Tyler Shape Depth

Davy Paindaveine[*]   and   Germain Van Bever[†]

[*] ECARES and Departement of Mathematics, Université libre de Bruxelles, Avenue F.D. Roosevelt, 50, CP114/04, B-1050, Brussels, Belgium

[†] Departement of Mathematics and Namur Institute for Complex Systems, Université de Namur, Rempart de la Vierge, 8, 5000, Namur, Belgium

Abstract

In many problems from multivariate analysis, the parameter of interest is a shape matrix, that is, a normalized version of the corresponding scatter or dispersion matrix. In this paper, we propose a depth concept for shape matrices that involves data points only through their directions from the center of the distribution. We use the terminology Tyler shape depth since the resulting estimator of shape, namely the deepest shape matrix, is the median-based counterpart of the M-estimator of shape of Tyler (1987). Beyond estimation, shape depth, like its Tyler antecedent, also allows hypothesis testing on shape. Its main benefit, however, lies in the ranking of shape matrices it provides, whose practical relevance is illustrated in principal component analysis and in shape-based outlier detection. We study the invariance, quasi-concavity and continuity properties of Tyler shape depth, the topological and boundedness properties of the corresponding depth regions, existence of a deepest shape matrix and prove Fisher consistency in the elliptical case. Finally, we derive a Glivenko–Cantelli-type result and establish almost sure consistency of the deepest shape matrix estimator.

Keywords: Elliptical distribution; Principal component analysis; Robustness; Shape matrix; Statistical depth; Test for sphericity.

1   Introduction

Location depths measure the centrality of an arbitrary k-vector θ with respect to a probability measure $P = P^X$ over $\mathbb{R}^k$. Letting $\mathcal{S}^{k-1} = \{x \in \mathbb{R}^k : \|x\|^2 = x^T x = 1\}$ denote the unit sphere in $\mathbb{R}^k$, the most famous instance is the Tukey (1975) halfspace depth

$$D(\theta, P) = \inf_{u \in \mathcal{S}^{k-1}} \text{pr}\{u^T(X - \theta) \geq 0\}; \qquad (1.1)$$

throughout, pr refers to probability under the probability measure P at hand. The halfspace depth regions $\{\theta \in \mathbb{R}^k : D(\theta, P) \geq \alpha\}$ form a family of nested convex subsets of $\mathbb{R}^k$. The Tukey median θ_P, defined as the barycenter of the innermost region $M_P = \{\theta \in \mathbb{R}^k : D(\theta, P) = \max_{\xi \in \mathbb{R}^k} D(\xi, P)\}$, extends the univariate median to the multivariate case and is a robust alternative to the expectation $E(X)$. Beyond location estimation, many inference problems can be tackled in a robust and nonparametric way by using the center-outward order resulting from depth (Liu et al., 1999). Adopting the

1

parametric depth approach from Mizera (2002), $D(\theta, P)$ can also be read as a measure of how well the location parameter value θ fits the probability measure P. In this spirit, possible outliers in a data set $X_1, \ldots, X_n$ will be flagged by low depth values $D(X_i, P_n)$, where P_n denotes the corresponding empirical probability measure.

In this paper, the focus is on multivariate dispersion parameters known as shape matrices. For simplicity, we restrict in this section to elliptical distributions. Let $\mathcal{P}_k$ be the collection of $k \times k$ symmetric positive definite matrices and write $A^{1/2}$, with $A \in \mathcal{P}_k$, for the unique square root of A in $\mathcal{P}_k$. We will say that $P = P^X$ is elliptical with location $\theta \in \mathbb{R}^k$, scatter $\Sigma \in \mathcal{P}_k$ and generating variate R if X has the same distribution as $\theta + R\Sigma^{1/2}U$, where U is uniformly distributed over $\mathcal{S}^{k-1}$ and is independent of the nonnegative scalar random variable R, which has unit median. This median constraint makes Σ identifiable without moment conditions. Under finite second-order moments, the resulting covariance matrix is $\Sigma_P = \{E(R^2)/k\}\Sigma$. Inference problems such as constructing confidence regions for θ require one to estimate the full scatter matrix Σ or the full covariance matrix Σ_P. However, in many other problems, it is sufficient to estimate the shape matrix, that is, the normalized scatter matrix

$$V = \frac{k}{\text{tr}(\Sigma)} \Sigma = \frac{k}{\text{tr}(\Sigma_P)} \Sigma_P.$$

This shape matrix V could be normalized, as in Paindaveine (2008), to have determinant one or upper-left entry one, which would not affect the results of the present paper. For instance, principal components may be equivalently computed from V, from Σ or, when it exists, from Σ_P, since proportional matrices have the same eigenvectors. Now, when it comes to fixing the number of principal components on which to base further analysis, one typically looks at the proportions of explained variances $p_m(\Sigma_P) = \sum_{\ell=1}^{m} \lambda_\ell(\Sigma_P) / \sum_{\ell=1}^{k} \lambda_\ell(\Sigma_P)$ $(m = 1, \ldots, k)$, where $\lambda_\ell(A)$ denotes the ℓth largest eigenvalue of A. Similarly to eigenvectors, these proportions remain unchanged if they are computed from V rather than from Σ or Σ_P. In principal component analysis it is thus sufficient to estimate, or know the value of, V.

There is a large literature on inference for shape. Our main contribution is to provide a depth concept for shape, measuring how well a given shape matrix V fits the probability measure P. While the proposed depth will lead to estimators and tests for shape, its main added value is the ordering of shape matrices resulting from depth. Here, we mention only two possible applications. The first is in principal component analysis, where a suitable estimator $\hat{V}$ is to be chosen. When it is suspected that there might be outliers, one might for instance consider the minimum covariance determinant estimates $\hat{V}_\gamma$, $\gamma \in [0.5, 1]$, trimming a proportion $1 - \gamma$ of the data; see § 5. Choosing γ should typically be done on the basis of the proportion of outliers, which is usually unknown. We will show that the shape depth of $\hat{V}_\gamma$ allows for an informed choice on γ. The second application concerns outlier detection in multivariate financial times series. Since volatility is key in finance, one might flag atypical days in such series by spotting days that associate a low depth to a shape estimator $\hat{V}_{\text{full}}$ computed from the full series.

Depth for a generic parameter has been discussed in Mizera (2002). Depth for scatter matrices, however, has only been considered in Zhang (2002), Chen et al. (2018) and Paindaveine and Van Bever (2018), and only the last considers depth for shape matrices.

2 Shape depth

Tyler (1987) introduced a shape notion extending the concept of shape outside the elliptical setup. Consider the multivariate sign $U_{\theta,V}$ defined as $V^{-1/2}(X-\theta)/\|V^{-1/2}(X-\theta)\|$ if $X \neq \theta$ and as 0 otherwise, where $V^{-1/2}$ is the inverse of $V^{1/2}$. Let also $W_{\theta,V} = \mathrm{vec}\{U_{\theta,V}U_{\theta,V}^T - (1/k)I_k\}$, where $\mathrm{vec}\,A$ stacks the columns of A on top of each other and where I_k is the $k \times k$ identity matrix. The Tyler shape of $P = P^X$, V_T say, is then the matrix $V \in \mathcal{P}_{k,\mathrm{tr}} = \{V \in \mathcal{P}_k : \mathrm{tr}(V) = k\}$ satisfying

$$E(W_{\theta,V}) = 0. \tag{2.1}$$

If P is smooth at θ, in the sense that no hyperplane containing θ has a strictly positive P-probability mass, then (2.1) admits a unique solution $V \in \mathcal{P}_{k,\mathrm{tr}}$ that agrees with the true shape if P is elliptical with location θ (Tyler, 1987; Kent and Tyler, 1988; Dümbgen, 1998). In essence, (2.1) identifies the shape V making the origin of $\mathbb{R}^{k^2}$ most central in an L_2-sense for the distribution $P^{W_{\theta,V}}$ of $W_{\theta,V}$, that is, it defines V_T as the solution of

$$0 = \arg \min_{m \in \mathbb{R}^{k^2}} E(\|W_{\theta,V} - m\|^2). \tag{2.2}$$

The present work finds its source in the idea that one may define the shape of P as the matrix $V \in \mathcal{P}_{k,\mathrm{tr}}$ making the origin of $\mathbb{R}^{k^2}$ most central for the distribution of $W_{\theta,V}$, in the halfspace depth sense, that is, as the value of V maximizing the following depth.

Definition 2.1 (Tyler shape depth). *Let $P = P^X$ be a probability measure over $\mathbb{R}^k$ and fix $V \in \mathcal{P}_{k,\mathrm{tr}}$. (i) For any $\theta \in \mathbb{R}^k$, the fixed-θ shape depth of V with respect to P is $D_\theta(V,P) = D(0, P^{W_{\theta,V}}) = \inf_{u \in \mathcal{S}^{k^2-1}} \mathrm{pr}(u^T W_{\theta,V} \geq 0)$. (ii) The shape depth of V with respect to P is $D(V,P) = D_{\theta_P}(V,P)$, where θ_P is the Tukey median of P.*

We will use the notation $D(\cdot, P)$ for both halfspace and Tyler shape depths, as the vector or matrix nature of the argument will remove any ambiguity. The fixed-θ shape depth can equivalently be defined as $D_\theta(V,P) = \inf_M \mathrm{pr}\{U_{\theta,V}^T M U_{\theta,V} - \mathrm{tr}(M)/k \geq 0\}$, where the infimum is over all $k \times k$ symmetric matrices M; see Lemma 1 in the Supplementary Material. While, in view of (2.2), V_T can be seen as a sign-based mean concept for shape, the maximizer of Tyler shape depth is of a median nature. The main benefit of the proposed depth does not come from the deepest shape itself but rather from the ranking of shapes it provides; see § 5.

Definition 2.1(ii) calls for some comments. Two approaches were considered in the literature for Tyler shape in the case of unspecified center: the Tyler (1987) plug-in approach, which replaces the unknown θ with some location functional, and the Hettmansperger and Randles (2002) approach, which jointly solves $E(U_{\theta,V}) = 0$ and $E(W_{\theta,V}) = 0$; existence of a unique solution to joint location and scatter M-estimating equations was studied in Maronna (1976) under ellipticity and in Tatsuoka and Tyler (2000) for non-elliptic cases. Both approaches provide two distinct shapes outside the elliptical setup. In contrast, for the proposed depth, the plug-in and joint maximization approaches always lead to the same shape: irrespective of λ, the objective function $(\theta, V) \mapsto D(0, P^{U_{\theta,V}}) + \lambda D(0, P^{W_{\theta,V}})$ is indeed maximized at $\theta = \theta_P$ and

$V = \arg\max_V D(0, P^{W_{\theta_P,V}})$, since $D\left(0, P^{U_{\theta,V}}\right) = D(0, P^{V^{-1/2}(X-\theta)}) = D(\theta, P^X)$ is, for any V, maximized at $\theta = \theta_P$.

An alternative way to obtain an unspecified location version of Tyler shape is to construct it on pairwise differences (Dümbgen, 1998). We will not investigate this for our shape depth, since the sample version of the resulting depth would lead to a much heavier computational burden.

3 Main properties

In this section, we study the main properties of the shape depth $D_\theta(V, P)$ and of the corresponding depth regions $R_\theta(\alpha, P) = \{V \in \mathcal{P}_{k,\mathrm{tr}} : D_\theta(V, P) \geq \alpha\}$. Topological statements for subsets of $\mathcal{P}_{k,\mathrm{tr}}$ and for functions defined on $\mathcal{P}_{k,\mathrm{tr}}$ will refer to the topology whose open sets are generated by balls of the form $B(V_0, r) = \{V \in \mathcal{P}_{k,\mathrm{tr}} : d(V, V_0) < r\}$, where d is the usual geodesic distance on $\mathcal{P}_k$: with the classical log mapping on $\mathcal{P}_k$, this distance is such that $d(V_a, V_b) = \|\log(V_a^{-1/2} V_b V_a^{-1/2})\|_F$, where $\|A\|_F = \{\mathrm{tr}(AA^T)\}^{1/2}$ is the Frobenius norm of A (Bhatia, 2007). We start with the following continuity result.

Theorem 3.1. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Then, (i) $V \mapsto D_\theta(V, P)$ is upper semicontinuous on $\mathcal{P}_{k,\mathrm{tr}}$; (ii) the depth region $R_\theta(\alpha, P)$ is closed for any $\alpha \geq 0$; (iii) if P is absolutely continuous with respect to the Lebesgue measure, then $V \mapsto D_\theta(V, P)$ is also lower semicontinuous, hence continuous, on $\mathcal{P}_{k,\mathrm{tr}}$.*

We will say that a subset R of $\mathcal{P}_{k,\mathrm{tr}}$ is bounded if and only if $R \subset B(I_k, r)$ for some $r > 0$; since d satisfies the triangle inequality, we need only consider balls centered at I_k. Moreover, we will say that P is smooth at θ if and only if $t_{\theta,P} = 0$, with $t_{\theta,P} = \sup_{u \in \mathcal{S}^{k-1}} \mathrm{pr}\{u^T(X - \theta) = 0\}$. We then have the following result.

Theorem 3.2. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Then the depth region $R_\theta(\alpha, P)$ is bounded and compact for any $\alpha > t_{\theta,P}$.*

The main reason to work with geodesic distance rather than Frobenius distance $d_F(V_1, V_2) = \|V_2 - V_1\|_F$ is that, unlike $(\mathcal{P}_{k,\mathrm{tr}}, d_F)$, the metric space $(\mathcal{P}_{k,\mathrm{tr}}, d)$ is complete; see, e.g., Proposition 10 in Bhatia and Holbrook (2006). This is what allows us to establish compacity in Theorem 3.2, which is the main ingredient for the following result.

