mathematics

MDPI

Article

On Johnson's "Sufficientness" Postulates for Feature-Sampling Models

Federico Camerlenghi [1,2,3,*] **and Stefano Favaro** [2,4,5]

1 Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy
2 Collegio Carlo Alberto, Piazza V. Arbarello 8, 10122 Torino, Italy; stefano.favaro@unito.it
3 BIDSA, Bocconi University, via Röntgen 1, 20136 Milano, Italy
4 Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy
5 IMATI-CNR "Enrico Magenes", 20133 Milano, Italy
* Correspondence: federico.camerlenghi@unimib.it

Abstract: In the 1920s, the English philosopher W.E. Johnson introduced a characterization of the symmetric Dirichlet prior distribution in terms of its predictive distribution. This is typically referred to as Johnson's "sufficientness" postulate, and it has been the subject of many contributions in Bayesian statistics, leading to predictive characterization for infinite-dimensional generalizations of the Dirichlet distribution, i.e., species-sampling models. In this paper, we review "sufficientness" postulates for species-sampling models, and then investigate analogous predictive characterizations for the more general feature-sampling models. In particular, we present a "sufficientness" postulate for a class of feature-sampling models referred to as Scaled Processes (SPs), and then discuss analogous characterizations in the general setup of feature-sampling models.

Keywords: Bayesian nonparametrics; exchangeability; feature-sampling model; de Finetti theorem; Johnson's "sufficientness" postulate; predictive distribution; scaled process prior; species-sampling model

Citation: Camerlenghi, F.; Favaro, S. On Johnson's "Sufficientness" Postulates for Feature-Sampling Models. *Mathematics* **2021**, *9*, 2891. https://doi.org/10.3390/math9222891

Academic Editors: Emanuele Dolera and Federico Bassetti

Received: 10 October 2021
Accepted: 10 November 2021
Published: 13 November 2021

1. Introduction

Exchangeability (de Finetti [1]) provides a natural modeling assumption in a large variety of statistical problems, and it amounts to the assumption that the order in which observations are recorded is not relevant. Consider a sequence of random variables $(Z_j)_{j\geq 1}$ defined on a common probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in an arbitrary space, which is assumed to be Polish. The sequence $(Z_j)_{j\geq 1}$ is exchangeable if and only if

$$(Z_1, \ldots, Z_n) \stackrel{\mathrm{d}}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})$$

for any permutation σ of the set $\{1, \ldots, n\}$ and any $n \geq 1$. By virtue of the celebrated de Finetti representation theorem, exchangeability of $(Z_j)_{j\geq 1}$ is tantamount to asserting the existence of a random element $\tilde{\mu}$, defined on a (parameter) space Θ, such that, conditionally on $\tilde{\mu}$, the Z_js are independent and identically distributed with common distribution $p_{\tilde{\mu}}$, i.e.,

$$\begin{aligned} Z_j \,|\, \tilde{\mu} &\stackrel{\text{iid}}{\sim} p_{\tilde{\mu}} \quad j \geq 1 \\ \tilde{\mu} &\sim \mathscr{M}, \end{aligned} \tag{1}$$

where $\mathscr{M}$ is the distribution of $\tilde{\mu}$. In a Bayesian setting, $\mathscr{M}$ takes on the interpretation of a prior distribution for the parameter object of interest. In this sense, the de Finetti representation theorem is a natural framework for Bayesian statistics. For mathematical convenience,

Θ is assumed to be a Polish space, equipped with the Borel σ-algebra $\mathscr{B}(\Theta)$. Hereafter, with the term *parameter*, we refer to both a finite- and an infinite-dimensional object.

Within the framework of exchangeability (1), a critical role is played by the predictive distributions, namely, the conditional distributions of the $(n+1)$th observation Z_{n+1} given $\boldsymbol{Z}_n := (Z_1, \ldots, Z_n)$. The problem of characterizing prior distributions $\mathscr{M}$ in terms of their predictive distributions has a long history in Bayesian statistics, starting from the seminal work of the English philosopher Johnson [2] who provided a predictive characterization of the symmetric Dirichlet prior distribution. Such a characterization is typically referred to as Johnson's "sufficientness" postulate. Species-sampling models (Pitman [3]) provide arguably the most popular infinite-dimensional generalization of the Dirichlet distribution. They form a broad class of nonparametric prior models that correspond to the assumption that $p_{\tilde{p}}$ in (1) is an almost surely discrete random probability measure

$$\tilde{p} = \sum_{i \geq 1} \tilde{p}_i \delta_{\tilde{z}_i}, \tag{2}$$

where: (i) $(\tilde{p}_i)_{i \geq 1}$ are non-negative random weights almost surely summing up to 1; (ii) $(\tilde{z}_i)_{i \geq 1}$ are random species' labels, independent of $(\tilde{p}_i)_{i \geq 1}$, and i.i.d. with common (non-atomic) distribution P. The term *species* refers to the fact that the law of $\tilde{p}$ is a prior distribution for the unknown species composition $(\tilde{p}_i)_{i \geq 1}$ of a population of individuals Z_js, with Z_j belonging to a species $\tilde{z}_i$ with probability $\tilde{p}_i$ for $j, i \geq 1$. In the context of species-sampling models, Regazzini [4] and Lo [5] provided a "sufficientness" postulate for the Dirichlet process (Ferguson [6]). Such a characterization was then extended by Zabell [7] to the Pitman–Yor process (Perman et al. [8], Pitman and Yor [9]) and by Bacallado et al. [10] to the more general Gibbs-type prior models (Gnedin and Pitman [11]).

In this paper, we introduce and discuss Johnson's "sufficientness" postulates in the feature-sampling setting, which generalizes the species-sampling setting by allowing each individual of the population to belong to multiple species, now called features. We point out that feature-sampling models are extremely important in different areas of application; see, e.g., Griffiths and Ghahramani [12], Ayed et al. [13] and the references therein. Under the framework of exchangeability (1), the feature-sampling setting assumes that

$$Z_j | \tilde{\mu} = \sum_{i \geq 1} A_{j,i} \delta_{\tilde{w}_i} \sim p_{\tilde{\mu}}, \tag{3}$$

and

$$\tilde{\mu} = \sum_{i \geq 1} \tilde{p}_i \delta_{\tilde{w}_i}$$

where: (i) conditionally on $\tilde{\mu}$, $(A_{j,i})_{i \geq 1}$ are independent Bernoulli random variables with parameters $(\tilde{p}_i)_{i \geq 1}$; (ii) $(\tilde{p}_i)_{i \geq 1}$ are $(0,1)$-valued random weights; (iii) $(\tilde{w}_i)_{i \geq 1}$ are random features' labels, independent of $(\tilde{p}_i)_{i \geq 1}$, and i.i.d. with common (non-atomic) distribution P. That is, individual Z_j displays feature $\tilde{w}_i$ if and only if $A_{j,i} = 1$, which happens with probability $\tilde{p}_i$. For example, if, conditionally on $\tilde{\mu}$, Z_j displays only two features, say $\tilde{w}_1$ and $\tilde{w}_5$, it equals the random measure $\delta_{\tilde{w}_1} + \delta_{\tilde{w}_5}$. The distribution $p_{\tilde{\mu}}$ is the law of a Bernoulli process with parameter $\tilde{\mu}$, which is denoted by $\text{BeP}(\tilde{\mu})$, whereas the law of $\tilde{\mu}$ is a nonparametric prior distribution for the unknown feature probabilities $(\tilde{p}_i)_{i \geq 1}$, i.e., a feature-sampling model. Here, we investigate the problem of characterizing prior distributions for $\tilde{\mu}$ in terms of their predictive distributions, with the goal of providing "sufficientness" postulates for feature-sampling models. We discuss such a problem and present partial results for a class of feature-sampling models referred to as Scaled Process (SP) priors for $\tilde{\mu}$ (James et al. [14], Camerlenghi et al. [15]). With these results, we aim at stimulating future research in this field to obtain "sufficientness" postulates for general feature-sampling models.

