Wasserstein Dictionary Learning

Morgan A. Schmitz*, Matthieu Heitz, Nicolas Bonneel, Fred Ngolè, David Coeurjolly, Marco Cuturi, Gabriel Peyré, Jean-Luc Starck*

*CosmoStat, Astrophysics Dept., IRFU, CEA Saclay | morgan.schmitz@cea.fr

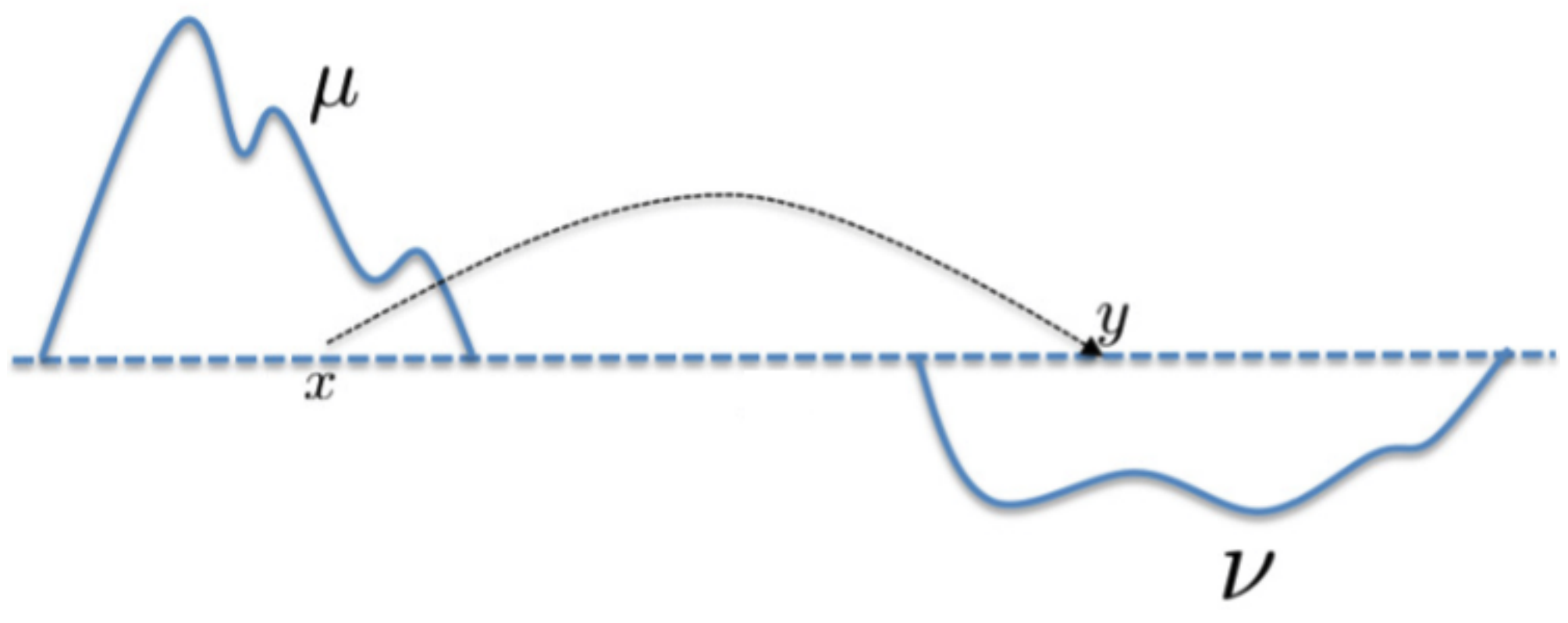
CosmoStat

Introduction

- Optimal Transport (OT) theory allows for the definition of a distance on all measures of a given set.
- In the discrete case, most data can be recast as histograms, *i.e.* discrete measures.
- By definition, OT distances capture the warping between two histograms.
- A new method, analogous to dictionary Learning, but making full use of the OT geometry, is introduced to obtain a non-linear representation of data that exploits the attractive properties of OT.

Optimal Transport distances

Overview



Graphical representation of the mass transportation problem: find the optimal way of moving a heap of sand μ into a hole ν knowing the cost of moving grains of sand to and from any position.

Wasserstein distance

- In the discrete case, histograms μ and ν are vectors in $\mathbb{R}^N$ and the cost function can be contained within a matrix $C \in \mathbb{R}^{N\times N}$.
- The solution to the mass transportation problem defines an OT distance:

$$W(\mu,\nu) := \min_{T\in\Pi(\mu,\nu)} \langle T, C\rangle.$$

- $\Pi(\mu,\nu)$ is the set of admissible transport plans, the discrete equivalent of bivariate measures with marginals μ,ν:

$$\Pi(\mu,\nu) := \left\{T \in \mathbb{R}_+^{N\times N}, T\mathbb{1}_N = \mu, T^\top \mathbb{1}_N = \nu\right\}.$$

- In the particular case where C corresponds to a metric on the grid, W is called Wasserstein distance.

