arXiv:2301.02750v2 [stat.ML] 9 Jul 2024

空间形式中的主成分分析

Puoya Tabaghi、Michael Khanzadeh、Yusu Wang 和 Siavash Mirarab

摘要——**主成分分析（PCA）是现代数据科学的主力工具。虽然 PCA 假设数据符合欧几里得几何，但对于某些特定类型的数据，例如层次结构数据和循环数据结构，其他空间更为合适。我们研究空间形式中的主成分分析，即具有恒定曲率的空间。在黎曼流形上的一个点处，我们可以根据一组切向量定义一个黎曼仿射子空间。对于空间形式中的给定点，寻找最优低维仿射子空间相当于进行降维。我们的空间形式主成分分析（SFPCA）旨在寻找能够以最小投影代价最佳表示一组流形值点的仿射子空间。我们提出了满足以下两项性质的适当代价函数：（1）其最优仿射子空间是某个特征方程的解；（2）不同维度的最优仿射子空间构成嵌套集合。这些性质改进了现有方法；现有方法大多是收敛速度慢、理论保证较弱的迭代算法。我们在球形空间和双曲空间的真实数据与模拟数据上评估了所提出的 SFPCA。结果表明，在模拟数据中估计真实子空间时，该方法在收敛速度或准确度方面（通常两者兼具）优于其他方法。**

索引词—**主成分分析、黎曼流形、双曲空间和球面空间。**

I. 引言

给定 IVEN 一组多变量点，主成分分析（PCA）会求出一组正交基向量，使数据的不同分量在新坐标系中彼此不相关，并使主要基所承载的点的投影方差最大。PCA 与因子分析 [49], Karhunen-Lo´eve 展开以及奇异值分解 [55] 有关，其历史可追溯至18世纪 [45]。现代 PCA 的形式化表述源自 Hotelling 的工作 [19]。凭借其可解释性和灵活性，PCA 一直是数据科学应用中不可或缺的工具 [23]。文献中对 PCA 的表述已进行了大量研究。Tipping 和 Bishop [50]在概率框架下建立了因子分析与 PCA 之间的联系。此后又提出了其他扩展 [52], , 例如高斯过程 [29] 以及合理的 [40],贝叶斯 [2], 稀疏 [3], [17], [24], [59],和稳健主成分分析 [56]。

主成分分析的主要特点是线性，以及不同维度子空间的嵌套最优性。主成分分析使用线性变换来提取特征。因此，将主成分分析应用于非欧几里得数据会忽略其几何结构，产生可能不属于原空间的点，并破坏依赖这种几何结构的下游应用 [12], [14], [22], [47]。

我们聚焦于空间形式：维度为 $d \geq 2$ 且曲率恒定的完备、单连通黎曼流形——包括球形、欧几里得或双曲空间（曲率分别为正、零和负）[31]。空间形式能够表示多种形式的数据，因此在机器学习领域受到广泛关注。双曲空间适用于层次结构 [7], [8], [44], [47],[48],生物数据 [26], [58],和系统发育树 [22]。球形空间则应用于文本嵌入 [35],纵向数据 [10], 以及图中的环结构 [16]。

为了解决欧几里得主成分分析在非欧几里得数据上的不足，许多作者提出了黎曼主成分分析方法 [1], [13], [14], [20], [30], [38], [39], [42], [43], [51]。黎曼流形通常缺乏向量空间结构 [46], , 这给定义主成分带来了挑战。一种常见的流形值数据降维方法依赖于切空间。Fletcher等人 [13]提出了一种用于量化黎曼仿射子空间质量的代价函数，但采用启发式方法——主测地线分析（PGA）——对其进行优化：(1) 基点（内在均值）是不动点问题的解；(2) 在切空间中进行欧几里得主成分分析，以估计低秩切向量。即使是理论基础更完善的方法，也难以直接得出分析大规模数据所需的易处理解 [43],, 球形和双曲主成分分析就是如此 [5], [10], [32]。

尽管近期取得了进展，空间形式中的主成分分析仍未得到充分探索。一般而言，基于代价函数的黎曼主成分分析（例如球形和双曲主成分分析）依赖于通过最小化一个非凸函数来寻找最优的黎曼仿射子空间。代价函数、代理目标或方法通常受 ℓ_2^2 cost 启发，却没有明确的理论依据 [5],[10], [13], [14], [20], [32]。这些算法依靠迭代方法来估计黎曼仿射子空间，例如梯度下降、不动点迭代和近端交替最小化；它们收敛速度较慢，且需要进行参数调优。此外，除非通过贪心（次优）的方式进行代价最小化，即基于先前估计的低维子空间构建高维子空间，否则无法保证估计出的黎曼仿射子空间在包含关系下构成全序（即高维最优子空间包含低维子空间）。值得注意的是，Chakraborty 等人提出了一种针对空间形式的贪心主测地线分析，每次估计一条主测地线 [4]。他们推导出一个解析公式，用于将点投影到参数化测地线上，从而简化了主测地线分析中的投影步骤。然而，在每次迭代中估计主测地线时，我们仍须求解一个非凸优化问题（且没有理论保证）。

我们针对空间形式中的主成分分析局限性，提出一种具有闭式解、理论上最优且计算高效的方法，可一次性求出所有主测地线。我们首先从微分几何视角探讨欧几里得主成分分析（第二节），随后对黎曼

P. Tabaghi 和 Y. Wang 任职于加州大学圣迭戈分校 Halıcıo˘glu 数据科学研究所（电子邮件：ptabaghi@ucsd.edu；yusuwang@ucsd.edu）。M. Khanzadeh 曾任职于加州大学圣迭戈分校计算机科学与工程系。S. Mirarab 任职于加州大学圣迭戈分校电气与计算机工程系（电子邮件：smirarabbaygi@ucsd.edu）。

主成分分析（第三节）进行一般性描述。从这一视角来看，合适的主成分分析代价函数必须（1）自然地为流形值点定义质心；（2）给出理论上最优且按包含关系构成全序的仿射子空间。我们为球形（第四节）和双曲（第五节）主成分分析问题引入合适的代价函数。最小化每个代价函数都会得到一个特征方程，而且可以高效求解。对于双曲主成分分析，最优仿射子空间通过求解洛伦兹空间中的特征方程得到，而该空间配备了不定内积。这些结果为我们提供了高效的算法，用于求出双曲主成分。证明见附录。

A. 预备知识与记号

设 $(\mathcal{M}, g)$ 为一个黎曼流形。切空间$T_p\mathcal{M}$ 是 $p \in \mathcal{M}$ 处所有切向量的集合。黎曼度量 g_p： $T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ 由一个正定内积给出，并随 p 光滑变化。我们用 g_p 来定义子空间、范数和角度等概念，这与内积空间中的定义类似。对于任意子空间 $H \subseteq T_p\mathcal{M}$，我们将其正交补 定义如下：

$$H^{\perp} = \{h' \in T_p\mathcal{M} : g_p(h, h') = 0,\ \forall h \in H\} \subseteq T_p\mathcal{M}. \quad (1)$$

$v \in T_p\mathcal{M}$ 的范数为 $\|v\| \overset{\text{def}}{=} \sqrt{g_p(v,v)}$。我们将光滑曲线 $\gamma : [0,1] \to \mathcal{M}$ 的长度记为 $L[\gamma] =: \int_0^1 \|\gamma'(t)\|\, dt$。测地线 $\gamma_{p_1 p_2}$ 是连接 p_1 与 $p_2 \in \mathcal{M}$ 的最短路径，即 $\gamma_{p_1,p_2} = \arg\min_{\gamma} L[\gamma] : \gamma(0) = p_1, \gamma(1) = p_2$。将参数 t 解释为时间，若测地线 $\gamma(t)$ 从 $\gamma(0) = p \in \mathcal{M}$ 出发，初速度为$\gamma'(0) = v \in T_p\mathcal{M}$，则指数映射 $\exp_p(v)$ 给出其在 $t = 1$ 时的位置。对于 p 和 $x \in \mathcal{M}$，对数映射 $\log_p(x)$ 给出沿测地线从 p 到 x、以恒定速度在一个时间步内移动所需的初速度。若黎曼流形 $\mathcal{M}$ 上的每个点 $p \in \mathcal{M}$处的指数映射和对数映射都是良定义的算子 [15]，则称该流形测地完备。黎曼流形 $(\mathcal{M}, g)$ 的子流形 $\mathcal{M}'$ 若满足：在 $\mathcal{M}'$ 上、相对于其诱导度量 g 的任意测地线也是 $(\mathcal{M}, g)$ 上的测地线，则称其为测地子流形。对于 $N \in \mathbb{N}$，令$[N] \overset{\text{def}}{=} \{1, \ldots, N\}$ 和 $[N]_0 \overset{\text{def}}{=} [N] \cup \{0\}$。变量 x_1是向量 $x = (x_0, \ldots, x_{D-1})^{\top} \in \mathbb{R}^D$ 的一个元素。它也可以是带索引的向量，例如 $x_1, \ldots, x_N \in \mathbb{R}^D$。这一区别将在上下文中说明。我们用 $\mathbb{E}_N[\cdot]$ 表示其输入中索引属于 $[N]$ 的经验均值。

II. 主成分分析——再探

与 Pearson 的观点类似， [37], 主成分分析旨在找到最优的低维仿射空间来表示数据。设 $p \in \mathbb{R}^D$，且 $H \in \mathbb{R}^{D \times K}$ 的列张成一个子空间。对于仿射子空间 $p + H$，主成分分析采用以下代价函数：

$$\operatorname{cost}(p + H | \mathcal{X}) = \mathbb{E}_N\Big[f \circ d\big(x_n, \mathcal{P}_{p+H}(x_n)\big)\Big],$$

其中 $\mathcal{X} = \{x_n \in \mathbb{R}^D : n \in [N]\}$，$\mathcal{P}_{p+H}(x_n) = \arg\min_{x \in p+H} d(x, x_n)$，$d(\cdot,\cdot)$ 计算距离，而失真函数 $f(x) = x^2$。这种形式依赖于：(1) 仿射子空间 $p + H$，(2) 投影算子 $\mathcal{P}_{p+H}$，以及 (3) 失真函数 f。为了将仿射子空间推广到黎曼流形，考虑参数曲线：

$$\gamma_{p,x}(t) = (1-t)p + tx, \quad \text{and} \quad \gamma_{h'}(t) = p + th', \quad (2)$$

其中 $h' \in H^{\perp}$是子空间H的正交补；见图 1（a， b）。我们将仿射子空间重新表述如下：

$$\begin{aligned} p + H &= \{x \in \mathbb{R}^D : \langle x - p, h' \rangle = 0, \text{ for all } h' \in H^{\perp}\} \\ &= \{x \in \mathbb{R}^D : \langle \gamma'_{p,x}(0), \gamma'_{h'}(0) \rangle = 0, \text{ for all } h' \in H^{\perp}\}, \end{aligned}$$

其中 $\langle \cdot,\ \cdot \rangle$ 表示点积，且 $\gamma'(t_0)$ 定义为 $= \frac{d}{dt}\gamma(t)|_{t=t_0}$。

定义 1。仿射子空间是点的集合，例如 x，其中存在 $p \in \mathbb{R}^D$，使得直线 $\gamma_{x,\ p}$（即 $\gamma'_{p,\ x}(0)$）的（切向量）与 p处的一组切向量正交。

定义 1 需要 $\dim(H^{\perp})$ 个参数来描述 $p + H$。例如，在 $\mathbb{R}^3$ 中，我们需要两个正交归一向量来表示一个一维仿射子空间；见图 1（a）。我们采用定义 1，因为它通过直线 和切向量来描述仿射子空间，而非使用全局坐标系。

III. 黎曼主成分分析

接下来，我们将介绍黎曼仿射子空间，并提出一个黎曼主成分分析的通用框架。

A. 黎曼仿射子空间

利用黎曼流形 $\mathcal{M}$ [13] 的切子空间，可以将仿射子空间的概念推广到黎曼流形。黎曼仿射子空间是 $T_p\mathcal{M}$ 中某个子空间在指数映射下的像，即

$$\mathcal{M}_H = \exp_p(H) \overset{\text{def}}{=} \{\exp_p(h) : h \in H\}, \quad (3)$$

其中 H 是 $T_p\mathcal{M}$ 的子空间，且 $p \in \mathcal{M}$。公式 3 等价于以下定义。

定义2。设 $(\mathcal{M}, g)$ 是一个测地完备的黎曼流形， $p \in \mathcal{M}$，且 $H \subseteq T_p\mathcal{M}$ 是子空间。令$\mathcal{M}_H = \{x \in \mathcal{M} : g_p(\log_p(x)\ h') = 0,\ \forall h' \in H^{\perp}\}$ 其中 $H^{\perp}$ 是 H 的正交补；参见公式 (1)。

集合 $\mathcal{M}_H$ 是由以下点构成的集合：这些点位于从 p 出发的测地线上，且其初速度与$H^{\perp} \subseteq T_p\mathcal{M}$中的向量正交，或者构成一个子空间 $H \subseteq T_p\mathcal{M}$；参见定义 1。

例 1。 当 H 是一维子空间时，$\mathcal{M}_H$ 包含经过 p 且初速度属于 H的测地线，即 $\mathcal{M}_H = \{\gamma(t)$：测地线 γ，其中 $\gamma(0) = p, \gamma'(0) \in H, t \in \mathbb{R}\}$。因此，根据方程(3)，测地线是一维黎曼仿射子空间。

例 2。 欧几里得指数映射为$\exp_p(h) = p + h$；参见表 *I*。因此，方程(3) 可恢复仿射子空间，即$\exp_p(H) = \{p + h : h \in H\}$ 定义$= p + H$。

回顾一下，非空集合 V 是欧几里得仿射子空间，当且仅当存在 $p \in V$，使得 $\alpha(v_1 - p) + p \in V$，并且对于所有 $\alpha \in \mathbb{R}$ 和 v_1，$v_2 \in V$，均有 $(v_1 - p) + (v_2 - p) + p \in V$。黎曼仿射子空间也有类似的定义。

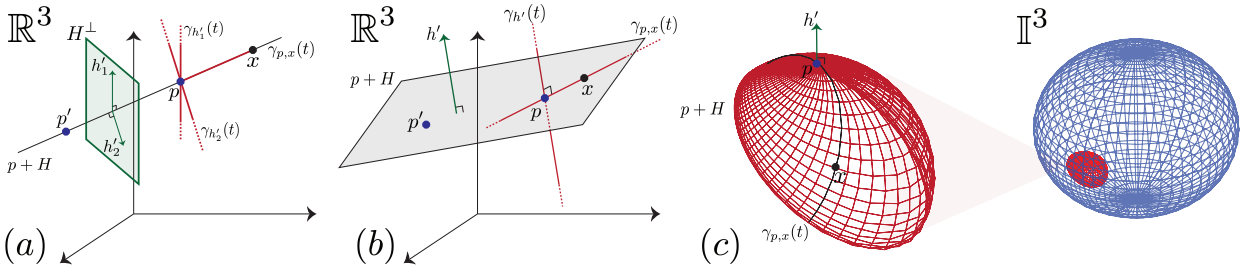


图 1。 （a，b）一维（a）和二维 （b）仿射子空间，位于 $\mathbb{R}^3$ 中。我们展示了经过点 p 而非原点的子空间（$H^\perp$）。我们可以使用其他基点来定义相同的黎曼仿射子空间，例如 基点 p'。（c）双曲空间（庞加莱模型）中的二维仿射子空间，其中 $h' \in T_p\mathbb{I}^3 = \mathbb{R}^3$.