The paper is structured as follows. In Section 2, we present a brief review on Johnson's "sufficientness" postulates for species-sampling models. Section 3 focuses on nonparametric

prior models for the Bernoulli process, i.e., feature-sampling models; we review their definitions, properties, and sampling structures. In Section 4, we present a "sufficientness" postulate for SPs. Section 5 concludes the paper by discussing our results and conjecturing analogous results for more general classes of feature-sampling models.

2. Species-Sampling Models

To introduce species-sampling models, we assume that the observations are $\mathbb{Z}$-valued random elements, and $\mathbb{Z}$ is supposed to be a Polish space whose Borel σ-algebra is denoted by $\mathscr{Z}$. Thus, $\mathbb{Z}$ contains all the possible species' labels of the populations. When we deal with species-sampling models, the hierarchical formulation (1) specializes as

$$\begin{aligned} Z_j|\tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \quad j \geq 1 \\ \tilde{p} &\sim \mathscr{M} \end{aligned} \tag{4}$$

where $\tilde{p} = \sum_{i\geq 1} \tilde{p}_i \delta_{\tilde{z}_i}$ is an almost surely discrete random probability measure on $\mathbb{Z}$, and $\mathscr{M}$ denotes its law. We also remind the reader that: (i) $(\tilde{p}_i)_{i\geq 1}$ are non-negative random weights almost surely summing up to 1; (ii) $(\tilde{z}_i)_{i\geq 1}$ are random species' labels, independent of $(\tilde{p}_i)_{i\geq 1}$, and i.i.d. as a common (non-atomic) distribution P. Using the terminology of Pitman [3], the discrete random probability measure $\tilde{p}$ is a *species-sampling model*. In Bayesian nonparametrics, popular examples of species-sampling models are: the Dirichlet process (Ferguson [6]), the Pitman–Yor process (Perman et al. [8], Pitman and Yor [9]), and the normalized generalized Gamma process (Brix [16], Lijoi et al. [17]). These are examples belonging to a peculiar subclass of species-sampling models, which are referred to as Gibbs-type prior models (Gnedin and Pitman [11], De Blasi et al. [18]). More general subclasses of species-sampling models are, e.g., the homogeneous normalized random measures (Regazzini et al. [19]) and the Poisson–Kingman models (Pitman [20,21]). We refer to Lijoi and Prünster [22] and Ghosal and van der Vaart [23] for a detailed and stimulating account on species-sampling models and their use in Bayesian nonparametrics.

Because of the almost sure discreteness of $\tilde{p}$ in (4), a random sample $\boldsymbol{Z}_n := (Z_1, \ldots, Z_n)$ from $\tilde{p}$ features ties, that is, $\mathbb{P}(Z_{j_1} = Z_{j_2}) > 0$ if $j_1 \neq j_2$. Thus, $\boldsymbol{Z}_n$ induces a random partition of the set $\{1, \ldots, n\}$ into $K_n = k \leq n$ blocks, labeled by $Z_1^*, \ldots, Z_{K_n}^*$, with corresponding frequencies $(N_{n,1}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)$, such that $N_{i,n} \geq 1$ and $\sum_{1\leq i\leq K_n} N_{i,n} = n$. From Pitman [3], the predictive distribution of $\tilde{p}$ is of the form

$$\mathbb{P}(Z_{n+1} \in A|\boldsymbol{Z}_n) = g(n,k,\boldsymbol{n})P(A) + \sum_{i=1}^{k} f_i(n,k,\boldsymbol{n})\delta_{Z_i^*}(A), \quad A \in \mathscr{Z}, \tag{5}$$

for any $n \geq 1$, having set $\boldsymbol{n} = (n_1, \ldots, n_k)$, with g and f_i being arbitrary non-negative functions that satisfy the constraint $g(n,k,\boldsymbol{n}) + \sum_{i=1}^{k} f_i(n,k,\boldsymbol{n}) = 1$. The predictive distribution (5) admits the following interpretation: (i) $g(n,k,\boldsymbol{n})$ corresponds to the probability that Z_{n+1} is a new species, that is, a species not observed in $\boldsymbol{Z}_n$; (ii) $f_i(n,k,\boldsymbol{n})$ corresponds to the probability that Z_{n+1} is a species Z_i^* in $\boldsymbol{Z}_n$. The functions g and f_i completely determine the distribution of the exchangeable sequence $(Z_j)_{j\geq 1}$ and, in turn, the distribution of the random partition of $\mathbb{N}$ induced by $(Z_j)_{j\geq 1}$. Predictive distributions of popular species-sampling models, e.g., the Dirichlet process, the Pitman–Yor process, and the normalized generalized Gamma process, are of the form (5) for suitable specification of the functions g and f_i. We refer to Pitman [21] for a detailed account of random partitions induced by species-sampling models and generalizations thereof.

Here, we recall the predictive distribution of Gibbs-type prior models (Gnedin and Pitman [11], De Blasi et al. [18]). Let us first introduce the definition of these processes.

Definition 1. *Let $\sigma \in (-\infty, 1)$ and let P be a (non-atomic) distribution on $(\mathbb{Z}, \mathscr{Z})$. A Gibbs-type prior model is a species-sampling model with a predictive distribution of the form*

$$\mathbb{P}(Z_{n+1} \in A | \boldsymbol{Z}_n) = \frac{V_{n+1,k+1}}{V_{n,k}} P(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{k} (n_i - \sigma) \delta_{Z_i^*}(A), \quad A \in \mathscr{Z}, \tag{6}$$

for any $n \geq 1$, where $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ is a collection of non-negative weights that satisfy the recurrence relation $V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1}$ for all $k = 1, \ldots, n$, $n \geq 1$, with the proviso $V_{1,1} = 1$.

Note that the Dirichlet process is a Gibbs-type prior model that corresponds to

$$V_{n,k} = \frac{\theta^k}{(\theta)_n}$$

for $\theta > 0$, where we have denoted by $(a)_b = \Gamma(a + b)/\Gamma(a)$ the Pochhammer symbol for the rising factorials. Moreover, the Pitman–Yor process is a Gibbs-type prior model corresponding to

$$V_{n,k} = \frac{\prod_{i=0}^{k-1} (\theta + i\sigma)}{(\theta)_n}$$

for $\sigma \in (0, 1)$ and $\theta > -\alpha$. We refer to Pitman [20] for other examples of Gibbs-type prior models and for a detailed account of the $V_{n,k}$s; see also Pitman [21] and the references therein.

Because of de Finetti's representation theorem, there exists a one-to-one correspondence between the functions g and f_i in the predictive distribution (5) and the law $\mathscr{M}$ of $\tilde{p}$, i.e., the de Finetti measure. This is at the basis of Johnson's "sufficientness" postulates, characterizing species-sampling models through their predictive distributions. Regazzini [4] and, later, Lo [5] provided the first "sufficientness" postulate for species-sampling models, showing that the Dirichlet process is the unique species-sampling model for which the function g depends on $\boldsymbol{Z}_n$ only through n, and the function f_i depends on $\boldsymbol{Z}_n$ only through n and n_i for $i \geq 1$. Such a result was extended in Zabell [24], providing the following "sufficientness" postulate for the Pitman–Yor process: The Pitman–Yor process is the unique species-sampling model for which the function g depends on $\boldsymbol{Z}_n$ only through n and k, and the function f_i depends on $\boldsymbol{Z}_n$ only through n and n_i for $i \geq 1$. Bacallado et al. [10] discussed the "sufficientness" postulate in the more general setting of Gibbs-type prior models, showing that Gibbs-type prior models are the sole species-sampling models for which the function g depends on $\boldsymbol{Z}_n$ only through n and k, and the function f_i depends on $\boldsymbol{Z}_n$ only through n, k, and n_i. This result shows a critical difference—at the sampling level—between the Pitman–Yor process and Gibbs-type prior models, which lies in the inclusion of the sampling information on the observed number of distinct species in the probability of observing, at the $(n + 1)$-th draw, a species already observed in the sample.

