# Genomic variety estimation with Bayesian nonparametric hierarchies

by

Lorenzo Masoero

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019



Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
January 10, 2019

Signature redacted

Certified by .....
Tamara Broderick
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Theses

# Genomic variety estimation with Bayesian nonparametric hierarchies

by

Lorenzo Masoero

Submitted to the Department of Electrical Engineering and Computer Science
on January 10, 2019, in partial fulfillment of the
requirements for the degree of
Masters of Science in Computer Science and Engineering

## Abstract

The recent availability of large genomic studies, with tens of thousands of observations, opens up the intriguing possibility to investigate and understand the effect of rare genetic variants in biological human evolution as well as their impact in the developement of rare diseases. To do so, it is imperative to develop a statistical framework to assess what fraction of the overall variation present in human genome is not yet captured by available datasets.
In this thesis we introduce a novel and rigorous methodology to estimate how many new variants are yet to be observed in the context of genomic projects using a nonparametric Bayesian hierarchical approach, which allows to perform prediction tasks which jointly handle multiple subpopulations at the same time. Moreover, our method performs well on extremely small as well as very large datasets, a desirable property given the variability in size of available datasets. As a byproduct of the Bayesian formulation, our estimation procedure also naturally provides uncertainty quantification of the estimates produced.

Thesis Supervisor: Tamara Broderick
Title: Assistant Professor of Electrical Engineering and Computer Science

# Contents

# List of Figures

# Chapter 1

# Introduction

The recent availability of large genomic studies, with tens of thousands of observations, opens up the intriguing possibility to investigate and understand the effect of rare genetic variants in biological human evolution as well as their impact in the developement of rare diseases. To do so, it is imperative to develop a statistical framework to assess what fraction of the overall variation present in human genome is not yet captured by available datasets. In this thesis we introduce a novel and rigorous methodology to estimate how many new variants are yet to be observed in the context of genomic projects using a nonparametric Bayesian hierarchical approach, which allows to perform prediction tasks which jointly handle multiple subpopulations at the same time.

In genomics, in the context of genetic variation discovery projects [Consortium, 2010, 2015, Eichler et al., 2007, Lek et al., 2016], scientists and practitioners have access to large datasets which contain information about larger and larger regions of the human genome. The ultimate goal of the studies is to produce freely available and high quality datasets, readily accessible to the scientific community, which provide an exhaustive map of the human genome, identifying and characterizing as much genetic information as possible present across different human populations.. This should in turn foster the understanding of how different genetic sequences impact the development of diseases and biological evolution [Ionita-Laza and Laird, 2010].

One way of describing each individual sequence, the statistical unit of the study or

*allele*, is through a binary vector encoding, in which a given coordinate denotes the presence or absence of variation with respect to a fixed, known underlying sequence, the "reference" genome. It is well established that structural variation in the genome, i.e. variation from the underlying reference genome, is a crucial factor for the understanding of biological human evolution, as well as for the development of rare Mendelian diseases [Bamshad et al., 2011, Li and Durbin, 2011, MacArthur et al., 2014, Stankiewicz and Lupski, 2010]. Historically, despite existing theories on the role of *rare* variants [Pritchard, 2001], empirical extensive study of genetic patterns and their relation to diseases had been limited to understanding the effect of *common* variants, i.e. those variants that appear relatively frequently in the observations (typically, variants which appear at rate greater than 5%). Indeed, until recently, available datasets were characterized by extremely limited sample sizes due to the prohibitive cost of genome sequencing, which made any attempt of studying *rare* variants hopeless. Recent technological breakthroughs, however, have made high-throughput DNA sequencing technologies largely available, allowing the sequencing of unprecented amounts of individuals. The 1000 Genomes Project [1KPG] [Consortium, 2010, 2015] was the first collective effort made by the scientific community to provide free access to a detailed catalogue of human genetic variation, and included genomic sequences from $N = 2{,}504$ individuals. More recently, the Exome Sequencing Project [ESP] [Fu et al., 2013] and Exome Aggregation Consortium [ExAC] [Lek et al., 2016] further scaled up the effort, making genomic sequences from $N = 60{,}076$ individuals freely available.

A crucial problem when working with such studies, is to develop a quantitative framework to reliably understand how much information about the underlying population is provided by the datasets used. Concretely, since the eventual goal is to characterize human genome in its entirety, we are interested in understanding how much variation in the genome is yet to be seen, and similarly, due to the costs associated to sequencing procedures, how many new variants would be observed if $M$ new alleles were added to the cohort of $N$ existing observations. It should be also noted that in these studies, the observations are typically subdivided into groups reflecting the different geographical origins of the individuals sequenced. Geographical factors are well known to have

an impact on the structure of genetic variations [Li et al., 2008, Novembre et al., 2008], and understanding how genomic variation differs across different populations is another key factor of interest.

Recent work has considered these and related problems: in particular, Ionita-Laza et al. [2009] proposed a parametric method to estimate the amount of yet unseen variants in the human genome, while, more recently, Zou et al. [2016] developed an algorithm that relies on a linear program to learn a histogram of the frequencies of all variants present in the human genome which can be used to estimate the same quantity. Despite their usefulsness, these methods suffer from substantial drawbacks and limitations. First of all, they don't provide a principled way to jointly handle data from multiple populations, in the sense that observations from different groups need to be either merged into a unique super-group, or considered separately. Moreover, as we show experimentally, the method of Ionita-Laza et al. [2009] struggles in the presence of large datasets in which the vast majority of variants are extremely rare. Modern datasets, in which the number of observations is in the order of the thousands, and the number of observations showing a given variant is characterized by a power-law behavior, make the approach of Ionita-Laza et al. [2009] inadequate. Viceversa, the method of Zou et al. [2016] can be inadequate for datasets of modest sample size, since it requires to specify as an input an arbitrary threshold whose optimal value depends on the sample size, and whose misspecification can drastically affect the quality of prediction when the sample size is small. While the sizes of available datasets are generally growing, obtaining genomic sequences from certain geographical regions or from patients associated with certain clinical conditions remains challenging, making the small data regime particularly relevant. Last, neither of these methods provides any form of direct uncertainty quantification for the estimates produced, and therefore bootstrapping is required in order to obtain uncertainty estimates.

In this thesis we develop a generic framework for the problem of estimating the variation in genomic sequences with the goal of addressing the limitations suffered from the existing methods. We do so by developing a hierarchical Bayesian nonparametric method. Bayesian methods stand out for their ability to formulate hierarchical models

which allow to share strength and statistical power across multiple datasets: we develop a model that can handle any number of subpopulations within a dataset jointly. Moreover, differently from the approaches of Ionita-Laza et al. [2009], Zou et al. [2016], we show by means of extensive synthetic experiments that the procedure we design works well both on datasets with large sample sizes, comparable to that of the ExAC dataset [Lek et al., 2016], with a vast proportion of extremely rare frequencies, as well as on small datasets, in which the majority of rare frequencies is unlikely to be observed. The nonparametric assumption, which entails that an ever growing number of features are going to be observed as new observations are collected, provides a convenient approximation. For any realistic sample size, indeed, it is reasonable to assume not to be nearly as close to have observed the "true" overall number of variants in the population. Therefore, it is reasonable to assume that the number of unique variants observed in a dataset keeps on growing with the number of datapoints observed. Last, uncertainty quantification naturally follows from the Bayesian formulation, in which quantities of interest are expressed in terms of posterior predictive distributions, naturally carrying a notion of uncertainty. In practice, this is useful as it allows to provide point estimates as well as credible intervals for the quantities of interest.

The rest of this manuscript is organized as follows: in Chapter 2 we provide a mathematical formulation of the problem of feature diversity and develop a suitable hierarchical Bayesian nonparametric framework. In Chapter 3 we prove the fundamental properties of the model, and use them to derive suitable estimators for the problem under consideration. In Chapter 4 we analyze practical aspects for the implementation of the prediction procedure developed. We conclude with experiments in Chapter 5, where we also investigate the critical aspects of the hierarchical method proposed, and discuss potential directions for improvements. Proofs and more details on the methods under investigations are deferred to the Appendix.

# Chapter 2

# Modeling genomic variation with nonparametric Bayesian methods

Genomic sequencing technologies allow to determine complete DNA sequences of an organism's genome. Roughly speaking, DNA sequencing is achieved by determining the order in which nuclotides and bases appear in the DNA. Recently released datasets, such as the ones of Consortium [2015], Lek et al. [2016], have been generated using modern high throughput sequencing technologies, which produce several reads of each individual genome. The multiple reads are aligned and mapped back to a fixed, underlying reference genome [Schbath et al., 2012]. Whenever the reads disagree with the reference genome, we say that a *variant* is observed. Genomic variation is quantified by an exhaustive comparison and analysis of the disalignment of an individual's genomic sequence from the underlying reference genome. While several different kinds of variation from the reference genome can be observed, e.g. specific deletions, inversions, transolcations or insertions of certain parts of the sequence [Feuk et al., 2006], we here ignore the differences between different forms of variation, and simply consider if variation from the reference genome is observed at a given *locus* in an individual's genome is observed.

Therefore, we imagine that the output of a sequencing procedure is a binary matrix $\boldsymbol{X} \in \{0,1\}^{N \times K_N}$. $N$ here denotes the number of *alleles*, i.e. the individual samples in the cohort under study, and $K_N$ denotes the number of *active* variants among the

$N$ observations, i.e. the number of specific variations from the underlying reference genome observed among the $N$ individuals sequenced.[1] Notice that at each individual location in the genome, possibly more than one variant could be observed. For simplicity, as done in Zou et al. [2016], we ignore the fact that the same locus can be associated with multiple variations, i.e. we ignore the fact that specific variants are mutually exclusive. It should be noted that this approximation might have some quantitative impact for the problem under consideration, since approximately 10% of the variants observed happen at loci where multiple variants are observed.

The kind of data we deal with has therefore a peculiar combinatorial structure: each observation – a row $X_n$ of the matrix $\boldsymbol{X}$ – is a binary sequence of length $K_N$, where the $k$-th coordinate $X_{n,k}$ of the sequence denotes the presence or absence of a variant at the genomic locus associated with the $k$-th column. This kind of structure is closely related to the concept of *feature allocation*. Formally, given a finite or countable set of $N$ objects $\Pi = \{\pi_1, \dots, \pi_N\}$, a feature allocation of $\Pi$ is a collection $\mathfrak{F} = \{F_1, \dots, F_{K_N}\}$ of subsets of $\Pi$, i.e. satisfying $F_k \subseteq \Pi$ for each $k \in [K_N] := \{1, \dots, K_N\}$, $K_N \in \{1, 2, \dots\}$, such that each object $\pi_n$ belongs to finitely many subsets, i.e. $\forall n, |\{k : \pi_n \in F_k\}| < \infty$. In our example, we can think of $\Pi$ as the set of all genomic sequences, with the genomic sequence of individual $n$ represented by the element $\pi_n \in \Pi$. Each set $F_k$ contains those genomic sequences which show variation at a given position $k$. Therefore, each binary matrix $\boldsymbol{X}$ uniquely defines a feature allocation.[2]

**Remark 1 (Relatives of feature allocations: partitions and trait allocations)** Feature allocations arise as a natural extension of partitions: a partition of a set $\Pi = \{\pi_1, \dots, \pi_N\}$ is a collection $\mathfrak{E} = \{E_1, \dots, E_{L_N}\}$ of subsets of $\Pi$, such that $E_l \subseteq \Pi$ for every $l$, $\cup_l E_l = \Pi$ and $E_l \cap E_{l'} = \emptyset$ for any $l \neq l'$. Models based on partitions have

[1] In practice, we only have access to datasets which report aggregated quantities, such as the column-wise sum of the matrix $\boldsymbol{X}$. We explain how we deal with this practical issue in Chapter 5.

[2] Notice that, while each binary matrix $\boldsymbol{X}$ uniquely identifies a feature allocation, the converse is not true. For example, by swapping two columns of $\boldsymbol{X}$, one still obtains the exact same feature allocation induced by the original matrix. A feature allocation is therefore represented by an equivalence class with respect to a function on binary matrices which is invariant to the ordering of the columns (see Griffiths and Ghahramani [2011], Section 4.2 for a thorough discussion of this point and the choice of such a function).

a long and successful history in statistics (see Hartigan [1990], Quintana and Iglesias [2003] for an overview). Probabilistic models for partitions have been particularly successful in the Bayesian nonparametric literature, in which, following the seminal paper of Ferguson [1973], a huge number of models based on the Dirichlet process and its generalization have been developed. These have successfully been applied in a large number of applications, and their properties extensively studied (see Ghosal and Van der Vaart [2017] for an introduction and review of nonparametric Bayesian methods and their application in partition models, and Pitman [2006] for extensive treatment of their theoretical properties).

Models which extend the notion of feature allocations to combinatorial structures in which each datapoint is allowed to belong to multiple groups with arbitrary degree of belonging have also been considered. These are usually referred to as trait allocation models [Campbell et al., 2018]. Formally, a trait allocation $\mathfrak{T}$ of the set $\Pi$ is a collection $\mathfrak{T} := \{T_1, \ldots, T_{Z_N}\}$ such that for each $z \in [Z_N]$, $T_z$ is a multiset whose distinct elements are elements of $\Pi$. The cardinality $c_{z,n} = |T_z \cap \{\pi_n\}|$ gives the degree of belonging of the element $\pi_n$ to the multiset $T_z$. As for the case of feature allocations, also a trait allocations must satisfy that each datapoint $\pi_n$ is assigned to finitely many traits, i.e. $\forall n, |\{z : \pi_n \in T_z\}| < \infty$.

## 2.1 A probabilistic Bayesian model for genetic variation

Following the observations made in Chapter 1 and leveraging the representation of genomic variants in terms of collections of binary vectors as described above, we now want to introduce a Bayesian nonparametric model for feature allocations, which requires specifying a generative model for observed data. This amounts to defining a joint probability distribution over the (observed) data and an unknown and random parameter governing the data generating process. In the problem under consideration, the data is given by $\boldsymbol{X} = X_1, \ldots, X_N$, where each $X_n$ is the binary vector encoding,

representing individual's $n$ genomic sequence. The parameter $\theta$ governs the probability that each entry of each binary vector is equal to one or to zero.

Mathematically, the Bayesian formulation can be broken down into two steps: first we identify a parameter space, $\Xi$, together with a prior distribution over possible values for the parameter, $p(\theta) for \theta \in \Xi$. Next, we define a likelihood model, i.e. a probability distribution for the data, conditionally on the parameter, $p(X_n \mid \theta)$. We assume throughout that the datapoints are obtained conditionally i.i.d. given the random parameter. This can be written as

$$\theta \sim p(\theta), \quad X_n \mid \theta \overset{\text{i.i.d.}}{\sim} p(X_n \mid \theta) \tag{2.1}$$

Bayes' rule then allows to derive the posterior distribution of the random parameter of interest:

$$p(\theta|\boldsymbol{X}) = \frac{p(\theta)p(\boldsymbol{X} \mid \theta)}{\int_\Theta p(\boldsymbol{X} \mid \theta)p(\theta)\mathrm{d}\theta} \tag{2.2}$$

The model we pick should match the desiderata outlined before: it should be amenable to a hierarchical formulation and flexible enough to capture the underlying generative structure in presence of scarce data while retaining desirable computational properties when the data grows large.

### 2.1.1 Poisson point processes and completely random measures

In this section we introduce the fundamental building blocks to define a nonparametric Bayesian model for feature allocations. The natural way to define a prior distribution which can accomodate an evergrowing number of points in the support, is to consider a nonparametric prior. We focus on a distinguished class of models which stand out for their tractability and flexibility, characterized by prior distributions of a very special nature, known in the probabilstic literature as completely random measures [CRMs]. These are a class of discrete random measures which have successfully been used as

prior distributions in the context of Bayesian nonparametric inference [Broderick et al., 2018, Campbell et al., In press, James, 2017, Lijoi and Prünster, 2010] and are tightly linked to Poisson point processes.

**Poisson point processes [PPPs]**: PPPs are a very broad class of stochastic point processes whose theoretical properties have been extensively studied (see Kingman [1992] for a detailed treatment). Due to their many interpretable properties and their overall analytical tractability, these processes have been used extensively in the statistics and machine learning communities to model a huge variety of real world phenomena (see, e.g., Lewis et al. [2012], Mohler et al. [2011], Papapantoleon [2008]). Informally, a Poisson point process on some space $\mathcal{S}$ is a stochastic process whose realizations are collections of points randomly distributed in $\mathcal{S}$. A draw from a Poisson process has the distinctive property that, given two disjoint subsets of the state space, the counts of the number of points in each subset are independent Poisson random variables. We now provide a very general measure-theoretic definition of these processes, following Kingman [1992].

**Definition 1 (Poisson process)** *Let $\mathcal{S}$ be a state space of interest endowed with a $\sigma$-algebra $\mathcal{B}$ such that the every set containing just the singleton $\{s\} \in \mathcal{S}$ is measurable. A Poisson Process on $\mathcal{S}$ is a random countable set $\Pi \subset \mathcal{S}$ such that*

***(P1)*** *Given measurable and disjoint $B_1, \ldots, B_K \in \mathcal{B}$ the random variables*

$$N(B_k) := |\Pi \cap B_k|, \tag{2.3}$$

*for $k = 1, \ldots, K$, are independent.*

***(P2)*** *For any $B \in \mathcal{B}$, $N(B) \sim \text{Pois}(\mu(B))$, where $\mu : \mathcal{B} \to \mathbb{R}_+$ is a measure.*

*We denote a draw from such a process as $\Pi \sim \text{PP}(\mu)$.*

Hence, a draw $\Pi \sim \text{PP}(\mu)$ almost surely yields a collection of points $\Pi$. From Definition 1, we see that if $\mu(\mathcal{S}) = \infty$, $|\Pi| = \infty$ with probability one, whereas if $\mu(\mathcal{S}) < \infty$, $|\Pi| < \infty$ almost surely. Moreover, for any $B \in \mathcal{B}$, if $\mu(B) = \infty$, $N(B) = \infty$

almost surely, while if $\mu(B) < \infty$, $N(B) < \infty$ with probability one and specifically, $N(B) \sim \text{Pois}(\mu(B))$ and if $\mu(B) = 0$, then $N(B) = 0$.

**Completely Random Measures [CRMs]**: Recall that a random measure is nothing but a random-valued measure, i.e. a random set function which takes non-negative values and satisfies the property of countable additivity. A completely random measure $\Theta$ on a measurable space $(\Psi, \mathcal{B})$, is simply a random measure such that for any finite collection of disjoint measurable subsets $\{B_1, \ldots, B_K\} \subset \mathcal{B}$, the random variables $\Theta(B_k), \Theta(B_{k'})$ are independent for $k \neq k'$.

It is a trivial observation that Poisson processes can be used to produce CRMs. Consider the following construction: let one of the axes of the state space be the positive real line, i.e. $\mathcal{S} = \mathbb{R}_+ \times \Psi$, for some space $\Psi$ of interest. Assume also to pair this space with a suitable product $\sigma$-algebra. It is clear that by drawing a Poisson process on $\mathcal{S}$ we obtain a random countable set $\Theta = \{\theta_k, \psi_k\}_{k \geq 1} \subset \mathbb{R}_+ \times \Psi$. For each point in $\mathcal{S}$, by treating the coordinate $\theta_k$ as a random rate, and the coordinate $\psi_k$ as a random location, one can produce a random measure on the space $\Psi$.

