Journal of Machine Learning Research 22 (2021) 1-43 Submitted 9/19; Revised 5/21; Published 6/21

On Efficient Multilevel Clustering via Wasserstein Distances

Viet Huynh* VIET.HUYNH@MONASH.EDU
Faculty of Information Technology, Monash University

Nhat Ho* MINHNHAT@UTEXAS.EDU
Department of Statistics and Data Sciences, University of Texas, Austin

Nhan Dam NHAN.DAM@MONASH.EDU
Faculty of Information Technology, Monash University

XuanLong Nguyen XUANLONG@UMICH.EDU
Department of Statistics, University of Michigan

Mikhail Yurochkin MIKHAIL.YUROCHKIN@IBM.COM
IBM Research

Hung Bui BUI.H.HUNG@GMAIL.COM
VinAI Research

Dinh Phung DINH.PHUNG@MONASH.EDU
Faculty of Information Technology, Monash University

Editor: Ryan Adams

Abstract

We propose a novel approach to the problem of multilevel clustering, which aims to simultaneously partition data in each group and discover grouping patterns among groups in a potentially large hierarchically structured corpus of data. Our method involves a joint optimization formulation over several spaces of discrete probability measures, which are endowed with Wasserstein distance metrics. We propose several variants of this problem, which admit fast optimization algorithms, by exploiting the connection to the problem of finding Wasserstein barycenters. Consistency properties are established for the estimates of both local and global clusters. Finally, experimental results with both synthetic and real data are presented to demonstrate the flexibility and scalability of the proposed approach.

Keywords: Optimal transport, multi-level clustering, Wasserstein barycenter

1. Introduction

In numerous applications in engineering and sciences, data are often organized in a multilevel structure. For instance, a typical structural view of text data in machine learning is to have words grouped into documents and documents grouped into corpora. A prominent strand of modeling and algorithmic work in the past couple of decades has been to discover latent multilevel structures from these hierarchically structured data. For specific clustering tasks, one may be interested in simultaneously partitioning the data in each group (to obtain local clusters) and partitioning a collection of data groups (to obtain global clusters). A concrete example is the problem of clustering images (i.e. global clusters) where each image

∗. Viet Huynh and Nhat Ho contributed equally to this work and are in random order.

contains multiple annotated regions (i.e. local clusters) (Oliva and Torralba, 2001). While hierarchical clustering techniques may be employed to find a tree-structured clustering given a collection of data points, they are not applicable to discovering the *nested* structure of multilevel data. Bayesian hierarchical models provide a powerful approach, exemplified by influential work such as (Blei et al., 2003; Pritchard et al., 2000; Teh et al., 2006). More specific to the simultaneous and multilevel clustering problem, we mention the paper of Rodriguez et al. (2008). In this interesting work, a Bayesian nonparametric model called the nested Dirichlet process (NDP) model was introduced that enables the inference of clustering of a collection of probability distributions from which different groups of data are drawn. With suitable extensions, this modeling framework has been further developed for simultaneous multilevel clustering, see, for instance, (Wulsin et al., 2016; Nguyen et al., 2014; Huynh et al., 2016).

The focus of this paper is on the multilevel clustering problem motivated in the aforementioned modeling papers, *but we shall take a pure optimization approach.* This paper includes substantially new results compared to our preliminary conference version (Ho et al., 2017). We aim to formulate optimization problems that enable the discovery of multilevel clustering structures hidden in grouped data. Our technical approach is inspired by the role of optimal transport distances in hierarchical modeling and clustering problems. The optimal transport distances, also known as Wasserstein distances (Villani, 2003), have been shown to be the natural distance metric for the convergence theory of latent mixing measures arising in both mixture models (Nguyen, 2013) and hierarchical models (Nguyen, 2016). They are also intimately connected to the problem of clustering — this relationship goes back at least to the work of (Pollard, 1982), where it is pointed out that the well-known K-means clustering algorithm can be directly linked to the quantization problem — the problem of determining an optimal finite discrete probability measure that minimizes its second-order Wasserstein distance from the empirical distribution of given data (Graf and Luschgy, 2000).

If one is to perform simultaneous K-means clustering on hierarchically grouped data, both at the global level (among groups), and local level (within each group), then this can be achieved by a joint optimization problem defined with suitable notions of Wasserstein distances inserted into the objective function. In particular, multilevel clustering requires to optimize in the space of probability measures defined in different levels of abstraction, including the space of measures of measures on the space of grouped data. Our goal, therefore, is to formulate this optimization precisely, to develop algorithms for solving the optimization problem efficiently, and to make sense of the obtained solutions in terms of statistical consistency.

The algorithms that we propose address directly a multilevel clustering problem formulated from a pure optimization viewpoint, but they may also be taken as a fast approximation to the inference of latent mixing measures that arises in the nested Dirichlet process in (Rodriguez et al., 2008). From a statistical viewpoint, we shall establish a consistency theory for our multilevel clustering problem in the manner achieved for K-means clustering (Pollard, 1982). From a computational viewpoint, quite interestingly, we will be able to explicate and exploit the connection between our optimization formulation and the problem of finding the Wasserstein barycenter (Agueh and Carlier, 2011), a computational problem

that has also attracted much recent interest, e.g., (Cuturi and Doucet, 2014; Lin et al., 2020).

In summary, the main contributions offered in this work include: *(i)* several new optimization formulations of the multilevel clustering problem using Wasserstein distances defined on different levels of the hierarchical data structure; *(ii)* fast algorithms by exploiting the connection of our formulation to the Wasserstein barycenter problem; *(iii)* consistency theorems established for the proposed estimation under a very mild condition of data distributions; *(iv)* several flexible alternatives by introducing constraints that encourage the borrowing of strength among local and global clusters; *(v)* finally, demonstration of efficiency and flexibility of our approach in a number of simulated and real datasets.

The paper is organized as follows. Section 2 provides the preliminary background on Wasserstein distances, Wasserstein barycenter, and the connection between K-means clustering and the quantization problem. Section 3 presents several optimization formulations of the multilevel clustering problem and the algorithms for solving them. Sections 4 and 5 present the alternatives of our proposed formulations under two scenarios: multilevel structure data with context and first order Wasserstein distance replacing second order Wasserstein distance for robust clustering. Section 6 establishes the consistency results for the estimators introduced in previous sections. Section 7 presents empirical studies with both synthetic and real datasets. Finally, we conclude the paper in Section 8. Additional technical details, including all proofs, are given in the appendices.

2. Background

For any given subset $\Theta \subset \mathbb{R}^d$, let $\mathcal{P}(\Theta)$ denote the space of Borel probability measures on Θ. The Wasserstein space of order $r \in [1, \infty)$ of probability measures on Θ is defined as $\mathcal{P}_r(\Theta) = \left\{ G \in \mathcal{P}(\Theta) : \int \|x\|^r \mathrm{d}G(x) < \infty \right\}$, where $\|.\|$ denotes the Euclidean metric in $\mathbb{R}^d$. Additionally, for any $k \geq 1$ the probability simplex is denoted by $\Delta_k = \left\{ u \in \mathbb{R}^k : \ u_i \geq 0, \ \sum_{i=1}^{k} u_i = 1 \right\}$. Finally, let $\mathcal{O}_k(\Theta)$ (resp., $\mathcal{E}_k(\Theta)$) be the set of probability measures with at most (resp., exactly) k support points in Θ.

Wasserstein distances: For any elements G and G' in $\mathcal{P}_r(\Theta)$ where $r \geq 1$, the Wasserstein distance of order r between G and G' is defined as (cf. (Villani, 2003)):

$$W_r(G, G') = \left(\inf_{\pi \in \Pi(G, G')} \int_{\Theta^2} \|x - y\|^r \mathrm{d}\pi(x, y) \right)^{1/r},$$

where $\Pi(G, G')$ is the set of all probability measures on $\Theta \times \Theta$ that have marginals G and G'. In words, $W_r(G, G')$ is the optimal cost of moving mass from G to G', where the cost of moving unit mass is proportional to r-power of Euclidean distance in Θ.

By the recursion of concepts, we can speak of measures of measures, and define a suitable distance metric on this abstract space: the space of Borel measures on $\mathcal{P}_r(\Theta)$, to be denoted by $\mathcal{P}_r(\mathcal{P}_r(\Theta))$. This is also a Polish space (i.e. complete and separable metric space) as $\mathcal{P}_r(\Theta)$ is a Polish space. It will be endowed with a Wasserstein metric of order r that is

induced by a metric W_r on $\mathcal{P}_r(\Theta)$ as follows (cf. Section 3 of (Nguyen, 2016)): for any $\mathcal{D}, \mathcal{D}' \in \mathcal{P}_r(\mathcal{P}_r(\Theta))$

$$W_r(\mathcal{D}, \mathcal{D}') := \Big(\inf_{\pi \in \Pi(D, D')} \int\limits_{\mathcal{P}_r(\Theta)^2} W_r^r(G, G') \mathrm{d}\pi(G, G') \Big)^{1/r},$$

where $\Pi(\mathcal{D}, \mathcal{D}')$ is the set of all probability measures on $\mathcal{P}_r(\Theta) \times \mathcal{P}_r(\Theta)$ that have marginals $\mathcal{D}$ and $\mathcal{D}'$. In words, $W_r(\mathcal{D}, \mathcal{D}')$ corresponds to the optimal cost of moving mass from $\mathcal{D}$ to $\mathcal{D}'$, where the cost of moving unit mass in its space of support $\mathcal{P}_r(\Theta)$ is proportional to the r-power of the W_r distance in $\mathcal{P}_r(\Theta)$. Note a slight abuse of notation — W_r is used for both $\mathcal{P}_r(\Theta)$ and $\mathcal{P}_r(\mathcal{P}_r(\Theta))$, but it should be clear which one is being used from context.

Entropic version of Wasserstein distances: The Wasserstein distance has been shown to have expensive computational complexity (Pele and Werman, 2009) when the probability measures are discrete. To account for the computational complexity, Cuturi (2013) proposed the entropic version of Wasserstein distances, which is given by:

$$\hat{W}_r^r(G, G') := \inf_{\pi \in \Pi(G, G')} \int\limits_{\Theta^2} \|x - y\|^r \mathrm{d}\pi(x, y) + \tau \mathrm{H}(\pi | G \otimes G'), \tag{1}$$

where $\tau > 0$ denotes the regularized parameter, $G \otimes G'$ denotes the product measure between G and G', and $\mathrm{H}(\pi | G \otimes G')$ denotes the relative entropy between π and $G \otimes G'$:

$$\mathrm{H}(\pi | G \otimes G') := \int \log \Big(\frac{d\pi}{dG \otimes G'}(x, y) \Big) d\pi(x, y),$$

where ${}^{d\pi}/_{dG \otimes G'}(x, y)$ denotes the density of π with respect to $G \otimes G'$.

When G and G' are discrete measures with at most k supports, we can compute the entropic version of Wasserstein distances via the Sinkhorn algorithm. Furthermore, by choosing the regularized parameter τ at the order of $\varepsilon / \log k$ where ε stands for the desired error, the computational complexity of the Sinkhorn algorithm for approximating the Wasserstein distance is of the order k^2/ε^2 (Dvurechensky et al., 2018). The recent works of Lin et al. (2019a,b) proposed the acceleration of the Sinkhorn algorithm with an improved complexity $k^{7/3}/\varepsilon^{4/3}$ in terms of ε. Due to the favorable performance of the Sinkhorn algorithm, throughout the paper we utilize that algorithm to compute the entropic regularized version of Wasserstein distance between the probability measures.

Wasserstein barycenter: Next, we present a brief overview of the Wasserstein barycenter problem, first studied in (Agueh and Carlier, 2011) and subsequentially many others (e.g. (Benamou et al., 2015; Solomon et al., 2015; Álvarez Estebana et al., 2016)). Given probability measures $P_1, P_2, \ldots, P_N \in \mathcal{P}_2(\Theta)$ for $N \geq 1$, their Wasserstein barycenter $\overline{P}_{N,\lambda}$ is such that

$$\overline{P}_{N,\lambda} = \arg\min_{P \in \mathcal{P}_2(\Theta)} \sum_{i=1}^{N} \lambda_i W_2^2(P, P_i), \tag{2}$$

where $\lambda \in \Delta_N$ denotes weights associated with $P_1, \ldots, P_N$. When $P_1, \ldots, P_N$ are discrete measures with finite number of atoms and the weights λ are uniform, it was shown in

(Anderes et al., 2015) that the problem of finding Wasserstein barycenter $\overline{P}_{N,\lambda}$ over the space $\mathcal{P}_2(\Theta)$ in equation (2) is reduced to the search over only a much simpler space $\mathcal{O}_l(\Theta)$ where $l = \sum_{i=1}^{N} s_i - N + 1$ and s_i is the number of components of P_i for all $1 \leq i \leq N$. Efficient algorithms for finding local solutions of the Wasserstein barycenter problem over $\mathcal{O}_k(\Theta)$ for some $k \geq 1$ have been studied recently in (Cuturi and Doucet, 2014). These algorithms will prove to be a useful building block for our method as we shall describe in the sequel.

K-means as quantization problem: The well-known K-means clustering algorithm can be viewed as solving an optimization problem that arises in the problem of quantization, a simple yet very useful connection (Pollard, 1982; Graf and Luschgy, 2000) as follows. Given n unlabeled samples $Y_1, \ldots, Y_n \in \Theta$, we assume that these data are associated with at most k clusters where $k \geq 1$ is some given number. The K-means problem finds the set S containing at most k elements $\theta_1, \ldots, \theta_k \in \Theta$ that satisfies the following objective

$$\inf_{S:|S|\leq k} \frac{1}{n} \sum_{i=1}^{n} d^2(Y_i, S), \tag{3}$$

where $d^2(Y_i, S) = \min_{\theta \in S} \|Y_i - \theta\|^2$ is the square Euclidean distance from sample Y_i to set S. Let $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$ be the empirical measure of data $Y_1, \ldots, Y_n$ where δ_Y denotes the Dirac measure centred on Y. Then, problem (3) is equivalent to finding a discrete probability measure G which has finite number of support points and solves:

$$\inf_{G \in \mathcal{O}_k(\Theta)} W_2^2(G, P_n). \tag{4}$$

Due to the inclusion of the Wasserstein metric in its formulation, we call this a *Wasserstein means problem*. This problem can be further thought of as a Wasserstein barycenter problem where $N = 1$. In light of this observation, as noted in (Cuturi and Doucet, 2014), the algorithm for finding the Wasserstein barycenter offers an alternative for the popular Loyd's algorithm for determining the local minimum of the K-means objective.

3. Clustering with multilevel structure data

Given m groups of n_j exchangeable data points $X_{j,i}$ where $1 \leq j \leq m, 1 \leq i \leq n_j$, i.e. data are represented in a two-level grouping structure, our goal is to learn about the two-level clustering structure of the data. We want to obtain simultaneously local clusters for each data group, and global clusters among all groups.

3.1 Multilevel Wasserstein means (MWM) algorithm

For any $j = 1, \ldots, m$, we denote the empirical measure for group j by $P_{n_j}^j := \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{X_{j,i}}$. Throughout this section, for simplicity of exposition, we assume that the numbers of both local and global clusters are either known or bounded above by a given number. In particular, for local clustering we allow group j to have at most k_j clusters for $j = 1, \ldots, m$.

For global clustering, we assume to have M group (Wasserstein) means among the m given groups. We now describe the high level idea of our proposed model, later elaborate its formal formulation, and demonstrate the connection to Bayesian hierarchical models.

High level idea: For local clustering, for each $j = 1, \ldots, m$, performing a K-means clustering for group j, as expressed by (4), can be viewed as finding a finite discrete measure $G_j \in \mathcal{O}_{k_j}(\Theta)$ that minimizes squared Wasserstein distance $W_2^2(G_j, P_{n_j}^j)$. For global clustering, we are interested in obtaining clusters out of m groups, each of which is now represented by the discrete measure G_j, for $j = 1, \ldots, m$. Adopting again the viewpoint of (4), provided that all of G_js are given, we can apply K-means quantization method to find their distributional clusters. The global clustering in the space of measures of measures on Θ can be succinctly expressed by

$$\inf_{\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))} W_2^2\bigg(\mathcal{H}, \frac{1}{m}\sum_{j=1}^m \delta_{G_j}\bigg).$$

However, G_js are not known — they have to be optimized through local clustering in each data group.

