A Convex Exemplar-based Approach to MAD-Bayes Dirichlet Process Mixture Models

Ian E. H. Yen [1] IANYEN@CS.UTEXAS.EDU
Xin Lin [1] JIMMYLIN@UTEXAS.EDU
Kai Zhong [2] ZHONGKAI@ICES.UTEXAS.EDU
Pradeep Ravikumar [1,2,3] PRADEEPR@CS.UTEXAS.EDU
Inderjit S. Dhillon [1,2] INDERJIT@CS.UTEXAS.EDU

[1] Department of Computer Science, University of Texas at Austin, TX 78712, USA.
[2] Institute for Computational Engineering and Sciences, University of Texas at Austin, TX 78712, USA.
[3] Department of Statistics and Data Sciences, University of Texas at Austin, TX 78712, USA.

Abstract

MAD-Bayes (MAP-based Asymptotic Derivations) has been recently proposed as a general technique to derive scalable algorithm for Bayesian Nonparametric models. However, the combinatorial nature of objective functions derived from MAD-Bayes results in hard optimization problem, for which current practice employs heuristic algorithms analogous to k-means to find local minimum. In this paper, we consider the exemplar-based version of MAD-Bayes formulation for DP and Hierarchical DP (HDP) mixture model. We show that an exemplar-based MAD-Bayes formulation can be relaxed to a convex structural-regularized program that, under cluster-separation conditions, shares the same optimal solution to its combinatorial counterpart. An algorithm based on Alternating Direction Method of Multiplier (ADMM) is then proposed to solve such program. In our experiments on several benchmark data sets, the proposed method finds optimal solution of the combinatorial problem and significantly improves existing methods in terms of the exemplar-based objective.

1. Introduction

In the recent years, MAD-Bayes (MAP-based Asymptotic Derivations) has been proposed as a technique to obtain scalable algorithm for Bayesian Nonparametric models such as Dirichlet Process (DP) Mixture, Latent Feature Allocation and Infinite Hidden Markov Model (Broderick et al., 2013; Kulis & Jordan, 2012; Roychowdhury et al., 2013; Jiang et al., 2012; Campbell et al., 2013). The objective function derived from MAD-Bayes, however, hard to optimize and existing approaches employ algorithms analogous to k-means that only guarantees to find local optimum. In this paper, we consider exemplar-based version of the MAD-Bayes objective, where the *mean parameters* of each mixture are restricted to be one of the data samples.

The exemplar-based clustering, namely k-medoids problem, has been investigated since 1987 and known for its ability to handle arbitrary dissimilarity measure (Kaufman & Rousseeuw, 1987). The k-medoids problem, however, is still NP-hard in general (Papadimitriou, 1981; Megiddo & Supowit, 1984) and the best known approximate algorithm guarantees an approximation ratio of $1 + \sqrt{3} + \epsilon \approx 2.732$ with theoretical limit being $1 + 2/e \approx 1.736$. Commonly used algorithms such as Partitioning Around Medoids (PAM) (Van der Laan et al., 2003) and Affinity Propagation (AP) (Frey & Dueck, 2007) only guarantee to find local optimum. On the other hand, a recent interest in the exemplar-based formulation arose due to its natural convex relaxation (Yaron & Shalev-Shwartz, 2010; Elhamifar et al., 2012; Awasthi et al., 2015; Nellore & Ward, 2013) . The convex relaxation, without additional assumptions, could lead to fractional solutions and thus does not guarantee to find optimum of the original problem (Yaron & Shalev-Shwartz, 2010). However, Elhamifar et al. (Elhamifar et al., 2012) and later on Nellore et al. (Nellore & Ward, 2013) proved that, when there exists clustering assignments satisfying certain separation requirement (that is, the dissimilarities within cluster are small enough compared to the dissimilarities between clusters), the convex relaxation guarantees to find the optimal solution of its com-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

binatorial counterpart.

In this paper, we show that the theory for convex relaxation of k-medoids also applies to a variety of Bayesian Nonparametric models through connection to the MAD-Bayes formulation. This results in an exemplar-based objective with convex structural-regularized relaxation that guarantees to recover the optimal solution under the separation requirement.

The paper will be organized as follows. In Section 2, we derive relaxation for the exemplar-based Dirichlet Process (DP) and Hierarchical DP (HDP) mixture models under MAD-Bayes formulation. Then in Section 3, we provide sufficient conditions under which one can recover the optimal solution from the relaxation. In Section 4, an algorithm based on Alternating Direction Method of Multiplier (ADMM) is introduced to efficiently solve the convex program. In Section 6, we experiment on several benchmark data sets, for which the proposed approach recovers the optimal solution of the combinatorial problem and significantly improves the objective achieved by existing methods.

2. Exemplar-Based MAD-Bayes

In this section, we derive convex relaxation for the exemplar-based MAD-Bayes estimation of DP and HDP mixture models.

2.1. Exemplar-based Dirichlet Process (DP) Mixture

Given a mixture model with DP prior $p(z)$ and some observation distribution $p(\boldsymbol{x}|z)$ belonging to the exponential family, the MAP inference of $\{z_i\}_{i=1}^N$ given observation $\{\boldsymbol{x}_i\}_{i=1}^N$ can be formulated via MAD-Bayes as

$$\min_{z_i\in[K],\boldsymbol{\mu}_k\in\mathbb{R}^p,K} \quad \sum_{i=1}^{N}\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{\mu}_{z_i})+\lambda K \tag{1}$$

by taking variance of $p(\boldsymbol{x}|z)$ asymptotically to 0, where $\mathbb{D}(.)$ is Bregman divergence associated with the partition function of $p(\boldsymbol{x}|z)$ (Kulis & Jordan, 2012; Broderick et al., 2013; Jiang et al., 2012). For $p(\boldsymbol{x}|z)$ being *Spherical Gaussian*, $\mathbb{D}(.)$ is simply the *square Euclidian distance* and for $p(\boldsymbol{x}|z)$ being *Multinouli* distribution, $\mathbb{D}(.)$ is the *KL-divergence*. Note that K in the formulation is also a variable to optimize, which, as in the spirit of Bayesian Nonparametrics, provides the flexibility to vary number of clusters based on the supports from the data. In (Kulis & Jordan, 2012), an algorithm analogous to k-means called *DP-means* was proposed to solve (1) with Gaussian observation distribution, which was then adopted and generalized in (Jiang et al., 2012) to solve instances of exponential family observation distribution, and improved in (Broderick et al., 2013) using idea of collapsed sampling. However, none of those algorithms provides guarantee for the quality of solution w.r.t. its objective.

Here we consider the exemplar-based version of (1), which confines the optimization space of $\boldsymbol{\mu}_k$ to a finite set of candidates (exemplars) $\mathcal{E}=\{\bar{\boldsymbol{\mu}}_j\}_{j=1}^J$ instead of the whole parameter space $\mathbb{R}^p$. For mixture model a natural choice of $\mathcal{E}$ is the sample set $\{\boldsymbol{x}_j\}_{j=1}^N$ itself. The problem can then be written as

$$\begin{aligned}\min_{w_{ij}\in\{0,1\}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{J} w_{ij}\mathbb{D}(\boldsymbol{x}_i,\bar{\boldsymbol{\mu}}_j)+\lambda\sum_{j=1}^{J}\max_{i\in[N]} w_{ij} \\ s.t. \quad & \sum_{j=1}^{J} w_{ij}=1,\forall i.\end{aligned} \tag{2}$$

where with $w_{ij}\in\{0,1\}$, the term $\sum_{j=1}^J \max_{i\in[N]} w_{ij}$ simply counts number of exemplars j assigned to *any* sample i. In contrast to formulation (1) where $\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{\mu}_j)$ must be a Bregman Divergence so the minimization w.r.t. $\boldsymbol{\mu}_j$ can be carried out efficiently [1], the exemplar-based objective (2) takes the advantage that $\mathbb{D}(\boldsymbol{x}_i,\bar{\boldsymbol{\mu}}_j)$ is a pre-computable constant that can be any measure of dissimilarity. By replacing integer constraint $w_{ij}\in\{0,1\}$ with nonnegative constraint $w_{ij}\geq 0$, we obtain a convex problem of linear loss function, group-sparse regularization, and simplex constraint:

$$\min_{W\in\mathbb{R}_+^{N\times J}:W\mathbf{1}=\mathbf{1}} \quad \mathbb{D}\circ W+\lambda\|W\|_{\infty,1} \tag{3}$$

where $\circ$ is the element-wise inner product, $\mathbb{D}$ is an N by N matrix with $\mathbb{D}_{ij}=\mathbb{D}(\boldsymbol{x}_i,\bar{\boldsymbol{\mu}}_j)$, and $\mathbf{1}$ is J by 1 vector of all elements equal to 1. Note without imposing additional conditions, (3) might have fractional solutions. Before discussion of the recovery conditions, we first introduce HDP mixture, a generalization of DP mixture when data come in groups.

