Applied Bayesian Nonparametric Mixture Modeling
Session 1 – Introduction/Dirichlet process priors

Athanasios Kottas (thanos@ams.ucsc.edu)
Abel Rodriguez (abel@ams.ucsc.edu)

University of California, Santa Cruz

2014 ISBA World Meeting
Cancún, Mexico
Sunday July 13, 2014
Outline

1. Introduction and motivation
2. Definition of the Dirichlet process
3. Constructive definition of the Dirichlet process
4. Pólya urn representation of the Dirichlet process
5. Posterior inference
6. Applications
Motivating example: ROC curves for biomarker evaluation

- Individuals are drawn from one of two populations (call them Healthy and Diseased).

- We measure a univariate biomarker $Y$ in order to diagnose a disease (e.g., prostate-specific antigen – PSA – to detect prostate cancer).

- Individuals with an “elevated” biomarker (i.e., $Y > t$ for some threshold $t$) are declared ”Diseased”.
  - Sensitivity (or true positive rate)
  - Specificity (or true negative rate)

- We assume that a *gold-standard* is available: disease status can be assumed known for the training set used to estimate the reliability of the diagnostic test.
Motivating example: ROC curves for biomarker evaluation

Let $F_H$ and $F_D$ be the c.d.f.s of the biomarker for the healthy and diseased groups.

- $FPR(t) = 1 - F_H(t) = \bar{F}_H(t)$.
- $FNR(t) = F_D(t) = 1 - \bar{F}_D(t)$.
- Selecting $t$ implies a tradeoff between FPR and FNR!
Motivating example: ROC curves for biomarker evaluation

That tradeoff is captured by the ROC curve, implicitly defined as

\[ \{(\text{FNR}(t), \text{TPR}(t)) : t \in (-\infty, \infty)\} \]

Alternatively

\[ \text{ROC}(u) = 1 - F_D(F_H^{-1}(1 - u)) \]

for \( u \in [0, 1] \).

Hence, estimating the ROC curve requires estimation of two distribution functions.
Motivating example: ROC curves for biomarker evaluation

- Blood PSA levels for 454 healthy and 229 diseased males, with ages ranging from 47 to 81 years (Etzioni et al., 1999).

- Two simple (frequentist) analyses for the data involve
  - Parametric: Assume normality of logscores (skewness?).
  - Nonparametric: Estimate c.d.f.s through the empirical c.d.f.s.
Motivating example: ROC curves for biomarker evaluation

- ROC curves and bootstrap-based pointwise confidence intervals for both simple analyses.
- As a Bayesian, obtaining an analog to the left-hand picture is easy. But how do we get a picture that is analogous to the one on the center?
Parametric vs. nonparametric Bayes: A simple example

- Let $y_i \in \mathcal{Y}$, $y_i \mid F \sim_{iid} F$, $F \in \mathcal{F}^*$,

$$\mathcal{F}^* = \{N(y \mid \mu, \tau^2), \mu \in \mathbb{R}, \tau \in \mathbb{R}^+\}.$$  

- In this **parametric** specification a prior on $\mathcal{F}^*$ boils down to a prior on $(\mu, \tau^2)$.

- However, $\mathcal{F}^*$ is tiny compared to

$$\mathcal{F} = \{\text{All distributions on } \mathcal{Y}\}.$$  

- **Nonparametric Bayes** involves priors on much larger subsets of $\mathcal{F}$ (infinite-dimensional spaces).

- One handy way to do this is to use stochastic processes.
Bayesian nonparametrics

- Bayesian nonparametrics $\Leftrightarrow$ priors on spaces of functions, $\{g(\cdot): g \in G\}$ (infinite dimensional spaces). An oxymoron?

- Even though we focus on distribution functions, the methods are much more widely useful: hazard or cumulative hazard function, link function, calibration function ...

- More generally, enriching usual parametric models, typically leading to semiparametric models.

- Wandering nonparametrically near a standard class.

- What makes a nonparametric model “good”? (e.g., Ferguson, 1973)
  - The model should be tractable, i.e., it should be easily computed, either analytically or through simulations.
  - The model should be rich, in the sense of having large support.
  - The hyperparameters in the model should be easily interpretable.
Some references

- General review papers on Bayesian nonparametrics: Walker, Damien, Laud and Smith (1999); Müller and Quintana (2004); Hanson, Branscum and Johnson (2005).

- Review papers on specific application areas of Bayesian nonparametric and semiparametric methods: Hjort (1996); Sinha and Dey (1997); Gelfand (1999).


