Powered Dirichlet Process

Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering

G. Poux-Médard, J. Velcin, S. Loudcher

Université de Lyon, France
Lyon 2, ERIC UR 3083

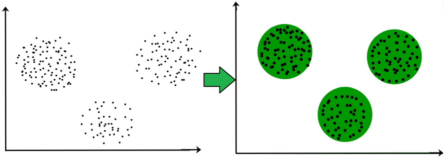
19th September 2023

Bayesian clustering

- Bayesian clustering approaches received a broad attention over the last decades
  - Medicine, natural language processing, genetics, recommender systems, sociology, etc.
- General idea (Bayes theorem):

$$\underbrace{P(\text{cluster}|\text{data})}_{\text{Posterior probability}} \propto \underbrace{P(\text{data}|\text{cluster})}_{\text{Likelihood}} \times \underbrace{P(\text{cluster})}_{\text{Prior probability}}$$

- In most cases, the prior is a Dirichlet distribution
  - Natural: yield an array whose entries sum to 1
  - Convenient: Can be expressed as a process



Dirichlet process

- Dirichlet process prior probability for the ith observation:

$$DP(C_i = c|\mathcal{H}, \alpha) = \begin{cases} \frac{N_c}{\alpha+N} & \text{if c = 1, ..., K} \\ \frac{\alpha}{\alpha+N} & \text{if c = K+1} \end{cases}$$

- Useful to model sequential data
- The number of clusters does not have to be specified in advance

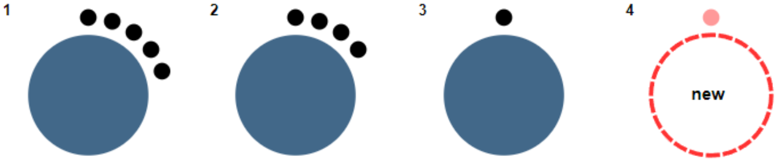
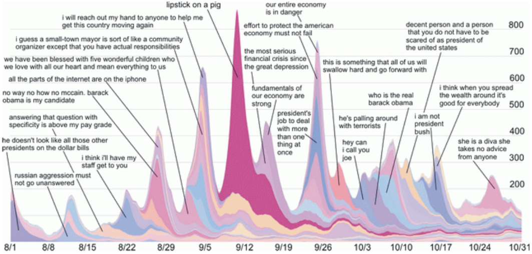


Figure 1: Example of a Dirichlet Process (10 steps)

Dirichlet Process - Why

- The DP exhibit a linear "rich-get-richer" property ($P(c) \propto N_c$)
  - Why $P(c)$ should linearly depend on N_c? (Lee and Sang, 2022)
  - Why $P(c)$ should depend on population at all? (Wallach et al., 2010)
- The "rich-get-richer" implies that the expected number of clusters K grows as $\alpha log(N)$ $\rightarrow$ No *a priori* reason for it to be true
  - Ex. news stream: new clusters may appear at a constant rate
  - Problem usually bypassed by fine-tuning α $\rightarrow$ We lose all the benefit of a sequential model.



To summarize

- Dirichlet Processes are an arbitrary choice.
  - Other priors are possible (Welling, 2006; Lee and Sang, 2022)
  - The choice of the prior matters (Wallach et al., 2009)
- Most variations still consider a linear dependence on N_c:
  - Pitman-Yor Process: $P(c) \propto N_c - \beta$
  - Generalized Gamma Process: $P(c) \propto N_c - \sigma$
  - Indian Buffet Process: $P(c) \propto N_c - \frac{\alpha}{K}$
- The Uniform Process gets rid of the dependence on N_c:
  - Uniform Process: $P(c) \propto 1$
  - Intermediate alternatives?

$\rightarrow$ Powered Dirichlet Process as an answer

Powered Dirichlet Process

- Powered Dirichlet Process (PDP):

$$PDP(C_i = c|\mathcal{H}, \alpha, r) = \begin{cases} \frac{N_c^r}{\alpha+\sum_k N_k^r} & \text{if c = 1, ..., K} \\ \frac{\alpha}{\alpha+\sum_k N_k^r} & \text{if c = K+1} \end{cases}$$

- Generalization of existing models:
  - $r < 0$: "rich-get-poorer"
  - $r = 0$: "rich-get-no-richer" (Uniform Process)
  - $0 < r < 1$: "rich-get-less-richer"
  - $r = 1$: "rich-get-richer" (Dirichlet Process)
  - $r > 1$: "rich-get-more-richer"

Implications of PDP

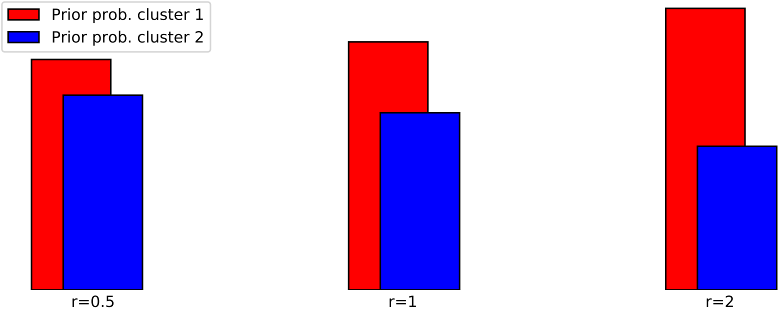


Figure 2: If $r < 1$, populated clusters have a smaller *a priori* probability to get chosen over smaller ones. If $r > 1$, populated clusters have an even greater *a priori* probability to get chosen over smaller ones.

