Bayesian Nonparametric Methods for Non-exchangeable Random Structures

Giuseppe Di Benedetto
St. Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2020

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. My personal contributions are as outlined in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text, acknowledgements, forms and papers.

Giuseppe Di Benedetto
Hilary 2020

This thesis is dedicated to Jonida
and to my family, Annamaria, Gabriele, and Valeria.

Acknowledgements

I would like to thank my supervisors François Caron and Yee Whye Teh for their great support during these years. I am extremely grateful to François Caron and Judith Rousseau for their patience, for guiding me along this journey and for being source of inspiration. This thesis would have not been possible without their valuable help.

Thanks to Paul Vanetti, Beniamino Hadj-Amar, Jack Jewson, Nathan Cunningham, Xenia Miscouridou, Fadhel Ayed, Marco Battiston, Kaspar Martens, Francesca Panero, Marco Norzi, and to all the collegues and friends from the Oxford Statistics Department, who made these experience so rich and enjoyable. Thanks to the Badminton Club: Claire Huang, Philippe Gagnon, Fadhel Ayed, Chris Carmona, Jotun Hein, Judith Rousseau. I am thankful to Long Nguyen and his students Aritra Guha, Rayleigh Lei, and Mikhail Yurochkin, for being so welcoming during my visit to Ann Arbor.

I owe much to Jonida, for always being a source of positivity.

Abstract

Bayesian Statistics has been increasingly popular in the last five decades. Besides having decision theoretic foundation, the Bayesian approach found popularity thanks to the simple and intuitive framework that it offers to model phenomena, allowing the statistician to include prior knowledge in the inference procedure. Each unknown quantity in the model is endowed with a prior distribution which encodes our prior belief, and Bayes' rule functions as a probabilistic tool to update the prior belief after observing the data. The result of the inference scheme is the posterior distribution, which, being a probability distribution on the unknown parameters, automatically provides a principled way to quantify the uncertainty of our estimation and make prediction. The choice of the prior distribution can be crucial for the performance of the Bayesian model. In particular, restricting the prior distribution to a finite-dimensional space can lead to a model which may not fit the data properly. Very frequently in applications, in order to capture complex patterns of the data, statistical models include infinite-dimensional parameters, and in such cases the model is referred to as nonparametric.

The first part of this thesis is dedicated to Bayesian Nonparametric modelling. In particular the focus will be on modelling two discrete random structures: random partitions and random feature allocations. These two objects are the building blocks for several models used for clustering, feature modeling, network modeling, regression, density estimation. They allow to describe dataset of increasing complexity, such as collections of labelled observations with an increasing number of labels, and observations showing features with the number of features growing over time. A recurrent assumption in Bayesian modelling is the exchangeability of the data. Despite being a reasonable and convenient assumption, it leads to strict asymptotic behaviour of the model. In clustering application for instance, exchangeability in random partition models implies that the size of each cluster grows linearly with the number of obser-

vations. Analogously, in feature modelling, exchangeable feature allocation models have the number of objects sharing the same feature grow linearly with the number of objects observed. Such restrictive asymptotic properties are intrinsic in exchangeable models and are not appropriate for some applications, therefore it is necessary to go beyond exchangeability in order to design priors with a broader range of asymptotic properties. Moreover, we show that the proposed priors retain good properties already achieved by exchangeable models, suche as the power-law property, which is ubiquitous in real-world applications.

On the one hand Bayesian modelling gives freedom by allowing to choose the prior distribution which best reflects our prior belief. On the other hand, especially in nonparametric settings, the design of such priors could be challenging and one would rather opt for standard and objective priors to use in practice. One way to validate and compare priors consists in a frequentist analysis of the posterior distribution. Consistency of the posterior distribution for example, ensures that the effect of the prior knowledge vanishes as an increasing amount of data is collected, while posterior contraction rates can be used to compare different priors. In the second part of the thesis we study asymptotic properties for different priors in Functional Data Analysis. The latter has gained an increasing attention in the last decades thanks to the raise of data collected over fine grids, making it reasonable to model them as functions. While Functional Data Analysis has been thoroughly studied in the frequentist literature and widely used in Bayesian modelling, asymptotic analysis of Bayesian methods are rather scarce. We will focus on two models: Functional Linear Regression and Functional Single-Index Model, both in the setting of functional predictor and scalar response. In particular we will show that for two classes of priors used for Bayesian Functional Linear Regression, minimax posterior contraction rates can be achieved. Moreover, posterior contraction rates are derived for a class of priors for the Bayesian Functional Single-Index Models.

