Exemplar-Based Nonparametric Bayesian as Convex Structural-Regularized Program

October 7, 2014

Abstract

MAD-Bayes (MAP-based Asymptotic Derivations) has been recently proposed as a general technique to cast nonparametric Bayesian inference into scalable optimization problem that has demonstrated success in applications ranging from Dirichlet Process mixture model to nonparametric Hidden Markov model. However, the asymptotic derivation yields combinatorial objective functions that till now can only be solved via hill-climbing algorithm analogous to k-means. In this write-up, we show the exemplar-based version of those combinatorial objectives can be relaxed to convex structural-regularized program that, under certain conditions, share the same optimal solution. A general, efficient algorithm based on Alternating Direction Method of Multiplier (ADMM) and Frank-Wolfe is proposed for solving such convex program. In our experiments, the global minimum achieved by our approach significantly improve existing methods in terms of the objective without sacrificing efficiency.

1 Introduction

2 Formulation

3 Recovery Guarantees

3.1 Notation

$\mathcal{N} = \{1, 2, ..., N\} =: [N]$ is the whole set of data points. $i, j \in \mathcal{N}$ denote points. D is the number set of datasets. $\mathcal{N}_d \subset \mathcal{N}$ denotes the set of points in the dth dataset, i.e. $\cup_{d=1}^{D} \mathcal{N}_d = \mathcal{N}$. $N_d = |\mathcal{N}_d|$ is the number of points in Dataset d. $d(i) \in [D]$ denotes the dataset index of Point i. $\mathcal{M} \subset \mathcal{N}$ 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 $\mathcal{D}_k \subset [D]$ denote the datasets contained or partially contained in Cluster k. Denote $\mathcal{S}_{k,d} := \mathcal{S}_k \cap \mathcal{N}_d$ 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 $d \in \mathcal{D}_k$.

3.2 Main Theorem

Theorem 1 *For DP, if*

$$\max_{k \in \mathcal{M}} \max_{i,j \in \mathcal{S}_k} N_k \left(d_{ij} - d_{ik}\right) \leq \min_{(k,l \in \mathcal{M}, k \neq l)} \min_{(i \in \mathcal{S}_k, j \in \mathcal{S}_l)} N_k \left(d_{ij} - d_{ik}\right) \tag{1}$$

then the optimal solution of integer programming is also the optimal solution of linear programming.

Example of DP
Assume $N_k = N/K, \forall k$, where K is the number of clusters. We assume the points in every cluster k are all in the ball of radius R with the center at Point k, i.e. $d_{ik} \leq R, \forall i \in \mathcal{S}_k$. Thus $d_{ij} - d_{ik} \leq d_{jk} \leq R$, $\forall i, j \in \mathcal{S}_k, \forall k \in \mathcal{M}$. Now we can satisfy Eq. (1) by assuming

$$d_{ij} \geq 2R, \ \forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l, \forall k, l \in \mathcal{M} \text{ and } k \neq l \tag{2}$$

This equation can be interpreted as: the distance between any two clusters should be greater than the diameter of the clusters.

Theorem 2 *For HDP, if there exist some θ and λ such that,*

$$\frac{\lambda}{N_k} + \frac{\theta}{N_{k,d(i)}} \geq d_{ij} - d_{ik},\ \forall i,j \in \mathcal{S}_k, \forall k \in \mathcal{M} \tag{3}$$

$$\lambda \geq \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} d_{ij} - d_{ik},\ \forall d \in \mathcal{D}_k, \forall j \in \mathcal{S}_k, \forall k \in \mathcal{M} \tag{4}$$

$$\frac{\lambda}{N_k} + \frac{\theta}{N_{k,d(i)}} \leq d_{ij} - d_{ik},\ \forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l, \forall k,l \in \mathcal{M}\ s.t.\ \mathcal{D}_k \cap \mathcal{D}_l \neq \phi\ and\ k \neq l \tag{5}$$

$$\frac{\lambda}{N_k} + \theta\left(\frac{1}{N_{k,d(i)}} - \frac{1}{N_{d(i)}}\right) \leq d_{ij} - d_{ik},\ \forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l, \forall k,l \in \mathcal{M}\ s.t.\ \mathcal{D}_k \cap \mathcal{D}_l = \phi \tag{6}$$

then the optimal solution of integer programming is also the optimal solution of linear programming.

