Entropic Wasserstein Component Analysis

Antoine Collas[1], Titouan Vayer[2], Rémi Flamary[3], Arnaud Breloy[4]

[1]University Paris-Saclay, Inria, CEA
[2]University Lyon, Inria, CNRS, ENS Lyon, UCB Lyon 1, LIP UMR 5668
[3]Ecole Polytechnique, Institut Polytechnique de Paris, CMAP, UMR 7641
[4]LEME, University Paris Nanterre

Available in Python Optimal Transport library:

Subspace learning

Given data $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \in (\mathbb{R}^d)^n$, the goal is to find a subspace $\boldsymbol{U}$ such that $\boldsymbol{x}_i \approx \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i$.

Principal Component Analysis (PCA):

$$\boldsymbol{U}^{\mathrm{PCA}} \in \underset{\boldsymbol{U} \in \mathrm{St}(d,k)}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i\|_2^2$$

with $\mathrm{St}(d,k) \triangleq \left\{\boldsymbol{U} \in \mathbb{R}^{d\times k} \mid \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_k\right\}$.

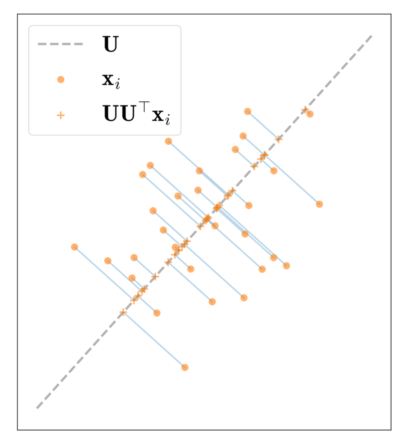

Figure 1. PCA illustration

Optimal transport (OT)

Given $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n), (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n) \in (\mathbb{R}^d)^n$ and their empirical measures

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i} \quad \text{and} \quad \nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{z}_i}$$

the squared 2-Wasserstein distance with the ℓ_2 metric is [1]

$$\mathcal{W}_2^2(\mu,\nu) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\text{minimize}} \sum_{i,j}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2$$

with

$$\boldsymbol{\Pi} \triangleq \left\{\boldsymbol{\pi} \in \mathbb{R}^{n\times n} \mid \pi_{ij} \geq 0,\, \boldsymbol{\pi}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n,\, \boldsymbol{\pi}^\top\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n\right\}.$$

Entropic regularized OT

Given the entropy $\mathrm{H}(\boldsymbol{\pi}) \triangleq -\sum_{i,j}^{n,n} \pi_{ij}\log\pi_{ij}$ and $\varepsilon > 0$,

$$\underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\text{minimize}} \sum_{i,j}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2 - \varepsilon\,\mathrm{H}(\boldsymbol{\pi}).$$

Solved with the Sinkhorn-Knopp algorithm.

