z}}\left[pu(\|\mathbf{z}\|) + u'(\|\mathbf{z}\|)\|\mathbf{z}\|\right]}.$$

In our case $u(\|\mathbf{z}_0\|) = (D^-_{\mathbf{Z}}(\mathbf{z}_0))^2$, thus the influence function of eigenvectors of the ADCM is a bounded increasing function of $\|\mathbf{z}_0\|$.

3.3.2 Asymptotic and finite-sample efficiencies

The asymptotic variance of the i^{th} eigenvector of the sample covariance matrix, say $\hat{\boldsymbol{\gamma}}_i$, is (Anderson, 2003):

$$A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_i) = \sum_{k=1; k\neq i}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T; \quad 1 \leq i \leq p. \tag{3.10}$$

Asymptotic relative efficiencies of eigenvectors from the sample DCM with respect to the sample covariance matrix can now be derived using (3.10) and (B.2) from Corollary B.2 from Appendix B:

$$\begin{aligned} ARE(\hat{\boldsymbol{\gamma}}_{D,i}, \hat{\boldsymbol{\gamma}}_i; \mathbb{F}) &= \frac{\text{Tr}(A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_i))}{\text{Tr}(A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_{D,i}))} \\ &= \left[\sum_{k=1; k\neq i}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2}\right] \left[\sum_{k=1; k\neq i}^{p} \frac{\mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4 (\mathbb{S}_{ik}(\Lambda^{1/2}\mathbf{z}; \mathbf{0}))^2]}{(\lambda_{D,i} - \lambda_{D,k})^2}\right]^{-1}. \end{aligned}$$

Obtaining ARE of the ADCM is, in comparison to DCM, more straightforward. The asymptotic covariance matrix of an eigenvector of the affine equivariant scatter functional C is given by:

$$A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_{C,j}) = A\mathbb{V}(C_{12}, \mathbb{F}_0) \sum_{k=1, k\neq i}^{p} \frac{\lambda_i \lambda_k}{\lambda_i - \lambda_k} . \boldsymbol{\gamma}_i \boldsymbol{\gamma}_k^T,$$

Distribution	PD-ACM				HSD-ACM			
	$p=2$	$p=5$	$p=10$	$p=20$	$p=2$	$p=5$	$p=10$	$p=20$
t_5	4.73	3.99	3.46	3.26	4.18	3.63	3.36	3.15
t_6	2.97	3.28	2.49	2.36	2.59	2.45	2.37	2.32
t_{10}	1.45	1.47	1.49	1.52	1.30	1.37	1.43	1.49
t_{15}	1.15	1.19	1.23	1.27	1.01	1.10	1.17	1.24
t_{25}	0.97	1.02	1.07	1.11	0.85	0.94	1.02	1.08
MVN	0.77	0.84	0.89	0.93	0.68	0.77	0.84	0.91

Table 3: Table of AREs of the ADCM for different choices of p and data-generating distributions, and two choices of depth functions

where $A\mathbb{V}(C_{12}, \mathbb{F}_0)$ is the asymptotic variance of an off-diagonal element of C when the underlying distribution is $\mathbb{F}_0 \equiv \mathcal{E}(\mathbf{0}, \mathbb{I}_p, g)$. Following Croux and Haesbroeck (2000) this equals

$$A\mathbb{V}(C_{12}, \mathbb{F}_0) = \mathbb{E}_{\mathbf{Z}}\left[(\alpha_C(\|\mathbf{z}\|)\mathbb{S}_{12}(\mathbf{z};\mathbf{0}))^2\right] = \mathbb{E}_{\mathbf{Z}}\alpha_C(\|\mathbf{z}\|)^2.\mathbb{E}_{\mathbf{Z}}(\mathbb{S}_{12}(\mathbf{z};\mathbf{0}))^2,$$

using the fact that $\|\mathbf{Z}\|$ and $\mathbf{S}(\mathbf{Z})$ are independent with $\mathbf{Z} \sim \mathbb{F}_0$. It now follows that

$$ARE(\hat{\boldsymbol{\gamma}}_{\Sigma_M,i}, \hat{\boldsymbol{\gamma}}_i; \mathbb{F}) = \frac{1}{A\mathbb{V}(C_{12}, \mathbb{F}_0)} = \frac{\left[\mathbb{E}_{\mathbf{Z}}(pu\|\mathbf{z}\| + u'(\|\mathbf{z}\|)\|\mathbf{z}\|)\right]^2}{p^2(p+2)^2\mathbb{E}_{\mathbf{Z}}(u(\|\mathbf{z}\|)^2\mathbb{E}_{\mathbf{Z}}(\mathbb{S}_{12}(\mathbf{z};\mathbf{0}))^2}. \tag{3.11}$$

Table 3 considers 6 different elliptic distributions (namely, p-variate t with df = 5, 6, 10, 15, 25 and multivariate normal) and summarizes ARE of the first eigenvectors for ADCMs corresponding to projection depth (PD-ACM) and halfspace depth (HSD-ACM). Due to difficulty in analytically obtaining the AREs, we calculate them using Monte-Carlo simulation of 10^6 samples and subsequent numerical integration. The ADCM seems to be particularly efficient in lower dimensions for distributions with heavier tails (t_5 and t_6), while for distributions with lighter tails, the AREs increase with data dimension. At higher values of p the ADCM is almost as efficient as the sample covarnace matrix when the data comes from multivariate normal distribution.

We also obtain finite sample efficiencies of the three DCMs and ADCMs by a simulation study for $p = 4$, and compare them with the same from the SCM and Tyler's scatter matrix. We consider the same 6 elliptical distributions considered in ARE calculations above, and from every distribution draw 10000 samples each for sample sizes $n = 20, 50, 100, 300, 500$. All distributions are centered at $\mathbf{0}_p$, and have covariance matrix $\Sigma = \text{diag}(4, 3, 2, 1)$.

We use the concept of principal angles (Miao and Ben-Israel, 1992) to find out error estimates for the first eigenvector of a scatter matrix. In our case, the first eigenvector is

$$\boldsymbol{\gamma}_1 = (1, \overbrace{0, ..., 0}^{p-1})^T.$$

We measure the prediction error for an eigenvector estimate by some method, say $\hat{\boldsymbol{\gamma}}_{E,1}$, by the smallest angle between the true and predicted vectors, i.e. $\cos^{-1}|\hat{\boldsymbol{\gamma}}_{E,1}^T\hat{\boldsymbol{\gamma}}_1|$. A small absolute value of this angle means to better prediction. We repeat this 10000 times and

calculate the **Mean Squared Prediction Angle**:

$$MSPA(\hat{\gamma}_{E,1}) = \frac{1}{10000}\sum_{m=1}^{10000}\left(\cos^{-1}\left|\gamma_1^T\hat{\gamma}_{E,1}^{(m)}\right|\right)^2.$$

The finite sample efficiency of this eigenvector estimate relative to that from the sample covariance matrix, i.e. $\hat{\gamma}_1$ is obtained as:

$$FSE(\hat{\gamma}_{E,1}, \hat{\gamma}_1) = \frac{MSPA(\hat{\gamma}_1)}{MSPA(\hat{\gamma}_{E,1})}.$$

As seen in Table 4 for $p = 4$, DCM- and ADCM-based estimators (columns 3-5) outperform SCM and Tyler's M-estimate of scatter. Among the 3 depth functions considered, projection depth gives the best results. Its finite sample performances are better than Tyler's and Huber's M-estimators of scatter, as well as their symmetrized counterparts that are much more computationally intensive (see Table 4 in Sirkiä et al. (2007)). The affine equivariant versions of all DCMs (columns 6-8) further improve their performance.

