中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors

Geert-Jan Huizingy*, Laura Cantini, Gabriel Peyre
CNRS and ENS, PSL University

(ICML'2022)

Fanmeng Wang

2022-07-21

Outline

- Introduction
- Method: Wasserstein Singular Vectors
- Improvement: Scale the method to large datasets
- Application: Study a single-cell RNA-sequencing dataset
- Conclusion

- Introduction
- Method: Wasserstein Singular Vectors
- Improvement: Scale the method to large datasets
- Application: Study a single-cell RNA-sequencing dataset
- Conclusion

Introduction

- **How to define distances between samples in a dataset?**
  - The dataset is represented as a data matrix $U \in R_{+}^{m \times n}$, with m rows (the features) and n columns (the samples).
  - Dening meaningful distances between samples, which are columns in a data matrix, is a fundamental problem in machine learning.
- **Optimal Transport (OT) defines geometrically meaningful distances between probability distributions.**

Introduction

Optimal Transport distances

- OT is parametrized by a distance between the features (the rows of the data matrix): the "ground cost".
- OT lifts a ground pairwise distance matrix A ∈ $R_{+}^{m\times m}$ between the m features, to the **Wasserstein OT distance** between normalized samples $a_i := \frac{U_{\cdot,i}}{\|U_{\cdot,i}\|_1}$ and $a_j := \frac{U_{\cdot,j}}{\|U_{\cdot,j}\|_1}$.

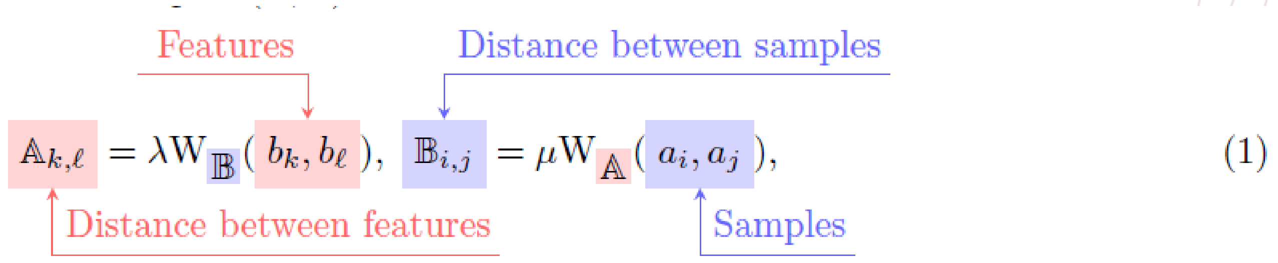
$$\mathrm{W}_{\mathbb{A}}(a_i,a_j) := \min_{P\in\mathbb{R}_{+}^{m\times m}} \sum\nolimits_{k,\ell} P_{k,\ell}\mathbb{A}_{k,\ell} \quad \text{s.t.} \quad \begin{cases} P\mathbb{1}_m = a_i,\\ P^{\top}\mathbb{1}_m = a_j.\end{cases}$$

- It is computed by optimizing a transport plan P encoding the displacement of mass between the two histograms

Introduction

From supervised to unsupervised ground metric learning

- We simultaneously compute an OT distance between the rows and between the columns of a data matrix.



$$\mathbb{A}_{k,\ell} = \lambda \mathrm{W}_{\mathbb{B}}(b_k, b_\ell), \quad \mathbb{B}_{i,j} = \mu \mathrm{W}_{\mathbb{A}}(a_i, a_j), \qquad (1)$$

- where $(\lambda, \mu) \in R_+^2$ are scaling factors, $a_i := \frac{U_{.,i}}{\|U_{.,i}\|_1}$, $a_j := \frac{U_{.,j}}{\|U_{.,j}\|_1}$, $b_i := \frac{U_{i,.}}{\|U_{i,.}\|_1}$, $b_j := \frac{U_{j,.}}{\|U_{j,.}\|_1}$

