A Wasserstein-type distance in the space of Gaussian mixture models

Julie Delon, Agnès Desolneux

CNRS and Ecole Normale Supérieure Paris-Saclay

GDR MIA Thematic Days on High-Dimensional Data Analysis

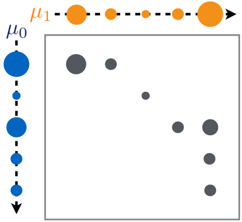
Marseille, vendredi 22 octobre 2021

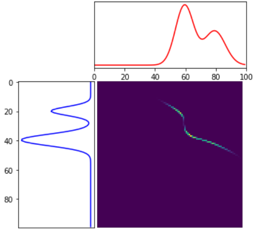
Wasserstein distance

Let μ_0 and μ_1 be two probability measures on $\mathbb{R}^d$, then the 2-Wasserstein distance between μ_0 and μ_1 is defined by

$$W_2^2(\mu_0, \mu_1) := \inf_{Y_0 \sim \mu_0; Y_1 \sim \mu_1} \mathbb{E}\left(\|Y_0 - Y_1\|^2\right) = \inf_{\gamma \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y_0 - y_1\|^2 d\gamma(y_0, y_1),$$

where $\Pi(\mu_0, \mu_1) \subset \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ is the subset of probability distributions γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distributions μ_0 and μ_1.





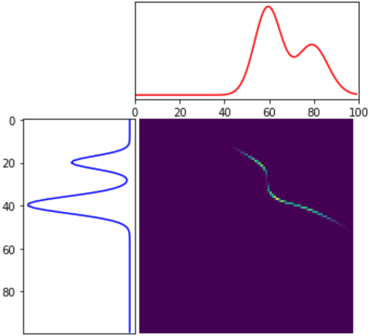
[1] C. Villani, *Optimal transport : old and new*, 2008.
[2] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, 2015.
[3] G. Peyré and M. Cuturi, *Computational optimal transport*, 2019.

Optimal transport

If μ_0 is absolutely continuous, then it can be shown that the optimal transport plan γ is unique and has the form

$$\gamma = (\mathrm{Id}, T)\#\mu_0,$$

where $T : \mathbb{R}^d \mapsto \mathbb{R}^d$ is an application called *optimal transport map* and satisfying $T\#\mu_0 = \mu_1$.



Notation : $T\#\mu(A) = \mu(T^{-1}(A))$.

Barycenter

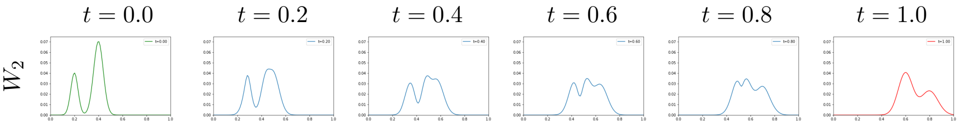
If γ is an optimal transport plan for W_2 between two probability distributions μ_0 and μ_1, the path $(\mu_t)_{t\in[0,1]}$ given by

$$\forall t \in [0,1], \quad \mu_t := P_t\#\gamma, \quad \text{where} \quad P_t(x,y) = (1-t)x + ty,$$

defines a geodesic in $\mathcal{P}_2(\mathbb{R}^d)$.

The path $(\mu_t)_{t\in[0,1]}$ is called the displacement interpolation between μ_0 and μ_1 and it satisfies

$$\mu_t \in \operatorname{argmin}_\rho \ (1-t)W_2(\mu_0,\rho)^2 + tW_2(\mu_1,\rho)^2.$$



Optimal transport between Gaussian distributions

If $\mu_i = \mathcal{N}(m_i, \Sigma_i)$, $i \in \{0,1\}$ are two Gaussian distributions on $\mathbb{R}^d$, then

$$W_2^2(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \mathrm{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right),$$

where, for every symmetric semi-definite positive matrix M, the matrix $M^{\frac{1}{2}}$ is its unique semi-definite positive square root.