Theorem 3.3. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. (i) If $R_\theta(t_{\theta,P}, P)$ is non-empty, then there exists a shape $V_* \in \mathcal{P}_{k,\mathrm{tr}}$ maximizing $D_\theta(V, P)$. In particular, (ii) if P is smooth at θ, then such a deepest shape V_* exists.*

While the previous result guarantees existence of a deepest shape for absolutely continuous probability measures, uniqueness is not guaranteed in general. Parallel to what is done for the Tukey median, we then define the fixed-θ shape matrix of P as the barycenter of the deepest shape region of P, that is, as the shape matrix $V_{\theta,P}$ satisfying

$$\mathrm{vec}\, V_{\theta,P} = \int_{\mathrm{vec}\, R_\theta(\alpha_*, P)} v \, dv \Big/ \int_{\mathrm{vec}\, R_\theta(\alpha_*, P)} dv, \tag{3.1}$$

with $\alpha_* = \max_V D_\theta(V, P)$. Two remarks are in order. First, the integrals in (3.1) exist and are finite since $\operatorname{vec} \mathcal{P}_{k,\mathrm{tr}}$ is a bounded subset of $\mathbb{R}^{k^2}$: $0 \leq V_{ij}^2 < V_{ii}V_{jj} \leq k^2$ for any $V \in \mathcal{P}_{k,\mathrm{tr}}$. Second, the following convexity result implies that $V_{\theta,P}$ has maximal depth.

Theorem 3.4. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Then, (i) $V \mapsto D_\theta(V, P)$ is quasi-concave: $D_\theta(V_t, P) \geq \min\{D_\theta(V_a, P), D_\theta(V_b, P)\}$ for $V_t = (1 - t)V_a + tV_b$ with $V_a, V_b \in \mathcal{P}_{k,\mathrm{tr}}$ and $t \in [0, 1]$; (ii) the region $R_\theta(\alpha, P)$ is convex for any $\alpha \geq 0$.*

This defines the fixed-θ shape of a probability measure P under the very mild condition that $R_\theta(t_{\theta,P}, P)$ is non-empty, hence in particular when P is smooth at θ. Of course, it is important that, under ellipticity, this agrees with the elliptical concept of shape provided in §1. The following Fisher consistency result confirms that this is the case.

Theorem 3.5. *Let P be an elliptical probability measure over $\mathbb{R}^k$ with location θ_0 and shape V_0. Then, $D_{\theta_0}(V_0, P) \geq D_{\theta_0}(V, P)$ for any $V \in \mathcal{P}_{k,\mathrm{tr}}$, and, provided that $\operatorname{pr}[\{\theta_0\}] < 1$, the equality holds if and only if $V = V_0$. Letting Y_k be Beta with parameters $1/2$ and $(k-1)/2$, the maximal depth is $D_{\theta_0}(V_0, P) = (1 - \operatorname{pr}[\{\theta_0\}])\operatorname{pr}(Y_k > 1/k)$.*

In this result, $\operatorname{pr}[\{\theta_0\}]$ equals the probability that the generating variate R associated to P is equal to zero. Lemma 2 in Paindaveine and Van Bever (2017) implies that the maximal depth in Theorem 3.5 is monotone decreasing in k if $\operatorname{pr}[\{\theta_0\}]$ does not depend on k, in which case the maximal depth is convergent as k goes to infinity. Since Y_k has the same distribution as $Z_1^2/(\sum_{\ell=1}^k Z_\ell^2)$, where $Z = (Z_1, \ldots, Z_k)^T$ is k-variate standard normal, the limit is equal to $\operatorname{pr}(Z_1^2 > 1) \approx 0.317$. The proof of Theorem 3.5 requires the following result.

Theorem 3.6. *Let $P = P^X$ be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Then, for any shape matrix V, any invertible $k \times k$ matrix A and any k-vector b,*

$$D_{A\theta+b}(V_A, P^{AX+b}) = D_\theta(V, P^X), \quad R_{A\theta+b}(\alpha, P^{AX+b}) = \{V_A : V \in R_\theta(\alpha, P)\},$$

where $V_A = kAVA^T/\operatorname{tr}(AVA^T)$ is the shape matrix proportional to AVA^T.

This shows that the fixed-θ shape depth and the corresponding regions behave well under affine transformations, and in particular under changes of the measurement units. Affine invariance is a classical requirement in location depth (Zuo and Serfling, 2000).

Tyler shape depth is a sign concept in the sense that it depends on the underlying random vector X only through its multivariate sign $U_{\theta,V}$. In the elliptical case, it follows that, if the distribution does not charge the center of the distribution, this depth does not depend on the distribution of the underlying generating variate R. More precisely, we have the following result.

Theorem 3.7. *Let P be an elliptical probability measure over $\mathbb{R}^k$ with location θ_0 and shape V_0. Then, (i) for some $h : \mathcal{P}_{k,\mathrm{tr}} \to [0, 1]$ that does not depend on V or on P,*

$$D_{\theta_0}(V, P) = (1 - \operatorname{pr}[\{\theta_0\}])\, h\left\{\frac{k(V_0^{-1/2}VV_0^{-1/2})}{\operatorname{tr}(V_0^{-1}V)}\right\}; \tag{3.2}$$

(ii) for $k = 2$,

$$D_{\theta_0}(V, P) = (1 - \mathrm{pr}[\{\theta_0\}]) \, \mathrm{pr}\left(Y_2 \geq \frac{1}{2} + \frac{1}{2}\left[1 - \det\left\{ \frac{2V_0^{-1}V}{\mathrm{tr}(V_0^{-1}V)} \right\} \right]^{1/2} \right), \qquad (3.3)$$

with Y_2 is Beta distributed with parameters $1/2$ and $1/2$.

The function h in this result does not depend on P, so that depth, under ellipticity, depends on P through V_0 and $\mathrm{pr}[\{\theta_0\}]$ only, with the dependence on $\mathrm{pr}[\{\theta_0\}]$ not affecting the induced ranking of shape matrices. It is easy to check that the explicit bivariate elliptical depth in (3.3) is compatible with the general results obtained above. While it seems very challenging to obtain an explicit expression for the function h in (3.2), numerical experiments lead us to conjecture that, irrespective of the dimension k, the mapping h is of the form $h(M) = g(\det M)$ for some function $g : \mathbb{R}^+ \to [0, 1]$.

The results of this section extend to the unspecified-location shape depth $D(V, P) = D_{\theta_P}(V, P)$ and to the corresponding regions $R(\alpha, P) = \{V \in \mathcal{P}_{k,\mathrm{tr}} : D(V, P) \geq \alpha\}$. Theorems 3.1 to 3.4 hold for any fixed θ and their unspecified-θ versions are simply obtained by substituting θ_P for θ throughout. In particular, the existence of an unspecified-location deepest shape matrix is guaranteed if P is smooth at θ_P, or, more generally, if $R(t_{\theta_P,P}, P)$ is non-empty. Under unspecified location, the shape V_P of P is then defined as the barycenter of the set of shape matrices maximizing $D(\cdot, P)$. In view of the affine equivariance of θ_P, i.e., $\theta_{P^{AX+B}} = A\theta_{P^X} + b$, the affine-invariance/equivariance properties

$$D\big(V_A, P^{AX+b}\big) = D(V, P^X), \quad R(\alpha, P^{AX+b}) = \big\{V_A : V \in R(\alpha, P)\big\}$$

follow directly from Theorem 3.6, to which we refer for the definition of V_A. Finally, Theorems 3.5 and 3.7 also readily extend to the unspecified-location case, since $\theta_P = \theta_0$ for any elliptical probability measure P with location θ_0. In particular, if P is elliptical with shape V_0, then the unspecified-θ shape depth $D(V, P)$ is uniquely maximized at $V = V_0$, if the distribution is not degenerate at a single point.

4 Consistency

When k-variate observations $X_1, \ldots, X_n$ are available, we define the sample fixed-θ depth of a shape matrix V as $D_\theta(V, P_n)$, where P_n is the empirical probability measure associated with $X_1, \ldots, X_n$, and its unspecified-location version as $D(V, P_n)$. In this section, we state a Glivenko–Cantelli-type result for these sample depths and investigate consistency of max-depth shape estimators.

Theorem 4.1. *Let P be a probability measure over $\mathbb{R}^k$ and let P_n denote the empirical probability measure associated with a random sample of size n from P. Then, (i) for any $\theta \in \mathbb{R}^k$, $\sup_{V \in \mathcal{P}_{k,\mathrm{tr}}} |D_\theta(V, P_n) - D_\theta(V, P)| \to 0$ almost surely as $n \to \infty$; (ii) if P is absolutely continuous with respect to the Lebesgue measure, then $\sup_{V \in \mathcal{P}_{k,\mathrm{tr}}} |D(V, P_n) - D(V, P)| \to 0$ almost surely as $n \to \infty$.*

We illustrate this result in the bivariate elliptical case associated with Theorem 3.7(ii). Figure 1 provides contour plots of $D_\theta(V, P)$ in terms of $V_{12}/(V_{11}V_{22})^{1/2}$ and V_{22}/V_{11}, for various bivariate, arbitrarily elliptical, probability measures. The sign nature of shape depth ensures that these contours, along with their empirical counterparts, are distribution-free in the class of elliptical distributions that do not charge the centre of symmetry. Figure 1 also reports the empirical contour plots obtained from a random sample of size $n = 800$ drawn from the corresponding bivariate normal distributions. Clearly, the results support the consistency in Theorem 4.1(i).

In § 3, the shape $V_{\theta,P}$ of P was defined as the barycenter of the collection of P-deepest shape matrices. In the empirical case, a natural estimator is the corresponding shape matrix V_{θ,P_n} computed from the empirical probability measure P_n associated with the sample at hand; existence here follows from the fact that $D_\theta(V, P_n)$ may only take values ℓ/n ($\ell = 0, 1, \ldots, n$). The same argument ensures the existence of the sample deepest shape V_{P_n} in the unspecified-location case. The sample Tukey median θ_{P_n} was one of the first affine-equivariant location estimators with a high breakdown point. It would therefore be interesting to investigate whether the affine-equivariant shape estimator V_{P_n}, parallel to the Maronna–Stahel–Yohai P-estimators of scatter, also has a high breakdown point (Tyler, 1994). Since this is beyond the scope of this paper, we focus on consistency of sample deepest shapes.

Theorem 4.2. *Let P be a probability measure over $\mathbb{R}^k$ and let P_n denote the empirical probability measure associated with a random sample of size n from P. (i) Fix $\theta \in \mathbb{R}^k$ and assume that $R_\theta(t_{\theta,P}, P)$ is non-empty. Then, $V_{\theta,P_n} \to V_{\theta,P}$ almost surely as $n \to \infty$. (ii) If P is absolutely continuous with respect to the Lebesgue measure, then $V_{P_n} \to V_P$ almost surely as $n \to \infty$.*

The specified-θ result in Theorem 4.2(i) holds in particular if P is smooth at θ. The unspecified-θ result requires a more stringent smoothness assumption, namely absolute continuity of P. This assumption, which is already present in Theorem 4.1(ii), is only needed to control the impact of replacing θ by θ_{P_n} in $D_\theta(V, P_n)$ and V_{θ,P_n}. Figure 1 also supports Theorem 4.2(i) since, in each sample considered, the sample deepest shape is close to its population counterpart.

5 Two applications

5.1 Choosing a shape matrix estimator in principal component analysis

There is a vast literature on scatter or shape estimation. Among the most famous estimators are the minimum covariance determinant scatters S_γ. Recall that, in the empirical case, S_γ is the covariance matrix with the smallest determinant among covariance matrices computed using only a proportion γ of the observations. The choice of the trimming proportion $1-\gamma$ is crucial, as the loss in efficiency can be very large if the trimming is excessive; see, for example, Croux and Haesbroeck (1999) or Paindaveine and Van Bever

(2014). Choosing γ is therefore difficult, as it should be taken large, but not so large as to incorporate outliers. In this section, we consider robust principal component analysis based on the shape estimators $\hat{V}_\gamma = kS_\gamma/\mathrm{tr}(S_\gamma)$ and show that Tyler shape depth allows the making of an informed choice of γ.

For several contamination proportions η, we independently generated $R = 500$ bivariate samples of $n = 800$ independent observations, each comprising $(1-\eta)n$ clean observations and ηn outliers. With X bivariate normal with zero mean and covariance matrix $\mathrm{diag}(4,1)$ and Y bivariate normal with mean $(0,\delta)^T$ and identity covariance matrix, the clean observations are equal to X in distribution, whereas the outliers are distributed, in equal proportions, as Y or $-Y$. Two simulations were conducted, one for $\delta = 4$ and one for $\delta = 5$; clearly, the former simulation provides a harder robustness problem than the latter. We consider estimating the first principal direction $e_1 = (1,0)^T$ of the uncontaminated distribution. For any $\gamma \in [0.5, 1]$, a natural estimator is, up to a sign, the first eigenvector $\hat{v}_\gamma$ of $\hat{V}_\gamma$. Denoting as $\hat{v}_{r,\gamma}$ this estimate in replication $r = 1, \ldots, R$, estimation performance can be measured through the mean squared error

$$\mathrm{MSE}_\gamma = \frac{1}{R} \sum_{r=1}^{R} (\Delta\alpha_{r,\gamma})^2,$$

where $\Delta\alpha_{r,\gamma} = \arccos(|e_1^T \hat{v}_{r,\gamma}|)$ is the angle between the population first eigendirection e_1 and its estimate $\hat{v}_{r,\gamma}$. Figure 2 plots MSE_γ as a function of γ; the Monte Carlo exercise was performed for every value of $\gamma \in \{0.5, 0.51, \ldots, 0.99, 1\}$. The results confirm that, for any contamination proportion η, a suitable value of γ should be identified. The optimal value $\gamma_0 = \arg\min_\gamma \mathrm{MSE}_\gamma$ basically coincides with $1-\eta$ in the easy case $\delta = 5$, whereas, in the harder one $\delta = 4$, γ_0 is slightly smaller than $1-\eta$ for large contaminations. This is no surprise: when outliers are hard to identify, the estimators $\hat{V}_\gamma$, with $\gamma \approx 1-\eta$, are likely to be based on some outliers, which will strongly affect the estimation performance.