3. Feature-Sampling Models

Feature-sampling models generalize species-sampling models by allowing each individual to belong to more than one species, which are now called features. To introduce feature-sampling models, we consider a space of features $\mathbb{W}$, which is assumed to be a Polish space, and we denote by $\mathscr{W}$ its Borel σ-field. Thus, $\mathbb{W}$ contains all the possible features' labels of the population. Observations are represented through the counting measure (3), whose parameter $\tilde{\mu}$ is an almost surely discrete measure with masses in $(0, 1)$. When we deal with feature-sampling models, the hierarchical formulation (1) specializes as

$$\begin{aligned} Z_j | \tilde{\mu} &\overset{\text{iid}}{\sim} \text{BeP}(\tilde{\mu}) \\ \tilde{\mu} &\sim \mathscr{M} \end{aligned} \tag{7}$$

where $\tilde{\mu} = \sum_{i\geq 1} \tilde{p}_i \delta_{\tilde{w}_i}$ is an almost surely discrete random measure on $\mathbb{W}$, and $\mathscr{M}$ denotes its law. We also remind the reader that: (i) conditionally on $\tilde{\mu}$, $(A_{j,i})_{i\geq 1}$ are independent Bernoulli random variables with parameters $(\tilde{p}_i)_{i\geq 1}$; (ii) $(\tilde{p}_i)_{i\geq 1}$ are $(0,1)$-valued random weights; (iii) $(\tilde{w}_i)_{i\geq 1}$ are random features' labels, independent of $(\tilde{p}_i)_{i\geq 1}$, and i.i.d. with common (non-atomic) distribution P. Completely random measures (CRMs) (Daley and Vere-Jones [25], Kingman [26]) provide a popular class of nonparametric priors $\mathscr{M}$, the most common examples of which are the Beta process prior and the stable Beta process prior (Teh and Gorur [27], James [28]); see also Broderick et al. [29] and the references therein for other examples of CRM priors and generalizations thereof. Recently, Camerlenghi et al. [15] investigated an alternative class of nonparametric priors $\mathscr{M}$, generalizing CRM priors and referring to these as Scaled Processes (SPs). SP priors first appeared in the work of James [28].

We assume a random sample $\boldsymbol{Z}_n := (Z_1, \ldots, Z_n)$ to be modeled as in (7), and we introduce the predictive distribution of $\tilde{\mu}$, that is, the conditional probability of Z_{n+1} given $\boldsymbol{Z}_n$. Note that, because of the pure discreteness of $\tilde{\mu}$, the observations $\boldsymbol{Z}_n$ may share a random number of distinct features, say $K_n = k$, denoted here as $W_1^*, \ldots, W_{K_n}^*$, and each feature W_i^* is displayed exactly by $M_{n,i} = m_i$ of the n individuals as $i = 1, \ldots, k$. Since the features' labels are immaterial and i.i.d. form the base measure P, the conditional distribution of Z_{n+1}, given $\boldsymbol{Z}_n$, may be equivalently characterized through the vector $(Y_{n+1}, A_{n+1,1}^*, \ldots, A_{n+1,K_n}^*)$, where: (i) Y_{n+1} is the number of new features displayed by the $(n+1)$th individual, namely, hitherto unobserved out of the sample $\boldsymbol{Z}_n$; (ii) $A_{n+1,i}^*$ is a $\{0,1\}$-valued random variable for any $i = 1, \ldots, K_n$, and $A_{n+1,i}^* = 1$ if the $(n+1)$th individual displays feature W_i^*; it equals 0 otherwise. Hence, the predictive distribution of $\tilde{\mu}$ is

$$\mathbb{P}((Y_{n+1}, A_{n+1,1}^*, \ldots, A_{n+1,K_n}^*) = (y, a_1, \ldots, a_{K_n}) | \boldsymbol{Z}_n) = f(y, a_1, \ldots, a_k; n, k, \boldsymbol{m}) \tag{8}$$

where we denote by f a probability distribution evaluated at $(y, a_1, \ldots, a_k)$, and where n, k and $\boldsymbol{m} := (m_1, \ldots, m_k)$ is the sampling information. In the rest of this section, we specify the function f under the assumption of a CRM prior and an SP prior, showing its dependence on n, K_n, and $(M_{n,1}, \ldots, M_{n,K_n})$. In particular, we show how SP priors allow one to enrich the predictive distribution of CRM priors by including additional sampling information in terms of the number of distinct features and their corresponding frequencies.

3.1. Priors Based on CRMs

Let $\mathsf{M}_{\mathbb{W}}$ denote the space of all bounded and finite measures on $(\mathbb{W}, \mathscr{W})$, that is to say, $\mu \in \mathsf{M}_{\mathbb{W}}$ iff $\mu(A) < +\infty$ for any bounded set $A \in \mathscr{W}$. Here, we recall the definition of a Completely Random Measure (CRM) (see, e.g., Daley and Vere-Jones [25]).

Definition 2. *A Completely Random Measure (CRM) $\tilde{\mu}$ on $(\mathbb{W}, \mathscr{W})$ is a random element taking values in the space $\mathsf{M}_{\mathbb{W}}$ such that the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are independent for any choice of bounded and disjoint sets $A_1, \ldots, A_n \in \mathscr{W}$ and for any $n \geq 1$.*

We remind the reader that Kingman [26] proved that a CRM may be decomposed as the sum of a deterministic drift and a purely atomic component. In Bayesian nonparametrics, it is common to consider purely atomic CRMs without fixed points of discontinuity, that is to say, $\tilde{\mu}$ may be represented as $\tilde{\mu} := \sum_{i\geq 1} \tilde{\eta}_i \delta_{\tilde{w}_i}$, where $(\tilde{\eta}_i)_{i\geq 1}$ is a sequence of random atoms and $(\tilde{w}_i)_{i\geq 1}$ are the random locations. An appealing property of purely atomic CRMs is the availability of their Laplace functional; indeed, for any measurable function $f : \mathbb{W} \to \mathbb{R}^+$, one has

$$\mathbb{E}\left[e^{-\int_{\mathbb{W}} f(w)\tilde{\mu}(\mathrm{d}w)}\right] = \exp\left\{-\int_{\mathbb{W}\times\mathbb{R}^+} (1 - e^{-sf(w)})\nu(\mathrm{d}w, \mathrm{d}s)\right\} \tag{9}$$

where ν is a measure on $\mathbb{W} \times \mathbb{R}^+$ called the Lévy intensity of the CRM $\tilde{\mu}$, and it is such that

$$\nu(\{w\} \times \mathbb{R}^+) = 0 \quad \forall w \in \mathbb{W}, \quad \text{and} \int_{A \times \mathbb{R}^+} \min\{s, 1\} \nu(\mathrm{d}w, \mathrm{d}s) < \infty \tag{10}$$

for any bounded Borel set A. Here, we focus on homogeneous CRMs by assuming that the atoms $\tilde{\eta}_i$s and the locations $\tilde{w}_i$s are independent; in this case, the Lévy measure may be written as

$$\nu(\mathrm{d}w, \mathrm{d}s) = \lambda(s)\mathrm{d}s P(\mathrm{d}w)$$

for some measurable function $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ and a probability measure P on $(\mathbb{W}, \mathscr{W})$, called the *base measure*, which is assumed to be diffuse. In this case, the distribution of $\tilde{\mu}$ will be denoted as $\mathrm{CRM}(\lambda; P)$, and the second integrability condition in (10) reduces to the following:

$$\int_{\mathbb{R}^+} \min\{s, 1\} \lambda(s) \mathrm{d}s < +\infty. \tag{11}$$

In the feature-sampling framework, $\tilde{\mu}$ may be used as a prior distribution if the sequence of atoms $(\tilde{\eta}_i)_{i \geq 1}$ is in between $[0, 1]$, which happens if the Lévy intensity has support on $\mathbb{W} \times [0, 1]$. A noteworthy example, widely used in this setting, is the stable Beta process prior (Teh and Gorur [27]). It is defined as a CRM with Lévy intensity

$$\lambda(s) = \alpha \cdot \frac{\Gamma(1 + c)}{\Gamma(1 - \sigma)\Gamma(c + \sigma)} s^{-1-\sigma}(1 - s)^{c+\sigma-1} \mathbb{1}_{(0,1)}(s) \tag{12}$$

where $c > 0$, $\sigma \in (0, 1)$, and $\alpha > 0$ (James [28], Masoero et al. [30]). Now, we describe the predictive distribution for an arbitrary CRM $\tilde{\mu}$. For the sake of clarity, we fix the following notation:

$$\mathrm{Poiss}(y; C) := \frac{C^y e^{-C}}{y!}, y \in \mathbb{N} \text{ and } \mathrm{Bern}(a; p) := p^a (1 - p)^{1-a}, a \in \{0, 1\}$$

to denote the probability mass functions of a Poisson with parameter $C > 0$ and a Bernoulli random variable with parameter $p \in [0, 1]$, respectively. We refer to James [28] for a detailed posterior analysis of CRM priors; see also Broderick et al. [29] and the references therein.

