MAD-Bayes: MAP-based Asymptotic Derivations from Bayes

Tamara Broderick, Brian Kulis, Michael I. Jordan

(ICML 2013)

Discussion by: Piyush Rai

March 28, 2014

Introduction

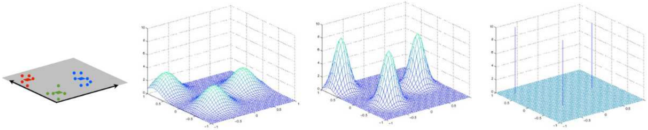
- Inference in NPBayes models can be computationally prohibitive
- Traditional approaches: MCMC, Variational Bayes (VB)
- Standard MCMC and VB can't cope with big data
- Lots of recent efforts on scaling up NPBayes for big data
  - Online/Stochastic methods: Sequential Monte Carlo, Particle MCMC, Stochastic Variational Inference
  - Parallel/Distributed versions of MCMC or VB
  - Point Estimation based methods (e.g., this paper)

Point Estimation for NPBayes

- Can be a quick-and-dirty way of finding a "reasonable" solution
- Point estimates can often be sensible initializers for MCMC/VB
- Some examples of point estimation based methods for NPBayes:
  - Greedy Search for DP mixture model (Wang & Dunson, JCGS 2011)
  - Beam Search for DP mixture model (Daumé III, AISTATS 2007)
  - Beam Search for IBP (Rai & Daumé III, ICML 2011)
  - Submodular Optimization for IBP (Reed & Ghahramani, ICML 2013)
  - Small-variance asymptotics
    - For DP and HDP mixture models (Kulis & Jordan, ICML 2012)
    - For Dependent DP mixture models (Campbell et al, NIPS 2013)
    - For HMM/infinite-HMM (Roychowdhury et al, NIPS 2013)
    - For DP and IBP (**Today's paper**)

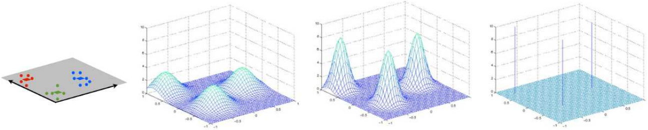
Paper Summary

- Deterministic, efficient, point estimation for two NPBayes models
  - CRP/DP based Gaussian mixture model
  - IBP/Beta-Bernoulli process based linear Gaussian model
- **Original motivation:** EM algorithm for inference in Gaussian mixture model (GMM) behaves like k-means as the mixture variances shrink to zero



Paper Summary

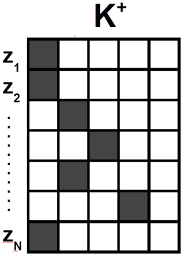
- Deterministic, efficient, point estimation for two NPBayes models
  - CRP/DP based Gaussian mixture model
  - IBP/Beta-Bernoulli process based linear Gaussian model
- **Original motivation:** EM algorithm for inference in Gaussian mixture model (GMM) behaves like k-means as the mixture variances shrink to zero



- **This paper:** Applies low-variance asymptotics on the **MAP objective** instead of on specific inference algorithms (EM, Gibbs sampling)
- **Key observation:** The negative log-likelihood of a GMM approaches k-means objective as the covariances of Gaussians tend to zero
- **Result:** Leads to objectives/algorithms for CRP/IBP that are reminiscent to that of k-means clustering

Clustering

- Consider data $x_1, \ldots, x_N$, where $x_n \in \mathbb{R}^D$
- Assume data can be grouped into K^+ clusters
- Cluster assignments: $z_1, \ldots, z_N$
- $z_{nk} = 1$ if x_n belongs to cluster k ($z_{nk'} = 0 \quad \forall k' \neq k$)



- **Chinese Restaurant Process:** Gives a prior on K^+ and $z_{1:N,1:K^+}$

Chinese Restaurant Process (CRP)

- **Analogy:** Each data point is a customer, each cluster is a table
- First customer sits on a new table
- Customer n sits on
  - An existing table k with probability $\propto S_{n-1,k} = \sum_{m=1}^{n-1} z_{m,k}$
  - A new table with probability $\propto \theta > 0$
- Probability of this clustering (the table assignments $z_{1:N,1:K^+}$)

$$\mathbb{P}(z_{1:N,1:K^+}) = \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k}-1)!$$