We can always find such θ and λ by assuming the minimal distance between different clusters greater enough than the maximal distance within the cluster. In Corollary 1, we will give a tight bound (maybe not the tightest) for the structure of d_{ij} such that the above conditions are satisfied.
Remark. If $\theta = 0$, this problem reduces to the DP-medoid problem and Eq. (3) will imply Eq. (4). Therefore, we don't specifically write the proof for Theorem 1.

Corollary 1 *If*

$$\lambda = \max_{k \in \mathcal{M}} \max_{j \in \mathcal{S}_k} \max_{d \in \mathcal{D}_k} \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} (d_{ij} - d_{ik})$$

$$\theta = \max_{k \in \mathcal{M}} \max_{i,j \in \mathcal{S}_k} N_{k,d(i)} \left(d_{ij} - d_{ik} - \frac{\lambda}{N_k}\right)$$

and Eq. (5) *and Eq.* (6) *are satisfied for the above λ and θ, then the solution of integer programming is also the solution of linear programming.*

Proof. Both λ and θ should be as small as possible. And λ should have a higher priority to make it small because a positive θ can relax Condition (6) to some extent compared with Condition (5). Thus we can first decide λ by Eq. (4) and then θ by Eq. (3),

$$\lambda = \max_{k \in \mathcal{M}} \max_{j \in \mathcal{S}_k} \max_{d \in \mathcal{D}_k} \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} (d_{ij} - d_{ik})$$

$$\theta = \max_{k \in \mathcal{M}} \max_{i,j \in \mathcal{S}_k} N_{k,d(i)} \left(d_{ij} - d_{ik} - \frac{\lambda}{N_k}\right)$$

We now prove that $\theta \geq 0$, so the choice of θ is justified. Let

$$(k^*, j^*, d^*) = \arg\max_{k \in \mathcal{M}} \max_{j \in \mathcal{S}_k} \max_{d \in \mathcal{D}_k} \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} (d_{ij} - d_{ik})$$

$$\begin{aligned}
\theta &\geq \max_{i \in \mathcal{S}_{k^*}} N_{k^*,d(i)} \left(d_{ij^*} - d_{ik^*} - \frac{\lambda}{N_{k^*}}\right) \\
&\geq \max_{i \in \mathcal{S}_{k^*,d^*}} N_{k^*,d^*} \left(d_{ij^*} - d_{ik^*} - \frac{\lambda}{N_{k^*}}\right) \\
&= \max_{i \in \mathcal{S}_{k^*,d^*}} N_{k^*,d^*} \left(d_{ij^*} - d_{ik^*}\right) - \sum_{i \in \mathcal{S}_{k^*,d^*}} (d_{i,j^*} - d_{ik^*}) \\
&\geq 0
\end{aligned}$$

□

Example of HDP
We consider a very special case. Assume $N_k = N/K, \forall k$, where K is the number of clusters and $|\mathcal{D}_k| = C$ is a constant for all k. And $N_{k,d} = N/(CK)$ for all k and $d \in \mathcal{D}_k$. Let

$$\lambda = C \max_{k \in \mathcal{M}} \max_{j \in \mathcal{S}_k} \max_{d \in \mathcal{D}_k} \sum_{i \in \mathcal{S}_{k,d}} (d_{ij} - d_{ik})$$

We assume the points in every cluster k are all in the ball of radius R with the center at Point k. Thus $d_{ij} - d_{ik} \leq R,\ \forall i,j \in \mathcal{S}_k, \forall k \in \mathcal{M}$. Now we can satisfy Eq. (3) by setting

$$\theta = \frac{NR}{CK} - \frac{\lambda}{C}$$

And Eq. (5) will hold if

$$d_{ij} \geq 2R,\ \forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l, \forall k, l \in \mathcal{M} \text{ s.t. } \mathcal{D}_k \cap \mathcal{D}_l \neq \phi \text{ and } k \neq l \tag{7}$$