3.4 Robust estimation of eigenvalues, and a plug-in estimator of Σ

As seen in Theorem 3.1, eigenvalues of the DCM are not same as the population eigenvalues, whereas the ADCM only gives back standardized eigenvalues. However, it is possible to robustly estimate the original eigenvalues by working with the individual columns of the robust score matrix. We do so using the following steps:

1. Randomly divide the sample indices $\{1, 2, ..., n\}$ into k disjoint groups $\{G_1, ..., G_k\}$ of size $\lfloor n/k \rfloor$ each;
2. Assume the data is centered. Transform the data matrix: $\mathbb{S} = \hat{\Gamma}_D^T\mathbb{X}$;
3. Calculate coordinate-wise variances for each group of indices G_j:

$$\hat{\lambda}_{i,j} = \frac{1}{|G_j|}\sum_{l\in G_j}(s_{li} - \bar{s}_{G_j,i})^2; \quad i = 1, ..., p; j = 1, ..., k$$

   where $\bar{\mathbf{s}}_{G_j} = (\bar{s}_{G_j,1}, ..., \bar{s}_{G_j,p})^T$ is the vector of column-wise means of $\mathbb{S}_{G_j}$, the submatrix of $\mathbb{S}$ with row indices in G_j.
4. Obtain estimates of eigenvalues by taking coordinate-wise medians of these variances:

$$\hat{\lambda}_i = \text{median}(\hat{\lambda}_{i,1}, ..., \hat{\lambda}_{i,k}); \quad i = 1, ..., p$$

The number of subgroups used to calculate this median-of-small-variances estimator can be determined following Minsker (2015). After this, we construct a consistent plug-in estimator of the population covariance matrix Σ:

Theorem 3.4. *Consider the estimates $\hat{\lambda}_i$ obtained from the above algorithm, and the matrix of eigenvectors $\hat{\Gamma}_D$ estimated using the sample DCM. Define $\hat{\Sigma} = \hat{\Gamma}_D\hat{\Lambda}\hat{\Gamma}_D^T; \hat{\Lambda} = \text{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_p)$. Then as $n \to \infty$,*

$$\|\hat{\Sigma} - \Sigma\|_F \xrightarrow{P} 0.$$

4-variate t_5	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	1.04	1.02	1.10	1.07	1.02	1.09	1.07	0.98
n=50	1.08	1.08	1.16	1.16	1.13	1.19	1.19	1.13
n=100	1.31	1.31	1.42	1.38	1.36	1.46	1.44	1.36
n=300	1.46	1.54	1.81	1.76	1.95	1.88	1.88	1.95
n=500	1.92	1.93	2.23	2.03	2.31	2.35	2.19	2.39
4-variate t_6	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	1.00	1.05	1.03	1.05	1.00	1.04	1.04	0.95
n=50	1.03	1.01	1.13	1.12	1.11	1.19	1.17	1.10
n=100	1.08	1.12	1.25	1.23	1.27	1.24	1.25	1.22
n=300	1.34	1.36	1.64	1.52	1.60	1.67	1.61	1.68
n=500	1.26	1.34	1.55	1.49	1.60	1.65	1.61	1.69
4-variate t_{10}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.90	0.89	0.95	0.98	0.98	0.96	1.01	0.95
n=50	0.90	0.91	1.01	0.98	0.98	1.03	1.04	0.99
n=100	0.87	0.87	0.93	0.95	1.01	0.99	1.01	1.05
n=300	0.87	0.87	1.09	1.09	1.17	1.14	1.16	1.23
n=500	0.88	0.92	1.10	1.10	1.23	1.19	1.22	1.29
4-variate t_{15}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.92	0.90	0.94	0.94	0.96	0.95	0.97	0.89
n=50	0.82	0.83	0.88	0.91	0.93	0.88	0.93	0.93
n=100	0.84	0.87	0.92	0.95	1.00	0.93	0.96	1.00
n=300	0.73	0.75	0.96	0.99	1.10	1.00	1.06	1.12
n=500	0.73	0.76	0.95	0.96	1.06	0.94	0.97	1.06
4-variate t_{25}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.89	0.92	0.92	0.92	0.90	0.96	0.95	0.89
n=50	0.82	0.84	0.89	0.90	0.91	0.93	0.96	0.92
n=100	0.77	0.76	0.90	0.90	0.96	0.94	0.98	1.04
n=300	0.73	0.77	0.93	0.91	0.98	1.00	0.98	1.03
n=500	0.67	0.71	0.83	0.83	0.96	0.88	0.90	1.00
4-variate Normal	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.82	0.84	0.87	0.90	0.91	0.89	0.93	0.89
n=50	0.80	0.81	0.87	0.88	0.88	0.88	0.92	0.88
n=100	0.68	0.71	0.80	0.85	0.91	0.82	0.86	0.92
n=300	0.61	0.63	0.82	0.85	0.93	0.86	0.91	0.96
n=500	0.60	0.64	0.77	0.80	0.90	0.82	0.86	0.96

Table 4: Finite sample efficiencies of several scatter matrices: $p = 4$. H, M or P after DCM or ADCM indicates the depth function used: H = halfspace depth, M = Mahalanobis depth, P = projection depth.

4 Robust PCA and supervised models

In the presence of a vector of univariate responses, say $\mathbf{y} = (y_1, \ldots, y_n)^T$, each y_i being i.i.d. relalizations of a univariate random variable Y that takes values in $\mathcal{Y} \subseteq \mathbb{R}$, there is substantial literature devoted to utilizing the subspace generated by the basis of $\Sigma = \mathbb{V}\mathbf{X}$ in modelling $\mathbb{E}(Y|\mathbf{X})$. This ranges from the simple Principal Components Regression (PCR) to Partial Least Squares (PLS) and Envelope methods (Cook et al., 2010). In this section, we concentrate on robust inference using Sufficient Dimension Reduction (SDR) (Adragni and Cook, 2009), mainly because it provides a general framework for reducing dimensionality of data directly using top eigenvectors of Σ (albeit in a different manner than PCR) or an appropriate affine transformation of it.

SDR attempts to find out a linear transformation $\mathbf{R}$ on $\mathbf{X}$ such that $\mathbb{E}(Y|\mathbf{X}) = \mathbb{E}(Y|\mathbb{R}(\mathbf{X}))$. Assuming that $\mathbf{R}(\mathbf{X})$ takes values in $\mathbb{R}^d, d \leq \min(n, p)$, this can be achieved

through an inverse regression model:

$$\mathbf{X}_y = \boldsymbol{\mu} + \Gamma_d \mathbf{v}_y + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{X}_y = \mathbf{X}|Y = y$, $\Gamma_d \in \mathbb{R}^{p\times d}$ is a semi-orthogonal basis for $\mathcal{S}_\Gamma$, the spanning subspace of $\{\mathbb{E}\mathbf{X}_y - \boldsymbol{\mu} | y \in \mathcal{Y}\}$, $\mathbf{v}_y = (\Gamma_d^T\Gamma_d)^{-1}\Gamma_d^T(\mathbb{E}\mathbf{X}_y - \boldsymbol{\mu}) \in \mathbb{R}^d$, and $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \Delta)$ for some positive-definite matix Δ. This formulation is straightforward to implement when Y is categorical, while for continuous responses, the vector $\mathbf{y}$ is divided into a number of slices.

Under this model the minimal sufficient transformation is $\mathbf{R}(\mathbf{X}) = \Gamma_d^T\Delta^{-1}\mathbf{X}$ (Adragni and Cook, 2009). The simplest case of this model is when $\Delta = \sigma^2\mathbb{I}_p$, for which the maximum likelihood estimator of $\mathbf{R}(\mathbf{X})$ turns out to be the first d PCs of Σ. Taking $\hat{\boldsymbol{\mu}}_y = \sum_{i=1}^n \mathbf{x}_i\mathbb{I}\{y_i = y\}/n$ and $\hat{\boldsymbol{\mu}} = \sum_{i=1}^n \mathbf{x}_i/n$, one can now estimate σ^2 as: $\hat{\sigma}^2 = \sum_{i=1}^p s_i/p$, where s_i is the i^{th} diagonal element of $\mathbb{C}_y^T\mathbb{C}_y/n$, with

$$\mathbb{C}_y = (\mathbf{c}_{y1}, \ldots, \mathbf{c}_{yn_y})^T; \quad n_y = \sum_{i=1}^n \mathbb{I}\{y_i = y\}, \mathbf{c}_{yi} = \mathbf{x}_{yi} - \hat{\boldsymbol{\mu}} - \hat{\Gamma}_d(\hat{\Gamma}_d^T\hat{\Gamma}_d)^{-1}\hat{\Gamma}_d^T(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}).$$

Following this, predictions for a new observation $\mathbf{x}$ is obtained as a weighted sum of the responses:

