Foundations of Nonparametric Bayesian Methods
Part II: Models on the Simplex

Peter Orbanz

http://mlg.eng.cam.ac.uk/porbanz/npb-tutorial.html
Tutorial Overview

► Part I: Basics
► Part II: Models on the simplex
  1. Bayesian models and de Finetti
  2. DP construction
  3. Generalization examples: Pólya trees, neutral processes
  4. Consistency of NP Bayesian estimates
► Part III: Construction of new models
Introduction: NP Bayes Motivation

Machine learning
Roughly: Flexible models that adapt well to data.

Bayesian statistics
Search for universal priors.

Bayesian modeling

▶ Assumption: Data i.i.d. from unknown model. (Essential!)
▶ De Finetti: Justified for exchangeable data.
▶ Theorem of de Finetti (in part) prompted search for NP priors.

Exchangeable sequence
Probability of data $X_1, X_2, \ldots$ invariant under permutation of sampling order.
De Finetti’s Theorem (1)

Theorem (De Finetti, 1937; Hewitt & Savage, 1955)

On a Polish space \((D, \mathcal{B}(D))\), RVs \(X_1, X_2, \ldots\) are exchangeable iff there is a random probability measure \(F\) on \(D\) such that \(X_1, X_2, \ldots\) are conditionally i.i.d. with distribution \(F\).

Additionally:

1. \(X_i\) exch.ble \(\Rightarrow\) distribution \(\mathcal{G}\) of \(F\) uniquely determined.

2. For \(B \in \mathcal{B}(D)\):

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_B(X_i) = \text{a.s.} \ F(B)
\]

That is: Empirical distribution \(\xrightarrow{n \to \infty} F\) almost surely
De Finetti’s Theorem (2)

Interpretation
If $X_i$ exchangeable, then:

1. There is RV $F$ with values $f \in \mathcal{M}_1^+(D)$
2. Distribution: $\mathbb{G} = F(\mathbb{P})$
3. $X_i$ are generated as:

$$f \sim \mathbb{G} \quad \text{and} \quad X_i \sim_{\text{i.i.d}} f$$

Consequence
For $A^\infty := \prod_{i=1}^{\infty} A_i$:

$$\mu_{X^\infty}(A^\infty) = \int_{\mathcal{M}_1^+(D)} \left( \prod_{i=1}^{\infty} f(A_i) \right) d\mathbb{G}(f)$$
Bayes and de Finetti

NP Bayesian density estimation

1. Start with a prior assumption (eg a DP) for $\mathcal{G}$.
2. Update prior given data.
3. Hopefully: Posterior will concentrate at true $f$.
   (We are not trying to estimate $\mathcal{G}$!)

Requirements

1. Prior sufficiently spread over $\mathcal{M}_1^+(D)$
2. Updating prior on $F$ given draws from $F$ must be possible

Terminology: Conjugacy

- Algebraic: Prior in $\mathcal{G}$, likelihood in $\mathcal{F}$ $\Rightarrow$ posterior in $\mathcal{G}$
- Functional: Prior, posterior in same parametric class
  Measurable map (prior par, data) $\mapsto$ posterior par
Prior Domains

Note
Parametric priors perfectly fit into de Finetti framework.
→ Why not use a parametric prior on $F$?

Example (Draper, 1999)

1. Start with Gaussian model, prior on $\mu, \sigma$.
2. Data comes in: Bimodal.
3. Two choices:
   - Keep model → poor fit
   - Change to bimodal model → confounded (data used twice)

Prior domain requirement
Prior should support sufficiently diverse models to roughly fit arbitrary data.
Process Models: Divergent Aims

NP Bayesian statistics
- Prior models with large support in $\mathcal{M}_1^+(D)$
- DP: Concentrates on discrete distributions in $\mathcal{M}_1^+(D)$
- Principal motivation of DP generalization: Overcome discreteness

Machine Learning
- Principal application: Clustering and mixtures
- Discreteness *required*
Construction Approaches

Kolmogorov Constructions
Construct process directly from marginals.

Generalization of DP
Consider all processes with a particular property of DP:
- Partitioning: Tailfree processes and Pólya trees
- Random CDFs: Neutral-to-the-right (NTTR) and Levy processes
- Related: DP mixtures (smoothed DPs)

De Finetti-based
Derive class of possible $\mathbb{G}$ from knowledge about $X_i$.
Example: $X_i$ follow generalized Pólya urn scheme $\Rightarrow \mathbb{G}$ is a DP
Dirichlet Process Construction

Recall: Two stage Construction

1. Extend marginals to process on \((\Omega^E, \mathcal{B}^E)\) (Kolmogorov)
2. Restrict process to subset \(\tilde{\Omega} \subset \Omega^E\) (Doob)

Kolmogorov’s theorem: Dirichlet process

- “Axes” \(\Omega_0 := [0, 1]\)
- Index set \(E = \mathcal{H}(D) \) "measurable partitions"

\[
\mathcal{H}(D) := \{ B_1, \ldots, B_d \mid \text{partition of } D, d \in \mathbb{N}, B_i \in \mathcal{B}(D) \}
\]

