Optimal transport for Gaussian mixture models

Yongxin Chen, Tryphon T. Georgiou and Allen Tannenbaum

Presented by: Zach Lucas

Intro and Motivation

A mixture model is a probabilistic model describing properties of populations with subpopulations.

To study OMT on certain submanifolds of probability densities. To retain the nice properties of OMT, herein, an explicit OMT framework on Gaussian mixture models is used.

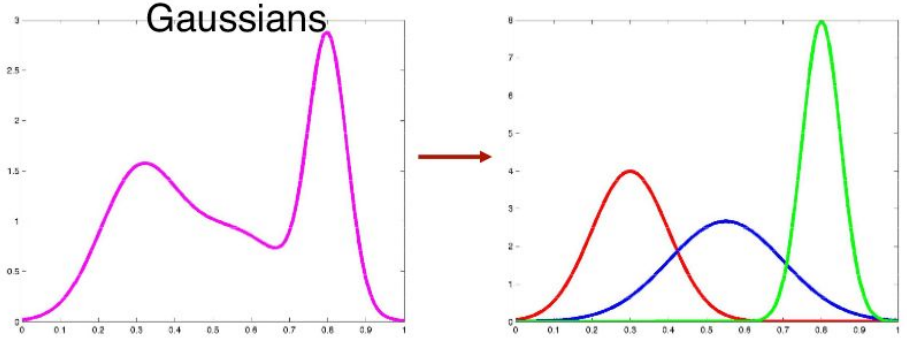
Data is sparsely distributed among subgroups. The difference between data within a subgroup is way less significant than that between subgroups.

Gaussian Mixture Model (GMM) Learning

Unsupervised clustering based on naive Bayes

$$P(X) = \sum_{c \in C} P(c)P(X \mid c)$$

- Can we recover the underlying Gaussians given some data?
  - Each data point is “generated” by one of the Gaussians



GMM: Expectation - Maximization (EM)

- Two parts, done over and over again
- Part 1: Expectation
  - What's our best guess for every data point as to which cluster it comes from
  - In general, compute the probability of hidden variables
- Part 2: Maximization:
  - Given our expectations, figure out the parameters for the gaussian distributions
  - In general, compute new parameters based on the probability of the hidden variables

GMM: Expectation

- Question: for every point X_j, what is the probability that class$_i$ generated that point?

$$P_{ij}(c_i \mid X_j) = \alpha P(X_j \mid c_i) P(c_i) = P_{ij}$$

$$N_i = \sum_j P_{ij}$$

GMM: Maximization

- For every class, compute a new class prior, mean, and standard deviation

$$\hat{\mu}_i = \frac{\sum_j P_{ij} x_j}{N_i}$$

new mean: weighted average of points assigned to class i

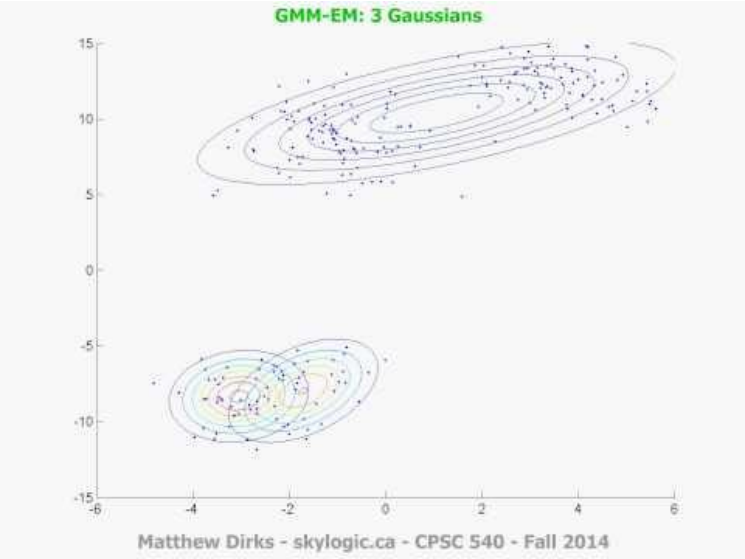
$$\hat{\sigma}_i = \sqrt{\frac{\sum_j P_{ij} x_j^2}{N_i} - \left(\frac{\sum_j P_{ij} x_j}{N_i}\right)^2}$$

new standard deviation: calculated in same weighted manner

$$\hat{P}(c_i) = \frac{N_i}{\sum_j N_j}$$

new class prior: proportion of weighted samples attributed to class

GMM: 2D example



https://www.youtube.com/watch?v=B36fzChfyGU

OMT Background

Consider two measures μ_0, μ_1 on $\mathbb{R}^n$ with equal total mass. Without loss of generality, we take μ_0 and μ_1 to be probability distributions. In the original formulation of OMT, a transport map

