Hyperbolic Procrustes Analysis Using Riemannian Geometry

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised alignment between datasets collected at different times, locations, or by different instruments is a fundamental scientific task. Hyperbolic spaces have recently provided a fruitful foundation for the development of informative representations of hierarchical data. Here, we take a purely geometric approach for unsupervised alignment of hierarchical datasets and introduce hyperbolic Procrustes analysis (HPA). HPA consists of new implementations of the three prototypical Procrustes analysis components: translation, scaling, and rotation, based on the Riemannian geometry of the Lorentz model of hyperbolic space. We analyze the proposed components, highlighting their useful properties for alignment. The efficacy of HPA, its analytic properties, stability and computational efficiency are demonstrated in simulations. In addition, we showcase its performance on three batch correction tasks involving gene expression and mass cytometry data. Specifically, we demonstrate high-quality unsupervised batch effect removal from data acquired at different sites and with different technologies that outperforms recent methods for unsupervised alignment in hyperbolic spaces.

1 Introduction

A key scientific task in modern data analysis is the alignment of data. The need for alignment often arises since data are acquired in multiple domains, under different environmental conditions, using various acquisition equipment, and at different sites. This paper focuses on the problem of unsupervised alignment of data embedded in hyperbolic spaces. Recently, hyperbolic spaces have accentuated in geometric representation learning. These non-Euclidean spaces have become popular since they provide a natural embedding of hierarchical data thanks to the exponential growth of the lengths of their geodesic paths [37, 38, 25, 12, 11, 27].

The problem of alignment of data embedded in hyperbolic spaces has been extensively studied, e.g., in the context of natural language processing [44], ontology matching [7], matching two data modalities [35], and improving the embedding in hyperbolic spaces [1]. A few of these studies are based on optimal transport (OT) [1, 18], a classical problem in mathematics [33] that has recently reemerged in modern data analysis, e.g., for domain adaptation [4]. Despite its increasing usage, OT for unsupervised alignment is fundamentally limited [49], since OT (as any density matching approach) cannot recover volume-preserving maps [2, 3, 31].

In this paper, we resort to Procrustes analysis (PA) [14, 15] that is based on purely geometric considerations. PA has been widely used for aligning datasets by eliminating the shift, scaling, and rotational factors. Over the years, it has been successfully applied to various applications, e.g., image registration [29], manifold alignment [47], shape matching [30], domain adaptation [42], and manifold learning [22], to name but a few. Here, we address the problem of unsupervised matching of hierarchical data embedded in hyperbolic spaces. We present hyperbolic Procrustes analysis

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

(HPA), a new PA method in the Lorentz model of hyperbolic geometry. The main novelty lies in the introduction of new implementations of the three prototypical PA components based on Riemannian geometry. Specifically, translation is viewed as a Riemannian mean alignment, implemented using parallel transport (PT). Scaling is determined with respect to geodesic paths. Rotation is considered as moment alignment on a projection of the tangent space of the manifold. Our analysis provides new derivations in the Riemannian geometry of the Lorentz model and specifies the commuting properties of the HPA components. We show that HPA, compared to existing baselines and OT-based methods, achieves improved alignment in a purely unsupervised setting. In addition, it has a natural and stable out-of-sample extension, it supports both small and big data, and it is computationally efficient.

We show application to batch correction in bioinformatics tasks. We present results on both gene expression and mass cytometry (CyTOF) data, exemplifying the generality and broad scope of our method. In contrast to recent works [23, 45], our method does not require landmark correspondence, which is often unavailable in many datasets or hard to obtain. Specifically, we show that batch effects caused by acquisition using different technologies, at different sites, and at different times can be accurately removed, while preserving the intrinsic structure of the data.

Our main contributions are as follows. (i) We present a new implementation of PA using the Riemannian geometry of the Lorentz model for unsupervised label-free hierarchical data alignment. (ii) We provide theoretical analysis and justification of our alignment method based on new derivations of Riemannian geometry operations in the Lorentz model. These derivations have their own merit as they could be used in other contexts. (iii) We show experimental results of accurate batch effect removal from several hierarchical bioinformatics datasets without landmark correspondence.

2 Background on hyperbolic geometry

Hyperbolic space is a non-Euclidean space with a negative constant sectional curvature and an underlying geometry that describes tree-like graphs with small distortions [41]. There exist four commonly-used models for hyperbolic spaces: Poincaré disk model, Lorentz model (hyperboloid model), Poincaré half-plane model, and Beltrami-Klein model. These four models are equivalent and there exist transformations between them. Here, we consider the Lorentz model, and specifically, the upper sheet of the hyperboloid model, because its basic Riemannian operations have simple closed-form expressions and the computation of the geodesic distances is stable [38, 25].

Formally, the upper sheet of the hyperboloid model in a d-dimensional hyperbolic space is defined by $\mathbb{L}^d := \{\boldsymbol{x} \in \mathbb{R}^{d+1} | \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1, \boldsymbol{x}(1) > 0\}$, where $\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = \boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x}$ is the Lorentzian inner product and $\boldsymbol{H} \in \mathbb{R}^{(d+1)\times(d+1)}$ is the hyperbolic metric tensor, defined by $\boldsymbol{H} = \begin{bmatrix} -1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{I}_d. \end{bmatrix}$. The Lorentzian norm of a hyperbolic vector $\boldsymbol{x} \in \mathbb{L}^d$ is denoted by $||\boldsymbol{x}||_{\mathcal{L}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}}}$, with the origin $\boldsymbol{\mu}_0 = [1, \boldsymbol{0}]^\top \in \mathbb{L}^d$. Let $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$ be the tangent space at $\boldsymbol{x} \in \mathbb{L}^d$, defined by $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d := \{\boldsymbol{v} | \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}} = 0\}$. Consider $\boldsymbol{x} \in \mathbb{L}^d$ and $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$, the geodesic path $\phi : \mathbb{R}_0^+ \to \mathbb{L}^d$ is defined by $\phi(t) = \cosh(||\boldsymbol{v}||_{\mathcal{L}} t)\boldsymbol{x} + \sinh(||\boldsymbol{v}||_{\mathcal{L}} t)\frac{\boldsymbol{v}}{||\boldsymbol{v}||_{\mathcal{L}}}$ with $\phi(0) = \boldsymbol{x}$ and initial velocity $\phi'(0) = \boldsymbol{v}$, where $\phi'(t) := \frac{d}{dt}\phi(t)$. In addition, the associated geodesic distance is $d_{\mathbb{L}^d}(\boldsymbol{x}, \phi_{\boldsymbol{v}}(t)) = \cosh^{-1}(-\langle \boldsymbol{x}, \phi_{\boldsymbol{v}}(t) \rangle_{\mathcal{L}})$.

