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 $\boldsymbol{X} = (X_1, \ldots, X_N)$ into clusters
  - Equivalent to identifying the set of parameters $\psi_n = \phi_k$ responsible for generating each observation X_N
  - Observations associated with ϕ_k form cluster k

Bayesian Mixture Models

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

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

Dirichlet Process

[Aldous, '85; Sethuraman, '94; Ishwaran & James, '01; etc.]

- 2 parameters:
  - Concentration parameter θ
  - Base distribution G_0
- $P(\psi_{N+1} \mid \psi_1, ..., \psi_N, \theta, G_0) =$

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

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

Pitman-Yor Process

[Pitman & Yor, '97]

- 3 parameters:
  - Concentration parameter θ
  - Discount parameter α
  - Base distribution G_0
- $P(\psi_{N+1} \mid \psi_1, \ldots, \psi_N, \theta, \alpha, G_0) =$

$$\begin{cases} \frac{N_k - \alpha}{N + \theta} & \psi_{N+1} = \phi_k \in \{\phi_1, \ldots, \phi_K\} \\ \frac{\theta + K\alpha}{N + \theta} & \psi_{N+1} \sim G_0 \end{cases}$$

Uniform Process

[Qin et al., '03]

- 2 parameters:
  - Concentration parameter θ
  - Base distribution G_0
- $P(\psi_{N+1} \,|\, \psi_1, ..., \psi_N, \theta, G_0) =$

$$\begin{cases} \frac{1}{K+\theta} & \psi_{N+1} = \phi_k \in \{\phi_1, ..., \phi_K\} \\ \frac{\theta}{K+\theta} & \psi_{N+1} \sim G_0 \end{cases}$$

- No “rich-get-richer” property

DP Asymptotics ($N \to \infty$)

[Arratia et al., '03]

- Expected number of unique clusters in a partition:

$$\mathbb{E}(K_N \,|\, \mathrm{DP}) = \sum_{n=1}^{N} \frac{\theta}{n-1+\theta} \simeq \theta \log N$$

- Expected number of clusters of size M:

$$\lim_{N\to\infty} \mathbb{E}(H_{M,N} \,|\, \mathrm{DP}) = \frac{\theta}{M}$$

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

PYP Asymptotics ($N \to \infty$)

[Pitman, '02]

- Expected number of unique clusters in a partition:

$$\mathbb{E}(K_N \,|\, \mathrm{PY}) \approx \frac{\Gamma(1+\theta)}{\alpha\Gamma(\alpha+\theta)} N^{\alpha}$$

- Expected number of clusters of size M:

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

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

UP Asymptotics ($N \to \infty$)

- Expected number of unique clusters in a partition:

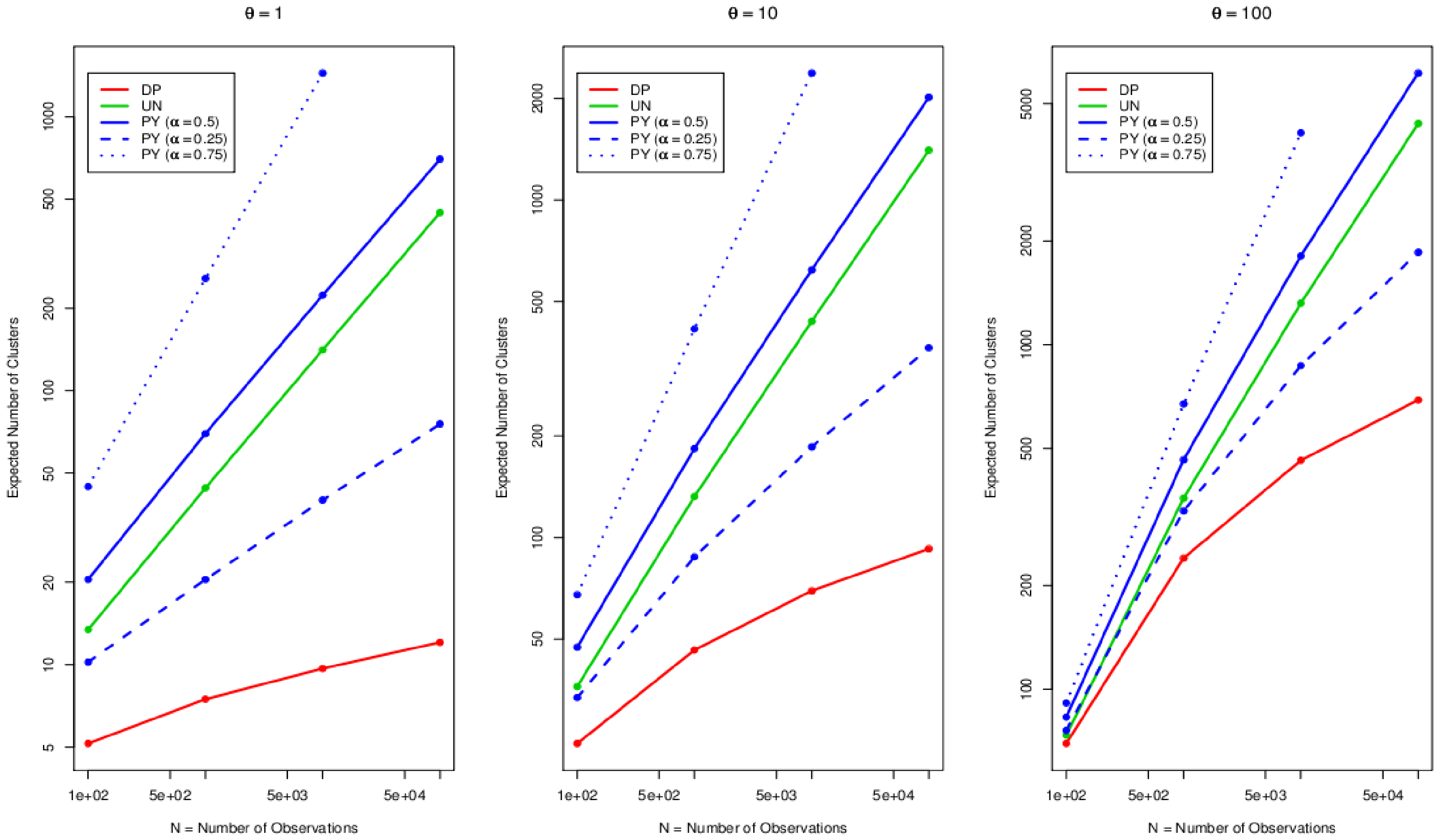
$$\mathbb{E}(K_N \,|\, \mathrm{UP}) \approx \sqrt{2\theta} \cdot N^{\frac{1}{2}}$$

- Expected number of clusters of size M:

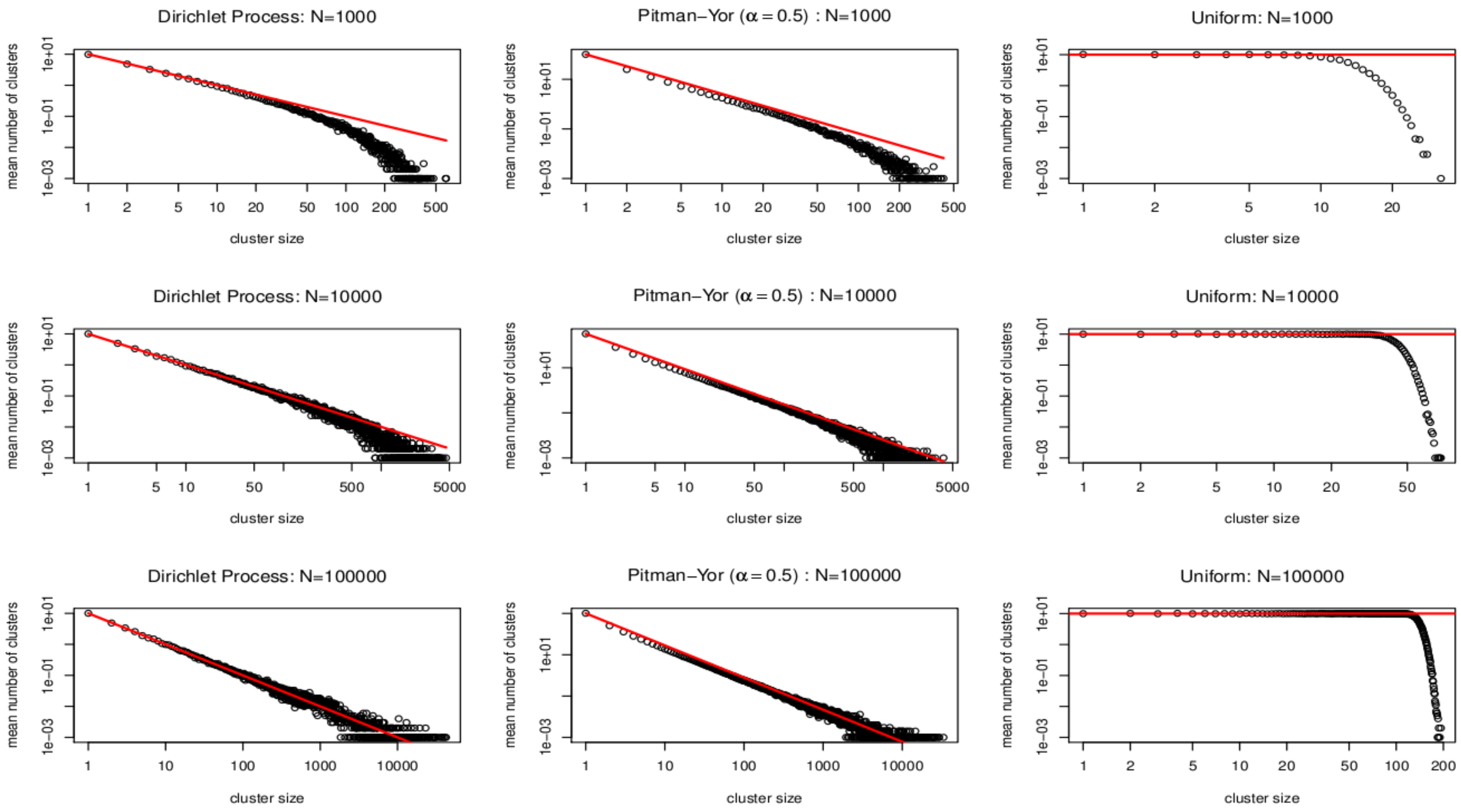
$$\mathbb{E}(H_{M,N} \,|\, \mathrm{UP}) \approx \theta$$