Implications of PDP

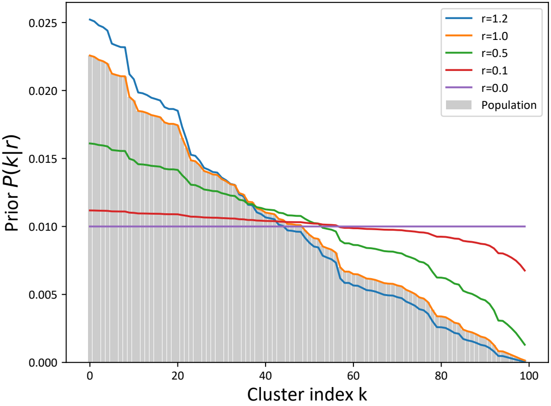
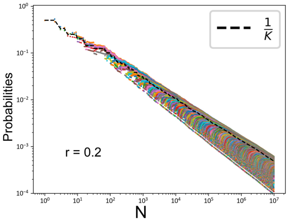


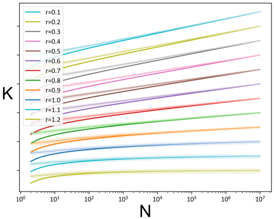
Figure 3: Same figure as before, but with 100 clusters. Spot the: "rich-get-no-richer", "rich-get-less-richer", "rich-get-richer", "rich-get-more-richer"

Elementary results

- Convergence of $PDP(c|\mathcal{H}, r)$ when $N \to \infty$:
  - $r < 1$: Uniform distribution
  - $r = 1$: Dirichlet distribution
  - $r > 1$: Dirac distribution
- Expected number of clusters $\mathbb{E}(K|N)$ when $N \gg 1$:
  - $r < 1$: $\mathbb{E}(K|N) \propto H_{\frac{r^2+1}{2}}(N) \propto N^{\frac{1-r^2}{2}}$
  - $r = 1$: $\mathbb{E}(K|N) \propto H_1(N) \quad \propto log(N)$
  - $r > 1$: $\mathbb{E}(K|N) \propto H_r(N) \quad \propto \zeta(\frac{r^2+1}{2})$

with $H_m(n) := \sum_{k=1}^{n} k^{-m}$ the generalized harmonic number.



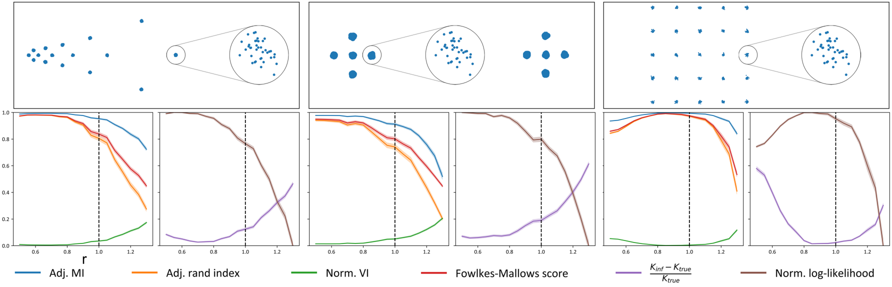


Synthetic data

- We couple PDP with an Infinite Gaussian Mixture Model

$$\rightarrow P(c|data) \propto IGMM(data|c) \times \begin{cases} N_c^r & \text{if c = 1, ..., K} \\ \alpha & \text{if c = K+1} \end{cases}$$

- We generate 3 types of datasets and run 100 experiments for each
- Results show that tuning r allows for better results

