Visualizing Grassmannians via Poincare Embeddings

Huanran Li[1], Daniel Pimentel-Alarcón[2]

Department of Electrical Engineering[1] / Biostatistics[2], Wisconsin Institute for Discovery[1,2], University of Wisconsin - Madison[1,2]

Problem: Visualizing Subspaces

Subspaces are a cornerstone of data analysis, with applications ranging from linear regression to principal component analysis (PCA), low-rank matrix completion (LRMC), computer vision, recommender systems, classification, and more. However, **there exist few tools** to visualize the Grassmann manifold $\mathbb{G}(m,r)$ of r-dimensional subspaces of $\mathbb{R}^m$. In this paper, we propose a method that **visualizes a collection of points in the Grassmannian (subspaces)** through an embedding onto the Poincaré disk $\mathbb{D} \subset \mathbb{R}^2$.

Classic Misleading Method for Grassmannian

The most intuitive of such visualizations is the representation of $\mathbb{G}(3,1)$ as the **closed half-sphere** where each point in the hemisphere represents the 1-dimensional subspace (line) in $\mathbb{R}^3$ that crosses that point and the origin (see Figure 1). While intuitive, this visualization bears certain **limitations**:

1. This representation *wraps* around the edge, so geodesic distances can be deceiving. For instance, two points (subspaces) that may appear **diametrically far** may in fact be **arbitrarily close** (see Figure 1).
2. But more importantly, the main caveat of this semi-sphere representation is that it is unclear how to generalize it to $m > 3$ or $r > 1$, which makes it quite restrictive, specially for analysis of modern **high-dimensional data**.

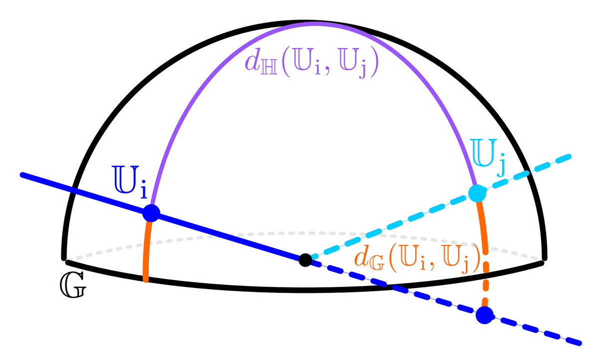

Fig. 1: Classical 3D Representation of the Grassmannian $\mathbb{G}(3,1)$.

Poincaré Disk

The Poincaré disk is a **2D hyperbolic geometric model**, usually displayed as a unit circle where the geodesic distance between two points in the disk is represented as the circular arc orthogonal to the unit circle, which corresponds to the projection of the hyperbolic arc of their geodesic. This unique feature brings several advantages:

1. It can accurately represent the global structure of **complex hierarchical data** while retaining its **local structures**. Since its hyperbolic arcs get **larger (tending to infinity)** as points approach the disk boundary, the disk can be viewed as a continuous embedding of tree nodes from the top of the **tree structure**.
2. It has a *Riemannian* manifold structure that allow us to perform **gradient-based optimization**, which is crucial to derive convergence guarantees, and for parallel training of large-scale dataset models.

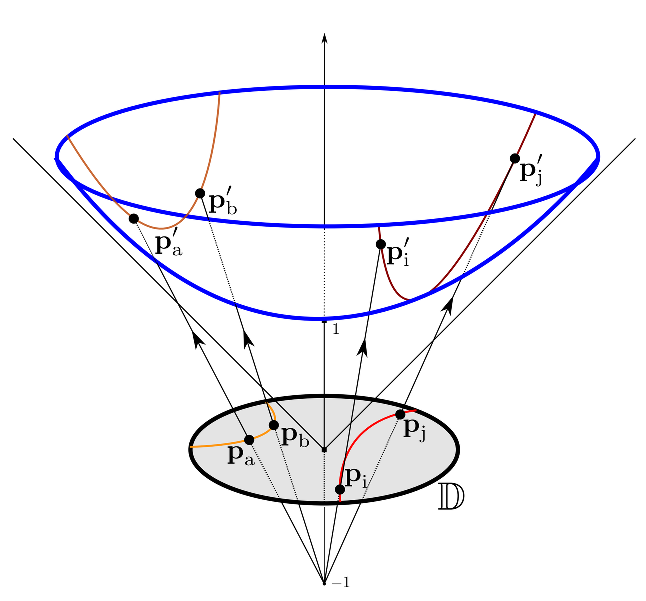

Fig. 2: Geodesics in the Poincaré disk $\mathbb{D}$.