$\Rightarrow$ Uniform distribution of cluster sizes

Simulation: Number of Clusters



Simulation: Cluster Sizes



Exchangeability

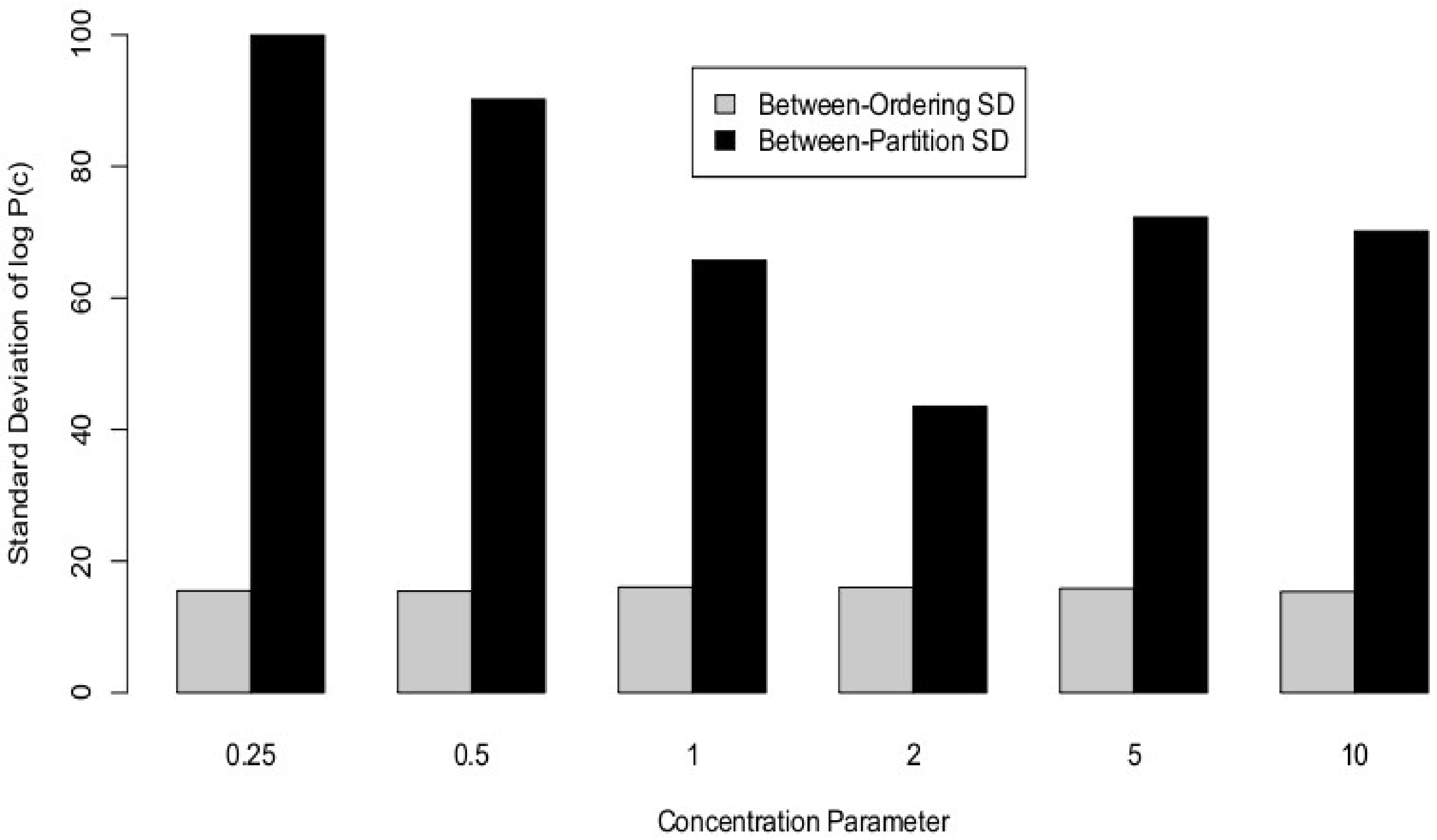
- Modeling tradeoffs: exchangeability vs. rich-get-richer
- The UP is not exchangeable over cluster assignments:
  - P(cluster assignments) is not invariant to permutations
- Previous work has not addressed this:
  - We present a new Gibbs sampling algorithm that is correct for a fixed ordering of cluster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

$$
\begin{aligned}
&P(c_n \mid \boldsymbol{c}_{\setminus n}, \boldsymbol{X}, \theta, \text{ ordering } 1, \ldots, N) \propto \\
&\qquad P(X_n \mid c_n, \boldsymbol{X}_{\setminus n}, \boldsymbol{c}_{\setminus n})\, P(c_n \mid c_1, \ldots, c_{n-1}, \theta) \\
&\qquad \prod_{m=n+1}^{N} P(c_m \mid c_1, \ldots, c_{m-1}, \theta)
\end{aligned}
$$

