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

Presenter: Jimmy Xin Lin

Joint work with Ian E.H. Yen, Kai Zhong, Pradeep Ravikumar and Inderjit S. Dhillon

The University of Texas At Austin

June 9, 2015
Abstract

- In this paper, we connect convex, exemplar approach to the literature of MAD-Bayes and demonstrate its effectiveness in theory and practice.

- MAD-Bayes: Estimating asymptotic Bayesian non-parametric models via optimization algorithm similar to k-means.

- Exemplar Approach: By restricting parameter space to a finite set of exemplars (as in k-medoids), it obtains (i) flexibility in choice of dissimilarity measure (beyond Bregman Divergence and convexity), and (ii) tight convex relaxation under mild conditions.

- Dirichlet Process Mixture Models (DP, HDP) have wide applications including clustering, hierarchical clustering and topic modeling.
# Table of Contents

1. MAD-Bayes: MAP-based Asymptotic Derivation from Bayes

2. Convex Exemplar-based Approach

3. Optimality Guarantees

4. ADMM for Structural-Regularized Programs

5. Results and Future Works
MAD-Bayes: Dirichlet Process Mixture

- MAD-Bayes derives the asymptotic log-likelihood of DP mixture when variance $\sigma^2_X$ of dist. $P(x_i|\mu_{z_i})$ approaches 0.

\[
DP(\alpha, G_0) \quad \xrightarrow{\sigma^2_X \to 0} \quad \alpha := \exp(-\lambda/\sigma^2_X) \to 0 \\
\text{Taking limit of 2nd moment} \\
\rightarrow \sum_{i=1}^{N} \mathbb{D}(x_i, \mu_{z_i}) + \lambda K
\]

where $\mathbb{D}(x_i, \mu_k)$ is the Bregman Divergence (corresponds to $P(x|z,\mu)$) between sample $x_i$ and $k$-th mean parameter $\mu_k$, and $K$ is the number of clusters.

- Inference (by MCMC etc.) is replaced by a MAP optimization program w.r.t. $z_i$, $\mu_k$, and $K$. 
MAD-Bayes: Dirichlet Process Mixture

- MAD-Bayes of DP mixture yields a combinatorial problem

\[
\min_{z_i \in [K], \mu_k \in \mathbb{R}^p, K} \sum_{i=1}^{N} \mathbb{D}(x_i, \mu_{z_i}) + \lambda K.
\]

- The approach extends to \textit{Hierarchical DP (HDP) mixture:}

\[
\min_{z_i \in [K_g], \mu_k, K_g, K_d} \sum_{d=1}^{D} \sum_{i \in d} \mathbb{D}(x_i, \mu_{z_i}) + \theta \sum_{d=1}^{D} K_d + \lambda K_g
\]

- Brian’s paper [1] proposed an \textit{K}-means-like algorithm (DP-means) to solve the DP optimization, where \textit{K} is incremented whenever doing so decreases objective, and another one (HDP-means) for the HDP optimization.

Table Of Contents

1 MAD-Bayes: MAP-based Asymptotic Derivation from Bayes

2 Convex Exemplar-based Approach

3 Optimality Guarantees

4 ADMM for Structural-Regularized Programs

5 Results and Future Works
Convex Exemplar-based Approach

Observation:

\[
\min_{z_i \in [K], \mu_k \in \mathbb{R}^p, K} \sum_{i=1}^{N} D(x_i, \mu_{z_i}) + \lambda K.
\]

is equivalent to

\[
\min_{W \in \{0,1\}^{N \times J}} D \circ W + \lambda \| W \|_{\infty,1}
\]

s.t. \( W1 = 1 \),

where \( D_{ij} = D(x_i, \mu_j) \),

assuming there is a Exemplar Set \( \mathcal{E} = \{\mu_j\}_j \) of possible mean parameters. The program becomes convex when we relax \( w_{i,j} \) from \( \{0,1\} \) to \([0,1] \). This concept of exemplar was also employed in \( K\)-medoid problem [1].

