Multilevel Clustering via Wasserstein Means

Nhat Ho [1] **XuanLong Nguyen** [1] **Mikhail Yurochkin** [1] **Hung Hai Bui** [2] **Viet Huynh** [3] **Dinh Phung** [3]

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 a number of 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, experiment results with both synthetic and real data are presented to demonstrate the flexibility and scalability of the proposed approach. [1]

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, documents are grouped into corpora. A prominent strand of modeling and algorithmic works in the past couple 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). Another concrete example is the problem of clustering images (i.e., global clusters) where each image contains partions of multiple annotated regions (i.e., local clusters) (Oliva and Torralba, 2001). While hierachical clustering techniques may be employed to find a tree-structed 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 works 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, namely 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 works, but we shall take a purely optimization approach. 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 for 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 the optimization in the space of probability mea-

[1] Department of Statistics, University of Michigan, USA. [2] Adobe Research. [3] Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia. Correspondence to: Nhat Ho <minhnhat@umich.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] Code is available at `https://github.com/moonfolk/Multilevel-Wasserstein-Means`

sures 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 purely optimization viewpoint, but they may also be taken as a fast approximation to the inference of latent mixing measures that arise in the nested Dirichlet process of (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 betwen our optimization and that of finding the Wasserstein barycenter (Agueh and Carlier, 2011), an interesting computational problem that have also attracted much recent interests, e.g., (Cuturi and Doucet, 2014).

In summary, the main contributions offered in this work include (i) a new optimization formulation to 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 proposed estimates under very mild condition of data's distributions; (iv) several flexibile alternatives by introducing constraints that encourage the borrowing of strength among local and global clusters, and (v) finally, demonstration of efficiency and flexibility of our approach in a number of simulated and real data sets.

The paper is organized as follows. Section 2 provides preliminary background on Wasserstein distance, Wasserstein barycenter, and the connection between K-means clustering and the quantization problem. Section 3 presents several optimization formulations of the multilevel clusering problem, and the algorithms for solving them. Section 4 establishes consistency results of the estimators introduced in Section 4. Section 5 presents careful simulation studies with both synthetic and real data. Finally, we conclude the paper with a discussion in Section 6. Additional technical details, including all proofs, are given in the Supplement.

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 $\mathcal{P}_r(\Theta) = \left\{ G \in \mathcal{P}(\Theta) : \int \|x\|^r dG(x) < \infty \right\}$, where $\|.\|$ denotes Euclidean metric in $\mathbb{R}^d$. Additionally, for any $k \geq 1$ the probability simplex is denoted by $\Delta_k = \left\{ u \in \mathbb{R}^k : u_i \geq 0, \sum\limits_{i=1}^{k} u_i = 1 \right\}$. 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)):

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

where $\Pi(G, G')$ is the set of all probability measures on $\Theta \times \Theta$ that have marginals G and G'. In words, $W_r^r(G, G')$ 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. When G and G' are two discrete measures with finite number of atoms, fast computation of $W_r(G, G')$ can be achieved (see, e.g., (Cuturi, 2013)). The details of this are deferred to the Supplement.

By a 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 (that is, 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 $\mathcal{D}, \mathcal{D}' \in \mathcal{P}_r(\mathcal{P}_r(\Theta))$

$$W_r(\mathcal{D}, \mathcal{D}') := \left(\inf \int_{\mathcal{P}_r(\Theta)^2} W_r^r(G, G') d\pi(G, G') \right)^{1/r}$$

where the infimum in the above ranges over all $\pi \in \Pi(\mathcal{D}, \mathcal{D}')$ such that $\Pi(\mathcal{D}, \mathcal{D}')$ is the set of all probability measures on $\mathcal{P}_r(\Theta) \times \mathcal{P}_r(\Theta)$ that has 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 notational abuse — 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.

Wasserstein barycenter Next, we present a brief overview of Wasserstein barycenter problem, first studied by (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

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

where $\lambda \in \Delta_N$ denote 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 by (Anderes et al., 2015) that the problem of finding Wasserstein barycenter $\overline{P}_{N,\lambda}$ over the space $\mathcal{P}_2(\Theta)$ in (1) is reduced to search only over 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. The notion of Wasserstein barycenter has been utilized for approximate Bayesian inference (Srivastava et al., 2015).

