entropy

MDPI

Article

Understanding Hierarchical Processes

Wray Buntine [1,2]

[1] College of Engineering and Computer Science, VinUniversity, Hanoi 100000, Vietnam; wray.b@vinuni.edu.vn
[2] Faculty of Data Science and AI, Monash University, Clayton, VIC 3800, Australia

Abstract: Hierarchical stochastic processes, such as the hierarchical Dirichlet process, hold an important position as a modelling tool in statistical machine learning, and are even used in deep neural networks. They allow, for instance, networks of probability vectors to be used in general statistical modelling, intrinsically supporting information sharing through the network. This paper presents a general theory of hierarchical stochastic processes and illustrates its use on the gamma process and the generalised gamma process. In general, most of the convenient properties of hierarchical Dirichlet processes extend to the broader family. The main construction for this corresponds to estimating the moments of an infinitely divisible distribution based on its cumulants. Various equivalences and relationships can then be applied to networks of hierarchical processes. Examples given demonstrate the duplication in non-parametric research, and presents plots of the Pitman–Yor distribution.

Keywords: Bayesian nonparametrics; Dirichet process; gamma process; Pitman–Yor process; hierarchical process; non-parametric LDA

check for updates

Citation: Buntine, W. Understanding Hierarchical Processes. *Entropy* **2022**, *24*, 1703. https://doi.org/10.3390/e24121703

Academic Editor: Narayanaswamy Balakrishnan

Received: 13 September 2022
Accepted: 18 November 2022
Published: 22 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The hierarchical Pitman–Yor process (HPYP) was first presented as a solution to n-gram language models [1] where it mimics the behavior of the Kneser–Ney algorithm [2]. It is an extension of the hierarchical Dirichlet process (HDP) [3]. The HPYP has since been used in a wide variety of ways, including for previously state-of-the-art and competitive algorithms for topic models [4] and text compression [5]. The HDP has been used for previously state-of-the-art and competitive algorithms for tweet clustering [6] and document segmentation [7]. Many more novel and creative uses of these processes exist, for instance, hierarchical topic models [8]. More general reviews are given by Teh and Jordan [9] and Jordan [10]. The gamma process can also be used hierarchically [11] and provides an alternative scheme for handling the HDP. The notion of hierarchical models fits in well with the computational approach to statistical modelling adopted in the machine learning community.

However, what exactly is the HPYP? A key concept for understanding the HDP and the HPYP is the notion of a discrete base probability measure. The base measure is a source measure for sampling points of the HDP or HPYP. These are discrete just when they have a countable number of possible points (the set on which the measure is based is countable). When finite, the base probability measure is just a probability distribution, usually represented as a vector. However, in non-parametric modelling, we seek to model structured objects for which the dimension may be unknown ahead of time: the number of clusters for points, the depth of a tree, the number of atoms in a molecule, the number of words in a sentence. Allowing the base measure to be countably infinite is a useful abstraction in this situation. Moreover, being able to generate an infinite discrete base probability measure provides us with the ability to model prior distributions for our structured objects without fixing dimensions ahead of time. The above models for text and clustering give examples.

It is known that the hierarchical Dirichlet process, when applied to a finite discrete base distribution, is just a Dirichlet distribution. Indeed, this property is the axiomatic

definition of the process [12]. So, applications of and inference with the HDP are really just using hierarchical Dirichlet distributions, requiring no non-parametric theory to describe, although algorithms may be using non-parametric methods.

So, there is a clear concept of what the HDP model is. What is the corresponding result for the hierarchical Pitman–Yor process? For all the algorithms using the HPYP, it would be nice to know what their actual model is! Teh first referred to the hierarchical version of the PYP as the Pitman–Yor distribution [in talks accompanying] [1], saying it has "no known analytical form". Moreover, is there a more general theory of hierarchical processes, and why does this case (the HDP) come out so neatly? These questions for hierarchical processes have been addressed in recent theory [13–15].

Note the Bayesian theory of non-hierarchical processes is extensive. A comprehensive analysis of different processes is developed by James [16], in the more general context of the generalised Indian buffet process [17]. The general posterior analysis of their normalised versions, including the DP, is developed by James et al. [18]. A useful review of theory and a slice sampler for the case of the normalised generalised gamma is given in Lomeli et al. [19]. A study of some of the processes considered here can also be found in Zhou and Carin [11], focusing on gamma processes and their relationships.

However, these treatments are grounded in extensive probability theory and assume the reader is already familiar with Poisson point processes, Lévy processes, subordinators and other advanced areas [20,21]. Some of these details are not strictly necessary for the understanding of the basic ideas. This paper presents the relevant background theory in a self-contained way to develop models for hierarchical processes generally based on the theory of subordinators and completely random measures [20,21]. The theory for the most part reinterprets results from the Bayesian non-parametric and statistical communities [18,22,23], though some related ideas can also be found in machine learning [11]. However, the answers to the questions about the nature of the HPYP and general application to hierarchical processes, networks of hierarchical processes and generalised Chinese restaurants are not well-known outside the Bayesian non-parametric community, so we present them here in a unified manner.

2. Background Theory

A formal theory of Poisson point processes (PPP), Lévy processes and completely random measures (CRMs) with treatment of measure theory is needed to rigorously cover this area [20,21]. Here, an informal summary is given, though trying to maintain a degree of precision, for instance keeping adequate rigor in the statement of results.

2.1. Completely Random Measures

A CRM is a discrete measure $\mu(\mathrm{d}x)$ on a space $\mathcal{X}$ constructed as

$$\mu(x) = C_0 + \sum_{i=1}^{\infty} \lambda_i \delta_{x_i}(x) \tag{1}$$

where the $x_i \in \mathcal{X}$ are called atoms and are assumed distinct, the $\lambda_i \in \mathcal{R}^+$ called jumps, and the background constant C_0 is zero in our use. This means that $\mu(x_i) = \lambda_i$, evaluated at atoms, and $\mu(x_i) = 0$ otherwise. The (λ_i, x_i) are mutually independent random variables, and a finite number of the x_i can also be fixed. These conditions ensure the measure is completely random, that is for $A, B \subset \mathcal{X}$, if $A \cap B = \emptyset$ then $\mu(A) \perp\!\!\!\perp \mu(B)$.

Moreover, suppose the class of CRMs where $C_0 = 0$ in Equation (1) can be normalised, so $\mu(\mathcal{X}) = \sum_{i=1}^{\infty} \lambda_i < \infty$. This yields discrete probability distributions on $\mathcal{X}$ represented as $\mu(x)/\mu(\mathcal{X})$. These are referred to as normalised random measures with independent increments (NRMIs) [18], a concept developed by Kingman [24], and are a general class of discrete probability distributions.

2.2. Poisson Point Process

A Poisson point process (PPP) is a stochastic process whose samples represent sets of independent events on a measurable space $\mathcal{X}$. For a sample, the count of events in $A \subseteq \mathcal{X}$ is denoted $N(A) \in \mathcal{N}$. Events are considered to be a countable subset of $\mathcal{X}$, only significant if $\mathcal{X}$ is not countable, for instance the real line. The PPP has complete independence, so for $A, B \subset \mathcal{X}$, if $A \cap B = \emptyset$ then $N(A) \perp\!\!\!\perp N(B)$ and $N(A \cup B) = N(A) + N(B)$. The sample is specified by a rate $\rho(\mathrm{d}x)$ which is any measure on $\mathcal{X}$. In PPP theory, the rate is referred to as a Lévy measure. The PPP has the defining property that $N(A) \sim \text{Poisson}(\rho(A))$, and samples can be generated from this by working with an ever finer partition of the space $\mathcal{X}$.

A special class of PPP can be used as a family of priors for a CRM. Assume a PPP has rate $\rho(\mathrm{d}\lambda)\mu(\mathrm{d}x)$ for $\lambda \in \mathcal{R}^+$ and $x \in \mathcal{X}$. This is called homogeneous because the terms in λ and x are independent [18]. In the case considered here, the $\mu(\mathrm{d}x)$ is a measure on $\mathcal{X}$ called a base measure, and the rate $\rho(\mathrm{d}\lambda)$ has the condition $\int_0^\infty \min(1, \lambda)\rho(\mathrm{d}\lambda) < \infty$ to make everything work neatly [20], as follows: This condition is equivalent to $\int_0^\infty \min(\epsilon, \lambda)\rho(\mathrm{d}\lambda) < \infty$ for any $0 < \epsilon < \infty$. As a consequence, $\rho([\epsilon, \infty))$ is bounded, meaning there will be a finite number of points with $\lambda > \epsilon$ in the sample of the PPP (within a finitely measured subset of $\mathcal{X}$) and $\int_0^\epsilon \lambda\rho(\mathrm{d}\lambda)$ is bounded, meaning the sum of the λ's less the ϵ in the sample of the PPP (within a finitely measured subset of $\mathcal{X}$) will be finite even if there is an infinite number of them. Then, a sample from the PPP is a countable set of points which can be used to constuct a CRM.

2.3. Example Processes

Consider a number of standard PPPs used to construct CRMs [21]: the generalised (three-parameter) beta process [25], the generalised (three-parameter) gamma process [26] and the stable process. These have the forms given in Table 1, where M is a constant background rate. They are given without specifying a base measure on $\mathcal{X}$, which could be given as a final parameter.

Table 1. General processes. Marginal is the corresponding infinitely divisible distribution for the total rate developed, for instance, using Theorem 1.

Name	Domain	Parameters	Rate (Lévy Measure)	Marginal
beP(M, α, β)	$0 < \lambda < 1$	$0 \le \alpha < 1, \beta > 0$	$M\frac{\lambda^{-\alpha-1}}{\Gamma(1-\alpha)}(1-\lambda)^{\alpha+\beta-1}$	for $\alpha = 0, \beta = 1$: Dickman(M)
GP(M, β)	$\lambda > 0$	$\beta > 0$	$M\lambda^{-1}e^{-\lambda\beta}$	gamma(M, β)
GGP(M, α, β)	$\lambda > 0$	$0 < \alpha < 1, \beta > 0$	$M\frac{\alpha}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}e^{-\lambda\beta}$	Twe$(\alpha, M^{1/\alpha}, \beta)$
staP(M, α)	$\lambda > 0$	$\alpha > 0$	$\frac{M\alpha}{\Gamma(1-\alpha)}\lambda^{-\alpha-1}$	pstable$(\alpha, M^{1/\alpha})$
PP(M)	$\lambda = 1$		M	Poisson(M)
NBP(M, ρ)	$\lambda \in \mathcal{N}^+$	$0 < \rho < 1$	$M\frac{-\rho^\lambda}{\lambda\log(1-\rho)}$	NB(M, ρ)

The Poisson process and the negative binomial process [11] are also included in Table 1. Both are used in the hierarchical context in Section 4.

The first three processes in Table 1 are widely used in various forms in the non-parametric Bayesian and machine learning communities. From a Bayesian perspective, they are best thought of as improper priors corresponding to the beta, gamma and gamma distributions, respectively. This analysis is presented later in Section 3.4.

NRMIs can be created by normalising CRMs. These are sometimes generated directly from distributions consisting of a normalised discrete set of weights as probabilities. So, generating the $\vec{\lambda}$ according to a generalised (or three parameter) gamma process, GGP(M, α, β), and then normalising yields, what is called a normalised generalised gamma process

(NGG). The normalised generalised gamma process (NGG) is constructed analogously to the Dirichlet process, which normalises the gamma process. They represent the main examples of NRMIs. These NRMIs, however, are not paired with base measures when forming a discrete process on $\mathcal{X}$, rather they need to be paired with base distributions $\Pr(x)$ since only one point is generated per sample. Denote the NGG process as $\text{NGG}(\alpha, \beta, M)$ or $\text{NGG}(\alpha, \beta, M, h(\cdot))$, where α, β, M are as described for the GGP, line 3 of Table 1, and $h(\cdot)$ is a base distribution. The DP is effectively the case when $\alpha = 0$.

Traditionally, the parameter vector part of the DP in Equation (1) is called a GEM distribution (specifically, when a size-biased order is used [27]), named after Griffiths, Engen and McCloskey [28]. This can be represented as an infinite vector $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots)$. Correspondingly, there is a two-parameter version of $\vec{\lambda}$ corresponding to the PYP, $\text{GEM}(\alpha, \beta)$, which has discount $0 \leq \alpha < 1$ and concentration $\beta > -\alpha$. Then, $\text{GEM}(0, \beta)$ is the original GEM. Including the base distribution $h(\cdot)$ yields $\text{DP}(\beta, h(\cdot))$ and $\text{PYP}(\alpha, \beta, h(\cdot))$.

