Case studies of discrete optimal transportation in geodesic spaces

1 Acknowledgements

I would like to thank Professor Tabak for his insight and guidance. His good humor, pleasant attitude and serenity were inspiring and always encouraging. He provided support well and beyond one student could have imagined. I learned a lot from working with him about Optimal transport and what it means doing research with rigor and yet inspiration. I want to also thank my partner who was willing to sacrifice many hours when attending to the needs of our new born beautiful daughter.

2 Introduction

Many problems in data analysis involve samples $\mathbf{x}_i$ with an unknown underlying probability distribution $\mu_s(x)$. An emerging methodology analyses this data through a transformation $y = T(x)$ that pushes forward $\mu_s(x)$ to a different distribution $\mu_t(y)$. For example, if $\mu_t(y)$ is a known proposed target distribution, then the knowledge of μ_t and of the map $y = T(x)$ yields $\mu_s(x)$, i.e. performs density estimation. In the case of conditional source distributions $\mu_s(x|z)$, mapping them all to a single target $\mu_t(y)$ through a z-dependent map $y = T(x; z)$ provides, in addition to conditional density estimation, a means for data-consolidation: in this case, z is the categorical variable describing the data source, and the map consolidates the data from all data-sources z into a single data base $\mathbf{y}_i$, which by construction is independent of z and suitable for statistical analysis.

Among the infinitely many maps $y = T(x)$ pushing forward μ_s to μ_t, it is convenient to chose the one that minimally distorts the data. This can be formulated as a problem in optimal transportation: given a cost function $c(x, y)$ that measures the distortion incurred upon by moving x to y, choose the map that minimizes the expected value of this distortion. In the z-dependent case, when one minimizes the expected cost not only over the maps $T(x; z)$ but also over the target distribution $\mu_t(y)$, the latter is denoted the Wasserstein barycenter of the distributions $\mu_s(x|z)$.

When the space X underlying the distributions μ_s and μ_t is Euclidean, a conventional choice for the cost function $c(x, y)$ is the squared distance between x and y, a choice similar in many ways to that of the squared error in regression. Yet this choice is not necessarily natural in data analysis, where the straight line joining two data points may go through areas that make little sense for the problem under consideration, such as combinations of pressure and temperature inconsistent with the material under study or of stock prices inconsistent with economics principles.

This thesis explores data-based, discrete optimal transport in situations where the cost itself is derived from the data. We consider two kinds of cost: one derived from the square of the geodesic distance along the manifold underlying the data –which must itself be estimated– and the other following a recent proposal ([1]) which considers not only the geometry of the manifold but also the sparsity of the data along it.

3 Settings

Given two sets of data points $\mathbf{X} = \{\mathbf{x}_{i=1}^N\}$ and $\mathbf{Y} = \{\mathbf{y}_{j=1}^M\}$ on the surface of manifold $\mathcal{M}$, the problem is, to discover a transformation which maps the cluster $\mathbf{X}$ to the cluster $\mathbf{Y}$. In the language of optimal transport, we are looking for the map **Y=T(X)**, which pushes forward the distribution $\mu_s(x)$ into the distribution $\mu_t(y)$.

We propose that $\mathbf{T}$ minimizes a transportation cost $C(T, \mathcal{M})$:

$$C(T) = \int c(x, T(x))\, d\mu_s$$

where C(T) can be interpreted as the minimal energy required to move a probability mass $\mu_s(x)$ from x to T(x).

The cost $c(x, y)$ is a measure of the distance between points x and y. In the classical optimal transport problem, the function c is given, as are the source and target distributions $\mu_s(x)$ and $\mu_t(y)$. The problem that we address here differs from the classical one in two main aspects:

1. The distributions $\mu_s(x)$ and $\mu_t(y)$ are only known through a finite set of sample points $\{x_i\}$ and $\{y_j\}$. This is the most typical scenario in applications; only in theoretical settings do we have access to close-form expressions for the distributions underlying the data.