$$T : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto T(x)$$

is sought that specifies where mass $\mu_0(dx)$ at x should be transported so as to match the final distribution in the sense that $T_\sharp \mu_0 = \mu_1$, i.e. μ_1 is the "push-forward" of μ_0 under T, meaning

$$\mu_1(B) = \mu_0(T^{-1}(B))$$

for every Borel set B in $\mathbb{R}^n$. Moreover, the map should achieve a minimum cost of transportation

$$\int_{\mathbb{R}^n} c(x, T(x))\mu_0(dx).$$

$$c(x, y) = \|x - y\|^2$$

OMT Background: Kantorovich

Coupling $\Pi(\mu_0, \mu_1)$ on $\mathbb{R}^n \times \mathbb{R}^n$,

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \pi(dxdy). \tag{1}$$

The unique optimal transport T is the gradient of a convex function

$$y = T(x) = \nabla \phi(x). \tag{2}$$

OMT Background: Kantorovich

The optimal coupling based on the transport map T in (2), where Id is the identity map.

$$\pi = (\mathrm{Id} \times T)_\sharp \mu_0,$$

The square root of the minimum of the cost defines a Riemannian metric on $P_2(\mathbb{R}^n)$ known as the Wasserstein metric W_2. On this Riemannian-type manifold, the geodesic curve is given by

$$\mu_t = (T_t)_\sharp \mu_0, \quad T_t(x) = (1-t)x + tT(x), \tag{3}$$

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1), \quad 0 \le s < t \le 1. \tag{4}$$

Displacement Interpolation

Gaussian marginal distributions

Denote the mean and covariance of $\mu_i, i = 0, 1$ by m_i and Σ_i
Let X, Y be two Gaussian random vectors associated with μ_0, μ_1. respectively.
Our new cost from (1) becomes

$$\mathbb{E}\{\|X - Y\|^2\} = \mathbb{E}\{\|\tilde{X} - \tilde{Y}\|^2\} + \|m_0 - m_1\|^2, \tag{5}$$

$$\tilde{X} = X - m_0, \tilde{Y} = Y - m_1$$

$$\min_S \left\{ \|m_0 - m_1\|^2 + \operatorname{trace}(\Sigma_0 + \Sigma_1 - 2S) \;\middle|\; \begin{bmatrix} \Sigma_0 & S \\ S^T & \Sigma_1 \end{bmatrix} \geq 0 \right\}, \tag{6}$$

$$S = \mathbb{E}\{\tilde{X}\tilde{Y}^T\}$$

Gaussian marginal distributions

The constraint is semidefinite constraint, so the (6) is a semidefinite programming (SDP). It turns out that the minimum is achieved by the unique minimizer in closed-form:

$$S = \Sigma_0^{1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}$$

With minimum value

$$W_2(\mu_0, \mu_1)^2 = \|m_0 - m_1\|^2 + \mathrm{trace}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2})$$

Gaussian marginal distributions

Displacement Interpolation as a Gaussian:

$$m_t = (1-t)m_0 + tm_1$$

$$\Sigma_t = \Sigma_0^{-1/2}\left((1-t)\Sigma_0 + t(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right)^2\Sigma_0^{-1/2}. \tag{7}$$

Wasserstein Distance can be extended to singular Gaussian distributions

$$W_2(\mu_0,\mu_1)^2 = \|m_0 - m_1\|^2 + \mathrm{trace}(\Sigma_0 + \Sigma_1 - 2\Sigma_0^{1/2}((\Sigma_0^{1/2})^\dagger\Sigma_1(\Sigma_0^{1/2})^\dagger)^{1/2}\Sigma_0^{1/2}). \tag{8}$$

when $\Sigma_0 = \Sigma_1 = 0$, $W_2(\mu_0,\mu_1) = \|m_0 - m_1\|$,

OMT for GMM

$$\mu = p^1\nu^1 + p^2\nu^2 + \cdots + p^N\nu^N$$

where each ν^k is a Gaussian distribution and $p = (p^1, p^2, \ldots, p^N)^T$ is a probability vector.

Space of distributions: $M(\mathbb{R}^n)$

We view it as a discrete distribution on the Wasserstein space of Gaussian distributions: $G(\mathbb{R}^n)$

OMT for GMM

Let μ_0, μ_1 be two Gaussian mixture models of the form

$$\mu_i = p_i^1 \nu_i^1 + p_i^2 \nu_i^2 + \cdots + p_i^{N_i} \nu_i^{N_i}, \quad i = 0, 1.$$

The discrete OMT problem:

$$\min_{\pi \in \Pi(p_0, p_1)} \sum_{i,j} c(i,j)\pi(i,j) \tag{9}$$

$$c(i,j) = W_2(\nu_0^i, \nu_1^j)^2.$$

Let π^* be a minimizer, and define

$$d(\mu_0, \mu_1) = \sqrt{\sum_{i,j} c(i,j)\pi^*(i,j)}. \tag{10}$$

Theorem 1: $d(\cdot,\cdot)$ defines a metric on $M(\mathbb{R}^n)$.