The Pitman–Yor process itself was developed by Pitman and Yor [28], and a general scheme for developing related models is by Pitman [29], called Poisson–Kingman models. However, as to be shown, the hierarchical PYP is very different from the PYP, so this theory is not entirely relevant for the hierarchical case. Alternatively, in Pitman and Yor [28] ([Proposition 21]), it was shown that a PYP can be developed by marginalising out a parameter of the NGG as follows.

Lemma 1. (Deriving a PYP from a NGG) *Let* $\mu(x) \sim NGG(\alpha, M, h(\cdot))$ *for* $\alpha, M > 0$ *and suppose* $M \sim gamma(\beta/\alpha, 1)$ *for* $\beta > 0$*, then it follows that* $\mu(x) \sim PDP(\alpha, \beta, h(\cdot))$.

The result is presented rather indirectly in Pitman and has been re-expressed by several authors [23] ([Section 3.1.1]), [30] ([Corollary 1]), and leads to a different class of models to the Poisson–Kingman models called Poisson-gamma models [23].

Notice the lemma restricts the PYP to the case where the concentration is positive. More generally, PYPs can have concentration $\beta > -\alpha$. When $\beta = 0$ and $\alpha > 0$, then the PYP is formed from normalising a positive stable distribution.

3. Defining Processes Axiomatically

This section gathers together some definitions and theory in order to present a general class of processes built on CRMs that can be treated hierarchically analogous to the Dirichlet process.

3.1. Subordinators

A simple useful case of these PPPs has the domain $\mathcal{X}$ being $\mathcal{R}^+$, the positive real line, and is constant for $\mathcal{X}$, so the rate is $\rho(\mathrm{d}\lambda)$ for $\lambda, x \in \mathcal{R}^+$. For this, define a new process for our case $C_0 = 0$ given by the cumulative values,

$$\sigma_t = \mu((0, t]) = \sum_{i=1}^{\infty} \lambda_i \delta_{x_i \leq t}$$

So, $\sigma_0 = 0$ and σ_t increases in steps as each distinct x_i is passed. This σ_t corresponds to the class of so-called pure jump driftless subordinators, which are a kind of nondecreasing Lévy process, which in turn are processes with stationary independent increments [20]. The key relationship that underlies the general theory of these processes is that σ_t is distributed according to a particular infinitely divisible non-negative distribution, explained in Theorem 1. Examples are given in Table 1. So, for instance, for the generalised gamma process with parameters (M, α, β), the total $\sigma_1 = \sum_{i=1}^{\infty} \lambda_i \delta_{x_i \leq 1}$ is distributed as a Tweedie distribution with parameters $(\alpha, M^{1/\alpha}, \beta)$.

The basic connection is given as follows, a special case of the Lévy–Khintchine formula for subordinators. This uses the Laplace exponent of a 1D random variable y defined as the function (of u) $\mathcal{E}[e^{-uy}]$, which is related to the characteristic function.

Theorem 1. *Consider σ_t defined as previously by a PPP with rate $\rho(d\lambda)$ for $\lambda, x \in \mathcal{R}^+$ and $\rho(d\lambda)$ satisfying $\int_0^\infty \min(1,\lambda)\rho(d\lambda) < \infty$. The Laplace exponent of σ_t is given by*

$$\mathcal{E}[e^{-u\sigma_t}] = e^{-t\psi(u)}$$

where $\psi(u) = \int_{(0,\infty)}(1-e^{u\lambda})\rho(d\lambda)$. This form means that σ_t has an infinitely divisible non-negative distribution. The t here can be referred to as the parameter for divisibility, occurring in any infinitely divisible distribution.

Thus, given a rate $\rho(\mathrm{d}\lambda)$ defining a particular σ_t, one can derive its Laplace exponent $\psi(u)$ and then infer the distribution on σ_t (where analytically possible). Note the scaling term M in Table 1 plays the role of t.

Some instances of this pairing, an infinitely divisible non-negative distribution with a corresponding rate are given in the last two columns of Table 1. Note that distributions corresponding to the generalised beta process are not well-known. Other distributions that could be included in the table are the inverse beta distribution (the beta distribution is not infinitely divisible but its inverse is), which includes the Pareto and F-distributions, and the generalised inverse gamma distribution [31].

3.2. Axiomatic Definitions

To extend Theorem 1 to broader classes of base distributions on general domains $\mathcal{X}$, not just the positive real line with constant measure used in subordinators, one can give an axiomatic definition of a process based on an infinitely divisible non-negative distribution:

1. The derived process is a CRM,
2. The process behaves like the given infinitely divisible distribution on subsets of $\mathcal{X}$.

Definition 1. (Axiomatic definition of a CRM process) *Consider an infinitely divisible non-negative distribution $\mathrm{G}(\mu)$, where μ is the parameter for divisibility. Further assume its Laplace exponent has zero drift. Given a measurable space $\mathcal{X}$, positive intensity M and measure $h(dx)$ on $\mathcal{X}$, consider a stochastic process denoted $\mathrm{GP}(M, h(\cdot))$ induced by $\mathrm{G}(\mu)$ as follows. $X \sim \mathrm{GP}(M, h(\cdot))$ yields a CRM on $\mathcal{X}$ such that*

1. *For $A, B \subset \mathcal{X}$, if $A \cap B = \emptyset$ then $X(A) \perp\!\!\!\perp X(B)$,*
2. *For $A \subseteq \mathcal{X}$, $X(A) \sim \mathrm{G}(M\,h(A))$.*

The first condition implies that the measures are CRMs as per Equation (1). The second condition implies one can construct the discrete measures iteratively, on an ever finer, nested sequence of partitions using the distribution $G()$. Alternatively, one can use the Lévy–Khintchine formula of Theorem 1 to show the existence of a corresponding rate yielding a CRM with rate $M\,h(\mathrm{d}\,x)\rho(\mathrm{d}\,\lambda)$ which must then satisfy the conditions.

Note that the Dirichlet process can be defined axiomatically [12], akin to Definition 1 with the Dirichlet distribution used instead of the gamma distribution, and base probability distribution used instead of a base measure. This axiomatic construction generalises for any infinitely divisible non-negative distribution as follows:

Definition 2. (Axiomatic definition of an NRMI process) *Consider an infinitely divisible non-negative distribution $\mathrm{G}(\mu)$, where μ is the parameter for divisibility. Further assume its Laplace exponent has zero drift. Consider as well the distribution on probability vectors induced by generating K values $\zeta_k \sim \mathrm{G}(\mu_k)$ and normalising to obtain*

$$\left(\frac{\zeta_1}{\sum_{k=1}^K \zeta_k}, \cdots, \frac{\zeta_K}{\sum_{k=1}^K \zeta_k}\right).$$

Denote this distribution by $\mathrm{NG}_K(\vec{\mu})$, where $\vec{\mu}$ is the vector of K value μ_k, given a measurable space $\mathcal{X}$, positive intensity M and probability distribution $h(dx)$ on $\mathcal{X}$. A process denoted

$NGP(M, h(\cdot))$, developed from $G(\mu)$, is defined as follows. It is a stochastic process whose sample is a probability measure on $\mathcal{X}$ such that if $C \sim NGP(M, h(\cdot))$ then for any finite partition $A_1, \ldots, A_K$ of $\mathcal{X}$, and count $N > 0$, $(C(A_1), \ldots, C(A_K)) \sim$ multinomial$(N, NG_K(Mh(A_1), \ldots, Mh(A_K)))$.

In this way, a multinomial process can be defined axiomatically, as done by Zhou and Carin [11] [Corollary IV2]. One uses $\mathrm{MP}(N, h(\cdot))$ where $N \in \mathcal{N}^+$ is the total count and $h(\cdot)$ a probability measure. The axiomatic part is $(X(A_1), \ldots, X(A_K)) \sim$ multinomial$(N, (h(A_1), \ldots, h(A_K)))$. Similarly, a Dirichlet compound multinomial (DCM) process can be defined, denoted as $\mathrm{DCMP}(N, h(\cdot))$, where the axiomatic part is $(X(A_1), \ldots, X(A_K)) \sim \mathrm{DCM}(N, (h(A_1), \ldots, h(A_K)))$. These correspond to a PPP and a NBP, respectively, both given in Table 1, where one has also conditioned on the total count being N.

3.3. On the Tweedie Distribution

From Table 1, the marginal distribution for the generalised gamma process is the Tweedie distribution [32] with exponent α, or sometimes expressed as index $p = 1 + \frac{1}{1-\alpha}$ which has $p > 2$ necessarily. For $\alpha = 0$, the Tweedie distribution becomes a gamma distribution.

The Tweedie distribution with exponent $0 < \alpha < 1$ is formed from a positive stable distribution defined in terms of the stable distribution with characteristic exponent α, scale parameter $s = M^{1/\alpha}$ location zero and symmetry one [33]. This distribution, denoted as $\mathrm{pstable}(\alpha, s)$, has the functional form [adding a scale to the standard formula of] [34] given by the remarkable formula

$$\begin{aligned} \Pr(x \mid \mathrm{pstable}(\alpha, s)) &= \frac{\alpha}{1-\alpha} \frac{1}{s\pi} (x/s)^{-\frac{1}{1-\alpha}} \int_0^\pi a_\alpha(\nu) e^{-(x/s)^{-\frac{\alpha}{1-\alpha}} a_\alpha(\nu)} \mathrm{d}\,\nu \\ \text{where } a_\alpha(\nu) &= \frac{\sin((1-\alpha)\nu)(\sin(\alpha\nu))^{\alpha/(1-\alpha)}}{\sin(\nu))^{1/(1-\alpha)}}, \end{aligned}$$

which yields a simple ingenious sampling formula [34]. To obtain a Tweedie distribution, "exponentially tilt" the $\mathrm{pstable}(\alpha, s)$, calculated by multiplying by $e^{-\beta x}$ and renormalising. The construction of exponentially tilting the distribution (see for instance Pitman [29]) gives the following:

$$\Pr(x \mid \mathrm{Twe}(\alpha, s, \beta)) = e^{(s\beta)^\alpha - \beta x} \Pr(x \mid \mathrm{pstable}(\alpha, s)) .$$

Here, the term $e^{(s\beta)^\alpha - \beta x}$ is added to achieve normalisation.

3.4. Bayesian Analysis

A complete Bayesian analysis of CRMs and NRMIs has been developed by James [16] and James et al. [18], respectively, in the non-hierarchical context. This models the standard framework in which hierarchical DPs or hierarchical PYPs are used, but also applies to the Indian buffet process [17]. This is informally developed below so that their theoretical results can be subsequently used. By Bayesian analysis, the following is meant: one has an infinitely divisible distribution suitable for use with Theorem 1. One samples a CRM from this with unknown parameters of rates $\vec{\lambda}$ and atoms x_i. Now, hierarchically sample sets of atoms from this CRM using a PPP. Each hierarchical sample from the CRM is a discrete set $A \subseteq \mathcal{X}$, and multiple samples are drawn. Then, the task is to estimate the parameters of the parent CRM.

A CRM is represented in the form $\mu(x) = \sum_{i=1}^\infty \lambda_i \delta_{x_i}(x)$ for $x \in \mathcal{X}$ where the x_i are distinct and is generated according to a homogeneous PPP with rate $\rho(\mathrm{d}\lambda)\omega(\mathrm{d}x)$ where $\rho(\mathrm{d}\lambda)$ is a rate satisfying the conditions of Theorem 1. One then takes J samples from this according to a PPP, so $\vec{n}_j \sim \mathrm{PPP}(\mu(\cdot))$ for $j = 1, \ldots, J$. Each sample will be a finite subset of the atoms, some possibly occurring multiple times. For representational purposes, post hoc reorder the atoms of $\mu(x)$ so that only the first I have non-zero counts. So, for $I < i \leq \infty$, none of the samples $\vec{n}_j$ contain x_i. The count of atom x_i in sample j is represented as $n_{j,i}$, so the condition $\vec{n}_{1:J,i} \neq \vec{0}$ means that at least one of the J samples contains an atom x_i.

