Bayesian Nonparametrics
Part I

Peter Orbanz
Today

1. Basic terminology
2. Clustering
3. Latent feature models

Tomorrow

5. Constructing nonparametric Bayesian models
6. Exchangeability
7. Asymptotics
Parameters

\[ P(X|\theta) = \text{Probability}[\text{data}|\text{pattern}] \]

Inference idea

\[ \text{data} = \text{underlying pattern} + \text{independent noise} \]
**TERMINOLOGY**

**Parametric model**
- Number of parameters fixed (or constantly bounded) w.r.t. sample size

**Nonparametric model**
- Number of parameters grows with sample size
- $\infty$-dimensional parameter space

**Example: Density estimation**

![Parametric](image1.png) ![Nonparametric](image2.png)
**Definition**

A nonparametric Bayesian model is a Bayesian model on an $\infty$-dimensional parameter space.

**Interpretation**

Parameter space $\mathcal{T} = \text{set of possible patterns}$, for example:

<table>
<thead>
<tr>
<th>Problem</th>
<th>$\mathcal{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density estimation</td>
<td>Probability distributions</td>
</tr>
<tr>
<td>Regression</td>
<td>Smooth functions</td>
</tr>
<tr>
<td>Clustering</td>
<td>Partitions</td>
</tr>
</tbody>
</table>

Solution to Bayesian problem = posterior distribution on patterns
Can we justify our assumptions?

Recall:

\[
\text{data} = \text{pattern} + \text{noise}
\]

In Bayes’ theorem:

\[
Q(d\theta|x_1, \ldots, x_n) = \frac{\prod_{j=1}^{n} p(x_j|\theta)}{p(x_1, \ldots, x_n)} Q(\theta)
\]

Definition

\(X_1, X_2, \ldots\) are exchangeable if \(P(X_1, X_2, \ldots)\) is invariant under any permutation \(\sigma\):

\[
P(X_1 = x_1, X_2 = x_2, \ldots) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots)
\]

In words:

Order of observations does not matter.
De Finetti’s Theorem

\[
P(X_1 = x_1, X_2 = x_2, \ldots) = \int_{\mathcal{M}(\mathcal{X})} \left( \prod_{j=1}^{\infty} \theta(X_j = x_j) \right) Q(d\theta)
\]

\[
\uparrow
\]

\[
X_1, X_2, \ldots \text{ exchangeable}
\]

where:

- \( \mathcal{M}(\mathcal{X}) \) is the set of probability measures on \( \mathcal{X} \)
- \( \theta \) are values of a random probability measure \( \Theta \) with distribution \( Q \)

Implications

- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: \( \theta \) is in general an \( \infty \)-dimensional quantity
CLUSTERING
Observations $X_1, X_2, \ldots$
- Each observation belongs to exactly one cluster
- Unknown pattern = partition of $\{1, \ldots, n\}$ or $\mathbb{N}$
Mixture models

\[ p(x|m) = \int_{\Omega_\theta} p(x|\theta)m(d\theta) \]

\( m \) is called the *mixing measure*

Two-stage sampling

Sample \( X \sim p( . | m) \) as:

1. \( \Theta \sim m \)
2. \( X \sim p( . | \theta) \)

Finite mixture model

\[ p(x|\theta, c) = \int_{\Omega_\theta} p(x|\theta)m(d\theta) \quad \text{with} \quad m( . ) = \sum_{k=1}^{K} c_k \delta_{\theta_k}( . ) \]
Random mixing measure

\[ M(\cdot) = \sum_{k=1}^{K} \lambda_k \delta_{\Theta_k} (\cdot) \]

Conjugate priors

A Bayesian model is *conjugate* if the posterior is an element of the same class of distributions as the prior ("closure under sampling").

\[
\begin{align*}
p(x|\theta) & \quad \text{conjugate prior} \\
\frac{1}{Z(\theta)} h(x) \exp(\langle S(x), \theta \rangle) & \quad \frac{1}{K(\lambda,y)} \exp(\langle \theta, y \rangle - \lambda \log Z(\theta))
\end{align*}
\]

Gaussian
multinomial
... 
Gaussian/inverse Wishart
Dirichlet
... 