This equation can be interpreted as: the distance between any two clusters which contain a same dataset should be greater than the diameter of the clusters.
To have Eq. (6) hold, we can require

$$d_{ij} \geq 2R - \frac{\theta}{N_d},\ \forall i \in \mathcal{S}_k, \forall j \in \mathcal{S}_l, \forall d \in \mathcal{D}_k \cup \mathcal{D}_l, \forall k, l \in \mathcal{M} \text{ s.t. } \mathcal{D}_k \cap \mathcal{D}_l = \phi$$

Compared with Eq. (7), the distance between any two clusters which **don't** contain a same dataset should be greater than a value less than $2R$.

3.3 Proof of Theorem 2

The KKT condition of the linear programming can be written as,

$$d_{ij} - \alpha_{ij} - \beta_i + \gamma_{ij} + \delta_{ij} = 0 \tag{8}$$

$$\theta = \sum_{i \in \mathcal{N}_d} \delta_{ij} \tag{9}$$

$$\lambda = \sum_i \gamma_{ij} \tag{10}$$

$$\delta_{ij}(w_{ij} - \zeta_{dj}) = 0 \tag{11}$$

$$\gamma_{ij}(w_{ij} - \xi_j) = 0 \tag{12}$$

$$\alpha_{ij} w_{ij} = 0 \tag{13}$$

$$\alpha_{ij} \geq 0 \tag{14}$$

$$\gamma_{ij} \geq 0 \tag{15}$$

$$\delta_{ij} \geq 0 \tag{16}$$

Our goal is to find the structure of d_{ij} under which there exists a set of α_{ij}, β_i, γ_{ij}, δ_{ij}, θ and λ such that the above equations hold, provided the solution of integer problem $\{\xi_j, \zeta_{dj}, w_{ij}\}$. We will discuss the cases entry-by-entry.

3.3.1 $\xi_j = 1,\ \zeta_{dj} = 1,\ w_{ij} = 1$

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

$$\gamma_{i,M(i)} + \delta_{i,M(i)} = \beta_i - d_{i,M(i)}, \quad \forall i \tag{17}$$

3.3.2 $\xi_j = 1,\ \zeta_{dj} = 1,\ w_{ij} = 0$

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

$$\beta_i \leq d_{ij}, \quad \forall j \in \mathcal{M} \text{ but } j \neq M(i) \text{ and } \mathcal{D}_j \cap \mathcal{D}_{M(i)} \neq \phi. \tag{18}$$

Summary of Section 3.3.1 and 3.3.2
We can set $\gamma_{i,M(i)} = \frac{\lambda}{N_{M(i)}}$ such that Eq. (10) holds and $\delta_{i,M(i)} = \frac{\theta}{N_{M(i),d(i)}}$ such that Eq. (9) holds. Thus,

$$\beta_i = \frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} \tag{19}$$

3.3.3 $\xi_j = 1,\ \zeta_{dj} = 0,\ w_{ij} = 0$

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

$$\delta_{ij} \geq \beta_i - d_{ij}$$
$$\delta_{ij} \geq 0$$

Thus

$$\theta = \sum_{i \in \mathcal{N}_d} \delta_{ij} \geq \sum_{i \in \mathcal{N}_d} (\beta_i - d_{ij})_+, \quad \forall d \notin \mathcal{D}_j, j \in \mathcal{M} \tag{20}$$

If we set $\beta_i - d_{ij} \leq \frac{\theta}{N_{d(i)}}$, Eq. (20) will be satisfied. That is

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} - d_{ij} \leq \frac{\theta}{N_{d(i)}} \tag{21}$$

3.3.4 $\xi_j = 0,\ \zeta_{dj} = 0,\ w_{ij} = 0$

$\alpha_{ij} = d_{ij} - \beta_i + \delta_{ij} + \gamma_{ij} \geq 0 \Rightarrow$

$$\gamma_{ij} \geq \beta_i - d_{ij} - \delta_{ij} \tag{22}$$

$$\lambda = \sum_i \gamma_{ij} \geq \sum_i (\beta_i - d_{ij} - \delta_{ij})_+, \quad \forall j \notin \mathcal{M} \tag{23}$$

$$\theta = \sum_{i \in \mathcal{N}_d} \delta_{ij}, \quad \forall d \in [D], \forall j \notin \mathcal{M} \tag{24}$$

