Report for CSE 5339 2018 — (OTMLSA)
Optimal Transport in Machine Learning and Shape Analysis

Optimal transport for Gaussian mixture models

Zachary Lucas

Abstract

Authors Yongxin Chen, Tryphon T. Georgiou and Allen Tannenbaum present a new optimal mass transport (OMT) framework specifically for use with Gaussian mixture models (GMM). The mixture models are treated as elements on submanifold of probability densities that are then embedded into the density spaces equipped with the Wasserstein metric. This method leads to lower computation complexity and naturally preserves the Gaussian structures of the model. The main idea of the paper is that if we study an optimal transport problem for probability densities with a specific structure (GMMs), that we can exploit that structure to lower computation costs and better preserve the structure over iterations of the transport problem.

1 Introduction

A mixture model is a probabilistic model describing properties of populations with subpopulations. The GMM is a specific instance of such models and is simply a weighted average of several Gaussian distributions. Each Gaussian component stands for a subpopulation. Optimal transport, or optimal mass transport as the authors refer to it, is an old problem that has been the primary focus of this class. The basic idea is to transport masses from an initial distribution to some terminal distribution, preserving the mass with minimum cost. The authors point out that when the unit cost is the square of the Euclidean distance, the OMT problem induces an extremely rich geometry for probability densities and endows a Riemannian metric on the space of probability densities. This the concept of geometry enables us to compare, interpolate and average probability densities in a very natural way, which is used in this application.

It is known that OMT on the entire manifold of probability densities is computationally expensive, but many of these densities have specific structures and can be parameterized. Thus, the authors seek an explicit OMT framework on Gaussian mixture models, with the hope that similar applications can follow with other density structures. This work on GMMs was motivated by the observation that data are sparsely distributed among subgroups when working with high dimensional data. The difference between data within a subgroup is way less significant than that between subgroups. For example, the difference between two dogs of the same breed is less than the difference between dogs of different breads.

2 GMM and Learning

The OMT problem discussed here already has a fully formed GMM however that is not necessarily the case when working with probability densities. GMMs may be learned through the expectation-maximization (EM) algorithm. At a high level we used unsupervised clustering based on Naïve Bayes to "recover" some underlying Gaussian structure. The algorithm is iterative with two parts: expectation and maximization. In expectation we compute the probabilities

of hidden variables, guessing which cluster each data point came from. Then in maximization we compute new parameters based on the newly computed probabilities of the hidden variables. For every point x_j we compute the probability that is belongs to some cluster c_i. Then we update the mean, standard deviation, and prior probability of each cluster.

$$P_{ij}(c_i|x_j) = aP(x_j|c_i)P(c_i) = P_{ij}$$

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

$$\mu_i = (\sum_j P_{ij}x_j)/N_i$$

$$\sigma_i = \sqrt{(\sum_j P_{ij}x_j^2)/N_i - ((\sum_j P_{ij}x_j)/N_i)^2}$$

$$P(c_i) = N_i/(\sum_j N_j)$$

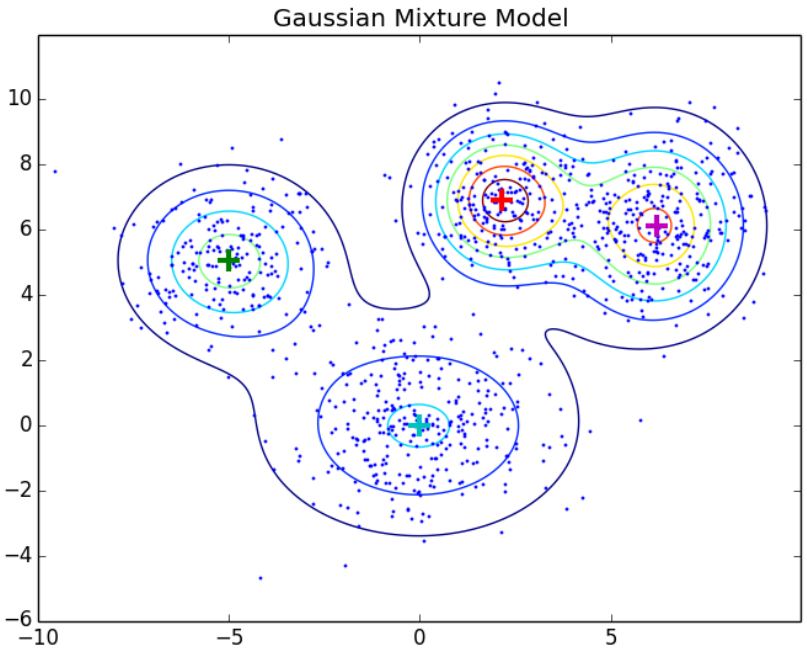


Fig. 0: A 2-D Gaussian mixture model with 4 clusters or sub-manifolds.

3 OMT Background

Let μ_0 and μ_1 be two probability distributions of equal mass on R^n and let T be a transport map from R^n to R^n. We seek a mass $\mu_0(dx)$ at x to be transported under T. Such a map should achieve a minimum transportation cost, c.

