



Check for updates

Bayesian Nonparametrics (BNP)

Chapter 7 by David B. Dunson

Zhengxiao Wei

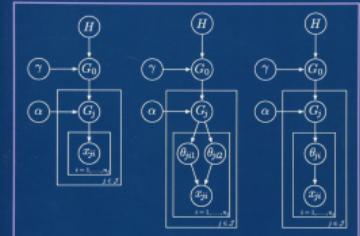
April 29, 2022 & November 7, 2023

Textbook

ISBN: 978-0-521-51346-3

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. *Bayesian Nonparametrics*, 28, 223-273.

Cambridge Series in Statistical and Probabilistic Mathematics



This section contains three diagrams illustrating hierarchical Bayesian models. Each diagram shows a variable H at the top, connected to a node G_0 . Below G_0 are nodes G_1 and G_2 , which are further connected to nodes θ_{ij} (with indices $i=1, \dots, n$ and $j=1, \dots, 2$). These θ_{ij} nodes are then connected to observed data nodes x_{ij} . Finally, there are prior nodes α connected to each of the G_i nodes.

Cambridge Series in Statistical and Probabilistic Mathematics

Editorial Board:

- Z. Ghahramani (Department of Engineering, University of Cambridge)
- F. P. Kelly (Mathematical Institute, Leiden University)
- F. W. Kader (Institutes of Pure Mathematics and Mathematical Statistics, University of Cambridge)
- B. D. Ripley (Department of Statistics, University of Oxford)
- S. Alpay (Department of Industrial and Systems Engineering, University of Southern California)
- M. Steele (Department of Mathematics, University of Chicago)

This series of high-quality open-access textbooks and monographs covers all areas of stochastic approximate mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

CAMBRIDGE UNIVERSITY PRESS
www.cambridge.org
 ISBN 978-0-521-51346-3
 9 780521 513463

Bayesian Nonparametrics
 Edited by Nils Lid Hjort, Chris Holmes,
 Peter Müller and Stephen G. Walker

Preliminary

Real Analysis: Lebesgue measure

Probability Theory: Borel σ -field, probability space (Ω, \mathcal{F}, P)

Stochastic Process: e.g., Chinese restaurant process, Indian buffet process

Experimental Design: linear mixed model

Generalized Linear Model: generalized estimating equations, generalized least squares

Bayesian Statistics: exchangeability, mixtures of g -priors, Markov chain Monte Carlo

Machine Learning: Gaussian mixture model

Nonparametric Bayes

Bayes' theorem: Posterior \propto Likelihood \times Prior

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

The diagram illustrates the components of Bayes' theorem. At the top, two boxes define 'LIKELIHOOD' (the probability of 'B' given 'A') and 'PRIOR' (the probability of 'A'). Arrows point from these definitions down to the term 'P(B|A).P(A)' in the numerator of the equation. Another arrow points from the term 'P(A)' in the numerator to the term 'P(B)' in the denominator. At the bottom, two boxes define 'POSTERIOR' (the probability of 'A' given 'B') and 'MARGINALIZATION' (the probability of 'B'). Arrows point from these definitions up to the term 'P(A)' in the numerator and the term 'P(B)' in the denominator respectively. The central equation is $P(A|B) = \frac{P(B|A).P(A)}{P(B)}$.

Image: Towards Data Science

Bayesian parametric model: A - parameters; B - data.

Bayesian nonparametric model: unbounded / growing / **infinite** number of parameters.

Repeated-Measures Design

- ▶ Longitudinal data: collected on the same subjects over time.
- ▶ Clustered data: measurements made on all units within a cluster in eq. 1 & 2.

Intraclass correlation $\rho = \frac{g}{g + \sigma^2}$

- proportion of total variance due to within-cluster variance;
- how strongly units within same cluster resemble each other.

$$\mathcal{M}_1: \quad y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad (1)$$

where y_{ij} is the j th observation within subject i ,
 μ_i is a subject-specific mean,
 ϵ_{ij} is an observation-specific residual,
 σ is the within-subject standard deviation,
 g is the between-subject variance on p. 7,
for $j = 1, \dots, n_i$ and $i = 1, \dots, n$.

