25

Nonparametric Bayes

David B. Dunson
Department of Statistical Science
Duke University, Durham, NC

I reflect on the past, present, and future of nonparametric Bayesian statistics. Current nonparametric Bayes research tends to be split between theoretical studies, seeking to understand relatively simple models, and machine learning, defining new models and computational algorithms motivated by practical performance. I comment on the current landscape, open problems and promising future directions in modern big data applications.

25.1 Introduction

25.1.1 Problems with parametric Bayes

In parametric Bayesian statistics, one chooses a likelihood function $L(y|\theta)$ for data y, which is parameterized in terms of a finite-dimensional unknown θ. Choosing a prior distribution for θ, one updates this prior with the likelihood $L(y|\theta)$ via Bayes' rule to obtain the posterior distribution $\pi(\theta|y)$ for θ. This framework has a number of highly appealing characteristics, ranging from flexibility to the ability to characterize uncertainty in θ in an intuitively appealing probabilistic manner. However, one unappealing aspect is the intrinsic assumption that the data were generated from a particular probability distribution (e.g., a Gaussian linear model).

There are a number of challenging questions that arise in considering, from both philosophical and practical perspectives, what happens when such an assumption is violated, as is arguably always the case in practice. From a philosophical viewpoint, if one takes a parametric Bayesian perspective, then a prior is being assumed that has support on a measure zero subset of the set of possible distributions that could have generated the data. Of course, as it is commonly accepted that all models are wrong, it seems that such a prior does not actually characterize any individual's prior beliefs, and one may question the meaning of the resulting posterior from a subjective Bayes perspective. It would seem that a rational subjectivist would assign positive prior probability to the case in which the presumed parametric model is wrong in unanticipated ways, and probability zero to the case in which the data are generated *exactly* from the presumed model. Objective Bayesians should similarly acknowledge that any parametric model is wrong, or at least has a positive probability of being wrong, in order to truly be objective. It seems odd to spend an enormous amount of effort showing that a particular prior satisfies various objectivity properties in a simple parametric model, as has been the focus of much of the Bayes literature.

The failure to define a framework for choosing priors in parametric models, which acknowledge that the "working" model is wrong, leads to some clear practical issues with parametric Bayesian inference. One of the major ones is the lack of a framework for model criticism and goodness-of-fit assessments. Parametric Bayesians assume prior knowledge of the true model which generated the data, and hence there is no allowance within the Bayesian framework for incorrect model choice. For this reason, the literature on Bayesian goodness-of-fit assessments remains under-developed, with most of the existing approaches relying on diagnostics that lack a Bayesian justification. A partial solution is to place a prior distribution over a list of possible models instead of assuming a single model is true *a priori*. However, such Bayesian model averaging/selection approaches assume that the true model is one of those in the list, the so-called M-closed viewpoint, and hence do not solve the fundamental problem.

An alternative pragmatic view is that it is often reasonable to operate under the working assumption that the presumed model is true. Certainly, parametric Bayesian and frequentist inferences often produce excellent results even when the true model deviates from the assumptions. In parametric Bayesian models, it tends to be the case that the posterior distribution for the unknown θ will concentrate at the value θ_0, which yields a sampling distribution that is as close as possible to the true data-generating model in terms of the Kullback–Leibler (KL) divergence. As long as the parametric model provides an "adequate" approximation, and this divergence is small, it is commonly believed that inferences will be "reliable." However, there has been some research suggesting that this common belief is often wrong, such as when the loss function is far from KL (Owhadi et al., 2013).

Results of this type have provided motivation for "quasi" Bayesian approaches, which replace the likelihood with other functions (Chernozhukov and Hong, 2003). For example, quantile-based substitution likelihoods have been proposed, which avoid specifying the density of the data between quantiles (Dunson and Taylor, 2005). Alternatively, motivated by avoiding specification of parametric marginal distributions in considering copula dependence models (Genest and Favre, 2007; Hoff, 2007; Genest and Nešlehová, 2012; Murray et al., 2013), use an extended rank-based likelihood. Recently, the idea of a Gibbs posterior (Jiang and Tanner, 2008; Chen et al., 2010) was introduced, providing a generalization of Bayesian inference using a loss-based pseudo likelihood. Appealing properties of this approach have been shown in various contexts, but it is still unclear whether such methods are appropriately calibrated so that the quasi posterior distributions obtained provide a valid measure of uncertainty. It may be the case that uncertainty intervals are systematically too wide or too narrow, with asymptotic properties such as consistency providing no reassurance that uncertainty is well characterized.

