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

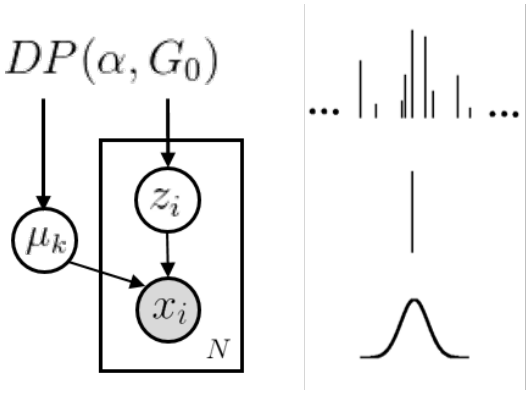
The University of Texas At Austin

June 9, 2015

Table of Contents

MAD-Bayes: Dirichlet Process Mixture

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



$$\sigma_X^2 \to 0, \quad \alpha := \exp(-\lambda/\sigma_X^2) \to 0$$

Taking limit of 2nd moment

$$-\log P(x|z,\mu) - \log P_\alpha(z) \to \sum_{i=1}^{N} \mathbb{D}(x_i, \mu_{z_i}) + \lambda K$$

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 μ_k, and K is the number of clusters.

- Inference (by MCMC etc.) is replaced by a MAP optimization program w.r.t. z_i, μ_k, and K.

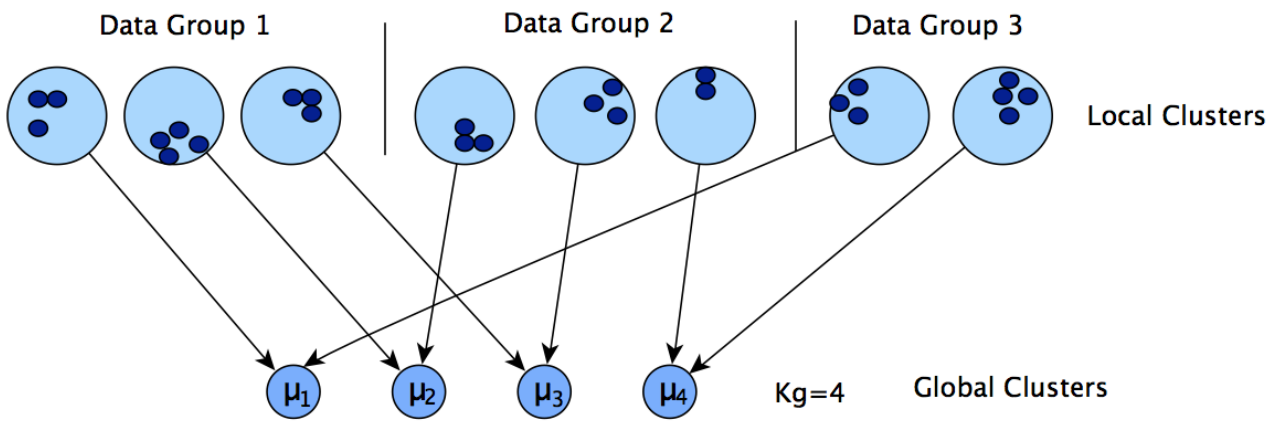
MAD-Bayes: Dirichlet Process Mixture

- MAD-Bayes of DP mixture yields a combinatorial problem

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

- Brian's paper [1] solves above via an algorithm similar to K-means, where K is incremented whenever doing so decreases objective.
- The approach extends to *Hierarchical DP (HDP) mixture*:

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



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

Table Of Contents

Convex Exemplar-based Approach

Observation:

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

is equivalent to

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

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

$$W = \begin{array}{c} \overbrace{}^{K=\|W\|_{\infty,1}=3} \\ \mu_1\ \mu_2\ \mu_3\ \cdots\cdots\ \mu_J \\ \begin{array}{c} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_N \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & 0 & \\ 1 & 0 & 0 & 0 & 0 & \cdots\cdots \\ 0 & 1 & 0 & 0 & 0 & \\ 0 & 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & 0 & \end{bmatrix} \end{array}$$

assuming there is a *Exemplar* Set $\mathscr{E} = \{\mu_j\}_j^J$ 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].

---

[1] Nellore, Abhinav and Ward, Rachel. Recovery Gurantees for exemplar-based clustering. *arXiv:1309.3256*, 2013.