- Software: functions for nonparametric Bayesian inference are spread over various R packages. Only comprehensive package we are aware of is the DPpackage.
The Dirichlet process as a model for random distributions

- A Bayesian nonparametric approach to modeling, say, distribution functions, requires priors for spaces of distribution functions.

- Formally, it requires stochastic processes with sample paths that are distribution functions defined on an appropriate sample space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbb{R}$, or $\mathbb{R}^+$, or $\mathbb{R}^d$), equipped with a $\sigma$-field $\mathcal{B}$ of subsets of $\mathcal{X}$ (e.g., the Borel $\sigma$-field for $\mathcal{X} \subseteq \mathbb{R}^d$)

- The **Dirichlet process** (DP), anticipated in the work of Freedman (1963) and Fabius (1964), and formally developed by Ferguson (1973, 1974), is the first prior defined for spaces of distribution functions.

- The DP is, formally, a (random) probability measure on the space of probability measures (distributions) on $(\mathcal{X}, \mathcal{B})$

- Hence, the DP generates random distributions on $(\mathcal{X}, \mathcal{B})$, and thus, for $\mathcal{X} \subseteq \mathbb{R}^d$, equivalently, random c.d.f.s on $\mathcal{X}$
Motivating the construction of the Dirichlet process

• Suppose you are dealing with a sample space with only two outcomes, say, $\mathcal{X} = \{0, 1\}$ and you are interested in estimating $x$, the probability of observing 1.

• A natural prior for $x$ is a beta distribution,

$$p(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1.$$  

• More generally, if $\mathcal{X}$ is finite with $q$ elements, the probability distribution over $\mathcal{X}$ is given by $q$ numbers $x_1, \ldots, x_q$ such that $\sum_{i=1}^{q} x_i = 1$. A natural prior for $(x_1, \ldots, x_q)$, which generalizes the Beta distribution, is the Dirichlet distribution (see the next two slides).

• With the Dirichlet process we further generalize to infinite but countable spaces.
Properties of the Dirichlet distribution

- Start with independent random variables

\[ Z_j \sim \text{Gam}(a_j, 1), \quad j = 1, \ldots, k, \]

with \( a_j > 0 \).

- Define

\[ Y_j = \frac{Z_j}{\sum_{\ell=1}^{k} Z_\ell}, \quad j = 1, \ldots, k. \]

- Then \((Y_1, \ldots, Y_k) \sim \text{Dirichlet}(a_1, \ldots, a_k)\).

- This distribution is singular w.r.t. Lebesgue measure on \( \mathbb{R}^k \), since \( \sum_{j=1}^{k} Y_j = 1 \).
Properties of the Dirichlet distribution

- \((Y_1, ..., Y_{k-1})\) has density
  \[
  \frac{\Gamma \left( \sum_{j=1}^{k} a_j \right)}{\prod_{j=1}^{k} \Gamma (a_j)} \left( 1 - \sum_{j=1}^{k-1} y_j \right)^{a_{k-1}} \prod_{j=1}^{k-1} y_j^{a_j-1}.
  \]

- Note that for \(k = 2\), Dirichlet\((a_1, a_2)\) \(\equiv\) Beta\((a_1, a_2)\).

- The moments of the Dirichlet distribution are:
  \[
  \begin{align*}
  \mathbb{E}(Y_j) &= \frac{a_j}{\sum_{\ell=1}^{k} a_\ell}, \\
  \mathbb{E}(Y_j^2) &= \frac{a_j(a_j + 1)}{\sum_{\ell=1}^{k} a_\ell(1 + \sum_{\ell=1}^{k} a_\ell)}, \\
  \mathbb{E}(Y_i Y_j) &= \frac{a_i a_j}{\sum_{\ell=1}^{k} a_\ell(1 + \sum_{\ell=1}^{k} a_\ell)}.
  \end{align*}
  \]

- We can think about the Dirichlet as having two parameters:
  - \(g = \{a_j/ \sum_{\ell=1}^{k} a_\ell\}\), the mean vector.
  - \(\alpha = \sum_{\ell=1}^{k} a_\ell\), a concentration parameter controlling its variance.
Definition of the Dirichlet process

- The DP is characterized by two parameters:
  - A positive scalar parameter $\alpha$.
  - A specified probability measure on $(\mathcal{X}, \mathcal{B})$, $Q_0$ (or, equivalently, a distribution function on $\mathcal{X}$, $G_0$).