定义 3。设$(\mathcal{M}, g)$ 是一个测地完备的黎曼流形。那么，非空集合 $V \subseteq \mathcal{M}$ 是仿射集，当且仅当存在 $p \in V$，使得 $\exp_p\big(\alpha \log_p(v_1)\big) \in V$，且对于所有 $\alpha \in \mathbb{R}$ 和 $v_1, v_2 \in V$，均有$\exp_p\big(\log_p(v_1) + \log_p(v_2)\big) \in V$。

B. 黎曼主成分分析的适当代价函数

与欧几里得主成分分析类似，黎曼主成分分析旨在寻找一个（黎曼）仿射子空间，使点与其投影之间的平均失真最小。

定义 *4.* 设 $(\mathcal{M}, g)$ 是一个测地完备的黎曼流形，配备距离函数 d,$p \in \mathcal{M}$，以及子空间 $H \subseteq T_p\mathcal{M}$。将 $x \in \mathcal{M}$ 测地投影到 $\mathcal{M}_H$ 上是指 $\mathcal{P}_H(x) \in \arg\min_{y\in\mathcal{M}_H} d(x, y)$。若 $\arg\min_{y\in\mathcal{M}_H} d(x, y) \neq \emptyset$，则对于任意测地投影 $\mathcal{P}_H(x)$，$\min_{y\in\mathcal{M}_H} d(x, y) = \|\log_{\mathcal{P}_H(x)}(x)\|$ 。

注 1. 将一个流形值点投影到黎曼仿射子空间并非易事，通常需要在子流形上求解一个非凸优化问题，即求解 $\arg\min_{y\in\mathcal{M}_H} d(x, y)$。定义 *4* 指出，如果解存在（但这并不总能保证），则投影距离必须等于 $\|\log_{\mathcal{P}_H(x)}(x)\|$。这还需要计算对数映射，而并非所有黎曼流形上都能定义该映射。

在欧几里得主成分分析中，最小化 ℓ_2^2 代价函数等价于最大化投影后数据的方差。为了避免围绕方差和质心概念产生混淆，我们用投影距离来形式化代价函数（以 $\mathcal{M}_H$ 为参数），即:

$$\mathrm{cost}(\mathcal{M}_H|\mathcal{X}) = \mathbb{E}_N\big[f(\|\log_{\mathcal{P}_H(x_n)}(x_n)\|)\big], \qquad (4)$$

其中，$\mathcal{X} = \{x_n \in \mathcal{M} : n \in [N]\}$，且 $f: \mathbb{R}^+ \to \mathbb{R}$ 是一个单调递增失真函数。投影点 $\mathcal{P}_H(x)$ 可能并不唯一。其极小化点（若存在）是表示流形值点的最佳仿射子空间。

定义 5。 黎曼主成分分析旨在最小化方程 (4) 中的代价函数——其中失真函数 f 取特定形式。

对 f 的选择。 最优欧几里得仿射子空间的闭式解源于令 $f(x) = x^2$。这是一个具有以下性质的适当代价函数:

1) 一致质心。 最优的 0 维仿射子空间（一个点）是数据点的质心，即 $p^* = \arg\min_{y\in\mathbb{R}^D} \mathbb{E}_N[f \circ d(x_n, y)] = \mathbb{E}_N[x_n]$。

2) 嵌套最优性。 最优仿射子空间构成一个嵌套集合，即 $p^* \subseteq \big(p + H_1\big)^* \subseteq \cdots$其中 $(p + H_d)^*$ 是最优的 d 维仿射子空间。

定义 6。 对于黎曼主成分分析，若 $\mathrm{cost}(\mathcal{M}_H|\mathcal{X})$ 的极小化点满足一致质心和嵌套最优性条件，我们就称其为适当的代价函数。

对于一般的黎曼流形而言，推导对数算子并非易事，例如秩亏半正定矩阵流形 [27]。针对常曲率黎曼流形，我们提出了失真函数；与现有方法不同，这些函数能够得到具有闭式最优解的适当代价函数。

IV. 球面主成分分析

考虑曲率为$C > 0$的球面流形（$\mathbb{S}^D$，$g^{\mathbb{S}}$），其中 $\mathbb{S}^D = \{x \in \mathbb{R}^{D+1} : \langle x, x\rangle = C^{-1}\}$是半径为 $C^{-\frac{1}{2}}$的球面，并且 $g_p^{\mathbb{S}}(u, v) = \langle u, v\rangle$ 计算 $u, v \in T_p\mathbb{S}^D = \{x \in \mathbb{R}^{D+1}$ 的点积: $\langle x, p\rangle = 0\}$ 定义= $p^\perp$。

A. 球面仿射子空间与投影算子

设 $p \in \mathbb{S}^D$ 以及子空间 $H \subseteq T_p\mathbb{S}^D = p^\perp$。根据定义2和表 I，球面仿射子空间为:

$$\mathbb{S}_H^D = \{x \in \mathbb{S}^D : \langle x, h'\rangle = 0,\ \forall h' \in H^\perp\} = \mathbb{S}^D \cap (p \oplus H),$$

其中，$\oplus$ 是直和算子，即 $p \oplus H = \{\alpha p + h : h \in H,\ \alpha \in \mathbb{R}\}$，而 $H^\perp$ 是 H的正交补；参见公式（1）。这与 Pennec 对指数重心球面子空间 [38]的度量完备化的概念相符。

命题1。 $\mathbb{S}_H^D$ 是一个测地子流形。

存在正交切向量 h_1'、.....、$h_{K'}'$，它们构成 $H^\perp$的完备基，即 $\langle h_i', h_j'\rangle = C\delta_{i,j}$，其中当 $i \neq j$ 时 $\delta_{i,j} = 0$，当 $i = j$时 $\delta_{i,j} = 1$ 。利用这些基向量，我们推导出投影距离的简单表达式。

命题 1。 对于任意 $\mathbb{S}_H^D$ 和 $x \in \mathbb{S}^D$，有

$$\min_{y\in\mathbb{S}_H^D} d(x, y) = C^{-\frac{1}{2}} \arccos\Bigg(\sqrt{1 - \sum_{k\in[K']} \langle x, h_k'\rangle^2}\Bigg),$$

其中，$\{h_{k'}'\}_{k'\in[K']}$ 是 $H^\perp$的完备正交基向量。$\mathbb{S}_H^D$ 和 $\mathbb{S}^D$ 的曲率 $C > 0$均固定。

表 I 欧几里得空间、球形空间和双曲空间中相关算子的汇总。

$\mathcal{M}$	$T_p\mathcal{M}$	$g_p(u,v)$	$\log_p(x):\ \theta = d(x,p)$	$\exp_p(v)$	$d(x,\ p)$
$\mathbb{R}^D$	$\mathbb{R}^D$	$\langle u, v\rangle$	$x - p$	$p + v$	$\|x-p\|_2$
$\mathbb{S}^D$	$p^\perp$	$\langle u, v\rangle$	$\frac{\theta}{\sin(\theta)}(x - \cos(\theta)p)$	$\cos(\sqrt{C}\|v\|)p + \sin(\sqrt{C}\|v\|)\ \frac{v}{\sqrt{C}\|v\|}$	$\frac{1}{\sqrt{C}}\arccos(C\langle x,\ p\rangle)$
$\mathbb{H}^D$	$p^\perp$	$[u, v]$	$\frac{\theta}{\sinh(\theta)}(x - \cosh(\theta)p)$	双曲余弦$(\sqrt{-C}\|v\|)p$ +双曲正弦$(\sqrt{-C}\|v\|)\ \frac{v}{\sqrt{-C}\|v\|}$	$\frac{1}{\sqrt{-C}}\operatorname{acosh}(C[x,p])$

对于距离 $\mathbb{S}^D_H$为 $C^{-\frac{1}{2}}\frac{\pi}{2}$的点，其在仿射子空间上的投影并不唯一。尽管如此，命题 1 给出了一个闭式表达式，可用 $H^\perp$的基向量表示投影距离。该距离会随着x 在 $H^\perp$上的残差长度单调递增，即 $\sum_{k\in[K']}\langle x,\ h'_k\rangle^2$。由于 $\dim(H) \ll D$ 较为常见，改用 H 的基有助于我们用更少的参数表示投影距离。

命题2。 对于任意 $\mathbb{S}^D_H$ 和 $x \in \mathbb{S}^D$，我们有

$$\min_{y\in\mathbb{S}^D_H} d(x,y) = C^{-\frac{1}{2}}\operatorname{acos}\Big(\sqrt{C^2\langle x,p\rangle^2 + \sum_{k\in[K]}\langle x,h_k\rangle^2}\Big),$$

其中，$\{h_k\}_{k\in[K]}$ 是 H的完备正交基向量。

接下来，我们推导出 $\mathbb{S}^D_H$ 与 $\mathbb{S}^{\dim(H)}$ 之间的一个等距映射，其中二者的曲率均为固定值 $C > 0$。

定理1. 等距映射 $\mathcal{Q}$: $\mathbb{S}^D_H \to \mathbb{S}^K$ 及其逆映射为

$$\mathcal{Q}(x) = C^{-\frac{1}{2}}\begin{bmatrix}C\langle x,p\rangle\\ \langle x,h_1\rangle\\ \vdots\\ \langle x,h_K\rangle\end{bmatrix},\ \mathcal{Q}^{-1}(y) = C^{-\frac{1}{2}}(y_0 Cp + \sum_{k=1}^{K} y_k h_k),$$

其中，$\{h_k\}_{k\in[K]}$ 是 H的一组完备正交基向量。

推论1. $\mathbb{S}^D_H$ 的维度为 $\dim(H)$。

最后，我们可以基于切片酉矩阵 $\{G \in\mathbb{R}^{(D+1)\times(K+1)}: G^\top G = I_{K+1}\}$，从另一角度看待球面仿射子空间。

命题2. 对于任意 $\mathbb{S}^D_H$，都存在一个切片酉算子 G: $\mathbb{S}^{\dim(H)} \to \mathbb{S}^D_H$，反之亦然。

B. 最小失真球面子空间

为了定义主成分，我们需要对失真函数 f 作出具体选择；见公式（4）。在介绍我们的选择之前，让我们先讨论已有研究中的代价函数。

1）已有研究综述：Dai 和 M¨uller 针对黎曼流形 $\mathcal{M}$ [10] 上的光滑函数型数据 $\mathcal{X}$，考虑了采用失真函数 $f(x) = x^2$ 的内在主成分分析，即

$$\operatorname{cost}_{\text{Dai.}}(\mathbb{S}^D_H|\mathcal{X}) = \mathbb{E}_N\big[f\big(\min_{y\in\mathbb{S}^D_H} d(x_n,y)\big)\big]. \tag{5}$$

他们的算法——黎曼函数主成分分析（RFPCA）——首先求解基点，即最优零维仿射子空间，也就是 Fr´echet 均值 $p^* = \arg\min_{p\in\mathbb{S}^D}\mathbb{E}_N\big[d(x_n,p)^2\big]$。随后，他们使用对数映射将每个点投影到 $T_{p^*}\mathcal{M}$。接着，他们对所得切向量进行主成分分析，以获得 K 维切子空间。最后，他们使用指数映射将投影后的切向量映射回 $\mathcal{M}$ （球面空间）。尽管方法简单，但存在四个缺点。(1) 球面数据的 Fr´echet 均值没有闭式表达式。(2) 目前尚不清楚估计 Fr´echet 均值的计算复杂度的理论分析；而且其计算涉及一个 argmin 操作，通常难以求导 [33]。(3) 即使 Fr´echet 均值估计得非常准确，也无法保证问题 (5) 的最优解就是 Fr´echet 均值。Huckemann 和 Ziezold [21] 证明，Fr´echet 均值可能不属于最优的一维仿射球面子空间。(4) 即使 Fr´echet 均值恰好是最优基点，在切空间中进行主成分分析也不是问题 (5) 的解。

Liu 等人提出了一个球面矩阵分解问题：

$$\min_{\substack{G\in\mathbb{R}^{(D+1)\times(K+1)}\\ \{y_n\in\mathbb{S}^K\}_{n\in[N]}}}\mathbb{E}_N\big[\|x_n - Gy_n\|_2^2\big]:\ G^\top G = I_{K+1}, \tag{6}$$

其中 $\mathcal{X} = \{x_n \in \mathbb{R}^{D+1}:\ n \in [N]\}$ 是测量集，$y_1,\ldots,y_N \in \mathbb{S}^K$ 是球面空间中的特征，且 $C = 1$ [32]。他们提出了一种近端算法来求解仿射子空间和特征。该形式化并不是球面主成分分析，因为测量值不属于球面空间。公式（6）中的目标旨在将欧几里得点最佳投影到低维球面仿射子空间，其依据是欧几里得距离平方 ——参见命题 2。然而，如果我们改变他们的形式化，并令输入空间为球面空间，就得到：

$$\begin{aligned}\operatorname{cost}_{\text{Liu}}(\mathbb{S}^D_H|\mathcal{X}) &= \mathbb{E}_N\big[-\cos\big(\min_{y_n\in\mathbb{S}^K} d(x_n,Gy_n)\big)\big]\\ &= \mathbb{E}_N\big[f\circ d(x_n,\mathcal{P}_H(x_n))\big],\end{aligned} \tag{7}$$

其中，H 是与 G 相对应的切子空间（见命题2），$\mathcal{P}_H$ 是球形投影算子。该形式化方法使用失真函数 $f(x) = -\cos(x)$。

Jung等人提出的嵌套球面 [25] 是拟合主嵌套球面的一种替代方法，用于迭代地降低数据的维度。该方法通过最小化以下代价函数，求得最优的（$D-1$）维子球面 $\mathcal{U}_{D-1}$，

$$\operatorname{cost}_{\text{Jung}}(\mathcal{U}_{D-1}|\mathcal{X}) = \mathbb{E}_N[(d(x_n,p) - r)^2],$$

其中 $\mathcal{U}_{D-1} = \{x \in \mathbb{S}^D:\ d(x,p) = r\}$ — 在 $p \in \mathbb{S}^D$ and $r \in \mathbb{R}^+$ 上。这是一个没有闭式解的约束非线性优化问题。估计出最优的 $\mathcal{U}_{D-1}$ 后，他们将每个点映射到低维球面空间 $\mathbb{S}^{D-1}$ — 并重复这一过程，直到达到目标维度。这些子球面不一定是大圆，因此这种分解不是测地的。

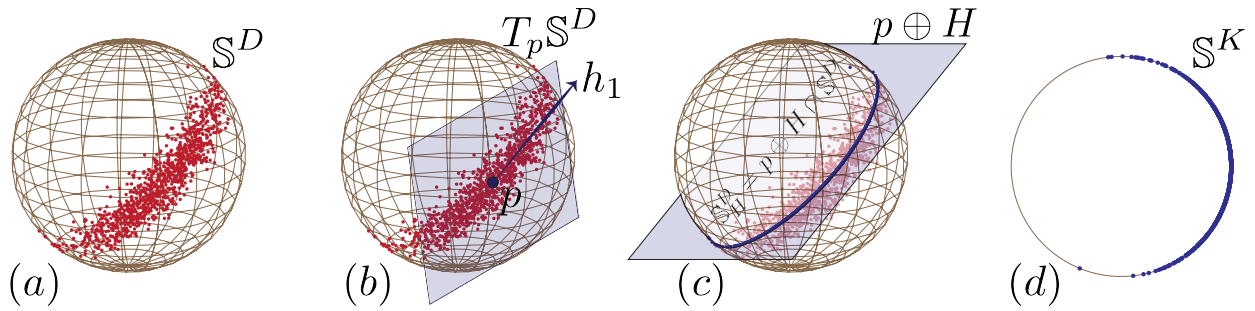


图 2。（a） $\mathbb{S}^D$中的一组数据点，其中 $D=2$。（b）基点 p和切子空间 $H=h_1\in T_p\mathbb{S}^D$的最佳估计——球面仿射子空间 $\mathbb{S}_H^D=(p\oplus H)\cap\mathbb{S}^D$。（$c$）点在 $\mathbb{S}_H^D$（$H=h_1$）上的投影。（d） $\mathbb{S}^K$中的低维特征，其中 $K=\dim(\mathbb{S}_H^D)=1$。

