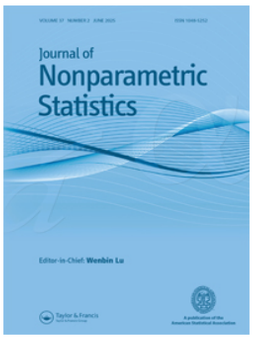

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: www.tandfonline.com/journals/gnst20

Measure of shape for object data

J. Virta

To cite this article: J. Virta (12 Jun 2025): Measure of shape for object data, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2025.2517775

To link to this article: https://doi.org/10.1080/10485252.2025.2517775

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2025.2517775

Taylor & Francis
Taylor & Francis Group

Measure of shape for object data

J. Virta

Department of Mathematics and Statistics, University of Turku, Turku, Finland

ABSTRACT

Object data analysis is concerned with statistical methodology for datasets whose elements reside in an arbitrary, unspecified metric space. In this work we propose the object shape, a novel measure of shape/symmetry for object data. The object shape is easy to compute and interpret, owing to its intuitive interpretation as interpolation between two extreme forms of symmetry. As one major part of this work, we apply object shape in various metric spaces and show that it manages to unify several pre-existing, classical forms of symmetry. We also propose a new visualisation tool called the peeling plot, which allows using the object shape for outlier detection and principal component analysis of object data.

ARTICLE HISTORY
Received 24 May 2024
Accepted 1 June 2025

KEYWORDS
Descriptive statistics; metric space; non-Euclidean data; object data analysis; symmetry

MATHEMATICS SUBJECT CLASSIFICATIONS
62G05; 62H05; 62H12

1. Introduction

Modern applications routinely produce datasets, such as images, functions or graphs, that do not take the familiar form of a $n \times p$ data matrix. An emerging trend in the literature is to approach the analysis of such data in a type-agnostic way, by not specifying the actual sample space, but simply assuming that the observed *objects*, $X_1, \ldots, X_n$, reside in a general metric space $(\mathcal{X}, d)$. Statistical methodology which depends on the object sample only through the interobject distances, $d(X_i, X_j)$, is collectively known as object data analysis, or metric statistics, see, e.g. (Bhattacharya and Patrangenaru 2003; Lyons 2013; Dubey and Müller 2019, 2022; Virta et al. 2022; Virta 2023; Zhu and Müller 2023; Zhou and Müller 2024; Bulté and Sørensen 2024; Dubey et al. 2024; Lin and Chen 2024) for works taking this viewpoint. The obvious advantage of such an approach is that any object data method is automatically applicable to all forms of object data, be they images, correlation matrices or graphs, making object data analysis highly universal.

The purpose of the current work is to propose a new descriptive statistic, the *object shape*, for measuring the shape (level of symmetry), of an object data distribution (sample). Given a distribution P taking values in a fixed metric space $(\mathcal{X}, d)$ we define the object shape of P as

$$S(P) = \frac{E\{d(X_1, X_2)^2 d(X_1, X_3)^2\}}{E\{d(X_1, X_2)^4\}}, \tag{1}$$

CONTACT J. Virta joni.virta@utu.fi, jomivi@utu.fi Department of Mathematics and Statistics, 20014 University of Turku, Turku, Finland

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/10485252.2025.2517775.

where X_1, X_2, X_3 are independent draws from P. While the quantity $S(P)$ appears rather simple, we show in the sequel that it both has a very intuitive interpretation and manages to unify several well-known, established forms of shape/symmetry. Note that we primarily use the word 'symmetry' in this work not in its standard statistical role as an antonym to 'skewness', but in a wider mathematical sense of an object (distribution) being invariant to a certain set of actions. In order to summarise our findings, we first define a terminology that will be used throughout the paper: we say that the three points $x_1, x_2, x_3 \in \mathcal{X}$ are d-linearly connected if the triangle inequality $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$ achieves equality for some ordering of the indices 1, 2, 3. In a Euclidean space, this definition thus corresponds to the standard concept of points on a line, whereas, e.g. on the unit sphere it is satisfied by points restricted on a great circle and sharing the same hemisphere. With this definition out of the way, as our main contributions about $S(P)$ we show in the current paper that:

(i) The object shape $S(P)$ takes values in the interval $[1/2, 1]$. Moreover, $S(P) = 1/2$ if and only if three i.i.d. objects drawn from P are almost surely d-linearly connected, and $S(P) = 1$ if and only if three i.i.d. objects drawn from P almost surely form an equilateral 'triangle' (in the metric d). In other words, the object shape interpolates between two extreme forms of symmetry: the smaller $S(P)$ is, the more 'one-dimensional' the distribution P is, and the larger $S(P)$ is, the more evenly spread out P is in the space $\mathcal{X}$.
(ii) The object shape $S(P)$ reduces to well-known measures of shape/symmetry under specific choices of $(\mathcal{X}, d)$ and P. For example, (a) for elliptical distributions in a Euclidean space, the object shape coincides with the classical notion of sphericity: the larger $S(P)$ is, the more the equidensity contours of P resemble spheres, (b) for discrete distributions over a finite set, the object shape reduces to measuring the uniformity of the distribution P, as in the classical Pearson's χ^2-test. Further examples are given in Section 3.
(iii) The scenario-specific upper bounds for the object shape $S(P)$ can be used to devise hypothesis tests for the related symmetries. For example, besides recovering the Pearson's χ^2-test for discrete data as mentioned already above, we obtain a new test of uniformity on the unit circle, see Section 3.

The diversity and difficulty of visualising object data means that data summaries such as the proposed object shape can be seen to be even more important for object data than in traditional data analysis. Thus far, the literature on descriptive tools for object data has focused heavily on location estimation, particularly in conjunction with statistical depth measures, see (Cholaquidis et al. 2023; Dai and Lopez-Pintado 2023; Geenens et al. 2023; Virta 2023). The concept of Frechét mean, i.e. the minimiser of the map $\mu \mapsto E\{d(X, \mu)^2\}$, also falls in this category, see, e.g. (Bhattacharya and Patrangenaru 2003; Dubey and Müller 2019). The most prominent examples of descriptive statistics measuring variation or co-variation of object data are the Frechét variance, i.e. $E\{d(X, \mu_0)^2\}$ where μ_0 is the Frechét mean, the distance covariance (Lyons 2013) which also allows characterising independence (i.e. lack of co-variation) in metric spaces of negative type, and metric covariance (Dubey and Müller 2020). In some sense multidimensional scaling (Kruskal and Wish 1978), a classical

method of object data visualisation, can also be seen as a member of this class as it is essentially based on principal component analysis, a second-order method.

Using classical statistical terminology, none of the methods listed in the previous paragraph thus measure *shape*, i.e. the properties of P beyond location or scale. Whereas, as we argue in Section 2, our object shape $S(P)$ is invariant to both location and scale (at least when $\mathcal{X}$ is structured enough to admit such concepts), meaning that calling it a measure of shape is warranted. Indeed, as far as we are aware, our object shape is the first intrinsic measure of shape for object data. By 'intrinsic' we mean that the object shape relies entirely on the geometry of the actual data space, and does not require the (approximate) embedding of the objects in an auxiliary Euclidean space. This, in conjunction with the clear interpretation and fast computation of $S(P)$ in practise, means that our proposal offers a valuable addition to the toolbox of descriptive statistical analysis of object data. This viewpoint is further explored in Section 5 where we compare object shape to various 'non-intrinsic' embedded measures of shape on a collection of 30 different real object data sets.

In addition to establishing the theoretical properties of $S(P)$ listed earlier, as one of our contributions we show how object shape gives rise to a new visualisation tool for object data which we call the peeling plot. This plot is constructed by removing observations one-by-one from an object data set in such a way that the object shape $S(P)$ is maximised/minimised at each step. The peeling plot is then obtained as the scatter plot between the indices of the removed observations and the value of $S(P)$ at each step. If minimisation is used, the peeling plot leads into a novel method for conducting principal component analysis for object data. Whereas, under maximisation, the peeling plot can be used for detecting outlying objects, see the examples in Section 4.

2. Main result

Let $(\Omega, \mathcal{F}, Q)$ be a probability space and let $(\mathcal{X}, d)$ be a complete and separable metric space. In the sequel, all probability distributions on $\mathcal{X}$ are taken to be measurable with respect to the Borel sets of $\mathcal{X}$. Let P denote a probability distribution on X. Throughout this work, we assume that the following condition holds true.

Assumption 2.1: *The distribution P is such that, for independent $X_1, X_2 \sim P$,*

(i) *There exists a point $a \in \mathcal{X}$ such that $E\{d(X_1, a)^4\} < \infty$.*
(ii) *The random variable $d(X_1, X_2)$ is not almost surely equal to zero.*

Assumption 2.1(i) can be seen as the object data equivalent of assuming that the fourth moment of a real random variable is finite. In particular, by Cauchy-Schwarz inequality and the triangle inequality, it guarantees that also the moments $E\{d(X_1, X_2)^2 d(X_1, X_3)^2\}$ and $E\{d(X_1, X_2)^4\}$ exist as finite. Whereas, Assumption 2.1(ii) simply requires that the distribution $\mathcal{P}$ is not a trivial Dirac point mass.