MWM problem formulation: We have arrived at an objective function for jointly optimizing over both local and global clusters

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta),\\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + \lambda W_2^2(\mathcal{H}, \frac{1}{m}\sum_{j=1}^m \delta_{G_j}), \tag{5}$$

where λ is a positive number used to balance the accumulative losses between the local and global clustering. We call the above optimization the problem of *Multilevel Wasserstein Means (MWM)*. The notable feature of MWM is that its loss function consists of two types of distances associated with the hierarchical data structure: one is the distance in the space of measures, i.e. $W_2^2(G_j, P_{n_j}^j)$, and the other in the space of measures of measures, i.e. $W_2^2(\mathcal{H}, \frac{1}{m}\sum_{j=1}^m \delta_{G_j})$. The global clustering in the space of measures of measures is also a special case of D2 clustering objective function in (Li and Wang, 2006). Our main difference with (Li and Wang, 2006) is the joint optimization between the local clustering and global clustering to encourage the grouping structures in the multilevel Wasserstein means.

By adopting K-means optimization to both local and global clustering, the MWM problem might look formidable at first sight. Fortunately, it is possible to simplify this original formulation substantially, by exploiting the structure of $\mathcal{H}$. Indeed, we can show that formulation (5) is equivalent to the following optimization problem, which looks much simpler as it involves only measures on Θ:

$$\inf_{G_j \in \mathcal{O}_{k_j}(\Theta), \boldsymbol{H}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{W_2}^2(G_j, \boldsymbol{H}), \tag{6}$$

where $d_{W_2}^2(G, \boldsymbol{H}) := \min_{1 \le i \le M} W_2^2(G, H_i)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$, with each $H_i \in \mathcal{P}_2(\Theta)$. The proof of this equivalence is deferred to Proposition B.4 in Appendix B.

Connection between MWM and Bayesian hierarchical models: The local measures obtained via Wasserstein means problems yield the local clustering of data in each group. Then, the K-means algorithm on these local measures leads to the global clustering of these groups. Therefore, the simultaneous local and global clusterings in the joint optimization formulation of MWM enable the discovery of nested multilevel structures hidden in grouped data. This property shares several similarities to Bayesian hierarchical models, such as the nested Dirichlet process (NDP) (Rodriguez et al., 2008).

Before going into the details of the algorithm for approximating the objective function (6) in Section 3.1.2, we shall present some simpler cases, which help to illustrate some properties of the optimal solutions of equation (6), while providing insights of subsequent developments of the MWM formulation.

3.1.1 Properties of MWM in special cases

Example 1. Suppose $k_j = 1$ and $n_j = n$ for all $1 \leq j \leq m$, and $M = 1$. Write $\boldsymbol{H} = H \in \mathcal{P}_2(\Theta)$. Under this setting, the objective function (6) can be rewritten as

$$\inf_{\substack{\theta_j \in \Theta, \\ H \in \mathcal{P}_2(\Theta)}} \sum_{j=1}^{m} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \lambda W_2^2(\delta_{\theta_j}, H)/m, \tag{7}$$

where $G_j = \delta_{\theta_j}$ for any $1 \leq j \leq m$. From the result of Theorem A.1 in the Supplement, the Wasserstein barycenter of G_j with uniform weight has exactly one component, i.e.,

$$\inf_{H \in \mathcal{P}_2(\Theta)} \sum_{j=1}^{m} W_2^2(\delta_{\theta_j}, H) \quad = \quad \inf_{H \in \mathcal{E}_1(\Theta)} \sum_{j=1}^{m} W_2^2(G_j, H) = \sum_{j=1}^{m} \|\theta_j - (\sum_{i=1}^{m} \theta_i)/m\|^2,$$

where the second infimum is achieved when $H = \delta_{(\sum_{j=1}^{m} \theta_j)/m}$. Thus, objective function (7) may be rewritten as

$$\inf_{\theta_j \in \Theta} \sum_{j=1}^{m} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \|m\theta_j - (\sum_{l=1}^{m} \theta_l)\|^2/m^3.$$

Write $\overline{X}_j = (\sum_{i=1}^{n} X_{j,i})/n$ for all $1 \leq j \leq m$. As $m \geq 2$, we can check that the unique optimal solutions for the above optimization problem are $\theta_j = \left((m^2 n + 1)\overline{X}_j + \sum_{i \neq j} \overline{X}_i\right)/(m^2 n + m)$ for any $1 \leq j \leq m$. If we further assume that our data $X_{j,i}$ are i.i.d samples from the probability measure P^j having mean $\mu_j = E_{X \sim P^j}(X)$ for any $1 \leq j \leq m$, the previous result implies that $\theta_i \not\to \theta_j$ for almost surely as long as $\mu_i \neq \mu_j$. As a consequence, if μ_j's are pairwise different, the multilevel Wasserstein means under that simple scenario of (6) will not have identical centers among local groups.

On the other hand, we have $W_2^2(G_i, G_j) = \|\theta_i - \theta_j\|^2 = \left(\frac{mn}{mn+1}\right)^2 \|\overline{X}_i - \overline{X}_j\|^2$. Now, from the definition of Wasserstein distance

$$W_2^2(P_n^i, P_n^j) \quad = \quad \min_{\sigma} \frac{1}{n} \sum_{l=1}^{n} \|X_{i,l} - X_{j,\sigma(l)}\|^2 \geq \|\overline{X}_i - \overline{X}_j\|^2,$$

where σ in the above sum varies over all the permutations of $\{1, 2, \ldots, n\}$ and the second inequality is due to Cauchy-Schwarz's inequality. It implies that as long as $W_2^2(P_n^i, P_n^j)$ is small, the optimal solution G_i and G_j of (7) will be sufficiently close to each other. By letting $n \to \infty$, we also achieve the same conclusion regarding the asymptotic behavior of G_i and G_j with respect to $W_2(P^i, P^j)$.

Example 2. Let $k_j = 1$ and $n_j = n$ for all $1 \leq j \leq m$ and $M = 2$. Write $\boldsymbol{H} = (H_1, H_2)$. Moreover, assume that there is a strict subset A of $\{1, 2, \ldots, m\}$ such that

$$C \cdot \max\left\{ \max_{i,j \in A} W_2(P_n^i, P_n^j), \max_{i,j \in A^c} W_2(P_n^i, P_n^j) \right\} \leq \min_{i \in A, j \in A^c} W_2(P_n^i, P_n^j),$$

where C is some sufficiently large positive constant, i.e., the distance of empirical measures P_n^i and P_n^j when i and j belong to the same set A or A^c is much smaller than that when i and j do not belong to the same set. Under this condition, by using the argument from Example 1 we can write the objective function (6) as

$$\inf_{\substack{\theta_j \in \Theta, \\ H_1 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_1)}{|A|} +$$

$$\inf_{\substack{\theta_j \in \Theta, \\ H_2 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A^c} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_2)}{|A^c|}.$$

The above objective function suggests that the optimal solutions θ_i, θ_j (equivalently, G_i and G_j) will not be close to each other as long as i and j do not belong to the same set A or A^c, i.e. P_n^i and P_n^j are very far. Therefore, the two groups of "local" measures G_js do not share atoms under that setting of empirical measures.

The examples examined above indicate that the MWM problem in general does not "encourage" the local measures G_js to share atoms among each other in its solution. Additionally, when the empirical measures of local groups are very close, it may also suggest that they belong to the same cluster and the distances among optimal local measures G_js can be very small.

3.1.2 Algorithm description

Now we are ready to describe our algorithm in the general case. As we mentioned in Section 2, direct computation of the second order Wasserstein distance is expensive. Therefore, we use the entropic regularized second order Wasserstein $\hat{W}_2$ to approximate the Wasserstein distance (see equation (1) for the definition). As the regularized parameter in the entropic version of second order Wasserstein distance is fixed in our experiment, we will not specifically mention that constant in $\hat{W}_2$ for the simplicity of the presentation. For the algorithmic development, we consider finding a local minimum of the entropic version of problem (6), which is given by:

$$\inf_{G_j \in \mathcal{O}_{k_j}(\Theta), \boldsymbol{H}} \sum_{j=1}^{m} \hat{W}_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{\hat{W}_2}^2(G_j, \boldsymbol{H}), \tag{8}$$

Algorithm 1 Multilevel Wasserstein Means (MWM)

Input: data $X_{j,i}$, parameters k_j and M.
Output: probability measures G_j and elements H_i of $\boldsymbol{H}$.
Initialize measures $G_j^{(0)}$, elements $H_i^{(0)}$ of $\boldsymbol{H}^{(0)}$, $t = 0$.
while $G_j^{(t)}, H_i^{(t)}$ have not converged **do**
1. Update $G_j^{(t)}$ for $1 \leq j \leq m$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1 \leq u \leq M}{\arg\min}\, \hat{W}_2^2(G_j^{(t)}, H_u^{(t)})$.
$G_j^{(t+1)} \leftarrow \underset{G_j \in \mathcal{O}_{k_j}(\Theta)}{\arg\min}\ \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda \hat{W}_2^2(G_j, H_{i_j}^{(t)})/m$.
end for
2. Update $H_i^{(t)}$ for $1 \leq i \leq M$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1 \leq u \leq M}{\arg\min}\, \hat{W}_2^2(G_j^{(t+1)}, H_u^{(t)})$.
end for
for $i = 1$ **to** M **do**
$C_i \leftarrow \{l : i_l = i\}$ for $1 \leq i \leq M$.
$H_i^{(t+1)} \leftarrow \underset{H_i \in \mathcal{P}_2(\Theta)}{\arg\min} \sum_{l \in C_i} \hat{W}_2^2(H_i, G_l^{(t+1)})$.
end for
3. $t \leftarrow t + 1$.
end while

where $d_{\hat{W}_2}^2(G, \boldsymbol{H}) := \min_{1 \leq i \leq M} \hat{W}_2^2(G, H_i)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$, with each $H_i \in \mathcal{P}_2(\Theta)$. We refer to objective function (8) as *entropic MWM*. The procedure for finding such local minimum of problem (8) is summarized in Algorithm 1. We explain the following details regarding the initialization and update steps required by the algorithm:

- The initialization of local measures $G_j^{(0)}$ (i.e., the initialization of their atoms and weights) can be obtained by performing K-means clustering on local data $X_{j,i}$ for $1 \leq j \leq m$. The initialization of elements $H_i^{(0)}$ of $\boldsymbol{H}^{(0)}$ is based on a simple extension of the K-means algorithm called three-stage K-means. Details are given in Algorithm 5 in Appendix C;

- The update of $G_j^{(t+1)}$ can be computed efficiently by simply using algorithms from (Cuturi and Doucet, 2014) to search for local solutions of these barycenter problems within the space $\mathcal{O}_{k_j}(\Theta)$ from the atoms and weights of $G_j^{(t)}$;

- Since all $G_j^{(t+1)}$s are finite discrete measures, finding the update for $H_i^{(t+1)}$ over the whole space $\mathcal{P}_2(\Theta)$ can be computationally expensive. We indeed utilize a result with Wasserstein barycenter that we can reduce this optimization problem with the Wasserstein barycenter to search for a local solution within the space $\mathcal{O}_{l^{(t)}}$, where

$l^{(t)} = \sum_{j \in C_i} |\text{supp}(G_j^{(t+1)})| - |C_i|$ from the global atoms $H_i^{(t)}$ of $\boldsymbol{H}^{(t)}$ (the justification of this reduction is derived from Theorem A.1 in Appendix A). Motivated from this result with the Wasserstein barycenter, we also only find the update for $H_i^{(t+1)}$ within the space $\mathcal{O}_{l^{(t)}}$ for the entropic version of Wasserstein barycenter. This again can be done by utilizing algorithms from (Cuturi and Doucet, 2014). Note that, as $l^{(t)}$ becomes very large when m is large, to speed up the computation of Algorithm 1 we indeed impose a threshold L, e.g., $L = 10$, for $l^{(t)}$ in the implementation.

The following guarantee for Algorithm 1 can be established:

Theorem 3.1. *For any $\lambda > 0$, the values of objective function (8) of the entropic MWM formulation at the updates $(G_1^{(t)}, G_2^{(t)}, \ldots, G_m^{(t)}, \mathbf{H}^{(t)})$ of Algorithm 1 are decreasing.*

The proof of Theorem 3.1 is in Appendix B.

3.2 Multilevel Wasserstein means with sharing

As we have observed from the analysis of several specific cases, the MWM formulation may not encourage sharing components locally among m groups in its solution. However, enforced sharing has been demonstrated to be a very useful technique, which leads to the "borrowing of strength" among different parts of the model, consequently improving the inference efficiency (Teh et al., 2006; Nguyen, 2016). In this section, we seek to encourage the borrowing of strength among groups by imposing additional constraints on the atoms of $G_1, \ldots, G_m$ in the original MWM formulation (5). Denote $\mathcal{A}_{M,\mathcal{S}_K} = \Big\{ G_j \in \mathcal{O}_K(\Theta) : \text{supp}(G_j) \subseteq \mathcal{S}_K \ \forall 1 \leq j \leq m \Big\}$ for any given $K, M \geq 1$, where the constraint set $\mathcal{S}_K$ has exactly K elements. To simplify the exposition, let us assume that $k_j = K$ for all $1 \leq j \leq m$. Consider the following locally constrained version of the MWM problem

$$\inf \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \lambda W_2^2(\mathcal{H}, \frac{1}{m}\sum_{j=1}^{m} \delta_{G_j}), \tag{9}$$

where $\mathcal{S}_K, G_j \in \mathcal{A}_{M,\mathcal{S}_K}$, $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ in the above infimum. We call the above optimization the problem of *Multilevel Wasserstein Means with Sharing (MWMS)*. The local constraint assumption $\text{supp}(G_j) \subseteq \mathcal{S}_K$ had been utilized previously in the literature — see, for example, (Kulis and Jordan, 2012), in which the authors developed an optimization-based approach to the inference of the hierarchical Dirichlet process (Teh et al., 2006), which also encourages explicitly the sharing of local group means among local clusters. Now, we can rewrite objective function (9) as follows

$$\inf \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{W_2}^2(G_j, \boldsymbol{H}), \tag{10}$$

where $\mathcal{S}_K, G_j \in \mathcal{B}_{M,\mathcal{S}_K}, \boldsymbol{H} = (H_1, \ldots, H_M)$ in the above infimum with $\mathcal{B}_{M,\mathcal{S}_K} = \Big\{ G_j \in \mathcal{O}_K(\Theta) : \text{supp}(G_j) \subseteq \mathcal{S}_K \ \forall 1 \leq j \leq m \Big\}$.

Similar to the multilevel Wasserstein means, for the algorithmic development, we also consider an entropic regularized version of MWMS, which is given by

$$\inf \sum_{j=1}^{m} \hat{W}_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{\hat{W}_2}^2(G_j, \boldsymbol{H}), \tag{11}$$

where $\mathcal{S}_K, G_j \in \mathcal{B}_{M,\mathcal{S}_K}, \boldsymbol{H} = (H_1, \ldots, H_M)$ in the infimum. We refer to objective function (11) as *entropic MWMS*.

The high level idea of finding local minimums of the objective function (11) is to first, update the elements of the constraint set $\mathcal{S}_K$ to provide the supports for local measures G_j's and then, obtain the weights of these measures as well as the elements of the global set H by computing the appropriate Wasserstein barycenters. We present the pseudocode of finding the local minimum of the entropic MWMS in Algorithm 2. We make the following remarks regarding the initialization and updates of Algorithm 2:

(i) An efficient way to initialize global set $S_K^{(0)} = \left\{a_1^{(0)}, \ldots, a_K^{(0)}\right\} \in \mathbb{R}^{d\times K}$ is to perform K-means on the whole data set $X_{j,i}$ for $1 \le j \le m, 1 \le i \le n_j$;

(ii) For any $1 \le j \le K$, the updates $a_j^{(t+1)}$ are indeed the solutions of the following optimization problems (see the proof of Theorem 3.2 in Appendix B for detailed argument about how these optimization problems are derived):

$$\min_{a_j} \left\{ m \sum_{u=1}^{m} \sum_{v=1}^{n_j} T_{j,v}^u \|a_j - X_{u,v}\|^2 + \lambda \sum_{u=1}^{m} \sum_{v} U_{j,v}^u \|a_j - h_{i_j,v}^{(t)}\|^2 \right\},$$

where T^j is an optimal coupling of $G_j^{(t)}$, P_n^j and U^j is an optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$. By taking the first order derivative of the above function with respect to $a_j^{(t)}$, we quickly achieve $a_j^{(t+1)}$ as the closed form minimum of that function;

(iii) Updating the local weights of $G_j^{(t+1)}$ is equivalent to updating $G_j^{(t+1)}$ as the atoms of $G_j^{(t+1)}$ are known to stem from $S_K^{(t+1)}$.