2.2. Exemplar-based Hierarchical Dirichlet Process (HDP) Mixture

In a HDP mixture model, samples $\{\boldsymbol{x}_i\}_{i=1}^N$ are grouped into D data sets $\mathcal{T}_1$,..,$\mathcal{T}_D$, each of which is modeled as a local DP mixture $DP(\alpha,G)$, while G is a base distribution drawn from a global DP shared among all data sets. The resulting MAD-Bayes estimation problem, as derived in (Kulis & Jordan, 2012; Jiang et al., 2012), has two penalty terms, one for the number of global clusters K_g and the

[1] The fact that *mean* serves as minimizer of a sum of Bregman Divergence w.r.t. the second parameter is used to derive the *DP-means* algorithm.

other for the number of clusters used by each data set K_d:

$$\begin{aligned} \min_{z_i\in[K_g],\boldsymbol{\mu}_k,K_g,K_d} \quad & \sum_{i=1}^{N}\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{\mu}_{z_i})+\theta K_l+\lambda K_g \\ s.t. \quad & K_l=\sum_{d=1}^{D}K_d \\ & K_d=\left|\{z_i|i\in\mathcal{T}_d\}\right|, d=1,..,D. \end{aligned} \tag{4}$$

where θ, λ are the hyper-parameters for local and global penalty respectively. To obtain the exemplar-based version, we constrain the space of $\boldsymbol{\mu}_k$ to a finite candidate set $\mathcal{E}=\{\bar{\boldsymbol{\mu}}_j\}_{j=1}^{J}$, so the problem can be written as

$$\begin{aligned} \min_{w_{ij}\in\{0,1\}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{J}w_{ij}\mathbb{D}(\boldsymbol{x}_i,\bar{\boldsymbol{\mu}}_j) \\ & +\theta\sum_{d=1}^{D}\sum_{j=1}^{J}\max_{i\in\mathcal{T}_d}w_{ij}+\lambda\sum_{j=1}^{J}\max_{i\in[N]}w_{ij} \\ s.t. \quad & \sum_{j=1}^{J}w_{ij}=1,\forall i. \end{aligned} \tag{5}$$

By replacing $w_{ij}\in\{0,1\}$ with $w_{ij}\geq 0$, we obtain the corresponding convex relaxation

$$\begin{aligned} \min_{W\in\mathbb{R}^{N\times J}} \quad & \mathbb{D}\circ W+\theta\|W\|_{\mathcal{G}}+\lambda\|W\|_{\infty,1} \\ s.t. \quad & W\succeq 0 \\ & W\mathbf{1}=\mathbf{1} \end{aligned} \tag{6}$$

where $\|.\|_{\mathcal{G}}$ is a latent group norm (Obozinski et al., 2011) defined by the collection of groups $\mathcal{G}=\{g_{d,j}|d\in[D],j\in[J]\}$, with each group of indices defined as $g_{d,j}=\{(i,j)|i\in\mathcal{T}_d\}$.

3. Optimality Guarantee

In this section, we give conditions under which the convex formulations (3), (6) share the same optimal solution with its combinatorial counterparts (2), (5). In particular, we show that if there exists a clustering that satisfies certain separation requirement, then it corresponds to an unique optimal solution W^* of both combinatorial and convex formulations with some λ, θ. Note an integer solution W^* of (2), (5) corresponds to a clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$, where $\mathcal{M}=\{j|\exists i,w_{ij}^*=1\}$ is the set of representative exemplars selected out of the candidate set $\mathcal{E}$, and $\mathcal{S}_k=\{i|w_{ik}^*=1\}$ is a set containing samples belonging to cluster represented by exemplar k. We use $M(i)$ to denote the representative i-th sample assigned to. Note each representative $k\in\mathcal{M}$ has minimal average dissimilarity to $\forall i\in\mathcal{S}_k$. In the following, we will assume exemplar set $\mathcal{E}$ to be the set of samples $\mathcal{E}=\{\boldsymbol{x}_j\}_{j=1}^{N}$. All proofs will be included in appendix.

Theorem 1. *Suppose there exists a clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ for which we can find λ such that*

$$\max_{k\in\mathcal{M}}\max_{i,j\in\mathcal{S}_k}N_k\delta_{ij}<\lambda<\min_{(k,l\in\mathcal{M},k\neq l)}\min_{(i\in\mathcal{S}_k,j\in\mathcal{S}_l)}N_k\delta_{ij} \tag{7}$$

where $N_k=|\mathcal{S}_k|$ and $\delta_{ij}=\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{x}_{M(i)})$, then the integer solution W^* realizing $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ is unique optimal solution to both (2) and (3).

Note that Theorem 1 is consistent with Corollary 7 given by (Nellore & Ward, 2013), where the Lagrange multiplier u in (Nellore & Ward, 2013) corresponds to our regularization parameter λ. Since DP is a special case of HDP, the proof for Theorem 1 is also a special case of that for Theorem 2. When each cluster has same size, Theorem 1 simply states a separation condition that requires the dissimilarity between points of different clusters, subtracted by the dissimilarity to cluster representative, larger than that between points in the same cluster. Intuitively, the separation condition is very reasonable. The fractional solution can be viewed as assigning a data point to multiple clusters. When clusters are separated with an enough distance, it is not likely for a data point being assigned to multiple clusters. When the measure of dissimilarity $\mathbb{D}(.)$ satisfies triangular inequality, we can further simplify the result as follows.

Corollary 1. *Assuming $\mathbb{D}(\cdot,\cdot)$ is a metric and $N_k=N/K,\forall k\in\mathcal{M}$, where K is the number of clusters, if there exists some R such that for any i*

$$\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{x}_{M(i)})<R,$$

and for $\forall i\in\mathcal{S}_k,\forall j\in\mathcal{S}_l$ with $k\neq l$,

$$\mathbb{D}(\boldsymbol{x}_i,\boldsymbol{x}_j)>2R,$$

then the optimal solutions of (2) and (3) are identical.

Note that, according to condition (7), a clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ with larger extent of separation corresponds to a larger range of λ. Solving (3) with this range of λ leads to the solution $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$. One can further show that $\|W^*\|_{\infty,1}$ is a non-increasing function of λ, and solutions with same $\|W^*\|_{\infty,1}$ must be the same. We have following proposition independent of Theorem 1.

Proposition 1. *For λ_1 and λ_2 with $\lambda_1<\lambda_2$ and their corresponding (unique) optimal solutions $W_1^*:=W^*(\lambda_1)$ and $W_2^*:=W^*(\lambda_2)$, we have $\|W_1^*\|_{\infty,1}\geq\|W_2^*\|_{\infty,1}$, and if $\|W_1^*\|_{\infty,1}=\|W_2^*\|_{\infty,1}$, then $W_1^*=W_2^*$, and $W^*(\lambda)=W_1^*$ for any $\lambda\in[\lambda_1,\lambda_2]$.*

For HDP mixture problem (6), the analysis becomes more involved, since the separation condition should be different for samples coming from same data set and that from different data sets.