Our main point of interest is the object shape $S(P)$, defined in (1), which is well-defined under Assumption 2.1. Before its proper study, the form of $S(P)$ already offers us some hints on its meaning and interpretation: Firstly, the denominator and numerator in (1) have the same 'degree' (four), implying that $S(P)$ is a dimensionless quantity. Secondly, if $(\mathcal{X}, d)$ is a normed space such that $d(X_1, X_2) = \|X_1 - X_2\|$ for some norm $\|\cdot\|$, it is

clear that $S(P)$ is invariant to translation and scaling of the distribution P (i.e. invariant to maps $X_i \mapsto aX_i + b$). These two observations together lead us to expect that S measures a deeper, scale and location-invariant aspect of the distribution P. Such quantities (that are essentially controlled by moments beyond the first two) are referred to as measures of 'shape' in classical statistics, hence our proposed name for the concept.

As our main result of this section, we next show that $S(P)$ is constrained to the interval $[1/2, 1]$ and give geometric characterisations for the endpoints of this interval.

Theorem 2.1: *Under Assumption* 2.1, $S(P) \in [1/2, 1]$. *Furthermore, letting* X_1, X_2, X_3 *be independently drawn from* P,

(i) $S(P) = 1/2$ *if and only if*

$$P(X_1, X_2, X_3 \text{ are } d\text{-linearly connected}) = 1.$$

(ii) $S(P) = 1$ *if and only if*

$$P(d(X_1, X_2) = d(X_2, X_3) = d(X_3, X_1)) = 1.$$

By Theorem 2.1, it is indeed reasonable to regard $S(P)$ as a measure of symmetry or shape of the distribution P. The quantity $S(P)$ achieves its maximal value if and only if three points drawn i.i.d. from P almost surely form an equilateral triangle, in the sense of the metric d. While such an event is unlikely in any practically reasonable scenario, intuitively it means that large values of $S(P)$ can only be achieved when P is spread out sufficiently evenly in the space $\mathcal{X}$. Later in Section 3 we show that in several situations S is maximised when the distribution in question is 'uniform' in some specific sense, implying that the previous intuition indeed holds true. Whereas, the minimal value of $S(P)$ is achieved precisely when three i.i.d. objects drawn from P are almost surely d-linearly connected (i.e. achieve equality in the triangle inequality). In particular, the constant value $S(P) = 1/2$ is obtained by every distribution P when $(\mathcal{X}, d)$ is the real line $\mathbb{R}$ equipped with the Euclidean distance. Hence, $S(P)$ can also be seen as a continuous relaxation of the concept of dimension: its minimal value corresponds to distributions P charging all their mass on a set whose points are d-linearly connected, i.e. essentially a one-dimensional object, whereas $S(P)$ is maximised when P spreads out to the full space $\mathcal{X}$, corresponding, in some sense, to a maximal dimension.

Being essentially a moment-based quantity, the sample estimation of (1) is simple. For a distribution P, we let the notation P_n denote the empirical distribution of a random sample $X_1, \ldots, X_n$ of size n drawn from P. Hence, a natural estimator of $S(P)$ is

$$S(P_n) = \frac{\frac{1}{n^3}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n d(X_i, X_j)^2 d(X_i, X_k)^2}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n d(X_i, X_j)^4} = \frac{1_n^{\mathrm{T}} B_n^2 1_n}{n\|B_n\|^2}, \tag{2}$$

where $1_n \in \mathbb{R}^n$ is a vector of ones, the $n \times n$ matrix $B_n = \{d(X_i, X_j)^2\}$ contains all pairwise squared distances between the observations and $\|\cdot\|$ denotes the Frobenius norm. Consequently, given a matrix B_n, computing $S(P_n)$ is an operation of complexity $\mathcal{O}(n^3)$. By the standard results on U-statistics (Lee 1990), it is simple to check that this estimator is consistent, $S(P_n) = S(P) + o_P(1)$. Furthermore, the convergence rate of $1/\sqrt{n}$ can

be obtained for the error term by taking on stronger moment conditions for $d(X_1, a)$ in Assumption 2.1.

3. Example scenarios

3.1. Introduction

In the following subsections, we illustrate the object shape under four different combinations of metric space and distributional family. In each case, we show that the range of values of $S(P)$ is actually a strict subset of the general interval $[1/2, 1]$ derived in Theorem 2.1. Additionally, we characterise the minimal and maximal values of $S(P)$ in terms of the parameters of the underlying distributional families and derive a hypothesis test of symmetry corresponding to the distribution with the largest value for $S(P)$ in each case. Our reasons for selecting these four particular scenarios are: (i) they demonstrate the wide range of situations $S(P)$ can be applied in, (ii) they allow for clear interpretations of the object shape in terms of the corresponding parameters, and (iii) closed-form solutions are available in them.

We are also aware that there are more specialised tools available for studying each of these data scenarios. Indeed, our task here is not to derive the most efficient solutions or tests for each given scenario, but rather to show that multiple seemingly disjoint concepts in classical statistics are unified by and expressible via the object shape $S(P)$, further highlighting its usefulness in the analysis of object data. Similarly, while we derive several new tests for various null hypotheses of symmetry in this section, we have not benchmarked them against competitors here. This is because our main interest is not on any specific test but rather on showing that, in the first place, various new tests can straightforwardly be constructed based on the object shape.

3.2. Elliptical distributions

In this subsection, we take $(\mathcal{X}, d)$ to be the p-dimensional Euclidean space equipped with the Euclidean metric. Let $R \sim F_R$ where F_R is some fixed distribution on the non-negative real numbers. Let U be independent of R and obey the uniform distribution on the unit sphere $\mathbb{S}^{p-1}$. We denote by $P_{\mu,\Sigma}$ the distribution of the random vector

$$X = \mu + R\Sigma^{1/2}U,$$

where $\mu \in \mathbb{R}^p$ and the matrix $\Sigma^{1/2} \in \mathbb{R}^{p\times p}$ is non-zero and positive semidefinite. Thus $P_{\mu,\Sigma}$ is a member of the family of non-degenerate *elliptical distributions* (Fang et al. 1990) with the fixed radial distribution F_R. This family of distributions has two parameters, μ and Σ, which can be interpreted similarly as the mean vector and the covariance matrix of the multivariate normal distribution. If F_R admits a density, then the equidensity contours of $P_{\mu,\Sigma}$ are ellipses with the directions and lengths of their axes determined by the eigenvectors and eigenvalues of Σ, respectively. The following theorem confirms the intuition that $P_{\mu,\Sigma}$ should be at its most symmetric when all eigenvalues of Σ are equal, i.e. when $P_{\mu,\Sigma}$ is *spherical* and its equidensity contours spheres.

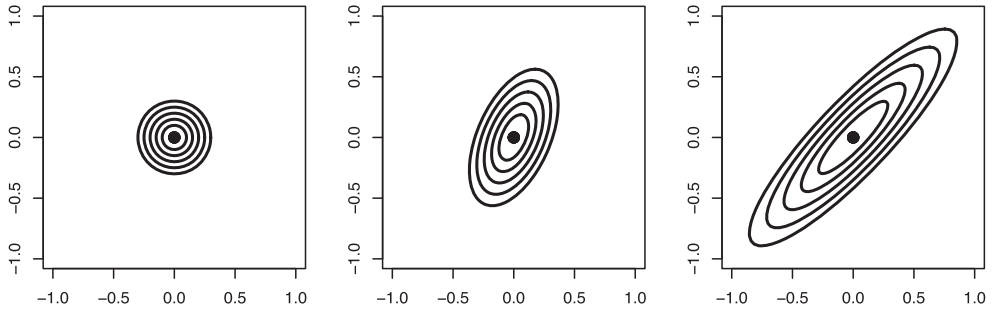

Figure 1. The equidensity contours of bivariate t-distribution with 8 degrees of freedom under three different covariance structures. The values of $S(P_{\mu,\Sigma})$ corresponding to the three panels are 0.600, 0.554 and 0.512, respectively.

Theorem 3.1: *Assume that* $E(R^4) < \infty$ *and denote*

$$u_R = \frac{1}{2} + \frac{p-1}{p(\beta_R + 1) + 2},$$

where $\beta_R = E(R^4)/\{E(R^2)\}^2$. *Then,* $(\mu, \Sigma) \mapsto S(P_{\mu,\Sigma})$ *takes values in* $[1/2, u_R]$ *and*

(i) $S(P_{\mu,\Sigma}) = 1/2$ *if and only if* Σ *has rank 1.*
(ii) $S(P_{\mu,\Sigma}) = u_R$ *if and only if* $\Sigma = \lambda I_p$ *for some* $\lambda > 0$.

By Theorem 3.1, $S(P_{\mu,\Sigma})$ essentially measures the relative contribution of the first principal component of $P_{\mu,\Sigma}$ to its total variation: The more concentrated $P_{\mu,\Sigma}$ is on a one-dimensional subspace, the smaller the value of $S(P_{\mu,\Sigma})$. Conversely, if $P_{\mu,\Sigma}$ does not favour any particular direction, the maximal value u_R is reached. These points are further illustrated in Figure 1 where we have assumed that $p = 2$ and $R^2 \sim F(2, 8)$, making X have a bivariate t-distribution with 8 degrees of freedom (Fang et al. 1990). The values of $S(P_{\mu,\Sigma})$ given in the figure caption clearly correspond to the shape of the ellipses in the manner described in Theorem 3.1.

