arXiv:2304.05244v2 [math.ST] 4 Sep 2024

用于组内及组间共享潜在特征的广义层次印度自助餐过程的贝叶斯分析

Lancelot F. James*香港科技大学 ISOM系 Juho Lee 韩国科学技术院人工智能研究生院
Abhinav Pandey 香港科技大学 ISOM系

2024年9月5日

摘要

贝叶斯非参数层次先验在构建灵活模型方面十分有效，能够刻画组内及组间共享信息的潜在数据结构。本文聚焦于潜在特征分配模型，其中数据结构对应多重集或无界稀疏矩阵，我们将其称为广义层次印度自助餐过程（HIBP）。这类模型基于广义尖峰与 slab 印度自助过程（IBP）的层次版本；其中，Thibaux 和 Jordan（2007年）提出的基于伯努利分布的 HIBP，是对 Griffiths 和 Ghahramani（2005年）提出的 IBP 的层次扩展。本文着重讨论贝叶斯推断，对 HIBP 的联合分布、边缘分布和后验分布给出了新的显式描述，显著加深了我们对这些过程的理解。我们的结果使得原本复杂的联合边缘分布能够进行精确采样。我们还对其后验分布作出一般性刻画，并指出实际实现中的瓶颈。随后，我们的重点转向泊松 HIBP 这一重要情形下的具体可处理结果；它们对应于混合泊松随机计数模型的推广，这些模型应用于遗传学、成像、主题建模、随机占据和物种抽样模型。我们还展示了这些模型与文献中出现的贝叶斯非参数潜在类别模型之间的重要联系。此外，我们证明所有一般层次印度自助餐过程都可以与泊松 HIBP 耦合，从而能够进一步分析这类过程。

关键词：贝叶斯非参数方法、贝叶斯统计机器学习、层次印度自助餐过程、潜在特征模型、尖峰与 slab 先验

*Supported in part by grants RGC-GRF 16301521, 16300217, and T31-604/18-N of the HKSAR.

1 引言

贝叶斯非参数层次先验在为潜在数据结构提供灵活模型方面非常有效，这些结构表现出组内及组间的信息共享。最为突出的是层次狄利克雷过程（HDP） [32], 及其后续变体，它们对组内及组间的潜在聚类/类别进行建模。请注意，在这些模型中，个体在组内及组间最多共享一个特征。

在本研究中，我们聚焦于层次化潜在特征分配模型的可处理贝叶斯分析，其中的数据结构对应于多重集或无界稀疏矩阵。在这方面的基础性发展是 [33] 提出的层次印度自助餐过程（HIBP）。HIBP 利用贝塔过程层次结构，对 J 个不同组之间潜在特征的共享进行建模；每个组 j（$j = 1, 2, ..., J$）包含 M_j 个客户。具体而言，每个组 j 都会根据 [11] 首次提出的贝塔-伯努利印度自助过程（IBP）先验，生成一个潜在二元随机矩阵，以反映组内特征共享。与 HDP 类似，这些信息会依据另一个共同的贝塔过程在各组之间合并／共享，而该过程是完全随机测度（CRM）的一种。

在此，我们将这一框架扩展到一般尖峰与厚尾印度自助过程的层次化形式，称为广义 HIBP，其中包括文献中已有的模型。我们尤其关注基于泊松过程的 HIBP；其中， [35, 37], 研究的单客户泊松 - 伽马 - 伽马模型，或等价的负二项分布 - 伽马随机计数模型，是值得注意的特例，同时具有丰富的建模能力和良好的性质。一般而言，泊松 HIBP 基于多元混合泊松过程生成复杂的随机计数矩阵，并对应于遗传学、成像、主题建模、随机占据以及物种抽样模型中的模型推广。我们还表明，它们与文献中的贝叶斯非参数潜在类别模型存在重要联系。本文论述的一个关键特点是，这些联系使得我们能够在分布意义上恢复各个潜在计数；而在非层次化情形下，这些计数原本是可观测量。此外，我们证明所有一般层次印度自助过程均可与泊松 HIBP 耦合，从而能够进一步分析此类过程。

基本的印度自助过程（IBP）最初由 [11], 提出，是一类灵活且易于处理的贝叶斯非参数先验，可用于涉及潜在特征的广泛统计应用。例如，参见 [24], ，了解其近期在遗传学问题上的一个有趣应用。有关 IBP 将特征分配给单个客户的顺序方案（最初由 [11], 描述），以及其与潜在类别模型中一般中国餐馆过程的比较，更多内容请参见 [3] 。IBP 激发了对多种推广形式的开发与研究，这些形式可产生非二元条目。值得一提的是， [34] 指出，扮演类似客户角色的图像中，同一特征会出现多次。因此，类比经典的伽马分布-泊松贝叶斯参数模型，他们提出在这种情形下采用带有伽马分布过程先验的泊松型 IBP。文献中还研究了其他分布形式。具体而言， [16] 为一般的尖峰与厚尾 IBP 型模型及其重要的多元扩展提供了统一框架。这些模型基于在零点处具有概率质量、但除此之外可产生任何随机变量对应的非零条目的随机变量，生成条目值任意的稀疏随机矩阵。此外，参见 [1] ，其中或许能找到关于具有一般尖峰与厚尾条目的矩阵的一些思路。在这方面值得注意的是，具有一般复合泊松条目的模型，以及复杂的随机计数矩阵，也会自然地出现在我们的研究背景中。 [5], 的工作还描述了对伯努利 IBP 的修改：不使用完全随机测度（CRM），而是采用缩放子ordinator [19], 以便更好地建模未观测特征/物种模型，例如遗传学中出现的此类模型。此修改可以轻易应用于本文的论述，但我们在此不作采用。具体而言， [2] 讨论了这一修改在一般 IBP 和层次潜在特征模型中的应用；后者由 [23] 提出，此外还包括我们所分析的 HIBP 过程。

此处。他们对HIBP 的看法基于一些不够完善的性质，而我们旨在在此予以纠正。此后，对于任意整数 n，我们将使用记号 $[n] := \{1, 2, ..., n\}$ 来表示从 1 到 n 的整数集合。

1.1 概述与贡献

正如 [23, 33, 37]的相关研究中已经讨论的那样，将印度自助过程类型的过程扩展为在组之间共享多项特征的过程，具有极其广泛的应用可能性。接下来需要进行详细分析，如同对狄利克雷过程 [9, 25] 以及随后得到广泛应用的所有贝叶斯非参数模型所进行的分析一样，从而更好地理解它们的结构/分布性质，进而为广泛的应用实现实用的建模，并有助于开发高效的推断计算方法。此外，建立与可能在其他语境中出现的现有模型之间的联系也十分重要。我们的目标就是为本文所描述的HIBP 模型完成这一工作。具体而言，在第 2 节中，我们首先更详细地介绍印度自助过程和 HIBP 的定义，以及基于完全随机测度的构建模块。我们首先在一般印度自助过程设定下重新审视 [16] 中的关键结果，并在命题 2.1中通过一种分解方法突出更易理解的视角。随后，我们结合完全随机测度的基本性质，为 HIBP 提供初步表示；接着分析一个关键但通常隐藏的多元泊松过程，该过程决定特征在各组之间的分配。值得注意的是，这一过程可以被视为 [26]中出现的过程的多元扩展。第 3节随后给出一个结果，以便从任意HIBP过程的边缘分布中进行精确采样。第3.1 节给出了泊松 HIBP的具体细节，并且重要的是证明了任意HIBP 都可以与泊松框架耦合。我们引入了更多变量，其解释见于 [26], ；这些变量将构成一种分布恢复方法的基础，用于后验分析中的关键量。第 4节给出了所有HIBP的后验分布的一般描述，重点说明了在恢复潜在量方面的一般障碍，而在非层次设定中这些量通常是可观测的。我们没有尝试一般性地讨论如何解决这一问题，而是将本文其余部分集中于一般混合泊松过程，即泊松 HIBP 的情形，以实现完全可处理性。我们还注意到，由于可以将一般 HIBP与这一框架耦合，因此在这些情形下也能获得更多细节。简而言之，聚焦于泊松 HIBP，在第5节中，我们获得了后验分布和预测描述的可处理描述，以及一个原本复杂的边缘分布描述。由于我们的表示，这一边缘分布可以看作永久过程相关的Gibbs（乘积类型）结构之和；例如，参见 [8], 中对多元负二项分布及更一般过程的研究，而这些过程通常难以进行采样。我们还为 HIBP 描述了一种高度可处理的简化形式；HIBP 原本展现出丰富的建模结构，并且可能与 Pitman-Yor过程及其有限维版本存在令人意外的联系，这些联系见于 [21, 22], 近期的工作。在那里，我们提供了从相应的 Pitman-Yor 多项过程进行精确采样的新方法，以及其他性质。

我们进一步指出，混合泊松框架与物种抽样框架有着紧密联系，因为后者涉及潜在类别/聚类模型。特别是，Pitman [29],在一份未发表的手稿中表明，混合泊松过程在 Fisher/McCloskey 的框架内已经具有一种解释：即在给定时间抽样/捕获的顾客/动物数量，以及这一事件发生时间的相应分布，参见[12][第 2.2.1 节。第 321] 页还提供了一些相关细节。特别是，与我们的论述相关的是，他指出了 [36] 的工作与 [18]中伽马技巧的使用之间的联系；前者研究的对象在分布上等价于单个泊松印度自助过程，即一个混合泊松过程。事实上，与伽马技巧相对应的变量可以直接解释为混合泊松过程相应的到达时间。这在导出我们的关键结果之一——命题 5.1——的过程中发挥了重要作用。我们的工作，

其中涉及 J 组，可用于扩展这些概念，使其产生超出本文论述范围的更广泛影响。

2 广义 HIBP

现在，我们首先回顾 [16] 中讨论的一般印度自助过程，再讨论广义 HIBP 的正式定义。我们先回顾完全随机测度的定义，并强调当下文定义的基测度B_0 为离散测度时，与本文论述相关的一些性质。对每个 $j \in [J]$，令 μ_j 为完全随机测度，定义为 $\mu_j = \sum_{k=1}^{\infty} s_{j,k}\ {}_k\delta_{w_{j,k}}$，${}_k \sim \text{CRM}(\rho_j, B_0)$，其中 $\rho_j(s)dsB_0(d\omega)$ 是泊松随机测度的均值测度；有关这一关系的更多内容，参见 [16]。它规定了 $(s_{j,k},w_{j,k})_{\{k\geq 1\}}$ 的分布如下：这里，$B_0(dw)$ 是波兰空间 Ω 上的有限测度，满足 $(w_{j,k})_{\{k\geq 1\}} \overset{iid}{\sim} \bar{B}_0(dw) := B_0(dw)/B_0(\Omega)$，而 $\rho_j(s)$ 是定义在 $(0,\infty)$ 上的 L´évy 密度，满足 $\int_0^{\infty} \min(s,1)\rho_j(s)ds < \infty$。$\rho_j$ 决定了 μ_j 的跳跃分布，记为 $(s_{j,k})_{\{k\geq 1\}}$，而 $\mu_j(\Omega) := \sum_{k=1}^{\infty} s_{j,k}$ 是一个正的无限可分连续随机变量，其拉普拉斯变换为 $\mathbb{E}[e^{-t\mu_j(\Omega)}] = e^{-B_0(\Omega)\psi_j(t)}$，拉普拉斯指数为 $\psi_j(t) := \int_0^{\infty}(1-e^{-ts})\rho_j(s)ds$. 此外，对于可数个互不相交的集合 $(Q_{j,l})_{l\geq 1}$，$\mu_j(Q_{j,l})$ 相互独立，其对应的分布由拉普拉斯指数$B_0(Q_{j,l})\psi_j(t)$（$l = 1, 2, \ldots$）表示。

注 2.1。后一个关于集合独立性的陈述是完全随机测度 的定义性性质，适用于 B_0的连续和离散情形，并将使我们在后一种情形中的一些初始表示和推导更加清晰。因此，对于每个j，跳跃 $\{s_{j,k}:\ w_{j,\ k} \in Q_{j,\ l}, k \geq 1\}$，对于 $l \in \{1, 2, \ldots\}$ 个可数的不相交集合，它们相互独立，并由$L\acute{e}vy$密度 $B_0\ (Q_{j,\ l})\ \rho_j$指定，且原子按照与 $\mathbb{I}_{\{\omega\in Q_{j,l}\}}B_0\ (d\omega)\ \ l = 1$成比例的方式独立同分布地选取，2，.....。

现在，根据 [16], 我们为每个组 j 的$i \in [M_j]$客户定义广义尖峰与厚尾印度自助过程 $Z_j^{(i)} = \sum_{k=1}^{\infty} A_j^{(i)}:\ {}_k\delta_{w_{j,\ k}},\ {}_k|\mu_j \overset{iid}{\sim}$印度自助过程$_{A_j}(\mu_j)$，其中 $\mu_j \sim$完全随机测度(ρ_j, B_0)。这里，印度自助过程$_{A_j}$表示这样的过程：对于每个 $j \in [J],\ \ A_j^{(i)},\ {}_k|s_{j,\ k} = s$都独立同分布地等于一个尖峰与厚尾变量 A_j $\mathbb{P}_{A_j}\ da|s\ = [1-\pi_j\ s\ ]\delta_0\ da\ +$, $\mathbb{I}_{a\neq 0}\mathbb{P}_{A_j}\ da|s\ ,\quad \mathbb{P}\ A_j = 0|s\ = 1-\pi_j\ s\ \quad \int_0^{\infty}\ 1-[1-\pi_j\ s\ ]^{M_j}\ \rho_j\ s\ ds$ 其分布为：( ) ( ) ( ) ( )，其中 ( ) ( )，且 (( )) ( ) 有限。因此 $\mathbb{P}((A^{(i)} = 0, i \in [M_j]|s) = [1-\pi_j(s)]^{M_j}$. 注意 $A_j \overset{d}{=} b_j\hat{}A_j$，其中 $b_j \sim$伯努利$(\pi_j(s))$，且 $\hat{}\ A_j|s$ 具有零截断分布 $\mathbb{I}_{a\neq 0}\mathbb{P}_{A_j}(a|s)/\pi_j(s)$。作为特殊情形，$A_j \overset{d}{=} b_j \sim$伯努利$(s)$产生 $Z_j^{(i)}|\mu_j \overset{iid}{\sim}$ 伯努利过程(μ_j)，表示伯努利过程。在此情形下，$\hat{}\ A_j = 1$. 如果 $A_j \sim$泊松$(r_j s)$，那么 $Z_j^{(i)}|\mu_j \sim \text{PoisP}(r_j\mu_j)$表示泊松过程，正如 [34, 36], 所述。有关这些及其他情形的更多细节，请参见 [16][第 4] 节。

与文献中讨论的经典伯努利情形 [33] 和泊松-伽马-伽马 HIBP [35, 37] 情形一样，我们通过指定 B_0 为离散随机测度来构建一般层次印度自助过程模型，从而促进各 J 组之间的共享。具体而言：$B_0 = \sum_{l=1}^{\infty} \lambda_l\delta_{Y_l} \sim \text{CRM}(\tau_0, F_0)$，其中 F_0 是 Ω 上一个假定为无原子的正规概率分布，且 $\mathbb{E}[e^{-tB_0(\Omega)}] = e^{-\Psi_0(t)}$，其中 $\Psi_0(t) = \int_0^{\infty}(1-e^{-t\lambda})\tau_0(\lambda)d\lambda$。由此得到一般层次印度自助过程：对于 $i \in [M_j], j \in [J]$：$Z_j^{(i)}|\mu_j \overset{iid}{\sim}$印度自助过程$_{A_j}(\mu_j)\ \ \mu_j|B_0 \sim \text{CRM}(\rho_j, B_0)\ \ B_0 \sim \text{CRM}(\tau_0, F_0)$。我们注意到，[37], 中的负二项-伽马过程是泊松 HIBP 情形，其中$Z_j^{(1)}|\mu_j \sim \text{PoisP}(r_j\mu_j)$，$\mu_j|B_0$ 是伽马完全随机测度，B_0 是另一个伽马完全随机测度。然而，在该设定下（$M_j = 1$, $j \in [J]$）。我们关注过程（$Z_j^{(1)}, j \in [J]$），并通过预测过程 $Z_{J+1}|(\hat{}Z_j^{(i)}, j \in [J])$，从 J开始逐步扩展随机矩阵，得到 $J+1$等。类似地，虽然这一点并不明显，也可能鲜为人知，但若 $Z_j^{(1)} \sim \text{BerP}(\mu_j)$，则（$Z_j^{(1)}, \in [J]$）与 [7], 中构造的过程在分布上等价；同样，我们关注在 $J = 1$、2、.....中的扩展。我们将证明，这两个特殊情形的后验分布都有精确且可实现的解。

以及在μ_j·的一般选择下的预测分布我们始终注意到，我们使用多项分布表示多项式分布，并使用 tP 表示零截断泊松分布或多元泊松分布；此外，令 **0** 表示全零向量，其维度可能会根据上下文而变化。

2.1 性质

我们现在继续对相关的边际分布、后验分布和预测分布进行描述。我们首先从解释 [16][Section 3] 中非层次化设定下的一些相关结果开始，这也将帮助我们介绍一些关键的量和分布，其中 B_0 被视为已知。下面的描述基于将相关空间分解为一个 M_j 客户不会从中进行采样的集合及其补集，其中μ_j 具有复合泊松分量，该分量确定了变量的适当边际分布、联合分布和条件分布。我们认为，这将更清晰地描述 [16], 中的结果及其底层机制，并且这些描述可以得到更广泛的应用。

命题 *2.1*。 *[16][*节 *3]* 对于每个 j，考虑 $Z_j^{(i)}|\mu_j \overset{iid}{\sim}$ 印度自助过程$A_j(\mu_j)$，对于 μ_j ∼完全随机测度(ρ_j, B_0)。设 $\Phi_j(M_j) = \int_0^\infty (1-[1-\pi_j(s)]^{M_j})\rho_j(s)ds$. 事件 $\{k: A_j^{(i)},_k = 0, i \in [M_j]\}$，其中 $\mathbb{P}(A_j^{(1)} = 0, \ldots, A^{(M_j)}{}_j = 0|s) = [1-\pi_j(s)]^{M_j}$ 诱导出 μ_j 的*L´évy* 密度ρ_j 如下分解：

1. $\rho_j(s) = \rho^{(M_j)}{}_j(s) + \Phi_j(M_j)f_S(s)$ 其中 $\rho^{(M_j)}{}_j(s) = [1-\pi_j(s)]^{M_j}\rho_j(s)$

2. $f_{S_j}(s) = (1-[1-\pi_j(s)]^{M_j})\rho_j(s)/\Phi_j(M_j)$ 是一个正规的概率密度函数（*pdf*）。

3. 因此存在一种分解，

$$\mu_j \overset{d}{=} \mu_{j,M_j} + \sum_{\ell=1}^{\xi_j} S_{j,\ell}\delta_{\omega_{j,\ell}} \tag{2.1}$$

其中 μ_{j,M_j} ∼ 完全随机测度$(\rho^{(M_j)}{}_j, B_0)$， ξ_j ∼ 泊松$(\Phi_j(M_j)B_0(\Omega))$，$\omega_j(\omega_{j,\ell})$ 是独立同分布的 $\bar{B}_0$，且 $S_j(S_{j,\ell})$ 是独立同分布的，其概率密度函数为 $f_{S_j}(s)$。

4. 对于每个 $\ell \in [\xi_j]$ 存在相互独立的向量 $(\tilde{A}_{j\,\ell}^{(1)},,\ldots,\tilde{A}^{(M_j)}{}_{j,\ell})|S_{j,\ell} = s$ 其分布为 $\mathbb{I}_{(\mathbf{a}_{j,\ell}\neq\mathbf{0})}\prod_{i=1}^{M_j}\mathbb{P}_{A_j}(da_j^{(i)},_\ell|s)/[1-(1-\pi_j(s))^{M_j}]$. 存在以下变量的联合分布$(\tilde{A}_{j\,\ell}^{(1)},,\ldots,\tilde{A}^{(M_j)}{}_{j,\ell})\,S_{j,\ell}$

$$\mathbb{I}_{(\mathbf{a}_{j,\ell}\neq\mathbf{0})}\prod_{i=1}^{M_j}\mathbb{P}_{A_j}(da_{j,\ell}^{(i)}|s)\rho_j(s)/\Phi_j(M_j) \tag{2.2}$$

5. 因此，对(2.2) 中的ρ_j 进行积分可得到边缘分布，其分布记为印度自助过程$(A_j^{(i)}, i \in [M_j], \rho_j, B_0)$，

$$(Z_j^{(i)}, i \in [M_j]) \overset{d}{=} (\sum_{\ell=1}^{\xi_j}\tilde{A}_{j,\ell}^{(i)}\delta_{\omega_{j,\ell}}, i \in [M_j]), \tag{2.3}$$

其中 $Z_j^{(i)}(\Omega) \overset{d}{=} \sum_{\ell=1}^{\xi_j}\tilde{A}_{j\,\ell}^{(i)}$，，逐分量且联合地。

6. $\mu_j|(Z_j^{(i)}, i \in [M_j])$ 的后验分布由(2.1)中的分解描述：固定 ξ_j 和$(\omega_{j,\ell}, \ell \in [\xi_j])$ 的取值，并对每个 ℓ 应用贝叶斯法则于(2.2)，以获得$S_{j,\ell}|(\tilde{A}_{j\,\ell}^{(1)},\ldots,\tilde{A}^{(M_j)}{}_{j,\ell})$的条件密度。由于没有从该过程中选取任何点，$\mu_{j,M_j}$ ∼$\mathrm{CRM}(\rho^{(M_j)}{}_j, B_0)$ 的分布保持不变。

2.2 HIBP 的描述

从技术角度看，这一结果依赖于跳跃$S_{j,\ell}$的唯一性，而非从 B_0.中抽取的伴随量$\omega_{j,\ell}$。因此，无论 B_0 是离散还是连续，上述描述均成立。现在，在我们的 HIBP 设定中，当 $B_0 = \sum_{l=1}^{\infty} \lambda_l \delta_{Y_l} \sim$ 为 CRM(τ_0, F_0)，因而同时具有离散性和随机性时，我们可以使用命题 2.1 和注 2.1，对每个 j，令$Q_{j,l} := \{\omega : \omega = Y_l\}$，其中 $B_0(Q_{j,l}) = \lambda_l$，从而得到更精细且重要的表示。具体如下：1. $\mu_j|B_0 \stackrel{d}{=} \sum_{l=1}^{\infty} \sigma_{j,l}(\lambda_l)\delta_{Y_l}$ ，其中，2. $\sigma_{j,l}(\lambda)$的分布由拉普拉斯指数 $\lambda\psi_j(t)$给出，其密度记为$\eta_j(t_j|\lambda)$；此外，$\sigma_{j,l}(\lambda) = \sum_{k=1}^{\infty} s_{j,k,l}$，其中$(s_{j,k,l})_{\{k\geq 1\}}$ 是 μ_j 在长度为 λ 的区间上的跳跃，其 L´évy 密度为 $\lambda\rho_j(s)$。3. $\xi_j \stackrel{d}{=} \sum_{l=1}^{\infty} \xi_{j,l}$，对于$\xi_{j,l} \sim$ ，服从泊松$(\Phi_j(M_j)\lambda_l)$。4. $(Z_j^{(i)}, i \in [M_j]) \stackrel{d}{=} (\sum_{l=1}^{\infty} [\sum_{k=1}^{\xi_{j,l}} \tilde{A}_{j,k,l}^{(i)}]\delta_{Y_l}, i \in [M_j])$。5. 其中（（$\tilde{A}_{j\,k}^{(i)}$, ,$l$, $i \in [M_j]$)$S_{j,k,l}$)的分布与(2.2)中的相同。使用(2.3)。

求和表示，尤其是对 $(\mu_j, j \in [J])$ 中跳跃的求和，类似于 HDP 类型设定下随机概率测度的概率质量所发生的情况，这表明存在隐藏的随机聚类机制。我们将看到，与印度自助过程设定中的情况一样，无限求和表示不会带来任何显著困难。此外，我们注意到，条目 $(\sum_{k=1}^{\xi_{j,\ l}} \tilde{A}_{j\,k}^{(i)}{}_{,,l}, i \in [M_j])$ 是复合泊松过程向量，因此具有无限可分性。我们论述的关键所在——但通常不易察觉——是多元泊松过程 $\mathbf{X} := (\sum_{l=1}^{\infty} \xi_{j,l}\delta_{Y_l}, j \in [J])$，它决定了分布于各组 $j \in [J]$ 中的特征及其数量。在下一个结果中，我们将描述这个多元泊松印度自助过程的性质。尽管这一点并不显而易见，但该过程在全文中都起着关键作用，因为它将 [26], 中的离散化过程推广到多元情形；另见 [16][备注 4.2]。我们再次采用分解方法来进行描述，而其余部分则依赖于 [16][节 5] 中关于多元印度自助过程的结果。

命题 2.2。考虑 $\mathbf{X} \stackrel{d}{=} (\sum_{l=1}^{\infty} \xi_{j,l}\delta_{Y_l}, j \in [J])$。则事件 $\{l : \xi_{j,l} = 0, j \in [J]\}$，其中 $\mathbb{P}((\xi_{j,\ l} = 0, j \in [J])|\lambda) = e^{-\lambda \sum_{j=1}^{J} \Phi_j(M_j)}$ 诱导出 *L´évy* 密度 τ_0 *of* B_0 的如下分解：

1. $\tau_0(\lambda) =$和$^{-\lambda \sum_{j=1}^{J} \Phi_j(M_j)}\tau_0(\lambda) + \Psi_0(\sum_{j=1}^{J} \Phi_j(M_j))f_H(\lambda)$.

