PROJECT MUSE®

Nonparametric Bayes: A Bridge Between Cultures

Arman Oganisian, Jason A. Roy

Observational Studies, Volume 7, Issue 1, 2021, pp. 175-178 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0005>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/799730>

Nonparametric Bayes: A Bridge Between Cultures

Arman Oganisian

*Department of Biostatistics
Brown University
Providence, RI 02912*

arman_oganisian@brown.edu

Jason A. Roy

*Department of Biostatistics and Epidemiology
Rutgers University
Piscataway, NJ 08854*

jason.roy@rutgers.edu

Abstract

In this commentary, we assess the cultural fit of Bayesian nonparametrics in light of advances in the field since Breiman’s 2001 article. We argue that Bayesian nonparametrics synthesizes desirable elements of the data modeling and algorithmic cultures to yield new insights and methodological improvements. We discuss how these methods have been combined with identification strategies from the causal inference literature to do flexible inference for interpretable target parameters.

Keywords: Bayesian Nonparametrics, Causal Inference, Probabilistic Machine Learning

Two decades ago, Breiman (2001) opined that the statistical community was split between the “data modeling” and “algorithmic modeling” cultures - with little overlap between the two. Broadly, the former was characterized by rigid parametric/distributional assumptions (an “a priori straight jacket”, to quote directly) imposed on the data generating mechanism, while the latter attempted to be less restrictive. A lot has changed since the 2001 paper. Surely, far greater than 2% of statisticians now develop and/or utilize machine learning algorithms in their own research. Breiman’s piece was therefore timely, being written right when a cultural shift was starting to take place.

In this piece, we reflect on the cultural fit of Bayesian nonparametrics (BNP) in light of the advances in the field since Breiman’s commentary. Parametric Bayesian methods fit neatly into Breiman’s “data modeling culture.” These typically involve specifying a data model indexed by finitely many parameter. For instance when estimating the conditional distribution $y \mid x \sim G_x$, we may assume G_x is Gaussian with x governing the mean in a linear, additive way according to slope and intercept parameters. After imposing these severe assumptions, posterior inference on G_x is done indirectly via a prior and posterior over the parameters assumed to be governing it. By contrast, BNP methods may avoid these constraints by placing a prior directly over the family of distributions $\{G_x : x \in \mathcal{X}\}$ - a very high-dimensional object. Therefore, on the one hand, BNP requires a probability model for proper posterior inference - endowing it with elements of the data modeling culture. On the other, the high dimensionality of these models enables them to capture

complex generative processes - tying it closely to the algorithmic culture. We argue that BNP does not fit neatly into either culture, but rather lies somewhere between the two and, perhaps, bridges them. In fact, many nonparametric Bayesian approaches can be seen as targeting some high-dimensional probabilistic data model underlying an algorithmic model. Casting algorithmic models as an approximation to a latent probabilistic model has made the former more amenable to formal study and helped shed light on why they work and, more importantly, when they can fail - often motivating useful extensions.

Consider penalized regression methods (e.g. Ridge, LASSO), which are now commonplace both in the academic literature, in graduate school machine learning (ML) curricula, and in applied ML. While the underlying probabilistic motivations (e.g. LASSO corresponding to posterior point inference under a double-exponential prior) are noted in popular statistical texts (Hastie et al., 2009), these methods still tend to be reasoned about in terms of the functional form of the penalties and the induced optimization constraint. The probabilistic motivations are often treated merely as intellectually interesting footnotes, even though they have yielded practical insights and new developments. For instance, these connections were central to the insights of Carvalho et al. (2010) who noted that several popular shrinkage priors including the LASSO can be characterized by the conditional distribution they induce on a “shrinkage coefficient.” This induced conditional provides a common yardstick for comparing these different methods on the basis of (1) how aggressively they shrink noise and (2) how well they detect real signals. They link performance on these two fronts to the tail-behavior under various priors. Analysis under this probabilistic lens revealed that the shrinkage of LASSO faces a trade-off between (1) and (2) and could lead to over-shrinkage of signals. They developed a horseshoe prior with more desirable tail-behavior and, therefore, shrinkage properties.