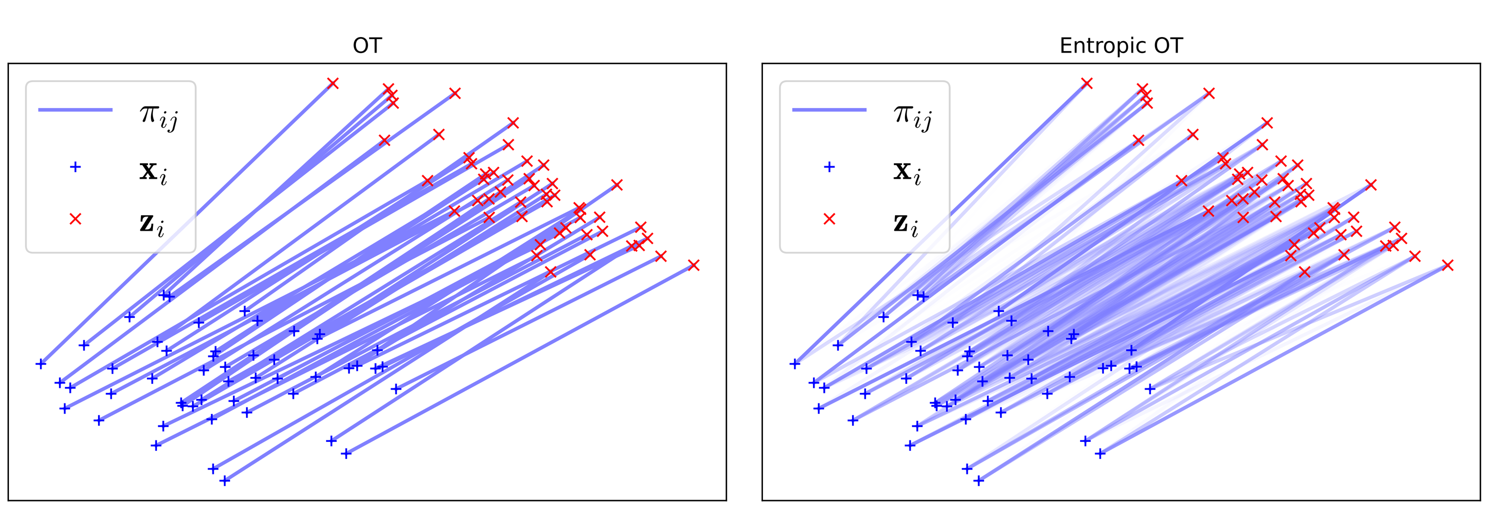

Figure 2. OT creates a one-to-one correspondence between the two datasets (left). Entropic regularization allows for non one-to-one correspondences (right). Figure adapted from POT library [2].

Motivation of the paper

Given the empirical measures

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i} \quad \text{and} \quad \nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_i},$$

$$\boldsymbol{U}^{\mathrm{PCA}} \in \underset{\boldsymbol{U} \in \mathrm{St}(d,k)}{\arg\min} \left\{\mathcal{W}_2^2(\mu,\nu) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}}{\text{minimize}} \sum_{i,j}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_j\|_2^2\right\}$$

and optimal coupling

$$\boldsymbol{\pi}^\star = \frac{1}{n}\boldsymbol{I}_n.$$

Motivation: add entropy to have $\boldsymbol{\pi}^\star \neq \frac{1}{n}\boldsymbol{I}_n$ and thus minimize reconstruction error of clusters (and not points).

Problem formulation

Entropic Wasserstein Component Analysis (EWCA):

$$\underset{\substack{\boldsymbol{\pi}\in\boldsymbol{\Pi}\\ \boldsymbol{U}\in\mathrm{St}(d,k)}}{\text{minimize}} \sum_{i,j=1}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_j\|_2^2 - \varepsilon\,\mathrm{H}(\boldsymbol{\pi}).$$