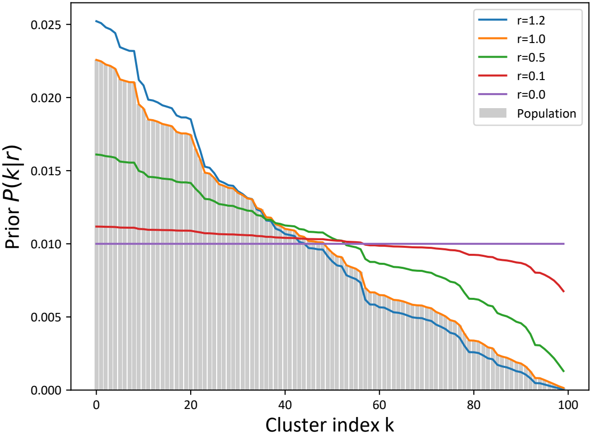


Real-world data

		Adj.MI (↑)	Adj.RI (↑)	Norm.VI (↓)	$\frac{K_{inf}-K_{true}}{K_{true}}$ (↓)
Iris	PDP (r=0.90)	**0.868(4)**	**0.866(7)**	**0.057(2)**	**0.000(0)**
	DP (r=1.00)	0.843(6)	0.820(12)	0.065(2)	0.030(10)
	UP (r=0.00)	0.544(2)	0.295(3)	0.303(2)	2.777(32)
Wines	PDP (r=0.10)	**0.712(15)**	**0.637(20)**	**0.102(5)**	**0.157(17)**
	DP (r=1.00)	0.589(19)	0.461(16)	0.128(4)	0.327(13)
	UP (r=0.00)	**0.713(17)**	**0.657(21)**	**0.103(5)**	**0.147(17)**
Cancer	PDP (r=0.10)	**0.254(17)**	**0.278(21)**	**0.118(1)**	**0.000(0)**
	DP (r=1.00)	0.085(16)	0.094(19)	0.108(2)	**0.000(0)**
	UP (r=0.00)	**0.271(17)**	**0.300(21)**	**0.118(1)**	**0.000(0)**
20-NG	PDP (r=0.80)	**0.421(4)**	**0.119(3)**	**0.477(3)**	-
	DP (r=1.00)	0.404(4)	0.105(4)	0.491(3)	-
	UP (r=0.00)	0.000(4)	0.000(0)	0.830(3)	-

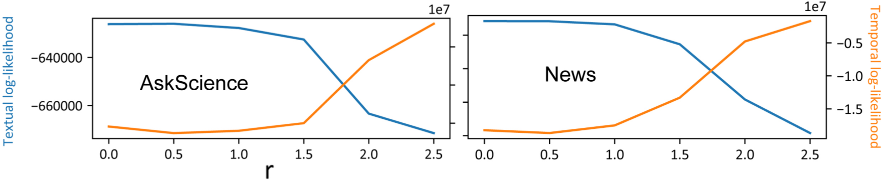
Control prior's informativeness

- Another way to look at PDP is as a way to control the prior's informativeness regarding the data it uses



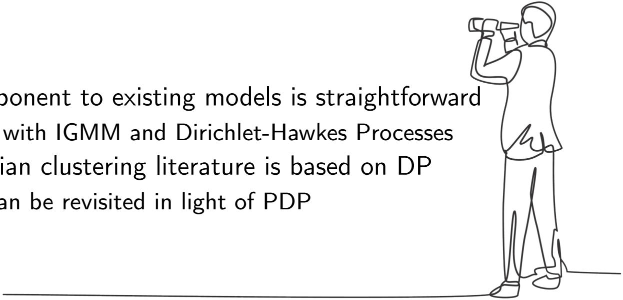
Perspectives

- In (Du et al., 2015): DP combined to Hawkes processes
  - The prior probability relies on the intensity of a temporal process
  - → How much should we rely on this temporal information?
- In (Poux-Médard et al., 2021), we explored this question using PDP
  - Depending on the situation, temporal information can be more or less relevant
  - → PDP allows to control its importance in the model

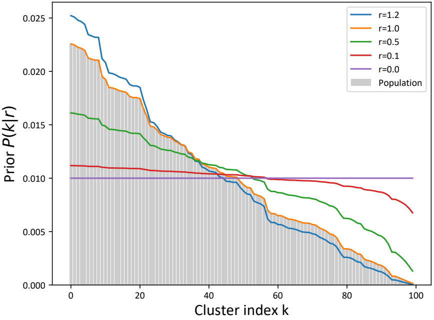


Conclusion and perspectives

- Conclusions:
  - Dirichlet Processes clustering priors come with a questionable "rich-get-richer" hypothesis
  - Powered Dirichlet Process allows for its direct control.
  - Powered Dirichlet Process yields better experimental results
- Perspectives:
  - Adding the exponent to existing models is straightforward
    - → Examples with IGMM and Dirichlet-Hawkes Processes
  - Most of Bayesian clustering literature is based on DP
    - All of it can be revisited in light of PDP



Thanks for your attention!



More here

- Webpage: `https://gaelpouxmedard.github.io/`
- Code and data: `https://github.com/GaelPouxMedard/PDP/`

Bibliographie I

Du, N., Farajtabar, M., Ahmed, A., Smola, A., and Song, L. (2015). Dirichlet-hawkes processes with applications to clustering continuous-time document streams. *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Lee, C. J. and Sang, H. (2022). Why the rich get richer? on the balancedness of random partition models. In *ICML*.

Poux-Médard, G., Velcin, J., and Loudcher, S. (2021). Powered hawkes-dirichlet process: Challenging textual clustering using a flexible temporal prior. *2021 IEEE International Conference on Data Mining (ICDM)*, pages 509–518.

Wallach, H., Jensen, S., Dicker, L., and Heller, K. (2010). An alternative prior process for nonparametric bayesian clustering. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 892–899.

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, 22.

Welling, M. (2006). Flexible priors for infinite mixture models. *Workshop on learning with non-parametric Bayesian methods*.