Fully Bayesian nonparametric methods require a full characterization of the likelihood, relying on models with infinitely-many parameters having carefully chosen priors that yield desirable properties. In the remainder of this chapter, I focus on such approaches.

25.1.2 What is nonparametric Bayes?

Nonparametric (NP) Bayes seeks to solve the above problems by choosing a highly flexible prior, which assigns positive probability to arbitrarily small neighborhoods around any true data-generating model f_0 in a large class. For example, as an illustration, consider the simple case in which $y_1, \ldots, y_n$ form a random sample from density f. A parametric Bayes approach would parameterize the density f in terms of finitely-many unknowns θ, and induce a prior for f through a prior for θ. Such a prior will in general have support on a vanishingly small subset of the set of possible densities $\mathcal{F}$ (e.g., with respect to Lebesgue measure on $\mathbb{R}$). NP Bayes instead lets $f \sim \Pi$, with Π a prior over $\mathcal{F}$ having large support, meaning that $\Pi\{f : d(f, f_0) < \epsilon\} > 0$ for some distance metric d, any $\epsilon > 0$, and any f_0 in a large subset of $\mathcal{F}$. Large support is the defining property of an NP Bayes approach, and means that realizations from the prior have a positive probability of being arbitrarily close to any f_0, perhaps ruling out some irregular ones (say with heavy tails).

In general, to satisfy the large support property, NP Bayes probability models include infinitely-many parameters and involve specifying stochastic processes for random functions. For example, in the density estimation example, a very popular prior is a Dirichlet process mixture (DPM) of Gaussians (Lo, 1984). Under the stick-breaking representation of the Dirichlet process (Sethuraman, 1994), such a prior lets

$$f(y) = \sum_{h=1}^{\infty} \pi_h \mathcal{N}(y; \mu_h, \tau_h^{-1}), \quad (\mu_h, \tau_h) \overset{\text{iid}}{\sim} P_0, \tag{25.1}$$

where the weights on the normal kernels follow a stick-breaking process, $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$, with $V_h \sim \mathcal{B}(1, \alpha)$ independently, α is the concentration parameter in the Dirichlet process, and P_0 is the base measure. Such kernel mixture priors satisfy the large support property and can be defined so that the resulting posterior concentrates around the unknown true f_0 at the minimax optimal rate up to a log factor (de Jonge and van Zanten, 2010).

Prior (25.1) is intuitively appealing in including infinitely-many Gaussian kernels having stochastically decreasing weights. In practice, there will tend to be a small number of kernels having large weights, with the remaining having vanishingly small weights. Only a modest number of kernels will be occupied by the subjects in a sample, so that the effective number of parameters may actually be quite small and is certainly not infinite, making posterior computation and inferences tractable. Of course, prior (25.1) is only one particularly simple example (to quote Andrew Gelman, "No one is really interested in density estimation"), and there has been an explosion of literature in recent years proposing an amazing variety of NP Bayes models for broad applications and data structures.

Section 25.2 contains an (absurdly) incomplete timeline of the history of NP Bayes up through the present. Section 25.3 comments briefly on interesting future directions.

25.2 A brief history of NP Bayes

Although there were many important earlier developments, the modern view of nonparametric Bayes statistics was essentially introduced in the papers of Ferguson (1973, 1974), which proposed the Dirichlet process (DP) along with several ideal criteria for a nonparametric Bayes approach including large support, interpretability and computational tractability. The DP provides a prior for a discrete probability measure with infinitely many atoms, and is broadly employed within Bayesian models as a prior for mixing distributions and for clustering. An equally popular prior is the Gaussian process (GP), which is instead used for random functions or surfaces. A non-neglible proportion of the nonparametric Bayes literature continues to focus on theoretical properties, computational algorithms and applications of DPs and GPs in various contexts.

In the 1970s and 1980s, NP Bayes research was primarily theoretical and conducted by a narrow community, with applications focused primarily on jointly conjugate priors, such as simple cases of the gamma process, DP and GP. Most research did not consider applications or data analysis at all, but instead delved into characterizations and probabilistic properties of stochastic processes, which could be employed as priors in NP Bayes models. These developments later had substantial applied implications in facilitating computation and the development of richer model classes.