Choice of priors in BMM

- Choose conjugate prior for each parameter
- In particular: Dirichlet prior on \((C_1, \ldots, C_k)\)
Dirichlet process

A Dirichlet process is a distribution on random probability measures of the form

\[ M(. \,) = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}(\cdot) \]

where

\[ \sum_{k=1}^{\infty} C_k = 1 \]

Constructive definition of DP \((\alpha, G_0)\)

\[ \Theta_k \sim_{iid} G_0 \]

\[ V_k \sim_{iid} \text{Beta}(1, \alpha) \]

Compute \(C_k\) as

\[ C_k := V_k \prod_{i=1}^{k-1} (1 - V_i) \]

"Stick-breaking construction"
DP Posterior

\[ \theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{n + \alpha} \sum_{j=1}^{n} \delta_{\theta_j}(\theta_{n+1}) + \frac{\alpha}{n + \alpha} G_0(\theta_{n+1}) \]

Mixture Posterior

\[ p(x_{n+1}|x_1, \ldots, x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n + \alpha} p(x_{n+1}|\theta^*_k) + \frac{\alpha}{n + \alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta \]

Conjugacy

- The posterior of DP \((\alpha, G_0)\) is DP \((\alpha + n, \frac{1}{n+\alpha} (\sum_k n_k \delta_{\theta^*_k} + \alpha G_0))\)
- Hence: The Dirichlet process is conjugate.
**Inference**

**Latent variables**

\[
p(x_{n+1}|x_1, \ldots, x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n + \alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n + \alpha} \int p(x_{n+1}|\theta)G_0(\theta)d\theta
\]

We do not actually observe the \( \Theta_j \) (they are latent). We observe \( X_j \).

**Assignment probabilities**

\[
\begin{pmatrix}
q_{10} & q_{11} & \cdots & q_{1K_n} \\
\vdots & \vdots & \ddots & \vdots \\
q_{n0} & q_{n1} & \cdots & q_{nK_n}
\end{pmatrix}
\]

Where:

\begin{itemize}
  \item \( q_{jk} \propto n_k p(x_j|\theta_k^*) \)
  \item \( q_{j0} \propto \alpha \int p(x_j|\theta)G_0(\theta)d\theta \)
\end{itemize}

**Gibbs Sampling**

Uses an assignment variable \( \phi_j \) for each observation \( X_j \).

\begin{itemize}
  \item Assignment step: Sample \( \phi_j \sim \text{Multinomial}(q_{j0}, \ldots, q_{jK_n}) \)
  \item Parameter sampling: \( \theta_k^* \sim G_0(\theta_k^*) \prod_{x_j \in \text{Cluster } k} p(x_j|\theta_k^*) \)
\end{itemize}
Dirichlet process

\[ K_n = \# \text{ clusters in sample of size } n \]

\[ \mathbb{E}[K_n] = O(\log(n)) \]

Modeling assumption

- Parametric clustering: \( K_\infty \) is finite (possibly unknown, but fixed).
- Nonparametric clustering: \( K_\infty \) is infinite

Rephrasing the question

- Estimate of \( K_n \) is controlled by distribution of the cluster sizes \( C_k \) in \( \sum_k C_k \delta_{\Theta_k} \).
- Ask instead: What should we assume about the distribution of \( C_k \)?
GENERALIZING THE DP

Pitman-Yor process

\[
p(x_{n+1} | x_1, \ldots, x_n) = \sum_{k=1}^{K_n} \frac{n_k - d}{n + \alpha} p(x_{n+1} | \theta_k^*) + \frac{\alpha + K_n \cdot d}{n + \alpha} \int p(x_{n+1} | \theta) G_0(\theta) d\theta
\]

Discount parameter \( d \in [0, 1] \).

Cluster sizes

![Cluster size graphs](image-url)
The distribution of cluster sizes is called a \textit{power law} if
\[
C_j \sim \gamma(\beta) \cdot j^{-\beta}
\]
for some $\beta \in [0, 1]$.

Examples of power laws

\begin{itemize}
\item Word frequencies
\item Popularity (number of friends) in social networks
\end{itemize}

Pitman-Yor language model
Discrete measures and partitions

Sampling from a discrete measure determines a *partition* of $\mathbb{N}$ into blocks $b_k$:

$$\Theta_n \sim_{iid} \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*}$$

and set $n \in b_k \iff \Theta_n = \theta_k^*$

As $n \to \infty$, the block proportions converge: $\frac{|b_k|}{n} \to c_k$

Induced random partition

The distribution of a random discrete measure $M = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}$ induces the distribution of a *random partition* $\Pi = (B_1, B_2, \ldots)$.