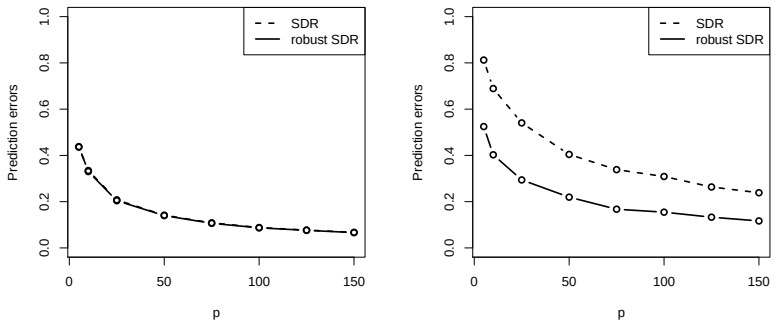
$$\hat{\mathbb{E}}(Y|\mathbf{X} = \mathbf{x}) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}; \quad w_i = \exp\left[-\frac{1}{\hat{\sigma}^2}\|\hat{\Gamma}_d^T(\mathbf{x} - \mathbf{x}_i)\|^2\right].$$

We formulate a robust version of the above procedure by estimating the quantities $\Gamma_d, \boldsymbol{\mu}, \boldsymbol{\mu}_y, \sigma^2$ by robust methods. Specifically, we take:

- $\tilde{\Gamma}_d$ = first d eigenvectors of the sample DCM;
- $\tilde{\boldsymbol{\mu}}$ = weighted spatial median of the rows of $\mathbb{X}$;
- $\tilde{\boldsymbol{\mu}}_y$ = weighted spatial median of the rows of $\mathbb{X}_y$, for each $y \in \mathcal{Y}$;
- $\tilde{\sigma}^2 = \sum_{i=1}^p \tilde{s}_i/p$, where $\tilde{s}_i$ is the i^{th} diagonal element of $\tilde{\mathbb{C}}_y^T\tilde{\mathbb{C}}_y/n$, $\tilde{\mathbb{C}}_y$ being the matrix composed of rows $\mathbf{x}_{yi} - \tilde{\boldsymbol{\mu}} - \tilde{\Gamma}_d(\tilde{\Gamma}_d^T\tilde{\Gamma}_d)^{-1}\tilde{\Gamma}_d^T(\tilde{\boldsymbol{\mu}}_y - \tilde{\boldsymbol{\mu}})$ for $i \in \{1, \ldots, n\}$ such that $y_i = y$.

We compare the performance of our robust SDR with the original method of Adragni and Cook (2009) with or without the presence of bad leverage points in Σ using a simulation study. For a given value of p, we take $n = 200, d = 1$, generate the responses $y_1, \ldots, y_n$ as independent standard normal, and the predictors as $\mathbf{X}_y = \mathbf{1}(y + y^2 + y^3) + \boldsymbol{\epsilon}$, with $\mathbf{1} \in \mathbb{R}^p$ consisting of all 1's, and $\mathbb{V}(\boldsymbol{\epsilon}) = 25\mathbb{I}_p$. We measure performances of the SDR models by their mean squared prediction error on another set of 200 observations $(\mathbf{y}^*, \mathbb{X}^*)$ generated similarly, and taking the average of these errors on 100 such training-test pairs of datasets. Finally we repeat the full setup for different choices of $p = 5, 10, 25, 50, 75, 100, 125, 150$.

Panel (a) of figure 3 compares prediction errors using robust and maximum likelihood SDR estimates when $\mathbb{X}$ contains no outliers: here the two methods are virtually indistinguishable. We now introduce outliers in each of the 100 datasets by adding 100 to first $p/5$ coordinates of the first 10 observations in $\mathbb{X}$, and repeat the analysis. Panel (b) of the figure shows that the robust SDR method remains more accurate in predicting out of sample observations for all values of p than standard SDR.



(a) (b)

Figure 3: Average prediction errors for two methods of SDR (a) in absence and (b) in presence of outliers

5 Robust inference with functional data

Detection of anomalous observations is of importance in real-life problems involving functional data analysis, and functional PCA is a widely used tool in this setting. In this section we utilize robust principal components from the DCM for this purpose. We use the approach of Boente and Salibian-Barrera (2015) for performing robust PCA on functional data using the estimated PCs from the DCM. Here we have a data matrix $\mathbb{H}$, that stores the values of a set of n curves, say $\mathcal{F} = \{f_1, \ldots, f_n\} \in L^2[0,1]$, each observed at a set of common design points $\{t_1, ..., t_m\}$. We model each of these functions as a linear combination of p mutually orthogonal B-spline basis functions $\mathcal{D} = \{\delta_1, ..., \delta_p\}$. Following this, we map data for each of the functions onto the coordinate system formed by the spline basis:

$$T(\mathbb{H}; \mathcal{F}, \mathcal{D})_{ij} = \sum_{l=2}^{m} f_i(t_l)\delta_j(t_l)(t_l - t_{l-1}); \quad 1 \leq i \leq n, 1 \leq j \leq p. \tag{5.1}$$

We now do depth-based PCA on the transformed $n \times p$ data matrix $T(\mathbb{H}; \mathcal{F}, \mathcal{D}) \equiv T(\mathbb{H})$, and obtain the rank-$q$ approximation ($q \leq p$) of the i^{th} observation using the robust $p \times q$ loading matrix $\tilde{\mathbf{P}}$ and robust $q \times 1$ score vector $\tilde{\mathbf{s}}_i$:

$$\widehat{T}(\mathbb{H})_i = \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{P}}\tilde{\mathbf{s}}_i,$$

with $\tilde{\boldsymbol{\mu}}$ being the spatial median of $T(\mathbb{H})$. Then we transform this approximation back to the original coordinates: $\hat{f}_i(t_l) = \sum_{j=1}^{p} \widehat{T}(\mathbb{H})_{ij}\delta_j(t_l)$.

We demonstrate the utility of our robust method for detecting functional outliers through two data examples. For any method of PCA with k components on a dataset of n observations and p variables, the *score distance* (SD) and *orthogonal distance* (OD) for

i^{th} observation $(i = 1, 2, ..., n)$ are defined as:

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{s_{ij}^2}{\hat{\lambda}_j}}; \quad OD_i = \|\mathbf{x}_i - \mathbf{P}\mathbf{s}_i^T\|,$$

where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are the score vectors, $\mathbf{P} \in \mathbb{R}^{p \times k}$ the loading matrix, and $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ are eigenvalues obtained from the PCA. From a practical standpoint, SD_i can be interpreted as a weighted norm of the projection of the i^{th} point on the hyperplane formed by first k principal components, and OD_i the orthogonal distance of point i from that hyperplane. For outlier detection, following Hubert et al. (2005) we set the upper cutoff values for score distances at $(\chi^2_{2,.975})^{1/2}$ and orthogonal distances at $[\text{median}(OD^{2/3}) + \text{MAD}(OD^{2/3})\Phi^{-1}(0.975)]^{3/2}$, where $\Phi(.)$ is the standard normal cumulative distribution function.

We consider the El-Niño data, which is part of a larger dataset on potential factors behind El-Niño oscillations in the tropical pacific available in `http://www.cpc.ncep.noaa.gov/data/indices`, as the first test case for outlier detection using our robust functional PCA. This records monthly average Sea Surface Temperatures from June 1970 to May 2004, and the yearly oscillations follow more or less the same pattern (see panel a of Figure 4). Using a cubic spline basis with knots at alternate months starting in June gives a close approximation of the yearly time series data (panel c), and performing depth-based PCA with $q = 1$ results in two points having their SD and OD larger than cutoff (panel e). These points correspond to the time periods June 1982 to May 1983 and June 1997 to May 1998 are marked by black curves in panels a and c, and pinpoint the two seasons with strongest El-Niño events.

Our second application is on the Octane data, which consists of 226 variables and 39 observations (Esbensen et al., 1994). Each sample is a gasoline compound with a certain octane number, and has its NIR absorbance spectra measured in 2 nm intervals between 1100 - 1550 nm. There are 6 outliers here: compounds 25, 26 and 36-39, which contain alcohol. We use the same basis structure as the one in El-Niño data here, and again the top robust PC turns out to be sufficient in identifying all 6 outliers (panels b, d and f of Figure 4).

6 Conclusion

In the above sections we elaborate on a proposed transformation based on the idea of combining sign functions in an inner product space and certain transformations of general peripherality functions defined using probability measures on the same space. Based on the conditions we impose in course of the chapter, we essentially end up using data depths as scalar multiples of the spatial sign in estimation and testing problems for the location parameter of an elliptical distribution in $\mathbb{R}^p$, and using inverse depth functions for robust estimation of different components of its covariance matrix. As demonstrated by several simulation studies and data examples, in all these situations the use of this composite transformation vector brings about efficiency gains, as well as favorable robustness properties in terms of bounded influence functions and deviations from Gaussianity.

