SNEkhorn

Dimension Reduction with Symmetric Entropic Affinities

Hugues Van Assel, Titouan Vayer, **Rémi Flamary**, Nicolas Courty

November 9 2023

CentraleSupelec, Université Paris-Saclay, France


H. Van Assel


T. Vayer

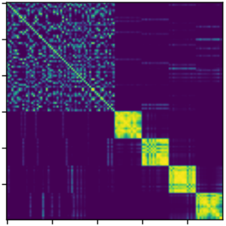

R. Flamary

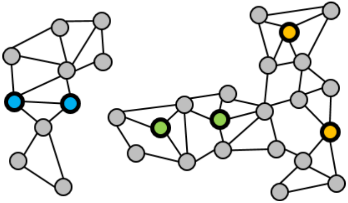

N. Courty

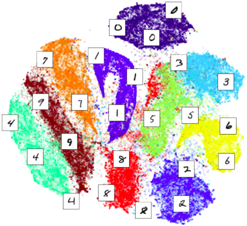
Table of content

Affinity matrices in machine learning

Affinity matrices in machine learning

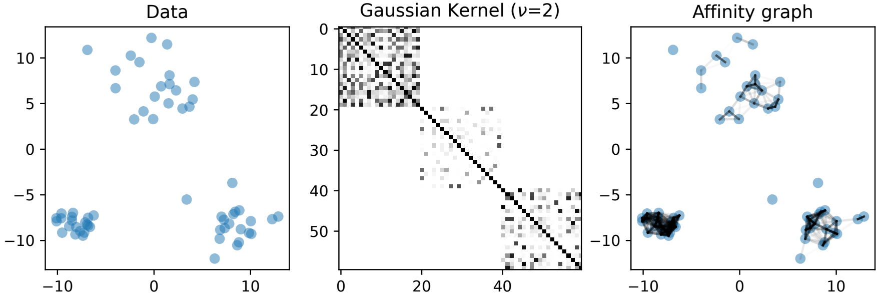






- Affinity matrix: symmetric and positive matrix encoding the relationship between data points (graph, kernel or similarity).
- Many ML methods rely on similarity/affinity matrices :
  - Kernel machines [Schölkopf and Smola, 2002].
  - Clustering (spectral, kernel) [Von Luxburg, 2007].
  - Semi-supervised learning[Zhou et al., 2003]
  - Self-supervised learning (Barlow twins [Zbontar et al., 2021]).
  - Dimensionality reduction (TNSE [Van der Maaten and Hinton, 2008]).

Kernel matrices

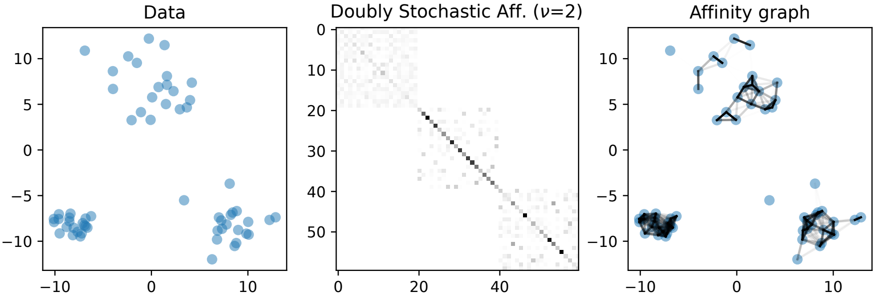


Gaussian (or Gibbs) kernel

$$K^g_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\nu) = \exp(-C_{ij}/\nu) \qquad \text{(Gaussian kernel)}$$

- $\{\mathbf{x}_i\}_{i=1,\dots,n}$ are data points in $\mathbb{R}^d$ and $C_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$.
- $\mathbf{K}$ is the kernel matrix of components K_{ij}.
- ν is a parameter tuning the neighborhood size.
- Used in support vector machines, spectral clustering, etc.
- Clusters have very different mass/impact on the kernel.

Doubly Stochastic affinity matrix



DS affinity matrix

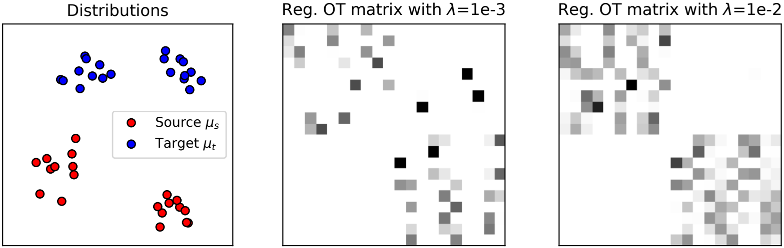
$$\mathbf{P}^{\mathrm{ds}} = \min_{\mathbf{P}\in\Pi} \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

- Projection of $\mathbf{K}$ on the set of doubly stochastic matrices Π.
- Can be solved (estimation of $\mathbf{f}$) using the Sinkhorn-Knopp algorithm with iterations $f_i^{t+1} \leftarrow \frac{1}{2}\left(f_i^t - \log \sum_k \exp\left(f_k^t - C_{ki}\right)\right)\ \forall i$ and solution :

$$P_{ij}^{\mathrm{ds}} = \exp\left((f_i + f_j - C_{ij})/\nu\right) \text{ where } \mathbf{f} \in \mathbb{R}^n . \tag{DS}$$

- Equivalent to self entropic regularized optimal transport [Cuturi, 2013].