The Exponential map, projecting the point $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$ to the manifold $\mathbb{L}^d$, is given by $\mathrm{Exp}_{\boldsymbol{x}}(\boldsymbol{v}) = \phi(1) = \cosh(||\boldsymbol{v}||_{\mathcal{L}})\boldsymbol{x} + \sinh(||\boldsymbol{v}||_{\mathcal{L}})\frac{\boldsymbol{v}}{||\boldsymbol{v}||_{\mathcal{L}}}$. The Logarithmic map, projecting a point $\boldsymbol{y} \in \mathbb{L}^d$ to the tangent space $\mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$ at $\boldsymbol{x}$, is defined by $\mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{\cosh^{-1}(\lambda)}{\sqrt{\lambda^2 - 1}}(\boldsymbol{y} - \lambda \boldsymbol{x})$, where $\lambda = -\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}$. The PT of a vector $\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{x}}\mathbb{L}^d$ along the geodesic path from $\boldsymbol{x} \in \mathbb{L}^d$ to $\boldsymbol{y} \in \mathbb{L}^d$ is defined by $\mathrm{PT}_{\boldsymbol{x}\to\boldsymbol{y}}(\boldsymbol{v}) = \boldsymbol{v} + \frac{\langle \boldsymbol{y} - \lambda \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathcal{L}}}{\lambda + 1}(\boldsymbol{x} + \boldsymbol{y})$, where $\lambda = -\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}$, keeping the metric tensor unchanged. The Riemannian mean $\overline{\boldsymbol{x}}_{\mathcal{X}}$ and the corresponding dispersion $d_{\mathcal{X}}$ of a set $\mathcal{X} = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{L}^d\}_{i=1}^n$ are defined using the Fréchet mean [10, 28] by

$$\overline{\boldsymbol{x}}_{\mathcal{X}} := m(\mathcal{X}) = \underset{\boldsymbol{x} \in \mathbb{L}^d}{\arg\min} \sum_{i=1}^{n} d^2_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{x}_i) \text{ and } d_{\mathcal{X}} := r(\mathcal{X}) = \min_{\boldsymbol{x} \in \mathbb{L}^d} \frac{1}{n} \sum_{i=1}^{n} d^2_{\mathbb{L}^d}(\boldsymbol{x}, \boldsymbol{x}_i), \tag{1}$$

where $m : \mathcal{X} \to \mathbb{L}^d$ and $r : \mathcal{X} \to \mathbb{R}^+$. The Fréchet mean is commonly computed by the Karcher Flow [20], which is computationally demanding. Importantly, in the considered hyperbolic space, the Fréchet mean can be efficiently obtained using the accurate gradient formulation [28].

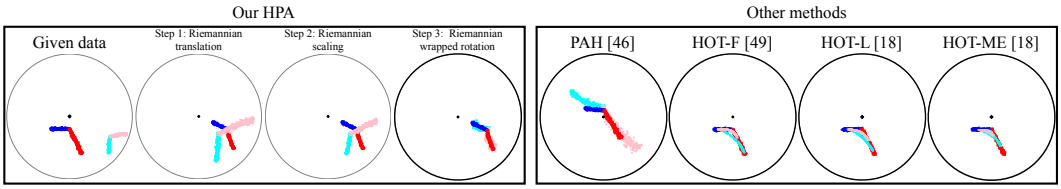

Figure 1: Illustration of our alignment method (HPA). Two sets of points in hyperbolic space are given, depicted in dark and light colors. Each point is associated with one of two labels: (i) blue/cyan, and (ii) red/pink. Left: Alignment results after each step of our HPA implemented in $\mathbb{L}^2$: Riemannian translation, Riemannian scaling, and Riemannian wrapped rotation. Note that these Riemannian operations in hyperbolic space are different than their Euclidean counterparts. Right: Alignment results of four methods: PAH [46] and HOT-F [49] applied in $\mathbb{L}^2$, and HOT-L [18] and HOT-ME [18] applied in the 2D Poincaré disk. For visualization, all points in $\mathbb{L}^2$ are transformed to the 2D Poincaré disk. The alignment after Step 3 of HPA (circled in black) is more accurate than the alignments obtained by the other methods.

Given a vector $\overline{\boldsymbol{x}} \in \mathbb{L}^d$ and a symmetric and positive-definite (SPD) matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, the wrapped normal distribution $\mathcal{G}(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma})$ provides a generative model of hyperbolic samples as follows [34, 8]. First, a vector $\boldsymbol{v}'$ is sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then, 0 is concatenated to the vector $\boldsymbol{v}'$ such that $\boldsymbol{v} = [0, \boldsymbol{v}']^\top \in \mathcal{T}_{\boldsymbol{\mu}_0}\mathbb{L}^d$. Finally, PT from the origin $\boldsymbol{\mu}_0$ to $\overline{\boldsymbol{x}}$ is applied to $\boldsymbol{v}$, and the resulting point is mapped to the manifold using the Exponential map at $\overline{\boldsymbol{x}}$. The probability density function of this model is given by $\log \mathcal{G}(\boldsymbol{y}|\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}) = \log \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) - (n-1)\log(\frac{\sinh ||\boldsymbol{v}'||_2}{||\boldsymbol{v}'||_2})$.

3 Hyperbolic Procrustes analysis

Existing methods for data alignment typically seek a function that minimizes a certain cost. A large body of work attempts to match the empirical densities of two datasets, e.g., by minimizing the maximum mean discrepancy (MMD) [43, 24] or solving OT problems [40, 1, 18]. Finding an effective cost function without labels or landmarks is challenging, and minimizing such costs directly often lead to poor alignment in practice (see illustration in Fig. 1). A different well-established approach that applies indirect alignment based on geometric considerations is PA. While preparing this manuscript, another method of PA in hyperbolic spaces (PAH) was presented for matching two sets, assuming that they consist of the same number of points and that there exists a point-wise isometric map between them [46]. We remark that the analysis we present here applies to broader settings and makes no such assumptions.

We consider two sets of points $\mathcal{H}^{(1)} = \{\boldsymbol{h}_i^{(1)}\}_{i=1}^{N_1}$ and $\mathcal{H}^{(2)} = \{\boldsymbol{h}_i^{(2)}\}_{i=1}^{N_2}$ in $\mathbb{L}^d$. Here, we aim to find a function $\zeta : \mathbb{L}^d \to \mathbb{L}^d$, consisting of three components: translation, scaling, and rotation, that aligns $\mathcal{H}^{(2)}$ with $\mathcal{H}^{(1)}$ in an unsupervised label-free manner as depicted in Fig. 1. Finding such a function can be viewed as an extension of classical PA from the Euclidean space $\mathbb{R}^{d+1}$ to the Lorentz model $\mathbb{L}^d$. A natural extension to multiple sets is described in Section 3.5. We remark that the statements are written in the context of the problem at hand. In Appendix A, we restate them more generally and present their proofs.