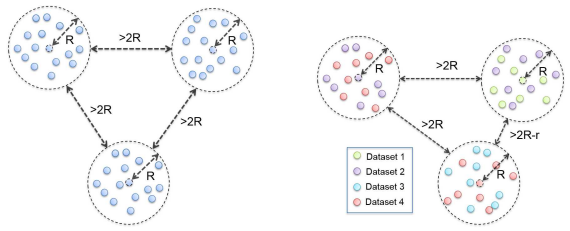


Figure 1. Examples of clustering described in Corollary 1 (left) and Corollary 2 (right).

Theorem 2. *Given a clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$, denote $\mathcal{S}_{k,d} = \mathcal{S}_k \cap \mathcal{T}_d$, $\mathcal{D}_k = \{d|\mathcal{S}_{k,d} \neq \phi\}$, and $d(i)$ the index of data set i belonging to. If we can find λ, θ such that, for $\forall i, j \in \mathcal{S}_k, \forall k \in \mathcal{M}$,*

$$\frac{\lambda}{N_k} + \frac{\theta}{N_{k,d(i)}} > \delta_{ij} \tag{8}$$

$$\frac{\lambda}{N_k} > \frac{1}{N_{k,d}} \sum_{i\in\mathcal{S}_{k,d}} \delta_{ij},\ \forall d \in \mathcal{D}_k \tag{9}$$

and for $\forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_{l\neq k}$,

$$\frac{\lambda}{N_k} + \frac{\theta}{N_{k,d(i)}} < \delta_{ij},\ \mathcal{D}_k \cap \mathcal{D}_l \neq \phi \tag{10}$$

$$\frac{\lambda}{N_k} + \theta\left(\frac{1}{N_{k,d(i)}} - \frac{1}{N_{d(i)}}\right) < \delta_{ij},\ \mathcal{D}_k \cap \mathcal{D}_l = \phi \tag{11}$$

, where $N_k = |\mathcal{S}_k|$, $N_d = |\mathcal{T}_d|$, and $N_{k,d} = |\mathcal{S}_k \cap \mathcal{T}_d|$, then the integer solution W^* realizing $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ is unique optimal solution to both (5) and (6).

For $\theta = 0$, (10) and (11) are equivalent to the RHS of (7), while (8) gives the LHS of (7) and implies (9), which then gives the same condition as in Theorem 1. For $\theta > 0$, inequality (11) requires less dissimilarity than that of inequality (10), which means the inter-cluster dissimilarity can be smaller if two clusters do not share samples from the same dataset. By tuning θ, the convex HDP formulation (6) can utilize structure of datasets to realize clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ not realizable by the DP formulation (3). For $\mathbb{D}(.,.)$ being a metric, we can further simplify result as follows.

Corollary 2. *Assuming $\mathbb{D}(\cdot,\cdot)$ is a metric, $N_k = N/K$, $|\mathcal{D}_k| = C$ is a constant, $N_{k,d} = N/(CK)$ for all $d \in \mathcal{D}_k$ and $K_d = CK/D$, then if there exists some R such that (a) for $\forall i$,*

$$\mathbb{D}(\boldsymbol{x}_i, \boldsymbol{x}_{M(i)}) < R,$$

(b) for $\forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l$ with $k \neq l$ and $\mathcal{D}_k \cap \mathcal{D}_l \neq \phi$,

$$\mathbb{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) > 2R,$$

(c) for $\forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l$ with $k \neq l$ and $\mathcal{D}_k \cap \mathcal{D}_l = \phi$,

$$\mathbb{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) > 2R - r,$$

Algorithm 1 ADMM for exemplar-based HDP mixture (6)

Initialize $t = 0$, $W^{(t)} = Y^{(t)} = Z^{(t)} \leftarrow 0$.
repeat
1. Solve each row of $W_1^{(t+1)}$ in (12) via Frank-Wolfe algorithm.
2. Solve each column of $W_2^{(t+1)}$ in (13) via proximal mapping.
3. $Z^{(t+1)} = (W_1^{(t+1)} + W_2^{(t+1)})/2$.
4. $Y_q^{(t+1)} \leftarrow Y_q^{(t)} + \alpha(W_q^{(t+1)} - Z^{(t+1)})$, for $q = 1, 2$.
5. $t = t + 1$.

until $\|Z^{(t)} - Z^{(t-1)}\| < \epsilon_1$ and $\|W_1^{(t)} - W_2^{(t)}\| < \epsilon_2$

where

$$r = \frac{1}{K_d}\left(R - \max_{k\in\mathcal{M}, j\in\mathcal{S}_k, d\in\mathcal{D}_k} \frac{1}{N_{k,d}} \sum_{i\in\mathcal{S}_{k,d}} \delta_{ij}\right)$$

then the optimal solutions of (5) and (6) are identical.

Figure 1 gives two examples that compare conditions of Corollary 1 and 2. For HDP, we also show in the following that $\|W^*\|_{\infty,1}$, $\|W^*\|_{\mathcal{G}}$ are monotonically decreasing with λ, θ respectively, and the optimal solutions in some triangular area of the (λ, θ) parameter space are the same if the corners of the triangular area lead to the same values of $\|W^*\|_{\infty,1}$, $\|W^*\|_{\mathcal{G}}$.

Proposition 2. *$\|W^*\|_{\infty,1}$, $\|W^*\|_{\mathcal{G}}$ are monotonically decreasing with λ, θ respectively, and for regularization parameters (λ_1, θ_1), (λ_1, θ_2) and (λ_2, θ_2) with $\lambda_1 < \lambda_2$ and $\theta_1 < \theta_2$, define their corresponding optimal solutions $W_1^* := W^*(\lambda_1, \theta_1)$, $W_{12}^* := W^*(\lambda_1, \theta_2)$ and $W_2^* := W^*(\lambda_2, \theta_2)$. Assuming unique optimal solution, then if*

$$\|W_1^*\|_{\mathcal{G}} = \|W_{12}^*\|_{\mathcal{G}} \text{ and } \|W_2^*\|_{\infty,1} = \|W_{12}^*\|_{\infty,1},$$

we have $W_1^* = W_{12}^* = W_2^*$ and furthermore for any

$$(\lambda, \theta) \in Conv\left((\lambda_1, \theta_1), (\lambda_1, \theta_2), (\lambda_2, \theta_2)\right)$$

the corresponding optimal solution

$$W^*(\lambda, \theta) = W_1^*,$$

where $Conv(\cdot)$ is the convex hull function.