2. The cost function depends on the distributions under consideration. Two related scenarios come to mind:

(a) The joint support of the distributions lies on a manifold with dimension smaller than the full space. Then the cost could be defined by the distance along the geodesics of the manifold.

(b) Distances are penalized along paths traversing sparsely populated areas, i.e. areas with small probability. For instance, for finite measures, we can conceive a graph where paths are composed of jumps between neighboring points.

Thus we will need to pose the problem in terms of the data points alone, first inferring from them the cost function $c(x, y)$ and then solving the corresponding optimal transport problem.

4 Discrete Optimal transport on a manifold

Given a set of points $Z = \{z_i\}$ (the union of the samples of all distributions being considered), we will compare two approaches to building a cost function $c_{ij} = c(z_i, z_j)$. Both are based on the matrix D_{ij} of the Euclidean distance between every pair of points.

In the first approach, we build "graph geodesics", replacing the distances between distant points by the minimal sum of distances along a chain of small-enough segments joining them. In the second approach, we use the distance defined in [1], which penalizes segments along sparsely populated domains.

4.1 Graph geodesics

In our first approach, starting with the matrix D_{ij} of the Euclidean distance between every pair of points (z_i, z_j), we replace every entry above a threshold ε by ∞, keeping finite only the distance among points relatively close to each other along the surface of the manifold.

We then determine the minimal distance between every point in the union data set using a shortest path algorithm [1].

At the core of the algorithm is an approximation of the distance between two points by replacing a large distance with a smaller distance using intermediate points: if $d_{(i,j)} > d_{(i,k)} + d_{(k,j)}$ then $d_{(i,j)} = d_{(i,k)} + d_{(k,j)}$

Depending on the distribution of the samples along the surface of the manifold, some points can still be unreachable from the others after the first two steps. For

[1] we use Floyd-Warshall $\mathcal{O}(V^3)$ but we could have used Dijkstra'algorithm which has a better performance $\mathcal{O}(V^2)$ where $|V|$ is the number of nodes

each isolated point x_k, we find the closest neighbor x_l, and we replace the infinite distance between these two points in the matrix D_{ij} by their original distance $\|x_k - x_l\|$. We then continue to apply the shortest path algorithm to the modified distance matrix D_{ij}. Following these changes, we reconnect each isolated point to the rest. We reiterate the same process over and over until every point in $\mathbf{Z}$ is no longer isolated.

We have considered two choices for the threshold ε:

$$\varepsilon = \max_i \min_j (D_{ij})$$

and

$$\varepsilon = \min_j (D_{ij})$$

Algorithm 1 Minimal distance between distributions on manifold $\mathcal{M}$

INPUT:
- Data sets: $\mathcal{D} = \{(x_i, y_j), i = 1, \cdots, n, j = 1, \cdots, m\}$, where $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$.
- The number of iteration: T
- Distance matrix A
- Minimal distance matrix D $\leftarrow$ A

OUTPUT:
- The matrix of minimal distances D_{ij} between $\{(x_i, y_j)\}$ along the manifold $\mathcal{M}$

Algorithm

Compute $\varepsilon = \max_i \min_j (D_{ij})$

Set to ∞ every $D(x_i, y_j) > \varepsilon$

$i \leftarrow 0$

while D_{ij} has ∞ entries || $i \leq T$ **do**

  $D \leftarrow$ compute minimal distances from D_{ij}

  if D has ∞ entries **then**

    Replace ∞ entries in D with corresponding smallest entries from A

  end if

  $i \leftarrow i + 1$

end while

return D.