- ➢ Introduction
- ➢ Method: Wasserstein Singular Vectors
- ➢ Improvement: Scale the method to large datasets
- ➢ Application: Study a single-cell RNA-sequencing dataset
- ➢ Conclusion

Method

Wasserstein distance map

- Wasserstein distance map computes the lifting from a ground metric $A \in D_m$ toward a pairwise distance matrix $\Phi_A(\mathbb{A}) \in D_n$.

$$\Phi_A(\mathbb{A})_{i,j} := \mathrm{W}_{\mathbb{A}}(a_i, a_j) + \tau\|\mathbb{A}\|_\infty R(a_i - a_j) \qquad (2)$$

- The map $\mathbb{B} \in \mathcal{D}_n \mapsto \Phi_B(\mathbb{B}) \in \mathcal{D}_m$ is defined in the same way.

Wasserstein singular vectors

- The singular vectors of the Wasserstein distance map
- Wasserstein Singular Vectors define natural Wasserstein distances (A,B) in an unsupervised manner.

Method

Wasserstein singular vectors

- Our ground metric learning operates by solving for a pair $A \in D_m$ and $B \in D_{\text{n}}$ of Wasserstein singular vectors satisfying

$$\exists(\lambda, \mu) \in {\mathbb{R}_+^*}^2 \text{ s.t. } \Phi_B(\mathbb{B}) = \lambda \mathbb{A},\ \Phi_A(\mathbb{A}) = \mu \mathbb{B}, \tag{3}$$

which corresponds to (1) when $\tau = 0$.

$$\Phi_A(\mathbb{A})_{i,j} := \mathrm{W}_{\mathbb{A}}(a_i, a_j) + \tau \|\mathbb{A}\|_\infty R(a_i - a_j) \tag{2}$$

$$\underbrace{\mathbb{A}_{k,\ell}}_{\text{Distance between features}} = \lambda \mathrm{W}_{\mathbb{B}}(\overbrace{b_k, b_\ell}^{\text{Features}}),\ \overbrace{\mathbb{B}_{i,j}}^{\text{Distance between samples}} = \mu \mathrm{W}_{\mathbb{A}}(\underbrace{a_i, a_j}_{\text{Samples}}), \tag{1}$$

Method

Power iterations algorithm

- The de-facto standard algorithm to extract singular vectors

$$\mathbb{A}_{t+1} := \frac{\Phi_B(\mathbb{B}_t)}{\|\Phi_B(\mathbb{B}_t)\|_\infty}, \quad \mathbb{B}_{t+1} := \frac{\Phi_A(\mathbb{A}_{t+1})}{\|\Phi_A(\mathbb{A}_{t+1})\|_\infty}. \tag{4}$$

- The complexity of performing a single power iteration is $O(n^2m^2$(n log(n) + mlog(m))), since the computation of a single Wasserstein distance in R_+^n is $O(n^3$ log n) .
- As n or m grows, the complexity of the power iterations (4) becomes prohibitive.

- Introduction
- Method: Wasserstein Singular Vectors
- Improvement: Scale the method to large datasets
- Application: Study a single-cell RNA-sequencing dataset
- Conclusion

Improvement

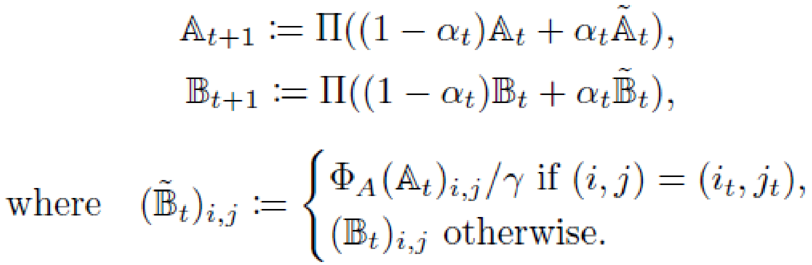
Large scale stochastic power iteration

- In order to work around the issue of (4)
- It is similar in spirit to **stochastic gradient descent**, which updates a single (or several if applied in a mini-batch setting) randomly chosen distance value at each step.
- This speeds up each iteration and leverages the correlations in the dataset.

