Powered Dirichlet Process - Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering

Gaël Poux-Médard, Julien Velcin, Sabine Loudcher

▶ **To cite this version:**

Gaël Poux-Médard, Julien Velcin, Sabine Loudcher. Powered Dirichlet Process - Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering. ECML-PKDD 2023, Sep 2023, Torino, Italy. hal-04171235

HAL Id: hal-04171235
https://hal.science/hal-04171235v1

Submitted on 26 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Powered Dirichlet Process - Controlling the "Rich-Get-Richer" Assumption in Bayesian Clustering

Gaël Poux-Médard [✉][1][0000−0002−0103−8778], Julien Velcin[1][0000−0002−2262−045X], and Sabine Loudcher[1][0000−0002−0494−0169]

[1] Université de Lyon, Lyon 2, ERIC UR 3083, 5 avenue Pierre Mendès France, F69676 Bron Cedex, France
gael.poux-medard@univ-lyon2.fr
julien.velcin@univ-lyon2.fr
sabine.loudcher@univ-lyon2.fr

Abstract. The Dirichlet process is one of the most widely used priors in Bayesian clustering. This process allows for a nonparametric estimation of the number of clusters when partitioning datasets. The "rich-get-richer" property is a key feature of this process, and transcribes that the *a priori* probability for a cluster to get selected dependent linearly on its population.
In this paper, we show that such hypothesis is not necessarily optimal. We derive the Powered Dirichlet Process as a generalization of the Dirichlet-Multinomial distribution as an answer to this problem. We then derive some of its fundamental properties (expected number of clusters, convergence). Unlike state-of-the-art efforts in this direction, this new formulation allows for direct control of the importance of the "rich-get-richer" prior. We confront our proposition to several simulated and real-world datasets, and confirm that our formulation allows for significantly better results in both cases.

Keywords: Dirichlet processes · Rich-get-richer · Discrete mathematics · Clustering · Bayesian prior

1 Introduction

The Bayesian clustering approach received a broad attention over the last decades. A non-exhaustive list of application includes medicine, [13], natural language processing [4, 33], genetics [16, 20, 23], recommender systems [1, 10, 22], sociology [6, 12], etc. The key idea is to generate a set of independent observations according to a set of latent variables (clusters). Given a set of existing observations, the prior probability that the next one is generated by any cluster depends on the number of observations they already generated. A very popular prior on clusters distributions that allows this is the Dirichlet distribution. It can be expressed as a process, the Dirichlet process, which allows new observations to be generated by yet unobserved clusters (that have not generated any observations).

However, the Dirichlet process' (and the related Pitman-Yor process') underlying hypothesis is that the prior probability depends linearly on the number of existing observations from a cluster: the *rich-get-richer* property [7]. While this seems a reasonable hypothesis in the complete absence of additional information on the generative process, it fails to describe situations where data is available beforehand. Depending on the data, there might not be any reason for clusters growth to rely linearly on their population, if at all [23, 30] In most cases, an ad-hoc solution is to fine-tune the Dirichlet process' concentration parameter α. However, this practice makes the resulting model unable to consider new data without fine tuning the α parameter again. This is a major problem due to most Dirichlet processes being used for online inference, where data is considered sequentially. Any new observation thus requires fitting the whole model once again. The need for alternatives to vanilla Dirichlet processes has already been pointed out in earlier works [31]. This problem is especially visible in the case of imbalanced data and scale-dependent clustering.

As an example of the imbalance problem, consider a case where data is treated sequentially –which is often the case when it comes to Dirichlet process. A new observation would have a much larger *a priori* probability to belong to a populated but irrelevant cluster, than to open a new one (this probability decreases as $\frac{1}{N_{obs}}$ in vanilla Dirichlet processes). In most situations where it is used, the "rich-get-richer" hypothesis does not transcribe the reality of a situation. For instance, when sampling topics from news streams [30, 32], there is no reason for a new topic to appear in the feed at a rate $\alpha \log N$ as in Dirichlet processes.

As for scale-dependent clustering, similar problems arise. Consider clustering people pinpointed on a map. Tiny clusters (at the scale of cities, for instance) might go unnoticed if clusters are created for larger scales (countries, for instance). The problem can be avoided by fine-tuning the α parameter so that city-scale clusters are found. But then, adding new observations would break the so-found balance on the clusters' scale, because of the rich-get-richer property. In vanilla Dirichlet processes, the number of clusters grow logarithmically with the number of observations ; for instance, if the number of cities grows sublogarithmically with the population instead, adding new observations would require fine-tuning α and fitting the whole model again to get relevant results. In this case, the "rich-get-richer" assumption as is may be too strong a hypothesis, but a "rich-get-no-richer" [30] might as well fail to capture any density-related effect; the optimal solution would be in-between these two priors, depending on the clustering objective. We explore such a case in Fig. 4.

We design a method to bridge the variety of possible priors between the Dirichlet process (DP) and the Uniform process (UP), in a continuous fashion. By generalizing these works, we show the existence of an unexplored class of behaviours, such as "rich-get-less-richer", "rich-get-more-richer" and "poor-get-richer". Little has been done in exploring alternative forms of priors for non-parameteric Bayesian modeling. In the present work, we propose to explicitly tune the importance of the "rich-get-richer" assumption. The resulting Powered Dirichlet Process (PDP) generalizes state-of-the-art works such as UP [30] and

DP. We show that controlling the "rich-get-richer" prior allows for better results on both synthetic and real-world datasets.

2 Background

2.1 Motivation

This work is motivated by the need to control the "rich-get-richer" assumption's importance in Dirichlet process priors. The "rich-get-richer" property of the DP may not always make it the suitable prior for modeling a given dataset. The usual motivation for using a DP prior is that a new observation has a prior probability of being assigned to any cluster proportional to its population. This leads to a prior probability of opening a new cluster decreasing as the inverse of the number of observations, which makes little sense in a number of real-world situations.