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 but very useful connection (Pollard, 1982; Graf and Luschgy, 2000). The connection is the following. Given n unlabelled samples $Y_1, \ldots, Y_n \in \Theta$. 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 minimizes the following objective

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

Let $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$ be the empirical measure of data $Y_1, \ldots, Y_n$. Then, problem (2) is equivalent to finding a discrete probability measure G which has finite number of support points and solves:

$$\inf_{G \in \mathcal{O}_k(\Theta)} W_2^2(G, P_n). \tag{3}$$

Due to the inclusion of 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 by (Cuturi and Doucet, 2014), the algorithm for finding the Wasserstein barycenter offers an alternative for the popular Loyd's algorithm for determing 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 presented 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 $P^j_{n_j} := \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{X_{j,i}}$. Throughout this section, for simplicity of exposition we assume that the number 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.

High level idea For local clustering, for each $j = 1, \ldots, m$, performing a K-means clustering for group j, as expressed by (3), can be viewed as finding a finite discrete measure $G_j \in \mathcal{O}_{k_j}(\Theta)$ that minimizes squared Wasserstein distance $W_2^2(G_j, P^j_{n_j})$. 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 Eq. (3), 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 succintly expressed by

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

However, G_j 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

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{j=1}^{m} W_2^2(G_j, P^j_{n_j}) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^{m} \delta_{G_j}). \tag{4}$$

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 distance in the space of measures, e.g., $W_2^2(G_j, P^j_{n_j})$, and the other in space of measures of measures, e.g., $W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^{m} \delta_{G_j})$. By adopting K-means optimization to both local and global clustering, the multilevel Wasserstein means problem might look formidable at the 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 (4) is equivalent to the following optimization problem, which looks much simpler as it involves only measures on Θ:

$$\inf_{G_j \in \mathcal{O}_{k_j}(\Theta), \boldsymbol{H}} \sum_{j=1}^{m} W_2^2(G_j, P^j_{n_j}) + \frac{d^2_{W_2}(G_j, \boldsymbol{H})}{m} \tag{5}$$

where $d^2_{W_2}(G, \boldsymbol{H}) := \min\limits_{1 \leq i \leq M} W_2^2(G, H_i)$ 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 the Supplement. Before going into to the details of the algorithm for solving (5) in Section 3.1.2, we shall present some simpler cases, which help to illustrate some properties of the optimal solutions of (5), while providing insights of subsequent developments of the MWM formulation. Readers may proceed directly to Section 3.1.2 for the description of the algorithm in the first reading.

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 (5) can be rewritten as

$$\inf_{\substack{\theta_j \in \Theta, \\ H \in \mathcal{P}_2(\Theta)}} \sum_{j=1}^{m} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + W_2^2(\delta_{\theta_j}, H)/m, \qquad (6)$$

where $G_j = \delta_{\theta_j}$ for any $1 \leq j \leq m$. From the result of Theorem A.1 in the Supplement,

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

where second infimum is achieved when $H = \delta_{(\sum\limits_{j=1}^{m} \theta_j)/m}$. Thus, objective function (6) 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 $\overline{X}_j = (\sum\limits_{i=1}^{n} X_{j,i})/n$ 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 $\theta_j = \left((m^2n + 1)\overline{X}_j + \sum\limits_{i \neq j} \overline{X}_i\right)/(m^2n + m)$ for any $1 \leq j \leq m$. If we further assume that our data $X_{j,i}$ are i.i.d samples from 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 are pairwise different, the multi-level Wasserstein means under that simple scenario of (5) will not have identical centers among local groups.

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

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

where σ in the above sum varies over all the permutation 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 (6) 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. $k_j = 1$ 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

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

i.e., the distances of empirical measures P_n^i and P_n^j when i and j belong to the same set A or A^c are much less than those when i and j do not belong to the same set. Under this condition, by using the argument from part (i) we can write the objective function (5) as

$$\begin{aligned} &\inf_{\substack{\theta_j \in \Theta, \\ H_1 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_1)}{|A|} + \\ &\inf_{\substack{\theta_j \in \Theta, \\ H_2 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A^c} \sum_{i=1}^{n} \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_2)}{|A^c|}. \end{aligned}$$

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_j do not share atoms under that setting of empirical measures.

The examples examined above indicate that the MWM problem in general do not "encourage" the local measures G_j 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_j can be very small.