3.1 Riemannian translation

Let $\overline{\boldsymbol{h}}^{(1)}$ and $\overline{\boldsymbol{h}}^{(2)}$ denote the Riemannian means of the sets $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, respectively. In this translation component, we find a map $\Gamma_{\overline{\boldsymbol{h}}^{(2)}\to\overline{\boldsymbol{h}}^{(1)}} : \mathbb{L}^d \to \mathbb{L}^d$ that aligns the Riemannian means of the sets. In the spirit of [48], we propose to construct $\Gamma_{\overline{\boldsymbol{h}}^{(2)}\to\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)})$ as the composition of three Riemannian operations in $\mathbb{L}^d$: the Logarithmic map applied to $\boldsymbol{h}_i^{(2)}$ at $\overline{\boldsymbol{h}}^{(2)}$, PT $\boldsymbol{h}_i^{(2)}$ from $\overline{\boldsymbol{h}}^{(2)}$ to $\overline{\boldsymbol{h}}^{(1)}$ along the geodesic path, and the Exponential map applied to the transported point at $\overline{\boldsymbol{h}}^{(1)}$:

$$\Gamma_{\overline{\boldsymbol{h}}^{(2)}\to\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)}) \coloneqq \mathrm{Exp}_{\overline{\boldsymbol{h}}^{(1)}}(\mathrm{PT}_{\overline{\boldsymbol{h}}^{(2)}\to\overline{\boldsymbol{h}}^{(1)}}(\mathrm{Log}_{\overline{\boldsymbol{h}}^{(2)}}(\boldsymbol{h}_i^{(2)}))). \tag{2}$$

Since the geodesic path between any two points in $\mathbb{L}^d$ is unique [41], $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ is well-defined. The rationale behind the combination of these three Riemannian operations is twofold. First, PT is a map that aligns the means of the sets, while preserving their internal structure. Second, the Logarithmic and Exponential maps compose a map whose domain and range are the Lorentz model $\mathbb{L}^d$ rather than the tangent space, as desired. We make these claims formal in the following results.

Proposition 1. *The map $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ defined in Eq. (2) aligns the means of the sets, i.e., it satisfies*

$$\overline{\boldsymbol{h}}^{(1)} = m(\{\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)})\}_{i=1}^{N_2}), \tag{3}$$

where m is the function defined in Eq. (1).

Proposition 2. *The map $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ can be recast as:*

$$\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)}) = \boldsymbol{h}_i^{(2)} - \beta(\boldsymbol{h}_i^{(2)})\overline{\boldsymbol{h}}^{(2)} + \gamma(\boldsymbol{h}_i^{(2)})\overline{\boldsymbol{h}}^{(1)}, \tag{4}$$

for all $\boldsymbol{h}_i^{(2)} \in \mathcal{H}^{(2)}$, where

$$0 < \beta(\boldsymbol{h}_i^{(2)}) = -\left\langle \frac{\overline{\boldsymbol{h}}^{(1)} + \overline{\boldsymbol{h}}^{(2)}}{\alpha+1}, \boldsymbol{h}_i^{(2)} \right\rangle_{\mathcal{L}}, \; 0 < \gamma(\boldsymbol{h}_i^{(2)}) = \left\langle \frac{\overline{\boldsymbol{h}}^{(1)} - (2\alpha+1)\overline{\boldsymbol{h}}^{(2)}}{\alpha+1}, \boldsymbol{h}_i^{(2)} \right\rangle_{\mathcal{L}}$$

and $0 < \alpha = -\langle \overline{\boldsymbol{h}}^{(1)}, \overline{\boldsymbol{h}}^{(2)} \rangle_{\mathcal{L}}$.

In addition to providing a compact closed-form expression, Proposition 2 gives the proposed translation based on Riemannian geometry an interpretation of standard mean alignment in linear vector spaces. It implies that the alignment is nothing but subtracting the mean of the source set $\overline{\boldsymbol{h}}^{(2)}$ from each vector in $\mathcal{H}^{(2)}$, and adding the mean of the target set $\overline{\boldsymbol{h}}^{(1)}$ (with the appropriate scales).

Proposition 3. *The map $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ preserves distances (i.e., it is an isometry):*

$$d_{\mathbb{L}^d}(\boldsymbol{h}_i^{(2)}, \boldsymbol{h}_j^{(2)}) = d_{\mathbb{L}^d}(\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(2)}), \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_j^{(2)})), \tag{5}$$

for any two points $\boldsymbol{h}_i^{(2)}, \boldsymbol{h}_j^{(2)} \in \mathcal{H}^{(2)}$.

Let $\psi(t)$ be the unique geodesic path from $\overline{\boldsymbol{h}}^{(2)}$ to $\overline{\boldsymbol{h}}^{(1)}$ such that $\psi(0) = \overline{\boldsymbol{h}}^{(2)}$ and $\psi(1) = \overline{\boldsymbol{h}}^{(1)}$, and let $\psi'(0) \in \mathcal{T}_{\overline{\boldsymbol{h}}^{(2)}}\mathbb{L}^d$ and $\psi'(1) \in \mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d$ be the corresponding velocities, respectively.

Proposition 4. *The map $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ aligns geodesic velocities, i.e., given the mapping of the geodesic velocities to the manifold $\mathbb{L}^d \ni \boldsymbol{v}_0 = \mathrm{Exp}_{\overline{\boldsymbol{h}}^{(2)}}(\psi'(0)) = \overline{\boldsymbol{h}}^{(1)}$ and $\mathbb{L}^d \ni \boldsymbol{v}_1 = \mathrm{Exp}_{\overline{\boldsymbol{h}}^{(1)}}(\psi'(1))$, we have*

$$\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{v}_0) = \boldsymbol{v}_1. \tag{6}$$

Isometry is determined up to rotation, a fact that can be problematic for alignment. For example, any H-unitary matrix [13] can be an isometric function in $\mathbb{L}^d$. When landmarks are given, they can be used to alleviate this redundancy. However, in the purely unsupervised setting we consider, other data-driven ques are required. Proposition 4 implies that the proposed translation based on PT fixes some of these rotational degrees of freedom by aligning the geodesic velocities. In Section 3.3 we revisit this issue. Now, with a slight abuse of notation, let $\widetilde{\mathcal{H}}^{(2)} = \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\mathcal{H}^{(2)})$.

Proposition 5. *Consider two subsets $\mathcal{A}, \mathcal{B} \subset \mathcal{H}^{(2)}$ and their translations $\widetilde{\mathcal{A}} = \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\mathcal{A}), \widetilde{\mathcal{B}} = \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}(\mathcal{B}) \subset \widetilde{\mathcal{H}}^{(2)}$. Let $\overline{\boldsymbol{a}} = m(\mathcal{A}), \overline{\boldsymbol{b}} = m(\mathcal{B}), \widetilde{\boldsymbol{a}} = m(\widetilde{\mathcal{A}})$, and $\widetilde{\boldsymbol{b}} = m(\widetilde{\mathcal{B}})$ be the Riemannian means of the subsets. Then,*

$$\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}} \circ \Gamma_{\overline{\boldsymbol{a}} \to \overline{\boldsymbol{b}}} = \Gamma_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}. \tag{7}$$

In the context of the alignment problem, the importance of Proposition 5 is the following. Suppose the two sets correspond to data measured at two labs (denoted with and without a tilde), and suppose each set was acquired by two types of equipment (denoted by $\mathcal{A}$ and $\mathcal{B}$). Proposition 5 implies that aligning data from the different labs and then aligning data acquired by the different equipment is equivalent to first aligning the different equipment and then the different labs, i.e., any order of the two alignments generates the same result. Seemingly, this is a natural property in Euclidean spaces. However, in a Riemannian manifold, it is not a trivial result and it holds for the transport along the geodesic path. See Appendix A for counter-examples.

