Scalable inference for Bayesian non-parametrics

Michael Zhang

The University of Texas at Austin

1 Introduction

Bayesian non-parametric based models are an elegant way for discovering underlying latent features within a data set. These types of models have proven useful for learning functions [13], clustering data [1], topic modeling [17], or learning low dimensional representations of data [3] without having to enforce strong assumptions on the model (like the parametric from of the function or the number of clusters with which to model data). However appealing Bayesian non-parametrics models are, though, inference for such models is difficult.

Inference in the Bayesian paradigm faces different challenges from the frequentist approach. Whereas under the frequentist ideology, we have to optimize a difficult objective function, in Bayesian statistics we have to integrate difficult functions for which closed from expressions usually do not exist. This procedure becomes even more complicated in the non-parametrics setting with infinite dimension priors. Furthermore, under the "big data" scenario we have models whose computation scales poorly with growing quantities of data. In total, these problems prevent Bayesian non-parametric methods from being popular models for the statistical and machine learning practitioner.

Over the course of my Ph.D. studies, I have mainly focused on the problem of developing fast and parallel Markov chain Monte Carlo (MCMC) inference techniques for Bayesian and Bayesian non-parametric models. But beyond just developing techniques to perform fast Bayesian inference, for the future direction of my research, I would like to show that through scalable inference we are able to develop more interesting models than would otherwise not be possible without such techniques.

2 Previous Work

For the duration of my doctoral studies, I have focused mainly on researching scalable inference techniques for Bayesian and Bayesian non-parametric models. Of the papers that I have written, I would like to highlight four that are particularly relevant for this discussion.

2.1 Latent Variable Models

2.1.1 Sampling Feature Allocation

Bayesian non-parametric latent variable models have received a lot of interest for the flexibility in modeling data. The Dirichlet process (DP) in particular is attractive because its

marginal Dirichlet property allows for tractable inference. [9] provides an overview of different MCMC sampling techniques for the Dirichlet process mixture model (DPMM). Sampling methods broadly fall into two categories: methods that integrate out the mixing parameter, π [9, 5], and methods that leave π instantiated [6, 18, 2].

Inference methods that allow π to be instantiated are inherently parallelizable since the cluster allocation probability is conditionally independent given the mixing proportion, but proposing new features under this setting is difficult due to the infinite dimension of the mixing proportion. On the other hand, integrating out π allows us to deal only with the cluster allocation and not the mixing proportion. This, in conjunction with a likelihood $P(X_i|\theta_{z_i})$, provides a full conditional distribution to be used in a Gibbs sampler. However, $\int P(Z|\pi,\alpha)P(\pi|\alpha)\,\mathrm{d}\pi$ means that the marginal distribution of z_i becomes dependent on all other cluster allocations z_{-i} which is unparallelizable without excessive and costly communication between processors.

To overcome this problem, we developed a method in [21] where we are trying to parallelize the assignment of the latent variable for non-parametric models in the family of completely random measures (CRMs), specifically for the DP and the Indian buffet process (IBP), which has far fewer fast inference techniques compared to the DP. To solve this problem, we partition the latent features into two spaces–a finite dimensional partition for popular instantiated features where we sample feature assignments with an instantiated mixing parameter and an infinite dimensional partition where we propose and sample new features according to a collapsed mixing parameter which solves the problem of proposing and sampling new features in a parallel setting. This is possible because disjoint subsets in CRMs, which the IBP or DP can be expressed as, are independent thus avoiding the need for excessive processor communication.

2.1.2 Proposing Latent Features

However, none of these techniques deal with the problem of sampling good feature locations in high dimensional space. The reason for this is that the base measure divides the available probability mass over the whole input space for a predefined likelihood model. It controls which models are feasible and which are not. For high-dimensional data, we would need to spread the probability mass thin which, even if we had the "right" base measure, might prevent finding the correct representation of our data. Most other distributed samplers so far have only sought to accomplish fast and correct [15, 20]. [21] proposed to allow all processors to propose new features to encourage better mixing of the MCMC sampler before proceeding with exact inference and only allowing one processor to propose new features.

But, the proposal method of new features itself poses a serious problem in the latent variable models. In proposing new features, we can either propose features from an *uncollapsed* representation–meaning, drawing features from the prior, $P(\theta)$, and assign observations to clusters with likelihood $P(X|\theta)$. Or, we can marginalize out the clusters and sample cluster assignments from a *collapsed* likelihood, $\int P(X|\theta)P(\theta)\,\mathrm{d}\theta$. In high dimensional settings, the uncollapsed sampler is likely to fail to find good locations for features. The collapsed sampler is more likely to sample better feature locations because it draws from the expectation of the likelihood with respect to the prior distribution on features but this requires us to obtain the collapsed distribution in closed form which, in general, is not available. Even if the

collapsed distribution were available in closed form, situations where the features occupy a sparse region of its domain will also lead the collapsed sampler to fail to find good features.