- Marginals: Dirichlet distributions (for \(I \subset E\))

\[
d\mu_{\Theta^I|\alpha,y^I}(\theta^I) = p^I(\theta^I|\alpha, y^I)d\lambda^I(\theta^I)
\]

where: \(p^I = \text{Dirichlet density}\)
\(\lambda^I = \text{Lebesgue measure restricted to } \text{Sim}(\mathbb{R}, |I|)\)
DP: Restriction Step

Restrict to $\tilde{\Omega} = \mathcal{M}_1^+(D)$

- Check: $\tilde{\Omega}$ Polish (in weak* topology; true if $D$ Polish)
- Check: $\mu^E,^*(\tilde{\Omega}) = 1$ (true for Dirichlet marginals)
- Doob’s theorem: Exists $\tilde{\mu}^E$ on $\tilde{\Omega}$ with same marginals as $\mu^E$.

We are done:

$\tilde{\mu}^E$ is our Dirichlet process $\tilde{\mu}^E = \text{DP}(\alpha G_0)$

Historical Note

- Ferguson’s (1973) construction: Only extension step
- Ghosh & Ramamoorthi (2002) note that singletons are not in $\mathcal{B}^E$, give equivalent construction on $\mathcal{Q}$
Bayesian Model with DP

Unobserved: \( G \sim \text{DP}(\cdot | \alpha_0 G_0) \) (draw from prior)

Observed: \( \theta \sim G \) (sample draw)

Consider Marginals

For any \( H = (B_1, \ldots, B_m) \in \mathcal{H} \): Exactly one bin \( B_k \) contains \( \theta \)

Posterior update:

- Prior parameters: \( y^H_0 = (G_0(B_1), \ldots, G_0(B_m)) \) and \( \alpha_0 \)
- Posterior: \( y^H_1 \propto (G_0(B_1), \ldots, G_0(B_k) + \frac{1}{\alpha_0}, \ldots, G_0(B_m)) \)
  \( \alpha_1 = \alpha_0 + 1 \)

DP posterior

Posterior is \( \text{DP}(\cdot | \alpha_1 G_1) \) with \( G_1 \propto \alpha_0 G_0 + \delta_\theta \) and \( \alpha_1 = \alpha_0 + 1 \)
Sampling Under DP Prior

Model

\[ G \sim \text{DP}(\alpha_0 G_0) \]
\[ \theta_1, \ldots, \theta_n \sim G \]

Marginalize out \( G \)

\[ \theta_1 \sim G_0 \]

Sequential Sampling

With posterior formula:

\[ \theta_{i+1} \sim \sum_{j=1}^{i} \delta_{\theta_j} + \alpha G_0 \]
Discreteness

Key Observation
If $G \sim \text{DP}(\alpha_0 G_0 + \delta_\theta)$, then $G(\{\theta\}) > 0$ a.s.

Consequence
Draw $G$ from DP discrete almost surely, so

$$G = \sum_{i=1}^{\infty} c_i \delta_{\theta_i}$$

“Stick-breaking” (Sethuraman, 1984)
Generative model for $c_i$ and $\theta_i$:

$$\theta_i \sim_{\text{iid}} G_0$$
$$c_i := \left(\prod_{j=1}^{i-1} (1 - Z_j)\right) Z_i \quad \text{with} \quad Z_i \sim_{\text{iid}} \text{Beta}(1, \alpha_0)$$
Other Properties

Undominated Model

- Family $\text{DP}(\cdot | \alpha_0 G_0 + \delta_\theta)$ of posteriors not dominated
- In particular: No Bayes equation
- Implication: Conjugacy essential

Prior support: “KL property”

DP posits non-zero mass on every Kullback-Leibler $\epsilon$-neighborhood of every measure in $\mathcal{M}_1^+(D)$. 
Now: Generalizations

(Potentially) Smooth models

- Dirichlet process mixtures
- Pólya trees, tailfree processes
- Neutral-to-the-right (NTTR) and Lévy processes
DP mixtures

Original motivation
Smooth DP by convolution with parametric model.

Mixture model
$q$ conditional density, $\mu_Z$ “mixing distribution”

$$p(x) = \int_{\Omega_z} q(x|z)d\mu_Z(z)$$

Example: $q$ and $\mu_Z$ Gaussian, $z$ variance $\rightarrow p$ Student

Finite mixture

$$\mu_Z(B) = \sum_{i=1}^{n} c_i \delta_{z_i}(B) \quad \text{for } B \in \mathcal{B}(\Omega_x)$$

DP mixture
Draw $\mu_Z$ from DP.
Pólya Trees

Idea (on $\mathbb{R}$)

- Binary splitting of $\mathbb{R}$ into measurable sets.
- Subdivision as tree:
  - Nodes = Sets with Children = Subsets
  - Root = $\mathbb{R}$
  - One random probability per edge
  - $P(B) = \text{product over probs on path Root } \rightarrow B$