Proof 1: Apparently, $d(\mu_0,\mu_1) \geq 0$ for any $\mu_0,\mu_1 \in M(\mathbb{R}^n)$ and $d(\mu_0,\mu_1) = 0$ if and only if $\mu_0 = \mu_1$. We next prove the triangular inequality, namely,

$$d(\mu_0,\mu_1) + d(\mu_1,\mu_2) \geq d(\mu_0,\mu_2)$$

for any $\mu_0,\mu_1,\mu_2 \in M(\mathbb{R}^n)$. Denote the probability vector associated with μ_0,μ_1,μ_2 by p_0,p_1,p_2 respectively. The Gaussian components of μ_i is denoted by ν_i^j. Let π_{01} (π_{12}) be the solution to (9) with marginals μ_0,μ_1 (μ_1,μ_2). Define π_{02} by

$$\pi_{02}(i,k) = \sum_j \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j}.$$

Clearly, π_{02} is a joint distribution between p_0 and p_2, namely, $\pi_{02} \in \Pi(p_0,p_2)$. It follows from direct calculation

$$\begin{aligned}
\sum_i \pi_{02}(i,k) &= \sum_{i,j} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} \\
&= \sum_j \frac{p_2^j \pi_{12}(j,k)}{p_1^j} \\
&= p_2^k.
\end{aligned}$$

Similarly, we have $\sum_k \pi_{02}(i,k) = p_0^i$. Therefore,

$$
\begin{aligned}
d(\mu_0,\mu_2) &\le \sqrt{\sum_{i,k} \pi_{02}(i,k) W_2(\nu_0^i,\nu_2^k)^2} \\
&= \sqrt{\sum_{i,j,k} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} W_2(\nu_0^i,\nu_2^k)^2} \\
&\le \sqrt{\sum_{i,j,k} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} (W_2(\nu_0^i,\nu_1^j) + W_2(\nu_1^i,\nu_2^k))^2} \\
&\le \sqrt{\sum_{i,j,k} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} W_2(\nu_0^i,\nu_1^j)^2} + \sqrt{\sum_{i,j,k} \frac{\pi_{01}(i,j)\pi_{12}(j,k)}{p_1^j} W_2(\nu_1^j,\nu_2^k)^2} \\
&= \sqrt{\sum_{i,j} \pi_{01}(i,j) W_2(\nu_0^i,\nu_1^j)^2} + \sqrt{\sum_{j,k} \pi_{12}(j,k) W_2(\nu_1^j,\nu_2^k)^2} \\
&= d(\mu_0,\mu_1) + d(\mu_1,\mu_2).
\end{aligned}
$$

In the above, the second inequality is due to the fact W_2 is a metric, and the third inequality is an application of the Minkowski inequality.

Geodesic

A geodesic on $M(\mathbb{R}^n)$ connecting μ_0 and μ_1 is given by

$$\mu_t = \sum_{i,j} \pi^*(i,j)\nu_t^{ij}, \tag{11}$$

where ν_t^{ij} is the displacement interpolation (see (7)) between ν_0^i and ν_1^j.

Theorem 2:

$$d(\mu_s, \mu_t) = (t-s)d(\mu_0, \mu_1), \quad 0 \le s < t \le 1. \tag{12}$$

Proof 2: For any $0 \le s \le t \le 1$, we have

$$\begin{aligned} d(\mu_s, \mu_t) &\le \sqrt{\sum_{i,j} \pi^*(i,j) W_2(\nu_s^{ij}, \nu_t^{ij})^2} \\ &= (t-s)\sqrt{\sum_{i,j} \pi^*(i,j) W_2(\nu_0^i, \nu_1^j)^2} = (t-s)d(\mu_0, \mu_1) \end{aligned}$$

where we have used the property (4) of W_2. It follows that

$$d(\mu_0, \mu_s) + d(\mu_s, \mu_t) + d(\mu_t, \mu_1) \le sd(\mu_0, \mu_1) + (t-s)d(\mu_0, \mu_1) + (1-t)d(\mu_0, \mu_1) = d(\mu_0, \mu_1).$$

On the other hand, by Theorem 1, we have

$$d(\mu_0, \mu_s) + d(\mu_s, \mu_t) + d(\mu_t, \mu_1) \ge d(\mu_0, \mu_1).$$

Combining these two, we obtain (12).