Same Mixed Model but Different Designs

- **Clustered data** (Dobson & Barnett, 2018, p. 259; Dunson, 2010, p. 224)

e.g., the income of the j th household in the randomly selected council area i

e.g., the weight of the j th piglet in the randomly selected litter i

$$\mathcal{M}_1: \quad y_{ij} = \mu + b_i + \epsilon_{ij} \quad \leftarrow \textcircled{P}$$

versus $\mathcal{M}_0: \quad y_{ij} = \mu + \epsilon_{ij}$

- **One-way within-subject data** (Rouder et al., 2012 & 2017)

e.g., the response time of the i th subject for the j th Stroop task

$$\mathcal{M}_1: \quad y_{ij} = \mu + b_i + t_j + \epsilon_{ij}$$

versus $\mathcal{M}_0: \quad y_{ij} = \mu + b_i + \epsilon_{ij} \quad \leftarrow \textcircled{b}$

Bayesian Parametric Modeling

$$\mathcal{M}_1: \quad y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

$$\mu_i \sim \mathcal{N}(\mu, g) \quad (2)$$

or $\mu_i = \begin{matrix} \mu \\ \text{fixed} \end{matrix} + \begin{matrix} b_i \\ \text{random effects} \end{matrix}, \quad b_i \sim \mathcal{N}(0, g)$ (3)

BNP Motivation

The normal distribution has light tails and does **not** allow some subjects to be very different from other subjects or to have groups of subjects that cluster close together. Hence, outlying subjects tend to have their means **over-shrunk** towards the population mean, and the data from such subjects may be overly-influential in estimation of the overall mean μ .

The t -distribution has heavier tails, but it still has a very restrictive unimodal and symmetric shape.

Bayesian Nonparametric Modeling

$$\mathcal{M}_1: \quad y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

$$\mu_i \sim P \quad (4)$$

$$P \sim DP(P_0, \alpha) \quad (5)$$

Choose a **Dirichlet process (DP)** prior for P .

P corresponds to a distribution function, or more formally, a random probability measure.

- ▶ (5) allows P to be an unknown distribution.
- ▶ P_0 is a fixed baseline probability measure (the expected value of the process), corresponding to one's best guess for P *a priori*,
e.g., P_0 is a normal distribution with the normal-inverse gamma hyperpriors.
- ▶ α is the concentration parameter ($\alpha > 0$), characterizing prior precision and clustering (expressing confidence in P guess),
e.g., $\alpha \sim \text{Gamma}(1, 1)$, a gamma hyperprior.

Hierarchical Dirichlet Process

$$\mu_i \mid P_i \sim P_i$$

$$P_i \mid P_0, \alpha_i \sim DP(P_0, \alpha_i)$$

$$P_0 \mid H, \alpha \sim DP(H, \alpha)$$

mixed model \neq mixture model

Dirichlet Distribution

The Dirichlet distribution, $\text{Dir}(\boldsymbol{\alpha})$, is a multivariate generalization of the beta distribution for $K \geq 2$. The PDF is

$$f(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1}, \quad (6)$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$,

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K), \quad \alpha_0 = \sum_{i=1}^K \alpha_i \quad (7)$$

$$\sum_{i=1}^K x_i = 1 \text{ and } x_i \geq 0, \quad \alpha_i > 0 \text{ for } i = 1, \dots, K.$$

- ▶ If $\alpha_1 = \dots = \alpha_K = \alpha$, the Dirichlet distribution will be symmetric.
- ▶ When $\alpha = 1$, it is a uniform distribution over the open standard $(K - 1)$ -simplex.