Now, similarly to Theorem 3.1, we also have the following theoretical guarantee regarding the behavior of Algorithm 2.

Theorem 3.2. *For any $\lambda > 0$, the values of objective function* (11) *of the entropic MWMS formulation at the updates $(S_K^{(t)}, G_1^{(t)}, G_2^{(t)}, \ldots, G_m^{(t)}, \mathbf{H}^{(t)})$ of Algorithm 2 are decreasing.*

The proof of Theorem 3.2 is in Appendix B.

4. Extension to multilevel structure data with context

So far, we have considered the setting of multilevel structure data without any additional structure. However, complex multilevel data in practice usually include more structures. In this section, we consider a specific setting of multilevel data with context (Nguyen et al.,

Algorithm 2 Multilevel Wasserstein Means with Sharing (MWMS)

Input: Data $X_{j,i}$, K, M, λ.
Output: global set S_K, local measures G_j, and elements H_i of $\boldsymbol{H}$.
Initialize $S_K^{(0)} = \left\{a_1^{(0)}, \ldots, a_K^{(0)}\right\}$, measures $G_j^{(0)}$ from $S_K^{(0)}$, elements $H_i^{(0)}$ of $\boldsymbol{H}^{(0)}$, and $t = 0$.
while $S_K^{(t)}, G_j^{(t)}, H_i^{(t)}$ have not converged **do**
1. Update global set $S_K^{(t)}$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1\le u\le M}{\arg\min}\, \hat{W}_2^2(G_j^{(t)}, H_u^{(t)})$.
$T^j \leftarrow$ optimal coupling of $G_j^{(t)}$, P_n^j, $U^j \leftarrow$ optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$.
end for
for $i = 1$ **to** M **do**
$h_i^{(t)} \leftarrow$ atoms of $H_i^{(t)}$ with $h_{i,v}^{(t)}$ as v-th column.
end for
for $i = 1$ **to** K **do**
$(m+\lambda)D \leftarrow m\sum_{u=1}^{m}\sum_{v=1}^{n_i} T_{i,v}^u + \lambda\sum_{u=1}^{m}\sum_{v\neq i} U_{i,v}^u$.
$a_i^{(t+1)} \leftarrow \left(m\sum_{u=1}^{m}\sum_{v=1}^{n_i} T_{i,v}^u X_{u,v} + \lambda\sum_{u=1}^{m}\sum_{v} U_{i,v}^u h_{j_u,v}^{(t)}\right)/mD$.
end for
2. Update $G_j^{(t)}$ for $1 \le j \le m$:
for $j = 1$ **to** m **do**
$G_j^{(t+1)} \leftarrow \underset{G_j:\mathrm{supp}(G_j)\equiv S_K^{(t+1)}}{\arg\min} \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda\hat{W}_2^2(G_j, H_{i_j}^{(t)})/m$.
end for
3. Update $H_i^{(t)}$ for $1 \le i \le M$ as in Algorithm 1.
4. $t \leftarrow t + 1$.
end while

2014; Huynh et al., 2016) and develop efficient methods to cluster these data based on the idea of Wasserstein means as in the previous sections. Assume now that we have m groups of n_j exchangeable data points $X_{j,i}$ where $1 \le j \le m, 1 \le i \le n_j$. For each group j, $\phi_j \in \mathbb{R}^{d_2}$ denotes the observed group-specific context. Our goal is to utilize the group-specific context to learn about the two-level clustering structure of the data. Similarly to the setting in Section 3, we assume that we have at most k_j clusters of group j for $j = 1, \ldots, m$ and M groups of global clustering and contexts.

Based on the idea of MWM developed earlier, regarding local clustering we perform K-means clustering for group j, for each $j = 1, \ldots, m$, to find a discrete measure $G_j \in \mathcal{O}_{k_j}(\Theta)$ that minimizes $W_2^2(G_j, P_{n_j}^j)$. Since we would like to incorporate group context ϕ_j to study the two-level clustering structure of the data, the global clustering can be expressed as

$$\inf_{\mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta)\times\mathbb{R}^{d_2})} W_2^2\left(\mathcal{H}, \frac{1}{m}\sum_{j=1}^{m}\delta_{(G_j,\phi_j)}\right).$$

By combining the losses from local and global clustering, we arrive at the following objective function

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta) \\ H \in \mathcal{E}_M(\mathcal{P}_2(\Theta)\times\mathbb{R}^{d_2})}} \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \lambda W_2^2(H, \frac{1}{m}\sum_{j=1}^{m} \delta_{(G_j,\phi_j)}), \tag{12}$$

where $\lambda > 0$ is a chosen penalty number. We call the above optimization the problem of *Multilevel Wasserstein Means with Context (MWMC)*. Similarly to the case of MWM, the objective function of MWMC can be rewritten as follows

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ H \in (\mathcal{P}_2(\Theta)\times\mathbb{R}^{d_2})^M}} \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{W_2}^2\bigg((G_j, \phi_j), \boldsymbol{H}\bigg), \tag{13}$$

where $d_{W_2}^2\left((G,\phi), \boldsymbol{H}\right) = \min\limits_{1\le i\le m} \left\{W_2^2(G, H_i) + \|\phi - \theta_i\|^2\right\}$ and $\boldsymbol{H} = \{(H_1,\theta_1),\ldots,(H_M,\theta_M)\}$ $\in (\mathcal{P}_2(\Theta)\times\mathbb{R}^{d_2})^M$.

Similar to both the MWM and MWMS, we also consider the entropic version of MWMC for the computational purpose, which we refer to as *entropic MWMC*. The objective function of the entropic MWMC is given by:

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ H \in (\mathcal{P}_2(\Theta)\times\mathbb{R}^{d_2})^M}} \sum_{j=1}^{m} \hat{W}_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{\hat{W}_2}^2\bigg((G_j, \phi_j), \boldsymbol{H}\bigg). \tag{14}$$

The procedure for finding a local minimum of the entropic MWMC objective function (14) is summarized in Algorithm 3. Here, we have the following comments regarding the initialization and update steps of that algorithm:

- The initialization of local measures $G_j^{(0)}$ and global set $\boldsymbol{H}^{(0)}$ can be carried out in a similar fashion as those in Algorithm 1. Furthermore, the initialization of $\theta_i^{(0)}$ can be obtained by performing K-means++ clustering proposed in (Arthur and Vassilvitskii, 2007) on the context data $\phi_1,\ldots,\phi_m$.
- The approaches to update $G_j^{(t+1)}$ and $\boldsymbol{H}^{(t+1)}$ are similar to those in Algorithm 1. The update of $\theta_i^{(t+1)}$ is to find an optimal center to minimize its distance to context ϕ_l for all $l \in C_i$.

Similar to Theorems 3.1 and 3.2, we also have the following guarantee regarding the performance of Algorithm 3:

Theorem 4.1. *For any $\lambda > 0$, the values of objective function* (14) *of the entropic MWMC formulation at the updates $(G_1^{(t)}, G_2^{(t)}, \ldots, G_m^{(t)}, \mathbf{H}^{(t)}, \theta_1^{(t)}, \ldots, \theta_M^{(t)})$ of Algorithm 3 are decreasing.*

The proof of Theorem 4.1 is similar to that of Theorem 3.1; therefore, it is omitted. Finally, we would like to remark that the extension of the MWMC problem to the setting in which local measures G_j share atoms among others can be carried out in a similar fashion as that in Section 3.2.

Algorithm 3 Multilevel Wasserstein Means with Context (MWMC)

Input: data $X_{j,i}$, context ϕ_j, parameters k_j and M.
Output: probability measures G_j and elements (H_i, θ_i) of $\boldsymbol{H}$.
Initialize measures $G_j^{(0)}$, elements $(H_i^{(0)}, \theta_i^{(0)})$ of $\boldsymbol{H}^{(0)}$, $t = 0$.
while $G_j^{(t)}, H_i^{(t)}, \theta_i^{(t)}$ have not converged **do**
1. Update $G_j^{(t)}$ for $1 \le j \le m$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1\le u\le M}{\arg\min}\, \hat{W}_2^2(G_j^{(t)}, H_u^{(t)}) + \|\phi_j - \theta_u^{(t)}\|^2$.
$G_j^{(t+1)} \leftarrow \underset{G_j\in\mathcal{O}_{k_j}(\Theta)}{\arg\min}\, \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda \hat{W}_2^2(G_j, H_{i_j}^{(t)})/m$.
end for
2. Update $(H_i^{(t)}, \theta_i^{(t)})$ for $1 \le i \le M$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1\le u\le M}{\arg\min}\, \hat{W}_2^2(G_j^{(t+1)}, H_u^{(t)}) + \|\phi_j - \theta_u^{(t)}\|^2$.
end for
for $i = 1$ **to** M **do**
$C_i \leftarrow \{l : i_l = i\}$ for $1 \le i \le M$.
$H_i^{(t+1)} \leftarrow \underset{H_i\in\mathcal{P}_2(\Theta)}{\arg\min} \sum_{l\in C_i} \hat{W}_2^2(H_i, G_l^{(t+1)})$.
$\theta_i^{(l+1)} = (\sum_{l\in C_i} \phi_l)/|C_i|$.
end for
3. $t \leftarrow t + 1$.
end while

5. Robust multilevel clustering with first order Wasserstein distance

In the previous sections, we develop our models based on W_2 metric. Nevertheless, these formulations do not directly account for the outliers in the data. In order to improve the robustness, it may be desirable to make use of the first order Wasserstein distance instead of the second order one. In particular, we reformulate MWM objective function in Section 3 based on W_1 distance as follows

$$\inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{j=1}^m W_1(G_j, P_{n_j}^j) + \lambda W_1(\mathcal{H}, \frac{1}{m}\sum_{j=1}^m \delta_{G_j}), \tag{15}$$

where λ is a positive number used to balance the accumulative losses between the local and global clustering. We call the above optimization the problem of *Multilevel Wasserstein Geometric Median* (MWGM). Similarly to the equivalence between objective functions (3) and (6), we can demonstrate that the objective function (15) is equivalent to the following simpler optimization problem

$$\inf_{G_j\in\mathcal{O}_{k_j}(\Theta),\boldsymbol{H}} \sum_{j=1}^m W_1(G_j, P_{n_j}^j) + \frac{\lambda}{m} d_{W_1}(G_j, \boldsymbol{H}), \tag{16}$$

where $d_{W_1}(G, \boldsymbol{H}) := \min_{1 \le i \le M} W_1(G, H_i)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$, with each $H_i \in \mathcal{P}_2(\Theta)$.

For the algorithmic development, we also consider an entropic version of the MWGM, which admits the following form:

$$\inf_{G_j \in \mathcal{O}_{k_j}(\Theta), \boldsymbol{H}} \sum_{j=1}^{m} \hat{W}_1(G_j, P^j_{n_j}) + \frac{\lambda}{m} d_{\hat{W}_1}(G_j, \boldsymbol{H}). \tag{17}$$

We refer to the objective function (17) as *entropic MWGM*. Unlike our previous algorithms, the algorithm to study objective function (17) relies on the update with Wasserstein barycenter under the entropic version of W_1 distance, which means we need to solve the following optimization problem

$$\overline{Q}_{N,\lambda} = \arg\min_{Q \in \mathcal{P}_1(\Theta)} \sum_{i=1}^{N} \eta_i \hat{W}_1(Q, Q_i), \tag{18}$$

where $Q_1, \ldots, Q_N \in \mathcal{P}_1(\Theta)$ for $N \ge 1$ and $\eta \in \Delta_N$ denotes weights associated with $Q_1, \ldots, Q_N$. In Appendix D, we provide an efficient algorithm to determine the Wasserstein barycenter over $\mathcal{O}_k(\Theta)$ under entropic version of W_1 distance when Q_i's are all discrete measures with finite number of atoms. The high level idea of our algorithm is to utilize the dual transportation formulation from (Cuturi and Doucet, 2014) to update weights of the barycenter while we use the idea of weighted geometric median to update the atoms of the barycenter. The algorithm for finding a local minimum of problem (17) is summarized in Algorithm 4.

Similarly to the previous algorithms, we also have the following guarantee regarding the performance of Algorithm 4.

Theorem 5.1. *For any $\lambda > 0$, the values of objective function* (17) *of the entropic MWGM formulation at the updates $(G_1^{(t)}, G_2^{(t)}, \ldots, G_m^{(t)}, \mathbf{H}^{(t)})$ of Algorithm 4 are decreasing.*

The proof of Theorem 5.1 is similar to that of Theorem 3.1; therefore, it is omitted.

6. Consistency results

We proceed to establish consistency for the estimators introduced in the previous sections. For the brevity of the presentation, we only focus on the MWM method while the consistency for MWMS, MWMC, and MWGM can be obtained in the similar fashion. The consistency of the estimators on the MWGM method is presented in Appendix E. Fix m and assume that P^j is the true distribution of data $X_{j,i}$ for $j = 1, \ldots, m$. Write $\boldsymbol{G} = (G_1, \ldots, G_m)$ and $\boldsymbol{n} = (n_1, \ldots, n_m)$. We say $\boldsymbol{n} \to \infty$ if $n_j \to \infty$ for $j = 1, \ldots, m$. Define the following functions

$$f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) = \sum_{j=1}^{m} W_2^2(G_j, P^j_{n_j}) + \lambda W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^{m} \delta_{G_j}),$$

$$f(\boldsymbol{G}, \mathcal{H}) = \sum_{j=1}^{m} W_2^2(G_j, P^j) + \lambda W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^{m} \delta_{G_j}),$$

Algorithm 4 Multilevel Wasserstein Geometric Median (MWGM)

Input: data $X_{j,i}$, parameters k_j and M.
Output: probability measures G_j and elements H_i of $\boldsymbol{H}$.
Initialize measures $G_j^{(0)}$, elements $H_i^{(0)}$ of $\boldsymbol{H}^{(0)}$, $t = 0$.
while $G_j^{(t)}, H_i^{(t)}$ have not converged **do**
1. Update $G_j^{(t)}$ for $1 \le j \le m$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1\le u\le M}{\arg\min}\, \hat{W}_1(G_j^{(t)}, H_u^{(t)})$.
$G_j^{(t+1)} \leftarrow \underset{G_j\in\mathcal{O}_{k_j}(\Theta)}{\arg\min}\ \hat{W}_1(G_j, P_{n_j}^j) + \lambda \hat{W}_1(G_j, H_{i_j}^{(t)})/m$.
end for
2. Update $H_i^{(t)}$ for $1 \le i \le M$:
for $j = 1$ **to** m **do**
$i_j \leftarrow \underset{1\le u\le M}{\arg\min}\, \hat{W}_1(G_j^{(t+1)}, H_u^{(t)})$.
end for
for $i = 1$ **to** M **do**
$C_i \leftarrow \{l : i_l = i\}$ for $1 \le i \le M$.
$H_i^{(t+1)} \leftarrow \underset{H_i\in\mathcal{P}_2(\Theta)}{\arg\min} \sum_{l\in C_i} \hat{W}_1(H_i, G_l^{(t+1)})$.
end for
3. $t \leftarrow t + 1$.
end while

where $G_j \in \mathcal{O}_{k_j}(\Theta)$, $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$ for $1 \le j \le m$. The first consistency property of the WMW formulation is as follows.

Theorem 6.1. *Given that $P^j \in \mathcal{P}_2(\Theta)$ for $1 \le j \le m$, then the following holds:*
(i) There holds almost surely, as $\boldsymbol{n} \to \infty$

$$\inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) - \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f(\boldsymbol{G}, \mathcal{H}) \to 0.$$

(ii) If Θ is bounded, then

$$\left| \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\boldsymbol{n}}^{1/2}(\boldsymbol{G}, \mathcal{H}) - \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} f^{1/2}(\boldsymbol{G}, \mathcal{H}) \right| = \begin{cases} O_P(m \cdot n_\vee^{-1/d}) \text{ when } d \ge 5, \\ O_P(m \cdot n_\vee^{-1/4}) \text{ when } d \le 4, \end{cases}$$

where $n_\vee = \max_{1\le i\le m} n_i$.

