Statistics and Computing (2024) 34:171
https://doi.org/10.1007/s11222-024-10473-x

ORIGINAL PAPER

Wasserstein principal component analysis for circular measures

Mario Beraha[1] · **Matteo Pegoraro**[2]

Received: 19 February 2024 / Accepted: 19 July 2024

Abstract

We consider the 2-Wasserstein space of probability measures supported on the unit-circle, and propose a framework for Principal Component Analysis (PCA) for data living in such a space. We build on a detailed investigation of the optimal transportation problem for measures on the unit-circle which might be of independent interest. In particular, building on previously obtained results, we derive an expression for optimal transport maps in (almost) closed form and propose an alternative definition of the tangent space at an absolutely continuous probability measure, together with fundamental characterizations of the associated exponential and logarithmic maps. PCA is performed by mapping data on the tangent space at the Wasserstein barycentre, which we approximate via an iterative scheme, and for which we establish a sufficient a posteriori condition to assess its convergence. Our methodology is illustrated on several simulated scenarios and a real data analysis of measurements of optical nerve thickness.

Keywords Optimal transport · Directional Data · PCA · Weak Riemannian structure · Distributional data analysis

1 Introduction

Distributional data analysis (DDA) is an emerging subfield of statistical learning dealing with probability distribution as data elements. DDA presents some unique challenges in that probability distributions are not easily embeddable in an Euclidean space, so that techniques developed for multivariate or functional data do not seamlessly translate to the case of DDA. Several recent papers (Bigot et al. 2017; Cazelles et al. 2018; Pegoraro and Beraha 2022; Chen et al. 2021; Zhang et al. 2020; Zhu and Müller 2023) proposed to carry out DDA for measures on the real line by considering data points as elements of the 2-Wasserstein space. Contributions include the definition of principal component analysis (PCA), linear regression and autoregressive models. In those works, the Wasserstein space is considered in close analogy to a "Riemannian" manifold and the characterisation of the tangent space at an absolutely continuous probability measure (Ambrosio et al. 2008) is exploited to perform statistical analysis.

The focus of this paper is principal component analysis for data living in the 2-Wasserstein space of probability measures supported on the unit-circle $\mathbb{S}_1 := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. PCA is popular among practitioners as it produces both a set of orthogonal directions, usually interpreted as the main directions of variability in the dataset, and a map from the space where data live onto the space generated by such directions. Hence, PCA can be used to visually interpret the variability in the dataset and to reduce the dimensionality of the data, by projecting data on their scores. In particular, classical multivariate statistical tools, such as linear regression or clustering, can be carried out by working on the scores. In the context of data living on non-linear spaces, this latter feature is particularly appealing, as it allows using out-of-the-box tools directly on the PCA scores.

The study of distributions on $\mathbb{S}_1$ is relevant in several applied fields including biology, meteorology, and environmental science to cite a few. In particular, directional statistics is an active area of research. See Batschelet (1981), Fisher (1995), Mardia and Jupp (2009), Landler et al. (2018), Pewsey and García-Portugués (2021) for an overview. In

Mario Beraha and Matteo Pegoraro have contributed equally to this work.

✉ Matteo Pegoraro
matteop@math.aau.dk

Mario Beraha
mario.beraha@polimi.it

[1] Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20123 Milan, Italy

[2] Department of Mathematical Sciences, Aalborg University, Skjernvej 4a, 9220 Aalborg, Denmark

Published online: 24 August 2024

this paper, we focus on a specific application in biological research which originates from the analysis in Ali et al. (2021) where measurements of the optical nerve head, obtained via Optical Coherence Tomography (OCT), are studied in connection to the development and progression of optic neuropathies such as glaucoma. The OCT produces a circular scan of the eye measuring neuroretinal rim (NRR) thickness, so that each datapoint can be considered as a function supported on $\mathbb{S}_1$. These are then normalized to eliminate undesired variability introduced by different magnitudes so that data can be considered as probability densities on $\mathbb{S}_1$. A clustering pipeline on the coefficients of the Fourier series expansion of the densities is then developed, thus taking into account the circular nature of the support but overlooking the compositional nature of the data.

1.1 Related works on optimal transport on $\mathbb{S}_1$

Optimal transport for periodic measures was initially investigated by Cordero-Erausquin (1999) where existence of optimal transport maps, as well as necessary conditions for the optimality of transport maps are established. See also Section 2.2 in Manole et al. (2021) for a summary of the main results. Manole et al. (2021) establishes minimax optimality for a class of plug-in estimates of the transport maps when measures are approximated with the empirical counterpart. By contrast, we assume here that measures are fully observed and obtain an analytical expression for the optimal transport maps. Delon et al. (2010) and Hundrieser et al. (2022) study the optimal transport cost for circular measures. Delon et al. (2010) propose efficient numerical algorithms for the computation of the Wasserstein distance with arbitrary cost function, as well as the definition of "locally optimal" transport plans. We build on these results to define the optimal transport maps in our setting. Hundrieser et al. (2022), instead, focuses on the Wasserstein distance where the cost is the geodesic distance on the circle (and not the squared one as considered here). They derive central limit theorems for the approximate transport cost obtained by replacing one measure with its empirical counterpart.

1.2 Related works on PCA in Wasserstein spaces

PCA for probability measures has been framed in different contexts, but, to the best of our knowledge, the focus has been either on analyzing histograms (or discrete measures) or measures supported in $\mathbb{R}$.

Different definitions of PCA (and related algorithms) for distributions under the Wasserstein metric have been proposed in Bigot et al. (2017), Cazelles et al. (2018), Pegoraro and Beraha (2022) and Campbell and Wong (2022). In these works, the space of square-integrable probability measures on the real line, endowed with the 2-Wasserstein metric (also called the Wasserstein space), is considered in close analogy to a "Riemannian" manifold and the characterization of the tangent space at an absolutely continuous probability measure (Ambrosio et al. 2008) is exploited to perform statistical analysis.

When the statistical units are not embedded in a linear space, classical tools from multivariate statistics need to be generalised to take into account the nonlinearity of the space. Think, for instance, about how the Frechét mean generalizes the notion of the sample mean. For data supported on manifolds, the statistical tools can be subdivided into *extrinsic* or *intrinsic* (Bhattacharya et al. 2012; Pennec 2006, 2008; Huckemann et al. 2010; Patrangenaru and Ellingson 2015; Fletcher 2013; Banerjee et al. 2015). The extrinsic approach consists of finding a linear space (usually a tangent space at a suitable centring point) that approximates the manifold (or the region of the manifold where data are located), and performing standard (Euclidean) PCA on the projection of data onto the linear space. In the intrinsic case, instead, the geodesic structure of the manifold is exploited to define a PCA based on the distance between datapoints and (geodesically) convex subsets of the manifold, whereby one considers convex subsets as the natural generalisation of linear subspaces. Note that extrinsic techniques introduce an approximation that might significantly impact the results if the manifold is not well approximated, while intrinsic techniques are usually computationally intensive and not suitable to analyse large datasets.

Focusing on the case of data in the 2-Wasserstein space of measures supported on $\mathbb{R}$, we can label the the *geodesic*-PCA in Bigot et al. (2017) as an intrinsic method, while the *log* PCA in Cazelles et al. (2018) and the *projected* one in Pegoraro and Beraha (2022) are extrinsic tools. These approaches are based on the explicit knowledge of optimal transport maps from an absolutely continuous measure to any other measure, which is a peculiarity of this particular setting. Moreover, Bigot et al. (2017), Cazelles et al. (2018), Pegoraro and Beraha (2022), Campbell and Wong (2022) exploit well-known isometric isomorphisms between the 2-Wasserstein space and closed convex cones in suitably defined L_2 spaces. Thus, the "manifold" nature of the space of probability measure is taken into account by considering the "cone constraints". The *log*-PCA in Cazelles et al. (2018) can be, in principle, applied to distributions over more complex domains. However, as discussed in Pegoraro and Beraha (2022), the *log*-PCA results in poor interpretability of the components and does not allow to work on the scores, which is usually a standard requirement for PCA.

1.3 Our contribution and outline

The aim of our paper is to build a framework for PCA for measures on $\mathbb{S}_1$. In particular, we propose an *extrinsic* PCA,

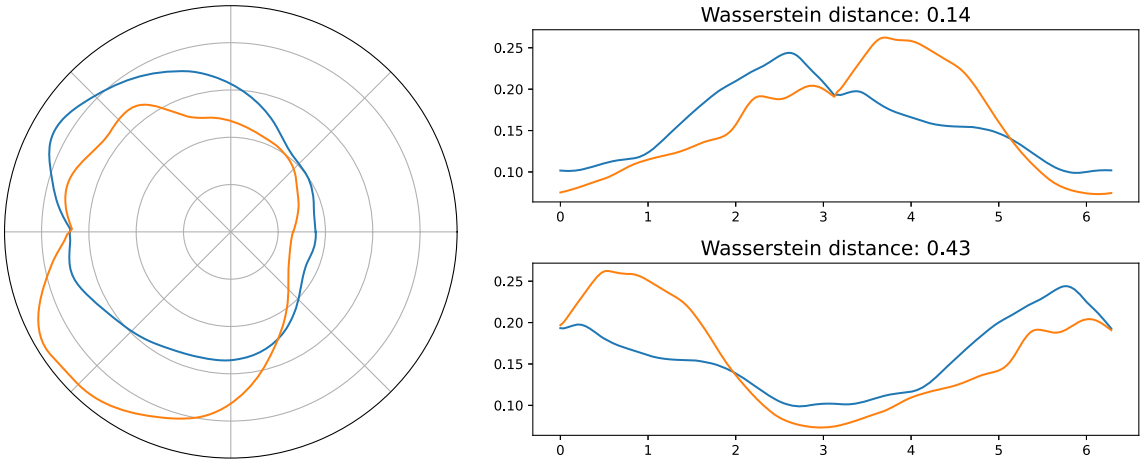

Fig. 1 Two OCT samples on $\mathbb{S}_1$ (left) and when unrolled on $[0, 2\pi]$ (left) starting from 0 (top) or from π (bottom) and the associated Wasserstein distances computed between the probability measures on $[0, 2\pi]$

which consists in choosing a suitable tangent space at a point $\bar{\mu}$, and analyse the transformed data obtained by mapping the observations to the tangent via the logarithmic map. The tangent space is a Hilbert space so that standard PCA could be carried out on the transformed data. However, we prove that the image of the logarithmic map is a convex cone inside the tangent space. We argue that such a constraint should be considered when performing PCA to obtain interpretable results. Indeed, as discussed in Cazelles et al. (2018); Pegoraro and Beraha (2022), failing to do so results in poor interpretability of the directions, and the impossibility to work in the scores. Essentially, both issues are due to the fact that the principal directions might not be orthogonal (or even geodesics) when seen as curves in the Wasserstein space. Following Bigot et al. (2017), we propose a nested PCA, that requires solving a variational problem over the space of probability measures to find the principal directions. Introducing a suitable B-spline approximation, we show how such an optimization problem can be translated into a finite-dimensional constrained optimization problem, whose solution can be approximated numerically using standard software for constrained optimization.