The following informal analysis is offered as an explanation, but formal proofs are in James [16]. To make analysis feasible, we have to convert the rate $\rho(\mathrm{d}\lambda)$ to one with finite total measure. James [16] ingeniously and elegantly presents this by viewing the posterior for μ_i after seeing the evidence of having at least one non-zero value in the J values, so $\vec{n}_{1:J,i} \neq \vec{0}$. For the particular sampling distribution of $n_{j,i}$, in our case a Poisson(λ_i),

$$\Pr(\vec{n}_{1:J,i} \neq \vec{0} \mid \lambda_i) = 1 - e^{-J\lambda_i}$$

which has a term in λ_i so the posterior rate $\Pr(\vec{n}_{1:J,i} \neq \vec{0} \mid \lambda_i)\rho(\mathrm{d}\lambda_i)$ obtains finite total measure. Denote this total by $\Psi_J = \int \Pr(\vec{n}_{1:J,i} \neq \vec{0} \mid \lambda_i)\rho(\mathrm{d}\lambda_i)$. Then, working entirely with finite PPPs, one can compute the marginal. First, we generate the number of non-zero atoms I (for the given sample count J) by a Poisson and then generate the vector of counts for each atom $\vec{n}_{1:J,i}$, like so

$$\begin{aligned}
\Pr(\vec{n}_1, \ldots, \vec{n}_J \mid \rho(\mathrm{d}\lambda), \mathrm{PPP}) &= e^{-\Psi_J} \frac{\Psi_J^I}{I!} \prod_{i=1}^{I} \Pr(\vec{n}_{1:J,i} \mid \vec{n}_{1:J,i} \neq \vec{0}) \\
&= e^{-\Psi_J} \frac{\Psi_J^I}{I!} \prod_{i=1}^{I} \frac{\int \Pr(\vec{n}_{1:J,i} \mid \lambda)\rho(\mathrm{d}\lambda)}{\int \Pr(\vec{n}_{1:J,i} \neq \vec{0} \mid \lambda)\rho(\mathrm{d}\lambda)} \\
&= e^{-\Psi_J} \frac{1}{I!} \prod_{i=1}^{I} \int \Pr(\vec{n}_{1:J,i} \mid \lambda)\rho(\mathrm{d}\lambda)\,,
\end{aligned} \tag{2}$$

where the term $I!$ can be removed if one considers that the atoms are ordered. With similar reasoning, one obtains:

the posterior rate of $\boldsymbol{\lambda_i}$: for $i \leq I$ has rate $\Pr(\vec{n}_{1:J,i} \mid \lambda_i)\rho(\mathrm{d}\lambda_i)$,

the posterior rate of the remainder CRM: $\mu_R(x) = \sum_{i=I+1}^{\infty} \lambda_i \delta_{x_i}(x)$, has rate $\Pr(\vec{n}_{1:J} \neq \vec{0} \mid \lambda)\rho(\mathrm{d}\lambda)\omega(\mathrm{d}x)$,

the total rate of the remainder CRM: $T_R = \sum_{i=I+1}^{\infty} \lambda_i$ as given by Theorem 1.

The key formula for this kind of analysis is given in our context in Table 2.

Table 2. Key formula for posterior analysis of CRMs, $\Psi_J = \int \Pr(\vec{n}_{1:J} \neq \vec{0} \mid \lambda)\rho(\mathrm{d}\lambda)$, and the distribution on the remainder $T_R = \sum_{i=I+1}^{\infty} \lambda_i$.

Name	$\boldsymbol{\Psi_J}$	**Remainder $\boldsymbol{T_R}$**
beP(M, α, β)-BP	$M \sum_{j=0}^{J-1} \frac{\Gamma(\alpha+\beta+j)}{\Gamma(1+\beta+j)}$	$\mu_R \sim \mathrm{beP}(M, \alpha, J+\beta)$
GP(M, β)-PP	$M(\log(J+\beta) - \log\beta)$	gamma$(M, J+\beta)$
GGP(M, α, β)-PP	$M((J+\beta)^\alpha - \beta^\alpha)$	Twe$(\alpha, M^{1/\alpha}, J+\beta)$
GP(M, β)-NBP(ρ)	$M\Big(\log(J\log\frac{1}{1-\rho}+\beta) - \log\beta\Big)$	gamma$(M, J\log\frac{1}{1-\rho}+\beta)$
GGP(M, α, β)-NBP(ρ)	$M\Big((J\log\frac{1}{1-\rho}+\beta)^\alpha - \beta^\alpha\Big)$	Twe$(\alpha, M^{1/\alpha}, J\log\frac{1}{1-\rho}+\beta)$
staP(M, α)-PP	MJ^α	Twe$(\alpha, M^{1/\alpha}, J+\beta)$
staP(M, α)-NBP(ρ)	$M(J\log\frac{1}{1-\rho})^\alpha$	Twe$(\alpha, M^{1/\alpha}, J\log\frac{1}{1-\rho})$

The first line, the beP-BP case, is the three parameter beta process with Bernoulli data, which is the three parameter Indian buffet process. The second line is the gamma process

with Poisson data. Note the data marginals $\int \Pr(\vec{n}_{1:J,i} \mid \lambda)\rho(\mathrm{d}\lambda)$ in our context can be obtained more directly, developed in Section 4.2, so formulas are not given.

4. Using Discrete Base Distributions

It is important to understand what happens when you use a discrete distribution as a base distribution to a CRM, since this is what happens when hierarchical constructions of these processes are made. Let the base measure on $\mathcal{X}$ have the form $\mu(x) = \sum_{i=1}^{\infty} \lambda_i \delta_{x_i}(x)$, and the CRM is constructed using a homogeneous PPP with rate $\rho(\mathrm{d}\lambda)\omega(\mathrm{d}x)$. What happens? This section considers various implications of this. Note different but more extensive treatment of this scenario for the results on moments, Section 4.2, and the generalised Chinese restaurant process, Section 4.4, is given by Camerlenghi et al. [14], Argiento et al. [15]. They also include example MCMC sampling algorithms.

4.1. General Results

Superposition of PPPs says to decompose a discrete CRM into a union of trivial PPPs each with rate in the form $\mu_i\rho(\lambda)\delta_{x_i}$, so the $\mathcal{X}$ component is a delta function. The resultant CRM is also trivial and takes the form, using Definition 1, $\Lambda\delta_{x_i}$, where Λ is the total of the λ_k generated using the rate $\mu_i\rho(\lambda)$. This total is distributed as the corresponding marginal distribution for the subordinator with intensity parameter μ_i, as per Theorem 1.

Lemma 2. (CRM when base measure is discrete) *Let a discrete measure on $\mathcal{X}$ have the form $\mu(x) = \sum_{i=1}^{\infty} \mu_i \delta_{x_i}(x)$ for $x \in \mathcal{X}$ where the x_i are distinct, and a homogeneous CRM is constructed by sampling using a PPP with rate $\rho(d\lambda)\mu(dx)$ on $\mathcal{R}^+ \times \mathcal{X}$. Let $\Gamma(t)$ be the marginal total distribution for the corresponding subordinator, where t is the parameter of divisability. Then, the CRM has the form*

$$\gamma(x) = \sum_{i=1}^{\infty} \gamma_i \delta_{x_i}(x) \tag{3}$$

where the random variable $\gamma_i \sim \Gamma(\mu_i)$, and the x_i are inherited from $\mu(\cdot)$.

The CRM $\mu(\cdot)$ when used as a base distribution for a PPP is mapped element-wise to form a new CRM $\gamma(\cdot)$. So, no PPP modelling is required if you know the form of the element-wise distribution.

There are a number of very convenient and well-known properties of the Dirichlet that allow it to be used in hierarchical contexts. As it happens, most of these properties also hold for other NRMIs with discrete base measures, and some for CRMs, so these results are developed here. The first property is aggregation. This has that if $(x_1, x_2, x_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$, then $(x_1, x_2 + x_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2 + \alpha_3)$, and this applies for a Dirichlet of any dimension. The second property is renormalisation and has that if $(x_1, x_2, x_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ then $(x_1, x_2)/(x_1 + x_2) \sim \text{Dirichlet}(\alpha_1, \alpha_2)$. Both properties clearly follow from the fact that a Dirichlet is a normalised Gamma, and by analogy hold for NRMIs too.

Definition 3. (Aggregation property) *Consider a process that takes a measure as an input parameter and outputs another measure. The process has the aggregation property if when $\sum_{i=1}^{\infty} \gamma_i \delta_{x_i}(x)$ is a sample from the process with a discrete input measure $\sum_{i=1}^{\infty} \mu_i \delta_{x_i}(x)$ where the x_i are distinct, then $\sum_{i=3}^{\infty} \gamma_i \delta_{x_i}(x) + (\gamma_1 + \gamma_2)\delta_{x_1}(x)$ is a sample from the process with input measure $\sum_{i=3}^{\infty} \mu_i \delta_{x_i}(x) + (\mu_1 + \mu_2)\delta_{x_1}(x)$.*

The aggregation property can be used to form arbitrary groupings of the dimensions.

Definition 4. (Renormalisation property) *Consider a process that takes a measure as an input parameter and outputs a probability measure. The process has the renormalisation property if when $\sum_{i=1}^{\infty} \gamma_i \delta_{x_i}(x)$ is a sample from the process with a discrete input measure $\sum_{i=1}^{\infty} \mu_i \delta_{x_i}(x)$ where*

the x_i are distinct, then $\frac{1}{\sum_{i=2}^{\infty}\gamma_i}\sum_{i=2}^{\infty}\gamma_i\delta_{x_i}(x)$ is a sample from the process with input measure $\sum_{i=2}^{\infty}\mu_i\delta_{x_i}(x)$.

The renormalisation property then yields probability measures on subsets of the discrete domain, so it can be used for incremental sampling.

Lemma 3. (Aggregation and renormalisation) *Consider the context of Lemma 2. The aggregation property holds for all CRMs and NRMIs. In the case of an NRMI, the renormalisation property holds. For the PYP, the aggregation property holds but not the renormalisation property.*

The results for the PYP can be developed using Lemma 1. The aggregation and renormalisation properties together mean that efficient size-biased samplers can be developed for NRMIs by sampling one dimension at a time according to a two-dimensional version of the NRMI, which is effectively the stick breaking construction (although, only a few explicit cases of this are known). Alternatively, one can sample the underlying CRM according to its corresponding infinitely divisible distribution.

A third property of the Dirichlet is neutrality, which applies in the context of renormalisation and requires that the part taken away is independent of the remainder: if $(x_1, x_2, x_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$, then $(x_1, x_2)/(x_1 + x_2)$ is independent of x_3.

Definition 5. (Neutrality property) *Consider a process that outputs a finite discrete probability measure, and without loss of generality let $\sum_{i=1}^{I}\gamma_i\delta_{x_i}(x)$ be a sample from the process where the x_i are distinct. The process is completely neutral if there exists mutually independent non-negative variables $\lambda_1, \ldots, \lambda_I$ such that $(\gamma_1, \ldots, \gamma_K)$ and $\left(\lambda_1, \lambda_2(1-\lambda_1), \ldots, \lambda_I\prod_{i=1}^{I-1}(1-\lambda_i)\right)$ have the same distribution.*

It is known that the only distribution on finite probability vectors with complete neutrality is the Dirichlet distribution [35].

4.2. Results on Moments

Moments of CRMs are critical quantities for their posterior analysis [18,36] to be developed in Section 5 and seen in Section 3.4. The generalised version is derived by unfolding the recursion that relates the moments of a distribution to its cumulants. In the context of Lemma 2, where $\gamma_i \sim \Gamma(\mu_i)$, various moments such as $\mathcal{E}[\gamma_i^n \mid \mu_i]$ and $\mathcal{E}[\gamma_i^n e^{-U\gamma_i} \mid \mu_i]$ can be computed recursively from the moments of the PPP rate $\rho(\mathrm{d}\lambda)$ [22] ([Section 1.3]) and its exponentially titled form. Note these moments compute the marginals one needs for multinomial and Poisson data, respectively, hence their importance.

In the theorem, the notation $\mathcal{P}^n$ is used to represent all possible non-empty partitions of n items, the set $\{1, \ldots, n\}$. As an example, $\mathcal{P}^3$ is the set

$$\left\{ \{\{1,2,3\}\},\ \{\{1\},\{2,3\}\},\ \{\{1,2\},\{3\}\},\ \{\{1,3\},\{2\}\},\ \{\{1\},\{2\},\{3\}\} \right\},$$

so it contains the partition $\{\{1\},\{2,3\}\}$ as an element, for instance. Moreover, $\mathcal{P}_K^n \subseteq \mathcal{P}^n$ are all members are of size K, so $|\mathcal{P}_1^n| = |\mathcal{P}_n^n| = 1$ and $|\mathcal{P}_2^3| = 3$.

The following Lemma is a corollary the major result by Pitman [22], and some related results appear in Camerlenghi et al. [14], as proven in Appendix A.

