Iterative collapsed MAP inference for Bayesian nonparametrics

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
maxl@mit.edu

Abstract

Despite the significant practical potential of Bayesian nonparametric (BNP) models, since they require computationally intensive inference, they are often restricted to applications where computational resources and time for inference are plentiful. Here, we study simple, yet rigorous and efficient, maximum a-posteriori (MAP) inference algorithms for BNP models. We show example algorithms that are as simple as K-means in clustering and Viterbi decoding in hidden Markov modelling, often performing as well as Gibbs sampling, while requiring only a fraction of the computational effort. Unlike related small variance asymptotics, these algorithms are non-degenerate, and so retain important properties such as the "rich get richer" property of the DP and hierarchical DP, and a closed-form likelihood enabling standard statistical methods such as cross-validation to be used. We present an application of the approach to the analysis of DNA copy number variation data.

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 infeasible 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. This is an important example of the emerging need for approaches to inference that simultaneously minimize both empirical risk and computational complexity [4]. Towards that end we present a simple, statistically rigorous and computationally efficient approach for the estimation of BNP models that will significantly reduce the computational burden involved, while keeping most of the model properties intact. In this work, to provide concrete examples, we concentrate on inference in *Dirichlet process mixture models* (DPMM) and its extension to sequential data, the *infinite hidden Markov* *model* (iHMM)[1](generalized as the HDP-HMM in [18]), but our arguments are general and can be extended to most BNP models.

Inference in probabilistic models for which closed-form statistical estimation is intractable, is often performed using computationally demanding *Markov-chain Monte Carlo* (MCMC) techniques ([13],[18],[19]), that simulate the distribution of the model parameters given the data. Despite the asymptotic convergence guarantees of MCMC, in practice MCMC often takes too long to converge and this can severely limit the range of applications. A popular alternative is to cast the inference problem as an optimization problem for which *variational Bayes* (VB) techniques can be used. [3] first introduced VB inference for the DPMM, but it involves truncating the variational distribution of the joint DPMM posterior. Later, collapsed variational methods [17] reduce the inevitable truncation error by working in a reduced-dimensional parameters space, but they are based on a sophisticated family of marginal likelihood bounds for which optimization is challenging. Streaming variational methods [5] obtain significant scaling by optimizing local variational bounds on batches of data visiting data points only once, but as a result they can easily become trapped at a poor fixed point. Similarly, stochastic variational methods [20] also allow for a single pass through the data, but sensitivity to initial conditions increases substantially. Alternatively, methods which learn memoized statistics of the data in a single pass ([9],[8]), have recently shown great promise.

While MCMC methods simulate the posterior of the parameters and VB techniques learn the model parameters that maximize some lower bound on the marginal likelihood, for many applications the posterior distribution of the parameters is not of explicit interest, only their mode. Indeed, directly finding a good approximation of the maximum-a-posterior (MAP) value of the parameters is a much simpler problem than estimating the entire posterior distribution. For example, [7] addresses this MAP problem with a combinatorial search approach but only for computationally tractable objective functions. Since the DPMM complete data likelihood is computationally intractable, their algorithm is only approximate, and this also makes it sample-order dependent. [21] presents another approach to fast inference for the DPMM that discards the exchangeability assumption of the data partitioning and instead assumes the data is in the correct ordering. Then a greedy, repeated "uniform resequencing" is proposed to maximize a pseudo-likelihood that approximates the DPMM complete data likelihood. The suggested procedure does not have any guarantees for convergence even to a local solution.

[6] proposes a more general approach to solving the MAP problem for wider set of models by forcing the likelihood variance of BNP models to zero. By making some additional simplifying assumptions, the approach reduces MCMC updates to a fast optimization algorithm that converges quickly to the approximate MAP solution. However, this *small variance asymptotic* (SVA) reasoning breaks many of the key properties of the underlying probabilistic model: SVA applied to DPMM ([12],[10]) loses the sequential reinforcement effect of the infinite clustering, as the prior term over the partition drops from the likelihood; and degeneracy in the likelihood forbids any kind of out-of-sample prediction and thus, for example, cross-validation. [15] suggests somewhat more flexible SVA assumptions to derive an optimization algorithm for inference in the iHMM. But although the approach overcomes some of the drawbacks of SVA [6], this algorithm completely departs from the assumptions of the underlying probabilistic graphical model. The method is shown to be efficient for clustering time dependent data, but essentially no longer has an underlying probabilistic model. Furthermore, [15] demonstrates that there is more than one way of applying SVA assumptions to a given probabilistic model, and therefore, by making different SVA assumptions, one obtains entirely different inference algorithms that emphasise different structures in the data, even though the underlying probabilistic model remains the same. For example, HDP-means [10] in the context of time series, and an alternative SVA approach[15], optimize very different objective functions.