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

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

The Kantorovich relaxation seeks a coupling $\Pi(\mu_0 \times \mu_1$. Then we solve

$$inf_{\pi \subset \Pi} \int_{R^n \times R^n} ||x-y||^2 \pi(dxdy)$$

The optimal coupling is $\pi = (Identitymap \times T)\mu_0$.The square root of the minimum of the cost defines a Riemannian metric on $P_2(R^n)$

When both probability distributions are Gaussian the problem can be simplified to a closed form. Denote the mean and covariance $\mu_i, i = 0, 1$ by m_i, Σ_i respectively and let X and Y be be two Gaussian random vectors associated with μ_0 and μ_1. Then the cost is defined as

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

where $\tilde{X} = X - m_0, \tilde{Y} = Y - m_1$ are zero mean versions of X and Y.

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

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

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

with minimum value

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

The consequent displacement interpolation μ_t is a Gaussian distribution with mean $m_t = (1-t)m_0 + tm_1$ and covariance

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

The Wasserstein distance can be extended to singular Gaussian distributions by replacing the inverse by the pseudoinverse †, which leads to

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

In particular, when $\Sigma_0 = \Sigma_1 = 0$, we have

$$W_2(\mu_0, \mu_1) = \|m_0 - m_1\|.$$

4 OMT for GMM

A GMM is of the form

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

where each v^k is a Gaussian distribution and p is a probability vector. N is the finite number of Gaussian clusters in the model or the number of component means. The space of the Gaussian mixture distribution is $M(R^n)$. Now we let μ_0 and μ_1 be two GMMs of the form described above. We solve the discrete OMT problem

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

Here $\Pi(p_0, p_1$ denote the space of joint distributions. The cost $c(i,j)$ is taken to be the square of the Wasserstein metric

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

By standard linear programming theory, the discrete OMT problem (9) always has at least one solution.Let π^* be a minimizer, and define

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

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

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

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

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

Theorem 2:

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

5 Examples

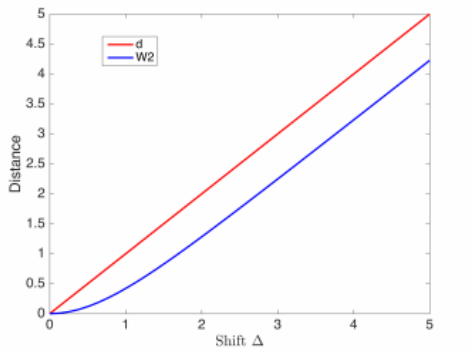
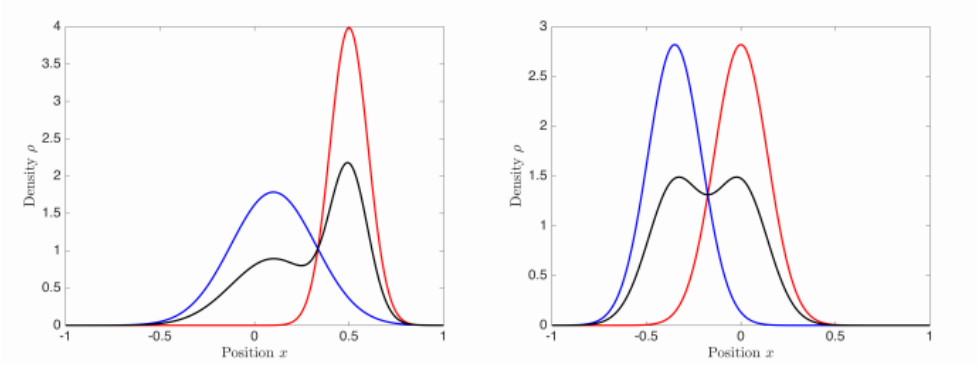
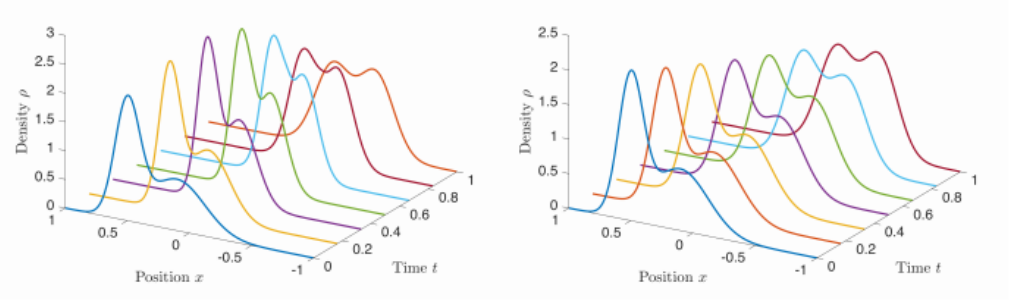


Fig. 1: d vs W_2



(a) μ_0 (b) μ_1

Fig. 2: Marginal distributions



(a) OMT (b) our framework

Fig. 3: Interpolations

Figure 1 shows that d is bounded below by W_2. Figure 2 depicts two example GMMs and Figure 3 shows the differences between the two different methods of interpolation. We can see that the author's interpolation preserves the structure of the Gaussian components better.

6 Conclusion

The authors have presented a computationally efficient method for the optimal transport of Gaussian mixture models. Specifically with the geodesic implementation we can see that the metric remains on the sub-manifold of the Gaussian mixture models' distributions, preserving the shape of the underlying data better than tradition optimal transport on the same distributions. This work seems very relevant to any application of OMT where the underlying distribution has a known structure, especially considering there are machine learning techniques to help uncover those models. Extension of this theory to more general mixture models is the direction of future works the authors would like to take.

References

[1] Y. Chen et al. "Optimal Transport for Gaussian Mixture Models." `LearningWassersteinEmbeddings`