Our Method: GrassCaré

Figure 3 is a demonstration of how GrassCaré can be applied to visualize high-dimensional object trajectory in the motion segmentation task.

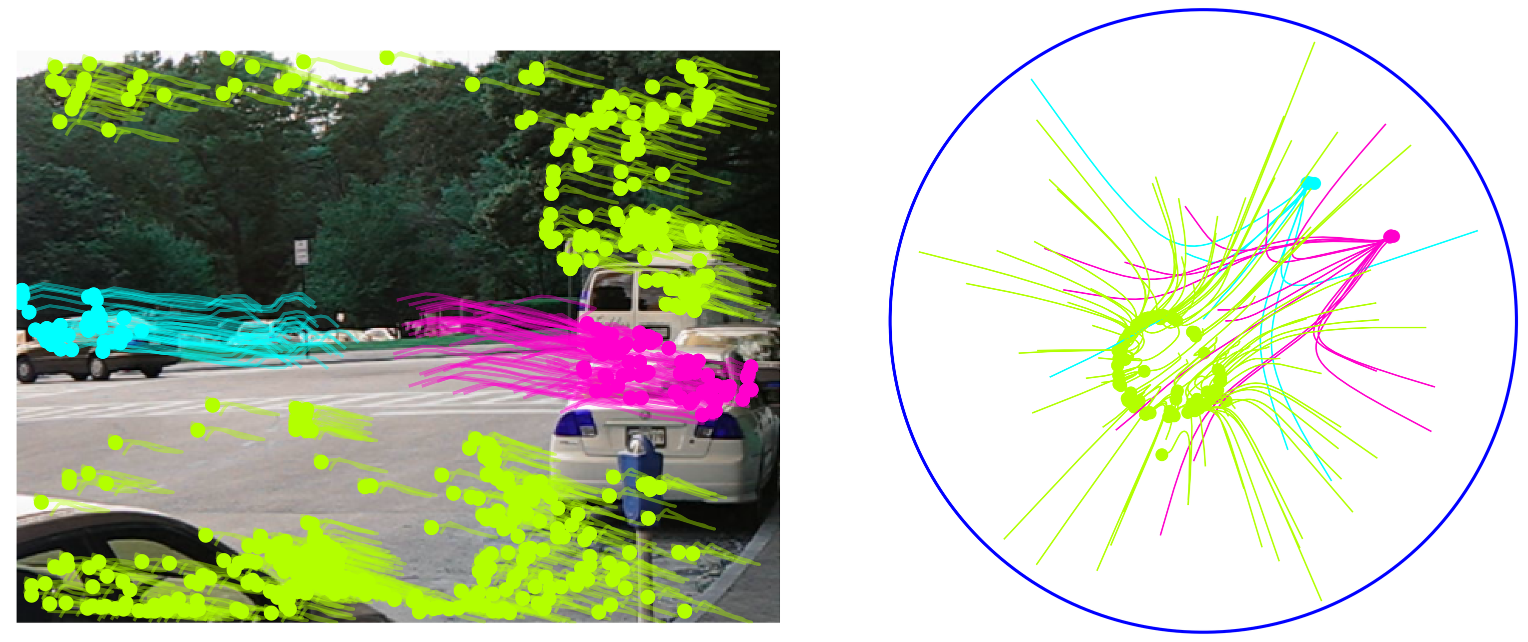

Fig. 3: GrassCaré on Motion Segmentation Dataset (Hopkins155). **Left:** Video of 2 objects (blue, fuchsia) and backgrounds (green). **Right:** GrassCaré Optimization Process (Random Initialization to Final embeddings).