Kingman [1967] showed that CRMs and Poisson processes are intimately related: in particular, a CRM can always be described as the sum of three components,

$$\Theta(\cdot) = \Theta_{\text{det}}(\cdot) + \Theta_{\text{fix}}(\cdot) + \Theta_{\text{ord}}(\cdot) \tag{2.4}$$

where

- $\Theta_{\text{det}}(\cdot)$ is a non-random measure on $\Psi$.
- $\Theta_{\text{fix}}(\cdot)$ is a discrete measure on $\Psi$ with support at deterministic locations $\{\psi_k^{(\text{fix})}\}_{k=1}^{L_f}$, and random and independent weights $\{\theta_k^{(\text{fix})}\}_{k=1}^{L_f}$.
- $\Theta_{\text{ord}}(\cdot)$ arises from a Poisson point process on $\mathbb{R}_+ \times \Psi$ as described in the previous subsections, i.e. it is a random subset of random cardinality of the form $\{\theta_k^{(\text{ord})}, \psi_k^{(\text{ord})}\}_{k=1}^{L_o}$.

For our purposes, we shall always assume that $\Theta_{\text{det}}(\cdot) = 0$. The fixed component, $\Theta_{\text{fix}}(\cdot)$, will play an important role in obtaining posterior representations of the random

measure $\Theta(\cdot)$, while the ordinary component $\Theta_{\text{ord}}(\cdot)$ will be the main focus of most of our analysis.

**Remark 2 (Normalized completely random measures)** Whenever the measure $\Theta(\cdot)$ is finite, i.e. $\Theta(\Psi) < \infty$, we can always consider its *normalization*,

$$\Xi(\cdot) := \frac{\Theta(\cdot)}{\Theta(\Psi)} \in [0, 1]. \tag{2.5}$$

$\Xi$ is a random probability measure on $\Psi$. Notice, however, that $\Xi$ is not a CRM: if $A, B \in \mathcal{B}$ are disjoint and $\Xi(A) > p$ for some $p > 0$, then necessarily $\Xi(B) < 1 - p$, that is to say the random weights of disjoint subsets is not independent.

### 2.1.2 BNP feature allocation models

We now have all the ingredients to introduce a framework for defining nonparametric Bayesian models for feature allocations which rely on CRM priors. From a generative standpoint, a feature allocation can be described in terms of allocation of datapoints to features. Specifically, we define generative models which, under suitable assumptions, produce countable collections of features and associated rates through the prior. Then, we "allocate" such features to datapoints through the likelihood.

**Remark 3** Following Broderick et al. [2018], James [2017], we adopt a very general framework, which can accomodate several combinatorial structures. Not only we can allow datapoints to be associated with multiple features, as in a feature allocation model (Section 2.2): the same machinery can be used to define partition models as well as trait allocation models. In partition models (see Appendix A.1) we require each datapoint to be associated to one and only one feature. Instead in trait allocation models we allow each datapoint to simultaneously belong to different features, each with different belonging degree, and we refer to such features as traits (see Appendix A.2).

The terms *traits* and *features* are often used interchangeably, even though it should be noticed that the word *feature* is originates in the context of binary-valued feature

allocation models, like the ones we consider here, while the word *trait* is typically used for generic integer-valued trait allocation models.

In full generality, the BNP models we consider are essentially made of two parts:

1. A model for a collection of tuples of traits with their rates. This is obtained in the prior via a CRM made only of an ordinary component. Given a feature space $\Psi$, together with a $\sigma$-algebra $\mathcal{B}$, the prior is almost surely a discrete random measure on $\Psi$ of the form

$$\Theta(\cdot) = \sum_k \theta_k \delta_{\psi_k}(\cdot). \tag{2.6}$$

   Since we have by assumption no fixed component, the prior is a Poisson process on $\mathbb{R}_+ \times \Psi$ with some mean measure $\mu(\mathrm{d}\theta \times \mathrm{d}\psi) = \nu(\mathrm{d}\theta) \times P_0(\mathrm{d}\psi)$, where $\nu(\mathrm{d}\theta) = \rho(\theta)\mathrm{d}\theta$ is a diffuse measure over $\mathbb{R}_+$ and $P_0(\mathrm{d}\psi)$ is a diffuse probability measure over $\Psi$, and write

$$\Theta \sim \mathrm{CRM}(\nu, P_0). \tag{2.7}$$

   The diffuseness assumption guarantees that all traits will almost surely be different.

2. A model for the allocation of each datapoint, conditionally on $\Theta$, to different traits via a likelihood process [LP]. The $n$-th individual is therefore described via a random measure on $\Psi$,

$$X_n(\cdot) = \sum_k x_{n,k} \delta_{\psi_k}(\cdot), \tag{2.8}$$

   in which the random weights $x_{n,k}$ are integer-valued. We obtain this measure by specifying a (conditional) distribution $H(\cdot \mid \theta)$ with a probability mass function $h(\cdot \mid \theta)$ supported at $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$ and by drawing each "count"

$x_{n,k} \mid \Theta \sim h(\cdot|\theta_k)$.We denote this as

$$X_n \mid \Theta \overset{\text{i.i.d.}}{\sim} \text{LP}(\Theta). \tag{2.9}$$

The fact that we allow for countably many traits in the prior is a distinctive property and benefit of using a nonparametric approach. In many real-world applications this property proves extremely useful, since we often don't know how many "active" features are going to be needed for the data at hand. For example, in the variant discovery task, we do not know beforehand how many "active" variants are going to be observed as the sample size increases: through a BNP approach we are not required to specify this quantity, or even just an upper bound beforehand. The model always allows us to discover new traits as more data is observed by instantiating new ones, making it a natural modeling framework for streaming and growing data.

**Remark 4** While we are interested in the setting in which a feature represents an *observed* variant in the genomic sequence, it should be noted that in most applications the likelihood process $X_n$ is typically a *latent* element of a more complex generative scheme. In these cases – in which we don't get to observe directly the likelihood processes, rather random functions of it – the goal is typically related to inferring the latent traits and the relevance of traits to each observation [Blei, 2012, Blei et al., 2003, Broderick et al., 2015, Ghahramani and Griffiths, 2006, Kemp et al., 2006, Zhou and Carin, 2015]. For example, in text analysis, we observe documents $Y_n$ and we model them through latent topics, the features of the model. To each document $Y_n$ we associate a latent binary vector $X_n$, in which the $k$-th component $x_{n,k}$ denotes the presence or absence of the latent topic $\psi_k$ in that specific document.

We now impose common assumptions on the prior $\Theta(\cdot)$ and on the likelihood processes $X_n$, which are needed in order for any BNP model to be useful and realistic (see Broderick et al. [2018], James [2017]).

**(A1)** *The prior supports a.s. countably many traits.* This assumption reflects our desire of specifying a *nonparametric* model. Mathematically, this is achieved

by imposing that the CRM is obtained via a Poisson process whose underlying rate-component measure $\nu$ is an infinite, but $\sigma$-finite measure. This implies that $\nu$ satisfies

$$\nu(\mathbb{R}_+) = +\infty, \tag{2.10}$$

and there exists a (finite or countable) measurable partition $\{A_k\}_{k\geq 1}$ of $\mathbb{R}_+$ for which

$$\nu(A_k) < \infty, \quad \forall k. \tag{2.11}$$

**(A2)** *Each datapoint is allocated to finitely many traits.* This assumption reflects the natural idea that we cannot associate an observation with infinitely many "attributes". In the genomic variation example, each genomic sequence can show variation only at finitely many loci of the genome. Mathematically, if we let $\nu_x(\mathrm{d}\theta) := \nu(\mathrm{d}\theta)h(x|\theta)$, this translates into

$$\sum_{x\geq 1} \nu_x(\mathbb{R}_+) < \infty. \tag{2.12}$$

## 2.2 Examples

With the previous definitions in place, we can readily obtain feature allocation models. This kind of approach has found widespread application in very diverse fields (see, for example, Chu et al. [2006], Doshi-Velez and Ghahramani [2009], Fox et al. [2009], Görür et al. [2006], Kemp et al. [2006], Lee et al. [2016], Miller et al. [2009, 2008], Navarro and Griffiths [2007], and Ghahramani et al. [2007], Griffiths and Ghahramani [2011] for a review).

**Example 1 (Beta-Bernoulli process)** *The beta-Bernouli process, first introduced by Hjort [1990] in the BNP literature for applications to survival analysis, is arguably the most popular BNP feature allocation model. This is obtained through the following*

*hierarchy:*

1. *a beta process prior,* $\Theta \sim \mathrm{BP}(\alpha; P_0)$*, i.e.* $\Theta = \sum_{k\geq 1} \theta_k \delta_{\psi_k}$*. From a CRM perspective, this is obtained through a completely random measure with an ordinary component characterized by the mean measure*

$$\mu(\mathrm{d}\theta \times \mathrm{d}\psi) = 1_{[0,1]}(\theta)\alpha\theta^{-1}(1-\theta)^{\alpha-1}\mathrm{d}\theta P_0(\mathrm{d}\psi), \tag{2.13}$$

   *where* $\alpha > 0$*.*

2. *a Bernoulli process likelihood, i.e.* $X_n \mid \Theta \overset{i.i.d.}{\sim} \mathrm{BeP}(\Theta)$*, where, for every* $n$ *and every* $k$*,* $x_{n,k} \mid \theta_k \overset{ind}{\sim} \mathrm{Ber}(\theta_k)$*.*

The theoretical properties of this model have been extensively studied [Broderick et al., 2012, Kim, 1999, Paisley and Carin, 2009, Thibaux and Jordan, 2007], and several extensions of the models have been considered. These include, but are not limited to, models which rely on alternative definitions of the beta process prior, to account for example for power-law type distributions of the features [Broderick et al., 2012, Teh and Gorur, 2009], or dependent versions which can take in consideration covariates [Gershman et al., 2015, Ren et al., 2011, Zhou et al., 2011].

**Remark 5 (Indian buffet process)** Once an ordering on the feature labels has been imposed, the outcome of the process described above can be organized in a binary random matrix $\boldsymbol{X} = [X_1 \dots X_N]^\top \in \{0,1\}^{N \times K_N}$, where $K_N$ counts the number of unique features observed across the $N$ observations. While the beta-Bernoulli construction expresses the binary draws $x_{n,k}$ via a conditional distribution, it is also possible to describe the *marginal* distribution of the random matrix $\boldsymbol{X}$, i.e. unconditionally on the underlying beta process $\Theta$. The access to the marginal distribution of this random matrix allows to describe a sequential, urn-alike, construction – known in the literature as the Indian buffet process [IBP]. Using the well known food metaphor, each row of the random matrix is associated to a customer, and each column of the matrix is associated with a dish. To obtain a draw from an IBP with parameter $\alpha$,

- the first customer tries Pois($\alpha$) dishes. This gives the count of nonzero entries in the first row of the matrix.

- subsequently, customer $n$ tries

  - dish $k$ with probability $\frac{m_k}{n}$, where $m_k$ is the number of previous customers who tried dish $k$

  - Pois($\alpha/n$) new dishes

## 2.3 An extension to hierarchies

While the models introduced in Section 2.1 form a vast and flexible set of tools for statistical analysis, often a more complex modeling framework is desired. For instance, the practical problems under consideration often have an inherent hierarchical structure. Concretely, we might have access to multiple collections of documents – such as different scientific journals or different periodicals. Or we might wish to analyze genetic information across multiple populations of individuals. Rather than naively merge all of these datasets together or analyze them separately, Bayesian hierarchical modeling is famous for allowing sharing of power across different datasets while maintaining their idiosyncrasies.

Hierarchies are well-developed in the case of partition models, where each data point belongs to one and only one group [Dunson, 2009, Lijoi et al., 2005, 2007, MacEachern, 2000, Teh and Jordan, 2010, Teh et al., 2005, Yau and Holmes, 2011]. In this section we introduce a general class of BNP *hierarchies* for feature and trait allocations models, which include partition and feature allocation models as a special case, with the goal of providing a general formalism for the hierarchical case, like the one developed by Broderick et al. [2018], James [2017] for non-hierarchical BNP trait models. In particular, we imagine that all traits are shared across each population but also that the traits occur with different rates within each population. For example the topic "winter sports" might appear both in a general-interest newspaper and in a sports magazine, but it may occur more frequently in the latter than the former. Similarly,

the same ancestral groups may be present across individuals who e.g. currently live in different countries, with the incidence of individual groups differing across different countries.

Assume we have $D$ distinct sets of observations or sub-populations, where the $d$ - th sub-population has $N_d$ individuals, $d \in [D] := \{1, \ldots, D\}$. First we draw a shared CRM base measure from the prior, $\Theta_0 \sim \text{CRM}(\nu_0, P_0)$. Next, given $\Theta_0$, we draw conditionally independent random measures $\Theta_d \mid \Theta_0 = \sum_k \theta_{d,k}\delta_{\psi_k}$, where for all $d \in [D]$, $\theta_{d,k} \mid \Theta_0 \sim^{ind} \rho_d(\cdot \mid \theta_{0,k}, r_d)$. Let $r_d$ be a fixed hyperparameter and $\rho_d$ be a density over $\mathbb{R}_+$. We then obtain a measure-valued random vector $\Theta_{1:D} := [\Theta_1 \ldots \Theta_D]$, in which the traits are shared but the rates differ across the components of the vector, and denote it as $\Theta_{1:D} \mid \Theta_0 \sim \text{hCRM}(\rho_d(\cdot|\theta_{0,k}, r_d)_{k\geq 1, d\in[D]})$. The hyperparameters $\{r_d\}_d$ can be used, for example, to control the variance of each sub-population from the base measure $\Theta_0$ (see Remark 6). In the context of BNP models, this hierarchical construction has been considered in the specific case of the beta process by Thibaux and Jordan [2007] (see Example 2). Instead, some hierarchical processes like the hierarchical Dirichlet process [Teh et al., 2005] – which are not hierarchical completely random measures – are obtained through a different scheme, in which at higher levels of the hierarchy the same trait can be assigned to multiple random weights (using the food analogy, in the Chinese restaurant franchise, the same dish can appear at multiple tables within the same restaurant). This is not allowed in our construction.

Just like in the models described in Section 2.1, to complete the construction we pair the hCRMs with likelihoods that allocate datapoints to traits. For each $d \in [D]$, we fix a probability mass function $h_d(\cdot \mid \theta_{d,k}, s_d)$, where $s_d$ is a fixed, sub-population specific hyperparameter with support on $\mathbb{Z}_+ \cup \{0\}$. We consider latent trait counts $x_{n,d,k} \mid \theta_{d,k} \sim h_d(\cdot|\theta_{d,k}, s_d)$ conditionally i.i.d. across $n \in [N_d]$ and conditionally independent across $d \in [D]$. The trait allocation can then be represented through the counting measure $X_{n,d} = \sum_{k\geq 1} x_{n,d,k}\delta_{\psi_k}$, which we denote as $X_{n,d} \mid \Theta_d \sim \text{LP}(\Theta_d, h_d, s_d)$. The role of the sub-population specific hyperparameter $s_d$ is similar to that of the parameter $r_d$, and can be used to model the relationship between the base measure $\Theta_d$ and the likelihood processes $\{X_{n,d}\}_{n\in[N_d]}$.

The counts $\{x_{n,d,k}\}$ are independent random variables conditionally on the vector of random measures $\Theta_{1:D}$, i.e., are modeled as partially exchangeable random variables conditionally on the vector of hCRMs. This reflects the idea that *within* each sub-population, observations are homogeneous and we can treat them as exchangeable, while *across* different sub-populations, the exchangeability assumption is not preserved. For example, if we have a collection of corpora of documents from different scientific journals, articles from the same journal would be mutually exchangeable, but articles from different journals would not. Similarly, if we had access to genetic data from multiple populations, individuals within the same population would be treated as exchangeable, but individuals from different population would not.

Let $\pi_d(\theta_{d,k}) := h_d(0|\theta_{d,k}, s_d)$ be the probability that a trait $\psi_k$ with rate $\theta_{d,k}$ in population $d$ is not part of a trait allocation for some datapoint. We impose

$$(\mathbf{A3}) \quad \int_{\mathbb{R}_+ \times \Psi} \int_{\mathbb{R}_+} (1 - \pi_d(t)) \rho_d(t \mid s, r_d) \mathrm{d}t \nu_0(\mathrm{d}s) P_0(\mathrm{d}\psi) < \infty, \tag{2.14}$$

$\forall\, d \in [D]$. This technical assumption is needed to ensure that each observation is associated a.s. with finitely many traits, equivalently, $\sum_k \mathbb{1}(x_{n,k,d} > 0) < \infty$ a.s. (see Lemma 3 for a simple proof of this fact). Because the traits are typically *latent* within the generative model, we complete the model by specifying a probabilty distribution $f$ from which each observed value $Y_{n,d}$ is drawn conditionally i.i.d. given the latent counting measure $X_{n,d}$. To sum up, the hierarchical construction we consider, for $d \in [D], n \in [N_d]$, is

$$\begin{aligned} \Theta_0 &\sim \mathrm{CRM}(\nu_0, P_0) \\ \Theta_{1:D} \mid \Theta_0 &\sim \mathrm{hCRM}(\{\rho_d(\cdot|\theta_{0,k}, r_d)\}_{k,d}) \\ X_{n,d} \mid \Theta_{1:D} &\sim \mathrm{LP}(\Theta_d, h_d, s_d) \\ Y_{n,d} \mid X_{n,d} &\sim f(\cdot \mid X_{n,d}) \end{aligned} \tag{2.15}$$

We present here as an example a natural extension to their hierarchical counterpart of the model previously introduced in Example 1. We omit the last layer of the hierarchy (the observations, $Y$), since this is arbitrary and can always be specified without

affecting the underlying combinatorial structure.

**Example 2 (Hierarchical beta-Bernoulli process (hBB))** This model, originally introduced in Thibaux and Jordan [2007], generalizes Example 1 to the hierarchical case.

$$\begin{aligned} \Theta_0 &\sim \mathrm{BP}(\alpha, P_0) \\ \Theta_{1:D} \mid \Theta_0 &\sim \mathrm{hCRM}(\{\mathrm{Beta}(r_d\theta_{0,k}, r_d(1-\theta_{0,k})\}_{k\geq 1, d=1,\ldots,D}) \\ X_{n,d} \mid \Theta_d &\overset{\text{i.i.d.}}{\sim} \mathrm{BeP}(\Theta_d) \end{aligned} \tag{2.16}$$

**Remark 6** The choices of the hyperparameters $r_d$, which regulate the relationship between the shared CRM base measure $\Theta_0$ and the population-level base measure $\Theta_d$, and of the hyperparameters $s_d$, which regulate the allocation process of datapoints to traits, can play a fundamental role in the modeling of different phenomena.
For example, in the hBB of Equation (2.16), we have fixed $s_d = 1$ for all $d$, while the parameters $r_d$ serve to modulate the variance of the rates of each measure $\Theta_d$. Indeed, for $\Theta_d = \sum_k \theta_{d,k}\delta_{\psi_k}$, we see that

$$\theta_{d,k} \mid \Theta_0 \sim \mathrm{Beta}(r_d\theta_{0,k}, r_d(1-\theta_{0,k})), \tag{2.17}$$

*i.e.*

$$\mathbb{E}[\theta_{d,k}] = \theta_{0,k}, \mathrm{Var}(\theta_{d,k}) = \frac{1}{r_d+1}(1-\theta_{0,k})\theta_{0,k} \tag{2.18}$$

# Chapter 3

# Theoretical results

In this section we present two novel results which characterize the properties of the hierarchical CRMs introduced in Section 2.3. In particular, in Lemma 1 we provide the joint distribution of the counts induced by any hCRM which fits in the previous framework. This is the crucial tool to characterize the posterior distribution of any hCRM, as done in Lemma 2.