The proof of Theorem 6.1 is in Appendix B. The next theorem establishes that the "true" global and local clusters can be recovered. To this end, assume that for each $\boldsymbol{n}$ there is an optimal solution $(\widehat{G}_1^{n_1}, \ldots, \widehat{G}_m^{n_m}, \widehat{\mathcal{H}}^{\boldsymbol{n}})$ or in short $(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \mathcal{H}^{\boldsymbol{n}})$ of the objective function (5). Moreover, there exist a (not necessarily unique) optimal solution minimizing $f(\boldsymbol{G}, \mathcal{H})$

over $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$. Let $\mathcal{F}$ be the collection of such optimal solutions. For any $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$, we define

$$d(\boldsymbol{G}, \mathcal{H}, \mathcal{F}) = \inf_{(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}} \sum_{j=1}^{m} W_2^2(G_j, G_j^0) + W_2^2(\mathcal{H}, \mathcal{H}^0).$$

Given the above assumptions, we have the following result regarding the convergence of $(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \mathcal{H}^{\boldsymbol{n}})$:

Theorem 6.2. *Assume that Θ is bounded and $P^j \in \mathcal{P}_2(\Theta)$ for all $1 \leq j \leq m$. Then, we have $d(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}, \mathcal{F}) \to 0$ as $\boldsymbol{n} \to \infty$ almost surely.*

The proof of Theorem 6.2 is in Appendix B.

Remark: (i) The assumption that Θ is bounded is just for the convenience of proof argument. From the work of Pollard (1981), the consistency of K-means clustering is established under no assumptions on the compactness of the parameters. The idea of this paper is to show that when the sample size is sufficiently large, the optimal clusters are contained in a ball centered at the origin of sufficiently large radius. From here, we can reduce the consistency analysis to compact subset argument with the cluster centers. To extend the result of Theorem 6.2 when $\Theta = \mathbb{R}^d$, we also need a similar step to prove that the supports of local and global measures from multilevel Wasserstein means will be in a ball of sufficiently large radius when the sample sizes of local groups and the number of groups are sufficiently large. It is beyond the proof technique of the current paper and we leave it for the future work. (ii) If $|\mathcal{F}| = 1$, i.e. there exists a unique optimal solution $(\boldsymbol{G}^0, \mathcal{H}^0)$ minimizing $f(\boldsymbol{G}, \mathcal{H})$ over $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$, the result of Theorem 6.2 implies that $W_2(\widehat{G}_j^{n_j}, G_j^0) \to 0$ for $1 \leq j \leq m$ and $W_2(\widehat{\mathcal{H}}^{\boldsymbol{n}}, \mathcal{H}^0) \to 0$ as $\boldsymbol{n} \to \infty$.

7. Empirical studies

In this section, we present extensive simulation studies with our models under both synthetic data (Section 7.2) and real data (Section 7.3)[1]. In our experiments, we used a fixed value of all entropic regularization parameters $\tau = 10$. For the regularized term λ, we heuristically choose to balance global and local terms, i.e., $\lambda \approx {W_2^2(\mathcal{H}, 1/m \sum_{j=1}^m \delta_{G_j})}/{\sum_{j=1}^m W_2^2(G_j, P_{n_j}^j)}$. Before presenting experimental results, we describe our methodology to scale up our algorithms using a distributed computation framework in the following section.

7.1 Parallel implementation with Apache Spark

The running time complexity of the MWM and MWMS algorithms will be increased linearly with m — the number of data groups. When the number of data groups is large (e.g., tens of thousands or millions), the running time for the learning routine in these algorithms is dramatically increased. One possible solution to adapt MWM and MWMS algorithms for large-scale settings is to parallelize the learning process. Fortunately, Apache Spark provides an elegant framework to help us to accelerate running time with the map-reduce

1. Code is available at `https://github.com/viethhuynh/wasserstein-means`

Algorithm 5 MapReduce for Multilevel Wasserstein Means (MWM)

method MAP (j,$G_j^{(t)}, \boldsymbol{H}^{(t)}$)

$i_j \leftarrow \arg\min_{1\le u\le M} \hat{W}_2^2(G_j^{(t)}, H_u^{(t)})$.

$G_j^{(t+1)} \leftarrow \arg\min_{G_j\in\mathcal{O}_{k_j}(\Theta)} \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda\hat{W}_2^2(G_j, H_{i_j}^{(t)})/m$.

return $(i_j, G_j^{(t+1)})$

method REDUCE ($\{i_j\}, \{G_j^{(t+1)}\}$)

for $i = 1$ **to** M

$C_i \leftarrow \{l : i_l = i\}$ for $1 \le i \le M$.

$H_i^{(t+1)} \leftarrow \arg\min_{H_i\in\mathcal{P}_2(\Theta)} \sum_{l\in C_i} \hat{W}_2^2(H_i, G_l^{(t+1)})$.

endfor

return $\boldsymbol{H}$

mechanism. We use the Apache Spark framework to simultaneously update the clustering index and atoms of each data group in these algorithms. We summarize the pseudo-code of Map-Reduce procedures for MWM in Algorithm 5. Our parallelized algorithm can speed up the learning algorithms up to some orders-of-magnitude, which allows us to scale up the learning problem toward large real-world datasets containing millions of groups. The orders-of-magnitude speedup depends on the number of processors (or cores) that the systems process. For instance, if we have a cluster with hundreds or thousands of cores, the orders of magnitude can be to 2 or 3. Our implementation in this paper focuses on a large number of groups, therefore, it does not help to improve running time when the number of supports is large. However, to deal with a large number of supports, we can use the GPU implementation of Cuturi's algorithms. Since Apache Spark supports distributed systems running with GPU platforms, we can obtain scalable algorithms with a large number of groups as well as the vast size of the supports.

7.2 Synthetic data

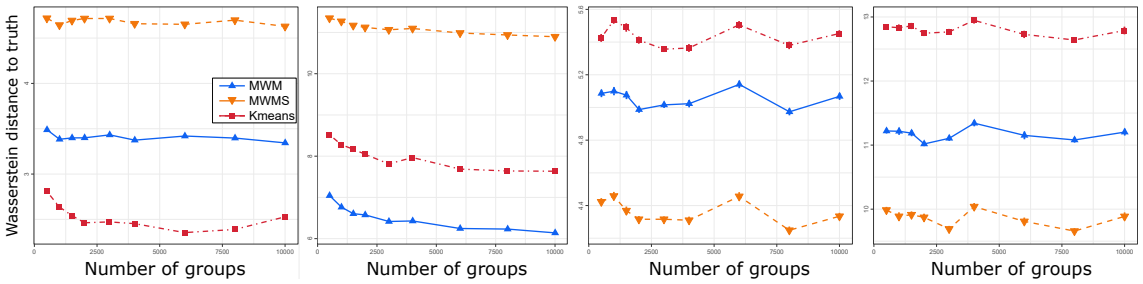


Figure 1: Data with a lot of small groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

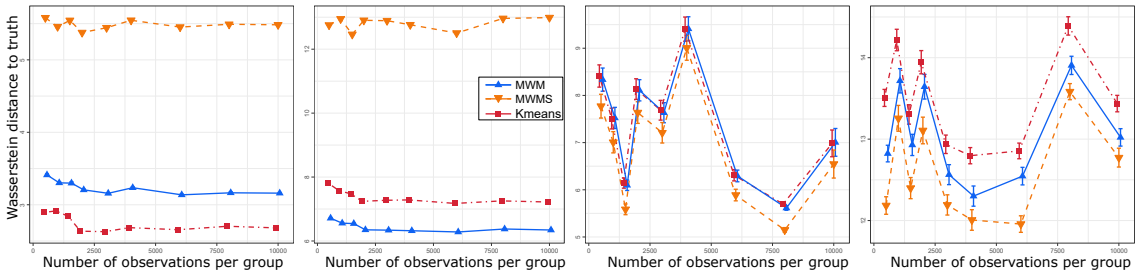


Figure 2: Data with a few big groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

First, we are interested in evaluating the effectiveness of all clustering algorithms in the paper by considering different synthetic data generating processes. Unless otherwise specified, we set the number of groups $m = 50$, number of observations per group $n_j = 50$, dimensions of each observation $d = 10$, number of global clusters $M = 5$ with 6 atoms.

Multilevel Wasserstein means: For Algorithm 1 (MWM), local measures G_j have 5 atoms each. We call this no-constraint (NC) data; for Algorithm 2 (MWMS) the number of atoms in the constraint set S_K is 50. These atoms are shared among local measures G_j each of which contains 5 atoms. We call this local-constraint (LC) data. As a benchmark for the comparison, we will use a basic 3-stage K-means approach, (the details of data generation and 3-stage K-means algorithm can be found in Appendix C). The Wasserstein distances between the estimated distributions (i.e. $\hat{G}_1, \ldots, \hat{G}_m$; $\hat{H}_1, \ldots, \hat{H}_M$) and the data generating ones will be used as the comparison metric.

Recall that the MWM formulation does not impose constraints on the atoms of G_i whilst the MWMS formulation explicitly enforces the sharing of atoms across these measures. We use multiple layers of mixtures while adding Gaussian noise at each layer to generate global and local clusters and the no-constraint (NC) data. We vary the number of groups m from 500 to 10,000. We notice that the 3-stage K-means algorithm performs best when there is no constraint structure *and* the variance is constant across all clusters (Figures 1a and 2a) — this is, not surprisingly, a favorable setting for the basic K-means method. As soon as we depart from the (unrealistic) constant variance no-sharing assumption, our MWM algorithm starts to outperform the basic 3-stage K-means (Figures 1b and 2b). The superior performance is most prominent with local-constraint (LC) data (with or without constant variance condition) (see Figures 1(c, d)). It is worth noting that even when the group variances are constant, the 3-stage K-means is no longer effective because it fails to account for the shared structure. When $m = 50$ and group sizes are larger, we set $S_K = 15$. The results reported in Figures 2(c, d) demonstrate the effectiveness and flexibility of our algorithms.

Multilevel Wasserstein means with context: Now, we demonstrate the capability of the MWMC framework to model the synthetic multilevel data with context. There are six clusters, each of which is denoted as a column in the ground truth row in Figure 3. In each cluster, the content data (bottom square) is selected as a mixture of three Gaussian components from a set of six shared Gaussian components whilst the context data

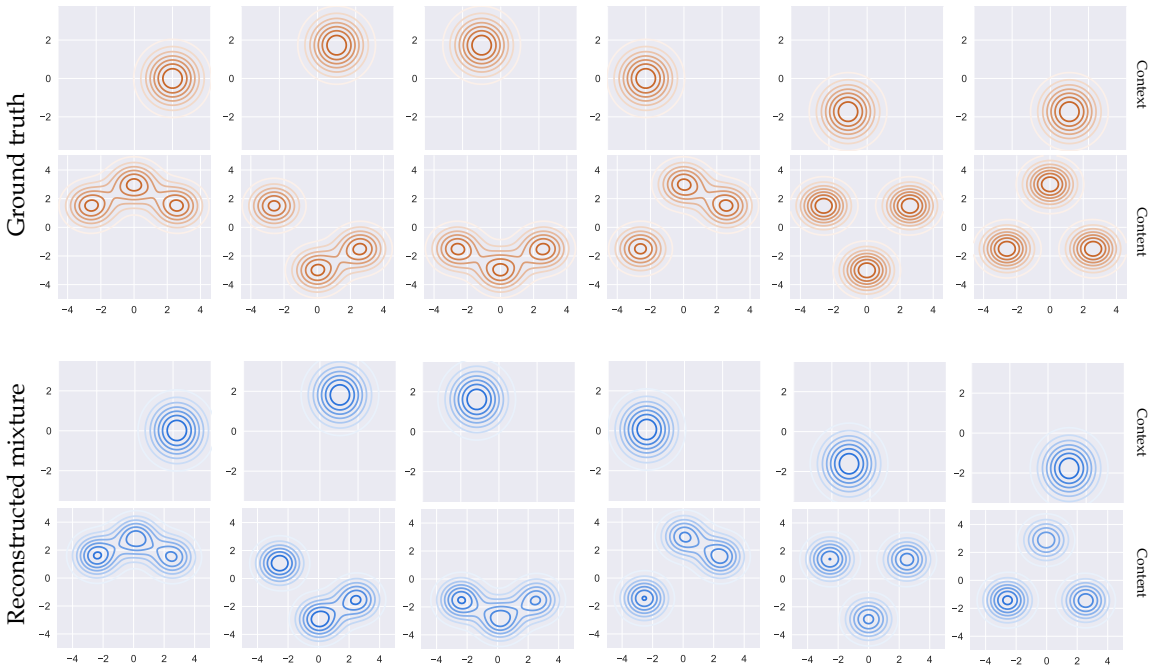


Figure 3: Synthetic data with context.

(top square) is a Gaussian distribution selected from six predefined Gaussian components. Visually, the top two rows in Figure 3 show the ground truth data including context (a Gaussian distribution) and content (a Gaussian mixture model). We uniformly generate 3000 groups of data from the above six clusters. Each group belongs to one of the six aforementioned clusters. Once the clustering index of a data group has been determined, we generate 100 data points from the corresponding mixture of Gaussians and a corresponding context observation. We run the synthetic data with the MWMC algorithm. The bottom two rows in Figure 3 depict the reconstructed context and content data which are similar to the ground truth.

7.3 Real-world data analysis

We now apply our multilevel clustering algorithms to two real-world datasets: LabelMe and StudentLife.

LabelMe dataset[2] consists of $2,688$ annotated images which are classified into 8 scene categories including *tall buildings, inside city, street, highway, coast, open country, mountain,* and *forest* (Oliva and Torralba, 2001). Each image contains multiple annotated regions. Each region, which is annotated by users, represents an object in the image. As shown in Figure 4, the left image is an example from *open country* category and contains 4 regions while the right panel shows an image of *tall buildings* category including 16 regions. Note that the regions in each image can be overlapped. Removing the images containing less than 4 regions, we obtain $1,800$ images for our experiments. We then extract the GIST

2. http://labelme.csail.mit.edu

Table 1: Clustering performance for LabelMe dataset. MWM = Multilevel Wasserstein Means and MWMS = MWM with shared atoms on the measure space.

Methods	NMI	ARI	AMI
K-means	0.370	0.282	0.365
TSK-means	0.203	0.101	0.170
MC2 (Huynh et al., 2016)	0.315	0.206	0.273
MWM	0.374	0.302	0.368
MWMS	**0.416**	**0.355**	**0.411**
MWM with context	0.662	0.580	0.654
MWMS with context	**0.675**	**0.603**	**0.666**

feature (Oliva and Torralba, 2001) for each region in an image. GIST is a 512-dimensional visual descriptor representing perceptual dimensions and oriented spatial structures of a scene. We further use PCA to project GIST features onto 30 dimensions. Consequently, we obtain $1,800$ "documents", each of which contains regions as observations. Each region is represented by a 30-dimensional vector. We now can perform clustering regions in every image since they are visually correlated. In the next level of clustering, we can cluster images into scene categories. Each image in the LabelMe dataset has associated tags. These tags of each image are represented as a count context vector

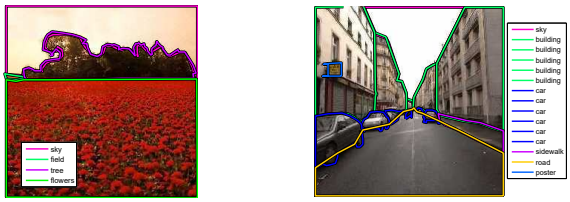
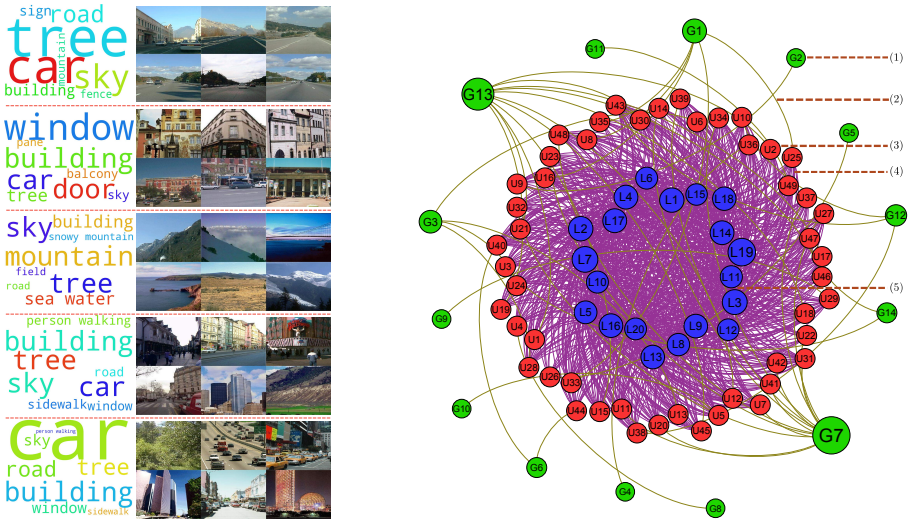


Figure 4: Sample images in LabelMe dataset. Each image consists of different number of annotated regions.