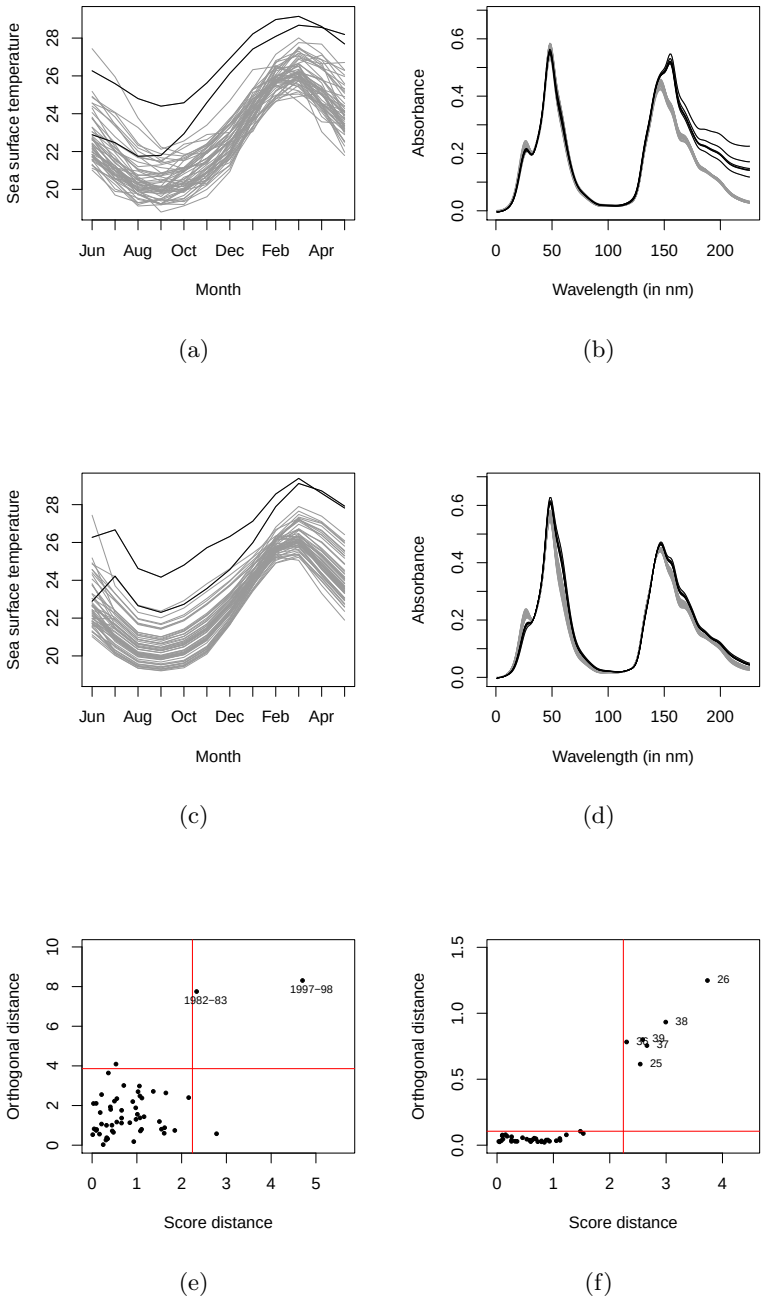


Figure 4: Actual sample curves, their spline approximations and diagnostic plots respectively for El-Niño (a,c,e) and Octane (b,d,f) datasets

Regarding the multivariate rank transformation we propose, to be noted is the fact that the mapping $\mathbf{X} \to \tilde{\mathbf{X}}$ is in fact one-to-one for elliptical distributions. Thus, such rank vectors can possibly be used for inference based on transformation-retransformation type techniques (e.g. Chakraborty and Chaudhuri (1996); Chakraborty et al. (1998)). We defer this to future research. Finally, while these rank vectors have excellent intuitive appeal in preserving the shape of a multivariate data cloud, it would be interesting to study the properties of transformations similar to (1.1) in general Hilbert spaces, as well as explore their applications in different data-analytic domains.

Appendix

A Form of $\mathbb{V}_D(\mathbb{F})$

First observe that for $\mathbb{F}$ having covariance matrix $\Sigma = \Gamma\Lambda\Gamma^T$, we have

$$\mathbb{V}_D(\mathbb{F}) = (\Gamma \otimes \Gamma)\mathbb{V}_D(\mathbb{F}_\Lambda)(\Gamma \otimes \Gamma)^T,$$

where $\mathbb{F}_\Lambda$ has the same elliptic distribution as $\mathbb{F}$, but with covariance matrix Λ. Now,

$$\begin{aligned}\mathbb{V}_D(\mathbb{F}_\Lambda) &= \mathbb{E}\left[\text{vec}\left\{\frac{(D^-_{\mathbf{Z}}(\mathbf{z}))^2\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}} - \Lambda_D\right\}\text{vec}^T\left\{\frac{(D^-_{\mathbf{Z}}(\mathbf{z}))^2\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}} - \Lambda_D\right\}\right] \\ &= \mathbb{E}\left[\text{vec}\left\{(D^-_{\mathbf{Z}}(\mathbf{z}))^2\mathbb{S}(\Lambda^{1/2}\mathbf{z};\mathbf{0})\right\}\text{vec}^T\left\{(D^-_{\mathbf{Z}}(\mathbf{z}))^2\mathbb{S}(\Lambda^{1/2}\mathbf{z};\mathbf{0})\right\}\right] - \text{vec}(\Lambda_D)\text{vec}^T(\Lambda_D)\end{aligned}$$

The matrix $\text{vec}(\Lambda_D)\text{vec}^T(\Lambda_D)$ consists of elements $\lambda_{D,i}\lambda_{D,j}$ at $(i,j)^{\text{th}}$ position of the $(i,j)^{\text{th}}$ block, and 0 otherwise. These positions correspond to variance and covariance components of on-diagonal elements. For the expectation matrix, all its elements are of the form $\mathbb{E}[\sqrt{\lambda_a\lambda_b\lambda_c\lambda_d}z_az_bz_cz_d.(D^-_{\mathbf{Z}}(\mathbf{z}))^4/(\mathbf{z}^T\Lambda\mathbf{z})^2]$, with $1 \le a,b,c,d \le p$. Since $(D^-_{\mathbf{Z}}(\mathbf{z}))^4/(\mathbf{z}^T\Lambda\mathbf{z})^2$ is even in $\mathbf{z}$, which has a sperically symmetric distribution, all such expectations will be 0 unless $a = b = c = d$, or they are pairwise equal. Following a similar derivation for spatial sign covariance matrices in Magyar and Tyler (2014), we collect the non-zero elements and write the matrix of expectations:

$$(\mathbb{I}_{p^2} + \mathbb{K}_{p,p})\left\{\sum_{a=1}^p\sum_{b=1}^p d_{ab}(\mathbf{e}_a\mathbf{e}_a^T \otimes \mathbf{e}_b\mathbf{e}_b^T) - \sum_{a=1}^p d_{aa}(\mathbf{e}_a\mathbf{e}_a^T \otimes \mathbf{e}_a\mathbf{e}_a^T)\right\} + \sum_{a=1}^p\sum_{b=1}^p d_{ab}(\mathbf{e}_a\mathbf{e}_b^T \otimes \mathbf{e}_a\mathbf{e}_b^T)$$

where $\mathbb{I}_k = (\mathbf{e}_1, ..., \mathbf{e}_k), \mathbb{K}_{m,n} = \sum_{i=1}^m\sum_{j=1}^n \mathbb{J}_{ij} \otimes \mathbb{J}_{ij}^T$ with $\mathbb{J}_{ij}$ the $m \times n$ matrix having 1 as $(i,j)^{\text{th}}$ element and 0 elsewhere, and $d_{mn} = \mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{mn}(\mathbf{z};\mathbf{0}))^2]; 1 \le m,n \le p$.