If Σ_0 is non-singular, then the optimal map T between μ_0 and μ_1 is affine and given by

$$\forall x \in \mathbb{R}^d, \quad T(x) = m_1 + \Sigma_0^{-\frac{1}{2}}\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x - m_0) = m_1 + \Sigma_0^{-1}(\Sigma_0\Sigma_1)^{\frac{1}{2}}(x - m_0),$$

and the optimal plan γ is then a degenerate Gaussian distribution on $\mathbb{R}^{2d}$, supported by the affine line $y = T(x)$.

Moreover, if Σ_0 and Σ_1 are non-degenerate, the geodesic path (μ_t), $t \in (0,1)$, between μ_0 and μ_1 is given by $\mu_t = \mathcal{N}(m_t, \Sigma_t)$ with $m_t = (1-t)m_0 + tm_1$ and

$$\Sigma_t = ((1-t)I_d + tC)\Sigma_0((1-t)I_d + tC),$$

with I_d the $d \times d$ identity matrix and $C = \Sigma_1^{\frac{1}{2}}\left(\Sigma_1^{\frac{1}{2}}\Sigma_0\Sigma_1^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Sigma_1^{\frac{1}{2}}$.

dowson1982frechet

Applications

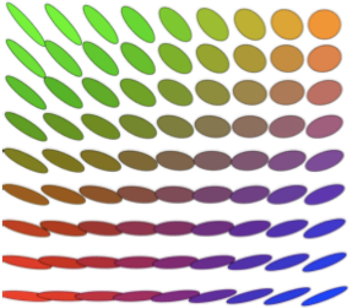
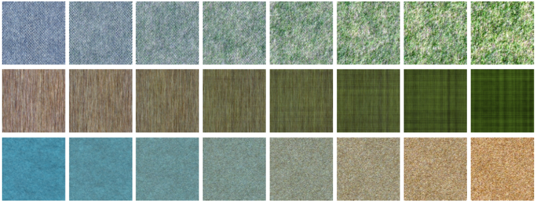


illustration: Cuturi, Peyré, OT book



Texture mixing [Xia et al, 2014]

Gaussian Mixture Models (GMM)

Definition :
Let $K \geqslant 1$ be an integer. A Gaussian mixture model of size K on $\mathbb{R}^d$ is a probability distribution μ on $\mathbb{R}^d$ that can be written

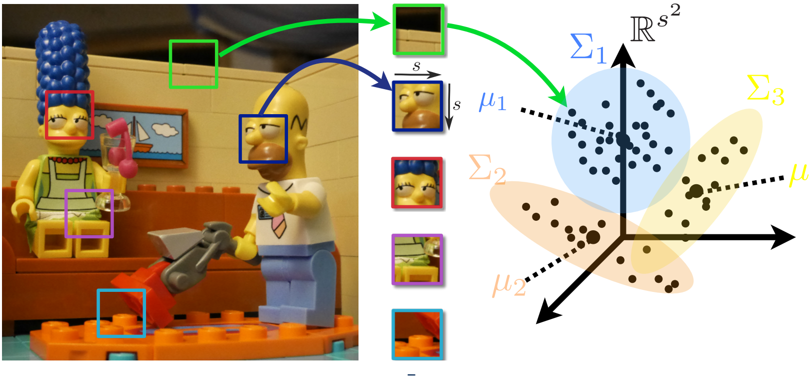
$$\mu = \sum_{k=1}^{K} \pi_k \mu_k \text{ where } \mu_k = \mathcal{N}(m_k, \Sigma_k) \text{ and } \pi \in \mathbb{R}_+^K, \sum_{k=1}^{K} \pi_k = 1.$$

Notation : This set is denoted $GMM_d(K)$, and let

$$GMM_d(\infty) = \cup_{K \geqslant 1} GMM_d(K).$$

Remark : Inference from samples via EM algorithm.

GMM on patches



$\longrightarrow$ Many applications for image restoration, image editing (style transfer, inpainting), texture synthesis, etc.

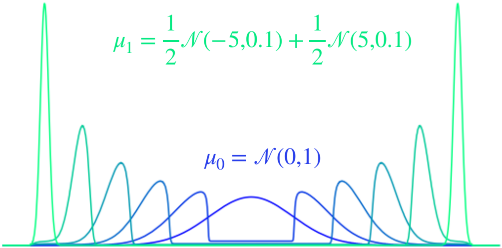
Optimal transport between GMM

OT plans between GMM : usually not GMM themselves. Same remark for barycenters !