2. 其中 $f_H(\lambda) = (1 - e^{-\lambda \sum_{j=1}^{J} \Phi_j(M_j))\tau_0(\lambda)\, \Psi_0(\sum_{j=1}^{J} \Phi_j(M_j))}$ 是一个正规的密度。

3. 因此，令 $\tau_{0,J}(\lambda) = e^{-\lambda \sum_{j=1}^{J} \Phi_j(M_j)}\tau_0(\lambda)$，由此得到 B_0: 的分解

$$B_0 \stackrel{d}{=} B_{0,J} + \sum_{l=1}^{\varphi} H_l \delta_{\tilde{Y}_l} \tag{2.4}$$

其中 $B_{0,J} \sim \mathrm{CRM}(\tau_{0,J}, F_0)$、$(\tilde{Y}_l)$ 为独立同分布的 F_0，表示由 $\sum_{j=1}^{J} M_j$ 客户选取的不同特征数量，其分布为 $\varphi \sim$ 泊松$(\Psi_0(\sum_{j=1}^{J} \Phi_j(M_j)))$。$(H_l, l \in [\varphi]) \overset{iid}{\sim} f_H$ 是 B_0. 的相应跳跃

4. 从 $\mathbf{X}$ 中抽样会选出 φ 个独立同分布的向量；对于每个 $l \in [\varphi]$，它们对应于$(X_{j,\ l} = x_{j,l}, j \in [J])|H_l = \lambda$ 多元零截断泊松分布，记为 $\mathrm{tP}(\lambda(\Phi_1(M_1) \ldots, \Phi_J(M_J))$。其中 $x_l := \sum_{j=1}^{J} x_{j,l} = 1$、2、...

5. 因此 $\tilde{X}_l := \sum_{j=1}^{J} X_{j,\ l}|H_l = \lambda \sim \mathrm{tP}(\lambda \sum_{j=1}^{J} \Phi_j(M_j))$，其边缘分布记为 $\mathrm{MtP}(\sum_{j=1}^{J} \Phi_j(M_j)\ \tau_0)$

6. $(X_{j,l}, j \in [J])|\tilde{X}_l = x_l \sim$多项分布$(x_l;\ q_1,\ \ldots\ldots,\ q_J)$，对于 $q_j = \frac{\Phi_j(M_j)}{\sum_{v=1}^{J} \Phi_v(M_v)}$.

接下来，我们描述 $B_0|\mathbf{X}$ 的后验分布，它将在后续内容中发挥几个重要作用。我们可以看到，其后验分布的形式与 [16][第 4 节2], 中的单变量泊松印度自助过程相同，因为它仅取决于 ($\sum_{l=1}^{\infty}[\sum_{j=1}^{J}\xi_{j,l}]\delta_{Y_l})|B_0 \sim \text{PoiP}([\sum_{j=1}^{J}\Phi_j(M_j)]B_0)$。

命题 2.3。 $B_0|\mathbf{X}$ 的后验分布仅取决于 $(X_l = x_l, Y_l, l \in [\varphi], \varphi = r)$ ，并遵循 B_0 在(A.5) 中的分解，给定量如所示，~分布 $B_{0,J} \sim$ 完全随机测度$(\tau_{0,J}, F_0)$ 保持不变，而 $H_l|X_l = x_l$ 的密度正比于 $\lambda^{x_l}e^{-\lambda\sum_{j=1}^{J}\Phi_j(M_j)}\tau_0(\lambda)$，且对于 $l \in [r]$. 条件独立。

3 HIBP边际过程的精确抽样

现在我们利用上述结果来描述广义HIBP 的边缘过程；在许多情况下，可以通过对上述在非层次化设定中出现的基本变量进行采样来实际实现这一过程。

定理 3.1. 参照命题 *2.1* 和 *2.2* 中的变量，*HIBP* 的边缘分布 $(Z_j^{(i)}, i \in [M_j], j \in [J])$ 等于以下过程的分布：

$$\left(\sum_{l=1}^{\varphi}\left[\sum_{k=1}^{X_{j,l}}\tilde{A}_{j,k,l}^{(i)}\right]\delta_{\tilde{Y}_l}, i \in [M_j], j \in [J]\right) \tag{3.1}$$

注意，尽管按照（3.1）进行采样相对简单，但由此可以对相当复杂的多维分布进行精确采样。此外，为了与已有结果进行比较， [37] 中对应于具有（$M_j = 1$， $j \in [J]$）、的泊松-伽马-伽马 HIBP 的随机计数矩阵，其矩阵元素属于 $\mathbb{Z}^{J\times\varphi}$，其中 $\mathbb{Z} = \{0, 1, 2, \ldots\}$, J是行数， φ 是与观测到的特征对应的列数~（Y_l， $l \in [\varphi]$）。在我们的一般混合泊松 HIBP 设置中，随机计数矩阵取值于 $\mathbb{Z}^{\sum_{j=1}^{J}M_j\times\varphi}$，并随（$M_j$， $j \in [J]$）变化，其中 μ_j, B_0可以有更加一般的取值。关于其看似复杂的联合分布的精确定义，请参见第 5.3 节。下一节将进一步介绍泊松情形的细节，并证明所有一般 HIBP 都可以作为与泊松 HIBP 框架耦合的过程进行采样。

3.1 通过泊松情形对每个 HIBP 边缘分布进行精确采样

虽然我们可以直接对 A_j 使用定理 3.1，但我们将说明，可以通过变换 $\mathbb{P}(A_j \neq 0|v) = \pi_j(v) = (1 - e^{-s})$ 将其与泊松 HIBP 耦合，从而从每个 HIBP 的边缘分布中采样，其中 π_j^{-1} 表示其逆变换。请注意，我们可以通过为 A_j 使用不同的分布，同时对每个 j 耦合多个此类过程。

首先，作为定理 3.1 的一个特例，我们介绍泊松 HIBP 采样所需的一些重要通用细节，这些内容将在本文后续部分中反复使用。下文中，为了区分泊松情形，我们将使用 $C_j := A_j \sim$ 泊松(s)，令$\tilde{C}_{j\,k}^{(i)}$， ，$_l = \tilde{A}_{j\,k}^{(i)}$， ，$_l$。此外，我们用 B_j 代替 μ_j，并令 $\rho_j = \tau_j$，其中 $\tau^{(M_j)}{}_j(s) = e^{-sM_j}\tau_j(s)$，且令 $\Phi_j(M_j) = \psi_j(M_j) = \int_0^{\infty}(1 - e^{-sM_j})\tau_j(s)ds$。这里，$f_{S_j}(s) = (1 - \mathrm{e}^{-M_j s})\tau_j(s)/\psi_j(M_j)$，它指定了相应跳跃分布 $S_{j,k,l}$ 的分布，其中 $(\tilde{C}_{j\,k}^{(i)}, , l, i \in [M_j])|S_{j,k,l} = s \sim \text{tP}(s(M_j, \ldots, M_j))$ 。更多细节请参见 [16][第 4 节2]。

特别是，对于我们的总体阐述而言，重要的是，对于 $\tilde{C}_{j,\ k,\ l} := \sum_{i=1}^{M_j}\tilde{C}_{j\,k\,l}^{(i)}$ ，，， ~ 的联合分布 $S_{j,\ k,\ l}$，$C_{j,\ k,\ l}$ 可以表示为

$$\frac{s^{c_{j,k,l}}e^{-sM_j}\rho_j(s)}{\psi_j^{(c_{j,k,l})}(M_j)} \times \frac{M_j^{c_{j,k,l}}\psi_j^{(c_{j,k,l})}(M_j)}{\psi_j(M_j)c_{j,k,l}!} \tag{3.2}$$

其中 $\psi^{(c_{j,k,l})}{}_j(M_j) = \int_0^\infty s^{c_{j,\,k,\,l}} e^{-sM_j} \rho_j(s)ds$，在这些泊松情形立场下，对（3.2）的解释见 [26]。此外，同样对我们的总体阐述很重要的是，[26][的命题 12、推论 13]，（如 [16][备注 4.1], 中所述）也为泊松情形下的变量 $(H_l,\ X_l)$赋予了更精确的含义，具体是通过分解~空间。从该研究可知，联合分布 $(H_l,\ X_l)$可以用~ $\sim H_l|X_l = x_l$ 和 X_l，表示为:

$$\frac{\lambda^{x_l} e^{-\lambda \sum_{j=1}^J \psi_j(M_j)} \tau_0(\lambda)}{\Psi_0^{(x_l)}(\sum_{j=1}^J \psi_j(M_j))} \times \frac{\left(\sum_{j=1}^J \psi_j(M_j)\right)^{x_l} \Psi_0^{(x_l)}(\sum_{j=1}^J \psi_j(M_j))}{x_l! \Psi_0(\sum_{j=1}^J \psi_j(M_j))} \tag{3.3}$$

对于 $\Psi^{(x_l)}{}_0(\sum_{j=1}^J \psi_j(M_j)) = \int_0^\infty \lambda^{x_l} e^{-\sum_{j=1}^J \psi_j(M_j)} \tau_0(\lambda) d\lambda$. 即 $\tilde{X}_l \sim$ MtP分布$(\sum_{j=1}^J \psi_j(M_j)\ \tau_0)$

命题 3.1。设对每个 j，$(Z_j^{(i)}, i \in [M_j])|B_j \overset{iid}{\sim} \mathrm{PoiP}(B_j)$，其中 $B_j|B_0 \sim$ 完全随机测度(τ_j, B_0)在给定 $j \in [J]$ 的条件下相互独立。接着考虑求和过程 $(\sum_{i=1}^{M_j} Z_j^{(i)}, j \in [J])$，其中 $\sum_{i=1}^{M_j} Z_j^{(i)}|\mu_j \sim \mathrm{PoiP}(M_j B_j)$。

1. 求和过程 $(\sum_{i=1}^{M_j} Z_j^{(i)}, i \in [M_j], j \in [J])$ 的联合边缘分布与其同分布于

$$(\sum_{l=1}^{\varphi} N_{j,l} \delta_{\tilde{Y}_l}, j \in [J]) \overset{d}{=} ((\sum_{l=1}^{\varphi} \sum_{k=1}^{X_{j,l}} \tilde{C}_{j,k,l} \delta_{\tilde{Y}_l}, j \in [J]) \tag{3.4}$$

2. 其中 $(\tilde{C}_{j,\,k,\,l} = \sum_{i=1}^{M_j} \tilde{C}_{j\,k\,,\ ,\ l}^{(i)}, k \in [X_{j,l}]) \overset{iid}{\sim}$ MtP分布(M_j, τ_j) 其中 $X_{j,\,l}$ 个分量在 l之间相互独立。

3. $\tilde{X}_l \sim$ MtP分布$(\sum_{j=1}^J \psi_j(M_j)\ \tau_0)$

4. $(X_{j,\,l,\ j \in [J]})|\tilde{X}_l = x_l \sim$多项分布$(x_l;\ q_1,\ \ldots,\ q_J)$，对于 $q_j = \frac{\psi_j(M_j)}{\sum_{v=1}^J \psi_v(M_v)}$.

5. $(\sum_{k=1}^{X_{j,\ \ l}} \tilde{C}_{j\,k}^{(i)}, \ , l, i \in [M_j])|N_{j,\,l} := \sum_{k=1}^{X_{j,\ k=1}{}^{l}} \tilde{C}_{j,\,k,\,l} = n_{j,\,l}$ 为 多项分布$(n_{j,\,l};\ 1/M_j, \ldots, 1/M_j)$，对于每个 j, l,

命题 3.1 表明，可以先按照式（3.4）从求和过程中抽样 $((N_{j,l}, j \in [J])\ Y_l, l \in [\varphi], \varphi)$，再按照第 5 项进行多项分布抽样，从而得到边缘分布 $(Z_j^{(i)}, i \in [M_j], j \in [J])$, 的样本。接下来的结果将说明如何通过变换跳跃，将任意一般层次印度自助餐过程与相应的泊松 HIBP 耦合并进行抽样。

命题 *3.2*。考虑变量定义如命题 *3.1* 的泊松 *HIBP* 设置。对于 $A_j|v \sim \mathbb{P}_{A_j}(da|v)$，其中 $\mathbb{P}(A_j \neq 0|v) = \pi_j(v)$，指定一个 *L´évy* 密度 $\rho_j(v)$，使得 $\psi_j(M_j) = \int_0^\infty [(1-(1-\pi_j(v))^{M_j}]\rho_j(v)dv = \int_0^\infty (1-e^{-sM_j})\tau_j(s)ds$。接着考虑变量 $(\hat{A}_{j\,k}^{(i)}, , l, i \in [M_j])|S_{j,k,l} = s$，其独立同分布的零截断分布与 $\mathbb{I}_{a \neq 0} \mathbb{P}_{A_j}(da|\pi_j^{-1}(1-e^{-s}))$ 成正比；其中 $S_{j,k,l}$ 独立同分布，且具有共同密度 $f_{S_j}(s) = (1-e^{-M_j s})\tau_j(s)/\psi_j(M_j)$

1. 因此，下式等于耦合泊松 *HIBP* 的边缘分布，

$$\left(\sum_{l=1}^{\varphi} \left[\sum_{k=1}^{X_{j,l}} \mathbb{I}_{\{\tilde{C}_{j,k,l}^{(i)} > 0\}} \hat{A}_{j,k,l}^{(i)} \right] \delta_{\tilde{Y}_l}, \sum_{l=1}^{\varphi} \left[\sum_{k=1}^{X_{j,l}} \tilde{C}_{j,k,l}^{(i)} \right] \delta_{\tilde{Y}_l}, i \in [M_j], j \in [J] \right)$$

2. 其中 $(\sum_{l=1}^{\varphi} \left[\sum_{k=1}^{X_{j,\ \ l}} \mathbb{I}_{\{\tilde{C}_{j,k,l}^{(i)} > 0\}} \hat{A}_{j\,k,\ ,\ l}^{(i)} \right] \delta_{\tilde{Y}_l}, i \in [M_j], j \in [J])$ 具有一般层次印度自助餐过程的边缘分布，其中 A_j、ρ_j、τ_0 如（3.1）所规定

3. （$\sum_{l=1}^{\varphi}\left[\sum_{k=1}^{X_j}{}_{,}{}^{l}\mathbb{I}_{\{\tilde{C}_{j,k,l}^{(i)}>0\}}\right]\delta_{\tilde{Y}_l}$、 $i\in[M_j]$、 $j\in[J]$）对应于伯努利层次印度自助过程，其中 $Z_j^{(i)}|\mu_j\sim$ 伯努利过程（μ_j）且 $\mu_j|B_0\sim$ 完全随机测度（ρ_j、 B_0）。

4 B_0和（μ_j， $j\in[J]$）的后验分布

虽然我们的结果表明，我们可以通过变量（X_j、ι、 （$A_{j\,k}^{(i)}$、 、ι、 $k\in[X_{j,l}]$、 $\tilde{Y}_l$、 $l\in[r]$、 $\varphi=r$、 $i\in[M_j]$、 $j\in J$））对过程（（$Z_j^{(i)}$、 $i\in[M_j]$） $j\in[J]$）进行精确采样，但如果假设我们只知道最终结果中的信息，那么这些信息可表示为（$\sum^{X_j}$、 $_{k=1}{}^{\iota}\tilde{A}_{j\,k}^{(i)}$、 、$\iota$、 $\tilde{Y}_l$、 $l\in[r]$、 $\varphi=r$、 $i\in[M_j]$、 $j\in J$）。现在我们来描述 B_0 和（μ_j、 $j\in[J]$）的后验分布。

定理 4.1。 $(\mu_j\in[J],B_0)|\mathbf{Z}_J$的后验分布描述如下： $\mathbf{Z}_J$： $=(\sum^{X_j}{}_{,}{}_{k=1}{}^{\iota}\tilde{A}_{j\,k}^{(i)},,\iota,\tilde{Y}_l,l\in[r],\varphi=r,i\in[M_j],j\in J)$，

1. $\mu_j\,j\in[J]|Z$ 的联合分布满足：逐分量地以及联合地，其分布等于，

$$(\tilde{\mu}_{j,M_j}+\sum_{l=1}^{r}\left[\tilde{\sigma}_{j,l}(H_l)+\sum_{k=1}^{X_{j,l}}S_{j,k,l}\right]\delta_{\tilde{Y}_l},j\in[J]),B_{0,J}+\sum_{l=1}^{r}H_l\delta_{\tilde{Y}_l}\qquad(4.1)$$

对于 $B_{0,J}\sim$完全随机测度$(\tau_{0,J},F_0)$ $\tilde{\mu}_{j,M_j}\sim$完全随机测度$(\rho^{(M_j)}{}_j,B_{0,J})$，以及 $(\tilde{\sigma}_{j,l})\overset{iid}{\sim}$完全随机测度$(\rho^{(M_j)}{}_j,\mathbb{L})$，其中 $\mathbb{L}(dt)=dt$ 在 $(0,\infty)$ 上，且与 (H_l) 和 $(S_{j,k,l})$ 相互独立。

2. $S_{jkl}|(\tilde{A}_{j\,k}^{(1)},,\iota,\ldots,\tilde{A}^{(M_j)}{}_{jkl})$ 的条件分布由 *(2.2)* 确定，且$H_l|\tilde{X}_l=x_l$的密度为 $\propto\lambda^{x_i}e^{-\lambda\sum_{j=1}^{J}\Phi_{j}M_j\tau_0\lambda}$,

3. 后验分布通过给出在 Z 条件下(

$X_{j\,l}\quad{}_{k\,l}\;k\in{}_{,}\,(\tilde{A}_{j}^{(i)}{}_{,,},,,,)$ 的描述而得出

$[X_{j,l}]\;l\in[r]\;i\in[M_j]\;j\in J$， $\quad{}_{J}$

定理 4.1 依赖于根据观测到的总和（,,(,,),,）推断潜变量的条件分布（ $X_{j\,l}\quad A_{j\,k\,l}^{(i)}\;k\in[X_{j,l}]\quad l\in[r]\;i\in[M_j]\;j\in J$， $\sum_{k=1}^{X_{j\,l}}\tilde{A}_{j\,k\,l}^{(i)}\;\tilde{Y}_l\;l\in$）。 $[r]\;\varphi=r\;i\in[M_j]\;j\in J$.

尽管有可能开发计算方法来利用这一概念，但总体方法仍不明确。此外，运用 [16][部分 5] 中的结果，将这些过程表示为多元 IBP，可以仅根据总和描述后验分布，而无需了解各个潜变量的条件分布。然而，只有在能够获得关于总和分布的大量信息时，这一策略才具有实际可行性。

带有 $(M_j=1,j\in[J])$ 的伯努利层次印度自助过程情形，在分布上对应于过程 $(\sum_{l=1}^{\infty}\xi_{j,l}\delta_{Y_l},j\in[J])$，对于$\xi_j$，$\iota\sim$Poisson$(\Phi_j(1)\lambda_l)$，这表明一种模型：人们只关注将 J 增加到 $J+1$ 等情况。事实上，这对应于 [7], 中的二部图模型，其中有 J 位读者、(Y_l) 本书以及 (λ_l)它们各自的流行度参数，并进一步设为 $\gamma_j=\Phi_j(1)$。定理 4.1 给出了一个完整且易处理的描述；否则，这一描述无法获得，并且对 [7]中γ_j 的任意选择都成立。特别地，由此可知 $\tilde{A}_{j\,k}^{(1)}$，、$\iota=1$，因此，在这种情况下，X_j、$\iota=\sum^{X_j}$、$_{k=1}{}^{\iota}\tilde{A}_{j\,k}^{(1)}$，、$\iota$被观测到，而 S_j、$_k$、$_\iota$ 的密度与$s\rho_j(s)$成正比，其中 s 属于(0,1)，且 $j=1$、2、.....。对于 $M_j>1$ 而言，伯努利情形更加复杂。相比之下，正如我们接下来将展示的，在泊松HIBP 情形下，利用 [18, 20, 26, 27, 36]中现有结果的新颖用法，可以以一般方式获得后验分布的完整且易处理的描述。反过来，由于命题 3.2，通过耦合，这可以用于为其他一般层次印度自助餐过程情形提供更多信息。

5 泊松 HIBP 的后验分布

现在我们来讨论泊松 *HIBP* 这一特殊情形。给定的结构为 $Z_J := (\sum_{k=1}^{X_j} \tilde{C}_{j,k}^{(i)}, l, \tilde{Y}_l, l \in [r], \varphi = r, i \in [M_j], j \in J)$。然而，根据命题 *3.1*，只需利用求和过程 $(N_{j,l}, Y_l, l \in [r], \varphi = r, j \in J)$ 中的信息，其中同样有 $N_{j,l} = \sum_{k=1}^{l} \tilde{C}_{j,k,l}$，用于描述后验分布。于是，任务变为描述在给定这些信息的条件下 $(\tilde{C}_{j,k,l}, k \in [X_{j,l}], X_{j,l})$ 的分布。有关单个计数 $\tilde{C}_{j,k}^{(i)}$，l 的分布恢复，见第 *5.5* 节。

我们在此的一项创新是，首先注意到，在给定 $(N_{j,l}, j \in [J])$ 和 $H_l = \lambda$ 的条件下，我们可以应用 [20, 26], 中的结果，也可以应用 [18, 36], 中的结果，直接推导出以下基于有限 Gibbs 交换分区概率函数，的关键结果。这些交换分区概率函数（EPPF）也具体见于 [18]命题 4]和 [36][第 2.3,3] 节。有关物种抽样模型的一般 EPPF，详见 [25, 27]。

命题 *5.1*。考虑命题 *3.1* 中的泊松 *HIBP* 设置。那么，对每个 $j \in [J]$，在给定 $N_{j,l} = n_{j,l}$ 和 $H_l = \lambda$ 的条件下，$(\tilde{C}_{j,k,l}, k \in [X_{j,l}], X_{j,l})$ 服从整数 $[n_{j,l}]$. 的随机分区分布，使得 $\sum_{k=1}^{X_j} \tilde{C}_{j,k,l} = n_{j,l}$，其中 $\tilde{C}_{j,k,l} = c_{j,k,l}$ 决定每个区块的大小，而 $X_{j,l} = x_{j,l}$ 对应随机区块数；该分布遵循 *[18, 36]* 中所得到的有限 *Gibbs* 交换分区概率函数，具体如下：

1. 对于 $\mathbf{c}_{j,l} = (c_{j,k,l}, k \in [x_{j,l}])$，以及 $\psi_j^{(c)}(M_j) = \int_0^\infty s^c e^{-sM_j} \tau_j(s)$ *{v31>*，指定这些条件分布的有限 *Gibbs* 分区分布为：

$$p^{[n_{j,l}]}(\mathbf{c}_{j,l} | \lambda \tau_j, M_j) = \frac{\lambda^{x_{j,l}} \prod_{k=1}^{x_{j,l}} \psi_j^{(c_{j,k,l})}(M_j)}{\Xi_j^{[n_{j,l}]}(\lambda, M_j)} \tag{5.1}$$