Theorem 1 (James [28])**.** *Let $Z_1, Z_2, \ldots$ be exchangeable random variables modeled as in (7), where $\mathscr{M}$ equals $\mathrm{CRM}(\lambda; P)$. If $\mathbf{Z}_n$ is a random sample that displays $K_n = k$ distinct features $\{W_1^*, \ldots, W_{K_n}^*\}$, and feature W_i^* appears exactly $M_{n,i} = m_i$ times in the samples, such as $i = 1, \ldots, K_n$, then*

$$\begin{aligned} &\mathbb{P}((Y_{n+1}, A_{n+1,1}^*, \ldots, A_{n+1,K_n}^*) = (y, a_1, \ldots, a_{K_n}) | \mathbf{Z}_n) \\ &\quad = \mathrm{Poiss}\left(y; \int_0^1 s(1 - s)^n \lambda(s) \mathrm{d}s\right) \prod_{i=1}^{k} \mathrm{Bern}(a_i; p_i^*) \end{aligned} \tag{13}$$

being

$$p_i^* := \frac{\int_0^1 s^{m_i+1}(1 - s)^{n-m_i} \lambda(s) \mathrm{d}s}{\int_0^1 s^{m_i}(1 - s)^{n-m_i} \lambda(s) \mathrm{d}s}.$$

Proof. We consider James [28] (Proposition 3.2) for Bernoulli product models (see also Camerlenghi et al. [15] (Proposition 1)); thus, the distribution of Z_{n+1}, given $\mathbf{Z}_n$, equals the distribution of

$$Z'_{n+1} + \sum_{i=1}^{K_n} A_{n+1,i}^* \delta_{W_i^*}, \tag{14}$$

where $Z'_{n+1}|\tilde{\mu}' = \sum_{i\geq 1} A'_{n+1,i}\delta_{\tilde{w}'_i} \sim \text{BeP}(\tilde{\mu}')$ such that $\tilde{\mu}' \sim \text{CRM}((1-s)^n\lambda; P)$, and $A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}$ are Bernoulli random variables with parameters $J_1, \ldots, J_{K_n}$, respectively, such that each J_i is a random variable whose distribution is with a density function of the form

$$f_{J_i}(s) \propto (1-s)^{n-m_i} s^{m_i}\lambda(s).$$

By exploiting the previous predictive characterization, we can derive the posterior distribution of Y_{n+1} given $\boldsymbol{Z}_n$ by means of a direct application of the Laplace functional. Indeed, the distribution of $Y_{n+1}|\boldsymbol{Z}_n$ equals $\sum_{i\geq 1} A'_{n+1,i}$. Thus, for any $t \in \mathbb{R}$, we have the following:

$$\begin{aligned}\mathbb{E}[e^{-tY_{n+1}}|\boldsymbol{Z}_n] &= \mathbb{E}[e^{-t\sum_{i\geq 1} A'_{n+1,i}}] = \mathbb{E}\Big[\prod_{i\geq 1} e^{-tA'_{n+1,i}}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\prod_{i\geq 1} e^{-tA'_{n+1,i}} \mid \tilde{\mu}'\Big]\Big] \\ &= \mathbb{E}\Big[\prod_{i\geq 1}\Big(e^{-t}\tilde{\eta}'_i + (1-\tilde{\eta}'_i)\Big)\Big],\end{aligned}$$

where we used the representation $\tilde{\mu}' = \sum_{i\geq 1}\tilde{\eta}'_i\delta_{\tilde{w}'_i}$ and the fact that the $A_{n+1,i}$s are independent Bernoulli random variables conditionally on $\tilde{\mu}'$. We now use the Laplace functional for $\tilde{\mu}'$ to get

$$\begin{aligned}\mathbb{E}[e^{-tY_{n+1}}|\boldsymbol{Z}_n] &= \mathbb{E}\left[\exp\left\{\sum_{i\geq 1}\log(1+\tilde{\eta}'_i(e^{-t}-1))\right\}\right] \\ &= \exp\left\{-(1-e^{-t})\int_0^1 (1-s)^n s\lambda(s)\mathrm{d}s\right\}.\end{aligned}$$

As a direct consequence, the posterior distribution of Y_{n+1} given $\boldsymbol{Z}_n$ is a Poisson distribution with mean $\int_0^1 (1-s)^n s\lambda(s)\mathrm{d}s$. Again, by exploiting the predictive representation (14), the posterior distribution of $A^*_{n+1,i}$, as $i = 1, \ldots, K_n$, is a Bernoulli with the following mean:

$$\mathbb{E}[J_i] = \int_0^1 s f_{J_i}(s)\mathrm{d}s = \frac{\int_0^1 (1-s)^{n-m_i}s^{m_i+1}\lambda(s)\mathrm{d}s}{\int_0^1 (1-s)^{n-m_i}s^{m_i}\lambda(s)\mathrm{d}s}.$$

□

Corollary 1. *Let $Z_1, Z_2, \ldots$ be exchangeable random variables modeled as in (7), where $\mathscr{M}$ is the law of the stable Beta process. If $\boldsymbol{Z}_n$ is a random sample that displays $K_n = k$ distinct features $\{W^*_1, \ldots, W^*_{K_n}\}$, and feature W^*_i appears exactly $M_{n,i} = m_i$ times in the samples, such as $i = 1, \ldots, K_n$, then*

$$\begin{aligned}&\mathbb{P}((Y_{n+1}, A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}) = (y, a_1, \ldots, a_{K_n})|\boldsymbol{Z}_n) \\ &\quad = \text{Poiss}\left(y; \alpha\frac{(c+\sigma)_n}{(c+1)_n}\right)\prod_{i=1}^{k}\text{Bern}\left(a_i; \frac{m_i-\sigma}{n+c}\right),\end{aligned} \tag{15}$$

where $(x)_y = \Gamma(x+y)/\Gamma(x)$ denotes the Pochhammer symbol for $x, y > 0$.

Proof. It is sufficient to specialize Theorem 1 for the stable Beta process. In particular, from Theorem 1, the posterior distribution of Y_{n+1} given $\boldsymbol{Z}_n$ is a Poisson distribution with mean

$$\int_0^1 s(1-s)^n\lambda(s)\mathrm{d}s \overset{(12)}{=} \frac{\alpha\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\int_0^1 s^{-\sigma}(1-s)^{n+c+\sigma}\mathrm{d}s = \alpha\frac{(c+\sigma)_n}{(c+1)_n}.$$