In extending the previously proposed approaches for Wasserstein PCA to measures on $\mathbb{S}_1$ we face several non-trivial issues. For measures on $\mathbb{R}$, the Wasserstein space has null curvature and is isometrically isomorphic to the *space of quantiles*, that is, the convex cone inside $L_2([0, 1])$ of non-decreasing functions. Thus, the manifold nature of the space of probability measure is taken into account by considering the cone constraints. Our setting is more challenging due to the non-Euclidean nature of $\mathbb{S}_1$. A trivial solution to analyzing distributions on $\mathbb{S}_1$ is to fix a point $\theta \in \mathbb{S}_1$ and "unroll" the circle starting from θ, which results in a bijection between $\mathbb{S}_1$ and $[0, 2\pi]$. Hence, it might be tempting to treat distributions on $\mathbb{S}_1$ as distributions on an interval of the real line. However, the Wasserstein metric is then dependent on the chosen θ as shown, for instance, in Fig. 1. More generally, optimal transport for measures supported on Riemannian manifolds is an active area of research (McCann 2001; Gigli 2011; Kim and Pass 2017). In particular, McCann (2001) provides a characterization of optimal transport maps while Gigli (2011) proposes a different definitions of tangent spaces based on the transport maps and plans. Due to the generality of their framework, the resulting expressions are not amenable for computations.

Therefore, prior to defining our PCA, we present a detailed investigation of optimal transport for measures supported on $\mathbb{S}_1$, which is also of independent (mathematical) interest. In particular, building on results in Delon et al. (2010), we derive an expression for optimal transport maps in (almost) closed form and propose an alternative definition of tangent space at any absolutely continuous probability measure. Observe that this is a crucial result for our methodology, as efficiently computing the optimal transport maps between the base point $\bar{\mu}$ and all observations is the first step of the PCA proposed here. Moreover, we specialize the theory developed by Cordero-Erausquin (1999) to the case of $\mathbb{S}_1$ and extend it by characterizing explicitly the image of the "logarithmic map" (i.e., the map from the Wasserstein to the tangent spaces). This characterization is the cornerstone of our methodology and was previously obtained only for measures on the real line (which is a much simpler scenario). Moreover, we establish an homeomorphism between the Wasserstein space and the image of the logarithmic map, ensuring coherence between the tangent space representation and the original measures living in the Wasserstein space space. In particular, the characterization of the image of the logarithmic map translates into linear constraints that define the space where we look for the principal components, making the optimization problem amenable to numerical software. On the other

hand, the continuity of the logarithmic map and its inverse are essential to motivate our approach: without continuity, performing the extrinsic PCA would not lead to any interpretable insight on the original dataset of probability measures.

Finally, we discuss an algorithm to approximate the Wasserstein barycentre and propose to use the output of such an algorithm as the centring point $\bar{\mu}$ for the PCA. Our algorithm follows the one in Zemel and Panaretos (2019), which requires explicit knowledge of optimal transport maps. We derive a sufficient *a posteriori* condition to assess its convergence to the barycentre, and validate it on several simulations, leaving a theoretical analysis for future works.

The paper is structured as follows. In Sect. 2 we cover the necessary background material on optimal transport. Section 3 contains the main results related to optimal transport for measures on $\mathbb{S}_1$ and Sect. 4 discusses our PCA framework and the numerical approximation of the Wasserstein barycentre. Numerical illustrations are presented in Sect. 5, where we discuss a simulation study for the PCA on location-scale families of distributions, highlighting the differences between the case of measures on $\mathbb{R}$ and $\mathbb{S}_1$. In Sect. 6 we present our analysis of the OCT measurements. Finally, we conclude the paper with a discussion on open problems and future work in Sect. 7. Proofs, further background material, and complementary results are deferred to the Supplementary Material. Code implementing the proposed methodologies is available at https://github.com/mberaha/WasserteinS1PCA.

2 Background on optimal transport and on manifold-valued data analysis

In this section, we provide a brief account of optimal transport and the Wasserstein distance for measures on compact manifolds. See, e.g., Ambrosio et al. (2008) for a detailed treatment. Technical details are deferred to Appendix A.

2.1 Riemannian manifolds

Informally, one can think of an n-dimensional smooth manifold M as a set which locally behaves like a Euclidean space: it can be covered with a collection of open sets $(U_i)_{i\geq 1}$ for which there exist homeomorphisms $\varphi : U_i \to \varphi(U_i) \subset \mathbb{R}^n$, called coordinate charts, which satisfy some compatibility conditions. We may refer to $(U_i, \varphi(U_i))$ as a *local parameterisation* of the manifold. A Riemannian manifold (M, g) of dimension n is a smooth manifold M endowed with (a smooth family of) inner products $g = (g_x)_{x\in M}$ on the tangent space $T_x M$ at each point $x \in M$. Its tangent bundle TM is defined as

$$TM := \coprod_{x\in M} T_x M = \bigcup_{x\in M} \{x\} \times T_x M. \tag{1}$$

Each $T_x M$ is a vector space of dimension n. The tangent bundle is itself a smooth manifold of dimension $2n$ with a standard smooth structure. See Lee (2013) for an introduction to Riemannian manifolds.

The *exponential* map at $z \in M$ denoted by $\exp_z : TM \to M$ allows us to map a tangent vector $v \in T_x M$ onto the manifold itself. Informally, $\exp_z(v)$ is the arrival point of the geodesic starting at z with direction v travelled for a unit of time. The *logarithmic* map $\log_z : M \to TM$, where it is defined, satisfies $\exp_z \circ \log_z(x) = x$. The inner product g induces the volume measure ω, which is locally (i.e., on a chart (U, φ)) given by

$$\mathcal{L}_M(A) = \int_{\varphi(A)} |\det(g(\varphi^{-1}(x)))|^{1/2}\, d\mathcal{L}(x) \tag{2}$$

for any measurable $A \subset U$ and with $\mathcal{L}$ being the Lebesgue measure. See Section A for measure theoretical details.

2.2 Wasserstein space

To define the Wasserstein metric, denote by $\mathcal{P}(M)$ the space of probability measures on M and let $c : M \times M \to \mathbb{R}_+$ be a cost function. The p-Wasserstein distance between two probability measures on M, say μ and ν, is

$$W_p(\mu, \nu)^p = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{M\times M} c(x, y)^p d\gamma(x, y), \tag{3}$$

where $\Gamma(\mu, \nu)$ is the set of all probability measures on $M \times M$ with marginals μ and ν. The existence of (at least one) optimal plan γ^o attaining the minimum in (3) is ensured if c is lower semicontinuous (Ambrosio et al. 2008). Definition (3) is due to Kantorovich and can be seen as the weak formulation of Monge's optimal transportation problem, i.e.

$$W_p(\mu, \nu)^p = \inf_{T:T\#\mu=\nu} \int_M c(x, T(x))^p d\mu(x),$$

where # denotes the pushforward operator: $T\#\mu(A) = \mu(T^{-1}(A))$ for all measurable A. It can be proven that when an optimal map exists, then it induces an optimal transport plan $\gamma^o = (\mathrm{Id}_M, T)\#\mu$ and the two formulations are equivalent. However, there are several situations in which Monge's problem has no solution.

In the following, we will always consider the Riemannian distance $\mathrm{d}_R(\cdot, \cdot)$ as cost function and set $p = 2$. We restrict our focus on measures in the 2-Wasserstein space, that is the subset of probability measures

$$\mathcal{W}_2(M) = \Big\{\mu \in \mathcal{P}(M) : \int_M \mathrm{d}_R(x, x_0)^2 d\mu(x) < \infty \quad \text{for every } x_0 \in M\Big\}.$$

This ensures that Wasserstein distance is always finite.

2.3 Geometry of the Wasserstein space

The Wasserstein space $(\mathcal{W}_2, W_2)$ can be endowed with a weak Riemannian structure induced by the tangent spaces of $\mathcal{W}_2$ at any absolutely continuous measure with respect to the volume measure (2). As in the case of measures supported in $\mathbb{R}^n$, the tangent spaces are subset of L^2 spaces of vector-valued functions defined on the ground space (in this case, M). Their definition needs some further background.

Consider a vector field $v : M \to TM$ such that for every $z \in M$, $v_z := v(z) \in T_z M$. To be more precise, denote by π the canonical projection map $\pi : TM \to M$, i.e. $\pi(z, v) = z \in M$, then v must be such that

$$\pi \circ v = \mathrm{Id}_M,$$

where Id_M is the identity map on M. Let $S(M)$ be the collection of all such vector fields. Then, for a measure $\mu \in \mathcal{P}(M)$ we can define L^2_μ as

$$L^2_\mu(M) = \left\{ v \in S(M) : \int g(v_z, v_z)^2 d\mu(z) < \infty \right\}. \tag{4}$$

See Appendix A in the appendix for further details. For $v \in S(M)$ we can define the map $\exp(v) : M \to M$ such that $\exp(v)(z) := \exp_z(v_z)$ for $z \in M$. With this notation, we can state a fundamental theorem in optimal transportation due to McCann (2001).

Theorem 1 *(Characterization of optimal transport plans) Let $\mu, \nu \in \mathcal{W}_2(M)$. If μ is absolutely continuous with respect to the volume measure (2), there exists a unique optimal transport plan, γ^o, from μ to ν, which has the form $\gamma^o = (Id_M, T)\#\mu$, where $T : M \to M$. Moreover, there exists a d^2_R-concave function ϕ such that $T = \exp(-\nabla\phi)$.*

The d^2_R-concavity condition is rather technical and not needed in the following, for this reason we report it only in Section A of appendix, see Gigli (2011) for further details. To make explicit the dependence of the transport map on the source and target measures, we will use notation T^ν_μ to refer to the optimal transport map (OTM) from μ to ν.

The existence and uniqueness of optimal transport maps suggest the following definition of tangent spaces (Corollary 6.4 of Gigli 2011)

$$\mathrm{Tan}_\mu(\mathcal{W}_2(M)) = \overline{\{v \in L^2_\mu(M) \mid \exists \varepsilon > 0 : \overline{(\mathrm{Id}_M, \exp(tv))\#\mu \text{ is optimal for } t \le \varepsilon\}}^{L^2_\mu}. \tag{5}$$

As in the case of Riemannian manifolds, we can define the exponential and logarithmic maps that allow to move from the tangent space $\mathrm{Tan}_\mu(\mathcal{W}_2(M))$ to the Wasserstein space and vice versa.

$$\begin{aligned} &\exp_\mu : L^2_\mu(M) \to \mathcal{W}_2(M),\ \exp_\mu(v) = \exp(v)\#\mu \\ &\log_\mu : \mathcal{W}_2(M) \to L^2_\mu(M),\ \log_\mu(\nu) = v \text{ s.t. } \exp(v) = T^\nu_\mu. \end{aligned} \tag{6}$$

This structure is usually referred to as the *weak Riemannian structure* of $\mathcal{W}_2(M)$.

3 Optimal transport on the circle

In this section, we specialise the general theory outlined in Sect. 2 to the case of measures supported on the unit-radius circle.

3.1 Geometry of $\mathbb{S}_1$

For our purposes, it is convenient to define the unit-radius circle as $\mathbb{S}_1 := \{z \in \mathbb{C} : \mid z \mid = 1\}$, where $\mid \cdot \mid$ denotes the modulus of a complex number. We first present the smooth (group) structure of $\mathbb{S}_1$ and then describe its Riemannian structure.

To endow $\mathbb{S}_1$ with a group structure, we start by considering the map $\exp_c : \mathbb{R} \to \mathbb{S}_1$ defined as $\exp_c(x) = e^{ix}$, and the map $\log_c : \mathbb{S}_1 \to \mathbb{R}$ defined as $\log_c(z) = x \in [0, 2\pi)$ such that $z = e^{ix}$. Note that log_c is right inverse of $\exp_c$, i.e., $\exp_c \circ \log_c = \mathrm{Id}_{\mathbb{S}_1}$. The exponential map $\exp_c$ is usually referred to as *universal covering* of $\mathbb{S}_1$ (Munkres 2000). Clearly, we take the multiplication between complex numbers as the group operation: $\cdot : \mathbb{S}_1 \times \mathbb{S}_1 \to \mathbb{S}_1$ given by $z \cdot w = \exp_c(\log_c(z) + \log_c(w))$. Informally speaking $\log_c(z)$ is the "angle" associated with the polar representation of z and $\cdot$ is the sum of the angles. It can be trivially seen that $(\mathbb{S}_1, \cdot)$ is a group and $\exp_c : (\mathbb{R}, +) \to (\mathbb{S}_1, \cdot)$ is a group homomorphism.