2）球面主成分分析的恰当代价函数：与 Liu等人 [32] 以及 Dai 和 M¨uller [10]，采用的失真 $f(x)=-\cos(x)$和 $f(x)=x^2$ 不同，我们选择 $f(x)=\sin^2(\sqrt{C}x)$。利用命题 1，我们得到：

$$\operatorname{cost}(\mathbb{S}_H^D|\mathcal{X})=\mathbb{E}_N\Big[\sum_{k\in[K']}\langle x_n,h'_k\rangle^2\Big], \tag{8}$$

即沿向量 $h'_1,\ldots,h'_{K'}\in T_p\mathbb{S}^D$的方向计算投影点的平均 ℓ_2^2 范数。表达式 **(8)** 导出了一个易于处理的约束优化问题。

主张 3. 设 $x_1,\ldots,x_N\in\mathbb{S}^D$。球面主成分分析方程(8) 旨在求出 $p\in\mathbb{S}^D$和正交的 $h'_1,\ldots,h'_{K'}\in T_p\mathbb{S}^D$，使 $\sum_{k\in[K']}{h'_k}^\top C_x h'_k$最小，其中 $C_x=\mathbb{E}_N[x_n x_n^\top]$。

主张 3 中问题的解是式 **(8)** 中代价函数的极小化点，它是一组正交向量 h'_1，……，$h'_{K'}\in p^\perp$；用量子物理的术语来说，这些向量捕获了 C_x可能达到的最低能量。

定理 2. 设 $x_1,\ldots,x_N\in\mathbb{S}^D$。则， p的最优解是 C_x的主特征向量，而 $h'_1,\ldots,h'_{K'}$（ $H^\perp$的基向量）则是对应于二阶矩矩阵 C_x最小的 K' 个特征值的特征向量。

推论 2. 最优子空间 H 由 C_x的 K个主特征向量张成，但不包括第一个。

主张 4. 式 (8)中的代价函数是适定的。

式 **(8)** 中的失真给出了球面数据点质心的闭式定义，即一个能最佳表示数据的零维仿射子空间。

定义 7. 对于点集 $\mathcal{X}$，球面均值 $\mu(\mathcal{X})$ 是满足 $\mathbb{E}_N[f\circ d\ x_n\ \mu\ \mathcal{X}\ ]=\min_{p\in\mathbb{S}^D}\mathbb{E}_N[f\circ(,())\min(,)$的任意点。该 $d\ x_n\ p\ ]$ 解是 C_x的一个缩放后的主特征向量。

将最优基点解释为定义 7 中的球面均值，表明我们的球面主成分分析具有一致质心和嵌套最优性：不同维度的最优球面仿射子空间在包含关系下形成一条链。然而，$\mu(\mathcal{X})$并不唯一，且只能确定点的主成分的**方向**；见图 2。

注 2. *PGA [13]*和 *RFPCA [10]* 涉及一项计算量很大的 *Fr´echet* 均值估计任务。这需要对非线性优化目标采用梯度下降或不动点迭代等迭代方法。已有研究探讨了 *Fr´echet* 均值计算的数值分析 *[33]*。另一方面，*SPCA [32]* 采用交替线性化最小化来估计最优子空间。相比之下，我们的方法（*SFPCA*）需要计算二阶矩矩阵 C_x，其复杂度为 $O(D^2N)$，并且涉及特征分解，其最坏情况复杂度为 $O(D^3)$。

V. 双曲主成分分析

我们先介绍洛伦兹空间。

定义 8. 洛伦兹$(D+1)$空间，记作 $\mathbb{R}^{1,D}$，是一个配备洛伦兹内积的向量空间

$$\forall x,y\in\mathbb{R}^{1,D}:[x,y]=x^\top J_D y,\ J_D=\begin{pmatrix}-1 & 0^\top\\ 0 & I_D\end{pmatrix},$$

其中，I_D 是<*style id='v2'*>单位矩阵。

具有曲率 $C<0$的 D维双曲流形（$\mathbb{H}^D$，$g^{\mathbb{H}}$），其中 $\mathbb{H}^D=\{x\in\mathbb{R}^{D+1}:[x,x]=C^{-1}$和 $x_0>0\}$，且指标 $g_p^{\mathbb{H}}(u,v)=[u,v]$适用于 $u,v\in T_p\mathbb{H}^D=\{x\in\mathbb{R}^{D+1}:[x,p]=0\}$ 定义$=p^\perp$；参见表 I。

A. 洛伦兹空间中的特征方程

与内积空间类似，我们在 $\mathbb{R}^{1,D}$中定义算子。

定义 9。 设 $A\in\mathbb{R}^{(D+1)\times(D+1)}$ 是 $\mathbb{R}^{1,D}$中的一个矩阵（算子）。当且仅当 $A^{[\top]}=J_D A^\top J_D$时，我们称 $A^{[\top]}$ 为 A 的 J_D-伴随。仅当 *(iff)* $A^{[-1]}J_D A=AJ_D A^{[-1]}=J_D$时，$A^{[-1]}$ 才是 A 的 J_D-逆映射。若可逆矩阵 A 满足 $A^\top J_D A=J_D$，则称其为 J_D-酉矩阵；更多详情请参见 *[18]*。

洛伦兹空间 $\mathbb{R}^{1,D}$ 配备了不定内积，即 $\exists x\in\mathbb{R}^{1,D}:[x,x]<0$。因此，需要根据其不定内积定义一种特征方程。为完整起见，我们提出如下复洛伦兹空间 $\mathbb{C}^{1,D}$中特征方程的定义。

定义10. 对于 $A\in\mathbb{C}^{(D+1)\times(D+1)}$, $v\in\mathbb{C}^{1,D}$ 是其 J_D-特征向量，λ 是对应的 J_D-特征值，若

$$AJ_D v=\operatorname{sgn}([v^*,v])\lambda v,\quad where\ \lambda\in\mathbb{C}, \tag{9}$$

且 v^* 是 v的复共轭。范数的符号，$\operatorname{sgn}([v^*,v])$，定义了正和负$J_D$-特征向量。

定义10与双曲特征方程 [41] 略有不同——它是（A，J）特征值分解的一种特殊情况。

我们更倾向于采用定义10，因为它将欧几里得空间中的熟悉结果推广到了洛伦兹空间。

命题3. 如果 $A = A^{[\top]}$，那么它的 J_D-特征值均为实数。

设 v 是 A 的一个 J_D-特征向量，其中 $|[v^*, v]| = 1$。那么，$[v^*, AJ_Dv] = \text{sgn}([v^*, v])\lambda[v^*, v] = \lambda$ 是 A 的 J_D-特征值；参见定义10。欧几里得特征方程与洛伦兹特征方程之间存在联系。也就是说，$AJ_D \in \mathbb{C}^{(D+1)\times(D+1)}$ 有特征向量 $v \in \mathbb{C}^{D+1}$，且 $[v^*, v] \neq 0$。那么，$\left(|[v^*, v]|^{-\frac{1}{2}} v, \text{sgn}([v^*, v])\lambda\right)$ 是 A 的一个 J_D-特征对。

主张 5. J_D 的特征向量与 A 的特征向量平行。

命题4表明，对于满秩矩阵，归一化因子是良定义的。

命题4. 若 A 是满秩的，则 $\{v \in \mathbb{R}^{D+1} : AJ_Dv = \lambda v \text{ 且 } [v^*, v] = 0\} = \emptyset$。

我们的算法利用了欧几里得特征方程与J_D-特征方程之间的联系。我们将可对角化性的概念加以扩展，以推导出最优仿射子空间；参见命题 7。

定义11. 当且仅当存在一个 J_D-可逆的 $V \in \mathbb{C}^{(D+1)\times(D+1)}$，使得 $AJ_DV = VJ_D\Lambda$，其中 Λ 是对角矩阵时，$A \in \mathbb{C}^{(D+1)\times(D+1)}$ 才是 J_D-可对角化的。

B. 双曲仿射子空间与投影算子

设 $p \in \mathbb{H}^D$ 和 $H^\perp$ 是$T_p\mathbb{H}^D = p^\perp$中的 K'维子空间。根据定义2和表 I，我们得到如下双曲仿射子空间的定义：

$$\mathbb{H}^D_H = \{x \in \mathbb{H}^D : [x, h'] = 0,\ \forall h' \in H^\perp\}, \tag{10}$$

其中，$H^\perp$ 是 H的正交补，即 $\mathbb{H}^D_H = \mathbb{H}^D \cap (p \oplus H)$。这也与指数重心双曲子空间 [38]的度量完备化相一致。

主张 6. 双曲子空间是一个测地子流形。

引理 1 表明，存在一组完备的正交切向量 $h'_1, \ldots, h'_{K'}$，其中$[h'_i, h'_j] = -C\delta_{i,j}$，且张成空间为 $H^\perp$。在命题 5 中，我们给出了关于 $H^\perp$的基、投影到 $\mathbb{H}^D_H$ 上的投影距离的闭式表达式。

命题 5. 对于任意 $\mathbb{H}^D_H$ 和 $x \in \mathbb{H}^D$，有

$$\min_{y \in \mathbb{H}^D_H} d(x, y) = |C|^{-\frac{1}{2}} \text{acosh}\Big(\sqrt{1 + \sum_{k \in [K']} [x, h'_k]^2}\Big),$$

其中 $\{h'_{k'}\}_{k' \in [K']}$ 是 $H^\perp$的完备正交基。

投影距离随 x 在 $H^\perp$上的残差范数单调增加，即 $\sum_{k \in [K']} [x, h'_k]^2$。命题 5 要求求出 $H^\perp$的正交基——通常是一个高维空间。我们可以使用 H 的基来计算投影距离。

命题 6. 对于任意 $\mathbb{H}^D_H$ 和 $x \in \mathbb{H}^D$，我们有

$$\min_{y \in \mathbb{H}^D_H} d(x, y) = |C|^{-\frac{1}{2}} \text{acosh}\Big(\sqrt{C^2[x, p]^2 - \sum_{k \in [K]} [x, h_k]^2}\Big),$$

其中 $\{h_k\}_{k \in [K]}$ 是 H的完备标准正交基。

我们将 $\mathbb{H}^D_H$ 中的点表示为基点与切向量的线性组合。给定这些 $K+1$ 向量，我们可以找到 $\mathbb{H}^D_H$中各点的低维表示，从而降低双曲数据点的维数。

定理 3。 等距映射 $\mathcal{Q}$： $\mathbb{H}^D_H \to \mathbb{H}^K$ 及其逆映射为

$$\mathcal{Q}(x) = \alpha \begin{bmatrix} C[x, p] \\ [x, h_1] \\ \vdots \\ [x, h_K] \end{bmatrix},\ \mathcal{Q}^{-1}(y) = \alpha(-y_0 Cp + \sum_{k=1}^{K} y_k h_k),$$

其中 $\alpha = |C|^{-\frac{1}{2}}$ 和 H 具有完备正交基向量 $\{h_k\}_{k \in [K]}$。$\mathbb{H}^D_H$ 和 $\mathbb{H}^K$ 的曲率均为 $C < 0$。

推论 3。 $\mathbb{H}^D_H$ 的仿射维数为 $\dim(H)$。

与球面情形类似，我们可以用切片 J_D-酉矩阵来刻画双曲仿射子空间——这为采用切片 J_D-酉矩阵上的约束优化方法求解双曲 PCA 问题铺平了道路。

主张 7. 对于任意 $\mathbb{H}^D_H$，存在一个切片 J_D-酉算子 G：$\mathbb{H}^{\dim(H)} \to \mathbb{H}^D_H$，即 $G^\top J_D G = J_{\dim(H)}$，反之亦然。

C. 最小失真双曲子空间

1) 现有工作的回顾：Chami 等人 提出了 HoroPCA[5]。他们将 GH$(p, q_1, \ldots, q_K)$ 定义为测地凸包$\gamma_1, \ldots, \gamma_K$，其中 γ_k 是一条测地线，满足 $\gamma_k(0) = p \in \mathbb{H}^D$，且对所有 $k \in [K]$，$\lim_{t \to +\infty} \gamma_k(t) = q_k \in \partial\mathbb{H}^D$。$\gamma_1, \ldots, \gamma_K$ 的测地凸包包含 $\gamma_k(t)$ 与 $\gamma_{k'}(t')$ 之间的直线，对所有 $t, t' \in \mathbb{R}$ 和 $k, k' \in [K]$ 均成立。

主张 8。 GH$(p, q_1, \ldots, q_K)$ 是一个双曲仿射子空间。

他们的目标是最大化投影方差的一个代理量：

$$\text{cost}_{\text{Chami}}(\mathbb{H}^D_H | \mathcal{X}) = -\mathbb{E}_N\Big[d\big(\widehat{\mathcal{P}}_H(x_n), \widehat{\mathcal{P}}_H(x_{n'})\big)^2\Big], \tag{11}$$

其中 $\mathcal{P}_H$ 是 horosphere 投影算子——它不是测地线（距离最小化）投影。他们提出了一种序贯算法，使用梯度下降法最小化公式 (11) 中的代价函数，步骤如下：(1) 通过梯度下降计算弗雷歇均值作为基点；(2) 根据低维最优仿射子空间估计更高维的仿射子空间。双曲主成分分析问题可以表述如下：

$$\min_{\substack{G \in \mathbb{R}^{(D+1)\times(K+1)} \\ \{y_n \in \mathbb{H}^K\}_{n \in [N]}}} \mathbb{E}_N\big[\|x_n - Gy_n\|_2^2\big] :\ G^\top J_D G = J_K, \tag{12}$$

其中，$x_1, \ldots, x_N \in \mathbb{R}^{D+1}$是测量值，$y_1, \ldots, y_N \in \mathbb{H}^K$是低维双曲特征。方程（12）中的表述将欧几里得矩阵分解为一个切片-J_D 酉矩阵和一个双曲矩阵——这是未来研究的一个课题。

2）双曲主成分分析的适当代价函数：我们选择$f(x)$ = 双曲正弦$^2(\sqrt{|C|x})$ 得到以下代价函数：

$$\mathrm{cost}(\mathbb{H}_H^D|\mathcal{X}) = \mathbb{E}_N\big[\sum_{k\in[K']}[x_n, h_k']^2\big]; \tag{13}$$

参见命题 5。我们将 cost($\mathbb{H}_H^D|\mathcal{X}$) 理解为沿法向切向量方向上各点累积的图像耗散量。如果 $x \in \mathbb{H}^D$ 在法向切向量方向上没有分量——即 $[x, h_k'] = 0$，其中 h_1'、.....、$h_{K'}'$是 $H^\perp$ 的正交基向量——那么 $\sum_{k\in[K']}[x, h_k']^2 = 0$。式（13）中的失真函数使双曲主成分分析可表述为如下约束优化问题：

问题1。 设 x_1、.....、$x_N \in \mathbb{H}^D$和$C_x = \mathbb{E}_N\big[x_n x_n^\top\big]$。双曲主成分分析旨在求出一个点 $p \in \mathbb{H}^D$ 及一组正交向量 h_1'、.....、$h_{K'}' \in T_p\mathbb{H}^D = p^\perp$，使函数 $\sum_{k\in[K']} {h_k'}^\top J_d C_x J_d h_k'$最小。

问题1旨在最小化式（13）中的代价函数，即

$$\mathrm{cost}(\mathbb{H}_H^D|\mathcal{X}) = \sum_{k\in[K']} {h_k'}^\top J_D C_x J_D h_k',$$

在 $p \in \mathbb{H}^D, H \subseteq p^\perp, \mathrm{dim}(H) = K$的条件下，对 $p \in \mathbb{H}^D, h_1', \ldots, h_{K'}' \in p^\perp, [h_i', h_j'] = -C\delta_{i,j}$ 和$C_x = \mathbb{E}_N\big[x_n x_n^\top\big]$进行优化。问题1要求在适当的切空间 $T_p\mathbb{H}^D$ 中寻找 C_x 的 J_D-正交向量 $h_1', \ldots, h_{K'}'$，使其关于洛伦兹标量积所捕获的能量尽可能少。