Numerical Optimal Transport

- Despite its simple formulation, practical computation of Wasserstein distances quickly reached a prohibitive cost until the recent introduction of numerical approximations.
- In particular, the addition of an entropic penalty term [Cuturi (2013)] to the definition of the Wasserstein distance yields:

$$W_\gamma(\mu,\nu) := \min_{T\in\Pi(\mu,\nu)} \langle T, C\rangle + \gamma H(T),$$

where $H(T) := \sum_{i,j} T_{ij}\log(T_{ij}-1)$.

- This makes the problem strictly convex and allows the use of the Sinkhorn algorithm [Sinkhorn (1967)] for linear convergence to W_γ by simple iterative matrix scalings.

Wasserstein barycenter

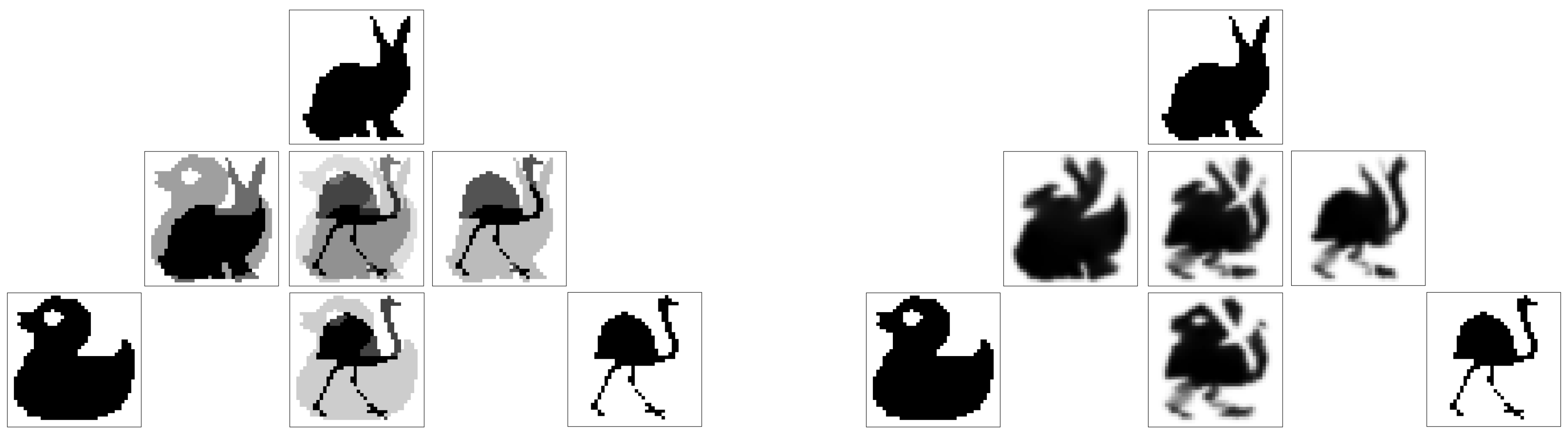
Definition

- By analogy with the Euclidean barycenter, for any input histograms $d_1,\dots,d_S$ and weights $\lambda_1,\dots,\lambda_S$, define [Agueh & Carlier (2011)] the Wasserstein barycenter as:

$$P(D,\lambda) = \underset{u}{\operatorname{argmin}} \sum_{s=1}^{S} \lambda_s W(u, d_s)$$

- When using the entropic penalty within that definition, a generalization of the Sinkhorn algorithm allows for fast computation of these barycenters by iterative scalings [Benamou et al. (2015)].