In what follows, assume $N := N_1 + \cdots + N_D$ observations are given across the $D$ populations. Let $\Psi^* := \{\psi_1^*, \ldots, \psi_{K_N}^*\} \subset \Psi$ denote the set of "active traits" among the $N$ observations, i.e., traits to which some data point belongs, and let $\boldsymbol{X}^* = \{X_{n,d}^*\}_{n,d}$ denote the associated assignments of data points to traits. These are counting measures of the form $X_{n,d}^* = \sum_k x_{n,d,k}^* \delta_{\psi_k^*}^*$. For $n \in [N_d], d \in [D], k \in [K_N]$, let $\boldsymbol{x}_{n,d}^* = [x_{n,d,1}^* \ldots x_{n,d,K_N}^*]$ denote the counts of the active traits and $\mathfrak{I}_{d,k} := \{n \in [N_d] : x_{n,d,k}^* := X_{n,d}(\psi_k^*) > 0\}$ the observations in population $d$ sharing trait $\psi_k^*$, with $m_{d,k} := |\mathfrak{I}_{d,k}|$. We now derive the joint distribution of the trait allocations and the locations, which describes the distribution of unique traits appearing across multiple populations.

**Lemma 1 (Characterization of the distribution of unique features)** *Consider the hierarchical CRM model of Equation* (2.15). *Let* $\mathcal{A} = \{A_1, \ldots, A_{K_N}\} \subseteq \mathcal{B}$ *and let* $P_{K_N}(\mathcal{A}; \mathfrak{I}, \boldsymbol{x}^*)$ *be the probability to observe exactly* $K_N$ *distinct features* $\psi_1^*, \ldots, \psi_{K_N}^*$, *each* $\psi_k^* \in A_k$ *and inducing the index set* $\mathfrak{I}$ *and observation counts* $\boldsymbol{x}^*$.

*For any $K_N \geq 1$, the probability distribution $P_{K_N}(\cdot; \mathfrak{I}, \boldsymbol{x}^*)$ is absolutely continuous with respect to $P_0^{\otimes K_N}(\cdot)$ with Radon-Nikodym derivative independent of the traits $\{\psi_k^*\}_k$ given by*

$$p_{K_N}(\psi_1^*, \ldots, \psi_{K_N}^*; \mathfrak{I}, \boldsymbol{x}^*) = p_{K_N}(\mathfrak{I}, \boldsymbol{x}^*) = e^{-\Phi(N_1,\ldots,N_D)} \times \tag{3.1}$$
$$\prod_{k=1}^{K_N} \left[ \int_{\mathbb{R}_+} \prod_{d=1}^{D} \int_{\mathbb{R}_+} \pi_d(t)^{N_d - m_{d,k}} \left( \prod_{n \in \mathfrak{I}_{d,k}} h_d(x_{n,d,k}^* \mid t, s_d) \rho_d(t \mid s, r_d) \mathrm{d}t \nu(\mathrm{d}s) \right) \right],$$

*with* $\Phi(N_1, \ldots, N_D) = \int_{\mathbb{R}_+} \left[ 1 - \prod_{d=1}^{D} \int_{\mathbb{R}_+} \pi_d(t)^{N_d} \rho(t \mid s, r_d) \mathrm{d}t \right] \nu(\mathrm{d}s)$.

We now obtain the posterior distribution of the hierarchical random vector $\Theta_{1:D}$.

**Lemma 2 (Characterization of the posterior distribution)** *Consider a general hCRM model, as defined in Equation* (2.15) *and let* $\boldsymbol{X}^* = \{X^*_{1:N,1:D,1:K_N}\}$ *be the integer value array of feature counts, in which we stack all observations from the different populations, and* $\boldsymbol{\xi}_0^* = (\xi^*_{0,1:K_N})$ *be a vector of latent jumps. The posterior distribution of the hierarchical vector* $\Theta_{1:D}$ *is given by*

$$\Theta_1 \ldots, \Theta_D \mid \boldsymbol{X}^*, \boldsymbol{\xi}_0^* \stackrel{d}{=} \left[ \Theta'_1 \quad \ldots \quad \Theta'_D \right] + \left[ \sum_{k=1}^{K_N} \xi^*_{1,k} \delta_{\psi_k^*} \quad \ldots \quad \sum_{k=1}^{K_N} \xi^*_{D,k} \delta_{\psi_k^*} \right] \tag{3.2}$$

*where*

- *the updated vector of hCRMs* $[\Theta'_1 \ldots \Theta'_D]$ *is obtained through the hierarchy*

$$\begin{aligned} \Theta'_0 &\sim \mathrm{CRM}(\nu'_0, P_0) \stackrel{d}{=} \sum_k \theta'_{0,k} \delta_{\psi_k} \\ \Theta'_1 \ldots \Theta'_D \mid \Theta'_0 &\sim \mathrm{hCRM}(\rho'_d, \Theta'_0) \stackrel{a.s.}{=} \sum_k \theta'_{d,k} \delta_{\psi_k}, \end{aligned} \tag{3.3}$$

  *with*

$$\nu'_0(\mathrm{d}\theta) = \prod_{d=1}^{D} \int_0^{\infty} \pi_d(t)^{N_d} \rho_d(t \mid \theta, r_d) \mathrm{d}t \nu_0(\mathrm{d}\theta)$$

  *and*

$$\theta'_{d,k} \mid \Theta'_0 \sim^{ind} \rho'_d(\cdot \mid \theta'_{0,k}, r_d) \propto \pi_d(\theta'_{d,k})^{N_d} \rho_d(\theta'_{d,k} \mid \theta'_{0,k}, r_d)$$

- *the collection* $(\xi^*_{d,k})_{d,k}$ *is made of independent jumps given the vector* $\boldsymbol{\xi}^*_0$*, with*

$$\xi^*_{d,k} \mid \xi^*_{0,k} \;\sim\; (\pi_d(\xi^*_{d,k}))^{N_d - m_{d,k}} \prod_{n \in \mathfrak{I}_{d,k}} h_d(x^*_{n,d,k} \mid \xi^*_{d,k}, s_d)\rho_d(\xi^*_{d,k} \mid \xi^*_{0,k}, r_d)\mathrm{d}\xi^*_{d,k}. \tag{3.4}$$

**Remark 7** It is also possible to characterize the posterior distribution of the latent jumps $\boldsymbol{\xi}^*_0$, conditionally on the array of counts $\boldsymbol{x}^*$. This is given by

$$p(\boldsymbol{\xi}^*_0 \mid \boldsymbol{x}^*) \propto \prod_{k=1}^{K_N} \left( \prod_{d=1}^{D} \int_{\mathbb{R}_+} \pi(t)^{N_d - m_{d.k}} \prod_{n \in \mathfrak{I}_{d,k}} h(x^*_{n,d,k} \mid t, s_d)\rho(t \mid \xi^*_{0,k}, r_d)\mathrm{d}\theta_{d,k} \right) \nu_0(\mathrm{d}\xi^*_{0,k}).$$

We can interpret the two elements in the posterior characterization of Equation (3.2) as follows:

- The vector $\Theta'_{1:D}$ provides the contribution to the posterior from the prior, and is a realization of an hCRM with "underweighted" shared based measure $\nu'_0(\mathrm{d}\theta) = \prod_d \ell_d(\theta)\nu_0(\mathrm{d}\theta)$ –where the downweighting factor is given by $\prod_{d=1}^D \ell_d(\theta)$, with
$$\ell_d(\theta) := \int_{\mathbb{R}_+} \pi_d(t)^{N_d}\rho_d(t \mid \theta, r_d)\mathrm{d}t \in (0,1).$$
- The vector $[\sum_{k=1}^{K_N} \xi^*_{1,k}\delta_{\psi^*_k} \cdots \sum_{k=1}^{K_N} \xi^*_{D,k}\delta_{\psi^*_k}]$ provides the contribution to the posterior from the observations, conditionally on the (latent) jumps $\boldsymbol{\xi}^*_0$. Each component of the vector is a random measure supported at the observed locations $\Psi^*$, whose distribution depends on the counts $x^*_{n,d,k}$, where $n = 1, \ldots, N_d$, and $k \geq 1$. For each trait $\psi^*_k$, datapoints $n \notin \mathfrak{I}_{d,k}$ – for which $x^*_{n,d,k} = 0$ – shift the distribution of $\xi^*_{d,k}$ towards 0. Viceversa, datapoints $n \in \mathfrak{I}_{d,k}$, shift the distribution of $\xi^*_{d,k}$ away from 0.

We now provide examples of the distributions derived in Lemma 1 and Lemma 2 for the model introduced before.

**Example 3 (Hierarchical 3beta-Bernoulli (h3BB))** *Consider the three-parameters generalization of the hierarchical beta process introduced in Example 2. This is to say,*

*replace the common base measure of Equation* (2.16) *with* $\Theta_0 \sim \mathrm{CRM}(\nu_0, P_0)$, *where* $\nu_0$ *is the measure over* $[0,1]$ *given by*

$$\nu_0(\mathrm{d}\theta) \;=\; \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \theta^{-(1+\sigma)}(1-\theta)^{c+\sigma-1} \mathbb{1}_{[0,1]}(\theta)\mathrm{d}\theta \tag{3.5}$$

*for* $\alpha > 0$, $\sigma \in (0,1)$ *and* $c > -\sigma$. *Let* $(a)_b := \Gamma(a+b)/\Gamma(a)$ *denote the ascending factorial of* $a$ *of order* $b$. *From Lemma 1,*

$$\begin{aligned} p_{K_N}(\mathfrak{I}, \boldsymbol{x}^*) = \exp\left\{ -\int_0^1 \left( 1 - \prod_{d=1}^{D} \frac{(r_d(1-\theta))_{N_d}}{(r_d)_{N_d}} \right) \nu_0(\mathrm{d}\theta) \right\} \\ \prod_{k=1}^{K_N} \int_0^1 \prod_{d=1}^{D} \frac{(r_d t)_{m_{d,k}} (r_d(1-t))_{N_d - m_{d,k}}}{(r_d)_{N_d}} \nu_0(\mathrm{d}t). \end{aligned} \tag{3.6}$$

*Equation* (3.6) *provides a generalization of the exchangeable feature probability function (EFPF) of the beta-Bernoulli process [Broderick et al., 2013] to the hierarchical, partially exchangeable setting. We can apply Lemma 2 to the h3BB case considered before. The updated rate measure of* $\Theta_0'$ *is given by*

$$\nu_0'(\mathrm{d}\theta) = \prod_{d=1}^{D} \frac{(r_d(1-\theta))_{N_d}}{(r_d)_{N_d}} \nu_0(\mathrm{d}\theta).$$

*The rates are conditionally independent with distribution*

$$\theta'_{d,k} \mid \Theta_0' \overset{ind}{\sim} \mathrm{Beta}(r_d\theta'_{0,k}, N_d + r_d(1-\theta'_{0,k})).$$

*Moreover, each jump* $\xi^*_{d,k} \mid \xi^*_{0,k} \sim \mathrm{Beta}(m_{d,k} + \xi^*_{0,k} r_d, N_d - m_{d,k} + r_d(1-\xi^*_{0,k}))$, *and the distribution of* $\xi^*_{0,k}$ *is given by*

$$p(\xi^*_{0,k} \mid \boldsymbol{x}^*) \propto (r_d \xi^*_{0,k})_{m_{d,k}} (r_d(1-\xi^*_{0,k}))_{N_d - m_{d,k}} (\xi^*_{0,k})^{-(1+\sigma)} (1-\xi^*_{0,k})^{c+\sigma-1}.$$

## 3.1 Estimation of diversity in single-population feature models

We now make use of results and machinery from the previous sections to derive estimators to quantify genomic variation in the case of a single population. Later, we extend the results to the hierarchical setting.

As described before, we assume to observe a binary matrix $\boldsymbol{X}^* = [x^*_{n,k}]_{n,k} \in \{0,1\}^{N \times K_N}$ of genomic sequences, where for every $n \in [N]$, $k \in [K_N]$, $x^*_{n,k} = 1$ if individual $n$ shows variation at locus $k$. Hence, $K_N$ denotes the number of distinct features appearing among the first $N$ observations. We denote with $\Psi^*_N = \{\psi^*_1, \ldots, \psi^*_{K_N}\} \subset \Psi$ the set of observed features up to step $N$, that is the set of locations in $\Psi$ for which there exists at least one index $n$ that assigns positive mass $x^*_{n,k} = 1$ for that trait. We can think each row $X^*_n$ as a binary counting measure on the feature space $\Psi$ of genomic loci,

$$X^*_n = \sum_{k=1}^{K_N} x^*_{n,k} \delta_{\psi^*_k}.$$

We model $X^*_n$ as a draw from the generative model of Example 1, where we replace the prior process $\Theta$ with its 3-parameters generalization introduced in Equation (3.5). In the following theorems, let $\boldsymbol{X}^* := X^*_1, \ldots, X^*_N$ denote the *observed* samples, and for any $M \geq 1$, let $\boldsymbol{X}' := X'_{N+1}, \ldots, X'_{N+M}$ denote $M$ additional samples, observed after the $N$ initial ones. Moreover, let $z_{N,k} := \sum_{n=1}^{N} x^*_{n,k}$.

**Theorem 1 (Number of old features in an additional sample)** *For every $k \in [K_N]$, let $O^{(M)}_{N,k}$ be the integer valued random variable which counts the number of additional samples which display feature $\psi^*_k$ when $M$ additional samples $\boldsymbol{X}'$ are provided after the $N$ initial ones $\boldsymbol{X}^*$, i.e.,*

$$O^{(M)}_{N,k} \mid \boldsymbol{X}^* \stackrel{d}{=} \sum_{m=1}^{M} x'_{N+m,k}, \tag{3.7}$$

*where the $x'_{N+m,k}$ are conditional independent random variables which appear in the posterior representation of eaxh $X'_{N+m}$.*

*For any* $M, N \in \mathbb{Z}_+, k \in [K_N], l \in \mathbb{Z}_+ \cup \{0\}$,

$$\mathbb{P}(O_{N,k}^{(M)} = l \mid \boldsymbol{X}^*) = \binom{M}{l} \frac{(z_{N,k} - \sigma)_l (N - z_{N,k} + c + \sigma)_{M-l}}{(N + c)_M}. \tag{3.8}$$

Another quantity of interest in several applications is the number of features which have not yet been observed in the sample.

**Theorem 2 (Number of new features in an additional sample)** *Let* $U_N^{(M)}$ *denote the random variable counting the number of yet unseen features which will be observed if* $M$ *additional samples* $\boldsymbol{X}'$ *are provided after the* $N$ *inital ones* $\boldsymbol{X}^*$, *i.e.*

$$U_N^{(M)} = \sum_{k \geq 1} \mathbb{1}\left( \sum_{m=1}^{M} x'_{n+m,k} > 0 \right) \mathbb{1}(z_{N,k} = 0). \tag{3.9}$$

*For any* $M, N \in \mathbb{Z}_+$, *it holds*

$$U_N^{(M)} \mid \boldsymbol{X}^* \sim \text{Pois}\left( \alpha \sum_{m=1}^{M} \frac{(c + \sigma)_{N+m-1}}{(c+1)_{N+m-1}} \right). \tag{3.10}$$

**Corollary 1 (Number of features appearing with a given frequency)** *We could also be interested in investigating how many features with a given frequency are going to appear in an additional sample, i.e. for some* $r \in \mathbb{Z}_+$, *we might be interested in the random variable* $U_{N,r}^{(M)}$ *defined as*

$$U_{N,r}^{(M)} = \sum_{k \geq 1} \mathbb{1}\left( \sum_{m=1}^{M} x'_{n+m,k} = r \right) \mathbb{1}(z_{N,k} = 0). \tag{3.11}$$

*Using the previous result, for any* $M, N \in \mathbb{Z}_+, r \leq M$, *it holds*

$$U_{N,r}^{(M)} \mid \boldsymbol{X}^* \sim \text{Pois}\left( \alpha \binom{M}{r} \frac{(1 - \sigma)_{r-1} (c + \sigma)_{N+M-r}}{(c+1)_{N+M-1}} \right). \tag{3.12}$$

## 3.2 Estimation of diversity for the multiple populations case

Assume to observe $D$ collections of samples, each containing $N_d$ observations for $d \in [D]$, drawn from the generative model of Example 3. This is to say the data is $\boldsymbol{X}^* = \{\boldsymbol{X}_1^*, \dots, \boldsymbol{X}_D^*\}$, where for every $d \in [D]$, $\boldsymbol{X}_d^* = \{X_{1,d}^* \dots, X_{N_d,d}^*\}$ is a collection of binary counting measures on the feature space $\Psi$, of the form $X_{n,d}^* = \sum_k x_{n,d,k}^* \delta_{\psi_k^*}$, as described before. We let $\boldsymbol{N} = [N_1 \dots N_D]$ be the vector of cardinalities of the $D$ populations, and $\boldsymbol{z}_{N,k} := \sum_d \sum_{n=1}^{N_d} x_{n,d,k}^*$, $z_{N_d,d,k} = \sum_{n=1}^{N_d} x_{n,d,k}^*$. Moreover, we let $N := N_1 + \cdots + N_D$ be the total number of observations, and $K_N$ denote the number of distinct features which appear among all the $D$ populations in the first $N$ observations. We let $\Psi*$ be the set of all observed atoms in $\Psi$ across the $D$ populations. We will denote with $M_d \in \mathbb{Z}_+$ the number of new samples observed in each population $d \in [D]$, and let $\boldsymbol{M} = [M_1 \dots M_D]$, $M = M_1 + \cdots + M_D$.