3.2 Riemannian scaling

Let $d^{(1)}$ and $d^{(2)}$ denote the Riemannian dispersions of $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$. By propositions 1 and 3, $\overline{\boldsymbol{h}}^{(1)}$ and $d^{(2)}$ are the mean and dispersion of $\widetilde{\mathcal{H}}^{(2)}$. Here, our goal is to align the Riemannian dispersions of $\mathcal{H}^{(1)}$ and $\widetilde{\mathcal{H}}^{(2)}$. For this purpose, we propose the scaling function $\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} : \mathbb{L}^d \to \mathbb{L}^d$, given by

$$\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}}(\widetilde{\boldsymbol{h}}_i^{(2)}) = \phi_i(s), \tag{8}$$

where $s = \sqrt{d^{(1)}/d^{(2)}}$ is the scaling factor and $\phi_i(t)$ is the geodesic path from $\overline{\boldsymbol{h}}^{(1)}$ to $\widetilde{\boldsymbol{h}}_i^{(2)}$ such that $\phi_i(0) = \overline{\boldsymbol{h}}^{(1)}$ and $\phi_i(1) = \widetilde{\boldsymbol{h}}_i^{(2)}$.

Proposition 6. *The dispersion of the rescaled set* $\widehat{\mathcal{H}}^{(2)} = \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}}(\widetilde{\mathcal{H}}^{(2)})$ *is* $d^{(1)}$.

3.3 Riemannian wrapped rotation

The purpose of this component is to align the orientation of the distributions of the two sets after translation and scaling, namely, after aligning their first and second moments. The proposed rotation function $\Theta_{\overline{\boldsymbol{h}}^{(1)}} : \mathbb{L}^d \to \mathbb{L}^d$ consists of mapping the points from the manifold $\mathbb{L}^d$ to the tangent space $\mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d$, projecting to $\mathbb{R}^d$, rotating in $\mathbb{R}^d$, and mapping back to the tangent space and then to the manifold. We perform the rotation in $\mathbb{R}^d$, which we term *wrapped rotation*, rather than a direct rotation on the manifold $\mathbb{L}^d$ or on the tangent space $\mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d$ for the following reasons. First, the frequently used rotation map in $\mathbb{L}^d$ [46] does not necessarily preserve the Riemannian mean, and in our context, it might reverse the mean alignment. Second, rotation applied directly to the tangent space $\mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d$ does not guarantee that the rotated points remain on the same tangent space. Third, applying the rotation in $\mathbb{R}^d$ allows us to use the standard Euclidean rotation using SVD. In Section 4, we empirically demonstrate the advantage of the proposed rotation compared to the alternatives.

Definition 1. *Let* $\mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}} : \mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d \to \mathbb{R}^d$ *be the following projection map defined on the tangent space at* $\overline{\boldsymbol{h}}^{(1)}$*:*

$$\mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{v}) := \left[\boldsymbol{v}(2), \boldsymbol{v}(3), \ldots, \boldsymbol{v}(d+1)\right]^\top \in \mathbb{R}^d, \tag{9}$$

and define its inverse map by

$$\mathcal{P}^{-1}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{s}) := \left[\frac{\langle \boldsymbol{s}, \mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\overline{\boldsymbol{h}}^{(1)})\rangle}{\overline{\boldsymbol{h}}^{(1)}(1)}, \boldsymbol{s}\right]^\top \in \mathcal{T}_{\overline{\boldsymbol{h}}^{(1)}}\mathbb{L}^d, \tag{10}$$

where $\boldsymbol{s} \in \mathbb{R}^d$ *and* $\langle \cdot, \cdot \rangle$ *is the standard Euclidean inner product.*

Note that removing the first element of $\boldsymbol{v}$ is valid due to the constraint imposed on the vector elements in the tangent space by definition. Indeed, no information is lost and the projection is invertible.

The first step in our rotation component is to map the points in $\mathcal{H}^{(1)}$ and in $\widehat{\mathcal{H}}^{(2)}$ to the tangent space at $\overline{\boldsymbol{h}}^{(1)}$: $\boldsymbol{v}_i^{(1)} = \mathrm{Log}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{h}_i^{(1)})$ for $i = 1, \ldots, N_1$, and $\boldsymbol{v}_i^{(2)} = \mathrm{Log}_{\overline{\boldsymbol{h}}^{(1)}}(\widehat{\boldsymbol{h}}_i^{(2)})$ for $i = 1, \ldots, N_2$. In the second step, we project the points by Definition 1 and re-center them:

$$\boldsymbol{s}_i^{(1)} = \mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{v}_i^{(1)}) - \overline{\boldsymbol{s}}^{(1)},\ i = 1, \ldots, N_1 \text{ and } \boldsymbol{s}_i^{(2)} = \mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{v}_i^{(2)}) - \overline{\boldsymbol{s}}^{(2)},\ i = 1, \ldots, N_2,$$

where $\overline{\boldsymbol{s}}^{(k)} = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{v}_i^{(k)})$ for $k = 1, 2$ is the mean vector of the projections. Then, the projected and centered points (in $\mathbb{R}^d$) are collected into matrices:

$$\boldsymbol{S}^{(k)} = \left[\boldsymbol{s}_1^{(k)}, \boldsymbol{s}_2^{(k)}, \ldots, \boldsymbol{s}_{N_k}^{(k)}\right] \in \mathbb{R}^{d \times N_k}. \tag{11}$$

In the third step, for each set $k = 1, 2$, we compute the rotation matrix $\boldsymbol{U}^{(k)} \in \mathbb{R}^{d \times d}$ by applying SVD to the matrix $\boldsymbol{S}^{(k)} = \boldsymbol{U}^{(k)}\boldsymbol{\Lambda}^{(k)}(\boldsymbol{E}^{(k)})^\top$. Since the left-singular vectors are determined up to a sign, we propose to align their signs as follows: $\boldsymbol{u}_i^{(2)} \leftarrow \mathrm{sign}(\langle \boldsymbol{u}_i^{(2)}, \boldsymbol{u}_i^{(1)}\rangle)\boldsymbol{u}_i^{(2)}$, where $\boldsymbol{u}_i^{(1)}$ and $\boldsymbol{u}_i^{(2)}$

are the i-th left-singular vector of the two sets, resulting in modified rotation matrices $\boldsymbol{U}^{(k)}$. Finally, we apply the rotation to $\widehat{\mathcal{H}}^{(2)}$ by

$$\Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}}(\widehat{\boldsymbol{h}}_i^{(2)}) = \mathrm{Exp}_{\overline{\boldsymbol{h}}^{(1)}}(\mathcal{P}^{-1}_{\overline{\boldsymbol{h}}^{(1)}}(\boldsymbol{U}^\top(\mathcal{P}_{\overline{\boldsymbol{h}}^{(1)}}(\mathrm{Log}_{\overline{\boldsymbol{h}}^{(1)}}(\widehat{\boldsymbol{h}}_i^{(2)})) - \overline{\boldsymbol{s}}^{(2)}) + \overline{\boldsymbol{s}}^{(2)})), \tag{12}$$

where $\boldsymbol{U} = \boldsymbol{U}^{(1)}(\boldsymbol{U}^{(2)})^\top$.

Proposition 7. *The wrapped rotation is bijective, and the inverse is given by*

$$(\Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}})^{-1} = \Theta^{\boldsymbol{U}^\top}_{\overline{\boldsymbol{h}}^{(1)}}. \tag{13}$$

3.4 Analysis

Putting all three components together, the proposed HPA that aligns $\mathcal{H}^{(2)}$ with $\mathcal{H}^{(1)}$ culminates in the composition of translation, scaling, and rotation:

$$\Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}. \tag{14}$$

As in most PA schemes, the order of the three components is important. Yet, the proposed components allow us a certain degree of freedom, as indicated in the following results.