To solve this issue, we developed a parallel inference technique for Bayesian non-parametric latent variable models [23]. In this approach, there is a serious problem of performing inference in high dimensions because draws from the prior cannot propose good features. Our inference strategy consists of two stages–one where we sample new features from data assigned to clusters with low likelihoods and the other where we run inference in the parallel uncollapsed MCMC sampling. Since MCMC samplers are valid from any starting location, this sampler is asymptotically guaranteed to converge to its correct limiting distribution. And because we no longer need to integrate the latent features from the likelihood for high-dimensional datasets, we show that we can apply our method to a more general class of models than previously possible.

2.2 Gaussian Processes

The challenge in Gaussian process (GP) inference comes in learning the hyperparameters that control the covariance function. Fitting these hyperparameters involves inverting an $N \times N$ covariance matrix obtained by evaluating at N inputs. In general, the computational cost of inverting this matrix is $O(N^3)$. One way to reduce this cost is use an easily-invertible class of covariance matrices. Two broad classes of methods have been proposed: "sparse" methods and "local" methods.

Sparse GP approximations [16] parameterize the covariance matrix of the GP model with M pseudo-inputs, where $M << N$, thereby requiring inversions of an $M \times M$ matrix. The pseudo-input locations are chosen so that the posterior function evaluated at these points is a good approximation to the true posterior. Though fast, sparse methods have a decreased ability to model fast moving functions, since the number of inducing points limits the amount of variation we can capture. Additionally, as with the full-covariance GP it approximates, the sparse GP cannot naturally model non-stationary data without resorting to a non-stationary kernel.

Local GP methods [4, 11] partition the data points into K groups, and learn a separate Gaussian process for each group. Conditioned on the partitioning, inference in the local GP scales approximately as $O(N^3/K^2)$, since we need to invert K matrices of average size $\frac{N}{K} \times \frac{N}{K}$. One advantage of local methods is that they allow us to use different covariance hyperparameters in different blocks, allowing us to capture behavior which is locally approximately stationary but where the lengthscale varies across partitions [12]. The disadvantage of the local methods, however, is that they risk ignoring important correlations since they assume zero correlation between different blocks in the partition. Thus, with either fast Gaussian process inference method we cannot capture all types of variation and correlations in the latent function.

Our approach to this problem is an idea that is both scalable and easily distributed [24]. We approximate the covariance matrix with a distribution over block diagonal matrices. In parallel, we multiple fit independent Gaussian processes to subsets of, allowing us to take advantage of the lower inversion cost. We then use importance sampling to combine these estimates in a principled manner–the only step in our algorithm requiring global communication. The resulting posterior predictive distribution has a dense covariance matrix, avoiding

edge effects common with local methods and allowing for an expressive covariance structure that can model both long- and short-range covariance.

2.3 Model Selection

In many modeling scenarios, many plausible models are available to fit to the data, each of which may result in drastically different predictions and conclusions. Being able to select the right model or to properly incorporate model uncertainty for inference is an crucial task. In the Bayesian setting, we have a natural probabilistic evaluation of models through posterior model probabilities. Depending on the objectives of the data analysis, we may be interested in selecting the "best" model or obtaining predictions with minimum error. Existing procedures to accomplish the aforementioned goals, however, will perform poorly under the presence of outliers. In addition, MCMC algorithms for these methods do not scale to "big data" situations.

To overcome these problems, we introduce a "divide and conquer" method and combine it with existing Bayesian model selection techniques for Bayesian models that is robust to outliers which, by consequence, will allow us to perform Bayesian model selection *in parallel*, thus also providing major reductions in computational burden with big data [22]. In our robust model selection strategy, we divide N observations into R subsets of roughly equal sample size. Once inference is built on each subset, the key step is to aggregate the subset models together into a final model. To aggregate our results, we collect the R number of subset models and find the geometric median [8] between these R elements.

Previous research in the notion of the geometric median and the median posterior has shown to be effective in robust and parallel statistical inference [19, 7]. These works have focused on the performance in terms of the errors relative to the true distribution. However, little research has been conducted on scalable methods for model selection or the utility of the geometric median for model selection. We show that, by aggregating with the geometric median, we obtain results that are robust to outliers and contamination of an unknown nature.

3 Future Research

3.1 Short Term Objectives

At the moment I am developing more parallel inference strategies for Bayesian models. Due to the costs involved with processor communication in parallel inference, we ideally would like to minimize the number of communication procedures to one if possible (also known as "embarrassingly parallel"). Other embarrassingly parallel methods exist [14, 10, 7], however one drawback of these methods is that we cannot quantify additional uncertainty as a result of the subposterior aggregation we can only claim the the resulting aggregated estimate is theoretically "close" to the true posterior nor are all of these strategies appropriate for all data modeling scenarios.