Through $\exp_c$ and $\log_c$ we can define the smooth structure of $\mathbb{S}_1$ by considering at each $z \in \mathbb{S}_1$ the map $\exp_z(x) := \exp_c(x + \log_c(z))$, that is the shifted version of the exponential map, and $\log_z(w) = y$ such that $y \in [-\pi/2, \pi/2)$ and $\exp_z(\log_z(w)) = w$. Letting $V_z := \mathbb{S}_1 \backslash \{-z\}$, we have that for each $z \in \mathbb{S}_1$ the couple $(V_z, \log_z)$ is a coordinate chart. With this differential structure $\mathbb{S}_1$ is a Lie Group and its tangent bundle is $T\mathbb{S}_1 = \{(x, v) \mid x \in \mathbb{S}_1 \text{ and } v \in T_x\mathbb{S}_1\} \simeq \mathbb{S}_1 \times \mathbb{R}$. We call 1 the point $(1, 0)$ which gives the neutral element in $\mathbb{S}_1$.

We consider the Riemannian metric g is induced by the embedding $\mathbb{S}_1 \hookrightarrow \mathbb{C} \simeq \mathbb{R}^2$, that is $g_z(x, y) = xy$ for $x, y \in T_z\mathbb{S}_1 \simeq \mathbb{R}$. This induces the arc-length distance d_R. Note that $det(g) \equiv 1$, so that $\mathcal{L}_{\mathbb{S}_1} = \exp_c \#\mathcal{L}$ or, equivalently,

$\log_c \#\mathcal{L}_{\mathbb{S}_1} = \mathcal{L}$. Thus for any $f : \mathbb{S}_1 \to \mathbb{R}$

$$\int_{\mathbb{S}_1} f(z) d\mathcal{L}_{\mathbb{S}_1}(z) = \int_{-\pi/2}^{\pi/2} f(\exp_c(x)) d\mathcal{L}(x). \tag{7}$$

See Appendix A for further details.

3.2 Optimal transport maps

With the notation introduced in the previous section we now focus on the optimal transportation problem on $M = \mathbb{S}_1$ endowed with its Riemannian distance d_R.

The fundamental observation is that a measure μ on $\mathbb{S}_1$ can be equivalently represented by a *periodic* measure on $\mathbb{R}$ defined as $\tilde{\mu}(A) := \mu(\exp_c(A))$ for any measurable A, which entails $\tilde{\mu}(A) = \tilde{\mu}(A + p)$ for any $p \in 2\pi\mathbb{Z}$, where $A + p$ amounts to shifting all the points in A by the amount p. Then we define the "periodic cumulative distribution function" associated with $\tilde{\mu}$ as $F_{\tilde{\mu}}(x) = \tilde{\mu}([0, x))$ for $x \in [0, 2\pi]$ and extend it over $\mathbb{R}$ via the rule $F_{\tilde{\mu}}(x + 2\pi) = F_{\tilde{\mu}}(x) + 1$. For $\theta \in \mathbb{R}$, let $F^{\theta}_{\tilde{\mu}}(x) = F_{\tilde{\mu}}(x) + \theta$ denote a vertical shift of the cumulative distribution function. Note that the measure induced by $F^{\theta}_{\tilde{\mu}}$ is independent from θ and is always $\tilde{\mu}$. This easily follows from, for instance, $\tilde{\mu}([a, b]) = F^{\theta}_{\tilde{\mu}}(b) - F^{\theta}_{\tilde{\mu}}(a) = F_{\tilde{\mu}}(b) - F_{\tilde{\mu}}(a)$.

Denote with $F^{-}_{\tilde{\mu}}$ the associated quantile function, i.e., the (generalised) inverse of $F_{\tilde{\mu}}$. We have that $(F^{\theta}_{\tilde{\mu}})^{-}(x) = F^{-}_{\tilde{\mu}}(x - \theta)$. Thus, θ acts as a rotation of the quantiles around the circle, by a factor of $z_{\theta}^{-1} = \exp_c(-\theta)$. Hence, the 0-th quantile $(F^{\theta}_{\tilde{\mu}})^{-}(0)$ is not 0 but z_{θ}^{-1}. Equivalently, $F^{\theta}_{\tilde{\mu}}(y) = \tilde{\mu}([z_{\theta}^{-1}, y))$.

Exploiting results contained in Delon et al. (2010), the following proposition provides an explicit characterisation for the optimal transport maps between two measures on $\mathbb{S}_1$.

Theorem 2 *Define θ^* as the solution of the following minimisation problem:*

$$\theta^* = \underset{\theta \in \mathbb{R}}{argmin} \int_0^1 \left(F^{-}_{\tilde{\mu}}(u) - (F^{\theta}_{\tilde{\nu}})^{-}(u) \right)^2 \mathrm{d}u. \tag{8}$$

Then the optimal transport map between μ and ν is

$$T^{\nu}_{\mu} := \exp_c \circ \left((F^{\theta^*}_{\tilde{\nu}})^{-} \circ F_{\tilde{\mu}} \right) \circ \log_c . \tag{9}$$

Note that (9) is closely related to the expression of optimal transport maps for measures on $\mathbb{R}$. In that case, setting $\exp_c = \log_c = \mathrm{Id}$ and $\theta^* = 0$ we recover the classical formulation of OTMs for measures on the real line. Observe that the expression of the OTM in (9) depends on solving the optimization problem in (8). However, it is easy to see that (8) is convex (Delon et al. 2010), and moreover it involves only one real variable. Therefore, its solution is unique and extremely fast to compute. In the following, we will write $\widetilde{T}^{\tilde{\nu}}_{\tilde{\mu}} := (F^{\theta^*}_{\tilde{\nu}})^{-} \circ F_{\tilde{\mu}}$ to denote the map between $\tilde{\mu}$ and $\tilde{\nu}$ associated with the optimal θ^* in (8). Although $\widetilde{T}^{\tilde{\nu}}_{\tilde{\mu}}$ is not "optimal" (since the cost associated to the transport of periodic measures is either zero or unbounded), we will refer to it as the optimal transport map between $\tilde{\mu}$ and $\tilde{\nu}$ in light with its connection with T^{ν}_{μ}.

Let us give some intuition behind the optimal transport map T^{ν}_{μ}. Observe that precomposing $(F^{\theta^*}_{\tilde{\nu}})^{-}$ with $\left(F_{\tilde{\mu}}\right)_{|[0,2\pi]}$, obtaining $\widetilde{T}^{\tilde{\nu}}_{\tilde{\mu}}$, means transporting quantiles identified by $F^{-}_{\tilde{\mu}}$ onto the corresponding shifted quantiles of $(F^{\theta^*}_{\tilde{\nu}})^{-}_{|[0,1]}$, in an anti-clockwise order (due to the definition of $\exp_c$). Note that $T^{\tilde{\nu}}_{\tilde{\mu}}((F_{\tilde{\mu}})^{-}(0)) = T^{\tilde{\nu}}_{\tilde{\mu}}(0) = F^{-}_{\tilde{\nu}}(-\theta^*) =: x_{-\theta^*}$ and

$$\begin{aligned} T^{\tilde{\mu}}_{\tilde{\nu}}((F_{\tilde{\mu}})^{-}(1)) &\leq T^{\tilde{\mu}}_{\tilde{\nu}}(2\pi) = (F^{\theta^*}_{\tilde{\nu}})^{-}(1) \\ &= F^{-}_{\tilde{\nu}}(1 - \theta^*) = 2\pi + F^{-}_{\tilde{\nu}}(-\theta^*) \\ &= 2\pi + x_{-\theta^*}, \end{aligned} \tag{10}$$

which means that the optimal transport maps sends $[0, 2\pi)$ into $[x_{-\theta^*}, 2\pi + x_{-\theta^*})$. As a consequence we can think at this situation as "unrolling" the circle in two different points, namely $z_{\theta^*}^{-1} = \exp_c(-\theta^*)$ for ν and $1 = \exp_c(0)$ for μ, and then matching the measures induced on $\mathbb{R}$. For instance, suppose μ and ν have densities f_{μ} and f_{ν} with respect to the Lebesgue measure on $\mathbb{S}_1$, $\mathcal{L}_{\mathbb{S}_1}$, then $(F^{\theta}_{\tilde{\nu}})^{-}_{|[0,1]}$ is the quantile function associated with the density $f_{\nu}(\exp_c(x))$ supported on $[x_{-\theta}, 2\pi + x_{-\theta}]$. Clearly no action is taken on μ and thus we transport $f_{\mu}(\exp_c(x))$ supported on $[0, 2\pi]$ onto $f_{\nu}(\exp_c(x))$ supported on $[x_{-\theta}, 2\pi + x_{-\theta}]$. The parameter θ^* then selects the optimal point from which to start unrolling the circle for ν.

Optimal transport maps are fundamental for the statistical methods we develop in the later sections: the optimal transport maps T_i from a reference distribution to the i-th datapoint will play the role of "tangent vectors", allowing us to approximate the Wasserstein space, with a space of functions. Thus, it is essential to characterise the optimal transport maps on $\mathbb{S}_1$, understanding their properties, and inspecting them assuming the perspective of the associated maps $\widetilde{T}$ between periodic measures on $\mathbb{R}$.

The following theorem proves a fundamental property of OTMs.

Theorem 3 *Given μ a.c. measure and $\nu \in \mathcal{W}_2(\mathbb{S}_1)$, $\widetilde{T} := (F^{\theta^*}_{\tilde{\nu}})^{-} \circ F_{\tilde{\mu}}$ is an optimal transport map if and only if:*

$$\int_0^{2\pi} \widetilde{T}(u) - u \, \mathrm{d}u = 0. \tag{11}$$

Comments on Theorem 3 will follow throughout the manuscript as it impacts many of the upcoming definitions and results. Here we just point out that Eq. 11 is independent of the measure μ and is a purely analytical/geometric condition on $\widetilde{T}$.

3.3 Weak Riemannian structure

As already mentioned, our aim is to exploit the weak-Riemannian structure of $\mathcal{W}_2(\mathbb{S}_1)$ to obtain a more tractable representation of a data set of probability measure, which enables the use of statistical tools. Thus, we now specialise the definition of $\mathrm{Tan}_\mu(\mathcal{W}_2(M))$ and the associated exponential and logarithmic maps when $M \equiv \mathbb{S}_1$, translating the original vector-field definition in terms of more tractable functions. Furthermore, we establish properties of the logarithmic map that will be fundamental to develop a coherent statistical framework for analysing probability measures in $\mathcal{W}_2(\mathbb{S}_1)$.

For our purposes, it is convenient to define $L^2_\mu(\mathbb{S}_1)$ as

$$\begin{aligned} L^2_\mu(\mathbb{S}_1) :&= \Big\{ v : \mathbb{S}_1 \to \mathbb{R} \text{ such that } \\ &\qquad \int_{\mathbb{S}_1} v^2(x) d\mu(x) < +\infty \Big\} \\ &= \Big\{ v : [0, 2\pi) \to \mathbb{R} \text{ such that } \\ &\qquad \int_0^{2\pi} v^2(x) d\tilde{\mu}(x) < +\infty \Big\}, \end{aligned}$$

where the second equality follows, with a slight abuse of notation, by considering $v \mapsto v \circ \log_c$. Observe that we recover the space in (4) by identifying $v(x)$ as an element of $T_x\mathbb{S}_1$. Then, if μ is an absolutely continuous measure, we have

$$\mathrm{Tan}_\mu(\mathcal{W}_2(\mathbb{S}_1)) = \overline{\{v : L^2_\mu(\mathbb{S}_1) \mid \exists \varepsilon > 0 : \overline{(\mathrm{Id}_{\mathbb{S}_1}, \exp(tv))\#\mu \text{ is optimal for } t \le \varepsilon\}}^{L^2_\mu}} \tag{12}$$

where we can interpret v as a function defined on $\mathbb{S}_1$ or $[0, 2\pi)$ according to our needs. Now we want to rewrite this definition to make it more easily tractable.