To accomplish that, the embeddings are calculated with those three steps:

1. Compute a **Grassmannian probability matrix** $P_{\mathbb{G}} \in [0,1]^{N\times N}$ whose $(i,j)^{th}$ entry represents the probability that subspace $\mathbb{U}_i$ is chosen as a nearest neighbor of $\mathbb{U}_j$:
$$[P_{\mathbb{G}}]_{ij} := \frac{1}{2N}\frac{\exp(-d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_j)^2/2\gamma_i^2)}{\sum_{k\neq i}\exp(-d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_k)^2/2\gamma_i^2)} + \frac{1}{2N}\frac{\exp(-d_{\mathbb{G}}(\mathbb{U}_j,\mathbb{U}_i)^2/2\gamma_j^2)}{\sum_{k\neq j}\exp(-d_{\mathbb{G}}(\mathbb{U}_j,\mathbb{U}_k)^2/2\gamma_j^2)},$$
where $d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_j)$ denotes the *geodesic distance* on Grassmannian, and γ_i is the variance of distances from point i to other points.
2. Create the **Poincaré probability matrix** $P_{\mathbb{D}} \in [0,1]^{N\times N}$, whose $(i,j)^{th}$ entry represents the probability that embedding point p_i is chosen as a nearest neighbor of p_j:
$$[P_{\mathbb{D}}]_{ij} := \frac{\exp(-d_{\mathbb{D}}(p_i,p_j)^2)}{\sum_{k\neq l}\exp(-d_{\mathbb{D}}(p_k,p_l)^2)}$$
where $d_{\mathbb{D}}(p_i,p_j)$ denotes the distance between p_i, p_j on the Poincaré disk.
3. **Maximize the similarity** between the two distributions $P_{\mathbb{G}}$ and $P_{\mathbb{D}}$, which we do by minimizing their *Kullback-Leibler* (KL) divergence:
$$\min_{p_1,\dots,p_n} \mathrm{KL}(P_{\mathbb{G}}\|P_{\mathbb{D}}) = \min_{p_1,\dots,p_n} \sum_{i,j}[P_{\mathbb{G}}]_{ij}\log\frac{[P_{\mathbb{G}}]_{ij}}{[P_{\mathbb{D}}]_{ij}}.$$

Theoretical Result

Theorem 1. Suppose $N > 3$. Define $\gamma := \min_i \gamma_i$ and $\Gamma := \max_i \gamma_i$. Let $\{\mathcal{U}_1,\dots,\mathcal{U}_K\}$ be a cluster partition of $\{\mathbb{U}_1,\dots,\mathbb{U}_N\}$ such that $|\mathcal{U}_k| \geq n_K > 1\ \forall$ k. Let
$$\delta := \frac{1}{\sqrt{2}\gamma}\max_k \max_{\mathbb{U}_i,\mathbb{U}_j\in\mathcal{U}_k} d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_j), \quad \Delta := \frac{1}{\sqrt{2}\Gamma}\min_{\substack{\mathbb{U}_i\in\mathcal{U}_k,\mathbb{U}_j\in\mathcal{U}_\ell:\\ k\neq\ell}} d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_j).$$
Then the **optimal loss** of GrassCaré is upper-bounded by:
$$\mathcal{L}^\star < \log D + \frac{5e^{\delta^2-\Delta^2}}{n_K-1},$$
where
$$D := N(n_K-1) + N(N-n_K)\cdot\exp\left(-\mathrm{arcosh}^2\left(1+\frac{2\sin(\pi/K)}{0.75^2}\right)\right).$$

Comparison

The **effectiveness** of a visualization can be evaluated using the concept of *representation error*, which is calculated as the Frobenius difference between the geodesic distance of the **subspaces** and the distance of corresponding points in the **embedding**. In the case of GrassCaré, distances in the embedding are measured according to the Poincaré geodesics, so the representation error of the GrassCaré embedding will be measured as:
$$\varepsilon^2(\mathbb{D}) = \sum_{i,j}\left(\frac{d_{\mathbb{G}}(\mathbb{U}_i,\mathbb{U}_j)}{Z_{\mathbb{G}}} - \frac{d_{\mathbb{D}}(p_i,p_j)}{Z_{\mathbb{D}}}\right)^2,$$
where $Z_{\mathbb{G}}^2 = \sum_{i,j} d_{\mathbb{G}}^2(\mathbb{U}_i,\mathbb{U}_j)$ and $Z_{\mathbb{D}}^2 = \sum_{i,j} d_{\mathbb{D}}^2(p_i,p_j)$ are normalization terms. Likewise, in all other methods where the objective is to minimize the Euclidean distance, $d_{\mathbb{D}}(p_i,p_j)$ is substituted with $\|v_i - v_j\|_2$. We generate 50 subspaces in 3 clusters with different values of **ambient dimension** m and **rank** r. The results of 100 trials are summarized in Figure 4, which confirms the superiority of our GrassCaré embedding.