Improvement

■ Large scale stochastic power iteration

- For some decreasing step size α_t and a scaling factor γ > 0, we define



$$\mathbb{A}_{t+1} := \Pi((1-\alpha_t)\mathbb{A}_t + \alpha_t\tilde{\mathbb{A}}_t),$$
$$\mathbb{B}_{t+1} := \Pi((1-\alpha_t)\mathbb{B}_t + \alpha_t\tilde{\mathbb{B}}_t),$$

where $(\tilde{\mathbb{B}}_t)_{i,j} := \begin{cases}\Phi_A(\mathbb{A}_t)_{i,j}/\gamma \text{ if } (i,j) = (i_t, j_t), \\ (\mathbb{B}_t)_{i,j} \text{ otherwise.}\end{cases}$

where $\prod(A)$:= A /$\|A\|_\infty$ and (i_t, j_t) is is drawn uniformly at random in $\{0, \dots, n\}^2$ and is the index updated at each step (the same update rule is applied to compute $\widetilde{A_{\mathrm{t}}}$)

Improvement

Large scale stochastic power iteration

- For $\alpha_t = 1 / \sqrt{t}$, $\gamma = \tau \min(min_{i\neq j}R(a_i - a_j), min_{k\neq l}R(b_k - \mathrm{b}_l))$ and τ large enough, **the stochastic power iterations** defined above converge to a pair (A,B) ∈ $D_m \times D_\mathrm{n}$ of positive singular vectors with a convergence rate of $O(\log(\mathrm{t}) / \sqrt{t})$

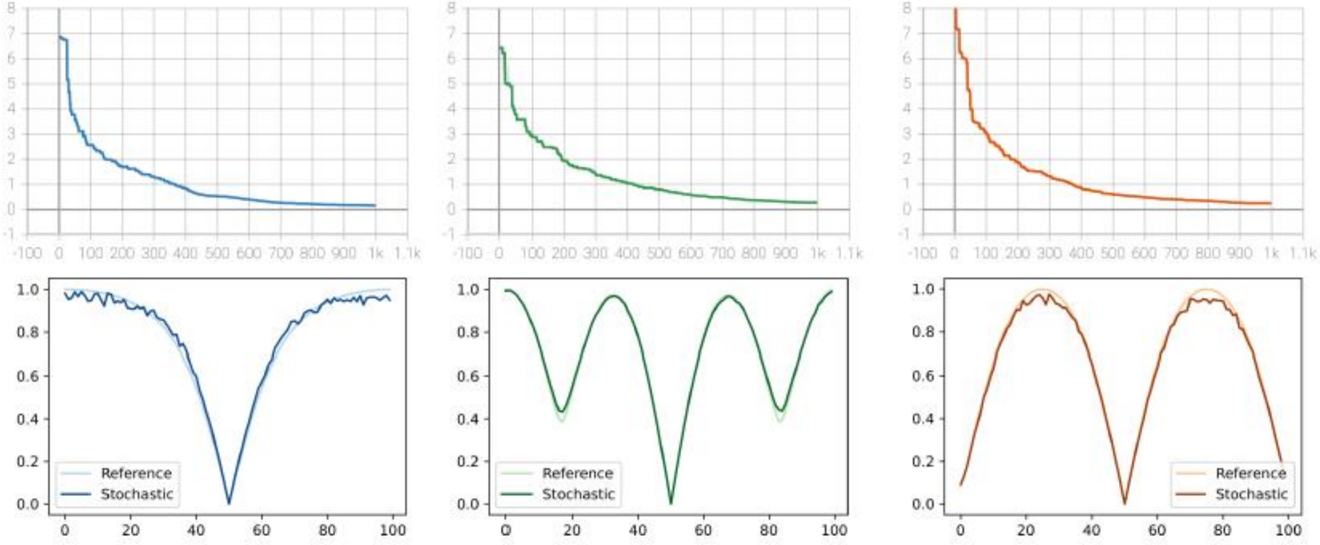


Figure 3: (top) Convergence rate $d_{\mathcal{H}}(\mathbb{B}_t, \mathbb{B}_\infty)$ (bottom) Comparison between the approximated singular vectors $(\mathbb{B}_p)_t$ and the reference singular vectors $\mathbb{B}_p$