- **DEFINITION** (Ferguson, 1973): The DP generates random probability measures (random distributions) $Q$ on $(\mathcal{X}, \mathcal{B})$ such that for any finite measurable partition $B_1, \ldots, B_k$ of $\mathcal{X}$,

$$(Q(B_1), \ldots, Q(B_k)) \sim \text{Dirichlet}(\alpha Q_0(B_1), \ldots, \alpha Q_0(B_k)).$$

- Here, $Q(B_i)$ (a random variable) and $Q_0(B_i)$ (a constant) denote the probability of set $B_i$ under $Q$ and $Q_0$, respectively.
- Also, the $B_i$, $i = 1, \ldots, k$, define a measurable partition if $B_i \in \mathcal{B}$, they are pairwise disjoint, and their union is $\mathcal{X}$.
Interpreting the parameters of the Dirichlet process

- For any measurable subset $B$ of $\mathcal{X}$, we have from the definition that $Q(B) \sim \text{beta}(\alpha Q_0(B), \alpha Q_0(B^c))$, and thus
  \[
  \mathbb{E}\{Q(B)\} = Q_0(B), \quad \text{Var}\{Q(B)\} = \frac{Q_0(B)\{1 - Q_0(B)\}}{\alpha + 1}
  \]

- $Q_0$ plays the role of the center of the DP (also referred to as baseline probability measure, or baseline distribution).

- $\alpha$ can be viewed as a precision parameter: for large $\alpha$ there is small variability in DP realizations; the larger $\alpha$ is, the closer we expect a realization $Q$ from the process to be to $Q_0$.

- See Ferguson (1973) for the role of $Q_0$ on more technical properties of the DP (e.g., Ferguson shows that the support of the DP contains all probability measures on $(\mathcal{X}, \mathcal{B})$ that are absolutely continuous w.r.t. $Q_0$).
Interpreting the parameters of the Dirichlet process

- Analogous definition for the random distribution function $G$ on $X$ generated from a DP with parameters $\alpha$ and $G_0$, a specified distribution function on $X$.

- For example, with $X = \mathbb{R}$, $B = (-\infty, x]$, $x \in \mathbb{R}$, and $Q(B) = G(x)$,

  $G(x) \sim \text{beta}(\alpha G_0(x), \alpha \{1 - G_0(x)\})$,

  and thus

  $$\mathbb{E}\{G(x)\} = G_0(x), \quad \text{Var}\{G(x)\} = \frac{G_0(x)\{1 - G_0(x)\}}{\alpha + 1}.$$  

- **Notation**: depending on the context, $G$ will denote either the random probability measure or the random distribution function.

- $G \sim \text{DP}(\alpha, G_0)$ will indicate that a DP prior is placed on $G$. 
Simulating c.d.f. realizations from a Dirichlet process

- Consider any grid of points $x_1 < x_2 < \ldots < x_k$ in $\mathcal{X} \subseteq \mathbb{R}$.
- Then, the random vector 

$$(G(x_1), G(x_2) - G(x_1), \ldots, G(x_k) - G(x_{k-1}), 1 - G(x_k))$$

follows a Dirichlet distribution with parameter vector 

$$(\alpha G_0(x_1), \alpha(G_0(x_2) - G_0(x_1)), \ldots, \alpha(G_0(x_k) - G_0(x_{k-1})), \alpha(1 - G_0(x_k)))$$

- Hence, if $(u_1, u_2, \ldots, u_k)$ is a draw from this Dirichlet distribution, then $(u_1, \ldots, \sum_{j=1}^{i} u_j, \ldots, \sum_{j=1}^{k} u_j)$ is a draw from the distribution of 

$$(G(x_1), \ldots, G(x_i), \ldots, G(x_k)).$$
library(MCMCpack)
alpha = 1
x = seq(0.005, 0.995, by=0.005)
k = length(x)
dG0 = punif(x[-1],0,1)-punif(x[-k],0,1)
dG = rdirichlet(1, alpha*dG0)
G = cumsum(dG)
Simulating c.d.f. realizations from a Dirichlet process

\[ \alpha = 0.1 \]  
\[ \alpha = 10 \]  
\[ \alpha = 100 \]

Figure 1.1: Realizations from a Dirichlet process with different values for the precision parameter. The solid black line corresponds to the baseline measure (a uniform on the interval \([0, 1]\) in this case), while the dashed colored lines represent multiple realizations.
Constructive definition of the DP

- Due to Sethuraman and Tiwari (1982) and Sethuraman (1994).