Exchangeable random partitions

- $\Pi$ is called *exchangeable* if its distribution depends only on the sizes of its blocks.
- All exchangeable random partitions, and only those, can be represented by a random discrete distribution as above (Kingman’s theorem).
Chinese Restaurant Process

The distribution of the random partition induced by the Dirichlet process is called the Chinese Restaurant Process.

"Customers and tables" analogy

Customers = observations (indices in $\mathbb{N}$)
Tables = clusters (blocks)

Historical remark

- Originally introduced by Dubins & Pitman as a distribution on infinite permutations
- A permutation of $n$ items defines a partition of $\{1, \ldots, n\}$ (regard cycles of permutation as blocks of partition)
- The induced distribution on partitions is the CRP we use in clustering
FAMILIES OF EXCHANGEABLE RANDOM PARTITIONS

- Poisson-Kingman
  - Normalized Completely Random Measure
    - Normalized Generalized Gamma
      - Normalized Inverse Gaussian
      - Normalized Stable
    - Dirichlet
  - Gibbs-type Measures
    - Pitman-Yor
      - Mixtures of Finite Dirichlets
      - Normalized Stable
      - Normalized Inverse Gaussian
## Random Discrete Measures

### Classification (due to Prünster)

<table>
<thead>
<tr>
<th>Class</th>
<th>Probability of new cluster</th>
<th>Prior Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>( P{ \Theta_{n+1} \in \text{new cluster}</td>
<td>\Theta^{(n)} } = f(n) )</td>
</tr>
<tr>
<td>II</td>
<td>( P{ \Theta_{n+1} \in \text{new cluster}</td>
<td>\Theta^{(n)} } = f(n, K_n) )</td>
</tr>
<tr>
<td>III</td>
<td>( P{ \Theta_{n+1} \in \text{new cluster}</td>
<td>\Theta^{(n)} } = f(n, K_n, n) )</td>
</tr>
</tbody>
</table>

### General partition priors

- Gibbs-type measures are completely classified [GP06b]
- Properties of some cases well-studied, e.g.:
  - Dirichlet process
  - Pitman-Yor process
  - Normalized inverse Gaussian process [LMP05b]
- In the future: We will have a range of models which express different prior assumptions on the distribution of cluster sizes.
## Summary: Clustering

### Nonparametric Bayesian clustering

- Infinite number of clusters, $K_n \leq n$ of which are observed.
- If partition exchangeable, it can be represented by a random discrete distribution.

### Inference

Latent variable algorithms, since assignments (≡ partition) not observed.

- Gibbs sampling
- Variational algorithms

### Prior assumption

- Distribution of cluster sizes.
- Implies prior assumption on number $K_n$ of clusters.
LATENT FEATURE MODELS
Indian Buffet Process

Latent feature models

- Grouping problem with overlapping clusters.
- Encode as binary matrix: Observation $n$ in cluster $k$ $\Leftrightarrow$ $X_{nk} = 1$
- Alternatively: Item $n$ possesses feature $k$ $\Leftrightarrow$ $X_{nk} = 1$

Indian buffet process (IBP)

1. Customer 1 tries Poisson($\alpha$) dishes.
2. Subsequent customer $n + 1$:
   - tries a previously tried dish $k$ with probability $\frac{n_k}{n + 1}$,
   - tries Poisson $\left( \frac{\alpha}{n + 1} \right)$ new dishes.

Properties

- An exchangeable distribution over finite sets (of dishes).
- Interpretation:
  Observation (= customer) $n$ in cluster (= dish) $k$ if customer “tries dish $k$”
Alternative description

1. Sample $w_1, \ldots, w_K \sim_{iid} \text{Beta}(1, \alpha/K)$
2. Sample $X_{1k}, \ldots, X_{nk} \sim_{iid} \text{Bernoulli}(w_k)$

We need some form of limit object for $\text{Beta}(1, \alpha/K)$ for $K \to \infty$.

Beta Process (BP)

Distribution on objects of the form

$$\theta = \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \quad \text{with } w_k \in [0, 1] .$$

- IBP matrix entries are sampled as $X_{nk} \sim_{iid} \text{Bernoulli}(w_k)$.
- Beta process is the de Finetti measure of the IBP, that is, $Q = \text{BP}$.
- $\theta$ is a random measure (but not normalized)
REFERENCES