Model components

$m = \text{level in tree}$

$\pi_m = \text{measurable partition}$

$\pi_{m+1} \text{ refines } \pi_m$

$V_{m,B} = \text{Beta RVs}$

$\alpha_{m,B} = \text{Beta parameters}$

![Diagram of Pólya Tree](image-url)
Pólya Tree Properties

Include discrete and continuous models

- Particular expectation $G_0$: Choose $\pi_m$ such that values $G_0(B)$ match quantiles of $G_0$

- $\alpha_{m,B} \propto m^2 \rightarrow$ draws absolutely continuous

- $\alpha_{m,B} \propto \frac{1}{m} \rightarrow$ DP

Conjugacy

Sample observation:

$$G \sim \text{PT}(\pi, \alpha)$$
$$Y \sim G$$

Posterior update:

$$\alpha_{m,B} \rightarrow \alpha_{m,B} + 1 \quad \text{for } B \text{ with } Y \in B.$$
The Tailfree Property

Pólya trees are example of “tailfree” processes.

**Def: Tailfree process**
Consider process defined like PT, but with *arbitrary* RVs $V_{m,B}$.

**Tailfree:**
If RV sets $\{V_{m,B} | B \in \pi_m\}$ independent between tree levels.

**Meaning: Tailfree model**
Prior makes no assumption about shape of tails outside region covered by samples.

**Freedman (1968)**
Countable domains: Tailfree processes have consistent posteriors.
Random CDFs on $\mathbb{R}$

Observation: For $D = \mathbb{R}$

- DP: normalized gamma process.
- Draw CDF as process trajectory.
- CDF properties are local:
  1. initialize at 0
  2. monotonicity
  → expressible as increments

Possible choice: Lévy processes

- Easily chosen as monotonic ("subordinators")
- Can be simulated numerically
- Best-understood class of stochastic processes
- Independence of components may be desirable
Lévy Processes

Def: Lévy process
Stochastic process with stationary independent increments and right-continuous pathes with left-hand limits (rcll).

Intuition: rcll
- Only countable number of jumps
- Jump at $x$: Value at $x$ “belongs to right-hand branch”

Decomposition property
Decomposes into a Brownian motion (continuous) and a Poisson (jump) component ($\rightarrow$ simulation).

De Finetti property (Bühlmann, 1960)
Process $(X_t)$ with $X_s \xrightarrow{P} X_t$ whenever $s \rightarrow t$. Increments:
- exchangeable $\Leftrightarrow$ conditionally stationary and independent
Lévy Processes and NPB

Direct approach
To generate CDF:
1. Draw trajectory from increasing process
2. Normalize

Dead end (James et al. 2005)
If process normalized subordinator and conjugate, it is a Dirichlet process.

Neutrality (NTTR processes)
Process is *neutral to the right* if cumulative hazard function (= 1-CDF) has independent increments.

\[ F_{NTTR} \Leftrightarrow -\log(1 - F(\cdot)) \text{ Lévy} \]

NTTR process properties
Conjugacy (algebraic), contain DP, range of consistency results
Lévy Processes and Kolmogorov Extensions

Construction of Lévy process

- Choose $\mu^I$ to guarantee independent stationary increments
- Apply Kolmogorov process $\mu^E$ on $\Omega^E$
- Restrict to $\tilde{\Omega} = \text{set of rcll functions}$

Convolution Semigroup

Set of measures $\nu_s$ ($s \in \mathbb{R}_+$) with

$$\nu_{s+t} = \nu_s \ast \nu_t \quad \text{for } s, t \in \mathbb{R}_+$$

Lévy marginals from convolution semigroup

For $I = \{t\}$ set $\mu^I := \nu_t$ with $\{\nu_s\}_{s \in \mathbb{R}_+}$ convolution semigroup.
Consistency

Consistency means:
If \( \theta_0 \) is “true” solution, posterior (regarded as random measure) converges to \( \delta_{\theta_0} \) a.s. as \( n \to \infty \).

Two notions of consistency

▶ “Frequentist” consistency: “a.s.” under \( \mu_X(X|\Theta = \theta_0) \).
▶ Bayesian consistency: “a.s.” under evidence → only ensures consistency for \( \theta_0 \) in prior domain

Different results

▶ Bayesian def: Under very general conditions, any Bayesian model consistent (Doob, 1949)
▶ Frequentist def: Models like DP can be completely wrong.
Example Results

Countable domain $D$ (Freedman, 1968)
Freedman gives example of Bayes-consistent model which converges to arbitrary prescribed distribution for any true value of $\theta_0$.
Remedy: Tailfree prior.

DP inconsistency (Diaconis & Freedman, 1986)
DP estimate of mean (of unknown distribution) can be inconsistent.

Recent results

- Quantify model complexity (learning-theory style)
- Under conditions on complexity: Consistency and convergence rates
- Controllable behavior under model misspecification
- Shen & Wasserman (1998); Ghosal, van der Vaart et al (2002 and later)
Part III

- Constructing *parameterized* models from parameterized marginals
- Conjugacy
- Sufficient statistics of process models
- What are suitable families of marginals?