- The above is known as Exchangeable Partition Probability Function (EPPF)

CRP Gaussian Mixture Model

- Assume data is generated from a mixture of K^+ Gaussians
- Mixture component k has mean μ_k and variance $\sigma^2 I_D$
- Likelihood: $\mathbb{P}(x|z,\mu) = \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \mathcal{N}or(x_n|\mu_k, \sigma^2 I_D)$
- Prior on component means: $\mathbb{P}(\mu_{1:K^+}) = \prod_{k=1}^{K^+} \mathcal{N}or(\mu_k|0, \rho^2 I_D)$
- Prior on cluster assignments: $\mathbb{P}(z_{1:N,1:K^+}) = \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k}-1)!$
- **Goal:** Find a point estimate of z and μ by maximizing the posterior

$$\underset{K^+,z,\mu}{\operatorname{argmax}} \mathbb{P}(z,\mu|x) \propto \underset{K^+,z,\mu}{\operatorname{argmin}} -\log \mathbb{P}(x,z,\mu)$$

MAP Objective: Small-Variance Asymptotics

- **Goal:** Solve $\text{argmin}_{K^+,z,\mu} - \log \mathbb{P}(x, z, \mu)$
- **Idea:** Take the MAP objective, set $\theta = \exp(-\lambda^2/(2\sigma^2))$ and consider limit $\sigma^2 \to 0$ (note that $\theta \to 0$ as $\sigma^2 \to 0$)

$$\begin{aligned}\mathbb{P}(x,z,\mu) &= \mathbb{P}(x|z,\mu)\mathbb{P}(z)\mathbb{P}(\mu) \\ &= \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \mathcal{N}(x_n|\mu_k,\sigma^2 I_D) \\ &\cdot \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k}-1)! \\ &\cdot \prod_{k=1}^{K^+} \mathcal{N}(\mu_k|0,\rho^2 I_D)\end{aligned}$$

$$\Longrightarrow$$

$$\begin{aligned}&-\log \mathbb{P}(x,z,\mu) \\ &= \sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} \left[O(\log\sigma^2) + \frac{1}{2\sigma^2}\|x_n-\mu_k\|^2 \right] \\ &+ (K^+-1)\frac{\lambda^2}{2\sigma^2} + O(1) \\ &+ O(1)\end{aligned}$$

- Therefore $-2\sigma^2 \log \mathbb{P}(x, z, \mu) =$
$\sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} ||x_n - \mu_k||^2 \quad + \quad (K^+ - 1)\lambda^2 + O(\sigma^2 \log(\sigma^2))$

DP-means algorithm

- Clustering objective function

$$\underset{K^+,z,\mu}{\operatorname{argmin}} \sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} ||x_n - \mu_k||^2 \quad + \quad (K^+ - 1)\lambda^2$$

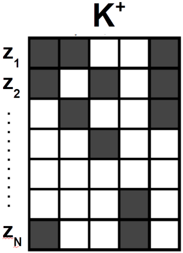
- Just like k-means, except every new cluster incurs λ^2 penalty

The algorithm:

- Iterate until no change
  - For each data point x_n:
    - Compute distance from cluster centers: $d_{nk} = ||x_n - \mu_k||^2, \quad k = 1, \ldots, K$
    - if $\min_k d_{nk} > \lambda^2$, set $K = K + 1$, $z_n = K$, and $\mu_K = x_n$
    - otherwise set $z_n = \operatorname{argmin}_k d_{nk}$
  - Update the cluster centers
- The algorithm converges to a local minimum (just like k-means)

Latent Feature Allocation

- Consider data $x_1, \ldots, x_N$, where $x_n \in \mathbb{R}^D$
- Assume data can be described using K^+ latent features
- Latent feature assignments: $z_1, \ldots, z_N$, where $z_n \in \{0,1\}^{K^+}$
- $z_{nk} = 1$ if latent feature k is present in x_n



- **Indian Buffet Process:** Gives a prior on K^+ and $z_{1:N,1:K^+}$

Indian Buffet Process (IBP)

