MAD-Bayes: MAP-based Asymptotic Derivations from Bayes

Tamara Broderick
UC Berkeley
tab@stat.berkeley.edu

Brian Kulis
Ohio State University
kulis@cse.ohio-state.edu

Michael I. Jordan
UC Berkeley
jordan@eecs.berkeley.edu

T. Broderick   B. Kulis   M. I. Jordan

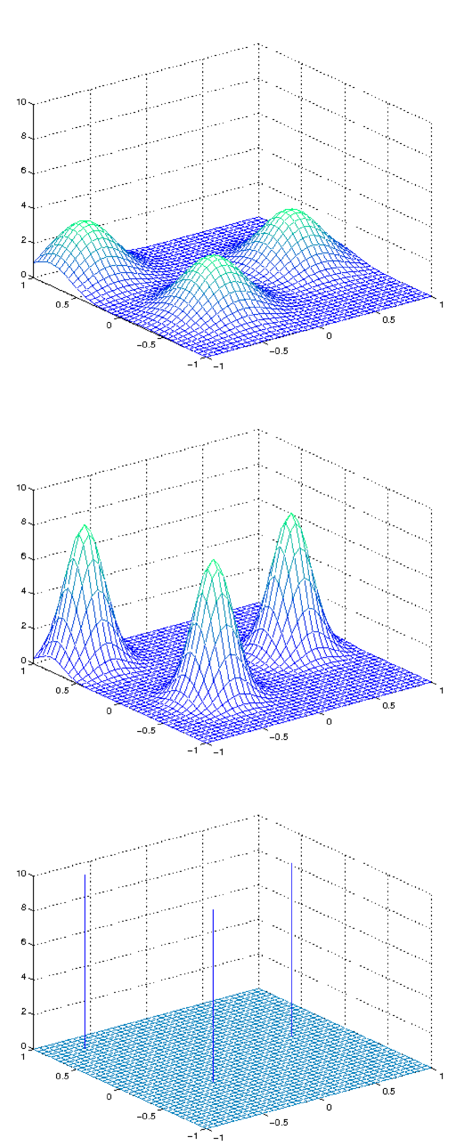
Background

- Finite mixture of Gaussians model with cluster-variance σ^2
  - Taking $\sigma^2 \to 0$, the negative log-likelihood of the mixture of Gaussians model approaches the K-means clustering objective
  - Taking $\sigma^2 \to 0$, the EM algorithm approaches the K-means clustering algorithm
- Dirichlet process (DP) mixture of Gaussians model with cluster-variance σ^2
  - Taking $\sigma^2 \to 0$, the Gibbs sampler approaches the DP-means clustering algorithm [2]

Our contributions

- We show that the DP-means objective can be obtained directly from the posterior, independent of any inference algorithm
- We show that this expanded perspective on *small-variance asymptotics* generalizes to a range of models beyond the DP mixture
- In particular, we find a K-means-like objective for *features*, a generalization of clusters that relaxes the exclusivity and exhaustivity assumptions
  - We apply small-variance asymptotics to the beta process (BP) with Bernoulli likelihood (equivalent to the Indian buffet process) with linear Gaussian likelihood to obtain a K-means-like objective for features: *BP-means*
- We show empirical results for BP-means

Small variance asymptotics: a cartoon



- We consider likelihood models that are Gaussian around some mean determined by the underlying combinatorial structure (e.g., clusters or features).
- *Small-variance asymptotics* takes the variance of these Gaussians to zero.
- We examine the effects of these limits on the model likelihood.

References

[1] T. Griffiths and Z. Ghahramani. The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, 12(April):1185–1224, 2011.

[2] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 23rd International Conference on Machine Learning*, 2012.

[3] C. E. Thomaz and G. A. Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902–913, June 2010. We use files `http://fei.edu.br/~cet/frontalimages_spatiallynormalized_partX.zip` with `X=1,2`.

DP-means objective

- Notation.
  - N data points x_n, each with dimension D.
  - $z_{nk} = 1$ if data point n belongs to cluster k and zero else.
  - K^+ is number of clusters (from generative model; not fixed).
  - μ_k is mean of cluster k.
  - λ^2 is a constant.
- Generative model: DP(θ) mixture of Gaussians with σ^2 variance.
- Small-variance limit.
  - $\operatorname{argmax}_{z,K^+,\mu} \mathbb{P}(z, \mu | x) = \operatorname{argmin}_{z,K^+,\mu} -2\sigma^2 \log \mathbb{P}(z, \mu, x)$
  - Taking $\sigma^2 \to 0$ and $\theta = \exp(-\lambda^2/2\sigma^2)$ yields DP-means problem:

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

BP-means objective

- Notation.
  - $z_{nk} = 1$ if data point n belongs to feature k and zero else.
  - μ_k is mean of feature k.
  - K^+ is number of features (from generative model; not fixed).
  - X is $N \times D$ matrix of the x_n; Z is $N \times K^+$ matrix of the z_n; A is $K^+ \times D$ matrix of the μ_k.
  - λ^2 is a constant.
- Generative model: BP/IBP(γ) features; linear-Gaussian likelihood with σ^2 variance
- Small-variance limit.
  - $\operatorname{argmax}_{Z,K^+,A} \mathbb{P}(Z, A | X) = \operatorname{argmin}_{Z,K^+,A} -2\sigma^2 \log \mathbb{P}(Z, A, X)$
  - Taking $\sigma^2 \to 0$ and $\gamma = \exp(-\lambda^2/2\sigma^2)$ yields BP-means objective:

$$\underset{Z,K^+,A}{\operatorname{argmin}} \boxed{\mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2}$$