- Let \( \{z_r : r = 1, 2, \ldots\} \) and \( \{\vartheta_\ell : \ell = 1, 2, \ldots\} \) be independent sequences of i.i.d. random variables
  - \( z_r \sim \text{beta}(1, \alpha), \ r = 1, 2, \ldots \)
  - \( \vartheta_\ell \sim G_0, \ \ell = 1, 2, \ldots \)

- Define \( \omega_1 = z_1 \) and \( \omega_\ell = z_\ell \prod_{r=1}^{\ell-1} (1 - z_r) \), for \( \ell = 2, 3, \ldots \)

- Then, a realization \( G \) from \( \text{DP}(\alpha, G_0) \) is (almost surely) of the form

\[
G(\cdot) = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\vartheta_\ell}(\cdot)
\]

where \( \delta_z(\cdot) \) denotes a point mass at \( z \).

- Note this definition implies \( \sum_{\ell=1}^{\infty} \omega_\ell = 1 \) (almost surely).
More on the constructive definition of the DP

- The DP generates distributions that have an (almost sure) representation as countable mixtures of point masses:
  - The locations $\vartheta_\ell$ are i.i.d. draws from the base distribution.
  - The associated weights $\omega_\ell$ are defined using the stick-breaking (sometimes called residual allocation) construction.

- This is not as restrictive as it might sound: Any distribution on $\mathbb{R}^d$ can be approximated arbitrarily well using a countable mixture of point masses.

- The realizations we showed before already hinted at this fact.
The stick-breaking construction

- Start with a stick of length 1 (representing the total probability to be distributed among the different atoms).

- Draw a random $z_1 \sim \text{beta}(1, \alpha)$, which defines the portion of the original stick assigned to atom 1, so that $\omega_1 = z_1$ — then, the remaining part of the stick has length $1 - z_1$.

- Draw a random $z_2 \sim \text{beta}(1, \alpha)$ (independently of $z_1$), which defines the portion of the remaining stick assigned to atom 2, therefore, $\omega_2 = z_2(1 - z_1)$ — now, the remaining part of the stick has length $(1 - z_2)(1 - z_1)$.

- Continue ad infinitum ....

- We denote the joint distribution on the vector of weights by $(\omega_1, \omega_2, \ldots) \sim \text{SB}(\alpha)$. 
The stick-breaking construction
More on the constructive definition of the DP

- The DP constructive definition yields another method to simulate from DP priors — in fact, it provides (up to a truncation approximation) the entire distribution $G$, not just c.d.f. sample paths.

- For example, a possible approximation is $G_J = \sum_{j=1}^{J} p_j \delta_{\vartheta_j}$, with $p_j = \omega_j$ for $j = 1, \ldots, J - 1$, and $p_J = 1 - \sum_{j=1}^{J-1} \omega_j = \prod_{r=1}^{J-1} (1 - z_r)$.

- To specify $J$, note that

$$E \left( \sum_{j=1}^{J} \omega_j \right) = 1 - \prod_{r=1}^{J} E(1 - z_r) = 1 - \prod_{r=1}^{J} \frac{\alpha}{\alpha + 1} = 1 - \left( \frac{\alpha}{\alpha + 1} \right)^J$$

Hence, $J$ could be chosen such that $(\alpha/\alpha+1)^J = \varepsilon$, for small $\varepsilon$. 
More on the constructive definition of the DP

Figure 1.2: Illustration for a DP with $G_0 = N(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 $N(0, 1)$ draws with heights given by the (truncated) stick-breaking weights. These spikes are then summed to generate one c.d.f. sample path. The right panel shows 8 such sample paths indicated by the lighter jagged lines. The heavy smooth line indicates the $N(0, 1)$ c.d.f.
Generalizing the DP

Many random probability measures can be defined by means of a stick-breaking construction – the $z_r$ are drawn independently from a distribution on $[0, 1]$.

- For example, the Beta two-parameter process (Ishwaran and Zarepour, 2000) is defined by choosing $z_r \sim \text{beta}(a, b)$.
- If $z_r \sim \text{beta}(1 - a, b + ra)$, for $r = 1, 2, \ldots$ and some $a \in [0, 1)$ and $b \in (-a, \infty)$ we obtain the two-parameter Poisson-Dirichlet process (e.g., Pitman and Yor, 1997).
- The general case, $z_r \sim \text{beta}(a_r, b_r)$ (Ishwaran and James, 2001).
- The probit stick-breaking process, where $z_r = \Phi(x_r)$ with $x_r \sim \mathcal{N}(\mu, \sigma^2)$ and $\Phi$ denoting the standard normal c.d.f. (Rodríguez and Dunson, 2011).
Further extensions based on the DP constructive definition

The constructive definition of the DP has motivated several of its extensions, including:

- $\epsilon$-DP (Muliere and Tardella, 1998), generalized DPs (Hjort, 2000); general stick-breaking priors (Ishwaran and James, 2001).