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 $S_{-M(j)} := \left(\cup_{d \in \mathcal{D}_{M(j)}} \mathcal{N}_d\right) \setminus S_{M(j)}$. The third part is all the points who don't have sister points in $S_{M(j)}$, denoted by $S_{--M(j)} := \cup_{d \in [D] \setminus \mathcal{D}_{M(j)}} \mathcal{N}_d$

$$\lambda \geq \sum_{i \in S_{M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ + \sum_{i \in S_{-M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ + \sum_{i \in S_{--M(j)}} (\beta_i - d_{ij} - \delta_{ij})_+ \tag{25}$$

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

$$\theta = \left(\sum_{i \in \mathcal{S}_{M(j),d}} \delta_{ij} \right), \quad \forall d \in \mathcal{D}_{M(j)}, \forall j \notin \mathcal{M} \tag{26}$$

$$\delta_{ij} = 0, \quad \forall i \in \mathcal{S}_{-M(j)}, \forall j \notin \mathcal{M} \tag{27}$$

$$\delta_{ij} = \frac{\theta}{N_{d(i)}}, \quad \forall i \in \mathcal{S}_{--M(j)}, \forall j \notin \mathcal{M} \tag{28}$$

such that Eq. (9) is satisfied.
Further more, if we can get the following equations satisfied,

$$\begin{aligned} &\beta_i - d_{ij} - \delta_{ij} \geq 0,\ \forall i \in \mathcal{S}_{M(j)} \\ &\beta_i - d_{ij} - \delta_{ij} \leq 0,\ \forall i \in \mathcal{S}_{-M(j)} \\ &\beta_i - d_{ij} - \delta_{ij} \leq 0,\ \forall i \in \mathcal{S}_{--M(j)} \end{aligned} \tag{29}$$

the only thing we need to show is

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

It is equivalent to

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

which is satisfied by medoid definition.
In the following, we analyze the conditions under which the three inequalities of Eq. (29) hold.

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

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

That is,

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} \geq d_{ij},\ \forall i, j \text{ s.t. } M(i) = M(j) \tag{30}$$

Then we can always find a δ_{ij} such that $0 \leq \delta_{ij} \leq \beta_i - d_{ij}$. To satisfy Eq. (26), we require

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

Equivalently, we have

$$\lambda \geq \frac{N_k}{N_{k,d}} \sum_{i \in \mathcal{S}_{k,d}} d_{ij} - d_{i,M(i)},\ \forall d \in \mathcal{D}_k, \forall j \in \mathcal{S}_k, \forall k \tag{31}$$

Second part $i \in S_{-M(j)}$ As set in Eq. (27), $\delta_{ij} = 0$, we require

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

That is,

$$\frac{\lambda}{N_{M(i)}} + \frac{\theta}{N_{M(i),d(i)}} + d_{i,M(i)} \leq d_{ij},\ \forall i,j \text{ s.t. } \mathcal{D}_{M(i)} \cap \mathcal{D}_{M(j)} \neq \phi \text{ and } M(i) \neq M(j) \tag{32}$$

This requirement also implies Eq. (18) will hold.

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

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

That is,

$$\frac{\lambda}{N_{M(i)}} + \theta\left(\frac{1}{N_{M(i),d(i)}} - \frac{1}{N_{d(i)}}\right) + d_{i,M(i)} \leq d_{ij},\ \forall i,j \text{ s.t. } \mathcal{D}_{M(i)} \cap \mathcal{D}_{M(j)} = \phi \tag{33}$$

This requirement also implies Eq. (21) will hold.

References

[1] T. Broderick, B. Kulis, and M. I. Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In Proceedings of the 30th International Conference on Machine Learning, 2013.

[2] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In Proceedings of the 29th International Conference on Machine Learning, 2012.

[3] A. Roychowdhury, K. Jiang, and B. Kulis, Small-variance asymptotics for hidden markov models, in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 21032111.