2. 其中，有恒等式：$\int_0^\infty t^{n_{j,l}} e^{-tM_j} \eta_j(t|\lambda)\, dt = e^{-\lambda \psi_j(M_j)} \Xi_j^{[n_{j,l}]}(\lambda, M_j)$ 对于 $\sum_*$，表示对所有满足 $\sum_{k=1}^{x_j} c_{j,k,l} = n_{j,l}$ 的正整数 $\mathbf{c}_j$, 求和：

$$\Xi_j^{[n_{j,l}]}(\lambda, M_j) = \sum_{x_{j,l}=1}^{n_{j,l}} \frac{n_{j,l}! \lambda^{x_{j,l}}}{x_{j,l}!} \sum_* \frac{\prod_{k=1}^{x_{j,l}} \psi_j^{(c_{j,k,l})}(M_j)}{\prod_{k=1}^{x_{j,l}} c_{j,k,l}!} \tag{5.2}$$

3. 区块数量的分布 $X_{j,l} = x_{j,l}$ 在 $(N_{j,l} = n_{j,l}, j \in [J], H_l = \lambda$ 条件下为，$X_{j,l} = 0$ 当 $n_{j,l} = 0$ 时，否则：

$$p^{[n_{j,l}]}(x_{j,l} | \lambda \tau_j, M_j) = \frac{n_{j,l}! \lambda^{x_{j,l}}}{x_{j,l}!} \sum_* \frac{\prod_{k=1}^{x_{j,l}} \psi_j^{(c_{j,k,l})}(M_j)}{\Xi_j^{[n_{j,l}]}(\lambda, M_j) \prod_{k=1}^{x_{j,l}} c_{j,k,l}!} \tag{5.3}$$

4. 如 *[20, 26],* (5.3) 和 (3.2) 所示，可得恒等式，$p^{[n_{j,l}]}(x_{j,l} | \lambda \tau_j, M_j) =$

$$\frac{n_{j,l}! \lambda^{x_{j,l}} (\psi_j(M_j))^{x_{j,l}}}{x_{j,l}! M_j^{n_{j,l}} \Xi_j^{[n_{j,l}]}(\lambda, M_j)} \mathbb{P}(\tilde{C}_{j,1,l} + \ldots + \tilde{C}_{j,x_{j,l},l} = n_{j,l}) \tag{5.4}$$

示例 *5.0.1*$(B_j | B_0 \sim G(\theta_j, \zeta; B_0))$。现在作为重要示例，参见 *[18]/p。 86],* 当 $\tau_j(s) = \theta_j s^{-1} e^{-s\zeta_j}$ 对应于伽马分布过程时，$\psi_j(M_j) = \theta_j [\log(\zeta_j + M_j) - \log(\zeta_j)]$，$\psi_j^{(c)}(M_j) = \theta_j \Gamma(c)(\zeta_j + M_j)^{-c}$，以及伽马分布密度，$\eta_j(t_j|\lambda) = t_j^{\theta_j \lambda_j - 1} e^{-t_j \zeta_j} \zeta^{\theta_j \lambda} / \Gamma(\theta_j \lambda)$ 导出 $\Xi_j^{[n_{j,l}]}(\lambda, M_j) = \Gamma(\theta_j \lambda + n_{j,l})(\zeta_j + M_j)^{n_j}$，$\Gamma(\theta_j \lambda) \cdot$ 由此，$p^{[n_{j,l}]}(\mathbf{c}_{j,l} | \lambda \tau_j, M_j)$ 等同于带参数 $\theta_j \lambda$ 的狄利克雷过程的众所周知的 *EPPF*：$p_{0,\theta_j \lambda}(\mathbf{c}_{j,l}) := \frac{\Gamma(\theta_j \lambda \theta_j^{x_j} \Gamma^{x_j})}{\Gamma(\theta_j \lambda + n_{j,l}) \prod_{k=1}^{x_j} \Gamma(c_{j,k,l})}$，，

且不依赖于(ζ_j, M_j)。现在利用这一点可知，在此情形下，$p^{[n_{j,l}]}(x_{j,l}|\lambda\tau_j, M_j)$与以下内容相同，

$$\mathbb{P}^{(n_{j,l})}_{0,\theta_j\lambda}(x_{j,l}) = \frac{\Gamma(\theta_j\lambda)\theta_j^{x_{j,l}}\lambda^{x_{j,l}}}{\Gamma(\theta_j\lambda + n_{j,l})}\begin{bmatrix} n_{j,l} \\ x_{j,l}\end{bmatrix} \tag{5.5}$$

其中，左侧表示狄利克雷过程下 $[n_{j,l}]$ 的划分区块数的分布，$\begin{bmatrix} n \\ k\end{bmatrix}$ 表示第一类无符号斯特林数，见 *[8, 31]*。由(5.4) 和(5.5)，可得等式

$$\mathbb{P}(\tilde{C}_{j,1,l} + \ldots + \tilde{C}_{j,x_{j,l},l} = n_{j,l}) = \frac{x_{j,l}!}{n_{j,l}!}\frac{(1 - \frac{\zeta_j}{\zeta_j + M_j})^{n_{j,l}}}{(\log(\zeta_j + M_j) - \log(\zeta_j))^{x_{j,l}}}\begin{bmatrix} n_{j,l} \\ x_{j,l}\end{bmatrix} \tag{5.6}$$

$\tilde{}$ $\tilde{}$表示参数为 $\mathbb{P}(C_{j,1,l}+\ldots+C_{j,x_{j,l},l} =n_{j,l})$ 的和对数分布的概率质量函数，其中 $\mathbb{P}(\tilde{C}_{j,k,l} = c) = (\quad \frac{M \quad c}{\zeta+M}\log(\zeta)]$，这为 *[35, 37]* 中的策略提供了依据。

注释 5.1。与 $0 < \alpha < 1$ 稳定过程相对应的情形 $\tau_j(s) = Cs^{\alpha-1}$ 在 *[26]*, 中进行了讨论，另见 *[28][* 第 *73]*, 页，稍后将以广义伽马过程的形式加以说明。

现在，我们使用 $p^{[n_{j,l}]}(\mathbf{c}_{j,\ l}|\tau_j,\ M_j)\Xi_j^{[n_{j,l}]}(\mathbf{1},\ M_j) = \prod_{k=1}^{x_{j,\ l}}\psi_j^{(c_{j,k,l})}(M_j)$。给出泊松 HIBP 后验分布的一个可实现描述。

定理 5.1。$B_j \in [J], B_0)|\mathbf{Z}_J$ 的后验分布如下所示，

1. $B_j, j \in [J])\ B_0|\mathbf{Z}_J$ 的联合分布满足：逐分量地以及整体上，二者均与以下分布等价：

$$(B_{j,M_j} + \sum_{l=1}^{r}\left[\tilde{\sigma}_{j,l}(H_l) + \sum_{k=1}^{X_{j,l}} S_{j,k,l}\right]\delta_{\tilde{Y}_l}, j \in [J]), B_{0,J} + \sum_{l=1}^{r} H_l\delta_{\tilde{Y}_l} \tag{5.7}$$

对于 $B_{0,\ J} \sim \mathrm{CRM}(\tau_{0,\ J}, F_0)$ $B_{j,\ M_j} \sim \mathrm{CRM}(\tau^{(M_j)j}, B_{0,\ J})$，以及

2. $(\tilde{\sigma}_{j,\ l}(H_l)\ j \in [J])|H_l = \lambda$ 相互独立，密度为 $\eta_j^{[0]}(t_j|\lambda, M_j) = e^{-t_jM_j}\eta_j(t_j|\lambda)e^{\psi_j(M_j)}$，且独立于$(S_{j,\ k,\ l})$。

3. $S_{j,k,l}|\tilde{C}_{j,k,l} = c_{j,k,l}$ 具有密度 $s^{c_{j,k,l}}e^{-sM_j\tau_j(s)}\psi^{(c_{j,k,l})}(M_j)$

4. $H_l|\tilde{X}_l = x_l$ 具有密度 $\lambda^{x_l}e^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)}\tau_0(\lambda)\,\Psi_0^{(x_l)}(\sum_{j=1}^{J}\psi_j(M_j))$，其中 $\tilde{X}_l = \sum_{j=1}^{J} X_{j,l}$

$\tilde{}$*5.* $(X_{j,l}, (C_{j,k,l}, k \in [X_{j,l}], l \in [r], j \in J)$，给定 $(N_{j,l}, j \in [J])\ l \in [r])$ 在 $l \in [r]$ 上条件独立，其分布与.....成正比

$$\frac{\Psi_0^{(x_l)}(\sum_{j=1}^{J}\psi_j(M_j))}{\Psi_0(\sum_{j=1}^{J}\psi_j(M_j))}\prod_{j=1}^{J} p^{[n_{j,l}]}(\mathbf{c}_{j,l}|\tau_j, M_j)\Xi_j^{[n_{j,l}]}(1, M_j) \tag{5.8}$$

其中 $x_l = \sum_{j=1}^{J} x_{j,\ l}$，并注意到 $X_{j,\ l} = 0$ 若 $N_{j,\ l} = 0.$

6. 特别地 $\tilde{X}_l|(N_{j,l} = n_{j,l}, j \in [J])\ l \in [r])$。对于 $x_l = \sum_{j=1}^{J} x_{j,l}$，当 $x_{j,l} = 0$ 时$n_{j,l} = 0$，由 $(X_{j,l} = x_{j,l}, j \in [J])$ 的联合分布决定，该分布正比于

$$\frac{\Psi_0^{(x_l)}(\sum_{j=1}^{J}\psi_j(M_j))}{\Psi_0(\sum_{j=1}^{J}\psi_j(M_j))}\prod_{j=1}^{J} p^{[n_{j,l}]}(x_{j,l}|\tau_j, M_j)\Xi_j^{[n_{j,l}]}(1, M_j)$$

以下命题表明，定理 5.1 中的后验分布仅直接取决于˜（$X_{j,l}$， $j \in [J]$）。换言之，无需知道总计数 $C_{j,k,l}$，尽管出于计算目的也可以将其纳入考虑。该结果源于 [18][定理 1],或等价地 [15][推论 5.1]中的总质量后验分布的随机化。另请参见 [27]。不过，若有兴趣˜获得计数 $C_{j,k,l}$.

命题 5.2。 考虑定理 *5.1* 中的泊松*HIBP* 设置。那么给定 $N_{j,\ l} = n_{j,\ l}, H_l = \lambda$，存在分布等价关系

$$\hat{\sigma}_{j,l}(\lambda) \stackrel{d}{=} \tilde{\sigma}_{j,l}(\lambda) + \sum_{k=1}^{X_{j,l}} S_{j,k,l} \tag{5.9}$$

其中 $\hat{\sigma}_{j,l}(\lambda)$ 的密度为 $\eta^{[n_{j,l}]}(t_j|\lambda, M_j) = \frac{t^{n_j}{}_{,j}{}^{l}\eta^{[0]}(t_j|\lambda, M_j)}{\Xi_j^{[n_{j,l}]}(\lambda, M_j)}$。在 $B_j \sim G(\theta_j, \zeta_j; B_0)$ 的情况下，可得$\hat{\sigma}_{j,l}(H_l) \sim$ 伽马分布$(\theta H_l + n_{j,l}, M_j + \zeta_j)$。

到目前为止，我们尚未讨论 B_0 的具体情形。在这里，我们描述 $B_0 \sim \mathrm{GG}(\alpha_0, \theta_0, \zeta_0, F_0)$ 这一情形的结果，其中GG表示如下所规定的广义伽马分布过程；我们还将考察 $B_j \sim \mathrm{GG}(\theta_j, \zeta_j, B_0)$这一情形。在此之前，对于 $0 < \alpha_j < 1$,令$P_{\alpha_j,0} \sim \mathscr{PY}(\alpha_j, 0, F_0)$表示（$\alpha_j, 0$）Pitman-Yor过程， [13, 25], 该过程由对一个稳定次协调过程进行归一化构造而成。于是，描述 $[n_{j,l}]$ 的随机分区分布的相应EPPF为：

$$p_{\alpha_j,0}(\mathbf{c}_{j,l}) = \frac{\alpha_j^{x_{j,l}-1}\Gamma(x_{j,l})}{\Gamma(n_{j,l})} \prod_{k=1}^{x_{j,l}} \frac{\Gamma(c_{j,k,l} - \alpha_j)}{\Gamma(1-\alpha_j)}.$$

由此可得区块数量的概率 $\mathbb{P}^{(n_{j,l})}{}_{\alpha_j,0}(k) = \frac{\alpha_j^{k-1}\Gamma(k)S_{\alpha_j}(n_{j,l},k)}{\Gamma(n_{j,l})}$，其中$S_{\alpha_j}(n_{j,l}, k)$ 表示第二类广义斯特林数，参见 [28][3.2.3, p. 65-66]。

命题 5.3. 设 $B_0 \sim \mathrm{GG}(\alpha_0, \theta_0, \zeta_0, F_0)$ 表示一个广义伽马过程完全随机测度，其 $0 < \alpha_0 < 1$的 *L´évy* 密度为： $\tau_0(\lambda) = \theta_0\alpha_0\lambda^{-\alpha_0-1}e^{-\zeta_0\lambda}/\Gamma(1-\alpha_0)$ 和 $\Psi_0(t) = \theta_0[(\zeta_0+t)^{\alpha_0} - \zeta_0^{\alpha_0}]$. 那么，给定$\mathbf{Z}_J$:

1. $B_{0,J} \sim \mathrm{GG}(\alpha_0, \theta_0, \zeta_0 + \sum_{j=1}^J \psi_j(M_j))$
2. $H_l|\tilde{X}_l \sim$ 伽马分布$(\tilde{X}_l - \alpha_0, \zeta_0 + \sum_{j=1}^J \psi_j(M_j))$ 和
3. 对于 $\tau_j(s) = \frac{\alpha_j\theta_j e^{-s\zeta_j}s^{-\alpha_j-1}}{\Gamma}(1-\alpha_j)$，对于 $j \in [J]$，$((\tilde{C}_{j,k,l} = c_{j,k,l}, k \in [X_{j,l}], X_{j,l} = x_{j,l})j \in J)$，给定 $(N_{j,l}, j \in [J])$，对于 $x_l = \sum_{j=1}^J x_{j、l}$，当 $x_{j、l} = 0$ 时$n_{j、l} = 0$，否则$x_{j、l} \in [n_{j,l}]$，其分布与.....成正比，

$$\frac{\Gamma(x_l - \alpha_0)}{\prod_{j=1}^J \Gamma(x_{j,l})} \frac{\prod_{j=1}^J \alpha_j\theta_j^{x_{j,l}}(\zeta_j + M_j)^{x_{j,l}\alpha_j}}{(\zeta_0 + \sum_{j=1}^J \theta_j[(\zeta_j + M_j)^{\alpha_j} - \zeta_j^{\alpha_j}])^{x_l - \alpha_0}} \prod_{j=1}^J p_{\alpha_j,0}(\mathbf{c}_{j,l}). \tag{5.10}$$

5.1 对任意 J的简化，以及 Poisson-GG-GG 示例

我们已经得出了相当一般的结果，提供了很大的灵活性。不过，就本文的具体阐述而言，我们希望对模型给出更简洁、更易读的描述与实现，同时确保模型足够丰富，能够满足实际应用需求，并具有显著的分布性质。这并不意味着不能使用更复杂的广义模型；这些模型允许拟合额外的参数，

以及过程的混合匹配。通过命题3.1以及即将在命题5.5中描述的预测规则，能够同时对所有组 J 进行采样，这使得对于所有 $j \in [J]$均满足 $M_j = M$、其中 $M = \max_{j\in[J]} M_j$的假设成为一种合理的简化。在此假设下，客户数少于 M 的组可以被视为拥有 $M - M_j$ 个额外的"幽灵"客户，这些客户未选择任何特征，或许相当于提交了空白文档。此外，对于所有 $j \in [J]$，我们设定$\tau_j(s) = \theta_j \tau(s)$，从而 $\psi_j(M) = \theta_j \psi(M)$，因此 $\sum_{j=1}^{J} \psi_j(M_j) = \psi(M) \sum_{j=1}^{J} \theta_j$，并且 $q_j = \theta_j / \sum_{v=1}^{J} \theta_v$,我们设定$\mathbf{q} := (q_1, \ldots, q_J)$.

5.2 泊松-GG-GG

对于简化的泊松-广义伽马-广义伽马模型，即式（5.10）中的泊松-广义伽马-广义伽马模型，$\alpha_0 = \beta/\alpha$ 对于 $0 < \beta < \alpha$, $\tau_j(s) = \theta_j \frac{\alpha s^{-\alpha-1} e^{-\zeta s}}{\Gamma\ 1-\alpha}$ ( ) 和 $\tau_0(\lambda) = \frac{\beta\lambda^{-\frac{\beta}{\alpha}-1} e^{-\lambda \sum_{j=1}^{J} \theta_j}}{\alpha\Gamma}$ ($_{1-\beta/\alpha}$)，以及 $M_j = M$.对于接下来的结果，需要回顾 [6], 中的 Pitman-Yor有限维分布概念，这也有助于将我们的工作与 [21, 22] 中近期开展的另一系列研究及相关参考文献联系起来。令 $(Q_1, \ldots, Q_J)$ 表示 Ω 的某个不相交划分，使得 $F_0(Q_j) := q_j \cdot$ 那么，对于 P_α, $_\theta \sim \mathscr{PY}(\alpha, \theta, F_0)$，其中 $\theta > -\alpha$ 表示基测度为 F_0 [13, 14, 25], 的双参数Pitman-Yor过程，我们定义其有限维分布 $(P_{\alpha,\theta}(Q_1) \ldots, P_{\alpha,\theta}(Q_J)) := (P_{\alpha,\theta}(q_j)\ j \in [J]) \sim \mathrm{RS}(\mathbf{q}\ \alpha, \theta)$。我们注意到，在纯稳定情形下（ $\tilde{P}_{\alpha,0}(q_j)\ j \in [J]) \overset{d}{=} \left(\frac{\sigma_j q_j^{1/\alpha}}{\sum_{v=1}^{J} \sigma_v q_v^{1/\alpha}}, j \in [J] \right)$，其中 $(\sigma_j, j \in [J])$ 是独立同分布的稳定 α 变量。

备注 5.2. 注意，过程 $\sum_{j=1}^{J} \tilde{P}_{\alpha,\theta(q_j)} \delta_{Y_j}$ 在 $q_j = 1/J$ 时等价于分布律为 $\mathrm{PY}(\alpha, \theta, J, F_0)$ 的 *Pitman-Yor* 多项过程，如 *[21][*命题 *1]*所述。对于一般的 q_j，我们可以将本例中的分布记为 $\mathrm{PY}(\alpha, \theta, \mathbf{q}\ F_0)$。

这将得到一个高度易处理的模型，最多包含 $J+2$ 个参数（$\alpha, \beta, \zeta, q_1, \ldots, q_J$)。为了从 Poisson-GG-GG 的边缘分布中采样：

1. $\varphi \sim$ 泊松$\left(\left(\sum_{j=1}^{J} \theta_j\right)^{\beta/\alpha} [(\zeta + M)^\beta - \zeta^\beta]\right)$。2. $\psi_j(M) = \theta_j[(\zeta + M)^\alpha - \zeta^\alpha]$3. $\tilde{C}_{j,k,l} \overset{iid}{\sim}MtP(M, \tau_j)$，其概率质量函数为 $(1-\frac{\zeta}{\zeta+M})^c (1-\frac{\zeta^\alpha}{(\zeta+M)^\alpha}) \times \frac{\alpha\Gamma(c-\alpha)}{c!\Gamma(1-\alpha)}$ 4. $\tilde{X}_l \sim$MtP$([\sum_j^J \theta_j][(\zeta + M)^\alpha - \zeta^\alpha], \tau_0)$，其概率质量函数为 $(1-\frac{\zeta^\alpha}{(\zeta+M)^\alpha})^{x_l} (1-\frac{\zeta^\beta}{(\zeta+M)^\beta}) \times \frac{\beta\Gamma(x_l-\frac{\beta}{\alpha})}{x_l!\alpha\Gamma(1-\frac{\beta}{\alpha})}$ 5. $X_{j,l}, j \in [J] | \tilde{X}_l = x_l \sim$多项分布$(x_l, q_1, \ldots, q_J)$，对于 $q_j = \theta_j / \sum_{v=1}^{J} \theta_v$6. $(\sum_{k=1}^{X_j} \tilde{C}_{j,k,l}, j \in [J]) \overset{d}{=} (N_{j,l}, j \in [J])$ 的联合概率质量函数适用于 $n_l = \sum_{j=1}^{J} n_{j,l} = 1, 2, \ldots$

$$\frac{(1-\frac{\zeta}{(\zeta+M)})^{n_l}}{(1-\frac{\zeta^\beta}{(\zeta+M)^\beta})} \frac{\beta\Gamma(n_l-\beta)}{n_l!\Gamma(1-\beta)} \times \frac{n_l!}{\prod_{j=1}^{J} n_{j,l}!} \mathbb{E}\left[\prod_{j=1}^{J} \tilde{P}_{\alpha,-\beta}^{n_{j,l}}(q_j)\right] \tag{5.11}$$

˜其中 $(P_{\alpha,-\beta}(q_j)\ j \in [J]) \sim$RS分布$(\mathbf{q}\ \alpha, -\beta)$,

˜因此，$((C_{j,k,l}, k \in [X_{j,l}])\ [X_{j,l}], j \in [J]), (N_{j,l}, j \in [J]))$ 的联合分布为：对于 $x_l = \sum_{j=1}^{J} x_{j,l}$, $\sum_{k=1}^{x_j} c_{j,k,l} = n_{j,l}$, 满足 $x_{j,l} = 0$ 对于 $n_{j,l} = 0$,:

$$\frac{\Gamma\ (x_l - \frac{\beta}{\alpha})\ \Gamma\ (1-\beta)\ \alpha^{x_l-1}}{\Gamma(1-\frac{\beta}{\alpha})\Gamma(n_l-\beta)} \prod_{j=1}^{J} q^{x_j}{}_j{}^{\ l} \prod_{k=1}^{x_j} \frac{\Gamma(c_{j,k,l}-\alpha)}{\Gamma(1-\alpha)} \times \frac{\beta\Gamma(n_l-\beta)}{n_l!\Gamma(1-\beta)} \frac{(1-\frac{\zeta}{(\zeta+M)})^{n_l}}{(1-\frac{\zeta^\beta}{(\zeta+M)^\beta})}. \tag{5.12}$$

关于给定$\mathbf{Z}_J$ 时后验分布的一些相关组成部分如下:

1. $\psi_j(M) = \theta_j[(\zeta+M)^\alpha - \zeta^\alpha]$, $\Psi_0(\sum_{j=1}^J \psi_j(M)) = (\sum_{j=1}^J \theta_j)^{\beta/\alpha}[(\zeta+M)^\beta - \zeta^\beta]$.

2. $B_{j,M}|B_{0,J} \sim \text{GG}(\alpha, \theta_j, \zeta+M; B_{0,J})$，以及 $B_{0,J} \sim \text{GG}(\beta/\alpha, 1, \sum_{j=1}^J \theta_j(\zeta+M)^\alpha; F_0)$。

3. $\sum_{j=1}^J \theta_j(\zeta+M)^\alpha H_l \overset{d}{=} G_{\tilde{X}_l - \frac{\beta}{\alpha}}|\tilde{X}_l \sim$伽马分布$(\tilde{X}_l - \frac{\beta}{\alpha}, 1)$。

4. 向量$(\hat{\sigma}_{j,l}(H_l),\ j \in [J])|(X_{j,l}, j \in [J])$ 在分布上等于

$$\left(\frac{T_\alpha^{(j,l)}(q_j G_{\tilde{X}_l - \frac{\beta}{\alpha}}) + G_{n_{j,l} - X_{j,l}\alpha}}{\zeta + M}, j \in [J] \right). \tag{5.13}$$

其中 $G_{n_{j,l}-X_{j,l}\alpha} \overset{ind}{\sim}$ 伽马分布$(n_{j,l} - X_{j,l}\alpha)$ 和 $T_\alpha^{(j,l)}(y)$ 是相互独立的广义伽马变量，使得 $\mathbb{E}[\mathrm{e}^{-tT_\alpha^{(j,l)}(y)}] = e^{-y[(1+t)^\alpha - 1]}$.