We additionally note the following consequences of Theorem 3.1: (i) When $p = 1$, the distribution $P_{\mu,\Sigma}$ equals that of its first principal component, making the two cases of Theorem 3.1 equal, $u_R = 1/2$ regardless of F_R, and $S(P_{\mu,\Sigma})$ a constant function. (ii) As a function of the distribution F_R, the upper bound u_R in Theorem 3.1 is maximal when $E(R^4) = \{E(R^2)\}^2$, i.e. when F_R is a Dirac point mass, in which case $u_R = p/(p+1)$. This maximal value is reached if and only if $P_{\mu,\Sigma}$ is a uniform distribution on a sphere of arbitrary radius in $\mathbb{R}^p$, making these distributions the most symmetric, in the sense of the object shape, out of all p-variate elliptical distributions.

We now move to consider a hypothesis test for the null hypothesis of sphericity, $\Sigma = \lambda I_p$ for some $\lambda > 0$, giving the asymptotic null distribution of $S(P_{\mu,\lambda I_p,n})$, the object shape of a sample of size n from $P_{\mu,\lambda I_p}$, as our next result.

Theorem 3.2: *Let* $\mu \in \mathbb{R}^p$, $\lambda > 0$ *be fixed and assume that* $E(R^8) < \infty$. *Then, as* $n \to \infty$,

$$\sqrt{n}\left\{u_R - S(P_{\mu,\lambda I_p,n})\right\} \rightsquigarrow \mathcal{N}(0, \sigma_R^2),$$

where u_R is as in Theorem 3.1 *and the constant σ_R^2 is a function of the moments of F_R and is given in the proof of this theorem.*

The test statistic in Theorem 3.2 is somewhat impractical in the sense that, for unknown F_R, its use requires estimating several higher moments of the radial variate R needed to compute u_R and σ_R^2. And while this is technically possible, it makes the test statistic $S(P_{\mu,\lambda I_p,n})$ less attractive in practise compared to its well-established competitors that do not necessitate this, see, e.g. (Hallin and Paindaveine 2006).

3.3. von Mises distribution on the circle

Let $\mathcal{X} = \mathbb{S}^1$ be the unit circle and let P_κ denote the centered von Mises distribution with the concentration parameter $\kappa \geq 0$, see, e.g. (Mardia and Jupp 2000). The probability density function of P_κ thus equals

$$f_\kappa(x) = \frac{1}{2\pi I_0(\kappa)} \mathrm{e}^{\kappa \cos(x)}, \quad x \in [0, 2\pi),$$

where I_0 is the modified Bessel function of the first kind and order 0. The density f_κ has been plotted in Figure 2 for various values of κ. A random variable $X \sim P_\kappa$ thus corresponds to a random point (its angle) on the unit circle and the larger the value of κ is, the more concentrated the distribution is around the zero angle. As $\kappa \to \infty$, the distribution approaches a point mass and, conversely, in the other extreme, $\kappa = 0$, we obtain the uniform distribution on the unit circle.

We next equip $\mathcal{X}$ with the metric $d(x, y) = \sqrt{1 - \cos(x - y)}$, see Lemma B.2 in the supplementary material for the proof that d satisfies the triangle inequality. By the preceding discussion, it is reasonable to expect that $\kappa \mapsto S(P_\kappa)$ is maximised at the uniform distribution, $\kappa = 0$, and the following result shows that this is indeed the case.

Theorem 3.3: *The map $\kappa \mapsto S(P_\kappa)$ is decreasing in $(0, \infty)$. Moreover, $S(P_0) = 2/3$ and $S(P_\kappa) \to 1/2$ as $\kappa \to \infty$.*

Theorem 3.3 reveals that (a) the least-dimensional member of the family P_κ (in the sense of having the smallest possible $S(P_\kappa)$) is the point mass distribution achieved in the limit $\kappa \to \infty$, and (b) the upper endpoint 2/3 of the range of $S(P_\kappa)$ is achieved at the uniform distribution. Hence, a test of uniformity on the unit circle can be devised based on the object shape of the empirical distribution $P_{\kappa,n}$ of a sample of size n from P_κ, whose null distribution we derive next.

Theorem 3.4: *We have, as $n \to \infty$,*

$$n \left\{ \frac{2}{3} - S(P_{0,n}) \right\} \rightsquigarrow \frac{2}{18} \chi_2^2 + \frac{1}{18} \chi_2^2,$$

where the two χ_2^2-variates are independent.

Up to our best knowledge, the hypothesis test corresponding to the null distribution in Theorem 3.4 appears to be novel and not equivalent to any of the classical tests of uniformity, see (Mardia and Jupp 2000, Section 6.3). This claim is further backed up by the

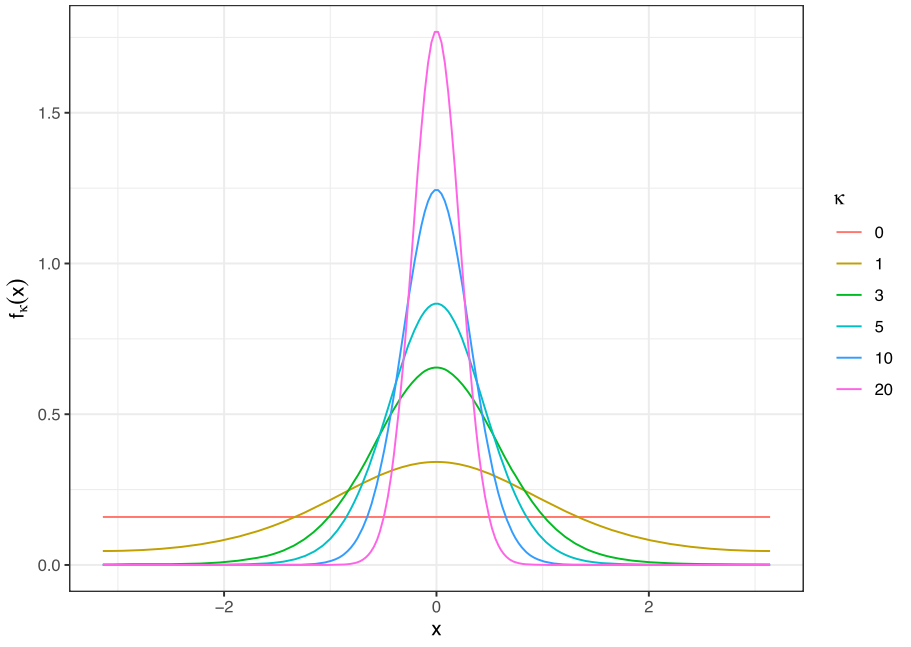

Figure 2. The density functions f_κ of the von Mises distribution with various values of the concentration parameter κ. The curves show that, for $\kappa = 0$, the distribution reduces to the uniform distribution on the unit sphere, whereas when $\kappa \to \infty$, the probability mass concentrates more and more around the zero angle.

fact that the test statistic $S(P_n)$, while simple to compute using the object data viewpoint in formula (2), has a reasonably complex expression as a function of the transformed variates $(\cos(X_i), \cos^2(X_i), \sin(X_i), \cos(X_i)\sin(X_i))$, see the proof of Theorem 3.4. Comparing the results of Theorems 3.2 and 3.4 we also observe that the convergence rate and the type of limiting distribution of $S(P_n)$ depends on the particular scenario we are in. In Theorem 3.2 the convergence is with rate $1/\sqrt{n}$ to a normal distribution and in Theorem 3.4, where the limiting distribution is a sum of scaled χ^2-distributions, the convergence is with rate $1/n$. The appearance of chi-squared distributions in the limit is natural in any context where the estimator converges to a constant which simultaneously acts as a bound for it. Correspondingly, the non-negativity of the limiting distribution in Theorem 3.4 provides evidence that $S(P_{0,n}) \leq 2/3$ might hold for all sample distributions $P_{0,n}$ which is a much stronger claim than Theorem 3.3 which states that the inequality holds for all von Mises distributions P_θ. Note that an analogous 'extension' cannot be obtained from Theorem 3.2 where the convergence to a limiting normal distribution indicates that $S(P_{\mu,\lambda I_p,n}) > u_R$ occurs with positive probability for large n. This is not in contradiction with Theorem 3.1 as the sample distributions $P_{\mu,\lambda I_p,n}$ are almost surely not perfectly elliptical.

