UNSUPERVISED GROUND METRIC LEARNING

Janis Auffenberg, Jonas Bresch, Oleh Melnyk, Gabriele Steidl
Technische Universität Berlin, Ludwig-Maximilians-Universität München

Key Takeaways

- Learning distances on samples and features via fixed points simultaneously.
- In [2], OT cost matrices are learned between samples and features.
- We propose a more general setting with extensions to Mahalanobis distances and graph Laplacians.

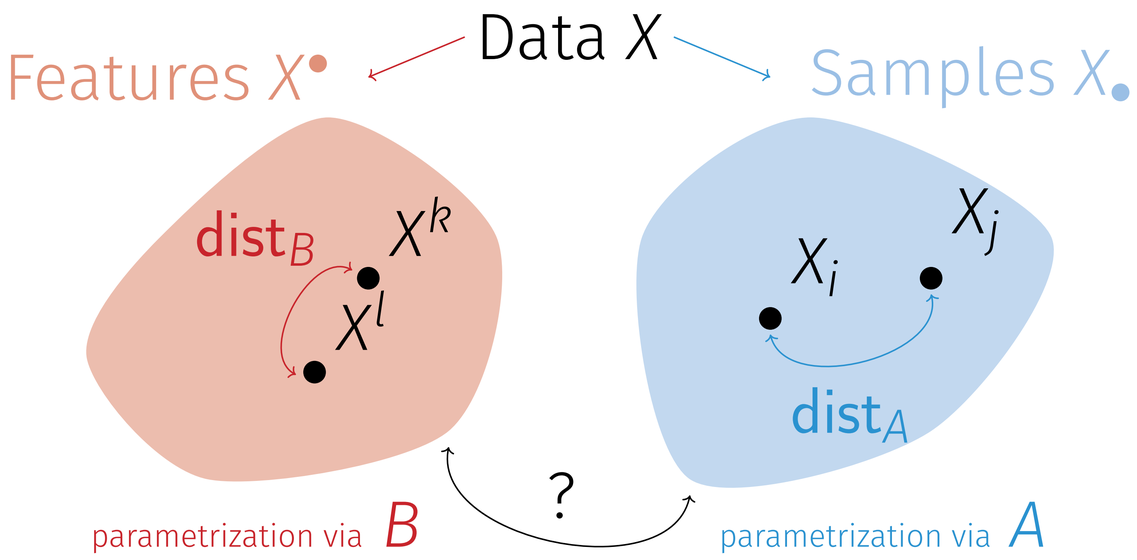
Setting

- Samples $X = (X_1 \dots X_n)$ and features $X^{\mathrm{T}} = (X^1, \dots, X^m)$.
- Assumption: $X_i \neq X_j, X^k \neq X^l$ for $i \neq j$ and $k \neq l$.
- In some applications, $\underline{X}$ and $\overline{X}$ are normalized:

$$\underline{X}\mathbf{1}_n = \mathbf{1}_m, \qquad \overline{X}^{\mathrm{T}}\mathbf{1}_m = \mathbf{1}_n.$$

Goal

Find distances $\mathbf{dist}_A$ and $\mathbf{dist}_B$ on the sample and feature spaces that can be used for clustering.



Main idea

Link parametrizations A and B as fixed points of certain mappings $\mathcal{F} : \mathbb{R}^{m\times m} \to \mathbb{R}^{n\times n}$ and $\mathcal{G} : \mathbb{R}^{n\times n} \to \mathbb{R}^{m\times m}$, so that

$$B = \mathcal{F}(A) \quad \text{and} \quad A = \mathcal{G}(B).$$

Examples

I) On pseudo-distance matrices A, B (for (1) in [2]):

$$\mathcal{F}(A) := \left(W_A(\underline{X}_i, \underline{X}_j)\right)_{i,j=1}^{n} + R_n, \qquad \mathcal{G}(B) := \left(W_B(\overline{X}^k, \overline{X}^l)\right)_{k,l=1}^{m} + R_m,$$

- where R_n, R_m are some regularizer and Wasserstein distance

$$W_\cdot(x,y) := \min_{P\in\Pi(x,y)} \langle \cdot, P\rangle, \quad \Pi(x,y) := \{P \in \mathbb{R}^{\star\times\star}_{\geq 0},\ P\mathbf{1}_\star = x,\ P^{\mathrm{T}}\mathbf{1}_\star = y\}$$

II) On semi-pseudo-distance matrices A, B with $\varepsilon > 0$:

$$\mathcal{F}(A) := S^\varepsilon_A(\underline{X}) + R_n \quad \text{and} \quad \mathcal{G}(B) := S^\varepsilon_B(\overline{X}) + R_m.$$

- with $W^\varepsilon_\cdot(x,y) := \min_{P\in\Pi(x,y)} \left\{\langle \cdot, P\rangle + \varepsilon\|\cdot\|_\infty \mathrm{KL}(P, xy^{\mathrm{T}})\right\}$
- and Sinkhorn divergence

$$S^\varepsilon_\cdot(x,y) := W^\varepsilon_\cdot(x,y) - \tfrac{1}{2}(W^\varepsilon_\cdot(x,x) + W^\varepsilon_\cdot(y,y)).$$

III) On symmetric-positive definite matrices A, B:

$$\mathcal{F}(A) := f\left(\left(M_A(X_i, X_j)\right)_{i,j=1}^{n}\right) + R_n, \quad \mathcal{G}(B) := g\left(\left(M_B(X^k, X^l)\right)_{k,l=1}^{m}\right) + R_m,$$

radially positive definite functions f, g.