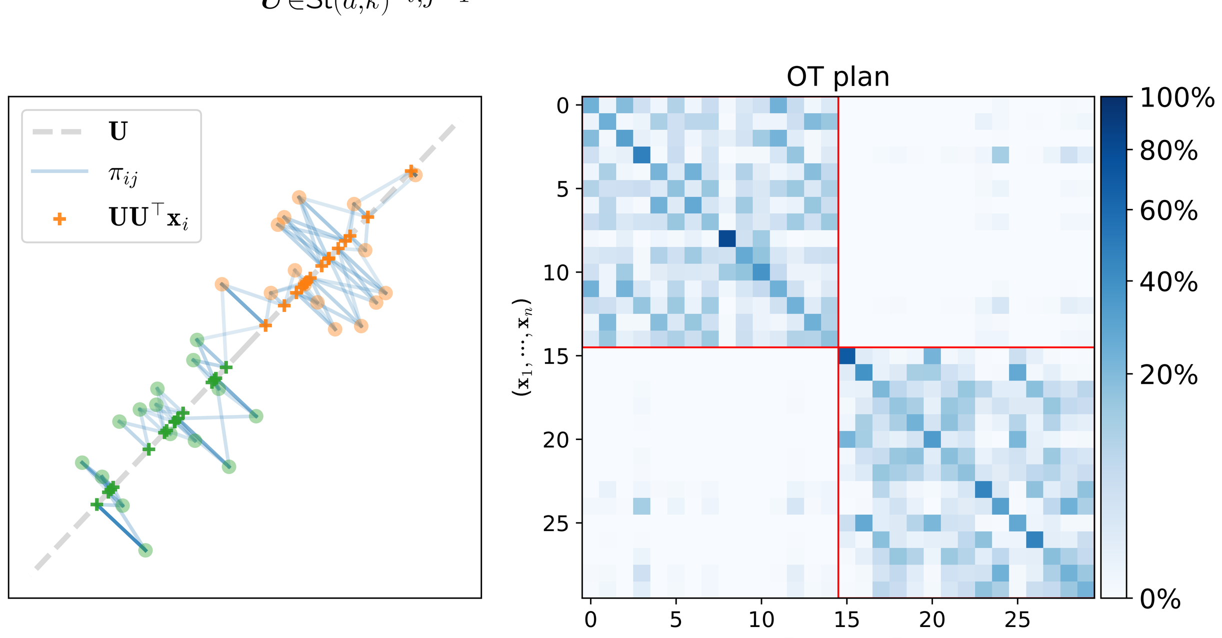

Figure 3. Illustration of EWCA on a two-class dataset. On the left are the samples and their 1D projections, and on the right is the corresponding OT plan.

Limit cases & interpretation

Denoting

$$(\boldsymbol{\pi}_\varepsilon, \boldsymbol{U}_\varepsilon) = \underset{\substack{\boldsymbol{\pi}\in\boldsymbol{\Pi}\\ \boldsymbol{U}\in\mathrm{St}(d,k)}}{\arg\min} \sum_{i,j=1}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_j\|_2^2 - \varepsilon\,\mathrm{H}(\boldsymbol{\pi})$$

- $\varepsilon \to 0 \implies \boldsymbol{\pi}_\varepsilon \to \frac{1}{n}\boldsymbol{I}_n$ and $\boldsymbol{U}_\varepsilon \to$ top k eigenvectors $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$; we recover PCA !
- $\varepsilon \to +\infty \implies \boldsymbol{\pi}_\varepsilon \to \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\boldsymbol{U}_\varepsilon \to$ last k eigenvectors of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$.

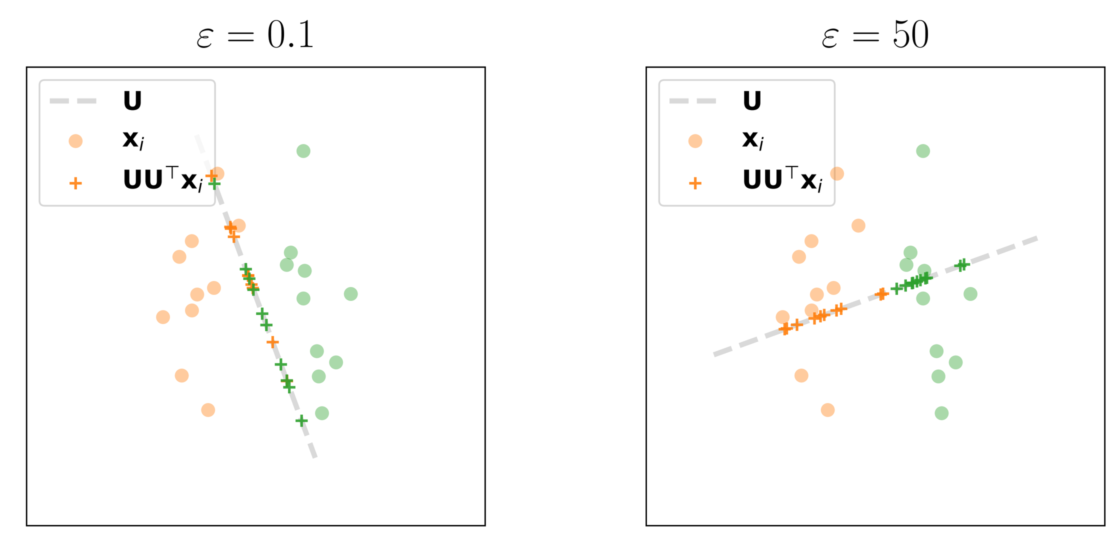

Figure 4. Illustration of EWCA for different values of ε on a two-class dataset.