Most state-of-the-art works rely on tuning a parameter α (see Eq. 1) to get the "right" number of clusters. This parameter shifts the distribution of the number of clusters as $\mathbb{E}(K|N) \propto \alpha \log N$ with K the number of clusters and N the number of observations. However, we argue this is a bad practice in some cases, typically when clusters size N_c grows sublinearly with the number of observations N [3]. For instance, tackling entity resolution problems need such sublinear growth [27, 26]. When data is treated sequentially, the α parameter has to be fine-tuned after the fit has been performed; because its value depends on the number of observations, it makes the model unsuitable to train on new data without fitting and fine-tuning α again.

To alleviate this problem, we derive a more general form of the DP process that allows for natural control of the "rich-get-richer" property.

2.2 Previous works

Dirichlet process A well-known metaphor for the Dirichlet process is referred to as "Chinese restaurant". The corresponding process is named "Chinese Restaurant Process" (CRP): if a n^{th} client enters a Chinese restaurant, they will sit at one of the K already occupied table with a probability proportional to the number of persons already sat at this table. They can also go to a new table and be the first client to sit there with a probability inversely proportional to the total number of clients in the restaurant. It can be written formally as:

$$CRP(C_i = c|\alpha, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c}{\alpha+N} \text{ if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+N} \text{ if c = K+1} \end{cases} \quad (1)$$

Where c is the cluster chosen by the i^{th} customer, N_k is the population of cluster k, K is the number of already occupied tables and α the concentration parameter. When the number of clients goes to infinity, this process is equivalent to a draw from a Dirichlet distribution over an infinite number of clusters with a

uniform concentration parameter α. It can be shown that the expected number of clusters after N observations evolves as $\log N$ [2].

The two best-known variations of the regular Dirichlet process that address the "rich-get-richer" property control are the seminal Pitman-Yor process and the Uniform process. Each of them can be expressed in a similar form as Eq. 1.

Uniform process The Uniform process has been used in some occasions [16, 23] without proper definition. More recently, it has been formalized and studied in comparison with the regular Dirichlet and Pitman-Yor processes [30]. It reads:

$$UP(C_i = c|\alpha, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{1}{\alpha+K} \text{ if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+K} \text{ if c = K+1} \end{cases} \quad (2)$$

Its formulation completely gets rid of the "rich-get-richer" property. The probability of a new client joining an occupied table is a uniform distribution over the number of occupied tables; it does not depend on the tables' population. In [30], it has been shown that the expected number of tables evolves with N as $\sqrt{N}$. Removing the "rich-get-richer" property leads to a flat prior. As we show later, our formulation allows to retrieve such flat priors and thus generalizes the Uniform Process.

The authors also address the non-exchangeability of this process; they argue that it plays a minor role in inference tasks when using Gibbs sampling algorithms. A recent extension of the Uniform process that guarantees its exchangeability has been proposed in [18]. In this work, the *a priori* probability of opening a new cluster is a constant anymore, and the *a priori* probability to belong to either cluster is constant as in [30]. However, it does not allow for direct control of the "rich-get-richer" property, which is absent of the proposed process.

Pitman-Yor process Following the Chinese Restaurant process metaphor, the Pitman-Yor process [21, 15] proposed to incorporate a *discount* β when a client opens a new table. Mathematically, the process can be formulated as:

$$PY(C_i = c|\alpha, \beta, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c-\beta}{\alpha+N} \text{ if c = 1, 2, ..., K} \\ \frac{\alpha+\beta K}{\alpha+N} \text{ if c = K+1} \end{cases} \quad (3)$$

The introduction of the discount parameter increases the probability of creating new clusters. A table with fewer customers has significantly less chances to gain new ones, while the probability of opening a new table increases significantly. It can be shown that the number of tables evolves with the number of clients N as N^β [11, 28]. However, this process does not control the arguable "rich-get-richer" hypothesis [31], since the relation to the population of a table remains linear; it only scales the linear dependence of a value β. The Pitman-Yor process thus comes with two limitations. First, since $\beta > 0$, it cannot be tuned to generate fewer clusters. Second, the discount parameter does not affect the linear dependence on previous observations for cluster allocations — rich still get richer.

Other rich-get-richer priors Another similar prior, the Power-law Indian Buffet Process, has been proposed so that a realization would yield a number of clusters obeying a power-law as the number of observations increases [29]. This formulation can be seen as a generalization of the Pitman-Yor process; it adds an additional parameter that sums with the number of observations. However, the posterior probability for a new customer to belong to a cluster depends linearly on each cluster's size, and the "rich-get-richer" hypothesis is preserved.

Finally, the Generalized Gamma Process proposed a similar discount idea to increase the probability of opening new clusters in [19]. The proposed prior [19]-Eq. 4 modifies a cluster's probability to get chosen by subtracting a constant term to each cluster's population. Thus, the "rich-get-richer" property is not alleviated in their approach either, since the dependence on cluster's population is still linear. As for the PY process, this formulation only allows to increase the number of clusters and does not alleviate the "rich-get-richer" hypothesis.

2.3 Contributions

In the present work, we derive the Powered Dirichlet Process (PDP) that allows controlling the "rich-get-richer" property while generalizing state-of-the-art works. This allows to define new classes of *a priori* hypotheses: poor-get-richer, rich-get-no-richer (Uniform process), rich-get-less-richer, rich-get-richer (DP), and rich-get-more-richer. We detail some key-properties of the Powered Dirichlet Process (convergence, expected number of clusters). Finally, we show that controlling the "rich-get-richer" prior of simple models yields better results on synthetic and real-world datasets.

