MAP for Exponential Family Dirichlet Process Mixture Models

Yordan P. Raykov
Department of Mathematics
Aston University
United Kingdom
yordan.raykov@gmail.com

Alexis Boukouvalas
Department of Molecular Sciences
University of Manchester
United Kingdom
alexis.boukouvalas@gmail.com

Max A. Little
Media Lab
Massachussetts Institute of Technology
United States
and
Department of Mathematics
Aston University
United Kingdom
max.little@mit.edu

Abstract

The *Dirichlet process mixture* (DPM) is a ubiquitous, flexible Bayesian nonparametric model. However, full probabilistic inference in this model is analytically intractable, so that computationally intensive techniques such as Gibb's sampling are required. As a result, DPM-based methods, which have considerable potential, are restricted to applications in which computational resources and time for inference is plentiful. For example, they would not be practical for digital signal processing on embedded hardware, where computational resources are at a serious premium. Here we develop simplified yet statistically rigorous approximate *maximum a-posteriori* (MAP) inference algorithms for DPMs, which we call *MAP-DP*. This algorithm is as simple as K-means clustering, performs in experiments as well as Gibb's sampling, while requiring only a fraction of the computational effort. Unlike related *small variance asymptotics*, our algorithm is non-degenerate and so inherits the "rich get richer" property of the Dirichlet process. It also retains a non-degenerate, closed-form likelihood which enables standard tools such as cross-validation to be used. This is a well-posed approximation to the MAP solution of the probabilistic DPM model.

1 Introduction

Bayesian nonparametric (BNP) models have been successfully applied to a wide range of domains but despite significant improvements in computational hardware, statistical inference in most BNP models remains unfeasible in the context of large datasets, or for moderate-sized datasets where computational resources are limited. The flexibility gained by such models is paid for with severe decreases in computational efficiency, and this makes these models somewhat impractical. Therefore, there is an emerging need for approaches that simultaneously minimize both empirical risk and computational complexity [3]. Typically, model inference is performed using Markov-chain Monte Carlo (MCMC) techniques, but their slow convergence and computational requirements can

severely limit the range of applications. That is often why we seek to convert the expensive problem of inference into the cheaper one of optimization.

One of the most popular ways to do so is *Variational Bayes* (VB). This relies on optimizing a lower bound of the complete data likelihood with respect to the posterior of the model parameters. To obtain good scaling, typically the range of posteriors over which optimization is performed is restricted, where restrictions often take the form of factorization assumptions ([10],[7]). As a result, by maximizing only the lower bound of the likelihood, VB techniques often lead to severe underestimates of the variance of the posterior and may require a significant number of iterations to converge. Although variational techniques are much less expensive than MCMC approaches, an optimization schedule is usually needed for each update of the variational bound which will make them, in general, quite slow. In the nonparametric setting, variational methods were first applied to DPMs in [2] where the variational distribution is not available in closed form and is approximated. Later, an approximation-free method was proposed in [19], but this requires additional Gibb's steps that reduce scalability.

[6] describe a related approach based on a combinatorial search that is guaranteed to find the mode for computationally tractable objective functions. As the DPM complete data likelihood is computationally intractable, their algorithm is only approximate, and this also makes it sample-order dependent. [5] on the other hand describes an algorithm that is guaranteed to find the global mode in $N\left(N+1\right)$ computations, but only in the case of univariate product partition models with non-overlapping clusters. Our approach does not make any further assumptions outside of the model structure and being derived from the Gibbs sampler, does not suffer from sample-order dependency. [20] present another approach for fast inference in DPMs that discards the exchangeability assumption of the data partitioning and instead assume the data is in the correct ordering. Then a greedy, repeated "uniform resequencing" is proposed to maximize a pseudo-likelihood that approximates the DPM complete data likelihood. The suggested procedure does not have any guarantees for convergence even to a local solution.