- Dependent DP priors (MacEachern, 1999, 2000; De Iorio et al., 2004; Griffin and Steel, 2006).

- Hierarchical DPs (Tomlinson and Escobar, 1999; Teh et al., 2006).

- Spatial DPs models (Gelfand, Kottas and MacEachern, 2005; Kottas, Duan and Gelfand, 2008; Duan, Guindani and Gelfand, 2007).

- Nested DPs (Rodriguez, Dunson and Gelfand, 2008).
Pólya urn characterization of the DP

- If, for $i = 1, \ldots, n$, $x_i \mid G$ are i.i.d. from $G$, and $G \sim \text{DP}(\alpha, G_0)$, the joint distribution for the $x_i$ induced by marginalizing $G$ over its DP prior is given by

  $$p(x_1, \ldots, x_n) = G_0(x_1) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(x_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{x_j}(x_i) \right\}$$

  (Blackwell and MacQueen, 1973).

- That is, the sequence of the $x_i$ follows a generalized Pólya urn scheme such that:
  - $x_1 \sim G_0$, and
  - for any $i = 2, \ldots, n$, $x_i \mid x_1, \ldots, x_{i-1}$ follows the mixed distribution that places point mass $(\alpha + i - 1)^{-1}$ at $x_j$, $j = 1, \ldots, i - 1$, and continuous mass $\alpha(\alpha + i - 1)^{-1}$ on $G_0$. 

Athanasios Kottas and Abel Rodriguez
Applied Bayesian Nonparametric Mixture Modeling – Session 1
The Chinese restaurant process

The Pólya urn characterization of the DP can be visualized using the Chinese restaurant analogy:

- A customer arriving at the restaurant joins a table that already has some customers, with probability proportional to the number of people in the table, or takes the first seat at a new table with probability proportional to $\alpha$.

- All customers sitting in the same table share a dish.
Prior to posterior updating with DP priors

- Let $G$ denote the random distribution function for the following results.
- Ferguson (1973) has shown that if the observations $y_i \mid G$ are i.i.d. from $G$, $i = 1, \ldots, n$, and $G \sim \text{DP}(\alpha, G_0)$, then the posterior distribution of $G$ is a $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$, with $\tilde{\alpha} = \alpha + n$, and

$$
\tilde{G}_0(t) = \frac{\alpha}{\alpha + n} G_0(t) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{[y_i, \infty)}(t)
$$

- Hence, the DP is a conjugate prior.
- All the results and properties developed for DPs can be used directly for the posterior distribution of $G$. 
Prior to posterior updating with DP priors

- For example, the posterior mean estimate for $G(t)$,

$$\mathbb{E}\{G(t) \mid y_1, \ldots, y_n\} = \frac{\alpha}{\alpha + n} G_0(t) + \frac{n}{\alpha + n} G_n(t)$$

where $G_n(t) = n^{-1} \sum_{i=1}^{n} 1_{[y_i, \infty)}(t)$ is the empirical distribution function of the data (the standard classical nonparametric estimator).

- For small $\alpha$ relative to $n$, little weight is placed on the prior guess $G_0$.
- For large $\alpha$ relative to $n$, little weight is placed on the data.
- Hence, $\alpha$ can be viewed as a measure of faith in the prior guess $G_0$ measured in units of number of observations (thus, $\alpha = 1$ indicates strength of belief in $G_0$ worth one observation).
- However, taking $\alpha$ very small in order to be “noninformative” is very dangerous.
C.d.f. estimation using DP priors

Figure 1.3: Estimating the distribution function under a DP prior, using simulated data. Both the true distribution generating the data and the baseline distribution are Gaussian. The left panel corresponds to a sample of $n = 10$ observations while the right panel corresponds to a sample of $n = 50$ observations.
A Bayesian nonparametric approach to ROC estimation

- Let $y_{i,H} \mid F_H \sim F_H$ for $i = 1, \ldots, n_H$, and $y_{j,D} \mid F_D \sim F_D$ for $j = 1, \ldots, n_D$.