StudentLife dataset[3] is a large dataset frequently used in pervasive and ubiquitous computing research. Data signals consist of multiple channels (e.g. WiFi signals and Bluetooth scan), which are collected from smartphones of 49 students at Dartmouth College over a 10-week spring term in 2013. However, in our experiments, we use only WiFi signal strengths. We apply a similar procedure described in (Nguyen et al., 2016) to pre-process the data. We aggregate the number of scans by each WiFi access point and select 500 WiFi IDs with the highest frequencies. Eventually, we obtain 49 "documents" with totally approximately 4.6 million 500-dimensional data points.

Clustering evaluation metrics: We use normalized Mutual Information (NMI), adjusted Rand index (ARI), and adjusted Mutual Information (AMI) to evaluate the performance metrics of clustering tasks. NMI measures mutual information between ground truth labels of data and predicted cluster labels normalized by the average entropy of these labels, $NMI(U,V) = {}^{2*MI(U,V)}/_{(H(U)+H(V))}$ where $MI(.,.)$ and $H(.)$ are respec-

3. https://studentlife.cs.dartmouth.edu/dataset.html



(a) (b)

Figure 5: Clustering representation for two datasets: (a) Five image clusters from *Labelme* data discovered by MWMS algorithm: tag-clouds on the left are accumulated from all images in the clusters while six images on the right are randomly chosen images in that cluster; (b) StudentLife discovered network with three node groups: (1) discovered student clusters, (3) student nodes, (5) discovered activity location (from WiFi data); and two edge groups: (2) Student to cluster assignment, (4) Student involved to activity location. Node sizes (of discovered nodes) depict the number of element in clusters while edge sizes between *Student* and *activity location* represent the popularity of student's activities.

tively mutual information and entropy. AMI is an adjustment of the mutual information score which is generally higher for two clusterings with a larger number of clusters $AMI(U,V) = {}^{MI(U,V)-E\{MI(U,V)\}}/_{\max\{H(U),H(V)\}-E\{MI(U,V)\}}$ where $E\{MI(.,.)\}$ is the expect ion of mutual information. ARI is an adjusted version of Rand index which computes similarity two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters (Hubert and Arabie, 1985).

Quantitative results: To quantitatively evaluate our proposed methods, we compare our algorithms with several baseline methods: K-means, 3-stage K-means (TSK-means) as described in Appendix C, MC2-SVI without context (Huynh et al., 2016). Clustering performance in Table 1 is evaluated with the image clustering problem on *LabelMe dataset*. With K-means, we average all data points to obtain a single vector for each image. K-means needs much less time to run since the number of data points is now reduced to $1,800$. For MC2-SVI, we used stochastic variational inference and parallelized Spark-based implementation in (Huynh et al., 2016) to carry out the experiments. This implementation has the advantage of making use of all of 16 cores on the test machine. In terms of clustering accuracy, MWMS and MWMS with context algorithms perform the best.

Figure 5a demonstrates five representative image clusters with six randomly chosen images in each (on the right) which are discovered by our MWMS algorithm. We also accumulate labeled tags from all images in each cluster to produce the tag cloud on the left,

Noise percentage	Metrics	Methods	
		MWGM	**MWM**
0%	NMI	0.536	0.473
	ARI	0.501	0.427
	AMI	0.530	0.465
0.5%	NMI	0.493	0.412
	ARI	0.440	0.340
	AMI	0.484	0.413
1%	NMI	0.553	0.501
	ARI	0.506	0.461
	AMI	0.533	0.486
5%	NMI	0.542	0.470
	ARI	0.508	0.409
	AMI	0.525	0.454

Table 2: Clustering performance for LabelMe dataset with noisy data.

which can be considered as the visual ground truth of clusters. Our algorithm can group images into clusters that are consistent with the tag cloud.

Qualitative results: We use the StudentLife dataset to demonstrate the capability of multilevel clustering with large-scale datasets. This dataset not only contains a large number of data points but presents in high dimension. Our algorithms need approximately 1 hour to perform multilevel clustering on this dataset. Figure 5b presents two levels of clusters discovered by our algorithms. The innermost (blue) and outermost (green) rings depict local and global clusters respectively. Global clusters represent groups of students while local clusters shared between students ("documents") may be used to infer locations of students' activities. From these clustering results, we can dissect students' shared location (activities), e.g. Student 49 ($U49$) mainly took part in activities at location 4 ($L4$).

Robust multilevel clustering: We now conduct experiments to demonstrate how *multilevel Wasserstein geometric median (MWGM)* algorithm can be robust with some proportions of "noise" data. We created their versions of "noise" LabelMe which we add into the original dataset three varying proportions of Gaussian noise including 0.5%, 1%, and 5% respectively. We now apply two algorithms, MWM and MWGM, with the contaminated LabelMe dataset. Table 2 shows the clustering performance of these algorithms. The clustering performance of the MWGM algorithm is more robust to the level of noise data added in compared with its second order counterpart, the MWM. It is also interesting that the clustering performance of MWGM is increasing then decreasing when more noise presenting in the data.

7.4 Wall-clock running time analysis

In this section, we illustrate the running time of sequential and parallel implementations of the proposed algorithms on both synthetic and real-world datasets. For synthetic data, we use datasets with different numbers of data groups (i.e. 1000, 2000, 4000, 8000, 16,000, and 20,000). With each dataset, we run four algorithms with/without local constraint and

with/without context observations on sequential and parallelized implementations. All experiments are conducted on the same machine (Windows 10 64-bit, core i7 3.4GHz CPU and 16GB RAM). We then observe the average running time of each iteration of serial and parallelized implementations of MWGM(S) (first order Wasserstein metric) and MWM(S) (second order Wasserstein metric) in Figures 6 and 7 respectively. The parallelized implementations have significantly reduced the wall-clock running time of the proposed algorithms especially when the number of groups is large. Since our experiments for parallelized implementations are conducted on a station machine with multiple processors, it is obvious the running time complexity will reduce more dramatically on cluster systems when the datasets contain an extremely large number of groups, e.g. millions. We now present the running

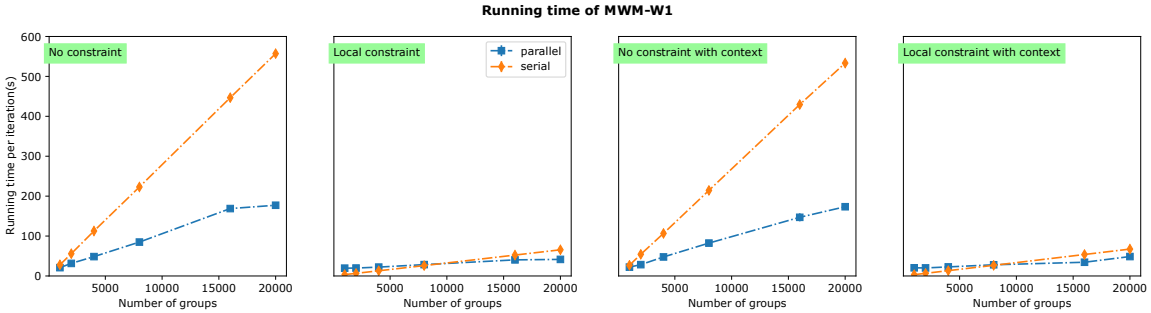


Figure 6: Comparison of running time between serial and parallelized implementations of Multilevel Wasserstein Geometric Median (MWGM) with 4 different settings: *no constraint, local constraint, no constraint with context, local constraint with context.*

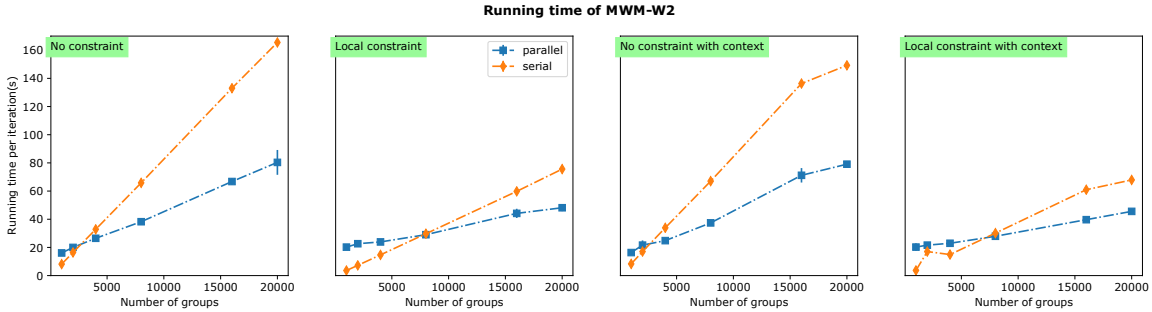


Figure 7: Comparison of running time between serial and parallelized implementations of Multilevel Wasserstein Mean (MWM) with 4 different settings: *no constraint, local constraint, no constraint with context, local constraint with context.*

time of the proposed algorithms on the real-world dataset LabelMe. Figure 8 depicts the running time for this dataset which shows the significant reduction in running time of the parallelized implementation compared with the serial version. Although there are less than 3000 groups of data in this dataset, the running time with the parallelized version of our proposed algorithms has reduced approximately twice. It is also noted that the MWGM algorithm takes more time to run since its inner loop approximates the geometric median instead of the closed-form in the MWM algorithm.

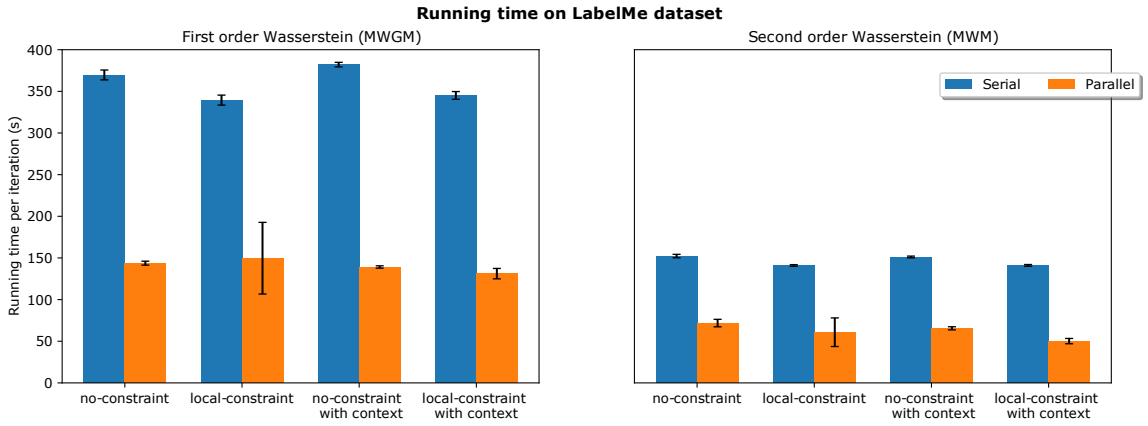


Figure 8: Comparison of running time between serial and parallelized implementations of MWM and MWGM with 4 different settings: *no constraint, local constraint, no constraint with context, local constraint with context* on LabelMe dataset.

8. Conclusion

We have proposed an optimization-based approach to multilevel clustering using Wasserstein metrics. There are several possible directions for extensions. Firstly, we have only considered continuous data. Hence, it is of interest to extend our formulation to discrete data. Secondly, our method requires knowledge of the numbers of clusters both in local and global clustering. When these numbers are unknown, it seems reasonable to incorporate a penalty on the model complexity.

Acknowledgments

This research is supported in part by grants NSF CAREER DMS-1351362, NSF CNS-1409303, the Margaret and Herman Sokol Faculty Award and research gift from Adobe Research (XN). DP, VH and ND gratefully acknowledge the support from the Australian Research Council (ARC) DP150100031 and DP160109394.

Appendix A. Wasserstein barycenter

In this appendix, we collect relevant information on the Wasserstein metric and Wasserstein barycenter problem, which were introduced in Section 2 in the paper. For any Borel map $g : \Theta \to \Theta$ and probability measure G on Θ, the push-forward measure of G through g, denoted by $g\#G$, is defined by the condition that $\int_\Theta f(y)\mathrm{d}(g\#G)(y) = \int_\Theta f(g(x))\mathrm{d}G(x)$ for every continuous bounded function f on Θ.

Wasserstein metric: We now provide further discussion about the formulation of Wasserstein metric when the probability measures are discrete. In particular, when $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$

and $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$ are discrete measures with finite support, i.e. k and k' are finite, the Wasserstein distance of order r between G and G' can be represented as

$$W_r^r(G, G') = \min_{T \in \Pi(G,G')} \langle T, M_{G,G'} \rangle, \tag{19}$$

where we have

$$\Pi(G, G') = \left\{ T \in \mathbb{R}_+^{k \times k'} : T\mathbb{1}_{k'} = \boldsymbol{p},\ T\mathbb{1}_k = \boldsymbol{p}' \right\}$$

such that $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ and $\boldsymbol{p}' = (p'_1, \ldots, p'_{k'})^T$, $M_{G,G'} = \left\{ \|\theta_i - \theta'_j\|^r \right\}_{i,j} \in \mathbb{R}_+^{k \times k'}$ is the cost matrix, i.e. matrix of pairwise distances of elements between G and G', and $\langle A, B \rangle = \mathrm{tr}(A^T B)$ is the Frobenius dot-product of matrices. The optimal $T \in \Pi(G, G')$ in optimization problem (19) is called the optimal coupling of G and G', representing the optimal transport between these two measures.

Wasserstein barycenter: As introduced in Section 2 in the paper, for any probability measures $P_1, P_2, \ldots, P_N \in \mathcal{P}_2(\Theta)$, their Wasserstein barycenter $\overline{P}_{N,\lambda}$ is such that

$$\overline{P}_{N,\lambda} = \arg\min_{P \in \mathcal{P}_2(\Theta)} \sum_{i=1}^{N} \lambda_i W_2^2(P, P_i),$$

where $\lambda \in \Delta_N$ denote weights associated with $P_1, \ldots, P_N$. According to (Agueh and Carlier, 2011), $P_{N,\lambda}$ can be obtained as a solution to so-called multi-marginal optimal transportation problem. In fact, if we denote T_k^1 as the measure preserving map from P_1 to P_k, i.e. $P_k = T_k^1 \# P_1$, for any $1 \leq k \leq N$ (note that, these maps exist as long as we assume for example that the probability measures $P_1, \ldots, P_N$ have density functions (Santambrogio, 2015)), then

$$\overline{P}_{N,\lambda} = \left(\sum_{k=1}^{N} \lambda_k T_k^1 \right) \# P_1.$$

Unfortunately, the forms of the maps T_k^1 are analytically intractable, especially if no special constraints on $P_1, \ldots, P_N$ are imposed.

Recently, Anderes et al. (2015) studied the Wasserstein barycenters $\overline{P}_{N,\lambda}$ when the probability measures $P_1, P_2, \ldots, P_N$ are finite discrete and $\lambda = \left(1/N, \ldots, 1/N\right)$. They demonstrate the following sharp result (cf. Theorem 2 in (Anderes et al., 2015)) regarding the number of atoms of $\overline{P}_{N,\lambda}$

Theorem A.1. *There exists a Wasserstein barycenter $\overline{P}_{N,\lambda}$ such that $|supp(\overline{P}_{N,\lambda})| \leq \sum_{i=1}^{N} s_i - N + 1$.*

Therefore, when $P_1, \ldots, P_N$ are indeed finite discrete measures and the weights are uniform, the problem of finding Wasserstein barycenter $\overline{P}_{N,\lambda}$ over the (computationally large) space $\mathcal{P}_2(\Theta)$ is reduced to a search over a smaller space $\mathcal{O}_l(\Theta)$ where $l = \sum_{i=1}^{N} s_i - N + 1$.