We remark that μ_t is a Gaussian mixture model since it is a weighted average of the Gaussian distributions ν_t^{ij}. Even though the solution to (9) is not unique in some instances, it is unique for generic $\mu_0, \mu_1 \in M(\mathbb{R}^n)$. Therefore, in most real applications, we need not worry about the uniqueness.

Notes

$$d(\mu_0, \mu_1) \geq W_2(\mu_0, \mu_1)$$

This is due to the fact that the restriction to the submanifold induces suboptimality in the transport plan.

it is unclear whether$d$ is the restriction of W_2 to $M(\mathbb{R}^n)$

d is a very good approximation of W2 if the variances of the Gaussian components are small compared with the differences between the means.

Only (9) must be solved to compute a new distance, which is extremely efficient with small distributions

Barycenter of GMM

$\mu_0, \mu_1, \ldots, \mu_L$

$$J(\mu) = \frac{1}{L} \sum_{k=1}^{L} W_2(\mu, \mu_k)^2. \tag{13}$$

$\frac{1}{L}(x_1 + x_2 + \cdots + x_L)$

$$J(x) = \frac{1}{L} \sum_{k=1} \|x - x_k\|^2.$$

$$\min_{\mu \in P_2(\mathbb{R}^n)} \sum_{k=1}^{L} \lambda_k W_2(\mu, \mu_k)^2. \tag{14}$$

$\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_L]$ is a probability vector

Barycenter of GMM

$$m = \sum_{k=1}^{L} \lambda_k m_k \tag{15}$$

$$\Sigma = \sum_{k=1}^{L} \lambda_k (\Sigma^{1/2} \Sigma_k \Sigma^{1/2})^{1/2}. \tag{16}$$

Solve with fixed point iteration: $(\Sigma)_{\text{next}} = \sum_{k=1}^{L} \lambda_k (\Sigma^{1/2} \Sigma_k \Sigma^{1/2})^{1/2}$

Remark: unrealistic to solve (14) for more than 3 dimensions for both general and gaussian distributions

Barycenter of GMM

Modified problem:

$$\min_{\mu \in M(\mathbb{R}^n)} \sum_{k=1}^{L} \lambda_k d(\mu, \mu_k)^2. \tag{17}$$

Let $\mu_k = p_k^1 \nu_k^1 + p_k^2 \nu_k^2 + \cdots + p_k^{N_k} \nu_k^{N_k}$ as a discrete measure on $G(\mathbb{R}^n)$

$$\operatorname{argmin}_{\nu} \sum_{k=1}^{L} \lambda_k W_2(\nu, \nu_k^{i_k})^2 \tag{18}$$

Barycenter of GMM

The optimal v is gaussian. Denote the set of all such minimerzers $\{\nu^1, \nu^2, \ldots, \nu^N\}$

$$\mu = p^1\nu^1 + p^2\nu^2 + \cdots + p^N\nu^N$$

For some probability vector $p = (p^1, p^2, \ldots, p^N)^T$

The number of element N is bounded above by $N_1 N_2 \cdots N_L$

Barycenter of GMM

$$\min_{\pi_1 \geq 0, \cdots, \pi_L \geq 0} \sum_{k=1}^{L} \sum_{i=1}^{N} \sum_{j_k=1}^{N_k} \lambda_k c_k(i, j_k) \pi_k(i, j_k) \tag{19a}$$

$$\sum_{i=1}^{N} \pi_k(i, j_k) = p_k^{j_k}, \quad \forall 1 \leq k \leq L, 1 \leq j_k \leq N_k \tag{19b}$$

$$\sum_{j_1=1}^{N_1} \pi_1(i, j_1) = \sum_{j_2=1}^{N_2} \pi_2(i, j_2) = \cdots = \sum_{j_L=1}^{N_L} \pi_L(i, j_L), \quad \forall 1 \leq i \leq N. \tag{19c}$$

$$c_k(i, j) = W_2(\nu^i, \nu_k^j)^2 \tag{20}$$

Barycenter $\mu = p^1 \nu^1 + p^2 \nu^2 + \cdots + p^N \nu^N$ with $p^i = \sum_{j=1}^{N_1} \pi_1(i, j) \quad 1 \leq i \leq N$