Convex Exemplar-based Approach

For HDP:

\[
\sum_{d=1}^{D} \sum_{i \in d} \mathbb{D}(x_i, \mu_{z_i}) + \theta \sum_{d=1}^{D} K_d + \lambda K_g
\]

is equivalent to

\[
\min_{W \in \{0,1\}^{N \times J}} \mathbb{D} \circ W + \theta \|W\|_\mathcal{G} + \lambda \|W\|_{\infty,1}
\]

s.t. \quad W1 = 1

where \( \mathbb{D}_{ij} = \mathbb{D}(x_i, \mu_j) \),

assuming there is a Exemplar Set \( \mathcal{E} = \{\mu_j\}_{j}^{J} \) of possible mean parameters. The program becomes convex when relaxing \( w_{i,j} \) from \( \{0,1\} \) to \([0,1]\).
Table Of Contents

1 MAD-Bayes: MAP-based Asymptotic Derivation from Bayes

2 Convex Exemplar-based Approach

3 Optimality Guarantees

4 ADMM for Structural-Regularized Programs

5 Results and Future Works
Optimality Guarantee

Suppose the convex relaxation

\[
\min_{W \in [0,1]^{N \times J}} \mathcal{D} \circ W + \lambda \|W\|_{\infty,1}
\]

\[\text{s.t.} \quad W1 = 1, \quad (1)\]

has integer solution, the solution is also optimal to the original combinatorial problem. The following shows a clustering satisfying a separation condition which leads to an Integer solution for the Convex Program (1) for a range of \( \lambda \).

**Theorem**

*Suppose there exists a clustering \( \{S_k\}_{k \in M} \) for which we can find \( \lambda \) such that*

\[
\max_{k \in M} \max_{i,j \in S_k} N_k \delta_{ij} < \lambda < \min_{(k,l \in M, k \neq l)} \min_{(i \in S_k, j \in S_l)} N_k \delta_{ij} \quad (2)
\]

*where \( N_k = |S_k| \) and \( \delta_{ij} = \mathcal{D}(x_i, x_j) - \mathcal{D}(x_i, x_{M(i)}) \), then the integer solution \( W^* \) realizing \( \{S_k\}_{k \in M} \) is unique optimal solution to (1).*
Optimality Guarantee

- According to the theorem, the larger extent of separation a clustering has, the wider range of $\lambda$ producing that clustering.

- (i) $K(\lambda) = \|W(\lambda)^*\|_{\infty,1}$ monotonically decreases with $\lambda$.

- (ii) $W^*(\lambda)$ and $K(\lambda)$ have one-one mapping.
Optimality Guarantee

- Similar guarantee can be obtained for HDP formulation, where clusters that do not share a data set can have smaller separation requirement.

- (i) $K_g$, $K_l$ monotonically decrease with $\lambda$, $\theta$ respectively.
- (ii) $W^*(\lambda, \theta)$ and $(K_g, K_l)$ have one-one mapping.
Table Of Contents

1. MAD-Bayes: MAP-based Asymptotic Derivation from Bayes

2. Convex Exemplar-based Approach

3. Optimality Guarantees

4. ADMM for Structural-Regularized Programs

5. Results and Future Works
Solving Structural-Regularized Programs

We employ an ADMM procedure that has linear convergence to the optimum of the given convex program

\[
\begin{align*}
\min_{W \in [0,1]^{N \times J}} & \quad \mathcal{D} \circ W + \theta \| W \|_\mathcal{G} + \lambda \| W \|_{\infty,1} \\
\text{s.t.} & \quad W1 = 1
\end{align*}
\]

by introducing dual variables $Y_1, Y_2$ and consensus variables $Z$ s.t. (3) can be decomposed into a sequence of sub-problem (14) with only simplex constraint

\[
W_1^{(t+1)} = \argmin_{W \in [0,1]^{N \times J}} \mathcal{D} \circ W + Y_1^{(t)} \circ W + \frac{\rho}{2} \| W - Z^{(t)} \|_2^2
\]

\[
\text{s.t.} \quad W1 = 1
\]

and sub-problem (14) with only structural regularizer

\[
W_2^{(t+1)} = \argmin_{W \in [0,1]^{N \times J}} \theta \| W \|_\mathcal{G} + \lambda \| W \|_{\infty,1} + Y_2^{(t)} \circ W + \frac{\rho}{2} \| W - Z^{(t)} \|_2^2.
\]
Solving Structural-Regularized Program

- The 1st sub-problem:

\[
W_1^{(t+1)} = \argmin_{W \in [0,1]^{N \times J}} W \circ D + Y_1^{(t)} \circ W + \frac{\rho}{2} \| W - Z(t) \|^2
\]

\[s.t. \quad W1 = 1\]

can be solved via Simplex Projection.