4.2 Formulation

Once we have computed the cost function **C(X,Y)** which is the minimal distance matrix between the two datasets **X** and **Y**, we use the squared-distance cost: $\hat{C}(x, y) = \frac{1}{2}\|C(x, y)^2\|$ and the Kantorovich formulation of the optimal transport reads:

$$C(\pi) = \int_{\mu_s \times \mu_t} \hat{C}(X, Y)\pi(x, y)\, dx\, dy \tag{1}$$

subject to

$$\int_{\mu_s \times \mu_t} \pi(x, y)\, dx = \mu_t(x)$$
$$\int_{\mu_s \times \mu_t} \pi(x, y)\, dy = \mu_s(y)$$

A transfer plan π^* is said to be optimal if it minimizes the total transportation cost $C(\pi)$.

4.3 Discrete Optimal transport on a circle

The optimal transfer plan π^* solution of (1) is:

$$\pi^* = \underset{\pi \in \Pi_{\mu_s, \mu_t}}{\operatorname{argmin}} C(\pi)$$

We are seeking an optimal assignment π_{ij} between two marginal distributions $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$. The solution of the problem can be expressed as:

$$C(\pi) = \operatorname{argmin} \sum_{i,j} c_{ij}\pi_{ij} \text{ subject to} \tag{2}$$

$$\sum_j \pi_{ij} = \frac{1}{n} \text{ , n=\#points}\{\mathbf{x}_i\}$$
$$\sum_i \pi_{ij} = \frac{1}{m} \text{ , m=\#points}\{\mathbf{y}_j\}$$

To simulate this problem we generate two data sets of points normally distributed along the unit circle.

$$\theta \sim \mathcal{N}(0,\, 1),\, \varphi \sim \mathcal{N}(\frac{\pi}{2},\, \frac{1}{2}),\, x_i(\cos(\theta), \sin(\theta)),\, y_j(\cos(\varphi), \sin(\varphi)).$$

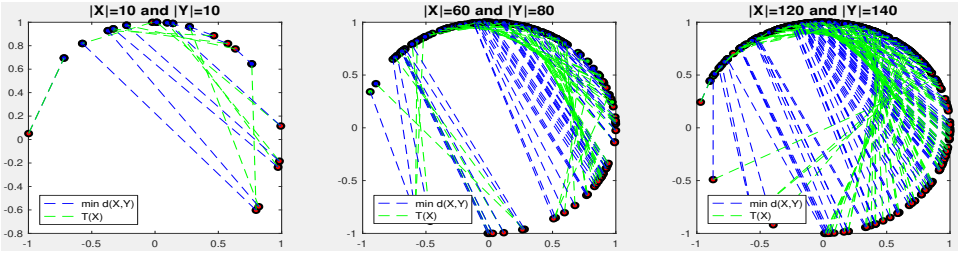

Figure 1: Example of estimated optimal map between two Gaussian distributions. (left) Same number of samples. (center) More points in each distributions. (right) The optimal mapping preserves the geometry of the manifold .

We observe that the set of the couplings generated by optimal transport follows the curvature of the circle. Differences between the closest neighbor points and the transported points are visualized by drawing lines connecting the source point to the target points, in different colors.

4.4 Discrete Wasserstein barycenter of normal distributions along a circle

The weighted barycenter of a set of points $\{\mathbf{x}_i\}$ with weights w_i is $\overline{x} = \sum_{i=1}^{N} \mathbf{w}_i \mathbf{x}_i$ which is also

$$\overline{x} = \underset{z}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbf{w}_i \|\mathbf{x}_i - z\|^2 \tag{3}$$

Similarly in the **Monge-Kantorovich** problem, the barycenter between a set of distributions $\mu_1, \cdots, \mu_N$ with weights $\mathbf{w}_1, \cdots, \mathbf{w}_N$ is the minimizer of the following problem:

$$\overline{\mu} = \underset{\rho \in P_2}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbf{w}_i W_2^2(\rho, \mu_i) \tag{4}$$

where P_2 is the metric in the space of measures provided by W_2 . And W_2 is the 2-Wasserstein distance:

$$W_2(\mu_s, \mu_t) = \left(\min_{\pi \in \Pi_{\mu_s, \mu_t}} \int_{\mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, d\pi(\mathbf{x}, \mathbf{y}) \right)^{\frac{1}{2}}$$