- with Mahalanobis distance

$$M_\cdot(x,y) := \sqrt{(x-y)^{\mathrm{T}} \cdot (x-y)}, \quad x, y \in \mathbb{R}^m.$$

IV) On positive semi-definite matrices A, B:

$$\mathcal{F}(A) := \mathrm{diag}\left(\mathcal{W}_A(\underline{X})\mathbb{1}_n\right) - \mathcal{W}_A(\underline{X}), \quad \mathcal{G}(B) := \mathrm{diag}\left(\mathcal{W}_B(\overline{X})\mathbb{1}_m\right) - \mathcal{W}_B(\overline{X}).$$

- with graph-Laplacian $\mathcal{W}_\cdot(X) := \left(M^2_\cdot(X_i, X_j)\right)_{i,j=1}^{n}$,
- using the Mahalanobis distance.

Fixed Point Algorithms

- Assumption: $\mathcal{F}$ and $\mathcal{G}$ are Lipschitz w.r.t. $\|\cdot\|_\infty$ with $L_\mathcal{F}, L_\mathcal{G} > 0$.
- Normalized iterations:

$$B^t = \tilde{\mathcal{F}}(A^t) := \frac{\mathcal{F}(A^t)}{\|\mathcal{F}(A^t)\|_\infty} \quad \text{and} \quad A^{t+1} = \tilde{\mathcal{G}}(B^t) := \frac{\mathcal{G}(B^t)}{\|\mathcal{G}(B^t)\|_\infty}. \tag{1}$$

- Scaled iterations:

$$B^t = \gamma_\mathcal{F}\mathcal{F}(A^t) \quad \text{and} \quad A^{t+1} = \gamma_\mathcal{G}\mathcal{G}(B^t), \quad \gamma_\mathcal{F}, \gamma_\mathcal{G} > 0. \tag{2}$$

- Iteration as a single operator $\tilde{\mathcal{T}} := \tilde{\mathcal{G}} \circ \tilde{\mathcal{F}}$ and

$$\mathcal{T}_\alpha(A) := (1-\alpha)A + \alpha\gamma_\mathcal{G}\mathcal{G}\left(\gamma_\mathcal{F}\mathcal{F}(A)\right), \qquad \alpha \in (0,1].$$

- Stochastic variant of (2): update single randomly chosen entry of A and B per iteration.

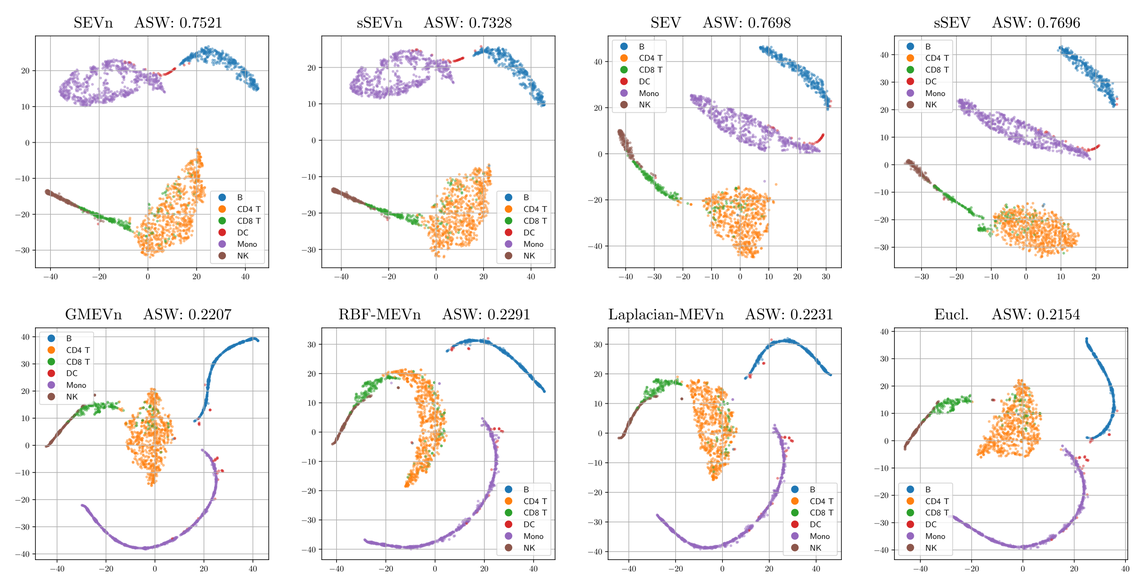
Theorem: existence and linear convergence

