Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Scaled Process Priors for Bayesian Nonparametric Estimation of the Unseen Genetic Variation

Federico Camerlenghi, Stefano Favaro, Lorenzo Masoero & Tamara Broderick

To cite this article: Federico Camerlenghi, Stefano Favaro, Lorenzo Masoero & Tamara Broderick (2022): Scaled Process Priors for Bayesian Nonparametric Estimation of the Unseen Genetic Variation, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2115918

To link to this article: https://doi.org/10.1080/01621459.2022.2115918

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2022.2115918

Scaled Process Priors for Bayesian Nonparametric Estimation of the Unseen Genetic Variation

Federico Camerlenghi[a,b,c], Stefano Favaro[b,d], Lorenzo Masoero[e], and Tamara Broderick[e]

[a]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy; [b]Collegio Carlo Alberto, Torino, Italy; [c]BIDSA, Bocconi University, Milano, Italy; [d]Department of Economics and Statistics, University of Torino, Torino, Italy; [e]Department of Electrical Engineering and Computer Science, CSAIL, Massachusetts Institute of Technology, Cambridge, MA

ABSTRACT

There is a growing interest in the estimation of the number of unseen features, mostly driven by biological applications. A recent work brought out a peculiar property of the popular completely random measures (CRMs) as prior models in Bayesian nonparametric (BNP) inference for the unseen-features problem: for fixed prior's parameters, they all lead to a Poisson posterior distribution for the number of unseen features, which depends on the sampling information only through the sample size. CRMs are thus not a flexible prior model for the unseen-features problem and, while the Poisson posterior distribution may be appealing for analytical tractability and ease of interpretability, its independence from the sampling information makes the BNP approach a questionable oversimplification, with posterior inferences being completely determined by the estimation of unknown prior's parameters. In this article, we introduce the stable-Beta scaled process (SB-SP) prior, and we show that it allows to enrich the posterior distribution of the number of unseen features arising under CRM priors, while maintaining its analytical tractability and interpretability. That is, the SB-SP prior leads to a negative Binomial posterior distribution, which depends on the sampling information through the sample size and the number of distinct features, with corresponding estimates being simple, linear in the sampling information and computationally efficient. We apply our BNP approach to synthetic data and to real cancer genomic data, showing that: (i) it outperforms the most popular parametric and nonparametric competitors in terms of estimation accuracy; (ii) it provides improved coverage for the estimation with respect to a BNP approach under CRM priors. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2021
Accepted August 2022

KEYWORDS
Bayesian nonparametrics; Beta process prior; Completely random measure; Genetic variation; Predictive distribution; Scaled process prior; Stable process; Unseen-features problem

1. Introduction

The problem of estimating the number of unseen features generalizes the popular unseen-species problem (Orlitsky, Suresh, and Wu 2016), and its importance has grown dramatically in recent years, driven by applications in biological sciences (Ionita-Laza, Lange, and Laird 2009; Gravel 2014; Zou et al. 2016; Chakraborty et al. 2019). Consider a generic population in which each individual is endowed with a finite collection of $\mathbb{W}$-valued features, with $\mathbb{W}$ possibly being an infinite space, and denote by p_i the probability that an individual has feature $w_i \in \mathbb{W}$ for $i \geq 1$. The unseen-features problem assumes $N \geq 1$ observable random samples $Z_{1:N} = (Z_1, \ldots, Z_N)$ from the population, such that $Z_n = (A_{n,i})_{i\geq 1}$ are independent Bernoulli random variables with unknown parameters $(p_i)_{i\geq 1}$. Then, the goal is to estimate the number of hitherto unseen features that would be observed if $M \geq 1$ additional samples were collected, that is,

$$U = \sum_{i\geq 1} \mathbb{1}\left(\sum_{n=1}^{N} A_{n,i} = 0\right) \mathbb{1}\left(\sum_{m=1}^{M} A_{N+m,i} > 0\right),$$

with $\mathbb{1}$ being the indicator function. The unseen-species problem arises under the assumption that each individual is endowed with only one feature, that is, a species. A wide range of approaches have been developed to estimate U, including Bayesian methods (Ionita-Laza, Lange, and Laird 2009; Masoero et al. 2021), jackknife (Gravel 2014), linear programming (Zou et al. 2016), and variations of Good-Toulmin estimators (Orlitsky, Suresh, and Wu 2016; Chakraborty et al. 2019).

In biological sciences, we may think of individuals as organisms and of features as groups to which organisms belong to, with each group being defined by any difference in the genome relative to a reference genome, that is, a (genetic) variant. In human biology, the estimation of U arises in the context of optimal allocation of resources between quantity and quality in genetic experiments: spending resources to sequence a greater number of genomes (quantity), which reveals more about variation across the population, or spending resources to sequence genomes with increased accuracy (quality), which reveals more about individual organisms' genomes. Accurate estimates of U are critical in the experimental pipeline toward the goal of maximizing the usefulness of experiments under the tradeoff

CONTACT Lorenzo Masoero lo.masoero@gmail.com Department of Electrical Engineering and Computer Science, CSAIL, Massachusetts Institute of Technology, Cambridge, MA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

between quantity and quality (Ionita-Laza and Laird 2010; Zou et al. 2016). While in human-biology the cost of sequencing has decreased in recent years (Schwarze et al. 2020), the expense remains nontrivial, and it is still critical in fields where scientists work with relatively budgets, for example, nonhuman and nonmodel organisms (Souza et al. 2017). Other applications arise in precision medicine (Momozawa and Mizukami 2020), microbiome analysis (Sanders et al. 2019), single-cell sequencing (Zhang, Ntranos, and Tse 2020) and wildlife monitoring (Johansson et al. 2020).

1.1. Our Contributions

We introduce a Bayesian nonparametric (BNP) approach to the unseen-features problem, which relies on a novel prior distribution for the unknown $(p_i)_{i\geq 1}$. Completely random measures (CRMs) (Kingman 1992) provide a broad class of nonparametric priors for feature sampling problems, the most popular being the stable-Beta process prior (James 2017; Broderick, Wilson, and Jordan 2018). In a recent work, Masoero et al. (2021) brought out a peculiar feature of CRM priors in the unseen-features problem: they all lead to a Poisson posterior distribution of U, given $Z_{1:N}$ and fixed prior's parameters, which depends on $Z_{1:N}$ only through the sample size N. Despite the broadness of the class of CRM priors, such a common Poisson posterior structure makes CRMs not a flexible prior model for the unseen-features problem. While the Poisson posterior distribution may be appealing in principle, making posterior inferences analytically tractable and easy to interpret, its independence from $Z_{1:N}$ makes the BNP approach a questionable oversimplification, with posterior inferences being completely determined by the estimation of the unknown prior's parameters. A somehow similar scenario occurs in BNP inference for the unseen-species problem under a Dirichlet process (DP) prior (Ferguson 1973), and led to the use of the Pitman-Yor process (PYP) prior (Pitman and Yor 1997) for enriching the posterior distribution of the number of unseen species, while maintaining analytical tractability and interpretability of the DP prior (Lijoi, Mena, and Prünster 2007).

We show that scaled process (SP) priors, first introduced in James, Orbanz, and Teh (2015), allow to enrich the posterior distribution of U arising under CRM priors. Under SP priors, we characterize the posterior distribution of U as a mixture of Poisson distributions that may include, through the mixing distribution, the whole sampling information in terms of the number of distinct features and their frequencies. While this is appealing in principle, it may be at stake with analytical tractability and interpretability, which are critical for a concrete use of SP priors. Then, we introduce the stable-Beta SP (SB-SP) prior, which provides a sensible tradeoff between the amount of sampling information introduced in the posterior distribution of U, and analytical tractability and interpretability of the posterior inferences. In particular, we characterize the SB-SP prior as the sole SP prior for which the posterior distribution of U, given $Z_{1:N}$ and fixed prior's parameters, depends on $Z_{1:N}$ through the sample size N and the number K_N of distinct features; the SB-SP may thus be considered as the natural counterpart of the PYP for the unseen-feature problem. Under the SB-SP prior, the posterior distribution of U, as well as of a refinement of U that deals with the number of unseen rare features, is a negative Binomial posterior distributions, whose parameters depend on N, K_N and the prior's parameters. Corresponding Bayesian estimates of U, with respect to a squared loss function, are simple, linear in K_N and computationally efficient.

We present an empirical validation of the effectiveness of our BNP methodology, both on synthetic and real data. As for real data, we consider cancer genomic data, where the goal is to estimate the number of new (genomic) variants to be discovered in future unobservable samples. In cancer genomics, accurate estimates of the number of new variants is of particular importance, as it might help practitioners understand the site of origin of cancers, as well as the clonal origin of metastasis, and in turn be a useful tool to develop effective clinical strategies (Chakraborty et al. 2019; Huyghe et al. 2019). We make use of data from the cancer genome atlas (TCGA), and focus on the challenging scenario in which the sample size N is particularly small, and also small with respect to the extrapolation size M. Such a scenario is of interest in genomic applications, where only few samples of rare cancer might be available. We show that our BNP methodology outperforms the most popular parametric and nonparametric competitors, both classical (frequentist) and Bayesian, in terms of estimation accuracy of U and a refinement of U for rare features. In addition, with respect to the BNP approach under the stable-Beta process prior (Masoero et al. 2021), our approach provides improved coverage for the estimation. This is an empirical evidence of the effectiveness of replacing the Poisson posterior distribution with the negative Binomial posterior distribution, which allows to better exploit the sampling information.

1.2. Organization of the Paper

In Section 2 we show how SP priors allow to enrich the posterior distribution of U arising under CRM priors. In Section 3 we introduce and investigate the SB-SP prior in the context of the unseen-features problem: (i) we characterize the SB-SP prior in the class of SP priors, providing its predictive distribution; (ii) we apply the SB-SP prior to the unseen-features problem, providing the posterior distribution of U and a BNP estimator. Section 4 contains illustrations of our method. In Section 5 we discuss our approach, a multivariate extension of it, and future research directions. Proofs and additional experiments are in the Appendix, supplementary materials.