Putting everything together, denote $\tilde{\mathbb{S}}(\mathbb{F}_\Lambda) = \sum_{i=1}^n(\tilde{D}^{n-}_{\mathbf{Z}}(\mathbf{z}_i))^2\mathbb{S}(\Lambda^{1/2}\mathbf{z}_i;\hat{\boldsymbol{\mu}}_n)/n$. Then the different types of elements in the matrix $\mathbb{V}_D(F_\Lambda)$ are as given below ($1 \le a,b,c,d \le p$):

- Variance of on-diagonal elements

$$A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{aa}(\mathbb{F}_\Lambda)) = \mathbb{E}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{aa}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2\right] - \lambda^2_{D,a}$$

- Variance of off-diagonal elements ($a \neq b$)

$$A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ab}(\mathbb{F}_\Lambda)) = \mathbb{E}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ab}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2\right]$$

- Covariance of two on-diagonal elements ($a \neq b$)

$$A\mathbb{C}(\sqrt{n}\tilde{\mathbb{S}}_{aa}(\mathbb{F}_\Lambda), \sqrt{n}\tilde{\mathbb{S}}_{bb}(\mathbb{F}_\Lambda)) = \mathbb{E}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ab}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2\right] - \lambda_{D,a}\lambda_{D,b}$$

- Covariance of two off-diagonal elements ($a \neq b \neq c \neq d$)

$$A\mathbb{C}(\sqrt{n}\tilde{\mathbb{S}}_{aa}(\mathbb{F}_\Lambda), \sqrt{n}\tilde{\mathbb{S}}_{bb}(\mathbb{F}_\Lambda)) = 0$$

- Covariance of one off-diagonal and one on-diagonal element ($a \neq b \neq c$)

$$A\mathbb{C}(\sqrt{n}\tilde{\mathbb{S}}_{ab}(\mathbb{F}_\Lambda), \sqrt{n}\tilde{\mathbb{S}}_{cc}(\mathbb{F}_\Lambda)) = 0$$

B Asymptotics of eigenvectors and eigenvalues

The following result allows us to obtain asymptotic joint distributions of eigenvectors and eigenvalues of the sample DCM, provided we know the limiting distribution of the sample DCM itself:

Theorem B.1. *(Taskinen et al., 2012) Let $\mathbb{F}_\Lambda$ be an elliptical distribution with a diagonal covariance matrix Λ, and $\hat{\mathbb{C}}$ be any positive definite symmetric $p \times p$ matrix such that at $\mathbb{F}_\Lambda$ the limiting distribution of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbb{C}}-\Lambda)$ is a p^2-variate (singular) normal distribution with mean zero. Write the spectral decomposition of $\hat{\mathbb{C}}$ as $\hat{\mathbb{C}} = \hat{\mathbb{P}}\hat{\mathbb{L}}\hat{\mathbb{P}}^T$. Then the limiting distributions of $\sqrt{n}vec(\hat{\mathbb{P}} - \mathbb{I}_p)$ and $\sqrt{n}vec(\hat{\mathbb{L}} - \Lambda)$ are multivariate (singular) normal and*

$$\mathrm{vec}(\hat{\mathbb{C}} - \Lambda) = [(\Lambda \otimes \mathbb{I}_p) - (\mathbb{I}_p \otimes \Lambda)]\,\mathrm{vec}(\hat{\mathbb{P}} - \mathbb{I}_p) + \mathrm{vec}(\hat{\Lambda} - \mathbb{L}) + o_P(n^{-1/2}) \tag{B.1}$$

The first matrix picks only off-diagonal elements of the LHS and the second one only diagonal elements. We now use the above result and the form of $\mathbb{V}_D(\mathbb{F})$ derived in Appendix A to obtain limiting variance and covariances of eigenvalues and eigenvectors.

Corollary B.2. *Consider the sample DCM $\tilde{\mathbb{S}} = \sum_{i=1}^n (D^{n-}_{\mathbf{X}}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)/n$ and its spectral decomposition $\tilde{\mathbb{S}} = \hat{\Gamma}_D\hat{\Lambda}_D\hat{\Gamma}_D^T$. Then the matrices $\mathbb{G}_D = \sqrt{n}(\hat{\Gamma}_D - \Gamma)$ and $\mathbb{L}_D = \sqrt{n}(\hat{\Lambda}_D - \Lambda_D)$ have independent distributions. The random variable $\mathrm{vec}(\mathbb{G}_D)$ asymptotically has a p^2-variate normal distribution with mean $\mathbf{0}_{p^2}$, and the asymptotic variance and covariance of different columns of $\mathbb{G}_D = (\mathbf{g}_1, ..., \mathbf{g}_p)$ are as follows:*

$$A\mathbb{V}(\mathbf{g}_i) = \sum_{k=1;k\neq i}^{p} \frac{\mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ik}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2]}{(\lambda_{D,i} - \lambda_{D,k})^2}\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^T \tag{B.2}$$

$$A\mathbb{C}(\mathbf{g}_i, \mathbf{g}_j) = -\frac{\mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ij}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2]}{(\lambda_{D,i} - \lambda_{D,j})^2}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_j^T; \quad i \neq j \tag{B.3}$$

The vector consisting of diagonal elements of $\mathbb{L}_D$, say $\ell = (l_1, ..., l_p)^T$ asymptotically has a p-variate normal distribution with mean $\mathbf{0}_p$ and variance-covariance elements:

$$A\mathbb{V}(l_i) = \mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ii}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2] - \lambda^2_{D,i}, \tag{B.4}$$

$$A\mathbb{C}(l_i, l_j) = \mathbb{E}[(D^-_{\mathbf{Z}}(\mathbf{z}))^4(\mathbb{S}_{ij}(\Lambda^{1/2}\mathbf{z};\mathbf{0}))^2] - \lambda_{D,i}\lambda_{D,j}; \quad i \neq j. \tag{B.5}$$

C Proofs

Proof of Proposition 2.1. Under contiguous alternatives $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the weighted sign test statistic $T_{n,w}$ has mean $\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$. For spherically symmetric $\mathbf{Z}$, $w(\mathbf{Z})$ depends on $\mathbf{Z}$ only through its norm. Since $\|\mathbf{Z}\|$ and $\mathbf{S}(\mathbf{Z})$ are independent, we get $\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z})) = \mathbb{E}w(\mathbf{Z}).\mathbb{E}\mathbf{S}(\mathbf{Z})$. A similar decomposition holds for $\mathbb{V}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$.

We can now simplify the approximate local power $\beta_{n,w}$ of a level-α, $0 < \alpha < 1$, test based on $T_{n,w}$:

$$\begin{aligned}\beta_{n,w} &= K_p\left(\chi^2_{p,\alpha} + n(\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))^T[\mathbb{E}(w^2(\mathbf{Z})\mathbf{S}(\mathbf{Z})\mathbf{S}(\mathbf{Z})^T)]^{-1}(\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))\right)\\ &= K_p\left(\chi^2_{p,\alpha} + \frac{\mathbb{E}^2 w(\mathbf{Z})}{\mathbb{E}w^2(\mathbf{Z})}.\mathbb{E}\mathbf{S}(\mathbf{Z})^T[\mathbb{V}\mathbf{S}(\mathbf{Z})]^{-1}\mathbb{E}\mathbf{S}(\mathbf{Z})\right),\end{aligned}$$

where K_p and $\chi^2_{p,\alpha}$ are distribution function and upper-α cutoff of a χ^2_p distribution, respectively. Since $\mathbb{E}^2 w(\mathbf{Z}) \leq \mathbb{E}w(\mathbf{Z})$, the largest possible value of $\beta_{n,w}$ is $K_p(\chi^2_{p,\alpha} + \mathbb{E}\mathbf{S}(\mathbf{Z})^T[\mathbb{V}\mathbf{S}(\mathbf{Z})]^{-1}\mathbb{E}\mathbf{S}(\mathbf{Z}))$, the approximate power of the unweighted sign test statistic. Equality is achieved when $w(\mathbf{Z})$ is a constant independent of $\mathbf{Z}$. □

Sketch of proofs for equations 2.3 and 2.4. A first step to obtain asymptotic normality for the high-dimensional location test statistic $C_{n,w}$ is obtaining an equivalent result of Lemma 2.1 in Wang et al. (2015):