Example : $\mu_0 = \mathcal{N}(0,1)$ and $\mu_1 = \frac{1}{2}(\delta_{-1} + \delta_1)$. Then μ_t has a density

$$f_t(x) = \frac{1}{1-t}\left(g\left(\frac{x+t}{1-t}\right) \mathbf{1}_{x<-t} + g\left(\frac{x-t}{1-t}\right) \mathbf{1}_{x>t} \right),$$

where g is the density of $\mathcal{N}(0,1)$.



Restricting the set of couplings : MW_2

Definition

Let μ_0 and μ_1 be two Gaussian mixture models. We define the Mixture-restricted Wasserstein distance by

$$MW_2^2(\mu_0, \mu_1) := \inf_{\gamma \in \Pi(\mu_0,\mu_1) \cap GMM_{2d}(\infty)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y_0 - y_1\|^2 d\gamma(y_0, y_1).$$

Properties of MW_2

Proposition

MW_2 has an equivalent discrete formulation, given by

$$MW_2^2(\mu_0, \mu_1) = \min_{w \in \Pi(\pi_0, \pi_1)} \sum_{k,l} w_{kl} W_2^2(\mu_0^k, \mu_1^l).$$

It happens that this discrete form has been recently proposed as an ingenious alternative to W_2 in the machine learning literature, both in [CGT19] and [CYL19].

Corollary

Let $\mu_0 = \sum_{k=1}^{K_0} \pi_0^k \mu_0^k$ and $\mu_1 = \sum_{k=1}^{K_1} \pi_1^k \mu_1^k$ be two Gaussian mixtures on $\mathbb{R}^d$, then the infimum in MW_2 is attained for a given

$$\gamma^* \in \Pi(\mu_0, \mu_1) \cap GMM_{2d}(K_0 + K_1 - 1).$$

[CGT19] Y. Chen, T. T. Georgiou, and A. Tannenbaum, Optimal Transport for Gaussian Mixture Models, *IEEE Access*, 2019.

[CYL19] Y. Chen, J. Ye, and J. Li, Aggregated Wasserstein Distance and State Registration for Hidden Markov Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Properties of MW_2 (continued)

Proposition

MW_2 defines a metric on $GMM_d(\infty)$ and the space $GMM_d(\infty)$ equipped with the distance MW_2 is a geodesic space.

Corollary

The barycenters between $\mu_0 = \sum_k \pi_0^k \mu_0^k$ and $\mu_1 = \sum_l \pi_1^l \mu_1^l$ all belong to $GMM_d(\infty)$ and can be written explicitely as

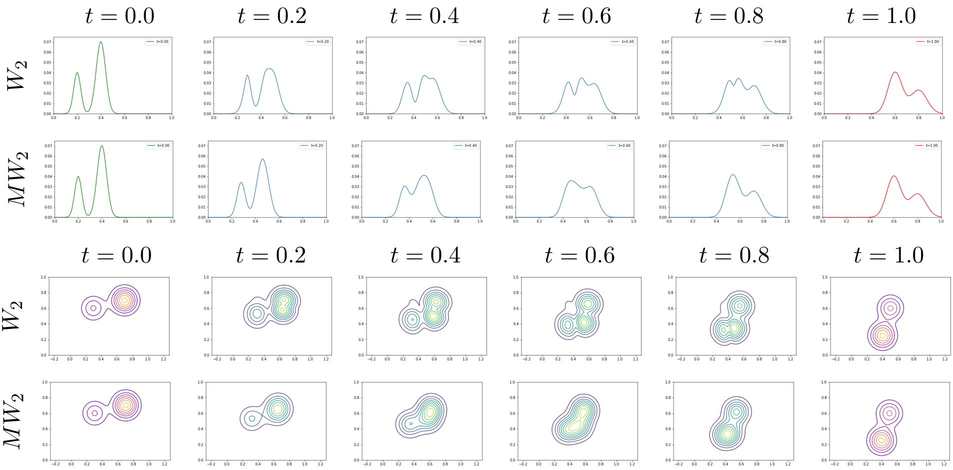
$$\forall t \in [0,1], \quad \mu_t = \mathrm{P}_t \# \gamma^* = \sum_{k,l} w_{k,l}^* \mu_t^{k,l},$$

where w^* is an optimal solution of the discrete formulation, and $\mu_t^{k,l}$ is the displacement interpolation between μ_0^k and μ_1^l. When Σ_0^k is non-singular, it is given by

$$\mu_t^{k,l} = ((1-t)\mathrm{Id} + tT_{k,l}) \# \mu_0^k,$$

with $T_{k,l}$ the affine transport map between μ_0^k and μ_1^l.
These barycenters have less than $K_0 + K_1 - 1$ components.