Lemma 4. (CRM moments when base measure is discrete) *Consider the context of Lemma 2. Let $\kappa_n = \int_0^{\infty}\lambda^n\rho(d\lambda)$ be the n-th moment for rate $\rho(\lambda)$, where it exists for $n \in \mathcal{N}^+$. Let $\psi(t)$ be the Laplace exponent for the rate. Then, the n-th cumulant of γ_i can be re-expressed as a moment of the original rate $\rho(\lambda)$, and the n-th moment of γ_i is computed recursively from it.*

$$\begin{aligned}
\kappa_n &= (-1)^{n+1} \frac{d^n \psi(t)}{d t^n}\Big|_{t=0} && (4)\\
cumulant_n(\gamma_i) &= \mu_i \kappa_n && (5)\\
\mathcal{E}[\gamma_i^n \mid \mu_i] &= \sum_{\Pi \in \mathcal{P}^n} \mu_i^{|\Pi|} \prod_{C \in \Pi} \kappa_{|C|} && (6)\\
&= \sum_{K=1}^{n} \mu_i^K T_K^n && (7)\\
where\ T_K^n &= \sum_{\Pi \in \mathcal{P}_K^n} \prod_{C \in \Pi} \kappa_{|C|}\\
&= \sum_{k=1}^{n-K+1} T_{K-1}^{n-k} \binom{n-1}{k-1} \kappa_k \,. && (8)
\end{aligned}$$

Note the recursion for T_K^n starts at $T_1^n = \kappa_n$ derived from the non-recursive form.

Thus, if the Laplace exponent is known, one can usually compute the moments of the process and hence the cumulants and evidence terms for its corresponding infinitely divisible distribution. When one has Poisson data, required moments need to include an exponential term, as proven in Appendix B.

Corollary 1. (Adding an exponential term) *Consider the context of Lemma 4 with rate $\rho(\lambda)$. To obtain exponentiated moments of the form $\mathcal{E}[\gamma_i^n e^{-U\gamma_i} \mid \mu_i]$, complete the following steps.*

1. *Use rate $e^{-U\lambda}\rho(\lambda)$, and the Laplace exponent is given by $\psi(U+t) - \psi(U)$, so the corresponding moments are given by*

$$\kappa_{n,U} = (-1)^{n+1} \frac{d^n \psi(t)}{d t^n}\Big|_{t=U}$$

2. *Obtain the corresponding T_K^n using Equation (8) with the $\kappa_{n,U}$, denoted $T_{K,U}^n$.*
3. *Consequently,*

$$\mathcal{E}[\gamma_i^n e^{-U\gamma_i} \mid \mu_i] = e^{-\mu_i \psi(U)} \sum_{K=1}^{n} \mu_i^K T_{K,U}^n \,.$$

The components from Lemma 4 for the processes in Table 1 are given in Table 3. These appear in various places in the broader statistical literature. The Laplace exponent is usually computed using integration by parts. The form $S_{s,\alpha}^n$ is the second order generalised Stirling number used in PYP inference [1,37], a generalized Stirling number of type $(-1, -d, 0)$ [38]. It can be verified using its recursion [37] with Equation (8).

Table 3. Properties of processes.

Name	$\boldsymbol{\kappa_n}$	$\boldsymbol{\psi(t)}$	$\boldsymbol{T_K^n}$
beP(M, α, β) (for $\beta > 1 - \alpha$)	$M \frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(n+\beta)}$	$M \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)\alpha} \Big({}_1F_1(1-\alpha, \beta, t) - 1 + \frac{1}{\beta} {}_1F_1(1-\alpha, \beta+1, t) \Big)$	use Equation (8)
GP(M, β)	$M \frac{\Gamma(n)}{\beta^n}$	$M \log(1 + t/\beta)$	$\frac{M^K}{\beta^n} S_K^n$
GGP(M, α, β)	$M \frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)} \frac{\alpha}{\beta^{n-\alpha}}$	$M((\beta+t)^\alpha - \beta^\alpha)$	$\frac{(M\alpha\beta^\alpha)^K}{\beta^n} S_{K,\alpha}^n$
staP(M, α)	NA	Mt^α	NA

Note the general beta process has no simple analytic form for either $\psi(t)$ or its marginal distribution. Fortunately, is is difficult to envisage a situation where it would be used hierarchically.

4.3. The Gamma Process

Let us consider the simple example of a gamma process, $\mathrm{GP}(M, \beta)$ and assume data yields Poisson likelihoods in the form $\prod_{i=1}^{I} \gamma_i^{n_i} e^{-U\gamma_i}$ for dimensions $i = 1, \ldots, I$ in the context of Lemma 2. The marginal likelihood then, for the data $= \{\, n_i, x_i : i = 1, \ldots, I \,\}$ is given by

$$\Pr(data \mid \mu(\cdot)) = \mathcal{E}[e^{-U\gamma_R} \mid \mu_R] \prod_{i=1}^{I} \mathcal{E}[\gamma_i^{n_i} e^{-U\gamma_i} \mid \mu_i]$$

where the expectation is taken with respect to $\gamma(\cdot) \sim \mathrm{GP}(M, \beta, \mu(\cdot))$, which has $\gamma_i \sim \mathrm{gamma}(M\mu_i, \beta)$. Note, in this case, the exact solution is known since the data marginals of the gamma distribution have a simple closed form,

$$\mathcal{E}[\gamma_i^{n_i} e^{-U\gamma_i} \mid \mu_i] = \int_0^\infty \gamma^{n_i} e^{-U\gamma} \frac{\beta^{M\mu_i}}{\Gamma(M\mu_i)} \gamma^{M\mu_i - 1} e^{-\beta\gamma} \mathrm{d}\,\gamma = \frac{\Gamma(M\mu_i + n_i)}{(U+\beta)^{M\mu_i + n_i}} \frac{\beta^{M\mu_i}}{\Gamma(M\mu_i)} \tag{9}$$

Consider, however, using Corollary 1. In this case, moments including $e^{-U\gamma_i}$ are found to be $\kappa_n = \int_0^\infty \gamma^n e^{-U\gamma} \rho(\mathrm{d}\,\gamma) = M \frac{\Gamma(n)}{(U+\beta)^n}$, and the Laplace exponent can be obtained using integration by parts as $M \log(1 + t/\beta)$. One can confirm that the corresponding index $T_K^n = \frac{1}{(U+\beta)^n} S_K^n M^K$ where S_K^n is an unsigned Stirling number of the first kind, an index that is found in collapsed versions of the CRP. Equation (8) yields the standard recurrence for it. So, by Equation (7), and adding back the term $e^{-\mu_i \psi(U)} = \left(\frac{\beta}{U+\beta}\right)^{M\mu_i}$ as per Corollary 1, obtain for atom index i the moment

$$\Pr(\gamma_i^n e^{-U\gamma_i} \mid \mu_i) = \mathcal{E}[\gamma_i^n e^{-U\gamma_i} \mid \mu_i] = \frac{\beta^{M\mu_i}}{(U+\beta)^{M\mu_i + n_i}} \sum_{K=1}^{n_i} S_K^{n_i} (M\mu_i)^K . \tag{10}$$

The sum can be converted using a standard identity [37] ([Lemma 16]) to get back The sum in Equation (10) has an interpretation as a form of Chinese restaurant process for the dimension i. Each partition of the set $\{1, \ldots, n_i\}$, given by $\Pi_i \in \mathcal{P}^{n_i}$ corresponds to a configuration of the n_i data in $|\Pi_i|$ tables. For any table with participants $C \in \Pi_i$, the probability of the table is $M\mu_i \frac{\Gamma(|C|)}{(U+\beta)^{|C|}}$. The probability of this configuration Π_i is then $\prod_{C \in \Pi_i} M\mu_i \frac{\Gamma(|C|)}{(U+\beta)^{|C|}}$. So, introducing the partition Π_i or its size as an additional variable,

$$\begin{aligned} \Pr(\gamma_i^n e^{-U\gamma_i}, \Pi_i \mid \mu_i) &= \frac{\beta^{M\mu_i}}{(U+\beta)^{M\mu_i + n_i}} (M\mu_i)^{|\Pi_i|} \prod_{C \in \Pi_i} \Gamma(|C|) \\ \Pr(\gamma_i^n e^{-U\gamma_i}, \ |\Pi_i| = K \mid \mu_i) &= \frac{\beta^{M\mu_i}}{(U+\beta)^{M\mu_i + n_i}} S_K^{n_i} (M\mu_i)^K . \end{aligned}$$

The second form, the probability of all configurations of size $K (= |\Pi_i|)$, follows from Equation (8).

4.4. General Chinese Restaurant Processes

Motivated by the gamma process example just given, now construct a generalised CRP interpretation of the results in Section 4.2. The marginals have an interpretation as generalised versions of Chinese restaurants, including the more efficient collapsed versions [6], both developed in this section. This is intended to complement the comprehensive Bayesian analysis already developed for the non-hierarchical cases by [16,18].

The significance of the formula in Lemma 4 is that the sum in Equation (6) is over partitions Π of the n data points, and $\kappa_{|C|}$ represents the probability of generating a single element C of size $|C|$ (in the partition Π) according to the rate $\rho(\lambda)$. The sum in Equation (7) is now over partition sizes K, and T_K^n is the probability of generating a partition of K non-empty sets according to the rate $\rho(\lambda)$.

Lemma 5. (General Chinese restaurant processes for CRMs) *Consider the posterior data marginal for $\gamma(\cdot)$, as in Corollary 2, where data is in the form of a Poisson likelihood with counts $n_i > 0$ at each atom x_i:*

$$\Pr(\{n_i, x_i \,:\, i = 1, \ldots, I\} \mid \gamma(\cdot), U) = \prod_{i=1}^{I} \gamma_i^{n_i} e^{-U\gamma_i}$$

One can treat $\Pi_i \in \mathcal{P}^{n_i}$ as a latent variable, which represents the seating configuration for instances of the atom. Then, the data marginal using $\Pi_1, \ldots, \Pi_I$ takes the form:

$$\Pr(\{n_i, x_i, \Pi_i \,:\, i = 1, \ldots, I\} \mid \mu(\cdot), U) = e^{-\psi(U)\sum_{i=1}^{\infty}\mu_i} \prod_{i=1}^{I} \left(\mu_i^{|\Pi_i|} \prod_{C \in \Pi_i} \kappa_{|C_i|,U} \right). \tag{11}$$

Moreover, for any j (including $j > I$),

$$\Pr(x_j \mid \{n_i, x_i, \Pi_i \,:\, i = 1, \ldots, I\}, \mu(\cdot)) = \mu_j \kappa_{1,U} + \sum_{C \in \Pi_j} \frac{\kappa_{|C|+1,U}}{\kappa_{|C|,U}}, \tag{12}$$

where the convention is used that $\Pi_j = \emptyset$ for $j > I$ (when there is no data). Alternatively, if K_i, the number of tables for atom index i is handled as a latent variable, then the data marginal given table numbers takes the form:

$$\Pr(\{n_i, x_i, K_i \,:\, i = 1, \ldots, I\} \mid \mu(\cdot), U) = e^{-\psi(U)\sum_{i=1}^{\infty}\mu_i} \prod_{i=1}^{I} \mu_i^{K_i} \, T_{K_i,U}^{n_i}. \tag{13}$$

Equation (12) is related to the generalized Blackwell–MacQueen sampling scheme by James et al. [18] [Section 3.3]. The data marginals in Equations (11) and (13) have a simple Poisson likelihood in $\vec{\mu}$. Thus, a CRP interpretation of a Gamma process can be used for hierarchical inference with a Gamma distribution, as used by Zhou and Carin [11], for instance.

To develop a corresponding formula for NRMIs where they are generated by normalising a CRM, we use an ingenious technique for normalising a CRM within a posterior analysis from [18] The basic idea is to convert multinomial sampling into Poisson sampling (without normalisation) but require some post manipulation to derive the results. A generative variation of this goes as follows:

1. For each multinomial $\vec{n}$ according to the unnormalised values $\vec{\lambda}$, introduce a scale-free latent relative mass denoted U, with the scale-invariant improper prior $\frac{\mathrm{d}\,U}{U}$.
2. Generate the data needed according to Poisson $n_i \sim \text{Poisson}(U\lambda_i)$ for $i = 1, \ldots, \infty$, noting that $n_i = 0$ for $i > I$.
3. Then, the joint posterior on $\vec{n}, \vec{\lambda}, U$ becomes quite concentrated for U and can be marginalised out.
4. To correct the formulas, multiply the marginal by $N = \sum_{i=1}^{I} n_i$ to obtain a conversion to a multinomial.

To see that this indeed does what is required, one needs to verify the following identity.

$$N \int_{\mathcal{R}^+} \prod_{i=1}^{\infty} \left(e^{-U\lambda_i} \frac{(U\lambda_i)^{n_i}}{n_i!} \right) \frac{\mathrm{d}\,U}{U} = \binom{N}{\vec{n}} \prod_{i=1}^{I} \left(\frac{\lambda_i}{\sum_i \lambda_i} \right)^{n_i}.$$

Note the product $\prod_{i=1}^{\infty}$ is well-defined because $\sum_{i=1}^{\infty} \lambda_i$ is finite.