While the metric $d(x, y) = \sqrt{1 - \cos(x - y)}$ was chosen for its tractable analytical properties, it is also of interest to examine how $S(P_\kappa)$ behaves when the metric is chosen to be something more conventional. Hence, we conclude this section by carrying out a Monte Carlo study where the values of $S(P_\kappa)$ for $\kappa = 0, 0.1, \ldots, 10$ are estimated with (2) from random samples of size $n = 5000$ and the standard angular distance (the shortest distance between two points along the circle) is used as a metric. The resulting plot of $S(P_\kappa)$

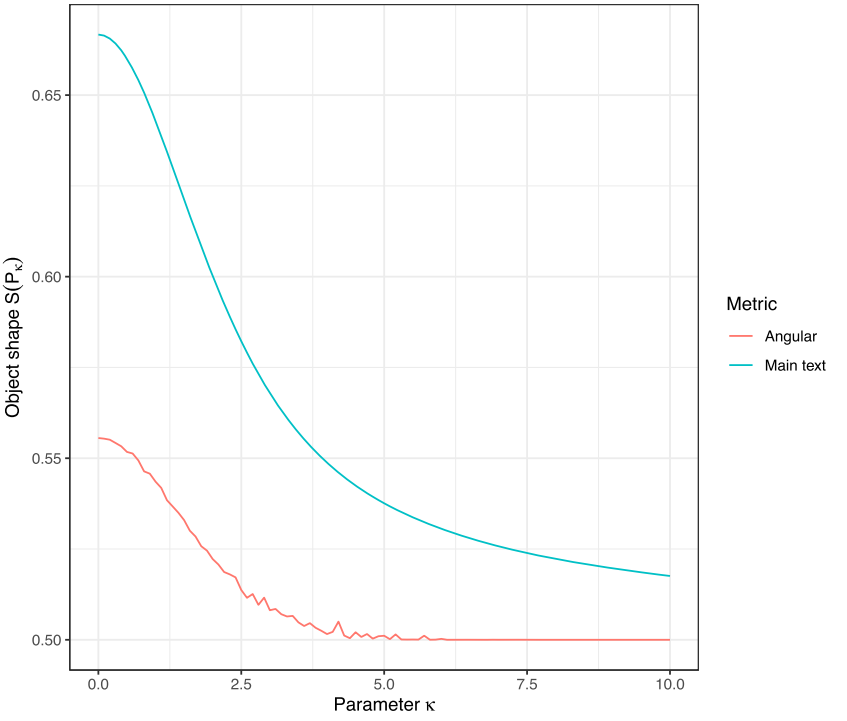

Figure 3. The values of the object shape $S(P_\kappa)$ as a function of the von Mises distribution parameter κ for two choices of metric, angular distance and the metric $d(x,y) = \sqrt{1-\cos(x-y)}$ used in the main text. The values for the former have been estimated using samples of size $n = 5000$.

versus κ is shown in Figure 3, along with the corresponding theoretical curve corresponding to the metric $d(x,y) = \sqrt{1-\cos(x-y)}$ (obtained from the proof of Theorem 3.3). We observe from the plot that both metrics yield equivalent behaviour for $S(P_\kappa)$, being, apart from the finite-sample fluctuation, strictly decreasing in κ and approaching 1/2 in the limit (a fact that follows also from Theorem 2.1). Hence, from a practical viewpoint, it matters little which metric one uses to construct S. Finally, note that any collection of three points from a single fixed semi-circle are d-linearly connected when d is taken to be the angular distance. Hence, if the full sample of n points lies on the same semi-circle, as often happens when the concentration parameter κ is large enough, then the object shape takes the exact value $S(P_\kappa) = 1/2$. In Figure 3, this occurs for several large values of κ.

3.4. Compositional data

Fix $p > 1$ and let $\Delta^p = \{x \in \mathbb{R}^p \mid 0 < x_1, \ldots x_p < 1, \sum_{j=1}^p x_j = 1\}$ denote the p-dimensional unit simplex. Data residing in the unit simplex are studied in compositional data analysis (Pawlowsky-Glahn and Buccianti 2011) and occur commonly, e.g. in biology. We equip Δ^p with the Aitchison metric,

$$d^2(x,y) = \frac{1}{2p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left\{\log\left(\frac{x_j}{x_k}\right) - \log\left(\frac{y_j}{y_k}\right)\right\}^2,$$

which is arguably the most commonly used distance in compositional data analysis. As is typical for compositional data, where the observations sum to unity, the Aitchison distance depends on the observed vector x only through the ratios of its components to one another, ignoring their absolute size.

Fix now a distribution F_Z taking values on the positive real line such that $\log Z$ is symmetrically distributed around zero for $Z \sim F_Z$. For $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\mu_1, \ldots, \mu_p > 0$, and $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\theta_1, \ldots, \theta_p \geq 0$, we let $P_{\mu,\theta}$ denote the distribution of the random composition

$$X = (X_1, \ldots, X_p) = \left(\frac{\mu_1 Z_1^{\theta_1}}{\mu_1 Z_1^{\theta_1} + \cdots + \mu_p Z_p^{\theta_p}}, \ldots, \frac{\mu_p Z_p^{\theta_p}}{\mu_1 Z_1^{\theta_1} + \cdots + \mu_p Z_p^{\theta_p}} \right), \tag{3}$$

where $Z_1, \ldots, Z_p$ are a random sample from F_Z. The distribution $P_{\mu,\theta}$ is similar in spirit to the Dirichlet distribution which is generated in the above manner but with the $\mu_j Z_j^{\theta_j}$ replaced by independent Gamma$(\theta_j, 1)$-variates.

A simple computation reveals that, for $X \sim P_{\mu,\theta}$, the map $h : \Delta^p \to \mathbb{R}$ defined as $h(a) = E\{d^2(X, a)\}$ is minimised uniquely at the vector $a = \mu/(\mu_1 + \cdots + \mu_p) \in \Delta^p$. Hence, μ is a location parameter and essentially the Frechét mean of the distribution $P_{\mu,\theta}$. The parameter θ, on the other hand, controls the dispersion of $P_{\mu,\theta}$. In Figure 4 we have illustrated the effect of θ with ternary diagrams (Pawlowsky-Glahn and Buccianti 2011) when $p = 3$ and $\log Z \sim \mathcal{N}(0, 1)$. When θ is a constant vector (top row), the resulting contours are symmetric w.r.t. the center of the simplex, whereas non-constant θ (bottom row) results in elongation along the axes corresponding to the largest elements of θ. The corresponding values of $S(P_{\mu,\theta})$ given in the caption of Figure 4 confirm the intuitive fact that, of the four distributions, the one in the bottom left panel is the least symmetric. We also note that, as with the Dirichlet distribution, certain values of the parameter θ lead into multimodal distributions (top right panel).

The following result shows that $S(P_{\mu,\theta})$ is invariant to μ and essentially measures how close to a constant vector the dispersion parameter θ is.

Theorem 3.5: *Assume that* $E\{(\log Z)^4\} < \infty$ *and denote*

$$u_Z = \frac{1}{2} + \frac{p(p-2)}{(p-1)(\gamma_Z + 2p + 1) + 4},$$

where $\gamma_Z = E\{(\log Z)^4\}/[E\{(\log Z)^2\}]^2$. *Then,* $(\mu, \theta) \mapsto S(P_{\mu,\theta})$ *takes values in* $[1/2, u_Z]$ *and*

(i) $S(P_{\mu,\theta}) = 1/2$ *if and only if exactly one of* $\theta_1, \ldots, \theta_p$ *is non-zero.*
(ii) $S(P_{\mu,\theta}) = u_Z$ *if and only if* $\theta_1 = \cdots = \theta_p$.

The implications of Theorem 3.5 include; (i) $S(P_{\mu,\theta})$ achieves its maximal value when the components of the composition have equal dispersion parameters θ_k, leading to the symmetry observed in the top panels of Figure 4. (ii) $S(P_{\mu,\theta}) = 1/2$ only when X is essentially one-dimensional. In terms of ternary plots, such a distribution would appear as a one-dimensional curve on the simplex. (iii) Comparison of Theorems 3.1 and 3.5 shows that