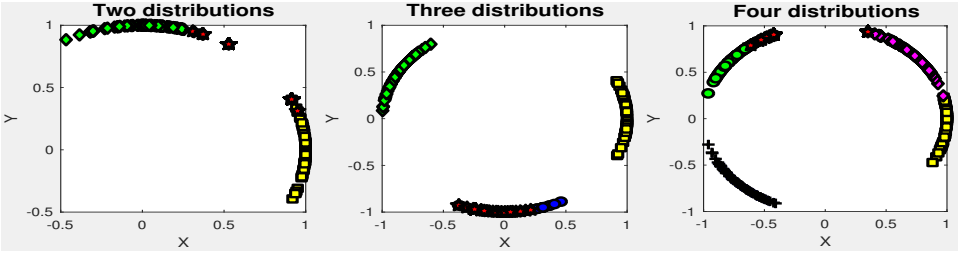

Figure 2: Example of barycenter between Gaussian distributions on a circle. The barycenter is in red .

For discrete distributions along the manifold $\mathcal{M}$, the barycenter can be characterized as:

$$\overline{\mu} = \underset{z}{\operatorname{argmin}} \sum_{ij} \mathbf{w}_i \pi_{ij} \|\mathbf{x}_i - \mathbf{z}\|^2 \tag{5}$$

where $\mathbf{x}$ are points of the distributions. Applying this result to N Gaussians distributions along a circle, and setting the weights to one, yields the Algorithm 2.

Algorithm 2 OT barycenter on manifold $\mathcal{M}$

INPUT:
- Data sets: N distributions $\{\mu_k\}(k = 1, \cdots, N)$
- The number of iteration: T
- Minimal distance matrix D between points in each distribution μ_k along the manifold $\mathcal{M}$

OUTPUT:
- barycenter $\overline{\mu}$

Algorithm

$\overline{\mu} \leftarrow$ arbitrary initial distribution μ_k

$i \leftarrow 0$

while updated $\overline{\mu}$ $||$ $i \leq T$ **do**

  find the optimal maps π_k from $\overline{\mu}$ to μ_k

  solve $\overline{\mu}$ = $\underset{z}{\operatorname{argmin}} \sum_{ij} \pi_{ij} D(x_i, z)$

  $i \leftarrow i + 1$

end while

return $\overline{\mu}$.

4.5 Experimental Results

We take our algorithm to practice applying it to the data publicly available in `https://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/2016/` consisting of hourly-measured temperatures at various stations across the United States, and in particular at one station Newton, Georgia in 2016.

The algorithm generates the monthly temperature barycenter, which we can see in green figure 3, is a smoothed version of all the twelve months for 2016. The barycenter is close to the median of the twelve months with the smallest variation.