Corollary 2. (General Chinese restaurant processes for NRMIs) *Consider the posterior data marginal for $\gamma(\cdot)$ as given in Lemma 2, where data is in the form of a multinomial likelihood with counts $n_i > 0$ at each atom x_i:*

$$\Pr(\{n_i, x_i \,:\, i = 1, \ldots, I\} \mid \gamma(\cdot)) = \prod_{i=1}^{I} \left(\frac{\gamma_i}{\sum_{i=1}^{\infty} \gamma_i} \right)^{n_i},$$

and let $N = \sum_{i=1}^{I} n_i$ *be the total count. Let* $U \sim gamma(N, \sum_{i=1}^{\infty} \gamma_i)$*. Then, the data marginal using* $\Pi_1, \ldots, \Pi_I$*, similarly to Lemma 5, takes the form:*

$$\Pr(\{n_i, x_i, \Pi_i : i = 1, \ldots, I\}, U \mid \mu(\cdot)) = \frac{U^{N-1}}{\Gamma(N)} e^{-\psi(U)\sum_{i=1}^{\infty}\mu_i} \prod_{i=1}^{I} \left(\mu_i^{|\Pi_i|} \prod_{C \in \Pi_i} \kappa_{|C_i|,U} \right) . \quad (14)$$

Moreover, for any j (including $j > I$*),*

$$\Pr(x_j \mid \{n_i, x_i, \Pi_i : i = 1, \ldots, I\}, U, \mu(\cdot)) = \mu_j \kappa_{1,U} + \sum_{C \in \Pi_j} \frac{\kappa_{|C|+1,U}}{\kappa_{|C|,U}} . \quad (15)$$

Alternatively, if each K_i *is handled as a latent variable, then the data marginal given table numbers takes the form:*

$$\Pr(\{n_i, x_i, K_i : i = 1, \ldots, I\}, U \mid \mu(\cdot)) = \frac{U^{N-1}}{\Gamma(N)} e^{-\psi(U)\sum_{i=1}^{\infty}\mu_i} \prod_{i=1}^{I} \mu_i^{K_i} T_{K_i,U}^{n_i} . \quad (16)$$

Note, to complete the analysis, one needs to model the unseen parts of the processes. So, while it is assumed μ_i for $i = 1, \ldots, I$ is being sampled or estimated, of μ_i and γ_i for $i = I+1, \ldots, \infty$ only a finite number, if any, can be sampled or estimated. Handling these is illustrated in Section 5 using a remainder term $\mu_R = \sum_{j=I+1}^{\infty} \mu_j$.

In general, then, there are two different levels of inference one can use when the marginal does not have a simple closed form and must instead be computed using the latent forms in Lemma 5 or Corollary 2:

Sampling over table configurations:

For the DP, this is exhibited by the standard CRP. One can see from Equations (6) and (12) that to resample which table a point belongs to, one would use the following proportionalities:

$$\Pr(C \mid \Pi, \mu_k, \ldots) \propto \begin{cases} \mu_k \kappa_1 & \text{start a new table} \\ \frac{\kappa_{|C|+1}}{\kappa_{|C|}} & \text{add to table } C \end{cases} . \quad (17)$$

Sampling over table sizes:

For the PYP, this is demonstrated by table indicator sampling methods [6,39] and "direct" Gibbs sampling of Gasthaus and Teh [5], though subsequently not used because in their context they needed to constantly resample discount α. This is a collapsed sampler that instead samples K, the number of tables using Equation (7):

$$\Pr(K \mid \mu_k, \ldots) \propto \mu_k^K T_K^n \quad (18)$$

This is only efficient when T_K^n can be tabulated. In the general case, this requires $O(n^2 K)$ steps to follow using Equation (8) and $O(nK)$ for cases such as the gamma process above where a simpler double recursion is available for T_K^n since they are generalised second-order Stirling numbers.

5. Variants of the Generalised Gamma Process

In this section, we develop both the CRM and NRMI variants of the generalised gamma process in the hierarchical context. Using the generalised gamma process in an NRMI yields an NGG or a PYP. When the NGG process and the PYP are supplied discrete base distributions as input, they behave analogously to the Dirichlet distribution, as illustrated with Lemma 3. In this discrete context, refer to the corresponding distributions as the NGG distribution and the Pitman–Yor distribution (PYD). Here analytical forms of the PY distribution are developed.

5.1. The Hierarchical Context

Consider an NRMI in the context of the base distribution $\mu(x)$, as before. Suppose multinomial type data is observed in the form of counts n_k associated with the atoms x_k of $\mu(\cdot)$ for $k = 1, \ldots, K$, with total count $N = \sum_{k=1}^{K} n_k$, where all others are zero. The latent relative mass trick of James et al. [18] can be used to include U as a latent variable in the likelihood for the NGG and the PYD. Setting $U = 1$ and dividing by N in this case restores the posterior to the original Poisson version. The likelihood for a PYD also includes M (via Lemma 1). To express this, the remainder terms for both the base distribution and the CRM need to be represented.

$$\begin{aligned} \lambda_R &= \sum_{k=K+1}^{\infty} \lambda_k = \Lambda - \sum_{k=1}^{K} \lambda_k \\ \mu_R &= \sum_{k=K+1}^{\infty} \mu_k \,. \end{aligned}$$

The joint posterior for the NGG is now

$$\begin{aligned} &\Pr(\{\lambda_k, n_k, x_k \,:\, k = 1, \ldots, K\}, U, \lambda_R \mid \mathrm{GGP}, M, \alpha, \beta, N, \mu(\cdot)) \\ &= \frac{1}{N^{1_{U\neq 1}}\Gamma(N)} e^{-U\Lambda} U^{N-1} \;\Pr(\lambda_R \mid \mathrm{Twe}(\alpha, (M\mu_R)^{1/\alpha}, 1)) \\ &\quad \prod_{k=1}^{K} \lambda_k^{n_k} \Pr(\lambda_k \mid \mathrm{Twe}(\alpha, (M\mu_k)^{1/\alpha}, 1)) \\ &= \frac{1}{\Gamma(N)} e^{-M((1+U)^{\alpha}-1)} U^{N-1} \;\Pr(\lambda_R \mid \mathrm{Twe}(\alpha, (M\mu_R)^{1/\alpha}, 1+U)) \\ &\quad \prod_{k=1}^{K} \lambda_k^{n_k} \Pr(\lambda_k \mid \mathrm{Twe}(\alpha, (M\mu_k)^{1/\alpha}, 1+U)) \,, \end{aligned} \tag{19}$$

where the second line is obtained by applying the exponential tilting formula. Note, Lemma 2 means element-wise application of a distribution to the parameter vector $\vec{\mu}$ inside $\mu(\cdot)$. Forms for the PYD are obtained by adding the prior for M. For the normalised stable process, denoted NSP, one obtains

$$\begin{aligned} &\Pr(\{\lambda_k, n_k, x_k \,:\, k = 1, \ldots, K\}, U, \lambda_R \mid \mathrm{NSP}, M, \alpha, N, \mu(\cdot)) \\ &= \frac{1}{\Gamma(N)} e^{-U\Lambda} U^{N-1} \;\Pr(\lambda_R \mid \mathrm{pstable}(\alpha, (M\mu_R)^{1/\alpha})) \\ &\quad \prod_{k=1}^{K} \lambda_k^{n_k} \Pr(\lambda_k \mid \mathrm{pstable}(\alpha, (M\mu_k)^{1/\alpha})) \\ &= \frac{1}{\Gamma(N)} e^{-MU^{\alpha}} U^{N-1} \;\Pr(\lambda_R \mid \mathrm{Twe}(\alpha, (M\mu_R)^{1/\alpha}, U)) \\ &\quad \prod_{k=1}^{K} \lambda_k^{n_k} \Pr(\lambda_k \mid \mathrm{Twe}(\alpha, (M\mu_k)^{1/\alpha}, U)) \,. \end{aligned} \tag{20}$$

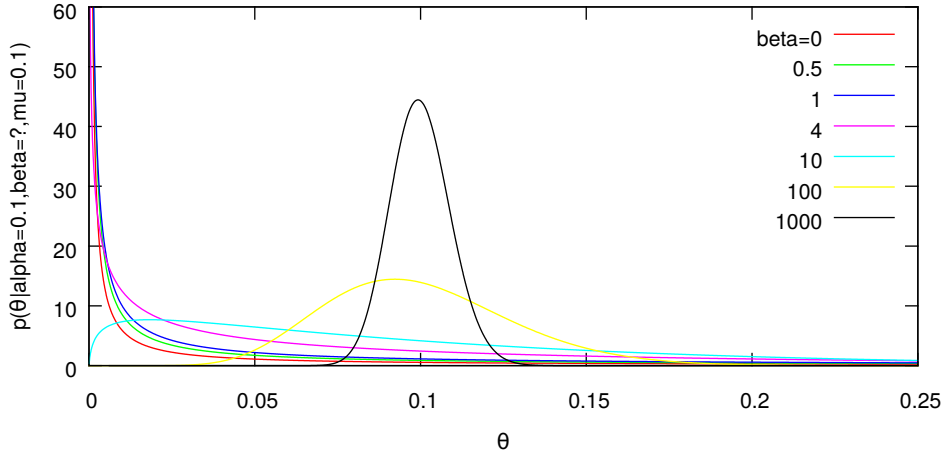
From this, one can derive an integral formula for the PYD. Details are in Appendix C, and the result is original.

Lemma 6. (Integral formula for the PY distribution) *Let $\vec{\mu}$ be a K-dimensional non-zero probability vector. Then, consider $\vec{\theta} \sim PYD(\alpha, \beta, \vec{\mu})$ for $\alpha > 0$ and $\beta \geq 0$. To express the probability of $\vec{\theta}$, introduce corresponding latent variables $\vec{v} = (v_1, \ldots, v_K) \in [0, \pi]^K$:*

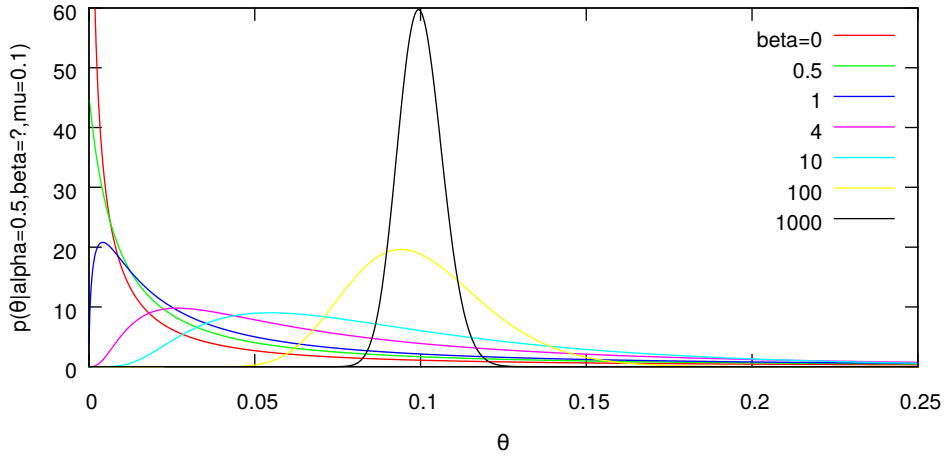
$$\Pr(\vec{\theta} \mid PYD, \alpha, \beta, \vec{\mu}) = \int_{[0,\pi]^K} \frac{\alpha^{K-1}\Gamma(1+\beta)}{(1-\alpha)^{K-1}\pi^K\Gamma(1+\beta/\alpha)} \frac{\Gamma(K+\beta(1-\alpha)/\alpha)\prod_{k=1}^{K} a_\alpha(v_k)\left(\frac{\mu_k}{\theta_k}\right)^{1/(1-\alpha)}}{\left(\sum_{k=1}^{K} a_\alpha(v_k)\left(\frac{\mu_k}{\theta_k}\right)^{1/(1-\alpha)}\theta_k\right)^{K+\beta(1-\alpha)/\alpha}} d\vec{v} \,. \tag{21}$$

This can be readily evaluated using numerical integration for small K. Plots of the marginal for θ_1 for different parameter settings are given in Figures 1 and 2.