4. Algorithm

In this section, we propose an algorithm based on ADMM (Alternating Direction Method of Multiplier) to solve the convex formulation of exemplar-based HDP mixture (6). The exemplar-based DP mixture (3) can be solved via the same algorithm by setting $\theta = 0$ and ignoring any group structure imposed by data sets. Each iteration of the

ADMM algorithm decomposes (6) into two sub-problems with augmented terms:

$$W_1^{(t+1)} = \underset{W\in\mathbb{R}_+^{N\times J}:W\mathbf{1}=\mathbf{1}}{argmin} \quad \mathbb{D}\circ W + Y_1^{(t)}\circ W + \frac{\rho}{2}\|W - Z^{(t)}\|^2 \tag{12}$$

and

$$W_2^{(t+1)} = \underset{W\in\mathbb{R}_+^{N\times J}}{argmin} \quad \theta\|W\|_{\mathcal{G}} + \lambda\|W\|_{\infty,1} + Y_2^{(t)}\circ W + \frac{\rho}{2}\|W - Z^{(t)}\|^2 \tag{13}$$

where $Z^{(t)}$ is a global consensus variable, and $Y_1^{(t)}$, $Y_2^{(t)}$ are dual variables for constraints $W_1 = Z$, $W_2 = Z$ respectively [2]. The two sub-problems are relatively easy to solve, since each row of W in (12) can be solved independently, and so as each column of W in (13). For each row of W in (12), we solve with a Frank-Wolfe algorithm specialized for simplex constraint (Jaggi, 2013), while for each column of W in (13), we solve with a closed-form proximal mapping. Since each group $g \in \mathcal{G}$ is a subset of a column of W, the proximal mapping $\mathbf{prox}(.)$ for each column of (13) can be decomposed as $\mathbf{prox}(.) = \mathbf{prox}_\lambda(\mathbf{prox}_\theta(.))$, where $\mathbf{prox}_\lambda(.)$ is the proximal-mapping for $\lambda\|W\|_{\infty,1}$ and $\mathbf{prox}_\theta(.)$ is the that for $\theta\|W\|_{\mathcal{G}}$, both of which can be computed in closed form as in (Liu et al., 2009). The overall algorithm 1 has following convergence guarantee.

Theorem 3. *Let $f(W)$ denotes the objective function of (6). Provided that α is sufficiently small, the sequence $\{f(W^{(t)})\}_{t=0}^{\infty}$ produced by Algorithm 1 converges to optimum f^* at a linear rate. In other words,*

$$t \geq c_1 \log(c_2/\epsilon)$$

iterations suffice to guarantee $f(W^{(t)}) - f^* \leq \epsilon$ and ϵ-infeasibility for some constants c_1, c_2 independent of t.

Since $\mathbb{D}\circ W$ and $\|W\|_{\infty,1}, \|W\|_{\mathcal{G}}$ are all *polyhedral functions*, Theorem 3 simply follows from Theorem 3.1 of (Hong & Luo, 2012).

5. Practical Issues

In this section, we discusses some implementation details and practical issues.

5.1. Solution Optimality and Perturbation

Whenever an integer optimal solution W^* is obtained from the convex relaxation, it is also optimal to the combinatorial problem. However, if there are multiple optimal solutions $W_1^*, W_2^*, ...W_M^*$ in the original problem (5), any fractional convex combination of them would also be optimal in the convex relaxation (6). One simple approach resolving this issue is to add a small perturbation to the matrix $\mathbb{D}$. Since in the combinatorial problem (5), function difference between optimal and sub-optimal solutions can be lower bounded by a finite constant, there exists small enough perturbation that makes one of the solutions become uniquely optimal.

5.2. Shrinking

Since many columns of W become 0 when iterate $W^{(t)}$ becomes close to optimum. We employ a *Shrinking* technique to shrink inactive columns of W that have primal and dual residual both equal, or close, to 0, that is, we shrink column j if

$$\|Z_{i,j}^{(t)} - Z_{i,j}^{(t-1)}\| \approx 0 \quad and \quad \|W_{1i,j}^{(t)} - W_{2i,j}^{(t)}\| \approx 0, \forall i.$$

The shrinked column will be set to 0 and never updated until stopping condition holds for all the other columns. Then we will check optimality of those shrinked columns and re-iterates if optimality does not hold. In practice, the technique reduces complexity for most iterations of Algorithm 1 from $O(N^2)$ to $O(NK^+)$, where $K^+ \ll N$ is the number of active columns.

5.3. Solving Subproblem Inexactly

The convergence of the ADMM algorithm does not require solving each subproblem exactly (Boyd et al., 2011; Hong & Luo, 2012). In practice, solving subproblem (12) by only a few iterations of Frank-Wolfe algorithm results in faster convergence. In our experiment, we fix maximum number of Frank-Wolfe iterations to 30.

5.4. Model Selection

One arguable advantage of Bayesian Nonparametrics is not fixing number of clusters K before learning, which however, comes with additional parameters λ. In practice, whether one way is better than the other depends on the prior knowledge one has. For some applications, the desired number of clusters is known a priori, and thus fixing K before training would be preferred. However, for situations where HDP applies, one can hardly expect how many local clusters each data set should have, and a single parameter θ, that encodes the additional loss reduction one should achieve to create a new local cluster, is preferred. As stated in Section 3, the parameters that result in the same optimal solution W^* form a convex region of parameters (a band of λ in case of DP mixture), and the larger extent of separation a clustering $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$ achieves in Theorem 1, 2, the larger the convex region. In practice, a clustering with better separation can be found in a wider range of parameters.

[2] More details about deriving the ADMM sub-problems can be found in, for example, (Boyd et al., 2011).

6. Experiment

In this section, we compare the convex, exemplar-based approach with existing approaches to MAD-Bayes DP mixture (1) and HDP mixture (4) on several public available data sets. We use square Euclidean distance as measure of dissimilarity $\mathbb{D}(.,.)$ so the compared algorithms are applicable (Kulis & Jordan, 2012; Jiang et al., 2012). The algorithms in comparison are listed as follows.

- *DP-means* exactly implements the k-means like algorithm proposed in (Kulis & Jordan, 2012) for the DP objective (1). Since the algorithm gives different result for different order of updating $\{z_i^{(t)}\}_{i=1}^N$, we run the algorithm for 1000 rounds with random permutation on the updating in order to achieve better local optimum. Global mean is used as initialization as specified in (Kulis & Jordan, 2012).
- *DP-medoids* applies the *DP-means* algorithm to the exemplar-based objective function (2). Since (2) is a special case of (1) that restricts the space of $\boldsymbol{\mu}_k$ to a set of exemplar $\mathcal{E}$, the resulting algorithm simply replaces the step of computing mean of cluster $\mathcal{S}_k$ by picking exemplar $\boldsymbol{\mu}_j \in \mathcal{E}$ of smallest dissimilarity to points in $\mathcal{S}_k$.
- *DP-convex* solves (3) using Algorithm 1 with $\theta = 0$.
- *DP-convex (means)* takes clustering assignments obtained from *DP-convex*, and compute *means* of each cluster $\mathcal{S}_k$ as the representative for $i \in \mathcal{S}_k$.
- *HDP-means* exactly implements the algorithm proposed in (Kulis & Jordan, 2012) for the HDP objective (4). As in *DP-means*, we run the algorithm for 1000 rounds with different random permutation of updating order to achieve better local optimum. Global mean is used as initialization as specified in (Kulis & Jordan, 2012).
- *HDP-medoids* applies the *HDP-means* algorithm to the exemplar-based objective function (5). Similar to *DP-medoids*, the algorithm simply replaces the step of computing mean w.r.t. a set of points $\mathcal{S}$ by picking exemplar $\boldsymbol{\mu}_j \in \mathcal{E}$ of smallest dissimilarity to points in $\mathcal{S}$.
- *HDP-convex* solves (6) using Algorithm 1.
- *HDP-convex (means)* takes clustering assignments obtained from *HDP-convex*, and compute *means* of each global cluster $\mathcal{S}_k$ as the representative for $i \in \mathcal{S}_k$.

Our experiments for DP mixture model are conducted on 5 publicly available data sets: *Iris*, *Glass*, *Wine*, *DNA* and *Segment*. For HDP mixture model, we experiment on *Wholesale* and *Water* data sets. The *Wholesale* data comprises the spending of customers from different regions and channels. We divide samples in *Wholesale* into 6 groups $\mathcal{T}_1, ..., \mathcal{T}_6$ based on region and channel. The *Water* data contains the daily measures of sensors in a urban waste water treatment plant. The data set is divided into 21 groups according to the month. For each group, we expect a smaller number of clusters than that appear across all groups. All of the data sets can be downloaded from UCI Machine Learning Repository

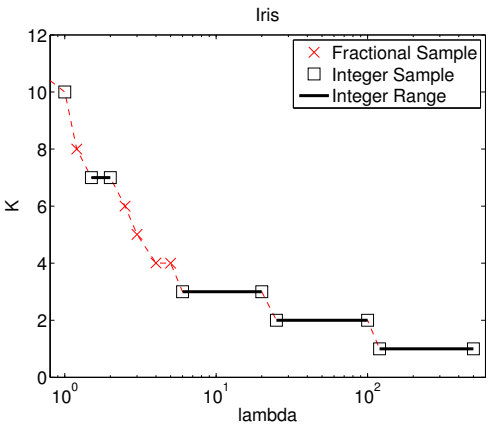
Table 1. Data sets employed in our experiments and their relevant statistics.