Contents

Chapter 1

Introduction

This thesis contains a collection of three research projects described in three main Chapters, two of which are presented in the form of papers. All the papers include literature review, bibliographic and supplementary materials. The paper in Chapter 2 is under major revision to the Annals of Statistics journal; the paper in Chapter 3 has been accepted for publication in the Proceedings of the International Conference on Artificial Intelligence and Statistics; Chapter 4 is a technical report which will be submitted in the future. This first Chapter covers some background material and gives motivation for the presented works. The last Chapter is a discussion about possible directions that could be explored and which are left for future work.

1.1 Bayesian Modeling with Discrete Random Structures

Discrete random structures play a central role in Statistics. Random partitions and random binary matrices are two remarkable examples which are the building blocks of several models used for clustering, classification, regression, dimensionality reduction, just to mention a few. In many application it is reasonable to assume that the data can be gathered into subgroups sharing specific features. The inference objective is to recover such underlying pattern in the data and predict how it can evolve as the number of observation in-

creases. Depending on the application under consideration, the subgroups we wish to unveil may be disjoint or overlapping. In the former case, the problem is referred to as clustering, while in the latter it is usually called feature allocation problem.

Random partition models

In clustering applications, observations belonging to the same subgroup are assumed to share similar characteristics, therefore they can be modelled as samples from a particular distribution, with different subgroups described by different distributions. Mixture models provide a natural way to solve the clustering. Assuming that n observations $X_1, \dots, X_n$ are gathered into $k < n$ disjoint clusters, the members of each cluster $j = 1, \dots, k$ are assumed to be sampled from a distribution $F(dX \,|\, \theta_j)$ characterised by a cluster parameter θ_j which belongs to some measurable space Θ. The distribution of the data can therefore be modelled as

$$P(dX) = \sum_{j=1}^{k} \omega_j \, F(dX \,|\, \theta_j)$$

where ω_j is the probability that the observation belongs to cluster j. Denoting the cluster label of X_i by c_i, we can write $\omega_j = \mathbb{P}(c_i = j)$. The parameters of interest in this model are the weights $(\omega_j)_{1:k}$ of the mixture, belonging to the $(k-1)$-dimentional simplex, and the kernels' parameters $(\theta_j)_{j=1:k}$. These two vectors of parameters can be interpreted as an atomic measure π over the space Θ, also referred to as mixing measure, and the mixture model can be written as follows

$$P(dX) = \int_{\Theta} F(dX \,|\, \theta) \, \pi(d\theta)$$
$$\pi(d\theta) = \sum_{j=1}^{k} \omega_j \, \delta_{\theta_j}(d\theta)$$

where δ_θ is the Dirac delta mass at θ. The space of parameters in this statistical model is the space of atomic measures with k atoms over the space Θ. For

example, if the observations' space is the real line $\mathcal{X} = \mathbb{R}$, one might consider the kernel to be a Gaussian distribution with location-scale parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Following a frequentist approach, the MLE for the parameters of the mixing measure can be found by an Expectation-Maximisation algorithm (see [Dempster et al., 1977] and [Friedman et al., 2001]). In a Bayesian framework instead, the user is allowed to place a prior distribution Π over the parameter π, and the statistical model can be described in the following hierarchical way:

$$\begin{aligned} \pi &\sim \Pi \\ \theta \,|\, \pi &\sim \pi \\ X \,|\, \theta &\sim F(\cdot \,|\, \theta). \end{aligned}$$

The weights $\boldsymbol{\omega} = (\omega_j)_{j=1:k}$ of π and its atoms' locations $\boldsymbol{\theta} = (\theta_j)_{j=1:k}$ can be chosen independently. Hence a prior Π_ω over the simplex and a prior Π_θ over Θ have to be specified, and the model takes this form

$$\begin{aligned} \boldsymbol{\omega} &\sim \Pi_\omega \\ \boldsymbol{\theta} &\sim \Pi_\theta \\ c \,|\, \boldsymbol{\omega} &\sim \text{Discrete}(\boldsymbol{\omega}) \\ X \,|\, c, \boldsymbol{\theta} &\sim F(\cdot \,|\, \theta_c). \end{aligned}$$

For instance Π_ω can be chosen to be a Dirichlet distribution, and Π_θ a Gaussian-Inverse-Gamma distribution in the Gaussian kernel example. After observing the data, the prior belief can be updated by using Gibbs sampler (see [Robert and Soubiran, 1993]). As a result, we are given an approximation of the posterior distribution over the parameters of interest, and over the clusters' membership parameters $(c_1, \ldots, c_n)$. The former are of particular interest if the objective is to approximate the distribution of data, while the latter are useful to cluster the observations.