2. Scaled Process Priors for Feature Sampling Problems

For a measurable space of features $\mathbb{W}$, we assume $N \geq 1$ observable individuals to be modeled as a random sample $Z_{1:N}$ from the $\{0,1\}$-valued stochastic process $Z(w) = \sum_{i\geq 1} A_i \delta_{w_i}(w)$, $w \in \mathbb{W}$, where $(w_i)_{i\geq 1}$ are features in $\mathbb{W}$ and $(A_i)_{i\geq 1}$ are independent Bernoulli random variables with unknown parameters $(p_i)_{i\geq 1}$, p_i being the probability that an individual has feature w_i, for $i \geq 1$. That is, Z is a Bernoulli process with parameter $\zeta = \sum_{i\geq 1} p_i \delta_{w_i}$, denoted as $\text{BeP}(\zeta)$. BNP inference for feature sampling problems relies on the specification of a prior distribution

on the discrete measure ζ, leading to the BNP-Bernoulli model,

$$\begin{aligned} Z_n \mid \zeta &\overset{\text{iid}}{\sim} \text{BeP}(\zeta) \qquad n = 1, \ldots, N, \\ \zeta &\sim \mathscr{Z}, \end{aligned} \tag{1}$$

namely ζ is a discrete random measure on $\mathbb{W}$ whose law $\mathscr{Z}$ takes on the interpretation of a prior distribution for the unknown feature's composition of the population. By de Finetti's theorem, the random variables Z_n's in (1) are exchangeable with directing measure $\mathscr{Z}$ (Aldous 1983). In this section, we show how SP priors for ζ (James, Orbanz, and Teh 2015) allow to enrich the posterior distribution of the number of unseen features arising under CRM priors.

2.1. CRM Priors for Bernoulli Processes

CRMs provide a standard tool to define nonparametric prior distributions on the parameter ζ of the Bernoulli process Z. Consider a homogeneous CRM μ_0 on $\mathbb{W}$, that is, $\mu_0 = \sum_{i\geq 1} \rho_i \delta_{W_i}$, where the ρ_i's are $(0,1)$-valued random atoms such that $\sum_{i\geq 1} \rho_i < +\infty$, while the W_i's are iid $\mathbb{W}$-valued random locations independent of the ρ_i's. The law of μ_0 is characterized, through Lévy-Khintchine formula, by the Lévy intensity measure $\nu_0(\mathrm{d}s, \mathrm{d}w) = \lambda_0(s)\mathrm{d}sP(\mathrm{d}w)$ on $(0,1)\times\mathbb{W}$, where: (i) λ_0 is a measure on $(0,1)$, which controls the distribution of the ρ_i's, and such that $\int_{(0,1)} \min\{s,1\}\lambda_0(s)\mathrm{d}s < +\infty$; ii) P is a nonatomic measure on $\mathbb{W}$, which controls the distribution of the W_i's. For short, $\mu_0 \sim \text{CRM}(\nu_0)$. See Appendix S1 for an account on CRMs (Kingman 1992, chap. 8). Note that, since P is nonatomic, the random atoms W_i's are almost surely distinct, that is to say the different features cannot coincide almost surely. The law of μ_0 provides a natural prior distribution for the parameter ζ of the Bernoulli process. The Beta and the stable-Beta processes are popular examples of $\mu_0 \sim \text{CRM}(\nu_0)$, for suitable specifications of ν_0. A comprehensive posterior analysis of CRM priors is presented in James (2017). In the next proposition, we recall the predictive distribution of CRM priors (James 2017, Proposition 3.2).

Proposition 1. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim \text{CRM}(\nu_0)$. If $Z_{1:N}$ displays $K_N = k$ distinct features $\{W_1^*, \ldots, W_{K_N}^*\}$, each feature W_i^* appearing exactly $M_{N,i} = m_i$ times, then the conditional distribution of Z_{N+1}, given $Z_{1:N}$, coincides with the distribution of

$$Z_{N+1} \mid Z_{1:N} \overset{d}{=} Z'_{N+1} + \sum_{i=1}^{K_N} A_{N+1,i}\delta_{W_i^*}, \tag{2}$$

where: (i) $Z'_{N+1} \mid \mu'_0 = \sum_{i\geq 1} A'_{N+1,i}\delta_{W'_i} \sim \text{BeP}(\mu'_0)$ and $\mu'_0 \sim \text{CRM}(\nu'_0)$, with $\nu'_0(\mathrm{d}s, \mathrm{d}w) = (1-s)^N\lambda_0(s)\mathrm{d}sP(\mathrm{d}w)$; (ii) the $A_{N+1,i}$'s are independent Bernoulli random variables with parameters J_i's, such that J_i is distributed according to the density function $f_{J_i}(s) \propto (1-s)^{N-m_i}s^{m_i}\lambda_0(s)$ for $i \geq 1$.

According to (2), Z_{N+1} displays "new" features W'_i's, that is, features not appearing in the initial sample $Z_{1:N}$, and "old" features W_i^*'s, that is, features appeared in the initial sample $Z_{1:N}$. The posterior distribution of statistics of "new" features is determined by the law of Z'_{N+1}, which depends on $Z_{1:N}$ only through the sample size N; the posterior distribution of statistics of "old" features is determined by the law of $\sum_{1\leq i\leq K_N} A_{N+1,i}\delta_{W_i^*}$, which depends on $Z_{1:N}$ through the sample size N, the number K_N of distinct features and their frequencies $(M_{N,1}, \ldots, M_{N,K_N})$. As a corollary of Proposition 1, the posterior distribution of the number of "new" features in $(Z_{N+1}, \ldots, Z_{N+M})$, given $Z_{1:N}$ and fixed prior's parameters, is a Poisson distribution that depends on $Z_{1:N}$ only through N (Masoero et al. 2021). Such a posterior structure is peculiar to CRM priors, being inherited by the Poisson process formulation of CRMs (Kingman 1992). That is, despite the broadness of the class of CRM priors, all CRM priors lead to the same Poisson posterior structure for the number of unseen features, which thus makes them not a flexible prior model for the unseen-features problem. While the Poisson posterior distribution may be appealing in principle, making the posterior inferences analytically tractable and of easy interpretability, its independence from $Z_{1:N}$ makes the BNP approach under CRM priors a questionable oversimplification, with posterior inferences being completely determined by the estimation of unknown prior's parameters.

Remark 1. For the sake of mathematical convenience, and in agreement with the work of James (2017), in the sequel we maintain the random measure formulation for both the prior model μ_0 and the Bernoulli processes Z_n. However, we point out that each Z_n is equivalently characterized by means of the Bernoulli variables $(A_{n,i})_{i\geq 1}$ and the random features $(W_i)_{i\geq 1}$. In other terms, there exits a one-to-one correspondence between Z_n and the sequence of points $\{(A_{n,i}, W_i)\}_{i\geq 1}$. Finally, note that, although the values of features' labels W_i are immaterial, the features W_i's are assumed to be random. This is in line with the BNP literature on species sampling models, where the species' labels are assumed to be random (Pitman 1996).

2.2. SP Priors for Bernoulli Processes

Consider a homogeneous CRM $\mu = \sum_{i\geq 1} \tau_i\delta_{W_i}$ on $\mathbb{W}$, where the τ_i's are nonnegative and such that $\sum_{i\geq 1}\tau_i < +\infty$, and the W_i's are iid and independent of the τ_i's. We denote by $\nu(\mathrm{d}s, \mathrm{d}w) = \lambda(s)\mathrm{d}sP(\mathrm{d}w)$ on $\mathbb{R}_+ \times \mathbb{W}$, with $\int_{\mathbb{R}_+} \min\{s,1\}\lambda(s)\mathrm{d}s < +\infty$, the Lévy intensity measure of μ. Let $\Delta_1 > \Delta_2 > \cdots$ be the decreasingly ordered τ_i's, and consider the discrete random measure

$$\mu_{\Delta_1} = \sum_{i\geq 1} \frac{\Delta_{i+1}}{\Delta_1}\delta_{W_{i+1}},$$

such that $\Delta_{i+1}/\Delta_1 \in (0,1)$, for $i \geq 1$, and $\sum_{i\geq 1}\Delta_{i+1}/\Delta_1 < +\infty$. A SP on $\mathbb{W}$ is defined from μ_{Δ_1} as follows. Let $F_{\Delta_1}(\mathrm{d}a) = \exp\left\{-\int_a^\infty \lambda(s)\mathrm{d}s\right\}\lambda(a)\mathrm{d}a$ be the distribution of Δ_1 (Ferguson and Klass 1972, p. 1636), and let G_a be the conditional distribution of $(\Delta_{i+1}/\Delta_1)_{i\geq 1}$ given $\Delta_1 = a$. Moreover, let $\Delta_{1,h}$ denote a random variable whose distribution has a density function $f_{\Delta_{1,h}}(a) = h(a)f_{\Delta_1}(a)$, where h is a nonnegative function and f_{Δ_1} is the density function of F_{Δ_1}. If $(\rho_i)_{i\geq 1}$ are $(0,1)$-valued random variables with distribution $G_{\Delta_{1,h}}$ then

$$\mu_{\Delta_{1,h}} = \sum_{i\geq 1} \rho_i\delta_{W_{i+1}}. \tag{3}$$

is a SP. For short, $\mu_{\Delta_{1,h}} \sim \text{SP}(\nu, h)$. The law of $\mu_{\Delta_{1,h}}$ is a prior distribution for the parameter ζ of the Bernoulli process. The next proposition characterizes the predictive distribution of SP priors. See also (James, Orbanz, and Teh 2015, Proposition 2.2) for a posterior analysis of SP priors.