In this framework, Tyler shape depth, as announced, may be very useful to select a suitable value of γ. We suggest choosing γ based on visual inspection of the curve $\mathcal{C} = \{(\gamma, D(\hat{V}_\gamma, P_{\gamma,n})) : \gamma \in [0.5, 1]\}$, where $P_{\gamma,n}$ denotes the empirical measure associated with the optimal subsample leading to $\hat{V}_\gamma$. The rationale is the following: for γ small, $D(\hat{V}_\gamma, P_{\gamma,n})$ will remain relatively high as long as no outlier is added to the optimal subsample. As γ increases and outliers are added in the computation of $\hat{V}_\gamma$, the depth $D(\hat{V}_\gamma, P_{\gamma,n})$ will sharply decrease, thereby forming a kink in $\mathcal{C}$. The selected γ for a given dataset, $\hat{\gamma}$, should therefore be the largest value for which $\mathcal{C}$ exhibit a stable behaviour. Figure 2 plots the curve $\mathcal{C}$ for the values of δ and η considered above and clearly illustrates the behaviour of the depth curves just described. When the outliers are easily identifiable, the kinks occur at γ_0, which coincides with $1-\eta$. In the harder case, where outliers and clean data tend to be mixed, the selected value $\hat{\gamma}$ is still remarkably close to γ_0. In conclusion, Tyler shape depth, and the ranking of shape matrices it provides, yield an effective visual tool that allows the selection of a sensible trimming proportion $1 - \gamma$ in a data-driven way when conducting, e.g., a principal component analysis.

8

5.2 Outlier detection

For each trading day between February 1st, 2015 and February 1st, 2017, we collected the Nasdaq Composite and S&P500 stock indices every five minutes and computed their returns, that is, the differences between two logs of consecutive index values. The returns on a given day form a bivariate dataset of usually 78 observations, though the number of observations varies due to missing values; days with fewer than 70 bivariate returns were discarded. The resulting dataset comprises $n = 38489$ observations on $D = 478$ trading days.

Our analysis studies the joint behaviour of the bivariate returns in order to determine which trading days are atypical. An important source of atypicality is associated with the overall scale of the bivariate returns, which alternate between periods of high and low volatility. Such deviations can easily be detected by comparing the trace of any scatter measure on intraday data with that on the whole dataset, so we focus instead on detecting atypical joint volatility, i.e., days on which the ratios of the marginal volatilities or the correlations between the returns deviate greatly from their global behaviour.

Let $\hat{V}_{\text{full}} = \hat{V}_{\hat{\gamma}}$ denote the minimum covariance determinant shape estimator computed from the full collection of n returns with maximal shape depth. More precisely, denoting as P_{full} the empirical distribution of the full collection of returns, let $\hat{\gamma} = \arg\max_{\gamma \in \Gamma} D(\hat{V}_\gamma, P_{\text{full}})$, for $\Gamma = \{0.5, 0.505, 0.51, \ldots, 0.995, 1\}$. The value obtained is $\hat{\gamma} = 0.825$, with corresponding depth $D(\hat{V}_{\hat{\gamma}}, P_{\text{full}}) = 0.497$. This high depth value ensures that $\hat{V}_{\text{full}}$ is an excellent proxy for the deepest shape matrix $\hat{V} = \arg\max_V D(V, P_{\text{full}})$, so the computation of $\hat{V}$ is unnecessary. Returns at the beginning of each trading period are known to be more volatile and should be discarded in shape estimation, so the robustness of $\hat{V}_{\text{full}}$ is an obvious asset: the value of $\hat{\gamma}$ allows us to adaptively discard days on which the volatility deviates from its global pattern. The procedure discarded more than half of the corresponding intra-day returns for 17 days, and, remarkably, 13 of these days lie within the two atypical periods mentioned in the next paragraph.

For each day $d = 1, \ldots, D$, we evaluated the depth $D(\hat{V}_{\text{full}}, P_d)$ of the global shape estimate with respect to the empirical distribution P_d of the bivariate returns on day d. The left panel of Figure 3 presents the depth values $D(\hat{V}_{\text{full}}, P_d)$. Vertical lines mark major events affecting the shape of the volatility, while the two greyed rectangles cover two periods during which the markets notoriously gave atypical returns: the first period follows the devaluation of the Yuan on August 11th, 2015 which saw rapid changes in the stock markets, including large devaluations on August 24th, event (a). The second period covers the beginning of 2016, when a slump in oil prices made stocks relying on oil very volatile compared to others. This resulted in atypical shape behaviour during January 22 – February 9; this last day, event (b), had the sharpest loss for the S&P500 index. The other events are (c) the decision of the European Central Bank on March 10th, 2016 to extend quantitative easing thereby slashing interest rates, which had a significant positive impact on both the Nasdaq and S&P500, but more pronounced for the latter, (d) the positive impact on the financial stocks following Fed officials' comments on the possibility of rate hike made on May 27, 2016, and (e) the aftermath of Donald Trump's election on November 9th. Detection of atypical observations was

9

achieved by flagging outliers with a depth so low that it is outside the box-and-whiskers plot. This resulted in 12 flagged days, each either being one of the events described above or lying in one of the greyed regions.

We also computed the halfspace shape depth $\text{HD}(\hat{V}_{\text{full}}, P_d)$ of the global estimate for each day d (Paindaveine and Van Bever, 2018). The right panel of Figure 3, a plot of $D(\hat{V}_{\text{full}}, P_d)$ versus $\text{HD}(\hat{V}_{\text{full}}, P_d)$, shows a clear positive association. Halfspace shape depth values seem to have a higher concentration than Tyler's, because the former maximizes a concept of scatter depth in scale and may be able to find scatter estimates better suited to the data. Indeed, a decrease in volatility in one of the marginals might be balanced by considering a scatter with a smaller scale which would have a large depth value. A byproduct of this is the fact that, when evaluating halfspace shape depth, the difficult maximisation step in scale seems to be crucial in correctly computing the depth ranking of the data, which can be affected by small deviations. More importantly, while events (a) and (b) receive low depth with respect to both concepts, only Tyler shape depth succeeds in flagging days associated with events (c) to (e) as outlying.

6 Hypothesis testing for shape

In the previous section, we presented two specific applications of shape depth. The concept also allows us to tackle more standard inference problems for shape, such as point estimation and hypothesis testing. Here, we consider testing $\mathcal{H}_0 : V = V_0$ against $\mathcal{H}_1 : V \neq V_0$ at level $\alpha \in (0,1)$, where $V_0 \in \mathcal{P}_{k,\text{tr}}$ is fixed, based on a random sample $X_1, \ldots, X_n$ from a k-variate elliptical distribution with known location θ and unknown shape V. In view of Theorem 3.5, a natural depth-based test, ϕ_D say, rejects the null for small values of $T_{\theta,n} = D_\theta(V_0, P_n)$, where P_n is the empirical distribution of $X_1, \ldots, X_n$. Since $T_{\theta,n}$ is discrete, achieving null size α in general requires randomization. The resulting test thus rejects the null hypothesis if $T_{\theta,n} < t_{\alpha,n}$, rejects the null hypothesis with probability $\gamma_{\alpha,n}$ if $T_{\theta,n} = t_{\alpha,n}$, and does not reject the null hypothesis if $T_{\theta,n} > t_{\alpha,n}$, where $t_{\alpha,n}$ is the null α-quantile of $T_{\theta,n}$ and $\gamma_{\alpha,n}$ is the amount of randomization. Under the assumption that P does not charge the center of the distribution, $T_{\theta,n}$ is distribution-free under the null hypothesis, which allows estimating $t_{\alpha,n}$ and $\gamma_{\alpha,n}$ arbitrarily well through simulations. Prior to applying the test below for $k = 2$ at level 5% with sample sizes $n = 200, 500$, these were estimated from 500,000 mutually independent standard normal samples for each sample size, yielding $\hat{t}_{0.05,200} = 0.40$, $\hat{\gamma}_{0.05,200} = 0.61$, $\hat{t}_{0.05,500} = 0.43$ and $\hat{\gamma}_{0.05,500} = 0.25$. Distribution-freeness of $T_{\theta,n}$ under the null hypothesis actually extends to the class of distributions with elliptical directions (Randles, 2000).

We performed two simulations in the bivariate case. The first considers the problem of testing the null hypothesis of sphericity $\mathcal{H}_0 : V_0 = I_2$ about $\theta = 0$ and compares the finite-sample powers of ϕ_D with those of some competitors. For each value of $\ell = 0, 1, \ldots, 6$ we generated $M = 3,000$ independent random samples $X_1, \ldots, X_n$ of size $n =$

500 from the normal distribution with location $\theta = 0$ and shape

$$V_{\ell,\xi} = I_2 + \ell\xi \begin{pmatrix} 1 & 0.5 \\ 0.5 & -1 \end{pmatrix}$$

and from the corresponding elliptical Cauchy distribution. The value $\ell = 0$ corresponds to the null hypothesis, whereas $\ell = 1, \ldots, 6$ provide increasingly severe alternatives. We took $\xi = 0.035$ and 0.045 for the normal and Cauchy samples in order to obtain roughly the same rejection frequencies in both cases.

For each sample, we carried out six tests at nominal level 5%: (i) the test ϕ_D described above; (ii) the Gaussian test from John (1972), or more precisely, its extension to elliptical distributions with finite fourth-order moments from Hallin and Paindaveine (2006); (iii) the sign test from Hallin and Paindaveine (2006); (iv) the Wald test based on the Tyler (1987) scatter matrix; (v)–(vi) the tests from Paindaveine and Van Bever (2014) based on the shape estimator $\hat{V}_\gamma$ in § 5, with $\gamma = 0.5$ and $\gamma = 0.8$. The tests (ii)–(vi) were performed based on their asymptotic null distribution. The rejection frequencies in Figure 4 reveal that ϕ_D performs very similarly to, although it may be slightly dominated by, the sign-based tests in (iii)–(iv) but performs very well under heavy tails, where it beats all other tests. As expected, the Gaussian test collapses under heavy tails and the minimum covariance determinant tests show low empirical power.

The second simulation tests $\mathcal{H}_0 : V = V_0$, with $V_0 = \text{diag}(2, 1/2)$ and specified location $\theta = 0$, and compares the tests above in terms of the level robustness (He et al., 1990). We considered mixture distributions $P^{X_{(\eta)}} = (1 - \eta)P^X + \eta P^Y$ with several contamination levels η. Here, X is a bivariate, normal or elliptical Cauchy, null random vector. The contamination random vector Y was chosen as follows: (a) Y has the same distribution as the vector obtained by rotating X about the origin by 45 degrees; (b) Y has the same elliptical distribution as X but its shape is $V = I_2$; (c) Y is obtained by multiplying the vector Y in (b) by four. The uncontaminated distribution P^X puts more mass along the horizontal axis. In (a), the contamination typically shows along the main bisector, whereas the contamination in (b) is uniformly distributed over the unit circle. As for (c), the contamination combines the directional feature of (b) with radial outlyingness. For each combination of distribution, normal or Cauchy, of contamination pattern, (a)–(c), and of contamination level, $\eta = 0, 0.025, 0.05, 0.1, 0.2, 0.25$ or 0.3, we generated 3,000 independent random samples $X_{(\eta)1}, \ldots, X_{(\eta)n}$ of size $n = 200$. Figure 5 plots the resulting rejection frequencies and reveals the very good robustness of the depth-based test ϕ_D; recall that, irrespective of η, the target rejection frequency is here 5%. In particular, ϕ_D always dominates its sign-based competitors (iii)–(iv). The minimum covariance determinant tests (v)–(vi) dominate ϕ_D in terms of robustness but exhibit poor finite-sample power. Radial outliers strongly affect the Gaussian test.

Summing up, the test associated with the proposed shape depth provides a good balance between efficiency and robustness. The improved robustness compared to its sign-based competitors is obtained at a very slight loss of power. Depth-based procedures can thus be defined for standard inference problems on shape, and will tend to perform as well as sign-based procedures. As shown in § 5, however, shape depth provides a

whole ranking of shape matrices that allows addressing less standard applications.

7 Perspectives for future research

The present work offers quite rich research perspectives. The asymptotic distributions of the sample depths $D_\theta(V, P_n)$ and $D(V, P_n)$ as well as those of the corresponding deepest shape estimators could be studied. Investigating the robustness properties of these shape estimators would also be of interest, in particular to see whether these estimators have a high breakdown point. Regarding hypothesis testing, it would be desirable to define depth-based tests for other shape problems, such as testing the null hypothesis that two populations share the same shape.

Another key point is related to computational aspects. Since Tyler shape depth was defined through halfspace depth, it can in principle be evaluated by using the numerous packages that are dedicated to halfspace depth. The definition of Tyler shape depth suggests that evaluation of this depth in dimension k requires the computation of halfspace depth in dimension k^2. Fortunately, redundancies in the random vector $W_{\theta,V}$ reduce the dimension from k^2 to $d_k = k(k+1)/2 - 1$ as shown by the following result.

Theorem 7.1. *Let $P = P^X$ be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Let $\mathrm{vech}(A)$ be the vector stacking the lower-diagonal entries of $A = (A_{ij})$ on top of each other and $\mathrm{vech}_0 A$ be $\mathrm{vech}(A)$ deprived of its first component. Then, $D_\theta(V, P) = D(0, P^{\tilde{W}_{\theta,V}}) = \inf_{u \in \mathcal{S}^{d_k-1}} \mathrm{pr}(u^T \tilde{W}_{\theta,V} \geq 0)$, with $\tilde{W}_{\theta,V} = \mathrm{vech}_0\{U_{\theta,V} U_{\theta,V}^T - (1/k) I_k\}$.*

It follows that, for $k = 2$ and 3, Tyler shape depth dominates its halfspace counterpart from Paindaveine and Van Bever (2018) from a computational point of view. There is, though, probably room for ad hoc algorithms to compute Tyler shape depth more efficiently. It would also be desirable to design iterative algorithms for the computation of deepest shape matrices.

