



EucApprox - Euclidean Embedding of Data Using Local Distances

Dimitris Arabadjis

Department of Statistics & Actuarial – Financial Mathematics
IDEALab

Motivation

Approximate the optimal Euclidean embedding functionally, using only local distance data and local operators

Motivation

Approximate the optimal Euclidean embedding functionally, using only local distance data and local operators

	Local Distances	Global Distances	Data Used	Data Processed
Global Methods	✓	✓	Dense	Dense / 1 step

Motivation

Approximate the optimal Euclidean embedding functionally, using only local distance data and local operators

	Local Distances	Global Distances	Data Used	Data Processed
Global Methods	✓	✓	Dense	Dense / 1 step
Local Methods	✓	✗	Sparse	Sparse / 1 step

Motivation

Approximate the optimal Euclidean embedding functionally, using only local distance data and local operators

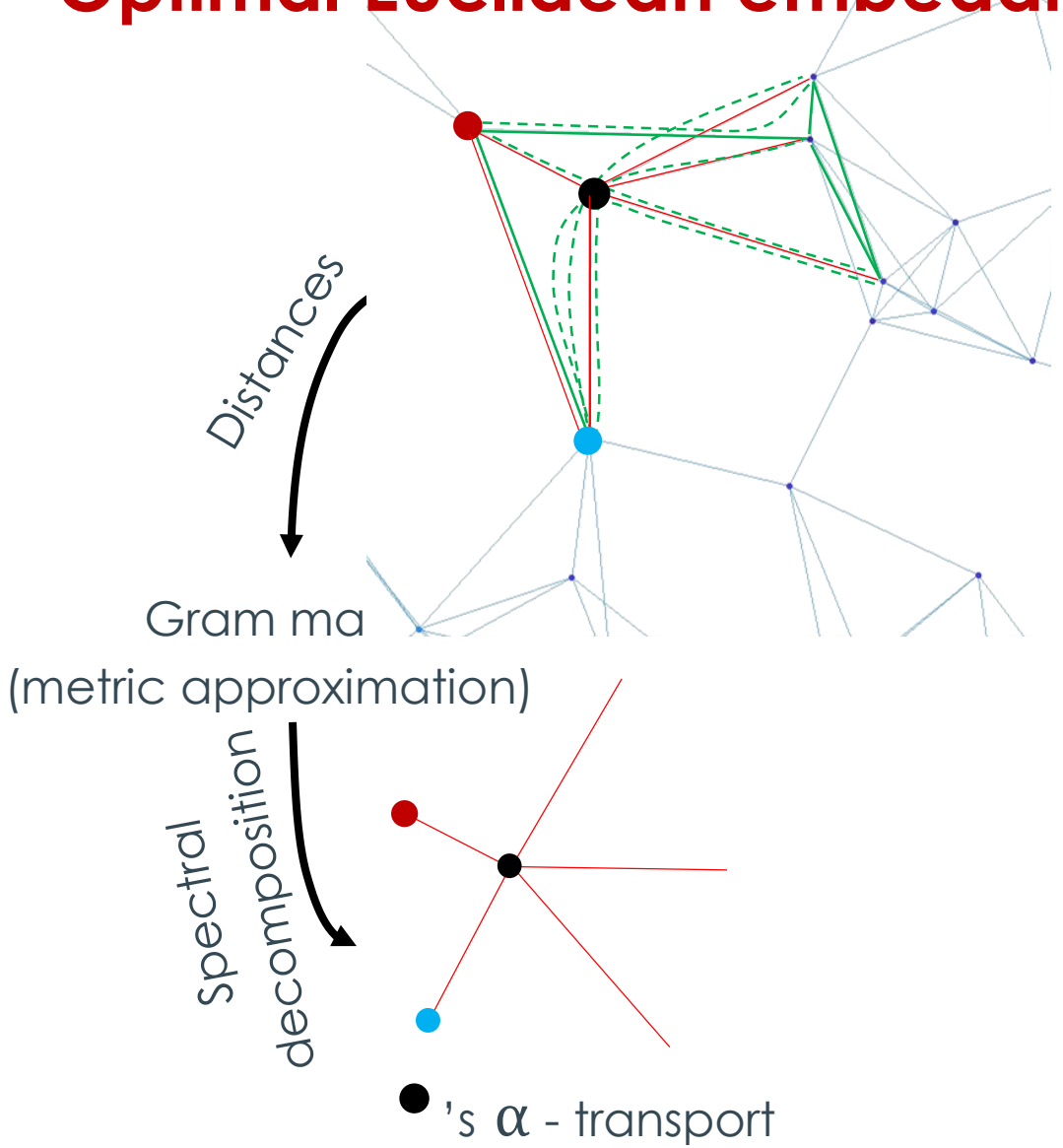
	Local Distances	Global Distances	Data Used	Data Processed	Optimization
Global Methods	✓	✓	Dense	Dense / 1 step	Pointwise
Local Methods	✓	✗	Sparse	Sparse / 1 step	Pointwise

Motivation

Approximate the optimal Euclidean embedding functionally, using only local distance data and local operators

	Local Distances	Global Distances	Data Used	Data Processed	Optimization
Global Methods	✓	✓	Dense	Dense / 1 step	Pointwise
Local Methods	✓	✗	Sparse	Sparse / 1 step	Pointwise
EucApprox	✓	✓	Sparse	Sparse / iterative	Functional