Numerical Examples

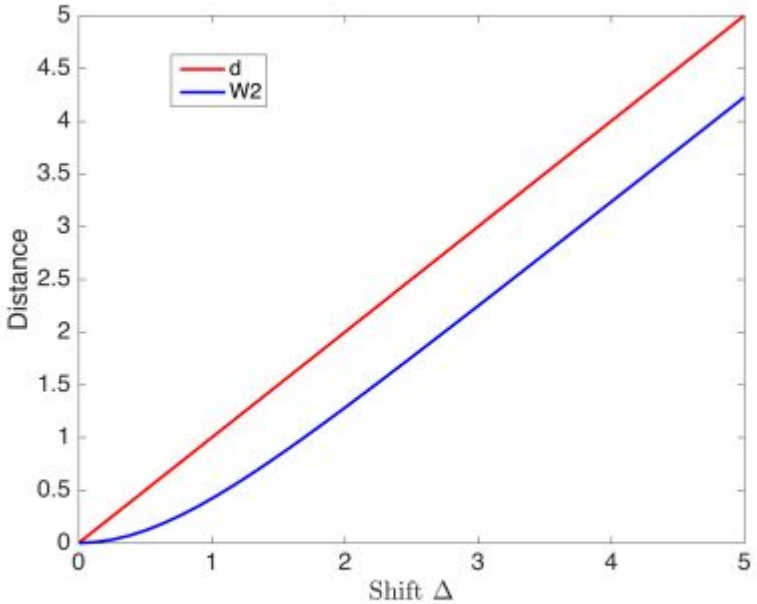
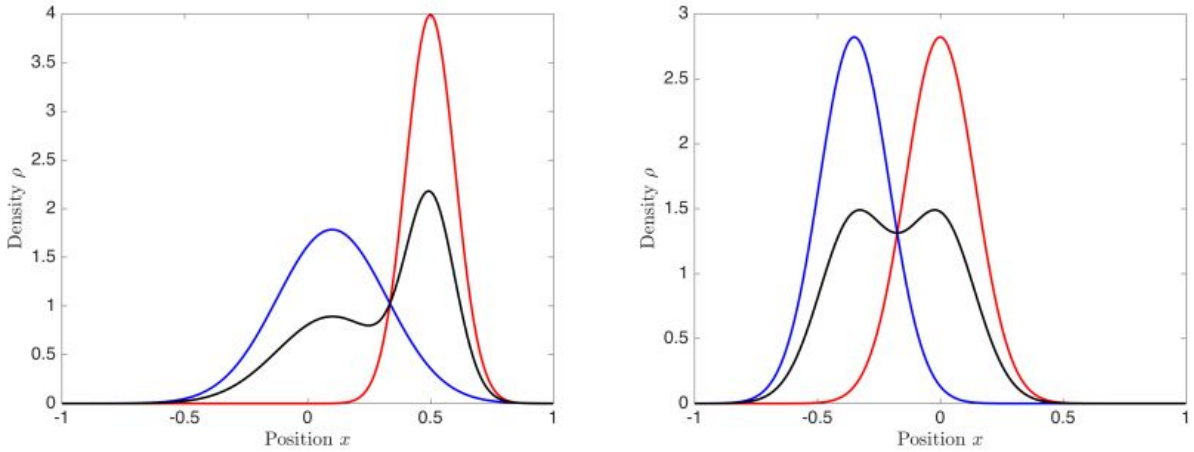