- Assign independent Dirichlet process priors to each $F_H$ and $F_D$.

  $$F_H \sim \text{DP}(\alpha_H, G_{0,H}), \quad F_D \sim \text{DP}(\alpha_D, G_{0,D}).$$

  (Neither $\alpha_H$ and $\alpha_D$ nor $G_{0,H}$ and $G_{0,D}$ need to be the same).

- Questions about the PSA example:
  - How would you choose the baseline measure?
  - How would you assess the impact of these choices in your analysis?
A Bayesian nonparametric approach to ROC estimation

- How would you choose the baseline measure?
  - In this case, it would make sense to take $G_{0,D} = G_{0,H} = \mathcal{N}(0, 1)$ (both from our knowledge of common PSA levels and from our descriptive analysis of the data).
  - It is less clear how to choose $\alpha_D$ and $\alpha_H$. Maybe $\alpha_D = \alpha_H = 5$? Recall the equivalent sample size interpretation!

- How would you assess the impact of these choices in your analysis?
  - We can simulate from this prior using the truncation approximation.
  - Since $G_{0,D} = G_{0,H}$, the mean ROC curve is a diagonal line (ignoring reflections).
  - Larger values of $\alpha_H$ and $\alpha_D$ imply a prior more concentrated around this mean.
A Bayesian nonparametric approach to ROC estimation

- Posterior distributions

\[ F_H \mid \{y_{i,H}\} \sim \text{DP} \left( \frac{\alpha_H}{\alpha_H + n_H} G_{0,H} + \frac{1}{\alpha_H + n_H} \sum_{i=1}^{n_H} 1[\{y_{i,H}\}, \infty) \right) \]

\[ F_D \mid \{y_{j,D}\} \sim \text{DP} \left( \frac{\alpha_D}{\alpha_D + n_D} G_{0,D} + \frac{1}{\alpha_D + n_D} \sum_{j=1}^{n_D} 1[\{y_{j,D}\}, \infty) \right) \]

- Inferences about the ROC curve can be obtained by sampling posterior realizations of \( F_H \) and \( F_D \) using a truncation.

  - Note that \( E(ROC(u) \mid \text{data}) \neq 1 - \tilde{F}_D(\tilde{F}_H^{-1}(1 - u))! \)

- Because sample sizes are fairly large, results are similar to those from the (frequentist) nonparametric model we presented in the introduction.
Dose-response modeling with Dirichlet process priors

- Study potency of a stimulus by administering it at $k$ dose levels to a number of subjects at each level.
  - $x_i$: dose levels (with $x_1 < x_2 < ... < x_k$).
  - $n_i$: number of subjects at dose level $i$.
  - $y_i$: number of positive responses at dose level $i$.
- $F(x) = \Pr(\text{positive response at dose level } x)$ (i.e., the potency of level $x$ of the stimulus).
- $F$ is referred to as the potency curve, or dose-response curve, or tolerance distribution.
- Standard assumption in bioassay settings: the probability of a positive response increases with the dose level, i.e., $F$ is a non-decreasing function, i.e., $F$ can be modeled as a c.d.f. on $\mathcal{X} \subseteq \mathbb{R}$.
Dose-response modeling with Dirichlet process priors

Questions of interest:
- Inference for $F(x)$ for specified dose levels $x$.
- Inference for unobserved dose level $x_0$ such that $F(x_0) = \gamma$ for specified $\gamma \in (0, 1)$.
- Optimal selection of $\{x_i, n_i\}$ to best accomplish goals 1 and 2 above (design problem).

Parametric modeling: $F$ is assumed to be a member of a parametric family of c.d.f.s (e.g., logit, or probit models).

Dose-response modeling with Dirichlet process priors

- Assuming (conditionally) independent outcomes at different dose levels, the likelihood is given by

\[ \prod_{i=1}^{k} p_i^{y_i} (1 - p_i)^{n_i - y_i}, \]

where \( p_i = F(x_i) \) for \( i = 1, \ldots, k \).

- If the prior for \( F \) is a DP with precision parameter \( \alpha > 0 \) and base c.d.f. \( F_0 \) (the prior guess for the potency curve), then a priori

\[
(p_1, p_2 - p_1, \ldots, p_k - p_{k-1}, 1 - p_k)
\]

follows a Dirichlet distribution with parameters

\[
(\alpha F_0(x_1), \alpha(F_0(x_2) - F_0(x_1)), \ldots, \alpha(F_0(x_k) - F_0(x_{k-1})), \alpha(1 - F_0(x_k))).
\]

- The posterior for \( F \) is a mixture of Dirichlet processes (Antoniak, 1974).
Mixtures of Dirichlet processes

- Extension of the DP to a hierarchical version: set

\[ G \mid \alpha, \psi \sim \text{DP}(\alpha, G_0(\cdot \mid \psi)), \]

where (parametric) priors are added to the precision parameter \( \alpha \) and/or the parameters of the base distribution, \( \psi \).