Moreover, the parameters of the Bernoulli random variables $A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}$ are equal to

$$p_i^* = \frac{\int_0^1 s^{m_i+1}(1-s)^{n-m_i}\lambda(s)\mathrm{d}s}{\int_0^1 s^{m_i}(1-s)^{n-m_i}\lambda(s)\mathrm{d}s} \overset{(12)}{=} \frac{B(m_i+1-\sigma, c+\sigma+n-m_i)}{B(m_i-\sigma, c+\sigma+n-m_i)} = \frac{m_i-\sigma}{n+c}$$

as $i = 1, \ldots, K_n$. □

3.2. SP Priors

From Theorem 1, under CRM priors, the distribution of the number of new features Y_{n+1} is a Poisson distribution that depends on the sampling information only through the sample size n. Moreover, the probability of observing a feature already observed in the sample, say W_i^*, depends only on the sample size n and the frequency m_i of feature W_i^* out of the initial sample. Camerlenghi et al. [15] showed that SP priors allow one to enrich the predictive structure of CRM priors, including additional sampling information in the probability of discovering new features. To introduce SP priors, consider a CRM $\tilde{\mu} = \sum_{i\geq 1} \tilde{\tau}_i \delta_{\tilde{w}_i}$ on $\mathbb{W}$, where $(\tilde{\tau}_i)_{i\geq 1}$ are positive random atoms and $(\tilde{w}_i)_{i\geq 1}$ are i.i.d. random atoms, with Lévy intensity $\nu(\mathrm{d}w, \mathrm{d}s) = \lambda(s)\mathrm{d}sP(\mathrm{d}w)$ satisfying

$$\int_0^\infty \min\{s, 1\}\lambda(s)\mathrm{d}s < +\infty. \tag{16}$$

Consider the ordered jumps $\Delta_1 > \Delta_2 > \cdots$ of the CRM $\tilde{\mu}$ and define the random measure

$$\tilde{\mu}_{\Delta_1} = \sum_{i\geq 1} \frac{\Delta_{i+1}}{\Delta_1}\delta_{\tilde{w}_i}$$

normalizing $\tilde{\mu}$ by the largest jump. The definition of SPs follows with a suitable change in the measure of Δ_1 (James et al. [14], Camerlenghi et al. [15]). Let us denote by $\mathscr{L}(\,\cdot\,, a)$ a regular version of the conditional probability distribution of $(\Delta_{i+1}/\Delta_1)_{i\geq 1}$ given $\Delta_1 = a$. Now denote by Ψ_1 a positive random variable with density function f_{Ψ_1} on $\mathbb{R}^+$ and define

$$\mathscr{L}(\,\cdot\,) := \int_{\mathbb{R}^+} \mathscr{L}(\,\cdot\,, a) f_{\Psi_1}(a)\mathrm{d}a$$

The distribution of $(\Delta_{i+1}/\Delta_1)_{i\geq 1}$ is obtained by mixing $\mathscr{L}(\,\cdot\,, a)$ with respect to the density function f_{Ψ_1}. Thus, we are ready to define an SP.

Definition 3. *A Scaled Process (SP) prior on $(\mathbb{W}, \mathscr{W})$ is defined as the almost surely discrete random measure*

$$\tilde{\mu}_{\Psi_1} := \sum_{i\geq 1} \tilde{\eta}_i \delta_{\tilde{w}_i}, \tag{17}$$

where $(\tilde{\eta}_i)_{i\geq 1}$ has distribution $\mathscr{L}$ and $(\tilde{w}_i)_{i\geq 1}$ is a sequence of independent random variables with common distribution P, also independent of $(\tilde{\eta}_i)_{i\geq 1}$. We will write $\tilde{\mu}_{\Psi_1} \sim \mathrm{SP}(\nu, f_{\Psi_1})$.

A thoughtful account with a complete posterior analysis for SPs is given in Camerlenghi et al. [15]. Here, we characterize the predictive distribution (8) of SPs.

Theorem 2 (Camerlenghi et al. [15], James [28])**.** *Let $Z_1, Z_2, \ldots$ be exchangeable random variables modeled as in (7), where $\mathscr{M}$ equals $\mathrm{SP}(\nu, f_{\Psi_1})$. If $\mathbf{Z}_n$ is a random sample that displays $K_n = k$ distinct features $\{W_1^*, \ldots, W_{K_n}^*\}$, and feature W_i^* appears exactly $M_{n,i} = m_i$ times in the samples, such as $i = 1, \ldots, K_n$, then the conditional distribution of Ψ_1, given $\mathbf{Z}_n$, has posterior density:*

$$f_{\Psi_1|\mathbf{Z}_n}(a) \propto e^{-\sum_{i=1}^n \int_0^1 s(1-s)^{n-1}a\lambda(as)\mathrm{d}s} \prod_{i=1}^k \int_0^1 s^{m_i}(1-s)^{n-m_i}a\lambda(as)\mathrm{d}s f_{\Psi_1}(a). \tag{18}$$

Moreover, conditionally on $\mathbf{Z}_n$ *and* Ψ_1,

$$\begin{aligned}\mathbb{P}((Y_{n+1}, A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}) = (y, a_1, \ldots, a_{K_n})|\mathbf{Z}_n, \Psi_1) \\ = \text{Poiss}\left(y; \int_0^1 s\Psi_1(1-s)^n\lambda(s\Psi_1)\mathrm{d}s\right)\prod_{i=1}^{k}\text{Bern}(a_i; p_i^*(\Psi_1))\end{aligned} \tag{19}$$

being

$$p_i^*(\Psi_1) := \frac{\int_0^1 s^{m_i+1}(1-s)^{n-m_i}\lambda(s\Psi_1)\mathrm{d}s}{\int_0^1 s^{m_i}(1-s)^{n-m_i}\lambda(s\Psi_1)\mathrm{d}s}.$$

Proof. The representation of the predictive distribution (19) follows from Camerlenghi et al. [15] (Proposition 2). Indeed, the posterior distribution of the largest jump directly follows from [15] (Equation (4)). In addition, the authors of [15] (Proposition 2) showed that the conditional distribution of Z_{n+1}, given $\mathbf{Z}_n$ and Ψ_1, equals the distribution of the following counting measure:

$$Z'_{n+1} + \sum_{i=1}^{K_n} A^*_{n+1,i}\delta_{W_i^*}, \tag{20}$$

where $Z'_{n+1}|\tilde{\mu}' = \sum_{i\geq 1} A'_{n+1,i}\delta_{\tilde{w}'_i} \sim \text{BeP}(\tilde{\mu}'_{\Psi_1})$ and $\tilde{\mu}'_{\Psi_1}$ is a CRM with Lévy intensity of the form

$$\nu'_{\Psi_1}(\mathrm{d}w, \mathrm{d}s) = (1-s)^n\Psi_1\lambda(\Psi_1 s)\mathbb{1}_{(0,1)}(s)\mathrm{d}sP(\mathrm{d}w).$$

Moreover, $A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}$ are Bernoulli random variables with parameters $J_1, \ldots, J_{K_n}$, respectively, such that conditionally on Ψ_1, each J_i has a distribution with a density function of the form

$$f_{J_i|\Psi_1}(s) \propto (1-s)^{n-m_i}s^{m_i}\Psi_1\lambda(\Psi_1 s) \quad \text{on } (0,1).$$

As in the proof of Theorem 1, we show that the distribution of $Y_{n+1}|(\Psi_1, \mathbf{Z}_n)$ equals $\sum_{i\geq 1} A'_{n+1,i}$. Thus, by the evaluation of the Laplace functional, one may easily realize that the last random sum has a Poisson distribution with mean $\int_0^1(1-s)^n s\Psi_1\lambda(\Psi_1 s)\mathrm{d}s$. Moreover, by exploiting the posterior representation (20), the variables $A^*_{n+1,i}$, such as $i = 1, \ldots, K_n$, conditionally on $\mathbf{Z}_n$ and Ψ, are independent and Bernoulli distributed with mean

$$\mathbb{E}[J_i|\Psi_1] = \int_0^1 s f_{J_i|\Psi_1}(s)\mathrm{d}s = \frac{\int_0^1(1-s)^{n-m_i}s^{m_i+1}\Psi_1\lambda(s\Psi_1)\mathrm{d}s}{\int_0^1(1-s)^{n-m_i}s^{m_i}\Psi_1\lambda(s\Psi_1)\mathrm{d}s}.$$

□

Remark 1. *According to* (18)*, the conditional distribution of* Ψ_1 *given* $\mathbf{Z}_n$ *may include the whole sampling information, depending on the specification of* ν *and* f_{Ψ_1}*, and hence, the conditional distribution of* Y_{n+1} *given* $\mathbf{Z}_n$ *may also include such sampling information. As a corollary of Theorem 2, the conditional distribution of* Y_{n+1} *given* $\mathbf{Z}_n$ *is a mixture of Poisson distributions that may include the whole sampling information; in particular, the amount of sampling information in the posterior distribution is uniquely determined by the mixing distribution, namely by the conditional distribution of* Ψ_1*, given* $\mathbf{Z}_n$*.*

Hereafter, we specialize Theorem 2 for the stable SP, that is, a peculiar SP defined through a CRM with a Lévy intensity ν such that $\lambda(s) = \sigma s^{-1-\sigma}$ for a parameter $\sigma \in (0,1)$. We refer to Camerlenghi et al. [15] for a detailed posterior analysis of the stable SP prior.