In this work, we present a simple scheme for finding the approximate solution of the MAP problem for BNP models without resorting to a degenerate likelihood. Our approach does not make any further assumptions beyond the model structure and being derived from the Gibbs sampler, does not suffer from sample-order dependency. This enables the presented algorithms to be more faithful to inference in the corresponding probabilistic models, and also enables the use of standard statistical tools such as out-of-sample prediction. We show how the *Chinese restaurant process* (CRP) may be exploited to produce simplified MAP inference algorithm for the fully collapsed DPMM and direct assignment representation from [18] may be used for efficient MAP inference in HDP-HMM. We now briefly discuss our approach to DPMM (Section 2) and iHMM (Section 3), followed by an example application of our MAP-iHMM in (Section 4).

2 Collapsed MAP inference for Dirichlet process mixture models

Here we show how to learn DPMMs by exploiting *iterated conditional modes* (ICM, see [11] and also [2, page 546]) and variable collapsing. The basic idea is to use conditional modal point estimates rather than samples from the conditional probabilities as in the popular Gibbs sampler. By applying MAP inference to the fully collapsed DPMM using the CRP to model the cluster indicators $z_1, \ldots, z_N$, the resulting algorithm finds good clustering solutions by integrating out the cluster parameters. Integration over the cluster parameters and mixture weights marginalizes over their uncertainty by introducing additional dependencies in the resulting graphical model. While these dependencies increase the computational effort per iteration, they substantially reduce the number of iterations required to reach convergence. The resulting algorithm is nearly as simple as the SVA approach in terms of both implementation and computational complexity, but it keeps the reinforcement effect of the DP, and retains a non-degenerate likelihood allowing for out-of sample estimation and principled model selection [14].

Assuming the data likelihood is an exponential family distribution and we have conjugate priors over the cluster parameters, this can be written in the form $p(\eta|\mathcal{X},\nu) \propto \exp\left[\eta^T\mathcal{X} - \nu\boldsymbol{a}(\eta) - \boldsymbol{a}_0(\mathcal{X},\nu)\right]$, where η, ν are the canonical parameters and where $\boldsymbol{a}$,$\boldsymbol{a}_0$ 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 $K+1$:

$$q_{i,k} = -\log p(\mathbf{x}_i|z_j = k, \nu, \mathcal{X}) \tag{1}$$

$$q_{i,K+1} = -\log p(\mathbf{x}_i|\nu, \mathcal{X}) \tag{2}$$

where the canonical parameters have been integrated out and we have omitted terms independent of k. Then 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 optimization step:

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

where N_k^{-i} is the number of data points assigned to cluster k, excluding data point $\mathbf{x}_i$ and $N_{K+1}^{-i} \equiv N_0$ denotes the concentration parameter of the underlying DP.

3 Collapsed MAP for the infinite hidden Markov model

Here we extend the approach above to address clustering of sequential data. One approach uses a *hierarchical Dirichlet process* (HDP) prior [18] applied to the transition matrix of the popular HMM, thereby obtaining the iHMM [1]. The standard HMM can be viewed as a generalization of the finite mixture model, where the exchangeability assumption between the cluster indicators is replaced with *Markov assumption*.

In this work we present a MAP-based inference method for the iHMM with collapsed local mixture weights, but retained shared base measure atoms, or we exploit the direct assignment construction of the HDP prior, again collapsing out the cluster parameters. The resulting MAP scheme mirrors the MAP-DP approach for the DPMM above. When $z_{t-1} \neq k$ and $z_{t+1} \neq k$, we can compute the state indicator for time point t by maximizing the following conditional:

$$p(z_t = k\,|z_{-t}, \mathbf{X}) \propto \begin{cases} \left(N_{z_{t-1}k} + \alpha\beta_k\right)\left(N_{kz_{t+1}} + \alpha\beta_{z_{t+1}}\right) p(\mathbf{x}_t|z_t = k, \nu, \mathcal{X}) & \text{for existing } k \\ N_0\beta_k\beta_{z_{t+1}} p(\mathbf{x}_t|\nu, \mathcal{X}) & \text{for } k = K+1 \end{cases} \tag{3}$$

while in cases where $z_{t-1} = k$ and/or $z_{t+1} = k$ slight adjustment has to be made to the corresponding transition counts and normalizing constants. Here $\beta_1, \ldots, \beta_K, \beta_{K+1}$ are draws from the global DP, $N_{z_{t-1}k}$ counts the number of times transition has occurred from the state of point $t-1$ to state k, $N_{kz_{t+1}}$ counts the number of times transition has occurred from state k to state pointed by z_{t+1} and N_0 is the concentration parameter of the local DP. Note that the emission probabilities $p(\mathbf{x}_t|z_j = k, \nu, \mathcal{X})$ and $p(\mathbf{x}_i|\nu, \mathcal{X})$ are computed in the same way as in (1) and (2) respectively.