Robustness to Orderings



Document Clustering

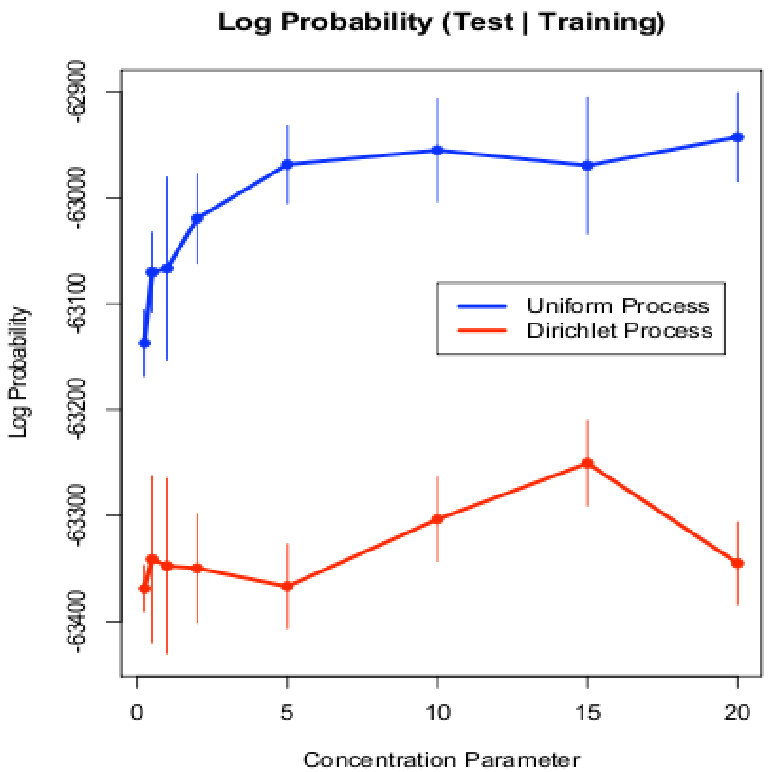
- No reason to expect rich-get-richer cluster usage
- Clustering model (generative process):

$$
\begin{aligned}
c_d \mid c_{<d} &\sim \begin{cases} \frac{1}{d-1+\theta} & c_d = k \in 1, \ldots, K \\ \frac{\theta}{d-1+\theta} & c_d = k_{\text{new}} \end{cases} \\
\boldsymbol{n}_k &\sim G_0 \\
\boldsymbol{\phi}_d &\sim \mathrm{Dir}(\boldsymbol{\phi}_d \mid \boldsymbol{n}_{c_d}, \beta) \\
\boldsymbol{w}_d &\sim \mathrm{Mult}(\boldsymbol{\phi}_d)
\end{aligned}
$$

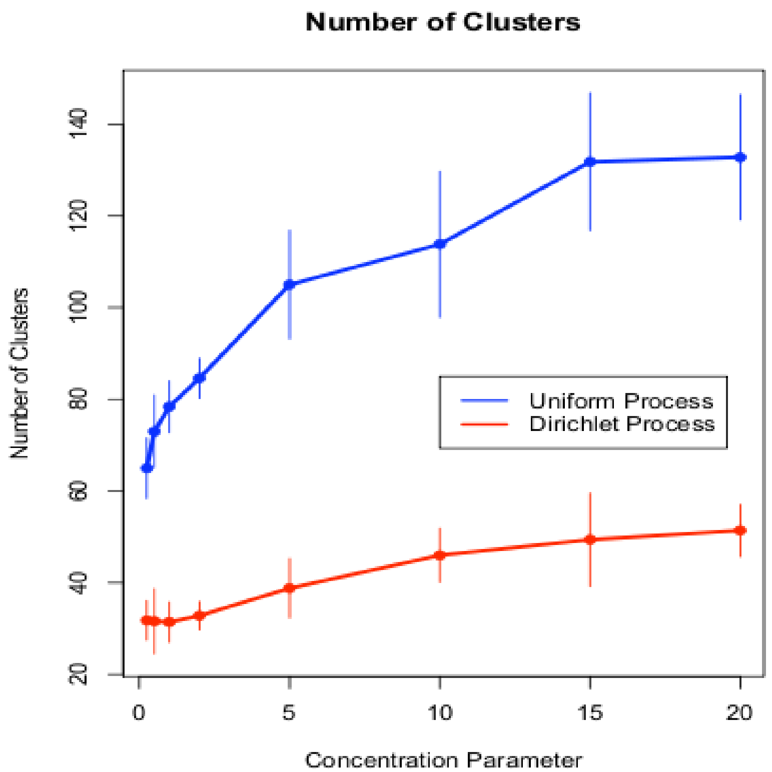
Experiments

- 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
-62900
-63000
-63100
-63200
-63300
-63400
0
5
10
15
20
Concentration Parameter
Uniform Process
Dirichlet Process



Number of Clusters
Number of Clusters
140
120
100
80
60
40
20
0
5
10
15
20
Concentration Parameter
Uniform Process
Dirichlet Process

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