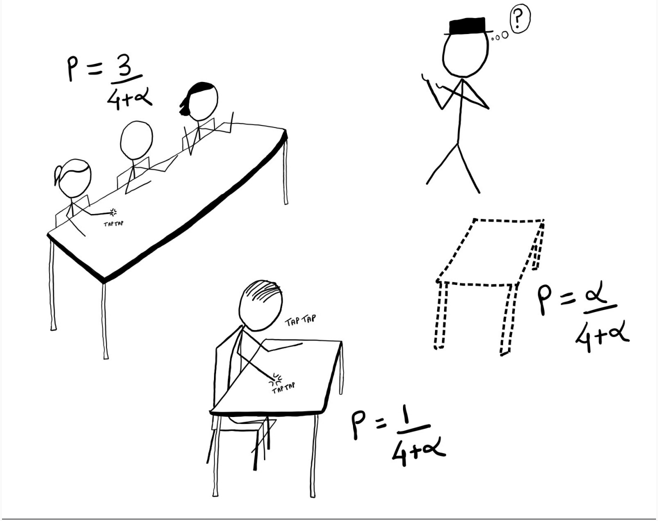
Efficient Online Inference for Nonparametric Mixture Models UAI 2021

Rylan Schaeffer, Blake Bordelon, Mikail Khona, Weiwei Pan, Ila Rani Fiete
Harvard SEAS, MIT BCS, MIT Physics

Motivation

- Clustering (mixture modeling) is a ubiquitous problem
- The Chinese Restaurant Process is a Bayesian Nonparametric model that allows the number of clusters to grow as more data are observed
- Common inference algorithms are formulated for the offline setting and scale poorly with the number of observations

Chinese Restaurant Process



Chinese Restaurant Process (CRP)

- $CRP(\alpha)$ is a single-parameter stochastic process that defines a distribution over **partitions of a set**
- CRP defines a conditional for t-th customer z_t given previous customers $z_{<t}$ and number of nonempty tables K_{t-1}:

$$P(z_t = k | z_{<t}, \alpha) = \begin{cases} \frac{\sum_{t'<t} \delta(z_{t'}=k)}{\alpha+t-1} & \text{if } 1 \leq k \leq K_{t-1} \\ \frac{\alpha}{\alpha+t-1} & \text{if } k = K_{t-1}+1 \\ 0 & \text{otherwise} \end{cases}$$

Research Objective

- Generative process: stream of discrete latent variables $z_{1:T}$ and observations $o_{1:T}$

$$z_{1:T} \sim CRP(\alpha)$$
$$o_t | z_t \sim p(o|z)$$

Research Objective

- Generative process: stream of discrete latent variables $z_{1:T}$ and observations $o_{1:T}$

$$z_{1:T} \sim CRP(\alpha)$$
$$o_t | z_t \sim p(o|z)$$

- Goal: infer (filter) $p(z_t | o_{\leq t})$, subject to two constraints:

Research Objective

- Generative process: stream of discrete latent variables $z_{1:T}$ and observations $o_{1:T}$

$$z_{1:T} \sim CRP(\alpha)$$
$$o_t | z_t \sim p(o|z)$$

- Goal: infer (filter) $p(z_t | o_{\leq t})$, subject to two constraints:
    1. Inference must be performed online, meaning the filter cannot make use of the (possibly) infinite past nor can the filter be used to revise the past.

Research Objective

- Generative process: stream of discrete latent variables $z_{1:T}$ and observations $o_{1:T}$

$$z_{1:T} \sim CRP(\alpha)$$
$$o_t | z_t \sim p(o|z)$$

- Goal: infer (filter) $p(z_t | o_{\leq t})$, subject to two constraints:
    1. Inference must be performed online, meaning the filter cannot make use of the (possibly) infinite past nor can the filter be used to revise the past.
    2. Inference must be efficient in the large t (sample) limit

Recursion for Online Filtering

- Streaming inference is difficult because CRP's conditional distribution $p(z_t|z_{<t}, \alpha)$ is dependent on the entire history