**Theorem 3 (Number of old features across multiple populations)** *Let $\psi_k^* \in \Psi$ be an observed feature among the first $N$ samples observed across the $D$ populations. Assume to have observed $N_d$ samples from population $d$ of which exactly $z_{d,N_d,k} \leq N_d$ displayed feature $\psi_k^*$. Now, assume additional $M_d$ samples are further observed from the same population. Let $O_{d,N_d,k}^{(M)}$ be the integer valued random variable which counts the number of samples among the $M_d$ additional ones which display feature $\psi_k^*$, i.e.*

$$O_{d,N_d,k}^{(M)} \mid \boldsymbol{X}_{\boldsymbol{N}^*} = \sum_{m=1}^{M} x'_{d,N_d+m,k}. \tag{3.13}$$

*For any $M_d, N_d \in \mathbb{Z}_+, k \in K_{\boldsymbol{N}}, l \in \mathbb{Z}_+ \cup \{0\}$,*

$$\begin{aligned} &\mathbb{P}\left(O_{d,N_d,k}^{(M)} = l \mid \boldsymbol{X}_{\boldsymbol{N}^*}\right) \\ &\quad = \frac{M_d!}{(M_d - l)!} \int_{[0,1]} \frac{(z_{d,N_d,k} + r_d\theta)_l (N_d - z_{d,N_d,k} + r_d(1-\theta))_{M_d - l}}{(N_d + r_d)_{M_d}} \nu_{3\mathrm{BP}}(\mathrm{d}\theta; \alpha, c, \sigma). \end{aligned} \tag{3.14}$$

**Theorem 4 (Number of new features across multiple populations)** *Let $U_{\boldsymbol{N}}^{(M)}$*

*be the integer valued random variable which counts how many features not yet seen in any of the $D$ populations are going to be observed if $M_d$ additional samples are observed for each $d \in [D]$, i.e.*

$$U_{\boldsymbol{N}}^{(\boldsymbol{M})} = \sum_{k \geq 1} \mathbb{1}\left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} x'_{d,N_d+m,k} > 0 \right) \mathbb{1}(\boldsymbol{z}_{N,k} = 0). \tag{3.15}$$

*The posterior predictive distribution of $U_{\boldsymbol{N}}^{(\boldsymbol{M})}$ is given by a Poisson distribution with parameter*

$$\begin{aligned} \upsilon_{\boldsymbol{N}}^{(\boldsymbol{M})} \mid \boldsymbol{X}_N^\star = \int_0^1 \Bigg\{ \Bigg( 1 - \prod_{d=1}^{D} \frac{\boldsymbol{B}(r_d\theta, N_d + M_d + r_d(1-\theta))}{\boldsymbol{B}(r_d\theta, N_d + r_d(1-\theta))} \Bigg) \\ \prod_{d=1}^{D} \left[ \frac{\boldsymbol{B}(r_d\theta, N_d + r_d(1-\theta))}{\boldsymbol{B}(r_d\theta, r_d(1-\theta))} \right] \nu_{\mathrm{3BP}}(\mathrm{d}\theta; \alpha, \sigma, c) \Bigg\}. \end{aligned} \tag{3.16}$$

**Corollary 2 (Number of features shared by two populations)** *Another interesting quantity to study is the number of features which are not observed in the initial $N$ samples, but are shared across multiple populations when new samples are introduced. Let $\mathfrak{J} \subset [D]$ be a collections of indices. Let $S_{\boldsymbol{N},\mathfrak{J}}^{(\boldsymbol{M})}$ be the integer valued random variable which counts how many features have not been observed among the first $N$ samples, but are observed and shared across (all) populations whose index is in $\mathfrak{J}$, i.e.*

$$S_{\boldsymbol{N},\mathfrak{J}}^{(\boldsymbol{M})} = \sum_{k \geq 1} \left( \prod_{i \in \mathfrak{J}} \mathbb{1}\left( \sum_{m=1}^{M_i} x'_{i,N_i+m,k} > 0 \right) \mathbb{1}(\boldsymbol{z}_{N,k} = 0) \right). \tag{3.17}$$

*The posterior predictive distribution of $S_{\boldsymbol{N},\mathfrak{J}}^{(\boldsymbol{M})}$ is given by a Poisson distribution with parameter*

$$\begin{aligned} \tau_{\boldsymbol{N},\mathfrak{J}}^{(\boldsymbol{M})} \mid \boldsymbol{X}_N^\star = \int_0^1 \Bigg\{ \prod_{i \in \mathfrak{J}} \Bigg( 1 - \frac{\boldsymbol{B}(r_i\theta, N_i + M_i + r_i(1-\theta))}{\boldsymbol{B}(r_i\theta, N_i + r_i(1-\theta))} \Bigg) \\ \prod_{d=1}^{D} \left[ \frac{\boldsymbol{B}(r_d\theta, N_d + r_d(1-\theta))}{\boldsymbol{B}(r_d\theta, r_d(1-\theta))} \right] \nu_{\mathrm{3BP}}(\mathrm{d}\theta; \alpha, \sigma, c) \Bigg\}. \end{aligned} \tag{3.18}$$

# Chapter 4

# Optimization

The estimators derived in Chapter 3 are all expressed in terms of posterior predictive distributions. As a direct consequence of the generative process chosen, they are all Poisson distributions whose parameters depends on the underlying parameters of the beta processe – $\alpha, \sigma, c$. Notably, none of these posterior predictive distributions depend *directly* on sample information, but only on the hyperparameters of the process.
In order to make practical use of the estimators described above, we therefore need an effective way of esimating these hyperparameters. A reasonable quantity to look for, is a likelihood criterion associated with the observations $\boldsymbol{X}^*$ under the true generative model.

## 4.1 Optimization in the single population case

In the single population case, where we have access to $\boldsymbol{X}^* \in \{0,1\}^{N \times K_N}$ – e.g. under the generative model outlined in Example 1 or in the case in which the prior is a CRM whose rates follow Equation (3.5) – the likelihood of the assignments is given by the exchangeable feature probability function [EFPF].

**Definition 2 (EFPF of the 3BB)** *Let $\boldsymbol{X}^* \in \{0,1\}^{N \times K_N}$ be a draw from Example 1, where we replaced the beta process prior with its three parameter generalization of Equation* (3.5).

*Let* $z_{N,k} := \sum_{n=1}^{N} x_{n,k}^*$ *for* $k = 1, \ldots, K_N$. *Then the EFPF takes the form*

$$\ell_{\alpha,\sigma,c}(\boldsymbol{X}^*) = \frac{1}{K_N!}\left(\frac{\alpha}{(c+1)_{N-1}}\right)^{K_N} \times \tag{4.1}$$
$$\times \exp\left\{-\alpha \sum_{n=1}^{N} \frac{(\sigma+c)_{n-1}}{(1+c)_{n-1}}\right\} \prod_{k=1}^{K_N} (1-\sigma)_{m_k-1}(c+\sigma)_{N-z_{N,k}},$$

*where* $(x)_s := \Gamma(x+s)/\Gamma(x) = x(x+1)\ldots(x+s-1)$ *is the rising factorial.*

The EFPF, or any monotonic transformation of it – such as its logarithm – can therefore be used as an objective function to find values of $\alpha, \sigma, c$. One option is to employ a fully Bayesian approach, i.e. specify priors $p_\alpha(\alpha), p_\sigma(\sigma), p_c(c)$ for the hyperparameters, and consider meaningful posterior values which arise from some inference scheme (e.g., MCMC) on these parameters, such as posterior means or posterior medians. Another approach is to use an emprical Bayes approach, and directly maximize the parameters from the data, i.e. find

$$\{\alpha^*, \sigma^*, c^*\} = \underset{\alpha\in\mathbb{R}_+, \sigma\in[0,1), c\in(-\sigma,\infty)}{\arg\max} \log(\ell_{\alpha,\sigma,c}(\boldsymbol{X}^*)). \tag{4.2}$$

We pursue the latter and notice that it is possible to compute in closed form the partial derivatives of $\log(\ell_{\alpha,\sigma,c})$ with respect of all the three parameters. These are given by

$$\frac{\partial}{\partial\alpha}\log(\ell_{\alpha,c,\sigma}(\boldsymbol{X}^*)) = \frac{K_N}{\alpha} - \sum_{n=1}^{N} \frac{(\sigma+1)_{n-1}}{(c+1)_{n-1}}, \tag{4.3}$$

$$\begin{aligned}\frac{\partial}{\partial c}\log(\ell_{\alpha,c,\sigma}(\boldsymbol{X}^*)) &= -K_N(\psi^{(0)}(c+N) - \psi^{(0)}(c+1) + \psi^{(0)}(c+\sigma)) \\ &\quad - \alpha \sum_{n=1}^{N} \frac{(\sigma+c)_{n-1}}{(1+c)_{n-1}} \times (\psi^{(0)}(\sigma+c+n-1) \\ &\quad + \psi^{(0)}(c+n) - \psi^{(0)}(\sigma+c) - \psi^{(0)}(c+1)),\end{aligned} \tag{4.4}$$

$$\begin{aligned}\frac{\partial}{\partial\sigma}\log(\ell_{\alpha,c,\sigma}(\boldsymbol{X}^*)) = -\alpha\sum_{n=1}^{N}\frac{(\sigma+c)_{n-1}}{(1+c)_{n-1}}(\psi^{(0)}(\sigma+c+n-1)-\psi^{(0)}(\sigma+c)) \\ +\sum_{k=1}K(\psi^{(0)}(m_k-\sigma)+\psi^{(0)}(c+\sigma+N-z_{N,k}) \\ -\psi^{(0)}(1-\sigma)-\psi^{(0)}(c+\sigma)),\end{aligned} \tag{4.5}$$

where $\psi^{(0)}(x) := \frac{d}{dx}\ln(\Gamma(x))$ is the digamma function. We can use these derivatives in a descent method to optimize Equation (4.1) and produce the estimators obtained in Chapter 3.

## 4.2 Optimization in the multiple population case

In the multiple population setting, the observations are given by collections of binary matrices. Under the generative model described in Example 3, Equation (3.6) is the counterpart of the EFPF defined in Equation (4.1). In principle a similar strategy to the one described in Section 4.1 could be pursued in order to maximize the parameters of the hierarchical process. In practice, the hierarchical EFPF (Equation (4.1)) is a complicated expression, expensive to evaluate even for a moderate value of distinct features $K_N$, since for each $k \in [K_N]$ an integral needs to be analytically computed. Morever, for the application under consideration, even for modest sample sizes, the number of distinct features is typically in the order of the hundreds of thousands or in the millions.

We notice that we can simplify the expression by considering ordered $D$-tuples $\boldsymbol{m} = [m_1,\ldots,m_D] \in \mathbb{Z}_+^D$. Indeed, if two features $k_1, k_2 \in [K_N]$ are such that $m_{d,k_1} = m_{d,k_2}$ for all $d \in D$, their contribution in the product of Equation (3.6) is the same. We can therefore re write the hierarchical EFPF as

$$\begin{aligned}p_{K_N}(\mathfrak{I},\boldsymbol{x}^*) = \exp\left\{-\int_0^1\left(1-\prod_{d=1}^{D}\frac{(r_d(1-\theta))_{N_d}}{(r_d)_{N_d}}\right)\nu_0(\mathrm{d}\theta)\right\} \\ \prod_{\boldsymbol{m}_k\in\mathfrak{M}}\left(\int_0^1\prod_{d=1}^{D}\frac{(r_d\theta)_{m_d}(r_d(1-\theta))_{N_d-m_d}}{(r_d)_{N_d}}\nu_0(\mathrm{d}\theta)\right)^{n(\boldsymbol{m}_k)},\end{aligned} \tag{4.6}$$

where $\mathfrak{M}$ is the collection of ordered cardinalities $\boldsymbol{m}_k = [m_{1,k} \dots m_{D,k}] \in \mathbb{Z}^D$ such that there exists at least one feature $k \in [K_N]$ observed exactly $m_{d,k}$ times in all populations, i.e. for all $d \in [D]$. $n(\boldsymbol{m}_k)$ counts how many features are observed exactly $m_{d,k}$ times in each population $d = 1, \dots, D$.

Despite this simplification, even the simple evalutation of Equation (4.6) for data like the ExAC of Lek et al. [2016], in which when the number of observations $N$ is in the order of the hundreds the number of distinct variants $K_N$ is in the order of the millions, is very computationally intensive.

## 4.3 Description of competing methods

We here provide a brief review of the alternative existing methods for genomic variety estimation that we consider in the experiments of Chapter 5. Neither of these methods is directly applicable to the hierarchical case, hence we will only describe estimators for the single population formulation of the problem.

### 4.3.1 Beta-Bernoulli product model [Ionita-Laza et al., 2009]

Ionita-Laza et al. [2009] considers the same problem of genomic variation described in Chapter 1. As usual, we are given data $\boldsymbol{X}^* \in \{0,1\}^{N \times K_N}$. The authors model the binary matrix $\boldsymbol{X}^*$ via a parametric beta-Bernoulli model. They assume that there exists a fixed, unknown number $K_\infty < \infty$ of possible variants. For each $k \in [K_\infty]$, they assume that the $k$-th feature is displayed by any observation with probability $\theta_k \in [0,1]$ distributed according to a beta distribution with parameters $a$, $b$, i.e.

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \dots & \theta_{K_\infty} \end{bmatrix}, \quad \text{with} \quad \theta_k \overset{\text{i.i.d.}}{\sim} \text{Beta}(a,b)\ \forall k, \tag{4.7}$$

and, conditionally on $\boldsymbol{\theta}$,

$$X_n^* = \begin{bmatrix} x_{n,1}^* & \dots & x_{n,K_\infty}^* \end{bmatrix}, \quad \text{with} \quad x_{n,k}^* \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta_k). \tag{4.8}$$

Under this model, the number of counts $z_{N,k}|\theta_k := \sum_{n=1}^{N} x^*_{n,k} \sim \text{Binom}(N, \theta_k)$. Letting $\hat{f}_x^{(N)} = \sum_{k=1}^{K_\infty} \mathbb{1}(z_{N,k} = x)$, the log likelihood is given by

$$\ell_{a,b}^{\text{BBPM}}(\boldsymbol{X}^*) = \sum_{n=1}^{N} z_{N,n} \log(\lambda_{n,N}), \tag{4.9}$$

where

$$\lambda_{n,N} = \frac{\binom{n}{N} \boldsymbol{B}(n+a, N-n+b)}{\sum_{n'=1}^{N} \binom{n'}{N} \boldsymbol{B}(n'+a, N-n'+b)} \tag{4.10}$$

Letting $M = tN$ be the number of additional samples to be observed, we can compute the expected number of new variants to be observed in additional $M$ samples after $N$ samples have been observed as

$$\begin{aligned} \Delta_N(M) &= \mathbb{E}\left[\sum_{k=1}^{K_\infty} \mathbb{1}\left(\sum_{n'=1}^{tN} x_{n,k} > 0\right) \mathbb{1}\left(\sum_{n=1}^{N} x_{n,k} = 0\right)\right] \\ &= \frac{\eta_1}{a} \frac{N+b-1}{N} \left[1 - \frac{\boldsymbol{B}(a, (t+1)N+b)}{\boldsymbol{B}(a, N+b)}\right], \end{aligned} \tag{4.11}$$

where $\eta_1 := \mathbb{E}[f_N^{(1)}]$ is the expected number of features which appear exactly one time in a sample of size $N$. To use the estimator $\Delta_N(M)$, Ionita-Laza et al. [2009] substitute $\eta_1$ with its empirical counterpart $\hat{f}_1^{(N)}$, the number of features which have been observed once in the sample $\boldsymbol{X}^*$. Then, similarly to our approach, they find $a, b$ via maximization the of the log-likelihood of the model,

$$a^*, b^* = \arg\max_{a>0, b>0} \ell_{a,b}^{\text{BBPM}}(\boldsymbol{X}^*) \tag{4.12}$$

**Remark 8 (Rare features)** Notice that the estimator Equation (4.11) crucially relies on the empirical frequency of features observed once among the first $N$ draws, $\hat{f}_1^{(N)}$. For example, if a dataset had $\hat{f}_1^{(N)} = 0$, $\Delta_N(M) = 0$ for every $M > 0$.

### 4.3.2 Linear program to estimate the frequencies of frequencies [Zou et al., 2016]

Zou et al. [2016] formalize the problem of feature variety estimation as that of recovering the distribution of frequencies of all the genetic variation in the population, including those features which have not yet been observed.

They assume that each possible variant in a sample is independent of the other variants, and that the $k$-th variant appears with a given probability $p_k$ conditionally i.i.d. across the $N$ individuals - i.e. the $p_k$ are the parameters of independent Bernoulli random variables. Therefore the binary matrix $\boldsymbol{X}^* \in \{0,1\}^{N \times K_N}$ of genetic variation, is modeled by a collection of independent Bernoulli random variables, which are also identically distributed along each column, and the sum $S_k := \sum_{n=1}^{N} X_{n,k} \sim \text{Binom}(N, p_k)$. From the frequencies $S_1, \ldots, S_{K_N}$ of the $K_N$ variants observed in $N$ samples, it is possible to compute the fingerprint of the sample, $\mathcal{F} = [\mathcal{F}_1 \ldots \mathcal{F}_N]$, where $\mathcal{F}_i := \#\{k \in [K_N] : S_k = i\}$. Given the fingerprint $\mathcal{F}$, the goal is to recover the population's histogram, which is a map quantifying, for every $x \in [0,1]$, the number of variants $k$ such that $p_k = x$. Formally,

$$h_P : (0,1] \to \mathbb{N} \cup \{0\} \tag{4.13}$$

In particular, because the empirical frequencies associated to more common variants should be well approximated by their empirical counterpart, they only consider the problem of estimating the histogram from the truncated fingerprint $\mathcal{F}^{(\kappa)} = \{\mathcal{F}_i : \frac{i}{N} \leq \kappa\}$. In their analysis, the authors set $\kappa = 0.01$. They further set a discretization factor $\delta$, and then set up a linear program in which the goal is to correctly estimate the popoulation histogram associated to the frequencies in the set $\mathcal{S} = \left\{\frac{1}{1000N}, \delta\frac{1}{1000N}, \ldots, \delta^i\frac{1}{1000N}, \ldots, \kappa\right\}$, which determines how many frequencies are going to be estimated in $(0, \kappa]$. Formally, they solve the optimization

$$\min_{h(s), s \in \mathcal{S}} \sum_{i : \frac{i}{N} \leq \kappa} \frac{1}{1 + \mathcal{F}_i} \left| \mathcal{F}_i - \sum_{s \in \mathcal{S}} h(s) \text{Binom}(N, s, i) \right| \tag{4.14}$$

subject to

$$h(s) \geq 0, \sum_{s \in \mathcal{S}} h(s) \leq K^{(\max)}, \sum_{s \in \mathcal{S}} s \cdot h(s) + \sum_{i: i/N > \kappa}^{K_N} \frac{i}{N} \mathcal{F}_i = \frac{K_N}{N} \tag{4.15}$$

where $K^{(\max)}$ is an upper bound on the total number of variants, and $\text{Binom}(N, s, i)$ is the probability that a Binomial draw with bias $s$ and $N$ rounds is equal to $i$. Given the estimated histogram $\hat{h}$, one can obtain an estimate of the number of unique variants at any sample size $M$ using

$$V(\hat{h}, M) = \sum_{s: \hat{h}(s) > 0} \hat{h}(s)(1 - (1 - s)^M). \tag{4.16}$$

Follwoing Zou et al. [2016], we refer to this estimator as the "unseenEST" estimator.

# Chapter 5

# Experiments

In order to test the estimators derived in Chapter 3, we consider both synthetic data and real data, with the goal of understanding how the predictive performances of the estimators change across different data generating regimes.