Entropic regularized optimal transport



Entropic regularized OT [Cuturi, 2013]

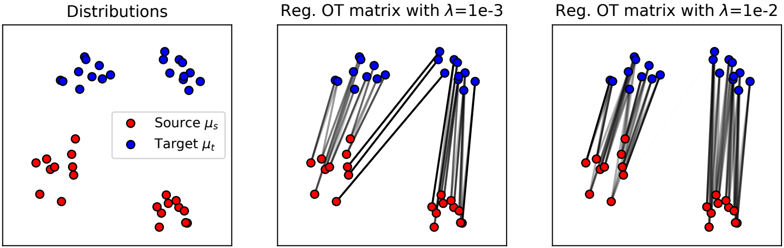
$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C}\rangle_F - \nu H(\mathbf{P})$$

- Regularization with the entropy of $H(\mathbf{P}) = -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$.
- Looses sparsity, gains stability, strictly convex, solved with Sinkhorn.
- Equivalent to the following problem (global constraint on entropy)

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C}\rangle_F \quad \text{s.t.} \quad H(\mathbf{P}) \geq \eta$$

- For symmetric $\mathbf{C}$ the OT plan $\mathbf{P}^{ds}$ is also symmetric.

Entropic regularized optimal transport



Entropic regularized OT [Cuturi, 2013]

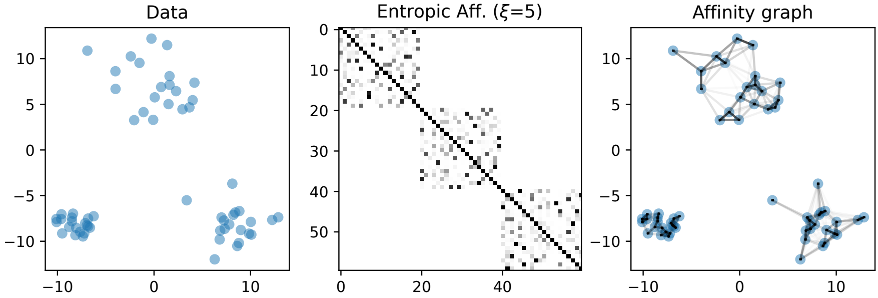
$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C}\rangle_F - \nu H(\mathbf{P})$$

- Regularization with the entropy of $H(\mathbf{P}) = -\sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$.
- Looses sparsity, gains stability, strictly convex, solved with Sinkhorn.
- Equivalent to the following problem (global constraint on entropy)

$$\mathbf{P}^{ds} = \underset{\mathbf{P}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{P}, \mathbf{C}\rangle_F \quad \text{s.t.} \quad H(\mathbf{P}) \geq \eta$$

- For symmetric $\mathbf{C}$ the OT plan $\mathbf{P}^{ds}$ is also symmetric.

Entropic Affinity matrices



Entropic affinities (perplexity ξ)

$$P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon_i^\star\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon_i^\star\right)} \quad \text{with } \varepsilon_i^\star \in \mathbb{R}_+^* \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log \xi + 1\,. \qquad \text{(EA)}$$

- $H(\mathbf{v}) = -\sum_i v_i(\log(v_i) - 1)$ is the entropy and $C_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$.
- Adaptive scaling $\varepsilon_i^\star$ per point to ensure an equivalent "spread" of the mass.
- $\mathbf{P}^{\mathrm{se}}$ is not symmetric. **Symmetric variant** : $\overline{\mathbf{P}^{\mathrm{e}}} = (\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})/2$.
- Used in Stochastic Neighbor Embedding (SNE) [Hinton and Roweis, 2002] and tSNE [Van der Maaten and Hinton, 2008] (symmetric variant).

Entropic affinities properties

EA matrices seen as an OT problem

Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ without constant rows. Then $\mathbf{P}^{\mathrm{e}}$ solves the entropic affinity problem (EA) with cost $\mathbf{C}$ if and only if $\mathbf{P}^{\mathrm{e}}$ is the unique solution of the convex problem

$$\mathbf{P}^{e} = \underset{\mathbf{P} \in \mathcal{H}_\xi}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{C} \rangle. \qquad \text{(EA as OT)}$$

- $\mathcal{H}_\xi$ is the set of doubly stochastic matrices with row entropic constraints.

$$\mathcal{H}_\xi = \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times n} \text{ s.t. } \mathbf{P}\mathbf{1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1 \right\}. \qquad (1)$$