5. 等价地，(5.13) 中的变量可表示为

$$\frac{G_{n_l - \beta}}{\zeta + M} \times \left(R_{X_l} \tilde{P}_{\alpha, \tilde{X}_l\alpha - \beta}(q_j) + \tilde{D}_{j,l}, j \in [J] \right) \tag{5.14}$$

其中 $G_{n_l-\beta} \sim$伽马分布$(n_l - \beta, 1)$ 独立于$(\tilde{P}_{\alpha,\tilde{X}_l\alpha-\beta}(q_j),\ j \in [J]) \sim$RS分布$(\alpha, X_l\alpha - \beta)$，$R_{\tilde{X}_l}$、$(\tilde{D}_{j,l},\ j \in [J]) \sim (X_l\alpha - \beta, n_{j,l} - X_{j,l}\alpha,\ j \in [J])$ 条件地独立于此，$(\tilde{P}_{\alpha,\tilde{X}_l\alpha-\beta}(q_j),\ j \in [J])$ 狄利克雷分布$(\alpha, \ldots, \alpha)$。

6. $(X_{j,l}, j \in [J])|(N_{j,l} = n_{j,l}, j \in [J])$ 根据（B.1）进一步进行采样。

注释 5.3。(5.13) 与(5.14) 之间的分布对应关系源于对 *[30]*[命题 *21]* 以及伽马分布变量和狄利克雷变量通常的贝塔-伽马代数的应用。有关(5.13)中所示形式的表示，另见 *[17]* 。

5.2.1 替代采样器

或者，也可以按如下方式精确抽样 $(N_{j,l},\ j \in [J])$。如果忽略 (5.12) 中的具体分组，则可对 $\mathbf{c}_l := (c_{k,l},\ k \in [x_l])$，使用以下方法，

$$p_{\alpha,-\beta}(\mathbf{c}_l) = \frac{\Gamma(x_l - \frac{\beta}{\alpha})\Gamma(1-\beta)}{\Gamma(1-\frac{\beta}{\alpha})\Gamma(n_l - \beta)} p_{\alpha,0}(\mathbf{c}_l)$$

表示 $[n_l]$ 的$(\alpha, -\beta)$ 划分的 EPPF，对应于从 $\tilde{P}_{\alpha,-\beta} \sim \mathscr{PY}(\alpha, -\beta, F_0)$ 中进行抽样，记为 $(C_{k,l}, k \in [\tilde{X}_l])|N_l = n_l$，其中 $\tilde{X}_l|N_l = \sum_{k=1}^{X_l} C_{k,l} = n_l$对应于 $\mathbb{P}(\tilde{X}_l = x_l|n_l) = \mathbb{P}^{(n_l)}{}_{\alpha,-\beta}(x_l)$。由此，$(C_{k,l}, k \in [X_l])$，$(X_{j,l}, j \in [J])|N_l = n_l$ 的分布为：

$$p_{\alpha,-\beta}(\mathbf{c}_l) \frac{x_l!}{\prod_{j=1}^J x_{j,l}!} \prod_{j=1}^J q_j^{x_{j,l}}.$$

更准确地说，给定 $(C_{k,l}, k \in [X_l], X_l = x_l)|N_l = n_l$、$[n_l]$ 的一个 $(\alpha, -\beta)$ 划分，存在独立同分布的分类变量 $(\mathscr{L}_1, \ldots, \mathscr{L}_{x_l})$，满足 $\mathbb{P}(\mathscr{L}_i = j) = q_j, j \in [J]$，以及配对 $((C_{k,l}, \mathscr{L}_k)\ k \in [x_l])$，其中集合 $((C_{k,l}:\ \mathscr{L}_k = j) \overset{d}{=} (\tilde{C}_{j,k,l}, k \in [X_{j,l}])$ 对于 $(X_{j,l} = \sum_{k=1}^{X_l} \mathbb{I}_{\{\mathscr{L}_k = j\}}, j \in [J])|\tilde{X}_l = x_l, N_l = n_l \sim$多项分布$(x_l, q_1, \ldots, q_J)$。因此，给定 $N_l = n_l$，

$$(\sum_{k=1}^{\tilde{X}_l} C_{k,l}\mathbb{I}_{\{\mathscr{L}_k = j\}}, j \in [J]) \overset{d}{=} (\sum_{k=1}^{X_{j,l}} \tilde{C}_{j,k,l}, j \in [J])$$

服从 $(N_{j,l}=n_{j,l},\ j\in[J])\ |N_l=n_l$，如（5.11）所示，其中$\mathbb{P}(N_l=n_l)\ =(1-\ \zeta^{\beta}/(\zeta+M)^{\beta})^{n_l}\frac{(1-\zeta^{\beta}/(\zeta+M)^{\beta})\beta\Gamma(n_l-\beta)}{n_l!\Gamma(1-\beta)}\cdot$

注记 5.4. 与 (5.12) 对比可知，$\mathbb{E}[\prod_{j=1}^{J}\tilde{P}_{\alpha,-\beta}^{n_{j,l}}(q_j)]$ 是 *[14]*[第 *3.2* 节，式 *(15)*]中所述一般联合矩公式的一个特例。

5.3 $(N_{j,l},j\in[J])$的边缘分布

现在我们给出一般边缘过程的描述。首先注意，$(N_{j,l},j\in[J])|(\hat{\sigma}_{j,l}(H_l)=t_j,j\in[J])$ H_l是 $\mathrm{tP}(t_1M_1,\ldots,t_JM_J)$，因此存在 $(N_{j,l},\hat{\sigma}_{j,l}(H_l)\ j\in[J])$ H_l，的联合分布，其中 $n_l:=\sum_{j=1}^{J}n_{j,l}=1,2,\ldots,$

$$\frac{\prod_{j=1}^{J}M_j^{n_{j,l}}t_j^{n_{j,l}}e^{-t_jM_j}}{(1-\mathrm{e}^{-\sum_{j=1}^{J}M_jt_j})\prod_{j=1}^{J}n_{j,l}!}\frac{(1-\mathrm{e}^{-\sum_{j=1}^{J}M_jt_j})\prod_{j=1}^{J}\eta_j(t_j|\lambda)}{(1-\mathrm{e}^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)})}f_{H_l}(\lambda).$$

对于下一个结果，定义向量 $\mathbf{k_J}=(k_1,k_2,\ldots,k_J)\in\mathbb{N}^J$ 和 $\mathbf{n}_l=(n_{1,l},\ldots,n_{J,l})$。其中，$k_j(n_{j,l})=0$ 当 $n_{j,l}=0$ 时成立，而 $k_j(n_{j,l})\in[n_{j,l}]$ 在其他情况下成立。将所有此类向量 $\mathbf{k_J}$ 组成的集合记为 $\mathcal{K}_J(\mathbf{n}_l)$。

现在回顾恒等式，$\int_0^{\infty}t^{n_{j,l}}e^{-tM_j}\eta_j(t|\lambda)\ dt=e^{-\lambda\psi_j(M_j)}\ \Xi_j^{[n_{j,l}]}(\lambda,\ M_j)$，以便用于以下结果。

命题 5.4。对于 $n_l=\sum_{j=1}^{J}n_{j,l}=1$、2、.....、，以及每个$l\in[\varphi],\ \varphi\sim$泊松$(\Psi_0(\sum_{j=1}^{J}\psi_j(M_j)))$，以下给出了 $(N_{j,l}=n_{j,l},j\in[J])$ 的边缘分布和条件分布的描述。

1. $(N_{j,l}=n_{j,l},j\in[J])\ |H_l=\lambda$ 的条件分布可表示为：

$$\frac{e^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)}\prod_{j=1}^{J}M_j^{n_{j,l}}\Xi_j^{[n_{j,l}]}(\lambda,M_j)}{(1-e^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)})\prod_{j=1}^{J}n_{j,l}!}$$

2. $(N_{j,l}=n_{j,l},j\in[J])$ 的边缘分布具有以下求和表示，其中$x_l:=\sum_{j=1}^{J}x_{j,l}$，并且 $\mathbf{x}_{j,l}=(x_{j,l},j\in[J])$：

$$\frac{\sum_{\mathbf{x}_{j,l}\in\mathcal{K}_J(\mathbf{n}_l)}\Psi_0^{(x_l)}(\sum_{j=1}^{J}\psi_j(M_j))\prod_{j=1}^{J}p^{[n_{j,l}]}(k_j|\tau_j,M_j)\Xi_j^{[n_{j,l}]}(1,M_j)M_j^{n_{j,l}}}{\Psi_0\left(\sum_{j=1}^{J}\psi_j(M_j)\right)\prod_{j=1}^{J}n_{j,l}!}$$

其中 $\Psi^{(x_l)}{}_0(\sum_{j=1}^{J}\psi_j(M_j))=\int_0^{\infty}\lambda^{x_l}e^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)}\tau_0(\lambda)d\lambda$ 可见于 *[26]*。此外，$p^{[n_{j,l}]}(\mathbf{c}_{j,l}|\tau_j,M_j)\Xi_j^{[n_{j,l}]}(1,M_j)=\prod_{k=1}^{x_{j,l}}\psi_j^{(c_{j,k,l})}(M_j)$。

5.4 预测规则

我们给出（$Z^{(M_j+1)}{}_j,j\in[J])|\mathbf{Z}_J$,每个组中一位客户所组成的向量的预测规则。其他描述可利用定理 5.1 和命题 3.1 得出。

命题 *5.5.* 对每个 $j\in[J]$，$Z^{(M_j+1)}{}_j\ \ |B_j\sim PoiP(B_j).B_j\sim$完全随机测度$(\tau_j,B_0)$ $B_0\sim$完全随机测度(τ_0,F_0)，则 $(Z^{(M_j+1)}{}_j,j\in[J])|\mathbf{Z}_J$的预测分布与 $(\tilde{Z}^{(M_j+1)}+\sum_{l=1}^{r}\mathscr{P}_{jl}^{(M_j+1)},(\sum_{k=1}^{X_j}S_{j,k,l})\delta_{\tilde{Y}_l},j\in[J])$ 的预测分布相同，其中 $\mathscr{P}^{(M_j+1)}{}_{j,l}(\lambda)\ \overset{ind}{\sim}$泊松$(\lambda)$，

$\tilde{Z}_j^{(M_j+1)}|B_{0,J}+\sum_{l=1}^r H_l\delta_{\tilde{Y}_l}$ ~印度自助过程(C_j, $\tau^{(M_j)}j$, $B_{0,\ J}+\sum_{l=1}^r H_l\delta_{Y_l}$) C_j ~泊松(s)。因此，给定 $\mathbf{Z}_J$,

$$(\tilde{Z}_j^{(M_j+1)}, j\in[J]) \overset{d}{=} (\sum_{v=1}^{\varphi^*}[\sum_{k=1}^{X^*_{j,l}}\tilde{C}^{(*,1)}_{j,k,v}]\delta_{Y_v}+\sum_{l=1}^{r}\left[\sum_{k=1}^{\mathscr{P}^*_{j,l}(H_l)}\tilde{C}^{(*,2)}_{j,k,l}\right]\delta_{\tilde{Y}_l}, j\in J) \tag{5.15}$$

其中 $(Y_v)\overset{iid}{\sim}F_0$, $\varphi^*\sim$泊松($[\Psi_0(\sum_{j=1}^J\psi_j(M_j+1))-\Psi_0(\sum_{j=1}^J\psi_j(M_j))]$),

1. $X_l^*:=\sum_{j=1}^J X^*_{j,l}\sim$MtP分布($\sum_{j=1}^J[\psi_j(M_j+1)-\psi_j(M_j)],\tau_{0,J}$) 概率质量函数为 $\frac{\sum_{j=1}^J[\psi_j(M_j+1)-\psi_j(M_j)]^{x_l}\Psi_0^{(x_l)}(\sum_{j=1}^J\psi_j(M_j+1))}{x_l![\Psi_0(\sum_{j=1}^J\psi_j(M_j+1))-\Psi_0(\sum_{j=1}^J\psi_j(M_j))]}$

2. $(X^*_{j,l}, j\in[J])|X_l^*=x_l^*\sim\mathrm{Multi}(x_l^*;q_1,\ldots,q_J)$, *for* $q_j=\frac{[\psi_j(M_j+1)-\psi_j(M_j)]}{\sum_{v=1}^J[\psi_v(M_v+1)-\psi_v(M_v)]}$.

3. 对每个 $j\in[J]$, $(\tilde{C}^{(*,1)}_{j,k,v},,)$和 $(\tilde{C}^{(*,2)}_{j,k,l},,))$均为独立同分布的 $\mathrm{MTP}(1,\tau_{(M_j)j})$随机变量，其概率质量函数为 $\frac{\psi_j^{(c)}(M_j+1)}{[\psi_j(M_j+1)-\psi_j(M_j)]c!}$

4. $\mathscr{P}_l^*(H_l)=\sum_{j=1}^J\mathscr{P}^*_{j,\ l}(H_l)\sim$泊松($H_l\sum_{j=1}^J[\psi_j(M_j+1)-\psi_j(M_j)]$)

5. $(\mathscr{P}^*_{j,\ l}(H_l)\ j\in[J])|\mathscr{P}_l^*(H_l)=n_l^*\sim$多项分布($n_l^*$; q_1, ..., q_J)

6. H_l, $(X_{j,\ l},j\in[J])|\mathbf{Z}_J$，如定理 *5.1* 所示。

如果希望增加组， J 到 $J+1$，如[37] 所示，预测规则如下

推论 *5.1.* 假设 $Z^{(1)}_{J+1}|B_{J+1}\sim$ *PoiP*(B_{J+1})且 $B_{J+1}\sim$ 完全随机测度(τ_{J+1}, B_0)。则$Z^{(1)}_{J+1}|\mathbf{Z}_J$ 在分布上等于 $\tilde{Z}^{(1)}_{J+1}$，可表示为（*A.5*）的单变量版本，其中以$M_{J+1}=0$替代 M_j、$X^*_{J+1,\ l}\sim$ *MtP*(ψ_{J+1}(1) $\tau_{0,J}$)，并且$\mathscr{P}^*_{J+1,\ l}(H_l)\sim$泊松($H_l\psi_{J+1}$(1))。将 φ^* 替换为 $\varphi^*_{J+1}\sim$泊松($\Psi_0($ ψ_{J+1}(1)$+\sum_{j=1}^J\psi_j(M_j))-\Psi_0(\sum_{j=1}^J\psi_j(M_j))$。要获得更简单的规则，即 *[37]*,其中 $\mathbf{Z}_J:=(Z_j^{(1)}$, $j\in[J])$，令$(M_j=1, j\in[J])$。

5.4.1 预测规则 Poisson-GG-GG

对于第5.2节中的 Poisson-GG-GG 模型，回忆一下 $M_j=M$，并令 NB(p, t) 表示参数为 (p, t) 的负二项分布变量，我们有

1. $\varphi^*\sim$泊松($(\sum_{j=1}^J\theta_j)^{\beta/\alpha}[(\zeta+M+1)^\beta-(\zeta+M)^\beta]$)，且对每个 $l\in[\varphi^*]$，2. $\mathscr{P}_l^*(H_l)\sim$负二项分布($(\zeta+M)^\alpha$ $(\zeta+M+1)^\alpha$, $\tilde{X}_l-\frac{\beta}{\alpha}$) 3. $\mathscr{P}^{(M_j+1)}_{j,l}(\sum_{k=1}^{X_j,l}S_{j,k,l})$ $\sim$ 负二项分布 ($(\overline{\zeta+M})$ $(\overline{\zeta+M+1})$ $n_{j,l}-X_{j,l}\alpha$)

---

4. $X_l^*\sim\mathrm{MtP}([\sum_{j=1}^J\theta_j[(\zeta+M+1)^\alpha-(\zeta+M)^\alpha],\tau_{0,J})$ with pmf $\frac{(1-\frac{(\zeta+M)^\alpha}{(\zeta+M+1)^\alpha})^{x_l}}{(1-\frac{(\zeta+M)^\beta}{(\zeta+M+1)^\beta})}\times\frac{\beta\Gamma(x_l-\frac{\beta}{\alpha})}{x_l!\alpha\Gamma(1-\frac{\beta}{\alpha})}$

5. ($\tilde{C}^{(*,1)}_{j,k,v},,$) 和 ($\tilde{C}^{(*,2)}_{j,k,l},,$)) 是 MtP分布(1, $\tau_j^{(M)}$) 变量的独立同分布（iid）集合，其概率质量函数为 $\frac{\alpha\Gamma(\zeta+M+1-\alpha)^{\alpha-c}}{c!\Gamma(c-\alpha)(1-\alpha)}$，且不依赖于 $j\in[J]$.

6. 参照 (5.12)，从与下式成比例的分布中抽取 $(X_{j,l},j\in[J])|(N_{j,l}=n_{j,l},j\in[J])$,$\Gamma(x_l-\frac{\beta}{\alpha})\alpha^{J-1}(1-\frac{\beta}{\alpha})(x_l)\Gamma(x_l)\prod_{j=1}^J\Gamma(x_{j,l})\prod_{j=1}^J q_j^{x_{j,l}}\mathbb{P}_{\alpha 0}^{(n_{j,l})},(x_{j,l})$

5.5 在泊松 HIBP 情形下恢复潜变量并耦合 HIBP

虽然命题 5.2 表明，在定理 5.1 中无需恢复个体变量（$\tilde{C}^{(i)}_{j\,k\,\cdot,l}$），但如前所述，它们在构造耦合HIBP方面具有实用价值，从而能够恢复在这些情况下描述后验分布所必需的潜变量；当然，对于任何关注个体信息恢复的应用也是如此。我们还注意到，从计算角度来看，纳入˜对（$C_{j,kl}$）进行采样可能更加容易。特别是，可以采用以下方案来恢复这些变量。

给定 ($N_{j,l} = n_{j,l}, j \in [J], H_l = \lambda$) 或 ($N_{j,l} = n_{j,l}, j \in [J]$)：

˜1. 按照命题 5.1 和式 (5.1) 中的描述，或按照定理 5.1，从其条件分布中抽取 ($C_{j,k,l}$, $k \in [X_{j,l}], X_{j,l}$)。

2. 然后，($\tilde{C}^{(i)}_{j\,k\,\cdot,l}, i \in [Mj]|\tilde{C}_{j,k,l} = c_{j,k,l}$) $\sim$多项分布($c_{j,k,l}$, $\frac{1}{M_j}, \ldots, \frac{1}{M_j}$)。这会将抽样得到的 $\tilde{C}_{j,k,l}$ 分配到各个 $\tilde{C}^{(i)}_{j\,k\,\cdot,l}$中。

3. 此外，从与 $s^{c_{j,k,l}} e^{-sM_j} \tau_j(s)$ 成正比的密度中抽取 $S_{j,k,l}|\tilde{C}_{j,k,l} = c_{j,k,l}$.

4. 耦合伯努利层次印度自助过程的后验量可通过 $\sum_{k=1}^{X_{j,\cdot,l}} \mathbb{I}_{\{\tilde{C}^{(i)}_{j,k,l}>0\}}$ 以及其他 HIBP，利用 $\mathbb{P}A_j(da|\pi^{-1}(1 - \exp(-S_{j,k,l})))$ 构造，如命题 3.2 所示。

6 模拟数据实验

我们已在第 5.2 节和第 5.4.1 节证明，简化的泊松 HIBP 包含与 [21, 22] 中出现的 Pitman-Yor 多项分布变体相对应的组成部分。读者可以参考这些研究，了解有趣的真实数据应用；这些应用清楚地展示了我们研究结果在实践中的广泛适用性，即使是在最简单的情形下。作为副产品，我们还提出了从这类过程中采样的新方法。这也应能启发人们进一步探索涉及我们潜在特征框架的应用。

本节中，我们针对命题 5.3 所述的更一般形式的泊松-GG-GG 提供模拟，并给出一种用于后验推断的采样算法。具体而言，我们介绍具有以下参数化形式的泊松-GG-GG 层次印度自助过程。

$$\tau_0(\lambda) = \mathrm{GG}(\lambda|\alpha_0, \theta_0/\alpha_0, \zeta_0) = \frac{\theta_0}{\Gamma(1-\alpha_0)} \lambda^{-\alpha_0-1} e^{-\zeta_0 \lambda}, \quad B_0 \sim \mathrm{CRM}(\tau_0, F_0),$$

$$\tau_j(s) = \mathrm{GG}(s|\alpha_j, \theta_j/\alpha_j, \zeta_j) = \frac{\theta_j}{\Gamma(1-\alpha_j)} s^{-\alpha_j-1} e^{-\zeta_j s}, \quad B_j \,|\, B_0 \sim \mathrm{CRM}(\tau_j, B_0) \text{ for } j \in [J],$$

$$Z_j^{(i)} \,|\, B_j \sim \mathrm{PoiP}(B_j) \text{ for } i \in [M_j], j \in [J].$$

在正文中，我们介绍了模拟及基于该模拟的合成文档分类结果。在补充材料的 C 节中，我们介绍了所有细节，另请参见命题 5.3，其中包括边缘分布、预测分布的推导，以及用于参数估计的吉布斯抽样步骤。

6.1 模拟

我们首先将模型生成的词频矩阵可视化，以观察模型参数如何影响生成的计数矩阵的整体模式。生成的词频矩阵如图1所示。对于由较大的 α_0 值生成的词频矩阵，这些值会促使词频呈现更强的幂律分布，因此词

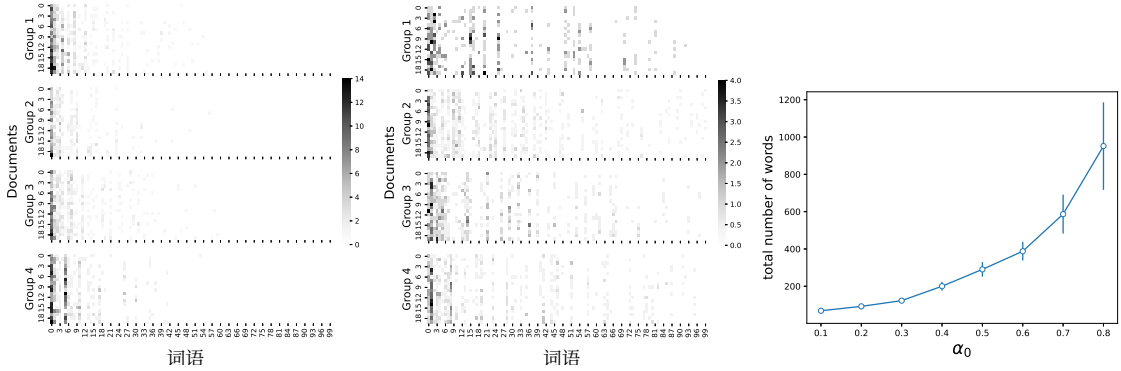

图1：（左、中）由参数 $J = 4$, $\theta_0 = 10$, $\zeta = 1$, $M_j = 20$, $\theta_j \sim$均匀分布(2，4), $\alpha_j \sim$均匀分布(0。2，0。5),$\zeta_j = 1$和 $\beta_j = 1$ 生成的 Poisson-GG-GG HIBP模型词频矩阵，适用于 $j \in [J]$。左侧矩阵由 $\alpha_0 = 0.2$ 生成，中间矩阵由 $\alpha_0 = 0.7$生成。仅展示排名前 100 的词语。（右）不同 α_0 值下各组的平均词语总数。我们针对每个 α_0 值生成 10 个矩阵并计算平均值。