Lemma C.1. *Under the conditions*
(C1) $\mathrm{Tr}(\Sigma^4) = o(\mathrm{Tr}^2(\Sigma^2))$,
(C2) $\mathrm{Tr}^4(\Sigma)/\mathrm{Tr}^2(\Sigma^2)\exp[-\mathrm{Tr}^2(\Sigma)/128p\lambda^2_{\max}(\Sigma)] = o(1)$

when H_0 is true we have

$$\begin{aligned}\mathbb{E}[(\boldsymbol{\epsilon}^T_{w1}\boldsymbol{\epsilon}_{w2})^4] &= O(1)\mathbb{E}^2[(\boldsymbol{\epsilon}^T_{w1}\boldsymbol{\epsilon}_{w2})^2], &\text{(C.1)}\\ \mathbb{E}[(\boldsymbol{\epsilon}^T_{w1}B_w\boldsymbol{\epsilon}_{w1})^2] &= O(1)\mathbb{E}^2[(\boldsymbol{\epsilon}^T_{w1}B_w\boldsymbol{\epsilon}_{w1})^2], &\text{(C.2)}\\ \mathbb{E}[(\boldsymbol{\epsilon}^T_{w1}B_w\boldsymbol{\epsilon}_{w2})^2] &= o(1)\mathbb{E}^2[(\boldsymbol{\epsilon}^T_{w1}B_w\boldsymbol{\epsilon}_{w1})^2], &\text{(C.3)}\end{aligned}$$

with $\boldsymbol{\epsilon} \sim \mathcal{E}(\mathbf{0}_p, \Lambda, g)$ and $\boldsymbol{\epsilon}_w = w(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})$.

A proof of this lemma is derived using results in section 3 of El Karoui (2009), noticing that any-scalar valued 1-Lipschitz function of $\boldsymbol{\epsilon}_w$ is a M_w-Lipschitz function of $\mathbf{S}(\boldsymbol{\epsilon})$, with $M_w = \sup_{\boldsymbol{\epsilon}} w(\boldsymbol{\epsilon})$. Same steps as in the proof of Theorem 2.2 in Wang et al. (2015) follow now, using the lemma above in place of Lemma 2.1 therein, to establish asymptotic normality of $C_{n,w}$ under H_0.

To derive the asymptotic distribution under contiguous alternatives we need the conditions (C3)-(C6) in Wang et al. (2015), as well as slightly modified versions of Lemmas A.4 and A.5:

Lemma C.2. *Given that condition (C3) holds, we have $\lambda_{\max}(\mathbb{B}_w) \leq 2\frac{\lambda_{\max}}{Tr(\Sigma)}(1 + o(1))$.*

Lemma C.3. *Define $\mathbb{D}_w = \mathbb{E}\left[\frac{w^2(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|^2}(\mathbb{I}_p - \mathbf{S}(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})^T)\right]$. Then $\lambda_{\max}(\mathbb{A}_w) \leq \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)$ and $\lambda_{\max}(\mathbb{D}_w) \leq \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)^2$. Further, if (C3) and (C4) hold then $\lambda_{\min}(\mathbb{A}_w) \geq \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)(1+o(1))/\sqrt{3}$.*

The proof now exactly follows steps in the proof of theorem 2.3 in Wang et al. (2015), replacing vector signs by weighted signs, using the fact that $w(\boldsymbol{\epsilon})$ is bounded above by M_w while applying conditions (C5)-(C6) and lemmas A.1, A.2, A.3, and finally using the above two lemmas in place of lemmas A.4 and A.5 respectively. □

Proof of Theorem 3.1. The proof follows directly from writing out the expression of $Cov(\tilde{\mathbf{X}})$:

$$\begin{aligned}
\mathbb{V}\tilde{\mathbf{X}} &= \mathbb{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T) - \mathbb{E}\tilde{\mathbf{X}}\mathbb{E}\tilde{\mathbf{X}}^T \\
&= \Gamma.\mathbb{E}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^2 \frac{\|\mathbf{z}\|^2}{\|\Lambda^{1/2}\mathbf{z}\|}\Lambda^{1/2}\mathbf{S}(\mathbf{z})\mathbf{S}(\mathbf{z})^T\Lambda^{1/2}\right]\Gamma^T - \mathbf{0}_p\mathbf{0}_p^T \\
&= \Gamma.\mathbb{E}\left[(D^-_{\mathbf{Z}}(\mathbf{z}))^2 \frac{\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}}\right]\Gamma^T
\end{aligned}$$

□

Proof of Lemma 3.2. For two positive definite matrices $\mathbb{A}, \mathbb{B}$, we denote by $\mathbb{A} > \mathbb{B}$ that $\mathbb{A} - \mathbb{B}$ is positive definite. Also, denote

$$\mathbb{S}_n = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|(D^{n-}_{\mathbf{X}}(\mathbf{x}_i))^2 - (D^-_{\mathbf{X}}(\mathbf{x}_i))^2\right|\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right].$$

Due to assumption (D1), given $\epsilon > 0$ we can find $N \in \mathbb{N}$ such that

$$|(D^{n_1-}_{\mathbf{X}}(\mathbf{x}_i))^2 - (D^-_{\mathbf{X}}(\mathbf{x}_i))^2| < \epsilon \tag{C.4}$$

for all $n_1 \geq N; i = 1, 2, ..., n_1$. This implies

$$\begin{aligned}
\mathbb{S}_{n_1} &< \epsilon\sqrt{n_1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{n_1})\right] \\
&= \epsilon\sqrt{n_1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{n_1}) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\} + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right] \qquad \text{(C.5)}
\end{aligned}$$

We now construct a sequence of positive definite matrices $\{a_k(B_k + C_k) : k \in \mathbb{N}\}$ so that

$$a_k = \frac{1}{k}, \quad \mathbb{B}_k = \sqrt{N_k}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\left\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{N_k}) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\}\right], \quad \mathbb{C}_k = \sqrt{N_k}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right]$$

where $N_k \in \mathbb{N}$ gives the relation (C.4) in place of N when we take $\epsilon = 1/k$. Under conditions $\mathbb{E}\|\mathbf{x} - \boldsymbol{\mu}\|^{-3/2} < \infty$ and $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = O_P(1)$, the sample SCM with unknown location parameter $\hat{\boldsymbol{\mu}}_n$ has the same asymptotic distribution as the SCM with known location $\boldsymbol{\mu}$ (Dürre et al., 2014), hence $\mathbb{B}_k = o_P(1)$, thus $a_k(\mathbb{B}_k + \mathbb{C}_k) \xrightarrow{P} 0$.

Now (C.5) implies that for any $\epsilon_1 > 0$, $\mathbb{S}_{N_k} > \epsilon_1 \Rightarrow a_k(\mathbb{B}_k + \mathbb{C}_k) > \epsilon_1$, which means $P(\mathbb{S}_{N_k} > \epsilon_1) < P(a_k(\mathbb{B}_k + \mathbb{C}_k) > \epsilon_1)$. Hence the subsequence $\{\mathbb{S}_{N_k}\} \xrightarrow{P} 0$. Since the main sequence $\{\mathbb{S}_k\}$ is bounded below by 0, this implies $\{\mathbb{S}_k\} \xrightarrow{P} 0$. Finally, we have that

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}(D_{\mathbf{X}}^{n-}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) - \frac{1}{n}\sum_{i=1}^{n}(D_{\mathbf{X}}^{-}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right] \leq \mathbb{S}_n + \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\}\right] \tag{C.6}$$

Since the second summand on the right hand side is $o_P(1)$ due to Dürre et al. (2014) as mentioned before, we have the needed. □

Proof of Theorem 3.3. The quantity in the statement of the theorem can be broken down as:

$$\sqrt{n}\left[\text{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D_{\mathbf{X}}^{n-}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right\} - \text{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D_{\mathbf{X}}^{-}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\}\right] + \sqrt{n}\left[\text{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D_{\mathbf{X}}^{-}(\mathbf{x}_i))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\} - \mathbb{E}\left[\text{vec}\left\{(D_{\mathbf{X}}^{-}(\mathbf{x}))^2\mathbb{S}(\mathbf{x};\boldsymbol{\mu})\right\}\right]\right]$$

The first part goes to 0 in probability by Lemma 3.2, and applying Slutsky's theorem we get the required convergence. □

Proof of Theorem 3.4. We are going to prove the following:

1. $\|\hat{\Gamma}_D - \Gamma\|_F \xrightarrow{P} 0$, and
2. $\|\hat{\Lambda} - \Lambda\|_F \xrightarrow{P} 0$,

as $n \to \infty$. For (1), we notice $\sqrt{n}\,\text{vec}(\hat{\Gamma}_D - \Gamma)$ asymptotically has a (singular) multivariate normal distribution following Corollary B.2, so that $\|\hat{\Gamma}_D - \Gamma\|_F = O_P(1/\sqrt{n})$ using Prokhorov's theorem.