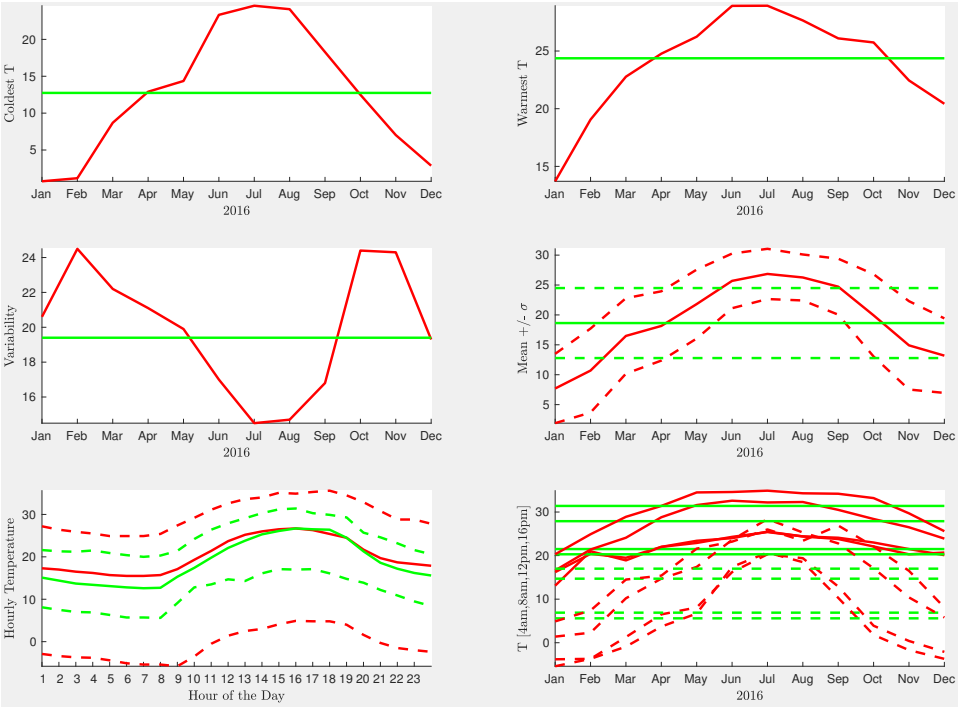

Figure 3: Monthly temperature barycenter is in green. (top): The coldest and warmest days (in average). (center left): Variability (Tmax - Tmin). (center right): Monthly mean +/- standard deviation. (bottom left): higher, median and lowest temperature at each hour.(bottom right) the min and max temperatures for each month at 4am,8am,12pm and 4pm.

Next we compare the barycenter of the initial distance matrix which is the difference in temperatures between the 366 days of 2016 and the one obtained by our algorithm which takes into account the constraints related to the manifold in which these temperatures lie.

Thus in order to capture the characteristics of the manifold supporting the data, we use two thresholds to capture the curvature of the manifold, the first one is the maximum of all the minimum temperature differences in the 366 by 366 day matrix and the other one is, given a specific day, the minimum of differences of temperatures between that day and the 365 other days.

They are close to each others with some variations which implies that the sampled temperatures in 2016 have low variance. We also use a d-distance estimator to compute the barycenter of the data distributions. The d-distance defined by (Sapienza et al. 2018 [1]) is the estimator:

For $d \geq 1$ and two points $\mathbf{p}, \mathbf{q} \in \mathcal{M} \subseteq \mathbb{R}^p$, $\mathcal{M}$ is a D-dimensional manifold, and typically $D << P$, $\mathcal{L}(.,.)$ is a distance defined on $\mathcal{M} \times \mathcal{M}$ (usually the Euclidean distance in $\mathbb{R}^p$), the d-distance estimator is defined as:

$$\mathcal{D}_{\mathbb{X}_N}(\mathbf{p}, \mathbf{q}) = \min_{(x_1,\ldots,x_K)\in\mathbb{X}_N^K} \sum_{i=1}^{K-1} \mathcal{L}(x_i, x_{i+1})^d$$

$(x_1, \ldots, x_K)$ are K sampled points from $\mathbb{X}_N$ which satisfy the minimization $x_1 = \text{argmin}_{x \in \mathbb{X}_N} \mathcal{L}(\mathbf{x}, \mathbf{p})$ and $x_K = \text{argmin}_{x \in \mathbb{X}_N} \mathcal{L}(\mathbf{x}, \mathbf{p})$

The $\mathcal{D}_{\mathbb{X}_N}$ estimator converges to the weighted geodesic distance in the sense of:

$$\lim_{N \to \infty} N^{\alpha} \mathcal{D}_{\mathcal{A}^N}(\mathbf{p}, \mathbf{q}) = K \cdot \inf_{\gamma \subset \mathcal{M}} \int_{\gamma} \frac{1}{f^{\alpha}} \tag{6}$$

where $\alpha = \frac{d-1}{D}$, K is a constant that depends on d and $\mathcal{D}$, $f : \mathcal{M} \to \mathbb{R}$ the density of the set of points $\mathbb{X}_N$ and the integral is performed over all rectifiable paths γ contained in the manifold $\mathcal{M}$.