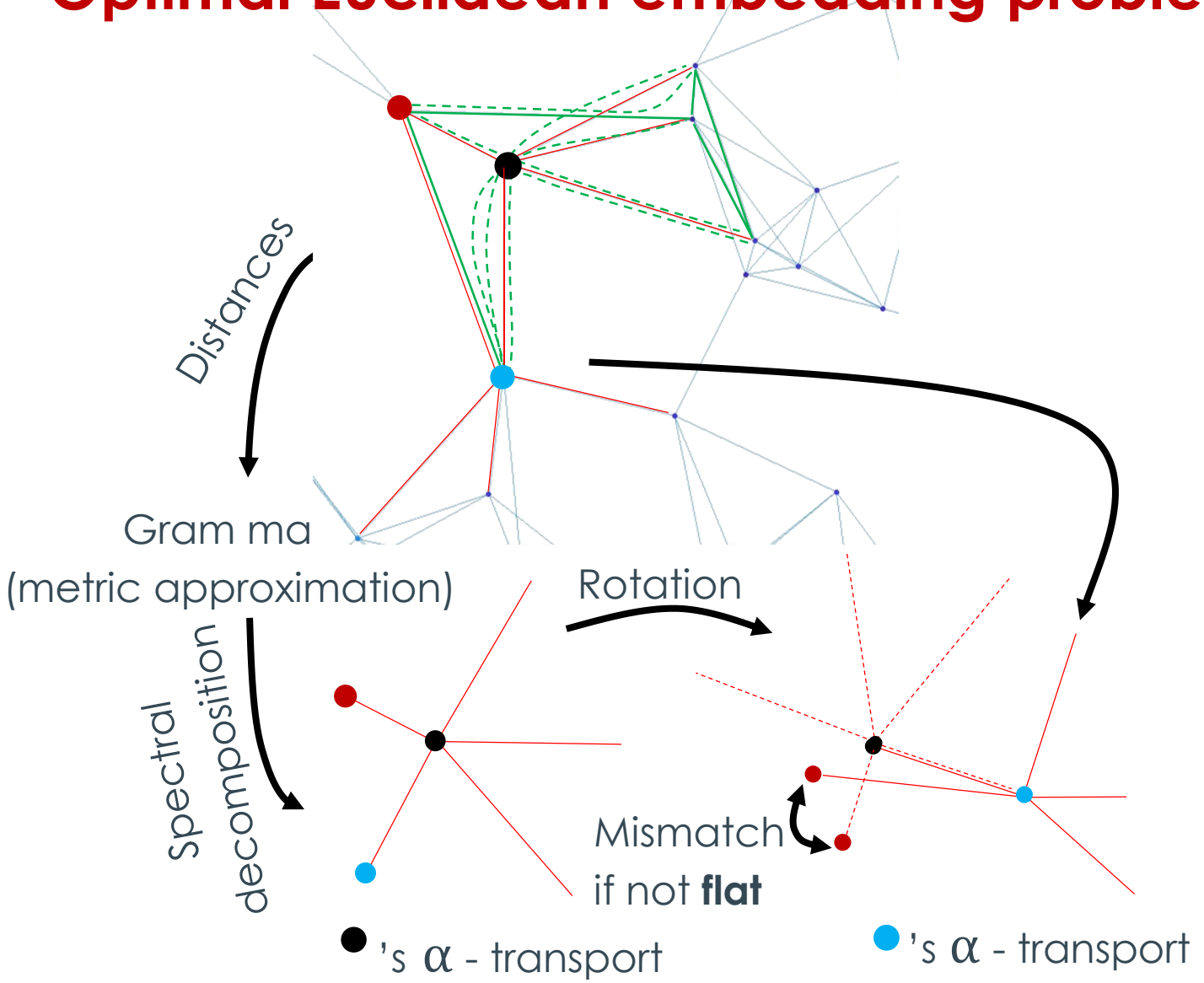
Optimal Euclidean embedding problem



Metric trivializing directions α: Infinitesimal transports that turn the metric into the identity matrix (spectral decomposition of the metric).