Corollary 2. *Let* $Z_1, Z_2, \ldots$ *be exchangeable random variables modeled as in* (7)*, where* $\mathscr{M}$ *equals* $\text{SP}(\nu, f_{\Psi_1})$*, with* $\lambda(s) = \sigma s^{-1-\sigma}$ *for some* $\sigma \in (0,1)$*. If* $\mathbf{Z}_n$ *is a random sample that displays* $K_n = k$ *distinct features* $\{W_1^*, \ldots, W_{K_n}^*\}$*, and feature* W_i^* *appears exactly* $M_{n,i} = m_i$ *times*

in the samples, such as $i = 1, \dots, K_n$*, then the conditional distribution of* Ψ_1*, given* $\boldsymbol{Z}_n$*, has posterior density:*

$$f_{\Psi_1|\boldsymbol{Z}_n}(a) \propto a^{-k\sigma} e^{-\sigma a^{-\sigma} \sum_{i=1}^{n} B(1-\sigma,i)} f_{\Psi_1}(a) \tag{21}$$

having denoted by $B(\,\cdot\,,\,\cdot\,)$ *the classical Euler Beta function. Moreover, conditionally on* $\boldsymbol{Z}_n$ *and* Ψ_1*,*

$$\begin{aligned} &\mathbb{P}((Y_{n+1}, A^*_{n+1,1}, \dots, A^*_{n+1,K_n}) = (y, a_1, \dots, a_{K_n}) | \boldsymbol{Z}_n, \Psi_1) \\ &\quad = \text{Poiss}\big(y; \sigma \Psi_1^{-\sigma} B(1-\sigma, n+1)\big) \prod_{i=1}^{k} \text{Bern}\left(a_i; \frac{m_i - \sigma}{n - \sigma + 1}\right). \end{aligned} \tag{22}$$

Proof. The proof is a plain application of Theorem 2 under the choice $\lambda(s) = \sigma s^{-1-\sigma}$. □

4. Predictive Characterizations for SPs

In this section, we introduce and discuss Johnson's "sufficientness" postulates in the context of feature-sampling models under the class of SP priors. According to Theorem 1, if the feature-sampling model is a CRM prior, then the conditional distribution of Y_{n+1}, given $\boldsymbol{Z}_n$, is a Poisson distribution that depends on the sampling information $\boldsymbol{Z}_n$ only through the sample size n. Moreover, the conditional probability of generating an old feature W_i^* given $\boldsymbol{Z}_n$ depends on the sampling information $\boldsymbol{Z}_n$ only through n and m_i. As shown in Theorem 2, SP priors enrich the predictive structure of CRM priors through the conditional distribution of the latent variable Ψ_1 given the observable sample $\boldsymbol{Z}_n$. In the next theorem, we characterize the class of SP priors for which the conditional distribution of Y_{n+1} given $\boldsymbol{Z}_n$ depends on the sampling information only through n.

Theorem 3. *Let* $Z_1, Z_2, \dots$ *be exchangeable random variables modeled as in* (7)*, where* $\mathscr{M}$ *equals* $\text{SP}(\nu, f_{\Psi_1})$ *and* $\nu(\mathrm{d}w, \mathrm{d}s) = \lambda(\mathrm{d}s)\mathrm{d}s P(\mathrm{d}w)$*. Moreover, suppose that* $\boldsymbol{Z}_n$ *is a random sample that displays* $K_n = k$ *distinct features* $\{W_1^*, \dots, W_{K_n}^*\}$*, and feature* W_i^* *appears exactly* $M_{n,i} = m_i$ *times in the samples, such as* $i = 1, \dots, K_n$*. If* $f_{\Psi_1} : (0, r) \to \mathbb{R}^+$ *is a continuous function on the compact support* $(0, r)$ *with* $r > 0$*, and the function* $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ *is continuous on its domain, then the conditional distribution of the latent variable* Ψ_1 *given* $\boldsymbol{Z}_n$ *depends on the sampling information* $\boldsymbol{Z}_n$ *only through n if and only if* $\lambda(s) = Cs^{-1}$ *on* $(0, r)$ *for some constant* $C > 0$*.*

Proof. First of all, if f_{Ψ_1} is defined on the compact support $(0, r)$ and if $\lambda(s) = Cs^{-1}$ on $(0, r)$ for some constant $C > 0$, then it is easy to see that the posterior distribution of Ψ_1 in (18) depends only on n and not on the other sample statistics. We now show the reverse implication. The posterior density of Ψ_1, conditionally on $\boldsymbol{Z}_n$, satisfies (18), and it is proportional to

$$f_{\Psi_1|\boldsymbol{Z}_n}(a) \propto \prod_{i=1}^{n} e^{-\phi_i(a)} \prod_{i=1}^{K_n} \int_0^1 s^{m_i} (1-s)^{n-m_i} a\lambda(as)\mathrm{d}s \, f_{\Psi_1}(a),$$

where $\phi_i(a) = \int_0^1 s(1-s)^{i-1} a\lambda(as)\mathrm{d}s$. Then, there exists $c(m_1, \dots, m_k, k, n)$ such that it holds that

$$f_{\Psi_1|\boldsymbol{Z}_n}(a) = \frac{\prod_{i=1}^{n} e^{-\phi_i(a)} \prod_{i=1}^{K_n} \int_0^1 s^{m_i} (1-s)^{n-m_i} a\lambda(as)\mathrm{d}s \, f_{\Psi_1}(a)}{c(m_1, \dots, m_k, k, n)}. \tag{23}$$

Because of the assumptions imposed, the distribution of $\Psi_1|\boldsymbol{Z}_n$ does not depend on K_n, nor on the corresponding sample frequencies $M_{n,1}, \dots, M_{n,K_n}$. Accordingly, the function

$$f_1(a, n) := f^{-1}_{\Psi_1|\boldsymbol{Z}_n}(a) \prod_{i=1}^{n} e^{-\phi_i(a)}, \quad a \in (0, r), \tag{24}$$

depends only on a and n, but not on k and $(m_1, \ldots, m_k)$. Then, putting together (23) and (24), it holds that

$$f_1(a,n) \cdot \prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{n-m_i} a\lambda(as)\mathrm{d}s = c(m_1, \ldots, m_k, n, k) \quad \forall a \in (0,r), \tag{25}$$

where c is the normalizing factor, and it does not depend on the variable a. By choosing $m_1 = \ldots = m_k = n \in \mathbb{N}$, thanks to Equation (25), we can state that the following function:

$$f_1(a,n)\left(\int_0^1 s^n a\lambda(as)\mathrm{d}s\right)^k, \tag{26}$$

which is defined for any $a \in (0,r)$ and does not depend on a, but only on k and n. Since the previous assertion is true for any $k \geq 1$, one may select $k = 1$, thus obtaining the following identity:

$$f_1(a,n) = c^*\left(\int_0^1 s^n a\lambda(as)\mathrm{d}s\right)^{-1} \tag{27}$$

for some constant c^*, independent of a, but that may depend on n. Substituting (27) into (26), we obtain that

$$c^*\left(\int_0^1 s^n a\lambda(as)\mathrm{d}s\right)^{k-1} \tag{28}$$

is a function that does not depend on a, but only on n and k. As a consequence, we have that

$$\int_0^1 s^n a\lambda(as)\mathrm{d}s = \int_0^a \frac{s^n}{a^n}\lambda(s)\mathrm{d}s = C^{**}$$

for a suitable constant C^{**}, which does not depend on $a \in (0,r)$. To conclude, we take a derivative of the previous expression with respect to a, and this allows us to show that

$$a^n\lambda(a) = na^{n-1}C^{**},$$

namely, $\lambda(a) = C/a$ for $a \in (0,r)$, where C is a positive constant. This is a Lévy intensity; indeed, it satisfies the condition (11). Outside the interval $(0,r)$, λ may be defined arbitrarily; indeed, the values of λ on $[r+\infty)$ do not affect the posterior distribution of Ψ_1 (18). □