Proposition 2. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim \text{SP}(\nu, h)$. If $Z_{1:N}$ displays $K_N = k$ distinct features $\{W_1^*, \ldots, W_{K_N}^*\}$, each feature W_i^* appearing exactly $M_{N,i} = m_i$ times, then the conditional distribution of $\Delta_{1,h}$, given $Z_{1:N}$, has a density function of the form

$$g_{\Delta_{1,h} \mid Z_{1:N}}(a) \propto \frac{\prod_{i=1}^{k} \int_0^1 s^{m_i}(1-s)^{N-m_i} a\lambda(as)\mathrm{d}s}{\exp\left\{\sum_{n=1}^{N} \int_0^1 s(1-s)^{n-1} a\lambda(as)\mathrm{d}s\right\}} f_{\Delta_{1,h}}(a). \tag{4}$$

Moreover, the conditional distribution of Z_{N+1}, given $(\Delta_{1,h}, Z_{1:N})$, coincides with the distribution of

$$Z_{N+1} \mid (\Delta_{1,h}, Z_{1:N}) \stackrel{d}{=} Z'_{N+1} + \sum_{i=1}^{K_N} A_{N+1,i}\delta_{W_i^*}, \tag{5}$$

where as (i) $Z'_{N+1} \mid \mu'_{\Delta_{1,h}} = \sum_{i \geq 1} A'_{N+1,i}\delta_{W'_i} \sim \text{BeP}(\mu'_{\Delta_{1,h}})$ and $\mu'_{\Delta_{1,h}} \mid \Delta_{1,h} \sim \text{CRM}(\nu'_{\Delta_{1,h}})$, with $\nu'_{\Delta_{1,h}}(\mathrm{d}s, \mathrm{d}w) = (1-s)^N \Delta_{1,h}\lambda(s\Delta_{1,h})\mathbb{1}_{(0,1)}(s)\mathrm{d}sP(\mathrm{d}w)$; (ii) the $A_{N+1,i}$'s are independent Bernoulli random variables with parameters J_i's, respectively, such that $J_i \mid \Delta_{1,h}$ is distributed according to the density function $f_{J_i \mid \Delta_{1,h}}(s) \propto (1-s)^{N-m_i} s^{m_i} \Delta_{1,h}\lambda(\Delta_{1,h}s)\mathbb{1}_{(0,1)}(s)\mathrm{d}s$ for $i \geq 1$.

See Appendix S2 for the proof of Proposition 2. The marginalization of (5) with respect to (4) leads to the predictive distribution of SP priors: (i) Z_{N+1} displays "new" features W'_i's, and the posterior distribution of statistics of "new" features, given $Z_{1:N}$, is determined by the law of $(\Delta_{1,h}, Z'_{N+1})$; (ii) Z_{N+1} displays "old" features W_i^*'s, and the posterior distribution of statistics of "old" features, given Z_{N+1}, is determined by the law of $(\Delta_{1,h}, \sum_{1 \leq i \leq K_N} A_{N+1,i}\delta_{W_i^*})$. Because of (4) and (5), the law of $(\Delta_{1,h}, Z'_{N+1})$ may include the whole sampling information, depending on the specification of ν and h, and hence the posterior distribution of statistics of "new" features, given $Z_{1:N}$, also includes such an information. As a corollary of Proposition 2, the posterior distribution of the number of unseen features, given $Z_{1:N}$ and fixed prior's parameters, 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 $\Delta_{1,h}$, given $Z_{1:N}$. SP priors thus allow to enrich the Poisson posterior structure arising from CRM priors, in terms of both a more flexible distribution and the inclusion of more sampling information than the sole sample size N, though they may lead to unwieldy posterior inferences due to the marginalization with respect to (4).

The use of the sampling information in the predictive structure of SPs somehow resembles that of Poisson-Kingman (PK) models (Pitman 2006). PK models form a broad class of nonparametric priors for species sampling problems. The DP prior is a PK model whose predictive distribution is such that: (i) the conditional probability that the $(N+1)$th draw is a "new" species, given N observable samples, depends only on the sample size; (ii) the conditional probability that the $(N+1)$th draw is an "old" species, given N observable samples, depends on the sample size, the number of distinct species and their frequencies. Such a behavior resembles that of CRM priors, that is, Proposition 1. PK models allow to include more sampling information in the probability of discovering a "new species" arising under the DP prior, which typically determines a loss of the analytical tractability of posterior inferences for the number of unseen species (Bacallado et al. 2017). Such a behavior resembles that of SP priors, that is, Proposition 2. The PYP prior is arguably the most popular PK model. It stands out for enriching the probability of discovering a "new" species arising under the DP prior, by including the sampling information on the number of distinct species, while maintaining the analytical tractability and interpretability of the DP prior.

3. Stable-Beta Scaled Process (SB-SP) Priors for the Unseen-Features Problem

In Section 2 we showed how SP priors allow to enrich the Poisson posterior structure of the number of unseen features arising under CRM priors, for example the Beta and the stable-Beta process priors. While this is an appealing property, it may lead to a lack of analytical tractability and interpretability of posterior inferences, thus, making SP priors not of practical interest in applications. In this section, we introduce and investigate a peculiar SP prior, which is referred to as the SB-SP prior, and we show that: (i) it leads to a negative Binomial posterior distribution for the number of unseen features, which generalizes the Poisson distribution while maintaining its analytical tractability and interpretability; (ii) it leads to a posterior distribution for the number of unseen features, which depends on the sampling through the sample size and the number of distinct features. The SB-SP prior thus provides a sensible tradeoff between the enrichment of the Poisson posterior structure of the number of unseen features arising under CRM priors and the analytical tractability and interpretability of posterior inferences. In particular, we characterize the SB-SP prior as the sole SP prior for which the posterior distribution of the number of unseen features depends on the observable sample only through the sample size and the number of distinct features. The SB-SP may thus be considered as a natural counterpart of the PYP for the unseen-feature problem.

3.1. SB-SP Priors for Bernoulli Processes

Stable scaled processes (S-SP) (James, Orbanz, and Teh 2015) form a subclass of SPs, and hence their definition follows from Section 2. In particular, for any $\sigma \in (0,1)$, let μ_σ be the σ-stable CRM on $\mathbb{W}$ (Kingman 1975), which is characterized by the Lévy intensity measure $\nu_\sigma(\mathrm{d}s, \mathrm{d}w) = \lambda_\sigma(s)\mathrm{d}sP(\mathrm{d}w)$ on $\mathbb{R}_+ \times \mathbb{W}$, with $\int_{\mathbb{R}_+} \min\{s,1\}\lambda_\sigma(s)\mathrm{d}s < +\infty$, where $\lambda_\sigma(s) = \sigma s^{-1-\sigma}$. We recall that the largest atom Δ_1 of μ_σ is distributed according to the density function

$$f_{\Delta_1}(a) = \sigma a^{-1-\sigma} \exp\left\{-a^{-\sigma}\right\}. \tag{6}$$

That is, $\Delta_1 = E^{-1/\sigma}$, where E denotes a negative exponential random variable with parameter 1. For any nonnegative function h, a S-SP on $\mathbb{W}$ is defined as the SP with law $\mathrm{SP}(\nu_\sigma, h)$. S-SP priors generalizes the Beta process prior, which is recovered by setting h to be the identity function, and then letting $\sigma \to 0$ (James, Orbanz, and Teh 2015). The predictive distribution of $\zeta \sim \mathrm{SP}(\nu_\sigma, h)$ is obtained from Proposition 2. In the next theorem, we characterize the S-SP priors as the sole SP priors for which the conditional distribution of $\Delta_{1,h}$, given $Z_{1:N}$, depends on $Z_{1:N}$ only through the sample size N and the number K_N of distinct features in $Z_{1:N}$.

Theorem 1. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim \mathrm{SP}(\nu, h)$, and let $Z_{1:N}$ displays K_N distinct features with corresponding frequencies $(M_{N,1}, \ldots, M_{N,K_N})$. Moreover, let $\nu(\mathrm{d}s, \mathrm{d}w) = \lambda(s)\mathrm{d}s P(\mathrm{d}w)$, and let $f_{\Delta_{1,h}}$ be the density function of $\Delta_{1,h}$. If $f_{\Delta_{1,h}} > 0$ on $\mathbb{R}_+$ and the functions λ and $f_{\Delta_{1,h}}$ are continuously differentiable, then the conditional distribution of $\Delta_{1,h}$, given $Z_{1:N}$, depends on $Z_{1:N}$ only through N and K_N if and only if $\nu = \nu_\sigma$.

See Appendix S3 for the proof of Theorem 1. We recall from Section 2 that the conditional distribution of $\Delta_{1,h}$, given $Z_{1,N}$, uniquely determines the amount of sampling information included in the posterior distribution of statistics of "new" features. Then, according to Theorem 1, S-SP priors are the sole SP priors for which the posterior distribution of the number of unseen features, given $Z_{1:N}$ and fixed prior's parameters, depends on $Z_{1:N}$ only through N and K_N. As a corollary of Theorem 1, the Beta process prior is the sole S-SP prior for which the posterior distribution of statistics of "new" features depends on $Z_{1:N}$ only through N. Analogous predictive characterizations are well-known in species sampling problems, and they are typically referred to as "sufficientness" postulates' (Bacallado et al. 2017). In particular, the DP prior is characterized as the sole species sampling prior for which the conditional probability that the $(N+1)$th draw is a "new" species, given N observable samples, depends only on the sample size (Regazzini 1978). Moreover, the PYP prior is characterized as the sole species sampling prior for which the conditional probability that the $(N+1)$th draw is a "new" species, given N observable samples, depends only on the sample size and the number of distinct species in the sample (Zabell 2005). Theorem 1 provides a "sufficientness" postulates' in the context of feature sampling problems.