However, in the mixture model just described, the number of components is assumed to be fixed apriori. In some situations the information about the

number of subgroups is given, but in most of the applications this quantity is unknown and has to be included in the inference scheme. Being it unknown, the number of clusters is considered to be random in the Bayesian paradigm, therefore the prior would concentrate on the product space of atomic measures with finite components ([Richardson and Green, 1997]). Nonetheless, in many applications the number of clusters is believed to grow as the number of data points increases. Under this assumption, it is more appropriate and convenient to define a prior on the atomic measures with infinite number of components leading to infinite mixture models ([Escobar and West, 1995], [Neal, 2000], [Rasmussen, 2000])

$$P(dX) = \sum_{j=1}^{\infty} \omega_j \, F(dX \mid \theta_j).$$

In this setting it is interesting to understand the properties of the prior we place on the sequence of weights of the mixing measure π. By choosing a prior, the user automatically imposes specific asymptotic behaviours on the partition structure of the data. As such, she needs to be aware of the prior properties since those should reflect her prior beliefs. In the nonparametric clustering problem it is essential to figure out how fast the number of clusters grow with the number of data points, as well as the asymptotic proportion of clusters of a given size, and how fast the clusters grow with respect to the size of the dataset. The Bayesian Nonparametrics' literature is rich with prior distributions on infinite-dimensional simplex which cover a wide range of asymptotic regimes for the partition's statistics mentioned above ([Aldous, 1985], [Pitman, 2003], [Lijoi et al., 2007], [Lijoi et al., 2005],[Lijoi et al., 2010]). In the following section we will recall more in detail the most popular Bayesian random partition models: the so called Chinese Restaurant Process and its two-parameter generalisation, the Pitman-Yor process. At the end of the section we will highlight the limitation of these models and how more flexibility can be achieved.

Feature allocation models

A generalisation of the clustering problem consists in allowing the same observation X_i to belong to multiple subgroups, which could represent features that an item possesses. Assuming there are $k \in \mathbb{N}$ possible features, the label parameter $c_i \in \{1, \dots, k\}$ is replaced by a binary vector $Z_i = (z_{ij})_{j=1:k} \in \{0,1\}^k$ indicating which features are present in X_i, and modelled as a Bernoulli random vector with parameters $\boldsymbol{p} = (p_1, \dots, p_k) \in [0,1]^k$. Therefore the binary matrix $Z = (z_{ij})_{i=1:n,\, j=1:k}$ is the discrete object that describes the features' allocation in a dataset of size n. Denoting the features' parameters by $\boldsymbol{\theta} = (\theta_j)_{j=1:k}$, a kernel which depends on the vector $Z_i \otimes \boldsymbol{\theta} = (z_{ij}\theta_j)_{j=1:k}$ describes the distribution of the observation X_i. This models are suitable for various applications such as image processing ([Dang and Chainais, 2017]), genetics ([Bernardo et al., 2003]), topic modelling ([Williamson et al., 2010], [Archambeau et al., 2014]). In the Bayesian framework, to complete the model specification, one needs to choose a prior distribution to place on the unknown Bernoulli parameters and on the parameter $\boldsymbol{\theta}$. Therefore the Bayesian factorial model, which generalises the mixture model, can be written as follows:

$$\begin{aligned} \boldsymbol{p} &\sim \Pi_p \\ \boldsymbol{\theta} &\sim \Pi_\theta \\ Z \,|\, \boldsymbol{p} &\sim \text{Bernoulli}(\boldsymbol{p}) \\ X \,|\, Z, \boldsymbol{\theta} &\sim F(\cdot \,|\, Z \otimes \boldsymbol{\theta}). \end{aligned}$$

An example of prior for the features' parameters p_j could be a Beta distribution, and it is worth noting that the vector $\boldsymbol{p}$ of features' probabilities does not need to sum to one, as in the mixture model case. As for the clustering problem, the number of features is usually unknown, and it has to be considered as a parameter to be inferred. Moreover, for some applications, the number of features increases with the number of observations. This requires to design a prior distribution on the space of binary matrices with infinite number of rows, and with each column representing the features displayed by each observation. Therefore the need of understanding the asymptotic prop-

erties of such distribution comes into play. The first requirement is sparsity: feature allocation models are often used to obtain a lower-dimensional representation of the data, and even if it is assumed that the complexity —in terms of number of features— increases with the size of the dataset, it is desirable to assume that each observation only displays a small number of features. Furthermore, as the number of data points increases, the asymptotic number of features shown by the observations, the proportions of features shared by a given number of objects, the number of objects displaying a specific feature, are of interest to design an appropriate prior for the problem under consideration. Over the years, priors for feature allocations have been studied in the Bayesian nonparametric setting. The most remarkable examples are the Indian Buffet Process and its three-parameter generalisation. Those priors will be recalled later in this chapter and their limitations will be discussed.