First, we consider the estimators for the single-level case. In particular, we focus on the task of estimating the number of distinct features observed if an additional sample of size $M$ is provided in addition to the $N_{\text{train}}$ samples already gathered – i.e. we compare the estimator provided in Theorem 2 to Equation (4.11) and Equation (4.16). To maximize Equation (4.16), we used the code released by the authors, which is freely available[1]. While Zou et al. [2016] consider learning only the population histogram for frequencies which are below a fixed threshold of $\kappa = 1\%$, we noticed that choosing different thresholds sometimes provided dramatic improvements in the prediction quality. For all the experiments, we therefore fixed three different thresholding levels, $\kappa \in \{1\%, 5\%, 10\%\}$ and ran different optimizations, one for each value of $\kappa$. Fixing this threshold at the correct level could represent a practical challenge in real world applications. Specifically, Zou et al. [2016] do not provide any insight on how to specify $\kappa$ in different data analysis problems. The only generic and intuitive guideline one can follow in fixing the threshold $\kappa$ is that for larger samples a smaller $\kappa$ should be used, and viceversa for smaller samples a larger $\kappa$ is preferable. Especially when the sample size is relatively small with respect to the size of the support, however,

[1]The code is available at `https://github.com/jameszou/unseenest`.

smaller values of $\kappa$ can affect quite substantially the quality of the prediction (see Figure 5-3). It should also be emphasized that in our experiments we have access to the true counts whose growth we want to predict, but in any realistic scenario these counts would be unknown, making the choice of the optimal threshold and even just a fair comparison of different values of $\kappa$ hard. Moreover, we modified the discretization factor $\delta$ from the default value 1.05 to 1.01, which allows to learn a "finer" histogram, with more frequencies and better predictive performance, at the cost of an increase in computation time.

The major issues associated to the optimizations of the BNP and the beta-Bernoulli product model estimators are related to numerical stability. Specifically, both optimization problems require computing gamma and beta functions of very large inputs, which require careful implementation to avoid overflows and underflows. To solve both optimization problems, after making sure that the implemented likelihood functions did not incur in numerical stability issues, we combined a grid search over the parameter space with first order descent methods, using the best results from the grid search as initializations. While we were able to obtain reliable values for the maximization of the BNP estimator, we noticed that when the sample size $N_{\text{train}}$ is in the order of the hundreds, the beta-Bernoulli product model seems to provide less reliable results. In particular, for datasets in which the distribution of the frequencies of the variants has a power-law behavior, the beta-Bernoulli model and related estimator seems to be inadequate.

## 5.1 Synthetic datasets

We start by considering in Section 5.1.1 datasets drawn from the three parameters Indian buffet process – the "true" data generating process under which the estimators are derived. This serves as a baseline, to check that, in the best case scenario, in which the data generating process matches the assumptions under which the BNP estimator is derived, we are able to obtain a good predictive performance.

### 5.1.1 Draws from the 3-IBP

To generate draws from the 3-IBP, we consider the generative scheme proposed in Teh and Gorur [2009]. This is the extension of the IBP scheme described in Remark 5 to the power law-case[2].

Given parameters $\alpha \in \mathbb{R}_+, \sigma \in [0,1), c > -\alpha$, we describe how to generate a random matrix with $N$ rows from a 3-IBP$(\alpha, c, \sigma)$ using the well known food analogy. Each row of the random matrix is associated to a "customer" which enters in an Indian restaurant that has a buffet serving infinitely many dishes. Each column of the random matrix is associated to a distinct dish. Then, to obtain a random matrix with $N$ rows,

- the first customer tries $L_1 \sim \text{Pois}(\alpha)$ dishes. That is to say, the first row of the matrix has $L_1$ entries equal to 1, followed by infinitely many zeros.
- subsequently, for $n = 2, \ldots, N$, customer $n$ tries
  - dish $k$ with probability $\frac{z_{n-1,k}-\sigma}{n-1+c}$, where $z_{n-1,k}$ is the number of previous customers who tried dish $k$. That is to say, for any column $k$ for which a previous row has at least one non-zero entry, the entry $(n,k)$ will be equal to 1 with probability $\frac{z_{n-1,k}-\sigma}{n-1+c}$
  - $L_n \sim \text{Pois}\left(\alpha \frac{(c+\sigma)_{n-1}}{(c+1)_{n-1}}\right)$ new dishes. That is to say, the $n$-th row of the matrix instantiates $L_n$ new columns at which it displays value 1.

As already discussed in Teh and Gorur [2009], the mass parameter $\alpha$ controls the total number of columns of the matrix generated by the random process. Indeed, we expect $K_N = \sum_{n=1}^{N} \alpha/n$ unique features, or, equivalenlty, columns to be displayed. The concentration parameter $c$ controls how many rows will tend to display each feature, while $\sigma$ determines the power-law behavior, and in particular if $\sigma = 0$ no power-law behavior is observed.

[2]One should notice that an equivalent way to obtain draws from the 3-IBP, would be to first obtain a draw $\Theta = \sum_k \theta_k \delta_{\psi_k}$ from a 3-Beta process prior, and then, conditionally on $\Theta$, draw independent Bernoulli process $X_1, \ldots, X_N$, in which $X_n = \sum_{k \geq 1} x_{n,k} \delta_{\psi_k}$, and $x_{n,k} | \Theta \sim \text{Ber}(\theta_k)$. In practice, this "conditional" representation, requires truncation of the CRM $\Theta$ to a finite approximation, in which only finitely many atoms $K$ are considered.

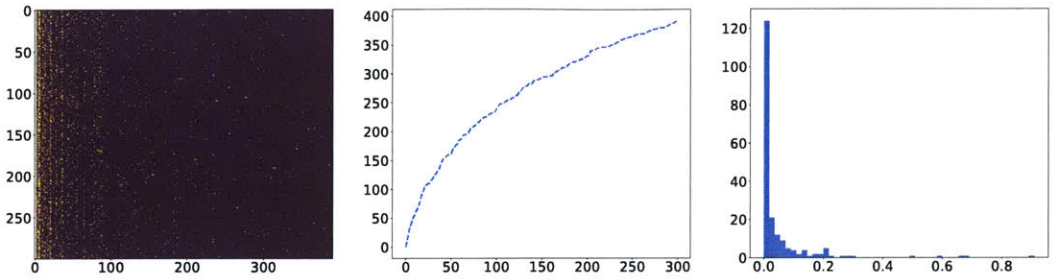



Figure 5-1: An example of a draw from a 3-IBP with parameters $\alpha = 10$, $c = 1$, $\sigma = 0.5$ for $N = 300$ draws. On the left, the binary matrix originating from the draw (zeros are purple, ones are yellow). In the center, the number of "active dishes" $K_N$ ($y$-axis) as a function of the number of customers $N$ ($x$-axis). On the right, the histogram of the counts $z_{300,k}$, for all the active dishes.

We generated synthetic data from different configurations of the hyperparameters and different sample sizes. In particular, we picked the mass parameter $\alpha \in \{1, 10, 100\}$, the concentration parameter $c \in \{1, 5, 10\}$ and the tail parameter $\sigma \in \{0.25, 0.50, 0.75\}$. For each of these configurations, we generated binary matrices with $N = 3000$ rows. Then, we retained the first $N_{\text{train}} \in \{150, 300, 600\}$ rows of these matrices, i.e. the 5%, 10% and 20% of the total observations respctively, and used them as training sets. For each of these datasets, we compared the performance of the different methods described in Chapter 4.

Under this favorable scenario, in which the data generating process is exactly the one under which the BNP estimator is derived, we observe that our method outperforms the alternative ones for all the configurations of the hyperparameters. We notice that the prediciton quality remains extremely high, even when the number of additional samples $M$ is considerably larger than $N$.

### Comparison to the Bernoulli-product model estimator

We compare the BNP estimator of Theorem 2 to the estimator of Equation (4.11) derived from the Bernoulli product model. Conceptually, the main difference between the two models lies in the prior specification for the variants' probabilities: while the BNP model relies on a nonparametric prior, which can fit power law behavior, the Bernoulli-product model assumes a parametric beta prior, which can't accomodate a

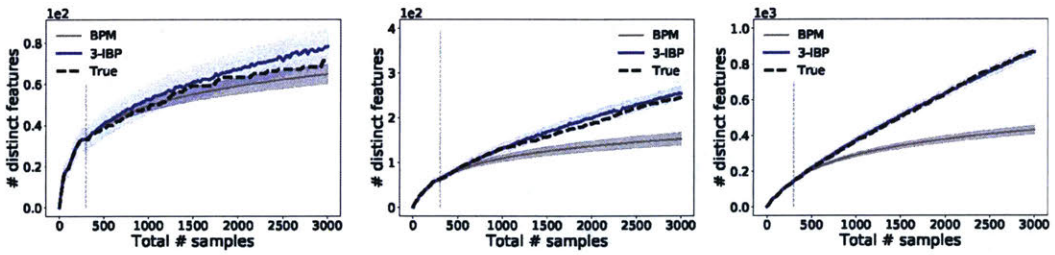



Figure 5-2: Comparison of the BNP estimator (blue) with the Bernoulli-product model estimator (grey). In the plots, we generate data from the 3-IBP and fix the mass parameter $\alpha = 1$ and the concentration parameter $c = 5$, while we vary the tail parameter $\sigma$ (Left: $\sigma = 0.25$, Center: $\sigma = 0.5$, Right: $\sigma = 0.75$). As the power law parameter increases, the predictive performance of the Bernoulli product model decreases.

power-law type distribution for the variants' frequencies. In our synthetic experiments, we find that as we increase the tail parameter $\sigma$, which governs the strength of the power-law behavior of the variants' probabilities, the performance of the estimator derived from the Bernoulli product model progressively worsens (see Figure 5-2).

### Comparison to the UnseenEST estimator

As already mentioned, an undesirable feature of the unseenEST estimator is that it requires to specify a threshold $\kappa$, which determines the fraction of variants' frequencies that are going to be learnt by the algorithm. We find in experiments that a correct tuning of this hyperparameter of the unseenEST estimator can be crucial for its predictive performance (see Figure 5-3). It is, however, unclear what is a principled way of picking the "right" or "best" threshold $\kappa$.

Another desirable property we observe of the IBP estimator, is its "sample efficiency". Under this data generating regime, the IBP estimator needs comparatively less samples in order to get accurate predictions, while the UnseenEST estimator need many samples to obtain accurate prediction (see Figure 5-5).

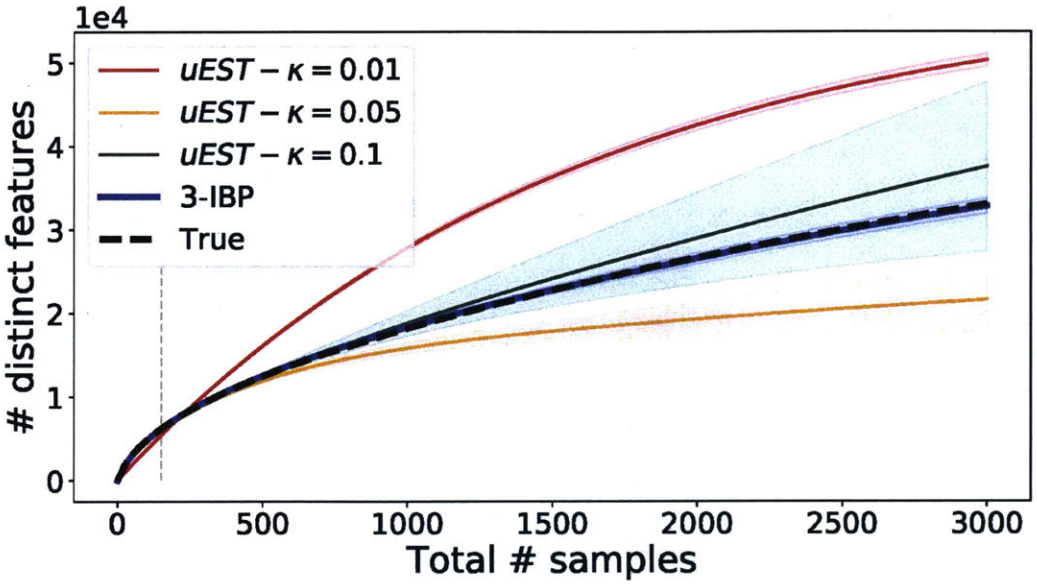
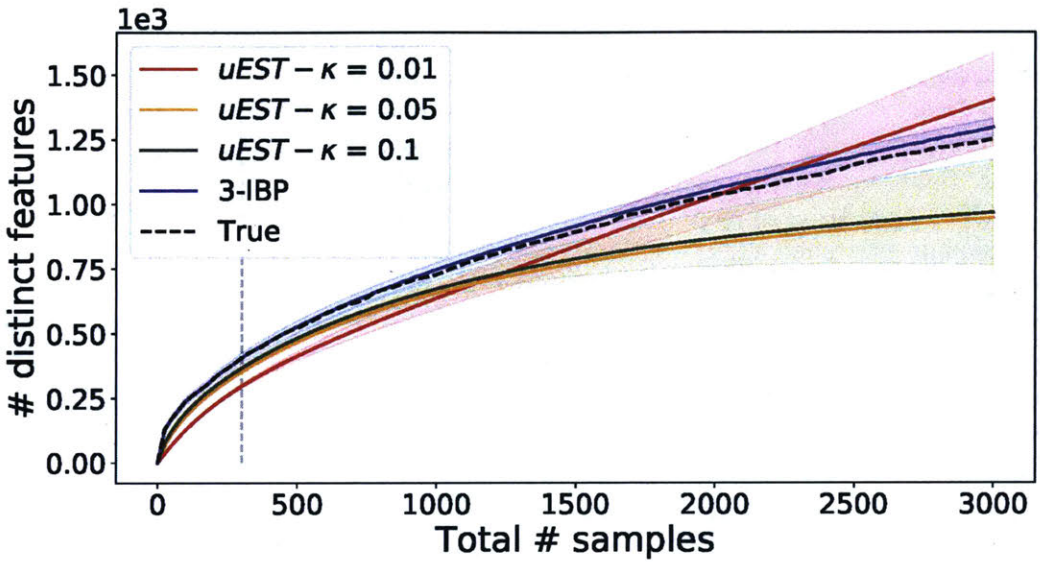



Figure 5-3: Comparing the BNP estimator (blue line) to the unseenEST (uEST) estimator of Zou et al. [2016]. The data is drawn from a 3-IBP with $\alpha = 100$, $\sigma = 0.5$, $c = 1$ with $N = 3000$ observations. 10 training sets of $N_{\text{train}} = 150$ (vertical dotted gray line) datapoints were used to estimate the population histograms and the parameters of the 3-IBP. The uncertainties are provided by averaging over the 10 runs, and plotting one standard deviation. In this case, different thresholds $\kappa$ dramatically change the quality of the performance. Here, we show the performance of unseenEST when $\kappa = 1\%$ (red line), 5% (yellow line) and 10% (green line). It should also be noted that, while in this case the quality of the estimator seems to improve as we increase the threshold level, we observed instances in which the opposite phenomenon is true (see Figure 5-4).

## 5.2 Real datasets

To test our method on real world data, we consider the ExAC dataset [Lek et al., 2016], which contains genetic information from $N = 60{,}076$ individuals recorded at $K_{\text{max}} = 1{,}195{,}872$ genomic loci. A relevant feature of this dataset, is that observations have been labeled according to one of seven categories or subpopulations, according to the geographical origin of the individual sampled: African/African American [AFR], Latino [AMR], East Asian [EAS], Finnish [FIN], Non-Finnish European [NFE], South Asian [SAS] and Other [OTH]. This feature makes it particularly interesting for the hierarchical setting.

Figure 5-4: Comparing the BNP estimator (blue line) to the unseenEST (uEST) estimator of Zou et al. [2016]. The data is drawn from a 3-IBP with $\alpha = 10$, $\sigma = 0.5$, $c = 1$ with $N = 3000$ observations. 10 training sets of $N_{\text{train}} = 300$ (vertical dotted gray line) datapoints were used to estimate the population histograms and the parameters of the 3-IBP. The uncertainties are provided by averaging over the 10 runs, and plotting one standard deviation.




Because of privacy reasons, the released dataset only provides aggregate statistics of the original genomic sequences. In particular, for each subpopulation $d \in [D]$ and for each locus $k \in [K_{\text{max}}]$, we have access to the aggregate number $L_{d,k}$ of times a given position has been succesfully read by the sequencing procedure, and the aggregate number $S_{d,k}$ of times in which variation from the underlying reference genome has been observed in population $d$ at position $k$, so that $L_{d,k} \geq S_{d,k}$ for all $d$, $k$. This allows us compute an empirical probability $p_{d,k} := \frac{S_{d,k}}{L_{d,k}} \in [0, 1]$ of observing variation at position $k$ in subpopulation $d$. We generate a binary matrix $\boldsymbol{X}^{(d)} \in \{0, 1\}^{N_d \times K_{\text{max}}}$, in which each row is a vector of $K_{\text{max}}$ independent Bernoulli random variables, and each entry $X_{n,k}^{(d)} \sim \text{Ber}(p_{d,k})$.

Interestingly, different subpopulations have very different properties, both in terms of sample size (e.g., the smallest subpopulation - FIN - has only $N = 4327$ observations, while the largest - NFE - has $N = 33370$ observations) and in terms of variation (e.g.,

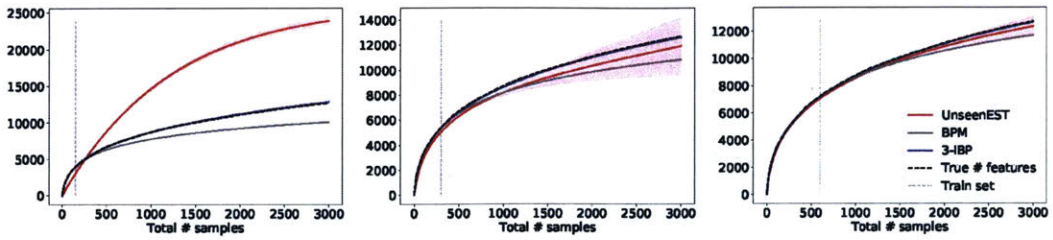



Figure 5-5: Comparing the BNP estimator (blue line) to the unseenEST (uEST) estimator of Zou et al. [2016] (red line) and the Bernoulli product model estimator of Ionita-Laza et al. [2009] (grey). The data is drawn from a 3-IBP with $\alpha = 100$, $\sigma = 0.25$, $c = 5$ with $N = 3000$ observations. 10 training sets of $N_{\text{train}} = 150, 300, 600$ (left, center, right respectively, vertical dotted gray line) datapoints were used to estimate the population histograms and the parameters of the 3-IBP. The uncertainties are provided by averaging over the 10 runs, and plotting one standard deviation.We picked $\kappa = 0.05$, which provided the best estimates. We see that the BNP estimator needs comparatively less samples to obtain precise estimates of the number of new features observed. In this case, $N_{\text{train}} = 150$ suffice to accurately predict up to $N = 3000$ number of samples. Viceversa, the unseenEST algorithm requires many more samples to obtain reliable estimates.

in the FIN subpopulation, variation is observed only in 5.5 % of the loci, while in the NFE subpopulation more than 51% of the loci show variation).

With the notable exception of the FIN dataset, both estimators provided similar results. Overall, the BNP provides very narrow posterior uncertainty estimates. This follows directly from the form of the predictive distribution of Theorem 2, which is a Poisson distribution – which implies that the variance of the estimator is of the same order of the mean. Since the number of distinct features is of the order of the hundreds of thousands for our application, one standard deviation will roughly be of the order of the hundreds. As already observed in the synthetic experiments, we noticed that specifying a correct value of $\kappa$ and $\delta$ for the unseenEST estimator plays a crucial role (see Figure 5-7).

