STAT 8570: Advanced Bayesian Analysis: Modeling

Nonparametric Bayesian Inference
Autumn 2014

Instructors: Steve MacEachern
Office: 205C Cockins Hall
Email: snm@stat.osu.edu
Office Hours: W 4-5pm and Th 10:30-11:30am

Xinyi Xu
Office: 440G Cockins Hall
Email: xinyi@stat.osu.edu
Office Hours: MW 12:30-1:30pm

Lecture Hours: MWF UH 0056, 3-3:55pm

Prerequisites:

Students are expected to be comfortable with probability and basic Bayesian methods, and to have some experience with programming in R.

Course Description:

Nonparametric Bayesian inference has drawn significant attention in recent years in both the statistics and machine learning communities. By placing probability distributions on families of functions or measures, it offers a way to make use of the Bayesian calculus without the parametric handcuffs. This course discusses the theoretical concepts, computational issues, and modelling uses of nonparametric Bayesian methods. The focus will be on Dirichlet process and its variants. Core topics include constructions and properties of Dirichlet Process, the suite of Monte Carlo methods for efficient computations, and the applications of nonparametric Bayesian methods to regression, density estimation, ANOVA, etc.

Text:

We will loosely follow the structure of the book *Nonparametric Bayesian Inference* by Peter Muller and Abel Rodriguez, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 9, 2013. This book is freely downloadable at the website http://projecteuclid.org/euclid.cbms/1362163742#info . The lectures will present additional motivation for the methods and will cover selected aspects of the models in greater depth than does the text.

In addition to the text, a number of papers will be used to supplement the lectures. We include a (partial) list of recommended reading materials and innovative papers at the end of this syllabus. Some of these papers are difficult to read, and you need not read them all during the course.

Grading:

Homework	40%
Final project	60%

Homework:

Homework will be collected approximately every three weeks (making for 4-5 homework assignments during the semester). Students are encouraged to work together on the problems, but each student must hand in his or her own work. Feel free to ask the instructors for help after making an attempt at the questions.

Final project:

Student are expected to work in small groups for the course project. Each group needs to submit a written project report and make an oral presentation at the end of the semester. Further details on the project will be given in class.

Recommended reading list:

Part I (Reviews): There are a number of review papers on nonparametric Bayesian methods. Some of the earlier reviews are now dated. Two good, recent ones follow. In reading these (and these you should read for the course), you will notice differences in view between the statistics community and the machine learning community.

1) Muller, P. and Mitra, R. (2013). Bayesian Nonparametric Inference – Why and How. Bayesian Analysis, 8, 269-302.

2) Teh, Y. W. (2010). Dirichlet Process. Encyclopedia of Machine Learning, 2010. Springer. http://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf

Part II (Classics): There are a number of classic papers on nonparametric Bayesian methods. A few of them are listed below. The first four provide the theoretical basis for the Dirichlet process.

3) Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics, 1, 209-230.

Note: This is the paper that defined the area. It contains the first rigorous definition of the Dirichlet process and is notable for the description of problems to which the methods can be applied toward the end of the paper. Connections are made to Dirichlet-multinomial distributions and gamma processes. Very hard reading to get the technical details.

4) Blackwell, D. and MacQueen, J.B. (1973). Ferguson Distributions Via Polya Urn Schemes. The Annals of Statistics, 1, 353-355.

Note: A refreshingly succinct and clear description of the Dirichlet process. Written contemporaneously with Ferguson's paper, this one describes the popular Polya urn scheme (essentially renamed as the Chinese Restaurant Process some time later) which leads directly to the stick-breaking construction of the Dirichlet process and consequently to the many variations on stick-breaking that have appeared in the recent literature.

5) Antoniak, C.E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. The Annals of Statistics, 2, 1152-1174.

Note: One of the most misunderstood papers in the area. This is the early paper on mixtures of Dirichlet processes which provide the key to nonparametric Bayesian prior distributions for continuous distributions. The work encompasses what has come to be called the Dirichlet process mixture model. This paper is hard to read. I suspect that many have only skimmed it and so missed important contents in it.

6) Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. Statistica Sinica, 4, 639-650.

Note: A clean presentation and extension of the stick-breaking representation that appeared earlier in work by Sethuraman and Tiwari. Truly beautiful mathematics.

<u>Part III (Applications):</u> The Dirichlet process and its variants have been applied to an extraordinary variety of problems. The early applications concentrated on problems that could be solved analytically or where one could grunt out a solution for a small problem. Many of the concepts and techniques used to improve computation here later resurfaced to make Markov chain Monte Carlo (MCMC) computation work effectively. Here are three that set the stage for much of the work that followed.

7) Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations. Journal of the American Statistical Association, 71, 897-902.