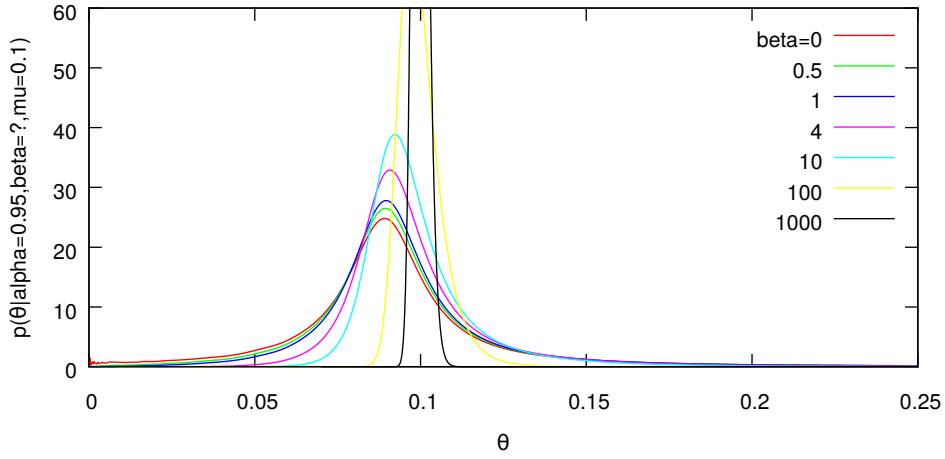
Due to the aggregation property of the PYD, these are representative marginals of the distribution for all dimensions. One can see the distributions becoming increasingly skewed as α increases.



(**a**) PDFs for $\alpha = 0.1$.

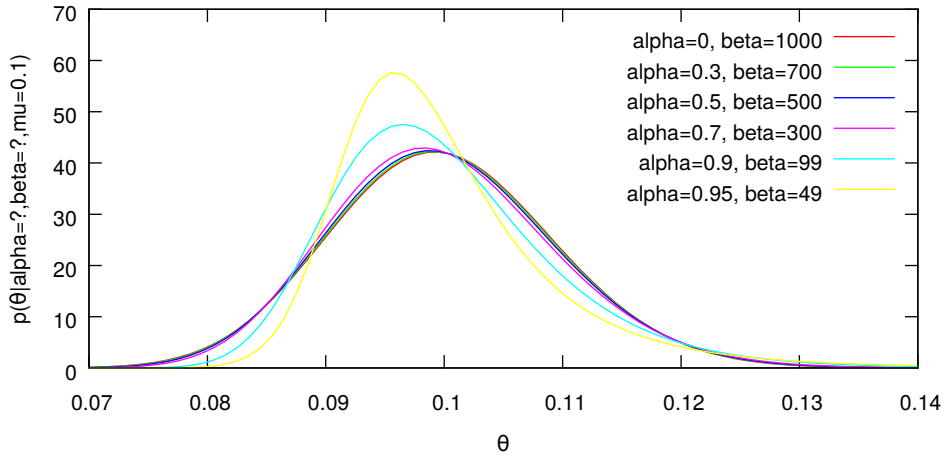


(**b**) PDFs for $\alpha = 0.5$.

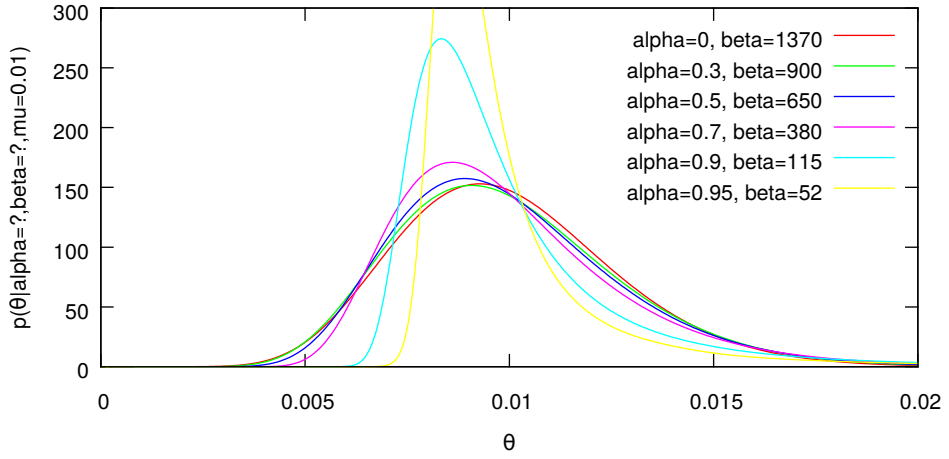


(**c**) PDFs for $\alpha = 0.95$.

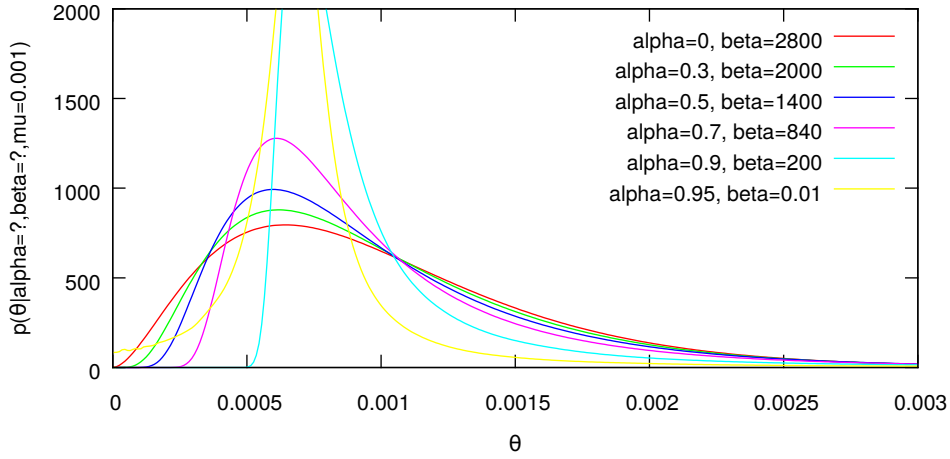
Figure 1. PDFs from Lemma 6 for location $\mu_1 = 0.1$ and fixed α.



(**a**) Variations with location $\mu_1 = 0.1$.



(**b**) Variations with location $\mu_1 = 0.01$.



(**c**) Variations with location $\mu_1 = 0.001$.

Figure 2. PDFs from Lemma 6 for variations with identical location and variance.

6. Networks of Processes

The next natural question to consider is how the above results apply to networks of processes. Several general schemes have been developed for inference in more general networks [3,6,11,19,39,40]. General networks for HPYPs have been demonstrated to scale [4,6], in contrast to earlier Gibbs schemes [3,40], and arguably the HGP has advantages over the HDP [11]. This section is a review of related material with regards to hierarchical processes.

6.1. Identifiability

One important question is the issue of statistical identifiability, and an underlying issue here is whether the parametric structure admits a unique representation [41]. In our case, some simple classes of non-uniqueness are easily identified and avoided. For instance, in Poisson matrix factorisation, if the matrix entry $x_{i,j} \sim \text{Poisson}\left(\sum_{k=1}^{K} \theta_{i,k}\phi_{k,j}\right)$, then one can insist that the scale of one of the matrices $\vec{\Theta}$ or $\vec{\Phi}$ (comprising the entries $\theta_{i,k}$ and $\phi_{k,j}$ respectively) needs to be anchored somehow so that the scale of the Poisson parameter is uniquely determined by just the other one. So, the rows of one of the matrices should normalise.

6.2. Equivalences

Another issue is that in some cases, networks can be transformed from one case to another. For instance, Zhou and Carin [11] ([Section VB]) show that a Poisson gamma-gamma process construction is equivalent to a HDP construction with an independent Poisson-gamma on the total. Given that there are significant differences between the corresponding algorithms in this case, and there are many more in the literature, what other equivalences are there?

Normalising processes are conducted to convert a CRM into an NRMI and in some cases, independence between the parts yields an equivalence between the CRM form and the NRMI form augmented with a total. This has major implications to networks of such processes, presented in the following subsection, so the results are summarised here.

The first results are on discrete processes and are well-known, some for instance reproduced by Zhou and Carin [11].

Lemma 7. (Equivalent processes) *Let $\vec{\mu}$ be a probability vector (possibly infinite), and M be a constant positive background rate. Let $X = \sum_{i=1}^{\infty} x_i$, the sum of entries of the non-negative integer vector $\vec{x}$. The following equivalences between (A) and (B) hold:*

- *Conditioning the PPP,*

$$(A)\ \vec{x} \sim PP(M\vec{\mu}) \quad (B)\ X \sim Poisson(M) \text{ and } \vec{x} \sim MP(X, \vec{\mu})\ .$$

- *NBP as a Poisson-gamma mixture,*

$$(A)\ \vec{x} \sim NBP(M, \rho, \vec{\mu}) \quad (B)\ \vec{x} \sim PP\left(GP\left(M, \frac{1-\rho}{\rho}, \vec{\mu}\right)\right)\ .$$

- *DCMP, given $X \in \mathcal{N}^+$, as a multinomial-Dirichlet mixture,*

$$(A)\ \vec{x} \sim DCMP(X, M\vec{\mu}) \quad (B)\ \vec{x} \sim MP(X, DP(M\vec{\mu}))\ .$$

- *Conditioning the NBP,*

$$(A)\ \vec{x} \sim NBP(M, \rho, \vec{\mu}) \quad (B)\ X \sim NB(M, \rho) \text{ and } \vec{x} \sim DCMP(X, M\vec{\mu})\ .$$

The conditioned versions of the PPP and NBP are used to decompose a likelihood into a total count and the vector of counts for atoms, given the total. Notice, while the conditioned version of the PPP yields a likelihood where the normalised measure ($\vec{\mu}$) and its total (M) are independent, the same does not hold for the conditioned NBP.

6.3. Normalisation and Independence

On non-discrete processes, some independences apply.

Lemma 8. (Normalised processes and independence) *Let $\Lambda = \sum_{i=1}^{\infty} \lambda_i$, the sum of entries of the infinite non-negative real vector $\vec{\lambda}$. The following two pairs (A) and (B) are equivalent:*

- *For the gamma process:*

(*A*) $\vec{\lambda} \sim GP(M, \beta)$;
(*B*) $\Lambda \sim ga(M, \beta)$ *and* $\vec{\lambda}/\Lambda \sim GEM(0, M)$, *where* Λ *and* $\vec{\lambda}/\Lambda$ *are independent.*

- *For the generalised gamma process where* $0 < \alpha < 1$, *marginalising* M
  (*A*) $\vec{\lambda} \sim GGP(M, \alpha, \beta)$ *where* $M \sim ga(\delta/\alpha, \beta^{\alpha})$;
  (*B*) $\Lambda \sim ga(\delta, \beta)$ *and* $\vec{\lambda}/\Lambda \sim GEM(\alpha, \delta)$, *where* Λ *and* $\vec{\lambda}/\Lambda$ *are independent.*

Moreover, the gamma process is the only case of such independence possible for pure NRMIs (this excludes the second case as it is marginalised).

Independence in the PYP case (represented as GEM(α, δ) in the lemma) is shown by Pitman and Yor [28] ([Proposition 21]).

That the gamma process is the only independence case for CRMs and their NRMIs is a result by Perman et al. [27] ([Corollary 2.3]). This is equivalent to the neutrality of the Dirichlet distribution, again the only distribution on probability vectors exhibiting neutrality. Neutrality and independence in this case can be shown to be equivalent properties. Independence in both these cases is also a consequence of the fact that so-called sized-biased sampling for the cases is independent of the total [27,29]. Independence properties such as in Lemma 8 do not hold generally, as indeed sized-biased sampling is not generally independent of the total.

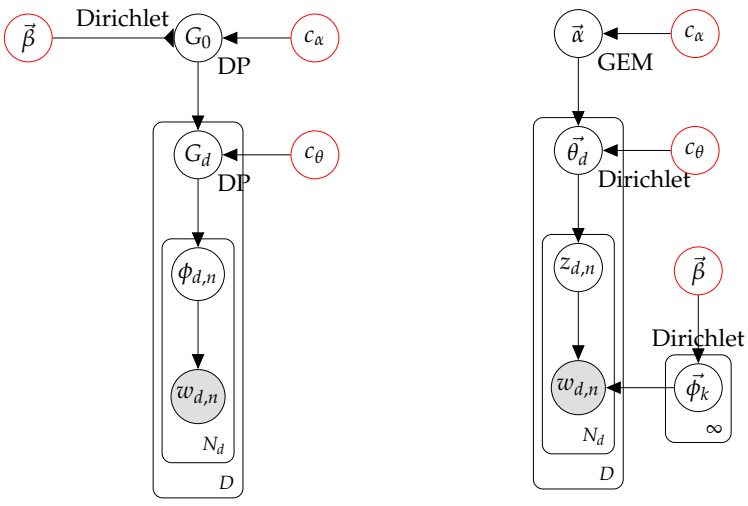
Lemma 9. (Normalisation of other process) *Let* $\Lambda = \sum_{i=1}^{\infty} \lambda_i$, *the sum of entries of the infinite vector* $\vec{\lambda}$.