- Not surprisingly, mixtures of Dirichlet processes are also conjugate priors.

- Mixtures of Dirichlet processes are different from Dirichlet process mixture models (which we will study in Session 2). However, there are important connections between the two.
Introduction

Definition of the DP

Constructive definition of the DP

Pólya urn

Posterior inference

Applications

---

Bayesian nonparametric modeling for cytogenetic dosimetry

- Cytogenetic dosimetry (in vitro setting): samples of cell cultures exposed to a range of doses of a given agent — in each sample, at each dose level, a measure of cell disability is recorded.

- Dose-response modeling framework, where “dose” is the form of exposure to radiation, and “response” is the measure of genetic aberration (in vivo setting, human exposures), or cell disability (in vitro setting, cell cultures of human lymphocytes).

- Focus on (ordered) polytomous categorical responses:
  - \( x_i \): dose levels (with \( x_1 < x_2 < \ldots < x_k \)).
  - \( n_i \): number of cells at dose level \( i \).
  - \( y_i = (y_{i1}, \ldots, y_{ir}) \): response vector (\( r \geq 2 \) classifications) at dose level \( i \).
  - Hence, now \( y_i | p_i \sim \text{Mult}(n_i, p_i) \), where \( p_i = (p_{i1}, \ldots, p_{ir}) \).
Bayesian nonparametric modeling for cytogenetic dosimetry

- Bayesian nonparametric modeling for polytomous response (Kottas, Branco and Gelfand, 2002).
- Consider simple case with $r = 3 \Rightarrow$ model for $p_{i1}$ and $p_{i2}$ is needed.
- Model $p_{i1} = F_1(x_i)$ and $p_{i1} + p_{i2} = F_2(x_i)$, and thus $F_1(\cdot) \leq F_2(\cdot)$.
- Bayesian nonparametric model requires prior on the space 
  $$ \{(F_1, F_2) : F_1(\cdot) \leq F_2(\cdot)\} $$
  of stochastically ordered pairs of c.d.f.s $(F_1, F_2)$.
- Constructive approach: $F_1(\cdot) = G_1(\cdot)G_2(\cdot)$, and $F_2(\cdot) = G_1(\cdot)$, with independent $\text{DP}(\alpha_\ell, G_{0\ell})$ priors for $G_\ell$, $\ell = 1, 2$.
- Induced prior for $q_\ell = (q_{\ell,1}, \ldots, q_{\ell,k})$, $\ell = 1, 2$, where $q_{\ell,i} = G_\ell(x_i)$. 

Athanasios Kottas and Abel Rodriguez

Applied Bayesian Nonparametric Mixture Modeling – Session 1
Combining with the likelihood, the posterior for \((q_1, q_2)\) is

\[
p(q_1, q_2 \mid \text{data}) \propto \prod_{i=1}^{k} \left\{ q_{1i}^{y_{i1} + y_{i2}} (1 - q_{1i})^{y_{i3}} q_{2i}^{y_{i1}} (1 - q_{2i})^{y_{i2}} \right\} 
\]

\[
\times q_{11}^{\gamma_1 - 1} (q_{12} - q_{11})^{\gamma_2 - 1} ... (q_{1k} - q_{1,k-1})^{\gamma_k - 1} (1 - q_{1k})^{\gamma_{k+1} - 1} 
\]

\[
\times q_{21}^{\delta_1 - 1} (q_{22} - q_{21})^{\delta_2 - 1} ... (q_{2k} - q_{2,k-1})^{\delta_k - 1} (1 - q_{2k})^{\delta_{k+1} - 1} 
\]

where

\[
\gamma_i = \alpha_1 (G_{01}(x_i) - G_{01}(x_{i-1})), \quad \delta_i = \alpha_2 (G_{02}(x_i) - G_{02}(x_{i-1})).
\]

Posterior for \(G_\ell(x_i)\) for \(\ell = 1, 2\) provide posteriors for \(F_1(x_i)\) and \(F_2(x_i)\), for all \(x_i\).
Bayesian nonparametric modeling for cytogenetic dosimetry