3 The model

3.1 The Dirichlet-Multinomial distribution

We recall:

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\prod_k p_k^{\alpha_k - 1}}{B(\boldsymbol{\alpha})} \qquad Mult(\boldsymbol{N}|N, \boldsymbol{p}) = \frac{\Gamma(n+1)}{\prod_k \Gamma(N_k+1)} \prod_k p_k^{N_k} \tag{4}$$

With $\boldsymbol{N} = (N_1, N_2, ..., N_K)$ where N_k is the integer number of draws assigned to cluster k, $N = \sum_k N_k$ the total number of draws, $\Gamma(x) = (x-1)!$ and $B(\boldsymbol{x}) = \prod_k \Gamma(x_k) / \Gamma(\sum_k x_k)$.

The regular Dirichlet process can be derived from the Dirichlet-Multinomial distribution. The Dirichlet-Multinomial distribution is defined as follows:

$$\begin{aligned} DirMult(\boldsymbol{N}|\boldsymbol{\alpha}, n) &= \int_{\boldsymbol{p}} Mult(\boldsymbol{N}|\boldsymbol{p}, n) Dir(\boldsymbol{p}|\boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \frac{B(\boldsymbol{\alpha}+\boldsymbol{N})\Gamma(n+1)}{B(\boldsymbol{\alpha})\prod_k \Gamma(N_k+1)} \int_{\boldsymbol{p}} \underbrace{\frac{\prod_k p_k^{N_k+\alpha_k-1}}{B(\boldsymbol{\alpha}+\boldsymbol{N})}}_{Dir(p|\boldsymbol{\alpha}+\boldsymbol{N})} d\boldsymbol{p} \\ &= \frac{\Gamma(\sum_k \alpha_k)\Gamma(n+1)}{\Gamma(\sum_k \alpha_k + N_k)} \prod_{k=1}^{K} \frac{\Gamma(N_k+\alpha_k)}{\Gamma(N_k+1)\Gamma(\alpha_k)} \end{aligned} \tag{5}$$

In Eq. 5, we sample n values over a space of K distinct clusters each with probability $\boldsymbol{p} = (p_1, p_2, ..., p_K)$, using a Dirichlet prior with parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$.

Now to express this as a Dirichlet Process, we need the probability for a new observation to belong to either cluster given the history of draws. Given Eq. 5, it is equivalent to drawing from a Categorical distribution with a Dirichlet prior with concentration parameter $\boldsymbol{\alpha} + \boldsymbol{N}$.

3.2 Powered conditional Dirichlet prior

In the the standard Dirichlet-Multinomial posterior predictive, the categorical distribution is coupled with a Dirichlet prior $Dir(\boldsymbol{p}|\boldsymbol{\alpha} + \boldsymbol{N})$. We propose to modify this prior by defining the Powered Dirichlet prior that has a nonlinear dependence on the history of draws:

$$Dir_r(\boldsymbol{p}|\boldsymbol{\alpha}, \boldsymbol{N}) = \frac{1}{B(\boldsymbol{\alpha}+\boldsymbol{N^r})} \prod_k p_k^{\alpha_k + N_k^r - 1} \tag{6}$$

where $\boldsymbol{N^r} = (N_1^r, ..., N_K^r)$ still represents the population of each cluster, but at the power r. The parameter $r \in \mathbb{R}$ controls the intensity of the shift on the concentration parameter. It is straightforward to demonstrate that the Powered Dirichlet distribution is a conjugate prior of the Multinomial distribution.

3.3 Posterior predictive

We are looking for the probability of the n^{th} draw belonging to cluster k. Let $\boldsymbol{c} = (c_1, ..., c_K)$ where $c_k = 1$ if the observation belongs to cluster k, and 0 otherwise.The probability of a draw from the Categorical distribution $Cat(\boldsymbol{c}|\boldsymbol{p}) = \prod_k p_k^{c_k}$ given a Powered Dirichlet prior as defined Eq. 6 reads:

$$\begin{aligned} DirCat_r(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{N}) &= \int_{\boldsymbol{p}} Cat(\boldsymbol{c}|\boldsymbol{p}) Dir_r(\boldsymbol{p}|\boldsymbol{\alpha}, \boldsymbol{N}) d\boldsymbol{p} \\ &= \int_{\boldsymbol{p}} \frac{1}{B(\boldsymbol{\alpha}+\boldsymbol{N^r})} \prod_k p_k^{c_k+\alpha_k+N_k^r-1} d\boldsymbol{p} = \frac{B(\boldsymbol{c}+\boldsymbol{\alpha}+\boldsymbol{N^r})}{B(\boldsymbol{\alpha}+\boldsymbol{N^r})} \end{aligned} \tag{7}$$

3.4 Powered Dirichlet Process

We finally derive an expression for the Powered Dirichlet Process from Eq. 7. Taking back the conditional probability for the n^{th} observation to belong to cluster c (Eq. 7), we have:

$$\begin{aligned} DirCat_r(\boldsymbol{c}|\boldsymbol{N},\boldsymbol{\alpha}) =& B(\boldsymbol{c}+\boldsymbol{\alpha}+\boldsymbol{N^r})/B(\boldsymbol{\alpha}+\boldsymbol{N^r}) \\ =& \frac{(N_c^r+\alpha_c)\prod_k \Gamma(N_k^r+\alpha_k)}{(\sum_k N_k^r+\alpha_k)\Gamma(\sum_k N_k^r+\alpha_k)} \cdot \frac{\Gamma(\sum_k N_k^r+\alpha_k)}{\prod_k \Gamma(N_k^r+\alpha_k)} \\ =& \frac{N_c^r+\alpha_c}{\sum_k N_k^r+\alpha_k} \end{aligned} \tag{8}$$

Now that we have the probability for each draw to belong to any cluster, we can iterate Eq. 8 as a process over K clusters. Finally, we assume an infinity of available clusters ($K \to \infty$). When considering a new observation, we must associate it to one of these clusters, that can be empty ($N_k = 0$) or non-empty ($N_k > 0$).