First, note that the optimality condition in (12) is equivalent to saying that there exist ν such that $\exp(tv)$ is an optimal transport map between μ and ν. Then, by Theorem 2 and the fact that $\exp_z(v_z) = \exp_c(\log_c(z) + v_z)$, the vector field v in (12) can be written as $tv(\log_c(x)) = \widetilde{T}(x) - x$, where $\widetilde{T}$ is as in Theorem 2, so that the OTM is $\exp_c(x + (\widetilde{T}(x) - x)) \equiv \exp_c(\widetilde{T}(x))$. Hence, we can restate the definition of tangent space in terms of the maps $\widetilde{T}$ as:

$$\mathrm{Tan}_\mu(\mathcal{W}_2(\mathbb{S}_1)) = \overline{\{\widetilde{T} : L^2_{\tilde{\mu}}([0, 2\pi]) \mid \exists \varepsilon > 0 : \overline{\exp_c(\mathrm{Id} + t(\widetilde{T} - \mathrm{Id})) \text{ is OTM for } t \le \varepsilon\}}^{L^2_{\tilde{\mu}}}}. \tag{13}$$

The definition of exponential and logarithmic map comes quite naturally:

$$\begin{aligned} &\exp_\mu : L^2_\mu(\mathbb{S}_1) \to \mathcal{W}_2(\mathbb{S}_1), \\ &\exp_\mu(\widetilde{T}) = \left(\exp_c \circ \widetilde{T} \circ \log_c\right)\#\mu \\ &\log_\mu : \mathcal{W}_2(\mathbb{S}_1) \to L^2_\mu(\mathbb{S}_1), \\ &\log_\mu(\nu) = \widetilde{T} \text{ s.t. } \widetilde{T}(x) = F^-_{\tilde{\nu}}(F_{\tilde{\mu}}(x) - \theta^*), \end{aligned} \tag{14}$$

where θ^* in the definition of the $\log_\mu$ map is as in Theorem 2. Observe that $\exp_c \circ \widetilde{T} \circ \log_c$ is an OTM between μ and ν. Furthermore, from Theorem 3 we note that the vector field $v : [0, 2\pi) \to \mathbb{R}$ induced by an optimal transport map $\widetilde{T}$ (i.e. $v(u) = \widetilde{T}(u) - u$) satisfying (11) has zero mean when integrated along $\mathbb{S}_1$ with respect to $\mathcal{L}_{\mathbb{S}_1}$. In particular, note that this condition does not depend on μ and gives a purely geometric characterisation of optimal transport maps. This is in accordance to other typically used optimality conditions such as cyclical monotonicity of the support of the transport plan and Brenier's characterisation of OTMs for measures on $\mathbb{R}^n$ (Ambrosio et al. 2008).

We now provide some further characterisations of the optimal transport maps in light of the pieces of notation we have just introduced, complementing the results in Cordero-Erausquin (1999), that only states some necessary conditions for transport maps to be optimal (which is insufficient for our purposes as explained below). These novel results will be of pivotal importance to investigate the map $\log_\mu$ and implementation of numerical algorithms.

Theorem 4 *Given μ a.c. measure, $\widetilde{T} : \mathbb{R} \to \mathbb{R}$ induces an optimal transport map between μ and $\nu := \exp_c \circ \widetilde{T} \circ \log_c \#\mu$ if and only if*

- *$\widetilde{T}$ is monotonically nondecreasing with $\widetilde{T}(x + p) = \widetilde{T}(x) + p$ for all $p \in 2\pi\mathbb{Z}$*
- *$\widetilde{T}$ satisfies (11)*
- *$|\widetilde{T}(x) - x| < \pi$ μ-a.e.*

From the previous result, it is immediate to prove the following.

Corollary 1 *Let μ be an a.c. measure on $\mathbb{S}_1$. Then the image of $\log_\mu$ defined in (14) is a convex set.*

Moreover, the following proposition establishes the continuity of both $\exp_\mu$ and $\log_\mu$.

Theorem 5 *Let μ be an a.c. measure on $\mathbb{S}_1$. Then:*

1. *for any* $\nu_1, \nu_2 \in \mathcal{W}(\mathbb{S}_1)$

$$W_2^2(\nu_1, \nu_2) \leq \int_{\mathbb{S}_1} d_R^2(T_\mu^{\nu_1}, T_\mu^{\nu_2}) d\mu \leq \| \log_\mu(\nu_1) - log_\mu(\nu_2) \|_{L^2_{\bar{\mu}}}^2 .$$

In particular, the $\exp_\mu$ *map is continuous;*

2. *if* $W_2(\nu, \nu_n) \to 0$ *in* $\mathcal{W}_2(\mathbb{S}_1)$ *then*

$$\| \log_\mu(\nu_n) - \log_\mu(\nu) \|_{L^2_{\bar{\mu}}} \to 0,$$

that is, the $\log_\mu$ *map is sequentially continuous. As a consequence, since in metric spaces sequential continuity is equivalent to continuity,* $\mathcal{W}_2(\mathbb{S}_1)$ *and* $\log_\mu(\mathcal{W}_2(\mathbb{S}_1))$ *are homeomorphic via* $\log_\mu$ *and* $\exp_\mu$.

Let us comment here on the importance of the results above in light of our goal of performing PCA. As already mentioned, our aim is to define PCA via projections in the tangent space. The directions found by the PCA need to be interpreted as directions inside $\mathcal{W}_2(\mathbb{S}_1)$, i.e., the results are mapped back into the Wasserstein space. Thus, the success of our methodology deeply relies on the image of the log map being a convex set and the coherence between the Wasserstein space and the chosen tangent space, in the same way the effectiveness of the distributional data analysis techniques developed for $\mathcal{W}_2(\mathbb{R})$ rely on the convex isometric embedding of $\mathcal{W}_2(\mathbb{R})$ into the tangent spaces. Theorem 4 guarantees convexity and Theorem 5 ensures that there is a high level of coherence between the measures in $\mathcal{W}_2(\mathbb{S}_1)$ and their representation via $\log_\mu(\mathcal{W}_2(\mathbb{S}_1))$. It is not an isometric representation as in the case $\mathcal{W}_2(\mathbb{R})$ (see Pegoraro and Beraha (2022)), but the continuity of the exponential and logarithmic maps implies a high level of interpretability.

To conclude this section, we present also another proof of Theorem 5, item 2. To be more precise, it is a proof for a weaker result, but which we believe can be generalised to other compact Riemannian manifolds, on the contrary of the proof of Theorem 5, item 2.

Proposition 1 *Let* σ *be an a.c. measure and* $\{\mu_t\}_t$ *be a sequence of a.c. measures such that* $\mu_t \to \mu_0$ *(in the Wasserstein metric) as* $t \to 0$. *Further assume that the support of* σ *and* μ_t *is (geodesically) convex and their density is bounded from above and strictly greater than zero. Then* $\| \widetilde{T}_\sigma^{\mu_t} - \widetilde{T}_\sigma^{\mu_0} \| \to 0$.

4 PCA for measures on $\mathbb{S}_1$

In this section, we demonstrate how the results obtained in Sect. 3 can be leveraged to develop a principal component analysis framework for measures on $\mathbb{S}_1$ in an extrinsic fashion, by considering $\mu_1, \dots, \mu_n \in \mathcal{W}_2(\mathbb{S}_1)$ in analogy to points of a Riemannian manifold, cf. Section 3.3. This parallelism was first exploited to perform inference on the Wasserstein space in Bigot et al. (2017), Cazelles et al. (2018), Pegoraro and Beraha (2022) to develop a PCA for probability measures on the real line, and in Chen et al. (2021) and Zhang et al. (2020) who propose linear regression and autoregressive models for measures on $\mathbb{R}$ respectively.

As already mentioned in the introduction, in the case of measures on the real line, the weak Riemannian structure of the Wasserstein space allows the definition of both intrinsic and extrinsic techniques (Bigot et al. 2017; Cazelles et al. 2018; Chen et al. 2021; Zhang et al. 2020; Pegoraro and Beraha 2022). In particular, since $\mathcal{W}_2(\mathbb{R})$ can be seen as a convex cone inside a suitably defined L_2 space (by identifying each measure with the associated optimal transport map), intrinsic methods simply need to take into account the "cone constraints" (Pegoraro and Beraha 2022). As noted above, such a cone representation does not hold in the case of $\mathcal{W}(\mathbb{S}_1)$. Therefore, developing intrinsic methods would require working with curves of probability measures. We believe that the results established in Sect. 3 could be the first building block of such intrinsic methods. However, supported by the continuity result in item (3.) of Theorem 5, we propose a *log* PCA, which is computed after mapping all datapoints onto a suitable tangent space. In fact, the continuity results suggest that the approximation we make when mapping data to the tangent space is not too coarse, or, at least, should always produce interpretable results. The numerical illustrations presented in Sect. 5 seem to validate this claim.

4.1 Log convex PCA on $\mathcal{W}_2(\mathbb{S}_1)$

As shown in Corollary 6.6 of Gigli (2011), the tangent space at absolutely continuous measures is Hilbert so that we could apply standard PCA techniques to $\log_{\bar{\mu}}(\mu_1), \dots, \log_{\bar{\mu}}(\mu_n)$, for some fixed measure $\bar{\mu}$. We call this approach "naive" log-PCA. However, as argued in Pegoraro and Beraha (2022), disregarding the fact that the image of the $\log_{\bar{\mu}}$ map is not the whole $\mathrm{Tan}_{\bar{\mu}}(\mathcal{W}_2(\mathbb{S}_1))$ tangent space, but only a convex subset, might produce misleading results. In particular, when two elements of the tangent space lie outside the image of $\log_{\bar{\mu}}$, returning to the Wasserstein space and then back to the tangent via $\log_{\bar{\mu}} \circ \exp_{\bar{\mu}}$ can produce undesired behaviours in terms of distances and angles. More in general, a principal direction is interpretable and captures meaningful variability only as long as it lies inside the convex subset. This fact undermines, for instance, the interpretability of scores and principal directions when they lie outside $\log_{\bar{\mu}}(\mathcal{W}_2(\mathbb{S}_1))$: directions may not be orthogonal and variance inside $\mathcal{W}_2(\mathbb{S}_1)$ may not be decomposed appropriately.

To avoid the problems with the "naive" log-PCA, we propose the following definition of log convex PCA, which

amounts to performing a convex PCA (Bigot et al. 2017) in the tangent space, thus taking into account the constraints enforced by the image of the log map. Let us introduce some notation first. Let $X := \log_{\bar{\mu}}(\mathcal{W}_2(\mathbb{S}_1))$, $H := \mathrm{Tan}_{\bar{\mu}}(\mathcal{W}_2(\mathbb{S}_1))$. For a closed convex set $C \subset X$ and a point $x \in X$ let $d(x, C) = \min_{y \in C} \|x - y\|_{L^2_{\bar{\mu}}}$ Let Sp denote the span of a set of vectors and $\mathcal{C}_{x_0}(U) := (x_0 + Sp(U)) \cap X$ for $x_0 \in X$ and $U \subset H$.