We now compare the properties of the d-distance to two other "inferred" distances:

- "distance 1" is obtained by the minimization of the distances for the set of data points $(x_1, \ldots, x_K)$ which keeps the largest minimal distance between these points
- "distance 2" is the minimization of the distances for the same set of data points $(x_1, \ldots, x_K)$ which only keeps the minimal distance between the same points

The d-distance will eliminate consecutive points with large $\mathcal{L}$ and distances 1-2 will be more tolerant when considering neighboring points. It is illustrated in figure 4 which shows two distributions lying on a circle with some density, d-distance does not pair points, point from the first distribution with point from the second distribution, close to each other, distances "1" and "2" happen to connect such points.

We can observe that the d-distribution is including the density of the circle albeit not following the curvature of the circle.

The clustering properties of the distances using a threshold and d-distance can be further revealed using t-sne plots, like in figure 5. We set the parameter d in the d-distance to 2 but it might depend on the specific application for which the d -distance is used for. As the authors of the d-distance pointed out, d is not a parameter of the algorithm, it characterizes the macroscopic distance we want to estimate.

We can notice that the clustering properties are similar in all three plots: gradual transition between seasons (change of colors). The scatter plot using the d-distance shows an overlapping cluster of points from different seasons (different shapes) but with similar size (similar amplitude in temperature). This is due to