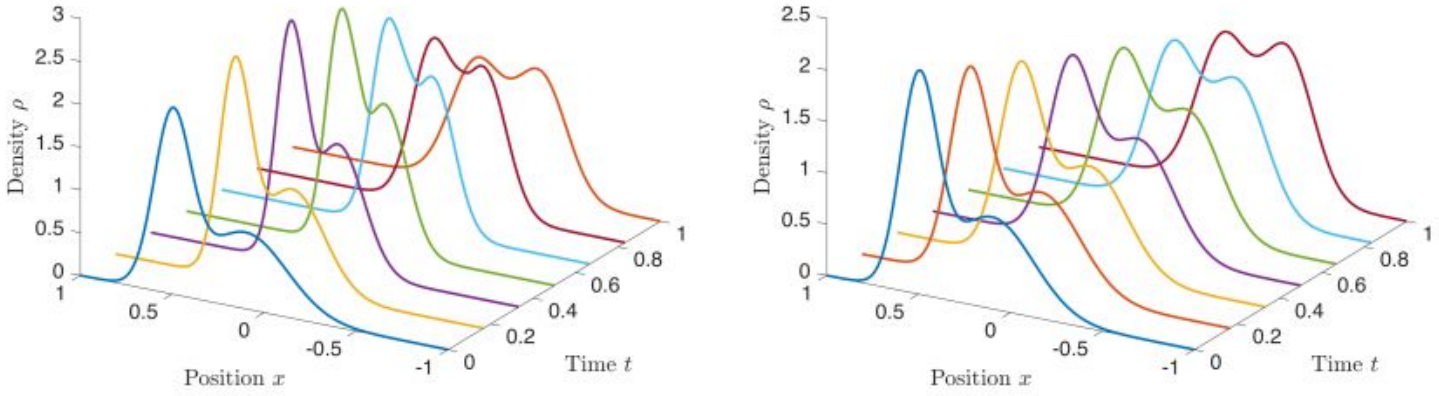
Fig. 1: d vs W_2

Geodesic



(a) μ_0 (b) μ_1

Fig. 2: Marginal distributions



(a) OMT

(b) our framework

Fig. 3: Interpolations

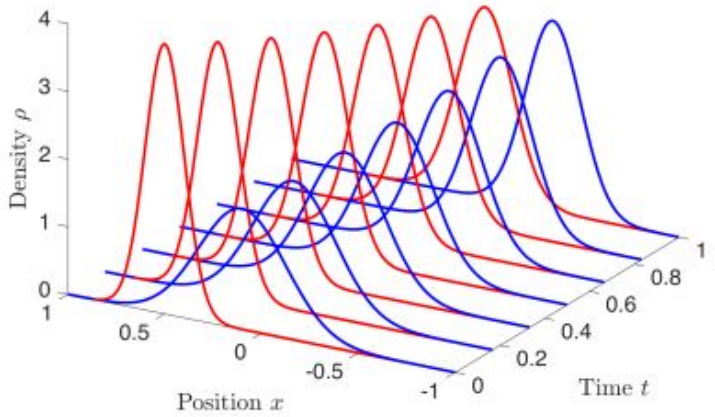
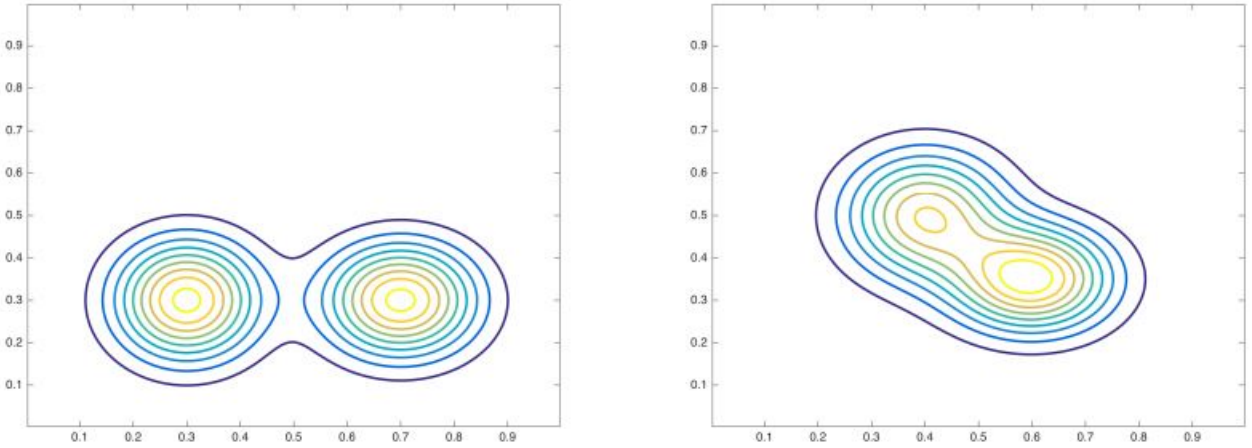