With the rise in computing power, development of Gibbs sampling and explosion in use of Markov chain Monte Carlo (MCMC) algorithms in the early 1990s, nonparametric Bayes methods started to become computationally tractable. By the late 1990s and early 2000s, there were a rich variety of inferential algorithms available for general DP mixtures and GP-based models in spatial statistics, computer experiments and beyond. These algorithms, combined with increasing knowledge of theoretical properties and characterizations, stimulated an explosion of modeling innovation starting in the early 2000s but really gaining steam by 2005. A key catalyst in this exponential growth of research activity and innovation in NP Bayes was the dependent Dirichlet process (DDP) of Steve MacEachern, which ironically was never published and is only available as a technical report. The DDP and other key modeling innovations were made possible by earlier theoretical work providing characterizations, such as stick-breaking (Sethuraman, 1994; Ishwaran and James, 2001) and the Polya urn scheme/Chinese restaurant process (Blackwell and MacQueen, 1973). Some of the circa 2005–10 innovations include the Indian buffet process (IBP) (Griffiths and Ghahramani, 2011), the hierarchical Dirichlet process (HDP) (Teh et al., 2006), the nested Dirichlet process (Rodríguez et al., 2008), and the kernel stick-breaking process (Dunson and Park, 2008).

One of the most exciting aspects of these new modeling innovations was the potential for major applied impact. I was fortunate to start working on NP Bayes just as this exponential growth started to take off. In the NP Bayes statistics community, this era of applied-driven modeling innovation peaked at the 2007 NP Bayes workshop at the Issac Newton Institute at Cambridge University. The Newton Institute is an outstanding facility and there was an energy and excited vibe permeating the workshop, with a wide variety of topics being covered, ranging from innovative modeling driven by biostatistical applications to theoretical advances on properties. One of the most exciting aspects of statistical research is the ability to fully engage in a significant applied problem, developing methods that really make a practical difference in inferences or predictions in the motivating application, as well as in other related applications. To me, it is ideal to start with an applied motivation, such as an important aspect of the data that is not captured by existing statistical approaches, and then attempt to build new models and computational algorithms that have theoretical support and make a positive difference to the bottom-line answers in the analysis. The flexibility of NP Bayes models makes this toolbox ideal for attacking challenging applied problems.

Although the expansion of the NP Bayes community and impact of the research has continued since the 2007 Newton workshop, the trajectory and flavor of the work has shifted substantially in recent years. This shift is due in part to the emergence of big data and to some important cultural hurdles, which have slowed the expansion of NP Bayes in statistics and scientific applications, while stimulating increasing growth in machine learning. Culturally, statisticians tend to be highly conservative, having a healthy skepticism of new approaches even if they seemingly improve practical performance in prediction and simulation studies. Many statisticians will not really trust an approach that lacks asymptotic justification, and there is a strong preference for simple methods that can be studied and understood more easily. This is perhaps one reason for the enormous statistical literature on minor variations of the lasso.

NP Bayes methods require more of a learning curve. Most graduate programs in statistics have perhaps one elective course on Bayesian statistics, and NP Bayes is not a simple conceptual modification of parametric Bayes. Often models are specified in terms of infinite-dimensional random probability measures and stochastic processes. On the surface, this seems daunting and the knee-jerk reaction by many statisticians is negative, mentioning unnecessary complexity, concerns about over-fitting, whether the data can really support such complexity, lack of interpretability, and limited understanding of theoretical properties such as asymptotic behavior. This reaction restricts entry into the field and makes it more difficult to get publications and grant funding.

However, these concerns are largely unfounded. In general, the perceived complexity of NP Bayes models is due to lack of familiarity. Canonical model classes, such as DPs and GPs, are really quite simple in their structure and tend to be no more difficult to implement than flexible parametric models. The intrinsic Bayesian penalty for model complexity tends to protect against over-fitting. For example, consider the DPM of Gaussians for density estimation shown in equation (25.1). The model is simple in structure, being a discrete mixture of normals, but the perceived complexity comes in through the incorporation of infinitely many components. For statisticians unfamiliar with the intricacies of such models, natural questions arise such as “how can the data inform about all these parameters” and “there certainly must be over-fitting and huge prior sensitivity.” However, in practice, the prior and the penalty that comes in through integrating over the prior in deriving the marginal likelihood tends to lead to allocation of all the individuals in the sample to relatively few clusters. Hence, even though there are infinitely many components, only a few of these are used and the model behaves like a finite mixture of Gaussians, with sieve behavior in terms of using more components as the sample size increases. Contrary to the concern about over-fitting, the tendency is instead to place a high posterior weight on very few components, potentially under-fitting in small sample sizes. DPMs are a simple example but the above story applies much more broadly.