Dataset	N	D	Mean-Dist	Std-Dist
Iris	150	4	2.1940	2.1475
Wine	178	13	4.2966	2.4209
Glass	214	9	2.1445	2.2049
DNA	2000	180	67.1564	7.2841
Segment	2310	19	6.3195	5.5270
Wholesale	440	6	0.4220	0.3444
Water	527	38	1.6073	1.5893

Figure 2 and 3 illustrate the sampled regularization path obtained from *DP-convex* on five data sets. For each data set, we observed several bands of λ that give integer (and thus optimal) solutions. Each band corresponds to a clustering assignment $\{\mathcal{S}_k\}_{k\in\mathcal{M}}$. Note there are several integer solutions marked by square mark '□' that has a band shorter than our sampled interval, which corresponds to clustering with small extent of separation and is usually not preferred in model selection. In Figure 4, we present the number of global clusters K_g and total number of local clusters K_l obtained from *HDP-convex* under different (λ, θ) pairs of parameter. Note for fractional solution, $K_g = \|W^*\|_{\infty,1}$ and $K_l = \|W^*\|_{\mathcal{G}}$, and integer solutions are marked with '□'.

In Table 2, we pick λ within a band in Figure 3 and Figure 2 for each data set, and compare the objective obtained from convex approach to existing approaches (Kulis & Jordan, 2012; Jiang et al., 2012) with 1000 random re-trials. Note that, for the same clustering, using *mean* as representative should always get lower objective than using *medoid* as representative since *mean* is the minimizer of square Euclidean distance. However, the result shows that the optimal *medoids* obtained from convex approach still achieves significantly lower objective than clustering found by local search method.

In addition, we observe that the number of clusters obtained from local methods are consistently smaller. The reason behind this is, the *DP-means* algorithm proposed in (Kulis & Jordan, 2012; Jiang et al., 2012) creates a new cluster



(a) Iris

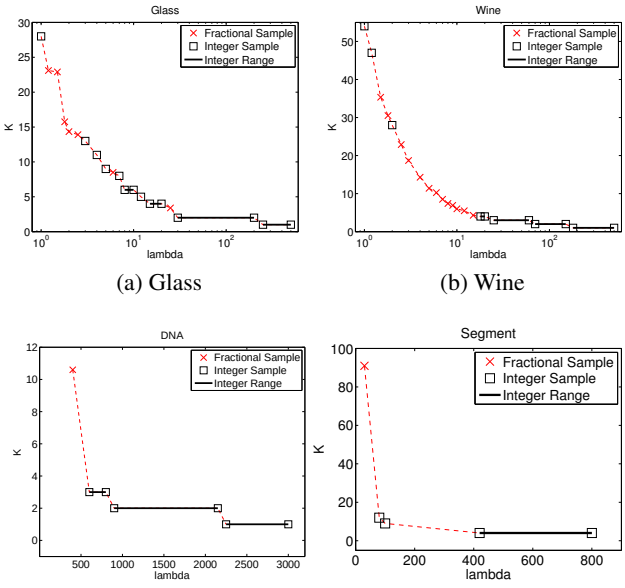
Figure 2. Number of clusters $K = \|W^*\|_{\infty,1}$ vs. λ obtained from DP-convex, where each $\square$ denotes an integer solution and a horizontal line between contiguous $\square$ shows band of λ yielding the same integer solution.

only when a single observation has square distance to its representative more than λ. However, it is often the case that only moving several points together to a new cluster can make a loss reduction more than λ. In that case, *DP-means* can easily get stuck at local minimum. For the data set *Wine*, the algorithm gets stuck even at initialization.

Figure 5 shows the lowest objective achieved by different methods over time. Generally, a single run of *DP-means*, *HDP-means* or their exemplar-based version converges much faster than the ADMM algorithm. However, even with large number of random re-trials, the objective achieved by local search methods could hard improve significantly. For *Wine* and *Glass* data set, all random retrials of *DP-means* and *DP-medoids* get stuck at the same local minimum.

Conclusion. In this paper, we show that the MAD-Bayes formulation for DP and HDP mixture models can be naturally relaxed to a convex domain in the exemplar-based setting, which is tight under cluster-separation condition. Similar convex formulation can be derived for more interesting applications such as Latent Feature Allocation (Broderick et al., 2013), Infinite Hidden Markov Models (Roychowdhury et al., 2013), and applications where exemplar comes in naturally, such as Multiple Sequence Alignment and Motif Finding. This will be an active research direction for us in the near future.

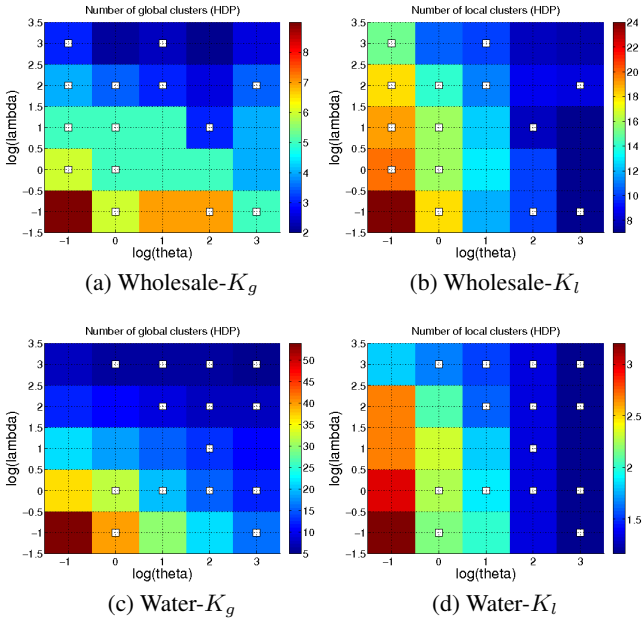
Acknowledgement. P. R. acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, IIS-1447574, and DMS-1264033. I. S. D. acknowledges the support of NSF via grants CCF-1320746 and CCF-1117055.



(a) Glass (b) Wine

(c) DNA (d) Segment

Figure 3. More results of K v.s. λ obtained from DP-convex program. Note that each $\square$ denotes an integer solution and a horizontal line between contiguous $\square$ shows band of λ yielding the same integer solution.