Properties of MW_2 (continued)



Properties of MW_2 (continued)

Proposition

Let $\mu_0 \in GMM_d(K_0)$ and $\mu_1 \in GMM_d(K_1)$ be two Gaussian mixtures. Then,

$$W_2(\mu_0, \mu_1) \leqslant MW_2(\mu_0, \mu_1) \leqslant W_2(\mu_0, \mu_1) + \sum_{i=0,1} \left(2 \sum_{k=1}^{K_i} \pi_i^k \mathrm{trace}(\Sigma_i^k) \right)^{\frac{1}{2}},$$

where the Σ_i^k are the covariance matrices of the components of μ_i.

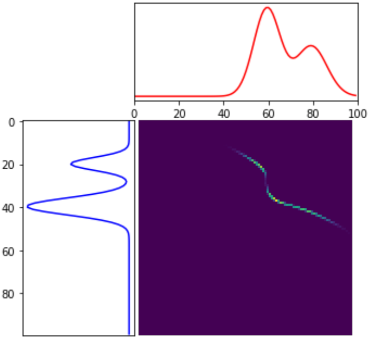
Using MW_2 on real data

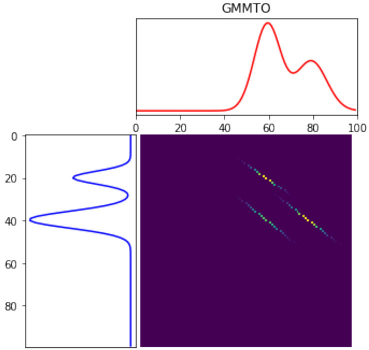
From a transport plan to a map :

Let μ_0 and μ_1 be two GMM. Then, the optimal transport plan between μ_0 and μ_1 for MW_2 is given by

$$\gamma(x,y) = \sum_{k,l} w^*_{k,l} g_{m_0^k,\Sigma_0^k}(x)\delta_{y=T_{k,l}(x)}.$$

It is not of the form $(\mathrm{Id}, T)\#\mu_0$





Using MW_2 on real data

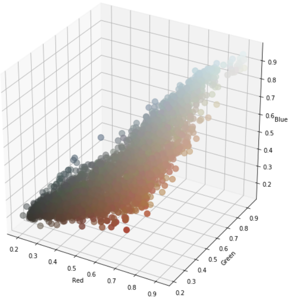
We can define two maps :

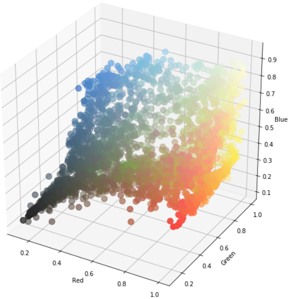
$$T_{mean}(x) = \mathbb{E}_\gamma(Y|X = x) = \frac{\sum_{k,l} w^*_{k,l}\, g_{m_0^k,\Sigma_0^k}(x) T_{k,l}(x)}{\sum_k \pi_0^k\, g_{m_0^k,\Sigma_0^k}(x)}.$$

$$T_{rand}(x) = T_{k,l}(x) \quad \text{with probability } p_{k,l}(x) = \frac{w^*_{k,l}\, g_{m_0^k,\Sigma_0^k}(x)}{\sum_j \pi_0^j\, g_{m_0^j,\Sigma_0^j}(x)}.$$

(It is not clear how to define a measurable random map from T_{rand}.)