As a noteworthy example of S-SPs, we introduce the SB-SP. The SB-SP is a S-SP obtained by a suitable specification of the nonnegative function h. In particular, for any $c, \beta > 0$ let

$$h_{c,\beta}(a) = \frac{\beta^{c+1}}{\Gamma(c+1)} a^{-c\sigma} \exp\left\{-(\beta-1)a^{-\sigma}\right\}, \qquad (7)$$

where $\Gamma(\cdot)$ denotes the Gamma function. Then a SB-SP on $\mathbb{W}$ is defined as the SP with law $\mathrm{SP}(\nu_\sigma, h_{c,\beta})$. For short, we denote the law of a SB-SP by SB-SP(σ, c, β). The SB-SP prior generalizes the Beta process prior, which is recovered by setting $c = 0$ and $\beta = 1$, and then letting $\sigma \to 0$. According to the construction of SPs, the distribution of $\Delta_{1,h_{c,\beta}}$ has a density function obtained by combining (6) and (7); this is a polynomial-exponential tilting of the density function (6). In particular, $\Delta_{1,h_{c,\beta}}^{-\sigma}$ is distributed as a Gamma distribution with shape $(c+1)$ and rate β. Such a straightforward distribution for $\Delta_{1,h_{c,\beta}}$ is at the core of the analytical tractability of posterior inferences under the SB-SP prior; this fact will be clear in the application of the SB-SP prior to the problem of estimating the number of unseen features. The next proposition characterizes the predictive distribution of the SB-SP prior.

Proposition 3. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim \text{SB-SP}(\sigma, c, \beta)$. If $Z_{1:N}$ displays $K_N = k$ distinct features $\{W_1^*, \ldots, W_{K_N}^*\}$, each feature W_i^* appearing exactly $M_{N,i} = m_i$ times, then the conditional distribution of $\Delta_{1,h_{c,\beta}}$, given $Z_{1:N}$, has a density function of the form

$$g_{\Delta_{1,h_{c,\beta}} \mid Z_{1:N}}(a) = \sigma \frac{(\beta + \gamma_0^{(N)})^{k+c+1}}{\Gamma(k+c+1)} a^{-k\sigma-(c+1)\sigma-1} \mathrm{e}^{-a^{-\sigma}(\beta+\gamma_0^{(N)})}, \quad (8)$$

where $\gamma_0^{(N)} = \sigma \sum_{1 \le i \le N} B(1-\sigma, i)$, with $B(\cdot,\cdot)$ being the (Euler) Beta function. Moreover, the conditional distribution of Z_{N+1}, given $(\Delta_{1,h_{c,\beta}}, Z_{1:N})$, coincides with the distribution of

$$Z_{N+1} \mid (\Delta_{1,h_{c,\beta}}, Z_{1:N}) \stackrel{d}{=} Z'_{N+1} + \sum_{i=1}^{K_N} A_{N+1,i}\delta_{W_i^*}, \qquad (9)$$

where:

(i) $Z'_{N+1} \mid \mu'_{\Delta_{1,h_{c,\beta}}} = \sum_{i \ge 1} A'_{N+1,i}\delta_{W'_i} \sim \mathrm{BeP}(\mu'_{\Delta_{1,h_{c,\beta}}})$ such that $\mu'_{\Delta_{1,h_{c,\beta}}} \mid \Delta_{1,h_{c,\beta}} \sim \mathrm{CRM}(\nu'_{\Delta_{1,h_{c,\beta}}})$, with

$$\nu'_{\Delta_{1,h_{c,\beta}}}(\mathrm{d}s, \mathrm{d}w) = \Delta_{1,\Delta_{1,h_{c,\beta}}}^{-\sigma}(1-s)^N \sigma s^{-1-\sigma} \mathbb{1}_{(0,1)}(s)\mathrm{d}s P(\mathrm{d}w);$$

(ii) the $A_{N+1,i}$'s are independent Bernoulli random variables with parameters J_i's, respectively, such that each $J_i \mid \Delta_{1,h_{c,\beta}}$ is distributed according to a density function of the form

$$f_{J_i \mid \Delta_{1,h_{c,\beta}}}(s) = \frac{1}{B(m_i - \sigma, N - m_i + 1)} s^{m_i-\sigma}(1-s)^{N-m_i+1}\mathbb{1}_{(0,1)}(s).$$

See Appendix S3 for the proof of Proposition 3. According to Equation (8), the conditional distribution of $\Delta_{1,h_{c,\beta}}$, given $Z_{1:N}$, depends on $Z_{1:N}$ only through the sample size N and the number K_N of distinct features in $Z_{1:N}$. This agrees with Theorem 1, implying that the posterior distribution of the number of unseen features, given $Z_{1:N}$ and fixed prior's parameters, depends on $Z_{1:N}$ only through N and K_N. Because of (8) and (9), the posterior distribution of statistics of "new" features stands out for analytical tractability, thus, being competitive with that arising from CRMs, for example, the Beta and the stable-Beta processes. In particular, from Equation (9), the conditional distribution of Z'_{N+1}, given $(\Delta_{1,h_{c,\beta}}, Z_{1:N})$ is a Poisson distribution that depends on $Z_{1:N}$ only through N. Then, from (8), its marginalization with respect to the conditional distribution of $\Delta_{1,h_{c,\beta}}$, given $Z_{1:N}$, leads to a negative Binomial posterior distribution. Such an appealing property arises from the peculiar form $h_{c,\beta}$ that, combined with ν_σ, leads to a conjugacy property for the conditional distribution of $\Delta_{1,h_{c,\beta}}$, given $Z_{1:N}$. That is, the conditional

distribution of $\Delta_{1,h_{c,\beta}}^{-\sigma}$, given $Z_{1:N}$, is a Gamma distribution with shape $(K_N + c + 1)$ and rate $\beta + \gamma_0^{(N)}$, which is the distribution $\Delta_{1,h_{c,\beta}}^{-\sigma}$ with shape and rate being updated through $Z_{1:N}$. The next proposition establishes the distribution of a random sample $Z_{1:N}$ from a SB-SP prior. See Appendix S3 for details.

Proposition 4. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim$ SB-SP(σ, c, β). The probability that $Z_{1:N}$ displays a particular feature allocation of k distinct features with frequencies $(m_1, \ldots, m_k)$ is

$$p_k^{(N)}(m_1, \ldots, m_k) = \frac{\frac{\sigma^k \beta^{c+1}}{(\beta+\gamma_0^{(N)})^{k+c+1}}}{\frac{\Gamma(c+1)}{\Gamma(k+c+1)}} \prod_{i=1}^{k} \times \frac{\Gamma(m_i - \sigma)\Gamma(N - m_i + 1)}{\Gamma(N - \sigma + 1)}. \quad (10)$$

3.2. BNP Inference for the Unseen-Features Problem

Now, we apply the SB-SP prior to the unseen-features problem. For any $N \geq 1$ let $Z_{1:N}$ be an observable sample modeled as the BNP Bernoulli model (1), with $\zeta \sim$ SB-SP(σ, c, β). Moreover, under the same model of the Z_n's, for any $M \geq 1$ let $(Z_{N+1}, \ldots, Z_{N+M})$ be additional unobservable sample. Then, the unseen-feature problem calls for the estimation of

$$U_N^{(M)} = \sum_{i \geq 1} \mathbb{1}\left(\sum_{m=1}^{M} A_{N+m,i} > 0\right) \mathbb{1}\left(\sum_{n=1}^{N} A_{n,i} = 0\right), \quad (11)$$

namely the number of hitherto unseen features that would be observed in $(Z_{N+1}, \ldots, Z_{N+M})$. As generalization of the unseen-feature problem (11), for $r \geq 1$ we consider the estimation of

$$U_N^{(M,r)} = \sum_{i \geq 1} \mathbb{1}\left(\sum_{m=1}^{M} A_{N+m,i} = r\right) \mathbb{1}\left(\sum_{n=1}^{N} A_{n,i} = 0\right), \quad (12)$$

namely the number of hitherto unseen features that would be observed with prevalence r in $(Z_{N+1}, \ldots, Z_{N+M})$. Of special interest is $r = 1$, which concerns rare (unique) features. The next theorem characterizes the posterior distributions of $U_N^{(M)}$ and $U_N^{(M,r)}$, given $Z_{1:N}$. We denote by NegativeBinonial(n, p) the negative Binomial distribution with parameter n and $p \in (0, 1)$.

Theorem 2. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim$ SB-SP(σ, c, β), and let $Z_{1:N}$ displays $K_N = k$ distinct features with frequencies $(M_{N,1}, \ldots, M_{N,K_N}) = (m_1, \ldots, m_k)$. Then, the posterior distributions of $U_N^{(M)}$ and of $U_N^{(M,r)}$, given $Z_{1:N}$, coincide with the distributions of

$$U_N^{(M)} \mid Z_{1:N} \sim \text{NegativeBinonial}\left(K_N + c + 1, \frac{\gamma_N^{(M)}}{\beta + \gamma_0^{(N+M)}}\right), \quad (13)$$

and

$$U_N^{(M,r)} \mid Z_{1:N} \sim \text{NegativeBinonial} \times \left(K_N + c + 1, \frac{\rho_N^{(M,r)}}{\beta + \gamma_0^{(N)} + \rho_N^{(M,r)}}\right), \quad (14)$$

for any index of prevalence $r \geq 1$, respectively, where $\gamma_N^{(M)} = \sigma \sum_{1 \leq i \leq M} B(1-\sigma, N+i)$ and where $\rho_N^{(M,r)} = \binom{M}{r} \sigma B(r-\sigma, N+M-r+1)$, with $B(\cdot, \cdot)$ denoting the (Euler) Beta function.

See Appendix S4 for the proof of Theorem 2. The posterior distributions (13) and (14) depend on $Z_{1:N}$ through the sample size N and the number K_N of distinct features. This is in contrast with the corresponding posterior distributions obtained under the Beta and the stable-Beta process priors, which are Poisson distributions that depend on $Z_{1:N}$ only through N (Masoero et al. 2021, Proposition 1). BNP estimators of $U_N^{(M)}$ and $U_N^{(M,r)}$, with respect to a squared loss function, are obtained as the posterior expectations of (13) and (14), that is,

$$\widehat{U}_N^{(M)} = (K_N + c + 1)\frac{\gamma_N^{(M)}}{\beta + \gamma_0^{(N+M)} - \gamma_N^{(M)}} \quad (15)$$

and

$$\widehat{U}_N^{(M,r)} = (K_N + c + 1)\frac{\rho_N^{(M,r)}}{\beta + \gamma_0^{(N)}}, \quad (16)$$

respectively. The estimators (15) and (16) are simple, linear in the sampling information and computationally efficient. In the next theorem we establish the large M asymptotic behavior of the posterior distributions (13) and (14), showing that the number of unseen features has a power-law growth in M. The same growth in M holds under the stable-Beta process prior (Masoero et al. 2021, Proposition 2), though the limiting distribution is degenerate.