- *For the generalised gamma process, if* $\vec{\lambda} \sim GGP(M, \alpha, \beta)$ *then* $\Lambda \sim Twe(\alpha, M^{1/\alpha}, \beta)$ *and* $\vec{\lambda}/\Lambda \sim NGG(\alpha, M)$.
- *For the stable process, if* $\vec{\lambda} \sim staP(M, \alpha)$ *then* $\Lambda \sim pstable(\alpha, M^{1/\alpha})$ *and* $\vec{\lambda}/\Lambda \sim PYP(\alpha, 0)$,

Λ *and* $\vec{\lambda}/\Lambda$ *are not independent in either case.*

6.4. Modelling LDA Using HDP

Consider models for the HDP variant of LDA [3], called HDP-LDA, which has been the subject of extensive research. There is a wide variation in the literature of how these are to be represented by graphical model and for statistical inference. Figure 3 shows two equivalent models for HDP-LDA. Figure 3a gives the original model as formulated by Teh et al. [3], and Figure 3b shows the modification used here. Authors sometimes use a more complicated formulation in terms of the underlying stick-breaking model.



(**a**) Graphical representation with DPs

(**b**) Modified representation with Dirichlets

Figure 3. Equivalent versions of HDP-LDA. In (**a**), the arc from $\vec{\beta}$ has a modified head to indicate that Dirichlet$(\vec{\beta})$ is used in a nested manner.

In this problem, there are D documents and N_d words in each document for $d = 1, \ldots, D$, where the words $w_{d,n}$ are modelled with an admixture. The probabilistic specification for the corresponding models are given in Figure 4.

$$\begin{aligned} G_0 &\sim \text{DP}(c_\alpha, \text{Dirichlet}(c_\beta \vec{\beta})) \\ G_d &\sim \text{DP}(c_\theta, G_0) \\ \vec{\phi}_{d,n} &\sim G_d \\ x_{d,n} &\sim \text{Categorical}(\vec{\phi}_{d,n}) \end{aligned}$$

(**a**) Original probabilistic model

$$\begin{aligned} \vec{\phi}_k &\sim \text{Dirichlet}(c_\beta \vec{\beta}) \\ \vec{\alpha} &\sim \text{GEM}(0, c_\alpha) \\ \vec{\theta}_d &\sim \text{Dirichlet}(c_\theta \vec{\alpha}) \\ z_{d,n} &\sim \text{Categorical}(\vec{\theta}_d) \\ x_{d,n} &\sim \text{Categorical}(\vec{\phi}_{z_{d,n}}) \end{aligned}$$

(**b**) Modified probabilistic model

Figure 4. Equivalent versions of HDP-LDA. Concentration parameters c_X treated as constants or estimated. Indices $d = 1, \ldots, D$, $n = 1, \ldots, N_d$ and $k = 1, \ldots, \infty$. The $\vec{\phi}$ are indexed differently in the two versions. The $\vec{\alpha}$ and $\vec{\theta}_d$ are infinite probability vectors in the CRM representation of G_0 and G_d, respectively.

Figure 4a shows the probabilistic specification with full base distributions. While this follows the theory directly, it is a fairly large departure from the original representation of LDA. The reformulation in Figure 4b is a direct analogue of the original representation of LDA with two modifications essential for the treatment of a HDP, discussed below as the root node and the non-root node.

The root node of the DP hierarchy is represented as a GEM, which generates the infinite vector. In practice, this can be represented using size-biased sampling [27] formulations, and in the simplest and popular cases this corresponds to stick-breaking methods [42]. In implementation, however, there is no need for this as posterior formulations for the processes are well understood and require no implicit ordering constraints as in stick-breaking.

Non-root nodes down the hierarchy are represented using their underlying infinitely divisible non-negative distribution, in this case the Dirichlet. Note, however, this extends the standard definition of a Dirichlet as the input parameter is an infinite dimensional vector. In implementation, this is no impediment as only a finite amount of data is ever dealt with, although it does require modelling the current number of non-empty dimensions. This can be readily handled using standard parametric techniques [6] or by using truncation [4].

Note Figure 4a also uses a nested construction [43] with the expression $\text{DP}(c_\alpha, \text{Dirichlet}(c_\beta \vec{\beta}))$. Here a distribution, in this case a Dirichlet, but it could also be a GP, a DP or any other process, is used as the base distribution. This nesting construction is exactly what is needed to model matrix and tensor factorisation using hierarchical processes.

The nested, hierarchical equivalent to Figure 5b is as follows:

$$\begin{aligned} \vec{\beta} &\sim \text{GEM}(d_\beta, c_\beta) \\ G_0 &\sim \text{GP}\Big(c_\alpha, 1, \text{PYD}\Big(d_\phi, c_\phi, \vec{\beta}\Big)\Big) \\ G_d &\sim \text{GP}(c_\theta, s_\theta, G_0) \\ \phi_{d,n} &\sim G_d \\ \vec{n}_d &\sim \text{Poisson}(\vec{\phi}_d) \end{aligned}$$

The background word probabilities $\vec{\beta}$ are generated, then used as the base distribution for a PYD which then creates variants $\vec{\phi}_k$ as each atom of the gamma process G_0. The mixture weights of G_0 correspond to $\vec{\alpha}$ from Figure 5b. Variants of this, G_d, are then created which modify the mixture weights $\vec{\alpha}$ but leave the atoms constant. So, G_d is a weighted sum of the original $\vec{\phi}_k$, as is the case in Figure 5b. This is very elegant, but Figure 5b better exposes the detail needed for implementation.

$$
\begin{aligned}
\vec{\alpha} &\sim \text{GEM}(0, c_\alpha) \\
\vec{\beta} &\sim \text{GEM}(d_\beta, c_\beta) \\
\vec{\theta}_d &\sim \text{Dirichlet}(c_\theta \vec{\alpha}) \\
\vec{\phi}_k &\sim \text{PYD}(d_\phi, c_\phi, \vec{\beta}) \\
\vec{n}_d &\sim \text{multinomial}\left(N_d, \sum_{k=1}^{\infty} \theta_{d,k} \vec{\phi}_k\right)
\end{aligned}
$$

(**a**) Non-parametric topic model

$$
\begin{aligned}
\vec{\alpha} &\sim \text{GP}(c_\alpha, 1) \\
\vec{\beta} &\sim \text{GEM}(d_\beta, c_\beta) \\
\vec{\theta}_d &\sim \text{gamma}(c_\theta \vec{\alpha}, s_\theta) \\
\vec{\phi}_k &\sim \text{PYD}(d_\phi, c_\phi, \vec{\beta}) \\
\vec{n}_d &\sim \text{Poisson}\left(\sum_{k=1}^{\infty} \theta_{d,k} \vec{\phi}_k\right)
\end{aligned}
$$

(**b**) Corresponding matrix factorisation

Figure 5. NP-LDA and its matrix factorisation counterpart. Concentration parameters c_X are constants or estimated, as are discounts d_X. Indices $d = 1, \ldots, D$ and $k = 1, \ldots, \infty$. Vector-wise versions of the gamma and Poisson represent the gamma process and Poisson process, respectively.

6.5. Example Equivalences with Non-Parametric LDA

Consider extending HDP-LDA to include a Pitman–Yor distribution on the word side. This model, termed NP-LDA Buntine and Mishra [4], has been demonstrated using a truncated approximation. To bring out equivalences, the multinomial form of the topic model is given, and both are defined in Figure 5.

The gamma scale parameter on α_0 is one as it has an equivalent affect to c_θ. So, it needs to be made a constant for identifiability. The equivalence is obtained by noting, from Lemmas 7 and 8, and many such results exist for the finite case, for instance by [44]. One can introduce a total rate for documents, Θ_d, and model the count, N_d, entirely independently:

$$
\begin{aligned}
\alpha_0 &\sim \text{gamma}(c_\alpha, 1) \\
\Theta_d &\sim \text{gamma}(c_\theta \alpha_0, s_\theta) \\
N_d &\sim \text{Poisson}(\Theta_d)\ .
\end{aligned}
$$

If the concentration parameters are estimated during learning, which is the common case, and recommended for topic models, then equivalence does not hold.

Experimental evidence [4] shows the following:

- The topic side, $\vec{\theta}_d$, is best not modelled using PYPs because experiments indicate that this gives no performance improvement. The non-Zipfian DPs work best, probably because of the smaller dimensions for number of topics.
- Modelling the word side, $\vec{\phi}_k$, using PYPs systematically outperforms HDP-LDA by a moderate margin in perplexity and yields more explainable topics because the overall "background" words are separately modelled using $\vec{\beta}$.

Several model equivalences hold with regard to these kinds of models.

- The asymmetric-symmetric version of LDA [45] is a truncated version, not well understood in the community.
- The asymmetric-asymmetric version of LDA, evaluated by Wallach et al. [45], is a truncated version of the model in Figure 5a.
- Hierarchical Poisson factorisation [46] (HPF) is a non-parametric formulation of Poisson-gamma matrix factorisation using stick-breaking, and thus is equivalent to HDP-LDA above (when augmented with a gamma model of the total counts).
- Robust (negative binomial) Poisson factorisation by Zhou et al. [47] is related (ignoring some issue with hyperparameters) to bursty topic models by Doyle and Elkan [48], which has a non-parametric extension in Buntine and Mishra [4].

7. Conclusions

Discrete base distributions make CRMs behave like vectors of infinitely divisible distributions, where application is element-wise without the non-parametrics. So, the gamma

process becomes an element-wise gamma distribution, and the generalised gamma process becomes an element-wise Tweedie distribution. This was presented in Lemma 2, Lemma 4 and Corollary 1 and accompanying tables. Similarly, discrete base distributions make NRMIs and related processes behave as normalised versions of the above, sharing some properties of the DP such as renormalisation. So, the HPYP becomes the PY distribution, whose form was developed in Section 5.

If closed forms for analysis of the infinitely divisible distributions do not exist, the generalised versions of Chinese restaurant process (CRP) sampling, given in Equation (17), can be used instead, including versions of the more recent, efficient collapsed samplers for CRPs [6,39], given in Equation (18). Similar formulations also appear in [14,15]. Note many of these quantities, for instance in Table 1, can be derived from the Laplace exponent of the CRM, so a convenient form of the distribution is not needed. The CRPs come about when unfolding the recursion that relates the cumulants of a distribution to the moments of the distribution, a simple result in basic statistics. In this way, known CRPs for the gamma process follow a general scheme that also applies for the generalised gamma process, the generalised beta process and others.

While most of these results follow fairly simply from general results in the non-parametric Bayesian community, some have not yet seen use in the Bayesian machine learning community.

As a specific example of hierarchical distributions, it was also shown in Section 5 that the NGG and PY distributions, for the case where discount $\alpha > 0$ and concentration $\beta > 0$, are behaving like normalised Tweedie variables, and for the case where concentration $\beta = 0$ like normalised positive stable variables. Moments of the Tweedie distribution show how the standard hierarchical likelihood for the HPYP used to date [5,37] can be directly derived from this framework without considering non-parametric theory. A novel integral expression for the PY distribution for discount $\alpha > 0$ and concentration $\beta \geq 0$ was also developed in Equation (21). This answers the question, "what is a hierarchical PYP"?