Proposition 8. *The Riemannian translation and the Riemannian scaling commute w.r.t. the Riemannian means $\overline{\boldsymbol{h}}^{(1)}$ and $\overline{\boldsymbol{h}}^{(2)}$:*

$$\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}} = \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}} \circ \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(2)}}. \tag{15}$$

Note that $\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}}$ and $\Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$ do not necessarily commute: $\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}} \neq \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}} \circ \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}}$.

Proposition 9. *The Riemannian scaling and the wrapped rotation commute:*

$$\Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}} = \Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}}. \tag{16}$$

We note that the rotation does not commute with the translation, because PT only preserves the local covariant derivative on the tangent space but might cause rotation and distortion along the transportation. Therefore, the rotation is required to be the last component of our HPA.

Thus far, we did not present a model for the discrepancy between the two sets, nor we presented the proposed HPA as optimal with respect to some criterion. In the following result, we show that if the discrepancy between the sets can be expressed as a composition of translation, scaling, and rotation, then the two sets can be perfectly aligned using HPA.

Proposition 10. *Let $\eta : \mathbb{L}^d \to \mathbb{L}^d$ be a map, given by $\eta = \Theta^{\boldsymbol{U}}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Upsilon^{s}_{\overline{\boldsymbol{h}}^{(1)}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(2)} \to \overline{\boldsymbol{h}}^{(1)}}$. If $\mathcal{H}^{(1)} = \{\boldsymbol{h}_i^{(1)} = \eta(\boldsymbol{h}_i^{(2)})\}_{i=1}^{N_2}$, then,*

$$\boldsymbol{h}_i^{(2)} = (\Theta^{\boldsymbol{U}'}_{\overline{\boldsymbol{h}}^{(2)}} \circ \Upsilon^{\frac{1}{s}}_{\overline{\boldsymbol{h}}^{(2)}} \circ \Gamma_{\overline{\boldsymbol{h}}^{(1)} \to \overline{\boldsymbol{h}}^{(2)}})(\boldsymbol{h}_i^{(1)}),\ i = 1, \ldots, N_2, \tag{17}$$

where $\boldsymbol{U}' \in \mathbb{O}(d)$.

Note that HPA consists of the sequence of Riemannian translation, Riemannian scaling and wrapped-rotation. The domain and range of each component is the manifold $\mathbb{L}^d$. Yet, the first and last operations of each component are the Logarithmic and Exponential maps that project a point from the manifold to the tangent space, and vice versa, respectively. This allows us to propose an efficient implementation of the sequence without the back and forth projections as described in Appendix B.

3.5 Extension to multiple sets

We can naturally scale up the setting to support the alignment of $K > 2$ sets, denoted by $\mathcal{H}^{(k)} = \{\boldsymbol{h}_i^{(k)}\}_{i=1}^{N_k}$, where $k \in \{1, 2, \ldots, K\}$. Let $\overline{\boldsymbol{h}}^{(k)}$ and $d^{(k)}$ be the Riemannian mean and dispersion of the k-th dataset, respectively. In addition, let $\overline{\boldsymbol{h}}$ be the global Riemannian mean of $\{\overline{\boldsymbol{h}}^{(k)}\}_{k=1}^{K}$. We propose to transport the points of the k-th set using $\Gamma_{\overline{\boldsymbol{h}}^{(k)} \to \overline{\boldsymbol{h}}}$. Next, the Riemannian dispersion of each set is set to 1 by applying $\Upsilon^{s}_{\overline{\boldsymbol{h}}}$ with $s = 1/\sqrt{d^{(k)}}$. Finally, the wrapped rotation is applied to

all the data sets on the projection of the tangent space $\mathcal{T}_{\overline{\boldsymbol{h}}}\mathbb{L}^d$ and then mapped back to the manifold $\mathbb{L}^d$. The first set is designated as the reference set, and all other rotation matrices $\boldsymbol{U}^{(k)}$ are updated according to $\boldsymbol{u}_i^{(k)} \leftarrow \text{sign}(\langle \boldsymbol{u}_i^{(k)}, \boldsymbol{u}_i^{(1)} \rangle)\boldsymbol{u}_i^{(k)}$. We remark that the choice of the reference set may be improved by choosing the set with the smallest sum of rotations w.r.t. the other sets. In addition, the point to which we align all the means need to be far from the boundary to minimize distortions caused by the hyperbolic metric. Indeed, $\overline{\boldsymbol{h}}$ lies in the convex hull of the means $\{\overline{\boldsymbol{h}}^{(k)}\}_{k=1}^K$, and therefore, is naturally away from the boundary. The proposed HPA is summarized in Algorithm 1.

Algorithm 1 Hyperbolic Procrustes analysis

Input: K sets of hyperbolic points $\mathcal{H}^{(1)} = \{\boldsymbol{h}_i^{(1)}\}_{i=1}^{N_1}, \ldots, \mathcal{H}^{(K)} = \{\boldsymbol{h}_i^{(K)}\}_{i=1}^{N_K}$

Output: K *aligned* sets of hyperbolic points $\breve{\mathcal{H}}^{(1)} = \{\breve{\boldsymbol{h}}_i^{(1)}\}_{i=1}^{N_1}, \ldots, \breve{\mathcal{H}}^{(K)} = \{\breve{\boldsymbol{h}}_i^{(K)}\}_{i=1}^{N_K}$

for each set $\mathcal{H}^{(k)}$ **do**
  compute the Riemannian mean $\overline{\boldsymbol{h}}^{(k)}$ and dispersion $d^{(k)}$
end for
compute $\overline{\boldsymbol{h}}$, the global Riemannian mean of $\{\overline{\boldsymbol{h}}^{(k)}\}_{k=1}^K$
for each set $\mathcal{H}^{(k)}$ **do**
  apply the Riemannian translation $\widetilde{\boldsymbol{h}}_i^{(k)} = \Gamma_{\overline{\boldsymbol{h}}^{(k)} \to \overline{\boldsymbol{h}}}(\boldsymbol{h}_i^{(k)})$ // Eq. (2)
  apply the Riemannian scaling $\widehat{\boldsymbol{h}}_i^{(k)} = \Upsilon_{\overline{\boldsymbol{h}}}^s(\widetilde{\boldsymbol{h}}_i^{(k)})$ with $s = 1/\sqrt{d^{(k)}}$ // Eq. (8)
  apply the wrapped rotation $\breve{\boldsymbol{h}}_i^{(k)} = \Theta_{\overline{\boldsymbol{h}}}^{\boldsymbol{U}}(\widehat{\boldsymbol{h}}_i^{(k)})$ with $\boldsymbol{U} = \boldsymbol{U}^{(1)}(\boldsymbol{U}^{(k)})^\top$ // Eq. (12)
end for

4 Experimental results

We apply HPA to simulations and to three biomedical datasets. In addition, we test HPA on MNIST [26] and USPS [19] datasets, which arguably do not have a distinct hierarchical structure. Nonetheless, we demonstrate in Appendix C that our HPA is highly effective in aligning these two datasets. All the experiments are label-free. We compare the obtained results to the following alignment methods: (i) PAH [46], which is applied only to the simulated data since it requires the existence of a one-to-one correspondence between the sets, (ii) only the Riemannian translation (RT), (iii) OT in hyperbolic space with the weighted Fréchet mean (HOT-F) extended to an unsupervised setting according to [49], (iv) OT with W-linear map (HOT-L) [18], and (v) hyperbolic mapping estimation (HOT-ME) [18]. As a baseline, we present the results obtained before the alignment (Baseline). For more details on the experimental setting, see Appendix B.