注释3。 矩阵的谱是其特征值的集合。从矩阵的特征值分解中舍弃一个特征值，会以与被舍弃特征值的大小（即被舍弃的能量）成正比的误差来近似该矩阵。类似地，可以将二阶矩矩阵 C_x 的 J_D-谱定义为其 J_D-特征值的集合。正如我们在数值实验中所展示的，我们使用 J_D-谱来识别是否存在离群双曲点。

这与欧几里得主成分分析类似：该子空间由协方差矩阵的主特征向量张成。然而，要证明双曲主成分分析定理，我们需要关于二阶矩矩阵 J_D-可对角化性的一个技术性结果。

命题 7. 若 $A \in \mathbb{R}^{(D+1)\times(D+1)}$ 是一个对称且 J_D-可对角化的矩阵，即 $AJ_DV = VJ_D\Lambda$，其 Λ的对角元素（绝对值）互不相同，则 $A = V\Lambda V^\top$，其中 V 是一个 J_D-酉矩阵。

根据命题 7，任何对称且 J_D-可对角化的矩阵，只要其 J_D-特征值（绝对值）互不相同，就有 D 个正特征向量和一个负 J_D-特征向量——它们彼此正交。

定理 4. 设 x_1、.....、$x_N \in \mathbb{H}^D$和 $C_x = \mathbb{E}_N\big[x_n x_n^\top\big]$是一个 J_D-可对角化矩阵。那么，点 p 的最优解是 C_x 的经缩放的负 J_D-特征向量，而最优的 h_1'、.....、$h_{K'}'$是与 C_x中最小 K' J_D-特征值对应的经缩放的正 J_D-特征向量。此外，H 由 C_x中与最大的 J_D-特征值对应的 $K = D - K'$ 个经缩放的正 J_D-特征向量张成。

J_D 的可对角化性条件要求 C_x 与某个对角矩阵相似。命题 7 给出了一个充分条件，确保其具有 J_D 可对角化性；事实上，我们猜想，即使它具有重复的 J_D 特征值，对称性仍然是充分条件。

主张 9。 方程 (13)中的代价函数是适当的。

方程 (13) 中的适当代价函数给出了双曲质心的以下闭式定义。

定义 12 $\mathcal{X}$ μ $\mathcal{X}$ $\mathbb{E}_N[f\circ$。 的双曲均值为 $\overset{d}{()}$，若 $\overset{\mathcal{X}}{(x_n, \mu())}] = \min_{p\in\mathbb{H}^D}\mathbb{E}_N[f\circ d(x_n, p)]$。该解是 C_x的缩放后的负 J_D特征向量。

注释 4。 空间形式（欧几里得、球面和双曲）主成分分析的形式体系通过使用（不）定特征方程呈现出相似性。这源于引入了适当代价函数，由此得到关于基点和切向量的二次代价函数。然而，这种方法不一定能推广到其他黎曼流形。这一局限性是由于一般黎曼流形（例如秩亏半正定矩阵流形 *[27]*）缺少 *(1)* 简单的黎曼度量、*(2)* 闭式距离函数，以及 *(3)* 闭式指数映射和对数映射。

VI. 数值结果

我们从准确度和速度两方面，将空间形式主成分分析算法（SFPCA）与其他领先算法进行比较。

A. 合成数据和实验设置

我们在已知（但随机生成的）球形和双曲仿射子空间上生成随机含噪点。随后，我们通过估计仿射子空间，应用主成分分析方法来恢复投影点。我们开展实验，考察点的数量 N、环境空间的维度 D、仿射子空间的维度 K 以及噪声水平 σ 对算法性能的影响。

1) 随机仿射子空间：对于固定的环境空间维度和子空间维度 D 和 K，我们从正态分布中采样并将其归一化，以得到球形（或双曲）点 p。接着，我们从标准正态分布中生成随机向量，并使用格拉姆–施密特过程构造切向量：将第一个随机向量投影到 $p^\perp$ 上并归一化为 $h_1 \in T_pS^D$，其中 $S \in \{\mathbb{S}, \mathbb{H}\}$。然后，我们将第二个随机向量投影到 $(p \oplus H)^\perp$ 上，其中 $H = h_1$，并将其归一化为 $h_2 \in T_pS^D$。重复此过程，直到在 T_pS^D 中构造出一个 K 维仿射子空间。

2）含噪点：令 $H \in \mathbb{R}^{(D+1)\times K}$ 为 T_pS^D的子空间。对于 $n \in [N]$，我们生成 $c_n \sim \mathcal{N}(\mathbf{0}, \alpha_S I_{K+1})$ 并令 $v_n = Hc_n$。为添加噪声，我们将$\nu_n \sim \mathcal{N}(\mathbf{0}, \alpha_S\sigma^2 I_{D+1})$ 投影到 T_pS^D上，即 $p^\perp v_n$。然后令 $x_n = \exp_p(v_n + p^\perp\nu_n)$ 为含噪点。最后，当 $S = \mathbb{S}$ 时令 $\alpha_S = \frac{\pi}{4}$，当 $S = \mathbb{H}$时令 $\alpha_S = 1$。

3) 对含噪数据进行主成分分析：我们使用每种算法估计一个仿射子空间 $S_{\widehat{H}}^D$，其中 $\widehat{H} \subseteq T_{\widehat{p}}S^D$ 和 $\widehat{p}$ 是估计参数。我们令 n_i定义为$= \mathbb{E}_N[d(x_n,\ \mathcal{P}_H(x_n))]$ 测量值与真实子空间之间距离的经验均值，并令 n_o定义为$= \mathbb{E}_N\Big[d\big(\mathcal{P}_{\widehat{H}}(x_n)\ \mathcal{P}_H(\mathcal{P}_{\widehat{H}}(x_n))\big)\Big]$ 为

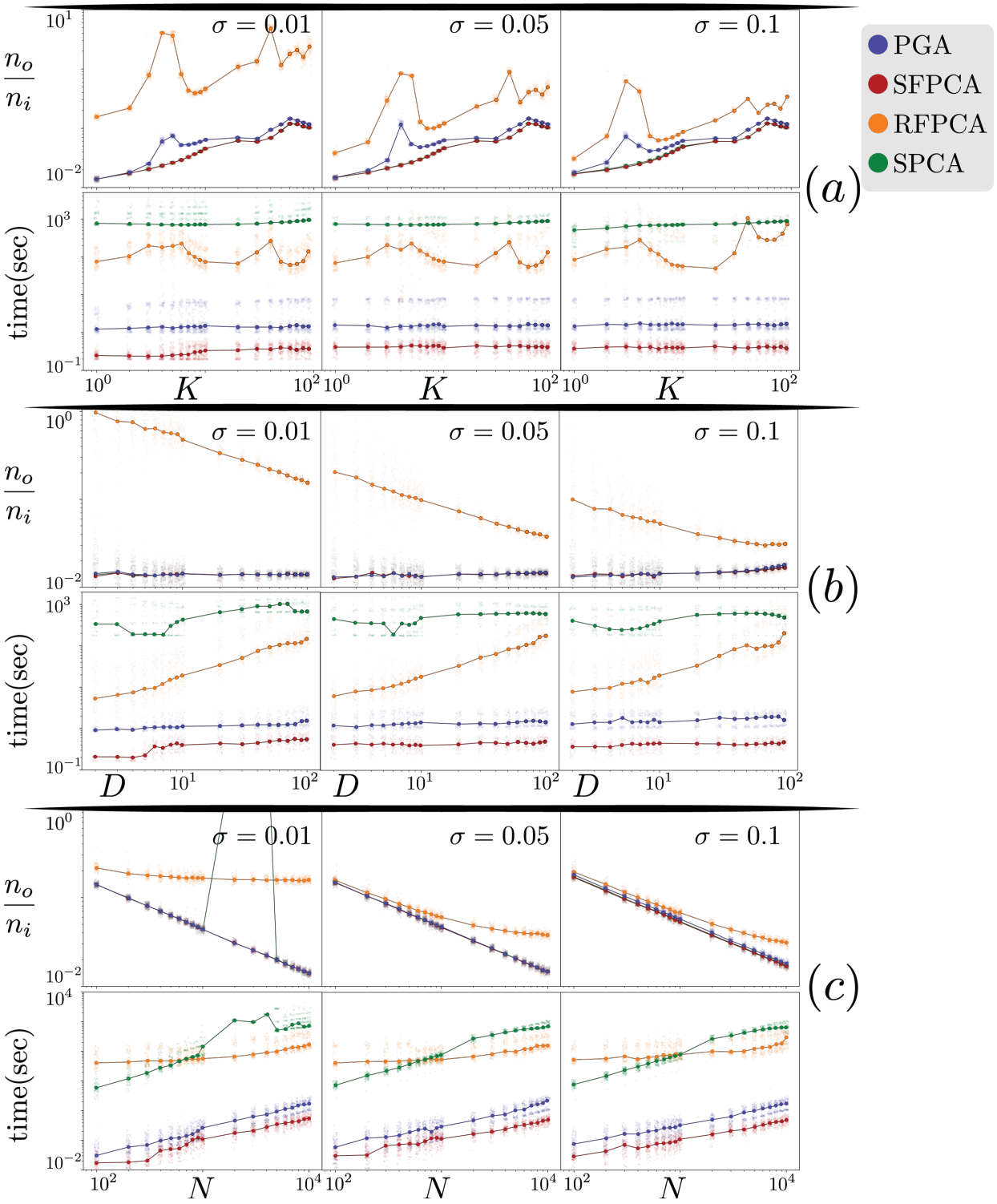


图 3. 对于每个球面实验，纵轴报告运行时间和归一化输出误差。每个点对应一次随机试验，连线圆圈表示所有试验的中位数。图（a，b，c）分别展示了 $\mathbb{S}(K_1)$ $\mathbb{S}(D_1)$ $\mathbb{S}(N_1)$ 的结果。所有坐标轴均采用对数刻度。

去噪点 $\{\mathcal{P}_{\widehat{H}}(x_n)\}_{n\in[N]}$ 与真实仿射子空间之间的平均距离。如果 $S^D_{\widehat{H}}$ 是对 S^D_H 的良好近似，那么 n_o 就很小。我们使用归一化输出误差 $\frac{n_o}{n_i}$ 来评估算法的性能。

注 5. n_o 与 n_i 的比值量化了估计点偏离真实子空间的程度，相较于相对于原始含噪点。该量经过归一化，即不受数据点尺度影响，因此非常适合在 D、K、σ 和 N 变化时比较结果。这个比值的合理上界为1——因为预期主成分分析能够通过为点集找到最优低维仿射子空间来去噪。

4）随机实验与算法：对于每个随机仿射子空间和含噪点集，我们报告每种算法的归一化误差和运行时间。随后，将每次随机试验重复100次。我们使用自己实现的主测地线分析（PGA）[13]。我们还为球面主成分分析 [10] 实现了黎曼函数主成分分析（RFPCA），并实现了球面主成分分析（SPCA） [32]。由于SPCA的计算开销较大，我们先运行自己的SFPCA，为其提供良好的初始参数。对于双曲实验，我们使用HoroPCA [5] 和Chami等人 [5]实现的重心子空间分析（BSA） [38], 。

B. 球面主成分分析

1) 实验 $\mathbb{S}(K_1)$：对于固定的 $D = 10^2$，$N = 10^4$，随着子空间维数增加 K，所有算法的归一化输出误差都会增大；见图 3（a）。RFPCA 的结果不可靠，而其他方法的误差降低模式相似。当 K 接近 D时，SFPCA 相较于 PGA 略有优势，但这一优势始终如一。SFPCA 的速度快于其他方法，并且 K 对运行时间的影响很小。

2) 实验 $\mathbb{S}(D_1)$ 和 $\mathbb{S}(D_2)$：在 $\mathbb{S}(D_1)$ 中——固定$K = 1$、$N = 10^4$ 并改变 D ——PGA、SFPCA 和 SPCA 的去噪性能相近，且不受 D影响；见图 3（b）。RFPCA 的输出误差高于其他方法。为进一步比较 SFPCA 及其有力竞争者 PGA，我们设计了具有挑战性的实验 $\mathbb{S}(D_2)$，其中包含 $K = 10$ 和 $N = 10^3$。在此设置下，SFPCA 在误差降低方面明显优于 PGA；见图 4。在这两种设置下，尽管 PGA 使用了热启动，SFPCA 在几乎所有条件下仍然是速度最快的方法。

3）实验 $\mathbb{S}$（N_1）：对于固定的 $K = 1$ 和 $D = 10^2$，改变 N 和 σ时，我们的 SFPCA 运行时间最快；在归一化输出误差最低这一点上，它与 SPCA 和 PGA 并列；见图 3（c）。正如预期，增大N 通常会使所有方法变慢，部分原因是计算 C_x 的复杂度为 O（N）。使用迭代计算在所有 N 点上求取基点 p，采用 N时十分耗时，而我们的 SFPCA 的最坏情况复杂度为 O（D^3）。由于为每种算法提供了过多的

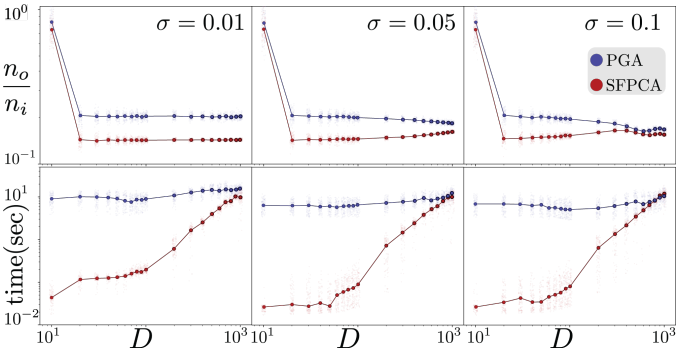


图4. 球面实验 $\mathbb{S}(D_2)$。y 轴表示运行时间和归一化输出误差。所有坐标轴均采用对数刻度。

点，SFPCA 的误差降低效果与其他方法相近。在某些情况下，SPCA 会失效，这从归一化输出误差的不稳定表现中显而易见。SPCA 每次试验处理 10^4 个点约需 15 分钟，而我们的 SFPCA 用时不到一秒。PGA 的归一化误差最接近，但速度约慢三倍。

C. 双曲主成分分析

1) 实验 $\mathbb{H}(K_1)$ 和 $\mathbb{H}(K_2)$：对于$\mathbb{H}(K_1)$（$D = 50$，$N = 51$）中的小型数据集，每次试验中，HoroPCA 和 BSA 都需要近一小时，而 SFPCA 和 PGA 只需几毫秒；见图 5（a）。增大 K 只会增加 BSA 和 HoroPCA 的运行时间，不会影响 SFPCA 和 PGA。这是意料之中的，因为它们采用贪心方式，每次估计一个维度的仿射子空间。关于误差降低，正如预期，随着 K 增大，所有方法的效果都会变差。对于较小的 σ，所有方法达到的归一化输出误差水平相近，PGA 和 SFPCA 略有优势。随着 σ 增大，与 BSA 和 SFPCA 相比，PGA 和 HoroPCA 的效果变差。对于较大的 σ，SFPCA 明显优于所有其他方法。在较大规模的$\mathbb{H}(K_2)$实验（$D = 10^2$，$N = 10^4$）中，我们比较了速度最快的两种方法 SFPCA 和 PGA；见图6（a）。当 σ较小时，两种方法在较小的 K下去噪性能相近；只有在 K较大时，SFPCA 才表现更好。随着 σ 增大，无论 K如何，SFPCA 都优于 PGA。