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

$$\begin{aligned}\min_{W}\quad & \mathbb{D}\circ W+\theta\|W\|_{\mathcal{G}}+\lambda\|W\|_{\infty,1}\\ s.t.\quad & W\mathbf{1}=\mathbf{1}\end{aligned}$$

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

$$\sum_{d=1}^{3}K_d=\|W\|_{\mathcal{G}}:=\sum_{d=1}^{3}\|W_{d,:}\|_{\infty,1}=4$$

$$W=\begin{array}{c} \\ z_1\\ z_2\\ z_3\\ z_4\\ z_5\\ z_6\end{array}\begin{array}{c}\overbrace{}^{\mu_1\ \mu_2\ \mu_3}\ \cdots\cdots\ \mu_J\\ \left[\begin{array}{ccccc}1&0&0&0&0\\1&0&0&0&0\\0&1&0&0&0\\0&1&0&0&0\\0&1&0&0&0\\0&0&1&0&0\end{array}\right.\ \cdots\cdots\ \left.\begin{array}{c}\\\\\\\\\\\end{array}\right]\end{array}$$

assuming there is a *Exemplar* Set $\mathscr{E}=\{\mu_j\}_j^J$ of possible mean parameters. The program becomes convex when relaxing $w_{i,j}$ from $\{0,1\}$ to $[0,1]$.

Table Of Contents

Optimality Guarantee

Suppose the convex relaxation

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

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 λ.

Theorem

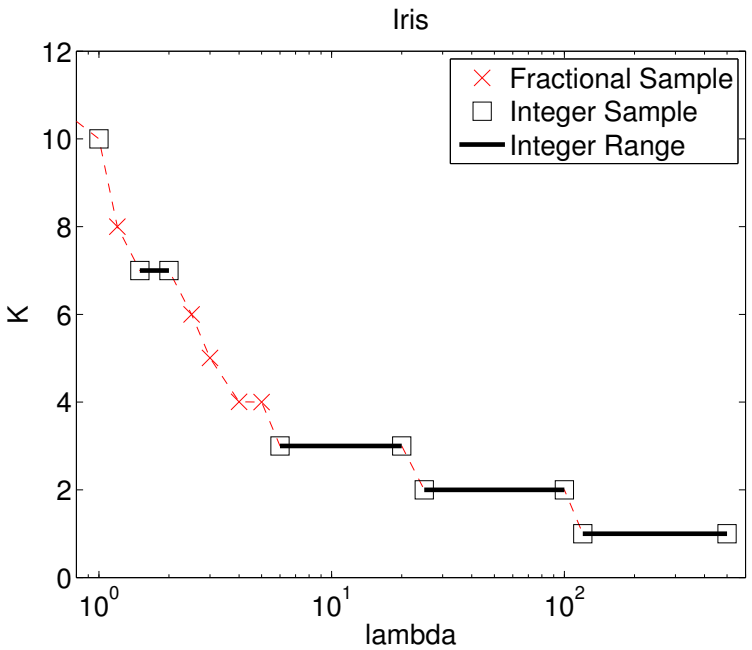
Suppose there exists a clustering $\{S_k\}_{k \in M}$ for which we can find λ such that

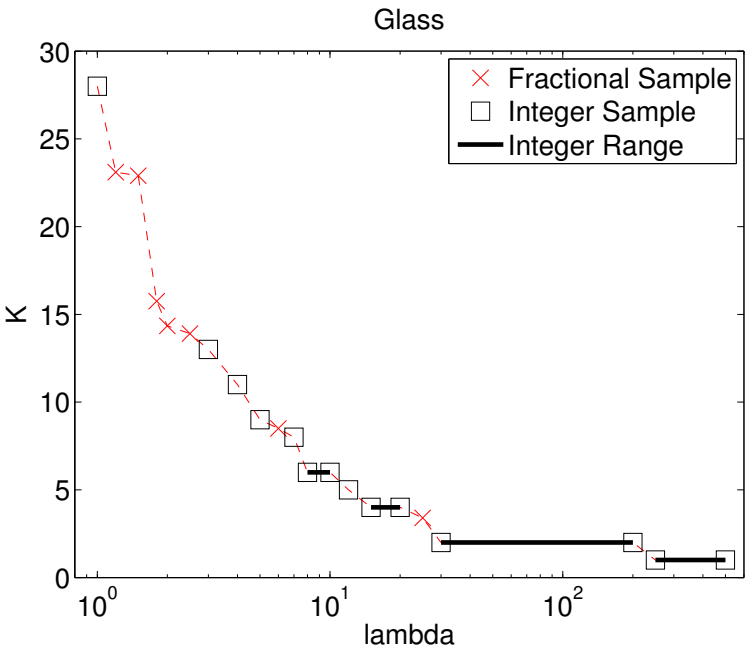
$$\max_{k \in M} \max_{i,j \in S_k} N_k \delta_{ij} < \lambda < \min_{(k,l \in M, k \neq l)} \min_{(i \in S_k, j \in S_l)} N_k \delta_{ij} \tag{2}$$