- Analogy: Each data point is a customer, each latent feature is a dish
- First customer selects $K_1^+ \sim \mathit{Poisson}(\gamma)$ dishes
- Customer n selects
  - An existing dish k with probability $\propto S_{n-1,k} = \sum_{m=1}^{n-1} z_{m,k}$
  - $K_n^+ \sim \mathit{Poisson}(\gamma/n)$ new dishes
- Probability of this latent feature allocation (dish assignments $z_{1:N,1:K^+}$)

$$\mathbb{P}(z_{1:N,1:K^+}) = \frac{\gamma^{K^+} \exp\{-\sum_{n=1}^{N} \frac{\gamma}{n}\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} S_{N,k}^{-1} \binom{N}{S_{N,k}}^{-1}$$

- The above is known as Exchangeable Feature Probability Function (EFPF)

IBP Linear Gaussian Model

- Assume data is additive combination of K^+ latent features $\mu_1, \ldots, \mu_{K^+}$

$$x_n \sim \mathcal{N}or(x_n | \sum_{k=1}^{K^+} z_{nk}\mu_k, \sigma^2 I_D)$$

- Notation: $X = [x_1, \ldots, x_N]^\top, Z = [z_1, \ldots, z_N]^\top, A = [\mu_1, \ldots, \mu_{K^+}]^\top$
- Likelihood: $\mathbb{P}(X|Z, A) = \frac{1}{(2\pi\sigma^2)^{(ND/2)}} \exp\{-\frac{\mathrm{tr}((X-ZA)^\top (X-ZA))}{2\sigma^2}\}$
- Prior on latent feature means $\mathbb{P}(A) = \prod_{k=1}^{K^+} \mathcal{N}or(\mu_k | 0, \rho^2 I_D)$
- Prior on latent feature allocations $\mathbb{P}(Z) = \frac{\gamma^{K^+} \exp\{-\sum_{n=1}^{N} \frac{\gamma}{n}\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} S_{N,k}^{-1} \binom{N}{S_{N,k}}^{-1}$
- **Goal:** Find a point estimate of Z and A by maximizing the posterior

$$\underset{K^+, Z, A}{\mathrm{argmax}}\, \mathbb{P}(Z, A|X) \propto \underset{K^+, Z, A}{\mathrm{argmin}} - \log \mathbb{P}(X, Z, A)$$

MAP Objective: Small-Variance Asymptotics

- **Goal:** Solve $\operatorname{argmin}_{K^+, Z, A} -\log \mathbb{P}(X, Z, A)$
- Set $\gamma = \exp(-\lambda^2/(2\sigma^2))$ for some constant λ^2 and consider limit $\sigma^2 \to 0$

$$
\begin{aligned}
&\mathbb{P}(X, Z, A) = \mathbb{P}(X|Z, A)\mathbb{P}(Z)\mathbb{P}(A) \\
&= \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\} \\
&\quad \cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\quad \cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{-\frac{1}{2\rho^2} A'A\right\}
\end{aligned}
$$

$$
\begin{aligned}
&-\log \mathbb{P}(X, Z, A) \\
&= O(\log \sigma^2) + \frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA)) \\
&+ K^+ \frac{\lambda^2}{2\sigma^2} + \exp(-\lambda^2/(2\sigma^2)) \sum_{n=1}^{N} n^{-1} + O(1) \\
&+ O(1)
\end{aligned}
$$

- Therefore $-2\sigma^2 \log \mathbb{P}(X, Z, A) =$
  $\mathrm{tr}((X - ZA)^\top (X - ZA)) \quad + \quad K^+\lambda^2 + O(\sigma^2 \exp(-\lambda^2/(2\sigma^2))) + O(\sigma^2 \log(\sigma^2))$

BP-means algorithm

- Feature allocation objective function

$$\underset{K^+,Z,A}{\operatorname{argmin}} \operatorname{tr}((X - ZA)^\top (X - ZA)) \quad + \quad K^+\lambda^2$$

The algorithm:

- Iterate until no change
  - For $n = 1, \ldots, N$
    - For $k = 1, \ldots, K^+$, choose the optimal value (0 or 1) of z_{nk}
    - Let Z' equal Z but with one new latent feature $K^+ + 1$ only for this data point (and set $A' = A$ but with an additional row $X_n - Z_n A$)
    - If the triplet $(K^+ + 1, Z', A')$ has lower objective than (K^+, Z, A), replace the latter with the former
  - Set $A \leftarrow (Z^\top Z)^{-1} Z^\top X$
- The algorithm converges to a local minimum (just like k-means)