频总数往往更多（图1右侧）。此外，这些矩阵呈现出更多样的词语占用模式，组间重叠更少。

6.2 参数的后验推断

给定一组观测文档 $\mathbf{Z}_J = ((Z_j^{(i)}, i \in [M_j])\ j \in [J])$，我们可能希望推断参数 $\phi := (\alpha_0, \zeta_0, \theta_0, ((\alpha_j, \zeta_j, \theta_j, \beta_j)\ j \in [J])$。为计算参数估计所需的边缘分布，我们可能需要 $\varphi = r$、$(X_j, \iota = x_j, \iota, j \in [J], l \in [r])$、$(\widetilde{C}_{j\,k}^{(i)}, \iota = c_{j\,k}^{(i)}, \iota, i \in [M_j], k \in [x_{j,l}], j \in [J], l \in [r])$，或聚合词频 $\widetilde{N}_j^{(i)}, \iota = \sum_{k=1}^{X_j} \widetilde{C}_{j\,k}^{(i)}, \iota$。在我们的实验中，我们使用 $(\varphi, N_{j,l}, X_{j,l})$ 的边缘分布来估计参数，该分布可由正文中的结果推导得出。Poisson-GG-GG HIBP模型的推导见 C 节，其中还给出了用于参数估计的吉布斯采样算法。注意，我们应恢复计数 $(X_{j,l})$，这可以根据定理 5.1 完成。

作为示例，我们生成文档 $\mathbf{Z}_J$，并将真实参数值设为$\theta_0 = 10, \zeta = 1, J = 4$, $M_j = 200, \theta_j \sim$均匀分布(2.0,4.0) $\alpha_j \sim$均匀分布(0.1, 0.6) $\zeta_j = 1, \beta_j =$1（对于 $j \in [J]$），以及 $\alpha_0 \in \{0.2, 0.6\}$。我们将 ζ_0、$(\zeta_j, j \in [J])$固定为 1，并推断 ϕ:=$(\theta_0, \alpha_0, ((\theta_j, \alpha_j)$ $j \in [J])$。在后验推断中，我们为所有参数设置非信息先验，例如为 θ_0 设置伽马分布(1, 1)，为 α_0设置 logit $\mathcal{N}$(0, 1)。详细的先验设置请参见补充材料中的 C.II 节。

对于两个数据集（一个使用 $\alpha_0 = 0.2$，另一个使用 $\alpha_0 = 0.6$），我们分别运行三条独立的 MCMC链，每条链运行 30,000 步，其中 15,000 个样本作为预热样本。图 2 展示了 θ_0 和 α_0 的后验分布样本。其余参数的结果见补充材料中的交叉验证部分。总体而言，所提出的 Gibbs 抽样器正确识别了集中在真实值附近的后验分布。

6.3 预测与分类

我们进一步展示了如何使用我们的模型，根据文档所属的组对文档进行分类。给定一组文档 $\mathbf{Z}_J$，可以使用命题 5.5 模拟或计算未来文档的预测概率。具体来说，对于一篇新文档 Z_*，对于每个 $j \in [J]$，我们计算 Z_*由组 j生成的概率 $\mathbb{P}(Z_* = Z^{(M_j+1)}{}_j \quad | \mathbf{Z}_J)$。

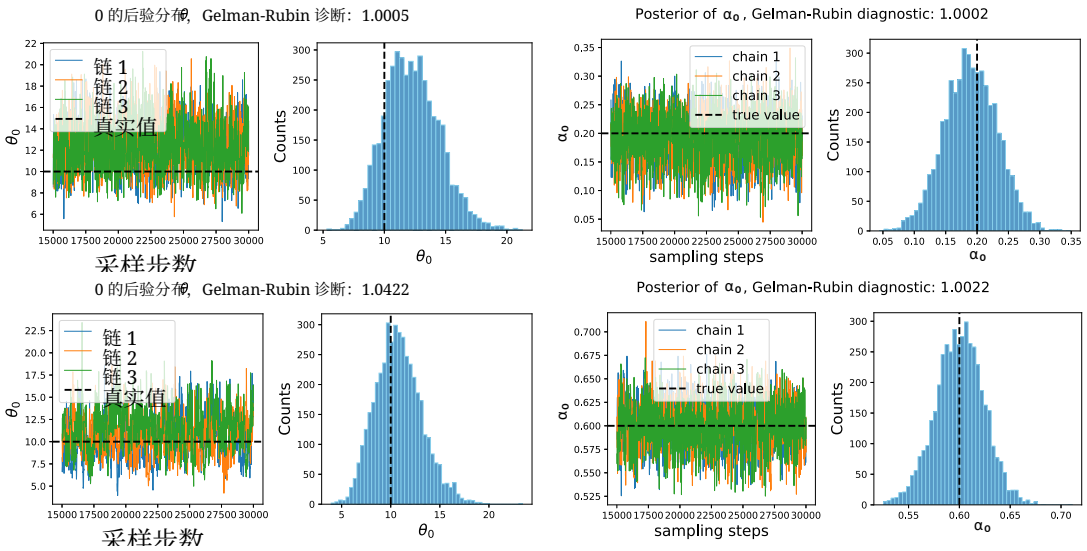

图 2： θ_0 和 α的后验分布样本。顶行显示使用 $\alpha_0 = 0.2$ 生成的数据的结果，底行显示使用 $\alpha_0 = 0.6$生成的数据的结果。我们展示了 Gelman-Rubin收敛诊断 $[4, 10]$; Gelman-Rubin 诊断值明显小于 1.2，表明已收敛。

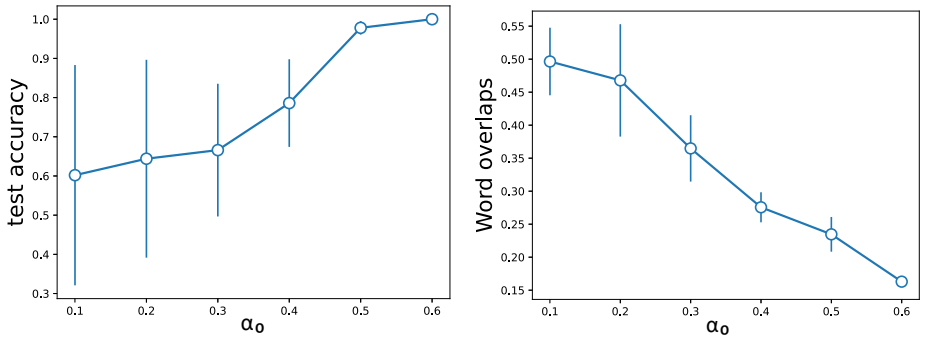

图3：左：针对以不同 α_0 值生成的测试文档的平均组分类准确率。右：针对以不同 α_0 值生成的数据，跨组的词语平均重叠度（定义见正文）。

然后，我们可以将 Z_* 分类到使 $\arg\max_j \mathbb{P}(Z_* = Z^{(M_j+1)}{}_j$ $|\mathbf{Z}_J)$ 取值的组中。

作为示例，我们根据上一节描述的模型生成训练文档$\mathbf{Z}_J$，并使用命题 5.5 中描述的预测规则，为每个组进一步生成 100 个测试文档。关于 Poisson-GG-GG HIBP 模型的预测规则的详细说明，请参见 C.III 节。随后，我们使用训练数据运行后验推断算法，收集估计参数，并计算每个测试文档属于 J 个组之一的预测概率。与训练过程类似，计算预测概率也需要处理聚合计数。在 C.IV 节中，我们给出一种吉布斯采样算法，用于恢复计算预测概率所需的潜变量。

除 α_0 外，参数设置与第 6.2 节相同，我们生成了 $\mathbf{Z}_J$，其中 $\alpha_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$。对于每个 α_0，我们生成五个数据集，总计 30 个数据集。（图 3，左）汇总了针对用于生成数据的 α_0 值计算的平均测试组分类准确率。我们发现，使用较大的 α_0 值生成的文档更容易分类；这可能是因为使用较大的 α_0 值生成的文档不太可能与其他组共享词语，因此具有各自独有的词语。为了定量

考察这一点，我们测量了各组之间词语的平均重叠度，

$$\mathrm{Overlap}(\mathbf{Z}_J \mid \varphi = r, ((X_{j,l} = x_{j,l}))) = \frac{1}{\binom{J}{2}} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{l=1}^{r} \frac{\mathbb{1}_{\{x_{j,l}>0 \wedge x_{j',l}>0\}}}{r},$$

即所有组对之间共同出现的词语所占的平均比例。（图3，右）展示了每个 α_0在五个数据集上的Overlap($\mathbf{Z}_J$)平均值。随着 α 增加，平均重叠度降低，表明来自不同组的文档变得更容易区分。

A 证明与讨论

A.I 第2节和第3节中的结果

命题 2.1，正如正文中所提到的，只是对 [16][第3]节结果的重新表述。然而，在将L´évy密度分解为 $\rho_j(s) = [1-\pi_j(s)]^{M_j}\rho_j(s) + (1-[1-\pi_j(s)]^{M_j})\rho_j(s)$，之后，读者不难直接核查所有内容，其中条件 $\Phi_j(M_j) = \int_0^\infty (1-[1-\pi_j(s)]^{M_j})\rho_j(s)ds < \infty$，诱导出$\mu_j$，的分解，其中有限分量诱导出一个基于分量$(\xi_j,\ S_{j,\ \ell},\ \omega_{j,\ \ell},\ \ell \in [\xi_j])$ 的复合泊松表示，如命题 2.1所述，客户将从中进行抽样。其余部分，正如所解释的，是对联合分布、条件分布和边缘分布的处理，其中涉及随机向量。

注 A.1. 需要注意的是，这些结果依赖于跳跃$(S_{j,\ell})$ 的唯一性，因此也依赖于唯一的数对$(S_{j,\ell}, \omega_{j,\ell})$，无论$(\omega_{j,\ell}, \ell \in [\xi_j])$ 是否存在并列值。

命题 2.2 和 2.3 也可作为 [16][第5节命题 5.1 和 5.2], 的特例得出。考虑一个长度为 J 的多元 IBP 过程，即 $Z_0^{(1)} = \mathbf{X}$ ，其基于单个单变量完全随机测度(τ_0, F_0) $\mu_0 := B_0$；此处左侧的记号在其他方面与[16][第5节命题 5.1 和 5.2]中的相应量一致。注意，F_0 在该文中扮演记作 B_0 的变量的角色。按照零集所决定的，τ_0分解，并适当地应用贝叶斯法则，可以更直观地看出这些结果。这里用到的是，对于 $j \in [J]$，$(\xi_1, \ldots, \xi_J)|\lambda$ 是条件独立的泊松$(\Phi_j(M_j)\lambda)$变量，使得

$$\mathbb{P}(\xi_1 = 0, \ldots, \xi_J = 0|\lambda) = e^{-\lambda\sum_{j=1}^J \Phi_j(M_j)}$$

表示尖峰分布，它决定了根据过程 X 抽取特征的区域或不抽取特征的区域，从而得到 $\tau_0(\lambda) = e^{-\lambda\sum_{j=1}^J \Phi_j(M_j)}\tau_0(\lambda) + (1-e^{-\lambda\sum_{j=1}^J \Phi_j(M_j)})\tau_0(\lambda)$ 的分解，该分解确定变量 H_l、φ 和 B_0、J. 的分布。接下来使用多元零截断泊松分布 $\mathrm{tP}(\lambda(\Phi_1(M_1)\ldots, \Phi_J(M_J)))$，使得对于 $l \in [\varphi]$，$(X_{j,l} = x_{j,l}, j \in [J])$ H_l 的联合分布为 $x_l := \sum_{j=1}^J x_{j,l} = 1, 2, \ldots,$

$$\frac{\prod_{j=1}^J \Phi_j^{x_{j,l}}(M_j)\lambda^{x_{j,l}}e^{-\lambda\Phi_j(M_j)}}{(1-e^{-\lambda\sum_{j=1}^J \Phi_j(M_j)})\prod_{j=1}^J x_{j,l}!} f_{H_l}(\lambda).$$

然后利用这一点得到适当的联合分布、边缘分布和后验分布，类似于命题2.1中的讨论。再次强调，供后续使用的一个关键观察是，˜变量（H_l，X_l）可以按照 [26] 中的方式解释。

A.II 定理 3.1 的证明

此处我们利用如下事实：给定 $B_0(Z_j^{(i)}, i \in [M_j], j \in [J])$ 与 $(\sum_{l=1}^\infty [\sum_{k=1}^{\xi_{j,l}} \tilde{A}_{j,k,l}^{(i)}]\delta_{Y_l}, i \in [M_j], j \in [J])$ 同分布，后者可视为一个多元 IBP，其零集事实上与过程 $\mathbf{X}$. 的零集相同。因此，τ_0 具有与命题 2.2 中相同的分解。随后，对于每个 l,，可以像命题 2.2 的情形那样，处理 $(\sum_{k=1}^{\xi_{j,\ l}} \tilde{A}_{j\ k}^{(i),,l,} i \in [M_j],$ $j \in [J])$ 和 $(\xi_{j,l}, j \in [J])|\lambda$的联合分布。接着，像命题 2.2 中那样对 $\mathbf{X}$ 的分布进行混合，即可完成证明。实际上，该结果是应用 [16][命题 5.2], 得到的；不过，不必精确求出这些和 $(\sum_{k=1}^{\xi_{j,\ l}} [\tilde{A}_{j,k,l}^{(i)}],$ $[M_j])$ 的条件分布，因为它们可能相当复杂。

□

关于第 3.1 节中的结果，命题 3.1 的详细内容可由 [16][第 4 节2], 中的非层次泊松设定以及 [26], 耦合的相应描述得出。

与命题 2.2一致。命题3.2直接源于基于完全随机测度的莱维密度 ρ_j 到τ_j 的适当变换，即通过所示映射对跳跃进行变换。

备注 *A.2.* 原则上，定理 *3.1* 的证明中基于 *[16][Section5]*，的设定足以通过考察 $(\sum^{X_j}$, $_{k=1}^{l}\tilde{A}_{j\,k}^{(i)}$, ,$_l$, $\tilde{Y}_l$, $l\in[r]$, $\varphi=r$, $i\in[M_j]$, $j\in J)$ 和 $(H_l, l\in[r])$ 的联合分布，并使用贝叶斯法则，得到 $B_0|(Z_j^{(i)}, i\in[M_j], j\in[J])$ 的后验分布表达式，从而获得给定 $(\sum^{X_j}$, $_{k=1}^{l}\tilde{A}_{j\,k}^{(i)}$, ,$_l$, $l\in[r]$, $i\in[M_j]$, $j\in J)$ 时跳跃 H_l, $l\in[r])$ 的后验分布，因为否则 $B_{0,J}$ 的分布保持不变。然而，正如正文中所述，我们寻求的是便于实际采样并有助于理解其结构性质的、可处理的后验分布描述。虽然得到的描述是正确的，但若没有关于这些和的分布的更多信息，它仍达不到这一标准。正如我们在泊松 *HIBP* 情形中所展示的，还需要进一步细化。

A.III 定理 4.1 的证明

如备注 A.2中所述，使用 [16][章节 5] 来获得$B_0|\left(\sum_{k=1}^{X_{j,l}}\tilde{A}_{j,k,l}^{(i)}, \tilde{Y}_l, l\in[r], \varphi=r, i\in[M_j], j\in[J]\right)$的描述是正确的；无论如何，$(H_l)$的后验分布通过$(X_{j,l})$依赖于这些和。相反，我们使用包含个体变量 $\left(X_{j,l}, (A_{j\,k}^{(i)}, ,_l, k\in[X_{j,l}], l\in[r], i\in[M_j], j\in[J])\right)$的扩展空间。显然，在给定这些信息的情况下， B_0 具有˜命题 2.3中给出的后验分布。给定 B_0 及这些变量，对于每个 $j\in[J]$，我们可以将$(S_{j,k,l}, Y_l, k\in[X_{j,l}], l\in[\varphi])$与$((S_{j,\ell}, \omega_{j,\ell})\ \ell\in[\xi_j])$等同，并在命题 2.1中进行适当代入。进一步注意到，在给定 B_0 及这一额外信息的情况下，μ_{j,M_j}是CRM($\rho^{(M_j)}$ B_0)，如命题

$$\begin{array}{ccccc} 2 & 1 & B_0 & X_{j\,l} & {}_{k\,l}\ k\in[X_{j,l}]\ l\in \\ [r]\ i\in[M_j] & j\in[J] & & \mu_{j\,M_j} & j\in[J] \end{array}$$

中所述。。然后，我们利用给定(, , $(A_j^{(i)}$,,, , ,,)的分布来证明(, ,)的联合分布是逐分量的CRM($\rho^{(M_j)}$ B_0, $_J+\sum_{l=1}^{r}H_l\delta_{\tilde{Y}_l}$)，从而$\mu_{j,M_j}\overset{d}{=}\tilde{\mu}_{j,M_j}+\sum_{l=1}^{r}\tilde{\sigma}_{j,l}(H_l)\delta_{\tilde{Y}_l}$。这些论证给出了定理 4.1的陈述 1和2中的表示。该结果的陈述 3则补充了相应的描述。

□

A.IV 第5节中泊松 HIBP 的结果

现在我们建立第5节中的结果。我们不使用定理 4.1，而是利用泊松变量的求和性质，并直接利用$\sigma_{j,l}(\lambda)$。的无限可分性，以一种更直接的方法处理定理 5.1及相关结果。我们首先总结泊松 HIBP的一些性质/表示，这些性质/表示展现了其混合泊松表示，并将$(B_j, j\in[J])$ B_0)描述为如 [16]中的多元完全随机测度。正如我们将展示的，利用这些表示，可以从 [16][Section5]得到另一种后验分析。然而，可处理性仍然需要使用增广。在这里，我们再次使用记号 $\mathscr{P}(\lambda)\sim$ 泊松(λ)

关于 $Z_j^{(i)}|B_j\sim$PoiP(B_j) $B_j|B_0\sim$完全随机测度(τ_j, B_0) $B_0\sim$完全随机测度(τ_0, F_0) 的事实：

1. 令 $\eta_j(t_j|\lambda)$ 表示 $\sigma_{j,l}(\lambda)=\sum_{k=1}^{\infty}s_{j,k,l}$ 的密度，其对应的拉普拉斯指数为 $\lambda\psi_j(t)$，且 ($s_{j,k,l})\{k\geq 1\}$ 由 $\lambda\tau_j(s)$。 确定。

2. $(B_j\in[J], B_0)$ 是 [16][第 5] 节意义下的多元完全随机测度，其联合 Lévy 密度为 ´evy density, $\prod_{j=1}^{J}\eta_j\ t_j|\lambda\ \tau_0\ \lambda$, $\sum_{l=1}^{\infty}\sigma_{j\,l}\ \lambda_l\ \delta_{Y_l}\ j\in(\ )\ (\ )$，并且与 (,( ),,) 分布等价。 $[J]\ \sum_{l=1}^{\infty}\lambda_l\delta_{Y_l}$

3.$\mathbf{Z}_J:=(Z_j^{(i)}, i\in[M_j], j\in[J])$在分布上等于

$$(\sum_{l=1}^{\infty}\mathscr{P}_{j,l}^{(i)}(\sigma_{j,l}(\lambda_l))\delta_{Y_l}, i\in[M_j], j\in[J]), \tag{A.1}$$

where $\mathscr{P}_{j,l}^{(i)}(\sigma_{j,l}(\lambda_l)) = \sum_{k=1}^{\infty} \mathscr{P}_{j,k,l}(s_{j,k,l}) \sim \text{Poisson}(\sigma_{j,l}(\lambda_l))$

4. $(B_j \in [J],\ B_0)|\mathbf{Z}$ 的后验分布仅取决于求和过程中的信息

$$(\sum_{l=1}^{\infty} \mathscr{P}_{j,l}(M_j \sigma_{j,l}(\lambda_l))\delta_{Y_l}, j \in [J])$$

其中 $\mathscr{P}_{j,\ l}(M_j\sigma_{j,\ l}(\lambda_l)) = \sum_{i=1}^{M_j} \mathscr{P}_{j\,l}^{(i)}{}_{,}\ (\sigma_{j,\ l}(\lambda_l))$。

我们现在仅利用该过程观测到的总和，来描述泊松过程的相关组成部分；再次采用 [16][Section 5], 中的框架，但旨在通过分解方法给出更清晰的说明。以 $\sigma_{j,l,}$ 表示的求和表示有助于采用这一方法。对于每个固定的 $l,$，给定 $(\sigma_{j,l}(\lambda) = t_j, j \in [J])\ \lambda_l = \lambda$，下面的零集由下式给出。

$$\mathbb{P}(\mathscr{P}_{1,l}(M_1 t_1) = 0, \ldots, \mathscr{P}_{J,l}(M_J t_J) = 0) = \mathrm{e}^{-\sum_{j=1}^{J} M_j t_j}$$

接下来的结果如下：

1. 事件 $\{l\colon\ \mathscr{P}_{j\,l}^{(i)}{}_{,}(\sigma_{j,l}(\lambda_l)) = 0)\ i \in [M_j], j \in [J]\}$ 诱导出 $(B_j, j \in [J], B_0)$ 的联合列维密度分解如下：

$$e^{-\sum_{j=1}^{J} M_j t_j} \prod_{j=1}^{J} \eta_j(t_j|\lambda)\tau(\lambda) + \Psi_0(\sum_{j=1}^{J} \psi_j(M_j))\varrho(t_1, \ldots, t_j|\lambda) f_H(\lambda). \tag{A.2}$$

2. 联合密度 $\varrho(\mathbf{t}|\lambda) f_H(\lambda)$ 是适当的，其中：

$$\varrho(\mathbf{t}|\lambda) = \frac{(1 - e^{-\sum_{j=1}^{J} M_j t_j}) \prod_{j=1}^{J} \eta_j(t_j|\lambda)}{(1 - e^{-\lambda \sum_{j=1}^{J} \psi_j(M_j)})}. \tag{A.3}$$

3. 联合密度表示变量 $(\hat{}\sigma_{j,\ l}(H_l)\ j \in [J],\ H_l)$ 的联合密度；对于 $l \in [\varphi]$，这些变量是独立同分布的，其中 (H_l) 是具有共同密度的独立同分布变量

$$f_H(\lambda) = \frac{(1 - e^{-\lambda \sum_{j=1}^{J} \psi_j(M_j)})\tau_0(\lambda)}{\Psi_0\left(\sum_{j=1}^{J} \psi_j(M_j)\right)}. \tag{A.4}$$

4. $(B_j, j \in [J], B_0)$可分解为：

$$\left(B_{j,M_j} + \sum_{l=1}^{\varphi} \hat{\sigma}_{j,l}(H_l)\delta_{\tilde{Y}_l}, j \in [J], B_{0,J} + \sum_{l=1}^{\varphi} H_l \delta_{\tilde{Y}_l}\right). \tag{A.5}$$

5.其中 $B_{0,J} \sim$完全随机测度$(\tau_{0,J}, F_0)$、$B_{j,M_j} \sim$完全随机测度$(\tau^{(M_j)\,j}, B_{0,J})$，并且具有与定理 5.1相同的分布。$(\tilde{Y}_l)$为独立同分布的 F_0，表示将由$\sum_{j=1}^{J} M_j$ 客户选取的不同特征的 $\varphi \sim$ 泊松 $(\Psi_0(\sum_{j=1}^{J} \psi_j(M_j)))$ 数量。

6. $(N_{j,l}, j \in [J])|(\hat{\sigma}_{j,l} = t_j, j \in [J]), (H_l) \sim \mathrm{tP}((t_j M_j, j \in [J]).$

要证明定理 5.1，剩下的工作是确定边际量和后验量，以描述基于 $(B_j, j \in [J], B_0)|\mathbf{Z}_J$ 的联合分布 $(N_{j,l}, j \in [J]),(\hat{}\sigma_{j,l} = t_j, j \in [J])\ H_l)$, 的整体后验分布，其中每个 $l \in [\varphi]$ 表示为

$$\frac{\prod_{j=1}^{J} M_j^{n_{j,l}} t_j^{n_{j,l}} e^{-t_j M_j}}{(1 - e^{-\sum_{j=1}^{J} M_j t_j}) \prod_{j=1}^{J} n_{j,l}!} \times \varrho(\mathbf{t}|\lambda) f_H(\lambda) \tag{A.6}$$

由此可得，$(N_{j,l}, j \in [J]), (\hat{}\sigma_{j,l}(H_l) = t_j, j \in [J])|H_l = \lambda$,的联合分布具有如下形式

$$\frac{\prod_{j=1}^{J} M_j^{n_{j,l}} t_j^{n_{j,l}} e^{-t_j M_j} \eta_j(t_j|\lambda)}{\prod_{j=1}^{J} n_{j,l}!} \tag{A.7}$$