2) 实验 $\mathbb{H}(D_1)$ 和 $\mathbb{H}(D_2)$：在 $\mathbb{H}(D_1)$ 中，我们固定 $K = 1$、$N = 101$；在 $\mathbb{H}(D_2)$ 中，我们令 $K = 1$、$N = 10^4$。改变 D 对各方法的影响各不相同，见图5（b）。随着D 增大，SFPCA 和 PGA 的运行时间逐渐增加，但仍明显快于另外两种方法，平均运行时间不到 0.1 秒。HoroPCA 的运行时间几乎不受 D 影响，因为其代价函数（投影方差）不含参数D。HoroPCA 和 BSA 在误差降低方面都不如 SFPCA。随着 D 增大，所有方法的误差降低性能都有所提升。当 σ较大时，SFPCA 在所有算法中实现的误差降低效果最佳。比较最快的两种方法 SFPCA 和 PGA，我们在 $\mathbb{H}$（D_1）和 $\mathbb{H}$（D_2）中观察到一致的规律：(1) 无论 D 如何，SFPCA 都更快，两种方法的速度差距最高可达 10 倍。(2) 当 $\sigma < 0.1$时，PGA 在降低误差方面略优于 SFPCA；在噪声最低的情况下（$\sigma = 0.01$），对D的所有取值取平均后，PGA 的准确度高出 17%。然而，随着 σ 增大，SFPCA 的效果更佳；在最高值处（$\sigma = 0.5$），对 D取平均后，SFPCA 的表现比 PGA 高出 40%；见图6（b）。

3) 实验 $\mathbb{H}(N_1)$ 和 $\mathbb{H}(N_2)$：在 $\mathbb{H}(N_1)$, ($K = 1$,$D = 10$) 中，由于需要计算 C_x，随着 N 增大，SFPCA 和 PGA 的运行时间会增加；参见图5（c）。尽管如此，这两种方法的速度都比 HoroPCA 和 BSA 快几个数量级。随着 N 增大，所有方法的误差降低效果都会提升。在较大的数据集上比较快速方法 SFPCA 和 PGA（ $\mathbb{H}(N_2)$ ($K = 1$, $D = 10^2$)）可见，SFPCA 始终更快，在小型 σ 上的输出误差略高，而在大型 σ 上则有显著改善；参见图6（c）。

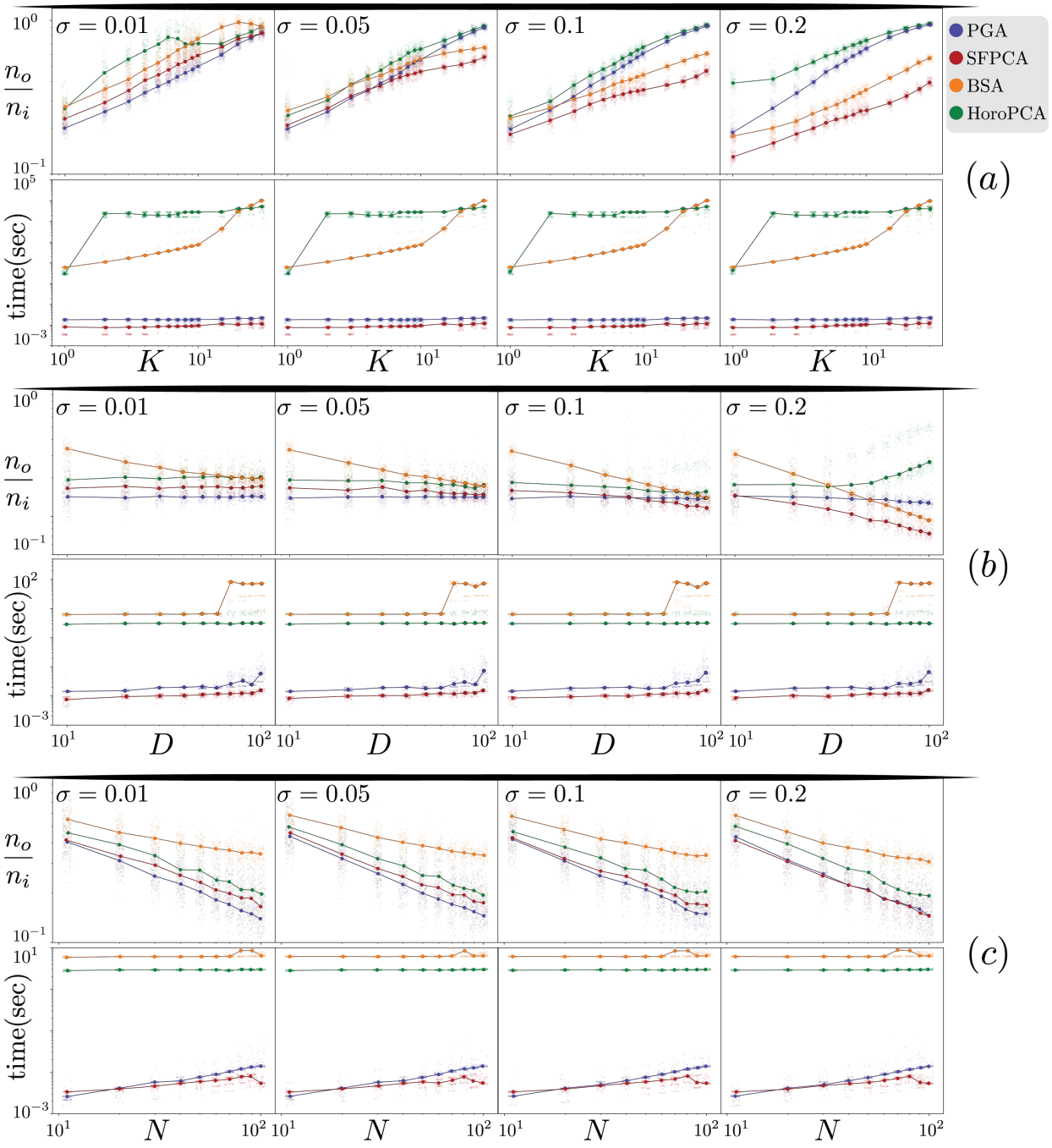


图 5. 对于每个缩小规模的双曲实验，纵轴以对数刻度报告运行时间和归一化输出误差。每个点对应一次随机试验，圆圈表示所有试验的中位数。第（a），（b）行以及（c）行中的图分别为 $\mathbb{H}$（K_1） $\mathbb{H}$（D_1）和 $\mathbb{H}$（N_1）。

D. 真实数据：球形空间

我们使用以下数据集评估主成分分析方法的性能：(1) 肠道微生物组：Lahti 等人[28] 分析了 $N = 1{,}006$ 名成年人的肠道微生物群，涵盖 $D = 130$ 个细菌群。该研究探讨了年龄、国籍、体质指数和 DNA提取方法对微生物组的影响。他们评估了年轻组（18–40 岁）、中年组（41–60 岁）和老年组（61–77 岁）之间微生物组组成的差异（三级分类问题）。(2) 咽喉微生物组：Charlson 等人 [6] 研究了吸烟对 29 人气道微生物群的影响吸烟者和33名非吸烟者（二分类问题），使用非培养法454焦磷酸测序对16S核糖体RNA进行测序。(3) 新闻组：使用 Python 的 scikit-learn 软件包，从 20新闻组数据集中保留属于两个不同类别的 $N = 400$个样本，将其简化为一个二分类问题。使用词频-逆文档频率进行特征降维，将特征缩减至 $D = 3000$ 个，以提高计算效率。每个数据集均经过标准预处理，例如归一化和平方根变换，以确保数据点呈球面分布。对于固定的子空间维数，我们估计球面仿射

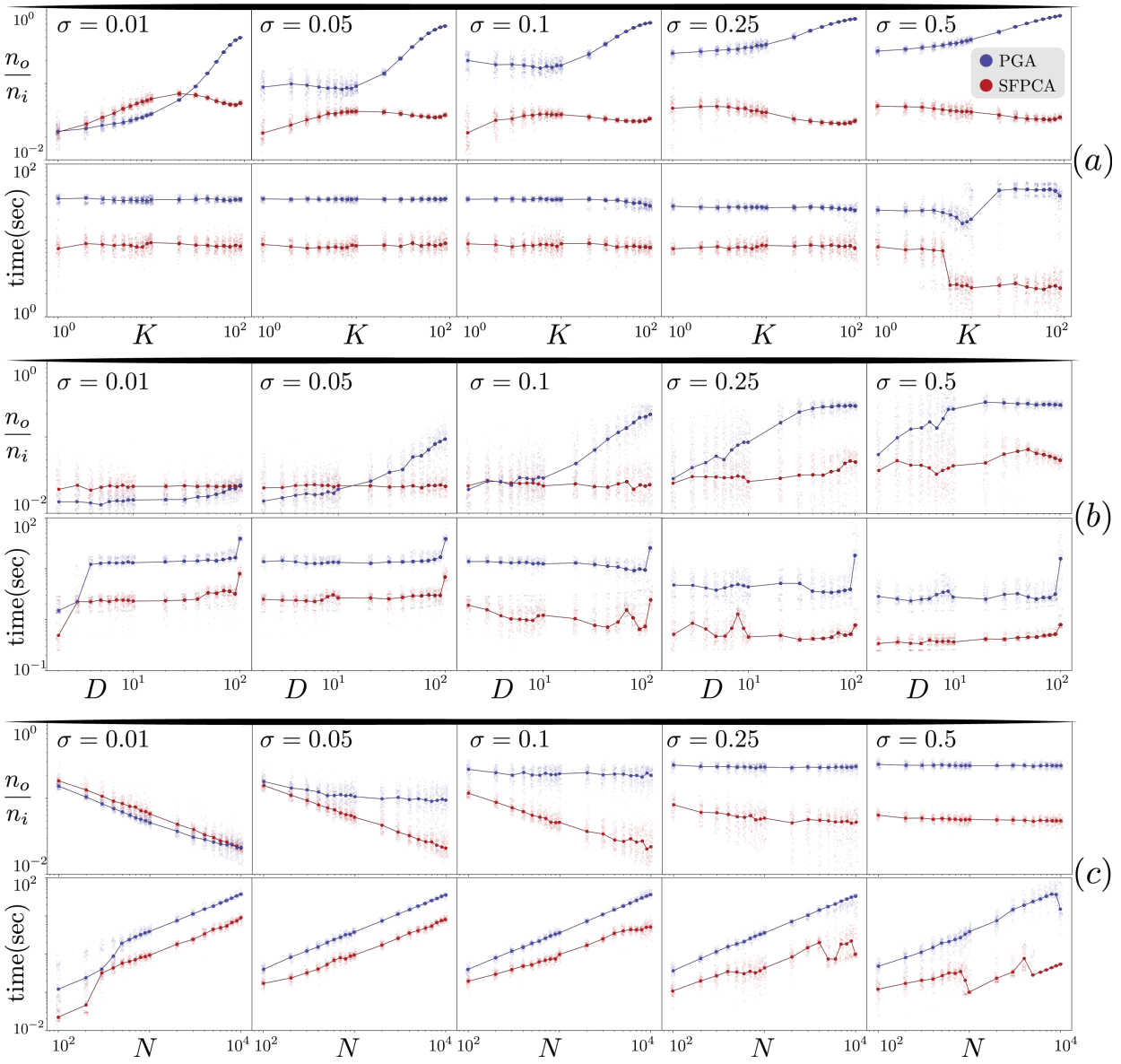


图 6. 对于每个全规模双曲实验，我们在 y 轴上以对数刻度报告运行时间和归一化输出误差。点表示一次随机试验，圆圈表示中位数。第 (*a*) 行、(*b*) 行和 (*c*) 行中的图分别为 $\mathbb{H}(K_2)$ $\mathbb{H}(D_2)$ 和 $\mathbb{H}(N_2)$。所有坐标轴均采用对数刻度。

子空间。然后，我们计算投影球面数据点，并对原始组成数据进行去噪。

失真分析。 对于组成数据，我们使用艾奇逊（$\mathbb{AI}$）、詹森–香农（$\mathbb{JS}$）和全变差（$\mathbb{TV}$）指标计算距离矩阵。我们还计算球面距离矩阵（$\mathbb{S}$）。对于每个嵌入维度 K，我们计算投影点集。随后，我们计算归一化误差；归一化误差的一个示例为 $\frac{\|D_{\mathbb{TV}}-\tilde{D}_{\mathbb{TV}}\|_F}{\|D_{\mathbb{TV}}\|_F}$，其中 $\tilde{D}_{\mathbb{TV}}$ 和 $D_{\mathbb{TV}}$ 分别是原始数据和估计数据的全变差距离矩阵。对于每种算法，我们再将这些归一化误差除以所有算法对应误差的平均值，从而得到相对度量；也就是说，如果所得相对误差大于1，则该算法的表现低于平均水平。随后，我们报告不同维度下相对误差的均值和标准差；见表 II。在所有数据集上，SFPCA 的表现均优于其他方法。新闻组由于 SPCA 和 RFPCA 的计算复杂度显著较高，实验仅比较 SFPCA 和 PGA。

分类性能。 对于每个 K，我们使用去噪组成数据，执行两种分类器：五层神经网络（$\mathbb{NN}$）和随机森林模型（$\mathbb{RF}$）。我们通过所有方法的平均准确度进行归一化，并报告分类准确度的均值和标准差。从表 II 可以看出，SFPCA 在肠道微生物组和新闻组上的性能优于竞争方法，不过准确度差异大多小于百分之一。在 A 节中，我们进一步比较这两种分类器在新闻组上的性能，并分析其与主成分分析的关系。

E. 真实数据：双曲空间

我们使用一个包含 103 种植物和藻类转录组的生物数据集 [54]。作者根据

表 II 均值和标准差形式的归一化距离误差，均除以其在各方法间的平均值。分类准确率表示相对于100% 的百分比偏差，其中100%代表各方法间的平均准确率。粗体和红色分别表示 SFPCA 和表现最佳的方法。较低的失真(↓) 和较高的准确率(↑) 更优。

指标（方法）	咽喉微生物组				肠道微生物组				新闻组	
	SFPCA	RFPCA	SPCA	PGA	SFPCA	RFPCA	SPCA	PGA	SFPCA	PGA
$\mathbb{S}(\downarrow)$	**0.88 ± 0.99**	1.13 ± 1.28	0.9 ± 0.99	1.1 ± 1.18	**0.75 ± 1.64**	0.93 ± 2.09	1.47 ± 2.07	0.85 ± 1.79	**0.77 ± 1.03**	1.23 ± 1.4
$\mathbb{AI}(\downarrow)$	**0.98 ± 0.48**	1.03 ± 0.57	0.99 ± 0.46	1.01 ± 0.52	**0.8 ± 1.06**	0.81 ± 1.03	1.55 ± 1.04	0.84 ± 1.14	**0.999 ± 0.4**	1.001 ± 0.5
$\mathbb{JS}(\downarrow)$	**0.91 ± 0.84**	1.08 ± 1.04	0.95 ± 0.84	1.06 ± 0.97	**0.75 ± 1.55**	0.93 ± 1.96	1.47 ± 1.93	0.85 ± 1.7	**0.9 ± 0.81**	1.1 ± 0.98
$\mathbb{TV}(\downarrow)$	**0.91 ± 1.0**	1.08 ± 1.21	0.94 ± 1.0	1.07 ± 1.14	**0.76 ± 1.65**	0.92 ± 2.09	1.49 ± 2.07	0.83 ± 1.75	**0.87 ± 0.88**	1.13 ± 1.12
$(\mathbb{NN})(\uparrow)$	**−0.06 ± 3.03**	0.15 ± 3.15	−0.26 ± 3.7	0.17 ± 3.0	**0.04 ± 0.36**	−0.05 ± 0.9	−0.02 ± 0.6	0.03 ± 0.3	**0.04 ± 1.6**	−0.04 ± 1.6
$(\mathbb{RF})(\uparrow)$	**0.59 ± 9.5**	1.2 ± 9.6	−1.9 ± 9.3	0.11 ± 9.8	**0.4 ± 2.3**	0.11 ± 2.7	−0.74 ± 2.8	0.2 ± 2.3	**0.3 ± 3.5**	−0.3 ± 4.4