[12] introduced an optimization approach, *DP-means*, later generalized by [4] to IBPs and by [9] to general exponential family mixtures. By shrinking the variance of the posterior of BNP models to zero, a fast optimization schedule is obtained from the resulting degenerate likelihood. The approach builds upon the connection between K-means and Gaussian mixture models [1, page 443] and extends it to the nonparametric setting. However, in applying *small variance asymptotic* (SVA) reasoning, it breaks some of the key properties of the underlying model, leading to a degenerate likelihood. Additionally, SVA applied to DPMs [8] loses the sequential reinforcement effect of the infinite clustering, as the prior term over the partition drops from the likelihood. Also, the input data is modeled as spherical. SVA techniques always lead to degeneracy in the likelihood and so any kind of out-of-sample prediction or cross-validation is impossible.

In this report, we present an algorithm, MAP-DP, for finding the solution of the MAP problem posed in [4] without resorting to a degenerate likelihood. Our algorithm is thus more faithful to inference in the corresponding probabilistic model, and also allows the use of standard statistical tools such as out-of-sample prediction for cross-validation. In Section 2 we present the new MAP-DP algorithm for exponential family distributions. In Section 3 we compare its performance against DP-means, variational DP and a collapsed Gibb's reference implementation. We conclude with a summary in Section 4.

2 MAP-DP: MAP for Exponential family Dirichlet process mixture models

Here, we propose a DPM inference algorithm based on *iterated conditional modes* (ICM, see [11] and also [1, page 546]). This is also called the *maximization-maximization* (M-M) algorithm by [21]. The basic idea is to use conditional modal point estimates rather than samples from the conditional probabilities as in Gibb's. The resulting algorithm is nearly as simple as the SVA approach in terms of both implementation and computation complexity, whilst retaining a non-degenerate likelihood. As a result, this method keeps the reinforcement effect of the DP and allows for out-of sample estimation and principled model selection [16]. We use the collapsed *Dirichlet process mixture model* with conjugate priors [13, 16]:

$$\begin{aligned} z_1, \ldots, z_N &\sim \text{CRP}\left(N_0, N\right) \\ \eta_k &\sim G_0 \\ \mathbf{x}_i &\sim F\left(\mathbf{x}_i | \eta_{z_i}\right) \end{aligned} \tag{1}$$

for $k = 1, \ldots, K$, and N_0 being the *Chinese restaurant process* (CRP) *concentration parameter*, and N the dataset size. The likelihood is $p\left(\mathbf{x}_i | \eta, z_i\right) \propto \prod_{k=1}^{K} \exp\left[\eta_k^T \boldsymbol{g}\left(\mathbf{x}_i\right) - \boldsymbol{a}\left(\eta_k\right)\right]^{\delta(z_i,k)}$ with the corresponding conjugate prior $p\left(\eta | \mathcal{X}, \nu\right) \propto \exp\left[\eta^T \mathcal{X} - \nu \boldsymbol{a}\left(\eta\right) - \boldsymbol{a}_0\left(\mathcal{X}, \nu\right)\right]$ where K denotes the unknown number of components of the CRP that have points assigned, $\delta\left(z_i, k\right)$ is the Kronecker delta function, η the natural parameters and $\boldsymbol{a}$, $\boldsymbol{a}_0$ are the log partition functions of the likelihood and prior, respectively. For each observation $\mathbf{x}_i$, we compute the negative log probability for each existing cluster k and for a new cluster $K+1$:

$$q_{i,k} = -\log p\left(\mathbf{x}_i | z_{-i}, \mathbf{X}_{-i}, z_i = k, \nu, \mathcal{X}\right) \tag{2}$$

$$q_{i,K+1} = -\log p\left(\mathbf{x}_i | \nu, \mathcal{X}\right) \tag{3}$$

where the natural parameters have been integrated out and we have omitted terms independent of k. For each observation $\mathbf{x}_i$ we compute the above $K+1$-dimensional vector $\boldsymbol{q}_i$ and select the cluster number according to the following scheme:

$$z_i = \underset{k \in \{1, \ldots, K, K+1\}}{\arg\min} \left[q_{i,k} - \log N_k^{-i}\right]$$

where N_k^{-i} is the number of data points assigned to cluster k, excluding data point $\mathbf{x}_i$ and, for notational convenience, we define $N_{K+1}^{-i} \equiv N_0$. The algorithm proceeds to the next observation $\mathbf{x}_{i+1}$ by updating the cluster component statistics to reflect the new value of the cluster assignment z_i and remove the effect of data point $\mathbf{x}_{i+1}$. To check convergence of the algorithm we compute the negative log of the complete data likelihood:

$$p\left(\mathbf{x}, z | N_0\right) = \left(\prod_{i=1}^{N} \prod_{k=1}^{K} p\left(\mathbf{x}_i | z_{-i}\right)^{\delta(z_i,k)}\right) p\left(z_1, \ldots, z_N\right) \tag{4}$$

where $p\left(z_1, \ldots z_N\right)$ is the CRP partition function [14]. We show in Algorithm 1 all the steps involved in optimizing the complete data likelihood.

Algorithm 1: MAP-DP: Exponential Families

Input: $\mathbf{x}_1, \ldots, \mathbf{x}_N$: data; $N_0 > 0$: concentration parameter, $\epsilon > 0$: convergence threshold; $(\mathcal{X}, \nu)$: cluster prior parameters; $\boldsymbol{a}_0(.)$: prior log partition function; $\boldsymbol{g}(.)$: sufficient statistics.
Output: $z_1, \ldots, z_N$: cluster assignments, K: number of clusters.

$K = 1, z_i = 1$, for all $i \in 1, \ldots, N$
$E_{\text{new}} = \infty$
repeat
  $E_{\text{old}} = E_{\text{new}}$
  for $i \in 1, \ldots, N$ **do**
    for $k \in 1, \ldots, K$ **do**
      $q_{i,k} = \boldsymbol{a}_0\left(\mathcal{X} + \sum_{j:z_j=k, j\neq i} g\left(\mathbf{x}_j\right), \nu + N_k^{-i}\right) - \boldsymbol{a}_0\left(\mathcal{X} + \sum_{j:z_j=k} g\left(\mathbf{x}_i\right), \nu + N_k\right)$
    $q_{i,K+1} = \boldsymbol{a}_0\left(\mathcal{X}, \nu\right) - \boldsymbol{a}_0\left(\mathcal{X} + g(\mathbf{x}_i), \nu + 1\right)$
    $z_i = \arg\min_{k \in 1, \ldots, K, K+1} \left(q_{i,k} - \log N_k^{-i}\right)$
    if $z_i = K + 1$ **then**
      $K = K + 1$
  $E_{\text{new}} = \sum_{k=1}^{K} \sum_{i:z_i=k} q_{i,k} - K \log N_0 - \sum_{k=1}^{K} \log \Gamma\left(N_k\right)$
until $E_{\text{old}} - E_{\text{new}} < \epsilon$

3 Synthetic CRP parameter estimation

We examine the performance of the MAP-DP, collapsed Gibbs, DP-means [12] and variational DP [2] on CRP-partitioned non-spherical Gaussian data in terms of estimation error and the computational effort needed. We measure clustering estimation accuracy using the sum *normalized mutual*

Table 1: Performance of clustering algorithms on the CRP mixture experiment (Section 3). Mean and standard deviation (in brackets) reported across the 100 CRP mixture samples. The range of the NMI is $[0, 1]$ with higher values reflecting lower clustering error.

	Gibbs-MAP	MAP-DPM	DP-Means	Variational
Training set NMI	0.81 (0.1)	0.82 (0.1)	0.68 (0.1)	0.75 (0.1)
Iterations	1395 (651)	10 (3)	18 (7)	45 (18)

information (NMI) [18]. The control parameter for the DP-Means algorithm is set using a binary search procedure such that the algorithm gives rise to the correct number of partitions. This approach favors the DP-means algorithm as it is given knowledge of the true number of clusters which is not available for the other approaches. For the variational DP we set the truncation limit to be ten times the number of clusters in the current CRP sample.