The lack of understanding in the broad statistical community of the behavior of NP Bayes procedures tempered some of the enthusiastic applications-driven modeling of the 2000s, motivating an emerging field focused on studying frequentist asymptotic properties. There is a long history of NP Bayes asymptotics, showing properties such as consistency and rates of concentration of the posterior around the true unknown distribution or function. In the past five years, this field has really taken off and there is now a rich literature showing strong properties ranging from minimax optimal adaptive rates of posterior concentration (Bhattacharya et al., 2013) to Bernstein–von Mises results characterizing the asymptotic distribution of functionals (Rivoirard and Rousseau, 2012). Such theorems can be used to justify many NP Bayes methods as also providing an optimal frequentist procedure, while allowing frequentist statisticians to exploit computational methods and probabilistic interpretations of Bayes methods. In addition, an appealing advantage of NP Bayes nonparametric methods is the allowance for uncertainty in tuning parameter choice through hyperpriors, bypassing the need for cross-validation. The 2013 NP Bayes conference in Amsterdam was notable in exhibiting a dramatic shift in topics compared with the 2007 Newton conference, away from applications-driven modeling and towards asymptotics.

The other thread that was very well represented in Amsterdam was NP Bayes machine learning, which has expanded into a dynamic and important area. The machine learning (ML) community is fundamentally different culturally from statistics, and has had a very different response to NP Bayes methods as a result. In particular, ML tends to be motivated by applications in which bottom-line performance in metrics, such as out-of-sample prediction, takes center stage. In addition, the ML community prefers peer-reviewed proceedings for conferences, such as Neural Information Processing Systems (NIPS) and the International Conference on Machine Learning Research (ICML), over journal publications. These conference proceedings are short papers, and there is an emphasis on innovative new ideas which improve bottom line performance. ML researchers tend to be aggressive and do not shy away from new approaches which can improve performance regardless of complexity. A substantial proportion of the novelty in NP Bayes modeling and computation has come out of the ML community in recent years. With the increased emphasis on big data across fields, the lines between ML and statistics have been blurring. However, publishing an initial idea in NIPS or ICML is completely different than publishing a well-developed and carefully thought out methods paper in a leading statistical theory and methods journal, such as the *Journal of the American Statistical Association*, *Biometrika* or the *Journal of the Royal Statistical Society*, *Series B*. My own research has greatly benefited by straddling the asymptotic, ML and applications-driven modeling threads, attempting to develop practically useful and innovative new NP Bayes statistical methods having strong asymptotic properties.

25.3 Gazing into the future

Moving into the future, NP Bayes methods have rich promise in terms of providing a framework for attacking a very broad class of ‘modern’ problems involving high-dimensional and complex data. In big complex data settings, it is much more challenging to do model checking and to carefully go through the traditional process of assessing the adequacy of a parametric model, making revisions to the model as appropriate. In addition, when the number of variables is really large, it becomes unlikely that a particular parametric model works well for all these variables. This is one of the reasons that ensemble approaches, which average across many models/algorithms, tend to produce state of the art performance in difficult prediction tasks. Combining many simple models, each able to express different characteristics of the data, is useful and similar conceptually to the idea of Bayesian model averaging (BMA), though BMA is typically only implemented within a narrow parametric class (e.g., normal linear regression).

In considering applications of NP Bayes in big data settings, several questions arise. The first is "Why bother?" In particular, what do we have to gain over the rich plethora of machine learning algorithms already available, and which are being refined and innovated upon daily by thousands of researchers? There are clear and compelling answers to this question. ML algorithms almost always rely on convex optimization to obtain a point estimate, and uncertainty is seldom of much interest in the ML community, given the types of applications they are faced with. In contrast, in most scientific applications, prediction is not the primary interest and one is usually focused on inferences that account for uncertainty. For example, the focus may be on assessing the conditional independence structure (graphical model) relating genetic variants, environmental exposures and cardiovascular disease outcomes (an application I'm currently working on). Obtaining a single estimate of the graph is clearly not sufficient, and would be essentially uninterpretable. Indeed, such graphs produced by ML methods such as graphical lasso have been deemed "ridiculograms." They critically depend on a tuning parameter that is difficult to choose objectively and produce a massive number of connections that cannot be effectively examined visually. Using an NP Bayes approach, we could instead make highly useful statements (at least according to my collaborators), such as (i) the posterior probability that genetic variants in a particular gene are associated with cardiovascular disease risk, adjusting for other factors, is $P\%$; or (ii) the posterior probability that air pollution exposure contributes to risk, adjusted for genetic variants and other factors, is $Q\%$. We can also obtain posterior probabilities of an edge between each variable without parametric assumptions, such as Gaussianity. This is just one example of the utility of probabilistic NP Bayes models; I could list dozens of others.