As in Pegoraro and Beraha (2022), we also make the following technical assumption: given a collection of probability measures $\bar{\mu}, \mu_0, \mu_1, \ldots, \mu_n \in \mathcal{W}(\mathbb{S}_1)$ we assume that $\log_{\bar{\mu}}(\mu_0)$ lies in the relative interior of the convex hull of $\{\log_{\bar{\mu}}(\mu_i)\}$. The most common choice for μ_0 is to be chosen as the "mean" of $\{\log_{\bar{\mu}}(\mu_i)\}$, which, being inside an Hilbert space, could violate our assumption in some pathological cases. However, in applications we always resort to a finite-dimensional approximation of $L^2_{\bar{\mu}}$, in which the assumption is always satisfied. For more details see Appendix A in Pegoraro and Beraha (2022).

Definition 1 Consider a collection of probability measures $\bar{\mu}, \mu_0, \mu_1, \ldots, \mu_n \in \mathcal{W}(\mathbb{S}_1)$. Let $\widetilde{T}_i = \log_{\bar{\mu}}(\mu_i) = \widetilde{T}^{\mu_i}_{\bar{\mu}}$, $i = 0, \ldots, n$. A $(k, \bar{\mu}, \mu_0)$ log convex principal component for $\mu_1, \ldots, \mu_n$ is the subset $C_k := \mathcal{C}_{\widetilde{T}_0}(\{w_1^*, \ldots, w_k^*\})$ such that

1. For $k = 1$,

$$w_1^* = \underset{w \in H, \|w\| = 1}{\mathrm{argmin}} \sum_{i=1}^{n} d\left(\widetilde{T}_i, \mathcal{C}_{\widetilde{T}_0}(\{w\})\right);$$

2. For $k > 1$,

$$w_k^* = \underset{w \in H, \|w\| = 1, w \perp Sp(\{w_1^*, \ldots, w_{k-1}^*\}}{\mathrm{argmin}} \sum_{i=1}^{n} d\left(\widetilde{T}_i, \mathcal{C}_{\widetilde{T}_0}(\{w\})\right).$$

Figure 2 exemplifies the difference between the naive L_2 and the convex one in a simpler example when $H = \mathbb{R}^2$ and X is a convex subset. When data are close to the border of X, the L_2 metric between data and the principal components capture variability that lies outside of the convex set. See also Pegoraro and Beraha (2022) for some indexes that quantify the loss of information of the L_2 PCA opposed to the convex one.

4.2 Computation of the log convex PCA via B-spline approximation

The definition of convex PCA translates into a constrained optimisation problem to find the directions $\{w_1^*, \ldots, w_k^*\}$. In Cazelles et al. (2018), the authors discretize the transport maps and solve the optimisation problem via a forward-backward algorithm. As discussed in Pegoraro and Beraha (2022), a more efficient approach consists in approximating the transport maps via quadratic B-splines and solving a constrained optimisation problem via an interior-point method. Here, we follow the second approach.

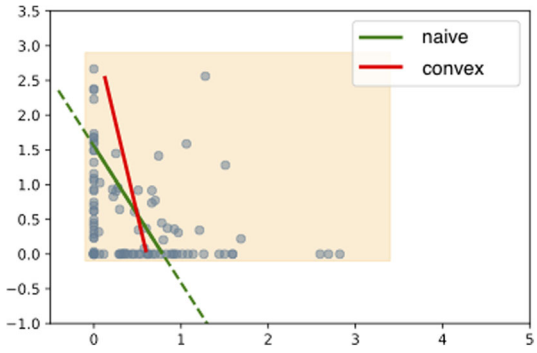

Fig. 2 First principal direction found by the naive L_2 and the convex PCA when the space $H = \mathbb{R}^2$ and X is the yellow rectangle. The blue dots denote observations. Image taken from Pegoraro and Beraha (2022)

Let $\{\psi_1, \ldots, \psi_J\}$ a B-spline basis on equispaced knots in $[0, 2\pi]$. We let $\widetilde{T}_i(x) \approx \sum_{j=1}^{J} a_{ij}\psi_j(x)$. Note that if the spline is quadratic then (i) the function $\sum_{j=1}^{J} a_j\psi_j(x)$ is monotonically nondecreasing if an only if the coefficients $a_1, \ldots, a_J$ are (see, e.g., Proposition 4 in Pegoraro and Beraha 2022,). Hence, from now on, we consider the ψ_j's to be quadratic spline basis functions on $[0, 2\pi]$. The spline basis expansion also allows for faster computations of L_2 inner products: let E be a $J \times J$ matrix with entries $E_{i,j} = \int_0^{2\pi} \psi_i(x)\psi_j(x)\mathrm{d}x$ and $\boldsymbol{a}_i = (a_{i,1}, \ldots, a_{i,J})$, we have $\langle \widetilde{T}_i, \widetilde{T}_j \rangle = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle_E := \boldsymbol{a}_i^T E \boldsymbol{a}_j$. We denote by $\|\cdot\|_E$ the associated norm.

Similarly to Proposition 6 in Pegoraro and Beraha (2022), we obtain that the k-th direction $\boldsymbol{w}_k$ and the associated scores $\lambda^k_{1:n} = \lambda_1, \ldots, \lambda_n$ (of the observations the k-th direction) of the log-convex PCA can be computed by solving a constrained optimisation problem. The objective function is:

$$\lambda^k_{1:n}, \boldsymbol{w}_k = \underset{\lambda_{1:n}, \boldsymbol{w}}{\mathrm{argmin}} \sum_{i=1}^{n} \|\boldsymbol{a}_i - \boldsymbol{a}_0 - \sum_{j=1}^{k} \lambda_i^k \boldsymbol{w}_k\|, \tag{15}$$

where $\lambda_i \in \mathbb{R}$ is the of score for the i-th datum along the k-th direction. Moreover, the usual orthogonality and unit-norm constraints must be satisfied:

$$\|\boldsymbol{w}\|_E = 1, \quad \langle \boldsymbol{w}_h, \boldsymbol{w} \rangle_E = 0, \quad h = 1, \ldots, k-1.$$

In addition to those, we must also require that $\sum w_j\psi_j$ belongs to $H := \mathrm{Tan}_{\bar{\mu}}(\mathcal{W}_2(\mathbb{S}_1))$. The monotonicity constraint is equivalent to

$$\lambda_i w_j + a_{0,j} - \lambda_i w_{j-1} - a_{0,j-1} \geq 0, \quad j = 2 \ldots J$$

that is the monotonicity of the spline coefficients (since the splines are quadratic. See, e.g., Proposition 4 in Pegoraro and Beraha (2022)). Moreover, the "periodicity" constraint is satisfied by design. To impose (11), let $M_j = \int \psi_j(u)\mathrm{d}u$, then (11) is equivalent to

$$\sum w_j M_j = 2\pi^2.$$

Finally, thanks to (11) it is sufficient to control the value of the function w at the initial point, i.e. $w_0 \in (-\pi/2, \pi/2)$.

We implement the resulting constrained optimisation problem using the `Python` package `pyomo` and approximate the solution using an interior point method using the Ipopt solver.

4.3 Wasserstein barycentre

We are left to discuss the choice of the base point μ_0 of the PCA as well as the measure $\bar{\mu}$ at which the tangent space is considered. A standard choice when performing PCA in non-Euclidean spaces, it to set both μ_0 and $\bar{\mu}$ equal the barycentre, that is the Fréchet mean. In our case, the barycentre minimizes the following Fréchet functional:

$$F(\nu; \mu_1, \dots, \mu_n) = \frac{1}{2n} \sum_{i=1}^{n} W_2^2(\nu, \mu_i). \tag{16}$$

While, in principle, the log-PCA can be carried out by working in the tangent space at any absolutely continuous measure, embedding the PCA in the tangent at the barycentre is to be preferred since, intuitively, this should result in the distances between datapoints in the tangent space (at the barycentre) to be more similar to the distances in the Wasserstein space. The quality of the approximation provided by tangent spaces decays as distances from the tangent point increase, and thus choosing as a tangent point the barycentre of the data set is a good choice for trying to minimize the average error produced by the approximations. As a consequence, the projections of the principal components can be interpreted as deviations from the "average" of the data set. Note that centering the PCA at the barycentre poses no conceptual problem in our case as the Wasserstein barycentre is unique if at least one of the measures μ_j is absolutely continuous. See Theorem 3.1 in Kim and Pass (2017). Similar results for measures supported on $\mathbb{R}^d$ have been developed in Agueh and Carlier (2011).

Numerical algorithms for computing the solution of (16) have been developed in Carlier et al. (2015), Srivastava et al. (2015) for the case of atomic measures, whereby the optimization can be reduced to a linear program. Zemel and Panaretos (2019) instead propose a procustes algorithm based on gradient descent, which works for general measures on $\mathbb{R}^d$ (of which one must be absolutely continuous).

Algorithm 1: Procrustes Barycentre

input Measures $\mu_1, \dots, \mu_n$, starting point ν, threshold ε.
repeat
  Compute the optimal transport maps $\widetilde{T}_{\nu}^{\mu_i}$ as in Theorem 2.
  Set

$$\widetilde{\nu}' := \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{T}_{\bar{\mu}}^{\mu_i} \right) \# \widetilde{\nu}$$

until $W_2(\nu, \nu') < \varepsilon$
Output $\bar{\mu} = \exp_c \circ(\widetilde{\nu}')$.
end

In a nutshell, the gradient descent algorithm in Zemel and Panaretos (2019) starts from an initial guess of the barycentre and updates it by pushing forward the current guess ν_r via the average of the transport maps between ν_r and all the measures. This procedure is guaranteed to converge to the barycentre under some technical conditions on the measures μ_i's. In particular, it converges in one iteration if the measures are *compatible* (see Section 2.3.2 in Panaretos and Zemel 2020). As a drawback, this approach requires solving n optimal transportation problems at each iteration, which might be challenging outside the case of measures supported on $\mathbb{R}$ or location-scatter families, for which explicit solutions exist (Alvarez-Esteban et al. 2018). Taking a different approach, Cuturi and Doucet (2014) propose an approximate solution to the Fréchet mean by introducing in (16) an "entropic regularisation" term, which makes optimization easier.

Here, we propose to use the gradient descent algorithm developed in Zemel and Panaretos (2019). Indeed, our Theorem 2 allows for (almost) explicit solutions to the optimal transportation problem. Moreover, as shown in Delon et al. (2010), the optimization problem in (8) is convex in θ so that finding θ^* is simple. We report the pseudocode in Algorithm 1.

We want to remark that we have not been able (yet) to prove either the convergence of the algorithm to the barycentre in the general case or if such procustes algorithm amounts to a gradient descent also in our framework. From the technical point of view, the proofs in Zemel and Panaretos (2019) do not hold in our case, since they are based on sub-differentiability and super-differentiability results of the Wasserstein distance as provided in Theorems 10.2.2 and 10.2.6 in Ambrosio et al. (2008) which are stated for measures on separable Hilbert spaces. Nonetheless, the following result establishes a sufficient condition for the convergence of Algorithm 1.

Proposition 2 *Let μ^* be an absolutely continuous measure in $\mathcal{W}(\mathbb{S}_1)$, and $\mu_1, \dots, \mu_n$ be measures in $\mathcal{W}(\mathbb{S}_1)$. If, for any*

$i, j = 1, \dots, n$

$$\| \log_{\mu^*}(\mu_i) - \log_{\mu^*}(\mu_j) \|_{L^2_{\mu^*}} = W_2(\mu_i, \mu_j),$$

then letting $\bar{T} := n^{-1} \sum_{i=1}^n T_{\mu^*}^{\mu_i}$ *be the barycentre of the* $\log_{\mu^*}(\mu_i)$*'s, we have that* $\bar{T}\#\mu^*$ *is the Wasserstein barycentre of* $\mu_1, \dots, \mu_n$.