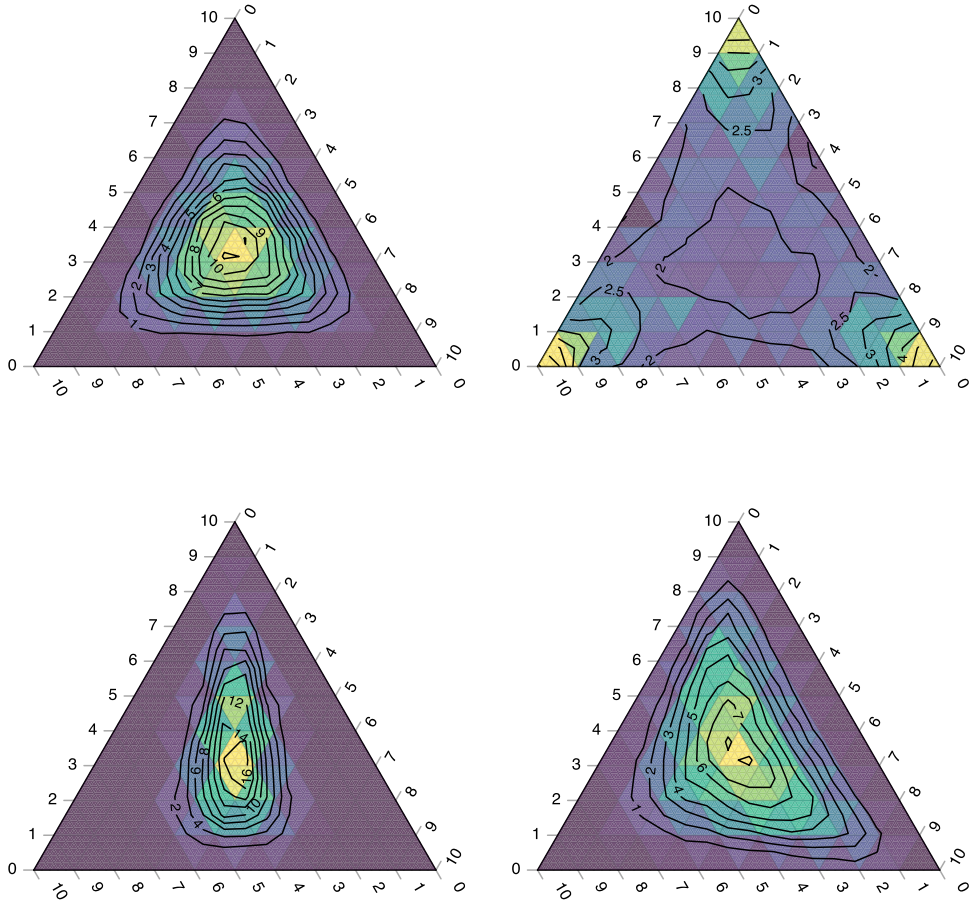

Figure 4. Density estimates of the distribution $P_{\mu,\theta}$ for $\mu = (1, 1, 1)^{\mathrm{T}}$ and four different values of θ, based on samples of size $n = 10000$. The four panels correspond, from left to right, top to bottom, to the values $\theta = (0.5, 0.5, 0.5)$, $(1.25, 1.25, 1.25)$, $(0.75, 0.25, 0.25)$, $(0.75, 0.75, 0.25)$. The measure $S(P_{\mu,\theta})$ takes values in $[0.5, 0.625]$ and in the four scenarios its population value is 0.625, 0.625, 0.547, 0.594, respectively. The plots were drawn using the `R`-package `Ternary` (Smith 2017).

their upper bounds and equality conditions share a certain resemblance. This is not a coincidence as the Aitchison distance between compositions $x, y \in \Delta^p$ is equal to the Euclidean distance between the centered logratio (clr) transformations of x and y.

We next derive a test for the null hypothesis of symmetry, $\theta_1 = \cdots = \theta_p$. For simplicity, we compute the limiting null distribution only in the special case where F_Z is the standard log-normal distribution, $\log(Z) \sim \mathcal{N}(0, 1)$. Equivalent results for other distributions F_Z could be derived using the same proof techniques, by computing the relevant moments up to the eighth order, see the proof of Theorem 3.6.

Theorem 3.6: *Let* $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\mu_1, \ldots, \mu_p > 0$, *and* $\theta_0 > 0$ *be fixed, denote* $1_p = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^p$ *and assume that* $\log(Z) \sim \mathcal{N}(0, 1)$. *Then, as* $n \to \infty$,

$$\sqrt{n}\left\{\frac{2p-1}{2p+2} - S(P_{\mu,\theta_0 1_p,n})\right\} \rightsquigarrow \mathcal{N}\left(0, \frac{(p-2)^2}{2(p+1)^3(p-1)}\right).$$

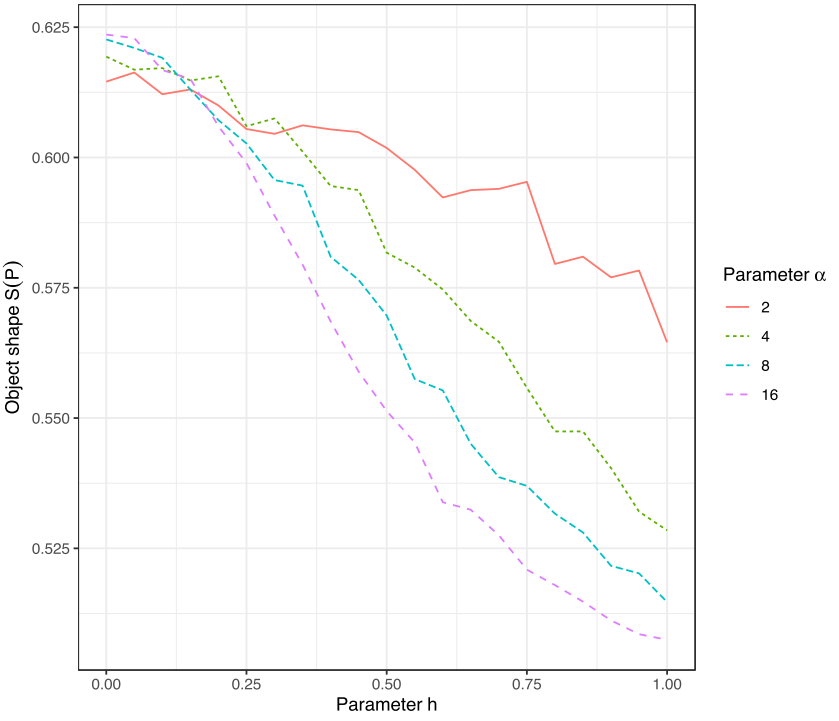

Figure 5. The values of the object shape $S(P)$ as a function of the parameters α, h for the Dirichlet$(\alpha^{1-h}, \alpha, \alpha^{1+h})$-distribution. The values have been estimated using samples of size $n = 5000$.

As with the circular data earlier, also the above test appears to be novel. However, its practical usefulness is limited by its parametric nature.

Since our chosen distributional family (3) does not contain the Dirichlet distribution, perhaps the most used compositional data generating process, we conclude this section with a Monte Carlo study estimating the values of $S(P)$ for Dirichlet$(\alpha^{1-h}, \alpha, \alpha^{1+h})$-distribution with $\alpha = 2, 4, 8, 16$ and $h = 0.00, 0.05, 0.10, \ldots, 1.00$, using the estimator (2) and sample size $n = 5000$. The results are shown in Figure 5 and are perfectly intuitive and in line with our results for the family (3): for each fixed α, the object shape is maximised when $h = 0$, that is, when the three components have the same shape parameter and the distribution is at its most symmetric. Moreover, when h increases, and the distribution becomes more and more concentrated on its third coordinate, we observe the object shape decreasing towards the limiting value $S(P) = 1/2$, corresponding to exact 'unidimensionality'. The decrease is the fastest in the case $\alpha = 16$ which is as expected since $h \mapsto \alpha^{1+h}$ grows faster for larger α.

3.5. Discrete metric in a finite space

Fix $p \geq 3$, and let Θ^p denote the set of all vectors $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in [0,1]^p$ with elements summing to one and having at least two non-zero elements. For $\theta \in \Theta^p$, we denote by P_θ the discrete distribution in $\{1, \ldots, p\}$ taking the value i with the probability θ_i, $i = 1, \ldots, p$. Thus, Θ^p indexes the set of all non-degenerate probability distributions on a p-element set.

We equip the support set $\{1, \ldots, p\}$ with the discrete metric, $d(i, j) = 1 - \mathbb{I}(i = j)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. While this metric is extremely simple, interestingly, it still leads to a meaningful and useful characterisation of shape among the distributions P_θ via the quantity $S(P_\theta)$, as evidenced by the following theorem, where $1_p \in \mathbb{R}^p$ denotes the vector of ones.

Theorem 3.7: *The map $\theta \mapsto S(P_\theta)$ takes values in $[1/2, 1 - 1/p]$ and*

(i) $S(P_\theta) = 1/2$ *if and only if θ has exactly two non-zero elements.*
(ii) $S(P_\theta) = 1 - 1/p$ *if and only if $\theta = (1/p)1_p$.*

By Theorem 3.7, $S(P_\theta)$ measures the uniformity of P_θ. The maximal value $1 - 1/p$ is reached precisely when each of the p objects has exactly the same probability mass. Now, given a sample distribution P_n and the associated observed relative frequencies $(n_1/n, \ldots, n_p/n)$, a classical way of testing the null hypothesis of uniformity is via Pearson's chi squared statistic,

$$T_n = \sum_{i=1}^{p} np \left(\frac{n_i}{n} - \frac{1}{p} \right)^2,$$

which satisfies $T_n \rightsquigarrow \chi^2_{p-1}$ when $n \to \infty$. Our next result shows that the sample object shape is actually asymptotically equivalent to T_n under the null hypothesis that $\theta = (1/p)1_p$. Hence, in this simple case, the object shape recovers the optimal test of symmetry.

Theorem 3.8: *For $\theta_0 = (1/p)1_p$, we have, as $n \to \infty$,*

$$\frac{p(p-1)}{p-2} n \left\{ 1 - \frac{1}{p} - S(P_{\theta_0,n}) \right\} = T_n + o_p(1) \rightsquigarrow \chi^2_{p-1}.$$

Note that the family P_θ contains all distributions on a p-element set, in particular the sample distributions $P_{\theta_0,n}$. As such, the limiting chi-squared distribution in Theorem 3.8 does not lead to a similar insight as with Theorem 3.4 earlier.

4. Peeling plot

Our proposed tool of object data visualisation, the peeling plot, admits two variants, maximisation and minimisation, and we next introduce them one-by-one.