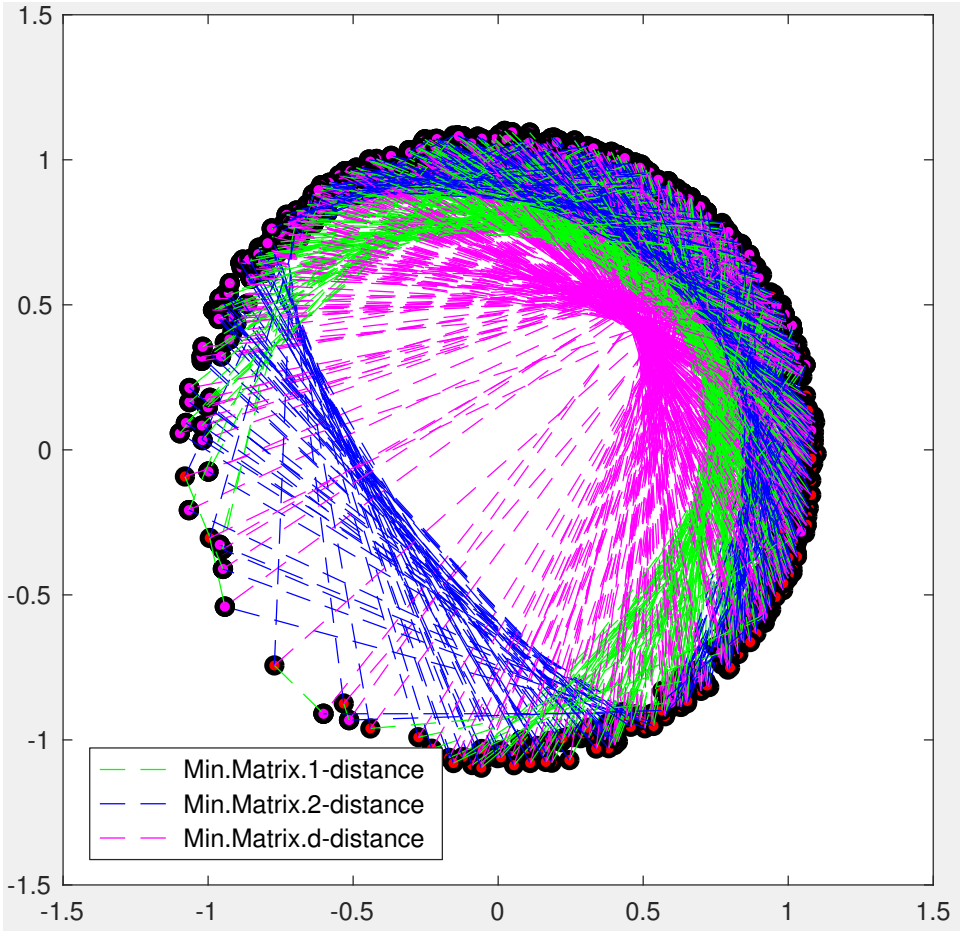

Figure 4: Optimal transport on a ”noisy” circle using as cost function 1-2 or d-distances.

the fact that the d-distance captures the differences in density of the manifold and groups together regions with equivalent densities.

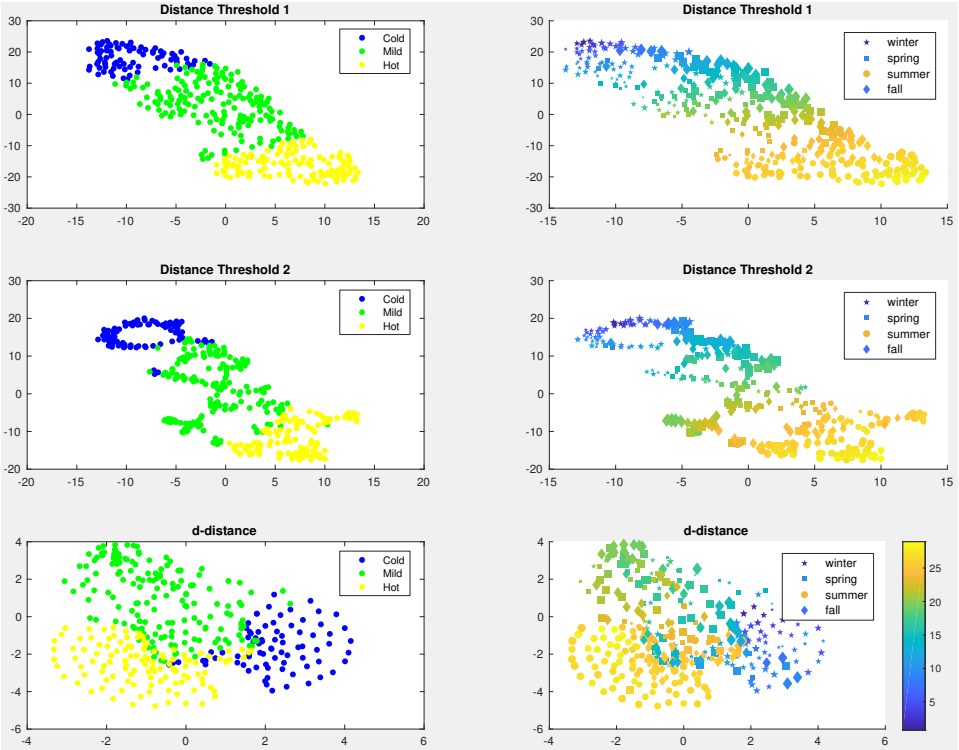

Figure 5: t-sne gscatter and scatter plots using as cost function 1-2 or d-distances. (left): clustering using the temperature categories: cold, mild, hot. (right): size of the point in the scatter plot is the max-min temperature for that day and the color is proportional to the average temperature the same day.

As we can see in figure 6, the $\mathcal{D}_{\mathcal{A}^N}$ estimator provides the optimal barycenter: it captures the properties of the manifold $\mathcal{M}$ supporting the data distribution.