Recursion for Online Filtering

- Streaming inference is difficult because CRP's conditional distribution $p(z_t|z_{<t}, \alpha)$ is dependent on the entire history
- Approach: replace conditional with a recursively computed marginal distribution $p(z_t|\alpha)$

Recursion for Online Filtering

- Streaming inference is difficult because CRP's conditional distribution $p(z_t|z_{<t}, \alpha)$ is dependent on the entire history
- Approach: replace conditional with a recursively computed marginal distribution $p(z_t|\alpha)$
- Bayes' Theorem

$$\underbrace{p(z_t = k|o_{\leq t})}_{\text{Latent Posterior}} = \frac{p(o_t|z_t = k)}{p(o_t|o_{<t})} \underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}}.$$

Recursion for Online Filtering

- Streaming inference is difficult because CRP's conditional distribution $p(z_t|z_{<t}, \alpha)$ is dependent on the entire history
- Approach: replace conditional with a recursively computed marginal distribution $p(z_t|\alpha)$
- Bayes' Theorem

$$\underbrace{p(z_t = k|o_{\leq t})}_{\text{Latent Posterior}} = \frac{p(o_t|z_t = k)}{p(o_t|o_{<t})} \underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}}.$$

- Latent prior is the expected conditional distribution, averaged over all possible paths

$$\underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}} = \mathbb{E}_{p(z_{<t}, K_{t-1}|o_{<t})}\Big[p(z_t = k|z_{<t}, K_{t-1}, o_{<t})\Big]$$

Recursion for Online Filtering

- Making one approximation, we obtain the R-CRP recursion:

$$\underbrace{p(z_t = k | o_{\leq t})}_{\text{Latent Posterior}} \approx \frac{p(o_t | z_t = k)}{p(o_t | o_{<t})} \Bigg[\frac{1}{\alpha + t - 1} \sum_{t' < t} \underbrace{p(z_{t'} = k | o_{\leq t'})}_{\text{Previous Posteriors}} + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1 | o_{<t}) \Bigg]$$

Recursion for Online Filtering

- Making one approximation, we obtain the R-CRP recursion:

$$\underbrace{p(z_t = k | o_{\leq t})}_{\text{Latent Posterior}} \approx \frac{p(o_t | z_t = k)}{p(o_t | o_{<t})} \Bigg[\frac{1}{\alpha + t - 1} \sum_{t' < t} \underbrace{p(z_{t'} = k | o_{\leq t'})}_{\text{Previous Posteriors}} + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1 | o_{<t}) \Bigg]$$

- Intuition:

Recursion for Online Filtering

- Making one approximation, we obtain the R-CRP recursion:

$$\underbrace{p(z_t = k | o_{\leq t})}_{\text{Latent Posterior}} \approx \frac{p(o_t | z_t = k)}{p(o_t | o_{<t})} \Bigg[\frac{1}{\alpha + t - 1} \sum_{t' < t} \underbrace{p(z_{t'} = k | o_{\leq t'})}_{\text{Previous Posteriors}} + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1 | o_{<t}) \Bigg]$$

- Intuition:
  - First term is running sum of preceding latents' posteriors' masses, which means commonly used clusters will probably generate the next observation

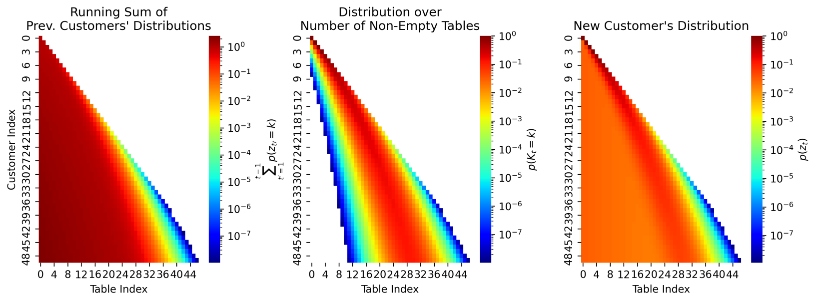
Recursion for Online Filtering

- Making one approximation, we obtain the R-CRP recursion:

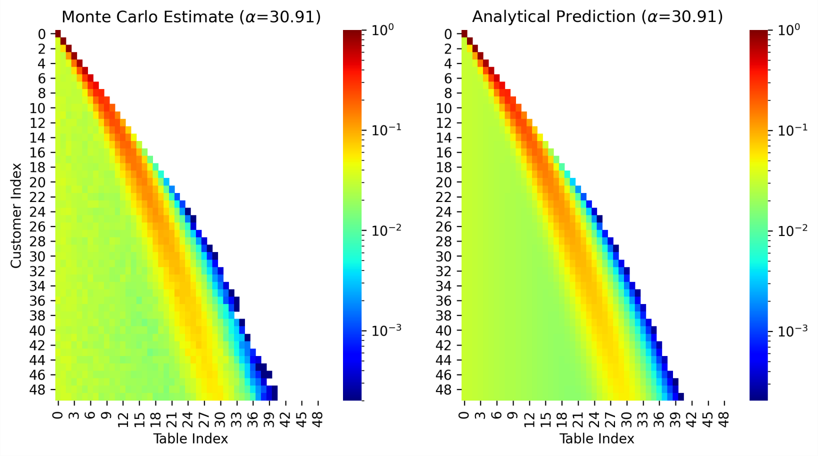
$$\underbrace{p(z_t = k | o_{\leq t})}_{\text{Latent Posterior}} \approx \frac{p(o_t | z_t = k)}{p(o_t | o_{<t})} \Bigg[\frac{1}{\alpha + t - 1} \sum_{t' < t} \underbrace{p(z_{t'} = k | o_{\leq t'})}_{\text{Previous Posteriors}} + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1 | o_{<t}) \Bigg]$$

- Intuition:
  - First term is running sum of preceding latents' posteriors' masses, which means commonly used clusters will probably generate the next observation
  - Second term is the number of clusters, which grows over time, incentivizing creation of new clusters

Recursion for Online Filtering

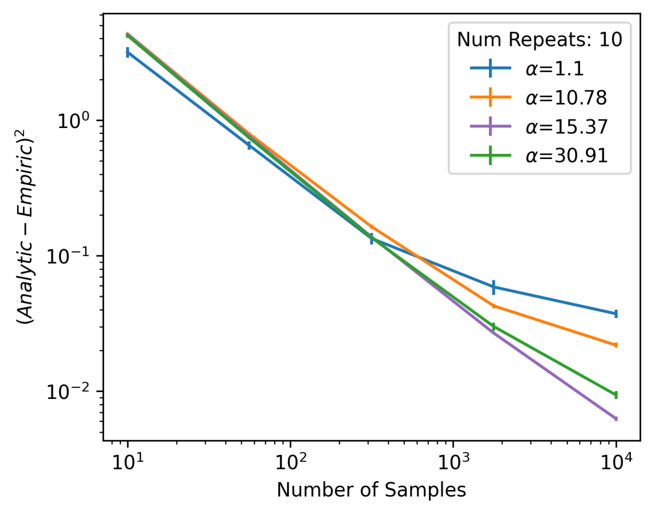


R-CRP is exact for CRP prior:



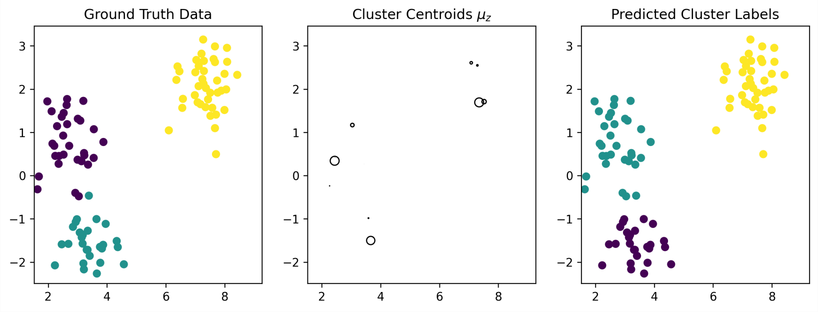
Experiments: CRP Prior

R-CRP is exact for CRP prior:



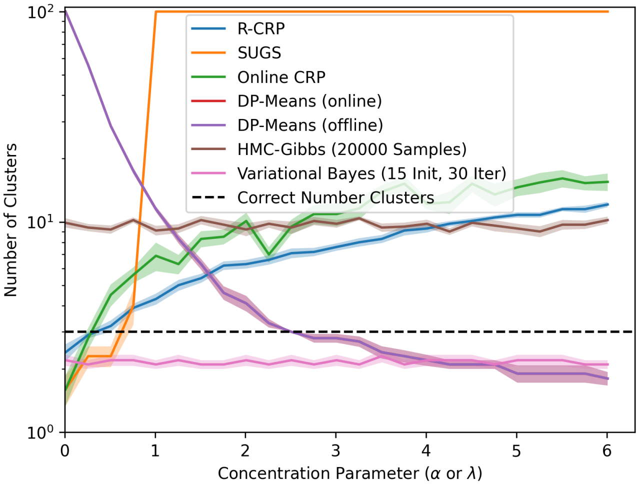
Experiment: Mixtures of Gaussians

R-CRP finds highly plausible centroids and cluster assignments



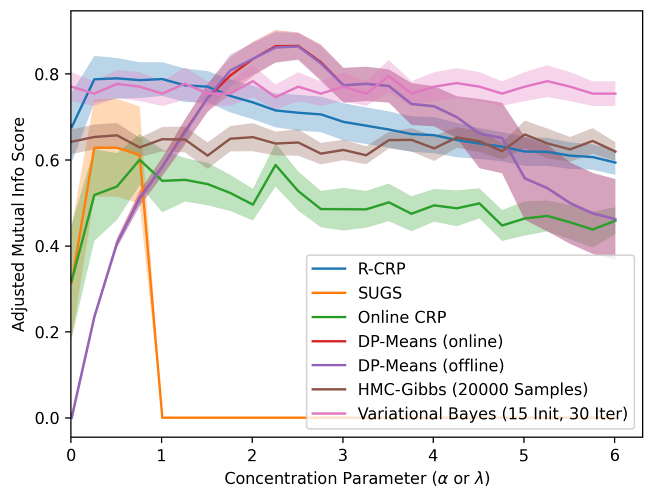
Experiment: Mixtures of Gaussians

R-CRP learns (close to) the correct number of clusters over wide range of concentration parameters



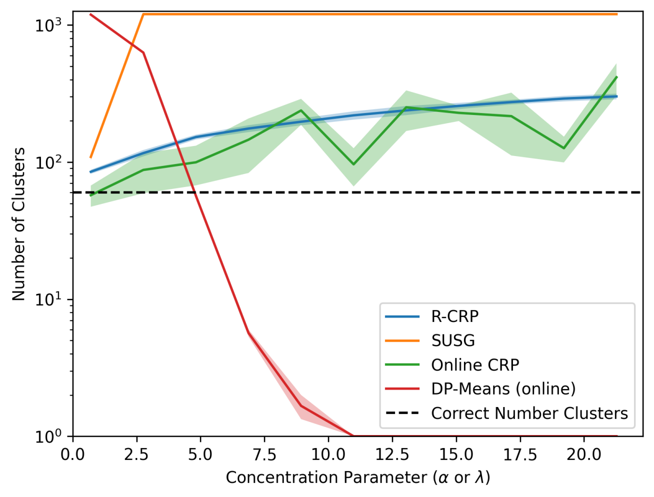
Experiment: Mixtures of Gaussians

R-CRP has higher adjusted mutual information with true cluster labels than most baselines over range of concentration patameters



Experiment: Handwritten Characters (Omniglot)

R-CRP learns (close to) the correct number of clusters over wide range of concentration parameters



Experiment: Handwritten Characters (Omniglot)

R-CRP has higher adjusted mutual information with true cluster labels than online baselines over range of concentration patameters