Extension: Collapsed Objective

- **Idea:** Integrate out the latent feature means A from the likelihood
- Collapsed likelihood $\mathbb{P}(X|Z)$ for the latent feature model

$$\frac{\exp\left\{-\frac{\mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)}{2\sigma^2}\right\}}{(2\pi\sigma^2)^{(ND/2}(\rho^2/\sigma^2)^{K^+D/2}|Z'Z + \frac{\sigma^2}{\rho^2}I_D|^{D/2}}$$

- The objective function after the small-variance asymptotics becomes

$$\operatorname*{argmin}_{K^+,Z} \operatorname{tr}(X^\top(I_N - Z(Z^\top Z)^{-1}Z^\top)X) \quad + \quad K^+\lambda^2$$

- **Note:** a similar objective obtained for the DP clustering case as well

Parametric Objective

- Prior $\mathbb{P}(Z)$ on latent feature allocations (fixed K) is:

$$\prod_{k=1}^{K} \left(\frac{\Gamma(S_{N,k} + \gamma)\Gamma(N - S_{N,k} + 1)}{\Gamma(N + \gamma + 1)} \frac{\Gamma(\gamma + 1)}{\Gamma(\gamma)\Gamma(1)} \right)$$

- Limiting behavior of the MAP objective as $\sigma^2 \to 0$

$$\underset{Z,A}{\operatorname{argmin}} \operatorname{tr}[(X - ZA)^\top (X - ZA)]$$

The K-features algorithm:

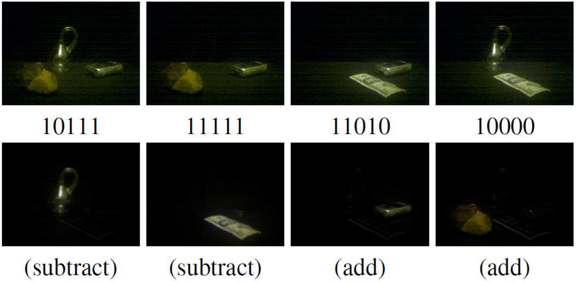
- Repeat until no change
  - For $n = 1, \ldots, N$
    - For $k = 1, \ldots, K$, set z_{nk} to minimize $||x_n - z_{n,1:K} A||^2$
  - Set $A = (Z^\top Z)^{-1} Z^\top X$

Experiments

- Experiments on the IBP based latent feature model
- Datasets considered: Tabletop data and Faces data (both are image datasets)
- Algorithms considered
  - Gibbs sampling for the IBP (Gibbs)
  - BP-means (BP-m)
  - Collapsed BP-means (Collap)
  - Stepwise K-features (FeatK)
- Note: BP-m, Collap, FeatK were run 1000 times with different initializations
- Hyperparameter λ^2 was set to a "reasonable" value

Tabletop Data

- 100 color images, each reduced to a 100 dimensional feature vector via PCA
- Example images and discovered latent features by BP-means

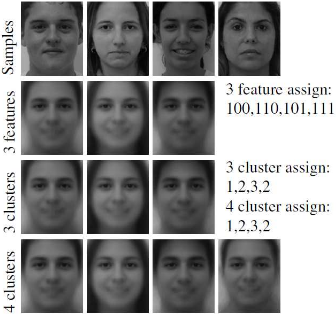


- Timing results and number of inferred latent features

Alg	Per run	Total	#
Gibbs	$8.5 \cdot 10^3$	—	10
Collap	11	$1.1 \cdot 10^4$	5
BP-m	0.36	$3.6 \cdot 10^2$	6
FeatK	0.10	$1.55 \cdot 10^2$	5

Faces Data

- 400 images of 200 people (2 images for each person - neutral and smiling), each image reduced to a 100 dimensional feature vector via PCA
- Only stepwise K-features and K-means compared



- A simple demonstration of why K latent features may be better than K (or more) clusters

Conclusions

- Point estimation algorithms via small-variance asymptotics on the MAP objective functions
- Resulting objectives are akin to those used in other model-selection methods such as AIC
- Algorithms similar to k-means, easy to implement (and have potential for some parallelization)
- Initialization can be an issue (tricks from k-means literature can be used)
- The algorithm critically depends on λ. It is unclear how to set it.

Thanks!