Let $X_1, \ldots, X_n$ be an observed object sample in some metric space $(\mathcal{X}, d)$. We construct the maximisation peeling plot of the sample as follows. Letting $P_{n,-i}$ denote the empirical distribution of the sample with the ith observation removed, we first identify the index $i = 1, \ldots, n$ for which $S(P_{n,-i})$ is maximised. Denoting this index by j_1, we then remove the j_1th observation from the sample and iteratively repeat this process. In the end, we obtain the vector of indices $(j_1, j_2, \ldots, j_n)$, giving the order in which the observations were peeled (removed), and the vector $s \in [1/2, 1]^n$ containing the successive values of the object shape produced by this process. Thus, in particular, $s_1 = S(P_{n,-j_1})$. The peeling plot is then the scatter plot of s versus $1, \ldots, n$, with the accompanying index values $(j_1, j_2, \ldots, j_n)$. For discrete sample space $\mathcal{X}$, it might happen that at some stage of the peeling process we are

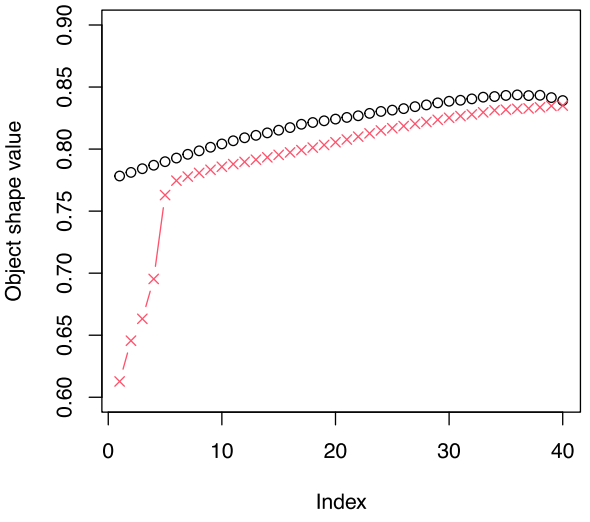

Figure 6. Examples of two superimposed maximisation peeling plots. The red crosses correspond to a sample having 5 outliers, the effect of their peeling being clearly visible as a steep increase in the plot.

left with a sample of identical objects, making the object shape undefined. To avoid such situations, we found it useful to carry out the process only until 80% of the total sample size n has been peeled. R-code for computing the peeling plot is available on the author's web page, https://users.utu.fi/jomivi/software/.

Based on Theorem 2.1, this process attempts to find, in a greedy fashion, subsamples that are essentially as symmetric as possible, in the sense of Theorem 2.1 and the metric d. While the finer interpretation of the plot is, in general, dependent on the particular setting at hand, we found that it is very useful in detecting outliers regardless of the metric space. This is because (a) having isolated outliers makes satisfying the condition in part (ii) of Theorem 2.1 very unlikely, and (b) the effect of outliers on $S(P)$ is rather strong due to the presence of the fourth moments of the distances in formula (1), implying that they get removed first in the peeling. We have demonstrated this in Figure 6, which shows the peeling plots of a particular sample of size $n = 50$ without outliers (black circles) and another sample of size $n = 50$ with five outliers (red crosses). The main difference between the two curves is the sharp increase in the red curve, which represents the sudden rise in $S(P)$ as soon as all five outliers have been peeled from the data.

Figure 6 reveals that it is possible for the object shape to decrease in the peeling process. This corresponds to cases where the current data configuration is approximately symmetric such that removing any observation breaks the symmetry. This mostly happens when either the remaining sample size is small (as in Figure 6) or the data space is highly discrete, both of which make symmetric patterns more likely to occur. Neither of these conditions intervenes with the main objective of the peeling plot, i.e. identifying outliers. This is because (a) the outliers are expected to be peeled in the beginning of the process and any possible decrease in the 'tail' of the peeling plot does not impair recognising them, and (b) if the data are very discrete, then there are more direct ways of detecting outliers, such as tabulating the values and their frequencies.

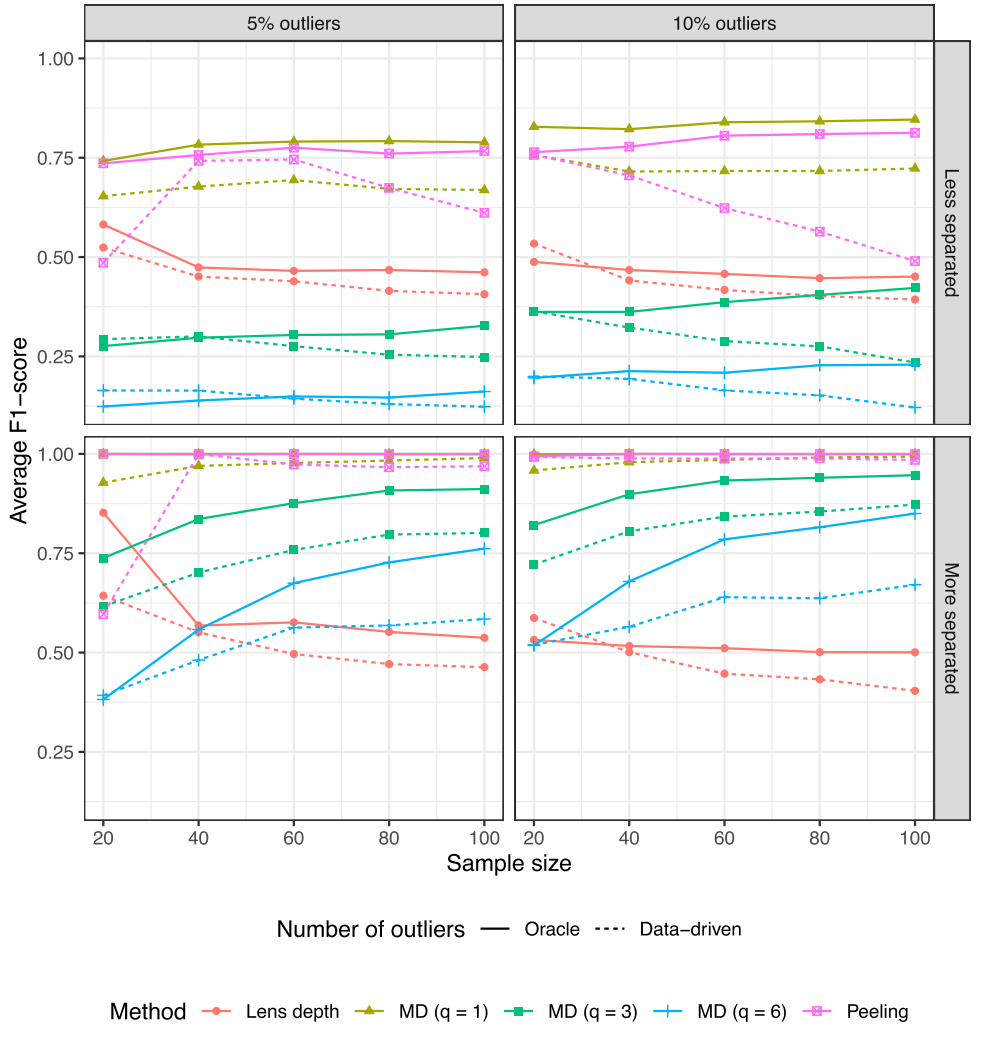

Figure 7. Average F1-scores in the outlier detection simulation, grouped according to sample size, proportion of outliers, separation of the groups, method and how the number of outliers is determined. MD = Mahalanobis distance.

We further investigated the outlier detection capabilities of the peeling plot in a simulation study. We took $(\mathcal{X}, d)$ to be the space of all positive definite 3×3 matrices equipped with the affine invariant metric (Bhatia 2009). We generated the observations as $X_i = \exp(\theta) U_i \mathrm{diag}\{\exp(z_{i1}), \exp(z_{i2}), \exp(z_{i3})\} U_i^{\mathrm{T}}$ where the 3×3 orthogonal matrix U_i is drawn uniformly w.r.t. the Haar measure and $z_{i1}, z_{i2}, z_{i3} \sim \mathcal{N}(0, 1)$, independently. The bulk of the data was generated using the value $\theta = 0$, whereas a small proportion $\varepsilon = 0.05, 0.10$ was taken to be outliers, generated either with $\theta = 2$ (less separated case) or $\theta = 4$ (more separated case). We considered the sample sizes $n = 20, 40, 60, 80, 100$ and in each replicate of the simulation attempted to detect the outlying observations using the following ten methods: (1) The peeling plot where all observations preceding the largest jump in the plot are taken as outliers. (2) The 'oracle' peeling plot where we label the first

$n\varepsilon$ peeled observations as outliers. (3) Another oracle-type estimator which computes metric lens depths (a measure of non-outlyingness) of the sample as described in Cholaquidis et al. (2023) and Geenens et al. (2023) and labels the $n\varepsilon$ objects with the smallest depths as outliers. (4) Similar as the previous method, but instead of knowing the exact amount of outliers, we order the metric lens depths of the sample in increasing order and label all points left of the single largest jump as outliers. (5–6) We embed the data to $q = 6$ dimensions using multidimensional scaling (MDS), compute the 50% breakdown point MCD-based Mahalanobis distances (Hubert and Debruyne 2010) of the embeddings using `OutlierMahdist` in the R-package `rrcovHD` (Todorov 2024), transform the distances as $D \to 1/(1 + D)$, and finally proceed as in (3) and (4) using the transformed Mahalanobis distances in place of the lens depths. (7–8) Similar as the previous method but using an embedding to $q = 3$ dimensions instead. (9–10) Similar as the previous method but using an embedding to $q = 1$ dimension instead. Each of the ten methods produces an index set of observations it labelled as outliers, and as our final evaluation criterion we use the F1-scores (harmonic mean of precision and recall) between these and the true set of outlier indices. Methods 1, 2, 3, 4 can be seen to utilise the natural geometry of the data metric space, whereas methods 5, 6, 7, 8, 9, 10 first obtain an Euclidean approximation of the data and then leverage the rich outlier detection methodology developed for Euclidean data.