Fig. 4: Two Gaussian components of the interpolation



(a) μ_0 (b) μ_1

Fig. 5: Marginal distributions

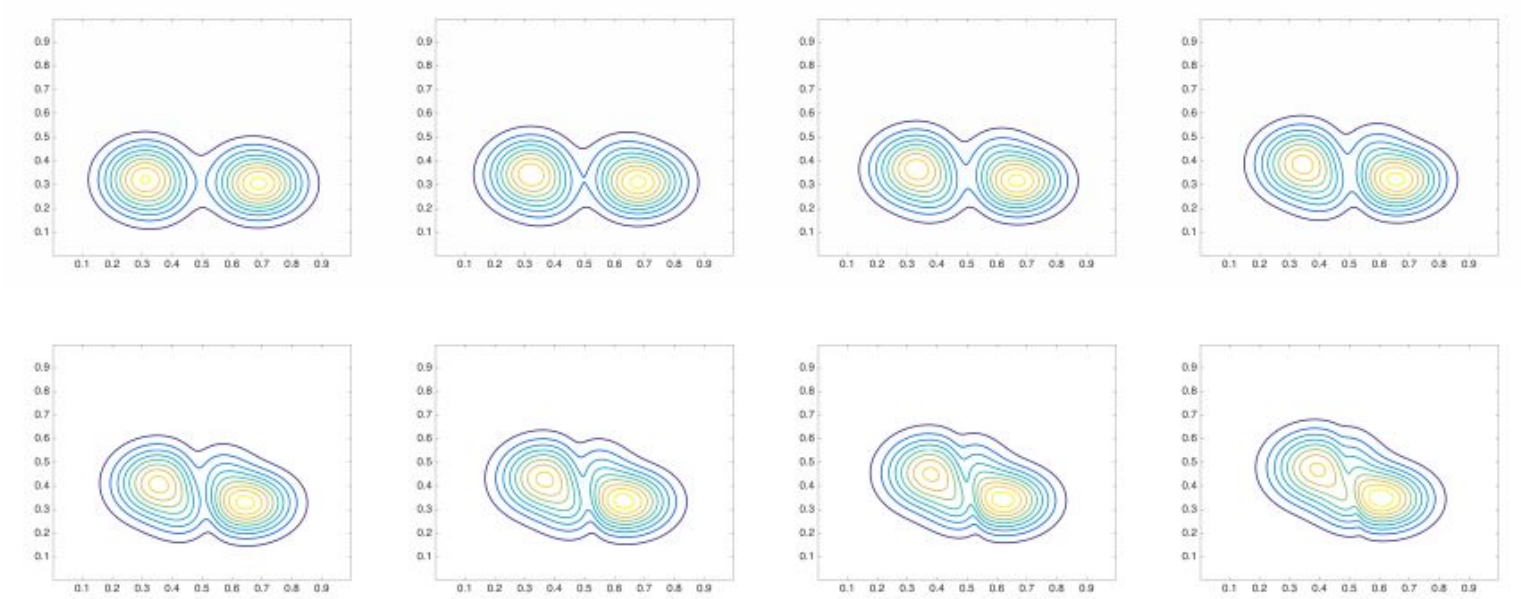


Fig. 6: OMT Interpolation

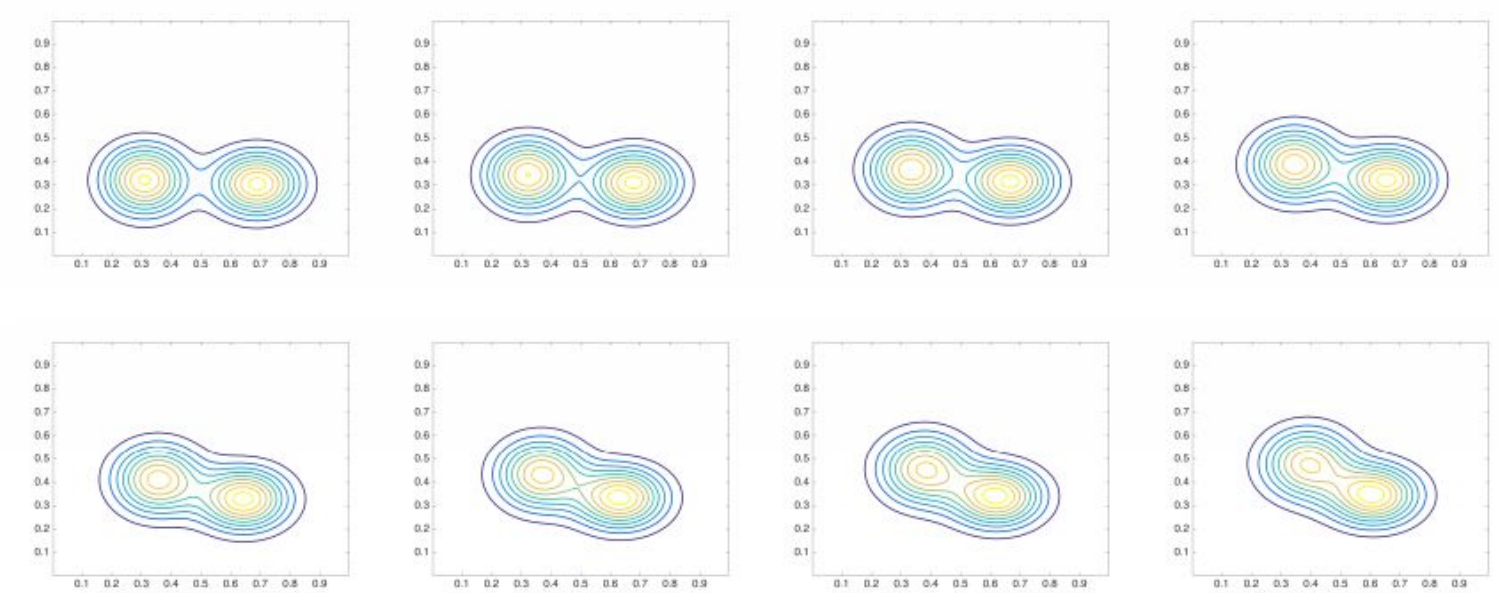
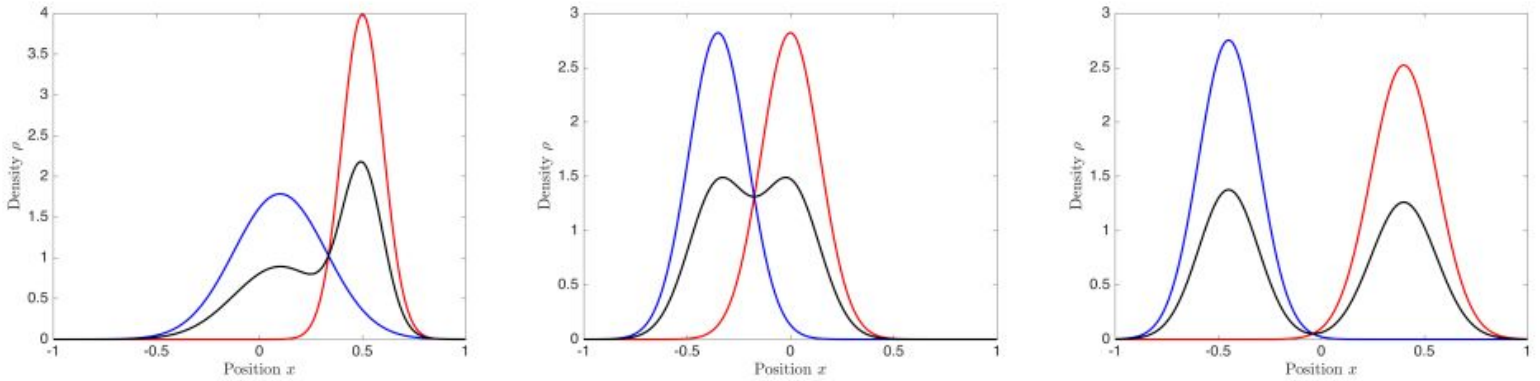