BP-means algorithm

Iterate until no changes are made:

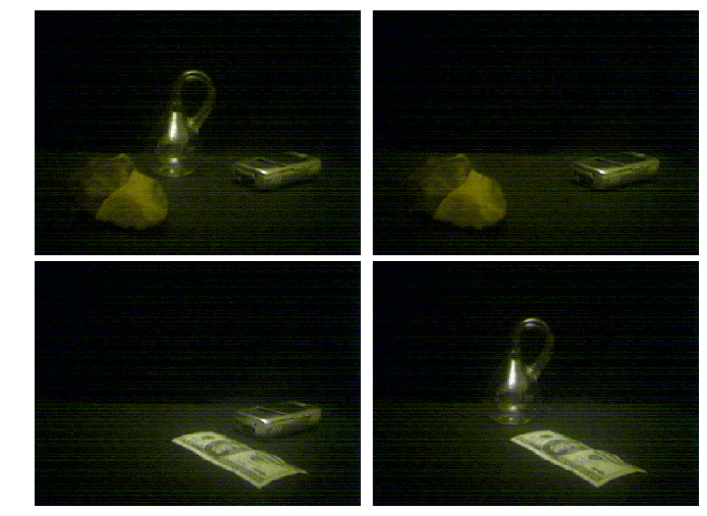
1. For $n = 1, \ldots, N$
   - For $k = 1, \ldots, K^+$, choose the optimal value (0 or 1) of z_{nk}.
   - Let Z' equal Z but with one new feature (labeled $K^+ + 1$) containing only data index n. Set $A' = A$ but with one new row: $A'_{K^+ + 1, \cdot} \leftarrow X_{n,\cdot} - Z_{n,\cdot} A$.
   - If the triplet $(K^+ + 1, Z', A')$ lowers the objective from the triplet (K^+, Z, A), replace the latter triplet with the former.
2. Set $A \leftarrow (Z'Z)^{-1} Z'X$.

Other objectives

Other feature models yield the *collapsed BP-means* and the finite *K-features* objectives $\mathbf{tr}[(X - ZA)'(X - ZA)]$. Let *stepwise K-features* denote dynamically solving the latter problem for each fixed K then iteratively incrementing K by one until the BP-means objective is not improved.

Tabletop photos and features

Data: 100 JPEG $240 \times 320 \times 3$ photos [1]; four sample photos at right.

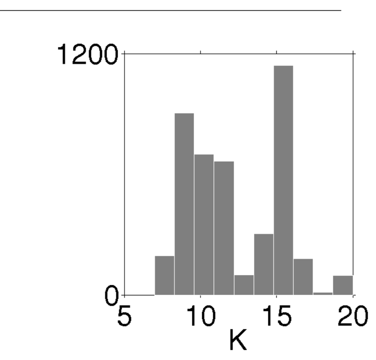


Stepwise K-features with $\lambda = 1$ identifies 5 features: the table and these four objects. The upper two features are subtracted; the lower two are added.

BP-means results: Tabletop photos

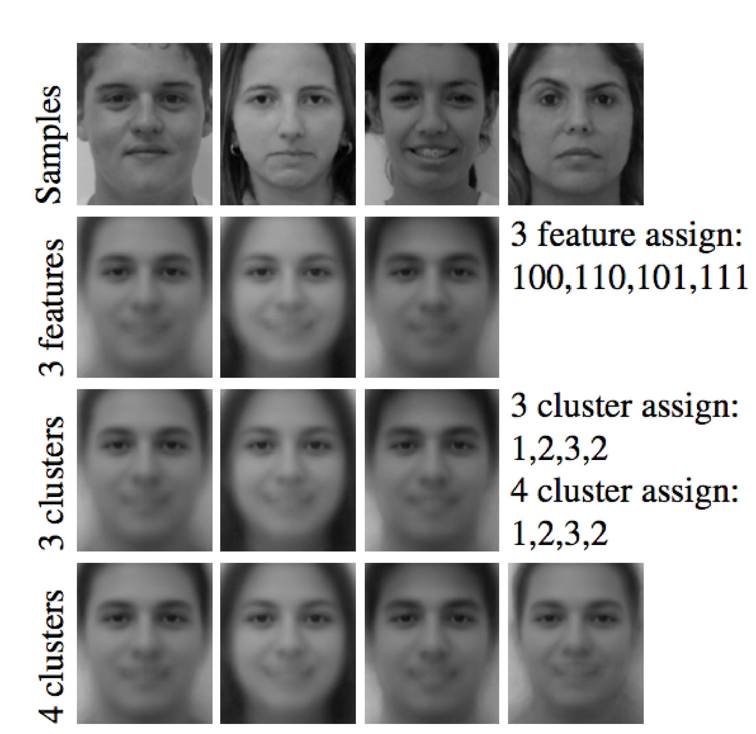
We compare an IBP Gibbs sampler [1], collapsed BP-means (Collap), the basic BP-means algorithm, and stepwise K-features (FeatK).

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



Above Left: *First column*: run time per run in sec. *Second column*: total running time (i.e., over multiple repeated runs for the final three). *Third column*: final number of features learned (the IBP # is stable for > 900 final iterations). *Above Right*: Histogram of collections of the final K values found by the IBP for a variety of initializations and parameter starting values.

BP-means results: Face photos



Row 1: 4 sample photos in a set of 400 [3]. *Rows 2*: Three features and assignments found using the BP-means objective. *Row 3*: Cluster centers and assignments using K-means with $K = 3$. *Row 4*: Same with $K = 4$.