A Appendix

As in the main manuscript, pr will refer to probability under the probability measure P at hand. However, it will sometimes be needed to emphasize the underlying probability measure, in which case we will write pr_P, pr_Q, pr_{P_n}, etc.

Many of the subsequent results require the following lemma.

Lemma A.1. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Write $C_{\theta,V}^M = \{x \in \mathbb{R}^k \setminus \{\theta\} : (u_{\theta,V}^x)^T M u_{\theta,V}^x \geq \mathrm{tr}(M)/k\}$ and $\tilde{C}_{\theta,V}^M = \{x \in \mathbb{R}^k : (u_{\theta,V}^x)^T M u_{\theta,V}^x \geq \mathrm{tr}(M)/k\}$, where $u_{\theta,V}^x$ is defined as $V^{-1/2}(x-\theta)/\|V^{-1/2}(x-\theta)\|$ if $x \neq \theta$ and as 0 otherwise. Then, for any $V \in \mathcal{P}_{k,\mathrm{tr}}$ and any $r \in \mathbb{R}$,*

$$D_\theta(V, P) = \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}(\tilde{C}_{\theta,V}^M) = \inf_{M \in \mathcal{M}_{k,F}^{\mathrm{all}}} \mathrm{pr}(\tilde{C}_{\theta,V}^M) = \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}(C_{\theta,V}^M) = \inf_{M \in \mathcal{M}_k^r} \mathrm{pr}(C_{\theta,V}^M),$$

where $\mathcal{M}_k^{\text{all}}$ collects the $k \times k$ symmetric matrices with arbitrary trace, $\mathcal{M}_k^r$ is the subset of $\mathcal{M}_k^{\text{all}}$ of matrices with trace r, and where $\mathcal{M}_{k,F}^{\text{all}}$ is the collection of matrices in $\mathcal{M}_k^{\text{all}}$ with Frobenius norm one.

Proof. It directly follows from the definition of Tyler shape depth that

$$D_\theta(V, P) = \inf_{v \in \mathbb{R}^{k^2}} \mathrm{pr}\big[\{x \in \mathbb{R}^k : v^T \mathrm{vec}\,\{u_{\theta,V}^x (u_{\theta,V}^x)^T - (1/k)I_k\} \geq 0\}\big].$$

When v runs over $\mathbb{R}^{k^2}$, the matrix M satisfying $v = \mathrm{vec}(M^T)$ runs over the collection $\mathcal{N}_k$ of $k \times k$ matrices. Since $(u_{\theta,V}^x)^T M u_{\theta,V}^x = (u_{\theta,V}^x)^T \{(M + M^T)/2\} u_{\theta,V}^x$ for any $M \in \mathcal{N}_k$, this yields

$$
\begin{aligned}
D_\theta(V, P) &= \inf_{M \in \mathcal{N}_k} \mathrm{pr}\big[\{x \in \mathbb{R}^{k^2} : \mathrm{tr}\big[M\{u_{\theta,V}^x (u_{\theta,V}^x)^T - (1/k)I_k\}\big] \geq 0\}\big] \\
&= \inf_{M \in \mathcal{N}_k} \mathrm{pr}\big(\tilde{C}_{\theta,V}^M\big) = \inf_{M \in \mathcal{M}_k^{\text{all}}} \mathrm{pr}\big(\tilde{C}_{\theta,V}^M\big). \quad\quad\quad (A.1)
\end{aligned}
$$

Letting $\mathbb{I}(A)$ be equal to one if condition A holds and to zero otherwise, this provides

$$D_\theta(V, P) = \inf_{M \in \mathcal{M}_k^{\text{all}}} \Big(\mathrm{pr}\big(C_{\theta,V}^M\big) + \mathrm{pr}[\{\theta\}]\mathbb{I}\{\mathrm{tr}(M) \leq 0\}\Big) = \inf_{M \in \mathcal{M}_k^{\text{all}}} \mathrm{pr}\big(C_{\theta,V}^M\big), \quad (A.2)$$

where we have used the fact that $\mathrm{pr}\big(C_{\theta,V}^M\big)$ is unchanged when M is replaced with $M + \lambda I_k$ for any $\lambda \in \mathbb{R}$. The same invariance property explains that the infimum over $\mathcal{M}_k^{\text{all}}$ in (A.2) may be replaced with an infimum over $\mathcal{M}_k^r$ for any r. Finally, the result for $\mathcal{M}_{k,F}^{\text{all}}$ follows from (A.1) by noting that $\tilde{C}_{\theta,V}^{\lambda M} = \tilde{C}_{\theta,V}^M$ for any $\lambda > 0$ and that $M = 0$ cannot provide the infimum in (A.1). The proof is complete. $\qquad\square$

Proof of Theorem 3.1. (i) Fix $M \in \mathcal{M}_k^{\text{all}}$ and consider $\tilde{C}^M = \tilde{C}_{0,I_k}^M$, where $\tilde{C}_{\theta,V}^M$ was defined in Lemma A.1. Since $\tilde{C}^M$ is closed, the mapping $P \mapsto \mathrm{pr}_P(\tilde{C}^M)$ is upper semi-continuous for weak convergence. Now, Slutzky's lemma entails that, as $d(V, V_0) \to 0$, the measure defined by $B \mapsto \mathrm{pr}(\theta + V^{1/2}B)$ converges weakly to the one defined by $B \mapsto \mathrm{pr}(\theta + V_0^{1/2}B)$. Therefore, $V \mapsto \mathrm{pr}(\theta + V^{1/2}\tilde{C}^M) = \mathrm{pr}(\tilde{C}_{\theta,V}^M)$ is upper semicontinuous at V_0. From Lemma A.1, we then obtain that

$$V \mapsto D_\theta(V, P) = \inf_{M \in \mathcal{M}_k^{\text{all}}} \mathrm{pr}(\tilde{C}_{\theta,V}^M),$$

is upper semicontinuous, as it is the infimum of a collection of upper semicontinuous functions. (ii) The result follows from the fact that the depth region $R_\theta(\alpha, P)$ is the inverse image of $[\alpha, \infty)$ by the upper semicontinuous function $V \mapsto D_\theta(V, P)$. (iii) Fix a sequence (V_n) in $\mathcal{P}_{k,\mathrm{tr}}$ such that $d(V_n, V_0) \to 0$. In view of Lemma A.1 again, we can, for any n, pick $M_n \in \mathcal{M}_{k,F}^{\text{all}}$ such that $P(\tilde{C}_{\theta,V_n}^{M_n}) \leq D_\theta(V_n, P) + \frac{1}{n}$. Compactness of $\mathcal{M}_{k,F}^{\text{all}}$ ensures that we can extract a subsequence (M_{n_ℓ}) of (M_n) that converges to $M_0 \in \mathcal{M}_{k,F}^{\text{all}}$. Writing $\mathbb{I}(B)$ for the indicator function of the set B, the dominated convergence theorem then yields that

$$\mathrm{pr}(\tilde{C}_{\theta,V_{n_\ell}}^{M_{n_\ell}}) - \mathrm{pr}(\tilde{C}_{\theta,V_0}^{M_0}) = \int_{\mathbb{R}^k} \big\{\mathbb{I}(\tilde{C}_{\theta,V_{n_\ell}}^{M_{n_\ell}}) - \mathbb{I}(\tilde{C}_{\theta,V_0}^{M_0})\big\} \, dP \to 0$$

13

as $\ell \to \infty$. The absolute continuity assumption on P guarantees that $\mathbb{I}(\tilde{C}_{\theta,V_{n_\ell}}^{M_{n_\ell}}) - \mathbb{I}(\tilde{C}_{\theta,V_0}^{M_0}) \to 0$ P-almost everywhere. Consequently,

$$\liminf_{n\to\infty} D_\theta(V_n, P) = \liminf_{n\to\infty} \mathrm{pr}\big(\tilde{C}_{\theta,V_n}^M\big) = \liminf_{\ell\to\infty} \mathrm{pr}\big(\tilde{C}_{\theta,V_{n_\ell}}^{M_{n_\ell}}\big) = \mathrm{pr}\big(\tilde{C}_{\theta,V_0}^{M_0}\big) \geq D_\theta(V_0, P).$$

We conclude that, if P is absolutely continuous with respect to the Lebesgue measure, then $V \mapsto D_\theta(V, P)$ is also lower semicontinuous, hence continuous. $\square$

The proof of Theorem 3.2 requires the following result.

Lemma A.2. *Let P be a probability measure over $\mathbb{R}^k$ and fix $\theta \in \mathbb{R}^k$. Write $u_\theta^x = (x - \theta)/\|x - \theta\|$ if $x \neq \theta$ and 0 otherwise. For any $c \geq 0$, further let $t_{\theta,P}(c) = \sup_{v\in\mathcal{S}^{k-1}} \mathrm{pr}(|v^T u_\theta^X| \leq c)$, so that $t_{\theta,P} = t_{\theta,P}(0) = \sup_{v\in\mathcal{S}^{k-1}} \mathrm{pr}\{v^T(X - \theta) = 0\}$. Then, $t_{\theta,P}(c) \to t_{\theta,P}$ as $c \to 0$.*

Proof of Lemma A.2. Since $t_{\theta,P}(c)$ is increasing in c over $[0, \infty)$ and is larger than or equal to $t_{\theta,P}$ for any positive c, we have that $\tilde{t}_{\theta,P} = \lim_{c\to 0} t_{\theta,P}(c)$ exists and is such that $\tilde{t}_{\theta,P} \geq t_{\theta,P}$. Now, fix a decreasing sequence (c_n) converging to 0 and consider an arbitrary sequence (v_n) such that

$$\mathrm{pr}(|v_n^T u_\theta^X| \leq c_n) \geq t_{\theta,P}(c_n) - (1/n).$$

Since $\mathcal{S}^{k-1}$ is compact, we can consider a subsequence (v_{n_ℓ}) that converges to $v_0 \in \mathcal{S}^{k-1}$; without loss of generality, we can of course assume that this subsequence is such that $(v_0^T v_{n_\ell})$ is an increasing sequence. Let then $C_\ell = \{v \in \mathcal{S}^{k-1} : v_0^T v \geq v_0^T v_{n_\ell}\}$. Clearly, C_ℓ is a decreasing sequence of sets with $\cap_\ell C_\ell = \{v_0\}$, so that

$$\lim_{\ell\to\infty} \mathrm{pr}\big[u_\theta^X \in \cup_{v\in C_\ell}\{y : |v^T y| \leq c_{n_\ell}\}\big] = \mathrm{pr}\big[u_\theta^X \in \{y : |v_0^T y| \leq 0\}\big] = \mathrm{pr}\big(|v_0^T u_\theta^X| = 0\big).$$

Now, for any ℓ, we have $\mathrm{pr}\big[u_\theta^X \in \cup_{v\in C_\ell}\{y : |v^T y| \leq c_{n_\ell}\}\big] \geq \mathrm{pr}(|v_{n_\ell}^T u_\theta^X| \leq c_{n_\ell}) \geq t_{\theta,P}(c_{n_\ell}) - (1/n_\ell)$, which implies that $t_{\theta,P} \geq \mathrm{pr}(|v_0^T u_\theta^X| = 0) \geq \tilde{t}_{\theta,P}$. $\square$

Proof of Theorem 3.2. Fix $V \in \mathcal{P}_{k,\mathrm{tr}}$ and denote as $\lambda_1(V)$ the largest eigenvalue of V. Similarly, denote $\lambda_k(V)$. Possible ties are unimportant below. Letting $v_1(V)$ and $v_k(V)$ be arbitrary corresponding unit eigenvectors, Lemma A.1 provides, with $M_V = v_1(V) v_1^T(V) \in \mathcal{M}_k^{\mathrm{all}}$

$$
\begin{aligned}
D_\theta(V, P) &\leq \mathrm{pr}\Big\{(u_{\theta,V}^X)^T M_V u_{\theta,V}^X \geq \mathrm{tr}(M_V)/k, X \neq \theta\Big\} \\
&= \mathrm{pr}\Big[k\lambda_1^{-1}(V)\{v_1^T(V)(X - \theta)\}^2 \geq \|V^{-1/2}(X - \theta)\|^2, X \neq \theta\Big] \\
&\leq \mathrm{pr}\Big\{\|V^{-1/2}(X - \theta)\|^2 \leq k\|X - \theta\|^2, X \neq \theta\Big\} \\
&= \mathrm{pr}\big(\|V^{-1/2} u_{\theta,I_k}^X\|^2 \leq k, u_\theta^X \neq 0\big),
\end{aligned}
$$

14

where we used the inequality $\lambda_1(V) \geq 1$ which follows from the constraint $\operatorname{tr}(V) = k$, and where u_θ^s is defined in Lemma A.2. Therefore,

$$D_\theta(V, P) \leq \operatorname{pr}\left[\lambda_1(V^{-1})\{v_1^T(V)u_\theta^X\}^2 \leq k\right] \leq t_{\theta,P}\left[\{k\lambda_k(V)\}^{1/2}\right]. \qquad (A.3)$$

Now, ad absurdum, take $\varepsilon > 0$ such that $R_\theta(t_{\theta,P} + \varepsilon, P)$ is unbounded. This implies that there exists a sequence (V_n) in $\mathcal{P}_{k,\mathrm{tr}}$ satisfying $D_\theta(V_n, P) \geq t_{\theta,P} + \varepsilon$ for any n and for which $d(V_n, I_k) \to \infty$. Since $\lambda_1(V_n) < \operatorname{tr}(V_n) = k$, we must have that $\lambda_k(V_n) \to 0$. Lemma A.2 and (A.3) then imply that $D_\theta(V_n, P) < t_{\theta,P} + \varepsilon$ for n large enough, a contradiction. Consequently, $R_\theta(\alpha, P)$ is bounded for any $\alpha > t_{\theta,P}$.