其等于

$$\prod_{j=1}^{J} \frac{t_j^{n_{j,l}} \eta_j^{[0]}(t_j|\lambda, M_j)}{\Xi_j^{[n_{j,l}]}(\lambda, M_j)} \times \frac{e^{-\lambda \sum_{j=1}^{J} \psi_j(M_j)} \prod_{j=1}^{J} M_j^{n_{j,l}} \Xi_j^{[n_{j,l}]}(\lambda, M_j)}{(1 - e^{-\lambda \sum_{j=1}^{J} \psi_j(M_j)}) \prod_{j=1}^{J} n_{j,l}!} \tag{A.8}$$

其中 $\eta_j^{[0]}(t_j|\lambda, M_j) := e^{-t_j M_j} \eta_j(t_j|\lambda) e^{\lambda \psi_j(M_j)}$，将$\hat{}\sigma_{j,l}(\lambda)|N_{j,l} = n_{j,l}$、$H_l = \lambda$ 的密度标识为 $\eta_j^{[n_{j,l}]}(t_j|\lambda, M_j)$，这与命题 5.2 中的描述一致。

A.IV.1 关于命题5.1的评论

也就是说，$\hat{}\sigma_{j,l}(\lambda)|N_{j,l} = n_{j,l}, H_l = \lambda$ 的分布等价于$\sigma_{j,l}(\lambda)|\mathscr{P}_{j,l}(M_j \sigma_{j,l}(\lambda)) = n_{j,l}$的分布，并且对于$n_{j,l} > 0$，其分布与$\sigma_{j,l}(\lambda)|\sum_{v=1}^{n_j} \mathsf{e}_{v,l}, ,() = M_j$的分布相同，其中 $(\mathsf{e}_{v,l})_{v \geq 1}$ 是一个由指数分布 (1) 随机变量组成的独立同分布集合。这~建立了变量$C_{j,k,l}, X_{j,l}$ 与 [18, 36], 中相应变量之间的关系；通过直接比较，或根据 [29] 和 [20, 26] 的研究，这些关系导向了命题 5.1 中的有限吉布斯划分。

注 A.3。顺便一提，从形式上看，我们可以将 $\sigma_{j,l}(\lambda) = \sum_{k=1}^{\infty} s_{j,k,l}$ 的跳跃$(s_{j,k,l})_{\{k \geq 1\}}$ 归一化，以形成 *[18, 27],* 中的泊松–金曼类，然后再以 $N_{j,l} = n_{j,l}$· 为条件。更细节的内容将在别处讨论。

A.IV.2 建立命题 5.2、5.4 及定理 5.1 的其余部分

如正文中所述，命题 5.2 随即由 [15, 17, 18], 中现有的上述结果推出，从而得到

$$\hat{\sigma}_{j,l}(\lambda) \stackrel{d}{=} \tilde{\sigma}_{j,l}(\lambda) + \sum_{k=1}^{X_{j,l}} S_{j,k,l}$$

(A.8) 中的另一项对应于命题 5.4 第 1 点中出现的 $(N_{j,l}, j \in [J])|H_l = \lambda$ 的联合分布。$\Xi_j^{[n_{j,l}]}(\lambda, M_j)$ 的表示，例如可由 [15, 18], 得出，是对具有指数倾斜密度 $\eta^{[0]}(t_j|\lambda, M_j)$ 的无限可分变量的 $n_{j,l}$ 阶矩进行累积量展开的结果。对 f_H 积分即可完成命题 5.4，同时牢记变量 $X_{j,l}, X_l, C_{j,k,l}$ 等的含义及其边缘分布；这些内容均在我们的论述~中由命题 3.1 推导得出。定理 5.1 第 5 点中出现的、给定 $(N_{j,l}, j \in [J])$ $l \in [r]$ 时 $(X_{j,l}, (C_{j,k,l}, k \in [X_{j,l}], l \in [r],$ $j \in J))$ 的联合分布，也可通过去掉以下等价表示中的求和项直接验证。该表示给出了带有 $x_l := \sum_{j=1}^{J} x_{j,l}$ 的 $(N_{j,l} = n_{j,l}, j \in [J])$ 的边缘分布（还需要再除以 $\prod_{j=1}^{J} n_{j,l}!$ ）。

$$\sum_{\mathbf{x}_{j,l} \in \mathcal{K}_J(\mathbf{n}_l)} \mathbb{P}(\tilde{X}_l = x_l) \mathbb{P}((X_{j,l} = x_{j,l}, j \in [J])|\tilde{X}_l = x_l) \prod_{j=1}^{J} \mathbb{P}(\sum_{i=1}^{x_{j,l}} \tilde{C}_{j,i,l} = n_{j,l}) \tag{A.9}$$

其中，如命题 3.1中所述，$\tilde{X}_l \sim$MtP分布$(\sum_{j=1}^{J} \psi_j(M_j)\ \tau_0)\ X_{j,\,l},\ j \in [J])|\tilde{X}_l - x_l \sim$多项分布$(x_l;\ q_1, \ldots,\ q_J)$，对于 $q_j = \psi_j(M_j) / \sum_{v=1}^{J} \psi_v(M_v)$ 和 $(\tilde{C}_{j,\,i,\,l},\ i \in [x_{j,l}) \overset{iid}{\sim}$MtP 分布 $(M_j,\ \tau_j)$。(A.9) 可通过先使用命题 5.4的第 2项，然后使用由命题 5.1 的第 4 项推导出的恒等式轻易得到。

A.IV.3 命题 5.5 中一般预测规则的结果

一旦按照定理 5.1 或上述讨论建立了 $(B_j,\ j\in[J],\ B_0)|\mathbf{Z}_J$ 的后验分布，就可以直接得到命题 5.5 中的预测规则。下面简要说明几处细节。

1. 由 (A.5) 可知 $B_{j,M_j}+\sum_{l=1}^{\varphi}\hat{}\sigma_{j,l}(H_l)\delta_{\tilde{Y}_l}$，因此 $Z^{(M_j+1)}{}_j \sim$ 泊松P$(B_{j,M_j}+\sum_{l=1}^{\varphi}\hat{}\sigma_{j,l}(H_l)\delta_{\tilde{Y}_l})$.

2. 因此，$(Z^{(M_j+1)}{}_j, j\in[J])|\mathbf{Z}_J$可分解为一个向量

$$(\hat{Z}_j, j\in[J])+(\sum_{l=1}^{r}\mathscr{P}_{j,l}(\hat{\sigma}_{j,l}(H_l))\delta_{\tilde{Y}_l}, j\in[J])$$

3. 进一步地，由命题 5.2 或直接由定理 5.1 可得，

$$\mathscr{P}_{j,l}(\hat{\sigma}_{j,l}(H_l))\overset{d}{=}\mathscr{P}^{(2)}_{j,l}(\tilde{\sigma}_{j,l}(\lambda))+\mathscr{P}_{j,l}(\sum_{k=1}^{X_{j,l}}S_{j,k,l})$$

其中 $\mathscr{P}^{(2)}_{j\,l}$,是相互独立的泊松变量，其随机强度如图所示

4. $\hat{}Z_j|B_{j,M_j}\sim$泊松P(B_{j,M_j})，且 $B_{j,M_j}|B_{0,J}\sim$完全随机测度$(\tau_{(M_j)\,j}, B_{0,J})$，其中 $B_{0,J}\sim CRM(\tau_{0,J}, F_0)$

$\hat{}$ 5. 因此，向量 $(Z_j, j\in[J])$, 具有形式 (A.1) 的表示，并按照命题 3.1 进行抽样，该表示基于变量 $C^{(*,1)}_{j\,k\,l}$,,,$X^*_{j,l}$，如命题 5.5 所示

6. $\sum_{l=1}^{r}\mathscr{P}^{(2)}_{j\,l}$,$(\tilde{\sigma}_{j,l}(\lambda))\delta_{\tilde{Y}_l}$服从 PoiP $\left(\sum_{l=1}^{\varphi}\tilde{\sigma}_{j,l}(H_l)\delta_{\tilde{Y}_l}\right)$ 分布。因此，该向量也可以基于 $C^{(*,2)}_{j\,k\,l}$,,,$\mathscr{P}_{j,l}(H_l)$，通过应用命题 3.1 生成。

7. 等价地，可以先在给定 B_0 的条件下考虑 $(\tilde{Z}^{(M_j+1)}{}_j, j\in[J])$；根据命题 2.1 的应用，$\tilde{Z}^{(M_j+1)}{}_j$ $|\mathbf{Z}_J, B_0$服从 IBP$(C_j, \tau^{(M_j)}{}_j, B_0)$ 分布，如命题 2.1 第 5 项所述。然后使用 B_0 的后验表示，并应用命题 3.1。

B 简化泊松-GG-G 的一些计算 G

尽管关于简化Poisson-GG-GG的结果在细节上应该更为简单，但精细化结果绝非显而易见。我们给出边缘分布及相关量的推导。

B.I 泊松-GG-GG 边缘/联合推导

对于简化的泊松-广义伽马-广义伽马模型，即泊松-GG-GG，回想一下，我们在命题 5.3 中设定：$\alpha_0=\beta/\alpha$ 对于 $0<\beta<\alpha$，$\tau_j(s)=\theta_j\frac{\alpha s^{-\alpha-1}e^{-\zeta s}}{\Gamma\ 1-\alpha}$()且 $\tau_0(\lambda)=\frac{\beta\lambda^{-\frac{\beta}{\alpha}-1}e^{-\lambda\sum_{j=1}^{J}\theta_j\zeta^{\alpha}}}{\alpha\Gamma}$($_{1-\beta/\alpha)}$,且 $M_j=\bar{M}$.回想一下，$(N_{j,\ l}$、$\hat{}\sigma_{j,\ l}(H_l)\ j\in[J])\ H_l$，在 $n_l:=\sum_{j=1}^{J}n_{j,\ l}=1$、2、.....、时的联合分布的一般形式为，

$$\frac{\prod_{j=1}^{J}M_j^{n_{j,l}}t_j^{n_{j,l}}e^{-t_jM_j}}{(1-\mathrm{e}^{-\sum_{j=1}^{J}M_jt_j})\prod_{j=1}^{J}n_{j,l}!}\frac{(1-\mathrm{e}^{-\sum_{j=1}^{J}M_jt_j})\prod_{j=1}^{J}\eta_j(t_j|\lambda)}{(1-\mathrm{e}^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)})}f_{H_l}(\lambda).$$

在此情形下

$$f_{H_l}(\lambda)=\frac{(1-e^{-\lambda\sum_{j=1}^{J}\theta_j[(\zeta+M)^\alpha-\zeta^\alpha]})\beta\lambda^{-\frac{\beta}{\alpha}-1}e^{-\lambda\sum_{j=1}^{J}\theta_j\zeta^\alpha}}{\alpha\Gamma(1-\frac{\beta}{\alpha})(\sum_{j=1}^{J}\theta_j)^{\frac{\beta}{\alpha}}[(\zeta+M)^\beta-\zeta^\beta]}$$

此外，(ˆ$\sigma_{j,l}(\lambda)$ $j\in[J]$) 的联合密度表示为

$$\frac{(1-\mathrm{e}^{-M\sum_{j=1}^{J}t_j})\prod_{j=1}^{J}\eta_j(t_j|\lambda)}{(1-e^{-\lambda\sum_{j=1}^{J}\theta_j[(\zeta+M)^\alpha-\zeta^\alpha]})}$$

其中

$$\eta_j(t_j|\lambda)dt_j=\mathrm{e}^{-t_j\zeta}\mathbb{P}(\sigma_j(\theta_j\lambda)\in dt_j)e^{\theta_j\lambda\zeta^\alpha}$$

其中 $\sigma_j(\theta_j\lambda)\overset{d}{=}\sigma_j\theta_j^{\frac{1}{\alpha}}\lambda^{\frac{1}{\alpha}}$ 其中 σ_j 是一个 α-稳定随机变量，其拉普拉斯变换为$\mathbb{E}[\mathrm{e}^{-t\sigma_j}]=\mathrm{e}^{-t^\alpha}$. 由此可知，我们总是可以写成

$$\frac{\prod_{j=1}^{J}M_j^{n_{j,l}}t_j^{n_{j,l}}e^{-t_jM_j}}{\prod_{j=1}^{J}n_{j,l}!}=\frac{M^{n_l}(\sum_{j=1}^{J}t_j)^{n_l}\mathrm{e}^{-M\sum_{j=1}^{J}t_j}}{n_l!}\times\frac{n_l!}{\prod_{j=1}^{J}n_{j,l}!}\prod_{j=1}^{J}p_j^{n_{j,l}}$$

其中 $p_j=t_j/\sum_{v=1}^{J}t_v$. 集合 $\sigma=\sum_{j=1}^{J}\sigma_j\theta_j^{\frac{1}{\alpha}}$. 则 ($N_{j\,l}$ $j\in$,) 的边缘分布可以表示为对.....的期望 $[J]$ $|H_l=\lambda$

$$\frac{M^{n_l}\sigma^{n_l}\lambda^{\frac{n_l}{\alpha}}\mathrm{e}^{-(M+\zeta)\sigma\lambda^{1/\alpha}}}{n_l!(1-\mathrm{e}^{-\lambda\sum_{j=1}^{J}\psi_j(M_j)})}\times\frac{n_l!}{\prod_{j=1}^{J}n_{j,l}!}\prod_{j=1}^{J}P_j^{n_{j,l}}\times\mathrm{e}^{\lambda\sum_{j=1}^{J}\theta_j\zeta^\alpha}$$

wher $_1\mathrm{e}$ $P_j=\sigma_j\theta_j^{\alpha}/\sigma$. 然而，我们可以先对 f_{H_l} 进行积分，这相当于积分 al

$$\frac{\int_0^\infty\lambda^{\frac{n_l}{\alpha}}\mathrm{e}^{-(M+\zeta)\sigma\lambda^{1/\alpha}}\beta\lambda^{-\frac{\beta}{\alpha}-1}d\lambda}{\alpha\Gamma(1-\frac{\beta}{\alpha})(\sum_{j=1}^{J}\theta_j)^{\frac{\beta}{\alpha}}[(\zeta+M)^\beta-\zeta^\beta]}=\frac{\beta\Gamma(n_l-\beta)\sigma^{-n_l+\beta}(M+\zeta)^{-n_l+\beta}}{\Gamma(1-\frac{\beta}{\alpha})(\sum_{j=1}^{J}\theta_j)^{\frac{\beta}{\alpha}}[(\zeta+M)^\beta-\zeta^\beta]}$$

请注意

$$\mathbb{E}[\sigma^\beta]=(\sum_{j=1}^{J}\theta_j)^{\beta/\alpha}\frac{\Gamma(1-\frac{\beta}{\alpha})}{\Gamma(1-\beta)}$$

且 $\tilde{\sigma}=\sigma/(\sum_{j=1}^{J}\theta_j)^{\frac{1}{\alpha}}$是一个尺度参数为 1 的 α 稳定变量。此外，我们可以写作 $P_j\overset{d}{=}\sigma_jq_j^{\frac{\alpha}{\cdot}}/\tilde{\sigma}$，其中，为了清晰起见

$$\mathbb{E}[\tilde{\sigma}^\beta]=\frac{\Gamma(1-\frac{\beta}{\alpha})}{\Gamma(1-\beta)}$$

并写作

$$\frac{\beta\Gamma(n_l-\beta)\sigma^{-n_l+\beta}(M+\zeta)^{-n_l+\beta}}{\Gamma(1-\frac{\beta}{\alpha})(\sum_{j=1}^{J}\theta_j)^{\frac{\beta}{\alpha}}[(\zeta+M)^\beta-\zeta^\beta]}=\frac{\beta\Gamma(n_l-\beta)\sigma^{-n_l}(M+\zeta)^{-n_l+\beta}}{\Gamma(1-\beta)[(\zeta+M)^\beta-\zeta^\beta]}\times\frac{\tilde{\sigma}^\beta}{\mathbb{E}[\tilde{\sigma}^\beta]}$$

因此，($N_{j,\ l},\ \ j\in[J]$) 的联合边缘分布可以精确表示为,

$$\frac{(1-\frac{\zeta}{(\zeta+M)})^{n_l}}{(1-\frac{\zeta^\beta}{(\zeta+M)^\beta})}\frac{\beta\Gamma(n_l-\beta)}{n_l!\Gamma(1-\beta)}\times\frac{n_l!}{\prod_{j=1}^{J}n_{j,l}!}\mathbb{E}[\frac{\tilde{\sigma}^\beta}{\mathbb{E}[\tilde{\sigma}^\beta]}\times\prod_{j=1}^{J}P_j^{n_{j,l}}]$$

其中，参见例如 [30],

$$\mathbb{E}\left[\frac{\tilde{\sigma}^{\beta}}{\mathbb{E}[\tilde{\sigma}^{\beta}]}\times\prod_{j=1}^{J}P_j^{n_{j,l}}\right]=\mathbb{E}[\prod_{j=1}^{J}\tilde{P}_{\alpha,-\beta}^{n_{j,l}}(q_j)]$$

˜因此，$((C_{j,\ k,\ l},\ \ k\in[X_{j,l}])\ [X_{j,l}],\ \ j\in[J])$，$(N_{j,\ l},\ \ j\in[J])$ 的联合分布为，对于$x_l=\sum_{j=1}^{J}x_{j,\ l},\ \ \sum^{x_{j,\ k=1}}{}^{l}c_{j,\ k,\ l}=n_{j,\ l}$，且 $x_{j,\ l}=0$ 对于 $n_{j,\ l}=0$，：

$$\frac{\Gamma(x_l-\frac{\beta}{\alpha})\Gamma(1-\beta)\alpha^{x_l-1}}{\Gamma(1-\frac{\beta}{\alpha})\Gamma(n_l-\beta)}\prod_{j=1}^{J}q_j^{x_{j,l}}\prod_{k=1}^{x_{j,l}}\frac{\Gamma(c_{j,k,l}-\alpha)}{\Gamma(1-\alpha)}\times\frac{\beta\Gamma(n_l-\beta)}{n_l!\Gamma(1-\beta)}\frac{(1-\frac{\zeta}{(\zeta+M)})^{n_l}}{(1-\frac{\zeta^{\beta}}{(\zeta+M)^{\beta}})}. \tag{B.1}$$

C 一般泊松-GG-GG 层次印度自助过程的后验推断算法详情

C.I 边缘分布的推导

在我们的模拟研究中，我们考虑以下采用通用参数化的泊松-GG-GG 层次印度自助过程模型,

$$\begin{aligned}
&\tau_0(\lambda)=\mathrm{GG}(\lambda|\alpha_0,\theta_0/\alpha_0,\zeta_0)=\frac{\theta_0}{\Gamma(1-\alpha_0)}\lambda^{-\alpha_0-1}e^{-\zeta_0\lambda},\quad B_0\sim\mathrm{CRM}(\tau_0,F_0),\\
&\tau_j(s)=\mathrm{GG}(s|\alpha_j,\theta_j/\alpha_j,\zeta_j),\quad B_j|B_0\sim\mathrm{CRM}(\tau_j,B_0)\text{ for }j\in[J],\\
&Z_j^{(i)}|B_j\sim\mathrm{PoiP}(B_j)\text{ for }i\in[M_j],j\in[J].
\end{aligned} \tag{C.1}$$

在继续之前，我们先计算描述所需的一些量,

$$\begin{aligned}
&\kappa_j:=\psi_j(M_j)=\int_0^{\infty}(1-e^{-sM_j})\tau_j(s)ds=\frac{\theta_j}{\alpha_j}((\zeta_j+M_j)^{\alpha_j}-\zeta_j^{\alpha_j}),\\
&\kappa_{\cdot}:=\sum_{j=1}^{J}\kappa_j,\quad q_j:=\frac{\kappa_j}{\kappa_{\cdot}},\quad \Psi_0(\kappa_{\cdot})=\frac{\theta_0}{\alpha_0}((\zeta_0+\kappa_{\cdot})^{\alpha_0}-\zeta_0^{\alpha_0}),
\end{aligned} \tag{C.2}$$

令 $Z_j:=\sum_{i=1}^{M_j}Z_j^{(i)}$和$\mathbf{Z}_J:=(Z_j,j\in[J])$。根据定理 3.1,

$$\mathbf{Z}_J\overset{d}{=}\left(\sum_{l=1}^{\varphi}\left[\sum_{k=1}^{X_{j,l}}\widetilde{C}_{j,k,l}\right]\delta_{\widetilde{Y}_l},i\in[M_j],j\in[J]\right), \tag{C.3}$$

其中，相应的边缘密度可以计算为:

$$\begin{aligned}
&\mathbb{P}(\varphi=r,(X_{j,l}=x_{j,l},\widetilde{C}_{j,k,l}=c_{j,k,l},j\in[J],k\in[x_{j,l}],l\in[r]))\\
&=\mathbb{P}(\varphi=r)\prod_{l=1}^{r}\mathbb{P}(X_l=x_l)\prod_{j=1}^{J}\mathbb{P}(X_{j,l}=x_{j,l}|X_l=x_l)\prod_{k=1}^{x_{j,l}}\mathbb{P}(C_{j,k,l}=c_{j,k,l}),
\end{aligned} \tag{C.4}$$

其中

$$\begin{aligned}
&\mathbb{P}(\varphi=r)=\frac{\Psi_0(\kappa_{\cdot})^re^{-\Psi_0(\kappa_{\cdot})}}{r!},\\
&\mathbb{P}(X_l=x_l)=\frac{\mathbb{1}_{\{x_l>0\}}p_0^{x_l}}{1-(1-p_0)^{\alpha_0}}\frac{\alpha_0\Gamma(x_l-\alpha_0)}{x_l!\Gamma(1-\alpha_0)}\text{ where }p_0:=\frac{\kappa_{\cdot}}{\zeta_0+\kappa_{\cdot}},\\
&\mathbb{P}((X_{j,l}=x_{j,l},j\in[J])|X_l=x)=\frac{x_l!}{\prod_{j=1}^{J}x_{j,l}!}\prod_{j=1}^{J}q_j^{x_{j,l}},\\
&\mathbb{P}(\widetilde{C}_{j,k,l}=c_{j,k,l})=\frac{\mathbb{1}_{\{c_{j,k,l}>0\}}p_j^{c_{j,k,l}}}{1-(1-p_j)^{\alpha_j}}\frac{\alpha_j\Gamma(c_{j,k,l}-\alpha_j)}{c_{j,k,l}!\Gamma(1-\alpha_j)}\text{ where }p_j=\frac{M_j}{\zeta_j+M_j}.
\end{aligned} \tag{C.5}$$

因此，边缘分布计算为，

$$\begin{aligned}
&\mathbb{P}(\varphi = r, (X_{j,l} = x_{j,l}, \widetilde{C}_{j,k,l} = c_{j,k,l}, j \in [J], k \in [x_{j,l}], l \in [r])) \\
&= \frac{\Psi_0(\kappa.)^r e^{-\Psi_0(\kappa.)}}{r!} \prod_{l=1}^{r} \frac{\mathbb{1}_{\{x_l>0\}} p_0^{x_l}}{1-(1-p_0)^{\alpha_0}} \frac{\alpha_0 \Gamma(x_l - \alpha_0)}{x_l! \Gamma(1-\alpha_0)} \frac{x_l!}{\prod_{j=1}^{J} x_{j,l}!} \\
&\times \prod_{j=1}^{J} \frac{q_j^{x_{j,l}} \mathbb{1}_{\{c_{j,k,l}>0\}} p_j^{c_{j,k,l}}}{1-(1-p_j)^{\alpha_j}} \frac{\alpha_j \Gamma(c_{j,k,l} - \alpha_j)}{c_{j,k,l}! \Gamma(1-\alpha_j)}.
\end{aligned} \tag{C.6}$$

