An Alternative Prior Process for Nonparametric Bayesian Clustering

Hanna Wallach (UMass Amherst)
Shane Jensen (UPenn)
Lee Dicker (Harvard)
Katherine Heller (Cambridge)
Nonparametric Bayesian Clustering

- Many uses: topic modeling, DNA motif clustering, etc.
- Underlying assumptions:
  - Set of RVs drawn from some unknown distribution
  - Unknown distribution is drawn from some prior
- Examples of nonparametric Bayesian priors:
  - Dirichlet process (DP): ubiquitous
  - Pitman-Yor process (PYP): generalization of the DP
Prior Assumptions

- DP & PYP both exhibit the "rich-get-richer" property
- Rich-get-richer implications:
  - Small # of large clusters
  - Large # of small clusters
- Rich-get-richer isn't always appropriate
- Want greater diversity of priors for clustering:
  - More choices for practitioners
The Uniform Process (UP)

- Introduced as an ad hoc prior for DNA motif clustering
  - Does not exhibit the rich-get-richer property
- We compare the UP to the DP & PYP in terms of:
  1. Asymptotic characteristics
  2. Characteristics for typical sample sizes
  3. Modeling trade-offs (e.g., exchangeability)
  4. Real-world clustering performance
Mixture Models for Clustering

• Mixture models:
  - Assume each $X_N$ was generated by one of $K$ mixture components characterized by parameters $\Phi = \{\phi_k\}_{k=1}^{K}$

• Clustering:
  - Goal: partition $\mathbf{X} = (X_1, \ldots, X_N)$ into clusters
  - Equivalent to identifying the set of parameters $\psi_n = \phi_k$ responsible for generating each observation $X_N$
  - Observations associated with $\phi_k$ form cluster $k$
Bayesian Mixture Models

- Bayesian mixture modeling:
  - Assume parameters $\Phi$ come from a prior $P(\Phi)$

- Nonparametric Bayesian mixture modeling
  - $P(\psi_N = \phi_k | \psi_1, \ldots, \psi_{N-1})$ is well-defined as $K \to \infty$
  - Model learns the "right" # of mixture components
  - Avoids costly model comparisons
Dirichlet Process

- 2 parameters:
  - Concentration parameter $\theta$
  - Base distribution $G_0$

- $P(\psi_{N+1} | \psi_1, \ldots, \psi_N, \theta, G_0) =$

\[
\begin{cases}
\frac{N_k}{N+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\
\frac{\theta}{N+\theta} & \psi_{N+1} \sim G_0
\end{cases}
\]

where $N_k = \sum_{n=1}^{N} I(\psi_n = \phi_k)$

[Aldous, '85; Sethuraman, '94; Ishwaran & James, '01; etc.]
Pitman-Yor Process

- 3 parameters:
  - Concentration parameter $\theta$
  - Discount parameter $\alpha$
  - Base distribution $G_0$

- $P(\psi_{N+1} | \psi_1, \ldots, \psi_N, \theta, \alpha, G_0) =$
  \[
  \begin{cases}
  \frac{N_k - \alpha}{N + \theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\
  \frac{1}{\theta + K\alpha} & \psi_{N+1} \sim G_0
  \end{cases}
  \]
Uniform Process

- 2 parameters:
  - Concentration parameter $\theta$
  - Base distribution $G_0$

- $P(\psi_{N+1} | \psi_1, \ldots, \psi_N, \theta, G_0) =$
  \[
  \begin{cases}
  \frac{1}{K+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\
  \frac{\theta}{K+\theta} & \psi_{N+1} \sim G_0
  \end{cases}
  \]

- No "rich-get-richer" property
DP Asymptotics ($N \to \infty$)

- Expected number of unique clusters in a partition:
  \[
  \mathbb{E}(K_N | \text{DP}) = \sum_{n=1}^{N} \frac{\theta}{n^{1+\theta}} \approx \theta \log N
  \]

- Expected number of clusters of size $M$:
  \[
  \lim_{N \to \infty} \mathbb{E}(H_{M,N} | \text{DP}) = \frac{\theta}{M}
  \]

$\Rightarrow$ Small # large clusters, large # small clusters

[Arratia et al., '03]
PYP Asymptotics ($N \to \infty$)

- Expected number of unique clusters in a partition:

  \[ \mathbb{E}(K_N | \text{PY}) \approx \frac{\Gamma(1+\theta)}{\alpha \Gamma(\alpha+\theta)} N^\alpha \]