Now, Lemma C.1 in Paindaveine and Van Bever (2018) readily implies that a bounded subset of $\mathcal{P}_{k,\mathrm{tr}}$ is also totally bounded, in the sense that, for any $\varepsilon > 0$, it can be covered by finitely many balls of the form $B(V, \varepsilon) = \{\tilde{V} \in \mathcal{P}_{k,\mathrm{tr}} : d(\tilde{V}, V) < \varepsilon\}$. Part (i) of the result and Theorem 3.1(ii) thus entail that, for any $\alpha > t_{\theta,P}$, the region $R_\theta(\alpha, P)$ is closed and totally bounded. The result then follows from the completeness of the metric space $(\mathcal{P}_{k,\mathrm{tr}}, d)$. □

Proof of Theorem 3.3. Let $\alpha_* = \sup_{V \in \mathcal{P}_{k,\mathrm{tr}}} D_\theta(V, P)$. By assumption, $R_\theta(t_{\theta,P}, P)$ is non-empty. Thus, $\alpha_* \geq t_{\theta,P}$ and the result holds if $\alpha_* = t_{\theta,P}$. We may therefore assume that $\alpha_* > t_{\theta,P}$. For any n, pick then V_n in $R_\theta(\alpha_* - 1/n, P)$, where $R_\theta(\alpha, P)$ is defined as $\mathcal{P}_{k,\mathrm{tr}}$ for $\alpha < 0$. Fix $\varepsilon \in (0, \alpha_* - t_{\theta,P})$. For n large enough, all terms of the sequence (V_n) belong to the compact set $R_\theta(\alpha_* - \varepsilon, P)$; see Theorem 3.2. Thus, there exists a subsequence (V_{n_k}) that converges in $R_\theta(\alpha_* - \varepsilon, P)$, to V_* say. For any $\varepsilon' \in (0, \varepsilon)$, all (V_{n_k}) eventually belong to the closed set $R_\theta(\alpha_* - \varepsilon', P)$, so that $V_* \in R_\theta(\alpha_* - \varepsilon', P)$. Therefore, $\alpha_* - \varepsilon' \leq D_\theta(V_*, P) \leq \alpha_*$ for any such ε', which establishes the result. □

The proof of Theorem 3.4 requires the following preliminary result.

Lemma A.3. *For any $y \in \mathbb{R}^k$ and any $k \times k$ symmetric matrix M, the mapping $V \mapsto \operatorname{tr}(MV)y^T V^{-1} y$ is quasi-convex, that is, for any $V_a, V_b \in \mathcal{P}_{k,\mathrm{tr}}$ and any $t \in [0,1]$, $\operatorname{tr}(MV_t)y^T V_t^{-1} y \leq \max\{\operatorname{tr}(MV_a)y^T V_a^{-1} y, \operatorname{tr}(MV_b)y^T V_b^{-1} y\}$, with $V_t = (1-t)V_a + tV_b$.*

Proof. We treat two cases separately. (i) Assume first that $\operatorname{tr}(MV_a)\operatorname{tr}(MV_b) > 0$. Write

$$\frac{V_t}{\operatorname{tr}(MV_t)} = (1 - s_t)\frac{V_a}{\operatorname{tr}(MV_a)} + s_t\frac{V_b}{\operatorname{tr}(MV_b)}, \quad \text{with } s_t = \frac{t\operatorname{tr}(MV_b)}{(1-t)\operatorname{tr}(MV_a) + t\operatorname{tr}(MV_b)}.$$

Since $s_t \in [0,1]$, the weighted harmonic-arithmetic matrix inequality then shows that, for any $y \in \mathbb{R}^k$,

$$y^T\left\{\frac{V_t}{\operatorname{tr}(MV_t)}\right\}^{-1} y \leq y^T\left[(1 - s_t)\left\{\frac{V_a}{\operatorname{tr}(MV_a)}\right\}^{-1} + s_t\left\{\frac{V_b}{\operatorname{tr}(MV_b)}\right\}^{-1}\right]y$$

$$\leq \max\left[y^T\left\{\frac{V_a}{\operatorname{tr}(MV_a)}\right\}^{-1} y, y^T\left\{\frac{V_b}{\operatorname{tr}(MV_b)}\right\}^{-1} y\right],$$

as was to be showed; we refer to Lemma 2.1(vii) in Lawson and Lim, 2013 for the aforementioned inequality. (ii) Assume then that $\operatorname{tr}(MV_a)\operatorname{tr}(MV_b) \leq 0$. Without loss

15

of generality, assume that $\mathrm{tr}(MV_a) \leq 0$ and $\mathrm{tr}(MV_b) \geq 0$. If $\mathrm{tr}(MV_a) = \mathrm{tr}(MV_b) = 0$, then $\mathrm{tr}(MV_t) = 0$ for any t and the result trivially holds. Hence, we may assume that $\mathrm{tr}(MV_a) \neq 0$ or $\mathrm{tr}(MV_b) \neq 0$, which implies that $\mathrm{tr}(MV_{t_0}) = 0$ for a unique $t_0 \in [0,1]$. From continuity, pick then $\delta \in (0, 1 - t_0)$ such that, for any $t \in [t_0, t_0 + \delta)$,

$$\mathrm{tr}(MV_t)y^T V_t^{-1} y \leq \mathrm{tr}(MV_b)y^T V_b^{-1} y$$
$$\leq \max\left\{\mathrm{tr}(MV_a)y^T V_a^{-1} y, \mathrm{tr}(MV_b)y^T V_b^{-1} y\right\}.$$

By applying Part (i) of the proof with $V_{t_0+\delta}$ and V_b, we obtain that, for any $t \in [t_0 + \delta, 1]$,

$$\mathrm{tr}(MV_t)y^T V_t^{-1} y \leq \max\left\{\mathrm{tr}(MV_{t_0+\delta})y^T V_{t_0+\delta}^{-1} y, \mathrm{tr}(MV_b)y^T V_b^{-1} y\right\}$$
$$\leq \max\left\{\mathrm{tr}(MV_a)y^T V_a^{-1} y, \mathrm{tr}(MV_b)y^T V_b^{-1} y\right\}.$$

Since $\mathrm{tr}(MV_t)y^T V_t^{-1} y \leq 0 \leq \max\left\{\mathrm{tr}(MV_a)y^T V_a^{-1} y, \mathrm{tr}(MV_b)y^T V_b^{-1} y\right\}$ for any $t \in [0, t_0]$, the result follows. $\square$

Proof of Theorem 3.4. (i) Write $V_t = (1-t)V_a + tV_b$, where $V_a, V_b \in \mathcal{P}_{k,\mathrm{tr}}$ and $t \in [0,1]$ are fixed. First note that, letting $d_\theta^2(V) = (X - \theta)^T V^{-1}(X - \theta)$, Lemma A.1 yields

$$D_\theta(V, P) = \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}\left\{(X - \theta)^T V^{-1/2} M V^{-1/2}(X - \theta) \geq (1/k)\mathrm{tr}(M)d_\theta^2(V), X \neq \theta\right\}$$

$$= \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}\left\{(X - \theta)^T M(X - \theta) \geq (1/k)\mathrm{tr}(MV)d_\theta^2(V), X \neq \theta\right\}. \qquad \text{(A.4)}$$

Writing again $V_t = (1-t)V_a + tV_b$, Lemma A.3 thus yields that, for any $M \in \mathcal{M}_k^{\mathrm{all}}$,

$$\mathrm{pr}\left\{(X - \theta)^T M(X - \theta) \geq (1/k)\mathrm{tr}(MV_t)d_\theta^2(V_t), X \neq \theta\right\}$$
$$\geq \mathrm{pr}\left[(X - \theta)^T M(X - \theta) \geq (1/k)\max\{\mathrm{tr}(MV_a)d_\theta^2(V_a), \mathrm{tr}(MV_b)d_\theta^2(V_b)\}, X \neq \theta\right]$$
$$= \min\left[\mathrm{pr}\left\{(X - \theta)^T M(X - \theta) \geq (1/k)\mathrm{tr}(MV_a)d_\theta^2(V_a), X \neq \theta\right\},\right.$$
$$\left.\mathrm{pr}\left\{(X - \theta)^T M(X - \theta) \geq (1/k)\mathrm{tr}(MV_b)d_\theta^2(V_b), X \neq \theta\right\}\right]$$
$$\geq \min\{D_\theta(V_a, P), D_\theta(V_b, P)\}.$$

The result then follows from (A.4). (ii) If $V_a, V_b \in R_\theta(\alpha, P)$, then Part (i) of the result entails that $D_\theta(V_t, P) \geq \min\{D_\theta(V_a, P), D_\theta(V_b, P)\} \geq \alpha$, so that $V_t \in R_\theta(\alpha, P)$. $\square$

The proof of Theorem 3.5 requires both following lemmas.

Lemma A.4. *Let P be elliptical over $\mathbb{R}^k$ with location 0 and shape I_k. Then, $D_0(I_k, P) = (1 - \mathrm{pr}[\{0\}])\mathrm{pr}(U_1^2 > 1/k)$, where $U = (U_1, \ldots, U_k)^T$ is uniformly distributed over the unit sphere $\mathcal{S}^{k-1}$.*

Lemma A.5. *Let P be elliptical over $\mathbb{R}^k$ with location 0 and shape I_k. Then, for any $V \in \mathcal{P}_{k,\mathrm{tr}} \setminus \{I_k\}$, $D_0(V, P) < (1 - \mathrm{pr}[\{0\}])\mathrm{pr}(U_1^2 > 1/k)$, where $U = (U_1, \ldots, U_k)^T$ is uniformly distributed over $\mathcal{S}^{k-1}$.*

Proof of Lemma A.4. In the spherical setup considered, we have that, for any $M \in \mathcal{M}_k^{\mathrm{all}}$,

$$\mathrm{pr}\{(X^T M X)/\|X\|^2 \geq \mathrm{tr}(M)/k, \ X \neq 0\} = \mathrm{pr}\{U^T M U \geq \mathrm{tr}(M)/k\}\mathrm{pr}(X \neq 0),$$

where $U = (U_1, \ldots, U_k)^T$ is uniform over $\mathcal{S}^{k-1}$. Lemma A.1 then entails that

$$D_0(I_k, P) = (1 - \mathrm{pr}[\{0\}]) \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}\{U^T M U \geq \mathrm{tr}(M)/k\}.$$

Decomposing M into $O \Lambda O^T$, where O is a $k \times k$ orthogonal matrix and where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is a diagonal matrix, this yields

$$D_0(I_k, P) = (1 - \mathrm{pr}[\{0\}]) \inf_{\lambda \in \mathbb{R}^k} \mathrm{pr}\left(\sum_{\ell=1}^k \lambda_\ell U_\ell^2 \geq k^{-1} \sum_{\ell=1}^k \lambda_\ell \right)$$

$$= (1 - \mathrm{pr}[\{0\}]) \inf_{\lambda \in \mathbb{R}^k} \mathrm{pr}\left\{ \sum_{\ell=1}^k \lambda_\ell (U_\ell^2 - k^{-1}) \geq 0 \right\} = (1 - \mathrm{pr}[\{0\}]) \inf_{\lambda \in \mathbb{R}^k} p(\lambda).$$

By using successively the facts that $p(0) = 1$ and $p(\lambda) = p(\lambda/\|\lambda\|)$ for any $\lambda \in \mathbb{R}^k \setminus \{0\}$, we obtain

$$D_0(I_k, P) = (1 - \mathrm{pr}[\{0\}]) \inf_{\lambda \in \mathbb{R}^k \setminus \{0\}} p(\lambda) = (1 - \mathrm{pr}[\{0\}]) \inf_{\lambda \in \mathcal{S}^{k-1}} p(\lambda). \quad (\mathrm{A}.5)$$

The result then follows from Theorem 2 from Paindaveine and Van Bever (2017), that states that the last infimum in (A.5) is equal to $\mathrm{pr}(U_1^2 > 1/k)$. $\qquad\square$

Proof of Lemma A.5. Fix $V \in \mathcal{P}_{k,\mathrm{tr}}$ and let X be a random k-vector with $P = P^X$. Write $V = O \Lambda O^T$, where O is a $k \times k$ orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is a diagonal matrix with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. The affine invariance property from Theorem 3.6 entails that

$$D_0(V, P^X) = D_0(O^T V O, P^{O^T X}) = D_0(\Lambda, P^X).$$

Denoting by e_1 the first vector of the canonical basis of $\mathbb{R}^{k^2}$, we then have

$$D_0(V, P^X) = D_0(\Lambda, P^X) \leq \mathrm{pr}[e_1^T \mathrm{vec}\{U_{0,\Lambda} U_{0,\Lambda}^T - (1/k)I_k\} \geq 0]$$

$$= \mathrm{pr}[\{(U_{0,\Lambda})_1\}^2 \geq 1/k] = \mathrm{pr}[\{(U_{0,\Lambda})_1\}^2 \geq 1/k, \ X \neq 0]$$

$$= \mathrm{pr}\{\lambda_1^{-1} X_1^2 \geq (1/k)\textstyle\sum_{\ell=1}^k \lambda_\ell^{-1} X_\ell^2, \ X \neq 0\} \leq \mathrm{pr}\{X_1^2 \geq (1/k)\textstyle\sum_{\ell=1}^k X_\ell^2, \ X \neq 0\} \quad (\mathrm{A}.6)$$

$$= \mathrm{pr}(X_1^2/\|X\|^2 \geq 1/k, \ X \neq 0) = \mathrm{pr}(X \neq 0)\mathrm{pr}(U_1^2 \geq 1/k), \quad (\mathrm{A}.7)$$

where $U = (U_1, \ldots, U_k)^T$ is uniform over $\mathcal{S}^{k-1}$. To have $D_0(V, P^X) = \mathrm{pr}(X \neq 0)\mathrm{pr}(U_1^2 \geq 1/k)$, the inequality in (A.7) needs to be an equality, which requires that $\lambda_\ell = \lambda_1$ for all ℓ, hence that $V = I_k$. $\qquad\square$

We can now prove Theorem 3.5.