I currently have two ideas to quantify aggregation uncertainty: One is an idea in a forthcoming paper to create noisy estimates of the log likelihood density function via local

polynomial regression of the accepted states of the MCMC sampler of each subposterior and the associated likelihood of the data given the parameters and sample a final MCMC chain using this estimated likelihood instead of the true likelihood function. Using a local regression model, we can propagate the uncertainty of the estimator to see its effect on the resulting noisy posterior. Also, since we no longer need the likelihood function for final MCMC sampling, this method is suitable for models with expensive likelihoods whereas others are not.

The second is to encode the parallel computation as part of the structure in a Bayesian hierarchical model. Here, the local components of the hierarchy represent different processors' parameters with global parameters representing the aggregated estimate of the local posteriors. We obtain uncertainty quantification naturally with the global posteriors. Because we have a method of linking local processor parameters to global ones, my hypothesis is that we can apply the embarrassingly parallel framework to mixture models, which suffer from label switching issues and cannot be with the aforementioned parallel techniques, by placing a global mixture model on the local mixture parameters.

3.2 Long Term Objectives

The next research directive that I hope to investigate as a post-doctoral researcher and later as a professor is developing fast inference methods for general types of Bayesian and Bayesian non-parametric models. We have seen in [23] that, as a result of our novel method, we can learn the features of simple black and white image datasets like the MNIST or Yale face dataset where typical MCMC sampling will fail. Furthermore, the learning of non-stationary functions through [24] means that we can learn non-stationary functions that, previously, we would have needed complicated kernels to learn. Now, we can instead use a composite of simpler covariance functions. As a long term goal, I am interested in developing such methods for learning deep Gaussian processes and fitting latent variable models to data that can only be modeled with more complicated likelihoods, like color images or video. Optimistically, I suspect that I can extend what I have already researched, especially with [23] and [24], to accomplish this goal.

4 Conclusion

Bayesian non-parametrics, while attractive on an intuitive level, still faces difficult challenges especially under the regime of "big data" due to inferential problems especially for MCMC methods. In contrast, deep learning models are difficult to interpret; but because fitting deep learning models is possible for complicated and large datasets, we have seen immense interest in the continued research and use for such models. Hopefully, with the advent of better inference, the Bayesian non-parametric community can also enjoy the popularity and interest that the deep learning community has earned itself. Furthermore, my biggest wish is that people in statistics and machine learning can use the theoretically appealing properties of Bayesian non-parametrics in actual practice for more complicated problems than the settings for which current inference methods have already been developed.

References

[1] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

[2] Hong Ge, Yutian Chen, Moquan Wan, and Zoubin Ghahramani. Distributed inference for Dirichlet process mixture models. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2276–2284, 2015.

[3] Zoubin Ghahramani and Thomas L. Griffiths. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482, 2006.

[4] Robert B. Gramacy and Herbert K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[5] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *JASA*, 96(453):161–173, 2001.

[6] Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

[7] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[8] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664, 2014.

[9] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Dept. of Statistics, University of Toronto, 1998.

[10] W. Neiswanger, C. Wang, and E. Xing. Asymptotically Exact, Embarrassingly Parallel MCMC. *ArXiv e-prints*, November 2013.

[11] Jun Wei Ng and Marc Peter Deisenroth. Hierarchical mixture-of-experts model for large-scale Gaussian process regression. *arXiv preprint arXiv:1412.3078*, 2014.

[12] Carl E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *Neural Information Processing Systems*, pages 881–888, 2002.

[13] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian processes for machine learning*, volume 1. MIT Press, 2006.

[14] Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

[15] Padhraic Smyth, Max Welling, and Arthur U. Asuncion. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems*, pages 81–88, 2009.

[16] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Neural Information Processing Systems*, pages 1257–1264, 2005.

[17] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[18] Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54, 2007.

[19] Xiangyu Wang, Peichao Peng, and David B. Dunson. Median selection subset aggregation for parallel inference. In *Advances in Neural Information Processing Systems*, pages 2195–2203, 2014.

[20] Sinead A. Williamson, Avinava Dubey, and Eric Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In *Proceedings of the 30th International Conference on Machine Learning*, pages 98–106, 2013.

[21] Michael Minyi Zhang, Avinava Dubey, and Sinead A. Williamson. Distributed inference in Bayesian nonparametric models. 2016. Working paper.

[22] Michael Minyi Zhang, Henry Lam, and Lizhen Lin. Robust and parallel Bayesian model selection. 2016. arXiv:1610.06194.

[23] Michael Minyi Zhang and Fernando Pérez-Cruz. Accelerated inference for latent variable models. 2017. arXiv:1705.07178.

[24] Michael Minyi Zhang and Sinead A. Williamson. Embarrassingly parallel inference for Gaussian processes. 2017. arXiv:1702.08420.