Improvement

Parallelization with entropic regularisation

- Entropic regularization can be computed eciently in O(n^2/ε^2) using Sinkhorn's algorithm at the expense of an approximation of the OT distances of order ε^2
- Further speed up the method
- Parallel computations of the distance map on GPUs

Improvement

Parallelization with entropic regularisation

- The entropic regularized OT is defined

$$\mathrm{W}^{\varepsilon}_{\mathbb{A}}(a_i,a_j) := \min_{P\in\Gamma(a_i,a_j)} \langle P,\mathbb{A}\rangle + \varepsilon\|\mathbb{A}\|_\infty\langle P,\log P\rangle.$$

- The bias is removed by using instead the Sinkhorn divergence

$$\bar{\mathrm{W}}^{\varepsilon}_{\mathbb{A}}(a_i,a_j) := \mathrm{W}^{\varepsilon}_{\mathbb{A}}(a_i,a_j) - \frac{1}{2}\left(\mathrm{W}^{\varepsilon}_{\mathbb{A}}(a_i,a_i)+\mathrm{W}^{\varepsilon}_{\mathbb{A}}(a_j,a_j)\right).$$

- The **entropic Wasserstein distance map** is then defined as

$$\Phi^{\varepsilon}_{A}(\mathbb{A}\in\mathcal{D}_m)_{i,j} := \bar{\mathrm{W}}^{\varepsilon}_{\mathbb{A}}(a_i,a_j)+\tau\|\mathbb{A}\|_\infty R(a_i-a_j),$$

It reduces to (2) when $\varepsilon = 0$

$$\Phi_A(\mathbb{A})_{i,j} := \mathrm{W}_{\mathbb{A}}(a_i,a_j)+\tau\|\mathbb{A}\|_\infty R(a_i-a_j) \tag{2}$$

- Introduction
- Method: Wasserstein Singular Vectors
- Improvement: Scale the method to large datasets
- Application: Study a single-cell RNA-sequencing dataset
- Conclusion

Application

Study a single-cell RNA-sequencing dataset

- We use the annotation on cells (resp. on marker genes) to evaluate the quality of distances between cells (resp. between marker genes).
- We report in Table 1 and Table 2 the Average Silhouette Width (ASW), computed using the function silhouette score of Scikit-learn.

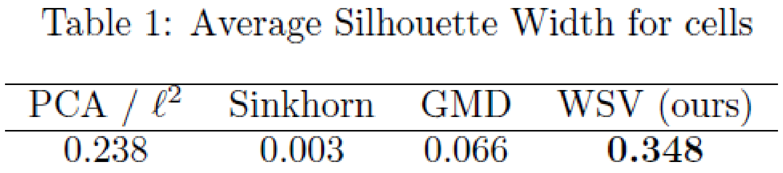


Table 1: Average Silhouette Width for cells

PCA / ℓ^2	Sinkhorn	GMD	WSV (ours)
0.238	0.003	0.066	**0.348**

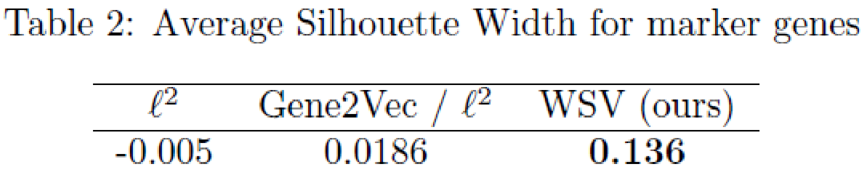


Table 2: Average Silhouette Width for marker genes

ℓ^2	Gene2Vec / ℓ^2	WSV (ours)
-0.005	0.0186	**0.136**

I. Euclidean distances on PCA embeddings
II. Sinkhorn divergence
III. GMD
IV. Wasserstein Singular Vectors

Application

Study a single-cell RNA-sequencing dataset

We visualize both of these distances using a 2-D UMAP projection.