全基因组基因推断系统发育树。树叶是现存物种，内部节点是祖先物种，分支权重表示进化距离。该数据集包含一个有 852 棵树的“未过滤”版本，以及一个移除易出错基因并过滤问题序列后得到的、包含 844 棵树的“已过滤”版本。错误会表现为离群值，先验上预计未过滤数据集中会有更多错误 先验。其他研究 [9], [34]也使用该数据集评估离群值检测方法。我们对每棵树进行预处理，将分支长度缩放至直径为 10，计算叶节点之间的距离矩阵，并使用半定规划将其嵌入 D 维（$D = 20$）双曲空间 [47]。我们使用两个指标评估主成分分析结果，然后将其用于离群值检测。

失真分析。 对于固定维度 K，我们估计双曲仿射子空间，计算投影后的双曲点及其双曲距离矩阵 $D_{\mathbb{H}}$。计算归一化距离误差 $\frac{\|D_{\mathbb{H}}-\widehat{D}_{\mathbb{H}}\|_F}{\|D_{\mathbb{H}}\|_F}$，其中 $D_{\mathbb{H}}$ 是原始距离矩阵。这些误差针对每种算法在$K \in [D]$ 上取平均，然后除以所有算法的平均归一化误差，从而得到相对误差。若相对误差大于 1，则该算法的表现低于平均水平。对于每种算法，我们报告这些相对误差在所有基因树上的均值和标准差。失真并非衡量主成分分析准确度的完美指标，因为含噪程度很高的数据在投影（去噪）步骤中理应产生较高的失真。在所有实验中，PGA 在失真方面均优于其他方法（表 III）。与合成实验不同，SFPCA 的距离保持性能处于平均水平。我们推测，这可能是因为这些树相对较小（参见图5中的缩小版双曲实验），也可能是因为较高的噪声水平使失真不适合作为准确度指标；此外，我们选择的失真函数$f = \cosh$（它会过度强调较大的距离）与距离失真度量之间也可能存在不一致。

四元组评分。为了采用一种具有生物学依据的准确度度量，我们使用四元组得分 [11]。对于目标维度K，将每种算法应用于嵌入的双曲点集，以计算投影（去噪）点。对于每组四个点，我们使用四点条件 [53]，寻找距离失真最小的最优树拓扑。对于 10^5组随机选取（但固定）的四个投影点，

表 III 归一化距离误差（$\mathbb{H}$）均值和标准差均除以各方法的平均值。四元组评分（$\mathbb{Q}$）为相对于100% （平均值）的百分比偏差。失真(↓) 越低、评分(↑) 越高越好。

方法	已筛选		未筛选	
	$\mathbb{Q}(\uparrow)$	$\mathbb{H}(\downarrow)$	$\mathbb{Q}(\uparrow)$	$\mathbb{H}(\downarrow)$
SFPCA	**1.50 ± 1.91**	**0.98 ± 0.23**	**1.10 ± 1.74**	**1.01 ± 0.25**
PGA	1.48 ± 1.47	0.55 ± 0.09	1.80 ± 1.45	0.53 ± 0.08
BSA	1.48 ± 1.61	1.45 ± 0.24	1.67 ± 1.91	1.45 ± 0.31
HoroPCA	−4.48 ± 2.79	1.01 ± 0.21	−4.57 ± 2.58	1.00 ± 0.20

估计其拓扑，并与基因树中的真实拓扑进行比较。对于每个维度 $K \in [D]$，我们计算正确估计的拓扑所占百分比，然后对所有维度取平均值。我们按照表 III 中的详细说明，以所有方法的平均得分为基准进行归一化，并报告四元组评分的均值和标准差。在这些实验中，与替代方法相比，PGA 和 SFPCA 表现最佳。这一点尤其值得注意，因为四元组得分衡量的是树拓扑准确性，而非距离。

1) 基于双曲谱的异常值检测：我们展示了 J_D 特征方程（定义 10）在物种树估计中的实用价值。如定理 4 所证明，主轴与 C_x 的前列 J_D 特征向量对齐。因此，最优 SFPCA 代价函数对应于其被忽略的 J_D 特征值之和。我们推测，含有离群值的树拥有更多异常的 J_D 特征值；见图7（a，b）。如果一棵树包含一个物种异常值集合（可能源自错误序列），那么它第二大的 J_D 特征值（λ_2）[1]会显著大于其余特征值。我们通过绘制其归一化保留能量 $\frac{\sum_{k=2}^{K}\lambda_k}{\sum_{d=2}^{D}\lambda_d}$（或累积谱）与归一化嵌入维度（或 J_D 特征值的数量） $x = K/D$ 之间的关系，并找出膝点，来量化这一现象。这使我们能够根据基因树的双曲谱对其进行排序。

排序后，我们使用膝点值最小（最不容易受离群值影响）的前 N 棵树，通过 ASTRAL [57] 构建物种拓扑。ASTRAL 输出四元组

[1]该 J_D特征值 λ_1 对应于基点。

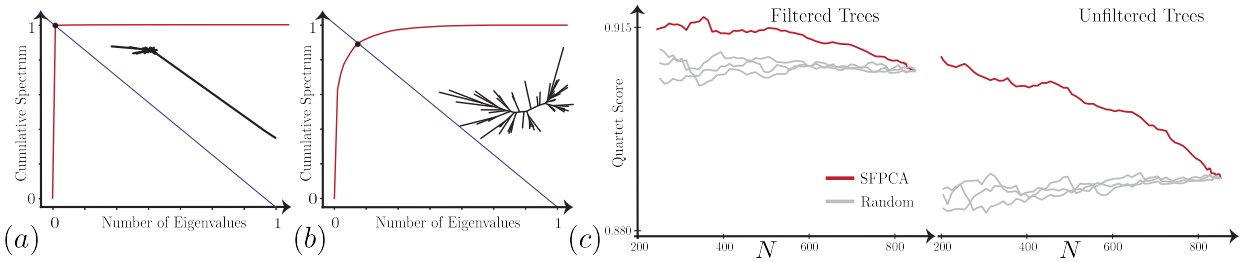


图7. (*a*) 和 (*b*)：归一化保留能量与两个基因树的归一化特征值数量之间的关系。膝点（与 $y = 1 - x$ 线的交点）对于含离群值的树趋近于 (0, 1). (*c*) 使用前 N 棵树（膝点值）构建的物种树的四元组得分与随机顺序的比较 .

得分，用于衡量估计的物种树与输入基因树之间的关系；得分越高，表示输入树之间的一致性越高。因此，过滤后得分升高，意味着离群基因树（可能是从问题序列中推断得出的）已被移除。我们的结果（图7（c））表明，只有我们的 SFPCA 提供的基于双曲谱的排序，能够有效识别与其他树最不相似的劣质树，而无需显式比较树拓扑。相比之下，随机排序会保持四元组得分不变。过滤后的树得分高于未过滤的树，因此进一步过滤带来的收益较小。值得注意的是，仅凭特征值，我们就能有效找出进化历史不一致的基因。

致谢

作者谨此感谢美国国立卫生研究院（NIH）提供经费支持。本研究部分由 NIHGrant 1R35GM142725 资助。

附录

2) 命题*1*：令 $x,\ y \in \mathbb{S}^D_H \subseteq \mathbb{S}^D$ 和 $\gamma_{x,\ y}(t)$ 为测地线，其中 $\gamma_{x,\ y}(0) = x$ 且 $\gamma_{x,\ y}(1) = y$。这条测地线位于 $\text{span}\{x, y\}$ 和 $\mathbb{S}^D$ 的交集中。由于 $x,\ y \in \mathbb{S}^D_H = p \oplus H \cap \mathbb{S}^D$ 且 $p \oplus H$ 是一个子空间，因此我们有 $\text{span}\{x, y\} \subseteq p \oplus H$。所以，对于所有 $t \in [0, 1]$，我们都有 $\gamma_{x,\ y}(t) \in \text{span}\{x, y\} \cap \mathbb{S}^D \subseteq p \oplus H \cap \mathbb{S}^D = \mathbb{S}^D_H$。

3) 命题*2*：对于球面仿射子空间 $\mathbb{S}^D_H$，我们定义切片酉矩阵 $G = \left[C^{\frac{1}{2}}p, C^{-\frac{1}{2}}h_1, \ldots, C^{-\frac{1}{2}}h_K\right]$，其中 p 是基点，h_1、……、h_K 是 H 的正交基。于是，对于所有 $y \in \mathbb{S}^K$，我们有 $Gy \in \mathbb{S}^D_H$。反过来，对于任意切片酉矩阵 $G = [g_0, g_1, \ldots, g_K]$，令 $p = C^{-\frac{1}{2}}g_0$，并对 $k \in [K]$ 取 $h_k = C^{\frac{1}{2}}g_k$。由于 h_k's 和 p 正交，我们可以定义该球面仿射子空间。

4) 主张 *3*：我们将其主成分分析代价函数简化为 $\text{cost}(\mathbb{S}^D_H|\mathcal{X}) = 1 - \mathbb{E}_N\left[\|P_H(\sum_{k\in[K']} h_k'^{\top} C_x h_k')\|_2^2\right] = {}_{(a)}= \mathbb{E}_N[\sum_{k\in[K']}\langle x_n, h_k'\rangle^2]_{(b)}$其中，(a) 根据命题得出，$h_1' \ldots h_{K'}'$ $H^{\perp} \subseteq T_p\mathbb{S}^D$,,是 的正交基，(b) 则根据迹的循环性质得出。

5) 主张 *4*：根据推论 2 和定义 7，最优的零维仿射子空间（一个点）是任何其他球面仿射子空间的子集。一般而言，$\mathbb{S}^D_{H_1} \subseteq \mathbb{S}^D_{H_2}$ $\mathbb{S}^D_{H_1} \leq \mathbb{S}^D_{H_2}$ 当且仅当 dim( ) dim( )。

6) 主张 *5*：设 $v \in \mathbb{C}^{D+1}$ 满足 $AJ_Dv = \text{sgn}([v^*, v])\lambda v$，其中 $|[v^*, v]| = 1$，则 $\|v\|_2^{-1}v$ 是 AJ_D 的特征向量，其特征值为 $\text{sgn}([v^*, v])\lambda$。

7) 主张 *6*：设 $x,\ y \in \mathbb{H}^D_H$ 和 $\gamma_{x,\ y}$ 是测地线，其中 $\gamma_{x,\ y}(0) = x,\ \gamma_{x,\ y}(1) = y$。这条测地线属于张成空间 $\{x, y\} \cap \mathbb{H}^D$。由于 $x, y \in \mathbb{H}^D_H = p \oplus H \cap \mathbb{H}^D$，且 $p \oplus H$ 是子空间，因此张成空间 $\{x, y\} \subseteq p \oplus H$。因此，对于所有 $t \in [0, 1]$，我们有 $\gamma_{x,\ y}(t) \in$ 张成空间 $\{x, y\} \cap \mathbb{H}^D \subseteq \mathbb{H}^D_H$。

8) 主张 *7*：对于双曲仿射子空间 $\mathbb{H}^D_H$，我们定义 $G = \left[|C|^{\frac{1}{2}}p, |C|^{-\frac{1}{2}}h_1, \ldots, |C|^{-\frac{1}{2}}h_K\right]$，其中 p 是基点，h_1、……、h_K 是 H 的正交基。我们有 $G^{\top}J_KG = J_K$ 和 $Gy \in \mathbb{H}^D$，对所有 $y \in \mathbb{H}^D_H$ 均成立。反过来，对于任意切片酉矩阵 $G = [g_0, g_1, \ldots, g_K]$，我们令 $p = |C|^{-\frac{1}{2}}g_0$ 和 $h_k = |C|^{-\frac{1}{2}}g_k$，其中 $k \in [K]$。由于 h_k 和 p 相互正交，我们可以定义球面仿射子空间。

9) 主张 8：设 $p \in \mathbb{H}^D$、q_1、……、$q_K \in \partial H^D$ 以及 γ_1、……、γ_K 为上述测地线。点 $x \in S \text{ def= } \text{GH}(p, q_1, \ldots, q_K)$ 属于一条测地线，其端点为 $\gamma_k(t)$ 和 $\gamma_{k'}(t')$，其中 t、$t' \in \mathbb{R}$ 以及 k、$k' \in [K]$。我们来证明，对于子空间 $H \subseteq T_p\mathbb{H}^D$，有 $x \in \mathbb{H}^D_H$。根据主张 6，$\mathbb{H}^D_H$ 是一个测地子流形。只需证明 γ_1、……、γ_K 属于 $\mathbb{H}^D_H$。令 $h_k = \gamma_k'(0) \in T_p\mathbb{H}^D$，对所有 $k \in [K]$ 均成立，并且 $H =$ 张成 $\{h_1, \ldots, h_K\}$。这就证明了 $S \subseteq \mathbb{H}^D_H$。反过来，令 $x \in \mathbb{H}^D_H$ 为按前述方式构造的双曲仿射子空间。由于 $\mathbb{H}^D_H$ 是一个测地子流形，x 属于一条测地线，其端点为 $\gamma_k(t)$ 和 $\gamma_{k'}(t')$，其中 t、$t' \in \mathbb{R}$ 以及 k、$k' \in [K]$，并按前述方式构造。由 $x \in \text{convhull}(\gamma_1, \ldots, \gamma_K)$，我们有 $\mathbb{H}^D_H \subseteq S$。

10) 主张 *9*：根据定理 4 和定义 12，最优零维仿射子空间是任意其他仿射子空间的子集。对于最优仿射子空间 $\{\mathbb{H}^D_{H_i}\}$，我们有 $\mathbb{H}^D_{H_1} \subseteq \mathbb{H}^D_{H_2}$ 当且仅当 $\dim(\mathbb{H}^D_{H_1}) \leq \dim(\mathbb{H}^D_{H_2})$。

11) 命题 *1*：考虑以下拉格朗日函数：

$$\mathcal{L}(y, \gamma, \Lambda) = \langle x, y\rangle + \gamma(\langle y, y\rangle - C^{-1}) + \sum_{k\in[K']} \lambda_k\langle y, h_k'\rangle, \quad (14)$$

其中 $\Lambda \text{ def= } \{\lambda_k : k \in [K']\}$，$\langle y, y\rangle = C^{-1}$ 且 $\langle y, h_k'\rangle = 0,\ \forall k \in [K']$。方程 (14) 的解具有以下形式：$\mathcal{P}_H(x) = \sum_{i\in[K']} \alpha_i h_i' + \beta x$，其中 $\{\alpha_i\}_{i\in[K']}$ 和 β 为标量。子空间条件 — $\langle \mathcal{P}_H(x)\ h_k'\rangle = 0$, $\forall k \in [K']$ — 给出 $\mathcal{P}_H(x) = \beta P_H(x)$，其中 $P_H(x) = x - C^{-1}\sum_{k\in[K']}\langle x, h_k'\rangle h_k'$。施加范数条件后，我们得到 $\mathcal{P}_H(x) = C^{-\frac{1}{2}}\|P_H(x)\|_2^{-1}P_H(x)$，其中 $\|P_H(x)\|_2^2 = C^{-1}(1 - \sum_{k\in[K']}\langle x, h_k'\rangle^2)$。于是，我们有 $d\ x\ \mathcal{P}_H\ x = C^{-\frac{1}{2}}\ \left(C\langle x\ C^{-\frac{1}{2}}\|P_H\ x\ \|_2^{-1}P_H\ x\ \right) = (\ ,\ (\ ))\ C^{-\frac{1}{2}}\ \left(C^{\frac{1}{2}}\|P_H\ x\ \|_2\right)\ P_H\ x = 0\ x \in \text{acos}\ (\ )\ (\ )\ \text{acos}\ (\ )$。若 ( ) — 即 $\text{span}(H^{\perp})$ — 则 $\mathcal{P}_H(x) \in \mathbb{S}^D_H$ 不唯一；但投影距离定义良好，为 $d(x, y) = C^{-\frac{1}{2}}\frac{\pi}{2}, \forall y \in \mathbb{S}^D_H$。

