University of Warwick – *Nonparametric Bayes Lectures*

David Dunson, Duke University dunson@duke.edu

April 4-6, 2016

Overview – This sequence of lectures will provide a practically motivated introduction to nonparametric Bayes thinking. The goal is to provide an initial philosophical motivation and overview of different ways of thinking "nonparametrically" within a Bayesian paradigm. We will then transition into introducing some canonical modeling classes, illustrating standard and simple approaches to posterior computation in such models. Toy examples will be used to illustrate concepts, but we will rapidly transition into attempting to solve real world applied problems using nonparametric Bayes machinery. A dual emphasis will be on providing an accessible entry point into using npB methods in difficult applications, while also motivating pushing forward the research envelope in this area. Given time the more advanced topics will be targeted to specific areas.

Lecture 1 – What & Why of Nonparametric Bayes?

In this lecture, I will first remind the students of the general Bayesian statistical approach and how it requires a complete specification of the likelihood function generating the observed data. This leads to a beautiful framework having all sorts of philosophical and practical appeal BUT the elephant in the room is that the underlying assumption is that the likelihood is correctly specified. As this assumption is clearly flawed in practice (famously all models are wrong), the question is (i) what is the practical impact of assuming a model is true when it is in fact wrong; (ii) what can we do to relax or remove this assumption.

In attempting to address (i)-(ii), there are several possibilities: (1) Bayesian model averaging (BMA), which places positive probability on a list of candidate models, with each of these candidates typically being parametric; (2) "robust" Bayesian models that attempt to allow heavy tails, outliers, etc; (3) pseudo-Bayesian approaches that avoid "modeling everything" and instead focus on various fake likelihood functions – for quantiles, based on loss functions (Gibbs posteriors), etc; and (4) "nonparametric" Bayesian approaches that choose priors having "large support" – meaning that the prior can generate sampling distributions within arbitrarily small neighborhoods of any true data-generating likelihood in a broad class.

(1) Bayesian model averaging: In using typical BMA, one chooses a list of K candidate models $M_1,\ldots,M_k$ & operates by placing prior probability on each of these models while also choosing prior probability distributions for the parameters in each model. It is then a simple matter to use Bayes rule to obtain posterior model probabilities, which provide a weight of evidence in favor of a given model relative to the alternatives synthesizing information in the prior and data. This is

a widely touted approach to account for model uncertainty, but in practice there are some clear drawbacks. Firstly, there is a well-known sensitivity to the prior on the coefficients within each model – e.g., under the Bartlett-Lindley paradox the smallest model in a nested list of models is increasingly favored as the prior variance for model-specific parameters increases. Also, there is a philosophical problem – in particular, in reality we don't really believe that any of the models in the list is exactly true but we are operating by placing positive prior probability on each model being true. This is flawed but **perhaps** is practically ok? One way to assess this is to examine asymptotics under the case in which the true model is not in the list – in this case, it is typically true that the model closest in Kullback-Leibler (KL) divergence to the true model has posterior model probability converging to one as the sample size increases. Arguably this is appealing, but there are several problems – (i) alternative ways to aggregate the models (e.g., by convex or linear combinations) may come closer to approximating the truth than asymptotically choosing the one model that is "best" [this is one reason frequentist ensembling approaches often outperform Bayes]; (ii) we end up being super confident in large samples that an incorrect model is exactly right; (iii) K-L may not be a good approximation of the actual loss function in an application. There have been various fix-ups to these problems by attempting to formally allow the M-open case in which the true model is outside the list, but this literature is under-developed & the methods seldom used in practice.

(2) Robust parametric Bayes: this type of approach typically focuses on using heavy-tailed priors and likelihood functions in an attempt to be less sensitive to the exact choices. For example, it is common to use t-distributions with low dgf, mixture models (e.g., local and/or scale mixtures of Gaussians), basis expansions in kernels or splines, etc. Such methods can improve substantially on inflexible parametric models, but substantial parametric assumptions can remain & one wonders about the practical ramifications of such assumptions on the results.

(3) Pseudo-Bayes: In frequentist semi- or nonparametric statistics, one seldom attempts to take a full likelihood-based approach to inferences but instead relies on some data reduction – e.g., through analyzing ranks, using loss functions without positing a full likelihood, etc. There is an increasingly rich literature "Bayesianizing" such ideas. In particular, whenever we have a fake likelihood function, we can potentially plug-in this bogus likelihood into Bayes rule in place of a likelihood & obtain a bogus posterior distribution that can be used in place of a true posterior. Such posteriors often inherit the frequentist asymptotic properties of the fully frequentist procedures being Bayesianized. However, the philosophical interpretation of the posterior summaries & uncertainty estimates can be challenging in these cases. In "modern" applications, it is increasing difficult to "model everything" about the data & hence some data reduction prior to defining a Bayesian model is very often necessary practically. There are several substantial success stories of such approaches – I would include Gibbs posteriors (incoming Yu & Moyeed's quantile regression approach) & Peter Hoff's extended rank likelihood. The main struggle is how to justify such methods in a Bayesian manner without relying on frequentist asymptotics? How can we interpret a "pseudo" posterior probability or credible interval in finite

samples? Perhaps we attempt a frequentist justification also in finite samples – both theoretically and through simulation studies, but is there a role for Bayesian thinking here?