- The 2nd sub-problem:

\[
W_2^{(t+1)} = \argmin_{W \in [0,1]^{N \times J}} \theta \| W \|_{\infty,1} + \lambda \| W \|_{\infty,1} + Y_2^{(t)} \circ W + \frac{\rho}{2} \| W - Z(t) \|^2.
\]

has closed-form solution via proximal mapping \(\text{prox}_\lambda(\text{prox}_\theta(\cdot))\).

- ADMM Update:

\[
Z^{(t+1)} \leftarrow (W_1^{(t+1)} + W_2^{(t+1)})/2
\]

\[
Y_q^{(t+1)} \leftarrow Y_q^{(t)} + \alpha (W_q^{(t+1)} - Z^{(t+1)}), \text{ for } q = 1, 2
\]
Table Of Contents

1 MAD-Bayes: MAP-based Asymptotic Derivation from Bayes

2 Convex Exemplar-based Approach

3 Optimality Guarantees

4 ADMM for Structural-Regularized Programs

5 Results and Future Works
Results

The MAD-Bayes objective function achieved by different approaches.

<table>
<thead>
<tr>
<th>Data set</th>
<th>DP-convex</th>
<th>DP-convex (means)</th>
<th>DP-medoids</th>
<th>DP-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris (λ = 2)</td>
<td>29.26 (K=7)</td>
<td>27.97 (K=7)</td>
<td>35.68 (K=3)</td>
<td>30.20 (K=4)</td>
</tr>
<tr>
<td>Glass (λ = 9)</td>
<td>137.40 (K=6)</td>
<td>128.13 (K=6)</td>
<td>175.42 (K=2)</td>
<td>154.66 (K=2)</td>
</tr>
<tr>
<td>Wine (λ = 20)</td>
<td>298.55 (K=4)</td>
<td>263.79 (K=4)</td>
<td>512.04 (K=1)</td>
<td>402.40 (K=1)</td>
</tr>
<tr>
<td>DNA (λ = 1000)</td>
<td>105947.0 (K=2)</td>
<td>68718.9 (K=2)</td>
<td>107211 (K=1)</td>
<td>68156.4 (K=1)</td>
</tr>
<tr>
<td>Segment (λ = 600)</td>
<td>4749.62 (K=4)</td>
<td>4572.3 (K=4)</td>
<td>8405.71 (K=1)</td>
<td>7898.98 (K=1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data set</th>
<th>HDP-convex</th>
<th>HDP-convex (means)</th>
<th>HDP-medoids</th>
<th>HDP-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wholesale (λ = 1.0, θ = 1.0)</td>
<td>56.28 (Kg=5, K1=16)</td>
<td>53.07 (Kg=5, K1=16)</td>
<td>83.35 (Kg=2, K1=6)</td>
<td>79.52 (Kg=2, K1=6)</td>
</tr>
<tr>
<td>Water (λ = 1.0, θ = 1.0)</td>
<td>244.59 (Kg=32, K1=81)</td>
<td>232.07 (Kg=34, K1=83)</td>
<td>256.41 (Kg=33, K1=74)</td>
<td>237.73 (Kg=37, K1=64)</td>
</tr>
</tbody>
</table>

- (H)DP-convex: the ADMM-based program proposed at the previous section.
- (H)DP-convex-mean: adapted our program by taking means of the resulted assignment as cluster centroids.
- (H)DP-medoids: adapted Brian’s algorithm [1] for the exemplar-based objective.
- (H)DP-means: an K-means-like algorithm from [1].

Results

The MAD-Bayes objective function achieved by different approaches.

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

- All $\lambda$s here are picked from integer bands.
- The objectives achieved by (H)DP-convex are significantly lower than those reached by (H)DP-medoids.
- It verifies that the (H)DP-convex program, if a suitable $\lambda$ is selected, always reached the global optimum.
Future Works

- **Extension to Hidden Markov Model.** The convex optimization approach applies to other exemplar-based unsupervised learning problems such as *Multiple Sequence Alignment* and *Sequence Motif Finding*, which can be seen as exemplar-based version of Hidden Markov Model.