The condition in Theorem 2 has the practical advantage that it can be easily checked after Algorithm 1 terminates. Indeed, if $\| \log_{\bar{\mu}}(\mu_i) - \log_{\bar{\mu}}(\mu_j) \|_{L^2_{\mu^*}} = W_2(\mu_i, \mu_j)$, where $\bar{\mu}$ is the output of Algorithm 1, we are sure that $\bar{\mu}$ is the barycentre. Intuitively, if the Wasserstein distances are similar to the distances in the tangent space, this means that, along the geodesics connecting the datapoints, the curvature is small. Hence, the problem of finding the Wasserstein barycentre reduces to averaging the quantiles. Therefore, the output of Algorithm 1 should be accurate. In the following section we provide empirical evidence of its convergence, by checking the condition in Theorem 2 and comparing the output of Algorithm 1 to the one of the Sinkhorn algorithm proposed in Cuturi and Doucet (2014).

Remark 1 Although stated for measures on $\mathbb{S}_1$, Theorem 2 is true for measures on general connected compact finite dimensional Riemannian manifolds whose exponential map is non-expansive. This is the case, for instance, of manifolds with positive curvature. In Appendix B.6 of the appendix we prove the result in this more general setting.

5 Numerical illustrations

In this section we present the numerical simulations dealing with the Wasserstein barycentre and the PCA defined in Sect. 4.

5.1 Simulations for the barycentre

Let us give an illustrative example of the peculiarities that may arise when considering distributions on $\mathbb{S}_1$. Consider the two measures on the leftmost panel in Fig. 3. When the transport cost is the Euclidean one, the resulting barycentre is the one displayed in the rightmost panel: it has unimodal density with the same scale of the two measures and is centred exactly in the middle of them. When the cost instead is computed on $\mathbb{S}_1$, the barycentre becomes bimodal as shown in the middle panel of Fig. 3. In this specific example, the cost (on $\mathbb{S}_1$) of transporting the "correct" barycentre on the two measures is 30% lower than the cost of transporting the "Euclidean" one.

We now give some examples of barycentres. In what follows, we use $\bar{\mu}$ to represent the measure on $\mathbb{S}_1$ returned from Algorithm 1 and $\widetilde{\bar{\mu}}$ the associated periodic measure on $\mathbb{R}$. In some cases, it is intuitive what should be the barycentre and we show that our algorithm correctly converges to it. In other ones, intuition fails but we still might get an idea of the goodness of the approximation of the barycentre by comparing the Wasserstein distances $W_2(\mu_i, \mu_j)$ with the distances in the tangent space as in Theorem 2. Moreover, we also compare the output of Algorithm 1 with the so-called Sinkhorn barycentre (Cuturi and Doucet 2014; Janati et al. 2020) as implemented in the `Python` package `ott-jax` (Cuturi et al. 2022). To compute the Sinkhorn barycentre, we approximate each measure with an atomic measure with 1,000 equispaced support points on $[0, 2\pi)$, equipped with the geodesic distance on $\mathbb{S}_1$, giving to each point x_i a weight proportional to $\mu(\mathrm{d}x_i)$. Informally, we should expect the Wasserstein and Sinkhorn barycentres to be similar, but the Sinkhorn barycentre should be smoother due to the regularisation term involved in the Sinkhorn divergence.

We consider three simulated datasets as follows. Let $\mathcal{U}(c, w)$ denote the uniform measure centred in c and with width w, i.e. the uniform measure over $(c - w/2, c + w/2)$. In the first example, the measures are

$$\begin{aligned} \tilde{\mu}_i &= \mathcal{U}\left(0.25, 0.1 + 0.05i\right), \quad i = 1, \dots, 5, \\ \tilde{\mu}_i &= \mathcal{U}\left(0.75, 0.1 + 0.05(i-5)\right), \quad i = 5, \dots, 10, \end{aligned}$$

and extended periodically over the whole $\mathbb{R}$. In the second one instead

$$\begin{aligned} \tilde{\mu}_i &= \mathcal{U}\left(0, 0.05 + 0.015i\right), \quad i = 1, \dots, 10, \\ \tilde{\mu}_i &= \mathcal{U}\left(1/3, 0.05 + 0.015(i-10)\right), \quad i = 11, \dots, 20, \\ \tilde{\mu}_i &= \mathcal{U}\left(2/3, 0.05 + 0.015(i-20)\right), \quad i = 21, \dots, 30. \end{aligned}$$

In the third case instead, we generate the $\tilde{\mu}_i$'s by first considering Beta distributions on $(0, 2\pi)$ with parameters $(a_i, 2)$ and then taking their periodic extension. Specifically, $a_i \sim \mathcal{U}(1.3, 0.2)$ for $i = 1, \dots, 10$ and $a_i \sim \mathcal{U}(2.6, 0.4)$ for $i = 11, \dots, 20$. Figure 4 reports the Wasserstein barycentres as found by Algorithm 1 and the Sinkhorn ones for three different simulated datasets. We can see that the Wasserstein and Sinkhorn barycentres agree and that the Sinkhorn ones are generally smoother. Moreover, in the first and third example the log and Wasserstein distances are indistinguishable which suggests the convergence of Algorithm 1, while in the second example there are some discrepancies. The third simulation allows us to gather some insights into the geometry of $\mathcal{W}_2(\mathbb{S}_1)$. Indeed, note how, despite all the measures $\tilde{\mu}_i$ being unimodal, the barycentre is bimodal. This clearly arises from the manifold structure of $\mathbb{S}_1$ and specifically because of mass going through 0 along the geodesics connected some measures.

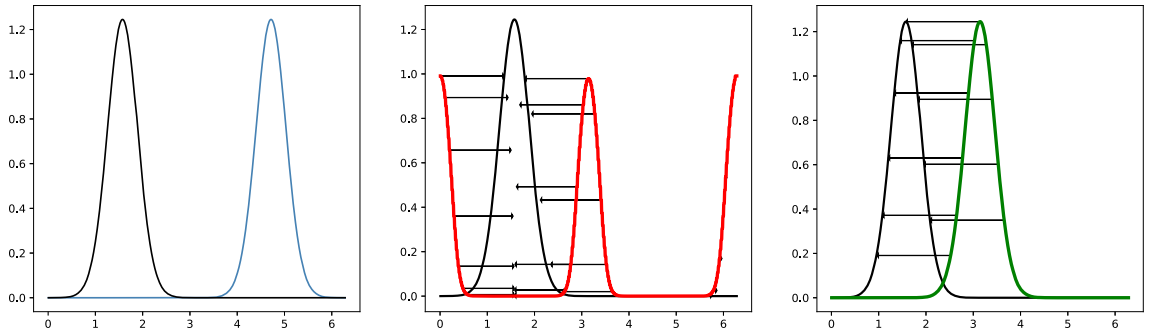

Fig. 3 From left to right: two measures on $\mathbb{S}_1$ (unrolled on $[0, 2\pi]$), the barycentre on $\mathbb{S}_1$ (red) and its transport to the leftmost measure, the barycentre on $\mathbb{R}$ and its transport to the leftmost measure

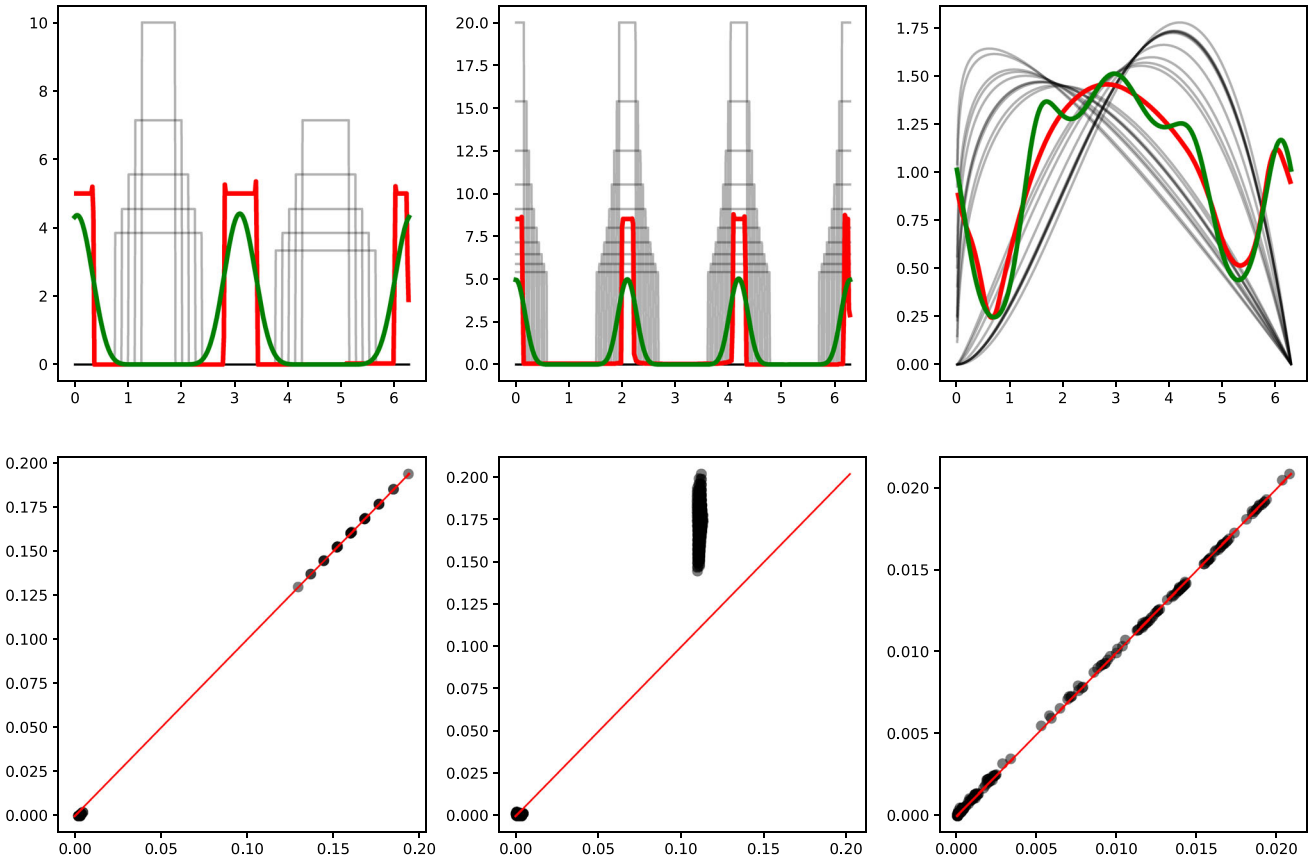

Fig. 4 Top row: denisties of the $\tilde{\mu}_j$'s on $[0, 2\pi]$, and of the Wasserstein and Sinkhorn barycentres (red and green line respectively). Bottom row: Wasserstein distance vs $d_{\log}$ for every possible couple of measures

5.2 Simulations for the PCA

In this section we analyse some simulated datasets which we use to showcase and interpret some behaviours of the PCA defined in previous sections. Another simulation with additional details and comparisons can be found in Appendix C of the appendix. To interpret the principal directions found by the PCA, we produce the plots of the densities of $\exp_{\bar{\mu}}(\log_{\bar{\mu}}(\mu_0) + \lambda w_k^*)$, where w_k^* is the k-th principal direction and λ varies in some range specified case-by-case. Unless otherwise stated, $\bar{\mu}$ and μ_0 are both equal to the Wasserstein barycentre approximated using Algorithm 1. In particular, note that the score λ represents the distance from the base point travelled along the geodesic whose direction is specified by the k-th principal direction. It is then possible to compare different values of λ across the simulations to interpret the distance from the barycentre after which some behaviours start to occur (for instance, it might happen that at a certain distance from the barycentre, the measures switch from unimodal to bimodal).

