Fast and Parallelizable Bayesian Inference

Michael Zhang

Princeton University

Analysis of complex phenomena requires models which carry a substantial computational burden. Computation is an ever-important problem for statistical inference, particularly for Bayesian models. 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. Bayesian inference therefore requires novel computational techniques when closed form expressions for calculating posteriors are unavailable. This becomes even more complicated as the dimensionality of the parameters and the size of the data grows, as we see in the Bayesian non-parametric setting with infinite dimension priors.

Bayesian non-parametric based models are an elegant way for discovering underlying latent features within a data set by assuming parameters are generated from an infinite dimensional distribution which thereby produces a posterior which is adaptive to the data at hand. These types of models have proven useful for modeling functions, clustering data, topic modeling, or learning low dimensional representations of data without having to enforce strong assumptions on the model (like a strict parametric form of the function or the number of clusters with which to model data, for example). The Bayesian non-parametric assumption that we may have infinite dimensional parameters a priori is only reasonable under "big data" situations, as we cannot assume infinite dimensional parameters if the dataset itself does not grow as well. However, the computational demands of BNP models quickly become unfeasible as the size of the data grow. *In total, these problems prevent Bayesian and Bayesian non-parametric methods from being popular models for the applied user.*

I have mainly focused on the problem of developing fast and parallel Markov chain Monte Carlo (MCMC) based computational techniques for Bayesian and Bayesian non-parametric models in collaboration with my doctoral advisor, Sinead Williamson, at the University of Texas at Austin. Beyond just developing fundamental techniques to perform fast Bayesian inference, **my research objective is to show that, through developing scalable inference, we are able to apply these flexible models in massive scenarios to make interesting, novel discoveries.** Lastly, I wish to continue pursuing interdisciplinary collaborations and substantive applications of scalable inference as part of my research agenda as a way to inform useful new directions for my work.

1 Advances in Bayesian Inference

1.1 Parallel Inference for Completely Random Measures

Bayesian non-parametric latent variable models have received a lot of interest for their flexibility in modeling data as they allow the practitioner to avoid specifying the exact number

of topics, clusters or latent factors. MCMC methods for such models typically demands a trade-off between the speed and quality of inference, depending on whether we decide to instantiate the infinite mixing measure, π. 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. Integrating out π allows us to deal only with the cluster allocation, z_i, and not the mixing proportion which simplifies the posterior update. However, the marginal distribution of z_i becomes dependent on all other cluster allocations z_{-i} which is unparallelizable without excessive processor communication.

To overcome this problem, we developed methods in [2, 9] where we parallelize the assignment of the latent variable for non-parametric models in the family of completely random measures (CRMs), which include the popular Dirichlet process, used in topic modeling and clustering, and the Indian buffet process, used in sparse factor modeling. We perform inference by splitting the latent features into a finite dimensional partition for popular instantiated features where we sample feature assignments with an instantiated mixing parameter and, on one processor, an infinite dimensional partition where we propose and sample new features according to a collapsed mixing parameter. This solves the problem of proposing and sampling new features in a parallel setting. This is possible because disjoint subsets in CRMs are independent thus avoiding the need for excessive, expensive inter-processor communication.

1.2 Parallel Inference for Gaussian Processes

Gaussian processes (GP) are very popular priors for non-linear regression and classification. In general, the computational cost of fitting GP models is $O(N^3)$, for N observations. Two broad classes of methods have been proposed to solve this problem: "sparse" methods and "local" methods. Sparse GPs approximates the posterior distribution via $M \ll N$ inducing points which represent pseudo-inputs to a GP model. This thereby reduces the complexity of fitting the model to $O(NM^2)$. Though fast, sparse methods cannot model fast-moving functions easily since the number of inducing points limits the amount of variation it can capture. In an example like modeling patient vital signs, we could have millions of observations with a fast moving latent function which renders sparse methods inappropriate.

Local GP methods partition the data points into K groups, and learn a separate Gaussian process for each group. This reduces the full covariance matrix into a block diagonal matrix. Assuming each block is size N/K, inference in the local GP scales approximately $O(N^3/K^2)$. The disadvantage of the local methods is that they risk ignoring important correlations since they assume zero correlation between different blocks. 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 [7, 8]. We approximate the covariance matrix with a mixture over block diagonal matrices. In parallel, we fit multiple independent Gaussian processes to partitions of the data, 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 non-stationary covariance behavior.