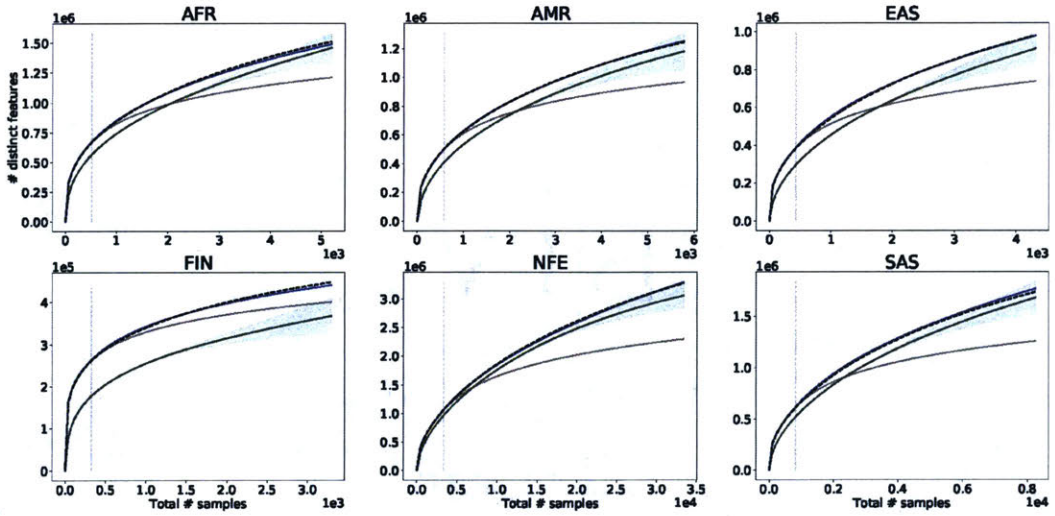



Figure 5-6: For the six main subpopulations present in the ExAC dataset, we report the results of the BNP estimator (blue line), the Bernoulli product model estimator (grey line), as well as the UnseenEST estimator (green line, where we have fixed $\kappa = 10\%$ and $\delta = 1.01$, which we found to perform best among all the hyperparameters tried). Both estimators were trained on 10 subsamples of the same size (vertical grey line). For both estimators, we plot the expected value of the number of distinct features as a function of the sample size, as well as one empirical standard deviation, obtained from the 10 different optimizations.

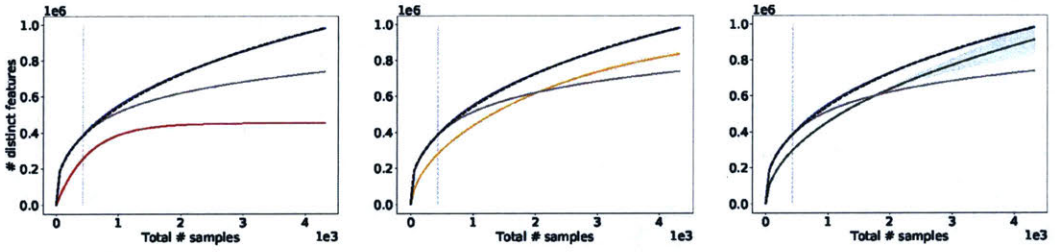



Figure 5-7: For the EAS dataset, we show the results of the BNP estimator (blue line) as well as unseenEST, trained with $\kappa = 0.01$ (left), $\kappa = 0.05$ (center) and $\kappa = 0.1$ (right). As already observed in the synthetic data experiments, different values of $\kappa$ can dramatically affect the quality of the prediction.

## 5.3 Hierarchical estimators: computational issues and future work

The estimators derived in Section 3.2, as already discussed in Section 4.2, require the estimation of the three parameters $\alpha, c, \sigma$ of the underlying beta-process prior, as well

as the parameters $r_d$, for $d = 1, \ldots, D$ which govern the dispersion of the frequencies in each sub-population from the underlying common base measure (see Remark 6). In the genetic datasets under consideration, using an empirical Bayes approach, i.e. a direct maximization of the hierarchical EFPF of Equation (3.6) in the same way as in the single population case is unfeasible. This is due to the fact that the simple evaluation of Equation (3.6) requires – in the naive case – to compute one integral for each feature. In all the datasets under consideration, the numbers of distinct features $K_N$ is in the order of the millions for $N$ in the order of the thousands. Even by using the simplified version Equation (4.6), evaluating the likelihood is too expensive. The study of alternative approaches to efficiently estimate the parameters needed for the hierarchial estimators of Section 3.2 will be part of future study and analysis.

# Appendix A

# Other models

## A.1 Partition models

Within the framework described in Section 2.1, *partition models* are an extremely popular class of models characterized by likelihood processes $X_n$ which allocate each point to one and only one trait. Because of this reason, partition models have found widespread application for the unsupervised problem of clustering. In these applications the likelihood processes $X_n$ are typically assumed to be latent within the generative model, and the goal is to learn posterior distributions over clusterings (see, for example, Antoniak [1974], Escobar and West [1995], Gelfand et al. [2005] and Gershman and Blei [2012], Hjort et al. [2010] for a review). This class of models has also been successfully employed for prediction tasks related to rare species discovery [Arbel et al., 2017, Cesari et al., 2014, Favaro et al., 2009, 2012a,b, 2016], in which, instead, we assume to have direct access to the likelihood processes and the goal is to learn prediction properties from the model.

First introduced by Ferguson [1973], the Dirichlet-multinomial process is an extremely popular partition model, whose theoretical properties have been extensively studied [Escobar and West, 1995, Ishwaran and Zarepour, 2002, Pitman, 2006, Sethuraman, 1994] and that has been widely and succesfully applied in several applied contexts.

**Example 4 (Dirichlet-multinomial process)** *The Dirichlet-multinomial process*

*is obtained by*

1. *a Dirichlet process prior,*

$$\Theta \sim \mathrm{DP}(\alpha, P_0) \overset{a.s.}{=} \sum_{k\geq 1} \theta_k \delta_{\psi_k}. \tag{A.1}$$

*From a CRM perspective, this is obtained by considering the normalized version (NCRM) of an ordinary-component-only CRM on $\mathbb{R}_+ \times \Psi$ with base measure*

$$\mu(\mathrm{d}\theta \times \mathrm{d}\psi) = \alpha\theta^{-1}e^{-\theta}\mathrm{d}\theta P_0(\mathrm{d}\psi),$$

*where $\alpha > 0$ is the concentration parameter and $P_0$ is a probability distribution on $\Psi$, often referred to as the base measure of the process.*

2. *Datapoints are allocated to a single trait through a multinomial likelihood process,*

$$X_n|\Theta \overset{i.i.d.}{\sim} \mathrm{Multi}(\Theta), \tag{A.2}$$

*that is to say, $X_n = \delta_{\psi_k}$ with probability $\theta_k$, $\forall k \geq 1$.*

Inspired by the Dirichlet-multinomial process, several others partition models have been developed; see Lijoi and Prünster [2010] for a review.

## A.2 Trait allocation models

A further generalization of the previous models can be obtained by considering a combinatorial structure in which each datapoint can belong to multiple traits, and to each trait with a different degree of belonging. The resulting "weighted" feature allocation is often referred to as trait allocation. Campbell et al. [2018] provides a theoretical treatment of this class of models, which have been applied to a several applications (see, for example, Broderick et al. [2015], Gupta et al. [2012], Titsias [2008], Zhou [2015, 2018]).

**Example 5 (Gamma-Poisson process)** *The gamma-Poisson process, considered e.g. in Titsias [2008] is obtained thorugh*

1. *A gamma process prior,*

$$\Theta \sim \Gamma\mathrm{P}(\alpha, c, P_0) = \sum_{k\geq 1} \theta_k \delta_{\psi_k}, \tag{A.3}$$

*where, as in the Dirichlet process case, $\Theta$ is an ordinary-component-only CRM on the space $[0,1] \times \Psi$, $\mu(\mathrm{d}\theta \times \mathrm{d}\psi) = \alpha\theta^{-1}e^{-\theta c}\mathrm{d}\theta P_0(\mathrm{d}\psi)$, where $\alpha > 0$ is the concentration parameter and $P_0$ is a probability distribution on $\Psi$*

2. *Datapoints are allocated to traits through a Poisson likelihood process $X_n|\Theta \sim \mathrm{PoP}(\Theta)$*

$$X_n|\Theta \overset{i.i.d.}{\sim} \mathrm{PoP}(\Theta), \tag{A.4}$$

*where, for every $n$, $x_{n,k}|\theta_k \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\theta_k)$.*

**Example 6 (Hierarchical Dirichlet-multinomial process (hDM))**

$$\begin{aligned} \Theta_0 &\sim \mathrm{DP}(\alpha, P_0) \\ \Theta_{1:D}|\Theta_0 &\sim \mathrm{hCRM}(\{\mathrm{Beta}(r_d\theta_{0,k}, r_d(1-\theta_{0,k})\}_{k\geq 1, d=1,\ldots,D}) \\ X_{n,d}|\Theta_d &\overset{\text{i.i.d.}}{\sim} \mathrm{Multi}(\Theta_d) \end{aligned} \tag{A.5}$$

**Example 7 (Hierarchical gamma-Poisson process (hGP) )**

$$\begin{aligned} \Theta_0 &\sim \Gamma\mathrm{P}(\alpha, c, P_0) \\ \Theta_{1:D}|\Theta_0 &\sim \mathrm{hCRM}(\{\mathrm{Gam}(\alpha_d\theta_{0,k}, c_d)\}_{k\geq 1, d=1,\ldots,D}) \\ X_{n,d}|\Theta_d &\overset{\text{i.i.d.}}{\sim} \mathrm{PoP}(s_d\Theta_d) \end{aligned} \tag{A.6}$$

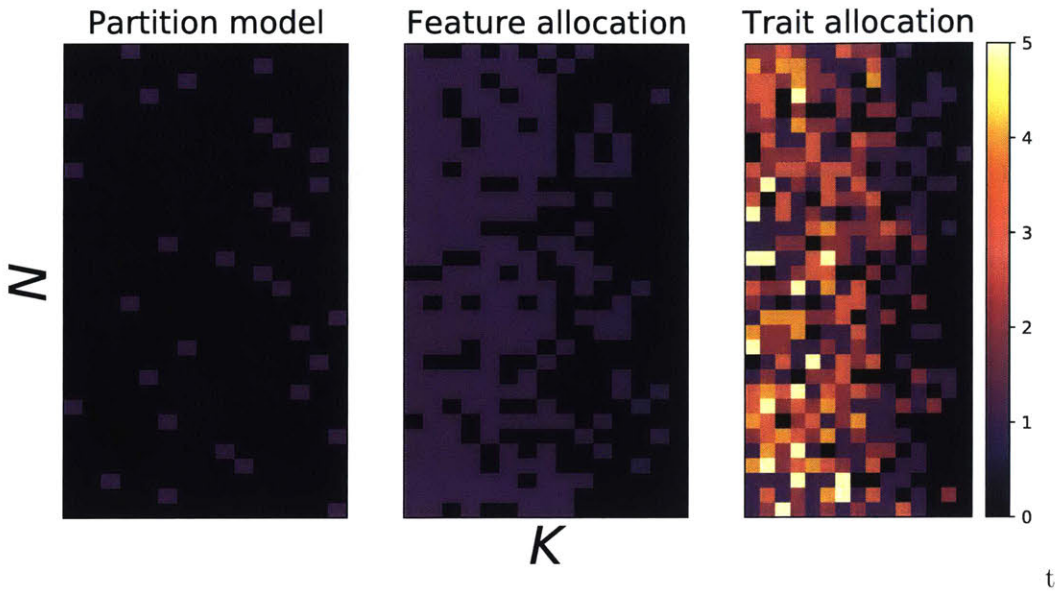



Figure A-1: Graphic representations of the three classes of models described above. In each subplot, each row $n$ represents an individual, and each column $k$ a trait. The colour of each entry $(n, k)$ denotes the allocation value of trait $k$ for individual $n$. In a partition model (left), there is exactly on non-zero column $k$ for each individual $n$, with value 1. In a feature allocation (center), for each individual $n$ there can be multiple traits with value 1. In a trait allocation model (right), we further allow for arbitrary integer values for the trait counts.

*We can also apply the reults of Chapter 3 to the hGP. Let* $m_{\bullet,d,k} = \sum_{n=1}^{N_d} x_{n,d,k}$. *Then,*

$$p_{K_N}(\psi_1, \ldots, \psi_{K_N}; \mathfrak{I}, \boldsymbol{x}^*) = \exp\left\{ -\int_0^\infty \left(1 - \prod_{d=1}^{D} \left(\frac{c_d}{c_d + N_d}\right)^{\theta\alpha_d}\right) \nu_0(\mathrm{d}\theta) \right\} \prod_{k=1}^{K_N} \int_0^1 \prod_{d=1}^{D} \frac{(\theta\alpha_d)_{m_{\bullet,d,k}}}{\prod_{n\in\mathfrak{I}_{d,k}} x^*_{n,d,k}!} \frac{c_d^{\theta\alpha_d}}{(N_d + c_d)^{\theta\alpha_d + m_{\bullet,d,k}}} \nu_0(\mathrm{d}\theta). \tag{A.7}$$

*The updated rate measure of* $\Theta'_0$ *is given by*

$$\nu'_0(\mathrm{d}\theta) = \prod_{d=1}^{D} \left(\frac{c_d}{s_d N_d + c_d}\right)^{\alpha_d\theta} \nu_0(\mathrm{d}\theta).$$

*The rates are conditionally independent with distribution* $\theta'_{d,k}|\Theta'_0 \overset{ind}{\sim} \mathrm{Gam}(\alpha_d\theta'_{0,k},\ c_d + s_d N_d)$. *Moreover, each jump* $\xi_{d,k} \sim \mathrm{Gam}\ (m_{\bullet,d,k} + \alpha_d\xi_{0,k}, s_d N_d + c_d)$.

# Appendix B

# Proofs

**Lemma 3** *Assume*

$$\int_{\mathbb{R}_+\times\Psi}\int_{\mathbb{R}_+}(1-\pi_d(t))\rho_d(t|s,r_d)\mathrm{d}t\nu_0(\mathrm{d}s)P_0(\mathrm{d}\psi)<\infty,\quad\forall\, d\in[D]. \tag{B.1}$$

*Then,*

$$\sum_{k\geq 1}\mathbb{1}(x_{n,d,k}>0)<\infty \tag{B.2}$$

*almost surely,* $\forall n\in[N_d],\forall d\in[D]$.

***Proof**: To show the thesis, we can equivalently show that*

$$\mathbb{E}\left[\sum_{k\geq 1}\mathbb{1}(x_{n,d,k}>0)\right]<\infty. \tag{B.3}$$

*We have*

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k\geq 1}\mathbb{1}(x_{n,d,k}>0)\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{k\geq 1}\mathbb{1}(x_{n,d,k}>0)\middle|\Theta_d\right]\right] \\
&= \mathbb{E}\left[\sum_{k\geq 1}\mathbb{P}(x_{n,d,k}>0|\Theta_d)\right] \\
&= \mathbb{E}\left[\sum_{k\geq 1}(1-\pi_d(\theta_{d,k}))\right] \\
&= \mathbb{E}\left[\sum_{k\geq 1}\mathbb{E}\left[(1-\pi_d(\theta_{d,k}))\middle|\Theta_0\right]\right] \\
&= \mathbb{E}\left[\sum_{k\geq 1}\int_{\mathbb{R}_+}(1-\pi_d(s))\rho_d(s|\theta_{0,k},r_d)\mathrm{d}\theta_{d,k}\right] \\
&= \int_{\mathbb{R}_+\times\Psi}\int_{\mathbb{R}_+}(1-\pi_d(s))\rho_d(s|t,r_d)\mathrm{d}s\nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) \\
&< \infty,
\end{aligned}
$$

*where the last step follows from the assumption of Equation* (B.1).

## B.1 Proof of Lemma 1

To determine the joint distribution eq. (3.1), we consider infinitesimally small balls centered around traits $\{\psi_k^*\}_k$ with radius $\epsilon > 0$ so that they are all mutually disjoint, $B(\psi_k^*,\epsilon)\cap B(\psi_{k'}^*,\epsilon)=\emptyset$ for all $k\neq k'\in[K_N]$. Let $\Psi^{**}:=\Psi\cap(\cup_{k=1}^{K_N}B(\psi_k^*,\epsilon))^C$ and define the following events:

$$E_1 := \{X_{n,d}(B(\psi_k^*,\epsilon)) = x_{n,d,k}^*, \forall n\in\mathcal{B}_{d,k}, d\in[D], k\in[K_N]\}, \tag{B.4}$$

$$E_2 := \{X_{n,d}(B(\psi_k^*,\epsilon)) = 0, \forall n\notin\mathcal{B}_{d,k}, d\in[D], k\in[K_N]\}, \tag{B.5}$$

$$E_3 := \{X_{n,d}(\Psi^{**}) = 0, \forall n\notin\mathcal{B}_{d,k}, d\in[D], k\in[K_N]\}. \tag{B.6}$$

Letting $E := E_1\cap E_2\cap E_3$, it is clear that

$$P_{K_N}(B(\psi_1^*,\epsilon),\ldots,B(\psi_{K_N}^*,\epsilon)) = \mathbb{P}(E),$$

and

$$\mathbb{P}(E) = \mathbb{E}\left[\prod_{d=1}^{D}\prod_{k=1}^{K_N}\left(\prod_{n\in\mathcal{B}_{d,k}}\mathbb{P}(X_{n,d}(B(\psi_k^*,\epsilon)) = x_{n,d,k}^*|\Theta_d)\right)\right.$$
$$\left.\left(\prod_{n\notin\mathcal{B}_{k,d}}\mathbb{P}(X_{n,d}(B(\psi_k^*,\epsilon)) = 0|\Theta_d)\right) \times \prod_{d=1}^{D}\prod_{n=1}^{N_d}\mathbb{P}(X_{n,d}(\Psi^{**}) = 0|\Theta_d)\right]. \tag{B.7}$$

From the hCRM construction, $\Theta_d|\Theta_0 = \sum_k \theta_{d,k}\delta_{\psi_k}$, where the atoms $\{\psi_k\}_k$ are almost surely disjoint if the base measure $P_0$ of $\Theta_0$ is diffuse. Hence, for $k \in [K_N]$,

$$\begin{aligned}
&\mathbb{P}(X_{n,d}(B(\psi_k^*,\epsilon)) = x_{n,d,k}^*|\Theta_d) \\
&\quad= \sum_{l\geq 1}\mathbb{P}(x_{n,d,l} = x_{n,d,k}^*)\delta_{\psi_l}(B(\psi_k^*,\epsilon)) + o\left(\prod_{k'=1}^{K_N}P_0(B(\psi_{k'}^*,\epsilon))\right) \\
&\quad= \sum_{l\geq 1}h_d(x_{n,d,l}^*|\theta_{d,l},s_d)\delta_{\psi_l}(B(\psi_k^*,\epsilon)) + o\left(\prod_{k'=1}^{K_N}P_0(B(\psi_{k'}^*,\epsilon))\right),
\end{aligned} \tag{B.8}$$

and similarly

$$\begin{aligned}
&\mathbb{P}(X_{n,d}(B(\psi_k^*,\epsilon)) = 0|\Theta_d) = \\
&\qquad\sum_{l\geq 1}\pi_d(\theta_{d,l})\delta_{\psi_l}(B(\psi_k^*,\epsilon)) + o\left(\prod_{k'=1}^{K_N}P_0(B(\psi_{k'}^*,\epsilon))\right).
\end{aligned} \tag{B.9}$$