Neighbors
Neighborhood's subgraph
Shortest paths

Optimal Euclidean embedding problem

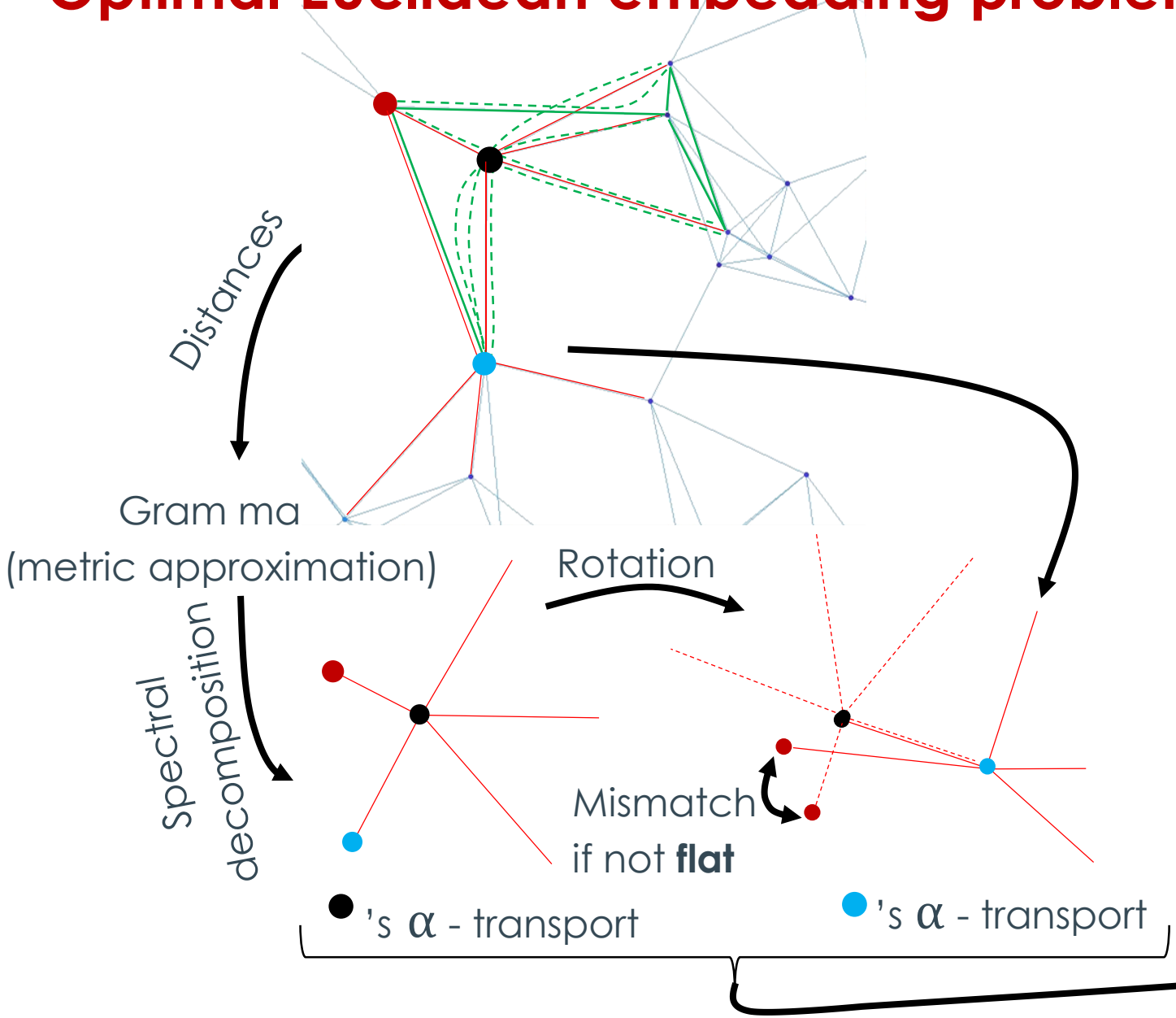


Metric trivializing directions α: Infinitesimal transports that turn the metric into the identity matrix (spectral decomposition of the metric).