The average F1-scores over 500 replicates of the simulation are shown in Figure 7, classified by whether they use the oracle information or select the number of outliers in a data-driven manner through the single largest jump. The main implications of the plots are: (i) The peeling plot is, in almost all settings, among the top performing methods. (ii) The lens depth, with or without oracle knowledge, gives subpar F1-scores. Closer inspection of the results reveals that the lens depth does identify many of the true outliers, but its F1-score is lowered by having large amounts of false positives interspersed with the actual outliers. (iii) Out of the Mahalanobis distances, the lowest-dimensional version with $q = 1$ works best, which is to be expected as the two groups in the model are linearly separable after a non-linear transformation (matrix logarithm). This method also manages to slightly outrank the peeling plot in the less separated case. However, we also observe that the success of the Mahalanobis distances depends greatly on the choice of q, the value $q = 6$ leading to severe underperformance, particularly in the less separated case. As such, using the Mahalanobis distance-based method can be problematic in practise if one has no preliminary information on an appropriate choice of the embedding dimension. Whereas, the peeling plot does not require this kind of tuning while still yielding similar (and in several cases better) performance. (iv) For all methods, the oracle version surpasses the data-driven approach, as expected. (v) The F1-score of the peeling plot without oracle knowledge (purple curve with plus-signs) deteriorates in the less separated case with growing sample size n. This non-intuitive phenomenon is caused by the fact that our criterion for selecting the outlier set (location of single largest jump) is very 'local'. That is, having a larger sample size makes it more likely that the true outlier set is masked by the bulk purely by chance. Better alternatives could likely be devised with, e.g. bootstrapping strategies that control for the randomness in the bulk. Nevertheless, even the current heuristic criterion appears to work extremely well, and object shape outperforms all other methods, except Mahalanobis distance with $q = 1$ in some scenarios. We also note that the same effect is present for the other methods as well, but, being robust, they resist it better.

Note finally that, for maximisation peeling to detect outliers, they must be such that removing them makes the distribution more symmetric. For example, if the bulk and the outliers both obey normal distributions with equal means and proportional covariance matrices, then the outliers have the same 'shape' as the bulk and our proposed prodedure is unable to detect them. As such, the maximisation peeling works best for small amounts of isolated outliers.

We next turn our attention to the minimisation peeling plot which is constructed in exact analogy to its maximisation counterpart, by simply minimising (instead of maximising) the object shape at every peeling. By Theorem 2.1, this process aims to find a subsample where equality is achieved in the triangle inequality as closely as possible for all triplets of points. In analogy to classical PCA, if we stop the peeling when $n - n_0$ observations are peeled, the remaining n_0 objects thus form a set of representatives of the first principal component 'direction' of the sample. As minimisation peeling is a descriptive method, we suggest selecting n_0 experimentally, by visualising the results for various choices and observing for possible structures. We have demonstrated the outcome in the case of Euclidean data in Figure 8, where the $n_0 = 20$ red stars which were left unpeeled in a sample of size $n = 100$ clearly capture the leading principal direction. In the plot we have further connected the crosses in an order which (approximately) minimises the 'surplus', $|d(x_1, x_2) + \cdots + d(x_{n_0-1}, x_{n_0}) - d(x_1, x_{n_0})|$, in the triangle inequality, using the implementation of the traveling salesman problem in the R-package `TSP` (Hahsler and Hornik 2007).

While the interpretation in the Euclidean setting in Figure 8 is clear, in general metric spaces the situation can be more elaborate. Nevertheless, the resulting set of n_0 observations can in any case be interpreted as an ordered sequence of objects corresponding to the single largest 'axis' of variation in the data. To further demonstrate this, we applied minimisation peeling to the hand-written digit data set available in the R-package `tensorBSS` (Virta et al. 2021). Each observation in the data is a greyscale image of size 16×16 and of these we took a random subsample of size $n = 100$ of digits 3 and 8 only. The $n_0 = 20$ representatives of the first principal direction extracted from this sample using minimisation peeling with the Manhattan distance are shown in Figure 9, ordered from left to right, top to bottom. The sequence of objects is dominated by the group structure separating the two digit classes. Within each group we further have a continuum corresponding to a smooth change in digit shape. For example, the threes transition from more thickly drawn digits into thinner ones. Finally, at the very end we have a single outlying digit 8 drawn differently from the rest, making it resemble more a three than an eight and lie on the other side of the threes group.

5. Comparison to descriptive statistics on embeddings

To still better allow interpreting object shape, we next compare it to several well-known Euclidean measures of distributional 'shape' in embedded spaces. We do this by taking a total of 30 object data sets of various sizes, and computing for each a total of 10 descriptive statistics. These statistics are: (1) object shape, (2–4) embedding the data to $q = 3, 6, 9$ dimensions using MDS and computing the eigenvalue-based sphericity measure (Ledoit and Wolf 2002) defined as $[p\mathrm{tr}(S^2)/\{\mathrm{tr}(S)\}^2]^{-1}$ where S is the covariance matrix of the embedding (the inversion makes sure that the quantity measures sphericity and not the

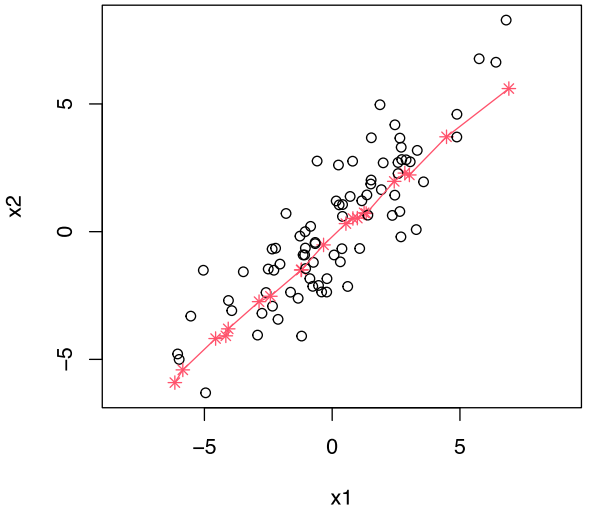

Figure 8. A simple bivariate dataset where the $n_0 = 20$ observations recovered by minimisation peeling (red stars) recover the first principal component direction in the data.

Figure 9. From left to right, top to bottom, the $n_0 = 20$ representatives of the first principal 'direction' in a dataset of size $n = 100$ consisting of greyscale images of digits 3 and 8.

lack of it), (5–7) as with sphericity, but instead we compute Mardia's multivariate kurtosis for each of the embeddings, (8–10) as with sphericity, but instead we compute Mardia's multivariate skewness for each of the embeddings.

We selected five publicly available real data sets from each of the following six types of object data, resulting into a total of 30 data sets: shape data with the Riemannian shape distance (Dryden and Mardia 2016), standard multivariate data with the ℓ_1-metric, functional data with the ℓ_2-metric, compositional data with the Aitchison distance, directional data with the angular distance (shortest path along the sphere) and histogram data with the 2-Wasserstein distance. The exact used data sets and their sources have been listed in Appendix A.

Computing the earlier 10 descriptive measures on these data results into a summary table of size 30×10 whose Spearman rank correlation matrix has been visualised as a heatmap in the left panel of Figure 10. The heatmap shows that the correlations within each of the three types of measures (sphericity, kurtosis, skewness) are quite strong, particularly between the 6 and 9-dimensional embeddings. Moreover, we observe that the kurtoses and skewnesses are positively correlated with each other which might be an indication that many of the data sets contain outlying observations. The object shape is positively correlated with sphericity and negatively correlated with skewness and kurtosis, which is expected based on Theorem 2.1, since one way to achieve high skewness/kurtosis is to have

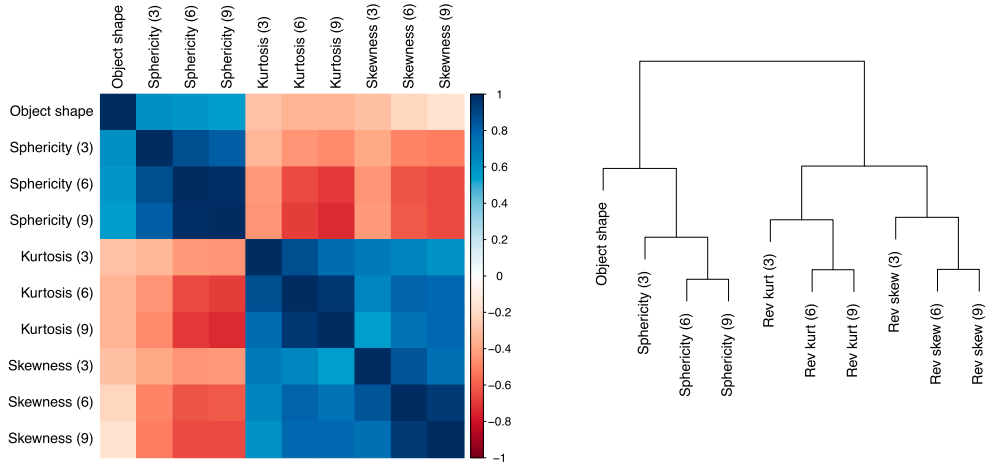

Figure 10. Left: Heatmap of the Spearman correlations between the 10 descriptive measures over the 30 used object data sets. Right: Dendrogram corresponding to the clustering of the 10 measures. 'Rev' (as in 'Rev kurt (3)') means that the sign of the measure has been flipped to make all correlations positive and to enable the clustering.

a small group of outliers sufficiently far from the bulk, which in turn makes the data appear globally concentrated on a single 'dimension'. Object shape is most strongly correlated with sphericity of the 3-dimensional embedding, being in line with our earlier interpretations in Section 3.