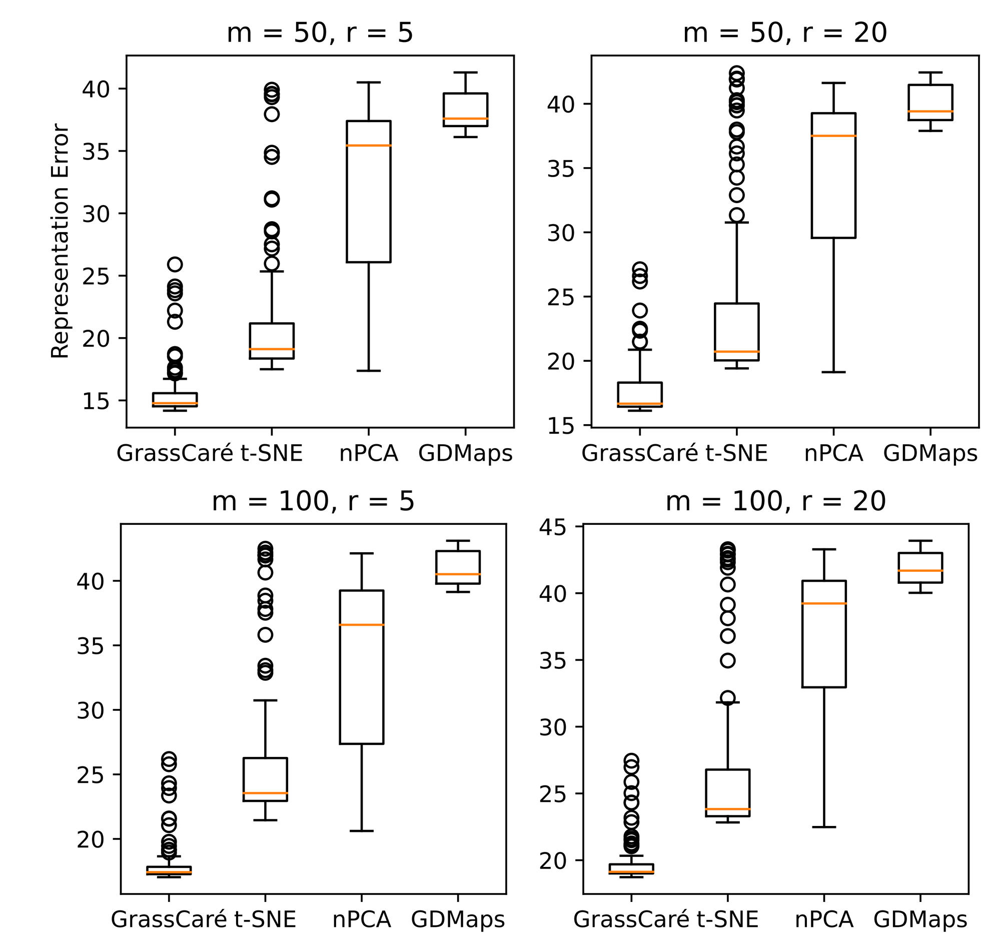

Fig. 4: Representation error of GrassCaré and other methods for high-dimensional Grassmannians $\mathbb{G}(m,r)$.

Remarks

We anticipate that **GrassCaré** will prove to be an effective tool for visualizing subspaces obtained from real-world data in high dimensions. In our experiments, we observed that GrassCaré is slightly slower than t-SNE and GDMaps, due to the additional calculations involved in computing distances in the Poincaré disk compared to Euclidean space. However, we believe that this added computation time is a reasonable trade-off for two reasons:

1. Our experiments have demonstrated that GrassCaré produces a **more accurate** visual representation. Moreover, its theoretical lower bound reinforces our belief in its accuracy.
2. GrassCaré makes better use of the unit circle's space, which eliminates the visual distortion caused by the different axis scales in GDMaps. This leads to a **more reliable and informative** visualization.

Publication & Award

The present study has been published and honored with the **"Best Student Paper Award"** at the 14th International Conference on Information Visualization Theory and Application (IVAPP 2023).