Theorem 3. Let $Z_{1:N}$ be a random sample from (1) with $\zeta \sim$ SB-SP(σ, c, β), and let $Z_{1:N}$ displays $K_N = k$ distinct features with frequencies $(M_{N,1}, \ldots, M_{N,K_N}) = (m_1, \ldots, m_k)$. As $M \to +\infty$

$$\frac{U_N^{(M)}}{M^\sigma} \mid Z_{1:N} \xrightarrow{\text{a.s.}} W_N, \quad (17)$$

where W_N is a Gamma random variable with shape $(K_N + c + 1)$ and rate $(\beta + \gamma_0^{(N)})/\Gamma(1-\sigma)$, and

$$\frac{U_N^{(M,r)}}{M^\sigma} \mid Z_{1:N} \xrightarrow{\text{a.s.}} W_{N,r}, \quad (18)$$

where $W_{N,r}$ is a Gamma random variable with shape $(K_N + c + 1)$ and rate $\Gamma(r+1)(\beta + \gamma_0^{(N)})/\sigma\Gamma(r-\sigma)$.

4. Experiments

Over the last decade, genomics has witnessed an extraordinary improvement in the data availability due to the advent of next generation sequencing technologies. Thanks to larger and richer datasets, researchers have started uncovering the role and impact of rare genetic variants in heritability and human disease (Hernandez et al. 2019; Momozawa and Mizukami 2020). The development of methods for estimating the number of new genomic variants to be observed in future studies is an active research area, as it can aid the design of effective clinical procedures in precision medicine (Ionita-Laza, Lange, and Laird 2009; Zou et al. 2016), enhance understanding of cancer biology

(Chakraborty et al. 2019), and help to optimize sequencing procedures (Rashkin et al. 2017; Masoero et al. 2021). Here, we consider datasets of individual genomic sequences. Following common practice, we assume that an underlying fixed and idealized genomic sequence (the "reference") is given. Then, each coordinate of an individual sequence reports the presence (1) or absence (0) of variation at a given locus with respect to the reference. All variants are treated equally, namely, any expression differing from the underlying reference at a given locus counts as a variant. We make use of our methodology to estimate the number of genomic loci at which variation was not observed in the original sample, and is going to be observed in (at least one of) M additional datapoints.

We find in our experiments that the estimates of the total number of new variants to be observed produced using the SB-SP-Bernoulli model, hereafter referred to as SSB, tend to be more accurate than other available methods in the literature. This phenomenon is particularly evident when the sample size N of the training set is small, and when the extrapolation size M is large with respect to N. Moreover, the SSB model is particularly effective in estimating the number of new rare variants, for example, variants appearing only once in the additional unobservable samples. Accurate estimation of rare variants is particularly important, as these are believed to be largely responsible for heritability of human disease (Rashkin et al. 2017; Chakraborty et al. 2019). To benchmark the quality of the SSB, we consider a number of competing methodologies for the feature prediction problem available in the literature: (i) Jackknife estimators (J) (Gravel 2014); (ii) a linear programming method (LP) (Zou et al. 2016) and variations of Good-Toulmin estimators (GT) (Chakraborty et al. 2019). We also compare our empirical findings to a BNP estimator obtained under the stable-Beta process prior (3BB), which has been introduced in Masoero et al. (2021). We complete our analysis with a thorough investigation on synthetic data in S6 and S7, as well as on additional real data from the gnomAD database (Karczewski et al. 2020) in S8.

https://github.com/lorenzomasoero/ScaledProcesses contains code and data to replicate all our analyses.

4.1. Empirics and Evaluation Metrics

For the SSB method to be useful, we need to estimate the underlying, unknown, parameters of the SB-SP prior. To learn these prior's parameters, we here adopt an empirical Bayes procedure, which consists in maximizing the marginal distribution (10). In particular, we maximize numerically Equation (10) with respect to the parameters $\beta > 0, c > 0$ and $\sigma \in (0, 1)$ of the SB-SP prior, and use the resulting values to produce our estimators. That is, we let

$$(\widehat{\beta}, \widehat{c}, \widehat{\sigma}) = \arg \max_{(\beta,c,\sigma)} \left\{ p_k^{(N)}(m_1, \ldots, m_k) \right\},$$

and plug these values in the BNP estimator (13) and (14). The resulting values provide our BNP estimates of the number $U_N^{(M)}$ of new variants and the number $U_N^{(M,r)}$ of new variants with prevalence r.

To assess the accuracy of our estimates, we consider the percent deviation of the estimate from the truth to be the achieved accuracy. That is, the accuracy of the estimator $\widehat{U}_N^{(M)}$ is defined as

$$v_N^{(M)} := 1 - \min \left\{ \frac{|U_N^{(M)} - \widehat{U}_N^{(M)}|}{U_N^{(M)}}, 1 \right\}. \tag{19}$$

In particular, the accuracy $v_N^{(M)}$ equals 1 when the estimate is perfect (no error is incurred), and decreases to 0 as the estimate deviates from the truth. The min operator in (19) ensures that $v_N^{(M)}$ lies in $[0, 1]$: we let the accuracy to be equal to 0 whenever there is a severe overestimation, and the percentage estimation error exceeds 100%, that is, when $\widehat{U}_N^{(M)} \geq 2 \times U_N^{(M)}$. The SSB, 3BB and LP methods also offer an estimate for the number of new features observed with a given prevalence r. We let $v_N^{(M,r)}$ be the accuracy metric, where we replace in (19) $U_N^{(M)}$ with $U_N^{(M,r)}$, the number of new features observed with prevalence r, and $\widehat{U}_N^{(M)}$ with $\widehat{U}_N^{(M,r)}$.

4.2. Estimating the Number of New Variants in Cancer Genomics

Following the empirical study of Chakraborty et al. (2019), we make use of data from the Cancer Genome Atlas (TCGA), the largest publicly available cancer genomics dataset, containing somatic mutations from 10,295 patients and spanning 33 different cancer types. We partition the samples into 33 smaller datasets according to cancer-type annotation of each patient. See Chakraborty et al. (2019) and (Masoero et al. 2021, Appendix F) for details on the data and the experimental setup. For each cancer type, we retain a small fraction of the data for purposes of training, and consider the task of estimating the number of new variants that will be observed in a follow-up sample given a pilot sample. We validate our estimates by comparing the estimate $\widehat{U}_N^{(M)}$ of the number of distinct variants to the true value, obtained by extrapolating to the remaining data. To assess the variability and error in our estimates, we repeat for every cancer type the experiment on $S = 1000$ subsets of the data, each obtained by randomly subsampling without replacement from the full sample.

We find that the SSB and 3BB methods perform particularly well when the training sample size N is small compared to the extrapolation sample size M. This setting is relevant in the context of cancer genomics, as scientists are interested in understanding the "unexploited potential" of the genetic information, especially for rare cancer subtypes (Chakraborty et al. 2019; Huyghe et al. 2019). To compare and quantify the performance of the available methodologies in this setting, we report in Figure 1 the distribution of the estimation accuracy when retaining only $N = 10$ samples for training and extrapolating to the largest possible sample size M for which we can compute the accuracy metric (Equation (19)). We report results for the 10 cancer types with the largest number of samples in the original dataset. For each cancer type and for each method, the distribution of the estimation accuracy is obtained by considering its performance across the $S = 1000$ replicates. Across all cancer types, the estimates obtained from the SSB method achieve higher accuracy.