4.1 Simulations

The synthetic data in $\mathbb{L}^d$ is generated using the sampling scheme described in Section 2 based on [34]. Given an arbitrary point $\boldsymbol{\mu} \in \mathbb{L}^d$ and an arbitrary SPD matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, we generate a set of N points $\mathcal{Q}^{(1)} = \{\boldsymbol{q}_i^{(1)}\}_{i=1}^N$ centered at $\boldsymbol{\mu}$ by $\mathbb{L}^d \ni \boldsymbol{q}_i^{(1)} = \text{Exp}_{\boldsymbol{\mu}}(\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\widetilde{\boldsymbol{v}}_i^{(1)}))$, where $\boldsymbol{\mu}_0 = [1, \mathbf{0}]^\top$ is the origin, $\boldsymbol{v}_i^{(1)} = [0, \widetilde{\boldsymbol{v}}_i^{(1)}]^\top$, and $\widetilde{\boldsymbol{v}}_i^{(1)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Next, we generate three noisy and distorted versions of $\mathcal{Q}^{(1)}$. The first noisy set $\mathcal{Q}^{(2)} = \{\boldsymbol{q}_i^{(2)}\}_{i=1}^N$ is generated as proposed in [46] by $\boldsymbol{q}_i^{(2)} = \boldsymbol{L}\boldsymbol{T}_{\epsilon_i}\boldsymbol{q}_i^{(1)}$, where $\boldsymbol{T}_{\epsilon_i}$ is a hyperbolic translation defined by $\boldsymbol{T}_{\boldsymbol{\epsilon}_i} = \begin{bmatrix} \sqrt{1 + \boldsymbol{\epsilon}_i^\top \boldsymbol{\epsilon}_i} & \boldsymbol{\epsilon}_i^\top \\ \boldsymbol{\epsilon}_i & (\boldsymbol{I} + \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^\top)^{\frac{1}{2}} \end{bmatrix}$, $\boldsymbol{\epsilon}_i$ is sampled from $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, σ^2 is the variance, and $\boldsymbol{L}$ is a random H-unitary matrix [13]. Another noisy set, denoted as $\mathcal{Q}^{(3)} = \{\boldsymbol{q}_i^{(3)}\}_{i=1}^N$, is generated by $\boldsymbol{q}_i^{(3)} = \boldsymbol{L}(\text{Exp}_{\boldsymbol{\mu}}(\text{PT}_{\boldsymbol{\mu}_0 \to \boldsymbol{\mu}}(\boldsymbol{u}_i^{(1)})))$, where $\boldsymbol{u}_i^{(1)} = [0, \widetilde{\boldsymbol{v}}_i^{(1)} + \boldsymbol{\epsilon}_i]^\top$. Here, the noise is added to the tangent space at $\boldsymbol{\mu}_0$. Finally, let $\mathcal{Q}^{(4)} = \{\boldsymbol{q}_i^{(4)}\}_{i=1}^N$ be a distorted set, given by $\boldsymbol{q}_i^{(4)} = f_{\boldsymbol{\mu}'}(\boldsymbol{q}_i^{(3)})$, where $f_{\boldsymbol{\mu}'}(\boldsymbol{x}) = \cosh(||\boldsymbol{u}||_{\mathcal{L}} t)\boldsymbol{\mu}' + \sinh(||\boldsymbol{u}||_{\mathcal{L}} t)\frac{\boldsymbol{u}}{||\boldsymbol{u}||_{\mathcal{L}}}$ and $\boldsymbol{u} = \text{Log}_{\boldsymbol{\mu}'}(\boldsymbol{x})$, for arbitrary (fixed) $\boldsymbol{\mu}' \in \mathbb{L}^d$ and $t > 0$.

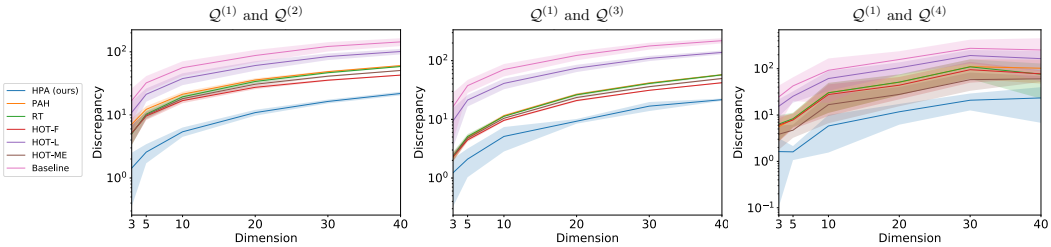

Figure 2: Alignment results of the simulated data.

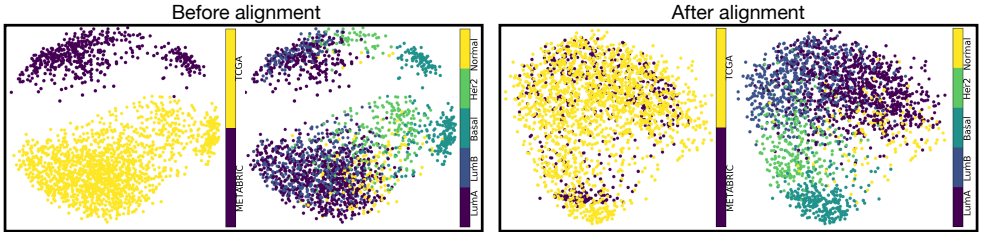

Figure 3: The visualizations of HPA applied to the two breast cancer gene expression datasets.

We apply Algorithm 1 to align the three pairs of sets $\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)}\}$, $\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(3)}\}$, and $\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(4)}\}$, setting $N = 100$, $\sigma = 1$, and $d \in \{3, 5, 10, 20, \ldots, 40\}$. Each experiment is repeated 10 times with different values of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}'$ and t. To evaluate the alignment, we use the pairwise discrepancy based on the *hidden* one-to-one correspondence, given by $\varepsilon(\mathcal{Q}^{(1)}, \mathcal{Q}^{(j)}) = \frac{1}{N}\sum_{i=1}^{N} d^2_{\mathbb{L}^d}(\boldsymbol{q}_i^{(1)}, \boldsymbol{q}_i^{(j)})$, where $j \in \{2, 3, 4\}$. The discrepancy as a function of the dimension d is shown in Fig. 2. We observe that the proposed HPA has lower discrepancy relative to the other label-free methods. Specifically, it outperforms OT-based methods that are designed to match the densities. Furthermore, the proposed HPA is stable, in contrast to HOT-L, which is highly sensitive to the noise and distortion introduced in $\mathcal{Q}^{(3)}$ and $\mathcal{Q}^{(4)}$. Interestingly, we remark that the discrepancies of RT and PAH are very close, empirically showing that RT alone is comparable to PAH. In addition, note that HPA is permutation-invariant and does not require one-to-one correspondence as PAH. We report the running-time in Appendix C and demonstrate that HPA is more efficient than HOT-F and HOT-ME.