The mixture of experts (MoEs) (Bishop, 2006) regression is another approach attributable to the algorithmic culture that has benefited from formal BNP analysis. The idea of MoEs is to partition the data space (joint distribution of features and outcome) into homogeneous regions. Data from each region can be used to train separate predictive models (“experts”) local to each region. One probabilistic analogue of MoEs in BNP is the Dirichlet process (DP) mixture of generalized linear models of Hannah et al. (2011). This models the data with an adaptive mixture of parametric experts, where the number of mixture components in the model is unspecified. Instead it is driven by the degree of heterogeneity in the data, allowing the model to grow in richness if needed. As with penalized regression, viewing this DP mixture model as a probabilistic MoE analogue lead to further model improvements. For instance, Wade et al. (2014) reveal that DP mixtures can produce poor predictions when the outcome model is conditioned on a very high dimensional set of features and develop an “enriched” DP prior with improved performance in these settings.

Many other examples of BNP analogues to products of the algorithmic culture come to mind, such as Bayesian principal component analysis (PCA) (Bishop, 1999). This extends probabilistic PCA (Tipping and Bishop, 1999) and classical PCA to handle several complexities by viewing it as Bayesian inference for a latent Gaussian generative model. Another well-known example is the Bayesian Additive Regression Trees (BART) model of Chipman et al. (2010) which connects with regression trees, boosting, and ensembles - approaches we would associate with the algorithmic culture. As a final example, posterior inference for

unknown functions under Gaussian process priors have become increasingly popular and have connections to classical ML methods such as kernel regression.

One of Breiman’s reservations about the data modeling culture was the use of parametric models. Conclusions regarding scientific hypotheses are made using estimates of the parameters. Breiman (2001) cautions that these “conclusions are about the model’s mechanism, and not about nature’s mechanism...If the model is a poor emulation of nature, the conclusions may be wrong.” We agree - especially because the choice of the scientific estimand is often driven by the convenience of these models, which puts the cart before the horse. BNP, when combined with identification strategies developed in the causal inference literature, has alleviated some of these concerns as well. The causal inference field has largely moved away from interpreting parameters from a parametric model and instead focuses on inference about “target parameters” of interest. In particular, we can identify scientifically meaningful and interpretable population-level estimands (e.g., an average causal effect) without reference to a particular parametric model. Once we identify these estimands in terms of observed data objects, we can leverage BNP to obtain a posterior over these objects and, therefore, the estimand of interest. Much progress has been made in such Bayesian nonparametric causal inference including applications of DP mixtures, BART, and other priors to complex estimands (see Oganisian and Roy (2021) for survey). In other words, Breiman’s “biostatistician friends” no longer have to worry that doctors cannot “interpret a black box containing fifty trees hooked together.”

Breiman (2001) lamented that developments in algorithmic models had “occurred largely outside statistics in a new community - often called machine learning - that is mostly young computer scientists.” However, things have been shifting and now contributions to algorithmic models are coming from many communities, as we partially enumerated above. Based on his other commentaries (Breiman, 1997), however, it seems Breiman did not imagine that a good deal of these contributions would be from BNP. There has even been a reversal of sorts, in recent years, with computer scientists and ML researchers turning to BNP. Indeed, the term “probabilistic machine learning” (Ghahramani, 2015; Murphy, 2021) has been popping up more in this literature. Though it is difficult to define probabilistic ML precisely, it makes heavy use of BNP to, in our view, combine the uncertainty quantification and probabilistic lens of the data modeling culture with the flexibility of the algorithmic culture. These methodological advances are accompanied by the explosion of available probabilistic programming software - such as Stan, PyMC3, and Turing in Julia - that facilitate computation. With these developments in mind, we think the future of BNP is bright and are excited by its capacity to unify and bring together ideas that are often treated in isolation.

References

- Christopher Bishop. Bayesian pca. In *Advances in Neural Information Processing Systems*, volume 11, pages 382–388. MIT Press, January 1999.
- Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- Leo Breiman. No bayesians in foxholes. *IEEE Expert*, 12(6):21–24, 1997.

- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3):199–231, 08 2001. doi: 10.1214/ss/1009213726.
- Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. ISSN 00063444.
- Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298, 03 2010. doi: 10.1214/09-AOAS285.
- Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.
- Lauren A Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(6), 2011.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2021. URL `probml.ai`.
- Arman Oganisian and Jason A Roy. A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2): 518–551, 2021. doi: <https://doi.org/10.1002/sim.8761>.
- Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999. ISSN 13697412, 14679868.
- Sara Wade, David B Dunson, Sonia Petrone, and Lorenzo Trippa. Improving prediction from dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research*, 15(1):1041–1071, 2014.