Proof of Theorem 3.5. Lemmas AA.4-AA.5 establish the result in the spherical case associated with $\theta_0 = 0$ and $V_0 = I_k$. For general values of θ_0 and V_0, note that $Y = V_0^{-1/2}(X - \theta_0)$ is elliptical with location 0, shape I_k, and satisfies $\mathrm{pr}(Y = 0) = \mathrm{pr}(X = \theta_0)$. Writing

$$W_0 = \frac{kV_0^{-1/2}VV_0^{-1/2}}{\mathrm{tr}(V_0^{-1/2}VV_0^{-1/2})}, \tag{A.8}$$

affine invariance then entails that

$$D_{\theta_0}(V, P^X) = D_0(W_0, P^Y) \leq D_0(I_k, P^Y) = D_0(W_0, P^Y) = D_{\theta_0}(V_0, P^X),$$

with equality if and only if $W_0 = I_k$, that is, if and only if $V = V_0$. □

Proof of Theorem 3.6. In the proof of Theorem 3.4, we showed that

$$D_\theta(V, P) = \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}\big\{(X - \theta)^T V^{-1/2} M V^{-1/2}(X - \theta)$$

$$\geq (1/k)\mathrm{tr}(M)(X - \theta)^T V^{-1}(X - \theta), X \neq \theta\big\}.$$

Using the fact that $V_A^{1/2} = k^{1/2}AV^{1/2}O/\{\mathrm{tr}(AVA^T)\}^{1/2}$ for some $k \times k$ orthogonal matrix O, this readily yields

$$D_{A\theta+b}(V_A, P) = \inf_{M \in \mathcal{M}_k^{\mathrm{all}}} \mathrm{pr}\big\{(X - \theta)^T V^{-1/2} O M O^T V^{-1/2}(X - \theta)$$

$$\geq (1/k)\mathrm{tr}(OMO^T)(X - \theta)^T V^{-1}(X - \theta), X \neq \theta\big\} = D_\theta(V, P),$$

as was to be shown. The affine-equivariance property of the depth regions readily follows. □

The proof of Theorem 3.7 requires the following lemma, whose proof is straightforward, hence is omitted.

Lemma A.6. *For any v_1, v_2 such that $v_1^2 + v_2^2 < 1$, we have*

$$\frac{(1 - v_1^2)^{1/2} - |v_2|}{(1 - v_1^2)^{1/2} + |v_2|} \leq \frac{(1 - v_1^2)^{1/2} + |v_2|}{(1 - v_1^2)^{1/2} - |v_2|} \leq \frac{1 + (v_1^2 + v_2^2)^{1/2}}{1 - (v_1^2 + v_2^2)^{1/2}}.$$

Proof of Theorem 3.7. (i) If $P = P^X$ is elliptical with location θ_0 and shape V_0, then $V_0^{-1/2}(X - \theta_0)$ is equal in distribution to RU, where U is uniformly distributed over the unit sphere $\mathcal{S}^{k-1}$ and is independent of the nonnegative random variable R. Theorem 3.6 then yields

$$D_{\theta_0}(V, P^X) = D_0(W_0, P^{RU}), \tag{A.9}$$

18

where W_0 is as in (A.8). Now, for any $\tilde{V} \in \mathcal{P}_{k,\mathrm{tr}}$, Lemma A.1 entails that

$$D_0(\tilde{V}, P^{RU}) = \inf_{M \in \mathcal{M}_k^0} \mathrm{pr}\big(U^T \tilde{V}^{-1/2} M \tilde{V}^{-1/2} U \geq 0, R > 0\big) \tag{A.10}$$

$$= \mathrm{pr}(R > 0) \inf_{M \in \mathcal{M}_k^0} \mathrm{pr}\big(U^T \tilde{V}^{-1/2} M \tilde{V}^{-1/2} U \geq 0\big) = \mathrm{pr}(R > 0) D_0(\tilde{V}, P^U).$$

Combining with (A.9), we obtain

$$D_{\theta_0}(V, P^X) = (1 - \mathrm{pr}_{P^X}[\{\theta_0\}]) D_0(W_0, P^U),$$

which establishes Part (i) of the result. (ii) Assume that $P = P^X$ is bivariate standard normal and fix $V \in \mathcal{P}_{2,\mathrm{tr}}$. We aim at evaluating

$$D_0(V, P^X) = \inf_{M \in \mathcal{M}_k^0} \mathrm{pr}\big(X^T V^{-1/2} M V^{-1/2} X \geq 0\big); \tag{A.11}$$

see (A.10). To do so, it will be convenient to parametrise V and the matrix M as

$$V = \begin{pmatrix} 1 + v_1 & v_2 \\ v_2 & 1 - v_1 \end{pmatrix} \quad \text{and} \quad M = m_1 \begin{pmatrix} 1 & m_2 \\ m_2 & -1 \end{pmatrix},$$

with $v_1^2 + v_2^2 < 1$ and $m_1 \neq 0$. Indeed, $m_1 = 0$ makes the probability in (A.11) equal to one, which cannot be the infimum. Decomposing $V^{-1/2} M V^{-1/2}$ into $O \Lambda O^T$, where O is a 2×2 orthogonal matrix and where $\Lambda = \mathrm{diag}\{\lambda_1(V^{-1}M), \lambda_2(V^{-1}M)\}$, with $\lambda_1(V^{-1}M) \geq \lambda_2(V^{-1}M)$, involves the eigenvalues of $V^{-1}M$ or, equivalently, of $V^{-1/2} M V^{-1/2}$, we have

$$D_0(V, P) = \inf_{(m_1, m_2) \in \mathbb{R}_0 \times \mathbb{R}} \mathrm{pr}\big\{\lambda_1(V^{-1}M)X_1^2 + \lambda_2(V^{-1}M)X_2^2 \geq 0\big\}, \tag{A.12}$$

where $X = (X_1, X_2)^T$ is still bivariate standard normal. Since $\lambda_1(-V^{-1}M) = -\lambda_2(V^{-1}M)$ for any $M \in \mathcal{M}_k^0$, we have

$$D_0(V, P) = \min \bigg[\inf_{(m_1, m_2) \in \mathbb{R}_0^+ \times \mathbb{R}} \mathrm{pr}\big\{\lambda_1(V^{-1}M)X_1^2 + \lambda_2(V^{-1}M)X_2^2 \geq 0\big\}, \tag{A.13}$$

$$\inf_{(m_1, m_2) \in \mathbb{R}_0^+ \times \mathbb{R}} \mathrm{pr}\big\{\lambda_1(V^{-1}M)X_1^2 + \lambda_2(V^{-1}M)X_2^2 \leq 0\big\}\bigg],$$

which allows us to restrict to positive values of m_1. We will show below that $\lambda_2(V^{-1}M) < 0 < \lambda_1(V^{-1}M)$ for any $M \in \mathcal{M}_k^0$. A direct computation shows that, for $m_1 > 0$,

$$\lambda_1(V^{-1}M) = \frac{m_1}{\det V}\Big[-(v_1 + m_2 v_2) + \{(v_1 + m_2 v_2)^2 + (1 + m_2^2)\det V\}^{1/2}\Big] > 0$$

and

$$\lambda_2(V^{-1}M) = \frac{m_1}{\det V}\Big[-(v_1 + m_2 v_2) - \{(v_1 + m_2 v_2)^2 + (1 + m_2^2)\det V\}^{1/2}\Big] < 0.$$

19

Since $f(m_2) = -\lambda_2(V^{-1}M)/\lambda_1(V^{-1}M)$ does not depend on m_1, (A.13) leads to

$$
\begin{aligned}
D_0(V,P) &= \min\left[\operatorname{pr}\left\{X_1^2/X_2^2 \ge \sup_{m_2 \in \mathbb{R}} f(m_2)\right\}, \operatorname{pr}\left\{X_1^2/X_2^2 \le \inf_{m_2 \in \mathbb{R}} f(m_2)\right\}\right] \\
&= \operatorname{pr}\left[X_1^2/X_2^2 \ge \max\left\{\sup_{m_2 \in \mathbb{R}} f(m_2),\ 1/\inf_{m_2 \in \mathbb{R}} f(m_2)\right\}\right]. \qquad \text{(A.14)}
\end{aligned}
$$

It is easy to check that f is differentiable over $\mathbb{R}$ with a derivative of the form $c_{v_1,v_2}(m_2)$ $(v_2 - v_1 m_2)$, where $c_{v_1,v_2}(m_2) > 0$ for any m_2, and that

$$
f(\pm\infty) = \lim_{m_2 \to \pm\infty} f(m_2) = \frac{(1 - v_1^2)^{1/2} \pm v_2}{(1 - v_1^2)^{1/2} \mp v_2}.
$$

We treat the cases $v_1 = 0$ and $v_1 \ne 0$ separately.

(a) Assume that $v_1 = 0$. If $v_2 = 0$, then $V = I_2$ and Theorem 3.5 establishes the result. If $v_2 \ne 0$, then f has no critical point and

$$
\sup_{m_2 \in \mathbb{R}} f(m_2) = \max\left\{f(-\infty), f(\infty)\right\} = \frac{1 + |v_2|}{1 - |v_2|}
$$

and

$$
\inf_{m_2 \in \mathbb{R}} f(m_2) = \min\left\{f(-\infty), f(\infty)\right\} = \frac{1 - |v_2|}{1 + |v_2|},
$$

so that (A.14) yields

$$
D_0(V,P) = \operatorname{pr}\left(\frac{X_1^2}{X_2^2} \ge \frac{1 + |v_2|}{1 - |v_2|}\right)
$$

$$
= \operatorname{pr}\left\{\frac{X_1^2}{X_2^2} \ge \frac{1 + (1 - \det V)^{1/2}}{1 - (1 - \det V)^{1/2}}\right\} = \operatorname{pr}\left\{Y_2 \ge \frac{1}{2} + \frac{1}{2}(1 - \det V)^{1/2}\right\},
$$

where we have used the fact that if Z has a $F(1,1)$ Fisher-Snedecor distribution, then $Z/(1+Z)$ has a Beta$(1/2, 1/2)$ distribution.

(b) Assume now that $v_1 \ne 0$. Then the only critical point of f is $m_2^{\mathrm{crit}} = v_2/v_1$, so that, irrespective of the fact that this critical point is a local minimum/maximum of f,

$$
\begin{aligned}
\sup_{m_2 \in \mathbb{R}} f(m_2) &= \max\left\{f(-\infty), f(\infty), f(m_2^{\mathrm{crit}})\right\} \\[2mm]
&= \max\left\{\frac{(1 - v_1^2)^{1/2} + |v_2|}{(1 - v_1^2)^{1/2} - |v_2|}, \frac{\operatorname{sign}(v_1) + (v_1^2 + v_2^2)^{1/2}}{\operatorname{sign}(v_1) - (v_1^2 + v_2^2)^{1/2}}\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
\inf_{m_2 \in \mathbb{R}} f(m_2) &= \min\left\{f(-\infty), f(\infty), f(m_2^{\mathrm{crit}})\right\} \\[2mm]
&= \min\left\{\frac{(1 - v_1^2)^{1/2} - |v_2|}{(1 - v_1^2)^{1/2} + |v_2|}, \frac{\operatorname{sign}(v_1) + (v_1^2 + v_2^2)^{1/2}}{\operatorname{sign}(v_1) - (v_1^2 + v_2^2)^{1/2}}\right\}.
\end{aligned}
$$

20

Lemma A.6 yields

$$\sup_{m_2 \in \mathbb{R}} f(m_2) = \frac{(1 - v_1^2)^{1/2} + |v_2|}{(1 - v_1^2)^{1/2} - |v_2|} \mathbb{I}(v_1 < 0) + \frac{1 + (v_1^2 + v_2^2)^{1/2}}{1 - (v_1^2 + v_2^2)^{1/2}} \mathbb{I}(v_1 > 0)$$

and

$$\inf_{m_2 \in \mathbb{R}} f(m_2) = \frac{-1 + (v_1^2 + v_2^2)^{1/2}}{-1 - (v_1^2 + v_2^2)^{1/2}} \mathbb{I}(v_1 < 0) + \frac{(1 - v_1^2)^{1/2} - |v_2|}{(1 - v_1^2)^{1/2} + |v_2|} \mathbb{I}(v_1 > 0),$$

hence also

$$\max\left\{ \sup_{m_2 \in \mathbb{R}} f(m_2) \,,\, 1 / \inf_{m_2 \in \mathbb{R}} f(m_2) \right\} = \frac{1 + (v_1^2 + v_2^2)^{1/2}}{1 - (v_1^2 + v_2^2)^{1/2}} = \frac{1 + (1 - \det V)^{1/2}}{1 - (1 - \det V)^{1/2}}.$$

Therefore, (A.14) finally provides

$$D_0(V, P) = \mathrm{pr}\left\{ \frac{X_1^2}{X_2^2} \geq \frac{1 + (1 - \det V)^{1/2}}{1 - (1 - \det V)^{1/2}} \right\} = \mathrm{pr}\left\{ Y_2 \geq \frac{1}{2} + \frac{1}{2}(1 - \det V)^{1/2} \right\}.$$

This proves the result for the case where P is bivariate standard normal. The general result then follows from Part (i) of the theorem. $\qquad\square$