4.2 Batch effect removal

We consider bioinformatics datasets consisting of gene expression data and CyTOF. Representing such data in hyperbolic spaces was shown to be informative and useful [21], implying that such data have an underlying inherent hierarchical structure. Batch effects [39] arise from experimental variations that can be attributed to the measurement device or other environmental factors. Batch correction is typically a critical precursor to any subsequent analysis and processing.

Three batch effect removal tasks are examined. The first task involves breast cancer (BC) gene expression data. We consider two publicly available datasets: METABRIC [5] and TCGA [36], consisting of samples from five breast cancer subtypes. The batch effect stems from different profiling techniques: gene expression microarray and RNA sequencing. In the second task, three cohorts of lung cancer (LC) gene expression data [17] are considered, consisting of samples from three lung cancer subtypes. The data were collected using gene expression microarrays at three different sites (a likely source of batch effects): Stanford University (ST), University of Michigan (UM), and Dana-Farber Cancer Institute (D-F). The last task involves CyTOF data [43] consisting of peripheral blood mononuclear cells (PBMCs) collected from two multiple sclerosis patients during four days: two day before treatment (BT) and two days after treatment (AT). These $8 = 2 \times 2 \times 2$ batches were collected with or without PMA/ionomycin stimulated PBMCs. We aim to remove the batch effects between two different days from the same condition (BT/AT) and from the same patient. In each batch removal task, we first learn an embedding of the data from all the batches into the Lorentz model $\mathbb{L}^d$ [38]. Then, HPA is applied to the embedded points in $\mathbb{L}^d$.

Table 1: The k-NN AUC-ROC from the best k in each method.

Datasets	Space	S-Baseline	Baseline	HPA	RT	HOT-F	HOT-L	HOT-ME
BC	$\mathbb{L}^{10}$	0.7716 ± 0.0088	0.5254 ± 0.0091	$\mathbf{0.7410 \pm 0.0354}$	0.6001 ± 0.0658	0.5838 ± 0.0121	0.4594 ± 0.0428	0.5600 ± 0.0147
LC	$\mathbb{L}^{20}$	0.9137 ± 0.0532	0.5521 ± 0.0991	$\mathbf{0.8316 \pm 0.0904}$	0.5318 ± 0.0370	0.5400 ± 0.0396	0.5150 ± 0.0654	0.4901 ± 0.0516
P1 BT	$\mathbb{L}^{20}$	0.9723 ± 0.0058	0.6646 ± 0.1556	$\mathbf{0.9401 \pm 0.0068}$	0.9020 ± 0.0086	0.8603 ± 0.0142	0.8855 ± 0.0057	0.8919 ± 0.0089
P1 AT	$\mathbb{L}^{20}$	0.9590 ± 0.0150	0.7656 ± 0.1564	$\mathbf{0.9329 \pm 0.0011}$	0.8270 ± 0.0880	0.8469 ± 0.0873	0.8914 ± 0.0175	0.8848 ± 0.0306
P2 BT	$\mathbb{L}^{20}$	0.9686 ± 0.0114	0.6971 ± 0.1335	$\mathbf{0.9329 \pm 0.0186}$	0.8830 ± 0.0142	0.8762 ± 0.0215	0.8566 ± 0.1720	0.8810 ± 0.0040
P2 AT	$\mathbb{L}^{20}$	0.9045 ± 0.0070	0.5688 ± 0.0688	$\mathbf{0.8453 \pm 0.0798}$	0.7190 ± 0.0439	0.7012 ± 0.0912	0.7136 ± 0.0012	0.7291 ± 0.0113

Table 2: The mean and std of MMD computed over ten random subsets of size 300 for BC, 15 for LC, and 1000 for CyTOF from the two datasets.

Datasets	Space	Baseline	HPA	RT	HOT-F	HOT-L	HOT-ME
BC	$\mathbb{L}^{10}$	0.2089 ± 0.0027	0.0013 ± 0.0004	0.0072 ± 0.0011	0.0009 ± 0.0001	0.0019 ± 0.0004	$\mathbf{0.0007 \pm 0.0002}$
ST&UM	$\mathbb{L}^{20}$	0.1072 ± 0.0051	0.0162 ± 0.0048	0.0250 ± 0.0049	0.0023 ± 0.0002	0.0014 ± 0.0007	$\mathbf{0.0011 \pm 0.0006}$
ST&D-F	$\mathbb{L}^{20}$	0.3213 ± 0.0152	0.0122 ± 0.0042	0.0150 ± 0.0106	0.0019 ± 0.0002	$\mathbf{0.0010 \pm 0.0004}$	0.0015 ± 0.0003
UM&D-M	$\mathbb{L}^{20}$	0.0790 ± 0.0071	0.0168 ± 0.0090	0.0168 ± 0.0032	0.0017 ± 0.0003	$\mathbf{0.0012 \pm 0.0006}$	0.0029 ± 0.0005
P1 BT	$\mathbb{L}^{20}$	0.0638 ± 0.0024	0.0012 ± 0.0002	0.0020 ± 0.0002	0.0004 ± 0.0002	$\mathbf{0.0002 \pm 0.0001}$	0.0009 ± 0.0002
P1 AT	$\mathbb{L}^{20}$	0.0598 ± 0.0014	0.0006 ± 0.0001	0.0015 ± 0.0001	0.0003 ± 0.0001	$\mathbf{0.0001 \pm 0.0001}$	0.0002 ± 0.0001
P2 BT	$\mathbb{L}^{20}$	0.0424 ± 0.0021	0.0012 ± 0.0001	0.0015 ± 0.0001	$\mathbf{0.0002 \pm 0.0001}$	0.0020 ± 0.0008	0.0009 ± 0.0003
P2 AT	$\mathbb{L}^{20}$	0.0758 ± 0.0053	0.0011 ± 0.0002	0.0013 ± 0.0002	$\mathbf{0.0002 \pm 0.0001}$	0.0015 ± 0.0008	0.0007 ± 0.0003

Fig. 3 shows a visualization of the embedding of the two breast cancer datasets before and after HPA. For visualization, we project the points in $\mathbb{L}^3$ to the 3D Poincaré ball. Before the alignment, the dominant factor separating between the patients' samples (points) is the batch. In contrast, after the alignment, the batch effect is substantially suppressed (visually) and the factors separating the points are dominated by the cancer subtype.

We evaluate the quality of the alignment in two aspects using objective measures: (i) k-NN classification, with leave-one-batch-out cross-validation, is utilized for assessing the alignment of the intrinsic structure, and (ii) MMD [16] is used for assessing the distribution alignment quality. For the classification, we view the five subtypes of BC, the three subtypes of LC, and the presence of stimulated cells in CyTOF as the labels in the respective tasks. In addition to the results of the different alignment methods, we report the k-NN classification based only on a single batch (*S-Baseline*), which indicates the adequacy of the representation in hyperbolic space to the task at hand.