From Eq. 8, the probability of choosing a non-empty cluster c linearly depends on $N_c^r + \alpha_c$; the probability of choosing *one* empty cluster linearly depends on α_c. However, all empty clusters are rigorously interchangeable, because they are fully characterized by their (null) population. We can therefore describe the initial probability of choosing *any* cluster with a single value $\alpha := \sum_k^\infty \alpha_k$. Because the number clusters is infinite, it follows that $\sum_k^K \alpha_k \to 0$ for any finite set of K clusters. Therefore, for the finite set of non-empty clusters, $\alpha_k = 0$. On the other hand, for the infinite set of empty clusters, the sum of their α_k goes to α.

In the following, we rewrite K the number of non-empty clusters, and $K+1$ any of the empty clusters. This transition between the finite Dirichlet-Categorical distribution and the infinite Dirichlet Process is standard in the literature [17, 8]. From these considerations and Eq. 8, the Powered Dirichlet Process follows:

$$PDP(C_n = c|\alpha, C_1, C_2, ..., C_{i-1}) = \begin{cases} \frac{N_c^r}{\alpha+\sum_k^K N_k^r} \text{ if c = 1, 2, ..., K} \\ \frac{\alpha}{\alpha+\sum_k^K N_k^r} \text{ if c = K+1} \end{cases} \tag{9}$$

This formulation generalizes the Uniform process when $r = 0$ and the Dirichlet process when $r = 1$.

We illustrate the change on prior probability for an existing cluster to get chosen due to the Powered Dirichlet Process in Fig. 1. This figure plots the population of clusters (grey bars) and their associated prior probability of being selected. When $r > 1$, the most populated clusters are associated with a higher prior probability than in the standard CRP, whereas the less populated ones have even less chances to get chosen; rich-get-more-richer. When $r < 1$, the exact opposite is observed; rich-get-less-richer. In the limit case $r = 0$, we recover the Uniform Process; rich get-no-richer.

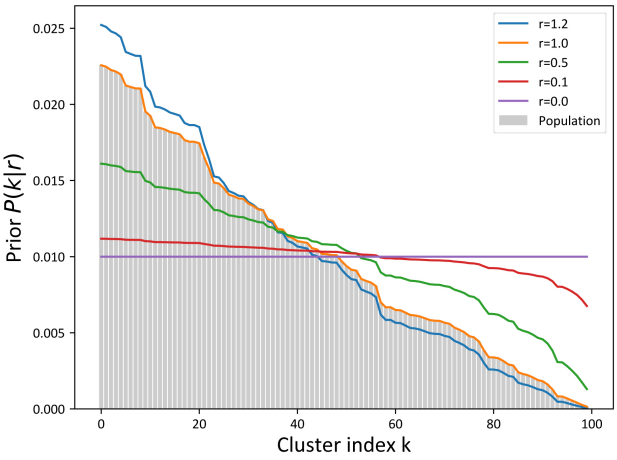

Fig. 1. Illustration of the effect of r on the Powered Dirichlet Process prior probability. Populations have been randomly sampled from a uniform distribution.

4 Properties of the Powered Dirichlet Process

4.1 Convergence

Proposition 1 *For $N \to \infty$, the Powered Dirichlet Process converges to a stationary distribution. When $r < 1$, it converges to a uniform distribution, and when $r > 1$, it converges to a Dirac distribution.*

Proof. We start with a simple situation where only 2 clusters are involved. The generalization to the case where $K \to \infty$ clusters are involved is straightforward. When clusters' population is large enough, we make the following Taylor approximation:

$$(N_i + 1)^r = N_i^r (1 + \frac{1}{N_i})^r = N_i^r + r N_i^{r-1} + \mathcal{O}(N^{r-2}) \tag{10}$$

Since the population of a cluster N_i is a non-decreasing function of N, we assume that first order Taylor approximation holds when $N \to \infty$. Given clusters population at the N^{th} observation, we perform a stability analysis of the gap between probabilities $\Delta p(N) = p_1(N) - p_2(N)$. We recall that the probability for cluster i to get selected at step N is $p_i(N) = N_i^r / (\sum_k N_k^r)$. Either cluster is selected with this probability at step $N+1$: $\Delta p(N+1) = p_1(N+1) - p_2(N)$ with probability $p_1(N)$, and $\Delta p(N+1) = p_1(N) - p_2(N+1)$ with probability $p_2(N)$). Explicitly, the variation of the gap between probabilities when N grows

is written as:

$$\begin{aligned}
&\frac{p_1(N)(p_1(N+1)-p_2(N))+p_2(N)(p_1(N)-p_2(N+1))-\Delta p(N)}{\Delta p(N)} \\
&\overset{\text{Eq. 10}}{\approx} \frac{1}{p_1(N)-p_2(N)} \times \left(p_1(N)\frac{N_1^r - N_2^r + rN_1^{r-1}}{N_1^r + N_2^r + rN_1^{r-1}} + p_2(N)\frac{N_1^r - N_2^r - rN_2^{r-1}}{N_1^r + N_2^r + rN_2^{r-1}} \right) \\
&= \frac{2rN_1^r N_2^r}{(N_1^r + N_2^r + rN_1^{r-1})(N_1^r + N_2^r + rN_2^{r-1})} \left(\frac{N_1^{r-1} - N_2^{r-1}}{N_1^r - N_2^r} \right)
\end{aligned} \tag{11}$$

We see in Eq. 11 that the sign of the variation of the gap between probabilities depend only on the term $\frac{N_1^{r-1}-N_2^{r-1}}{N_1^r-N_2^r}$. We can therefore perform a stability analysis of the Powered Dirichlet Process using only this expression.