Relation to the Gamma Distribution

For K independently distributed gamma distributions,

$$Y_i \stackrel{\text{ind.}}{\sim} \text{Gamma}(\alpha_i, \beta) \quad \text{for } i = 1, \dots, K,$$

it can be proved that the sum is also a gamma distribution,

$$Y = \sum_{i=1}^K Y_i \sim \text{Gamma}(\alpha_0, \beta). \quad (8)$$

Then, the K -dimensional Dirichlet distributed random vector is

$$\mathbf{X} = (X_1, \dots, X_K) = \left(\frac{Y_1}{Y}, \dots, \frac{Y_K}{Y} \right) \sim \text{Dir}(\boldsymbol{\alpha}). \quad (9)$$

$$\text{Cov}(X_i, X_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2 (\alpha_0 + 1)} \text{ for } i \neq j, \text{ not independent.}$$

The marginal distributions are $X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$. Exponential family.

R Scripts

```
options(digits=3); set.seed(277)
# LaplacesDemon::rdirichlet()
(pis <- gtools::rdirichlet(n=3, #number of random vectors to generate
                           alpha=rep(1,5)))
#      [,1]   [,2]   [,3]   [,4]   [,5]
# [1,] 0.35982 0.2053 0.0148 0.07106 0.3490
# [2,] 0.00204 0.2274 0.6151 0.00293 0.1525
# [3,] 0.61949 0.0075 0.2448 0.06353 0.0647

rowSums(pis) # 1 1 1
gtools::rdirichlet(3, rep(1000,5))
#      [,1]   [,2]   [,3]   [,4]   [,5]
# [1,] 0.208 0.200 0.207 0.193 0.193
# [2,] 0.190 0.210 0.199 0.199 0.203
# [3,] 0.194 0.207 0.195 0.206 0.198
```

Dir(α) is the canonical Bayesian distribution for the parameter estimates of a multinomial distribution.

Transitioning to the Dirichlet Process

For any finite partition S_1, \dots, S_K of the measurable set S $\left(\sum_{i=1}^K P_0(S_i) = 1 \right),$

$$P \sim DP(P_0, \alpha),$$



$$(P(S_1), \dots, P(S_K)) \sim \text{Dir}(\alpha P_0(S_1), \dots, \alpha P_0(S_K)).$$

Recall marginal betas.

Note that $\mathbb{E}[P(S_i)] = P_0(S_i)$ and $\text{Var}(P(S_i)) = \frac{P_0(S_i)(1 - P_0(S_i))}{\alpha + 1}$.
 α determines the variance of the random probability measure.

Dirichlet Process

The infinite-dimensional generalization of the Dirichlet distribution is the Dirichlet process $DP(P_0, \alpha)$.

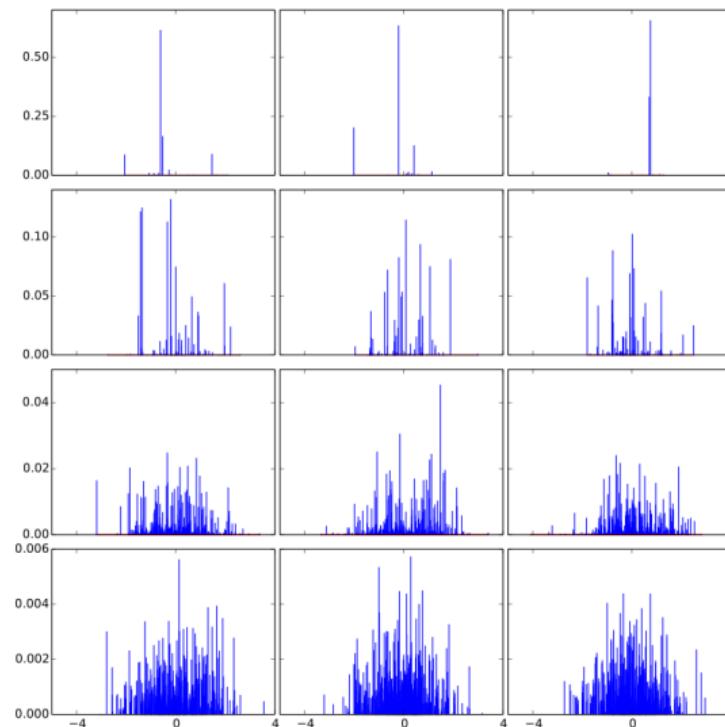
- ▶ Even if P_0 is continuous, the distributions drawn from the Dirichlet process are almost surely **discrete**.
- ▶ α specifies how strong this discretization is.
 - As $\alpha \rightarrow 0$, the realizations are all concentrated at a single value.
 - As $\alpha \rightarrow +\infty$, the realizations become continuous.