2 Current Research

2.1 Applications

One direction that I would like to continue as an assistant professor is collaborative research with the broader scientific community. Previously, I have collaborated with scientists in biology [4, 5], electrical engineering [3] and operations research [6]. Recently, I have been working with sociologists at Princeton to improve the computational and memory usage performance of a particular topic model that is particularly useful to social scientists, the Structural Topic Model (STM), because it can incorporate covariate information in the document-topic and the word-topic distributions. Though this topic model can capture much richer details, it is difficult to apply STM in large-scale situations as we cannot use typical fast variational inference techniques due to non-conjugate priors in the model.

In addition, I have also been working a on large scale data deduplication problem. The Eviction Lab at Princeton has collected extensive property ownership records from major US cities over time and are interested in investigating the claim that the ownership structure of residential properties in the US is being monopolized over time. However, we have a complicated merging process because we have temporal dependence in the linkage structure, for which we lack existing methods. Thus, we are developing deduplication models based on hierarchical modeling and BNP clustering that are highly scalable as well, considering that we are analyzing millions of property records.

Another applied research direction that I am pursuing is applications of Gaussian process modeling for medical data. GP models have proven to be very useful in monitoring patient health, with examples in jointly modeling patient vital signs and in sepsis detection. We can extend the fast inference method for non-stationary functions developed in [8] and its online variant [7], in conjunction with the multi-output GPs to track multiple patient vital signals in real-time. In particular, we look at real patient data in the MIMIC-III dataset to help develop our approach. We can further enrich this model by modeling latent force effects, as we did in [1], to create a toolset of fast and flexible GP models for medical applications and aid better real-time modeling and decision making in patient health.

3 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. My hope is that statistical practitioners can use the theoretically appealing properties of Bayesian non-parametrics in practice, for more complicated problems as a result of having highly scalable inference methods.

References

[1] L.-F. Cheng, B. Dumitrascu, M. M. Zhang, C. Chivers, K. Li, and B. E. Engelhardt. Personalized effects of medication on patients using latent force models with Gaussian processes. 2019. In review.

[2] A. Dubey, M. M. Zhang, E. P. Xing, and S. A. Williamson. Distributed, partially collapsed MCMC for Bayesian nonparametrics. 2019. In review. Joint first author.

[3] F. Pérez-Cruz, P. M. Olmos, M. M. Zhang, and H. Huang. Probabilistic time of arrival localization. *IEEE Signal Processing Letters*, 26(11):1683–1687, 2019. arXiv:1910.06569.

[4] Z. I. Phillips, M. M. Zhang, and U. G. Müller. Dispersal of *Attaphila fungicola* (Blattodea: Ectobiidae), a symbiotic cockroach of leafcutter ants (Hymenoptera: Formicidae). *Insectes Sociaux*, 64(2):277–284, 2017.

[5] Z. I. Phillips, M. M. Zhang, and L. Reding. Social immune tolerance as a special protection of the queen. 2018. In review.

[6] S. A. Williamson, M. M. Zhang, and P. Damien. A new class of time-dependent latent factor models with applications. 2019. arXiv:1904.08548. To appear in Journal of Machine Learning Research.

[7] M. M. Zhang, B. Dumitrascu, S. A. Williamson, and B. E. Engelhardt. Sequential Gaussian processes for online learning of nonstationary functions. 2019. arxiv:1905.10003. In review.

[8] M. M. Zhang and S. A. Williamson. Embarrassingly parallel inference for Gaussian processes. 2019. arXiv:1702.08420. To appear in the Journal of Machine Learning Research.

[9] M. M. Zhang, S. A. Williamson, and F. Pérez-Cruz. Accelerated parallel non-conjugate sampling for Bayesian non-parametric models. 2019. arXiv:1705.07178. In review, revise and resubmit. Appeared in "BNP@NeurIPS 2018" as workshop paper. Previously known as "Accelerated Inference for Latent Variable Models".