- Expected number of clusters of size $M$:

  \[ \mathbb{E}(H_{M,N} | \text{PY}) \approx \frac{\Gamma(1+\theta) \prod_{m=1}^{M-1} (m-\alpha)}{\Gamma(\alpha+\theta) M!} N^\alpha \]

⇒ Small # large clusters, large # small clusters

[Ref: Pitman, '02]
UP Asymptotics ($N \to \infty$)

- Expected number of unique clusters in a partition:
  \[ \mathbb{E}(K_N | \text{UP}) \approx \sqrt{2\theta} \cdot N^{\frac{1}{2}} \]

- Expected number of clusters of size $M$:
  \[ \mathbb{E}(H_{M,N} | \text{UP}) \approx \theta \]

$\Rightarrow$ Uniform distribution of cluster sizes
Simulation: Number of Clusters
Simulation: Cluster Sizes

Dirichlet Process: N=1000

Pitman–Yor ($\alpha = 0.5$) : N=1000

Uniform: N=1000

Dirichlet Process: N=10000

Pitman–Yor ($\alpha = 0.5$) : N=10000

Uniform: N=10000

Dirichlet Process: N=100000

Pitman–Yor ($\alpha = 0.5$) : N=100000

Uniform: N=100000
Exchangeability

- Modeling tradeoffs: exchangeability vs. rich-get-richer
- The UP is not exchangeable over cluster assignments:
  - \( P(\text{cluster assignments}) \) is not invariant to permutations
- Previous work has not addressed this:
  - We present a new Gibbs sampling algorithm that is correct for a fixed ordering of cluster assignments
  - We demonstrate that \( P(\text{cluster assignments}) \) is highly robust to permutations of the cluster assignments
Gibbs Sampler

• Let $c_n$ be the cluster assignment for $X_n$:
  - $c_n = k$ implies $\psi_n = \phi_k$

• Given an ordering of observations:

$$P(c_n | c_{\setminus n}, X, \theta, \text{ordering } 1, \ldots, N) \propto P(X_n | c_n, X_{\setminus n}, c_{\setminus n}) P(c_n | c_1, \ldots, c_{n-1}, \theta) \prod_{m=n+1}^{N} P(c_m | c_1, \ldots, c_{m-1}, \theta)$$
Robustness to Orderings

The diagram illustrates the standard deviation of the log posterior probability $\log P(c)$ for different concentration parameters. The concentration parameters are 0.25, 0.5, 1, 2, 5, and 10. The bars represent the standard deviation for both between-ordering and between-partition scenarios.

- The height of the bars indicates higher standard deviations for lower concentration parameters, especially for between-ordering SD.
- The pattern suggests that the robustness to orderings improves with higher concentration parameters.

This analysis supports the use of higher concentration parameters in Bayesian clustering for improved robustness to different orderings of the data.
Document Clustering

- No reason to expect rich-get-richer cluster usage
- Clustering model (generative process):

\[ c_d \mid c_{<d} \sim \begin{cases} \frac{1}{d^{-1+\theta}} & c_d = k \in 1, \ldots, K \\ \frac{\theta}{d-1+\theta} & c_d = k_{\text{new}} \end{cases} \]

\[ n_k \sim G_0 \]

\[ \phi_d \sim \text{Dir}(\phi_d \mid n_{cd}, \beta) \]

\[ w_d \sim \text{Mult}(\phi_d) \]
Experiments

- 1200 carbon nanotechnology patent abstracts:
  - 1000 training abstracts, 200 test abstracts
  - Single, fixed ordering

- Compare predictive performance with DP and UP priors:
  - 5 Gibbs sampling runs
  - 8 concentration parameter values
  - Compute (approximate) probability of test documents
Results

Log Probability (Test | Training)

- Log Probability
- Concentration Parameter

Number of Clusters

- Uniform Process
- Dirichlet Process

Concentration Parameter

An Alternative Prior Process for Bayesian Clustering
Summary

● DP & PYP both lead to a “rich-get-richer” property
  – Not always appropriate/desirable
● We compared the UP to the DP & PYP in terms of:
  1. Asymptotic characteristics
  2. Characteristics for typical sample sizes
  3. Modeling trade-offs (e.g., exchangeability)
  4. Real-world clustering performance