- **Reduction of Time Complexity.** For DP mixture models, the current algorithm runs in $O(N^2)$ time. A greedy *Column Generation* method that reduces complexity from $O(N^2)$ to $O(NK)$ is desired.
Optimality Guarantee

Suppose the convex relaxation
\[
\min_{W \in [0,1]^n \cdot} D \circ W + \lambda \|W\|_{\infty,1}
\text{ s.t. } W 1 = 1,
\]
has integer solution, the solution is also optimal to the original combinatorial problem. The following shows a clustering satisfying a separation condition which leads to an integer solution for the Convex Program (1) for a range of \( \lambda \).

Theorem

Suppose there exists a clustering \( \{S_k\}_{k \in M} \) for which we can find \( \lambda \) such that
\[
\max_{k \in M} \min_{j \in S_k} N_k \delta_j < \lambda < \min_{(k,l) \in M \times J} \min_{j \in S_k \cup S_l} N_k \delta_j
\]
where \( N_k = |S_k| \) and \( \delta_j = D(x_i, x_j) - D(x_i, x_M(j)) \), then the integer solution \( W^* \) realizing \( \{S_k\}_{k \in M} \) is unique optimal solution to (1).

Solving Structural-Regularized Program

- The 1st sub-problem:
  \[
  W_1^{(t+1)} = \arg\min_{W \in [0,1]^{n \times r}} D \circ W + \|W\|_{\infty,1} + \lambda \|W\|_{\infty,1} + \sum_{k=1}^K \frac{1}{2} \|W - Z^{(t)}\|^2
  \text{ s.t. } W 1 = 1
\]
can be solved via Simplex Projection.

- The 2nd sub-problem:
  \[
  W_2^{(t+1)} = \arg\min_{W \in [0,1]^{n \times r}} \|W\|_{\infty} + \|W - Z^{(t)}\|^2
\]
has closed-form solution via proximal mapping \( \text{prox}_{\lambda} \phi(\cdot) \).

- ADMM Update:
  \[
  Z^{(t+1)} \leftarrow (W_1^{(t+1)} + W_2^{(t+1)}) / 2
  \]
  \[
  Y_0^{(t+1)} \leftarrow Y_0^{(t)} + \alpha \left( W_2^{(t+1)} - Z^{(t+1)} \right), \text{ for } q = 1, 2
  \]

<table>
<thead>
<tr>
<th>Data set</th>
<th>DP-convex</th>
<th>DP-convex (means)</th>
<th>DP-medoids</th>
<th>DP-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris ( (\lambda = 2) )</td>
<td>29.26 (K=7)</td>
<td>27.97 (K=7)</td>
<td>35.68 (K=3)</td>
<td>30.20 (K=4)</td>
</tr>
<tr>
<td>Glass ( (\lambda = 9) )</td>
<td>137.40 (K=6)</td>
<td>128.13 (K=6)</td>
<td>175.42 (K=2)</td>
<td>154.66 (K=2)</td>
</tr>
<tr>
<td>Wine ( (\lambda = 20) )</td>
<td>298.55 (K=4)</td>
<td>263.79 (K=4)</td>
<td>512.04 (K=1)</td>
<td>402.40 (K=1)</td>
</tr>
<tr>
<td>DNA ( (\lambda = 1000) )</td>
<td>105947.0 (K=2)</td>
<td>68718.9 (K=2)</td>
<td>107211 (K=1)</td>
<td>68156.4 (K=1)</td>
</tr>
<tr>
<td>Segment ( (\lambda = 600) )</td>
<td>4749.62 (K=4)</td>
<td>4572.3 (K=4)</td>
<td>8405.71 (K=1)</td>
<td>7898.98 (K=1)</td>
</tr>
<tr>
<td>Data set</td>
<td>HDP-convex</td>
<td>HDP-convex (means)</td>
<td>HDP-medoids</td>
<td>HDP-means</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Wholesale ( (\lambda = 1.0, \theta = 1.0) ) ( (K_p=5, K_i=16) )</td>
<td>56.28</td>
<td>53.07</td>
<td>83.35</td>
<td>79.52</td>
</tr>
<tr>
<td>Water ( (\lambda = 1.0, \theta = 1.0) ) ( (K_p=32, K_i=81) )</td>
<td>244.59</td>
<td>232.07</td>
<td>256.41</td>
<td>237.73</td>
</tr>
</tbody>
</table>

I. En-Hsu Yen, X. Lin, K. Zhong, P. Ravikumar, I. Dhillon

The University of Texas At Austin
Thank you!