Proof of Theorem 4.1. (i) Let P and Q be two probability measures over $\mathbb{R}^k$ and fix $V \in \mathcal{P}_{k,\mathrm{tr}}$. Fix $\varepsilon > 0$ and assume, without loss of generality, that $D_\theta(V, P) \leq D_\theta(V, Q)$. Lemma A.1 entails that there exists $M_0 \in \mathcal{M}_k^0$ such that $\mathrm{pr}_P\big(C_{\theta,V}^{M_0}\big) \leq D_\theta(V, P) + \varepsilon$, where we still use the notation $C_{\theta,V}^M = \big\{ x \in \mathbb{R}^k \setminus \{\theta\} : (u_{\theta,V}^x)^T M u_{\theta,V}^x \geq \mathrm{tr}(M)/k \big\}$. Consequently, using Lemma A.1 again,

$$|D_\theta(V, Q) - D_\theta(V, P)| = D_\theta(V, Q) - D_\theta(V, P)$$
$$\leq \mathrm{pr}_Q\big(C_{\theta,V}^{M_0}\big) - \mathrm{pr}_P\big(C_{\theta,V}^{M_0}\big) + \varepsilon \leq \sup_{C \in \mathcal{C}_\theta} |\mathrm{pr}_Q(C) - \mathrm{pr}_P(C)| + \varepsilon,$$

with $\mathcal{C}_\theta = \{C_{\theta,V}^M : M \in \mathcal{M}_k^0, V \in \mathcal{P}_{k,\mathrm{tr}}\}$. Since this holds for any $\varepsilon > 0$ and for any $V \in \mathcal{P}_{k,\mathrm{tr}}$, we have

$$\sup_{V \in \mathcal{P}_{k,\mathrm{tr}}} |D_\theta(V, Q) - D_\theta(V, P)| \leq \sup_{C \in \mathcal{C}_\theta} |\mathrm{pr}_Q(C) - \mathrm{pr}_P(C)|.$$

It thus only remains to show that $\mathcal{C}_\theta$ is a Vapnik-Chervonenkis class. To do so, note that $C_{\theta,V}^M = \big\{ x \in \mathbb{R}^k \setminus \{\theta\} : (x - \theta)^T V^{-1/2} M V^{-1/2}(x - \theta) \geq 0 \big\}$, so that $\mathcal{C}_\theta \subset \{D_{\theta,A} \cap (\mathbb{R}^k \setminus \{\theta\}) : A \in \mathcal{M}_k^{\mathrm{all}}\}$, with $D_{\theta,A} = \big\{ x \in \mathbb{R}^k : (x - \theta)^T A(x - \theta) \geq 0 \big\}$. Theorem 4.6 from Dudley (2014) implies that $\{D_{\theta,A} : A \in \mathcal{M}_k^{\mathrm{all}}\}$ is a Vapnik-Chervonenkis class $\mathcal{D}_\theta$. It then follows from Lemma 2.6.17(ii) in van der Vaart and Wellner (1996) that $\{D_{\theta,A} \cap (\mathbb{R}^k \setminus \{\theta\}) : A \in \mathcal{M}_k^{\mathrm{all}}\}$, hence also $\mathcal{C}_\theta$, is a Vapnik-Chervonenkis class. (ii) The proof is long and technical, but follows along the same lines as the proof of Theorem 2.2 in Paindaveine and Van Bever (2018), hence is omitted for the sake of brevity. $\qquad\square$

Proof of Theorem 4.2. (i) Recall from (3.1) that $V_{\theta,P}$ is defined as the barycentre of $R_\theta(\alpha_*, P)$, with $\alpha_* = \max_V D_\theta(V, P)$. The mapping $V \mapsto D_\theta(V, P)$ is upper semi-continuous (Theorem 3.1) and constant over $R_\theta(\alpha_*, P)$. Clearly, it is easy to define a mapping $V \mapsto \tilde{D}_\theta(V, P)$ that is upper semicontinuous, agrees with $V \mapsto D_\theta(V, P)$ in the complement of $R_\theta(\alpha_*, P)$, and for which $V_{\theta,P}$ is the unique maximizer. By using Theorem 4.1, it follows from Theorem 2.12 and Lemma 14.3 in Kosorok (2008) that $d(V_{\theta,P_n}, V_{\theta,P}) \to 0$ almost surely as $n \to \infty$. Part (i) of the result then follows from the fact that, in neighbourhoods of the form $\{V : d(V, V_{\theta,P}) < \varepsilon\}$, there exists a constant $C = C_\varepsilon$ such that $d_F(V, V_{\theta,P}) < C d(V, V_{\theta,P})$, where d_F is the Frobenius distance. (ii) The proof is entirely similar, hence is omitted. $\square$

Proof of Theorem 7.1. Let $L_{\theta,V} = U_{\theta,V} U_{\theta,V}^T - (1/k) I_k$. Since $(L_{\theta,V})_{11} = -\sum_{\ell=2}^{k} (L_{\theta,V})_{\ell\ell}$, there exists a $(d_k + 1) \times d_k$ full-rank matrix H_0 such that $\mathrm{vech}(L_{\theta,V}) = H_0 \,\mathrm{vech}_0(L_{\theta,V})$. Therefore, there exists a $k^2 \times d_k$ full-rank matrix H such that $W_{\theta,V} = \mathrm{vec}(L_{\theta,V}) = H\tilde{W}_{\theta,V}$. One can, for example, take $H = DH_0$, where D is the usual duplication matrix. It follows that

$$D_\theta(V, P) = D(0, P^{W_{\theta,V}}) = \inf_{u \in \mathbb{R}^{k^2}} \mathrm{pr}\big(u^T W_{\theta,V} \geq 0\big)$$

$$= \inf_{u \in \mathbb{R}^{k^2}} \mathrm{pr}\big\{(H^T u)^T \tilde{W}_{\theta,V} \geq 0\big\} = \inf_{v \in \mathbb{R}^{d_k}} \mathrm{pr}\big(v^T \tilde{W}_{\theta,V} \geq 0\big) = D(0, P^{\tilde{W}_{\theta,V}}),$$

where we used the fact that H^T has full column rank. $\square$

References

Bhatia, R. (2007) *Positive Definite Matrices*. Princeton, NJ: Princeton University Press.

Bhatia, R. and Holbrook, J. (2006) Riemannian geometry and matrix geometric means. *Linear Algebra Appl.*, **413**, 594–618.

Chen, M., Gao, C. and Ren, Z. (2018) Robust covariance and scatter matrix estimation under huber's contamination model. *Ann. Statist.*, **46**, 1932–1960.

Croux, C. and Haesbroeck, G. (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Anal.*, **71**, 161–190.

Dudley, R. M. (2014) *Uniform Central Limit Theorems*. Cambridge University Press, 2nd edition edn.

Dümbgen, L. (1998) On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.*, **50**, 471–491.

Hallin, M. and Paindaveine, D. (2006) Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Ann. Statist.*, **34**, 2707–2756.

He, X., Simpson, D. and Portnoy, S. (1990) Breakdown robustness of tests. *J. Amer. Statist. Assoc.*, **85**, 446–452.

Hettmansperger, T. P. and Randles, R. H. (2002) A practical affine equivariant multivariate median. *Biometrika*, **89**, 851–860.

John, S. (1972) The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, **59**, 169–173.

Kent, J. and Tyler, D. E. (1988) Maximum likelihood estimation for the wrapped Cauchy distribution. *J. Appl. Statist.*, **15**, 247–254.

Kosorok, M. R. (2008) *Introduction to Empirical Processes and Semiparametric Inference.* Springer Series in Statistics. New York: Springer.

Lawson, J. and Lim, Y. (2013) Weighted means and karcher equations of positive operators. *Proc. Natl. Acad. Sci. USA*, **110**, 15626–15632.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Ann. Statist.*, **27**, 783–858.

Maronna, R. A. (1976) Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, **4**, 51–67.

Mizera, I. (2002) On depth and deep points: a calculus. *Ann. Statist.*, **30**, 1681–1736.

Paindaveine, D. (2008) A canonical definition of shape. *Statist. Probab. Lett.*, **78**, 2240–2247.

Paindaveine, D. and Van Bever, G. (2014) Inference on the shape of elliptical distributions based on the MCD. *J. Multivariate Anal.*, **129**, 125–144.

— (2017) On the maximal halfspace depth of permutation-invariant distributions on the simplex. *Statist. Probab. Lett.*, **129**, 335–339.

— (2018) Halfspace depth for scatter, concentration and shape matrices. *Ann. Statist.*, **46**, 3276–3307.

Randles, R. H. (2000) A simpler, affine-invariant, multivariate, distribution-free sign test. *J. Amer. Statist. Assoc.*, **95**, 1263–1268.

Sirkiä, S., Taskinen, S., Oja, H. and Tyler, D. E. (2009) Multivariate tests and estimates for shape based on spatial signs and ranks. *J. Nonparametr. Stat.*, **21**, 155–176.

Tatsuoka, K. S. and Tyler, D. E. (2000) On the uniqueness of *S*-functionals and *M*-functionals under nonelliptical distributions. *Ann. Statist.*, **28**, 1219–1243.

Tukey, J. W. (1975) Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, 523–531. Canad. Math. Congress, Montreal, Que.

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

— (1994) Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.*, **22**, 1024–1044.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

Zhang, J. (2002) Some extensions of Tukey's depth function. *J. Multivariate Anal.*, **82**, 134–165.

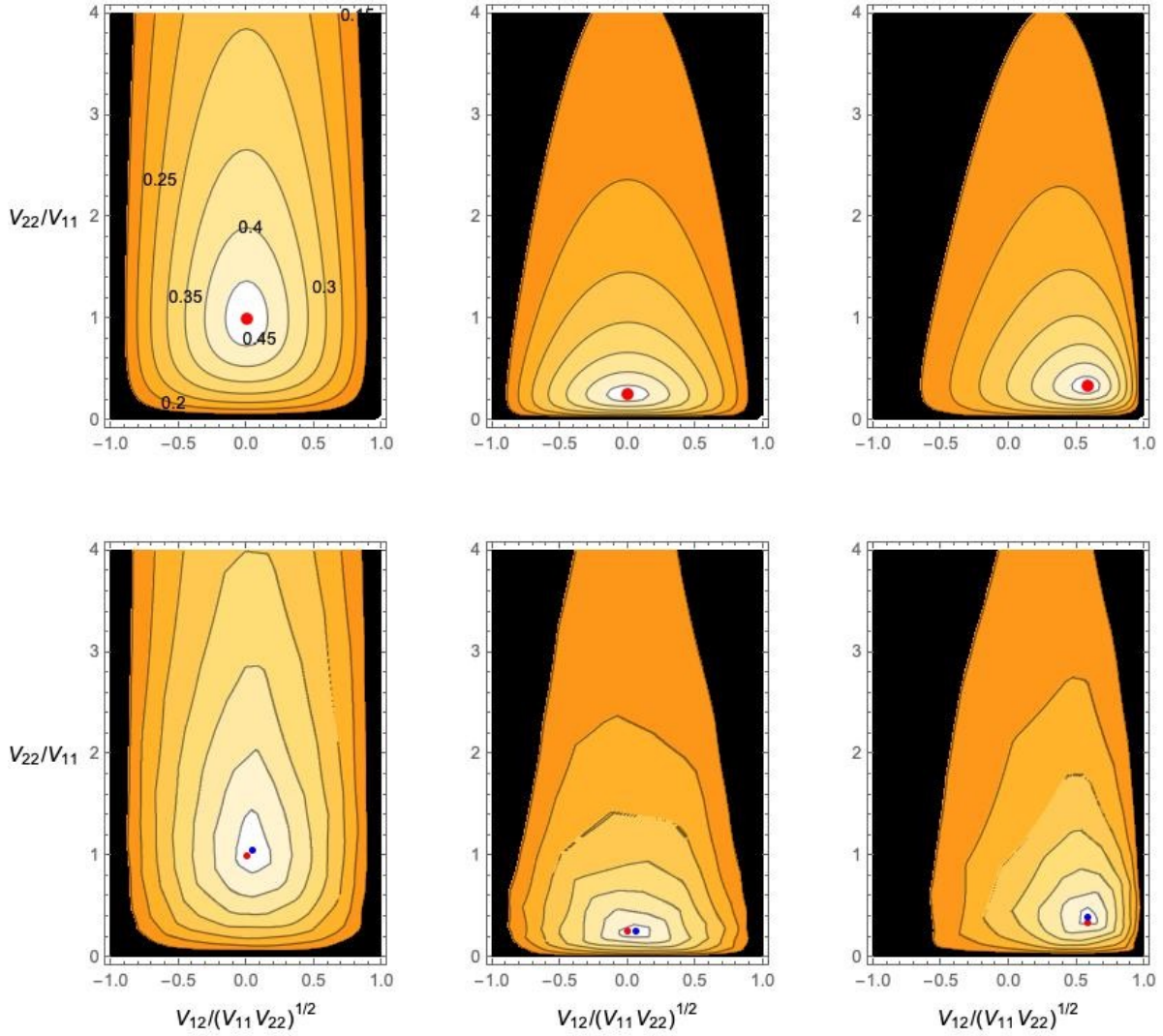Zuo, Y. and Serfling, R. (2000) General notions of statistical depth function. *Ann. Statist.*, **28**, 461–482.