— For $0 < r < 1$ the following relation holds: $N_1^{r-1} - N_2^{r-1} < 0 \Leftrightarrow N_1^r - N_2^r > 0 \; \forall N_1, N_2$, and conversely. That makes right hand side of Eq. 11 negative. Therefore adding a new observation statistically reduces the gap between the probabilities of the two clusters. We could forecast this prediction from Eq. 10. We see that the more a cluster is populated, the less a new observation increases its probability at the next step – rich-get-less-richer. Moreover, we see from Eq. 10 that a crowded cluster (such as $N_1^r \gg N_2^r$) see its probability evolve as N^{r-1}. Asymptotically, the only fixed point of Eq. 11 when $N \to \infty$ is $N_1 \to N_2$, which implies a uniform distribution. We verify this result numerically in Fig. 2-left.

— For $r > 1$ the following relation holds: $N_1^{r-1} - N_2^{r-1} > 0 \Leftrightarrow N_1^r - N_2^r > 0 \; \forall N_1, N_2$; ; that makes right hand side of Eq. 11 positive. Adding a new observation statistically increases the gap between probabilities. From Eq. 10, we see that the more a cluster is populated, the more a new observation increases its probability at the next step – rich-get-more-richer. In this case, Eq. 11 has $K+1$ fixed points, with K the number of clusters. The uniform distribution is an unstable fixed point, while K Dirac distributions (each on one cluster) are stable fixed points of the system. It means the gap converges to 1, that is a probability of 1 for one cluster and a probability of 0 for the others.

— For $r = 1$, the right hand side of Eq. 11 is null. It means the gap remains statistically constant $\forall N_i$, which is a classical result for the regular Dirichlet process. This convergence has already been studied on many occasions [7, 2].

— For $r \to 0$, Eq. 11 is not defined anymore. That is because the probability for a cluster to be chosen does not depend on its population anymore. In this case, $p_1(N) - p_2(N) \propto N_1^0 - N_2^0 = 0$: the probability for any cluster to be chosen is equal at all times, hence the Uniform process – "rich-get-no-richer".

4.2 Expected number of tables

Proposition 2 *When N is large, $\sum_k N_k^r$ varies with N as $N^{\frac{r^2+1}{2}}$ when $r < 1$, and with N^r when $r \geq 1$.*

Proof. Taking back Eq. 9, we are interested in the variation of $p_i = \frac{N_i^r}{\sum_k N_k^r}$ according to N when N_i^r is large. Since N_i is either way a non-decreasing function of N, we reformulate the constraint N_i^r large into N^r large:

$$p_i(N+1) - p_i(N) = \begin{cases} \frac{rN_i^{r-1} + \mathcal{O}(N^{r-2})}{\sum_k N_k^r} & \text{if } N_i \text{ grows} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

— For $r < 1$, the larger N_i the slower the variation of p_i. It means that for large N_i^r, we can write $N_i \propto Np_i$, with p_i being now independent N.

— For $r > 1$, the probability p_i varies greatly with N and quickly converges to 1 for large N (see Proposition 1), and so $N_i \approx N$ for cluster i and $N_{j\neq i} \ll N_i \; \forall j$.

Because the sum $\sum_k N_k^r$ mostly varies according to large N_k, we approximate $\sum_k N_k^r \approx N^r \sum_k p_k^r$ for large N^r.

Besides, we showed in Proposition 1 that for large N the process converges to a uniform distribution for $r < 1$ and to a Dirac distribution when $r > 1$. Therefore, we can express $\sum_k^K p_k^r$ as:

$$\sum_k^K p_k^r \overset{N \gg 1}{\approx} \begin{cases} K \cdot (\frac{1}{K})^r = K^{1-r} & \text{for } r < 1 \\ 1 & \text{for } r \geq 1 \end{cases} \tag{13}$$

Based on the demonstration of Eq.4 in [30], we assume that K evolves with N as $N^{\frac{1-r}{2}}$ when $r < 1$. We verify that this assumption holds in the Experiment section, Fig. 2-middle.

Therefore, we can write:

$$\sum_k N_k^r \approx N^r \sum_k^K p_k^r \approx \begin{cases} N^r \left(N^{\frac{1-r}{2}}\right)^{1-r} = N^{\frac{1+r^2}{2}} & \text{for } r < 1 \\ N^r & \text{for } r \geq 1 \end{cases} \tag{14}$$

Proposition 3 *For $N \gg 1$, the expected number of tables of the Powered Dirichlet Process grows with N as $H_{\frac{r^2+1}{2}}(N)$ for $r < 1$ and as $H_r(N)$ when $r \geq 1$, where $H_m(n)$ is the generalized harmonic number.*

Proof. In general, the expected number of clusters at the N^{th} step can be written as:

$$\mathbb{E}(K|N,r) = \sum_1^N \frac{\alpha}{\sum_k N_k^r + \alpha} \overset{N^r \gg 1}{\propto} \sum_1^N \frac{1}{\sum_k N_k^r} \tag{15}$$

We showed in Proposition 2 that we can rewrite $\sum_k N_k^r \propto N^{\frac{r^2+1}{2}}$ when $r < 1$ and $\sum_k N_k^r \propto N^r$ when $r \geq 1$. Injecting this result in Eq. 15, we get:

$$\mathbb{E}(K|N,r) \overset{N^r \gg 1}{\propto} \begin{cases} \sum_1^N N^{-\frac{r^2+1}{2}} = H_{\frac{r^2+1}{2}}(N) & \text{for } r < 1 \\ \sum_1^N N^{-r} = H_r(N) & \text{for } r \geq 1 \end{cases} \tag{16}$$

This result is verified numerically in Fig. 2-right.