Only flat manifolds allow to isometrically deform α into integrable entities

Neighbors
Neighborhood's subgraph
Shortest paths

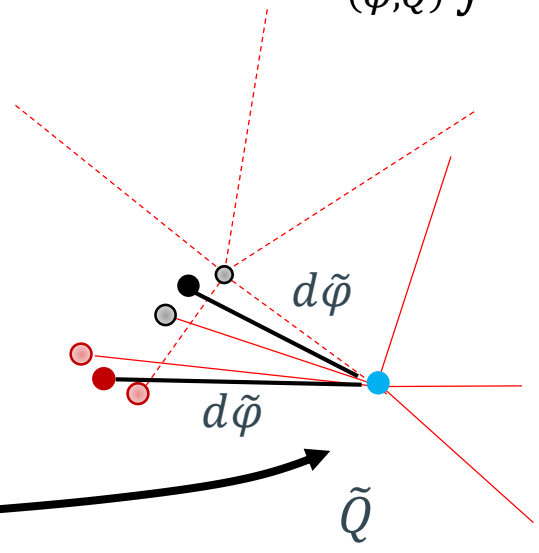
Optimal Euclidean embedding problem



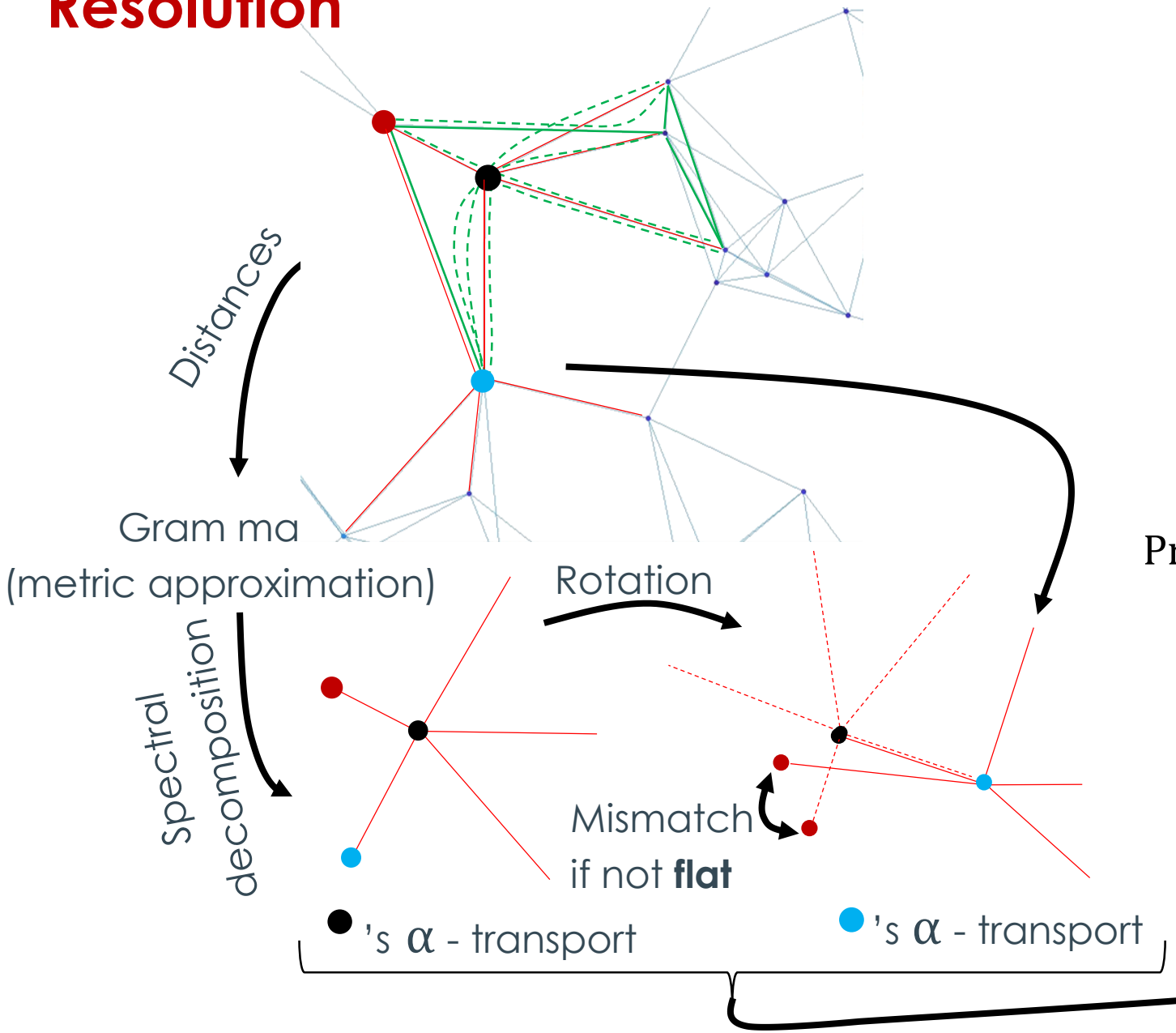
Core idea: Determine the closest integrable entities to the isometries of α.

n – dimensional optimal Euclidean embedding: Find **functions** $\tilde{\varphi} \in \mathbb{R}^N$ and $\tilde{Q} \in O(N)$ so that

$$(\tilde{\varphi}, \tilde{Q}) = \arg\min_{(\varphi,Q)} \int \|d\varphi - \alpha Q\|^2$$



Resolution



Euler-Lagrange equations :

$$\Delta\tilde{\varphi} = \mathrm{div}\left(\alpha\tilde{Q}\right), \qquad \iota_{\partial\Omega}(d\tilde{\varphi}) = \iota_{\partial\Omega}\left(\alpha\tilde{Q}\right)$$

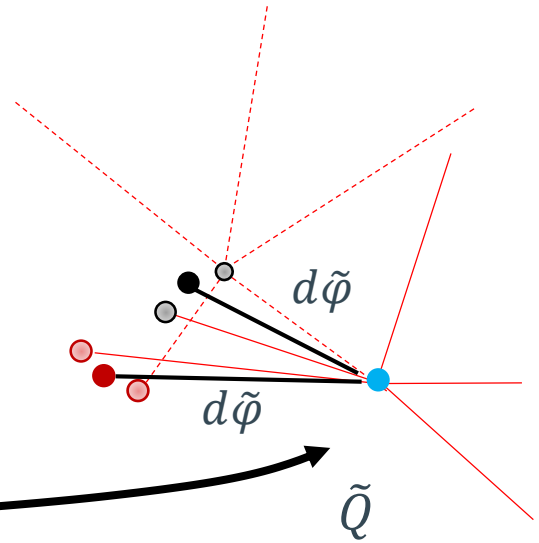
$$\tilde{Q} = \mathrm{Procrustes}(d\tilde{\varphi}, \alpha)$$