(4) Nonparametric Bayes: This brings up to the general approach, which is referred to as Nonparametric Bayes or NpBayes for short. In this strategy, one attempts to define a prior distribution with "large support." For example, consider the simple case in which data consist of a scalar continuous variable y_i for subjects i=1,…,n & the density function f is unknown. From a frequentist nonparametric perspective, we could define some density estimator – e.g., through kernel smoothing. From a parametric Bayes perspective, we would choose some parametric form for the density having finitely many parameters θ, and we'd induce a prior on the density f through a prior for θ. Let $\mathcal{F}$ denote the set of all densities on the real line with respect to Lesbesgue measure, suppose that the true density that generated the data is f_0, and define neighborhoods around f_0 using some distance $d(f,f_0)$. Parametric priors for f will in general always generate densities on a vanishingly small subset of $\mathcal{F}$. If the true density does not exactly follow the parametric form, then the parametric prior assigns probability zero to small neighborhoods around f_0. The idea of large support priors is to define a prior that assigns non-zero probability around f_0 for any f_0 in a "large" subset of $\mathcal{F}$ (perhaps only ruling out weird or irregular densities) & for any neighborhood size (even tiny ones). The "game" in NpBayes is to define large support priors that are as simple and interpretable as possible, correspond to prior knowledge to the extent possible, and that lead to tractable (ideally efficient & easy) posterior computation. In the next lecture, we will detail some "canonical" priors starting in simple cases, building to more exciting applications in the subsequent lectures.

Some Relevant References:

"Bayesian Data Analysis", Gelman et al, 2013, 3rd edition – *read nonparametric Bayes chapters.*

Yang Y, Dunson DB (2014) "Minimax optimal Bayesian aggregation", arXiV:1403.1345. *Refer to the motivation and discussion on limitations of Bayesian model averaging.*

Ferguson TS (1973) "A Bayesian analysis of some nonparametric problems" Annals of Statistics 1(2):209-230. *Introduces the Bayesian nonparametrics philosophy and Dirichlet process prior.*

Chernozhukov & Hong (2003) "An MCMC approach to classical estimation", Journal of Econometrics. *An article on using Bayesian machinery with bogus likelihoods.*

Hoff PD (2007), "Extending the rank likelihood for semiparametric copula estimation," *Annals of Applied Statistics*, 1(1):265-83. *Avoids modeling marginals in doing Bayes inference.*

Miller J, Dunson DB (2015) Robust Bayesian inference via coarsening. arXiV:1506.06101. *An approach for robust parametric inference accommodating model misspecification – supposes observed data not generated exactly from assumed model.*

Lecture 2 – Trying out Some Priors

In this lecture, we will start out by considering the simple univariate density estimation problem. Using this problem as motivation, we will review finite mixture models including their implementation from a Bayesian perspective, and will additionally introduce Dirichlet process priors & then Dirichlet process mixtures. "Deep" theoretically consideration of these models (e.g., posterior consistency, convergence rates, etc) will be left to Harry & I will focus on practical issues in defining priors, performing posterior computation, dealing with label switching, interpreting the results from analyses, advantages relative to frequentist kernel smoothing, etc. Steps of a practical algorithm for computation in approximations to DP mixtures of Gaussian kernels will be described, and the students will be encouraged to try out the algorithms as a homework exercise.

Some Relevant References:

Lo AY (1984), "On a class of Bayesian nonparametric estimates. I. Density estimates." Annals of Statistics, 12(1):351-7. *Article introducing DP mixtures for densities.*

Escobar & West (1995), "Bayesian density estimation and inference using mixtures," JASA. *Article developing practical methods for implementing DP mixtures.*

Sethuraman J (1994), "A constructive definition of Dirichlet priors," Statistica Sinica 4:639-50. *Article defining stick-breaking representative of DP, which is quite useful.*

Lecture 3 – Getting more Real

Univariate density estimation is of course a toy example and in this lecture we will transition into more complex models that accommodate multivariate structure, covariates and hierarchical dependence. Due to time constraints I'll focus specifically on an application to multivariate categorical data analysis in which the data can be organized as a many way contingency table. Such data arise routinely in survey applications, epidemiology, analysis of gene sequences, etc. The resulting model we focus on also has an interpretation as shrinking a high-dimensional tensor towards a low rank structure. This will lead into a discussion of the general approach of using np Bayes mixtures to shrink towards low-dimensions while inducing a sieve-type structure that adds components slowly as sample size increases.

Some Relevant References:

Dunson DB, Xing C (2009), “Nonparametric Bayes modeling of multivariate categorical data”, JASA, 104(487):1042-51.

Kunihama T, Dunson DB (2013), “Bayesian modeling of temporal dependence in large sparse contingency tables,” JASA, 108(504):1324-38.

Zhou et al (2015), “Bayesian factorizations of big sparse tensors”, JASA, 110(512):1562-1576.

Johndrow J, Bhattacharya A, Dunson DB (2016), “Tensor decompositions and sparse log linear models,” Annals of Statistics, to appear.

Lecture 4 – Student Choice – BNP Models in Cool Apps

Unless we’ve completely run out of time going through the above material, this lecture will focus on a topic chosen by popular vote among the following candidates:

(I) Nonparametric Bayes modeling of sequences (apps to animal vocalizations)
(II) Nonparametric Bayes modeling of network/graph-valued data (brains)
(III) Nonparametric Bayes screening (apps to epigenomics)
(IV) OR free form discussion of ongoing interesting topics

<u>Relevant key reference to each of the above</u>:
(I) Sarkar A, Dunson DB (2015), “Bayesian nonparametric modeling of higher order Markov chains”, arXiV:1506.06268 [jasa, to appear]
(II) Durante D, Dunson DB, Vogelstein J (2015), “Nonparametric Bayes modeling of populations of networks,” arXiV:1406.7851.
(III) Lock E, Dunson DB (2015), “Shared kernel Bayesian screening,” Biometrika.
(IV) Papers to be written