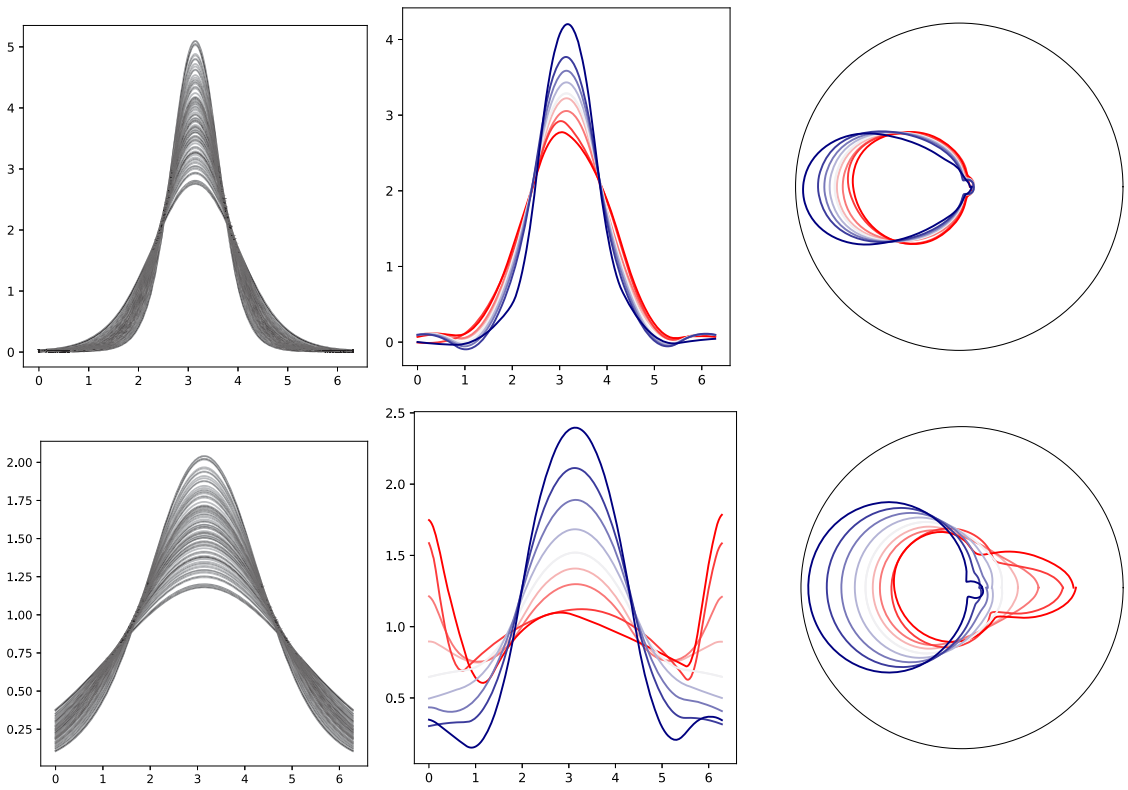

Fig. 5 Data and first principal direction for the Von Mises simulation. The second and third column represent densities along the first principal direction as λ varies between -0.1 (darkest blue) to 0.1 (darkest red), plotted as distributions on $[0, 2\pi)$ and on $\mathbb{S}_1$ respectively

First, we consider a sample from the von Mises distribution with location π and scale α, whose density function on $[0, 2\pi]$ is

$$f(x;\alpha) = \frac{\exp\left(3\cos\left(\frac{x-\pi}{\alpha}\right)\right)}{2\pi I_0(3)}, \tag{17}$$

where I_0 is the modified Bessel function of order zero. We simulate two datasets of $n = 100$ measures from (17), by considering $\alpha \sim \mathcal{U}(0.8, 1.5)$ and $\alpha \sim \mathcal{U}(2, 3.5)$ respectively. Data and the first principal direction are shown in Fig. 5. In the first case, the measures are sufficiently concentrated so that, in the neighbourhood of the barycentre associated to the grid of values for λ, the periodicity of $\mathbb{S}_1$ is effectively irrelevant, and the first principal direction reflects the change in scale of the distribution. On the other hand, in the second case, we have a good amount of mass around 0 for all distributions in the data set, and the variance of such distributions ranges over a bigger interval compared to the first data set. As a consequence, moving along the first principal direction (with the same scale as in the previous example), we keep pushing the mass on "the sides" at faster rates, so that it concentrates even more around 0 and we go from a unimodal to a bimodal density.

Although not shown here, when the same measures are considered as points in $\mathcal{W}_2(\mathbb{R})$, in both cases the first principal direction is associated with a change in the scale of the measures, while the location is kept fixed.

Next, we consider the same dataset as in the third simulation of Sect. 5.1. Figure 6 reports the first two principal directions. The first one corresponds mostly to a shift on the location but simultaneously it also captures the decrease of the density around the second mode that is located in 0 (see the barycentre in Fig. 4). Starting from the barycentre (white), if we go towards the red densities we see that the mode in zero gradually is absorbed the main mode; while if we go towards the blue ones the mode in 0 crosses the circle and it merge on the main mode, but on the right side of the plot. According to the geodesic structure of $\mathcal{W}(\mathbb{S}_1)$. The second direction, instead, is more clearly focused on separating distribution with significant amount of mass close to 0 (blue), from the measures which, instead, have all their mass away from 0 (red).

In summary, these simulations help us understand the geometry of $\mathcal{W}_2(\mathbb{S}_1)$ and, in particular, the differences with $\mathcal{W}_2(\mathbb{R})$. Indeed, it is well-known that, for measures on $\mathbb{R}$, the Wasserstein geodesics of location-scale families are obtained by lifting the Euclidean geodesics in the location-scale plane to the Wasserstein space. Hence, the Wasserstein PCA will

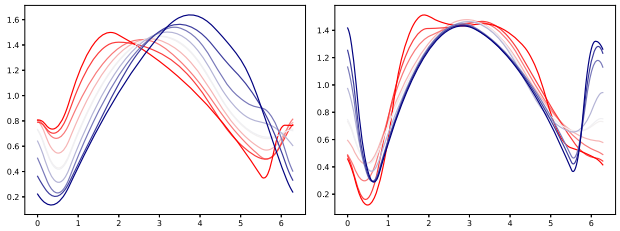

Fig. 6 Densities along the first two principal directions for the Beta distribution as in Sect. 5.1, as λ varies between -0.05 (darkest blue) to 0.05 (darkest red), plotted as distributions on $[0, 2\pi)$

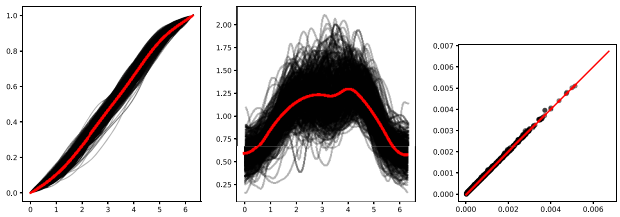

Fig. 7 From left to right: (a subsample of) cdfs of the eye's dataset measures (red line denotes the barycentre), pdfs of the eye's dataset measures (red line denotes the barycentre), Wasserstein distance against $d_{\log}$ in the tangent space at the barycentre

disentangle the effect of the location and the effect of the scale. Instead, as shown by our simulations, when measures are supported on $\mathbb{S}_1$ it is not possible to completely separate the effects of location and scale. Moreover, even if the datapoints are unimodal, it is often the case that the barycentre is multimodal. Multimodality is inherited by the measures along the principal directions, which might make the interpretation cumbersome. In Appendix C of the appendix we report an additional simulation for the PCA, where we discuss the choice of the point $\bar{\mu}$ (at which the tangent is attached) and its impact on the interpretability of the directions. In particular, we consider a dataset of truncated Gaussians, for which the barycentre has three modes. Instead, if $\bar{\mu}$ is chosen to be equal to one of the datapoints, then moving along the principal directions results in unimodal densities for which interpretation is easy. Of course, this poses a conceptual issue as the principal directions are not the "main directions of variability" per se, but the main directions of variability starting from one particular $\bar{\mu}$.

6 Case study: eye dataset

We present here the results of applying PCA to the dataset of OCT (Optical Coherence Tomography) measurements of NRR (neuroretinal rim) thicnkess in Ali et al. (2021), available in their supplementary materials, which contains the OCT measurements of 3973 patients, stratified according to their age groups. In particular, we assess the adequacy of Wasserstein PCA by interpreting the principal direction and performing clustering on the scores, showing how these clusters meaningfully capture shape patterns in data. Data are displayed in Fig. 7 together with the Wasserstein barycentre found via Algorithm 1. In the rightmost plot, we show how the Wasserstein and L_2 distances in the tangent space at the barycentre agree for almost all the couples of datapoints, thereby validating the use of the red measure in Fig. 7 as centering point for our PCA.

The first two principal directions—which, by construction, are the two directions capturing most variability—are reported in Fig. 8. We can clearly see that these decouple the shape variability along the horizontal and vertical axes. In particular, this implies that most of the variability in the data set is made by variations (in the distribution of the) of thickness of the optical nerve, along the horizontal axis. To assess the adequacy of Wasserstein PCA for this dataset, we compute the *average normalised reconstruction error* as a function of the number of directions k used for the PCA.

$$\mathrm{ANRE}_k := \frac{1}{n}\sum_{i=1}^{n} \frac{W_2^2(\mu_i, \mu_i^k)}{W_2^2(\bar{\mu}, \mu_i)},$$

where μ_i^k is the projection on the first k principal components of the measure μ_i. The ANRE index measures the approximation error, normalising by the deviation of the datapoints from the centre of the PCA, in close analogy with the decomposition of variance in the case of PCA in Euclidean spaces. Figure 9 (left plot) reports the ANRE index as a function of k, as well as the (normalised) eigenvalues of the L_2 PCA in the tangent space. Both measures show how the first $k = 5$ directions are enough to capture the variability of the dataset. Moreover, the L_2 variance decreases faster than ANRE. This is expected since L_2 PCA ignores that data are constrained on the image of $\log_{\bar{\mu}}$, and "captures variability" also outside this set. Lastly, we believe that the ANRE in stabilises to a positive (small) number due to numerical errors. In Appendix D of the appendix, we report the scatter plot of the scores along the first two directions, stratified by age groups. From the plot, it is clear that, on the first two components, there is no evident effect of age alone on the shape of the optical nerve.