Resolution: block coordinate descent

Given the estimate $(\boldsymbol{\pi}^{(t)}, \boldsymbol{U}^{(t)})$,

- $\boldsymbol{\pi}$-step: compute $\boldsymbol{\pi}^{(t+1)}$ using Sinkhorn-Knopp algorithm,
- $\boldsymbol{U}$-step: compute $\boldsymbol{U}^{(t+1)}$ as the k first eigenvectors of

$$\boldsymbol{X}\left(2\,\mathrm{sym}(\boldsymbol{\pi}^{(t+1)}) - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\boldsymbol{X}^\top.$$

Block-majorization-minimization algorithm to avoid SVD of a $d\times d$ matrix in the paper.

Application

Datasets of gene expressions:

- Breast: $d = 54675$, $n = 151$, and 6 classes [3],
- Khan2001: $d = 2308$, $n = 63$, and 4 classes [4].

Classification setup:

- 1-Nearest neighbor classifier on the projected data $\boldsymbol{U}^\top\boldsymbol{x}_i$,
- evaluation over 100 random splits of the data (50% training, 50% testing),
- hyperparameter ε tuned by cross-validation on the training set.

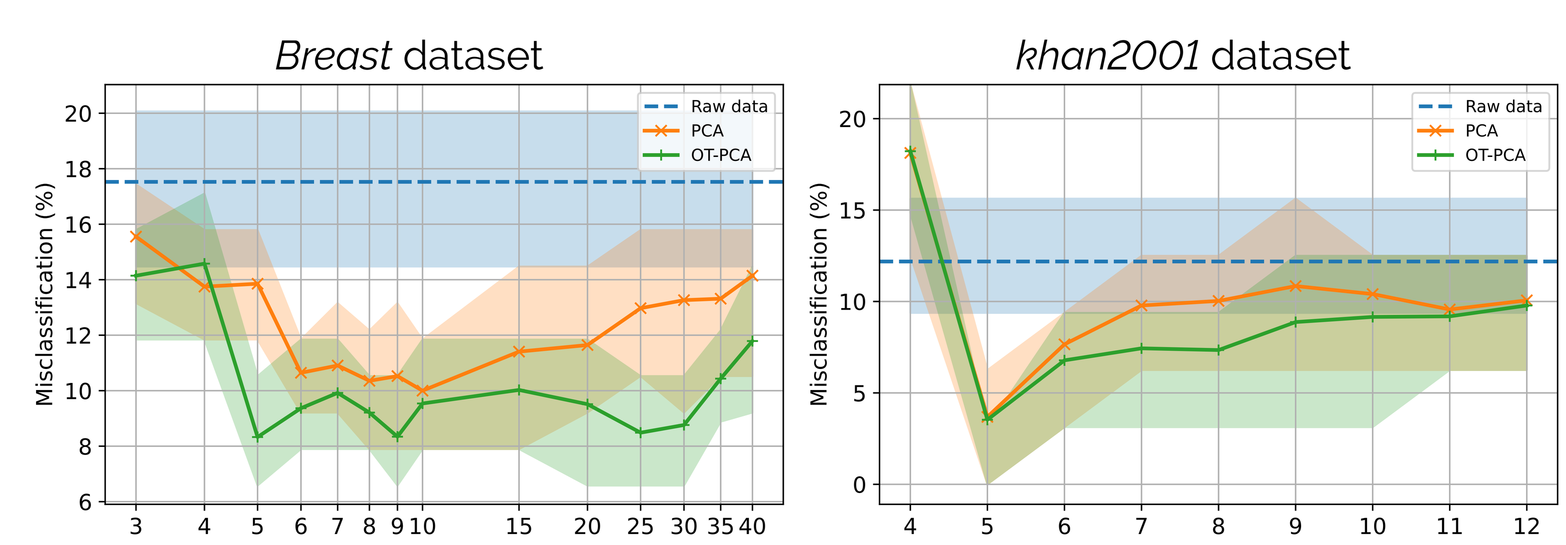

Figure 5. Misclassification rate (%) versus subspace dimension k (the lower the better). Mean, 1st and 3rd quartiles are reported.