Example : color transfer





Example : color transfer

Result of T_{mean}

Example : color transfer

Result of T_{rand}

Example : color transfer

Result of Sliced OT

Example : color transfer

Result of Separable OT

Example : color transfer

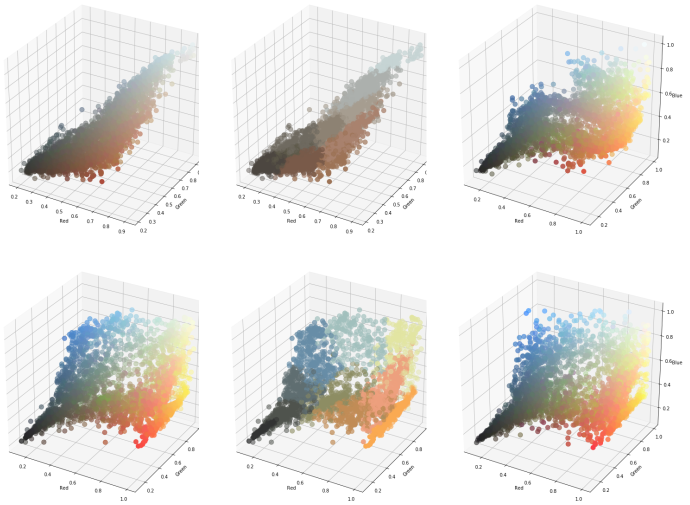


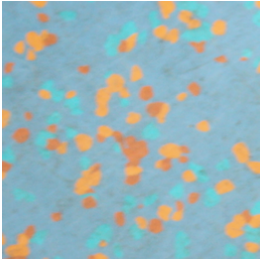
FIGURE – First line : color distribution of the image u_0, the 10 classes found by the EM algorithm, and color distribution of $T_{mean}(u_0)$. Second line : color distribution of the image u_1, the 10 classes found by the EM algorithm, and color distribution of $T_{rand}(u_0)$.

Example : color transfer

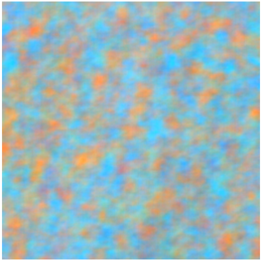
$K = 1$ $K = 3$ $K = 10$

FIGURE – The left-most image is the “red mountain” image, and its color distribution is modified to match the one of the right-most image (the “white mountain” image) with MW_2 using respectively $K = 1$, $K = 3$ and $K = 10$ components in the Gaussian mixtures.

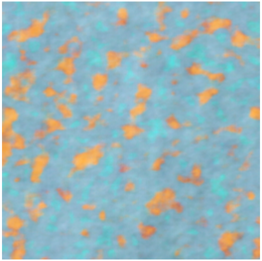
Example : Texture Synthesis



Original texture u



$\mathrm{ADSN}(u)$



Synthesized texture

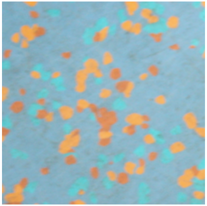
where $\mathrm{ADSN}(u)$ is a stationary Gaussian field that has same mean and same covariance as u.

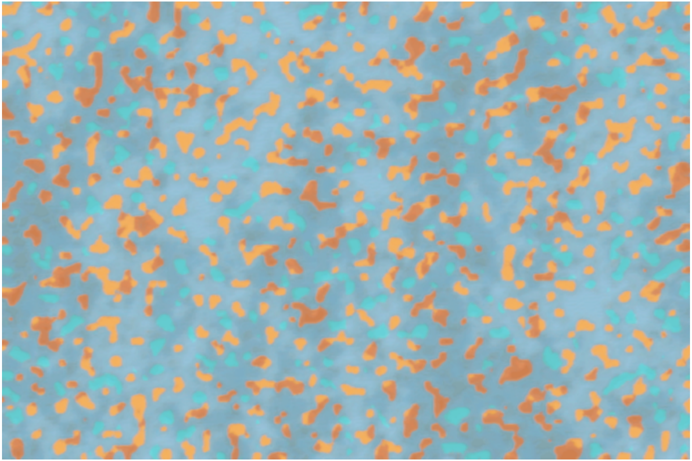
Texture synthesis algorithm :

- decompose u and $\mathrm{ADSN}(u)$ into two sets of patches
- compute the optimal plan (for MW_2) between corresponding GMMs
- replace patches from $\mathrm{ADSN}(u)$ with matching patches in u.