We cluster the datapoints via a hierarchical clustering algorithm with ward linkage working on the scores along the first $k = 5$ principal directions. In Appendix D of the appendix we show the dendrogram, while the two main clusters found are shown in Fig. 9. Figure 3 in the appendix reports a refined clustering obtained by cutting the dendrogram to get 7 clusters. We have reported in red the barycentres of the clusters, which may be of some help in interpreting the clusters, even though our clustering pipeline is not barycentre-driven like a K-means algorithm. When looking at the two clusters in Fig. 9b, it is clear that they identify two

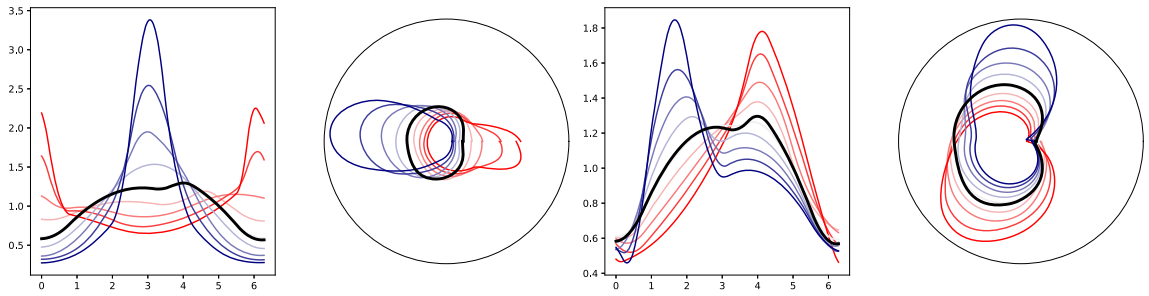

Fig. 8 First (left plots) and second (right plots) principal directions: we report the pdfs on [0, 1] (first and third panels) and in a polar plot (second and fourth panels). The black line denotes the barycentre

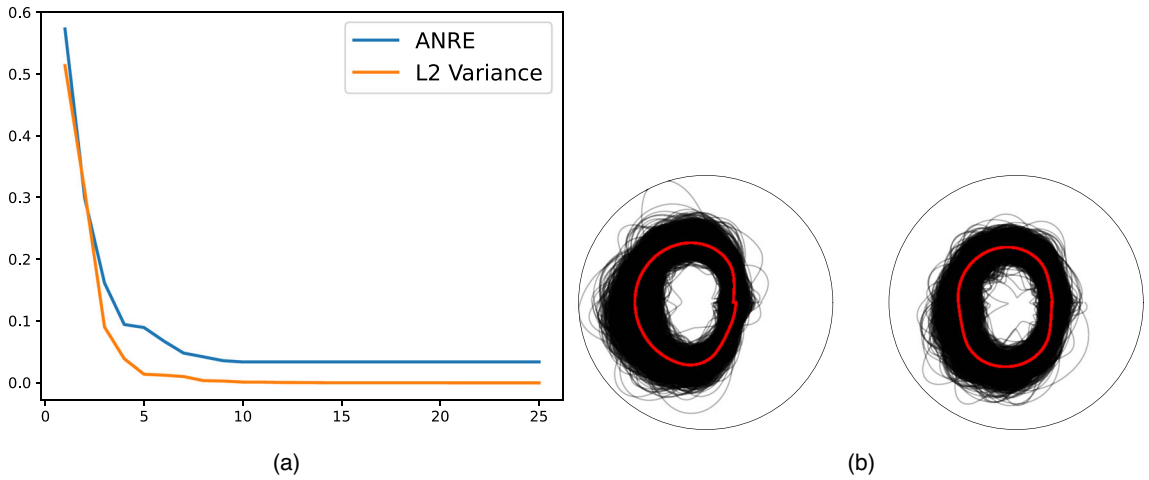

Fig. 9 ANRE index as a function of the dimension and fraction of (L_2) variance explained by each component (left plot) and data subdivided in two clusters (right), with the corresponding Wasserstein barycentre (red lines)

different shapes of the optical nerve with the left one being characterised by a clear bump in the left side. The refined clusters in Fig. 3 in the appendix show interesting patterns as well, see the appendix for further details.

We close this section by highlighting that, as mentioned in the introduction, a very important byproduct of PCA is that classical tools from multivariate statistics can be applied to our dataset after projecting data on the principal components. We leave it to future works to complement our unsupervised analysis with an investigation involving the covariates contained in the original dataset.

7 Discussion

In this paper, we tackle the problem of analysing distributional data supported on the circle. Following recent trends in statistics and machine learning, we set out to use the Wasserstein distance to compare probability distributions. To this end, we study the optimal transportation problem on $\mathbb{S}_1$ and establish several new theoretical results, which could also be of independent interest. In particular, we provide an explicit characterisation of the optimal transport maps. This result is rather surprising given that optimal transport on Riemannian manifolds is not well established and that the only case where such explicit formulas exist is for measures on the real line. We further explored the weak Riemannian structure of the Wasserstein space and establish strong continuity results for the exponential and logarithmic maps, as well as an explicit characterisation of the image of the logarithmic map.

Building on our theoretical findings, we propose a counterpart of the convex PCA in Bigot et al. (2017) for measures on $\mathbb{S}_1$. Following the approach in Pegoraro and Beraha (2022), we propose a numerical method to compute the principal directions by means of a B-spline expansion, which leads to an easily implementable numerical algorithm.

Our definition of PCA requires a "central point", which is usually set equal to the barycentre. We use the algorithm in Zemel and Panaretos (2019) to approximate the Wasserstein barycentre. However, we have not been able to prove the convergence of this algorithm in our setting. Despite numerical simulations do seem to validate the use of Algorithm 1, its theoretical analysis is still an open problem.

Our investigation paves the way to several interesting extensions. First, it is natural to consider the problem of Wasserstein regression. Thanks to the expression for the optimal transport maps, the geodesic regression in Fletcher (2013) can be defined in an analogous way for measure-valued dependent random variables. Similarly, our definition of tangent space is amenable to the definition of a log regression for measures on $\mathbb{S}_1$. For measures on $\mathbb{R}$, Pegoraro and Beraha (2022) proposed to map both dependent and independent variables onto the same tangent space, given that the Wasserstein space is isomorphic to any tangent. Here, it would be more suitable to consider two tangent planes: one for the independent and one for the dependent variables, centred at the respective barycentres, similarly to Chen et al. (2021).

More broadly, we believe that the interplay between optimal transport and distributional data analysis can nourish further developments of both fields. Specialising the treatment of the optimal transportation theory to specific cases of statistical interest, such as the sphere, could lead on one hand to a better understanding of how the properties of tangent spaces relate to the base manifold, and on the other hand to data analysis frameworks which can extract insights for instance from earth-related distributions and other relevant data which are nowadays collected.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10473-x.

Acknowledgements M.B. acknowledges the support by MUR, grant Dipartimento di Eccellenza 2023–2027, and received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257, M.P. received funding from the Independent Research Fund Denmark (1026-00037).

Author Contributions Both authors contributed equally to the manuscript.

Funding Open access funding provided by Aalborg University

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM J. Math. Anal. **43**(2), 904–924 (2011)

Ali, M., Wainwright, B., Petersen, A., Jonnadula, G.B., Desai, M., Rao, H.L., Srinivas, M., Jammalamadaka, S.R., Senthil, S., Pyne, S.: Circular functional analysis of oct data for precise identification of structural phenotypes in the eye. Sci. Rep. **11**(1), 23336 (2021)

Alvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matran, C.: Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. Bernoulli **24**(4A), 3147–3179 (2018)

Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows: In Metric Spaces and in the Space of Probability Measures. Springer Science & Business Media, Berlin (2008)

Banerjee, M., Chakraborty, R., Ofori, E., Vaillancourt, D., Vemuri, B.C.: Nonlinear regression on Riemannian manifolds and its applications to neuro-image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2015)

Batschelet, E.: Circular Statistics in Biology. Academic Press, New York (1981)

Bhattacharya, R.N., Ellingson, L., Liu, X., Patrangenaru, V., Crane, M.: Extrinsic analysis on manifolds is computationally faster than intrinsic analysis with applications to quality control by machine vision. Appl. Stoch. Models Bus. Ind. **28**(3), 222–235 (2012)

Bigot, J., Gouet, R., Klein, T., López, A.: Geodesic PCA in the Wasserstein space by convex PCA. Ann. Inst. Henri Poincare (B) **53**, 1–26 (2017)

Campbell, S., Wong, T.-K.L.: Efficient convex PCA with applications to Wasserstein geodesic PCA and ranked data. arXiv preprint arXiv:2211.02990 (2022)

Carlier, G., Oberman, A., Oudet, E.: Numerical methods for matching for teams and Wasserstein barycenters. ESAIM Math. Model. Numer. Anal. **49**(6), 1621–1642 (2015)

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., Papadakis, N.: Geodesic PCA versus log-PCA of histograms in the Wasserstein space. SIAM J. Sci. Comput. **40**(2), B429–B456 (2018)

Chen, Y., Lin, Z., Müller, H.-G.: Wasserstein regression. J. Am. Stat. Assoc. **118**, 1–40 (2021)

Cordero-Erausquin, D.: Sur le transport de mesures périodiques. C. R. Acad. Sci. Ser. I Math. **329**(3), 199–202 (1999)

Cuturi, M., Doucet, A.: Fast computation of Wasserstein barycenters. In: International Conference on Machine Learning, pp. 685–693. PMLR (2014)

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., Teboul, O.: Optimal transport tools (OTT): a JAX toolbox for all things Wasserstein. arXiv preprint arXiv:2201.12324 (2022)

Delon, J., Salomon, J., Sobolevski, A.: Fast transport optimization for Monge costs on the circle. SIAM J. Appl. Math. **70**(7), 2239–2258 (2010)

Fisher, N.I.: Statistical Analysis of Circular Data. Cambridge University Press, Cambridge (1995)

Fletcher, P.: Geodesic regression and the theory of least squares on Riemannian manifolds. Int. J. Comput. Vis. **105**, 171–185 (2013)

Gigli, N.: On the inverse implication of Brenier-Mccann theorems and the structure of $(P_2(M), W_2)$. Methods Appl. Anal. **18**(2), 127–158 (2011)

Huckemann, S., Hotzand, T., Munk, A.: Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric lie group actions. Stat. Sin. **20**, 1–58 (2010)

Hundrieser, S., Klatt, M., Munk, A.: The statistics of circular optimal transport. In: Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale, pp. 57–82. Springer, Berlin (2022)

Janati, H., Cuturi, M., Gramfort, A.: Debiased sinkhorn barycenters. In: International Conference on Machine Learning, pp. 4692–4701. PMLR (2020)

Kim, Y.-H., Pass, B.: Wasserstein barycenters over Riemannian manifolds. Adv. Math. **307**, 640–683 (2017)

Landler, L., Ruxton, G.D., Malkemper, E.P.: Circular data in biology: advice for effectively implementing statistical procedures. Behav. Ecol. Sociobiol. **72**, 1–10 (2018)

Lee, J.M.: Introduction to Smooth Manifold. Graduate Texts in Mathematics, vol. 218, 2nd edn. Springer, New York (2013)

Manole, T., Balakrishnan, S., Niles-Weed, J., Wasserman, L.: Plugin estimation of smooth optimal transport maps. arXiv preprint arXiv:2107.12364 (2021)

Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley & Sons, Hoboken (2009)

McCann, R.J.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. **11**(3), 589–608 (2001)

Munkres, J.R.: Topology, 2n edn. Prentice Hall Inc., Upper Saddle River (2000)

Panaretos, V.M., Zemel, Y.: An Invitation to Statistics in Wasserstein Space. Springer Nature, Berlin (2020)

Patrangenaru, V., Ellingson, L.: Nonparametric Statistics on Manifolds and Their Application to Object Data Analysis. CRC Press, Boca Raton (2015)

Pegoraro, M., Beraha, M.: Projected statistical methods for distributional data on the real line with the Wasserstein metric. J. Mach. Learn. Res. **23**(37), 1–59 (2022)

Pennec, X.: Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. J. Math. Imaging Vis. **25**, 127–154 (2006)

Pennec, X.: Statistical computing on manifolds: from Riemannian geometry to computational anatomy. In: LIX Fall Colloquium on Emerging Trends in Visual Computing, pp. 347–386. Springer, Berlin (2008)

Pewsey, A., García-Portugués, E.: Recent advances in directional statistics. TEST **30**(1), 1–58 (2021)

Srivastava, S., Cevher, V., Dinh, Q., Dunson, D.: WASP: scalable Bayes via barycenters of subset posteriors. In: Lebanon, G., Vishwanathan, S.V.N. (eds.) Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, vol. 38 (2015)

Zemel, Y., Panaretos, V.M.: Fréchet means and procrustes analysis in Wasserstein space. Bernoulli **25**(2), 932–976 (2019)

Zhang, C., Kokoszka, P., Petersen, A.: Wasserstein autoregressive models for density time series. arXiv preprint arXiv:2006.12640 (2020)

Zhu, C., Müller, H.-G.: Autoregressive optimal transport models. J. R. Stat. Soc. Ser. B Stat. Methodol. **85**(3), 1012–1033 (2023)