Δ – Laplacian

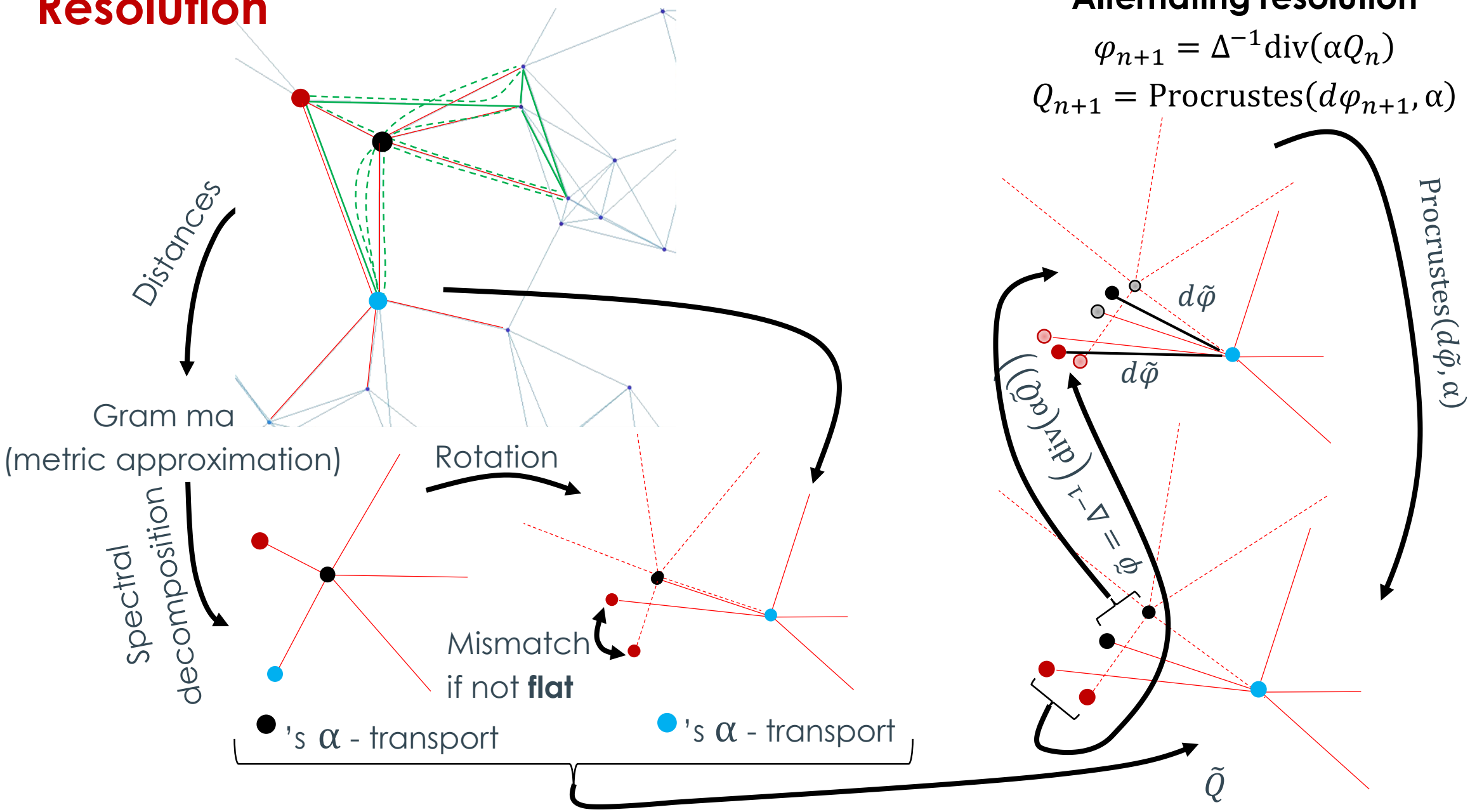
div – divergence

$\iota_{\partial\Omega}(\cdot)$ - inner product with the manifold's boundary normal