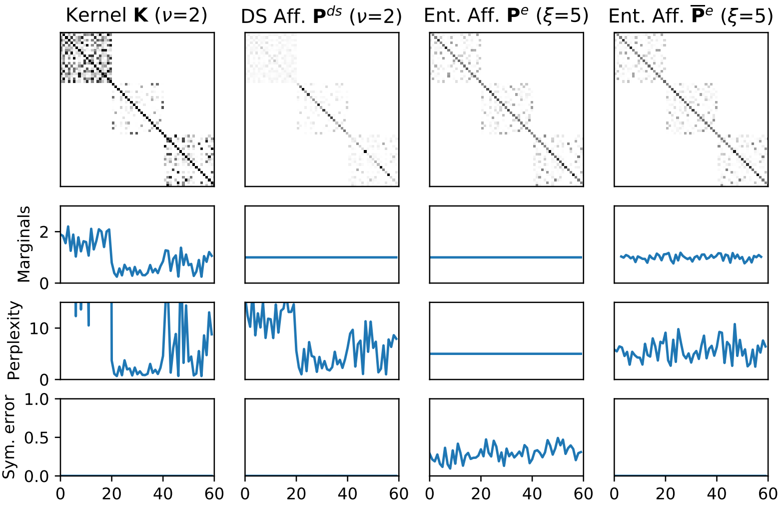
- EA matrix computation is a semi-relaxed OT with line entropy constraints.
- The solution $\mathbf{P}^e$ has saturated entropy with equality in the constraints.
- Can be solved with n independent root-finding algorithms.

Symmetric variant post-processing

$$\overline{\mathbf{P}^{\mathrm{e}}} = (\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})/2$$

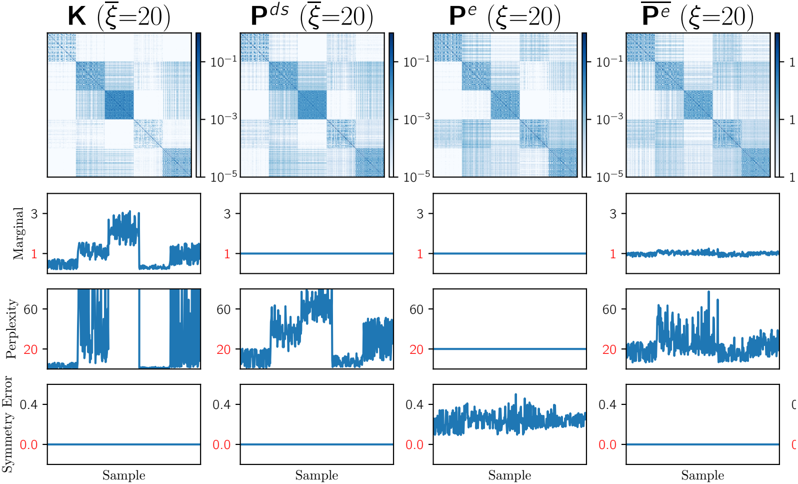
- Orthogonal (L2) projection of on the set of symmetric matrices $\mathcal{S}$.
- Mixture of L2 and KL geometry and last projection do not preserved entropic constraints.

Comparison between all affinity matrices



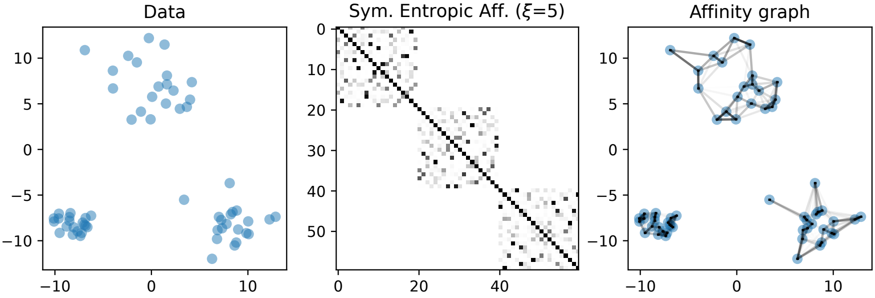
- We compute the marginals, the entropy and the L1 symmetry error.
- Comparison on 2D example (3 classes) and COIL (5 classes) dataset.

Comparison between all affinity matrices



- We compute the marginals, the entropy and the L1 symmetry error.
- Comparison on 2D example (3 classes) and COIL (5 classes) dataset.