Remark 2. *Note that in Theorem 3, we have supposed that f_{Ψ_1} has a compact support on $(0,r)$; thus, we are interested in defining λ on $(0,r)$; outside the interval, λ can be defined arbitrarily because it does not affect the posterior distribution (18) of Ψ_1. From the proof of Theorem 3, it becomes apparent that if the support of f_{Ψ_1} is the entire positive real line $\mathbb{R}^+$, the posterior distribution of the largest jump depends only on n if and only if $\lambda(s) = Cs^{-1}$ on $\mathbb{R}^+$ for some constant $C > 0$. However, in this case, λ does not meet the integrability condition (11); hence, this can only considered a limiting case. It is interesting to observe that such a limiting situation, with the additional assumption $f_{\Psi_1} = f_{\Delta_1}$, corresponds to the Beta process case with $\sigma = 0$ and $c = 1$ (Griffiths and Ghahramani [12]).*

Now, we characterize SPs for which the posterior distribution of Ψ_1 depends only on n and K_n, but not on the sample frequencies of the different features $\boldsymbol{m}$. Here, we assume that f_{Ψ_1} has full support a priori. The following characterization has been provided in Camerlenghi et al. [15] (Theorem 3), but for completeness, we report the proof.

Theorem 4 (Camerlenghi et al. [15])**.** *Let $Z_1, Z_2, \ldots$ be exchangeable random variables modeled as in (7), where $\mathscr{M}$ equals $\mathrm{SP}(\nu, f_{\Psi_1})$ and $\nu(\mathrm{d}w, \mathrm{d}s) = \lambda(\mathrm{d}s)\mathrm{d}sP(\mathrm{d}w)$. Suppose that $\boldsymbol{Z}_n$ is a random sample that displays $K_n = k$ distinct features $\{W_1^*, \ldots, W_{K_n}^*\}$, and feature W_i^* appears exactly $M_{n,i} = m_i$ times in the sample, such as $i = 1, \ldots, K_n$. If $f_{\Psi_1} : \mathbb{R}^+ \to \mathbb{R}^+$ is a strictly*

positive function on $\mathbb{R}^+$ *and continuously differentiable, and* λ *is continuously differentiable, then the conditional distribution of the latent variable* Ψ_1*, given* $\mathbf{Z}_n$*, depends on* $\mathbf{Z}_n$ *only through* n *and* K_n *if and only if* $\lambda(s) = Cs^{-1-\sigma}$ *on* $\mathbb{R}^+$ *for some constant* $C > 0$ *and* $\sigma \in (0, 1)$*.*

Proof. By arguing as in the proof of Theorem 3, the posterior density of Ψ_1 given $\mathbf{Z}_n$ is proportional to

$$\prod_{i=1}^{n} e^{-\phi_i(a)} \prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{n-m_i} a\lambda(as)\mathrm{d}s \, f_{\Psi_1}(a),$$

where $\phi_i(a) = \int_0^1 s(1-s)^{i-1} a\lambda(as)\mathrm{d}s$. Then, there exists $c(m_1, \ldots, m_k, n, k)$ such that it holds that

$$f_{\Psi_1|\mathbf{Z}_n}(a) = \frac{\prod_{i=1}^{n} e^{-\phi_i(a)} \prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{n-m_i} a\lambda(as)\mathrm{d}s \, f_{\Psi_1}(a)}{c(m_1, \ldots, m_k, n, k)}.$$

As a consequence,

$$f^{-1}_{\Psi_1|\mathbf{Z}_n}(a) \prod_{i=1}^{n} e^{-\phi_i(a)} \prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{n-m_i} a\lambda(as)\mathrm{d}s \, f_{\Psi_1}(a) = c(m_1, \ldots, m_k, n, k). \tag{29}$$

If the density function $f_{\Psi_1|\mathbf{Z}_n}(a)$ does not depend on $m_1, \ldots, m_k$, then the following function

$$f^{-1}_{\Psi_1|\mathbf{Z}_n}(a) \prod_{i=1}^{n} e^{-\phi_i(a)} \, f_{\Psi_1}(a) = f_1(a, k, n)$$

depends only on k, n and a, but not on the frequency counts. Therefore, (29) boils down to

$$f_1(a, k, n) \cdot \prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{n-m_i} a\lambda(as)\mathrm{d}s = c(m_1, \ldots, m_k, n, k). \tag{30}$$

where the function on the right-hand side of (30) is independent of a for any choice of the vector of sampling information $(m_1, \ldots, m_k, n, k)$. Now, since the vector $(m_1, \ldots, m_k, n, k)$ can be chosen arbitrarily, we can make the choice $m_1 = \cdots = m_k = m > 0$, such that the function

$$\left[w(a, k, n) \int_0^1 s^{m}(1-s)^{n-m} a\lambda(as)\mathrm{d}s \right]^k \tag{31}$$

does not depend on $a \in \mathbb{R}^+$, where $w(a, k, n) = \sqrt[k]{f_1(a, k, n)}$. Moreover, suppose that $m = n$; thus,

$$w(a, k, n) \int_0^1 s^{n} a\lambda(as)\mathrm{d}s \tag{32}$$

does not depend on $a \in \mathbb{R}^+$, which implies that

$$w(a, k, n) = c^* \left(\int_0^1 s^{n} a\lambda(as)\mathrm{d}s \right)^{-1} \tag{33}$$

for a constant $c^* > 0$ with respect to a, which can only depend on k and n. By substituting (33) into (31), we obtain

$$\left[\frac{c^*}{\int_0^1 s^{n}\lambda(as)\mathrm{d}s} \cdot \int_0^1 s^{m}(1-s)^{n-m} \lambda(as)\mathrm{d}s \right]^k,$$

which is independent of $a \in \mathbb{R}^+$. Now, it is possible to choose $m = n - 1$ in the previous function. Therefore, there exists a constant c^{**} independent of a such that the following identity holds:

$$\int_0^1 s^{n-1}\lambda(as)\mathrm{d}s - \int_0^1 s^n\lambda(as)\mathrm{d}s = c^{**}\int_0^1 s^n\lambda(as)\mathrm{d}s.$$

By taking the derivative of the previous equation two times with respect to a, one obtains

$$\lambda(a)(1 - nc^{**}) = a\lambda'(a)c^{**},$$

which is an ordinary differential equation in λ that can be solved by separation of variables. In particular, we obtain

$$\lambda(a) = Ca^{(1-nc^{**})/c^{**}}, \quad \text{for } C > 0. \tag{34}$$

To conclude, observe that the exponent of a in (34) should satisfy the integrability condition (11) for homogeneous CRMs. Accordingly, it is easy to see that we must consider

$$\lambda(a) = C\frac{1}{a^{1+\sigma}}$$

where $C > 0$ and $\sigma \in (0, 1)$. The reverse implication of the theorem is trivially satisfied; hence, the proof is completed. □

We recall from Theorem 2 that the conditional distribution of Ψ_1 given $\boldsymbol{Z}_n$ uniquely determines the amount of sampling information included in the conditional distribution of the number of new features Y_{n+1} given $\boldsymbol{Z}_n$. Such sampling information may range from the whole information, in terms of n, K_n, and $(M_{1,n}, \ldots, M_{K_n,n})$, to the sole information on the sample size n. According to Theorem 4, the stable SP prior of Corollary 2 is the sole SP prior for which the conditional distribution of the number of new features Y_{n+1} given $\boldsymbol{Z}_n$ depends on the sampling information $\boldsymbol{Z}_n$ only on n and K_n. Moreover, according to Theorem 3, the Beta process prior is the sole SP prior for which the conditional distribution of the number of new features Y_{n+1} given $\boldsymbol{Z}_n$ depends on the sampling information $\boldsymbol{Z}_n$ only on n. In particular, Theorems 3 and 4 show that the Beta process prior and the stable SP prior may be considered, to some extent, the feature sampling counterparts of the Dirichlet process prior the Pitman–Yor process prior.