$\mathrm{Procrustes}$ – orthogonal Procrustes alignment



Resolution



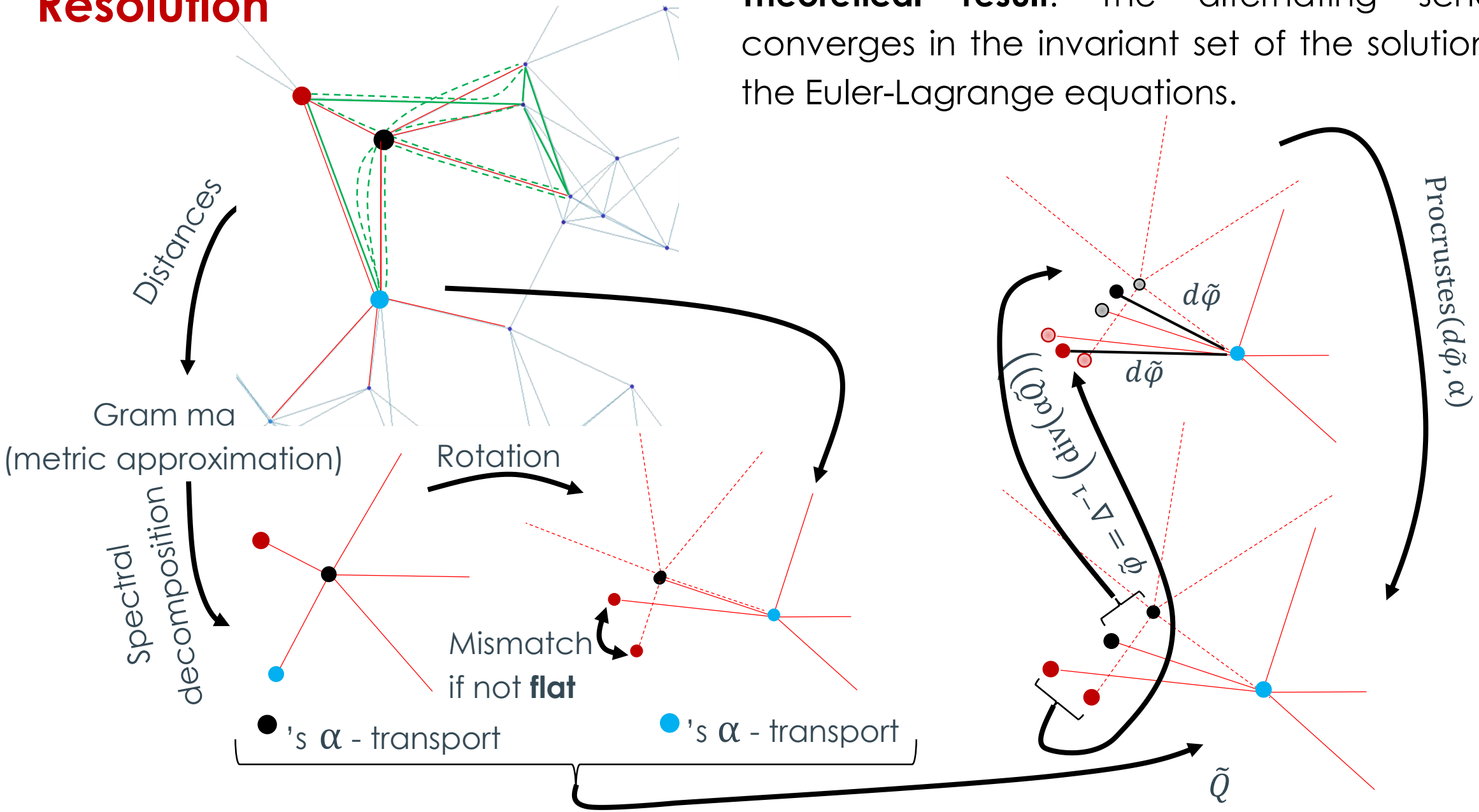
Alternating resolution

$$\varphi_{n+1} = \Delta^{-1}\mathrm{div}(\alpha Q_n)$$

$$Q_{n+1} = \mathrm{Procrustes}(d\varphi_{n+1}, \alpha)$$

Resolution

Theoretical result: The alternating scheme converges in the invariant set of the solutions of the Euler-Lagrange equations.



Implementation on distance graphs

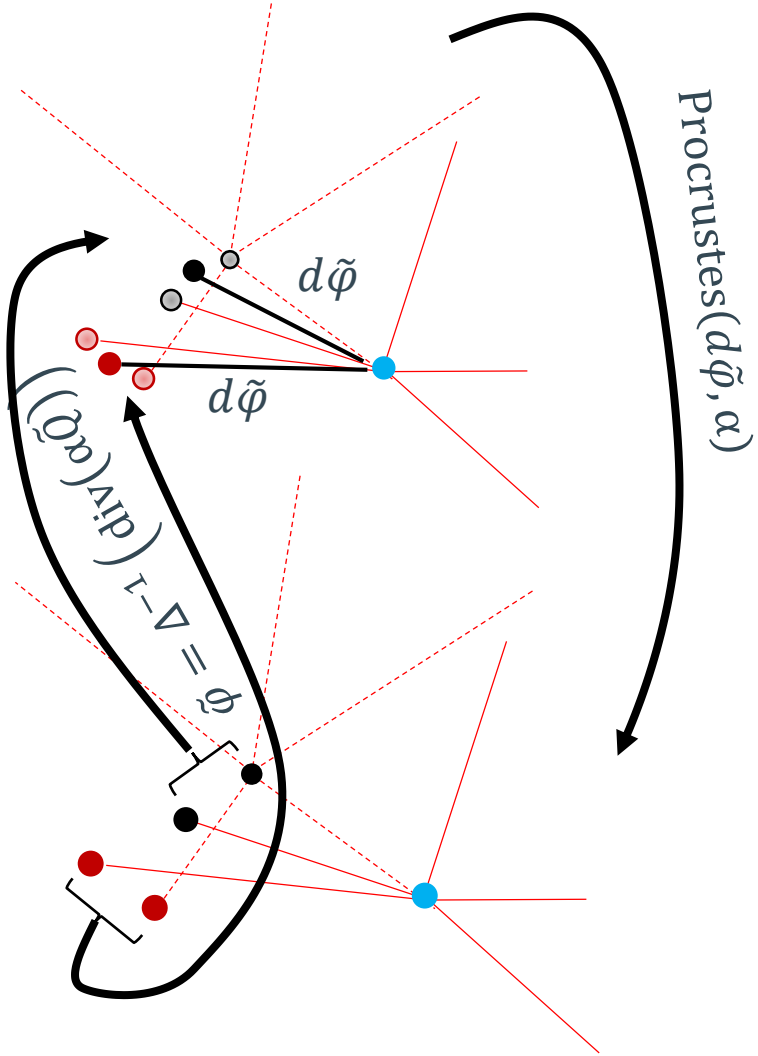
Differential 1-forms : Linear combinations of neighborhoods' barycentric coordinates' differentials.