Table 1 depicts the k-NN classification obtained for the best k per method, and Table 2 shows the MMD. In each task, we set the dimension of the Lorentz model d to the dimension in which the best empirical single-task performance is obtained (S-Baseline). We note that similar results and trends are obtained for various dimensions. The k-NN performance for various k values is reported in Appendix C. Although the OT-based methods obtain the best matching between the distributions of the batches, HPA outperforms in all three tasks in terms of classification (see Table 1). We note that in the two gene expression tasks, where the data have multiple labels and we align multiple batches, the advantage of HPA compared to the other methods is particularly significant. In Appendix C, we demonstrate HPA's out-of-sample capabilities on the CyTOF data by learning the batch correction map between the different days from one patient and applying it to the data of the other patient.

5 Conclusion

We introduced HPA for unsupervised alignment of data in the Lorentz model. Based on Riemannian geometry, we presented new translation and scaling operations that align the first and second Riemannian moments as well as a wrapped rotation that aligns the orientation in the hyperboloid model. Our theoretical analysis provides further insight and highlights properties that may be useful for practitioners. We empirically showed in simulations that HPA is stable under noise and distortions. Experimental results involving unsupervised batch correction of multiple bioinformatics datasets with multiple labels is demonstrated. Beyond alignment and batch effect removal, our method can be viewed as a type of domain adaptation or a precursor of transfer learning that relies on purely geometric considerations, exploiting the geometric structure of data as well as the geometric properties of the space of the data. In addition, it can be utilized for multimodal data fusion and geometric registration of shapes with hierarchical structure.

References

[1] D. Alvarez-Melis, Y. Mroueh, and T. Jaakkola. Unsupervised hierarchy matching with optimal transport over hyperbolic spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1606–1617. PMLR, 2020.

[2] Y. Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

[3] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[4] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[5] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[6] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[7] J. Euzenat and P. Shvaiko. Ontology matching: State of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.

[8] L. Falorsi, P. de Haan, T. R. Davidson, and P. Forré. Reparameterizing distributions on lie groups. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3244–3253. PMLR, 2019.

[9] R. Flamary and N. Courty. Pot python optimal transport library. *GitHub: https://github.com/rflamary/POT*, page 144, 2017.

[10] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

[11] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018.

[12] O. Ganea, G. Becigneul, and T. Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, page 5350–5360, 2018.

[13] I. Gohberg, P. Lancaster, and L. Rodman. Matrices and indefinite scalar products. 1983.

[14] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[15] J. C. Gower, G. B. Dijksterhuis, et al. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[17] D. N. Hayes, S. Monti, G. Parmigiani, C. B. Gilks, K. Naoki, A. Bhattacharjee, M. A. Socinski, C. Perou, and M. Meyerson. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *Journal of Clinical Oncology*, 24(31):5079–5090, 2006.

[18] A. Hoyos-Idrobo. Aligning hyperbolic representations: an optimal transport-based approach. *arXiv preprint arXiv:2012.01089*, 2020.

[19] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[20] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

[21] A. Klimovskaia, D. Lopez-Paz, L. Bottou, and M. Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):1–9, 2020.

[22] D. Kohli, A. Cloninger, and G. Mishne. LDLE: Low distortion local eigenmaps. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*, 2021.

[23] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[24] A. Kumagai and T. Iwata. Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4106–4113, 2019.

[25] M. Law, R. Liao, J. Snell, and R. Zemel. Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2019.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] Q. Liu, M. Nickel, and D. Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 8228–8239, 2019.

[28] A. Lou, I. Katsman, Q. Jiang, S. Belongie, S.-N. Lim, and C. De Sa. Differentiating through the fréchet mean. In *International Conference on Machine Learning (ICML)*, pages 6393–6403. PMLR, 2020.

[29] B. Luo and E. Hancock. Feature matching with procrustes alignment and graph editing. 1999.

[30] H. Maron, N. Dym, I. Kezurer, S. Kovalsky, and Y. Lipman. Point registration via efficient convex relaxation. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[31] R. J. McCann. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.

[32] H. H. Milioli, R. Vimieiro, I. Tishchenko, C. Riveros, R. Berretta, and P. Moscato. Iteratively refining breast cancer intrinsic subtypes in the metabric dataset. *BioData mining*, 9(1):1–8, 2016.

[33] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[34] Y. Nagano, S. Yamaguchi, Y. Fujita, and M. Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning (ICML)*, pages 4693–4702. PMLR, 2019.

[35] I. Naim, Y. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea. Unsupervised alignment of natural language instructions with video segments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[36] C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[37] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NIPS)*, page 6341–6350, 2017.

[38] M. Nickel and D. Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning (ICML)*, pages 3779–3788. PMLR, 2018.

[39] V. Nygaard, E. A. Rødland, and E. Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[40] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[41] J. G. Ratcliffe, S. Axler, and K. Ribet. *Foundations of hyperbolic manifolds*, volume 149. Springer, 1994.

[42] P. L. C. Rodrigues, C. Jutten, and M. Congedo. Riemannian procrustes analysis: Transfer learning for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2018.

[43] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

[44] D. Spohr, L. Hollink, and P. Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In *International Semantic Web Conference*, pages 665–680. Springer, 2011.

[45] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[46] P. Tabaghi and I. Dokmanic. On procrustes analysis in hyperbolic space. *arXiv preprint arXiv:2102.03723*, 2021.

[47] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *In International Conference on Machine Learning (ICML)*, pages 1120–1127, 2008.

[48] O. Yair, M. Ben-Chen, and R. Talmon. Parallel transport on the cone manifold of spd matrices for domain adaptation. *IEEE Transactions on Signal Processing*, 67(7):1797–1811, 2019.

[49] O. Yair, F. Dietrich, R. Talmon, and I. G. Kevrekidis. Domain adaptation with optimal transport on the manifold of spd matrices. *arXiv e-prints*, pages arXiv–1906, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   [Yes] Our claims in the Abstract and Section 1 are supported in the paper.
   (b) Did you describe the limitations of your work?
   [Yes] Our alignment method is analyzed in Section 3, showing both its pros and cons.
   (c) Did you discuss any potential negative societal impacts of your work?
   [N/A] We believe that our work does not involve any potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   [Yes] The authors have read the ethics review guidelines and ensured that this paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   [Yes] All the assumptions of the claims made in Section 3 are stated.
   (b) Did you include complete proofs of all theoretical results?
   [Yes] All the proofs appear in Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] Our implementation is included in the supplemental material. We also attach the code generating the synthetic examples. The URLs to the publicly available datasets appear in Appendix B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] All the implementation details are stated in Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   [Yes] The results in Section 4 and Appendix C are reported as mean values with standard deviation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   [Yes] It is reported in Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   [Yes] The used modules are cited in Section 4 and Appendix B.
   (b) Did you mention the license of the assets?
   [Yes] They are stated in Section 3, Section 4 and Appendix B.
   (c) Did you include any new assets either in the supplemental material or as a URL?
   [Yes] Our code is attached as part of the supplementary material. It will be publicly available on Github upon the conference decision.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   [Yes] The used bioinofrmatic datasets are publicly available. It is stated in Section 4.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
   [Yes] The used data are completely anonymized and do not include personally identifiable information. It is stated in Appendix B.