3.1.2. Algorithm Description

Now we are ready to describe our algorithm in the general case. This is a procedure for finding a local minimum of Problem (5) and is summarized in Algorithm 1. We prepare the following details regarding the initialization and updating steps required by the algorithm:

Algorithm 1 Multilevel Wasserstein Means (MWM)

```
Input: Data $X_{j,i}$, Parameters $k_j$, $M$.
Output: prob. 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 $Y_j^{(t)}, b_j^{(t)}, H_i^{(t)}$ have not converged do
  1. Update $Y_j^{(t)}$ and $b_j^{(t)}$ for $1 \leq j \leq m$:
  for $j = 1$ to $m$ do
    $i_j \leftarrow \underset{1 \leq u \leq M}{\arg\min}\, W_2^2(G_j^{(t)}, H_u^{(t)})$.
    $G_j^{(t+1)} \leftarrow \underset{G_j \in \mathcal{O}_{k_j}(\Theta)}{\arg\min}\ W_2^2(G_j, P_{n_j}^j) +$
    $+ W_2^2(G_j, H_{i_j}^{(t)})/m$.
  end for
  2. Update $H_i^{(t)}$ for $1 \leq i \leq M$:
  for $j = 1$ to $m$ do
    $i_j \leftarrow \underset{1 \leq u \leq M}{\arg\min}\, W_2^2(G_j^{(t+1)}, H_u^{(t)})$.
  end for
  for $i = 1$ to $M$ do
    $C_i \leftarrow \{l : i_l = i\}$ for $1 \leq i \leq M$.
    $H_i^{(t+1)} \leftarrow \underset{H_i \in \mathcal{P}_2(\Theta)}{\arg\min} \sum_{l \in C_i} W_2^2(H_i, G_l^{(t+1)})$.
  end for
  3. $t \leftarrow t + 1$.
end while
```

- 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 $H^{(0)}$ is based on a simple extension of the K-means algorithm. Details are given in Algorithm 3 in the Supplement;

- The updates $G_j^{(t+1)}$ 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 $G_j^{(t+1)}$ are finite discrete measures, finding the updates for $H_i^{(t+1)}$ over the whole space $\mathcal{P}_2(\Theta)$ can be reduced to searching for a local solution within 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 $\boldsymbol{H}^{(t)}$ (Justification of this reduction is derived from Theorem A.1 in the Supplement). 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 impose a threshold L, e.g., $L = 10$, for $l^{(t)}$ in its implementation.

The following guarantee for Algorithm 1 can be established:

Theorem 3.1. *Algorithm 1 monotonically decreases the objective function* (4) *of the MWM formulation.*

3.2. Multilevel Wasserstein Means with Sharing

As we have observed from the analysis of several specific cases, the **multilevel Waserstein means** formulation may not encourage the 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, consequentially improving the inferential 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 (4). Denote $\mathcal{A}_{M,\mathcal{S}_K} = \Big\{ G_j \in \mathcal{O}_K(\Theta),\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta)) : \text{supp}(G_j) \subseteq \mathcal{S}_K\ \forall 1 \leq j \leq m \Big\}$ 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 multilevel Wasserstein means problem

$$\inf \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^{m} \delta_{G_j}). \tag{7}$$

where $\mathcal{S}_K$, $G_j, \mathcal{H} \in \mathcal{A}_{M,\mathcal{S}_K}$ 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 the work of (Kulis and Jordan, 2012), who developed an optimization-based approach to the inference of the HDP (Teh et al., 2006), which also encourages explicitly the sharing of local group means among local clusters. Now, we can rewrite objective function (7) as follows

$$\inf_{\mathcal{S}_K, G_j, \boldsymbol{H} \in \mathcal{B}_{M,\mathcal{S}_K}} \sum_{j=1}^{m} W_2^2(G_j, P_{n_j}^j) + \frac{d_{W_2}^2(G_j, \boldsymbol{H})}{m} \tag{8}$$

where $\mathcal{B}_{M,\mathcal{S}_K} = \Big\{ G_j \in \mathcal{O}_K(\Theta),\ \boldsymbol{H} = (H_1, \ldots, H_M) : \text{supp}(G_j) \subseteq \mathcal{S}_K\ \forall 1 \leq j \leq m \Big\}$. The high level idea of finding local minimums of objective function (8) is to first, update the elements of constraint set $\mathcal{S}_K$ to provide the supports for local measures G_j and then, obtain the weights of these measures as well as the elements of global set H by computing appropriate Wasserstein barycenters. Due to space constraint, the details of these steps of the MWMS Algorithm (Algorithm 2) are deferred to the Supplement.

4. Consistency results

We proceed to establish consistency for the estimators introduced in the previous section. For the brevity of the presentation, we only focus on the MWM method; consistency for MWMS can be obtained in a similar fashion. 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

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

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

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

Theorem 4.1. *Given that $P^j \in \mathcal{P}_2(\Theta)$ for $1 \leq j \leq m$. Then, there holds almost surely, as $\boldsymbol{n} \to \infty$*

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

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 (4). 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 $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$, 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 4.2. *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.*

Remark: (i) The assumption Θ is bounded is just for the convenience of proof argument. We believe that the conclusion of this theorem may still hold when $\Theta = \mathbb{R}^d$. (ii) If $|\mathcal{F}| = 1$, i.e., there exists an 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 4.2 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 $\boldsymbol{n} \to \infty$.