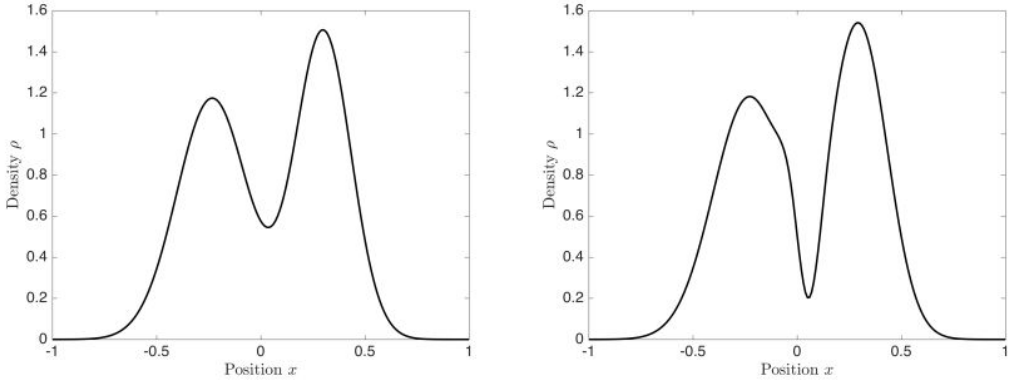
Fig. 7: Our Interpolation

Barycenter



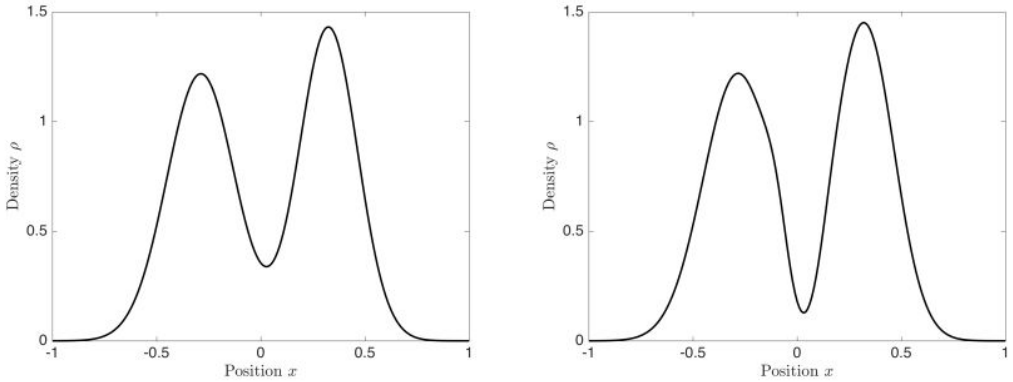
(a) μ_1 (b) μ_2 (c) μ_3

Fig. 8: Marginal distributions



(a) our method (b) optimal transport

Fig. 9: Barycenters with $\lambda = (1/3, 1/3, 1/3)$



(a) our method (b) optimal transport

Fig. 10: Barycenters with $\lambda = (1/4, 1/4, 1/2)$