- Inference based on an MCMC algorithm.
  - For any unobserved dose level $x_0$, the posterior (predictive) distribution for $q_{\ell,0} = G_{\ell}(x_0)$ for $\ell = 1, 2$, is given by
    
    $$p(q_{\ell,0} \mid \text{data}) = \int p(q_{\ell,0} \mid q_{\ell})p(q_{\ell} \mid \text{data})dq_{\ell}$$

    where $p(q_{\ell,0} \mid q_{\ell})$ is a rescaled Beta distribution.
  - The inversion problem (inference for an unknown $x_0$ for specified response values $y_0 = (y_{01}, y_{02}, y_{03})$) can be handled by extending the MCMC algorithm to the augmented posterior that includes the additional parameter vector $(x_0, q_{10}, q_{20})$.

- For the data illustrations, we compare with a parametric logit model,
  
  $$\log \frac{p_{ij}}{p_{i3}} = \beta_{1j} + \beta_{2j}x_i, \quad i = 1, \ldots, k, \quad j = 1, 2$$

  (model fitting, prediction, and inversion are straightforward under this model).
Cytogenetic dosimetry: simulated data examples

Figure 1.4: Two data sets generated from the parametric model. Posterior inference for $F_1$ (upper panels) and $F_2$ (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. “o” denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.
Cytogenetic dosimetry: simulated data examples

Figure 1.5: Two data sets generated using non-standard (bimodal) shapes for $F_1$ and $F_2$. Posterior inference for $F_1$ (upper panels) and $F_2$ (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. “o” denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.
Semiparametric regression for categorical responses

- Application of DP-based modeling to semiparametric regression with categorical responses.

- Categorical responses $y_i$, $i = 1, ..., N$ (e.g., counts or proportions).

- Covariate vector $x_i$ for the $i$-th response, comprising either categorical predictors or quantitative predictors with a finite set of possible values.

- $K \leq N$ predictor profiles (cells), where each cell $k$ ($k = 1, ..., K$) is a combination of observed predictor values — $k(i)$ denotes the cell corresponding to the $i$-th response.

- Assume that all responses in a cell are exchangeable with distribution $F_k$, $k = 1, ..., K$. 
Semiparametric regression for categorical responses

- **Product of mixtures of Dirichlet processes prior** (Cifarelli and Regazzini, 1978) for the cell-specific random distributions $F_k, k = 1, ..., K$:
  -conditionally on hyperparameters $\alpha_k$ and $\theta_k$, the $F_k$ are assigned independent DP($\alpha_k, F_{0k}(\cdot; \theta_k)$) priors, where, in general, $\theta_k = (\theta_{1k}, ..., \theta_{Dk})$
  -the $F_k$ are related by modeling the $\alpha_k$ ($k = 1, ..., K$) and/or the $\theta_{dk}$ ($k = 1, ..., K; d = 1, ..., D$) as linear combinations of the predictors (through specified link functions $h_d, d = 0, 1, ..., D$)
  -$h_0(\alpha_k) = x_k^T \gamma, k = 1, ..., K$
  -$h_d(\theta_{dk}) = x_k^T \beta_d, k = 1, ..., K; d = 1, ..., D$
  -(parametric) priors for the vectors of regression coefficients $\gamma$ and $\beta_d$

- DP-based prior model that induces dependence in the finite collection of distributions $\{F_1, ..., F_K\}$, though a weaker type of dependence than dependent DP priors (MacEachern, 2000). [Dependent nonparametric prior models will be studied in Session 4.]
Semiparametric regression for categorical responses

- Semiparametric structure centered around a *parametric backbone* defined by the $F_{0k}(\cdot; \theta_k)$ – useful interpretation and connections with parametric regression models.

- Example: regression model for counts (Carota and Parmigiani, 2002)

$$y_i \mid \{F_1, \ldots, F_K\} \sim \prod_{i=1}^{N} F_{k(i)}(y_i)$$

$$F_k \mid \alpha_k, \theta_k \overset{ind.}{\sim} \text{DP}(\alpha_k, \text{Poisson}(\cdot; \theta_k)), \ k = 1, \ldots, K$$

$$\log(\alpha_k) = x^T_k \gamma$$

$$\log(\theta_k) = x^T_k \beta, \ k = 1, \ldots, K$$

with priors for $\beta$ and $\gamma$

- Related work for: change-point problems (Mira and Petrone, 1996); dose-response modeling for toxicology data (Dominici and Parmigiani, 2001); variable selection in survival analysis (Giudici, Mezzetti and Muliere, 2003).