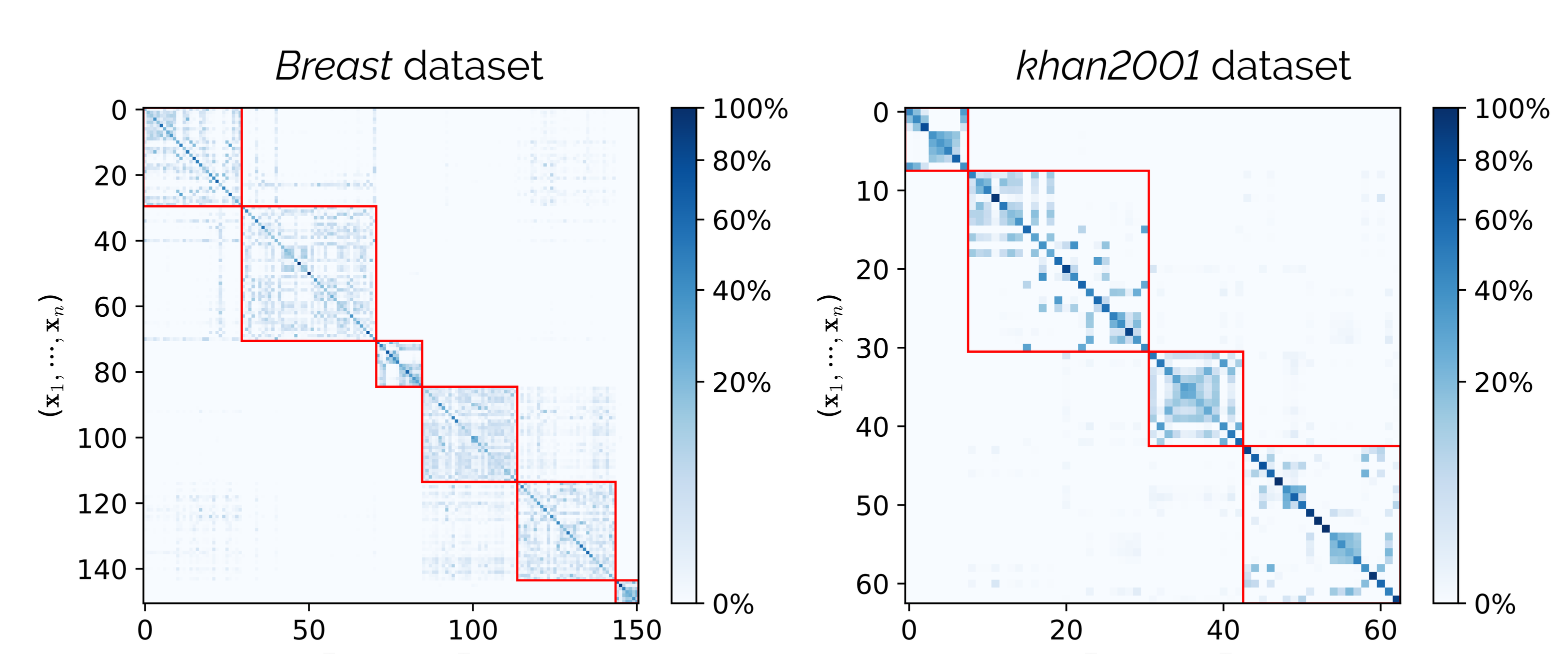

Figure 6. OT plan $\boldsymbol{\pi}$ (%) computed with EWCA ($k = 5$). The red squares enclose the data belonging to the same class.

References

[1] G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science.* 2019.

[2] R. Flamary et al. "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8. URL: `http://jmlr.org/papers/v22/20-451.html`.

[3] B. C. Feltes et al. "Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research". In: *Journal of Computational Biology* 26.4 (2019), pp. 376–386.

[4] J. Khan et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks". In: *Nat Med* 7.6 (2001), pp. 673–679. ISSN: 10788956. URL: `http://dx.doi.org/10.1038/89044`.

antoine.collas@inria.fr