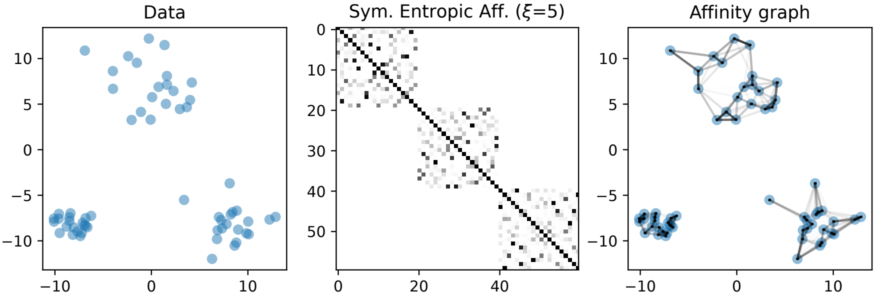
Symmetric Entropic Affinities (SEA)



Problem formulation for SEA

$$\mathbf{P}^{se} = \underset{\mathbf{P}\in\mathcal{H}_\xi\cap\mathcal{S}}{\operatorname{argmin}} \ \langle \mathbf{P}, \mathbf{C} \rangle \ . \tag{SEA}$$

- $\mathcal{S} = \{\mathbf{P} \in \mathbb{R}_+^{n\times n}$ s.t. $\mathbf{P} = \mathbf{P}^\top\}$ is the set of symmetric matrices.
- $\mathbf{P}^{se}$ is the unique solution of the convex problem (SEA) and has at least $n-1$ saturated entropy constraints (in practice we have n).



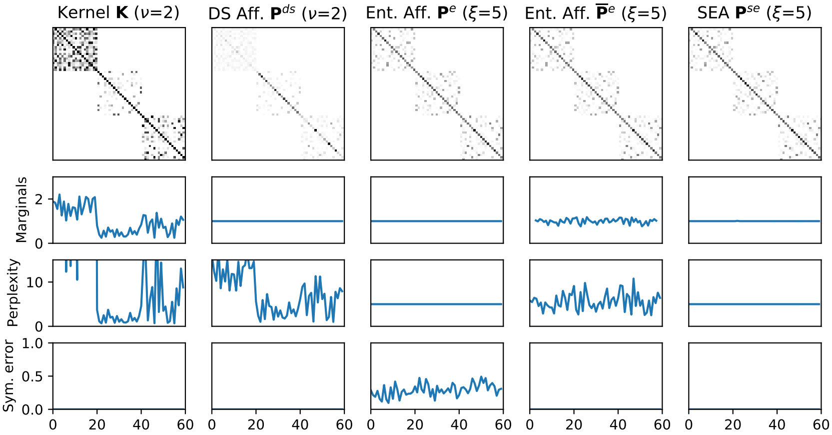
Solving for SEA

- Strong duality holds and the dual problem is

$$\max_{\boldsymbol{\gamma}>0,\boldsymbol{\lambda}} \quad \langle \mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}),\mathbf{C}\rangle + \langle \boldsymbol{\gamma},(\log\xi+1)\mathbf{1} - \mathrm{H}_{\mathrm{r}}(\mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}))\rangle + \langle \boldsymbol{\lambda},\mathbf{1}-\mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda})\mathbf{1}\rangle$$

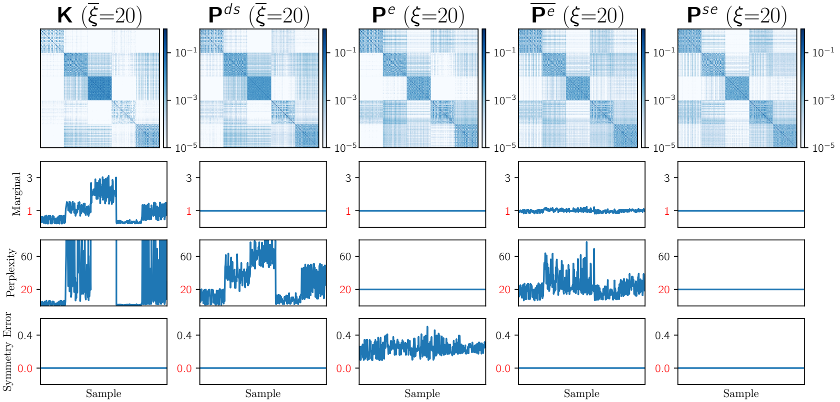
  where $\mathbf{P}(\boldsymbol{\gamma},\boldsymbol{\lambda}) = \exp\left((\boldsymbol{\lambda}\oplus\boldsymbol{\lambda} - 2\mathbf{C}) \oslash (\boldsymbol{\gamma}\oplus\boldsymbol{\gamma})\right)$.
- Solution is $\mathbf{P}^{se} = \mathbf{P}(\boldsymbol{\gamma}^\star,\boldsymbol{\lambda}^\star)$ for optimal dual variables $\boldsymbol{\gamma}^\star,\boldsymbol{\lambda}^\star$.
- Dual optimizer (L-BFGS, ADAM, etc.) is used to solve the dual problem in practice.

Comparison between all affinity matrices



- We compute the marginals, the entropy and the L1 symmetry error.
- Comparison on 2D example (3 classes) and COIL (5 classes) dataset.
- Only our approach provides controlled entropy and symmetry.

Comparison between all affinity matrices



- We compute the marginals, the entropy and the L1 symmetry error.
- Comparison on 2D example (3 classes) and COIL (5 classes) dataset.
- Only our approach provides controlled entropy and symmetry.