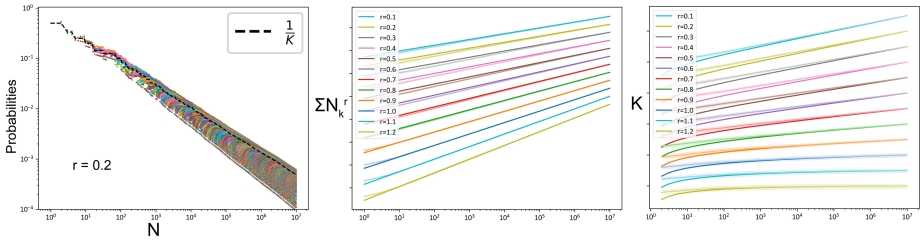

Fig. 2. Numerical validation of Propositions 1 (left), 2 (middle), 3 (right). In the first plot, K is the number of non-empty clusters. In the second and third plots, the theoretical results are the solid lines and the associated numerical results are the transparent lines of same color. Except for small N, the difference between theory and experiments is almost indistinguishable.

— For $r = 1$, $\mathbb{E}(K|N, r = 1) \propto H_1(N) \approx \gamma + \log(N)$ where γ is the Euler–Mascheroni constant, which is a classical result for the regular Dirichlet process.

— For $r > 1$ and $N \to \infty$, the term $H_{\frac{r^2+1}{2}}(N)$ converges towards a finite value and the sum $\sum_k p_k^r$ goes to 1 (see Proposition 1). By definition $\mathbb{E}(K|N, r > 1) \overset{N\to\infty}{\propto} \zeta(\frac{r^2+1}{2})$, where ζ is the Riemann Zeta function.

— For $r < 1$, we need to approximate the harmonic number in a continuous setting. We rewrite Eq. 16 as:

$$\sum_{n=1}^{N} n^{-\frac{r^2+1}{2}} \overset{N^r \gg 1}{\approx} \int_1^N n^{-\frac{r^2+1}{2}} dn = \frac{2}{1-r^2}(N^{\frac{1-r^2}{2}} - 1) \tag{17}$$

One can show that $\frac{N^{1-x}-1}{1-x} = H_x(N) + \mathcal{O}(\frac{1}{N^x})$. Therefore, the number of expected clusters in the Powered Dirichlet Process exhibits a power-law behaviour, similar to the Pitman-Yor process for $r = \sqrt{1-2\beta}$ for $0 < r < 1$. For values of $r > 1 \Leftrightarrow \beta < 0$, the equivalent Pitman-Yor process is not defined. Note that there is no *a priori* reason for r to be constrained in the domain of positive real number. Complex and negative analysis of the process might be an interesting lead for future works.

5 Experiments

5.1 Use case: infinite Gaussian mixture model

We consider a classical infinite Gaussian mixture model coupled with a Powered Dirichlet Process prior. We choose this application to ease visual understanding of the implications of the PDP, but the argument holds for other models using DP priors as well (text modeling, gene expression clustering, etc.). We fit the data using a standard collapsed Gibbs sampling algorithm for IGMM [24, 30, 33], with a Normal Inverse Wishart prior on the Gaussians' parameters. The order

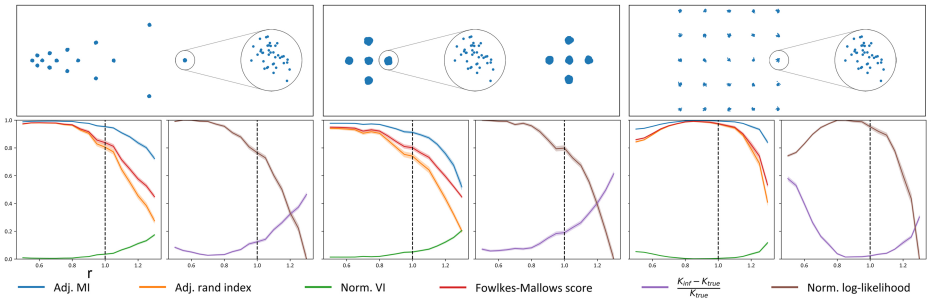

Fig. 3. Application on synthetic data. (Top) Original datasets used for the experiments (Density, Diamond and Grid). (Bottom) Results for various values of r; the x and y axes all the same. The dashed line indicates the regular DP prior as $r = 1$. The error correspond to the standard error of the mean over all runs.

in which input data is processes is set at random at each iteration, so that we reduce the ordering bias from the dataset [30]. Note that we cannot completely get rid of it because the Powered Dirichlet Process is not exchangeable for all r. The problem has been addressed on numerous occasions (Uniform process [30], distance-dependent CRP [5, 9], spectral CRP [25], balance-neutral partition [18]) and shown to induce negligible variations of results in the case of Gibbs sampling. We stop the sampler once the likelihood reaches stability; we perform 100 runs for each value of r. Finally, the parameter α is set to 1.[1]

Note that we choose not to compare to other types of clustering algorithms. This section aims to demonstrate the usefulness of a generalized form of the Dirichlet process with respect to the vanilla one. The argument on a simple model (here a regular DP combined with IGMM) extends to other priors built on Dirichlet processes (Hierarchical and Nested Dirichlet processes). Comparison of DP-based priors to other clustering methods (KNN, DBScan, Spectral clustering, etc.) has already been done numerous times and is out of our scope.