Consider now the probability generating function of the measure $X_{n,d}(\Psi^{**})$ conditionally on the measure $\Theta_d$. This is

$$\begin{aligned}
G_{X_{n,d}(\Psi^{**})|\Theta_d}(t) &= \mathbb{E}[t^{X_{n,d}(\Psi^{**})}|\Theta_d] \\
&= \prod_{l\geq 1}\mathbb{E}[\exp\{x_{n,d,l}\delta_{\psi_l(\Psi^{**})}\log(t)\}|\Theta_d] \\
&= \prod_{l\geq 1}(\delta_{\psi_l}(\Psi^{**})\mathbb{E}[t^{x_{n,d,l}}|\Theta_d] + 1 - \delta_{\psi_l}(\Psi^{**})) \\
&= \prod_{l\geq 1}\left(\delta_{\psi_l}(\Psi^{**})\left(\pi_d(\theta_{d,k}) + \sum_{x\geq 1}t^x h_d(x|\theta_{d,l},s_d)\right) + 1 - \delta_{\psi_l}(\Psi^{**})\right),
\end{aligned}$$

and so

$$\mathbb{P}(X_{n,d}(\Psi^{**}) = 0|\Theta_d) = G_{X_{n,d}(\Psi^*)|\Theta_d}(0) = \prod_{k\geq 1}(1-(1-\pi_d(\theta_{d,k}))\delta_{\psi_k}(\Psi^{**}))\,. \quad \text{(B.10)}$$

Adding together eq. (B.8), eq. (B.9) and eq. (B.10), using the fact that the atoms $\{\psi_l\}$ are a.s. disjoint,

$$\mathbb{P}(E) = \mathbb{E}\left[\prod_{d=1}^{D}\prod_{k=1}^{K_N}\left(\sum_{l\geq 1}(1-\pi_d(\theta_{d,k}))^{m_{d,k}}\pi_d(\theta_{d,k})^{N_d-m_{d,k}}\prod_{n\in\mathcal{B}_{d,k}}h_d(x^*_{n,d,k}|\theta_{d,l})\delta_{\psi_l}(B(\psi^*_k,\epsilon))\right)\right.$$
$$\left.\prod_{d=1}^{D}\prod_{l\geq 1}(\pi_d(\theta_{d,l})\delta_{\psi_l}(\Psi^*))^{N_d}\right] + o\left(\prod_{k=1}^{K_N}P_0(B(\psi^*_k,\epsilon))\right).$$

Now, conditioning on the measure $\Theta_0$, using conditional independence of the components of the hCRM

$$\begin{aligned}
\mathbb{P}(E) &= \mathbb{E}\left[\mathbb{E}\left[\prod_{d=1}^{D}\prod_{k=1}^{K_N}\left(\sum_{l\geq 1}\pi_d(\theta_{d,l})^{N_d-m_{d,k}}\prod_{n\in\mathcal{B}_{d,k}}h_d(x^*_{n,d,k}|\theta_{d,l})\delta_{\psi_l}(B(\psi^*_k,\epsilon))\right)\right.\right.\\
&\quad\left.\left.\prod_{d=1}^{D}\prod_{l\geq 1}(\pi_d(\theta_{d,l})\delta_{\psi_l}(\Psi^*))^{N_d}\right]\middle|\Theta_0\right] + o\left(\prod_{k=1}^{K_N}P_0(B(\psi^*_k,\epsilon))\right)\\
&= \mathbb{E}\left[\prod_{d=1}^{D}\prod_{k=1}^{K_N}\mathbb{E}\left[\left(\sum_{l\geq 1}\pi_d(\theta_{d,l})^{N_d-m_{d,k}}\prod_{n\in\mathcal{B}_{d,k}}h_d(x^*_{n,d,k}|\theta_{d,l})\delta_{\psi_l}(B(\psi^*_k,\epsilon))\right)\middle|\Theta_0\right]\right.\\
&\quad\left.\prod_{d=1}^{D}\prod_{l\geq 1}\mathbb{E}\left[(\pi_d(\theta_{d,l})\delta_{\psi_l}(\Psi^*))^{N_d}\middle|\Theta_0\right]\right] + o\left(\prod_{k=1}^{K_N}P_0(B(\psi^*_k,\epsilon))\right)\\
&= \mathbb{E}\left[\prod_{d=1}^{D}\prod_{k=1}^{K_N}\sum_{l\geq 1}\int_{\mathbb{R}_+}\pi_d(s)^{N_d-m_{d,k}}\prod_{n\in\mathcal{B}_{d,k}}h_d(x^*_{n,d,k}|s)\delta_{\psi_l}(B(\psi^*_k,\epsilon))\rho_d(s|r_d,\theta_{0,k})\mathrm{d}s\right.\\
&\quad\left.\prod_{d=1}^{D}\prod_{l\geq 1}\int_{\mathbb{R}_+}(\pi_d(s)\delta_{\psi_l}(\Psi^*))^{N_d}\,\rho_d(s|r_d,\theta_{0,l})\mathrm{d}s\right] + o\left(\prod_{k=1}^{K_N}P_0(B(\psi^*_k,\epsilon))\right).
\end{aligned}$$

Next, using the fact that the balls $\{B(\psi_k^*, \epsilon)\}_{k=1}^{K_N}$ are disjoint, and the increments of the CRM are independent,

$$\begin{aligned}
\mathbb{P}(E) = &\prod_{k=1}^{K_N} \mathbb{E}\left[\prod_{d=1}^{D}\sum_{l\geq 1}\int_{\mathbb{R}_+} \pi_d(s)^{N_d - m_{d,k}} \prod_{n\in\mathcal{B}_{d,k}} h_d(x_{n,d,k}^*|s)\delta_{\psi_l}(B(\psi_k^*,\epsilon))\rho_d(s|r_d,\theta_{0,l})\mathrm{d}s\right] \\
&\times \mathbb{E}\left[\prod_{l\geq 1}\prod_{d=1}^{D}\int_{\mathbb{R}_+} (\pi_d(s)\delta_{\psi_l}(\Psi^*))^{N_d}\,\rho_d(s|r_d,\theta_{0,l})\mathrm{d}s\right] \\
&+ o\left(\prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon))\right) \\
= &\prod_{k=1}^{K_N}\int_{B(\psi_k^*,\epsilon)\times\mathbb{R}_+} \\
&\prod_{d=1}^{D}\left(\int_{\mathbb{R}_+} \pi_d(s)^{N_d - m_{d,k}} \prod_{n\in\mathcal{B}_{d,k}} h_d(x_{n,d,k}^*|s)\delta_{\psi_l}(B(\psi_k^*,\epsilon))\rho_d(s|r_d,t)\mathrm{d}s\right)\nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) \\
&\times \mathbb{E}\left[\exp\left\{\sum_{l\geq 1}\log\left(\prod_{d=1}^{D}\int_{\mathbb{R}_+} (\pi_d(t)\delta_{\psi_l}(\Psi^*))^{N_d}\,\rho_d(s|r_d,t)\mathrm{d}s\right)\right\}\right] \\
&+ o\left(\prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon))\right).
\end{aligned}$$

Evaluating the Laplace functional of $\Theta_0$, and using the fact that $P_0$ is non-atomic, we get

$$
\begin{aligned}
\mathbb{P}(E) = & \prod_{k=1}^{K_N} \int_{B(\psi_k^*,\epsilon)\times\mathbb{R}_+} \\
& \prod_{d=1}^{D} \left( \int_{\mathbb{R}_+} \pi_d(s)^{N_d - m_{d,k}} \prod_{n\in\mathcal{B}_{d,k}} h_d(x^*_{n,d,k}|s)\delta_{\psi_l}(B(\psi_k^*,\epsilon))\rho_d(s|r_d,t)\mathrm{d}s \right) \\
& \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) \times \exp\left\{ -\int_{\mathbb{R}_+} \left( 1 - \prod_{d=1}^{D} \int_{\mathbb{R}_+} \pi(s)^{N_d}\rho_d(s|r_d,t)\mathrm{d}s \right) \nu_0(\mathrm{d}t) \right\} \\
& + o\left( \prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon)) \right) \\
= & \, e^{-\Phi(N_1,\ldots,N_D)} \prod_{k=1}^{K_N} \int_{B(\psi_k^*,\epsilon)\times\mathbb{R}_+} \prod_{d=1}^{D} \left( \int_{\mathbb{R}_+} \pi_d(s)^{N_d - m_{d,k}} \prod_{n\in\mathcal{B}_{d,k}} h_d(x^*_{n,d,k}|s) \right. \\
& \left. \rho_d(s|r_d,t)\mathrm{d}s \right) \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) + o\left( \prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon)) \right). \qquad \text{(B.11)}
\end{aligned}
$$

Now, using Lebesgue differentiation theorem, the density with respect to the $K_N$ fold product $P_0^{\otimes K_N}$ is given by

$$
\begin{aligned}
p_{K_N}(\psi_1^*,\ldots,\psi_{K_N}^*;\mathcal{B},\boldsymbol{x}^*) &= \lim_{\epsilon\downarrow 0} \frac{P_{K_N}(B(\psi_1^*,\epsilon),\ldots,B(\psi_{K_N}^*,\epsilon);\mathcal{B},\boldsymbol{x}^*)}{\prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon))} \qquad \text{(B.12)} \\
&= \lim_{\epsilon\downarrow 0} \frac{\mathbb{P}(E)}{\prod_{k=1}^{K_N} P_0(B(\psi_k^*,\epsilon))}, \qquad \text{(B.13)}
\end{aligned}
$$

for $P_0^{\otimes K_N}$ almost every point $\boldsymbol{\psi}$ in the product space $\Psi^{K_N}$.

**Lemma 4** *Let $\Theta_1,\ldots,\Theta_D$ be a vector of random measures on $\Psi$, and let $f_1,\ldots,f_D : \Psi \to \mathbb{R}_+$ be measurable.*

*Denote $\Theta_d(f_d) := \sum_k \theta_{d,k} f(\psi_k)$. Then if $\mathcal{L}_{\Theta_{1:D}}(f_1,\ldots,f_D) := \mathbb{E}[e^{-\sum_{d=1}^{D} f_d(\Theta_d)}]$ is the joint Laplace functional of $\Theta_{1:D}$ for $f_{1:D}$, it holds*

$$
\mathcal{L}_{\Theta_{1:D}}(f_1,\ldots,f_D) = \exp\left\{ -\int_{\Psi\times\mathbb{R}_+} \left( 1 - \prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-sf_d(\psi)}\rho_d(s|\rho_d,t)\mathrm{d}s \right) \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) \right\}.
$$

***Proof:*** *Let* $\Theta_1, \ldots, \Theta_D | \Theta_0 \sim \text{hCRM}((\rho_d(\cdot|r_d, \theta_{0,k})_{k,d}),\ \Theta_0 \sim \text{CRM}(\nu_0, P_0)$.

$$\begin{aligned}
\mathcal{L}_{\Theta_1,\ldots,\Theta_D}(f_1, \ldots, f_D) &= \mathbb{E}\left[\prod_{d=1}^{D} \mathbb{E}\left[e^{\Theta_d(f_d)}\middle|\Theta_0\right]\right] = \mathbb{E}\left[\prod_{d=1}^{D}\prod_{k\geq 1} \mathbb{E}\left[\exp\left\{-\theta_{d,k} f_d(\psi_k)\right\}\right]\right] \\
&= \mathbb{E}\left[\prod_{k\geq 1}\prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-s f_d(\psi_k)} \rho_d(s|r_d, \theta_{0,k}) \mathrm{d}s\right] \\
&= \mathbb{E}\left[\exp\left\{\sum_{k\geq 1} \log\left(\prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-s f_d(\psi_k)} \rho_d(s|r_d, \theta_{0,k}) \mathrm{d}s\right)\right\}\right],
\end{aligned}$$

*and using the Laplace functional evaluation of* $\Theta_0$ *yields the thesis.*

## B.2 Proof of Lemma 2

To show the equivalence in distribution of eq. (3.2) we prove that the posterior Laplace functional of the right handside coincides with the one of the left handside. Fix $D$ measurable functions $f_1, \ldots, f_D : \Psi \to \mathbb{R}_+$. We would like to evaluate

$$\mathcal{L}_{\Theta_{1:D}|\boldsymbol{x}^*, \boldsymbol{\xi}_0^*}(f_1, \ldots, f_D) := \mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D} \Theta_d(f_d)\right\}\middle|\boldsymbol{x}^*, \boldsymbol{\xi}_0^*\right]. \tag{B.13}$$

Now, consider the product space space $(\mathbb{R}_+ \times \Psi)^{K_N}$ and endow it with its Borel sigma algebra, and let $P_{K_N}(\cdot; \mathcal{B}, \boldsymbol{x}^*)$ be the joint density on this space with respect to the $K_N$-fold product measure $(\text{Leb}(\mathbb{R}_+) \times P_0)^{\otimes K_N}$. The previous equation is thus equivalent to computing the limit

$$\mathcal{L}_{\Theta_{1:D}|\boldsymbol{x}^*, \boldsymbol{\xi}_0^*}(f_{1:D}) = \lim_{\epsilon \downarrow 0} \frac{\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D} \Theta_d(f_d)\right\} P_{K_N}\left(\cup_{k=1}^{K_N}\{B(\psi_k^*, \epsilon) \times B(\xi_k^*, \epsilon)\}; \mathcal{B}, \boldsymbol{x}^*\right)\right]}{P_{K_N}\left(\cup_{k=1}^{K_N}\{B(\psi_k^*, \epsilon) \times B(\xi_k^*, \epsilon)\}; \mathcal{B}, \boldsymbol{x}^*\right)}.$$

Using the results from Lemma 1, neglecting superior order terms, we obtain

$$\mathcal{L}_{\Theta_{1:D}|\boldsymbol{x}^*,\boldsymbol{\xi}_0^*}(f_{1:D})$$
$$= \lim_{\epsilon \downarrow 0} \left[ \frac{e^{-\Xi(N_{1:D};f_{1:D})}}{e^{-\Phi(N_1,\ldots,N_D)}} \frac{\prod_{k=1}^{K_N} \int_{B((\psi_k^*,\xi_k^*),\epsilon)} \prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-sf_d(\psi_k^*)} I_{d,k}(s|t)\mathrm{d}s\ \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi)}{\prod_{k=1}^{K_N} \int_{B((\psi_k^*,\xi_k^*),\epsilon)} \prod_{d=1}^{D} \int_{\mathbb{R}_+} I_{d,k}(s|t)\mathrm{d}s\ \nu(\mathrm{d}t)P_0(\mathrm{d}\psi)} \right], \tag{B.14}$$

where

$$I_{d,k}(s|t) = \pi(s)^{N_d - m_{d,k}} \prod_{n \in \mathcal{B}_{d,k}} h_d(x_{n,d,k}^*|\theta_d, s_d)\rho_d(s|t, r_d),$$

and

$$\Xi(N_{1:D}, f_{1:D}) = \int_{\mathbb{R}_+ \times \Psi} \left( 1 - \prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-sf_d(\psi)} \pi_d(s)^{N_d} \rho_d(s|t, r_d)\mathrm{d}s \right) \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi).$$

The ratios of the exponentials in eq. (B.14) yields

$$\begin{aligned}
&\exp\left\{-\left[\Xi(N_1,\ldots,N_D,f_1,\ldots,f_D) - \Phi(N_1,\ldots,N_D)\right]\right\} \\
&\quad = \exp\left\{ -\int_{\Psi\times\mathbb{R}_+} 1 - \prod_{d=1}^{D} \int_{\mathbb{R}_+} e^{-sf_d(\psi)} \frac{\pi_d(s)^{N_d}\rho_d(s|t,r_d)\mathrm{d}s}{\int_{\mathbb{R}_+} \pi_d(s')^{N_d}\rho_d(s'|t,r_d)\mathrm{d}s'} \right. \\
&\quad \left. \left( \prod_{d=1}^{D} \int_{\mathbb{R}_+} \pi_d(s)^{N_d}\rho_d(s|t,r_d)\mathrm{d}s \right) \nu_0(\mathrm{d}t)P_0(\mathrm{d}\psi) \right\},
\end{aligned}$$

which is the Laplace functional of a vector of hCRM (see Lemma 4).

The ratio of integrals in Equation (B.14) is the Laplace transform of the vector of jumps $(\xi_{d,k}^*)_{d,k}$, hence the thesis follows.

## B.3 Proofs of Chapter 3

### B.3.1 Proof of Theorem 1

The crucial tool is the probability generating function [PGF] to obtain the posterior predictive distribution of $O_{N,k}^{(M)}$. Let $t \in \mathbb{R}$, then

$$G_{O_{N,k}^{(M)}}(t) := \mathbb{E}\left[t^{O_{N,k}^{(M)}} \middle| \boldsymbol{X}_N^*\right] = \mathbb{E}[t^{\sum_{m=1}^{M} x'_{N+m,k}}] = \mathbb{E}\left[\mathbb{E}[t^{x'_{N+m,k}}|\Theta]\right] \tag{B.15}$$

$$= \mathbb{E}\left[\prod_{m=1}^{M}\left(t\mathbb{P}(x'_{N+m} = 1|\Theta) + \mathbb{P}(x'_{N+m,k} = 0|\Theta)\right)\right] \tag{B.16}$$

$$= \mathbb{E}\left[\prod_{m=1}^{M}\left(t + (1-t)(1-\theta_k)\right)\right] \tag{B.17}$$

$$= \mathbb{E}[(t + (1-t)(1-\theta_k))^M]. \tag{B.18}$$

Now, using the fact that for any discrete random variable $Z$ it holds

$$\mathbb{P}(Z = l) = \frac{\mathrm{d}^l}{t^l}\frac{1}{l!}G_Z(t)|_{t=0}, \tag{B.19}$$

it follows that

$$\mathbb{P}(O_{N,k}^{(M)} = l|\boldsymbol{X}_N^*) = \frac{\mathrm{d}}{\mathrm{d}t^l}\frac{1}{l!}G_{O_{N,k}^{(M)}}(t)|_{t=0}. \tag{B.20}$$

The $l$-th derivative of the PGF is given by

$$\frac{\mathrm{d}^l}{\mathrm{d}t^l}G_{N,k}^{(M)}(t) = \frac{\mathrm{d}}{\mathrm{d}t^l}\mathbb{E}[(t + (1-t)(1-\theta_k))^M] \tag{B.21}$$

$$= M(M-1)\dots(M-l+1)\mathbb{E}[\theta_k^l(t + (1-t)(1-\theta_k))^{M-l}] \tag{B.22}$$

$$= \frac{M!}{(M-l)!}\mathbb{E}[\theta_k^l(t + (1-t)(1-\theta_k))^{M-l}]. \tag{B.23}$$

Now, plugging the value $t = 0$ in the previous expression we obtain

$$\mathbb{P}(O_{N,k}^{(M)} = l|\boldsymbol{X}_N^*) = \binom{M}{l}\mathbb{E}\left[\theta_k^l(1-\theta_k)^{M-l}\right] \tag{B.24}$$

$$= \binom{M}{l}\frac{\boldsymbol{B}(l + z_{N,k} - \sigma, M - l + N - z_{N,k} + c + \sigma)}{\boldsymbol{B}(z_{N,k} - \sigma, N - z_{N,k} + c + \sigma)}. \tag{B.25}$$

where we are using the posterior characterization of the Beta-Bernoulli process given in Example 3, i.e. the fact that the jumps $\theta_k \sim \text{Beta}(z_{N,k} - \sigma, N - z_{N,k} + c + \sigma)$.