- **Scalar function differentials** : 1-forms with coefficients the function's differences on the graph's edges (2nd order approximation due to the barycentric constraint)
  - **Inner products of 1-forms**: The inner product of barycentric coordinates' differentials is the volume form – scaled metric pseudo-inverse.
    - **Divergence**: Inner product – dual of scalar function differential.
      - **Laplacian**: Divergence of scalar function differential

Alternating resolution

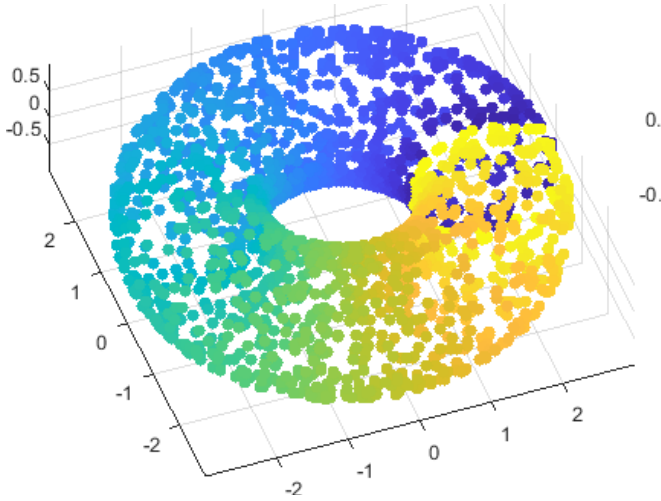
$$\varphi_{n+1} = \Delta^{-1}\mathrm{div}(\alpha Q_n)$$