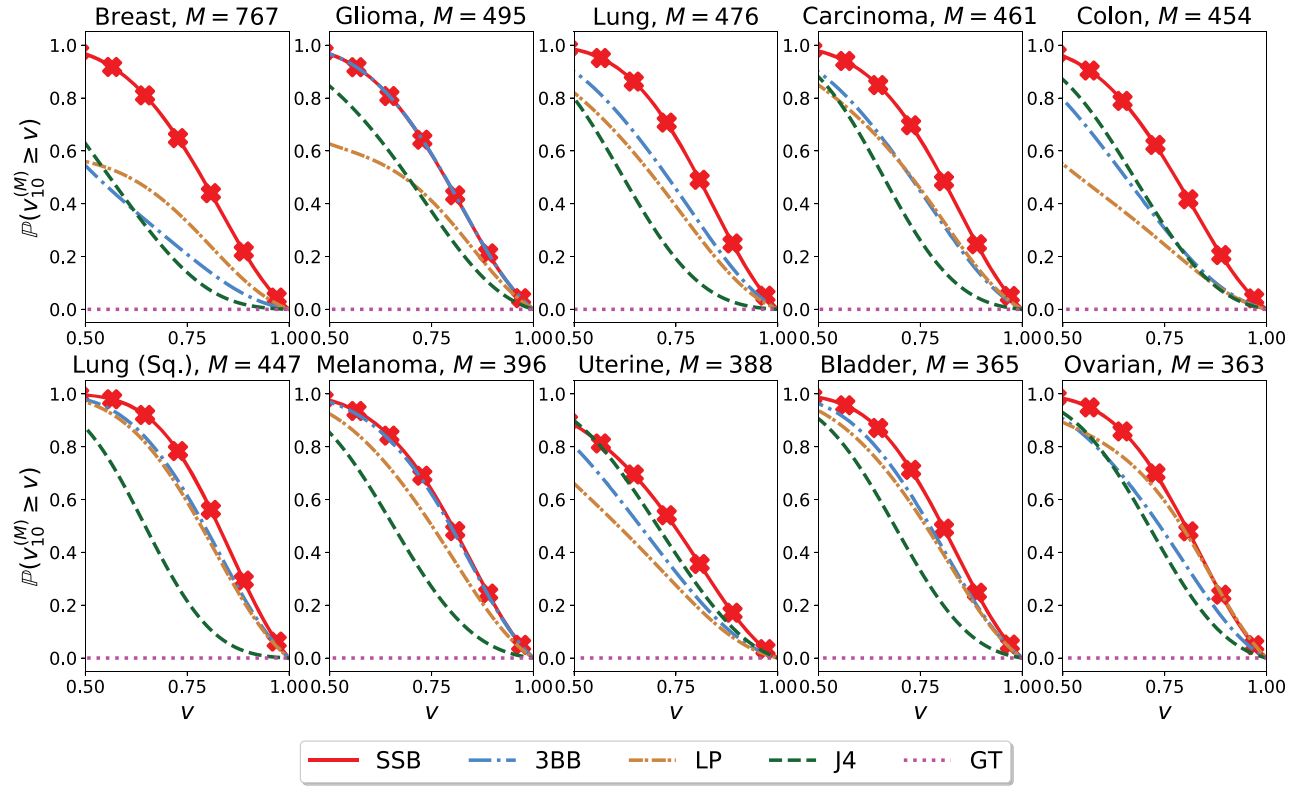

Figure 1. Estimation accuracy $v_N^{(M)}$ for the number of new genomic variants $\widehat{U}_{10}^{(M)}$. For each method and each cancer type, we retain $N = 10$ random samples and use them to estimate up to M total observations, where $N + M$ is the size of the original sample.

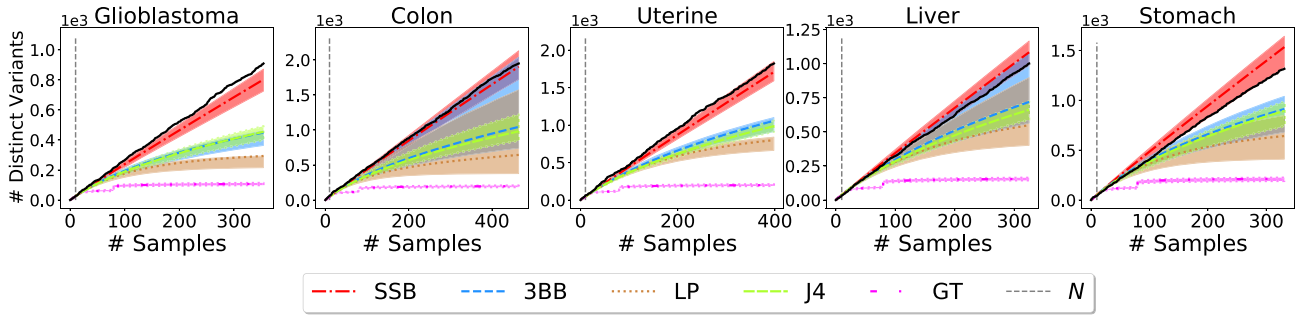

Figure 2. Estimation of the number of new genomic variants $\widehat{U}_N^{(i)}$, for $i = 1, \ldots, M$. For each method and cancer type, we retain $N = 10$ random samples and use them to estimate up to the largest possible size. We fit each model on the full sample, as well as $N = 10$ additional times by iteratively leaving one datapoint out from the training sample. The solid black line is the true number of features that would have been observed (vertical axis) for any extrapolation size $N+M$ (horizontal axis), for a fixed ordering of the data. Shaded regions report the prediction range obtained from the estimates from the leave-one-out fits.

We show in Figure 2 the behavior of $\widehat{U}_N^{(i)}$ for five different cancer types as $i = 1, \ldots, M$. Again, we let $N = 10$, and M be the largest possible extrapolation value, as dictated by the dataset size. We report the estimates obtained from a fixed sample of size $N = 10$, as well as the variability around such estimates obtained by refitting each model, iteratively leaving one datapoint out from the sample. In this setting, the SSB method outperforms competing methods in terms of estimation accuracy. Moreover, the variability in the estimates arising from refitting the model on subsets of the data provides a useful measure of uncertainty in such estimation.

4.3. Estimating the Number of New Rare Variants in Cancer Genomics

In recent years, the cancer genomics research community has become increasingly interested in studying and understanding the role of extremely rare variants, such as singletons, that is, observed in only one patient. Evidence suggests that rare deleterious variants can have far stronger effect sizes than common variants (Rasnic, Linial, and Linial 2020) and can play an important role in the development of cancer. For example, in breast cancer, it is well accepted that the risk of a variant is inversely proportional with respect to its prevalence: the rarer the variant, the higher the risk (Wendt and Margolin 2019). Therefore, effective identification and discovery of rare variants is an active, is an ongoing research area (Lawrenson et al. 2016; Lee et al. 2019). This phenomenon is not limited to breast cancer, but is progressively being studied across different cancer types. See, for example, the recent works on ovarian (Phelan et al. 2017), skin (Goldstein et al. 2017), prostate (Nguyen-Dumont et al. 2020) and lung (Liu et al. 2021) cancers and references therein. In downstream analysis, these estimates could be useful

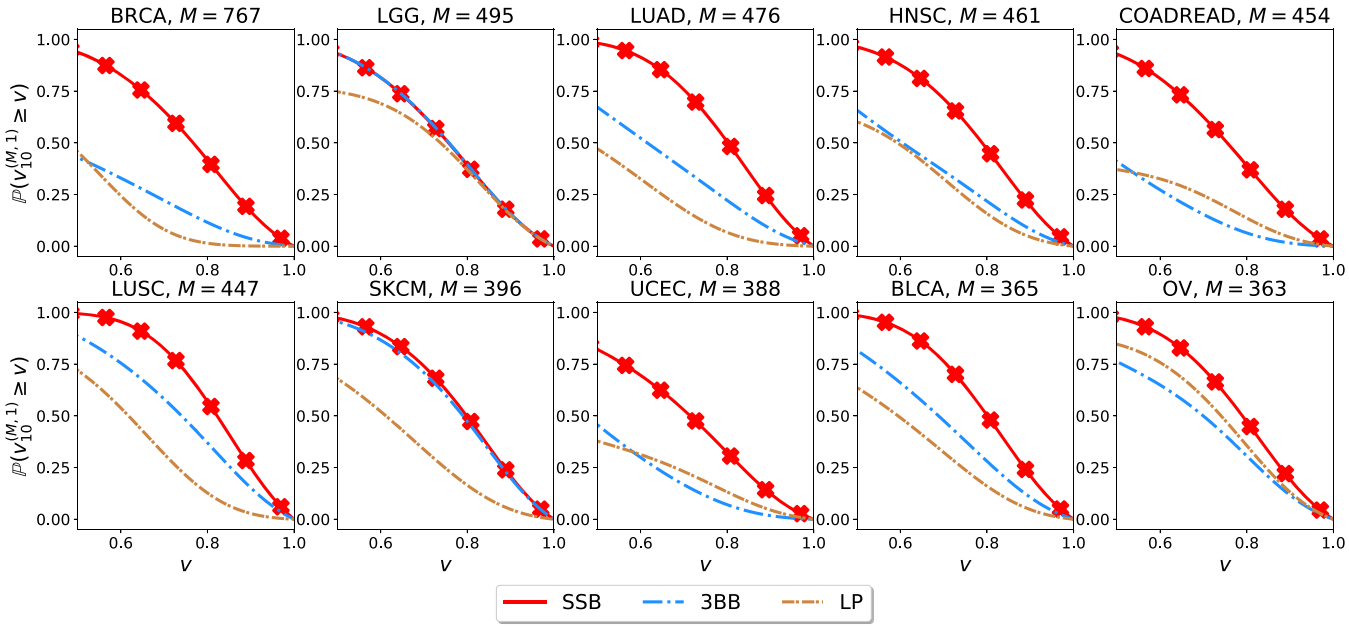

Figure 3. Estimation accuracy $v_N^{(M,1)}$ for new variants appearing with prevalence one in future unobservable samples for different cancer types. For each method and each cancer, we retain $N = 10$ random samples and use them to estimate up to the largest possible size.

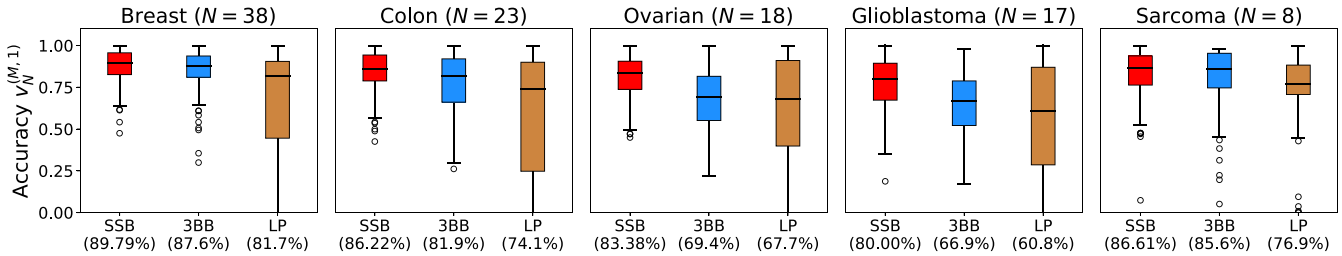

Figure 4. Estimation accuracy $v_N^{(M,1)}$ for new variants appearing with prevalence one in future samples. For each method and different cancer types, we retain a random sample of size $N = 5\%$ of the available dataset, and use it to estimate up to the largest possible size.

for planning and designing future experiments, for example, informing scientists on the number of new samples to be collected in order to observe a target number of new variants, or for power analysis considerations in rare variants association tests (Rashkin et al. 2017).