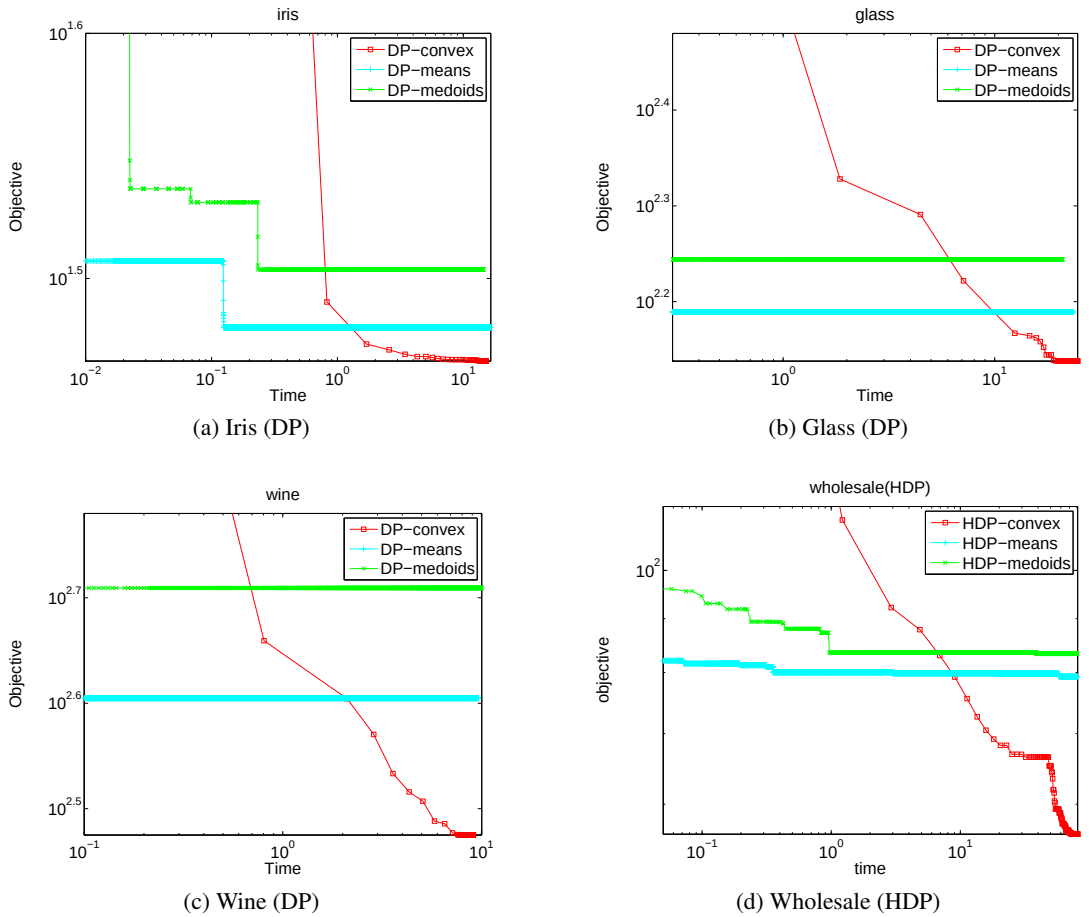
(a) Wholesale-K_g (b) Wholesale-K_l

(c) Water-K_g (d) Water-K_l

Figure 4. Number of global and local clusters obtained from HDP-convex, where each '$\square$' denotes an integer solution. (a) $K_g = \|W^*\|_{\infty,1}$, the number of global clusters in HDP. (b) $K_l = \|W^*\|_{\mathcal{G}}$, the total number of local clusters for HDP.

Table 2. Objective value obtained from different methods, where *DP-convex (means)* uses mean instead of medoid as representative of each cluster $\mathcal{S}_k$ obtained from W^*.

Data set	DP-convex	DP-convex (means)	DP-medoids	DP-means
Iris ($\lambda = 2$)	29.26 (K=7)	*27.97* (K=7)	35.68 (K=3)	30.20 (K=4)
Glass ($\lambda = 9$)	137.40 (K=6)	*128.13* (K=6)	175.42 (K=2)	154.66 (K=2)
Wine ($\lambda = 20$)	298.55 (K=4)	*263.79* (K=4)	512.04 (K=1)	402.40 (K=1)
DNA ($\lambda = 1000$)	105947.0 (K=2)	68718.9 (K=2)	107211(K=1)	68156.4 (K=1)
Segment ($\lambda = 600$)	4749.62 (K=4)	4572.3 (K=4)	8405.71(K=1)	7898.98 (K=1)
Data set	HDP-convex	HDP-convex (means)	HDP-medoids	HDP-means
Wholesale ($\lambda = 1.0, \theta = 1.0$)	56.28 (K_g=5, K_l=16)	*53.07* (K_g=5, K_l=16)	83.35 (K_g = 2, K_l = 6)	79.52 (K_g=2, K_l = 6)
Water ($\lambda = 1.0, \theta = 1.0$)	244.59 (K_g=32, K_l=81)	232.07 (K_g=34, K_l=83)	256.41 (K_g=33, K_l=74)	237.73 (K_g=37, K_l=64)



(a) Iris (DP)

(b) Glass (DP)

(c) Wine (DP)

(d) Wholesale (HDP)

Figure 5. Objective v.s. Time (in seconds). (a)-(c) show the lowest objective achieved by DP-convex, DP-means and DP-medoids over time on three data sets: (a) Iris. (b) Glass (c) Wine. Note we run DP-means and DP-medoids for 1000 rounds with random permutation on the updating order. (d) Running Time Comparison between HDP-convex, HDP-means and HDP-medoids on the Wholesale data set.

References

Awasthi, Pranjal, Bandeira, Afonso S, Charikar, Moses, Krishnaswamy, Ravishankar, Villar, Soledad, and Ward, Rachel. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 191–200. ACM, 2015.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Broderick, Tamara, Kulis, Brian, and Jordan, Michael I. Mad-bayes: Map-based asymptotic derivations from bayes. *ICML*, 2013.

Campbell, Trevor, Liu, Miao, Kulis, Brian, How, Jonathan P, and Carin, Lawrence. Dynamic clustering via asymptotics of the dependent dirichlet process mixture. In *Advances in Neural Information Processing Systems*, pp. 449–457, 2013.

Elhamifar, Ehsan, Sapiro, Guillermo, and Vidal, Rene. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in Neural Information Processing Systems*, pp. 19–27, 2012.

Frey, Brendan J and Dueck, Delbert. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

Hong, Mingyi and Luo, Zhi-Quan. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 427–435, 2013.

Jiang, Ke, Kulis, Brian, and Jordan, Michael I. Small-variance asymptotics for exponential family dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pp. 3158–3166, 2012.

Kaufman, Leonard and Rousseeuw, Peter. *Clustering by means of medoids*. North-Holland, 1987.

Kulis, Brian and Jordan, Michael I. Revisiting k-means: New algorithms via bayesian nonparametrics. *ICML*, 2012.

Liu, Han, Palatucci, Mark, and Zhang, Jian. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 649–656. ACM, 2009.

Megiddo, Nimrod and Supowit, Kenneth J. On the complexity of some common geometric location problems. *SIAM journal on computing*, 13(1):182–196, 1984.

Nellore, Abhinav and Ward, Rachel. Recovery guarantees for exemplar-based clustering. *arXiv preprint arXiv:1309.3256*, 2013.

Obozinski, Guillaume, Jacob, Laurent, and Vert, Jean-Philippe. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

Papadimitriou, Christos H. Worst-case and probabilistic analysis of a geometric location problem. *SIAM Journal on Computing*, 10(3):542–557, 1981.

Roychowdhury, Anirban, Jiang, Ke, and Kulis, Brian. Small-variance asymptotics for hidden markov models. In *Advances in Neural Information Processing Systems*, pp. 2103–2111, 2013.

Van der Laan, Mark, Pollard, Katherine, and Bryan, Jennifer. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73 (8):575–584, 2003.

Yaron, Margalit and Shalev-Shwartz, Shai. A convex relaxation to solve the k-medoids problem. *Hebrew University*, 2010.