Illustration



(a) Euclidean simplex (b) Wasserstein simplex

Wasserstein Dictionary Learning

Rationale

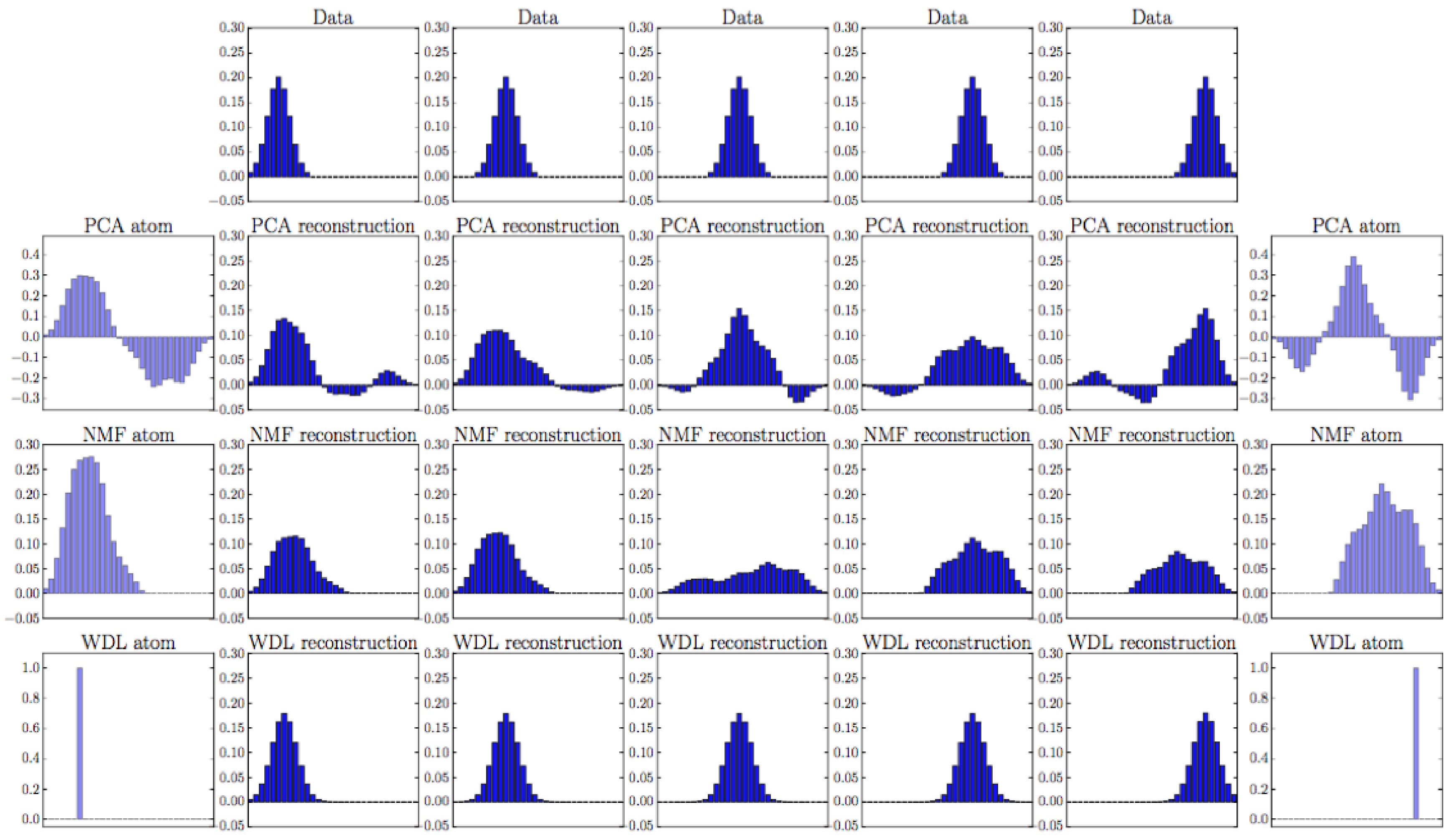
- Usual dictionary learning aims at representing data X using a dictionary, D, and a set of codes Λ so that $X \approx D\Lambda$.
- Adding constraints on either or both of these components can give the learned representation desirable properties: sparsity, positivity (NMF), etc.
- Ultimately, the relationship between the reconstructed data and the dictionary atoms remains linear.
- Our method breaks free from this constraint by replacing the matrix dot-product with the Wasserstein barycenter operator, *i.e.* we learn a representation such that $X \approx P(D,\Lambda)$.
- This not only allows for a non-linear dictionary learning method, but also one that leverages the natural OT property of accounting for the warping of histograms.

Automatic Differentiation

- The learning stage is performed using a descent method to minimize some arbitrary similarity criterion.
- The gradients in dictionary and atoms are obtained through automatic differentiation [Griewank & Walther (2008)].
- The algorithm is differentiated instead of the actual barycenter operator, allowing for computation by repeated applications of the chain rule.
- This approach in our case is very close to backward propagation, as made popular by deep learning.

Application

- Dataset consists of translated, discretized 1D Gaussians on a small grid.
- PCA, NMF and our approach are applied to learn only 2 components/atoms.



- Our method reconstructs Gaussians, as opposed to the linear approaches wherein neither the atoms nor the reconstructions are histograms.

Conclusion

- We introduce a new unsupervised method, analogous to dictionary learning.
- Because we learn our representation using the OT geometry (in particular, Wasserstein barycenters), our approach is non-linear and captures the warping between datapoints.

References

Agueh M. & Carlier G. 2011, SIAM Journal on Mathematical Analysis, 43, 904

Benamou J.-D., Carlier G., Cuturi M., Nenna L. & Peyré G. 2015, SIAM Journal on Scientific Computing, 37, A1111

Cuturi M. 2013, in Advances in Neural Information Processing Systems, 2292–2300

Griewank A. & Walther A. 2008, Evaluating derivatives: principles and techniques of algorithmic differentiation (SIAM)

Sinkhorn R. 1967, The American Mathematical Monthly, 74, 402

Acknowledgments

This work was supported by the Centre National d'Etudes Spatiales (CNES) and the European Community through the grant DEDALE (contract no. 665044) within the H2020 Framework Program.

Data Science Summer School, Palaiseau, France — August 28th–September 1st, 2017