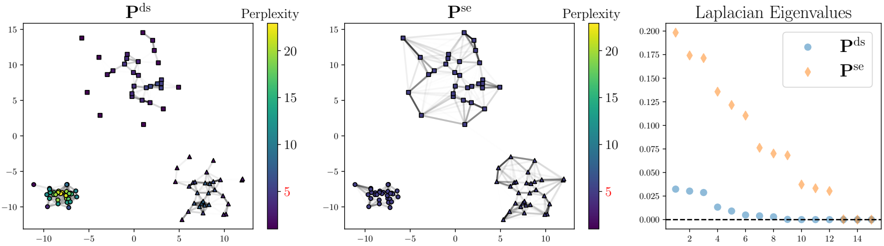
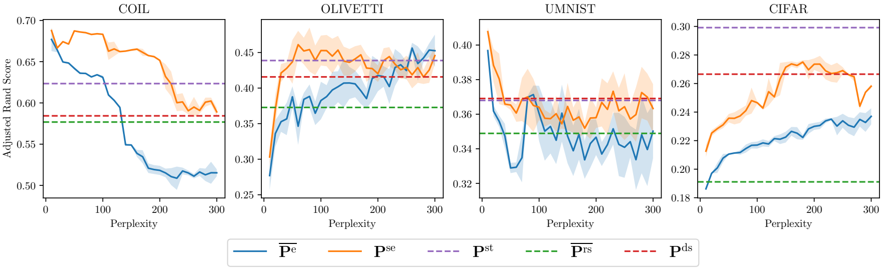


Illustration on 2D example (mixture of Gaussians)

- Comparison between Doubly Stochastic affinity and symmetric entropic affinity.
- Symmetric entropic affinity has a constant perplexity.
- Fixed perplexity adapt better to cluster of different sizes (local density).
- Eigenvalues of the Laplacian are much more separated (better clustering).



Experiment on image datasets

- Compute Adjusted Rand Index (ARI) for different affinities.
- Plot evolution of ARI as a function of perplexity.
- $\overline{\mathbf{P}^{rs}}$ is L2 symmetrized row stochastic Gaussian kernel.
- $\mathbf{P}^{st}$ is self-tuning affinity [Zelnik-Manor and Perona, 2004].
- SEA is state of the art except on CIFAR.

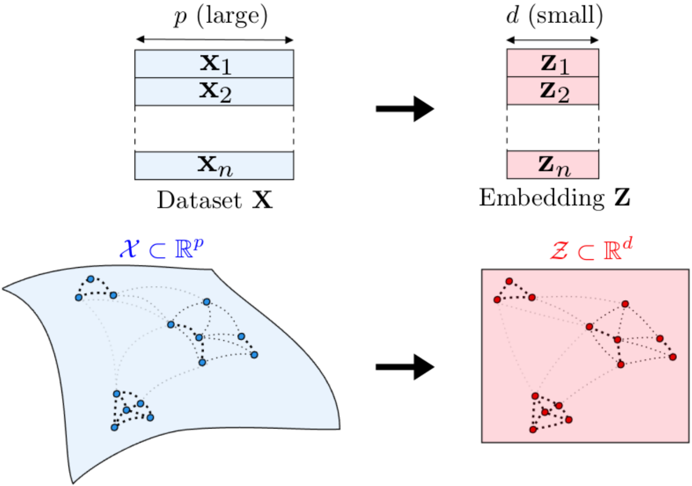
SEA for spectral clustering on Curated Microarray Database (CuMiDa)

Data set	$\overline{\mathbf{P}^{\text{rs}}}$	$\mathbf{P}^{\text{ds}}$	$\mathbf{P}^{\text{st}}$	$\overline{\mathbf{P}^{\text{e}}}$	$\mathbf{P}^{\text{se}}$
Liver (14520)	75.8	75.8	84.9	80.8	**85.9**
Breast (70947)	**30.0**	**30.0**	26.5	23.5	28.5
Leukemia (28497)	43.7	44.1	49.7	42.5	**50.6**
Colorectal (44076)	**95.9**	**95.9**	93.9	**95.9**	**95.9**
Liver (76427)	76.7	76.7	**83.3**	81.1	81.1
Breast (45827)	43.6	53.8	74.7	71.5	**77.0**
Colorectal (21510)	57.6	57.6	54.7	**94.0**	79.3
Renal (53757)	47.6	47.6	**49.5**	**49.5**	**49.5**
Prostate (6919)	12.0	13.0	13.2	16.3	**17.4**
Throat (42743)	9.29	9.29	11.4	11.8	**44.2**
scGEM	57.3	58.5	**74.8**	69.9	71.6
SNAREseq	8.89	9.95	46.3	55.4	**96.6**

Numerical experiments

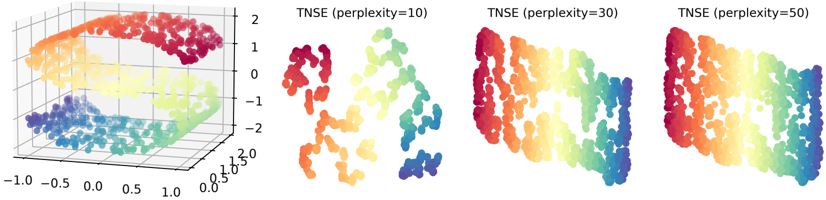
- ARI ($\times$100) for spectral clustering reported on CuMiDa datasets.
- Curated Microarray Database [Feltes et al., 2019].
- SEA is state of the art on 8/12 datasets.