12) 命题*2*：令 $A = [Cp, h_1, \ldots, h_K, h'_1, \ldots, h'_{K'}] \in \mathbb{R}^{(D+1)\times(D+1)}$ 其中 $K + K' = D$。A 的不同列彼此正交 — 即 $A^\top A = CI_{D+1}$。因此，$p, h_1, \ldots, h_K, h'_1, \ldots, h'_{K'}$ 线性无关。因此，我们有 $P_H(x - C^{-1}\sum_{k\in[K']}\langle x\ h'_k\rangle h'_k$ $\{\alpha_k\}_{k\in[K]}$ $\{\beta_k\}_{k\in[K']})$

$= \sum_{k\in[K]}\alpha_k h_k + \sum_{k\in[K']}\beta_k h'_k + \gamma p$ (a)，其中 β_k, γ 为标量，而 (a) 由命题 1 得出。因此，我们有 $\beta_k = C^{-1}\langle P(x)\ h'_k\rangle = 0$，$\alpha_k = C^{-1}\langle P(x)\ h_k\rangle = C^{-1}\langle x, h_k\rangle$ 以及 $\gamma = C\langle P(x)\ p\rangle = C\langle x, p\rangle$。据此可计算 $\|P_H(x)\|_2$，并证明该命题。

13) 命题*3*：设 A 是一个实矩阵，使得 $A = A^{[\top]}$，即 $A = J_D A^\top J_D$。设 (v, λ) 是 A 的一个特征向量–特征值对。那么，我们有 $AJ_D v = \mathrm{sgn}([v^*, v])\lambda v$ 和 $AJ_D v^* = \mathrm{sgn}([v^*, v])\lambda^* v^*$。我们有 $\lambda^* = \lambda$，因为

$$\begin{aligned}\lambda[v^*, v] &= (v^*)^\top J_D A J_D v\ \mathrm{sgn}\big([v^*, v]\big)\\ &= (AJ_D v^*)^\top J_D v\ \mathrm{sgn}\big([v^*, v]\big)\\ &= \mathrm{sgn}\big([v^*, v]\big)\lambda^* v^H J_D v\ \mathrm{sgn}\big([v^*, v]\big) = \lambda^*[v^*, v].\end{aligned}$$

14) 命题*4*：设 A 是一个满秩矩阵，且 $v \in \mathbb{C}^{1,D}$ 满足 $AJ_D v = \lambda v$ 和 $[v^*, v] = 0$。那么，我们有 $v^H J_D A J_D v = v^H J_D \lambda v = \lambda[v^*, v] = 0$。这与 A 是满秩矩阵这一假设矛盾。

15) 命题 *5*：考虑以下拉格朗日函数：

$$\mathcal{L}(y, \gamma, \Lambda) = [x, y] + \gamma([y, y] - C^{-1}) + \sum_{k\in[K']}\lambda_k[y, h'_k], \quad (15)$$

其中 $\Lambda = \{\lambda_k\}_{k\in[K']}$，对于标量 $\{\alpha_i\}_{i\in[K']}$ 和 β，解为 $\mathcal{P}_H(x) = \sum_{i\in[K']}\alpha_i h'_i + \beta x$。子空间条件，$[\mathcal{P}_H(x)\ h'_k] = 0$，$\forall k \in [K']$，给出 $\mathcal{P}_H(x) = \beta P_H(x)$，其中 $P_H(x) = x + C^{-1}\sum_{k\in[K']}[x, h'_k]h'_k$。施加范数条件后，得到 $\mathcal{P}_H(x) = |C|^{-\frac{1}{2}}\|P_H(x)\|^{-1}P_H(x)$，$\|P_H(x)\|_{\text{def}}=\sqrt{-[P_H(x)\ P_H(x)]} = |C|^{-\frac{1}{2}}\sqrt{1 + \sum_{k\in[K']}[x, h'_k]^2}$。我们有 $d\ x\ \mathcal{P}_H\ x = |C|^{-\frac{1}{2}}\ (C[x, \mathcal{P}_H\ x]) = (,\ ())\ |C|^{-\frac{1}{2}}\ (|C|^{\frac{1}{2}}\|P_H\ x\|)$ acosh() acosh()。

16) 命题 *6*：令 $A = [Cp, h_1, \ldots, h_K, h'_1, \ldots, h'_{K'}] \in \mathbb{R}^{(D+1)\times(D+1)}$，其中 $K + K' = D$。A 的列两两 J- 正交，即 $A^\top J_D A = |C|J_D$。由于我们有 $J_D A^\top J_D A_D = |C|I_{D+1}$，$p, h_1, \ldots, h_K, h'_1, \ldots, h'_{K'}$ 线性无关，因此有 $P_H(x) = \sum_{k\in[K]}\alpha_k h_k + \sum_{k\in[K']}\beta_k h'_k + \gamma p$ (a) $= x + C^{-1}\sum_{k\in[K']}[x, h'_k]h'_k$，其中 $\{\alpha_k\}_{k\in[K]}$，$\{\beta_k\}_{k\in[K']}$，γ 为标量，而 (a) 根据命题 5 得出。所以我们有 $\beta_k = -C^{-1}[P(x)\ h'_k] = 0$，$\alpha_k = -C^{-1}[x, h_k]$，以及 $\gamma = C[P(x)\ p] = C[x, p]$，即 $P_H(x) = C[x, p]p - C^{-1}\sum_{k\in[K]}[x, h_k]h_k$。然后，我们可以计算 $\|P_H(x)\|$ 并证明该命题。

17）命题 *7*：令 $A = V\Lambda V^\top$，其中 V 是 J_D-酉的。由于 Λ 是一个对角矩阵，我们令 $A = VJ_D\Lambda J_D V^\top$。我们有 $AJ_D V = VJ_D\Lambda J_D\ (V^\top J_D V) = VJ_D\Lambda$，即 A 是 J_D-可对角化的。

令 $A \in \mathbb{R}^{(D+1)\times(D+1)}$ 为对称的，且 $AJ_D V = VJ_D\Lambda$ 对于一个 J_D-可逆的 V 以及具有互不相同（绝对值）的元素的 $\Lambda = \mathrm{diag}(\lambda_d)_{d\in[D+1]}$ 成立。于是，我们有

$$AJ_D v_d = \begin{cases}-\lambda_d v_d, & \text{if } d = 1\\ \lambda_d v_d, & \text{if } d \neq 1,\end{cases} \quad (16)$$

其中，v_d 是 V 的第 d 列。在特征方程 (16) 中，负号（正号）对应范数为负（正）的特征向量。对于不同的 i，j，我们有

$$\begin{aligned}\big|\lambda_i[v_i, v_j]\big| &\overset{\text{(a)}}{=} \big|[AJ_D v_i, v_j]\big| = \big|v_i^\top J_D A^\top J_D v_j\big|\\ &= \big|v_i^\top J_D A J_D v_j\big| = \big|[v_i, AJ_D v_j]\big| = \big|\lambda_j[v_i, v_j]\big|,\end{aligned}$$

其中，(a) 是由特征方程 (16) 得出的。由于 $|\lambda_i| \neq |\lambda_j|$，因此我们必有 $[v_i, v_j] = 0$。不失一般性，J_D 个特征向量经过缩放，使得对 $d \in [D+1]$ 有 $|[v_d, v_d]| = 1$。引理 1 表明，对于 $d > 1$，$[v_1, v_1] = -1$ 且 $[v_d, v_d] = 1$。

Lemma 1. $V^\top J_D V = J_D$.

证明。令 A，其中 $AJ_D V = VJ_D\Lambda$，且 V J_D-对角化 $A^\top J_D A$，即 $V^\top J_D(A^\top J_D A)J_D V\Lambda$ (a)$=(VJ_D\Lambda)^\top J_D(VJ_D\Lambda)$ (b) $= \Lambda J_D\Lambda' J_D\Lambda = \Lambda^2\Lambda'$，其中 (a) 由 $AJ_D V = VJ_D\Lambda$ 得出，(b) 对于 $i \neq j$ 由 $[v_i, v_j] = 0$ 得出，且 Λ' 是一个对角矩阵。这表明 V 对角化了 $B \overset{\text{定义}}{=} (AJ_D)^\top J_D(AJ_D)$，即 $V^\top(AJ_D)^\top J_D(AJ_D)V$ 是一个对角矩阵。然而，B 是一个仅有一个负特征值 [47] 的对称矩阵。因此，不失一般性，Λ' 的第一个对角元素为负。 □

由引理 1 可得 $V^{-1} = J_D V^\top J_D$ 和 $A = AJ_D V(J_D V)^{-1} = VJ_D\Lambda V^{-1}J_D = VJ_D\Lambda(J_D V^\top J_D)J_D = VJ_D\Lambda J_D V^\top = V\Lambda V^\top$。

18) 定理*1*：考虑具有正交切向量 $h_1, \ldots, h_K$ 的 $\mathbb{S}_H^D$。对于 $x \in \mathbb{S}_H^D$，我们有 $P_H(x) = x$，$\|P_H(x)\|_2^2 = C^{-1}$，以及 $\|\mathcal{Q}(x)\|_2^2 = C^{-1}$，即 $\mathcal{Q}(x) \in \mathbb{S}^K$，且 $\mathcal{Q}$ 是从 $\mathbb{S}_H^D$ 到 $\mathbb{S}^K$ 的映射；参见命题2的证明。我们还有 $x = P_H\ x = C\langle x\ p\rangle p + C^{-1}\sum_{k\in[K]}\langle x\ h_k\rangle h_k = (\ )\ ,(\ )$ 对所有 $\mathcal{Q}^{-1}\circ\mathcal{Q}\ x$ $x \in \mathbb{S}_H^D$ $\mathcal{Q}^{-1}$ 都成立。因此，是$\mathcal{Q}$ 的逆映射——一个双射。最后，$\mathcal{Q}$ 是 $\mathbb{S}_H^D$ 与$\mathbb{S}^K$之间的等距映射，因为对于所有 $x_1, x_2 \in \mathbb{S}_H^D$，$d(x_1, x_2) = d(\mathcal{Q}(x_1)\ \mathcal{Q}(x_2))$。

19 *2* $\sum_{k\in[K']}h'^\top_k C_x h'_k \geq$*)* 定 $\sum_{k\in[K']}C\lambda_{D+1-k}\ C_x$ $\lambda_d\ C_x$ d 理：我们有 ( )，其中 ( ) 是 C_x 的第 -大特征值。若令 $h'_k = C^{\frac{1}{2}}v_{D+1-k}(C_x)$，其中 $k \in [K']$，即可达到下界。最优基点是张成空间 $\{h'_1, \ldots, h'_{K'}\}^\perp$ 中任意范数为 $C^{-\frac{1}{2}}$ 的向量，即 $p = C^{-\frac{1}{2}}v_1(C_x)$ 允许嵌套仿射子空间。

20）定理 3：考虑具有正交切向量 h_1、……、h_K 的 $\mathbb{H}_H^D$。对于 $x \in \mathbb{H}_H^D$，我们有 $P_H(x) = x$，$[P_H(x)\ P_H(x)] = C^{-1}$，且 $[\mathcal{Q}(x)\ \mathcal{Q}(x)] = C^{-1}$，即 $\mathcal{Q}(x) \in \mathbb{H}^K$，并且 $\mathcal{Q}$ 是从 $\mathbb{H}_H^D$ 到 $\mathbb{H}^K$ 的映射；参见命题 6 的证明。我们还有，对于所有 $x \in \mathbb{S}_H^D$，$x = P_H(x) = C[x, p]p - C^{-1}\sum_{k\in[K]}[x, h_k]h_k = \mathcal{Q}^{-1}\circ\mathcal{Q}(x)$。因此，$\mathcal{Q}^{-1}$ 是 $\mathcal{Q}$ 的逆映射——一个双射。最后，$\mathcal{Q}$ 是 $\mathbb{H}_H^D$ 与 $\mathbb{S}^K$ 之间的等距映射，因为对于所有 x_1、$x_2 \in \mathbb{H}_H^D$，$d(x_1, x_2) = d(\mathcal{Q}(x_1)\ \mathcal{Q}(x_2))$。

21）定理 4：不失一般性，我们对 $\{h'_{k'}\}_{k'}$ 和 H'定义 $[h'_1,\dots,h'_{K'}] \in \mathbb{R}^{(D+1)\times K'}$ $[h'_i,h'_j] = \delta_{ij}$ 进行缩放，使得，，即 $H'^{\top} J_D H' = I_{K'}$。代价函数为：

$$\begin{aligned}\mathrm{cost}(\mathbb{H}_H^D|\mathcal{X}) &= \sum_{k\in[K']} h'^{\top}_k J_D C_x J_D h'_k \\ &= \mathrm{Tr}\{H'^{\top} J_D C_x J_D H'\} = \mathrm{Tr}\{H'^{\top} J_D V \Lambda V^{\top} J_D H'\} \\ &= \mathrm{Tr}\{W^{\top}\Lambda W\} \overset{(a)}{=} \mathrm{Tr}\{WW^{\top} J_D \Lambda J_D\} = \mathrm{Tr}\{\mathcal{W}\Lambda J_D\},\end{aligned}$$

其中 $C_x = V\Lambda V^{\top}$ 可由 J_D对角化，$W = V^{\top} J_D H' \in \mathbb{R}^{(D+1)\times K'}$，（a）由 Λ 为对角矩阵，即 $\Lambda = J_D \Lambda J_D$，以及 $\mathcal{W}$ 定义= $WW^{\top} J_D$得出。

引理 2。 $W^{\top} J_D W = I_{K'}$。

证明。 $W^{\top} J_D W = H'^{\top} J_D (V J_D V^{\top}) J_D H' \overset{(a)}{=} H'^{\top} J_D H' = I_{K'}$，其中（a）由 $V J_D V^{\top} = J_D$得出。之所以如此，是因为根据定义，我们有 $V^{\top} J_D V = J_D$，或者$J_D V^{\top} J_D V_D = I_{D+1}$，即 $J_D V^{\top} J_D = V^{-1}$。因此，我们有 $V J_D V^{\top} = J_D$。最后，$H'^{\top} J_D H' = I_{K'}$是基向量 h'_1，……，$h'_{K'}$正交性的直接结果。□

我们将代价函数写成如下形式：

$$\mathrm{cost}(\mathbb{H}_H^D|\mathcal{X}) = \mathrm{Tr}\{\mathcal{W}\Lambda J_D\} = -\mathcal{W}_{11}\lambda_1 + \sum_{d=2}^{D+1} \mathcal{W}_{d,d}\lambda_d$$

其中 $\sum_{d=1}^{D+1} \mathcal{W}_{d\,d} = -\sum_{d=2}^{D+1} \mathcal{W}_{d\,d} + K'$ $\{W^{\top} J_D W\} = \{I_{K'}\}$ $W \in \mathbb{R}^{(D+1)\times K'}$ $\mathcal{W}_{11} =$，Tr Tr，即，。令其如下：

$$W = \begin{bmatrix} \sqrt{\|w_1\|_2^2 - 1} & \dots & \sqrt{\|w_{K'}\|_2^2 - 1} \\ w_1 & \dots & w_{K'} \end{bmatrix}, \quad (17)$$

对于范数大于或等于 1 的向量 w_1、……、$w_{K'} \in \mathbb{R}^D$ ℓ_2——注意 $W^{\top} J_D W = I_{K'}$ 和引理 2。根据 $\mathcal{W} = WW^{\top} J_D$ 和等式 (17)，我们有：

$$\mathcal{W}_{11} = -1 \sum_{k\in[K']} (\|w_k\|^2 - 1) \le 0 : \|w_k\|_2 \ge 1, \forall k \in [K'].$$

对于 $d \ge 2$，$\mathcal{W}_{d,\,d}$是 W 的第 d 行的范数平方。令 $W_c \in \mathbb{R}^{D\times K'}$，其中其第 d 行等于 W 的第 $(d+1)$ 行。因此，我们有

$$\sum_{d=2}^{D+1} \mathcal{W}_{d,d} = \mathrm{Tr}\{W_c^{\top} W_c\} = \sum_{k=1}^{K'} \|w_k\|_2^2. \quad (18)$$