The BNP framework considered here allows us to estimate the number of new rare variants to be discovered (Figure 3). While Zou et al. (2016) did not consider the problem of estimating rare variants, it is straightforward to obtain an estimate for this quantity from their framework. Indeed, for every prevalence $x \in [0,1]$, the LP estimates the histogram $h(x)$, which counts the number of variants appearing with prevalence x in the population, and the number of variants appearing with prevalence r follows from the binomial sampling model assumption, namely $\widehat{U}_N^{(M,r)} = \sum_x h(x) \left\{ \binom{N+M}{r} x^r (1-x)^{N+M-r} - \binom{N}{r} x^r (1-x)^{N-r} \right\}$. We show in Figure 4 that the SSB method provides better estimates than the 3BB and LP methods.

4.4. Coverage and Calibrated Uncertainties

One of the benefits of the BNP approach is that it automatically yields a notion of variability of the estimate of U via posterior credible intervals. We here check whether these intervals produce a useful notion of uncertainty, by investigating their calibration. For $\alpha \in (0,1)$, we say that a $100 \times \alpha\%$ credible interval is calibrated if it contains the true value of interest, arising from hypothetical repeated draws, $100 \times \alpha\%$ of the times. We here assess the calibration of a $100 \times \alpha\%$ credible interval for $U_N^{(M)}$ conditionally given $Z_{1:N}$ as follows. Let S be a large number ($S = 1000$ in our experiments). For each $s = 1, \ldots, S$, we retain a random subset of the data of size N, and estimate the corresponding parameters $\widehat{\beta}, \widehat{c}, \widehat{\sigma}$ as discussed in Section 4.1. Then, we let $\widehat{W}_{N,s,\text{low}}^{(M)}(\alpha)$, $\widehat{W}_{N,s,\text{hi}}^{(M)}(\alpha)$ be the endpoints of a $100 \times \alpha\%$ credible interval for the distribution of the number of new features, as given by Equation (13), centered around the posterior predictive mean. We compute coverage calibration via

$$w_N^{(M)}(\alpha) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1} \left\{ \widehat{W}_{N,s,\text{low}}^{(M)}(\alpha) \leq K_{N+M} \leq \widehat{W}_{N,s,\text{hi}}^{(M)}(\alpha) \right\}.$$

This is the fraction of the S experiments in which the true value was contained by a $100 \times \alpha\%$ credible interval. The closer $w_N^{(M)}(\alpha)$ to α, the better calibrated the credible intervals. We compute the same quantity for the 3BB method using the results in Masoero et al. (2021). Although still not perfect, we find that the posterior predictive intervals obtained from the SSB method

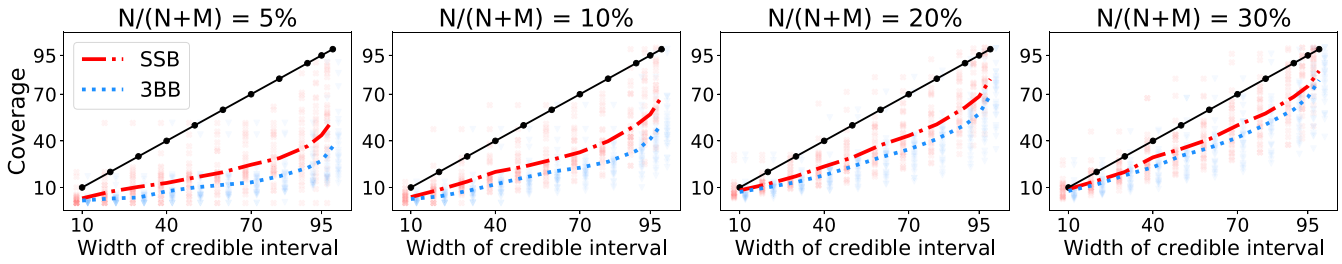

Figure 5. Coverage calibration of BNP estimators for number of new variants in future samples across all cancer types in TCGA. Different subplots refer to different ratios of the training N with respect to the extrapolation M. For each cancer, we retain a training sample of size $N \in \{5\%, 10\%, 20\%, 30\%\}$ of the total available dataset, and extrapolate up to the largest available M. Colored lines report the average coverage $w_N^{(M)}(\alpha)$ across all cancer types (y-axis) as a function of α (x-axis). Faded dots refer to coverage for individual cancer types.

are better calibrated than the ones under the 3BB method (see Figure 5).

5. Discussion

Masoero et al. (2021) first applied CRM priors to the unseen-features problem, showing that: (i) despite the broadness of the class of CRM priors, all CRM priors lead to the same Poisson posterior structure for the number of unseen features, which thus makes them not a flexible prior model for the unseen-features problem; (ii) while the Poisson posterior distribution may be appealing in principle, making the posterior inferences analytically tractable and of easy interpretability, its independence from $Z_{1:N}$ makes the BNP approach a questionable over-simplification, with posterior inferences being completely determined by the estimation of unknown prior's parameters. In this article, we introduced the SB-SP prior, and showed that: (i) it enriches the posterior distribution of the number of unseen features arising under CRM priors, which results in a negative Binomial distribution whose parameters depend on the sample size and the number of distinct features; (ii) it maintains the same analytical tractability and interpretability as CRM priors, which results in BNP estimators that are simple, linear in the sampling information and computationally efficient. The effectiveness of the SB-SP prior is showcased through an empirical analysis on synthetic and real data. Under the SB-SP prior, we found that estimates of the unseen number of features are accurate, and they outperform the most popular competitors in the challenging scenario where the sample size N is particularly small, and also small with respect to the extrapolation size M.

Our approach admits an extension to the multiple-feature setting, which takes into account of the many forms of variation, for example, single nucleotide changes, tandem repeats, insertions and deletions, copy number variations (Zou et al. 2016). We briefly describe the multiple-feature setting, and defer to Appendix S5 for details. It is assumed that a feature w_i comes with a characteristic, that is, the form of variation, chosen among $q > 1$ characteristics. For $N \geq 1$, the observable sample $\boldsymbol{Z}_{1:N} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ is modeled as a $\{0,1\}^q$-valued stochastic process $\boldsymbol{Z} = \sum_{i \geq 1} \boldsymbol{A}_i \delta_{w_i}$, where $\boldsymbol{A}_i := (A_{i,1}, \ldots, A_{i,q})$ is a Multinomial random variable with parameter $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,q})$ such that $|\boldsymbol{p}_i| = \sum_{1 \leq j \leq q} p_{i,j} < 1$, and the $\boldsymbol{A}_i$'s are iid. That is, for any $i \geq 1$ all the $A_{i,j}$'s are equal to 0 with probability $(1 - |\boldsymbol{p}_i|)$, that is, w_i does not display variation, or only one $A_{i,j}$'s is equal to 1 with probability $p_{i,j}$, that is, w_i displays variation with characteristic j. $\boldsymbol{Z}$ is a multivariate Bernoulli process with parameter $\boldsymbol{\zeta} = \sum_{i \geq 1} \boldsymbol{p}_i \delta_{w_i}$. The stable-Beta-Dirichlet process prior for $\boldsymbol{\zeta}$ is a multivariate generalization of the stable-Beta process prior (James 2017), and it leads to a Poisson posterior distribution for the number of unseen features, given $\boldsymbol{Z}_{1:N}$, which depends on $\boldsymbol{Z}_{1:N}$ only through N. In Appendix S5 we introduce a scaled version of the stable-Beta-Dirichlet process, and show that it leads to a negative Binomial posterior distribution for the number of unseen features, which depends on $\boldsymbol{Z}_{1:N}$ through N and the number of distinct features in $\boldsymbol{Z}_{1:N}$.

SP priors have been introduced in James, Orbanz, and Teh (2015) and, to the best of our knowledge, since then no other works have further investigated such a class of priors. To date, the peculiar predictive properties of SP priors appear to be unknown in the BNP literature. Our work on the unseen-features problem is the first to highlight the great potential of SP priors in BNPs, showing that they provide a critical tool for enriching the predictive structure of the popular CRM priors (James 2017; Broderick, Wilson, and Jordan 2018). We believe that SPs may be of interest beyond the unseen-features problem, and more generally beyond the broad class of feature sampling problems. CRM priors, and in particular the Beta and stable-Beta process priors, have been widely used in several contexts, with a broad range of applications in topic modeling, analysis of social networks, binary matrix factorization for dyadic data, analysis of choice behavior arising from psychology and marketing surveys, graphical models, and analysis of similarity judgment matrices. See Griffiths and Ghahramani (2011) and references therein for details. In all these contexts, SP priors may be more effective than CRM priors, as they allow to better exploit the sampling information in posterior inferences.

Among applications of SP priors beyond features sampling problems, it is worth mentioning the use of SP priors as hierarchical (or latent) priors in models of unsupervised learning (Griffiths and Ghahramani 2011, sec. 5), the most popular being Gaussian latent feature modeling. Differently from features sampling problems, where the values of features' labels W_is are immaterial, in Gaussian latent feature modeling the values the W_i's become material. That is, under the Gaussian latent feature model with a SP prior, observations are assumed to modeled as a multivariate Gaussian distribution, whose mean depends on latent features that are modeled with a SP prior, thus, making the values of features' labels W_i's of critical importance for the analysis. Bayesian factor analysis (Knowles and Ghahramani

2011) provides another context where SP priors may be usefully applied as hierarchical priors. Within the context of factor analysis, we also mention the work of Ayed and Caron (2021) with applications to network analysis. There, the authors exploit CRM priors to recover the latent community structure in a network between individuals, and the features' labels describe the level of affiliation of a certain individual to a latent community. In such a context, we believe that SP priors may be used in place of CRM priors, with the advantage of introducing richer predictive structure. In this respect, our work paves the way to promising directions of future research, in terms of both methods and applications.

Funding

Federico Camerlenghi and Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 817257. Federico Camerlenghi and Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, "Dipartimenti di Eccellenza" grant 2018-2022. Lorenzo Masoero and Tamara Broderick were supported in part by the DARPA I2O LwLL program, an NSF CAREER Award, and ONR award N00014-17-1-2072.

