Nonparametric Bayesian Statistics

Debdeep Pati
Florida State University

October 14, 2014

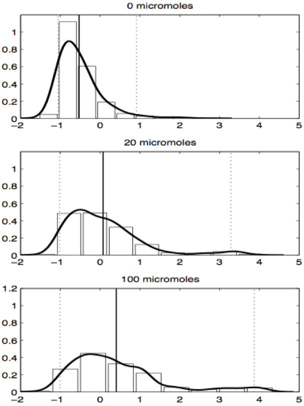
Dependent random probability measures - Motivating application

- Inferences on changes in response distributions across experimental conditions
- Assessing changes in a health response with continuous predictors (dose of treatment or environmental exposure)
- Clustering hospitals based on quality of care

Example 1: Modeling $y \mid x$, x: categorical

- Interest in studying changes in a response distribution across experimental conditions
- In genotoxicity experiments, y_i = measure of DNA damage, $x_i \in \{1, \ldots, T\}$ = dose group
- Interest in assessing how density of y changes with dose

Changes in DNA damage with hydrogen peroxide



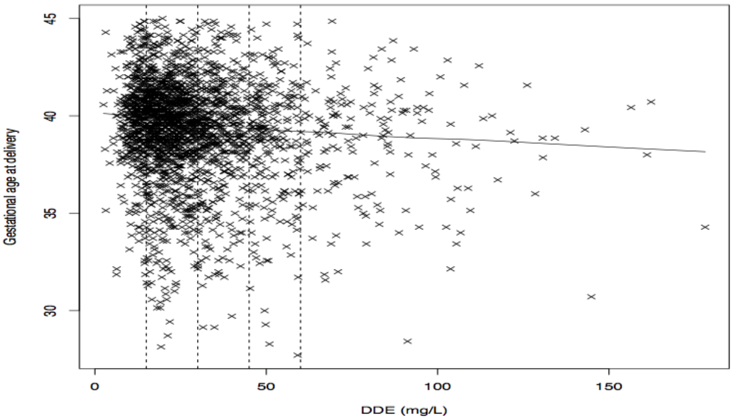
Example 2: Modeling $y \mid x$, x: possibly continuous

- Commonly interest in studying changes in distribution of a health response with predictors
- Premature delivery is major public health problem
- Current epidemiologic practice dichotomizes preterm birth at 37 weeks
- Let y_i=gestational age at delivery, x_i=predictors
- How to assess changes in distribution of Y with X?

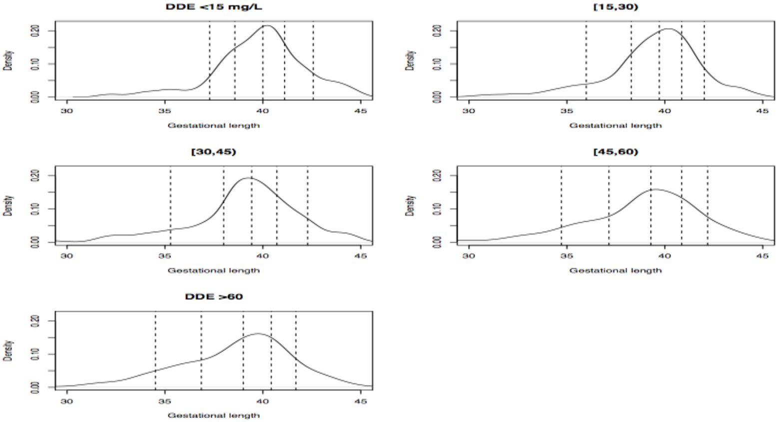
Example 2: Modeling length of gestation

- Preterm birth is a major public health problem leading to substantial mortality & short and long-term morbidity
- Preterm birth is typically defined as a delivery occurring prior to 37 weeks of completed gestation
- This cutoff is somewhat arbitrary & the shorter the length of gestation, the more adverse the associated health effects
- Appealing to model the distribution of gestational age at delivery as unknown & then allow predictors to impact this distribution

Gestational Length vs. DDE(mg/L)



Gestational Length Densities within DDE Categories



Comments on Gestational Length Data

- Data are non-Gaussian with a left skew
- Not straightforward to transform the data to approximate normality
- A different transformation would be needed within each DDE category
- Question: how to characterize gestational age at delivery distribution without given predictor $X = x$?

Motivating Example 3: Clustering health centers

- Goal is clustering of health centers and identification of outlying centers
- y_{ij} = proportion of patients given most appropriate antibiotic in hospital j of state i
- How to nonparametrically estimate distribution of Y for each state, while clustering states & borrowing information?

Dependent Random Probability Measures (RPM)

- In examples 1-3 we have multiple unknown distributions
- These unknown distributions are related, so it is appealing to characterize dependence through the prior
- Allows borrowing of information and reduction of parameters
- Requires alternatives to traditional DP formulations

Dependent RPMs

- We have focused on the case in which we have a single RPM, P
- For example, P may correspond to an unknown random effects distribution
- In many settings, it is of interest to consider a dependent collection of RPMs, $P_{\mathcal{X}} = \{P_x : x \in \mathcal{X}\}$.
- $P_x =$ RPM specific to index x
- x may correspond to time, space, or predictors

Definition of the Dependent DP (DDP)

- MacEachern (1999, 2001) proposed the following formulation,

$$P_x = \sum_{h=1}^{\infty} \pi_h(x)\delta_{\Theta_h(x)}, \quad \Theta_h \sim P_0, h = 1, \ldots, \infty$$

- To obtain $P_x \sim DP(\alpha P_{0x})$ marginally at each $x \in X$, let

$$\pi_h(x) = V_h(x) \prod_{l<h}(1 - V_l(x)), V_h(x) \sim \text{Beta}(1, \alpha)$$

- P_0 is a stochastic process over $\mathcal{X}$ - for example, a Gaussian process

Fixed π−DDP

- It is not obvious how to define a predictor-dependent stick-breaking process having the appropriate properties
- Typical focus is on "fixed-π" DDP:

$$P_x = \sum_{h=1}^{\infty} \pi_h \delta_{\Theta_h(x)}, \Theta_h \sim P_0, h = 1, \ldots, \infty$$

- π_h have typical DP stick-breaking form
- De Iorio et al. 2004 define an ANOVA DDP model & Gelfand et al. (2005) proposed a spatial DDP

Back to Example 2: Quality of Care

- $y_{ij} \sim P_i$, with P_i distribution of quality of care measure across hospitals in state i
- Goal is to cluster states in terms of quality of care of the hospitals
- Important to not just cluster mean or a single attribute of distribution
- Two states with the same mean may have very different tails

Nested Dirichlet Process (nDP) (Rodriguez et al., 08)

- Let f_i = density of the outcome measure in state i,

$$
\begin{aligned}
f_i(y) &= \int N(y;\mu,\tau)dP_i(\mu,\tau) \\
P_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\delta_{P_h^*}, V_h \sim \text{Beta}(1,\alpha) \\
P_h^* &\sim DP(\beta P_0)
\end{aligned}
$$

- DP Atoms in dependent Dirichlet process

Some Comments on nDP

- Provides an approach for non-parametric clustering of distributions instead of parameters
- Hospitals having the same distribution of patient outcomes will be clustered together
- Even if no interest in clustering, nDP useful as an approach for borrowing information
- Application of nDP to sequential data - Ni et al. (2007, ICML)

Comparison between HDP and nDP