Synthetic data Synthetic datasets are represented in Fig. 3-top, and comprise N=1000 observations each, that have been generated by sampling from 2D Gaussian distributions. We present the results on synthetic data in Fig. 3-bottom and in Table 1. We consider standard metrics in clustering evaluation with a non-fixed number of clusters: mutual information score and rand index both adjusted for chance (**Adj.MI** and **Adj.RI**, higher is better), normalized variation of information (**Norm.VI**, lower is better), Fowlkes-Mallow score (higher is better), marginal likelihood (normalized for visualization, higher is better) and absolute relative variation of the inferred number of clusters according to the number used in the generation process ($\frac{\mathbf{K_{inf}-K_{true}}}{\mathbf{K_{true}}}$, lower is better). The datasets are designed to investigate the effect varying r when clustering can take place on different scales.

[1] Codes and datasets available at https://github.com/GaelPouxMedard/PDP

Table 1. Numerical results of the various priors coupled to a standard IGMM. PDP allows to outperform the baselines consistently. The standard error on the last digit(s) over 100 runs is given in shorthand notation ($0.123(12) \Leftrightarrow 0.123 \pm 0.012$).

		Adj.MI (↑)	Adj.RI (↑)	Norm.VI (↓)	$\frac{K_{inf}-K_{true}}{K_{true}}$ (↓)
Density	PDP (r=0.60)	**0.992(1)**	**0.980(2)**	**0.006(1)**	**0.045(5)**
	DP (r=1.00)	0.951(4)	0.797(17)	0.037(3)	0.128(10)
	UP (r=0.00)	0.939(2)	0.854(4)	0.050(1)	0.548(1)
Diamond	PDP (r=0.50)	**0.982(2)**	**0.956(5)**	**0.011(1)**	**0.063(7)**
	DP (r=1.00)	0.909(7)	0.731(19)	0.053(4)	0.202(12)
	UP (r=0.00)	0.927(2)	0.844(6)	0.051(2)	0.544(2)
Grid	PDP (r=0.85)	**0.997(1)**	**0.990(2)**	**0.003(1)**	**0.014(2)**
	DP (r=1.00)	0.995(1)	0.977(4)	**0.004(1)**	**0.018(3)**
	UP (r=0.00)	0.811(1)	0.517(3)	0.154(1)	2.120(1)
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)	-

Real-world data In Table 1, we report the results for the 20Newsgroup (20-NG) real-world dataset, which is a collection of 18 000 users posts published on Usenet, organized in 20 Newsgroup (which are our target thematic clusters). As a model, we consider a modified version of LDA [4] that uses a PDP prior instead of DP in the words sampling step. Note that because the number of clusters must be provided to LDA, we do not compute $\frac{\mathbf{K_{inf}}-\mathbf{K_{true}}}{\mathbf{K_{true}}}$. We also run additional experiments on well known real-world datasets from sklearn: Iris (4 attributes, 3 classes), Wines (13 attributes, 3 classes), and Cancer (30 attributes, 2 classes). We see PDP allows for improved performances on every dataset.

We now illustrate the interest of using an alternate form of prior for the IGMM on real-world data. We consider a dataset of 4.300 roman sepulchral inscriptions comprising the substring "Antoni" that have been dated between 150AC and 200AC and assigned with map coordinates. The dates correspond to the reign of Antoninus Pius over the Roman empire. The dataset is available on Clauss-Slaby repository[2]. It was common to give children or slaves the name of the emperor; the dataset gives a global idea of the main areas of the roman empire at that time [14]. The task is to discover spatial clusters of individuals named after the emperor. We present the results in Fig. 4.

[2] http://www.manfredclauss.de/fr/index.html

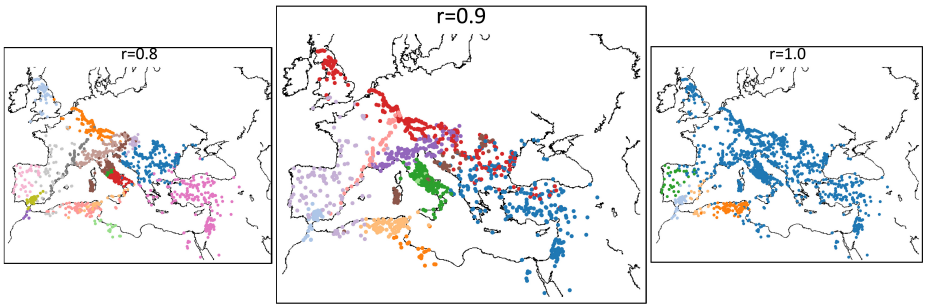

Fig. 4. Application to spatial clustering on geolocated data for $r = 0.8$ (left), $r = 1$ (right) and $r = 0.9$ (middle). We see that the model using a PDP prior for $r = 0.9$ and $r = 0.8$ describes the data better than the same model using a DP prior ($r = 1$).

We see that when $r = 1$, the classical DP prior is not fit for describing this dataset, as it misses most of the clusters. The problem could be solved by fine-tuning the α parameter, but such model would not be hold if we added new observations. On the other hand, when $r = 0.9$, the infinite Gaussian mixture model retrieves relevant clusters. It also finds some clusters that were not expected, such as the north Italian cluster or the long cluster going through Spain and France that corresponds to roman roads layout (via Augusta and via Agrippa; it was common to bury the dead on roads edges). Finally, when $r = 0.8$, some of the main clusters are broken into smaller ones (Italy breaks into Rome, North Italy, and South Italy; Britain becomes an independent cluster, etc.). In this case, tuning r controls the level of details of the clustering.

6 Conclusion

We discussed the need for controlling the "rich-get-richer" property that arises from the usual Dirichlet Process. We then derived the Powered Dirichlet Process to allow for its control. This formulation allows reducing the expected number of clusters, which is not possible with existing processes, while generalizing two of them. We derived elementary results on convergence and expected number of clusters of the PDP. Finally, we showed that it yields better results on both synthetic and real-world data. For future works, it might be interesting to investigate cases where r takes non-positive values (which might lead to a "poor-get-richer" kind of process), and to develop a procedure to infer this parameter based on data (by minimizing a dispersion criterion, for instance).