作为注释，回顾正文，对于 $0 < \alpha_j < 1$，令 $P_{\alpha_j,0} \sim \mathscr{PY}(\alpha_j, 0, F_0)$ 表示 $(\alpha_j, 0)$ Pitman-Yor过程， [13, 25], 其由对一个 α_j 稳定子过程进行归一化构造而成。描述 $[n_{j,l}]$ 的随机分区分布的相应 EPPF 给出为：$p_{\alpha_j,0}(c_{j,l}) = \frac{\alpha_j^{x_{j,l}}}{\Gamma(n_{j,l})} \prod_{k=1}^{x_{j,l}} \Gamma(c_{j,k,l} - \alpha_j)$ $(1-\alpha_j)$ · 由此，区块数量 $\mathbb{P}^{(n_{j,l})}_{\alpha_j,0}(k) = \frac{\alpha_j^{k-1}}{\Gamma(k)} S_{\alpha_j}(n_{j,l}, k)$，其中 $S_{\alpha_j}(n_{j,l}, k)$ 表示第二类广义 Stirling 数，如 [28][3.2.3, p. 65-66] 所述。

在实践中，由于我们只能观察到聚合计数 $N_{j,\,l} = \sum_{k=1}^{X_{j,\,l}} \widetilde{C}_{j,\,k,\,l}$，因此我们必须处理涉及 $(N_{j,\,l},\ j \in [J],\ l \in [r])$ 的边缘分布。根据命题 5.4 和 [26],

$$\begin{aligned}
\mathbb{P}(\widetilde{N}_{j,l} = n_{j,l} | X_{j,l} = x_{j,l}) &= \mathbb{1}_{\{n_{j,l} \geq x_{j,l}\}} \sum_{*} \prod_{k=1}^{x_{j,l}} \mathbb{P}(\widetilde{C}_{j,k,l} = c_{j,k,l}) \\
&= \mathbb{1}_{\{n_{j,l} \geq x_{j,l}\}} \frac{\alpha_j^{x_{j,l}} p_j^{n_{j,l}}}{(1-(1-p_j)^{\alpha_j})^{x_{j,l}}} \sum_{*} \prod_{k=1}^{x_{j,l}} \frac{\Gamma(c_{j,k,l} - \alpha_j)}{c_{j,k,l}! \Gamma(1-\alpha_j)} \\
&= \frac{\mathbb{1}_{\{n_{j,l} \geq x_{j,l}\}} \alpha_j^{x_{j,l}} p_j^{n_{j,l}}}{(1-(1-p_j)^{\alpha_j})^{x_{j,l}}} \frac{x_{j,l}! S_{\alpha_j}(n_{j,l}, x_{j,l})}{n_{j,l}!},
\end{aligned} \tag{C.7}$$

其中，求和 $\sum_*$ 遍历所有满足将 $n_{j,l}$ 划分为正整数集合的集合 $c_{j,\,1,\,l}, \ldots\ldots, c_{j,\,x_{j,\,l},\,l}$。第二类广义斯特林数 $S_\alpha(n, x)$ 定义如下：

$$S_\alpha(n, x) = \frac{n!}{x!} \sum_{*} \prod_{k=1}^{x} \frac{\Gamma(c_k - \alpha)}{c_k! \Gamma(1-\alpha)}, \tag{C.8}$$

并且可以通过以下递归公式轻松计算，

$$\begin{aligned}
&S_\alpha(n, 1) = \frac{\Gamma(n-\alpha)}{\Gamma(1-\alpha)}, \quad S_\alpha(n, n) = 1, \\
&S_\alpha(n+1, x) = (n - \alpha x) S_\alpha(n, x) + S_\alpha(n, x-1).
\end{aligned} \tag{C.9}$$

Based 基于这一恒等式，我们将使用的边缘分布计算为 a s,

$$\begin{aligned}
&\mathbb{P}(\varphi = r, (X_{j,l} = x_{j,l}, N_{j,l} = n_{j,l}, j \in [J], l \in [r])) \\
&= \frac{\Psi_0(\kappa.)^r e^{-\Psi_0(\kappa.)}}{r!} \prod_{l=1}^{r} \frac{\mathbb{1}_{\{x_l>0\}} p_0^{x_l}}{1-(1-p_0)^{\alpha_0}} \frac{\alpha_0 \Gamma(x_l - \alpha_0)}{\Gamma(1-\alpha_0)} \prod_{j=1}^{J} \frac{q_j^{x_{j,l}} \mathbb{1}_{\{n_{j,l} \geq x_{j,l}\}} \alpha_j^{x_{j,l}} p_j^{n_{j,l}}}{(1-(1-p_j)^{\alpha_j})^{x_{j,l}}} \frac{S_{\alpha_j}(n_{j,l}, x_{j,l})}{n_{j,l}!} \\
&= \frac{\theta_0^r e^{-\Psi_0(\kappa.)}}{r! (\zeta_0 + \kappa.)^{x. - \alpha_0 r}} \prod_{l=1}^{r} \frac{\mathbb{1}_{\{x_l>0\}} \Gamma(x_l - \alpha_0)}{\Gamma(1-\alpha_0)} \prod_{j=1}^{J} \frac{\mathbb{1}_{\{n_{j,l} \geq x_{j,l}\}} \theta_j^{x_{j,l}} M_j^{n_{j,l}}}{(\zeta_j + M_j)^{n_{j,l} - \alpha_j x_{j,l}}} \frac{S_{\alpha_j}(n_{j,l}, x_{j,l})}{n_{j,l}!},
\end{aligned} \tag{C.10}$$

其中 $x. := \sum_{l=1}^{r} x_l = \sum_{l=1}^{r} \sum_{j=1}^{J} x_{j,\,l}$。

C.II 抽样器的详细描述

令$\mathbf{n}_J := (n_{j,l}, j \in [J], l \in [r])$ 表示观测到的词计数集合。我们的目标是估计参数 ϕ:
$= \{\alpha_0, \theta_0, \{\alpha_j, \theta_j\}_{j=1}^J\}$，并将参数 ζ_0 和 $\{\zeta_j\}_{j=1}^J$ 固定为 1。我们对这些参数引入模糊先验，

$$\begin{aligned}
\pi_0(\alpha_0) &:= \operatorname{logit}\mathcal{N}(\alpha_0|0,1) \propto \frac{1}{\alpha_0(1-\alpha_0)} \exp\Big(- \frac{\Big(\log\frac{\alpha_0}{1-\alpha_0}\Big)^2}{2}\Big), \\
\pi_0(\theta_0) &:= \operatorname{Gamma}(\theta_0|1,1), \\
\pi_0(\alpha_j) &:= \operatorname{logit}\mathcal{N}(\alpha_j|0,1) \text{ for } j \in [J], \\
\pi_0(\theta_j) &:= \log \mathcal{N}(\theta_j|0,1) \propto \frac{1}{\theta_j} \exp\Big(- (\log\theta_j)^2\Big) \text{ for } j \in [J], \\
\pi_0(\phi) &:= \pi_0(\alpha_0)\pi_0(\theta_0) \prod_{j=1}^J \pi_0(\alpha_j)\pi_0(\theta_j).
\end{aligned} \tag{C.11}$$

待模拟的目标后验分布定义如下:

$$\pi(\phi, \mathbf{X}_J|\mathbf{n}_J) \propto \mathbb{P}(\varphi = r, (X_{j,l}, \widetilde{N}_{j,l} = n_{j,l}, j \in [J], l \in [r]))\pi_0(\phi), \tag{C.12}$$

其中 $\mathbf{X}_J := (X_{j,\,l},\ j \in [J],\ l \in [r])$。我们通过对潜变量和参数 ϕ进行吉布斯抽样来继续。

对 X 进行抽样$_J$。X 中每个条目 $X_{j,\,l}$ 的条件分布可简单表示为一个离散分布:

$$\mathbb{P}(X_{j,l} = x|\text{others}) \propto \frac{\Gamma(x + x_l^{-j} - \alpha_0)\theta_j^x(\zeta_j + M_j)^{\alpha_j x} S_{\alpha_j}(n_{j,l}, x)}{(\zeta_0 + \kappa_\cdot)^x}, \quad x \in [n_{j,l}], \tag{C.13}$$

其中 $x_l^{-j} := \sum_{j' \neq j} x_{j',\,l}$。

抽样 θ_0。给定伽马分布先验，θ_0 的后验分布也服从伽马分布。

$$\mathbb{P}(\theta_0 \in d\theta|\text{others}) = \operatorname{Gamma}\Big(\theta \,\Big|\, r + a, b + \frac{(\zeta_0 + \kappa_\cdot)^{\alpha_0} - \zeta_0^{\alpha_0}}{\alpha_0}\Big). \tag{C.14}$$

对其他参数进行采样。对于除 θ_0 之外的参数，遗憾的是，其条件分布没有闭式解。因此，我们使用随机游走 Metropolis–Hastings 方法，通过以下提议分布对其进行采样。

$$\begin{aligned}
q(\alpha_0'|\alpha_0) &= \operatorname{logit}\mathcal{N}\Big(\alpha_0' \,\Big|\, \log\frac{\alpha_0}{1-\alpha_0}, (0.01)^2\Big) \text{ for } j \in \{0\} \cup [J], \\
q(\theta_j'|\theta_j) &= \log\mathcal{N}(\theta_j'|\log\theta_j, (0.01)^2) \text{ for } j \in [J].
\end{aligned} \tag{C.15}$$

C.III 预测规则

设 $Z_j^* := Z^{(M_j+1)}{}_j$ 是一个要添加到 $\mathbf{Z}_J$ 中的组 j 的新文档。根据命题 5.5，对记号稍作调整后，

$$Z_j^*|\mathbf{Z}_J \overset{d}{=} \underbrace{\sum_{v=1}^{\varphi_j^*}\Big[\sum_{k=1}^{X_{j,v}^{(1)}} \widetilde{C}_{j,k,v}^{(1)}\Big]\delta_{Y_v}}_{:=①} + \underbrace{\sum_{l=1}^{r}\Big[\sum_{k=1}^{X_{j,l}^{(2)}} \widetilde{C}_{j,k,l}^{(2)}\Big]\delta_{\widetilde{Y}_l}}_{:=②} + \underbrace{\sum_{l=1}^{r}\Big[\sum_{k=1}^{x_{j,l}} \widetilde{C}_{j,k,l}^{(3)}\Big]\delta_{\widetilde{Y}_l}}_{:=③}. \tag{C.16}$$

下面，我们将详细介绍各个组成部分。

①. 记

$$\Delta_j := \psi_j(M_j+1) - \psi_j(M_j) = \frac{\theta_j}{\alpha_j}((\zeta_j+M_j+1)^{\alpha_j} - (\zeta_j+M_j)^{\alpha_j}). \tag{C.17}$$

Then

$$\begin{aligned}
\mathbb{P}(\varphi_j^* = r_j^*) &\sim \text{Poisson}\left(\Psi_0(\kappa_\cdot + \Delta_j) - \Psi_0(\kappa_\cdot)\right) \\
&= \frac{(\Psi_0(\kappa_\cdot+\Delta_j) - \Psi_0(\kappa_\cdot))^{r_j^*} e^{-\Psi_0(\kappa_\cdot+\Delta_j)+\Psi_0(\kappa_\cdot)}}{(r_j^*)!}, \\
\mathbb{P}(X_{j,v}^{(1)} = x_{j,v}^{(1)}) &= \int_0^\infty \frac{\theta_0 \lambda^{-\alpha_0-1} e^{-(\zeta_0+\kappa_\cdot)\lambda}(1-e^{-\Delta_j\lambda})}{\Gamma(1-\alpha_0)(\Psi_0(\kappa_\cdot+\Delta_j)-\Psi_0(\kappa_\cdot))} \frac{\mathbb{1}_{\{x_{j,v}^{(1)}>0\}}(\Delta_j\lambda)^{x_{j,v}^{(1)}} e^{-\Delta_j\lambda}}{(x_{j,v}^{(1)})!(1-e^{-\Delta_j\lambda})} d\lambda \\
&= \frac{\mathbb{1}_{\{x_{j,v}^{(1)}>0\}}(p_{0,j}^*)^{x_{j,v}^{(1)}}}{1-(1-p_{0,j}^*)^{\alpha_0}} \frac{\alpha_0\Gamma(x_{j,v}^{(1)}-\alpha_0)}{(x_{j,v}^{(1)})!\Gamma(1-\alpha_0)} \text{ where } p_{0,j}^* := \frac{\Delta_j}{\zeta_0+\kappa_\cdot+\Delta_j}, \\
\mathbb{P}(\widetilde{C}_{j,k,v}^{(1)} = c_{j,k,v}^{(1)}) &= \int_0^\infty \frac{\theta_j s^{-\alpha_j-1} e^{-(\zeta_j+M_j)s}(1-e^{-s})}{\Gamma(1-\alpha_j)\Delta_j} \frac{\mathbb{1}_{\{c_{j,k,v}^{(1)}>0\}} s^{c_{j,k,v}^{(1)}} e^{-s}}{(1-e^{-s})} ds \\
&= \frac{\mathbb{1}_{\{c_{j,k,v}^{(1)}>0\}}(p_j^*)^{c_{j,k,v}^{(1)}}}{1-(1-p_j^*)^{\alpha_j}} \frac{\alpha_j\Gamma(c_{j,k,v}^{(1)}-\alpha_j)}{(c_{j,k,v}^{(1)})!\Gamma(1-\alpha_j)} \text{ where } p_j^* := \frac{1}{\zeta_j+M_j+1}.
\end{aligned} \tag{C.18}$$

对于 $\widetilde{N}_{j,v}^{(1)} := \sum_{k=1}^{X_{j,k}^{(1)}} \widetilde{C}_{j,k,v}^{(1)}$，我们可以如下计算概率质量函数。

$$\mathbb{P}(\widetilde{N}_{j,v}^{(1)} = n_{j,v}^{(1)} | X_{j,v}^{(1)} = x_{j,v}^{(1)}) = \frac{\mathbb{1}_{\{n_{j,v}^{(1)} \geq x_{j,v}^{(1)}\}} \alpha_j^{x_{j,v}^{(1)}} (p_j^*)^{n_{j,v}^{(1)}}}{(1-(1-p_j^*)^{\alpha_j})^{x_{j,v}^{(1)}}} \frac{(x_{j,v}^{(1)})! S_{\alpha_j}(n_{j,v}^{(1)}, x_{j,v}^{(1)})}{n_{j,v}^{(1)}!}. \tag{C.19}$$

②. 根据命题 5.5，我们有

$$\begin{aligned}
\mathbb{P}(X_{j,l}^{(2)} = x_{j,l}^{(2)}) &= \int_0^\infty \text{Gamma}(\lambda | x_l - \alpha_0, \zeta_0+\kappa_\cdot) \text{Poisson}(x_{j,l}^{(2)} | \Delta_j\lambda) d\lambda \\
&= \frac{\Gamma(x_{j,l}^{(2)} + x_l - \alpha_0)}{(x_{j,l}^{(2)})!\Gamma(x_l-\alpha_0)} (p_{0,j}^*)^{x_{j,l}^{(2)}} (1-p_{0,j}^*)^{x_l-\alpha_0} \\
&= \text{NB}(x_{j,l}^{(2)} | x_l - \alpha_0, p_{0,j}^*), \\
\mathbb{P}(\widetilde{C}_{j,k,l}^{(2)} = c_{j,k,l}^{(2)}) &= \frac{\mathbb{1}_{\{c_{j,k,l}^{(2)}>0\}}(p_j^*)^{c_{j,k,l}^{(2)}}}{1-(1-p_j^*)^{\alpha_j}} \frac{\alpha_j\Gamma(c_{j,k,l}^{(2)}-\alpha_j)}{(c_{j,k,l}^{(2)})!\Gamma(1-\alpha_j)}, \\
\mathbb{P}(\widetilde{N}_{j,l}^{(2)} = n_{j,l}^{(2)} | X_{j,l}^{(2)} = x_{j,l}^{(2)}) &= \frac{\mathbb{1}_{\{n_{j,l}^{(2)} \geq x_{j,l}^{(2)}\}} \alpha_j^{x_{j,l}^{(2)}} (p_j^*)^{n_{j,l}^{(2)}}}{(1-(1-p_j^*)^{\alpha_j})^{x_{j,l}^{(2)}}} \frac{(x_{j,l}^{(2)})! S_{\alpha_j}(n_{j,l}^{(2)}, x_{j,l}^{(2)})}{n_{j,l}^{(2)}!}.
\end{aligned} \tag{C.20}$$

③. 根据命题 5.5，对于 $k \in [x_{j,l}]$,

$$\mathbb{P}(\widetilde{C}_{j,k,l}^{(3)} = c_{j,k,l}^{(3)}) = \text{NB}(c_{j,k,l}^{(3)} | c_{j,k,l} - \alpha_j, p_j^*)^{\mathbb{1}_{\{x_{j,l}>0\}}}. \tag{C.21}$$

由于具有相同成功概率的负二项变量之和仍服从负二项分布，

$$\mathbb{P}(\widetilde{N}_{j,l}^{(3)} = n_{j,l}^{(3)} | X_{j,l} = x_{j,l}) = \text{NB}(n_{j,l}^{(3)} | n_{j,l} - \alpha_j x_{j,l}, p_j^*)^{\mathbb{1}_{\{x_{j,l}>0\}}}. \tag{C.22}$$

总之，根据聚合计数计算的新文档 Z_j^* 的预测概率为

$$
\begin{aligned}
&\mathbb{P}(\varphi_j^*=r_j^*,(\widetilde{N}_{j,v}^{(1)}=n_{j,v}^{(1)},X_{j,v}^{(1)}=x_{j,v}^{(1)},v\in[r_j^*]),(\widetilde{N}_{j,l}^{(2)}=n_{j,l}^{(2)},X_{j,l}^{(2)}=x_{j,l}^{(2)},\widetilde{N}_{j,l}^{(3)}=n_{j,l}^{(3)},l\in[r])|\mathbf{Z}_J)\\
&=\frac{(\Psi_0(\kappa.+\Delta_j)-\Psi_0(\kappa.))^{r_j^*}e^{-\Psi_0(\kappa.+\Delta_j)+\Phi_0(\kappa.)}}{(r_j^*)!}\\
&\times\prod_{v=1}^{r_j^*}\frac{\mathbb{1}_{\{x_{j,v}^{(1)}>0\}}(p_{0,j}^*)^{x_{j,v}^{(1)}}}{1-(1-p_{0,j}^*)^{\alpha_0}}\frac{\alpha_0\Gamma(x_{j,v}^{(1)}-\alpha_0)}{(x_{j,v}^{(1)})!\Gamma(1-\alpha_0)}\frac{\mathbb{1}_{\{n_{j,v}^{(1)}\geq x_{j,v}^{(1)}\}}\alpha_j^{x_{j,v}^{(1)}}(p_j^*)^{n_{j,v}^{(1)}}}{(1-(1-p_j^*)^{\alpha_j})^{x_{j,v}^{(1)}}}\frac{(x_{j,v}^{(1)})!S_{\alpha_j}(n_{j,v}^{(1)},x_{j,v}^{(1)})}{n_{j,v}^{(1)}!}\\
&\times\prod_{l=1}^{r}\frac{\Gamma(x_{j,l}^{(2)}+x_l-\alpha_0)}{(x_{j,l}^{(2)})!\Gamma(x_l-\alpha_0)}(p_{0,j}^*)^{x_{j,l}^{(2)}}(1-p_{0,j}^*)^{x_l-\alpha_0}\left[\frac{\mathbb{1}_{\{n_{j,l}^{(2)}\geq x_{j,l}^{(2)}\}}\alpha_j^{x_{j,l}^{(2)}}(p_j^*)^{n_{j,l}^{(2)}}}{(1-(1-p_j^*)^{\alpha_j})^{x_{j,l}^{(2)}}}\frac{(x_{j,l}^{(2)})!S_{\alpha_j}(n_{j,l}^{(2)},x_{j,l}^{(2)})}{n_{j,l}^{(2)}!}\right]^{\mathbb{1}_{\{x_{j,l}^{(2)}>0\}}}\\
&\times\left[\frac{\Gamma(n_{j,l}^{(3)}+n_{j,l}-\alpha_j x_{j,l})}{(n_{j,l}^{(3)})!\Gamma(n_{j,l}-\alpha_j x_{j,l})}(p_j^*)^{n_{j,l}^{(3)}}(1-p_j^*)^{n_{j,l}-\alpha_j x_{j,l}}\right]^{\mathbb{1}_{\{x_{j,l}>0\}}}.
\end{aligned}
\tag{C.23}
$$

C.IV 计算测试文档的预测概率

现在假设我们已经观察到一个测试文档。为了计算 (C.23)，我们需要三类计数（$\widetilde{N}_{j\,v}^{(1)}$,, $\widetilde{N}_{j\,l}^{(2)}$,, $\widetilde{N}_{(3)\,j,l}$）和潜变量（$X_{j\,v}^{(1)}$,, $X_{j\,l}^{(2)}$,, $X_{j\,l}^{(3)}$,），但从该文档中我们只能获得这些计数

$$
\mathbf{n}^*:=(n_{j,v}^{(1)},v\in[r_j^*]),(n_{j,l}^{(2,3)},l\in[r]),
\tag{C.24}
$$

其中 $n_{j}^{(2,3)}{}_{,\,l}=n_{j}^{(2)}{}_{,\,l}+n_{j\,l}^{(3)}{}_{,}$。因此，为了计算给定$\mathbf{Z}_J$时$\mathbf{n}^*$ 的预测概率，我们应先通过吉布斯抽样恢复潜变量和计数。

采样 $X_{j\,v}^{(1)}$,。 $X_{j\,v}^{(1)}$,的条件分布给定其他变量后是一个离散分布，

$$
\mathbb{P}(X_{j,v}^{(1)}=x|\text{others})\propto\frac{\Gamma(x-\alpha_0)\theta_j^x(\zeta_j+M_j+1)^{\alpha_j x}S_{\alpha_j}(n_{j,v}^{(1)},x)}{(\zeta_0+\kappa.+\Delta_j)^x}\text{ for }x\in[n_{j,v}^{(1)}].
\tag{C.25}
$$

采样 $X_{j\,l}^{(2)}$,, $\widetilde{N}_{(2)\,j,l}$**和** $\widetilde{N}_{j\,l}^{(3)}$,.注意， $\widetilde{N}_{(2)\,j,l}+\widetilde{N}_{j}^{(3)}{}_{,l}=n_{j\,l}^{(2,3)}$,。我们有三种可能的情况:

1.$x_{j,l}=0$ 和 $n_{j}^{(2,3)}{}_{,l}=0$: 在这种情况下，测试文档不包含第 l 个词，因此 $X_{j}^{(2)}{}_{,l}=\widetilde{N}_{(2)\,j,l}=\widetilde{N}_{j}^{(3)}{}_{,l}=0$的概率为 1。

2. $x_{j,\,l}=0$ 和 $n_{j}^{(2,3)}{}_{,\,l}>0$: 测试文档中的第 l 个词仅来自 $\widetilde{N}_{j\,l}^{(2)}{}_{,}$，因此 $\widetilde{N}_{j}^{(2)}{}_{,l}=n_{j\,l}^{(2,3)}{}_{,}$ 以及 $\widetilde{N}_{(3)\,j,\,l}=0$的概率为 1。给定 $\widetilde{N}_{j}^{(2)}{}_{,\,l}=n_{j\,l}^{(2,3)}{}_{,}$， $X_{j\,l}^{(2)}{}_{,}$ 的条件分布是一个离散分布。

$$
\begin{aligned}
&\mathbb{P}(X_{j,l}^{(2)}=x|\widetilde{N}_{j,l}^{(2)}=n_{j,l}^{(2,3)},\text{others})\\
&\propto\frac{\Gamma(x+x_l-\alpha_0)\theta_j^x(\zeta_j+M_j+1)^{\alpha_j x}S_{\alpha_j}(n_{j,l}^{(2,3)},x)}{(\zeta_0+\kappa.+\Delta_j)^x}\text{ for }x\in[n_{j,l}^{(2,3)}].
\end{aligned}
\tag{C.26}
$$

3. $x_{j,l} > 0$ 和 $n_j^{(2,3)}{}_{,l} > 0$：测试文档中的第 l 个词要么同时来自$\widetilde{N}_{j\,l}^{(2)}$，和 $\widetilde{N}_{(3)\,j,l}$，要么仅来自 $\widetilde{N}_{(2)\,j,l}$和 $\widetilde{N}_{(3)\,j,l}$中的一个。$X_{j\,l}^{(2)}$，的条件分布是一个离散分布，

$$\begin{aligned}&\mathbb{P}(X_{j,l}^{(2)} = x \,|\, \widetilde{N}_{j,l}^{(2)} + \widetilde{N}_{j,l}^{(3)} = n_{j,l}^{(2,3)}, \text{others})\\ &\propto \frac{\Gamma(x + x_l - \alpha_0)\theta_j^x(\zeta_j + M_j + 1)^{\alpha_j x}(S_{\alpha_j}(n_{j,l}^{(2,3)}, x))^{\mathbb{1}_{\{x>0\}}}}{(\zeta_0 + \kappa_. + \Delta_j)^x} \text{ for } x \in \{0\} \cup [n_{j,l}^{(2,3)}].\end{aligned} \tag{C.27}$$

如果从上述条件分布中抽取的 $x_{j\,l}^{(2)}$，为零，则以概率 1 有 $\widetilde{N}_j^{(2)}{}_{,l} = 0$ 和 $\widetilde{N}_j^{(3)}{}_{,l} = n^{2}{}_{,j\,l}{}^{3}{}_{,}$。否则，$\widetilde{N}_{j\,l}^{(2)}$，的条件分布为，

$$\begin{aligned}&\mathbb{P}(\widetilde{N}_{j,l}^{(2)} = n \,|\, \widetilde{N}_{j,l}^{(2)} + \widetilde{N}_{j,l}^{(3)} = n_{j,l}^{(2,3)}, X_{j,l}^{(2)} = x_{j,l}^{(2)}, \text{others})\\ &\propto \frac{S_{\alpha_j}(n, x_{j,l}^{(2)})}{n!}\frac{\Gamma(n_{j,l}^{(2,3)} - n + n_{j,l} - \alpha_j x_{j,l})}{(n_{j,l}^{(2,3)} - n)!} \text{ for } n \in [n_{j,l}^{(2,3)}].\end{aligned} \tag{C.28}$$

给定 $n_{j\,l}^{(2)}$,，从上述条件分布中抽取，我们以概率 1有 $n_j^{(3)}{}_{,l} = n_j^{(2,3)}{}_{,l} - n_{j\,l}^{(2)}{}_{,}$。

交叉验证剩余参数的估计结果

在此我们展示参数（α_j，θ_j，{v7]).