A. Notation

$\mathcal{N} = \{1, 2, ..., N\} =: [N]$ is the whole set of data points. $i, j \in \mathcal{N}$ denote points. $d_{ij} := \mathbb{D}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. D is the number of data sets. $\mathcal{T}_d \subseteq \mathcal{N}$ denotes the set of points in the d-th dataset, i.e. $\cup_{d=1}^{D} \mathcal{T}_d = \mathcal{N}$. $N_d = |\mathcal{T}_d|$ is the number of points in Dataset d. $d(i) \in [D]$ denotes the dataset index of Point i. $\mathcal{M} \subseteq \mathcal{N}$ is the set of medoids. $k, l \in \mathcal{M}$ denote clusters and themselves are medoids. $\mathcal{S}_k$ is the set of points in Cluster k. $N_k = |\mathcal{S}_k|$ is the number of points in Cluster k. $M(i) \in \mathcal{M}$ denotes the cluster/representative of Point i. Let $\mathcal{D}_k \subseteq [D]$ denote the data sets contained or partially contained in Cluster k. Denote $\mathcal{S}_{k,d} := \mathcal{S}_k \cap \mathcal{T}_d$ for $d \in \mathcal{D}_k$. Thus $\cup_{d \in \mathcal{D}_k} \mathcal{S}_{k,d} = \mathcal{S}_k$. Denote $N_{k,d} := |\mathcal{S}_{k,d}|$ for $d \in \mathcal{D}_k$.

B. Proof of Theorem 1

Theorem 1 is a direct corollary of Theorem 2, by setting $\theta = 0$.

C. Proof of Theorem 2

First, the convex program (6) has same set of optimal solutions with the following linear program

$$\begin{aligned} \min_{w_{ij} \geq 0, \zeta_{d,j}, \xi_j} \quad & \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} w_{ij} + \theta \sum_{d=1}^{D} \sum_{j=1}^{N} \zeta_{d,j} + \lambda \sum_{j=1}^{N} \xi_j \\ s.t. \quad & \sum_{j=1}^{N} w_{ij} = 1 \\ & w_{ij} \leq \zeta_{d,j}, \ \forall i \in \mathcal{T}_d \\ & w_{ij} \leq \xi_j, \ \forall i \in [N]. \end{aligned} \tag{14}$$

The KKT condition of the linear programming can be written as

$$d_{ij} - \alpha_{ij} - \beta_i + \gamma_{ij} + \delta_{ij} = 0 \tag{15}$$

$$\theta = \sum_{i \in \mathcal{T}_d} \delta_{ij} \tag{16}$$

$$\lambda = \sum_{i} \gamma_{ij} \tag{17}$$

$$\delta_{ij}(w_{ij} - \zeta_{dj}) = 0 \tag{18}$$

$$\gamma_{ij}(w_{ij} - \xi_j) = 0 \tag{19}$$

$$\alpha_{ij} w_{ij} = 0 \tag{20}$$

$$\alpha_{ij} \geq 0 \tag{21}$$

$$\gamma_{ij} \geq 0 \tag{22}$$

$$\delta_{ij} \geq 0. \tag{23}$$

Our goal is to find a structure of d_{ij}, for which there exists a set of α_{ij}, β_i, γ_{ij}, δ_{ij}, θ and λ satisfying the above conditions (with $\alpha_{ij}, \gamma_{ij}, \delta_{ij}$ strictly positive for binding constraints). Then a clustering $\{\mathcal{S}_k\}_{k \in \mathcal{M}}$ with such structure will be an unique solution to (14). We will discuss the cases entry-by-entry.

C.1. $\xi_j = 1, \ \zeta_{dj} = 1, \ w_{ij} = 1$

$j = M(i), \alpha_{i,M(i)} = 0$

$$\gamma_{i,M(i)} + \delta_{i,M(i)} = \beta_i - d_{i,M(i)}, \quad \forall i \tag{24}$$

C.2. $\xi_j = 1, \ \zeta_{dj} = 1, \ w_{ij} = 0$

$j \in \mathcal{M}$, but $j \neq M(i)$
$\delta_{ij} = 0, \gamma_{ij} = 0 \Rightarrow \alpha_{ij} = d_{ij} - \beta_i > 0$, i.e.,

$$\beta_i < d_{ij}, \quad \forall j \in \mathcal{M} \text{ but } j \neq M(i) \text{ and } \mathcal{D}_j \cap \mathcal{D}_{M(i)} \neq \phi. \tag{25}$$

Summary of Section C.1 and C.2
We can set $\gamma_{i,M(i)} = \frac{\lambda}{N_{M(i)}}$ such that Eq. (17) holds and $\delta_{i,M(i)} = \frac{\theta}{N_{M(i),d(i)}}$ such that Eq. (16) holds. Thus,

$$\beta_i = \frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} \tag{26}$$

C.3. $\xi_j = 1,\ \zeta_{dj} = 0,\ w_{ij} = 0$

$j \in \mathcal{M}$, but $j \neq M(i)$
$\gamma_{ij} = 0 \Rightarrow \alpha_{ij} = d_{ij} - \beta_i + \delta_{ij} > 0$. Now we have

$$\begin{aligned} \delta_{ij} &> \beta_i - d_{ij} \\ \delta_{ij} &> 0 \end{aligned}$$

Thus

$$\theta = \sum_{i \in \mathcal{T}_d} \delta_{ij} > \sum_{i \in \mathcal{T}_d} (\beta_i - d_{ij})_+, \quad \forall d \notin \mathcal{D}_j, j \in \mathcal{M} \tag{27}$$

If we set $\beta_i - d_{ij} < \frac{\theta}{N_{d(i)}}$, Eq. (27) will be satisfied. That is

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} - d_{ij} < \frac{\theta}{N_{d(i)}} \tag{28}$$

C.4. $\xi_j = 0,\ \zeta_{dj} = 0,\ w_{ij} = 0$

In this case, we have $\alpha_{ij} = d_{ij} - \beta_i + \delta_{ij} + \gamma_{ij} > 0$, that is,

$$\gamma_{ij} > \beta_i - d_{ij} - \delta_{ij} \tag{29}$$

$$\lambda = \sum_i \gamma_{ij} > \sum_i (\beta_i - d_{ij} - \delta_{ij})_+, \quad \forall j \notin \mathcal{M} \tag{30}$$

$$\theta = \sum_{i \in \mathcal{T}_d} \delta_{ij}, \quad \forall d \in [D], \forall j \notin \mathcal{M} \tag{31}$$

To analyze this case, we divide $i \in [N]$ into three parts. The first part is the points in the same cluster as j denoted by $S_{M(j)}$. The second part is the points who have sister points (sister points mean they belong to the same dataset) in $S_{M(j)}$ but themselves are not in $S_{M(j)}$, denoted by $S_{1,M(j)} := \left(\cup_{d \in \mathcal{D}_{M(j)}} \mathcal{T}_d\right) \setminus S_{M(j)}$. The third part is all the points who don't have sister points in $S_{M(j)}$, denoted by $S_{2,M(j)} := \cup_{d \in [D] \setminus \mathcal{D}_{M(j)}} \mathcal{T}_d$

$$\begin{aligned} \lambda > & \sum_{i \in S_{M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \\ & + \sum_{i \in S_{1,M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \\ & + \sum_{i \in S_{2,M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \end{aligned} \tag{32}$$

In the following we will show our strategy to make this inequality hold.
If we set δ_{ij} to be

$$\theta = \left(\sum_{i \in \mathcal{S}_{M(j),d}} \delta_{ij} \right), \quad \forall d \in \mathcal{D}_{M(j)}, \forall j \notin \mathcal{M} \tag{33}$$

$$\delta_{ij} = 0, \quad \forall i \in \mathcal{S}_{1,M(j)}, \forall j \notin \mathcal{M} \tag{34}$$

$$\delta_{ij} = \frac{\theta}{N_{d(i)}}, \quad \forall i \in \mathcal{S}_{2,M(j)}, \forall j \notin \mathcal{M} \tag{35}$$

such that Eq. (16) is satisfied.
Further more, if we can get the following equations satisfied,

$$\begin{aligned} &\beta_i - d_{ij} - \delta_{ij} \geq 0,\ \forall i \in \mathcal{S}_{M(j)} \\ &\beta_i - d_{ij} - \delta_{ij} < 0,\ \forall i \in \mathcal{S}_{1,M(j)} \\ &\beta_i - d_{ij} - \delta_{ij} < 0,\ \forall i \in \mathcal{S}_{2,M(j)} \end{aligned} \tag{36}$$

the only thing we need to show is

$$\begin{aligned} \lambda &> \sum_{i \in S_{M(j)}} (\beta_i - d_{ij} - \delta_{ij}) \\ &= \sum_{i \in S_{M(j)}} (\frac{\lambda}{N_{M(i)}} + d_{i,M(i)} - d_{ij}) \end{aligned}$$

It is equivalent to

$$\sum_{i \in S_{M(j)}} d_{i,M(i)} < \sum_{i \in S_{M(j)}} d_{ij},$$

which is satisfied by medoid definition.
In the following, we analyze the conditions under which the three inequalities of Eq. (36) hold.