Dimensionality reduction with SNEkhorn



Objective: find low dimensional representation $\mathbf{Z} \in \mathbb{R}^{n \times q}$ of the data that preserves the geometry of the data.

Stochastic Neighbor Embedding (SNE) and tSNE



Symmetric SNE [Van der Maaten and Hinton, 2008]

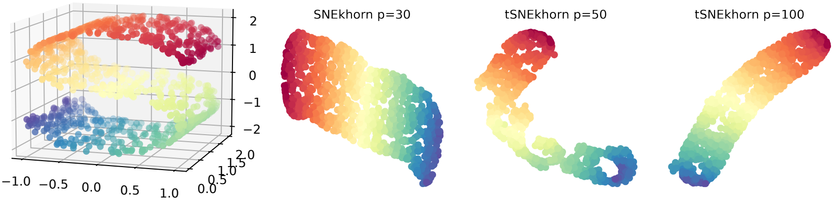
$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} \mathrm{KL}(\overline{\mathbf{P}^{\mathrm{e}}}|\widetilde{\mathbf{Q}}_{\mathbf{Z}}) \quad \text{where} \quad \overline{\mathbf{P}^{\mathrm{e}}} = \frac{1}{2}(\mathbf{P}^{\mathrm{e}} + \mathbf{P}^{\mathrm{e}\top})\,. \qquad \text{(Symmetric-SNE)}$$

with $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = \exp(-[\mathbf{C}_{\mathbf{Z}}]_{ij})/(\sum_{\ell t}\exp(-[\mathbf{C}_{\mathbf{Z}}]_{\ell t}))$ and $q \leq d$.

- Minimize the Kullback-Leibler divergence between the affinities of the data in the original space and the affinities of the embedded data.
- Embedding $\mathbf{Z}$ computed by gradient descent.

Other variants

- Original **SNE** [Hinton and Roweis, 2002] uses $\mathbf{P}^{\mathrm{e}}$ and $\mathbf{Q}_{\mathbf{Z}}$ normalized by row.
- **tSNE** uses $\overline{\mathbf{P}^{\mathrm{e}}}$ and $[\widetilde{\mathbf{Q}}_{\mathbf{Z}}]_{ij} = (1 + [\mathbf{C}_{\mathbf{Z}}]_{ij})^{-1}/\sum_{\ell,t}(1 + [\mathbf{C}_{\mathbf{Z}}]_{\ell t})^{-1}$

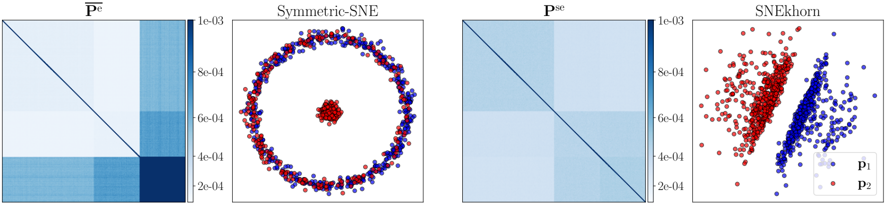


SNEkhorn

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} \mathrm{KL}\left(\mathbf{P}^{\mathrm{se}}|\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}\right), \qquad \text{(SNEkhorn)}$$

- $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}} = \exp\left(\mathbf{f}_{\mathbf{Z}} \oplus \mathbf{f}_{\mathbf{Z}} - \mathbf{C}_{\mathbf{Z}}\right)$ is computed with Sinkhorn algorithm with $\nu = 1$.
- Optimized with gradient descent and fast computation/update of Sinkhorn dual variables $\mathbf{f}_{\mathbf{Z}}$ with warm starting strategy.
- Variants:
  - **SNEkhorn** : $[\mathbf{C}_{\mathbf{Z}}]_{ij} = \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2$
  - **tSNEkhorn** : $[\mathbf{C}_{\mathbf{Z}}]_{ij} = \log(1 + \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)$
  - **Simple (t)SNEkhorn** : use $\widetilde{\mathbf{Q}}_{\mathbf{Z}}$ instead of $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}$.

