Optimal transport and dimension reduction: Entropic Wasserstein Component Analysis

Antoine Collas

Postdoc supervised by Alexandre Gramfort and Rémi Flamary

Paper published at MLSP (Machine Learning for Signal Processing) 2023, Rome

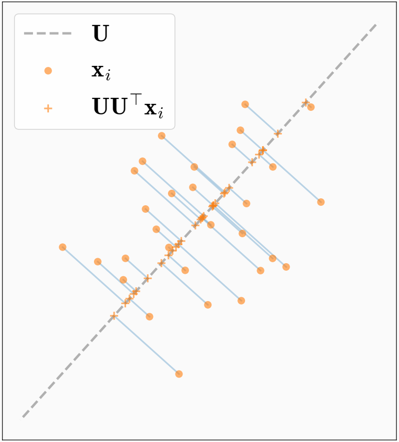
Work done with Titouan Vayer and Arnaud Breloy

A few reminders

Principal Component Analysis (PCA)

Subspace learning from data: $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) \in (\mathbb{R}^d)^n$.

Goal: find a subspace $\boldsymbol{U}$ such that $\boldsymbol{x}_i \approx \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i$.



Principal Component Analysis (PCA)

PCA: solution to

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

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

Solution computation:

$$\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \overset{\mathrm{SVD}}{=} \left[\boldsymbol{U}^{\mathrm{PCA}} \middle| \, \boldsymbol{U}_\perp \right] \boldsymbol{\Sigma} \boldsymbol{V}^\top$$

Optimal Transport (OT): Wasserstein distance

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

$$\mathcal{W}_2^2(\mu, \nu) = \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}(\frac{1}{n}\mathbf{1}_n, \frac{1}{n}\mathbf{1}_n)}{\text{minimize}} \sum_{i,j}^{n,n} \pi_{ij} \|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2$$

with

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

[Peyré et al. 2019]

Optimal Transport (OT): entropic regularization

Entropic regularized OT:

$$\underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\frac{1}{n}\mathbf{1}_n,\frac{1}{n}\mathbf{1}_n)}{\text{minimize}} \sum_{i,j}^{n,n} \pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2 - \varepsilon\,\mathrm{H}(\boldsymbol{\pi})$$

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

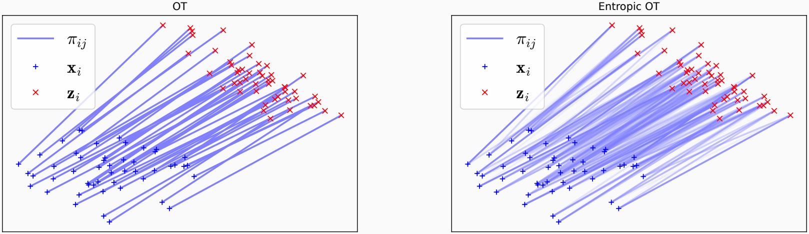


Figure adapted from POT library [Flamary et al. 2021]

Optimal Transport (OT): Sinkhorn-Knopp algorithm

Solution to the entropic regularized OT problem:

$$\boldsymbol{\pi} = \mathrm{diag}(\boldsymbol{u}) \boldsymbol{K} \, \mathrm{diag}(\boldsymbol{v})$$

with

$$K_{ij} \triangleq \exp(-\|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2 / \varepsilon)$$

and $\boldsymbol{u}$ and $\boldsymbol{v}$ obtained by iterating

$$\boldsymbol{u} \leftarrow \frac{1}{n} \mathbf{1}_n \oslash \boldsymbol{K} \boldsymbol{v}$$
$$\boldsymbol{v} \leftarrow \frac{1}{n} \mathbf{1}_n \oslash \boldsymbol{K}^\top \boldsymbol{u}.$$

[Cuturi 2013]

Entropic Wasserstein Component Analysis (EWCA)

Motivation

Given the empirical measures

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

we have

$$\boldsymbol{U}^{\text{PCA}} \in \underset{\boldsymbol{U}\in\text{St}(d,k)}{\arg\min}\left\{\mathcal{W}_2^2(\mu,\nu) = \underset{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\frac{1}{n}\mathbf{1}_n,\frac{1}{n}\mathbf{1}_n)}{\text{minimize}}\sum_{i,j}^{n,n}\pi_{ij}\|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_j\|_2^2\right\}$$

and optimal coupling

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

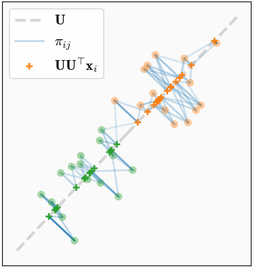
Motivation: add entropy to have $\boldsymbol{\pi}^\star \neq \frac{1}{n}\boldsymbol{I}_n$ and thus minimize reconstruction error of clusters (and not points).