5. Discussion and Conclusions

In this paper, we have introduced and discussed Johnson's "sufficientness" postulates in the context of feature-sampling models. "Sufficientness" postulates have been investigated extensively in the context of species-sampling models, providing an effective classification of species-sampling models on the basis of the form of their corresponding predictive distributions. Here, we made a first step towards the problem of providing an analogous classification for feature-sampling models. In particular, we obtained Johnson's "sufficientness" postulates when the class of feature-sampling models is restricted to the class of scaled process priors. However, the results presented in the paper remain preliminary, and do not at all provide a complete answer to the characterization problem within the general class of feature-sampling models. This problem remains open.

Within the feature-sampling setting, the predictive distribution is of the form (8), though for the purpose of providing "sufficientness" postulates, one may focus on feature-sampling models exhibiting a general predictive distribution of the following type:

$$\begin{aligned}\mathbb{P}((Y_{n+1}, A^*_{n+1,1}, \ldots, A^*_{n+1,K_n}) = (y, a_1, \ldots, a_{K_n})|\boldsymbol{Z}_n) \\ = g(y; n, k, \boldsymbol{m})\prod_{i=1}^{k} f_i(a_i; n, k, \boldsymbol{m}).\end{aligned} \tag{35}$$

Note that (35) is a probability distribution, and it must satisfy a consistency condition, as usual. Among all the feature-sampling models whose predictive distribution can be written in the form (35), we are interested in characterizing nonparametric priors such that: (i) The function g depends on the sampling information only through n, and the function f_i depends only on (n, m_i); (ii) g depends only on (n, k) and f_i depends only on (n, m_i); (iii) g depends only on (n, k) and f_i depends only on (n, k, m_i). In our view, these characterizations may provide a complete picture of sufficientness postulates within the feature setting, and they are also fundamental to guiding the selection of the prior distribution. We conjecture that CRMs are the nonparametric priors satisfying the characterization (i), the SP with a stable Lévy measure is an example of prior satisfying (ii), and no examples satisfying (iii) have been considered in the current literature. Results in this direction are in Battiston et al. [31], where the authors characterize an exchangeable feature allocation probability function (Broderick et al. [32]) in product forms; this could be a stimulating point of departure to study the characterization problem depicted above.

Author Contributions: Writing–original draft, F.C. and S.F.; writing–review and editing, F.C. and S.F. The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No. 817257.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: F.C. is extremely grateful to Eugenio Regazzini for the time spent at the Department of Mathematics of University of Pavia during his Ph.D. studies in Mathematical Statistics; F.C. wants to especially thank Eugenio Regazzini for having introduced him to the study of Bayesian Statistics with a stimulating Ph.D. course held together with Antonio Lijoi. S.F. wishes to express his gratitude to Eugenio Regazzini, whose fundamental contributions to Bayesian statistics have always been a great source of inspiration, transmitting enthusiasm and methods for the development of his own research. The authors gratefully acknowledge the financial support from the Italian Ministry of Education, University, and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. F.C. is a member of the *Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni* (GNAMPA) of the *Istituto Nazionale di Alta Matematica* (INdAM).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. de Finetti, B. La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **1937**, *7*, 1–68.
2. Johnson, W.E. Probability: The Deductive and Inductive Problems. *Mind* **1932**, *41*, 409–423. [CrossRef]
3. Pitman, J. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*; IMS Lecture Notes Monograph Series; Institute of Mathematical Statistics: Hayward, CA, USA, 1996; Volume 30, pp. 245–267. [CrossRef]
4. Regazzini, E. Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giornale dell'Istituto Italiano degli Attuari* **1978**, *41*, 77–89.
5. Lo, A.Y. A characterization of the Dirichlet process. *Stat. Probab. Lett.* **1991**, *12*, 185–187. [CrossRef]
6. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1973**, *1*, 209–230. [CrossRef]
7. Zabell, S.L. Symmetry and its discontents. In *Cambridge Studies in Probability, Induction, and Decision Theory*; Essays on the history of inductive probability, with a preface by Brian Skyrms; Cambridge University Press: New York, NY, USA, 2005; p. xii+279.
8. Perman, M.; Pitman, J.; Yor, M. Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Relat. Fields* **1992**, *92*, 21–39. [CrossRef]
9. Pitman, J.; Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **1997**, *25*, 855–900. [CrossRef]
10. Bacallado, S.; Battiston, M.; Favaro, S.; Trippa, L. Sufficientness postulates for Gibbs-type priors and hierarchical generalizations. *Stat. Sci.* **2017**, *32*, 487–500. [CrossRef]
11. Gnedin, A.; Pitman, J. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **2005**, *325*, 83–102. 244–245. [CrossRef]

12. Griffiths, T.L.; Ghahramani, Z. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **2011**, *12*, 1185–1224.
13. Ayed, F.; Battiston, M.; Camerlenghi, F.; Favaro, S. Consistent estimation of small masses in feature sampling. *J. Mach. Learn. Res.* **2021**, *22*, 1–28.
14. James, L.F.; Orbanz, P.; Teh, Y.W. Scaled subordinators and generalizations of the Indian buffet process. *arXiv* **2015**, arXiv:1510.07309.
15. Camerlenghi, F.; Favaro, S.; Masoero, L.; Broderick, T. Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *arXiv* **2021**, arXiv:2106.15480.
16. Brix, A. Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **1999**, *31*, 929–953. [CrossRef]
17. Lijoi, A.; Mena, R.H.; Prünster, I. Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 715–740. [CrossRef]
18. De Blasi, P.; Favaro, S.; Lijoi, A.; Mena, R.H.; Prunster, I.; Ruggiero, M. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 212–229. [CrossRef]
19. Regazzini, E.; Lijoi, A.; Prünster, I. Distributional results for means of normalized random measures with independent increments. *Ann. Stat.* **2003**, *31*, 560–585. [CrossRef]
20. Pitman, J. *Poisson-Kingman Partitions*; Lecture Notes-Monograph Series; Institute of Mathematical Statistics: Beachwood, OH, USA, 2003; pp. 1–34.
21. Pitman, J. *Combinatorial Stochastic Processes*; Lecture Notes in Mathematics; Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, 7–24 July 2002, with a foreword by Jean Picard; Springer: Berlin, Germany, 2006; Volume 1875, p. x+256.
22. Lijoi, A.; Prünster, I. Models beyond the Dirichlet process. In *Bayesian Nonparametrics*; Hjort, N.L., Holmes, C., Müller, P., Walker, S., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 80–136.
23. Ghosal, S.; van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2017; Volume 44, p. xxiv+646.
24. Zabell, S.L. The continuum of inductive methods revisited. In *The Cosmos of Science: Essays of Exploration*; University of Pittsburgh Press: Pittsburgh, PA, USA, 1997; pp. 351–385.
25. Daley, D.J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure (Probability and Its Applications)*, 2nd ed.; Springer: New York, NY, USA, 2008; p. xviii+573,
26. Kingman, J. Completely random measures. *Pac. J. Math.* **1967**, *21*, 59–78. [CrossRef]
27. Teh, Y.; Gorur, D. Indian buffet processes with power-law behavior. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1838–1846.
28. James, L.F. Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *Ann. Stat.* **2017**, *45*, 2016–2045. [CrossRef]
29. Broderick, T.; Wilson, A.C.; Jordan, M.I. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli* **2018**, *24*, 3181–3221. [CrossRef]
30. Masoero, L.; Camerlenghi, F.; Favaro, S.; Broderick, T. More for less: Predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika* **2021**, asab012, [CrossRef]
31. Battiston, M.; Favaro, S.; Roy, D.M.; Teh, Y.W. A characterization of product-form exchangeable feature probability functions. *Ann. Appl. Probab.* **2018**, *28*, 1423–1448. [CrossRef]
32. Broderick, T.; Pitman, J.; Jordan, M.I. Feature allocations, probability functions, and paintboxes. *Bayesian Anal.* **2013**, *8*, 801–836. [CrossRef]