First part $i \in S_{M(j)}$ In this part we try to let $\beta_i - d_{ij} - \delta_{ij} \geq 0$. As $\delta_{ij} > 0$, we require

$$\beta_i - d_{ij} > 0, \quad \forall i \in S_{M(j)}$$

That is, for $\forall i, j$ s.t. $M(i) = M(j)$,

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} > d_{ij}, \tag{37}$$

Then we can always find a δ_{ij} such that $0 < \delta_{ij} < \beta_i - d_{ij}$. To satisfy Eq. (33), we require

$$\theta < \sum_{i \in \mathcal{S}_{k,d}} \beta_i - d_{ij}, \quad \forall d \in \mathcal{D}_k, k = M(j)$$

Equivalently, we have

$$\lambda > \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} d_{ij} - d_{i,M(i)},\ \forall d \in \mathcal{D}_k, \forall j \in \mathcal{S}_k, \forall k \tag{38}$$

Second part $i \in S_{1,M(j)}$ As set in Eq. (34), $\delta_{ij} = 0$, we require

$$\beta_i - d_{ij} < 0,\ \forall i \in \mathcal{S}_{1,M(j)}$$

That is, for $\forall i, j$ s.t. $\mathcal{D}_{M(i)} \cap \mathcal{D}_{M(j)} \neq \phi$ and $M(i) \neq M(j)$

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} < d_{ij}. \tag{39}$$

This requirement also implies Eq. (25) will hold.

Third part $i \in S_{2,M(j)}$ For this part,

$$\beta_i - d_{ij} < \frac{\theta}{N_{d(i)}},\ \forall i \in \mathcal{S}_{2,M(j)}$$

That is, for $\forall i, j$ s.t. $\mathcal{D}_{M(i)} \cap \mathcal{D}_{M(j)} = \phi$,

$$\frac{\lambda}{N_{M(i)}} + \theta \left(\frac{1}{N_{M(i),d(i)}} - \frac{1}{N_{d(i)}} \right) + d_{i,M(i)} < d_{ij}, \tag{40}$$

This requirement also implies Eq. (28) will hold. □

D. Proof of Proposition 1

Given the conditions in the proposition, we have

$$\begin{aligned} & \mathbb{D} \circ W_1^* + \lambda_1 \|W_1^*\|_{\infty,1} \\ \leq\ & \mathbb{D} \circ W_2^* + \lambda_1 \|W_2^*\|_{\infty,1} \\ <\ & \mathbb{D} \circ W_2^* + \lambda_2 \|W_2^*\|_{\infty,1} \\ \leq\ & \mathbb{D} \circ W_1^* + \lambda_2 \|W_1^*\|_{\infty,1} \end{aligned} \tag{41}$$

So we have

$$\begin{aligned} \mathbb{D} \circ W_1^* \leq \mathbb{D} \circ W_2^* \\ \mathbb{D} \circ W_2^* \leq \mathbb{D} \circ W_1^* \end{aligned}$$

And under the unique optimum assumption, we have $W_1^* = W_2^*$.
For the rest of the proof, we first prove that $\|W^*(\lambda)\|_{\infty,1}$ is a non-increasing function. From Eq. (41),

$$\lambda_2 \|W_2^*\|_{\infty,1} - \lambda_1 \|W_2^*\|_{\infty,1} \leq \lambda_2 \|W_1^*\|_{\infty,1} - \lambda_1 \|W_1^*\|_{\infty,1}$$

that is,

$$(\lambda_2 - \lambda_1)(\|W_2^*\|_{\infty,1} - \|W_1^*\|_{\infty,1}) \leq 0$$

Therefore, for any $\lambda_1 < \lambda_2$, we have $\|W_2^*\|_{\infty,1} \leq \|W_1^*\|_{\infty,1}$. Now for any $\lambda \in [\lambda_1, \lambda_2]$, because $\|W^*(\lambda_1)\|_{\infty,1} = \|W^*(\lambda_2)\|_{\infty,1}$, we have $\|W^*(\lambda)\|_{\infty,1} = \|W_1^*\|_{\infty,1}$, and further under the unique optimum assumption,

$$W^*(\lambda) = W_1^*$$

□

E. Proof of Proposition 2

According to Proposition 1, given $\|W_1^*\|_{\mathcal{G}} = \|W_{12}^*\|_{\mathcal{G}}$, we have $W_1^* = W_{12}^*$ and for any $\theta \in [\theta_1, \theta_2]$, $W^*(\lambda_1, \theta) = W_{12}^*$.
Given $\|W_2^*\|_{\infty,1} = \|W_{12}^*\|_{\infty,1}$, we have, for any $\lambda \in [\lambda_1, \lambda_2]$, $W^*(\lambda, \theta_2) = W_{12}^*$.
Now we prove for any (λ, θ) on the line between point (λ_1, θ_1) and point (λ_2, θ_2) (defined by $\mathcal{L}_{12}$), $W^*(\lambda, \theta) = W_1^*$. We can write

$$\begin{aligned} (\lambda, \theta) &= (1-\alpha)(\lambda_1, \theta_1) + \alpha(\lambda_2, \theta_2) \\ &= (\lambda_1 + \alpha(\lambda_2 - \lambda_1), \theta_1 + \alpha(\theta_2 - \theta_1)) \end{aligned} \tag{42}$$

where $\alpha \in [0, 1]$.
Define

$$\begin{aligned} f(\alpha, W) =& \mathbb{D} \circ W + \theta_1 \|W\|_{\mathcal{G}} + \lambda_1 \|W\|_{\infty,1} \\ & + \alpha \left((\theta_2 - \theta_1) \|W\|_{\mathcal{G}} + (\lambda_2 - \lambda_1) \|W\|_{\infty,1} \right) \end{aligned}$$

If we see $(\theta_2 - \theta_1)\|W\|_{\mathcal{G}} + (\lambda_2 - \lambda_1)\|W\|_{\infty,1}$ as the new regularization term, according to Proposition 1 and $\operatorname{argmin}_W f(0, W) = \operatorname{argmin}_W f(1, W)$, we have for any $\alpha \in [0, 1]$, $\operatorname{argmin}_W f(\alpha, W) = W_1^*$.
So now we proved that the optimal solutions corresponding to the regularization parameters on the line $\mathcal{L}_{12}$ are identical.
For any

$$(\lambda, \theta) \in Conv\left((\lambda_1, \theta_1), (\lambda_1, \theta_2), (\lambda_2, \theta_2)\right),$$

we can find two points: one is $A := (\lambda, \theta_2)$ on the line between point (λ_1, θ_2) and (λ_2, θ_2); the other is $B := (\lambda, \frac{\lambda_2 - \lambda}{\lambda_2 - \lambda_1}\theta_1 + \frac{\lambda - \lambda_1}{\lambda_2 - \lambda_1}\theta_2)$ which is on the line $\mathcal{L}_{12}$. Similarly, we obtain that the optimal solutions corresponding to any points on the line between points A and B are identical. Therefore, we finish the proof. □