Consider $\mathcal{F}$ and $\mathcal{G}$ from Examples I, II, III or IV. Under mild assumptions, we have:

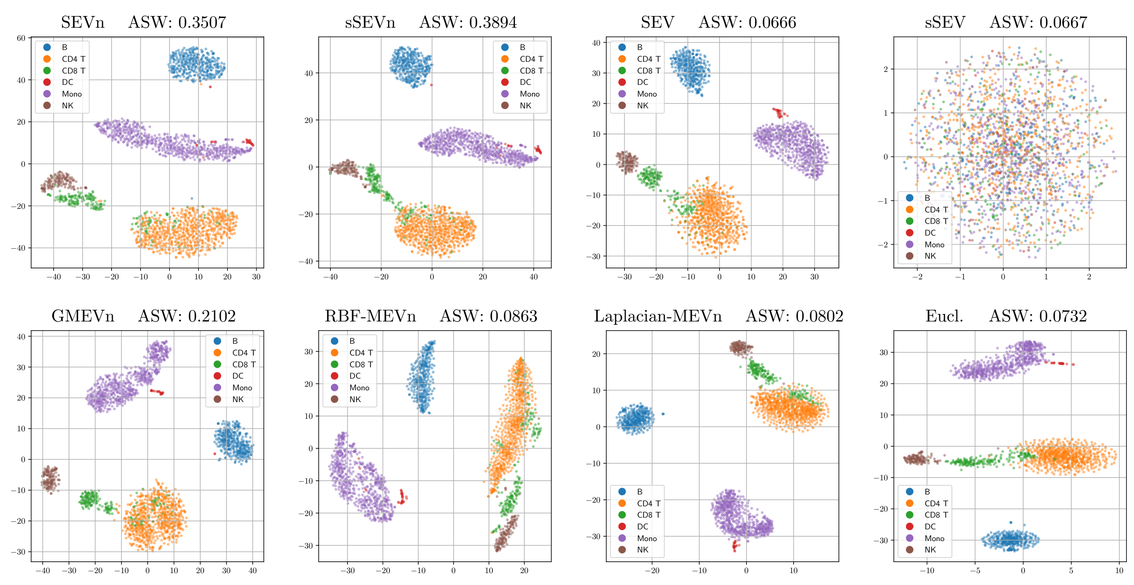
- Iterations (1) and (2) admit unique fixed points A and B.
- Linear convergence rate to fixed points.
- Stochastic version converges almost surely at a linear rate on average.

Numerical Experiments

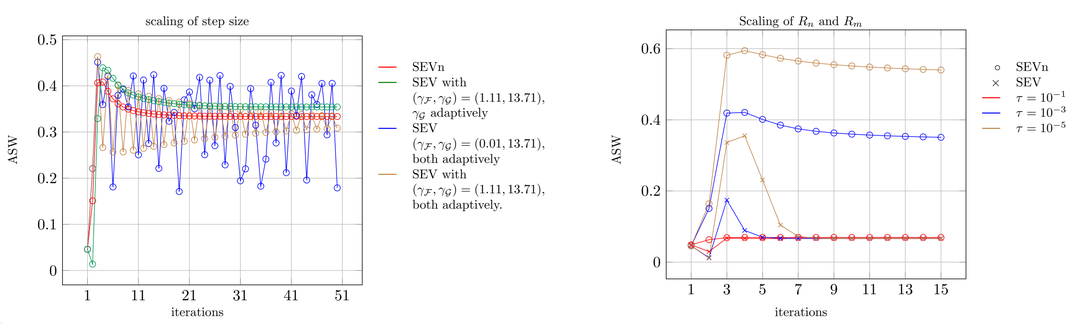
Reduced (PCA, with dimension 10) Single-cell RNA Sequencing



Single-cell RNA Sequencing



Impact of parameters



References

[1] J. Auffenberg, J. Bresch, O. Melnyk and G. Steidl: Unsupervised Ground Metric Learning. *arXiv* (2025) – arXiv:2507.13094.

[2] N. Hermer, and D. Russell and A. Sturm: Random function iterations for consistent stochastic feasibility. *Numerical Functional Analysis and Optimization* (2019).

[3] N. Hermer, and D. Russell and A. Sturm: Random function iterations for stochastic fixed point problems. *arXiv* (2020) – arXiv:2410.16149.

[4] G.-J. Huizing, L. Cantini and G. Peyré: Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors. *PMLR* (2022).

bresch@math.tu-berlin.de, melnyk@math.tu-berlin.de, steidl@math.tu-berlin.de