Next we proceed by re-computing the shared base measure parameters $\boldsymbol{\beta}$. The global DP is parameterized by a concentration parameter M_0 and the counts $M_1, \dots, M_K$ where M_k counts how many times state k has been chosen from the global DP. To update those we use m_{rk} to count how many times transition to state k has been sampled using the global DP from some state r, obviously having $M_k = \sum_r m_{rk}$. To compute the counts m_{rk} on the other hand we can use the CRP conditional of z_t given the preceding state:

$$p\left(z_t = k \left| Z_{rk}\right.\right) \propto \begin{cases} N_{rk}^{-z_t} & \text{we use existing } k \\ \alpha\beta_k & \text{we draw } k \text{ from global DP} \end{cases} \tag{4}$$

where $Z_{rk} = \{z_i : z_{i-1} = r, z_i = k\}_{i=1}^{n}$. In sequential manner, if we increase the number of considered points n in (4) from 1 to N we can count how many times the state indicator z_t would be sampled from the global DP. Once the counts m_{rk} are computed, we evaluate the counts $M_1, \dots, M_K$ and compute the base measure parameters $\beta_1, \dots, \beta_K, \beta_{K+1}$ from the mode of the corresponding Dirichlet posterior with distribution Dir $(M_1, \dots, M_K, M_0)$.

4 Genomic hybridization and DNA copy number variation

We examine the performance of the collapsed MAP-iHMM algorithm for segmentation of time series data that is assumed to have approximately piecewise constant behavior. The data is collected from DNA copy number ratios from the genomic hybridization study in [16]. We assume that the noise is independent and stationary, but that there are a few outliers. Further, we suppose we are interested in detecting a few, large jumps between different copy-number regimes.

Reconstruction using MAP-iHMM is shown in Figure 1. As expected, for fixed prior over the state parameters, different choices of the concentration parameters N_0 and M_0 return piecewise constant reconstructions of the underlying copy number variations at different levels of detail, with a few jumps breaking the genome sequence into different groups. We use *Normalize Mutual Information* (NMI) to compare the obtained MAP-iHMM solution with the MAP solution provided by corresponding Gibbs sampler. MAP-iHMM scored similarity of NMI=0.58 after 24 iterations, while Gibbs sampler was ran for 2500 iterations to obtain a converging chain and its best fit was obtained after 1256 iterations.

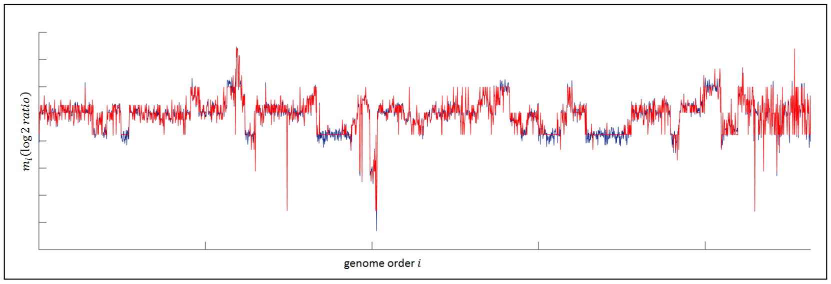

Figure 1: 2316 BAC clones (blue) for measurement of DNA copy number variation across the human genome. The red line is the reconstruction obtained using the MAP-iHMM algorithm.

5 Conclusions

We have presented a simple approach to estimation in Bayesian nonparameteric models, demonstrating example cases of algorithms for the DPMM and HDP-HMM that allow for efficient MAP-based estimation. This same approach can be readily applied whenever we have closed form conditional probabilities for each of the random variables in the underlying model. The conceptual and computational simplicity of the resulting algorithms demonstrate that this MAP-based approach can be an appropriate starting point for a wide range of Bayesian nonparametric problems where computational resources are at premium, or quick results are needed within seconds. Useful future directions would be extending our approach to online inference using memoized representation of the data [9], and for parallel, distributed implementations.

References

[1] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002. 1, 3

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

[3] 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

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

[5] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, and Michael Jordan. Streaming variational bayes. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1727–1735. 2013. 1

[6] 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

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

[8] Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth. Reliable and scalable variational inference for the hierarchical dirichlet process. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015. 1

[9] Michael C. Hughes and Erik B. Sudderth. Memoized online variational inference for dirichlet process mixture models. In Christopher J. C. Burges, Leon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NIPS*, pages 1133–1141, 2013. 1, 5

[10] 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

[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

[13] Radford Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000. 1

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

[15] Anirban Roychowdhury, Ke Jiang, and Brian Kulis. Small-variance asymptotics for hidden markov models. In Christopher J. C. Burges, Leon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *NIPS*, pages 2103–2111, 2013. 1

[16] Antoine M. Snijders, Norma Nowak, Richard Segraves, Stephanie Blackwood, Nils Brown, Jeffrey Conroy, Greg Hamilton, Anna Katherine Hindle, Bing Huey, Karen Kimura, Sindy Law, Ken Myambo, Joel Palmer, Bauke Ylstra, Jingzhu Pearl Yue, Joe W. Gray, Ajay N. Jain, Daniel Pinkel, and Donna G. Albertson. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29(3):263–264, 2001. 4

[17] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems*, volume 20, 2008. 1

[18] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006. 1, 3

[19] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1088–1095, New York, NY, USA, 2008. ACM. 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] Lianming Wang and David Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011. 1