现在我们将代价函数简化如下：

$$\mathrm{cost}(\mathbb{H}_H^D) = \sum_{d=2}^{D+1} \mathcal{W}_{d,d}(\lambda_1 + \lambda_d) - K'D\lambda_1. \quad (19)$$

引理 3。 对于所有 $d \ge 2$，我们有 $\lambda_d + \lambda_1 \ge 0$。

证明。令 v_d 为 C_x的第 d个 J_D特征向量。对于所有情况，我们有 $\lambda_d = v_d^{\top} J_D(\ [v_d, v_d]\ \lambda_d v_d) = v_d^{\top} J_D C_x J_D v_d$ $= \mathrm{sgn}(\,)$。证毕。$N^{-1}\sum_{n\in[N]}[x_n, v_d]^2 \ge 0$ $d \ge 1$ □

根据引理 3、公式 (18) 以及 $\|w_k\|_2 \ge 1$ 对所有$k \in [K']$成立，公式 (19) 中代价函数的最小值仅在 $\sum_{d=2}^{D+1} \mathcal{W}_{d,\,d} = \sum_{k=1}^{K'} \|w_k\|_2^2 = K'$，即$\|w_1\|_2 = \cdots = \|w_{K'}\|_2 = 1$时取得。因此，我们有 $W =$

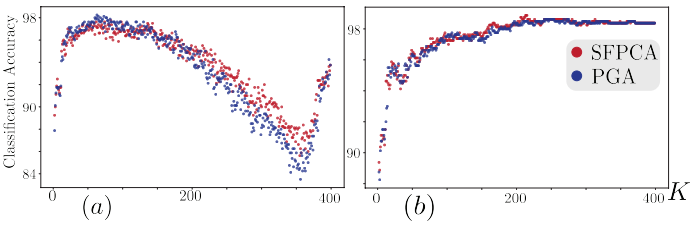


图 8. 在重构的组成数据上，PCA 降至维度 K 后的分类准确率（%）：(*a*) 随机森林和 (*b*) 神经网络。

$V^{\top} J_D H' = \begin{bmatrix} 0 & \cdots & 0 \\ w_1 & \cdots & w_{K'} \end{bmatrix}$。$V^{\top} J_D H'$的第一行对应于 $h'_1,\dots,h'_{K'}$与 V的第一列的洛伦兹内积，即 v_1。因此，我们有 $h'_1,\dots,h'_{K'} \in v_1^{\perp}$。由于 v_1 是 C_x唯一的负 J_D-特征向量，因此我们有 $p = v_1$。根据 $w_1,\dots,w_{K'}$的单位范数约束S，我们有 $W^{\top} W = I_{K'}$，即 W 是一个切片酉矩阵，且 $WW^{\top}$的特征值为 0 或 1。代价函数 cost($\mathbb{H}_H^D|\mathcal{X}$) $= \mathrm{Tr}\{W^{\top}\Lambda W\} = \mathrm{Tr}\{\Lambda WW^{\top}\}$当且仅当 $WW^{\top}$ 的非零奇异值与 Λ 中最小的 K' 个对角元素对齐时达到最小值——由冯·诺依曼迹不等式 [36]可知。设 $\lambda_2 \le \lambda_3 \le \cdots$。若对于 $i = 2,\dots,K'+1$，$w_i = e_i$成立，则代价函数达到最小值，也就是说，$h'_1,\dots,h'_{K'}$是 C_x中与最小 J_D-特征值配对的 K' 个负 J_D-特征向量。

A. 补充实验

我们在新闻组数据集上进行了实验，以展示球面主成分分析对分类准确率的影响。使用随机森林和五层神经网络分类器，并按 90% 训练集、10% 测试集划分数据；平均准确度基于 20 次随机划分计算得出。如图 8 所示，在 K （目标维度）组实验中，SFPCA 的平均准确度优于 PGA。这两种方法都显著提升了随机森林的性能，并适度提升了神经网络的性能。这可能归因于神经网络的去噪能力。当 K 较大时，随机森林的准确率会下降，而神经网络则不会。这可能是因为新闻组数据集具有较高的稀疏性。舍弃二阶矩矩阵中较小的特征值会显著改变数据的稀疏性，从而对随机森林的准确率产生不利影响。

参考文献

[1] R. ANIRUDH、P. TURAGA、J. SU 和 A. SRIVASTAVA，人体动作的弹性函数编码：从向量场到潜变量，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第3147–3155页。[2] C. BISHOP，贝叶斯主成分分析，《神经信息处理系统进展》，11（1998）。[3] T. T. CAI、Z. MA 和 Y. WU，稀疏主成分分析：最优速率与自适应估计，《统计年鉴》，41（2013），第3074–3110页。[4] R. CHAKRABORTY、D. SEO 和 B. C. VEMURI，一种用于常曲率流形的高效精确 *PGA* 算法，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第3976–3984页。[5] I. CHAMI、A. GU、D. P. NGUYEN 和 C. R´E，*HoroPCA*：通过 *horosphere* 投影进行双曲降维，载于国际机器学习会议，PMLR，2021年，第1419–1429页。

[6] E. S. CHARLSON、J. CHEN、R. CUSTERS-ALLEN、K. BITTINGER、H. LI、R. SINHA、J. HWANG、F. D. BUSHMAN 和 R. G. COLLMAN，吸烟者上呼吸道中的微生物群落失调，《PLOS ONE》，5（2010），第e15216页。[7] E. CHIEN、C. PAN、P. TABAGHI 和 O. MILENKOVIC，在庞加莱球中的高可扩展性且可证明准确的分类，载于2021年IEEE国际数据挖掘会议，IEEE，2021年，第61–70页。[8] E. CHIEN、P. TABAGHI 和 O. MILENKOVIC，Hyperaid：用于树拟合和层次聚类的双曲空间去噪，载于第28届ACM SIGKDD知识发现与数据挖掘会议论文集，2022年，第201–211页。[9] A. COMTE、T. TRICOU、E. TANNIER、J. JOSEPH、A. SIBERCHICOT、S. PENEL、R. ALLIO、F. DELSUC、S. DRAY 和 D. M. DE VIENNE，PhylteR：高效识别系统发育基因组数据集中的离群序列，《分子生物学与进化》，40（2023），第msad234页。[10] X. DAI 和 H.-G. M ¨ULLER，黎曼流形和球面上函数型数据的主成分分析，《统计年鉴》，46（2018），第3334–3361页。[11] G. F. ESTABROOK、F. R. MCMORRIS 和 C. A. MEACHAM，基于四个进化单元子树的无向系统发育树比较，《系统生物学》，34（1985），第193–200页。[12] P. T. FLETCHER 和 S. JOSHI，对称空间上的主测地线分析：扩散张量的统计学，载于《计算机视觉与医学和生物医学图像分析数学方法》，施普林格，2004年，第87–98页。[13] P. T. FLETCHER、C. LU 和 S. JOSHI，李群上的形状统计：主测地线分析，载于2003年IEEE计算机学会计算机视觉与模式识别会议论文集，第1卷，IEEE，2003年，第I–I页。[14] P. T. FLETCHER、C. LU、S. M. PIZER 和 S. JOSHI，用于研究形状非线性统计的主测地线分析，《IEEE医学影像汇刊》，23（2004），第995–1005页。[15] J. GALLIER 和 J. QUAINTANCE，《微分几何与李群：计算视角》，第12卷，Springer Nature，2020年。[16] A. GU、F. SALA、B. GUNEL 和 C. R´E，在乘积空间中学习混合曲率表征，载于国际学习表征会议，2018年。[17] Y. GUAN 和 J. DY，稀疏概率主成分分析，载于人工智能与统计，PMLR，2009年，第185–192页。[18] N. J. HIGHAM，J-正交矩阵：性质与生成，《SIAM评论》，45（2003），第504–519页。[19] H. HOTELLING，将一组统计变量分解为主成分的分析，《教育心理学杂志》，24（1933），第417页。[20] S. HUCKEMANN、T. HOTZ 和 A. MUNK，内蕴形状分析：模去等距李群作用的黎曼流形上的测地线主成分分析，《中华统计学志》，（2010），第1–58页。[21] S. HUCKEMANN 和 H. ZIEZOLD，黎曼流形上的主成分分析及其在三角形状空间中的应用，《应用概率进展》，38（2006），第299–319页。[22] Y. JIANG、P. TABAGHI 和 S. MIRARAB，用于系统发育树定位与更新的双曲嵌入学习，《生物学》，11（2022），第1256页。[23] I. T. JOLLIFFE，《主成分分析》，施普林格，2002年。[24] I. T. JOLLIFFE、N. T. TRENDAFILOV 和 M. UDDIN，一种基于套索回归的改进主成分技术，《计算与图形统计杂志》，12（2003），第531–547页。[25] S. JUNG、I. L. DRYDEN 和 J. S. MARRON，主嵌套球面分析，《生物统计学》，99（2012），第551–568页。[26] A. KLIMOVSKAIA、D. LOPEZ-PAZ、L. BOTTOU 和 M. NICKEL，用庞加莱映射分析单细胞数据中的复杂层级结构，《自然通讯》，11（2020），第1–9页。[27] A. LAHAV 和 R. TALMON，半正定矩阵流形上的普罗克鲁斯特斯分析及数据集对齐，《IEEE信号处理汇刊》，（2023）。[28] L. LAHTI、J. SALOJ ¨ARVI、A. SALONEN、M. SCHEFFER 和 W. M. DE VOS，人体肠道生态系统中的临界要素，《自然通讯》，5（2014），第4344页。[29] N. LAWRENCE，用于高维数据可视化的高斯过程潜变量模型，《神经信息处理系统进展》，16（2003）。[30] D. LAZAR 和 L. LIN，主测地线分析中的尺度与曲率效应，《多元分析杂志》，153（2017），第64–82页。[31] J. M. LEE，《黎曼流形：曲率导论》，第176卷，Springer Science & Business Media，2006年。

[32] K. LIU、Q. LI、H. WANG 和 G. TANG，《球面主成分分析》，载于《2019年SIAM国际数据挖掘会议论文集》，SIAM，2019年，第387–395页。[33] A. LOU、I. KATSMAN、Q. JIANG、S. BELONGIE、S.-N. LIM 和 C. DE SA，《对 Fr´echet 均值求导》，载于《国际机器学习会议》，PMLR，2020年，第6393–6403页。[34] U. MAI 和 S. MIRARAB，《TreeShrink：快速准确地检测系统发育树集合中的离群长枝》，《BMC基因组学》，19（2018），第272页。ISBN：9783319679785。[35] Y. MENG、J. HUANG、G. WANG、C. ZHANG、H. ZHUANG、L. KAPLAN 和 J. HAN，《球面文本嵌入》，《神经信息处理系统进展》，32（2019）。[36] L. MIRSKY，《约翰·冯·诺伊曼的一个迹不等式》，《数学月刊》，79（1975），第303–306页。[37] K. PEARSON，《LIII. 关于空间中点集的最佳拟合直线和平面》，《伦敦、爱丁堡和都柏林哲学杂志与科学期刊》，2（1901），第559–572页。[38] X. PENNEC，《流形上的重心子空间分析》，《统计年鉴》，46（2018），第2711–2746页。[39] I. U. RAHMAN、I. DRORI、V. C. STODDEN、D. L. DONOHO 和 P. SCHR ¨ODER，《流形值数据的多尺度表示》，《多尺度建模与仿真》，4（2005），第1201–1232页。[40] S. ROWEIS，《用于主成分分析和稀疏主成分分析的期望最大化算法》，《神经信息处理系统进展》，10（1997）。[41] I. SLAPNI ˇCAR 和 K. VESELI´C，《双曲特征向量矩阵条件数的一个界》，《线性代数及其应用》，290（1999），第247–255页。[42] S. SOMMER、F. LAUZE、S. HAUBERG 和 M. NIELSEN，《流形值统计、精确主测地线分析及线性近似的影响》，载于《计算机视觉—ECCV 2010：第11届欧洲计算机视觉会议，希腊克里特岛伊拉克利翁，2010年9月5日至11日，论文集，第VI卷》，施普林格，2010年，第43–56页。[43] S. SOMMER、F. LAUZE 和 M. NIELSEN，《精确主测地线分析的测地线优化》，《计算数学进展》，40（2014），第283–313页。[44] R. SONTHALIA 和 A. GILBERT，《树！我不是树！我是低维双曲嵌入》，《神经信息处理系统进展》，33（2020），第845–856页。[45] G. W. STEWART，《奇异值分解的早期历史》，《SIAM评论》，35（1993），第551–566页。[46] P. TABAGHI、E. CHIEN、C. PAN、J. PENG 和 O. MILENKOVI´C，《乘积空间形式中的线性分类器》，arXiv预印本 arXiv:2102.10204，（2021）。[47] P. TABAGHI 和 I. DOKMANI´C，《双曲距离矩阵》，载于《第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，2020年，第1728–1738页。[48] P. TABAGHI、J. PENG、O. MILENKOVIC 和 I. DOKMANI´C，《相似性比较的几何》，arXiv预印本 arXiv:2006.09858，（2020）。[49] L. L. THURSTONE，《多因子分析》，《心理学评论》，38（1931），第406页。[50] M. E. TIPPING 和 C. M. BISHOP，《概率主成分分析》，《皇家统计学会期刊：B辑（统计方法学）》，61（1999），第611–622页。[51] M. TOURNIER、X. WU、N. COURTY、E. ARNAUD 和 L. REVERET，《利用主测地线分析进行运动压缩》，载于《计算机图形论坛》，第28卷，Wiley在线图书馆，2009年，第355–364页。[52] R. VIDAL、Y. MA 和 S. SASTRY，《广义主成分分析（GPCA）》，《IEEE模式分析与机器智能汇刊》，27（2005），第1945–1959页。[53] T. WARNOW，《计算系统发育学：系统发育推断方法设计导论》，剑桥大学出版社，剑桥，英国，2017年。地址：英国剑桥。[54] N. J. WICKETT、S. MIRARAB、N. NGUYEN、T. WARNOW、E. J. CAR-PENTER、N. MATASCI、S. AYYAMPALAYAM、M. S. BARKER、J. G. BURLEIGH、M. A. GITZENDANNER、B. R. RUHFEL、E. WAFULA、J. P. DER、S. W. GRAHAM、S. MATHEWS、M. MELKONIAN、D. E. SOLTIS、P. S. SOLTIS、N. W. MILES、C. J. ROTHFELS、L. POKO-RNY、A. J. SHAW、L. DEGIRONIMO、D. W. STEVENSON、B. SUREK、J. C. VILLARREAL、B. ROURE、H. PHILIPPE、C. W. DEPAMPHILIS、T. CHEN、M. K. DEYHOLOS、R. S. BAUCOM、T. M. KUTCHAN、M. M. AUGUSTIN、J. J. WANG、Y. ZHANG、Z. TIAN、Z. YAN、X. WU、X. SUN、G. K.-S. WONG 和 J. J. LEEBENS-MACK，《陆生植物起源与早期分化的系统发育转录组学分析》，《美国国家科学院院刊》，111（2014），第E4859–4868页。arXiv：1301.5277v2 ISBN：0027-8424。

[55] S. WOLD、K. ESBENSEN 和 P. GELADI，主成分分析，《化学计量学与智能实验室系统》，2（1987），第37–52页。[56] H. XU、C. CARAMANIS 和 S. SANGHAVI，通过异常值追踪实现稳健主成分分析，《神经信息处理系统进展》，23（2010）。[57] C. ZHANG、M. RABIEE、E. SAYYARI 和 S. MIRARAB，*ASTRAL-III*：从部分解析的基因树中以多项式时间重建物种树，《BMC生物信息学》，19（2018），第153页。[58] Y. ZHOU 和 T. O. SHARPEE，基因表达的双曲几何，《iScience》，24（2021），第102225页。[59] H. ZOU、T. HASTIE 和 R. TIBSHIRANI，稀疏主成分分析，《计算与图形统计学杂志》，15（2006），第265–286页。