$$Q_{n+1} = \mathrm{Procrustes}(d\varphi_{n+1}, \alpha)$$

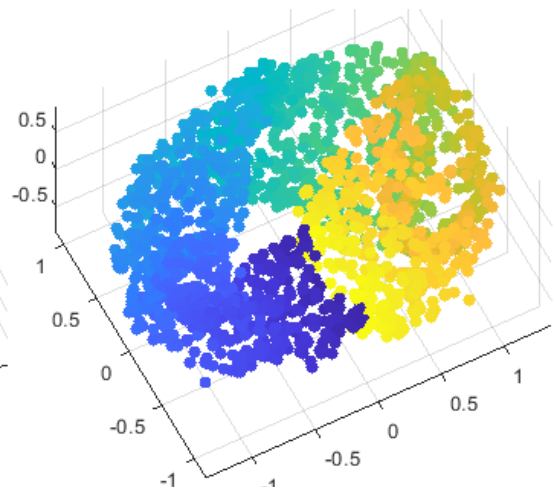


Evaluation

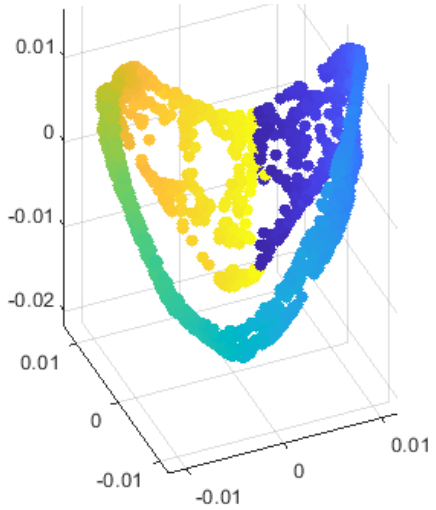
Original



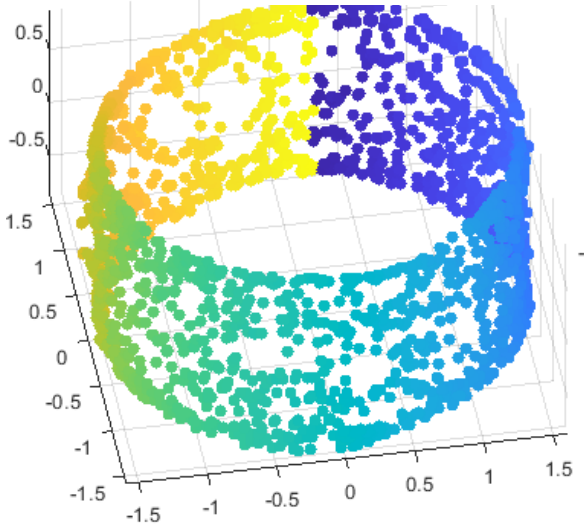
EucApprox



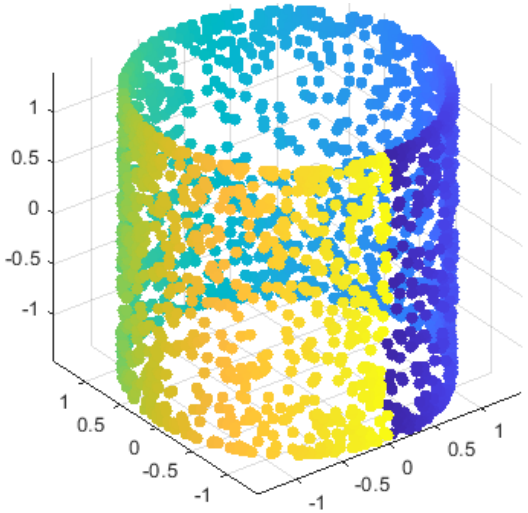
Laplacian Eigenmaps



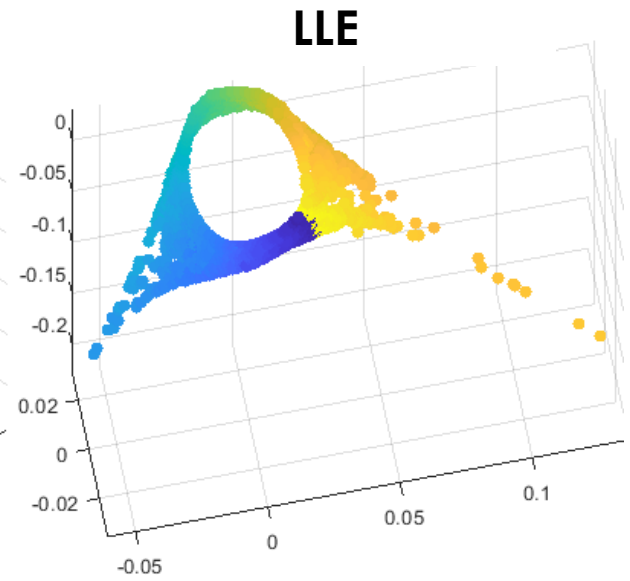
IsoMap



Diffusion Maps



LLE



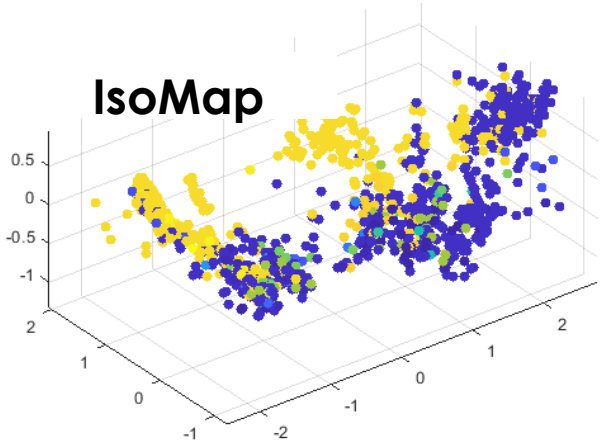
Flat torus short embedding in $\mathbb{R}^3$: Visually, EucApprox best approximates the support of the $\mathbb{R}^3$ isometry of the flat torus, while the baseline ones, either local or global, flatten one principal direction.

Evaluation

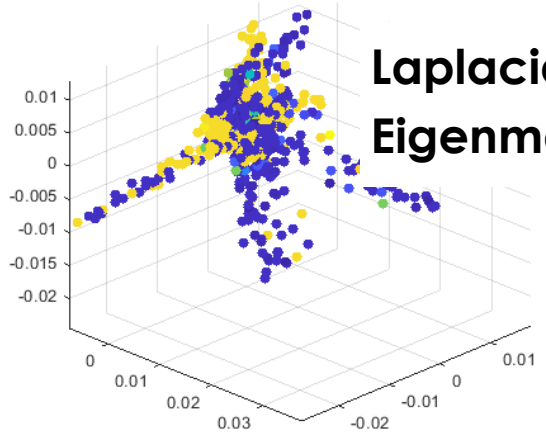
Table 2: Evaluation of the clusters' representation for the RNA-seq data set, related to lung cancer

	Diagnosis classification within embeddins' neighborhoods							
	LLE	LTSA	LplcMps	IsoMap	PHATE	t-SNE	UMAP	proposed
ACC ↑	0.6954	0.6640	0.6985	0.6825	0.6745	0.6917	0.7040	0.7126
NMI ↑	0.2934	0.2660	0.3006	0.2811	0.2952	0.3048	0.3213	0.3298
ARI ↑	0.3356	0.2525	0.3525	0.3020	0.2721	0.3461	0.3742	0.3821

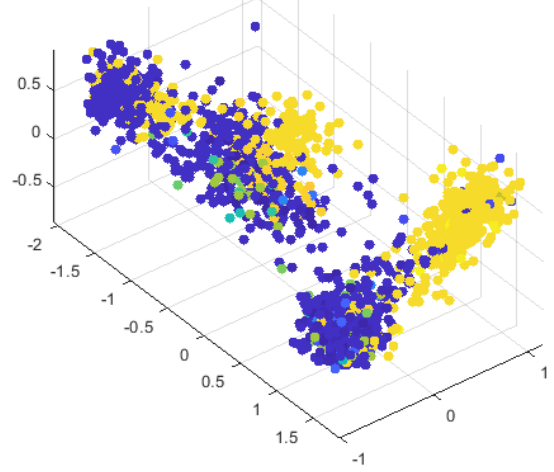
IsoMap



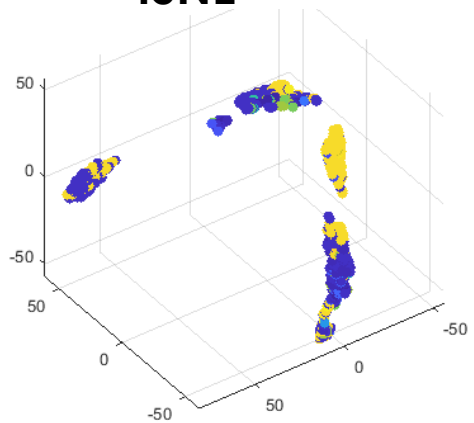
Laplacian Eigenmaps



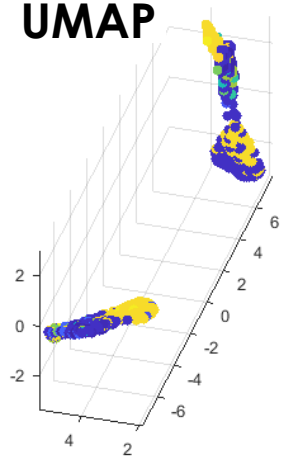
EucApprox



tSNE



UMAP



Lung cancer diagnoses' discrimination from single cell RNA data: 3D projection of the embeddings in $\mathbb{R}^{30}$. Coloring according to the lung cancer diagnosis. EucApprox respects the global distances and enhances the data discrimination by retaining local structures, up to integrability.