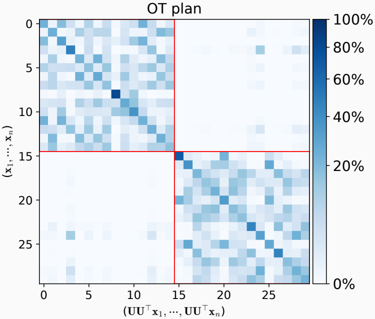
Entropic Wasserstein Component Analysis (EWCA) problem

Entropic Wasserstein Component Analysis (EWCA):

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

where H is the entropy function.



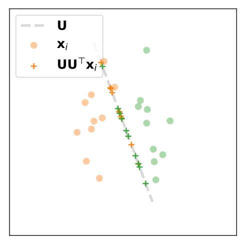


Entropic Wasserstein Component Analysis (EWCA) problem

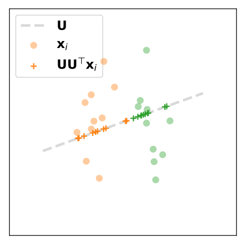
Limit cases:

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

- $\varepsilon \to 0 \implies \boldsymbol{\pi}_\varepsilon \to \frac{1}{n}\boldsymbol{I}_n$ 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 $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$.



(a) $\varepsilon = 0.1$



(b) $\varepsilon = 50$

Block coordinate descent algorithm

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

Given the current 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

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

Problem: $\boldsymbol{U}$-step requires SVD of a $d \times d$ matrix.

Majorization-minimization (MM) algorithm

Majorization-minimization over the Stiefel manifold

$$\underset{\boldsymbol{U}\in\mathrm{St}(d,k)}{\text{minimize}}\ f(\boldsymbol{U})$$

Given iterate $\boldsymbol{U}^{(t)}$,

- Majorization:

$$f(\boldsymbol{U}) \leq g(\boldsymbol{U}|\boldsymbol{U}^{(t)}) \text{ for all } \boldsymbol{U} \in \mathrm{St}(d,k)$$

such that

$$g(\boldsymbol{U}|\boldsymbol{U}^{(t)}) = 2\,\mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{M}\boldsymbol{U}^{(t)}) + \text{const. (linearity)}$$

for some $\boldsymbol{M} \in \mathbb{R}^{d\times d}$.

- Minimization:

$$\boldsymbol{U}^{(t+1)} = \mathrm{pf}(-\boldsymbol{M}\boldsymbol{U}^{(t)}) = \underset{\boldsymbol{U}\in\mathrm{St}(d,k)}{\arg\min}\ g(\boldsymbol{U}|\boldsymbol{U}^{(t)})$$

where pf returns the orthogonal factor of the polar decomposition.

[Breloy et al. 2021]

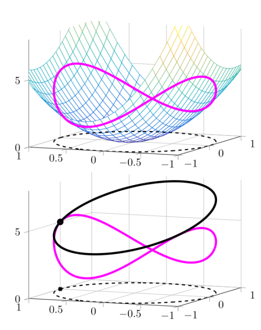


Figure 2: A quadratic form over $\mathrm{St}(2,1)$ (pink) and its surrogate (black). Figure from [Breloy et al. 2021].

Majorization-minimization over the Stiefel manifold

$\boldsymbol{U}$-step:

$$\underset{\boldsymbol{U} \in \mathrm{St}(d,k)}{\text{minimize}} \left\{ \sum_{i,j=1}^{n,n} \pi_{ij} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_j\|_2^2 \propto \mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{M} \boldsymbol{U}) \right\}$$

for some $\boldsymbol{M}^\top = \boldsymbol{M}$ and $\boldsymbol{M} \preccurlyeq \mathbf{0}$ (negative semi-definite).

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

- Majorization (by concavity):

$$\mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{M} \boldsymbol{U}) \leq 2\,\mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{M} \boldsymbol{U}^{(t)}) + \text{const.},$$

- Minimization:

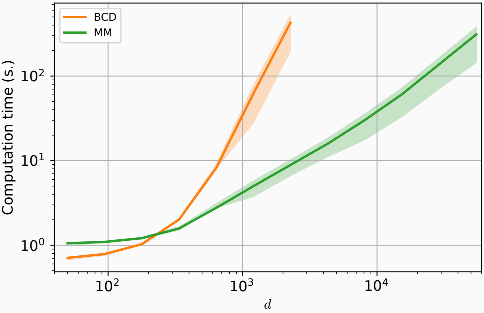
$$\boldsymbol{U}^{(t+1)} = \mathrm{pf}(-\boldsymbol{M}\boldsymbol{U}^{(t)})$$

BCD vs block-MM: computational complexity

Overall computational complexity per iteration:

- BCD: $\mathcal{O}(n^2 d + nd^2 + d^3)$,
- Block-MM: $\mathcal{O}(n^2 d)$.

Lower complexity of the block-MM but requires more iterations...