SNE vs SNEkhorn



Experiment on simulated data

- Simulated data with heteroscedastic noise.
- Two classes from multinomial distribution with different probability vectors.

$$\mathbf{x}_i = \tilde{\mathbf{x}}_i / (\sum_j \tilde{x}_{ij}), \quad \tilde{\mathbf{x}}_i \sim \begin{cases} \mathcal{M}(1000, \mathbf{p}_1), & 1 \leq i \leq 500 \\ \mathcal{M}(1000, \mathbf{p}_2), & 501 \leq i \leq 750 \\ \mathcal{M}(2000, \mathbf{p}_2), & 751 \leq i \leq 1000\,. \end{cases}$$

- Second class $\mathbf{p}_2$ has samples with either low or high variance.
- SNE is mislead by the batch effect unlike SNEkhorn.

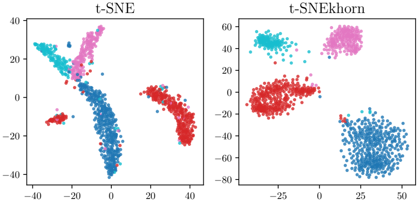
Dimensionality reduction performance

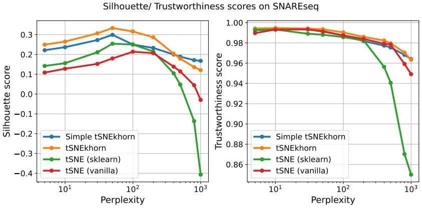
	Silhouette ($\times 100$)			Trustworthiness ($\times 100$)		
	UMAP	t-SNE	t-SNEkhorn	UMAP	t-SNE	t-SNEkhorn
COIL	20.4 ± 3.3	30.7 ± 6.9	$\mathbf{52.3 \pm 1.1}$	99.6 ± 0.1	99.6 ± 0.1	$\mathbf{99.9 \pm 0.1}$
OLIVETTI	6.4 ± 4.2	4.5 ± 3.1	$\mathbf{15.7 \pm 2.2}$	96.5 ± 1.3	96.2 ± 0.6	$\mathbf{98.0 \pm 0.4}$
UMNIST	-1.4 ± 2.7	-0.2 ± 1.5	$\mathbf{25.4 \pm 4.9}$	93.0 ± 0.4	99.6 ± 0.2	$\mathbf{99.8 \pm 0.1}$
CIFAR	13.6 ± 2.4	18.3 ± 0.8	$\mathbf{31.5 \pm 1.3}$	90.2 ± 0.8	90.1 ± 0.4	$\mathbf{92.4 \pm 0.3}$
Liver (14520)	49.7 ± 1.3	50.9 ± 0.7	$\mathbf{61.1 \pm 0.3}$	89.2 ± 0.7	90.4 ± 0.4	$\mathbf{92.3 \pm 0.3}$
Breast (70947)	28.6 ± 0.8	29.0 ± 0.2	$\mathbf{31.2 \pm 0.2}$	90.9 ± 0.5	91.3 ± 0.3	$\mathbf{93.2 \pm 0.4}$
Leukemia (28497)	22.3 ± 0.7	20.6 ± 0.7	$\mathbf{26.2 \pm 2.3}$	90.4 ± 1.1	92.3 ± 0.8	$\mathbf{94.3 \pm 0.5}$
Colorectal (44076)	67.6 ± 2.2	69.5 ± 0.5	$\mathbf{74.8 \pm 0.4}$	93.2 ± 0.7	93.7 ± 0.5	$\mathbf{94.3 \pm 0.6}$
Liver (76427)	39.4 ± 4.3	38.3 ± 0.9	$\mathbf{51.2 \pm 2.5}$	85.9 ± 0.4	89.4 ± 1.0	$\mathbf{92.0 \pm 1.0}$
Breast (45827)	35.4 ± 3.3	39.5 ± 1.9	$\mathbf{44.4 \pm 0.5}$	93.2 ± 0.4	94.3 ± 0.2	$\mathbf{94.7 \pm 0.3}$
Colorectal (21510)	38.0 ± 1.3	$\mathbf{42.3 \pm 0.6}$	35.1 ± 2.1	85.6 ± 0.7	$\mathbf{88.3 \pm 0.9}$	88.2 ± 0.7
Renal (53757)	44.4 ± 1.5	45.9 ± 0.3	$\mathbf{47.8 \pm 0.1}$	93.9 ± 0.2	$\mathbf{94.6 \pm 0.2}$	94.0 ± 0.2
Prostate (6919)	5.4 ± 2.7	8.1 ± 0.2	$\mathbf{9.1 \pm 0.1}$	77.6 ± 1.8	$\mathbf{80.6 \pm 0.2}$	73.1 ± 0.5
Throat (42743)	26.7 ± 2.4	28.0 ± 0.3	$\mathbf{32.3 \pm 0.1}$	$\mathbf{91.5 \pm 1.3}$	88.6 ± 0.8	86.8 ± 1.0
scGEM	26.9 ± 3.7	33.0 ± 1.1	$\mathbf{39.3 \pm 0.7}$	95.0 ± 1.3	96.2 ± 0.6	$\mathbf{96.8ß \pm 0.3}$
SNAREseq	6.8 ± 6.0	35.8 ± 5.2	$\mathbf{67.9ß \pm 1.2}$	93.1 ± 2.8	99.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$

- Comparison for different DR methods.
- Silhouette (clustering) and Trustworthiness (spatial relations) scores reported.
- t-SNEkhorn is state of the art on majority of criterion/datasets.