### B.3.2 Proof of Theorem 2

Again, we consider the PGF of the random variable, $U_M^{(N)}$. For any $t \in \mathbb{R}$,

$$G_{U_N^{(M)}} = \mathbb{E}\left[t^{U_N^{(M)}}|\boldsymbol{X}_N^*\right] = \mathbb{E}\left[t^{\sum_{k\geq 1}\mathbb{1}(\sum_{m=1}^M x'_{n+m,k}>0)}|\boldsymbol{X}_N^*\right] \tag{B.26}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\mathbb{E}\left(t^{\mathbb{1}(\sum_{m=1}^M x'_{N+m.k}>0)}|\Theta'_0\right)\right] \tag{B.27}$$

$$= \mathbb{E}\left[\prod_{k\geq 1} t\mathbb{P}\left(\sum_{m=1}^M x'_{N+m,k} > 0|\Theta'_0\right) + \mathbb{P}\left(\sum_{m=1}^M x'_{N+m,k} = 0|\Theta'_0\right)\right], \tag{B.28}$$

and since the sum $x'_{N+1,k} + \cdots + x'_{N+M,k} = 0$ only if $x'_{N+m,k} = 0$ for all $m \in [M]$, we get

$$= \mathbb{E}\left[\prod_{k\geq 1}\left(t + (1-t)\prod_{m=1}^M \mathbb{P}(x'_{N+m,k} = 0|\Theta'_0)\right)\right] \tag{B.29}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\left(t + (1-t)(1-\xi'_k)^M\right)\right] \tag{B.30}$$

$$= \mathbb{E}\left[\exp\left\{\sum_{k\geq 1}\log(t + (1-t)(1-\xi'_k)^M)\right\}\right] \tag{B.31}$$

$$= \exp\left\{-(1-t)\int_0^1 (1-(1-s)^M)\rho(s)(1-s)^N ds\right\}, \tag{B.32}$$

where we used the Laplace functional of $\Theta_0'$. Now we can rewrite

$$1-(1-s)^M = s\sum_{m=0}^{M-1}(1-s)^m, \tag{B.33}$$

and letting $\rho_N'(s) = (1-s)^N\rho(s)$ we get

$$G_{U_N^{(M)}} = \exp\left\{-(1-t)\sum_{m=0}^{M-1}\int_{[0,1]}(1-s)^m s\rho_N'(s)\mathrm{d}s\right\} \tag{B.34}$$

$$= \exp\left\{-(1-t)\sum_{m=0}^{M-1}\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\int_{[0,1]}s^{1-\sigma-1}(1-s)^{N+c+\sigma+m-1}\mathrm{d}s\right\} \tag{B.35}$$

$$= \exp\left\{-(1-t)\int_{[0,1]}(1-(1-s)^M)\rho_N'(s)\mathrm{d}s\right\} \tag{B.36}$$

$$= \exp\left\{-(1-t)\alpha\sum_{m=0}^{M-1}\frac{(c+\sigma)_{N+m}}{(c+1)_{N+m}}\right\}. \tag{B.37}$$

Now, taking the derivative

$$\frac{\mathrm{d}^l}{\mathrm{d}t^l}G_{U_N^{(M)}}(t) = \left(\alpha\frac{(c+\sigma)_{N+m}}{(c+1)_{N+m}}\right)^l \exp\left\{-(1-t)\alpha\sum_{m=0}^{M-1}\frac{(c+\sigma)_{N+m}}{(c+1)_{N+m}}\right\}, \tag{B.38}$$

and evaluating this as $t=0$ we get

$$\mathbb{P}(U_N^{(M)} = l|\boldsymbol{X}_N^*) = \frac{1}{l!}\left(\alpha\sum_{m=1}^{M}\frac{(c+\sigma)_{N+m-1}}{(c+1)_{N+m-1}}\right)^l \exp\left\{-\alpha\sum_{m=1}^{M}\frac{(c+\sigma)_{N+m-1}}{(c+1)_{N+m-1}}\right\}, \tag{B.39}$$

which is a Poisson random variable with parameter $\alpha\sum_{m=1}^{M}\frac{(c+\sigma)_{N+m-1}}{(c+1)_{N+m-1}}$.

### B.3.3 Proof of Corollary 1

Using the same strategy as the previous proof, we can write the PGF of $U_{N,r}^{(M)}$ for $t\in\mathbb{R}$ as

$$G_{U_{N,r}^{(M)}}(t) = \mathbb{E}\left[t^{U_{N,r}^{(M)}} | \boldsymbol{X}_N^*\right] \tag{B.40}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}(t-1)\mathbb{P}\left(\sum_{m=1}^{M} x'_{N+m,k} = r|\Theta_0\right) + 1\right]. \tag{B.41}$$

Since conditionally on $\Theta_0$, $x'_{N+1,k} + \cdots + x'_{N+M,k} \sim \text{Binom}(N, \theta'_k)$, we have

$$\mathbb{P}\left(\sum_{m=1}^{M} x'_{N+m,k} = r|\Theta_0\right) = \binom{M}{r}(\theta'_k)^r(1-\theta'_k)^{M-r}, \tag{B.42}$$

from which

$$G_{U_{N,r}^{(M)}}(t) = \mathbb{E}\left[\prod_{k\geq 1}\left((t-1)\binom{m}{r}(\theta'_k)^r(1-\theta'_k)^{M-r} + 1\right)\right] \tag{B.43}$$

$$= \mathbb{E}\left[\exp\left\{\sum_{k\geq 1}\log\left((t-1)\binom{M}{r}(\theta'_k)^r(1-\theta'_k)^{M-r} + 1\right)\right\}\right] \tag{B.44}$$

$$= \exp\left\{-\alpha(1-t)\binom{M}{r}\int_{[0,1]}\theta^r(1-\theta)^{M-r}\rho'_N(\theta)\mathrm{d}\theta\right\} \tag{B.45}$$

$$= \exp\left\{-\alpha(1-t)\binom{M}{r}\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\int_{[0,1]}\theta^{r-\sigma-1}(1-\theta)^{N+c+\sigma+M-r-1}\mathrm{d}\theta\right\} \tag{B.46}$$

$$= \exp\left\{-\alpha(1-t)\binom{M}{r}\frac{(1-\sigma)_{r-1}(c+\sigma)_{N+M-r}}{(c+1)_{N+M-1}}\right\}. \tag{B.47}$$

### B.3.4 Proof of Theorem 3

$$G_{O^{(M_d)}_{d,N_d,k}}(t) := \mathbb{E}\left[t^{O^{(M_d)}_{d,N_d,k}} \middle| \boldsymbol{X}^*_{\boldsymbol{N}}\right] = \mathbb{E}[t^{\sum_{m=1}^{M_d} x'_{d,N_d+m,k}}] = \mathbb{E}\left[\mathbb{E}[t^{x'_{d,N_d+m,k}}|\Theta_0]\right] \tag{B.48}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[t^{x'_{d,N_d+m,k}}|\Theta_d]|\Theta_0\right]\right] \tag{B.49}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\prod_{m=1}^{M_d}\left(t\mathbb{P}(x'_{N+m}=1|\Theta_d) + \mathbb{P}(x'_{N+m,k}=0|\Theta_d)\right)|\Theta_0\right]\right] \tag{B.50}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\prod_{m=1}^{M}\left(t+(1-t)(1-\theta_{d,k})\right)|\Theta_0\right]\right] \tag{B.51}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(t+(1-t)(1-\theta_{d,k})\right)^M|\Theta_0\right]\right]. \tag{B.52}$$

Now, the $l$-th derivative of the PGF is given by

$$\frac{\mathrm{d}^l}{\mathrm{d}t^l}O^{(M_d)}_{d,N_d,k}(t) = \frac{\mathrm{d}}{\mathrm{d}t^l}\mathbb{E}[(t+(1-t)(1-\theta_{d,k}))^M] \tag{B.53}$$

$$= M_d(M_d-1)\ldots(M_d-l+1)\mathbb{E}[\theta^l_{d,k}(t+(1-t)(1-\theta_{d,k}))^{M_d-l}] \tag{B.54}$$

$$= \frac{m!}{(m-l)!}\mathbb{E}\left[\mathbb{E}\left[\theta^l_{d,k}(t+(1-t)(1-\theta_{d,k}))^{M_d-l}|\Theta_0\right]\right]. \tag{B.55}$$

Plugging the value $t=0$ in the previous expression we obtain

$$\mathbb{P}(O^{(M_d)}_{d,N_d,k}|\boldsymbol{X}^*_{\boldsymbol{N}}) = \binom{M_d}{l}\mathbb{E}\left[\mathbb{E}\left[\theta^l_{d,k}(1-\theta_{d,k})^{M_d-l}|\Theta_0\right]\right] \tag{B.56}$$

$$= \binom{M_d}{l}\int_{[0,1]}\frac{(z_{d,N_d,k}+r_d\theta)_l(N-z_{N_d,d,k}+r_d(1-\theta))_{M_d-l}}{(N_d+r_d)_{M_d}}\nu_{3\mathrm{BP}}(\mathrm{d}\theta;\alpha,\sigma,c), \tag{B.57}$$

where we are using the posterior characterization of the Beta-Bernoulli process given in Example 3, i.e. the fact that the jumps $\theta_k \sim \mathrm{Beta}(z_{N,k}-\sigma, N-z_{N,k}+c+\sigma)$.

### B.3.5 Proof of Proposition 4

We consider the PGF of $U_{\boldsymbol{N}}^{(\boldsymbol{M})}$, for $t \in \mathbb{R}$,

$$G_{U_{\boldsymbol{N}}^{(\boldsymbol{M})}}(t) = \mathbb{E}\left[t^{U_{\boldsymbol{N}}^{(\boldsymbol{M})}} | \boldsymbol{X}_{\boldsymbol{N}}^{*}\right] \tag{B.58}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\left[t^{\mathbb{1}\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} x'_{d,N_d+m,k}>0\right)} | \Theta_1, \ldots, \Theta_D\right]\right] \tag{B.59}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\left[t + (1-t)\prod_{d=1}^{D}\prod_{m=1}^{M_d}\mathbb{P}(x'_{d,N_d+m,k} = 0|\Theta_d)|\Theta_1, \ldots, \Theta_D\right]\right] \tag{B.60}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\left[t + (1-t)\prod_{d=1}^{D}(1-\theta_{d,k})^{M_d}\right]\right] \tag{B.61}$$

$$= \mathbb{E}\left[\prod_{k\geq 1}\left[t + (1-t)\prod_{d=1}^{D}\mathbb{E}\left[(1-\theta_{d,k})^{M_d}|\Theta_0\right]\right]\right], \tag{B.62}$$

$$\tag{B.63}$$

where we have used the fact that the random flips $x'_{d,n,k}$ are conditionally independent Bernoulli random variables. Now, since $\theta'_{d,k}|\theta_{0,k} \sim \text{Beta}(r_d\theta_{0,k}, N_d + r_d(1-\theta_{0,k}))$,

$$\mathbb{E}\left[(1-\theta_{d,k})^{M_d}|\Theta_0\right] = \frac{\Gamma(N_d + r_d)}{\Gamma(r_d\theta_{0,k})\Gamma(N_d + r_d(1-\theta_{0,k}))} \times \tag{B.64}$$

$$\times \int_{[0,1]} (1-\theta)^{M_d}\,\theta^{r_d\theta_{0,k}-1}(1-\theta)^{N_d+r_d(1-\theta_{0,k})-1}\mathrm{d}\theta \tag{B.65}$$

$$= \frac{(N_d + r_d(1-\theta_{0,k}))_{M_d}}{(N_d + r_d)_{M_d}}. \tag{B.66}$$

Now combining this with the Laplace functional of the Beta process,

$$G_{U_{\boldsymbol{N}}^{(\boldsymbol{M})}}(t) = \mathbb{E}\left[\exp\left\{\sum_{k\geq 1}\log\left(t + (1-t)\prod_{d=1}^{D}\frac{(N_d + r_d(1-\theta_{0,k}))_{M_d}}{(N_d + r_d)_{M_d}}\right)\right\}\right] \tag{B.67}$$

$$= \exp\left\{-\alpha(1-t)\int_{[0,1]}\left[1 - \prod_{d=1}^{D}\frac{(N_d + r_d(1-\theta))_{M_d}}{(N_d + r_d)_{M_d}}\right]\right. \tag{B.68}$$

$$\left.\frac{(N_d + r_d(1-\theta))_{M_d}}{(N_d + r_d)_{M_d}}\nu_{\text{3BP}}(\mathrm{d}\theta; \alpha, \sigma, c)\right\} \tag{B.69}$$

Following the same steps as in the previous proof, the results follows.

# Bibliography

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, pages 839–858, 2017.

M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745, 2011.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 2012.

T. Broderick, J. Pitman, and M. I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):290–306, 2015.

T. Broderick, A. C. Wilson, and M. I. Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 24(4B):3181–3221, 2018.

T. Campbell, D. Cai, and T. Broderick. Exchangeable trait allocations. *Electronic Journal of Statistics*, 12(2):2290–2322, 2018.

T. Campbell, J. H. Huggins, J. P. How, and T. Broderick. Truncated random measures. *Bernoulli*, In press.

O. Cesari, S. Favaro, and B. Nipoti. Posterior analysis of rare variants in Gibbs-type species sampling models. *Journal of Multivariate Analysis*, 131:79–98, 2014.

W. Chu, Z. Ghahramani, R. Krause, and D. L. Wild. Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *Biocomputing 2006*, pages 231–242. World Scientific, 2006.

. G. P. Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010.

. G. P. Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68, 2015.

F. Doshi-Velez and Z. Ghahramani. Correlated non-parametric latent feature models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 143–150. AUAI Press, 2009.

D. B. Dunson. Bayesian nonparametric hierarchical modeling. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):273–284, 2009.

E. E. Eichler, D. A. Nickerson, D. Altshuler, A. M. Bowcock, L. D. Brooks, N. P. Carter, D. M. Church, A. Felsenfeld, M. Guyer, and C. Lee. Completing the map of human genetic variation. *Nature*, 447(7141):161, 2007.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):993–1008, 2009.

S. Favaro, A. Lijoi, and I. Prünster. Asymptotics for a Bayesian nonparametric estimator of species variety. *Bernoulli*, 18(4):1267–1283, 2012a.

S. Favaro, A. Lijoi, and I. Prünster. A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196, 2012b.

S. Favaro, B. Nipoti, and Y. W. Teh. Rediscovery of Good–Turing estimators via Bayesian nonparametrics. *Biometrics*, 72(1):136–145, 2016.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85, 2006.

E. Fox, M. I. Jordan, E. B. Sudderth, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems*, pages 549–557, 2009.

W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, and J. Shendure. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216, 2013.

A. E. Gelfand, A. Kottas, and S. N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

S. J. Gershman, P. I. Frazier, and D. M. Blei. Distance dependent infinite latent feature models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):334–345, 2015.

Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482, 2006.

Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. *Bayesian Statistics*, 8, 2007.

S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

D. Görür, F. Jäkel, and C. E. Rasmussen. A choice model with infinitely many latent features. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 361–368. ACM, 2006.

T. L. Griffiths and Z. Ghahramani. The Indian Buffet Process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224, 2011.

S. K. Gupta, D. Phung, and S. Venkatesh. A nonparametric Bayesian Poisson gamma model for count data. In *21st International Conference on Pattern Recognition (ICPR)*, pages 1815–1818. IEEE, 2012.

J. A. Hartigan. Partition models. *Communications in Statistics-Theory and methods*, 19(8):2745–2756, 1990.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*, volume 28. Cambridge University Press, 2010.

I. Ionita-Laza and N. M. Laird. On the optimal design of genetic variant discovery studies. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

I. Ionita-Laza, C. Lange, and N. M. Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13): 5008–5013, 2009.

H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

L. F. James. Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *The Annals of Statistics*, 45(5):2016–2045, 2017.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

Y. Kim. Nonparametric bayesian estimators for counting processes. *The Annals of Statistics*, 27(2):562–588, 04 1999.

J. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1): 59–78, 1967.

J. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992.

J. Lee, P. Müller, S. Sengupta, K. Gulukota, and Y. Ji. Bayesian feature allocation models for tumor heterogeneity. In *Statistical Analysis for High-Dimensional Data*, pages 211–232. Springer, 2016.

M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, and B. B. Cummings. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25(3):244–264, 2012.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493, 2011.

J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, and L. L. Cavalli-Sforza. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866): 1100–1104, 2008.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press, 2010.

A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472): 1278–1291, 2005.

A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

D. MacArthur, T. Manolio, D. Dimmock, H. Rehm, J. Shendure, G. Abecasis, D. Adams, R. Altman, S. Antonarakis, and E. Ashley. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469, 2014.

S. MacEachern. Dependent nonparametric processes. *Proceedings of the Section on Bayesian Statistical Science*, 2000.

K. Miller, M. I. Jordan, and T. L. Griffiths. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pages 1276–1284, 2009.

K. T. Miller, T. L. Griffiths, and M. I. Jordan. The phylogenetic Indian buffet process: a non-exchangeable nonparametric prior for latent features. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 403–410. AUAI Press, 2008.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

D. J. Navarro and T. L. Griffiths. A nonparametric Bayesian method for inferring features from similarity judgments. In *Advances in Neural Information Processing Systems*, pages 1033–1040, 2007.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, and M. R. Nelson. Genes mirror geography within Europe. *Nature*, 456(7218):98, 2008.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784. ACM, 2009.

A. Papapantoleon. An introduction to Levy processes with applications in finance. *TU Vienna, Lecture Notes, arXiv preprint arXiv:0804.0482*, 2008.

J. Pitman. Combinatorial stochastic processes. *Lecture Notes in Mathematics*, 1875, 2006.

J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137, 2001.

F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574, 2003.

L. Ren, Y. Wang, L. Carin, and D. B. Dunson. The kernel beta process. In *Advances in Neural Information Processing Systems*, pages 963–971, 2011.

S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, and J.-F. Gibrat. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19(6):796–813, 2012.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61:437–455, 2010.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, pages 1838–1846, 2009.

Y. W. Teh and M. I. Jordan. *Hierarchical Bayesian nonparametric models with applications*, pages 158–207. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392, 2005.

R. Thibaux and M. I. Jordan. Hierarchical Beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571, 2007.

M. K. Titsias. The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*, pages 1513–1520, 2008.

C. Yau and C. Holmes. Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, 6(2):329, 2011.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial Intelligence and Statistics*, pages 1135–1143, 2015.

M. Zhou. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, 2018.

M. Zhou and L. Carin. Negative Binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin. Dependent hierarchical beta process for image interpolation and denoising. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 883–891, 2011.

J. Zou, G. Valiant, P. Valiant, K. Karczewski, S. O. Chan, K. Samocha, M. Lek, S. Sunyaev, M. Daly, and D. G. MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293, 2016.