$DP(P_0, \alpha)$ is often used in Bayesian inference to describe the prior knowledge about the distribution of random variables — how likely it is that the random variables are distributed according to one or another particular distribution.

Example: $DP(\mathcal{N}(0, 1), \alpha)$

$\alpha = 1, 10, 100, 1,000$ from top to bottom.

Each row contains three repetitions of the same process.



Source: Wikipedia

Stick-Breaking Representation

Question: How do we draw a random distribution P from the Dirichlet process?

$P \sim DP(P_0, \alpha)$ is equivalent to **assigning weights to point masses at atoms** θ_h ,

$$P = p(\theta) = \sum_{h=1}^{\infty} \pi_h \cdot \delta_{\theta_h}(\theta), \quad \theta_h \stackrel{\text{i.i.d.}}{\sim} P_0, \quad (10)$$

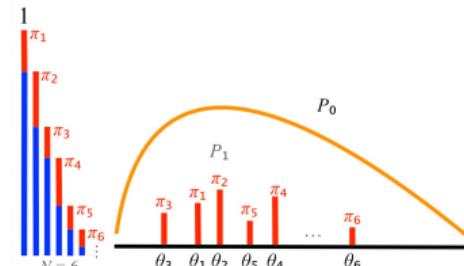
PMF

where $\pi_1 = V_1$ and $\pi_h = V_h \prod_{l < h} (1 - V_l)$ are probability weights that are formulated

from a stick-breaking process with $V_h \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$ for $h = 1, \dots, \infty$,

and $\delta_{\theta_h}(\cdot)$ is the indicator function $\delta_{\theta_h}(\theta_h) = 1$, otherwise 0.

In practice, we choose some N to truncate the series.



R Scripts (continued)

```
stick_breaking_process <- function(N, alpha) {  
  #' Input -  
  #' N:      number of weights (stick-breaks)  
  #' alpha:   concentration  
  #' Output - a vector of weights  
  V <- rbeta(N, 1, alpha)  
  V * c(1, cumprod(1-V))[1:N]  
}  
  
N <- 1000; alpha <- 100; set.seed(277)  
# Each random number from the base measure  $N(0,1)$  is  
# replicated a number of times corresponding to its weight.  
draws <- rep(rnorm(N),  
              round(stick_breaking_process(N, alpha) * 10000))  
  
hist(draws, prob=T, col="white", yaxt="n", ylab="", xlab=expression(theta),  
      main="Random Distribution From the Dirichlet Process")  
lines(density(draws), col="red", lwd=2)
```

"The rich get richer" (Shelley, 1818): Frequently sampled values in the past are more likely to be selected again.

Spike-and-Slab

$$\begin{aligned} M_i &\sim P \quad \text{for } i = 1, \dots, n \\ P &\sim DP(P_0, \alpha) \end{aligned}$$

Given observed data and a DP prior, we update our beliefs about the underlying distribution based on that data.

The [posterior](#) of P is

$$P | \mathbf{M} \sim DP\left(\frac{\text{PDF } P_0 + \sum_{i=1}^n \delta_{M_i}}{\alpha + n}, \alpha + n\right). \quad (11)$$

The base probability measure for the posterior DP can involve a mix of the continuous distribution (the “slab”) and the point masses (the “spikes”).

See also the Bayesian variable selection.

Thank you.

Contact: zhengxiao@uvic.ca