Figure 1: (First row:) Contour plots of $D_\theta(V, P)$ in terms of $V_{12}/(V_{11}V_{22})^{1/2}$ and V_{22}/V_{11}, where P refers to an arbitrary bivariate elliptical probability measure with location θ and with shape $V_A = \mathrm{diag}(1, 1)$ (left), $V_B = \mathrm{diag}(1.6, 0.4)$ (center), or V_C with diagonal vector $(1.5, 0.5)^T$ and off-diagonal elements 0.5 (right), that is so that $\mathrm{pr}[\{\theta\}] = 0$. (Second row:) The corresponding contour plots of $D_0(V, P_n)$, where P_n is the empirical probability measure associated with a random sample of size $n = 800$ from the centered bivariate normal with shape V_A (left), V_B (center), and V_C (right). The true shapes $V_{0,P}$ and sample deepest shapes V_{0,P_n} are marked in red and in blue, respectively.
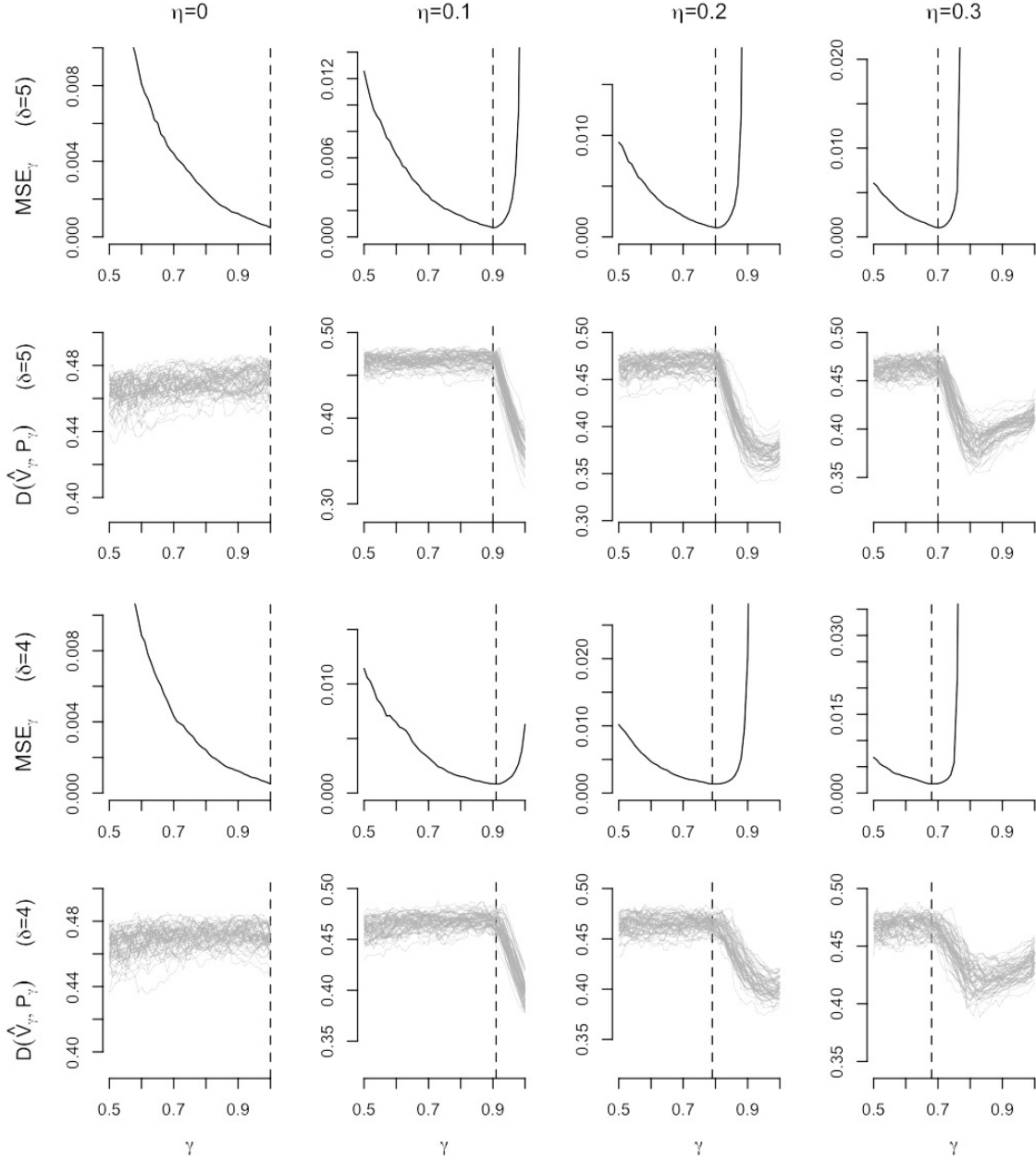
25

Figure 2: (First row:) Plot of the mapping $\gamma \mapsto \text{MSE}_\gamma$ in the easy case $\delta = 5$ and for contamination proportions $\eta \in \{0, 0.1, 0.2, 0.3\}$. (Second row). Plot of 50 random curves $\mathcal{C} = \{(\gamma, D(\hat{V}_\gamma, P_{\gamma,n})) : \gamma \in [0.5, 1]\}$, still for $\delta = 5$ and for the same contamination proportions. (Third and fourth rows:) The corresponding plots for the harder case associated with $\delta = 4$; see § 5 for details. Each panel shows a vertical line at $\gamma_0 = \arg\min_\gamma \text{MSE}_\gamma$.
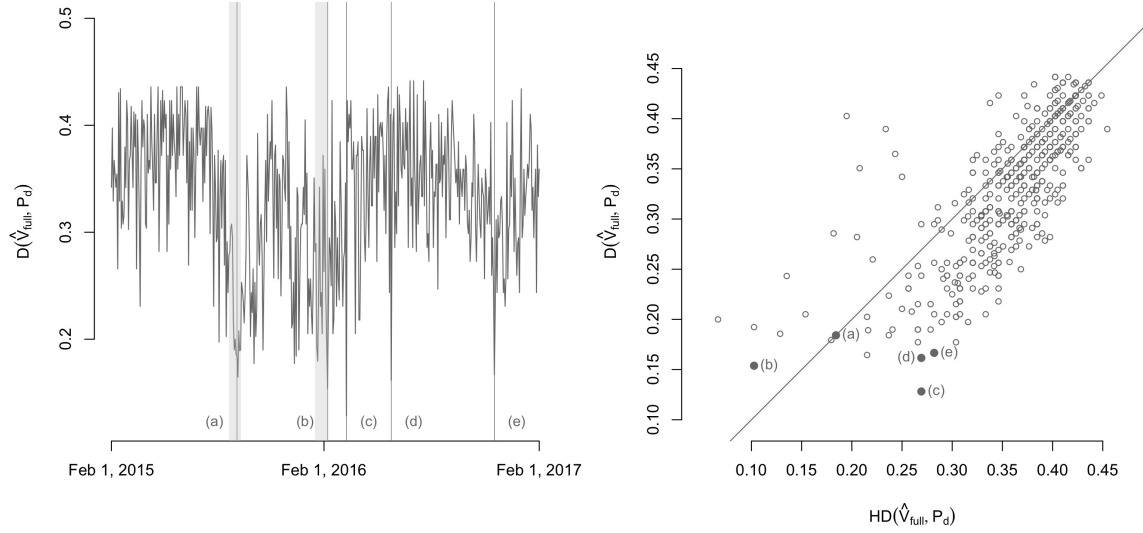
Figure 3: (Left:) Plot of $D(\hat{V}_{\text{full}}, P_d)$ as a function of d. Events (a) to (e) are described in § 5.2. (Right:) Plot of $D(\hat{V}_{\text{full}}, P_d)$ vs $HD(\hat{V}_{\text{full}}, P_d)$ for each trading day d. Events from the left panel are highlighted using the same colour.
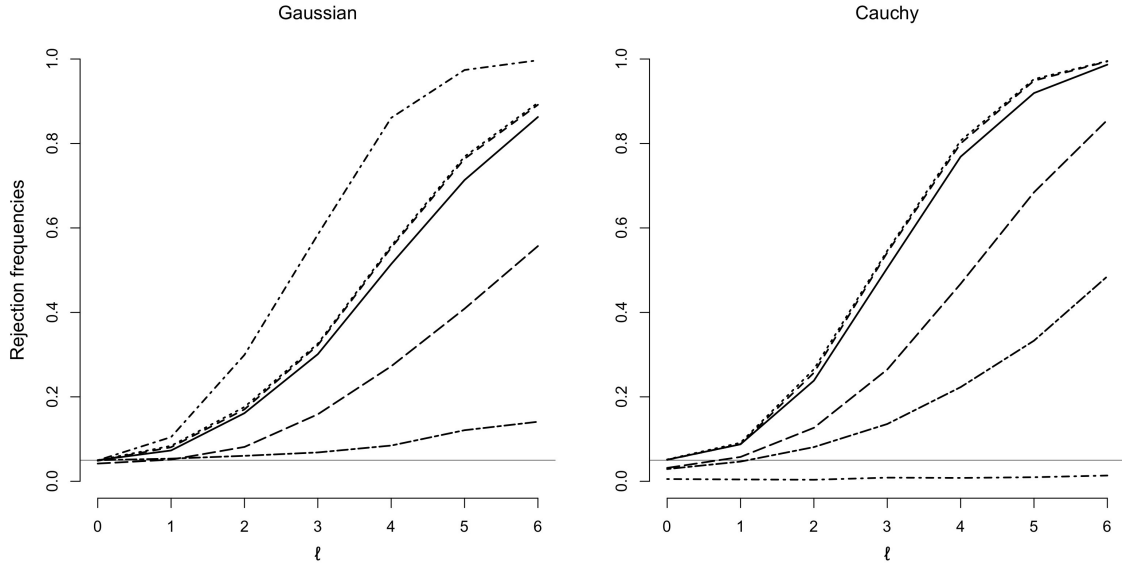

Figure 4: Rejection frequencies, under bivariate normal (left) and elliptical Cauchy (right) densities, of six tests of sphericity: the Gaussian test (dot-dashed curve), the sign test (dotted curve), the test based on Tyler's scatter matrix (dashed curve), the depth-based test (solid curve), and two minimum covariance determinant-based tests based on different trimming proportions (long dashed and short-long dashed curves for trimming proportion 0.2 and 0.5, respectively). Results are based on 3,000 replications and the sample size is $n = 500$. See § 6 for details.
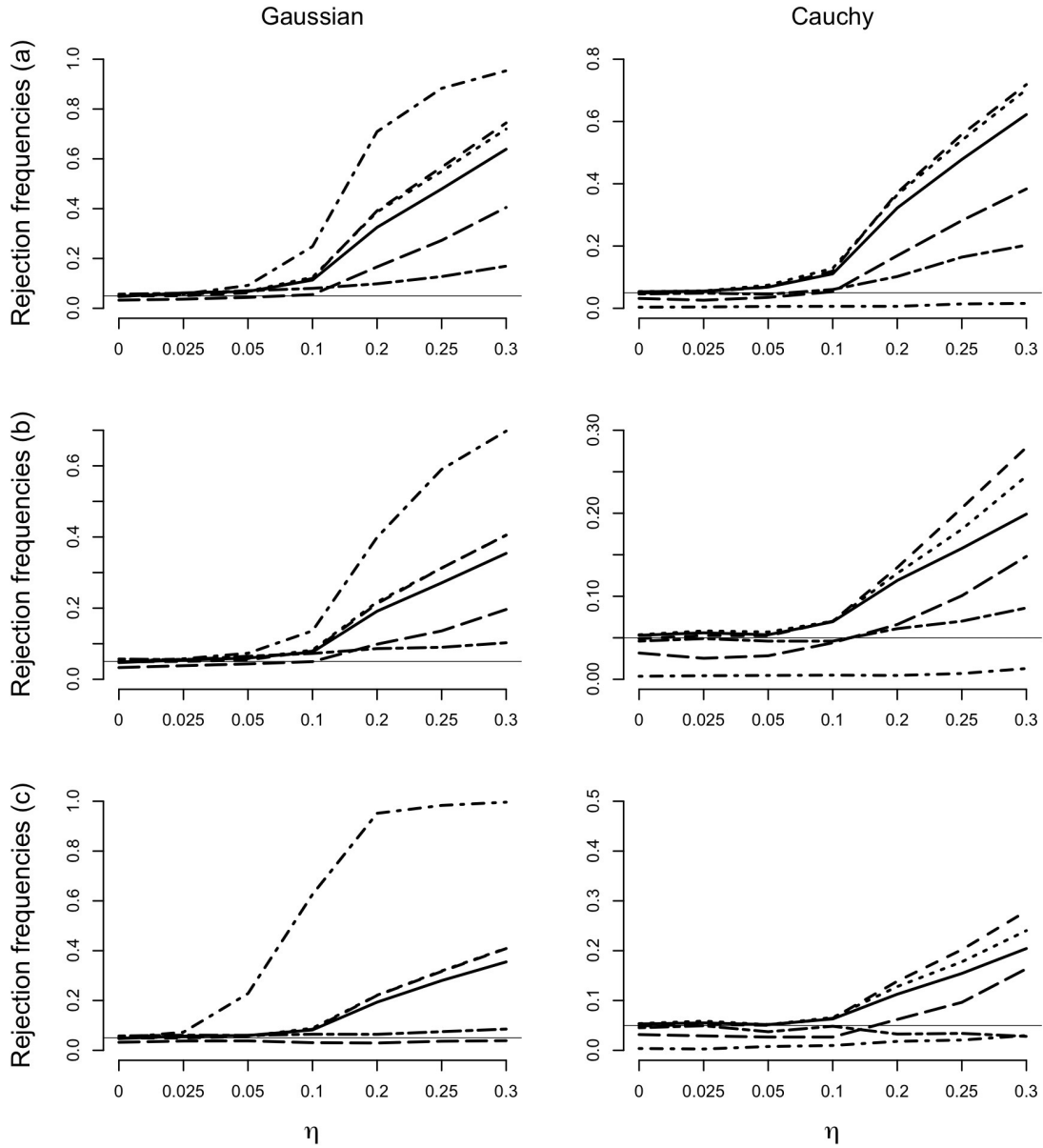
Figure 5: Null rejection frequencies, as a function of the contamination level η, of the same six tests (using the same line types) as in Figure 4, under bivariate normal (left) and elliptical Cauchy (right) densities. The labels (a)–(c) refer to the three contaminations patterns considered; see § 6 for details. Results are based on 3,000 replications and the sample size is $n = 200$.