5. Empirical studies

5.1. Synthetic data

In this section, we are interested in evaluating the effectiveness of both MWM and MWMS clustering algorithms by considering different synthetic data generating processes. Unless otherwise specified, we set the number of groups $m = 50$, number of observations per group $n_j = 50$ in $d = 10$ dimensions, number of global clusters $M = 5$ with 6 atoms. For Algorithm 1 (MWM) local measures G_j have 5 atoms each; for Algorithm 2 (MWMS) number of atoms in constraint set S_K is 50. As a benchmark for the comparison we will use a basic 3-stage K-means approach (the details of which can be found in the Supplement). The Wasserstein distance 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, while the MWMS formulation explicitly enforces the sharing of atoms across these measures. We used multiple layers of mixtures while adding Gaussian noise at each layer to generate global and local clusters and the no-constraint (NC) data. We varied number of groups m from 500 to 10000. We notice that the 3-stage K-means algorithm performs the best when there is no constraint structure *and* variance is constant across clusters (Fig. 1(a) and 2(a)) — 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, both of our algorithms start to outperform the basic three-stage K-means. The superior performance is most pronounced with local-constraint (LC) data (with or without constant variance conditions). See Fig. 1(c,d). It is worth noting that even when group variances are constant, the 3-stage K-means is no longer longer effective because now fails to account for the shared structure. When $m = 50$ and group sizes are larger, we set $S_K = 15$. Results are reported in Fig. 2 (c), (d). These results demonstrate the effectiveness and flexibility of our both algorithms.

5.2. Real data analysis

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

LabelMe dataset 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 image from *open country* category and contains 4 regions while the right panel denotes an image of *tall buildings* category

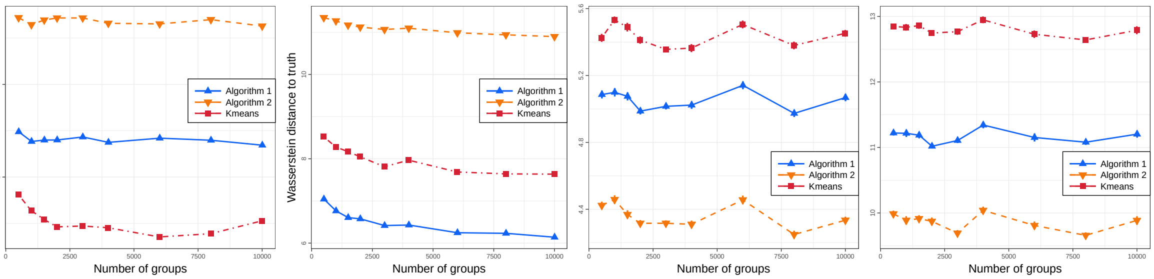


Figure 1: Data with a lot of small groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

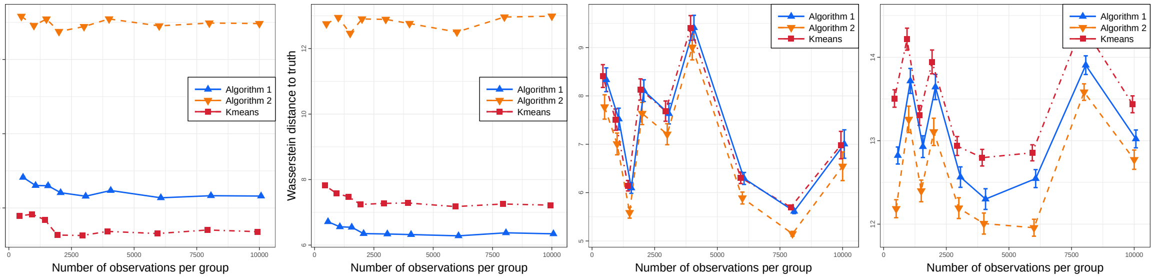


Figure 2: Data with few big groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

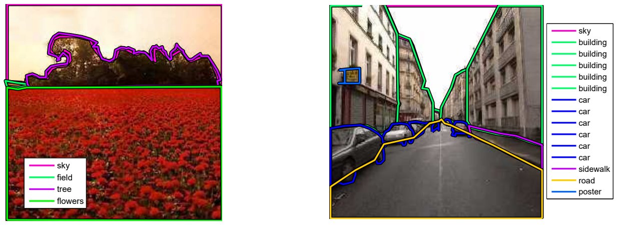


Figure 4: Examples of images used in LabelMe dataset. Each image consists of different annotated regions.