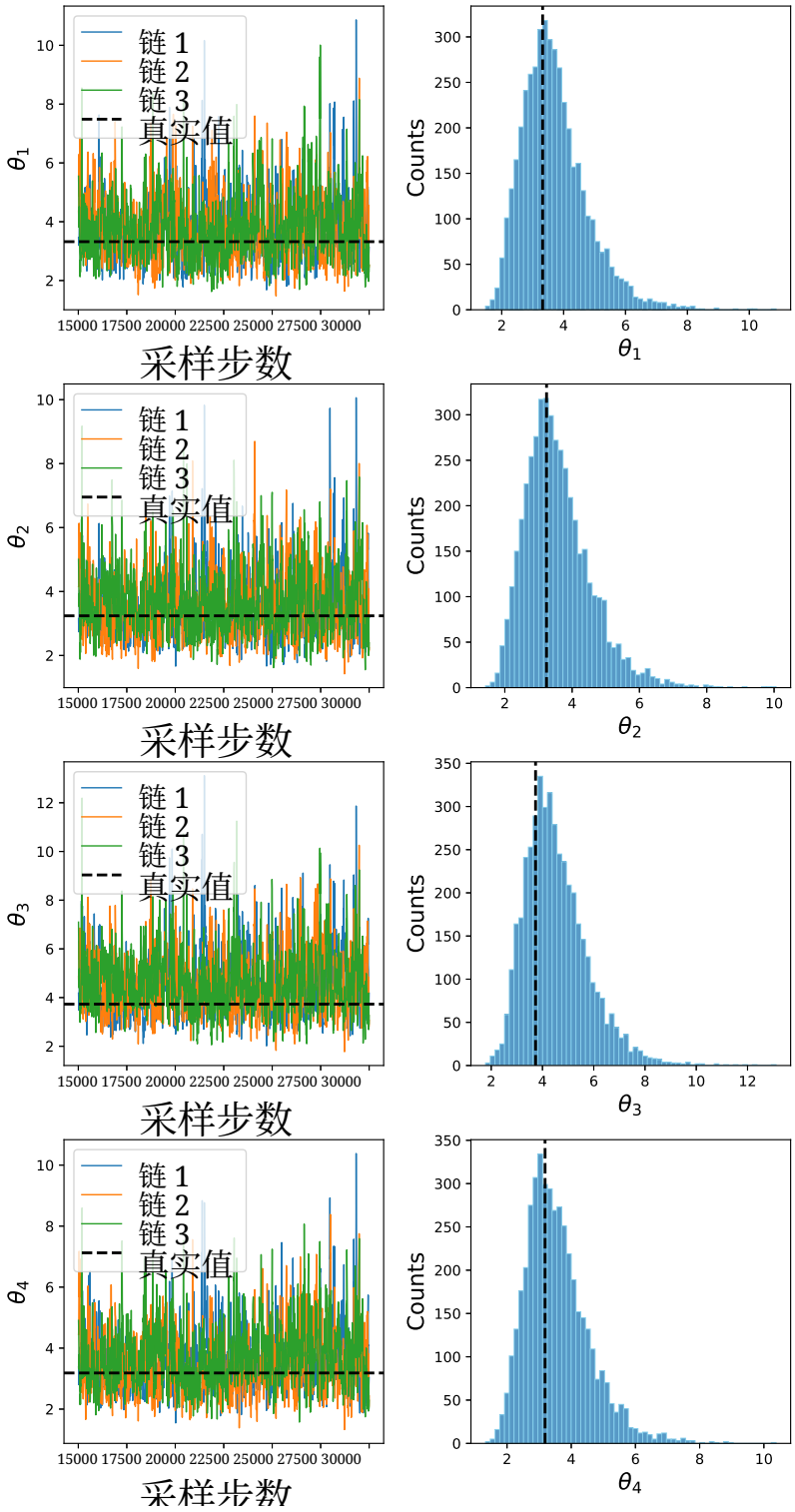

图 4：使用 $\alpha = 0.2$ 生成的数据的 θ_j 后验分布样本。

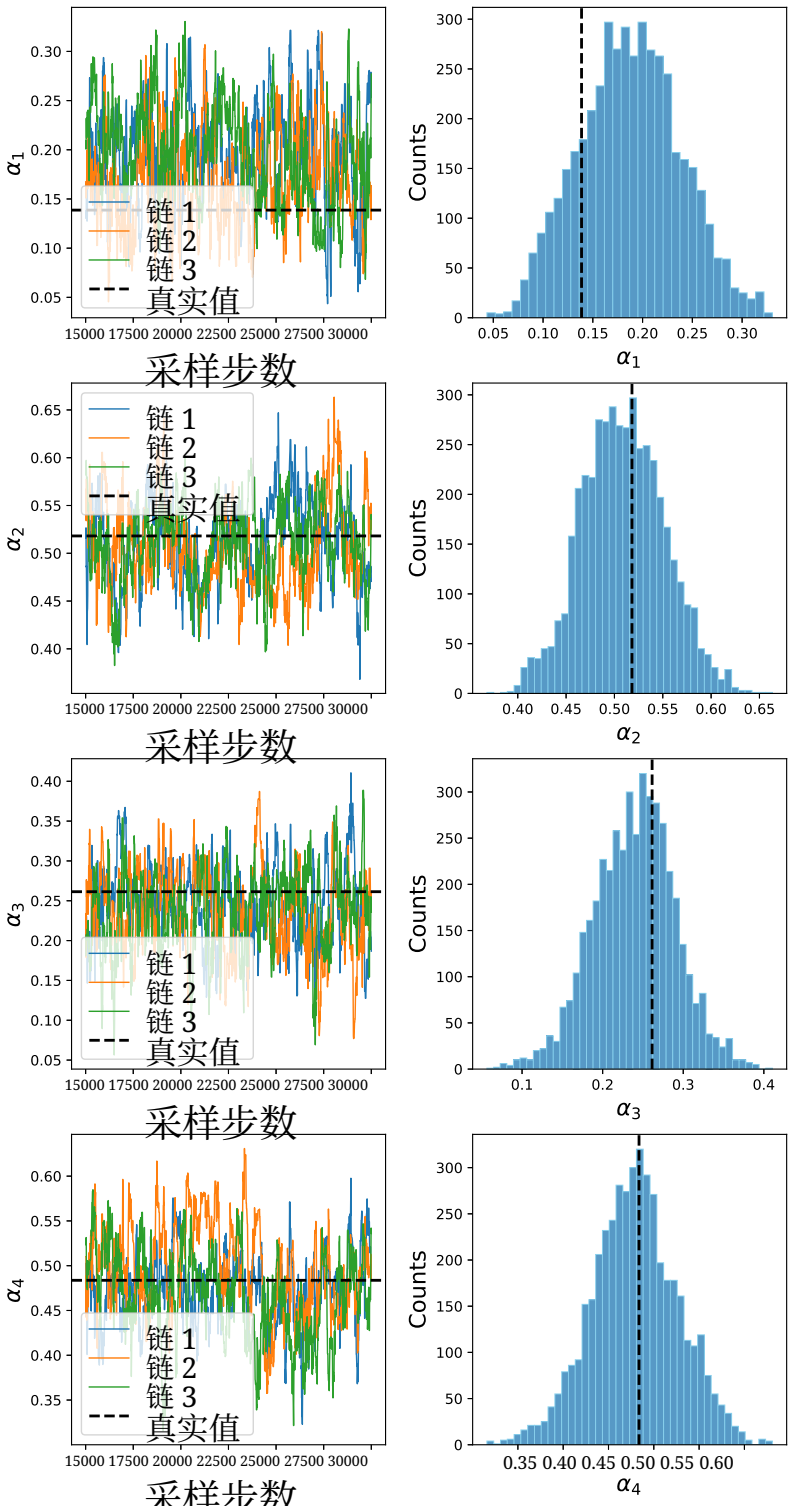

图 5：使用 $\alpha = 0.2$ 生成的数据的 α_j 后验样本。

《style id='1'>j 的后验分布，Gelman-Rubin 诊断：1.0339

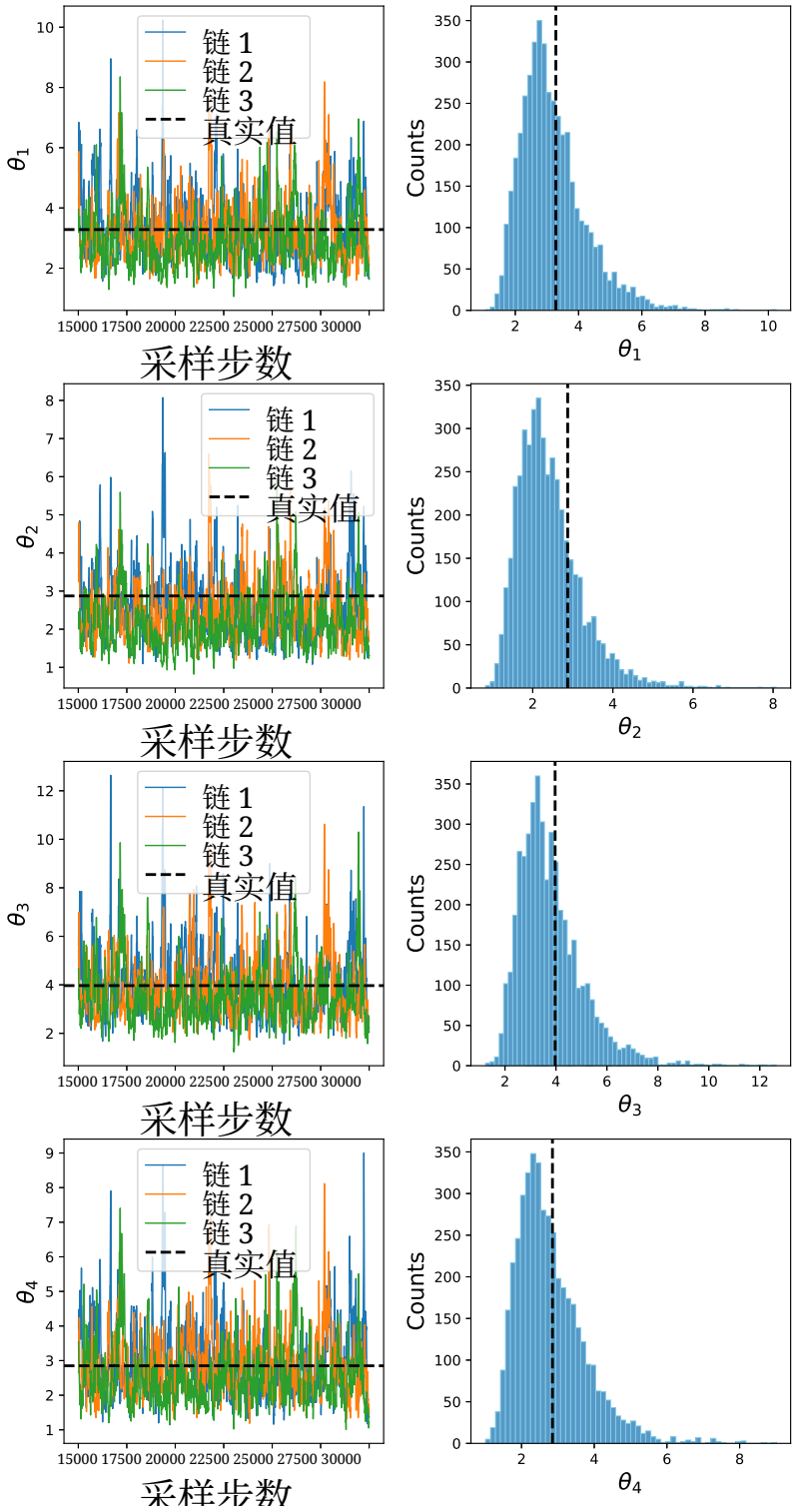

图 6：使用 $\alpha = 0.6$ 生成的数据中 θ_j 的后验样本。

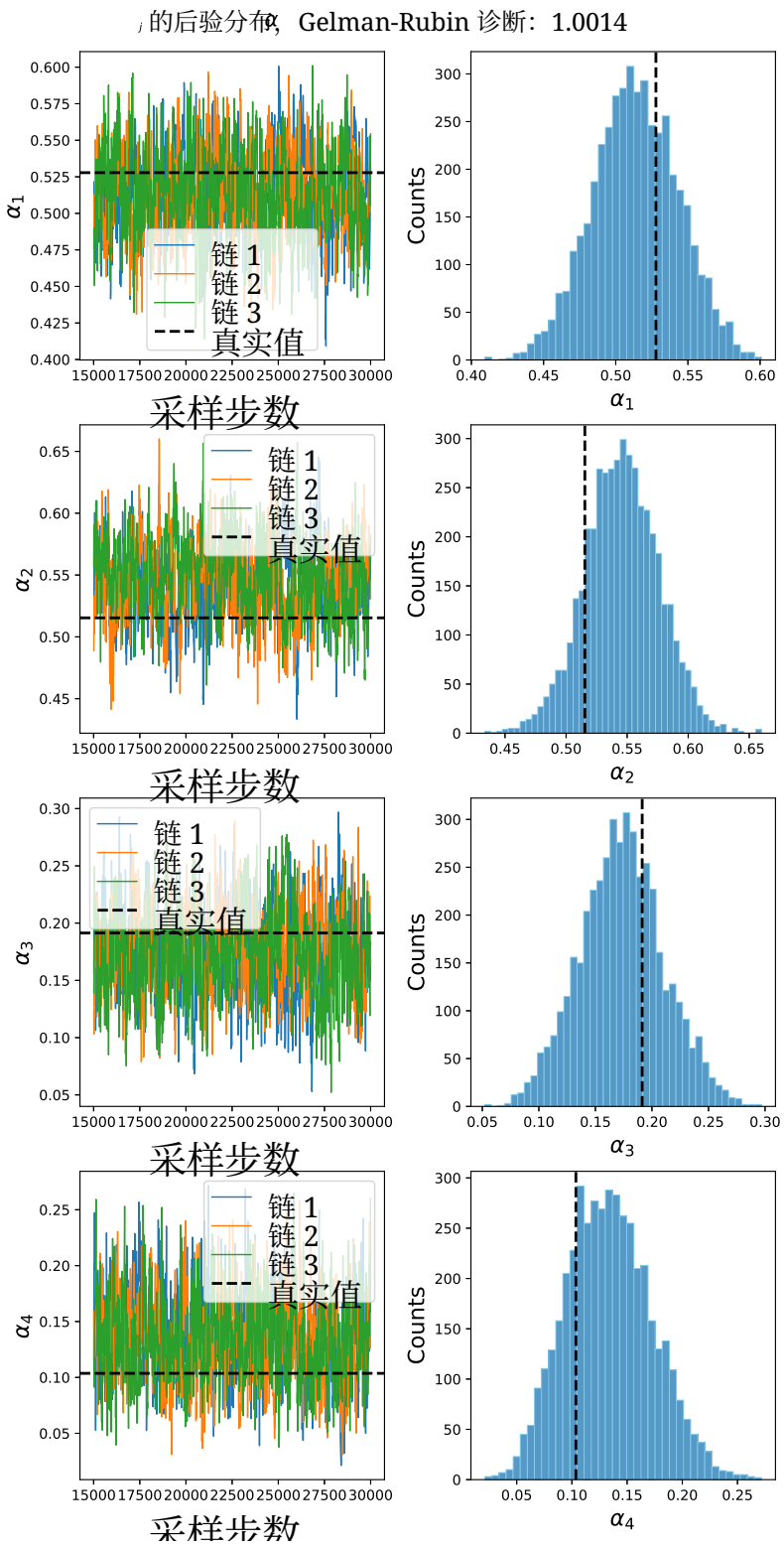

图 7：使用 $\alpha = 0.6$ 生成的数据的 α_j 后验样本。

参考文献

[1] Basbug, M. 和 Engelhardt, B. [2016], 《层次复合泊松因子分解》，载于《国际机器学习会议》，PMLR，第 1795–1803 页。[2] Beraha, M. 和 Favaro, S. [2023], 《贝叶斯非参数方法中用于性状分配的变换缩放过程先验》，arXiv 预印本 arXiv:2303.17844。[3] Broderick, T., Jordan, M. I. 和 Pitman, J. [2013], 《组合随机过程产生的簇与特征》，Statist. Sci. 28, 289–312。[4] Brooks, S. P. 和 Gelman, A. [1997], 《迭代模拟收敛监测的一般方法》，Journal of Computational and Graphical Statistics 7, 434–455。[5] Camerlenghi, F., Favaro, S., Masoero, L. 和 Broderick, T. [2024], 《用于未观测遗传变异贝叶斯非参数估计的缩放过程先验》，Journal of the American Statistical Association 119(545), 320–331。[6] Carlton, M. A. [2002], 《由三参数狄利克雷过程导出的一族密度》，Journal of Applied Probability 39, 764–774。[7] Caron, F. [2012], 《二部图的贝叶斯非参数模型》，Neural Information Processing Systems (NIPS 2012)，太浩湖，美国。[8]Crane, H. [2016], 《无处不在的 Ewens抽样公式》，Statistical Science 31(1), 1–19。[9] Ferguson, T. S. [1973], 《若干非参数问题的贝叶斯分析》，Ann. Statist. 1, 209–230。[10] Gelman, A. 和 Rubin, B. [1992], 《利用多序列迭代模拟进行推断》，Statistical Science 7, 457–511。[11] Ghahramani, Z. 和 Griffiths, T. [2005], 《无限潜在特征模型与印度自助餐过程》，神经信息处理系统进展 18。[12] Ho, M.-W., James, L. F. 和 Lau, J. W. [2021], 《吉布斯划分、黎曼–刘维尔分数阶算子、米塔格–勒夫勒函数，以及由稳定次 subordinators 导出的碎裂》，Journal of Applied Probability 58(2), 314–334。[13] Ishwaran, H. 和 James, L. F. [2001], 《截断棍棒先验的吉布斯抽样方法》，J. Amer. Stat. Assoc. 96, 161–173。[14] Ishwaran, H. 和 James, L. F. [2003], 《物种抽样混合模型的广义加权中国餐馆过程》，Statist. Sinica 13, 1211–1235。[15] James, L. F. [2002], 《泊松过程划分演算及其在可交换模型与贝叶斯非参数方法中的应用》，技术报告，未发表手稿。ArXiv math. PR/0205093。[16] James, L. F. [2017], 《通过广义印度自助餐过程先验进行潜在特征建模的贝叶斯泊松演算》，Ann. Stat. 45, 2016–2045。[17] James, L. F. [2019], 《给定物种抽样规模的截断棍棒 Pitman-Yor过程》，arXiv 预印本 arXiv:1908.07186。[18] James, L. F., Lijoi, A. 和 Pr¨ unster, I. [2009], 《具有独立增量的归一化随机测度的后验分析》，Scand. J. Stat. 36, 76–97。

[19] James,L.、Orbanz, P. 和 Teh,Y. [2015], 《缩放次 subordinators 与印度自助餐过程的推广》，arXiv 预印本arXiv:1510.07309。[20] Kolchin, V. F. [1986], 《随机映射》，《数学与工程中的翻译丛书》，Optimization SoftwareInc. 出版部，纽约。[21] Lijoi, A.、Pr¨unster, I.和 Rigon, T. [2020], 《用于混合建模的 Pitman–Yor多项分布过程》，Biometrika107，891–906。[22] Lijoi, A.、Pr¨unster,I.和 Rigon, T. [2024], 《有限维离散随机结构与贝叶斯聚类》，J.Amer. Stat.Assoc. 119，929–941。[23] Masoero, L.、Camerlenghi, F.、Favaro,S. 和 Broderick, T. [2018], 《性状分配模型中层次化完全随机测度的后验分布表示》，BNP-NeurIPS。2018。[24] Masoero, L.、Camerlenghi, F.、Favaro,S.和 Broderick,T. [2022], 《以少博多：通过贝叶斯非参数方法预测并最大化基因组变异发现》，Biometrika 109(1)，17–32。[25] Pitman,J. [1996], 《Blackwell-MacQueen urn 方案的一些发展》，《统计学、概率与博弈论》IMS Lecture Notes Monogr. Ser.，30，Inst. Math. Statist.，Hayward，CA，245–267。[26] Pitman, J. [1997], 《源自布朗运动与稳定次 subordinators 的分区结构》，伯努利 3，79–96。[27] Pitman,J. [2003], Poisson-Kingman 分区，载于 R. Goldstein 编，《Institute of Mathematical Statistics》，Hayward，加利福尼亚州，第 1–34 页。[28] Pitman, J. [2006], 《组合随机过程》，2002 年 7 月7–24 日在Saint-Flour 举行的第 32届概率论暑期学校课程讲义。Jean Picard作序。《数学讲义》，1875，Springer-Verlag，柏林。[29] Pitman, J. [2017], 用于聚类与物种抽样的混合泊松和负二项分布模型。准备中的手稿。[30] Pitman, J. 和 Yor, M. [1997], 《源自稳定次subordinator 的双参数泊松–狄利克雷分布》，Ann. Probab. 25，855–900。[31] Tavar´e,S. [2021], 《神奇的Ewens抽样公式》，《伦敦数学学会通报》53(6)，1563–1582。[32] Teh, Y. W.、Jordan, M. I.、Beal, M. J. 和 Blei, D. M. [2006], 《层次狄利克雷过程》，Journal of the AmericanStatistical Association。101，1566–1581。[33] Thibaux, R.和 Jordan, M. I. [2007], 层次贝塔过程与印度自助餐过程，载于《人工智能与统计国际会议》，第 564–571页。[34] Titsias, M.K. [2008], 《无限伽马分布–泊松特征模型》，神经信息处理系统进展。2008。[35] Zhou,M.和 Carin, L. [2015], 《负二项分布过程计数与混合建模》，IEEE Transactions on Pattern AnalysisandMachine Intelligence37，307–320。[36] Zhou,M.、Favaro,S. 和Walker,S. G. [2017], 《频数的频数分布与规模依赖的可交换随机分区》，Journal of the American Statistical Association 112(520)，1623–1635。

[37] Zhou, M., Madrid-Padilla,O. H. 和 Scott,J. G. [2016], ‘由一族负二项分布过程导出的随机计数矩阵的先验。’,《美国统计学会杂志》**111**,1144–1156。