In Table 1 a range of performance metrics are shown. Both MAP-DPM and Gibbs achieve similar clustering performance in terms of NMI whilst variational and DP-means have lower scores. The MAP-DPM algorithm requires the smallest number of iterations to converge with the Gibbs sampler requiring, on average, 140 more iterations and DP-means 1.8 times. In Figure 1 the median partitioning is shown in terms of the partitioning N_k/N and the number of clusters. MAP-DPM and variational DP fail to identify the smaller clusters whereas the Gibbs sampler is able to do so to a greater extent. This is a form of underfitting where the algorithm captures the mode of the partitioning distribution but fails to put enough mass on the tails (the smaller clusters). The NMI scores do not reflect this effect as the impact of the smaller clusters on the overall measure is minimal. The poorer performance of the DP-means algorithm can be attributed to the non-spherical nature of the data as well as the lack of reinforcement effect that leads to underestimation of the larger clusters and overestimation of the smaller clusters.

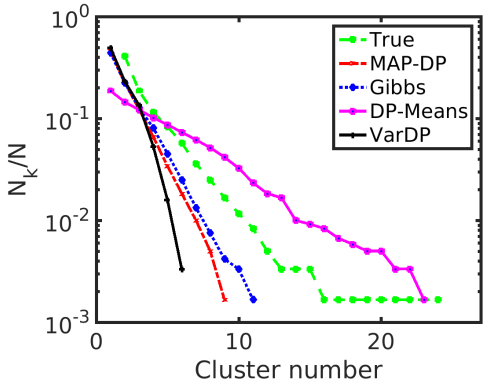

Figure 1: CRP mixture experiment; distribution of cluster sizes, actual and estimated using different methods. Cluster number ordered by decreasing size (horizontal axis) vs $\frac{N_k}{N}$ (vertical axis).

4 Conclusions

We have presented a simple algorithm for inference in DPMs based on non-degenerate MAP, and demonstrated its efficiency and accuracy by comparison to the ubiquitous Gibb's sampler, and a simple alternative, the SVA approach. Our algorithm can be readily applied wherever Gibb's sampling is applicable and the base measure is conjugate. The generality and simplicity of our approach makes it reasonable to adapt to other BNP models, for example the Pitman-Yor process which generalizes the DP [15] and the hierarchical DP [17]. Another useful direction, for large-scale datasets in particular, would be to extend our approach to perform inference that does not need to sweep through the entire data set in each iteration, for increased efficiency [22].

References

[1] Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 1, 2

[2] David Blei and Michael Jordan. Variational methods for the Dirichlet process. In *Proceedings of the 21th International Conference on Machine Learning (ICML)*, page 12. ACM, 2004. 1, 3

[3] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008. 1

[4] Tamara Broderick, Brian Kulis, and Michael Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 28(3):226–234, 2013. 1

[5] David B Dahl et al. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243–264, 2009. 1

[6] Daumé Hal. Fast search for Dirichlet process mixture models. In *Artificial Intelligence and Statistics, San Juan, Puerto Rico*, 2007. 1

[7] Tommi S. Jaakkola. Tutorial on variational approximation methods. In *IN ADVANCED MEAN FIELD METHODS: THEORY AND PRACTICE*, pages 129–159. MIT Press, 2000. 1

[8] Ke Jiang, Brian Kulis, and Michael Jordan. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3158–3166, 2012. 1

[9] Ke Jiang, Brian Kulis, and Michael I Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3158–3166, 2012. 1

[10] Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, USA, 1999. 1

[11] Josef Kittler and Janos Föglein. Contextual classification of multispectral pixel data. *Image and Vision Computing*, 2(1):13–29, 1984. 2

[12] Brian Kulis and Michael Jordan. Revisiting K-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 513–520, 2012. 1, 3

[13] Radford Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. 2

[14] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995. 2

[15] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997. 4

[16] Yordan Raykov, Alexis Boukouvalas, and Max A. Little. Simple approximate MAP Inference for Dirichlet processes. *arXiv:1411.0939*, 2014. 2

[17] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006. 4

[18] Nguyen Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010. 3

[19] Chong Wang and David M. Blei. Truncation-free online variational inference for bayesian nonparametric models. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 413–421. Curran Associates, Inc., 2012. 1

[20] Chong Wang, John William Paisley, and David M. Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 752–760, 2011. 1

[21] Max Welling and Kenichi Kurihara. Bayesian K-Means as a Maximization-Expectation algorithm. In *SDM*, pages 474–478. SIAM, 2006. 2

[22] Max Welling and Yee Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011. 4