Table 1: Clustering performance for LabelMe dataset.

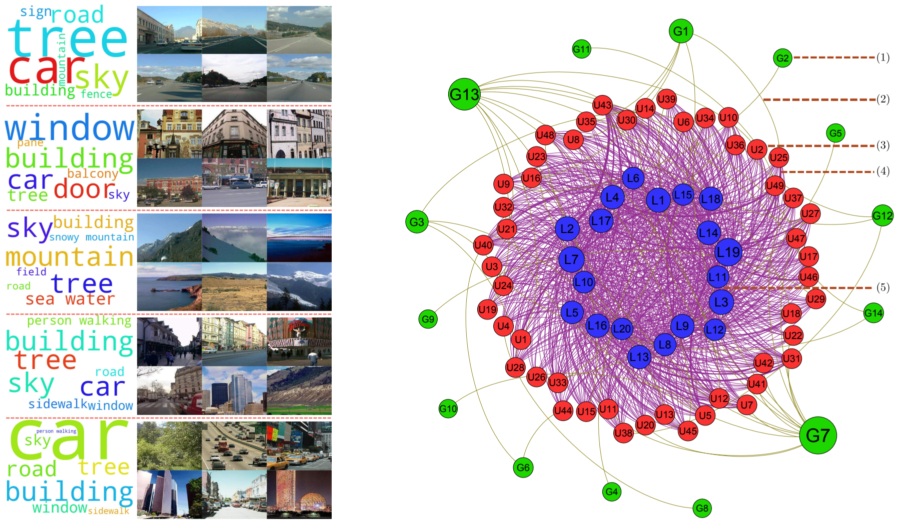
Methods	NMI	ARI	AMI	Time (s)
K-means	0.349	0.237	0.324	**0.3**
TSK-means	0.236	0.112	0.22	218
MC2	0.315	0.206	0.273	4.2
MWM	0.373	0.263	0.352	332
MWMS	**0.391**	**0.284**	**0.368**	544

including 16 regions. Note that the regions in each image can be overlapped. We remove the images containing less then 4 regions and obtain $1,800$ images.

We then extract GIST feature (Oliva and Torralba, 2001) for each region in a image. GIST is a visual descriptor to represent perceptual dimensions and oriented spatial structures of a scene. Each GIST descriptor is a 512-dimensional vector. We further use PCA to project GIST features into 30 dimensions. Finally, we obtain $1,800$ "documents", each of which contains regions as observations. Each region now 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.

StudentLife dataset is a large dataset frequently used in pervasive and ubiquitous computing research. Data signals consist of multiple channels (e.g., WiFi signals, Bluetooth scan, etc.), 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 applied 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.

Experimental results. To quantitatively evaluate our proposed methods, we compare our algorithms with several base-line methods: K-means, three-stage K-means (TSK-means) as described in the Supplement, MC2-SVI without context (Huynh et al., 2016). Clustering performance in Table 1 is evaluated with the image clustering problem for *LabelMe dataset*. With *K-means*, we average all data points



(a)

(b)

Figure 3: 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.

to obtain a single vector for each images. 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 varitational and a parallelized Spark-based implementation in (Huynh et al., 2016) to carry out experiments. This implementation has the advantage of making use of all of 16 cores on the test machine. The running time for MC2-SVI is reported after scanning one epoch. In terms of clustering accuracy, MWM and MWMS algorithms perform the best.

Fig. 3a 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. These tag-clouds can be considered as visual ground truth of clusters. Our algorithm can group images into clusters which are consistent with their tag-clouds.

We use 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. Fig. 3b 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 clusteing we can dissect students' shared location (activities), e.g. Student 49 (*U49*) mainly takes part in activity location 4 (*L4*).

6. Discussion

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; 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 penalty on the model complexity. Thirdly, our formulation does not directly account for the "noise" distribution away from the (Wasserstein) means. To improve the robustness, it may be desirable to make use of the first-order Wasserstein metric instead of the second-order one. Finally, we are interested in extending our approach to richer settings of hierarchical data, such as one when group level-context is available.

Acknowledgement. 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 gratefully acknowledges the partial support from the Australian Research Council (ARC) and AOARD (FA2386-16-1-4138).

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.

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.

S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag, New York, 2000.

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.

Thanh-Binh Nguyen, Vu Nguyen, Svetha Venkatesh, and Dinh 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. 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.

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.

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.

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

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

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

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.