It is now enough to prove convergence in probability of the individual eigenvalue estimates $\hat{\lambda}_i; i = 1, ..., p$. For this, define estimates $\tilde{\lambda}_i$ as median-of-small-variances estimator of the *true* score vectors $\Gamma^T\mathbf{X}$. For this we have

$$|\tilde{\lambda}_i - \lambda_i| \xrightarrow{P} 0 \tag{C.7}$$

using Theorem 3.1 of Minsker (2015), with $\mu = \lambda_i$. Now $\hat{\lambda}_i = \text{median}_j(\mathbb{S}(\mathbb{X}_{G_j}\hat{\boldsymbol{\gamma}}_{D,i}))$ and $\tilde{\lambda}_i = \text{median}_j(\mathbb{S}(\mathbb{X}_{G_j}\boldsymbol{\gamma}_i))$, so that

$$\begin{aligned} |\hat{\lambda}_i - \tilde{\lambda}_i| &\leq \text{median}_j\left[\mathbb{S}(\mathbb{X}_{G_j}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i))\right] \\ &\leq \|\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i\|^2\text{median}_j\left[\text{Tr}(\mathbb{S}(\mathbb{X}_{G_j}))\right] \end{aligned}$$

using Cauchy-Schwarz inequality. Combining the facts $\|\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i\| = O_P(1/\sqrt{n})$ and $\text{median}_j[\text{Tr}(\mathbb{S}(X_{G_j}))] \xrightarrow{P} \text{Tr}(\Sigma)$ (Minsker, 2015) with (C.7), we get the needed.

□

Proof of Corollary B.2. In spirit, this corollary is similar to Theorem 13.5.1 in Anderson (2003). Due to the decomposition (B.1) we have, for the distribution F_Λ, the following relation between any off-diagonal element of $\tilde{\mathbb{S}}(\mathbb{F}_\Lambda)$ and the corresponding element in the estimate of eigenvectors $\hat{\Gamma}_D(\mathbb{F}_\Lambda)$:

$$\sqrt{n}\hat{\gamma}_{D,ij}(\mathbb{F}_\Lambda) = \sqrt{n}\frac{\tilde{\mathbb{S}}_{ij}(\mathbb{F}_\Lambda)}{\lambda_{D,i} - \lambda_{D,j}}; \quad i \neq j$$

So that for eigenvector estimates of the original F we have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i) &= \sqrt{n}\Gamma(\hat{\boldsymbol{\gamma}}_{D,i}(\mathbb{F}_\Lambda) - \mathbf{e}_i) \\ &= \sqrt{n}\left[\sum_{k=1;k\neq i}^{p} \hat{\gamma}_{D,ik}(\mathbb{F}_\Lambda)\boldsymbol{\gamma}_k + (\hat{\gamma}_{D,ii}(\mathbb{F}_\Lambda) - 1)\boldsymbol{\gamma}_i\right]. \qquad \text{(C.8)}\end{aligned}$$

By assumption we now have $\sqrt{n}(\hat{\gamma}_{D,ii}(\mathbb{F}_\Lambda)-1) = o_P(1)$ and $A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ik}(\mathbb{F}_\Lambda), \sqrt{n}\tilde{\mathbb{S}}_{il}(\mathbb{F}_\Lambda)) = 0$ for $k \neq l$, so the above equation implies

$$A\mathbb{V}(\mathbf{g}_i) = A\mathbb{V}(\sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i)) = \sum_{k=1;k\neq i}^{p} \frac{A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ik}(\mathbb{F}_\Lambda))}{(\lambda_{D,i} - \lambda_{D,k})^2}\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^T$$

For the covariance terms, from (C.8) we get, for $i \neq j$,

$$\begin{aligned}A\mathbb{V}(\mathbf{g}_i, \mathbf{g}_j) &= A\mathbb{V}(\sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i), \sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,j} - \boldsymbol{\gamma}_j)) \\ &= A\mathbb{V}\left(\sum_{k=1;k\neq i}^{p} \sqrt{n}\hat{\gamma}_{D,ik}(\mathbb{F}_\Lambda)\boldsymbol{\gamma}_k, \sum_{k=1;k\neq j}^{p} \sqrt{n}\hat{\gamma}_{D,jk}(\mathbb{F}_\Lambda)\boldsymbol{\gamma}_k\right) \\ &= A\mathbb{V}\left(\sqrt{n}\hat{\gamma}_{D,ij}(\mathbb{F}_\Lambda)\boldsymbol{\gamma}_j, \sqrt{n}\hat{\gamma}_{D,ji}(\mathbb{F}_\Lambda)\boldsymbol{\gamma}_i\right) \\ &= -\frac{A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ij}(\mathbb{F}_\Lambda))}{(\lambda_{D,i} - \lambda_{D,j})^2}\boldsymbol{\gamma}_j\boldsymbol{\gamma}_i^T.\end{aligned}$$

The exact forms given in the statement of the corollary now follows from the Form of $\mathbb{V}_D(\mathbb{F})$ in Appendix A.

For the on-diagonal elements of $\tilde{\mathbb{S}}(\mathbb{F}_\Lambda)$, Theorem B.1 gives us $\sqrt{n}\hat{\lambda}_{D,i}(\mathbb{F}_\Lambda) = \sqrt{n}\tilde{\mathbb{S}}_{ii}(\mathbb{F}_\Lambda)$ for $i = 1, ..., p$. Hence

$$\begin{aligned}A\mathbb{V}(l_i) &= A\mathbb{V}(\sqrt{n}\hat{\lambda}_{D,i} - \sqrt{n}\lambda_{D,i}) \\ &= A\mathbb{V}(\sqrt{n}\hat{\lambda}_{D,i}(\mathbb{F}_\Lambda) - \sqrt{n}\lambda_{D,i}(\mathbb{F}_\Lambda)) \\ &= A\mathbb{V}(\sqrt{n}\mathbb{S}_{ii}(\mathbb{F}_\Lambda)).\end{aligned}$$

A similar derivation gives the expression for $A\mathbb{V}(l_i, l_j); i \neq j$. Finally, since the asymptotic covariance between an on-diagonal and an off-diagonal element of $\tilde{\mathbb{S}}(\mathbb{F}_\Lambda)$, it follows that the elements of $\mathbb{G}_D$ and diagonal elements of $\mathbb{L}_D$ are independent. □

F = Bivariate t_5	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.80	0.83	0.95	0.95	0.89	1.00	0.96	0.89
n=50	0.86	0.90	1.25	1.10	1.21	1.32	1.13	1.25
n=100	1.02	1.04	1.58	1.20	1.54	1.67	1.24	1.63
n=300	1.24	1.28	1.81	1.36	1.82	1.93	1.44	1.95
n=500	1.25	1.29	1.80	1.33	1.84	1.91	1.39	1.97
F = Bivariate t_6	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.77	0.79	0.92	0.92	0.86	0.96	0.92	0.85
n=50	0.76	0.78	1.11	1.00	1.08	1.17	1.03	1.13
n=100	0.78	0.79	1.27	1.06	1.33	1.35	1.11	1.41
n=300	0.88	0.91	1.29	1.09	1.35	1.38	1.15	1.45
n=500	0.93	0.96	1.37	1.13	1.40	1.44	1.19	1.48
F = Bivariate t_{10}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.70	0.72	0.83	0.84	0.77	0.89	0.87	0.79
n=50	0.58	0.60	0.90	0.84	0.86	0.95	0.88	0.91
n=100	0.57	0.59	0.92	0.87	0.97	0.98	0.90	1.03
n=300	0.62	0.64	0.93	0.85	0.99	0.99	0.91	1.06
n=500	0.62	0.65	0.93	0.86	1.00	1.00	0.92	1.08
F = Bivariate t_{15}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.63	0.66	0.76	0.78	0.72	0.81	0.81	0.73
n=50	0.52	0.52	0.79	0.75	0.80	0.84	0.79	0.85
n=100	0.51	0.52	0.83	0.77	0.88	0.88	0.81	0.94
n=300	0.55	0.56	0.84	0.79	0.91	0.89	0.84	0.98
n=500	0.56	0.59	0.85	0.80	0.93	0.91	0.86	0.99
F = Bivariate t_{25}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.63	0.65	0.77	0.79	0.74	0.80	0.81	0.74
n=50	0.49	0.50	0.73	0.71	0.76	0.78	0.75	0.80
n=100	0.45	0.46	0.73	0.69	0.81	0.78	0.73	0.87
n=300	0.51	0.52	0.78	0.75	0.87	0.83	0.79	0.94
n=500	0.53	0.55	0.79	0.75	0.87	0.84	0.80	0.94
F = BVN	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.56	0.60	0.69	0.71	0.67	0.73	0.74	0.68
n=50	0.42	0.43	0.66	0.66	0.70	0.71	0.69	0.75
n=100	0.42	0.43	0.69	0.66	0.77	0.74	0.71	0.83
n=300	0.47	0.49	0.71	0.69	0.82	0.76	0.73	0.88
n=500	0.48	0.50	0.73	0.71	0.83	0.78	0.76	0.89

Table 5: Finite sample efficiencies of several scatter matrices: $p = 2$

D Additional simulations

We list the finite sample efficiencies of DCM and ADCMs for $p = 2, 3$ in tables 5 and 6, respectively.