Appendix B. Proofs of theorems

In this appendix, we provide proofs for the remaining results in the paper. We start by giving a proof for the transition from multilevel Wasserstein means objective function (5) to objective function (6) in Section 3.1 in the paper. All the notations in this appendix are similar to those in the main text. For each closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$, we denote the Voronoi region generated by $\mathcal{S}$ on the space $\mathcal{P}_2(\Theta)$ by the collection of subsets $\{V_P\}_{P\in\mathcal{S}}$, where $V_P := \{Q \in \mathcal{P}_2(\Theta) : W_2^2(Q,P) = \min_{G\in\mathcal{S}} W_2^2(Q,G)\}$. We define the projection mapping $\pi_{\mathcal{S}}$ as: $\pi_{\mathcal{S}} : \mathcal{P}_2(\Theta) \to \mathcal{S}$ where $\pi_{\mathcal{S}}(Q) = P$ as $Q \in V_P$. Note that, for any $P_1, P_2 \in \mathcal{S}$ such that V_{P_1} and V_{P_2} share the boundary, the values of π_S at the elements in that boundary can be chosen to be either P_1 or P_2. Now, we start with the following useful lemmas.

Lemma B.1. *For any closed subset $\mathcal{S}$ on $\mathcal{P}_2(\Theta)$, if $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$, then $E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) = W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$ where $d^2_{W_2}(X,\mathcal{S}) = \inf_{P\in\mathcal{S}} W_2^2(X,P)$.*

Proof. For any element $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$:

$$\begin{aligned}\int W_2^2(P,G)\mathrm{d}\pi(P,G) &\geq \int d^2_{W_2}(P,\mathcal{S})\mathrm{d}\pi(P,G)\\ &= \int d^2_{W_2}(P,\mathcal{S})\mathrm{d}\mathcal{Q}(P) = E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})),\end{aligned}$$

where the integrations in the first two terms range over $\mathcal{P}_2(\Theta)\times\mathcal{S}$ while that in the final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we obtain

$$W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q}) = \inf \int_{\mathcal{P}_2(\Theta)\times\mathcal{S}} W_2^2(P,G)\mathrm{d}\pi(P,G) \geq E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})), \tag{20}$$

where the infimum in the first equality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$.

On the other hand, let $g : \mathcal{P}_2(\Theta) \to \mathcal{P}_2(\Theta)\times\mathcal{S}$ such that $g(P) = (P, \pi_{\mathcal{S}}(P))$ for all $P \in \mathcal{P}_2(\Theta)$. Additionally, let $\mu_{\pi_{\mathcal{S}}} = g\#\mathcal{Q}$, the push-forward measure of $\mathcal{Q}$ under mapping g. It is clear that $\mu_{\pi_{\mathcal{S}}}$ is a coupling between $\mathcal{Q}$ and $\pi_{\mathcal{S}}\#\mathcal{Q}$. Under this construction, we obtain for any $X \sim \mathcal{Q}$ that

$$\begin{aligned}E\left(W_2^2(X, \pi_{\mathcal{S}}(X))\right) &= \int W_2^2(P,G)\mathrm{d}\mu_{\pi_{\mathcal{S}}}(P,G)\\ &\geq \inf \int W_2^2(P,G)\mathrm{d}\pi(P,G) = W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q}),\end{aligned} \tag{21}$$

where the infimum in the second inequality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$ and the integrations range over $\mathcal{P}_2(\Theta)\times\mathcal{S}$. Now, from the definition of $\pi_{\mathcal{S}}$

$$\begin{aligned}E(W_2^2(X, \pi_{\mathcal{S}}(X))) &= \int W_2^2(P, \pi_{\mathcal{S}}(P))\mathrm{d}\mathcal{Q}(P)\\ &= \int d^2_{W_2}(P,\mathcal{S})\mathrm{d}\mathcal{Q}(P) = E(d^2_{W_2}(X,\mathcal{S})),\end{aligned} \tag{22}$$

where the integrations in the above equations range over $\mathcal{P}_2(\Theta)$. By combining (21) and (22), we would obtain that

$$E_{X\sim\mathcal{Q}}(d^2_{W_2}(X,\mathcal{S})) \geq W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q}). \tag{23}$$

From (20) and (23), it is straightforward that $E_{X\sim Q}(d(X,S)^2) = W_2^2(Q, \pi_S\#Q)$. Therefore, we achieve the conclusion of the lemma. □

Lemma B.2. *For any closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$ and $\mu \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$ with supp$(\mu) \subseteq \mathcal{S}$, there holds $W_2^2(\mathcal{Q}, \mu) \geq W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q})$ for any $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$.*

Proof. Since $\text{supp}(\mu) \subseteq \mathcal{S}$, it is clear that $W_2^2(\mathcal{Q}, \mu) = \inf\limits_{\pi\in\Pi(\mathcal{Q},\mu)} \int\limits_{\mathcal{P}_2(\Theta)\times\mathcal{S}} W_2^2(P,G)\mathrm{d}\pi(P,G)$.

Additionally, we have

$$\begin{aligned}\int W_2^2(P,G)\mathrm{d}\pi(P,G) &\geq \int d^2_{W_2}(P,\mathcal{S})\mathrm{d}\pi(P,G) = \int d^2_{W_2}(P,\mathcal{S})\mathrm{d}\mathcal{Q}(P)\\ &= E_{X\sim Q}(d^2_{W_2}(X,S)) = W_2^2(\mathcal{Q}, \pi_{\mathcal{S}}\#\mathcal{Q}),\end{aligned}$$

where the last inequality is due to Lemma B.1 and the integrations in the first two terms range over $\mathcal{P}_2(\Theta)\times\mathcal{S}$ while that in the final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we achieve the conclusion of the lemma. □

Equipped with Lemma B.1 and Lemma B.2, we are ready to establish the equivalence between multilevel Wasserstein means objective function (6) and objective function (5) in Section 3.1 in the main text.

Lemma B.3. *For any given positive integers m and M, we have*

$$A := \inf_{\mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))} W_2^2(\mathcal{H}, \frac{1}{m}\sum_{j=1}^m \delta_{G_j}) = \frac{1}{m}\inf_{\boldsymbol{H}=(H_1,\ldots,H_M)} \sum_{j=1}^m d^2_{W_2}(G_j, \boldsymbol{H}) := B.$$

Proof. Write $\mathcal{Q} = \dfrac{1}{m}\sum\limits_{j=1}^m \delta_{G_j}$. From the definition of B, for any $\epsilon > 0$, we can find $\overline{\boldsymbol{H}}$ such that

$$B \geq \frac{1}{m}\sum_{j=1}^m d^2_{W_2}(G_j, \overline{\boldsymbol{H}}) - \epsilon = E_{X\sim\mathcal{Q}}(d^2_{W_2}(X, \overline{\boldsymbol{H}})) - \epsilon = W_2^2(\mathcal{Q}, \pi_{\overline{\boldsymbol{H}}}\#\mathcal{Q}) - \epsilon \geq A - \epsilon,$$

where the second equality in the above display is due to Lemma B.1 while the last inequality is from the fact that $\pi_{\overline{\boldsymbol{H}}}\#\mathcal{Q}$ is a discrete probability measure in $\mathcal{P}_2(\mathcal{P}_2(\Theta))$ with exactly M support points. Since the inequality in the above display holds for any ϵ, it implies that $B \geq A$. On the other hand, from the formation of A, for any $\epsilon > 0$, we also can find $\mathcal{H}' \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$ such that

$$A \geq W_2^2(\mathcal{H}', \mathcal{Q}) - \epsilon \geq W_2^2(\mathcal{Q}, \pi_{\boldsymbol{H}'}\#\mathcal{Q}) - \epsilon = \frac{1}{m}\sum_{j=1}^m d^2_{W_2}(G_j, \boldsymbol{H}') - \epsilon \geq B - \epsilon,$$

where $\boldsymbol{H}' = \text{supp}(\mathcal{H}')$, the second inequality is due to Lemma B.2, and the third equality is due to Lemma B.1. Therefore, it means that $A \geq B$. We achieve the conclusion of the lemma. □

Proposition B.4. *For any positive integer numbers m, M and k_j as $1 \leq j \leq m$, we denote*

$$C := \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta)\ \forall 1\leq j\leq m,\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{i=1}^{m} W_2^2(G_j, P_{n_j}^j) + \lambda W_2^2(\mathcal{H}, \frac{1}{m}\sum_{i=1}^{m}\delta_{G_i})$$

$$D := \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta)\ \forall 1\leq j\leq m,\\ \boldsymbol{H}=(H_1,\ldots,H_M)}} \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \frac{\lambda\ d_{W_2}^2(G_j, \boldsymbol{H})}{m}.$$

Then, we have $C = D$.

Proof. The proof of this proposition is a straightforward application of Lemma B.3. Indeed, for each fixed $(G_1, \ldots, G_m)$ the infimum w.r.t to $\mathcal{H}$ in C leads to the same infimum w.r.t to $\boldsymbol{H}$ in D, according to Lemma B.3. Now, by taking the infimum w.r.t to $(G_1, \ldots, G_m)$ on both sides, we achieve the conclusion of the proposition. □

In the remainder of this Appendix, we present the proofs for all remaining theorems stated in the main text.

PROOF OF THEOREM 3.1 Recall that, we use the notation $\hat{W}_2$ to denote the entropic regularized second-order Wasserstein with some given regularized parameter (see equation (1) for the details). Now, for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$, we denote the function

$$f(\boldsymbol{G}, \boldsymbol{H}) = \sum_{j=1}^{m} \hat{W}_2^2(G_j, P_n^j) + \frac{\lambda\ d_{\hat{W}_2}^2(G_j, \boldsymbol{H})}{m},$$

where $\boldsymbol{G} = (G_1, \ldots, G_m)$. To obtain the conclusion of this theorem, it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$$

unless $(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$. It is clear that $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) = f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ when $(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$. Therefore, we will only consider the setting when $(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \neq (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$. There are two cases: $\boldsymbol{G}^{(t+1)} \neq \boldsymbol{G}^{(t)}$ or $\boldsymbol{H}^{(t+1)} \neq \boldsymbol{H}^{(t)}$.

Case 1: We first consider the setting when $\boldsymbol{G}^{(t+1)} \neq \boldsymbol{G}^{(t)}$. It means that there exists $j \in \{1, 2, \ldots, m\}$ such that $G_j^{(t+1)} \neq G_j^{(t)}$. Without loss of generality, we assume that $j = 1$. We show that $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ when $\boldsymbol{G}^{(t+1)} \neq \boldsymbol{G}^{(t)}$. From the updates of $\boldsymbol{G}^{(t+1)}$ in Algorithm 1, for any $1 \leq j \leq m$ we have

$$G_j^{(t+1)} \in \operatorname*{arg\,min}_{G_j \in \mathcal{O}_{k_j}(\Theta)} \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda \hat{W}_2^2(G_j, H_{i_j}^{(t)})/m$$

for all $1 \leq j \leq m$. It demonstrates that

$$\hat{W}_2^2(G_j^{(t+1)}, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j^{(t+1)}, H_{i_j}^{(t)}) \leq \min_{G_j \in \mathcal{O}_{k_j}(\Theta):\mathrm{supp}(G_j)\in\mathrm{supp}(G_j^{(t)})} \hat{W}_2^2(G_j, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j, H_{i_j}^{(t)}).$$

The RHS of the above inequality means that we only find the optimal measure $\bar{G}_j \in \mathcal{O}_{k_j}(\Theta)$ such that its supports lie in the set of supports of $G_j^{(t)}$. Therefore, finding the optimal measure $\bar{G}_j$ of that objective function is equivalent to find its optimal masses, which is a strongly convex problem. It proves that for any $1 \leq j \leq m$

$$\hat{W}_2^2(G_j^{(t+1)}, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j^{(t+1)}, H_{i_j}^{(t)}) \leq \hat{W}_2^2(G_j^{(t)}, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j^{(t)}, H_{i_j}^{(t)})$$

and the equality only holds when $G_j^{(t+1)} \equiv G_j^{(t)}$. Since $G_1^{(t+1)} \neq G_1^{(t)}$, we have

$$\hat{W}_2^2(G_j^{(t+1)}, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j^{(t+1)}, H_{i_j}^{(t)}) < \hat{W}_2^2(G_j^{(t)}, P_{n_j}^j) + \frac{\lambda}{m}\hat{W}_2^2(G_j^{(t)}, H_{i_j}^{(t)}).$$

Putting the above results together, we obtain $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ when $\boldsymbol{G}^{(t+1)} \neq \boldsymbol{G}^{(t)}$.

Case 2: We now consider the setting when $\boldsymbol{H}^{(t+1)} \neq \boldsymbol{H}^{(t)}$. Without loss of generality, we assume that $H_1^{(t+1)} \neq H_1^{(t)}$. We now prove that $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)})$. Indeed, recall that $C_i = \{l : i_l = i\}$ for $1 \leq i \leq M$ where $i_j = \arg\min_{1 \leq u \leq M} \hat{W}_2^2(G_j^{(t+1)}, H_u^{(t)})$ and

$$H_i^{(t+1)} \in \arg\min_{H_i \in \mathcal{P}_2(\Theta)} \sum_{l \in C_i} \hat{W}_2^2(H_i, G_l^{(t+1)}).$$

We denote by $\mathcal{A}_i$ the set of all supports of $H_i^{(t)}$ for $l \in C_i$ for any $1 \leq i \leq M$. From the formulation of $H_i^{(t+1)}$, we have

$$\sum_{l \in C_i} \hat{W}_2^2(H_i^{(t+1)}, G_l^{(t+1)}) \leq \min_{H_i:\ \mathrm{supp}(H_i) \equiv \mathcal{A}_i} \sum_{l \in C_i} \hat{W}_2^2(H_i, G_l^{(t+1)}).$$

The RHS objective function means that we only search for the optimal measure $\bar{H}_i$ that has supports in the set $\mathcal{A}_i$. It is equivalent to search for the optimal masses of $\bar{H}_i$ as the supports of $\bar{H}_i$ are fixed. The RHS objective function is strongly convex with respect to the masses of $\bar{H}_i$. It demonstrates that for any $i \in \{1, 2, \ldots, M\}$

$$\sum_{l \in C_i} \hat{W}_2^2(H_i^{(t+1)}, G_l^{(t+1)}) \leq \sum_{l \in C_i} \hat{W}_2^2(H_i^{(t)}, G_l^{(t+1)}),$$

and the equality only holds when $H_i^{(t+1)} \equiv H_i^{(t)}$. Since $H_1^{(t+1)} \neq H_1^{(t)}$, we have

$$\sum_{l \in C_i} \hat{W}_2^2(H_1^{(t+1)}, G_l^{(t+1)}) < \sum_{l \in C_i} \hat{W}_2^2(H_1^{(t)}, G_l^{(t+1)}).$$

Collecting the above results leads to $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)})$ when $\boldsymbol{H}^{(t+1)} \neq \boldsymbol{H}^{(t)}$.

In summary, the results from Cases 1 and 2 prove that $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ unless $(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ for any $t \geq 0$. Furthermore, the argument of these cases suggests that if there exists $\bar{t}$ such that $f(\boldsymbol{G}^{(\bar{t}+1)}, \boldsymbol{H}^{(\bar{t}+1)}) = f(\boldsymbol{G}^{(\bar{t})}, \boldsymbol{H}^{(\bar{t})})$, then

$(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ for all $t \geq \bar{t}$. It suggests that there exists $(\bar{\boldsymbol{G}}, \bar{\boldsymbol{H}})$ such that $f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ converges to $f(\bar{\boldsymbol{G}}, \bar{\boldsymbol{H}})$ and $(\bar{\boldsymbol{G}}, \bar{\boldsymbol{H}})$ satisfies that

$$\bar{G}_j \in \underset{G_j \in \mathcal{O}_{k_j}(\Theta)}{\arg\min}\ \hat{W}_2^2(G_j, P_{n_j}^j) + \lambda \hat{W}_2^2(G_j, \bar{H}_{\bar{i}_j})/m,$$

$$\bar{H}_i \in \underset{H_i \in \mathcal{P}_2(\Theta)}{\arg\min} \sum_{l \in \bar{C}_i} \hat{W}_2^2(H_i, \bar{G}_l),$$

where $\bar{i}_j = \underset{1 \leq u \leq M}{\arg\min}\, \hat{W}_2^2(\bar{G}_j, \bar{H}_u)$ for any $1 \leq j \leq m$ and $\bar{C}_i = \{l : \bar{i}_l = i\}$ for $1 \leq i \leq M$. As a consequence, we obtain the conclusion of the theorem.