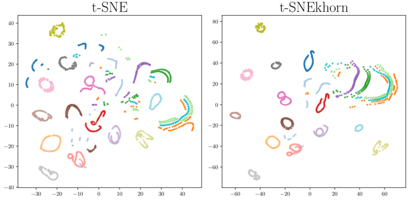
DR vizualisation on real life datasets

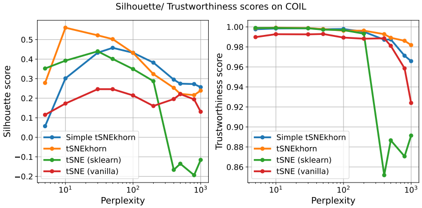
SNAREseq Sincle Cell dataset





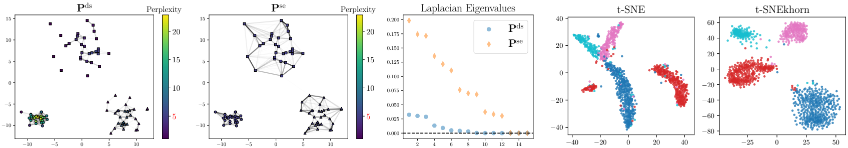
COIL 20 Image dataset





Conclusion

Conclusion



Symmetric entropic affinities and SNEkhorn

- We propose a symmetric affinity matrix that controls both the L1 norm and entropy of the rows/columns.
- We show its robustness to heteroscedastic noise (important for single cell data).
- Based on this affinity, we propose a new DR method : (t)SNEkhorn.
- Python code available at : `https://github.com/PythonOT/SNEkhorn`

Future works

- Implement SNEkhorn with all the (t)SNE accelerations.
- OT with point-wize entropy constraint [Van Assel et al., 2023a]
- Relations between Gromov-Wasserstein and DR [Van Assel et al., 2023b].

Thank you

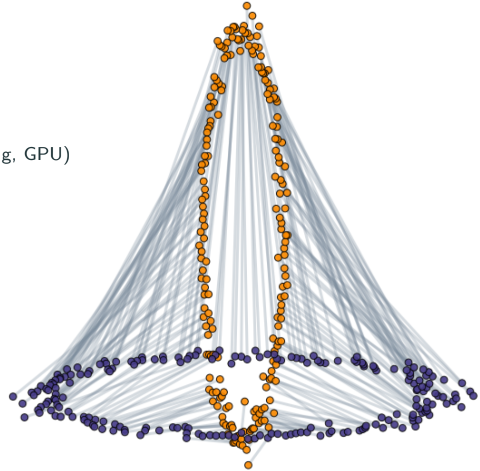
Python code available on GitHub:

Python code available on GitHub:
https://github.com/PythonOT/POT

- OT LP solver, Sinkhorn (stabilized, $\epsilon-$scaling, GPU)
- Domain adaptation with OT.
- Barycenters, Wasserstein unmixing.
- Wasserstein Discriminant Analysis.

Tutorial on OT for ML:
http://tinyurl.com/otml-isbi

Papers available on my website:
https://remi.flamary.com/



References i

Cuturi, M. (2013).

Sinkhorn distances: Lightspeed computation of optimal transportation.

In *Neural Information Processing Systems (NIPS)*, pages 2292–2300.

Feltes, B. C., Chandelier, E. B., Grisci, B. I., and Dorn, M. (2019).

Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research.

Journal of Computational Biology, 26(4):376–386.

PMID: 30789283.

Hinton, G. E. and Roweis, S. (2002).

Stochastic neighbor embedding.

Advances in neural information processing systems, 15.

Schölkopf, B. and Smola, A. J. (2002).

Learning with kernels: Support vector machines, regularization, optimization, and beyond.

MIT press.

Van Assel, H., Vayer, T., Flamary, R., and Courty, N. (2023a).
Optimal transport with adaptive regularisation.

Van Assel, H., Vincent-Cuaz, C., Vayer, T., Flamary, R., and Courty, N. (2023b).
Interpolating between clustering and dimensionality reduction with gromov-wasserstein.

Van der Maaten, L. and Hinton, G. (2008).
Visualizing data using t-sne.
Journal of Machine Learning Research, 9(2579-2605):85.

Von Luxburg, U. (2007).
A tutorial on spectral clustering.
Statistics and computing, 17:395–416.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021).
Barlow twins: Self-supervised learning via redundancy reduction.
In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Zelnik-Manor, L. and Perona, P. (2004).
Self-tuning spectral clustering.
Advances in neural information processing systems, 17.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003).
Learning with local and global consistency.
Neural Information Processing Systems (NeurIPS), 16.