References

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Phil. Trans. R. Soc. A*, 367:4385–4405.

Anderson, T. (3rd ed. 2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, Hoboken, NJ.

Boente, G. and Salibian-Barrera, M. (2015). S-Estimators for Functional Principal Component Analysis. *J. Amer. Statist. Assoc.*, 110:1100–1111.

Brown, B. (1983). Statistical use of the spatial median. *J. R. Statist. Soc. B*, 45:25–30.

3-variate t_5	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.96	0.97	1.06	1.03	0.99	1.07	1.06	0.97
n=50	1.07	1.08	1.28	1.20	1.18	1.33	1.23	1.20
n=100	1.12	1.15	1.49	1.31	1.40	1.57	1.38	1.48
n=300	1.49	1.54	2.09	1.82	2.07	2.19	1.93	2.18
n=500	1.60	1.66	2.18	1.87	2.21	2.27	1.95	2.30
3-variate t_6	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.90	0.92	1.00	0.99	0.95	1.02	1.01	0.94
n=50	0.95	0.96	1.16	1.09	1.09	1.21	1.14	1.11
n=100	0.98	0.99	1.32	1.22	1.25	1.38	1.27	1.29
n=300	1.10	1.14	1.57	1.40	1.58	1.62	1.47	1.64
n=500	1.17	1.20	1.57	1.43	1.60	1.63	1.51	1.67
3-variate t_{10}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.87	0.88	0.95	0.94	0.90	0.97	0.98	0.89
n=50	0.77	0.79	0.96	0.92	0.94	0.99	0.96	0.95
n=100	0.75	0.76	1.02	0.95	1.01	1.06	1.00	1.05
n=300	0.73	0.75	1.03	0.98	1.10	1.08	1.03	1.15
n=500	0.73	0.76	1.02	0.98	1.09	1.06	1.02	1.14
3-variate t_{15}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.84	0.86	0.92	0.92	0.89	0.94	0.94	0.87
n=50	0.75	0.76	0.92	0.90	0.90	0.96	0.94	0.93
n=100	0.66	0.67	0.91	0.87	0.95	0.96	0.92	1.00
n=300	0.61	0.64	0.90	0.87	1.00	0.93	0.91	1.04
n=500	0.65	0.67	0.89	0.87	0.99	0.93	0.91	1.03
3-variate t_{25}	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.78	0.79	0.87	0.89	0.87	0.89	0.92	0.86
n=50	0.70	0.71	0.88	0.86	0.88	0.91	0.90	0.90
n=100	0.61	0.63	0.86	0.83	0.89	0.90	0.88	0.94
n=300	0.58	0.59	0.83	0.80	0.92	0.87	0.85	0.98
n=500	0.62	0.64	0.83	0.82	0.94	0.88	0.87	0.99
3-variate Normal	SCM	Tyler	DCM-H	DCM-M	DCM-P	ADCM-H	ADCM-M	ADCM-P
n=20	0.76	0.78	0.85	0.87	0.84	0.87	0.90	0.83
n=50	0.66	0.67	0.82	0.81	0.84	0.86	0.86	0.86
n=100	0.56	0.58	0.77	0.75	0.83	0.82	0.79	0.87
n=300	0.53	0.55	0.75	0.74	0.85	0.79	0.78	0.90
n=500	0.56	0.58	0.76	0.76	0.87	0.80	0.80	0.92

Table 6: Finite sample efficiencies of several scatter matrices: $p = 3$

Chakraborty, B. and Chaudhuri, P. (1996). On a Transformation and Re-Transformation Technique for Constructing an Affine Equivariant Multivariate Median. *Proc. Amer. Math. Soc.*, 124:2539–2547.

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

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91:862–872.

Chen, S. X. and Qin, Y. L. (2010). A Two-sample Test for High-dimensional Data with Application to Gene-Set Testing. *Ann. Statist.*, 38:808–835.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45:223–256.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Biometrika*, 20:927–1010.

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.

Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statist. Probab. Lett.*, 14:119–128.

Dürre, A., Vogel, D., and Tyler, D. (2014). The spatial sign covariance matrix with unknown location. *J. Mult. Anal.*, 130:107–117.

Dutta, S. and Genton, M. G. (2017). Depthweighted robust multivariate regression with application to sparse data. *Can. J. Stat.*, 45(2):164–184.

Dutta, S. and Ghosh, A. (2012). On robust classification using projection depth. *Ann. Inst. Stat. Math.*, 64-3:657–676.

El Karoui, N. (2009). Concentration of Measure and Spectra of Random Matrices: with Applications to Correlation Matrices, Elliptical Distributions and Beyond. *Ann. Appl. Probab.*, 19:2362–2405.

Esbensen, K. H., Schönkopf, S., and Midtgaard, T. (1994). *Multivariate Analysis in Practice*. CAMO, Trondheim, Germany.

Fang, K. T., .Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Monographs on Statistics and Applied Probability 36. Chapman and Hall Ltd., London, United Kingdom.

Ghosh, A. and Chaudhuri, P. (2005). On Maximum Depth and Related Classifiers. *Scand. J. Statist.*, 32:327–350.

Haldane, J. B. S. (1948). Note on the Median of a Multivariate Distribution. *Biometrika*, 35:414–415.

Hallin, M. and Paindaveine, D. (2002). Optimal tests for multivariate loca-tion based on interdirections and pseudo-Mahalanobis ranks. *Ann. Statist.*, 30:1103–1133.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Staehl, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Huber, P. J. (1981). *Robust Statistics*. Wiley series in probability and mathematical statistics. Wiley.

Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47-1:64–79.

Jornsten, R. (2004). Clustering and classification based on the l_1 depth. *J. Mult. Anal.*, 90-1:67–89.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., and Cohen, K. (1999). Robust principal components of functional data. *TEST*, 8:1–73.

Magyar, A. and Tyler, D. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, 101:673–688.

Majumdar, S. and Chatterjee, S. (2018). Non-convex penalized multitask regression using data depth-based penalties. *Stat*, 7:e174.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in $\mathbb{R}^n$. *Linear Algebra Appl.*, 171:81–98.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21:2308–2335.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametric Stat.*, 5:201–213.

Oja, H. (1983). Descriptive Statistics for Multivariate Distributions. *Statist. and Prob. Lett.*, 1:327–332.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Lecture Notes in Statistics. Springer.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York, NY.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *TEST*, 23-4:725–750.

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrised M-estimators of scatter. *J. Mult. Anal.*, 98:1611–1629.

Taskinen, S., Koch, I., and Oja, H. (2012). Robustifying principal component analysis with spatial sign vectors. *Statist. Probab. Lett.*, 82:765–774.

Tyler, D. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, 15:234–251.

Wang, L., Peng, B., and Li, R. (2015). A High-Dimensional Nonparametric Multivariate Test for Mean Vector. *J. Amer. Statist. Assoc.*, 110:1658–1669.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490.

Zuo, Y. and Cui, M. (2005). Depth weighted scatter estimators. *Ann. Statist.*, 33-1:381–413.

Zuo, Y., Cui, M., and He, X. (2004). On the Staehl-Donoho estimator and depth-weighted means of multivariate data. *Ann. Statist.*, 32-1:167–188.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482.