[Ongoing work with A. Leclaire], inspired by [Leclaire, Galerne, Rabin, 2018]

Multiscale texture synthesis





[Ongoing work with A. Leclaire]

Extension 1 : Mixing EM and MW_2 ?

Instead of a two step formulation (first EM, then MW_2), we propose here a relaxed formulation combining directly MW_2 with EM.

Let ν_0 and ν_1 be two probability measures on $\mathbb{R}^d$, we define

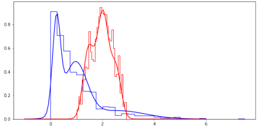
$$E_{K,\lambda}(\nu_0,\nu_1) = \min_{\gamma \in GMM_{2d}(K)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y_0 - y_1\|^2 d\gamma(y_0,y_1) - \lambda \mathbb{E}_{\nu_0}[\log P_0\#\gamma] - \lambda \mathbb{E}_{\nu_1}[\log P_1\#\gamma],$$

where $\lambda > 0$ is a parameter.

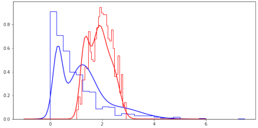
Remarks :

- Generally not a distance
- If ν_i has a density, then $\mathbb{E}_{\nu_i}[\log P_i\#\gamma] = -KL(\nu_i, P_i\#\gamma) - H(\nu_i)$, where $H(\nu_i)$ is the differential entropy of ν_i
  $\longrightarrow$ link with unbalanced transport [Chizat et. al]
- Use of automatic differentiation

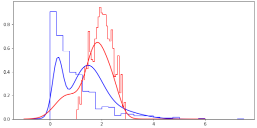
Mixing EM and MW_2



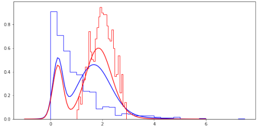
(a) $\lambda = 10$



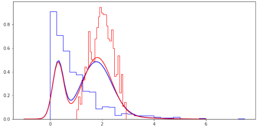
(b) $\lambda = 2$



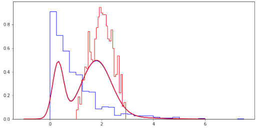
(c) $\lambda = 1$



(d) $\lambda = 0.5$



(e) $\lambda = 0.1$

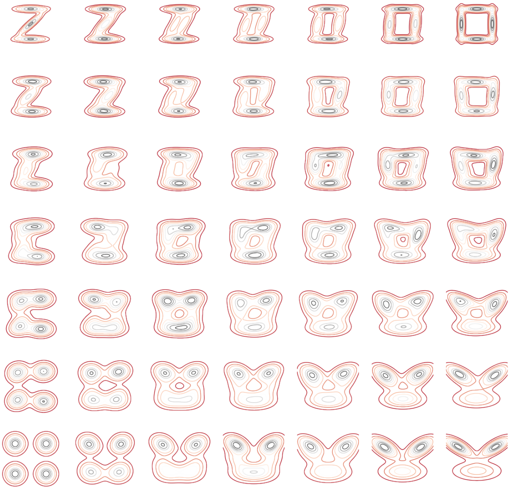


(f) $\lambda = 0.01$

FIGURE – The distributions ν_0 and ν_1 are 1d discrete distributions, plotted as the red and blue discrete histograms. The red and blue plain curves represent the final distributions $P_0\#\gamma$ and $P_1\#\gamma$. In this experiment, we use $K = 3$ Gaussian components for γ.

Extension 2 : multi-marginal formulation

$$\inf_{\nu \in GMM_d(\infty)} \sum_{j=0}^{J-1} \lambda_j MW_2^2(\mu_j, \nu)$$



Conclusion

- MW_2 : a distance on GMMs suited for high dimensional data
- Reduced complexity : Optimal Transport for a $K_0 \times K_1$ problem
- Relevant for data structured in classes
- Limitation : use of EM
- Extension to data living in spaces of different dimension ? (Gromov-Wasserstein)

J. Delon and A. Desolneux, A Wasserstein-type distance in the space of Gaussian Mixture Models, *SIAM Journal on Imaging Sciences*, Vol. 13(2), pp. 936-970, 2020.
https://hal.archives-ouvertes.fr/hal-02178204
https://github.com/judelo/gmmot