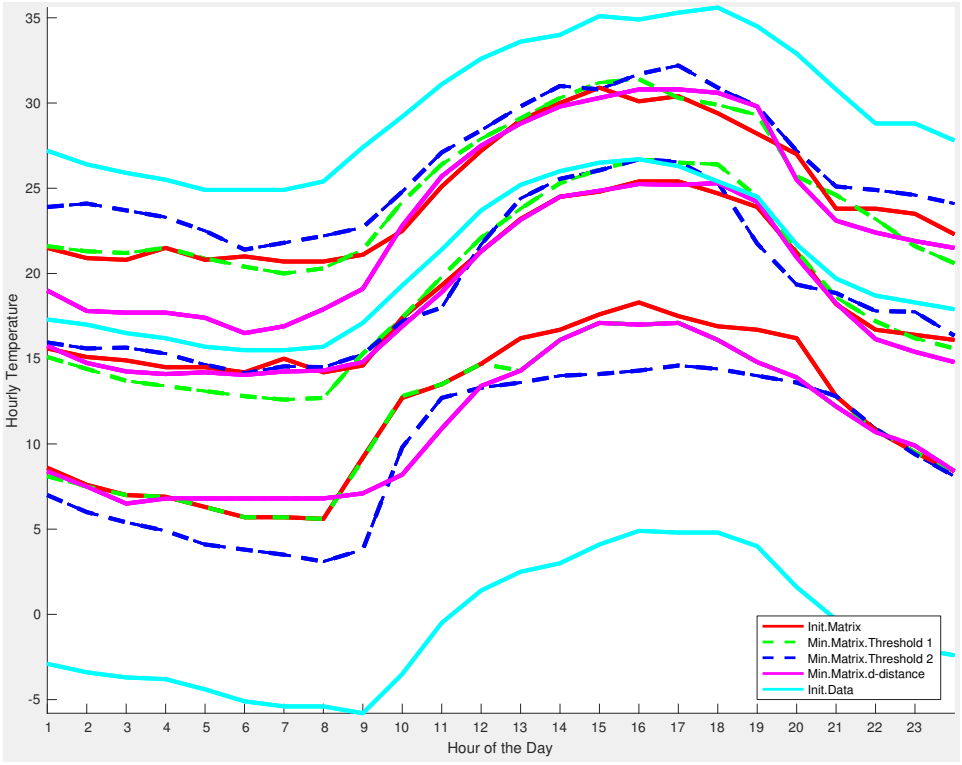

Figure 6: In light blue the min, median and max hourly temperature of the hourly temperatures in 2016. In red the min, median and max hourly temperature of the barycenter of the 2016 hourly temperatures. In bold magenta: the min, median and max hourly temperatures of the barycenter obtained using the d-distance estimator. In green and blue min, median and max hourly temperature of the barycenter for the minimum distance temperature matrices using different thresholds keeping only the closest neighbor in temperature.

5 Conclusion

This short investigation of the Optimal Transport between two data distributions constrained by the manifold characteristics has shown that the distance used to capture the cost of moving from one distribution to the other is crucial. The problem related to distribution known by a set of finite samples has been addressed in different situations. The work performed yet simple, eluded to more in depth potential research directions. Future work would be to look into the same constructs in continuous theoretical settings involving Riemann manifolds and smoothness of

the cost function.

	coldestDay	warmestDay	Max-Min (on Avg.)	Mean	Std
Initial matrix	13.1	24.1	19.4	18.7	5.3
Minimal dist. d-distance Matrix	11.8	22.7	18.1	18.1	5.2
Minimal dist. Matrix threshold 1	12.7	24.3	19.4	18.6	5.8
Minimal dist. Matrix threshold 2	13.4	24.9	24.4	18.5	6.4

Table 1: Various statistics regarding the barycenters for the initial temperatures matrix and the ”manifold” temperature matrices.

References

[1] Facundo Sapienza - 2018 *Weighted Geodesic Distance Following Fermat's Principle*.