There are a rich number of variations of matrix factorisation and topic models that exist, for instance, see ([11] Table 1) with seven different versions of negative binomial matrix factorisation, and the software used in Buntine and Mishra [4] has seven different non-parametric versions of LDA. This is ignoring the extensions of the model where the problem is changed significantly: document segmentation [7], hierarchical topics [8], supervised topic models, etc., and these extensions no doubt have their own rich variety of versions and equivalences. Moreover, some of the known equivalences between processes, when applied in the hierarchical case, yield relationships between models and algorithms in the machine learning community that deserve further investigation, discussed in Section 6. This is confounded by the fact that variants are evaluated using significantly different methodologies; compare, for instance, topic modelling evaluation with recommender systems evaluation. It is an open question as to what other significant equivalences exist in the literature, and the implications this has to the algorithms one can use.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Lemma 4

Proof. The major result is by Pitman [22]. Equation (5) is obtained by differentiating inside the integral of the Laplace exponent. Note that when they both exist, cumulants κ_n and central moments c_n are related by the following recursive formula

$$c_n = \kappa_n + \sum_{k=1}^{n-1} \binom{n-1}{k-1} \kappa_k c_{n-k} \,.$$

One can expand this iteratively to remove the recursion on moments. While $\mathcal{P}^n$ represents the set of all non-empty partitions of n objects, let $\mathcal{S}^n$ denote the set of all vectors representing the sizes of non-empty partitions of n. So, if $\vec{m} \in \mathcal{S}^n$ then $m_l > 0$ for $l = 1, \ldots, |\vec{m}|$ and $\sum_{l=1}^{|\vec{m}|} m_l = n$. One obtain s the following:

$$\text{moment}_n(\gamma_k) = \sum_{\vec{m} \in \mathcal{S}^n} \prod_{l=1}^{|\vec{m}|} \text{cumulant}_{m_l}(\gamma_k) \binom{n - \sum_{j<l} m_j - 1}{m_k - 1}.$$

This is the same form of expression used in defining the generalised Stirling numbers [37] ([Lemma 16]). The significance is that the sum is over the sizes of the partitions of n, and the product of choose expressions represents the number of partitions with those sizes. Thus, this can be re-expressed as Equation (6). The recursion of Equation (8) can be obtained from the original recursion on c_n and reformulation. □

Appendix B. Proof of Corollary 1

Proof. This is based on the following result: Suppose a Poisson process has rate $\mu_k \rho(\lambda)$, and the distribution of the total $T = \sum_{k=1}^{\infty} \lambda_k$ from a sample has distribution $\Pr(T \mid \mu_k)$. Then, it follows that given rate $e^{-U\lambda}\rho(\lambda)$, the distribution of the total becomes $e^{\mu_k \psi(U) - UT} \Pr(T \mid \mu_k)$ where $\psi(\cdot)$ is the Laplace exponent of $\rho(\lambda)$. The result follows by using the constant $e^{-\mu_k \psi(U)}$ to adjust the moments to those desired. □

Appendix C. Proof of Lemma 6

Proof. Start with Equation (19) with no data, so $n_k = 0$ and U can be dropped. Substituting terms, and letting λ_R be λ_0:

$$\begin{aligned} &= e^M M^{(K+1)/(1-\alpha)} e^{-\sum_{k=0}^{K} \lambda_k} \left(\frac{\alpha}{(1-\alpha)\pi} \right)^{K+1} \\ &\quad \prod_{k=0}^{K} a(\nu_k) \lambda_k^{-1/(1-\alpha)} \mu_k^{1/(1-\alpha)} e^{-\lambda_k^{-\alpha/(1-\alpha)} M^{1/(1-\alpha)} \mu_k^{1/(1-\alpha)} a(\nu_k)}. \end{aligned}$$

Note it can be seen that conditionally $M^{1/(1-\alpha)}$ has a gamma distribution. Conditionally, the variables λ_k are log concave and vanishing to zero at the limits $(0, \infty)$. Moreover, by the transformation $\lambda'_k = 1/(1+\lambda_k)$, the transformed Hessian is only non-negative when the derivative is positive, so the function of $\lambda' \in [0, 1]$ is unimodal and suitable for slice sampling. Moreover, it can also be shown that conditionally the auxiliary variables ν_k are unimodal and bounded so are readily sampled using efficient slice sampling (as for instance used in a related context by Lomeli et al. [19]).

Marginalising out M by adding its prior and then using the change of variables $m = M^{1/(1-\alpha)}$,

$$\begin{aligned} &= \left(\frac{\alpha}{(1-\alpha)\pi} \right)^{K+1} \frac{(1-\alpha) e^{-\sum_{k=0}^{K} \lambda_k}}{\Gamma(\beta/\alpha)} \prod_{k=0}^{K} a(\nu_k) \lambda_k^{-1/(1-\alpha)} \mu_k^{1/(1-\alpha)} \\ &\quad \frac{\Gamma(K + 1 + \beta(1-\alpha)/\alpha)}{\left(\sum_{k=0}^{K} \lambda_k^{-\alpha/(1-\alpha)} \mu_k^{1/(1-\alpha)} a(\nu_k) \right)^{K+1+\beta(1-\alpha)/\alpha}}, \end{aligned}$$

then conduct a change of variables from $(\lambda_0, \lambda_1, \ldots, \lambda_K)$ to $(\Lambda, \theta_1, \ldots, \theta_K)$ where Λ is the sum and $\theta_k = \lambda_k / \Lambda$. The determinant of the Hessian is Λ^K. This results in an independent term in Λ in the form $e^{-\Lambda} \Lambda^{\beta - 1}$ which integrates leaving $\Gamma(\beta)$. This results in the data likelihood as given in Equation (21), though the dimension has also been changed from $K+1$ to K for simplicity. Moreover, $\frac{(1-\alpha)\Gamma(\beta)}{\Gamma(\beta/\alpha)}$ has been re-expressed as $\frac{1-\alpha}{\alpha} \frac{\Gamma(1+\beta)}{\Gamma(1+\beta/\alpha)}$, so it is well-defined when $\beta = 0$.

The derivation for the $\beta = 0$ case is similar, starting from Equation (20), but again there is no data and $U = 0$. Introduce the integral expression for the pstable(α, s), perform a change of variables from $(\lambda_0, \lambda_1, \ldots, \lambda_K)$ to $(\Lambda, \theta_1, \ldots, \theta_K)$ and then marginalise out Λ. At this point, the terms in M will cancel. □

References

1. Teh, Y. A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL; ACL '06, Sydney, Australia, 17–21 July 2006; pp. 985–992.
2. Kneser, R.; Ney, H. Improved backing-off for m-gram language modeling. In Proceedings of the Internatinal Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995; Volume 1, pp. 181–184.
3. Teh, Y.; Jordan, M.; Beal, M.; Blei, D. Hierarchical Dirichlet Processes. *J. ASA* **2006**, *101*, 1566–1581. [CrossRef]
4. Buntine, W.; Mishra, S. Experiments with Non-parametric Topic Models. In Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014.
5. Gasthaus, J.; Teh, Y. Improvements to the Sequence Memoizer. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 685–693
6. Lim, K.; Buntine, W.; Chen, C.; Du, L. Nonparametric Bayesian Topic Modelling with the Hierarchical Pitman-Yor Processes. *Int. J. Approx. Reason.* **2016**, *78*, 172–191. [CrossRef]
7. Du, L.; Buntine, W.; Johnson, M. Topic Segmentation with a Structured Topic Model. In Proceedings of the NAACL-HLT, Atlanta, GA, USA, 13 June 2013; pp. 190–200.
8. Paisley, J.; Wang, C.; Blei, D.; Jordan, M. Nested hierarchical Dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 256–270. [CrossRef] [PubMed]
9. Teh, Y.; Jordan, M. Hierarchical Bayesian Nonparametric Models with Applications. In *Bayesian Nonparametrics*; Hjort, N., Holmes, C., Müller, P., Walker, S., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 158–206.
10. Jordan, M. Hierarchical models, nested models and completely random measures. In *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*; Springer: New York, NY, USA, 2010; pp. 207–218.
11. Zhou, M.; Carin, L. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 307–320. [CrossRef]
12. Ferguson, T. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1973**, *1*, 209–230. [CrossRef]
13. Buntine, W. Constructing Poisson Process Models. Monash University, Clayton, Victoria, Australia Unpublished Report. 2018.
14. Camerlenghi, F.; Lijoi, A.; Orbanz, P.; Prünster, I. Distribution theory for hierarchical processes. *Ann. Stat.* **2019**, *47*, 67–92. [CrossRef]
15. Argiento, R.; Cremaschi, A.; Vannucci, M. Hierarchical Normalized Completely Random Measures to Cluster Grouped Data. *J. Am. Stat. Assoc.* **2020**, *115*, 318–333. doi: [CrossRef]
16. James, L. Bayesian Poisson Calculus for Latent Feature Modeling via Generalized Indian Buffet Process Priors. *Ann. Stat.* **2016**, *45*, 2016–2045. [CrossRef]
17. Griffiths, T.; Ghahramani, Z. The Indian Buffet Process: An Introduction and Review. *J. Mach. Learn. Res.* **2011**, *12*, 1185–1224.
18. James, L.; Lijoi, A.; Prünster, I. Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **2009**, *36*, 76–97. [CrossRef]
19. Lomeli, M.; Favaro, S.; Teh, Y. A marginal sampler for σ-stable Poisson-Kingman mixture models. *J. Comput. Graph. Stat.* **2015**, *9*, 44–53. [CrossRef]
20. Sato, K.I. Basic Results on Lévy Processes. In *Lévy Processes: Theory and Applications*; Barndorff-Nielsen, O., Mikosch, T., Resnick, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 3–37.
21. Lijoi, A.; Prünster, I. Models beyond the Dirichlet process. In *Bayesian Nonparametrics*; Hjort, N., Holmes, C., Müller, P., Walker, S., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 80–135.
22. Pitman, J. *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*; Springer: Berlin/Heidelberg, Germany, 2006.
23. James, L. Stick-breaking PG(α, ζ)-Generalized Gamma Processes. *arXiv* **2013**, arXiv:1308.6570.
24. Kingman, J. Random Discrete Distributions. *J. R. Stat. Soc. Ser. (Methodological)* **1975**, *37*, 1–15. [CrossRef]
25. Broderick, T.; Jordan, M.; Pitman, J. Beta processes, stick-breaking and power laws. *Bayesian Anal.* **2012**, *7*, 439–476. [CrossRef]
26. Brix, A. Generalized Gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **1999**, *31*, 929–953. [CrossRef]
27. Perman, M.; Pitman, J.; Yor, M. Size-biased Sampling of Poisson Point Processes and Excursions. *Probab. Theory Relat. Fields* **1992**, *92*, 21–39. [CrossRef]
28. Pitman, J.; Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **1997**, *25*, 855–900. [CrossRef]
29. Pitman, J. Poisson-Kingman partitions. In *Statistics and science: A Festschrift for Terry Speed*; Goldstein, D., Ed.; Lecture Notes–Monograph Series; Institute of Mathematical Statistics: Tachikawa, Tokyo, 2003; Volume 40, pp. 1–34.
30. Chen, C.; Buntine, W.; Ding, N. Theory of dependent hierarchical normalized random measures. *arXiv* **2012**, arXiv:1205.4159.
31. Steutel, F.; van Harn, K. *Infinite Divisibility of Probability Distributions on the Real Line*; Chapman & Hall/CRC Pure and Applied Mathematics; CRC Press: Boca Raton, FL, USA, 2003.

32. Hofert, M. Sampling Exponentially Tilted Stable Distributions. *ACM Trans. Model. Comput. Simul.* **2011**, *22*, 3:1–3:11. [CrossRef]
33. Nolan, J. Maximum likelihood estimation and diagnostics for stable distributions. In *Lévy Processes: Theory and Applications*; Barndorff-Nielsen, O., Mikosch, T., Resnick, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 379–400.
34. Chambers, J.; Mallows, C.; Stuck, B. A method for simulating stable random variables. *J. ASA* **1976**, *71*, 340–344. [CrossRef]
35. James, I.R.; Mosimann, J. A New Characterization of the Dirichlet Distribution Through Neutrality. *Ann. Stat.* **1980**, *8*, 183–189. [CrossRef]
36. ames, L.F.; Lijoi, A.; Prünster, I. Conjugacy as a Distinctive Feature of the Dirichlet Process. *Scand. J. Stat.* **2006**, *33*, 105–120. [CrossRef]
37. Buntine, W.; Hutter, M. A Bayesian View of the Poisson-Dirichlet Process. *arXiv* **2012**, arXiv:1007.0296v2.
38. Hsu, L.; Shiue, P.S. A unified approach to generalized Stirling numbers. *Adv. Appl. Math.* **1998**, *20*, 366–384. [CrossRef]
39. Chen, C.; Du, L.; Buntine, W. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 296–311.
40. Wood, F.; Teh, Y.W. A Hierarchical Nonparametric Bayesian Approach to Statistical Language Model Domain Adaptation. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, FL, USA, 16–18 April 2009.
41. Cole, D.; Morgan, B.; Titterington, D. Determining the parametric structure of models. *Math. Biosci.* **2010**, *228*, 16–30. [CrossRef]
42. Ishwaran, H.; James, L. Gibbs Sampling Methods for Stick-Breaking Priors. *J. ASA* **2001**, *96*, 161–173. [CrossRef]
43. Rodriguez, A.; Dunson, D.; Gelfand, A. The nested Dirichlet process. *J. ASA* **2008**, *103*, 1131–1154. [CrossRef]
44. Ding, C.; Li, T.; Peng, W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.* **2008**, *52*, 3913–3927. [CrossRef]
45. Wallach, H.; Mimno, D.; McCallum, A. Rethinking LDA: Why priors matter. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1973–1981.
46. Gopalan, P.; Ruiz, F.J.R.; Ranganath, R.; Blei, D.M. Bayesian Nonparametric Poisson Factorization for Recommendation Systems. In Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), Reykjavic, Iceland, 22–25 April 2014.
47. Zhou, M.; Hannah, L.; Dunson, D.; Carin, L. Beta-negative binomial process and Poisson factor analysis. In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Canary Islands, 21–23 April 2012.
48. Doyle, G.; Elkan, C. Accounting for burstiness in topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Montreal, QC, Canada, 14–18 June 2009; pp. 281–288.