Appendix

A. Notation

$\mathcal{N} = \{1, 2, \ldots, N\} =: [N]$ is the whole set of data points. $i, j \in \mathcal{N}$ denote points. $d_{ij} := D(x_i, x_j)$. $D$ is the number of data sets. $\mathcal{T}_d \subseteq \mathcal{N}$ denotes the set of points in the $d$-th dataset, i.e., $\bigcup_{d=1}^{D} \mathcal{T}_d = \mathcal{N}$. $N_d = |\mathcal{T}_d|$ 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 $\bigcup_{d \in \mathcal{D}_k} \mathcal{S}_{k,d} = \mathcal{S}_k$. 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

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

(14)

The KKT condition of the linear programming can be written as

$$\begin{align*}
d_{ij} - \alpha_{ij} - \beta_i + \gamma_{ij} + \delta_{ij} &= 0 \quad (15) \\
\theta &= \sum_{i \in \mathcal{T}_d} \delta_{ij} \quad (16) \\
\lambda &= \sum_{i} \gamma_{ij} \quad (17) \\
\delta_{ij}(w_{ij} - \zeta_{dj}) &= 0 \quad (18) \\
\gamma_{ij}(w_{ij} - \xi_j) &= 0 \quad (19) \\
\alpha_{ij} w_{ij} &= 0 \quad (20) \\
\alpha_{ij} &\geq 0 \quad (21) \\
\gamma_{ij} &\geq 0 \quad (22) \\
\delta_{ij} &\geq 0. \quad (23)
\end{align*}$$

Our goal is to find a structure of $d_{ij}$, for which there exists a set of $\alpha_{ij}, \beta_{i}, \gamma_{ij}, \delta_{ij}, \theta$ and $\lambda$ 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. $\xi_j = 1, \zeta_{dj} = 1, w_{ij} = 1$

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

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

(24)
C.2. \( \xi_j = 1, \ \zeta_{dj} = 1, \ w_{ij} = 0 \)

\( j \in M, \text{but } j \neq M(i) \)

\( \delta_{ij} = 0, \ \gamma_{ij} = 0 \Rightarrow \alpha_{ij} = d_{ij} - \beta_i > 0, \text{ i.e.,} \)

\[ \beta_i < d_{ij}, \ \forall j \in M \text{ but } j \neq M(i) \text{ and } D_j \cap D_{M(i)} \neq \emptyset. \]  

(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,

\[ \beta_i = \frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} \]  

(26)

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

\( j \in M, \text{but } j \neq M(i) \)

\( \gamma_{ij} = 0 \Rightarrow \alpha_{ij} = d_{ij} - \beta_i + \delta_{ij} > 0. \text{ Now we have} \)

\[ \delta_{ij} > \beta_i - d_{ij} \]

\[ \delta_{ij} > 0 \]

Thus

\[ \theta = \sum_{i \in T_d} \delta_{ij} > \sum_{i \in T_d} (\beta_i - d_{ij})_+, \ \forall d \notin D_j, j \in M \]  

(27)

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

\[ \frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} - d_{ij} < \frac{\theta}{N_{d(i)}} \]  

(28)

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

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

\[ \gamma_{ij} > \beta_i - d_{ij} - \delta_{ij} \]  

(29)

\[ \lambda = \sum_i \gamma_{ij} > \sum_i (\beta_i - d_{ij} - \delta_{ij})_+, \ \forall j \notin M \]  

(30)

\[ \theta = \sum_{i \in T_d} \delta_{ij}, \ \forall d \in [D], \forall j \notin M \]  

(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 \( S_{1,M(j)} := (\cup_{d \in D_{M(j)}} T_d) \setminus S_{M(j)} \). The third part is all the points who don’t have sister points in \( S_{M(j)} \), denoted by \( S_{2,M(j)} := \cup_{d \in [D] \setminus D_{M(j)}} T_d \)

\[ \lambda > \sum_{i \in S_{M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \]

\[ + \sum_{i \in S_{1,M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \]

\[ + \sum_{i \in S_{2,M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \]  

(32)
In the following we will show our strategy to make this inequality hold. If we set $\delta_{ij}$ to be

$$\theta = \left( \sum_{i \in S_{M(j),d}} \delta_{ij} \right), \quad \forall d \in D_{M(j)}, \forall j \notin M$$

$$\delta_{ij} = 0, \quad \forall i \in S_{1,M(j)}, \forall j \notin M$$

$$\delta_{ij} = \frac{\theta}{N_{d(i)}}, \quad \forall i \in S_{2,M(j)}, \forall j \notin M$$

such that Eq. (16) is satisfied.