PROOF OF THEOREM 3.2 The proof is quite similar to the proof of Theorem 3.1. In fact, recall from the proof of Theorem 3.1 that for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\boldsymbol{H} = (H_1, \ldots, H_M)$ we denote the function

$$f(\boldsymbol{G}, \boldsymbol{H}) = \sum_{j=1}^{m} \hat{W}_2^2(G_j, P_n^j) + \frac{\lambda}{m} d_{\hat{W}_2}^2(G_j, \boldsymbol{H}),$$

where $\boldsymbol{G} = (G_1, \ldots, G_m)$. Now it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}),$$

unless $(S_K^{(t+1)}, \boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (S_K^{(t)}, \boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$. Here, we only give the proof for that inequality under the setting when $S_K^{(t+1)} \neq S_K^{(t)}$ as the the proof argument for that inequality when $S_K^{(t+1)} \equiv S_K^{(t)}$ and $(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)})) \neq (\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ is similar to the proof argument of Theorem 3.1. Indeed, by the definition of Wasserstein distances, we have

$$E = m \cdot f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}) = \sum_{u=1}^{m} \sum_{j,v} m T_{j,v}^u \|a_j^{(t)} - X_{u,v}\|^2 + \lambda U_{j,v}^u \|a_j^{(t)} - h_{i_u,v}^{(t)}\|^2.$$

Therefore, the update of $S_K^{(t+1)}$ from Algorithm 2 and the assumption that $S_K^{(t+1)} \neq S_K^{(t)}$ leads to

$$\begin{aligned}
E &> \sum_{u=1}^{m} \sum_{j,v} m T_{j,v}^u \|a_j^{(t+1)} - X_{u,v}\|^2 + \lambda U_{j,v}^u \|a_j^{(t+1)} - h_{i_u,v}^{(t)}\|^2 \\
&\geq m \sum_{j=1}^{m} \hat{W}_2^2(G_j^{(t)'}, P_n^j) + \lambda \sum_{j=1}^{m} \hat{W}_2^2(G_j^{(t)'}, H_{i_j}^{(t)}) \\
&\geq m \sum_{j=1}^{m} \hat{W}_2^2(G_j^{(t)'}, P_n^j) + \lambda \sum_{j=1}^{m} d_{\hat{W}_2}^2(G_j^{(t)'}, \boldsymbol{H}^{(t)}) \\
&= m f(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}),
\end{aligned}$$

where $\boldsymbol{G'}^{(t)} = (G_1^{(t)'}, \ldots, G_m^{(t)'})$, $G_j^{(t)'}$ are formed by replacing the atoms of $G_j^{(t)}$ by the elements of $S_K^{(t+1)}$, noting that $\mathrm{supp}(G_j^{(t)'}) \subseteq \mathcal{S}_K^{(t+1)}$ as $1 \leq j \leq m$, and the second inequality comes directly from the definition of Wasserstein distance. Hence, we obtain

$$f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}) > f(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}). \tag{24}$$

From the definition of $G_j^{(t+1)}$ as $1 \leq j \leq m$, we get

$$\sum_{j=1}^{m} d_{W_2}^2(G_j^{(t+1)}, \boldsymbol{H}^{(t)}) \leq \sum_{j=1}^{m} d_{W_2}^2(G_j^{(t)'}, \boldsymbol{H}^{(t)}).$$

Thus, it leads to

$$f(\boldsymbol{G'}^{(t)}, \boldsymbol{H}^{(t)}) \geq f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}). \tag{25}$$

Finally, from the definition of $H_1^{(t+1)}, \ldots, H_M^{(t+1)}$, we have

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t)}) \geq f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}). \tag{26}$$

By combining (24), (25), and (26), we arrive at the inequality $f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$ when $S_K^{(t+1)} \neq S_K^{(t)}$.

In summary, we have

$$f(\boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) < f(\boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)}),$$

unless $(S_K^{(t+1)}, \boldsymbol{G}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \equiv (S_K^{(t)}, \boldsymbol{G}^{(t)}, \boldsymbol{H}^{(t)})$. From here, using the similar argument as that of the proof of Theorem 3.1, we reach the conclusion of the theorem.

PROOF OF THEOREM 6.1 To simplify notation, writing

$$L_{\boldsymbol{n}} = \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) \text{ and } L_0 = \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f(\boldsymbol{G}, \mathcal{H}).$$

(i) For any $\epsilon > 0$, from the definition of L_0, we can find $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ such that $f(\boldsymbol{G}, \mathcal{H})^{1/2} \leq L_0^{1/2} + \epsilon$. Therefore, we would have

$$\begin{aligned} L_{\boldsymbol{n}}^{1/2} - L_0^{1/2} &\leq L_n^{1/2} - f(\boldsymbol{G}, \mathcal{H})^{1/2} + \epsilon \leq f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H})^{1/2} - f(\boldsymbol{G}, \mathcal{H})^{1/2} + \epsilon \\ &= \frac{f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) - f(\boldsymbol{G}, \mathcal{H})}{f_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H})^{1/2} + f(\boldsymbol{G}, \mathcal{H})^{1/2}} + \epsilon \leq \sum_{j=1}^{m} \frac{|W_2^2(G_j, P_{n_j}^j) - W_2^2(G_j, P^j)|}{W_2(G_j, P_{n_j}^j) + W_2(G_j, P^j)} + \epsilon \\ &\leq \sum_{j=1}^{m} W_2(P_{n_j}^j, P^j) + \epsilon. \end{aligned}$$

By reversing the direction, we also obtain the inequality $L_n^{1/2} - L_0^{1/2} \geq \sum_{j=1}^{m} W_2(P_{n_j}^j, P^j) - \epsilon$. Hence, $|L_n^{1/2} - L_0^{1/2} - \sum_{j=1}^{m} W_2(P_{n_j}^j, P^j)| \leq \epsilon$ for any $\epsilon > 0$. Since $P^j \in \mathcal{P}_2(\Theta)$ for all $1 \leq j \leq m$, we obtain that $W_2(P_{n_j}^j, P^j) \to 0$ almost surely as $n_j \to \infty$ (see for example Theorem 6.9 in (Villani, 2009)). As a consequence, we obtain the conclusion of part (i).

(ii) Based on the results of Fournier and Guillin (2015), as the probability measures P^j are compactly supported in $\mathbb{R}^d$ for $1 \leq j \leq m$, we have $W_2(P_{n_j}^j, P^j) = O_P(n_j^{-1/d})$ when

$d \geq 5$ and $W_2(P^j_{n_j}, P^j) = O_P(n_j^{-1/4})$ when $d \leq 4$ for any $1 \leq j \leq m$. Combining these results with the results of part (i), we obtain that

$$|L_{\boldsymbol{n}}^{1/2} - L_0^{1/2}| = \begin{cases} O_P(m \cdot n_\vee^{-1/d}) \text{ when } d \geq 5, \\ O_P(m \cdot n_\vee^{-1/4}) \text{ when } d \leq 4. \end{cases}$$

where $n_\vee = \max_{1 \leq i \leq m} n_i$. As a consequence, we reach the conclusion of part (ii).

PROOF OF THEOREM 6.2 For any $\epsilon > 0$, we denote

$$\mathcal{A}(\epsilon) = \Big\{ G_i \in \mathcal{O}_{k_i}(\Theta), \mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta)) : d(\boldsymbol{G}, \mathcal{H}, \mathcal{F}) \geq \epsilon \Big\}.$$

Since Θ is a compact set, we also have $\mathcal{O}_{k_j}(\Theta)$ and $\mathcal{E}_M(\mathcal{P}_2(\Theta))$ are compact for any $1 \leq i \leq m$. As a consequence, $\mathcal{A}(\epsilon)$ is also a compact set. For any $(\boldsymbol{G}, \mathcal{H}) \in \mathcal{A}(\epsilon)$, by the definition of $\mathcal{F}$ we would have $f(\boldsymbol{G}, \mathcal{H}) > f(\boldsymbol{G}^0, \mathcal{H}^0)$ for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Since $\mathcal{A}(\epsilon)$ is compact, it leads to

$$\inf_{(\boldsymbol{G}, \mathcal{H}) \in A(\epsilon)} f(\boldsymbol{G}, \mathcal{H}) > f(\boldsymbol{G}^0, \mathcal{H}^0),$$

for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. From the formulation of $f_{\boldsymbol{n}}$ as in the proof of Theorem 6.1, we can verify that $\lim_{\boldsymbol{n} \to \infty} f_{\boldsymbol{n}}(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}) = \lim_{\boldsymbol{n} \to \infty} f(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}})$ almost surely as $\boldsymbol{n} \to \infty$. Combining this result with that of Theorem 6.1, we obtain $f(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}) \to f(\boldsymbol{G}^0, \mathcal{H}^0)$ as $\boldsymbol{n} \to \infty$ for any $(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Therefore, for any $\epsilon > 0$, as $\boldsymbol{n}$ is large enough, we have $d(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}, \mathcal{F}) < \epsilon$. As a consequence, we achieve the conclusion regarding the consistency of the mixing measures.

Appendix C. Data generation processes in the simulation studies

In this appendix, we offer details on the data generation processes utilized in the simulation studies presented in Section 7 in the main text. The notions of m, n, d, M are given in the main text. Let K_i be the number of supporting atoms of H_i and k_j the number of atoms of G_j. For any $d \geq 1$, we denote $\mathbf{1}_d$ to be d dimensional vector with all components to be 1. Furthermore, $\mathcal{I}_d$ is an identity matrix with d dimensions.

Comparison metric (Wasserstein distance to truth)

$$\mathrm{W} := \frac{1}{m} \sum_{j=1}^{m} W_2(\hat{G}_j, G_j) + d_{\mathcal{M}}(\hat{\boldsymbol{H}}, \boldsymbol{H}),$$

where $\hat{\boldsymbol{H}} := \{\hat{H}_1, \ldots, \hat{H}_M\}$, $\boldsymbol{H} := \{H_1, \ldots, H_M\}$ and $d_{\mathcal{M}}(\hat{H}, H)$ is a minimum-matching distance (Tang et al., 2014; Nguyen, 2015):

$$d_{\mathcal{M}}(\hat{\boldsymbol{H}}, \boldsymbol{H}) := \max\{\overline{d}(\hat{\boldsymbol{H}}, \boldsymbol{H}), \overline{d}(\boldsymbol{H}, \hat{\boldsymbol{H}})\},$$

where

$$\overline{d}(\hat{\boldsymbol{H}}, \boldsymbol{H}) := \max_{1 \leq i \leq M} \min_{1 \leq j \leq M} W_2(H_i, \hat{H}_j).$$

Multilevel Wasserstein means setting The global clusters are generated as follows:

- Means for atoms $\mu_i := 5(i-1)$ for all $i = 1, \ldots, M$.
- Atoms of H_i: $\phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d)$ for all $j = 1, \ldots, K_i$.
- Weights of atoms: $\pi_i \sim \text{Dirichlet}(\mathbf{1}_{K_i})$.
- Let $H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}$.

For each group $j = 1, \ldots, m$, generate local measures and data as follows:

- Pick cluster label $z_j \sim \text{Uniform}(\{1, \ldots, M\})$.
- Mean for atoms: $\tau_{ji} \sim H_{z_j}$ for all $i = 1, \ldots, k_j$.
- Atoms of G_j: $\theta_{ji} \sim \mathcal{N}(\tau_{ji}, \mathcal{I}_d)$ for all $i = 1, \ldots, k_j$.
- Weights of atoms $p_j \sim \text{Dirichlet}(\mathbf{1}_{k_j})$.
- Let $G_j := \sum_{i=1}^{k_j} p_{ji} \delta_{\theta_{ji}}$.
- Data means $\mu_i \sim G_j$ for all $i = 1, \ldots, n_j$.
- Observations $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_{ji} of G_j is set to be proportional to global cluster label z_j assigned to G_j.

Multilevel Wasserstein means with sharing setting The global clusters are generated as follows:

- Means for atoms $\mu_i := 5(i-1)$ for all $i = 1, \ldots, M$.
- Atoms of $H_i : \phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d)$ for all $j = 1, \ldots, K_i$.
- Weights of atoms $\pi_i \sim \text{Dirichlet}(\mathbf{1}_{K_i})$.
- Let $H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}$.

For each shared atom $k = 1, \ldots, K$:

- Pick cluster label $z_k \sim \text{Uniform}(\{1, \ldots, M\})$.
- Mean for atoms: $\tau_k \sim H_{z_k}$.
- Atoms of S_K: $\theta_k \sim \mathcal{N}(\tau_k, \mathcal{I}_d)$.

For each group $j = 1, \ldots, m$ generate local measures and data as follows:

- Pick cluster label $\tilde{z}_j \sim \text{Uniform}(\{1, \ldots, M\})$.
- Select shared atoms $s_j = \{k : z_k = \tilde{z}_j\}$.

- Weights of atoms $p_{s_j} \sim \text{Dirichlet}(\mathbf{1}_{|s_j|})$; $\quad G_j := \sum_{i \in s_j} p_i \delta_{\theta_i}$.
- Data means $\mu_i \sim G_j$ for all $i = 1, \ldots, n_j$.
- Observations $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_i of G_j where $i \in s_j$ is set to be proportional to global cluster label $\tilde{z}_j$ assigned to G_j.

Three-stage K-means First, we estimate G_j for each group $1 \leq j \leq m$ by using K-means algorithm with k_j clusters. Then, we cluster labels using the K-means algorithm with M clusters based on the collection of all atoms of G_j's. Finally, we estimate the atoms of each H_i via K-means algorithm with exactly L clusters for each group of local atoms. Here, L is some given threshold being used in Algorithm 1 in Section 3.1 in the main text to speed up the computation (see final remark regarding Algorithm 1 in Section 3.1). The three-stage K-means algorithm is summarized in Algorithm 5.

Algorithm 5 Three-stage K-means

Input: Data $X_{j,i}$, k_j, M, L.
Output: local measures G_j and global elements H_i of $\boldsymbol{H}$.
Stage 1
for $j = 1$ **to** m **do**
  $G_j \leftarrow k_j$ clusters of group j with K-means (atoms as centroids and weights as label frequencies).
end for
$\mathcal{C} \leftarrow$ collection of all atoms of G_j.
Stage 2
$\{D_1, \ldots, D_M\} \leftarrow M$ clusters from K-means on $\mathcal{C}$.
Stage 3
for $i = 1$ **to** M **do**
  $H_i \leftarrow L$ clusters of D_i with K-means (atoms as centroids and weights as label frequencies).
end for

Appendix D. Computational aspects of Wasserstein barycenter under W_1 metric

In this appendix, we provide a fast and efficient algorithm to compute the Wasserstein barycenter under W_1 metric. In particular, we focus on the setup when $Q_1, \ldots, Q_N \in \mathcal{P}_1(\Theta)$ for $N \geq 1$ are finite discrete measures and we would like to determine the local Wasserstein barycenter of (18) within the space $\mathcal{O}_k(\Theta)$ for some given $k \geq 1$.

Weighted geometric median: Let $X_1, \ldots, X_m \in \mathbb{R}^d$ be m distinct points and $\eta_1, \ldots, \eta_m$ be m positive numbers. The weighted geometric median $X^* \in \mathbb{R}^d$ is the optimal solution

of the following convex optimization problem

$$\min_{X\in\mathbb{R}^d}\sum_{i=1}^{m}\eta_i\|X_i-X\|.$$

To the best of our knowledge, no explicit formula for X^* is available, therefore, we will utilize an iterative procedure to calculate an approximation for X^*. The most common approach for such procedure is Weiszfeld's algorithm (Weiszfeld, 1937); however, this approach has been shown to be unstable when the update is identical to one of the given points X_i for some $1 \le i \le m$. To account for this instability of Weiszfeld's algorithm, Vardi and Zhang (2000) introduces a solution for the setting when the update falls to the set of given points. In particular, their iterative algorithm can be summarized as follows

$$X^{(i+1)} = \left(1-\frac{\eta(X^{(i)})}{r(X^{(i)})}\right)^{+}\widetilde{T}(X^{(i)}) + \min\left\{1,\frac{\eta(X^{(i)})}{r(X^{(i)})}\right\}X^{(i)},$$

where $\eta(x) = \begin{cases}\eta_k \text{ if } x = X_k,\ k=1,\dots,m\\ 0 \text{ otherwise}\end{cases}$, $r(x) = \|\widetilde{R}(x)\|$ with $\widetilde{R}(x) = \sum\limits_{X_i\neq x}\eta_i\dfrac{X_i-x}{\|X_i-x\|}$, and $\widetilde{T}(x) = \left\{\sum\limits_{X_i\neq x}\dfrac{\eta_i}{\|X_i-x\|}\right\}^{-1}\sum\limits_{X_i\neq x}\dfrac{\eta_i X_i}{\|X_i-x\|}$ for all $x\in\mathbb{R}^d$. Here, we take the convention that $0/0 = 0$. For the convenience of argument later, we refer to this algorithm as the VZ algorithm. As being shown in (Vardi and Zhang, 2000), the VZ algorithm converges quickly to the global minimum of weighted geometric median problem. Due to its simplicity and efficiency in terms of computation, we will use the VZ algorithm for the updates of Wasserstein barycenter under W_1 metric.