The question then is why aren't more people using and working on the development of NP Bayes methods? The answer to the first part of this question is clearly computational speed, simplicity and accessibility. As mentioned above, there is somewhat of a learning curve involved in NP Bayes, which is not covered in most graduate curriculums. In contrast, penalized optimization methods, such as the lasso, are both simple and very widely taught. In addition, convex optimization algorithms for very rapidly implementing penalized optimization, especially in big data settings, have been highly optimized and refined in countless publications by leading researchers. This has led to simple methods that are scalable to big data, and which can exploit distributed computing architectures to further scale up to enormous settings. Researchers working on these types of methods often have a computer science or engineering background, and in the applications they face, speed is everything and characterizing uncertainty in inference or testing is just not a problem they encounter. In fact, ML researchers working on NP Bayes methods seldom report inferences or use uncertainty in their analyses; they instead use NP Bayes methods combined with approximations, such as variational Bayes or expectation propagation, to improve performance on ML tasks, such as prediction. Often predictive performance can be improved, while avoiding cross-validation for tuning parameter selection, and these gains have partly led to the relative popularity of NP Bayes in machine learning.

It is amazing to me how many fascinating and important unsolved problems remain in NP Bayes, with the solutions having the potential to substantially impact practice in analyzing and interpreting data in many fields. For example, there is no work on the above nonparametric Bayes graphical modeling problem, though we have developed an initial approach we will submit for publication soon. There is very limited work on fast and scalable approximations to the posterior distribution in Bayesian nonparametric models. Markov chain Monte Carlo (MCMC) algorithms are still routinely used despite their problems with scalability due to the lack of decent alternatives. Variational Bayes and expectation propagation algorithms developed in ML lack theoretical guarantees and often perform poorly, particularly when the focus goes beyond obtaining a point estimate for prediction. Sequential Monte Carlo (SMC) algorithms face similar scalability problems to MCMC, with a daunting number of particles needed to obtain adequate approximations for high-dimensional models. There is a clear need for new models for flexible dimensionality reduction in broad settings. There is a clear lack of approaches for complex non-Euclidean data structures, such as shapes, trees, networks and other object data.

I hope that this chapter inspires at least a few young researchers to focus on improving the state of the art in NP Bayes statistics. The most effective path to success and high impact in my view is to focus on challenging real-world applications in which current methods have obvious inadequacies. Define innovative probability models for these data, develop new scalable approximations and computational algorithms, study the theoretical properties, implement the methods on real data, and provide software packages for routine use. Given how few people are working in such areas, there are many low hanging fruit and the clear possibility of major breakthroughs, which are harder to achieve when jumping on bandwagons.

References

Bhattacharya, A., Pati, D., and Dunson, D.B. (2013). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals* *of Statistics*, in press.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355.

Chen, K., Jiang, W., and Tanner, M. (2010). A note on some algorithms for the Gibbs posterior. *Statistics & Probability Letters*, 80:1234–1241.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346.

de Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38:3300–3320.

Dunson, D.B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, 95:307–323.

Dunson, D.B. and Taylor, J. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, 17:385–400.

Ferguson, T.S. (1973). Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–629.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368.

Genest, C. and Nešlehová, J. (2012). Copulas and copula models. *Encyclopedia of Environmetrics*, 2nd edition. Wiley, Chichester, 2:541–553.

Griffiths, T. and Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224.

Hoff, P. (2007). Extending the rank likelihood for semi parametric copula estimation. *The Annals of Applied Statistics*, 1:265–283.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

Jiang, W. and Tanner, M. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36:2207–2231.

Lo, A. (1984). On a class of Bayesian nonparametric estimates. 1. density estimates. *The Annals of Statistics*, 12:351–357.

Murray, J.S., Dunson, D.B., Carin, L., and Lucas, J.E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108:656–665.

Owhadi, H., Scovel, C., and Sullivan, T. (2013). Bayesian brittleness: Why no Bayesian model is "good enough." *arXiv:1304.6772* .

Rivoirard, V. and Rousseau, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40:1489–1523.

Rodríguez, A., Dunson, D.B., and Gelfand, A. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103:1131–1144.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.