To conclude the experiment, we still clustered the 10 descriptive measures by applying hierarchical clustering with complete linkage by using the values $1-JCJ$ as similarity measures, where 1 is a matrix of ones, C is the earlier Spearman correlation matrix and J is a diagonal sign-change matrix that flips the signs of the kurtoses and skewnesses (to make all correlations positive). The resulting dendrogram is shown in the right panel of Figure 10 and further confirms our earlier interpretations: sphericity, skewness and kurtosis each form their own cluster, and of these three the object shape is the closest to the first one.

To summarise the results of the experiment, we observed that the object shape has connections to several standard statistics of shape (of embedded data), showing that it measures partially same aspects as these classical quantities. However, none of the correlations was extremely high, meaning that the object shape still carries also information (in particular, the characterisation in Theorem 2.1) that is not available from the other measures.

6. Discussion

To conclude, based on the previous sections, we can confidently claim that the proposed concept of object shape measures a specific form of symmetry, whose exact form depends on the actual metric space and distribution in question. In several cases, the object shape was seen to measure something that could be termed 'sphericity', but we have avoided using this terminology in the paper (outside of Euclidean cases) due to it evoking strong connections to Euclidean geometry. And while the wide scope of object shape inevitably

makes it more difficult to understand than some more standard descriptive measures, our numerous examples reveal that the object shape is also highly intuitive as a concept. That is, after studying our examples and given a new, suitably well-behaving family of probability distributions, it would likely be easy to formulate an accurate educated guess on what $S(P)$ measures in this case.

A natural continuation of this work would be to investigate how efficient the hypothesis tests obtained using object shape are compared to their parametrically optimal counterparts. It is clear that, as a 'payment' for the generality of our proposed concept, the obtained procedures cannot be expected to be optimal, but how much efficiency exactly is lost should be studied. As remarked in Section 3, in the discrete case we indeed recover the optimal test, but this is likely an artifact of the extreme simplicity of the scenario. Connected to this, another interesting question is whether, by restricting to some particular structured subset of metric spaces (such as Riemannian manifolds), generic tests for the null hypothesis $H_0 : S(P) = \max_Q S(Q)$ could be devised.

A question which we have ignored thus far is whether the extremal value $S(P) = 1$ can actually be reached in practise. This is indeed possible under very extreme scenarios. Take, e.g. a uniform distribution on the unit circle in $\mathbb{R}^2$ equipped with the railroad metric (all paths between points on distinct origin-centered rays go through the origin). Then, $d(X_1, X_2) = d(X_2, X_3) = d(X_3, X_1) = 1$ almost surely, giving $S(P) = 1$ by part (ii) of Theorem 2.1.

It is well-known that a square root $\sqrt{d}$ of a metric d is itself a metric. This viewpoint allows connecting the object shape to metric covariance (Dubey and Müller 2020), a measure of association for a pair of random objects. Namely, if one computes the object shape of $X \sim P$ using $\sqrt{d}$ as a metric, the denominator of (1) takes the form $\mathrm{E}\{d(X_1, X_2)^2\}$. This quantity is proportional to the metric covariance of X with itself (i.e. metric variance) with respect to using d as a metric, showing that the metric variance acts, in a sense, as a standardising quantity in the definition of object shape. However, one has to be careful in interpretations such as this, due to the above discrepancy in the used metrics.

Acknowledgements

The author is grateful to the Associate Editor and the two anonymous Reviewers whose comments greatly helped improve the contents and quality of the manuscript. The author wishes to thank H. Saarinen for a discussion regarding modified Bessel functions of the first kind.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Research Council of Finland under Grants 347501, 353769.

References

Bhatia, R. (2009), *Positive Definite Matrices*, Princeton: Princeton University Press.

Bhattacharya, R., and Patrangenaru, V. (2003), 'Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds', *Annals of Statistics*, 31(1), 1–29.

Bulté, M., and Sørensen, H. (2024), *An autoregressive model for time series of random objects*. Available at *arXiv preprint arXiv:2405.03778*.
Cholaquidis, A., Fraiman, R., Gamboa, F., and Moreno, L. (2023), 'Weighted Lens Depth: Some Applications to Supervised Classification', *Canadian Journal of Statistics*, 51(2), 652–673.
Dai, X., and Lopez-Pintado, S. (2023), 'Tukey's Depth for Object Data', *Journal of the American Statistical Association*, 118(543), 1760–1772.
Dryden, I.L., and Mardia, K.V. (2016), *Statistical Shape Analysis: With Applications in R*, Chichester: John Wiley & Sons.
Dubey, P., Chen, Y., and Müller, H.-G. (2024), 'Metric Statistics: Exploration and Inference for Random Objects with Distance Profiles', *Annals of Statistics*, 52(2), 757–792.
Dubey, P., and Müller, H.-G. (2019), 'Fréchet Analysis of Variance for Random Objects', *Biometrika*, 106(4), 803–821.
Dubey, P., and Müller, H.-G. (2020), 'Functional Models for Time-Varying Random Objects', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2), 275–327.
Dubey, P., and Müller, H.-G. (2022), 'Modeling Time-Varying Random Objects and Dynamic Networks', *Journal of the American Statistical Association*, 117(540), 2252–2267.
Fang, K.-T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions*, London: Chapman and Hall.
Geenens, G., Nieto-Reyes, A., and Francisci, G. (2023), 'Statistical Depth in Abstract Metric Spaces', *Statistics and Computing*, 33(2), 46.
Hahsler, M., and Hornik, K. (2007), 'TSP – Infrastructure for the Traveling Salesperson Problem', *Journal of Statistical Software*, 23(2), 1–21.
Hallin, M., and Paindaveine, D. (2006), 'Semiparametrically Efficient Rank-Based Inference for Shape I. Optimal Rank-Based Tests for Sphericity', *Annals of Statistics*, 34(6), 2707–2756.
Hubert, M., and Debruyne, M. (2010), 'Minimum Covariance Determinant', *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 36–43.
Kruskal, J.B., and Wish, M. (1978), *Multidimensional Scaling*. Number 11. Sage.
Ledoit, O., and Wolf, M. (2002), 'Some Hypothesis Tests for the Covariance Matrix when the Dimension is Large Compared to the Sample Size', *Annals of Statistics*, 30(4), 1081–1102.
Lee, A.J. (1990), *U-statistics: Theory and Practice*, New York: Routledge.
Lin, L., and Chen, Z. (2024), *A type of nonlinear Fréchet regressions*. Available at *arXiv preprint arXiv:2403.17481*.
Lyons, R. (2013), 'Distance Covariance in Metric Spaces', *Annals of Probability*, 41(5), 3284–3305.
Mardia, K.V., and Jupp, P.E. (2000), *Directional Statistics* (Vol. 2), Chichester: Wiley Online Library.
Pawlowsky-Glahn, V., and Buccianti, A. (2011), *Compositional Data Analysis*, Chichester: Wiley Online Library.
Smith, M.R. (2017), *Ternary: An R package for creating ternary plots*. R package version 2.3.3.
Todorov, V. (2024), *rrcovHD: Robust multivariate methods for high dimensional data*. R package version 0.3-1.
Virta, J. (2023), *Spatial depth for data in metric spaces*. Available at *arXiv preprint arXiv:2306.09740*.
Virta, J., Koesner, C.L., Li, B., Nordhausen, K., Oja, H., and Radojicic, U. (2021), *tensorBSS: Blind source separation methods for tensor-valued observations*. R package version 0.3.8.
Virta, J., Lee, K.-Y., and Li, L. (2022), 'Sliced Inverse Regression in Metric Spaces', *Statistica Sinica*, 32, 2315–2337.
Zhou, H., and Müller, H.-G. (2024), *Conformal inference for random objects*. Available at *arXiv preprint arXiv:2405.00294*.
Zhu, C., and Müller, H.-G. (2023), *Geodesic optimal transport regression*. Available at *arXiv preprint arXiv:2312.15376*.