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

Optimality Guarantee

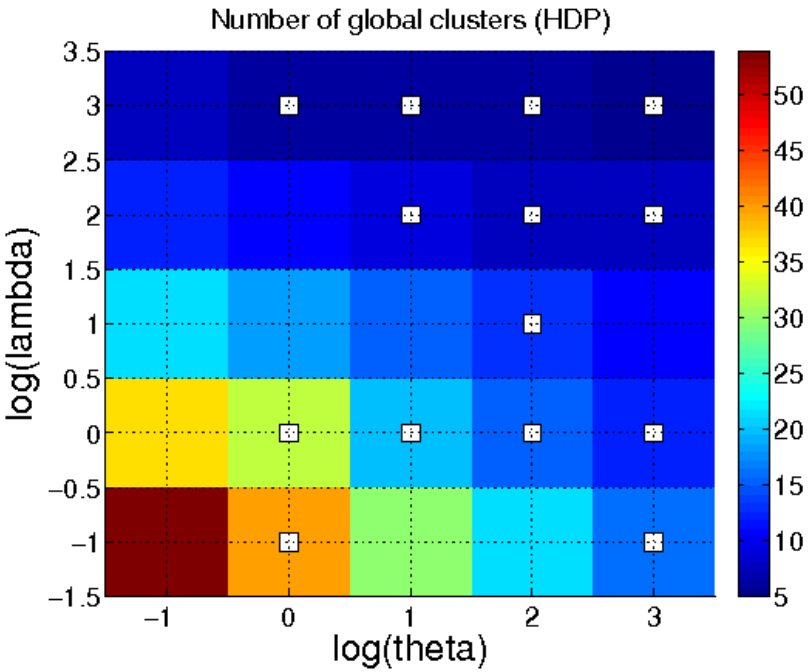
- According to the theorem, the larger extent of separation a clustering has, the wider range of λ producing that clustering.
- (i) $K(\lambda) = \|W(\lambda)^*\|_{\infty,1}$ monotonically decreases with λ.
  (ii) $W^*(\lambda)$ and $K(\lambda)$ have one-one mapping.





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 λ, θ respectively.
  (ii) $W^*(\lambda, \theta)$ and (K_g, K_l) have one-one mapping.



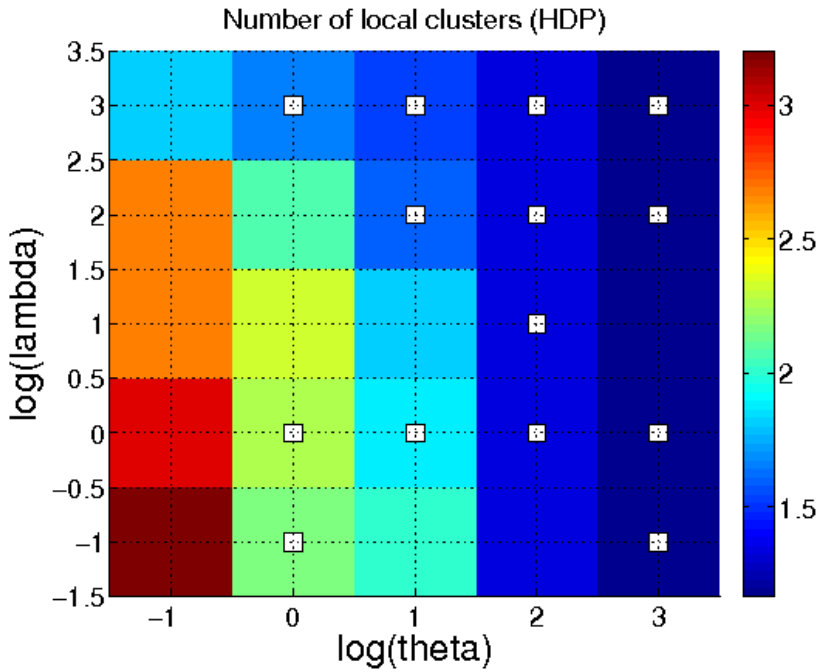


Table Of Contents

Solving Structural-Regularized Programs

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

$$\begin{aligned} \min_{W} \quad & \mathbb{D} \circ W + \theta \|W\|_{\mathscr{G}} + \lambda \|W\|_{\infty,1} \\ s.t. \quad & W\mathbf{1} = \mathbf{1} \end{aligned} \tag{3}$$

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

$$\begin{aligned} W_1^{(t+1)} = \underset{W \in [0,1]^{N \times J}}{argmin} \quad & \mathbb{D} \circ W + Y_1^{(t)} \circ W + \frac{\rho}{2} \|W - Z^{(t)}\|^2 \\ s.t. \quad & W\mathbf{1} = \mathbf{1} \end{aligned}$$

and (ii) sub-problem with only structural regularizer

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

Solving Structural-Regularized Program

- The 1st sub-problem:

$$W_1^{(t+1)} = \underset{W \in [0,1]^{N \times J}}{argmin} \quad \mathbb{D} \circ W + Y_1^{(t)} \circ W + \frac{\rho}{2} \| W - Z^{(t)} \|^2$$
$$s.t. \qquad W\mathbf{1} = \mathbf{1}$$

  can be solved via Simplex Projection.