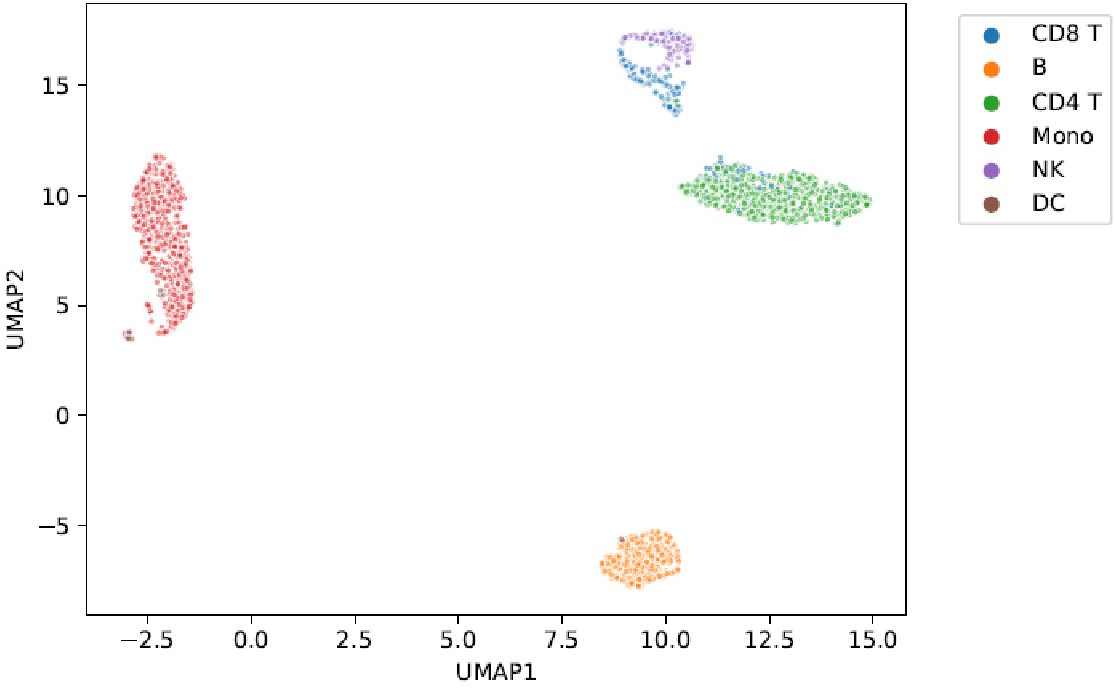


Figure 4: UMAP projection of the cells of scRNA-seq data using $\mathbb{B}$, with cells colored by cell type.