Note: This paper provides a Bayesian version of the famous Kaplan-Meier estimator for observed and right-censored event times. The paper stands out for its closed-form expression for the posterior (a mixture of Dirichlet processes) and suggested later developments which (loosely) pursue the "multinomial-product beta" version of conjugacy rather than the more restrictive "multinomial-Dirichlet distribution" version of conjugacy. Not done in this paper, but easily handled with a simple numerical integration, is a distribution over base measures.

8) Berry, D.A. and Christensen, R. (1979). Empirical Bayes Estimation of a Binomial Parameter Via Mixtures of Dirichlet Processes. The Annals of Statistics, 7, 558-568.

Note: The paper from the most remarkable Master's thesis I know of. This paper applies the Dirichlet process in a hierarchical fashion to address the compound decision problem (known then as the empirical Bayes problem, though the methods in the paper are purely Bayesian). Far ahead of its time in using the Dirichlet process for a clear modelling objective.

9) Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. The Annals of Statistics 12, 351-357.

Note: Develops the view that the Dirichlet process can be used to produce Bayesian density estimates when convolved with a continuous smoothing kernel. Although the methods could not be implemented in realistic problems at the time, the perspective conveyed in the paper is timeless.

<u>Part IV (Computation):</u> Modern use of the methods took off with the development of MCMC simulation methods. The early development was in Mike Escobar's PhD 1988 dissertation. Remarkably, this predates the famous 'Gibbs sampler' paper of Gelfand and Smith.

10) Escobar, M.D. (1994). Estimating Normal Means with a Dirichlet Process Prior. Journal of the American Statistical Association, 89, 268-277.

Note: The main paper from Escobar's dissertation. It lays out a basic MCMC method for a mixture of Dirichlet processes (MDP) model along the lines of Berry and Christensen's model. The benefits of using the better model are so great that that the MDP estimator trounces parametric Bayes and empirical Bayes methods, even with MCMC methods that are, by today's standards, based on too-short runs

11) MacEachern, S.N. (1998). Computational Methods for Mixture of Dirichlet Process Models, in "Practical Nonparametric and Semiparametric Bayesian Statistics", D. Dey, P. Mueller, D. Sinha (eds.), 23-44. Springer-Verlag.

Note: This book chapter describes computational innovations from a 1994 paper by MacEachern in the Communications in Statistics and a 1996 Biometrika paper MacEachern and Bush. The innovations (fully exploiting conjugacy; introducing a `remixing' step through reparameterization) greatly improve mixing of MCMC methods. There is also a bit more in the chapter on preintegration over the mass parameter of the Dirichlet process. Preintegration later proved useful for models that go beyond the Dirichlet process (e.g., the Pitman-Yor process). See also Radford Neal's work on improving mixing.

12) MacEachern, S.N., and Mueller, P. (1998). Estimating Mixture of Dirichlet Process Models. Journal of Computational and Graphical Statistics, 7, 223-238.

Note: The first paper to provide an MCMC method for nonconjugate MDP models. One algorithm appears in the paper, a second was trimmed during the editorial process. To this day, we get questions about one tricky step in the sampler which suggests to me that we had to trim too much from the original submission.

13) Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, 9, 249-265.

Note: Further computational strategies for nonconjugate MDP models. The methods fall under the Metropolis-Hastings framework for MCMC and many find them easier to follow than those in MacEachern and Muller (1998).

Subsequent computational developments include split-merge moves (Radford Neal and Sonia Jain; David Dahl), variational approximations (Michael Jordan and a suite of coauthors), and many more. Computational development is active at the moment, with scalability of the algorithm as the main theme. Not surprisingly, this includes investigation into parallelizing algorithms and into the use of approximations—by which we mean approximations other than those inherent in MCMC.

Part V (Other): The development of computational methods was accompanied by use of the models in an astonishing variety of problems. The few papers below are categorized by statistical content rather than by area of application. They represent basic uses of the models. Far more complicated models have been developed when needed by the application.

14) Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association, 90, 577-588.

Note: At long last, Lo (1984)'s density estimates, implemented via MCMC. An interesting perspective in this and further papers, especially by Mike West, is that (with some exaggeration), all of statistics can be reduced to density estimation. For example, regression? Estimate the joint density of (x,y) and extract the conditional mean of y|x.

15) Bush, C.A. and MacEachern, S.N. (1996). A Semiparametric Bayesian Model for Randomised Block Designs. Biometrika, 83, 275-285.

Note: This one makes use of covariates in the model and draws a distinction between fixed effects and random effects from the Bayesian perspective. The early use of a mix of parametric and nonparametric techniques for the mixed model.

More papers will be added over the next few days. Yet to come are more applications and a variety of extensions of the Dirichlet process.