Wasserstein barycenter under the entropic version of W_1 distance: For the simplicity of the paper, we only focus on determining Wasserstein barycenter under the entropic W_1 over the set of discrete probability measures with at most $k \ge 1$ components, i.e., we develop an efficient algorithm to estimate the optimal solution of the following optimization problem

$$\overline{Q}_{N,\lambda} = \operatorname*{arg\,min}_{Q\in\mathcal{O}_k(\Theta)}\sum_{i=1}^{N}\lambda_i\hat{W}_1(Q,Q_i), \tag{27}$$

where $Q_i \in \mathcal{O}_{k_i}(\Theta)$ for given $k_i \ge 1$ as $1 \le i \le N$ and $\lambda \in \Delta_N$ denotes weights associated with $Q_1,\dots,Q_N$.

Algorithm for Wasserstein barycenter under the entropic version of W_1 distance: The algorithm for determining Wasserstein barycenter of equation (27) will follow those in (Cuturi and Doucet, 2014) with the only modification regarding updating the atoms of $\overline{Q}_{N,\lambda}$ in terms of VZ algorithm for the geometric median. We summarize that algorithm in Algorithm 6.

Algorithm 6 Wasserstein barycenter under the entropic version of W_1 metric

Input: Atoms $Y_i \in \mathbb{R}^{d\times k_i}$ of Q_i, weights b_i of Q_i, and $\lambda \in \Delta_N$.
Output: Atoms $X \in \mathbb{R}^{d\times k}$ and weights a of $\overline{Q}_{N,\lambda}$.
Initialize atoms $X^{(0)}$, weights $a^{(0)}$ of $\overline{Q}^{(0)}_{N,\lambda}$, and $t = 0$.
while $X^{(t)}$ and $a^{(t)}$ have not converged **do**
  Update $a^{(t)}$ using Algorithm 1 in (Cuturi and Doucet, 2014).
  for $i = 1$ **to** N **do**
    $T^i \leftarrow$ optimal coupling of $\overline{Q}^{(t)}_{N,\lambda}$, Q_i.
  end for
  for $i = 1$ **to** k **do**
    $\widetilde{X}^{(t)}_i \leftarrow \underset{X_i\in\mathbb{R}^d}{\arg\min} \sum_{j=1}^{N} \lambda_j \sum_{u=1}^{k_j} T^j_{i,u}\|X_i - Y_{j,u}\|$.
  end for
  $\widetilde{X}^{(t)} \leftarrow [\widetilde{X}^{(t)}_1, \ldots, \widetilde{X}^{(t)}_k]$.
  $X^{(t)} \leftarrow (1-\theta)X^{(t)} + \theta\widetilde{X}^{(t)}$ where $\theta \in [0,1]$ is a line-search or preset value.
  $t \leftarrow t + 1$.
end while

Appendix E. Consistency of estimators from Multilevel Wasserstein Geometric Median (MWGM) method

In this appendix, we provide consistency of the objective function and the estimators of the MWGM method. To simplify the presentation, we also fix m and assume that P^j is the true distribution of data $X_{j,i}$ for $j = 1, \ldots, m$. Similar to Section 6, we denote $\boldsymbol{G} = (G_1, \ldots, G_m)$ and $\boldsymbol{n} = (n_1, \ldots, n_m)$. We say that $\boldsymbol{n} \to \infty$ if $n_j \to \infty$ for $j = 1, \ldots, m$. Define the following functions associated with the MWGM method and its population version:

$$g_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) = \sum_{j=1}^{m} W_1(G_j, P^j_{n_j}) + \lambda W_1(\mathcal{H}, \frac{1}{m}\sum_{j=1}^{m}\delta_{G_j}),$$
$$g(\boldsymbol{G}, \mathcal{H}) = \sum_{j=1}^{m} W_1(G_j, P^j) + \lambda W_1(\mathcal{H}, \frac{1}{m}\sum_{j=1}^{m}\delta_{G_j}),$$

where $G_j \in \mathcal{O}_{k_j}(\Theta)$, $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_1(\Theta))$ for $1 \le j \le m$. We first establish the first consistency property of the MWGM method.

Theorem E.1. *Assume that $P^j \in \mathcal{P}_1(\Theta)$ for $1 \le j \le m$. Then, as $\boldsymbol{n} \to \infty$*

$$\inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_1(\Theta))}} g_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) - \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_1(\Theta))}} g(\boldsymbol{G}, \mathcal{H}) \to 0$$

almost surely.

Proof. The proof of Theorem E.1 follows the same proof argument as that of Theorem 6.1. Here, we provide the proof for the completeness. We denote

$$L_{\boldsymbol{n}} = \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_1(\Theta))}} g_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) \text{ and } L_0 = \inf_{\substack{G_j\in\mathcal{O}_{k_j}(\Theta),\\ \mathcal{H}\in\mathcal{E}_M(\mathcal{P}_1(\Theta))}} g(\boldsymbol{G}, \mathcal{H}).$$

For any $\epsilon > 0$, from the definition of L_0, we can find $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ such that $g(\boldsymbol{G}, \mathcal{H}) \leq L_0 + \epsilon$. An application of triangle inequality leads to

$$\begin{aligned} L_{\boldsymbol{n}} - L_0 &\leq L_n - g(\boldsymbol{G}, \mathcal{H}) + \epsilon \leq g_{\boldsymbol{n}}(\boldsymbol{G}, \mathcal{H}) - g(\boldsymbol{G}, \mathcal{H}) + \epsilon \\ &\leq \sum_{j=1}^{m} W_1(P_{n_j}^j, P^j) + \epsilon. \end{aligned}$$

By reversing the direction, we also obtain the inequality $L_n - L_0 \geq \sum_{j=1}^{m} W_1(P_{n_j}^j, P^j) - \epsilon$. Putting the two inequalities together, we have

$$|L_n - L_0 - \sum_{j=1}^{m} W_1(P_{n_j}^j, P^j)| \leq \epsilon$$

for any $\epsilon > 0$. According to the hypothesis, $P^j \in \mathcal{P}_1(\Theta)$ for all $1 \leq j \leq m$. Therefore, we obtain that $W_1(P_{n_j}^j, P^j) \to 0$ almost surely as $n_j \to \infty$. As a consequence, we obtain the conclusion of the theorem. □

Our next result establishes that the consistency of the estimators from the MWGM method to those in the population version of MWGM. To ease the presentation, we assume that for each $\boldsymbol{n}$ there is an optimal solution $(\widehat{G}_1^{n_1}, \ldots, \widehat{G}_m^{n_m}, \widehat{\mathcal{H}}^{\boldsymbol{n}})$ (or in short $(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \mathcal{H}^{\boldsymbol{n}})$) of the MWGM objective function (15). Furthermore, we can find optimal solution minimizing $f(\boldsymbol{G}, \mathcal{H})$ over $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_1(\Theta))$. We denote $\mathcal{F}$ the collection of these optimal solutions. For any $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_1(\Theta))$, we define

$$\bar{d}(\boldsymbol{G}, \mathcal{H}, \mathcal{F}) = \inf_{(\boldsymbol{G}^0, \mathcal{H}^0) \in \mathcal{F}} \sum_{j=1}^{m} W_1(G_j, G_j^0) + W_1(\mathcal{H}, \mathcal{H}^0).$$

Given the above assumptions and the consistency of the objective function of the MWGM method, we have the following result regarding the convergence of $(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \mathcal{H}^{\boldsymbol{n}})$.

Theorem E.2. *Assume that Θ is bounded and $P^j \in \mathcal{P}_1(\Theta)$ for all $1 \leq j \leq m$. Then, we have $\bar{d}(\widehat{\boldsymbol{G}}^{\boldsymbol{n}}, \widehat{\mathcal{H}}^{\boldsymbol{n}}, \mathcal{F}) \to 0$ as $\boldsymbol{n} \to \infty$ almost surely.*

The proof of Theorem E.2 is a direct combination of the proof argument of Theorem 6.2 and the consistency of objective function of MWGM method in Theorem E.1; therefore, it is omitted.

Appendix F. Algorithms for computing the entropic Wasserstein barycenters

Algorithm 7 Smoothed Primal T^*_γ and Dual b^*_γ Optima

Input: $M = M_{XY}, \gamma, a, b$.
Output: T^*_γ and b^*_γ.
$K = \exp(-\gamma M)$
$\tilde{K} = \mathrm{diag}(b^{-1})K$
Initialize $u = \mathrm{ones}(k, 1)/k$
while u changed **do**
  $u = 1./(\tilde{K}(b./(K^\top u)))$
end while
$v = b./(K^\top u)$.
$b^*_\gamma = \frac{1}{\gamma}\log u + \frac{\log u^\top \mathbf{1}_k}{\gamma k}\mathbf{1}_k$.
$T^*_\gamma = \mathrm{diag}(u)K\mathrm{diag}(v)$.
return T^*_γ, b^*_γ

In this appendix, we include three algorithms for computing entropic Wasserstein barycenter in (Cuturi and Doucet, 2014) to make our manuscript self-contained. Before describing the details of the algorithms, we summarize the notations as follows. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ (resp., $Y_1, \ldots, Y_k \in \mathbb{R}^d$) be n (resp., k) atom points of discrete measure $P = \sum_{i=1}^{n} a_i \delta_{X_i}$ (resp., $Q = \sum_{i=1}^{k} b_i \delta_{Y_i}$) where $a = (a_1, \ldots, a_n) \in \Delta_n$ (resp., $b = (b_1, \ldots, b_k) \in \Delta_k$). The matrix M_{XY} of pairwise distances between elements of X_i and Y_j raised to the power p, i.e. $M_{XY} \in \mathbb{R}^{n\times k}$. The transportation polytope is denoted as $U(a, b) = \{\mathbb{R}^{n\times k}_+ \mid T\mathbf{1}_k = a, T^\top \mathbf{1}_n = b\}$. The Wasserstein distance raised to power 2 now reads

$$W_2^2(P, Q) = \min_{T\in U(a,b)} \mathrm{tr}\,(T^\top M_{XY}),$$

which has its entropic relaxed formula

$$\hat{W}_2^2(P, Q) = \min_{T\in U(a,b)} \mathrm{tr}\,(T^\top M_{XY}) + \gamma \sum_{i,j=1}^{n,k} T_{ij}\log T_{ij}. \tag{28}$$

The Wasserstein barycenter up to k support points of P_j, i.e. $P_j = \sum_{i=1}^{m} a_{ji}\delta_{X_{ji}}$, for $1 \le j \le N$ with weights of $\omega = (\omega_1, \ldots, \omega_N) \in \Delta_N$ is the minimizer of the following problem

$$Q = \min_{b,\boldsymbol{Y}} \sum_{j=1}^{N} \omega_j \,\mathrm{tr}\,(T^\top M_{X_jY}), \tag{29}$$

where $b \in \Delta_k$ and $\boldsymbol{Y} = Y_1, \ldots, Y_k$.

Algorithm 7 is designed to compute the optimal transport plan T^*_γ and dual optima b^*_γ (aka sub-gradient) of relaxed Wasserstein distance with respect to b in equation (28). Algorithm 8 describes routines for computing the weights of the barycenter in (29). The procedure for computing free support barycenter in (29) is summarized in Algorithm 9.

Algorithm 8 Fix-support Wasserstein barycenter

Input: Atoms $X_j \in \mathbb{R}^{d\times n_j}$ and weights a_j of P_j, cost matrices $M_j = M_{X_jY}$ and $\omega \in \Delta_N$, $\gamma, t_0 > 0$.
Output: Weights b of Q_N.
Initialize $\tilde{b}$ and set $\hat{b} = \tilde{b}$.
Set $t = 0$.
while $\hat{b}$ has not converged **do**
  $\beta = (t+1)/2$, $b = (1-\beta^{-1})\hat{b} + \beta^{-1}\tilde{b}$.
  Compute sub-gradient α_js using Algorithm 7 using M_j, b and a_j
  $\alpha = \sum_{j=1}^{N} \omega_j \alpha_j$
  $\tilde{b} = \tilde{b} * \exp(-t_0\beta\alpha)$, $\tilde{b} = \tilde{b}/\tilde{b}^\top \mathbf{1}_k$
  $\hat{b} = (1-\beta^{-1})\hat{b} + \beta^{-1}\tilde{b}$
end while
return $\hat{b}$

Algorithm 9 Free-support Wasserstein barycenter

Input: Atoms $X_j \in \mathbb{R}^{d\times n_j}$ and weights a_j of P_j, cost matrices $M_j = M_{X_jY}$ and $\omega \in \Delta_N$, $\gamma, t_0 > 0$, k – the number of support points of barycenter.
Output: Atoms $Y \in \mathbb{R}^{d\times k}$ and weights b of Q_N.
Initialize Y and b.
Set $t = 0$.
while Y and b have not converged **do**
  Computing b using Algorithm 8
  for $j = 1$ **to** N **do**
    Computing T_j using Algorithm 7
  end for
  Setting $\theta \in [0, 1]$ with line search or a preset value.
  $X = (1-\theta)X + \theta(\sum_{j=1}^{N} \omega_j X_j T_j^\top)\mathrm{diag}(b^{-1})$
end while
return X, b

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43:904–924, 2011.

E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *http://arxiv.org/abs/1507.07218*, 2015.

D. Arthur and S. Vassilvitskii. K-means++: The avantages of careful seeding. *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

J. D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Payré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 2: 1111–1138, 2015.

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res*, 3: 993–1022, 2003.

M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems 26*, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *ICML*, pages 1366–1375, 2018.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, New York, 2000.

N. Ho, X. Nguyen, M. Yurochkin, H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

V. Huynh, D. Phung, S. Venkatesh, X. Nguyen, M. Hoffman, and H. Bui. Scalable nonparametric Bayesian multilevel clustering. *Proceedings of Uncertainty in Artificial Intelligence*, 2016.

B. Kulis and M. I. Jordan. Revisiting K-means: new algorithms via Bayesian nonparametrics. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

J. Li and J. Z. Wang. Real-time computerized annotation of pictures. In *Proceedings of the ACM Multimedia Conference*, pages 911–920, 2006.

T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: an analysis of greedy and accelerated mirror descent algorithms. In *ICML*, pages 3982–3991, 2019a.

T. Lin, N. Ho, and M. I. Jordan. On the efficiency of the Sinkhorn and Greenkhorn algorithms and their acceleration for optimal transport. *ArXiv Preprint: 1906.01437*, 2019b.

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. I. Jordan. Fixed-support Wasserstein barycenters: computational hardness and fast algorithm. In *NeurIPS*, 2020.

T. B. Nguyen, V. Nguyen, S. Venkatesh, and D. Phung. MCNC: Multi-channel nonparametric clustering from heterogeneous data. In *Proceedings of ICPR*, 2016.

V. Nguyen, D. Phung, X. Nguyen, S. Venkatesh, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013.

X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21:618–646, 2015.

X. Nguyen. Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22:1535–1571, 2016.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

O. Pele and M. Werman. Fast and robust earth mover's distance. In *ICCV*. IEEE, 2009.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9:135–140, 1981.

D. Pollard. Quantization and the method of K-means. *IEEE Transactions on Information Theory*, 28:199–205, 1982.

J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

A. Rodriguez, D. Dunson, and A.E. Gelfand. The nested Dirichlet process. *J. Amer. Statist. Assoc.*, 103(483):1131–1154, 2008.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser Basel, 2015.

J. Solomon, G. Fernando, G. Payré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. In *The International Conference and Exhibition on Computer Graphics and Interactive Techniques*, 2015.

J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198. ACM, 2014.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

Y. Vardi and C. H. Zhang. The multivariate ℓ_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97:1423–1426, 2000.

C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemtical Sciences]*. Springer, Berlin, 2009.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.

D. F. Wulsin, S. T. Jensen, and B. Litt. Nonparametric multi-level clustering of human epilepsy seizures. *Annals of Applied Statistics*, 10:667–689, 2016.

P. C. Álvarez Estebana, E. del Barrioa, J.A. Cuesta-Albertosb, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441:744–762, 2016.