Supplementary Materials

All proofs of our statements, as well as additional experiments can be found in the Appendix. Code to replicate all our analyses is available at *https://github.com/lorenzomasoero/ScaledProcesses/*.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor, and two anonymous Referees for their comments and suggestions, which greatly improved the contents and presentation of the article. The authors thank Joshua Schraiber for useful discussions.

ORCID

Lorenzo Masoero http://orcid.org/0000-0002-6200-511X

References

Aldous, D. (1983), *Exchangeability and Related Topics*, volume 1117 of Lecture Notes in Mathematics, Berlin: Springer. [3]

Ayed, F., and Caron, F. (2021), "Nonnegative Bayesian Nonparametric Factor Models with Completely Random Measures," *Statistics and Computing*, 31, Paper No. 63. [11]

Bacallado, S., Battiston, M., Favaro, S., and Trippa, L. (2017), "Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations," *Statistical Science*, 32, 487–500. [4,5]

Broderick, T., Wilson, A. C., and Jordan, M. I. (2018), "Posteriors, Conjugacy, and Exponential Families for Completely Random Measures," *Bernoulli*, 24, 3181–3221. [2,10]

Chakraborty, S., Arora, A., Begg, C. B., and Shen, R. (2019), "Using Somatic Variant Richness to Mine Signals from Rare Variants in the Cancer Genome," *Nature Communications*, 10, 5506–5509. [1,2,7]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [2]

Ferguson, T. S., and Klass, M. J. (1972), "A Representation of Independent Increment Processes without Gaussian Components," *Annals of Mathematical Statistics*, 43, 1634–1643. [3]

Goldstein, A. M., Xiao, Y., Sampson, J., Zhu, B., Rotunno, M., Bennett, H., Wen, Y., Jones, K., Vogt, A., and Burdette, L. (2017), "Rare Germline Variants in Known Melanoma Susceptibility Genes in Familial Melanoma," *Human Molecular Genetics*, 26, 4886–4895. [8]

Gravel, S. (2014), "Predicting Discovery Rates of Genomic Features," *Genetics*, 197, 601–610. [1,7]

Griffiths, T. L., and Ghahramani, Z. (2011), "The Indian Buffet Process: An Introduction and Review," *Journal of Machine Learning Research*, 12, 1185–1224. [10]

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019), "Ultrarare Variants Drive Substantial CIS Heritability of Human Gene Expression," *Nature Genetics*, 51, 1349–1355. [6]

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., Conti, D. V., Qu, C., Jeon, J., and Edlund, C. K. (2019), "Discovery of Common and Rare Genetic Risk Variants for Colorectal Cancer," *Nature Genetics*, 51, 76–87. [2,7]

Ionita-Laza, I., and Laird, N. M. (2010), "On the Optimal Design of Genetic Variant Discovery Studies," *Statistical Applications in Genetics and Molecular Biology*, 9, 1–17. [2]

Ionita-Laza, I., Lange, C., and Laird, N. M. (2009), "Estimating the Number of Unseen Variants in the Human Genome," *Proceedings of the National Academy of Sciences*, 106, 5008–5013. [1,6]

James, L. F. (2017), "Bayesian Poisson Calculus for Latent Feature Modeling via Generalized Indian Buffet Process Priors," *The Annals of Statistics*, 45, 2016–2045. [2,3,10]

James, L. F., Orbanz, P., and Teh, Y. W. (2015), "Scaled Subordinators and Generalizations of the Indian Buffet Process," arXiv preprint arXiv:1510.07309. [2,3,4,5,10]

Johansson, Ö, Samelius, G., Wikberg, E., Chapron, G., Mishra, C., and Low, M. (2020), "Identification Errors in Camera-Trap Studies Result in Systematic Population Overestimation," *Scientific Reports*, 10, 1–10. [2]

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., and Birnbaum, D. P. (2020), "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans," *Nature*, 581, 434–443. [7]

Kingman, J. (1992), *Poisson Processes*. Oxford Studies in Probability, Oxford: Clarendon Press. [2,3]

Kingman, J. F. (1975), "Random Discrete Distributions," *Journal of the Royal Statistical Society*, Series B, 37, 1–15. [4]

Knowles, D., and Ghahramani, Z. (2011), "Nonparametric Bayesian Sparse Factor Models with Application to Gene Expression Modeling," *The Annals of Applied Statistics*, 5, 1534–1552. [11]

Lawrenson, K., Kar, S., McCue, K., Kuchenbaeker, K., Michailidou, K., Tyrer, J., Beesley, J., Ramus, S. J., Li, Q., and Delgado, M. K. (2016), "Functional Mechanisms Underlying Pleiotropic Risk Alleles at the 19p13. 1 Breast–Ovarian Cancer Susceptibility Locus," *Nature Communications*, 7, 1–22. [8]

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., de Villiers, C. B., Izquierdo, A., Simard, J., and Schmidt, M. K. (2019), "Boadicea: A Comprehensive Breast Cancer Risk Prediction Model Incorporating Genetic and Nongenetic Risk Factors," *Genetics in Medicine*, 21, 1708–1718. [8]

Lijoi, A., Mena, R. H., and Prünster, I. (2007), "Bayesian Nonparametric Estimation of the Probability of Discovering New Species," *Biometrika*, 94, 769–786. [2]

Liu, Y., Xia, J., McKay, J., Tsavachidis, S., Xiao, X., Spitz, M. R., Cheng, C., Byun, J., Hong, W., and Li, Y. (2021), "Rare Deleterious Germline Variants and Risk of Lung Cancer," *NPJ Precision Oncology*, 5, 1–12. [8]

Masoero, L., Camerlenghi, F., Favaro, S., and Broderick, T. (2021), "More for Less: Predicting and Maximizing Genomic Variant Discovery via Bayesian Nonparametrics," *Biometrika*, 109, 17–32. doi: 10.1093/biomet/asab012. [1,2,3,6,7,9,10]

Momozawa, Y., and Mizukami, K. (2020), "Unique Roles of Rare Variants in the Genetics of Complex Diseases in Humans," *Journal of Human Genetics*, 66, 11–23. [2,6]

Nguyen-Dumont, T., MacInnis, R. J., Steen, J. A., Theys, D., Tsimiklis, H., Hammet, F., Mahmoodi, M., Pope, B. J., Park, D. J., and Mahmood, K. (2020), "Rare Germline Genetic Variants and Risk of Aggressive Prostate Cancer," *International Journal of Cancer*, 147, 2142–2149. [8]

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016), "Optimal Prediction of the Number of Unseen Species," *Proceedings of the National Academy of Sciences*, 113, 13283–13288. [1]

Phelan, C. M., Kuchenbaecker, K. B., Tyrer, J. P., Kar, S. P., Lawrenson, K., Winham, S. J., Dennis, J., Pirie, A., Riggan, M. J., and Chornokur, G. (2017), "Identification of 12 New Susceptibility Loci for Different Histotypes of Epithelial Ovarian Cancer," *Nature Genetics*, 49, 680–691. [8]

Pitman, J. (1996), "Some Developments of the Blackwell-MacQueen urn Scheme," in *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes Monograph Series*, eds. T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, pp. 245–267, Hayward, CA: Institute of Mathematical Statistics. [3]

——— (2006), *Combinatorial Stochastic Processes*, volume 1875 of Lecture Notes in Mathematics, Berlin: Springer. [4]

Pitman, J., and Yor, M. (1997), "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator," *Annals of Probability*, 25, 855–900. [2]

Rashkin, S., Jun, G., Chen, S., Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), and Abecasis, G. R. (2017), "Genetics, and E. of Colorectal Cancer Consortium. Optimal Sequencing Strategies for Identifying Disease-Associated Singletons," *PLoS Genetics*, 13, e1006811. [7,9]

Rasnic, R., Linial, N., and Linial, M. (2020), "Expanding Cancer Predisposition Genes with Ultra-Rare Cancer-Exclusive Human Variations," *Scientific Reports*, 10, 1–9. [8]

Regazzini, E. (1978), "Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilià," *Giornale dell'Istituto Italiano degli Attuari*, 41, 77–89. [5]

Sanders, J. G., Nurk, S., Salido, R. A., Minich, J., Xu, Z. Z., Zhu, Q., Martino, C., Fedarko, M., Arthur, T. D., and Chen, F. (2019), "Optimizing Sequencing Protocols for Leaderboard Metagenomics by Combining Long and Short Reads," *Genome Biology*, 20, 1–14. [2]

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J., Camps, C., and Pentony, M. M. (2020), "The Complete Costs of Genome Sequencing: A Microcosting Study in Cancer and Rare Diseases from a Single Center in the United Kingdom," *Genetics in Medicine*, 22, 85–94. [2]

Souza, C. A., Murphy, N., Villacorta-Rath, C., Woodings, L. N., Ilyushkina, I., Hernandez, C. E., Green, B. S., Bell, J. J., and Strugnell, J. M. (2017), "Efficiency of ddrad Target Enriched Sequencing across Spiny Rock Lobster Species (palinuridae: Jasus)," *Scientific Reports*, 7, 1–14. [2]

Wendt, C., and Margolin, S. (2019), "Identifying Breast Cancer Susceptibility Genes–A Review of the Genetic Background in Familial Breast Cancer," *Acta Oncologica*, 58, 135–146. [8]

Zabell, S. (2005), *The Continuum of Inductive Methods Revisited*, Cambridge: Cambridge University Press. [5]

Zhang, M. J., Ntranos, V., and Tse, D. (2020), "Determining Sequencing Depth in a Single-Cell RNA-Seq Experiment," *Nature Communications*, 11, 1–11. [2]

Zou, J., Valiant, G., Valiant, P., Karczewski, K., Chan, S. O., Samocha, K., Lek, M., Sunyaev, S., Daly, M., and MacArthur, D. G. (2016), "Quantifying Unobserved Protein-Coding Variants in Human Populations Provides a Roadmap for Large-Scale Sequencing Projects," *Nature Communications*, 7, 13293. [1,2,6,7,9,10]