Numerical experiments

Numerical experiments: classification

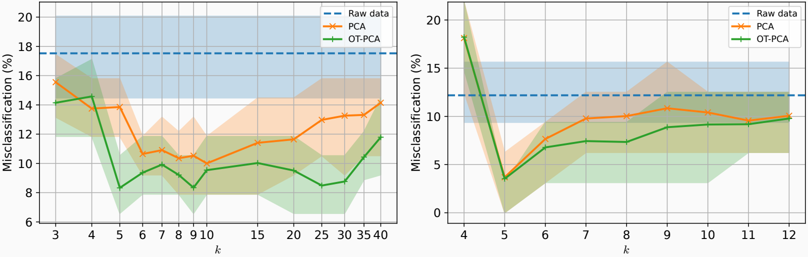
Datasets of gene expressions:

- Breast: $d = 54675$, $n = 151$, and 6 classes [Feltes et al. 2019],
- Khan2001: $d = 2308$, $n = 63$, and 4 classes [Khan et al. 2001].

Classification pipeline:

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

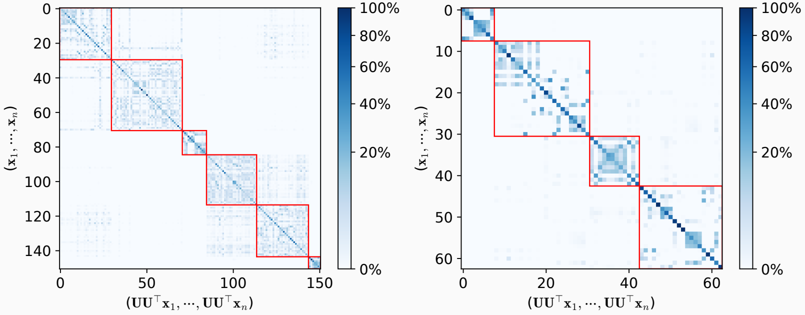
Numerical experiments: classification



(a) *Breast* dataset **(b)** *khan2001* dataset

Figure 3: Misclassification rate (%) versus subspace dimension k (the lower the better). Mean, 1$^{\text{st}}$ and 3$^{\text{rd}}$ quartiles are reported.

Numerical experiments: transport plan



(a) *Breast* dataset

(b) *khan2001* dataset

Figure 4: Transport plan π (%) computed with EWCA ($k = 5$). The red squares enclose the data belonging to the same class.

Numerical experiments: TSNE

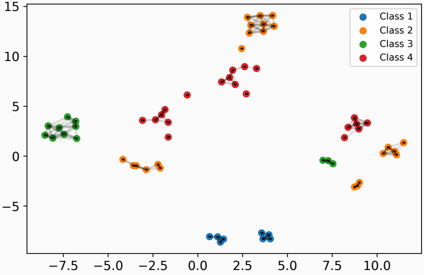


Figure 5: TSNE of the projected data $(\boldsymbol{U}^\top \boldsymbol{x}_1, \cdots, \boldsymbol{U}^\top \boldsymbol{x}_n)$ ($k = 5$) computed with EWCA on the *Khan2001* dataset. The grey links represent the intensity of the values of the transport plan.

Numerical experiments: sensitivity to entropy regularization

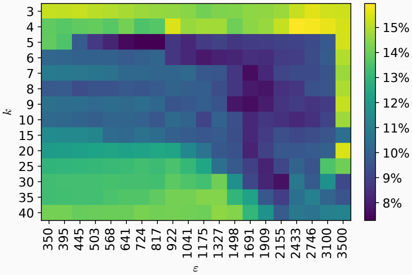


Figure 6: Misclassification rate (%) versus subspace dimension k and entropy intensity ε on the *Breast* dataset (the lower the better).

Conclusions

- Generalization of PCA that takes into account the neigbourhood of data,
- optimization with a BCD and a block-MM,
- use in place of PCA in a classification pipeline on two gene expressions datasets.

Preprint available at

https://arxiv.org/abs/2303.05119

Code available at

github.com/antoinecollas/Entropic_Wasserstein_Component_Analysis

Implemented in POT library

https://pythonot.github.io/master/auto_examples/others/plot_EWCA.html

References

Breloy, A. et al. (2021). “Majorization-minimization on the Stiefel manifold with application to robust sparse PCA”. In: *IEEE Transactions on Signal Processing* 69, pp. 1507–1520.

Cuturi, M. (2013). “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in neural information processing systems* 26.

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

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

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

Peyré, G. et al. (2019). *Computational Optimal Transport: With Applications to Data Science*.

Optimal transport and dimension reduction: Entropic Wasserstein Component Analysis

Antoine Collas

Postdoc supervised by Alexandre Gramfort and Rémi Flamary

Paper published at MLSP (Machine Learning for Signal Processing) 2023, Rome

Work done with Titouan Vayer and Arnaud Breloy