- The 2nd sub-problem:

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

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

- ADMM Update:

$$Z^{(t+1)} \leftarrow (W_1^{(t+1)} + W_2^{(t+1)})/2$$
$$Y_q^{(t+1)} \leftarrow Y_q^{(t)} + \alpha(W_q^{(t+1)} - Z^{(t+1)}), \ for \ q = 1, 2$$

Table Of Contents

Results

The MAD-Bayes objective function achieved by different approaches.

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)

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

Future Works

- 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.
- 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.

Question Time

Optimality Guarantee

Suppose the convex relaxation

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

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 λ.

Theorem

Suppose there exists a clustering $\{S_k\}_{k\in M}$ for which we can find λ such that

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

where $N_k=|S_k|$ and $\delta_{ij}=\mathbb{D}(x_i,x_j)-\mathbb{D}(x_i,x_{M(i)})$, 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:

$$\begin{aligned} W_1^{(t+1)} = \underset{W\in[0,1]^{N\times J}}{argmin} \quad & \mathbb{D}\circ W+Y_1^{(t)}\circ W+\frac{\rho}{2}\|W-Z^{(t)}\|^2 \\ s.t. \quad & W\mathbf{1}=\mathbf{1} \end{aligned}$$

  can be solved via Simplex Projection.

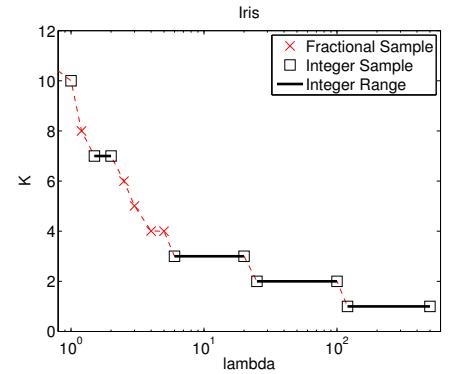
- The 2nd sub-problem:

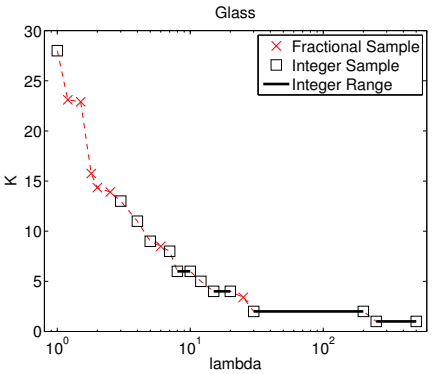
$$W_2^{(t+1)} = \underset{W\in[0,1]^{N\times J}}{argmin} \quad \theta\|W\|_{\mathscr{G}}+\lambda\|W\|_{\infty,1}+Y_2^{(t)}\circ W+\frac{\rho}{2}\|W-Z^{(t)}\|^2.$$

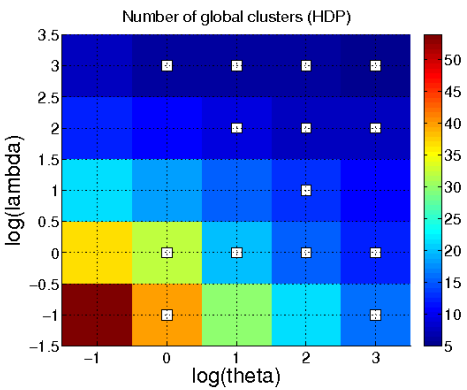
  has closed-form solution via proximal mapping $\mathbf{prox}_\lambda(\mathbf{prox}_\theta(.))$.

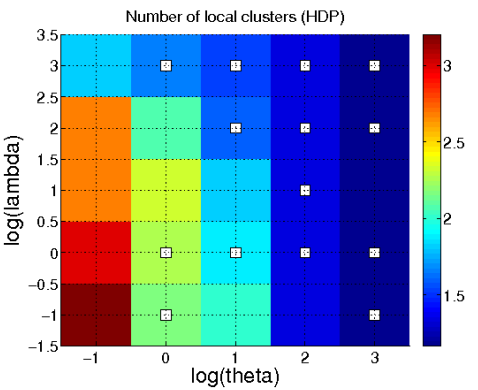
- ADMM Update:

$$\begin{aligned} Z^{(t+1)} &\leftarrow (W_1^{(t+1)}+W_2^{(t+1)})/2 \\ Y_q^{(t+1)} &\leftarrow Y_q^{(t)}+\alpha(W_q^{(t+1)}-Z^{(t+1)}),\ \textit{for } q=1,2 \end{aligned}$$









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)

Thank you!