The regular Dirichlet Process has been used for decades as a powerful prior in many real-world applications. However, alternate forms for this prior have been little explored. It would be very interesting to study the changes brought to state-of-the-art models based on Dirichlet priors by varying the importance of the "rich-get-richer" assumption as proposed in this paper.

References

1. Airoldi, E., Blei, D., Fienberg, S., Xing, E.: Mixed membership stochastic blockmodels. Journal of Machine Learning Research **9**, 1991–1992 (2008)
2. Arratia, R., Barbour, A.D., Tavaré, S.: Poisson process approximations for the ewens sampling formula. The Annals of Applied Probability **2**(3), 519–535 (1992)
3. Betancourt, B., Zanella, G., Miller, J.W., Wallach, H., Zaidi, A., Steorts, R.C.: Flexible models for microclustering with application to entity resolution **29** (2016), https://proceedings.neurips.cc/paper/2016/file/670e8a43b246801ca1eaca97b3e19189-Paper.pdf
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Blei, D., Frazier, P.: Distance dependent chinese restaurant processes. Journal of Machine Learning Research **12**, 2461–2488 (08 2011)
6. Cobo-López S., Godoy-Lorite A., D.J.: Optimal prediction of decisions and model selection in social dilemmas using block models. EPJ Data Sci **7(48)** (2018)
7. Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics **1**(2), 209 – 230 (1973)
8. Frigyik, A.B., Kapila, A., Gupta, M.R.: Introduction to the dirichlet distribution and related processes (2010)
9. Ghosh, S., Raptis, M., Sigal, L., Sudderth, E.B.: Nonparametric clustering with distance dependent hierarchies. p. 260–269. UAI'14 (2014)
10. Godoy-Lorite, A., Guimerà, R., Moore, C., Sales-Pardo, M.: Accurate and scalable social recommendation using mixed-membership stochastic block models. PNAS **113**(50), 14207–14212 (2016)
11. Goldwater, S., Griffiths, T.L., Johnson, M.: Producing power-law distributions and damping word frequencies with two-stage language models. JMLR **12**(68) (2011)
12. Guimera, R., Llorente, A., Sales-Pardo, M.: Predicting human preferences using the block structure of complex social networks. PLOS One **7**(9) (2012)
13. Guimerà, R., Sales-Pardo, M.: A network inference method for large-scale unsupervised identification of novel drug-drug interactions. PLoS Comput Biol (2013)
14. Hanson, J.W., Ortman, S.G., Lobo, J.: Urbanism and the division of labour in the roman empire. Journal of The Royal Society Interface **14**(136), 20170367 (2017)
15. Ishwaran, H., James, L.: Generalized weighted chinese restaurant processes for species sampling mixture models. Statistica Sinica **13**, 1211–1235 (10 2003)
16. Jensen, S., Liu, J.: Bayesian clustering of transcription factor binding motifs. In: Journal of the American Statistical Association. vol. 103, p. 188–200 (2008)
17. Jordan, M.: Dirchlet processes, chinese restaurant processes and all that. ICML (2005)
18. Lee, C.J., Sang, H.: Why the rich get richer? On the balancedness of random partition models. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12521–12541. PMLR (17–23 Jul 2022)
19. Lijoi, A., Mena, R.H., Prünster, I.: Controlling the reinforcement in bayesian non-parametric mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(4), 715–740 (2007). https://doi.org/https://doi.org/10.1111/j.1467-9868.2007.00609.x, https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00609.x

20. McDowell, I.C., Manandhar, D., Vockley, C.M., Schmid, A.K., Reddy, T.E., Engelhardt, B.E.: Clustering gene expression time series data using an infinite gaussian process mixture model. PLoS computational biology **14**(1), e1005896 (2018)
21. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. The Annals of Probability **25**(2), 855 – 900 (1997)
22. Poux-Médard, G., Velcin, J., Loudcher, S.: Interactions in information spread: quantification and interpretation using stochastic block models. arXiv (2020)
23. Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., Lawrence, C.E., Liu, J.S.: Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. In: Nature Biotechnology. vol. 21, p. 435–439 (2003)
24. Rasmussen, C.E.: The infinite gaussian mixture model. p. 554–560. NIPS'99, MIT Press (1999)
25. Socher, R., Maas, A., Manning, C.: Spectral chinese restaurant processes: Nonparametric clustering based on similarities. JMLR - Proceedings **15**, 698–706 (2011)
26. Steorts, R.C.: Entity resolution with empirically motivated priors **10**, 849 (2015)
27. Steorts, R.C., Hall, R., Fienberg, S.E.: Smered: A bayesian approach to graphical record linkage and de-duplication **33**, 922–930 (2014)
28. Sudderth, E., Jordan, M.: Shared segmentation of natural scenes using dependent pitman-yor processes. In: NIPS. vol. 21 (2009)
29. Teh, Y., Gorur, D.: Indian buffet processes with power-law behavior **22** (2009)
30. Wallach, H., Jensen, S., Dicker, L., Heller, K.: An alternative prior process for nonparametric bayesian clustering. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 892–899. JMLR (2010)
31. Welling, M.: Flexible priors for infinite mixture models. In: Workshop on learning with non-parametric Bayesian methods (2006)
32. Xu, W., Li, Y., Qiang, J.: Dynamic clustering for short text stream based on dirichlet process. Appl Intell (2021). https://doi.org/10.1007/s10489-021-02263-z
33. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 233–242. KDD '14, Association for Computing Machinery, New York, NY, USA (2014)