Application

■ Study a single-cell RNA-sequencing dataset

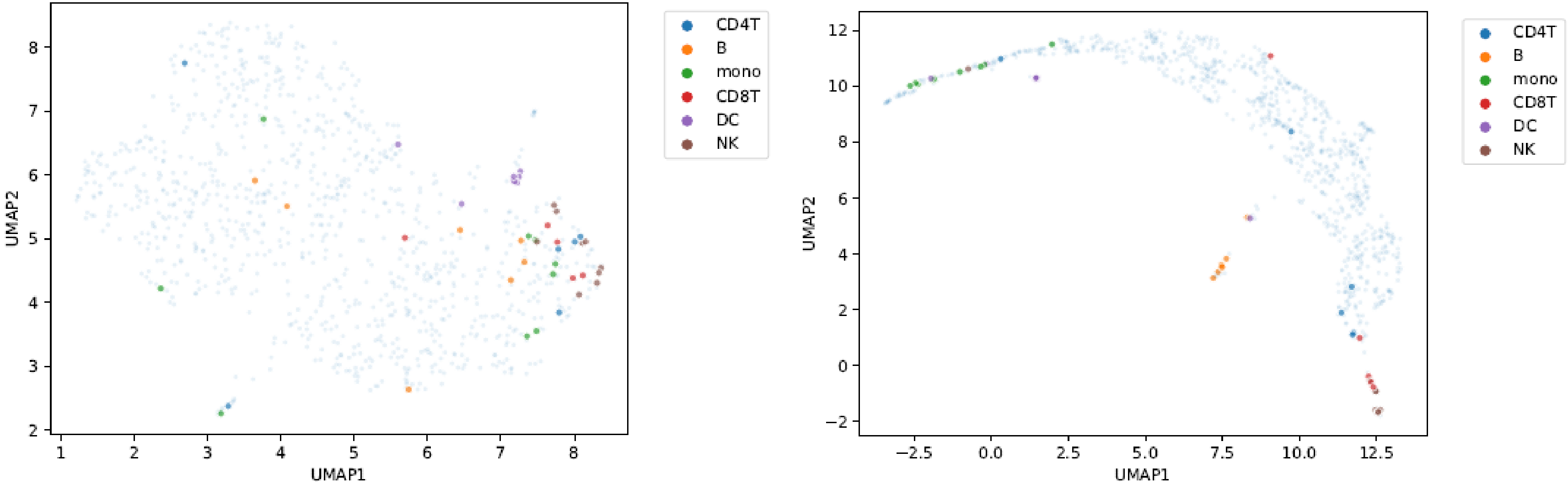
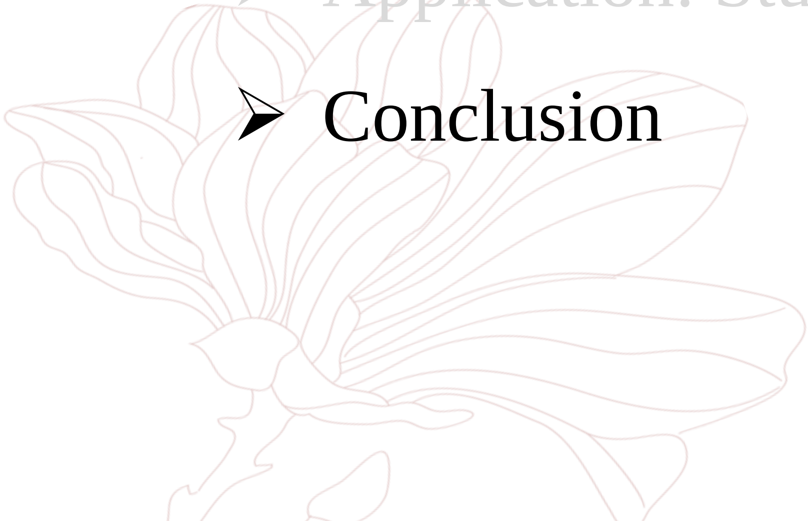


Figure 5: 2-D UMAP projection of marker genes in a scRNA-seq dataset, using the computed distances. Marker genes are colored by associated cell type, and other genes are faded out. Left, the euclidean distance on Gene2Vec [15] embeddings. Right, the singular vector $\mathbb{A}$.

- ➢ Introduction
- ➢ Method: Wasserstein Singular Vectors
- ➢ Improvement: Scale the method to large datasets
- ➢ Application: Study a single-cell RNA-sequencing dataset
- ➢ Conclusion



Conclusion

- Wasserstein singular vectors as the positive singular vectors of monotone homogeneous "distance maps“, defining a pair of " intrinsic" ground metrics associated to a given dataset.
- This solves in an elegant way the problem of unsupervised ground metric learning, without resorting to some ad-hoc embeddings.

Thank You for listening!

2022-07-21