Furthermore, if we can get the following equations satisfied,

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

$$\beta_i - d_{ij} - \delta_{ij} < 0, \quad \forall i \in S_{1,M(j)}$$

$$\beta_i - d_{ij} - \delta_{ij} < 0, \quad \forall i \in S_{2,M(j)}$$

the only thing we need to show is

$$\lambda > \sum_{i \in S_{M(j)}} (\beta_i - d_{ij} - \delta_{ij})$$

$$= \sum_{i \in S_{M(j)}} \left( \frac{\lambda}{N_{M(i)}} + d_{i,M(i)} - d_{ij} \right)$$

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. $M(i) = M(j)$,

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

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

$$\theta < \sum_{i \in S_{k,d}} \beta_k - d_{ij}, \quad \forall d \in D_{k}, k = M(j)$$

Equivalently, we have

$$\lambda > \frac{N_k}{N_{k,d}} \sum_{i \in S_{k,d}} d_{ij} - d_{i,M(i)}, \quad \forall d \in D_{k}, \forall j \in S_{k}, \forall k$$

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

$$\beta_i - d_{ij} < 0, \quad \forall i \in S_{1,M(j)}$$

That is, for $\forall i,j$ s.t. $D_{M(i)} \cap D_{M(j)} \neq \emptyset$ and $M(i) \neq M(j)$

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

This requirement also implies Eq. (25) will hold.
A Convex Exemplar-based Approach to MAD-Bayes Dirichlet Process Mixture Models

Third part \( i \in S_{2,M(j)} \)  For this part, \[
\beta_i - d_{ij} < \frac{\theta}{N_{d(i)}}, \quad \forall i \in S_{2,M(j)}
\]
That is, for \( \forall i, j \) s.t. \( D_{M(i)} \cap D_{M(j)} = \emptyset \),
\[
\frac{\lambda}{N_{M(i)}} + \theta \left( \frac{1}{N_{M(i),d(i)}} - \frac{1}{N_{d(i)}} \right) + d_{i,M(i)} < d_{ij},
\]
(40)
This requirement also implies Eq. (28) will hold.

D. Proof of Proposition 1

Given the conditions in the proposition, we have
\[
\mathbb{D} \circ W^*_1 + \lambda_1 ||W^*_1||_{\infty,1} \\
\leq \mathbb{D} \circ W^*_2 + \lambda_1 ||W^*_2||_{\infty,1} \\
< \mathbb{D} \circ W^*_2 + \lambda_2 ||W^*_2||_{\infty,1} \\
\leq \mathbb{D} \circ W^*_1 + \lambda_2 ||W^*_1||_{\infty,1}
\]
(41)
So we have
\[
\mathbb{D} \circ W^*_1 \leq \mathbb{D} \circ W^*_2
\]
\[
\mathbb{D} \circ W^*_2 \leq \mathbb{D} \circ W^*_1
\]
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),
\[
\lambda_2 ||W^*_2||_{\infty,1} - \lambda_1 ||W^*_1||_{\infty,1} \leq \lambda_2 ||W^*_1||_{\infty,1} - \lambda_1 ||W^*_1||_{\infty,1}
\]
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 \( ||W^*(\lambda)||_{\infty,1} = ||W^*(\lambda_2)||_{\infty,1} \), 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 \( ||W^*_1||_g = ||W^*_2||_g \), we have \( W^*_1 = W^*_2 \) and for any \( \theta \in [\theta_1, \theta_2] \), \( W^*(\lambda_1, \theta) = W^*_1 \).
Given \( ||W^*_2||_{\infty,1} = ||W^*_1||_{\infty,1} \), we have, for any \( \lambda \in [\lambda_1, \lambda_2] \), \( W^*(\lambda, \theta) = W^*_1 \).
Now we prove for any \( (\lambda, \theta) \) on the line between point \( (\lambda_1, \theta_1) \) and point \( (\lambda_2, \theta_2) \) (defined by \( L_{12} \)), \( W^*(\lambda, \theta) = W^*_1 \). We can write
\[
(\lambda, \theta) = (1-\alpha)(\lambda_1, \theta_1) + \alpha(\lambda_2, \theta_2) \\
= (\lambda_1 + \alpha(\lambda_2 - \lambda_1), \theta_1 + \alpha(\theta_2 - \theta_1))
\]
(42)
where \( \alpha \in [0, 1] \).
Define
\[
f(\alpha, W) = \mathbb{D} \circ W + \theta_1 ||W||_g + \lambda_1 ||W||_{\infty,1} \\
+ \alpha ((\theta_2 - \theta_1)||W||_g + (\lambda_2 - \lambda_1)||W||_{\infty,1})
\]
If we see \( (\theta_2 - \theta_1)||W||_g + (\lambda_2 - \lambda_1)||W||_{\infty,1} \) as the new regularization term, according to Proposition 1 and \( \arg\min_W f(0, W) = \arg\min_W f(1, W) \), we have for any \( \alpha \in [0, 1] \), \( \arg\min_W f(\alpha, W) = W^*_1 \).
So now we proved that the optimal solutions corresponding to the regularization parameters on the line $L_{12}$ are identical. For any 
$$(\lambda, \theta) \in \text{Conv} \left( \langle \lambda_1, \theta_1 \rangle, \langle \lambda_2, \theta_2 \rangle \right),$$
we can find two points: one is $A := (\lambda, \theta_2)$ on the line between point $\langle \lambda_1, \theta_2 \rangle$ and $\langle \lambda_2, \theta_2 \rangle$; 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 $L_{12}$. 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. $\square$