1.1.1 Poisson Random Measures

Most of the Bayesian nonparametric priors for discrete random structures build on Poisson random measures. We present here the main definitions and some results. For a complete and detailed description we refer to the books [Kingman, 1992], [Daley and Vere-Jones, 2007], [Ghosal and Van der Vaart, 2017].

Definition 1 (Poisson Random Measure)**.** *Let* $(\mathbb{X}, \mathcal{X})$ *a Polish space and* ν *a measure on it. A Poisson random set* Π *is subset of at most countably many point such that* $N(A) := \#(\Pi \cap A)$ *is a random variable for every* $A \in \mathcal{X}$*, and such that* $N(A_i) \overset{ind}{\sim} \text{Poi}(\nu(A_i))$ *for each collection of disjoint subsets* $A_1, \ldots, A_k \in \mathcal{X}$*. A Poisson random measure* N *with mean measure* ν *is the counting measure of* Π*.*

Poisson random measures are characterised by the Laplace functional, which can be written for every function $f : \mathbb{X} :\to \mathbb{R}$ as follows

$$\mathcal{L}_N(f \,|\, \nu) = \int_{\mathcal{M}} e^{-\int f \, dN} \mathcal{P}(dN \,|\, \nu) = \exp\left(- \int (1 - e^{-f(x)}) \nu(dx) \right)$$

where $\mathcal{M}$ is the space of measures on $(\mathbb{X}, \mathcal{X})$ and $\mathcal{P}(dN \,|\, \nu)$ is the distribution of the Poisson random measure N with mean measure ν.

Completely Random Measures

In Bayesian Nonparametrics, of particular interest are the functionals of the Poisson random measures of the following form, when $\mathbb{X} = \mathbb{R}_+ \times \Theta$:

$$\mu(d\theta) = \int \omega \, N(d\omega, d\theta) = \sum_j \omega_j \delta_{\theta_j}(d\theta)$$

for the mean measure $\nu(d\omega, d\theta) = \rho(d\omega)\alpha(d\theta)$, with α a sigma-finite and atomless measure and ρ satisfying $\int \min(1, \omega)\rho(d\omega) < \infty$. This defines a random discrete measure called *Completely Random Measure* (CRM) without fixed atoms (see [Kingman, 1967] or [Kingman, 1992] for a more general definition). CRMs are ubiquitous in Bayesian nonparametric modelling and they represent the most used tool to build models based on discrete random structures.
Some remarkable examples are the following:

- *Genearlised Gamma Process* denoted by $\mu \sim \text{GGP}(\alpha, \sigma, \zeta)$ characterised by the mean measure

  $$\nu(d\omega, d\theta) = \frac{\omega^{-1-\sigma}}{\Gamma(1-\sigma)} e^{-\zeta\omega} d\omega \, \alpha(d\theta)$$

  with α a measure on Θ, $(\sigma, \zeta) \in (-\infty, 0] \times (0, \infty)$ or $(\sigma, \zeta) \in (0, 1) \times [0, \infty)$. Noteworthy subcases are the Gamma process obtained for $\sigma = 0$ and $\zeta > 0$, the Stable process for $\sigma \in (0, 1)$ and $\zeta = 0$, and the Inverse-Gaussian process for $\sigma = 1/2$ and $\zeta > 0$.

- *Stable-Beta process* denoted by $\mu \sim \text{SBP}(\alpha, \sigma, c)$ with mean measure

  $$\nu(d\omega, d\theta) = \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \omega^{-1-\sigma}(1-\omega)^{c+\sigma-1} d\omega \, \alpha(d\theta)$$

  with α a measure on Θ, $\sigma \in [0, 1)$, $c > -\sigma$. The Beta process corresponds to the case $\sigma = 0$.

Both these examples are used to define some of the random partition models and feature allocation models popular in Bayesian Nonparametrics, and later

we will see that the case $\sigma > 0$ is particularly interesting, since it gives to these structures the so called *power-law* property.

Posterior charcterisation

The Poisson Partition Calculus, developed in [James, 2002], represents an intuitive framework to derive conditional distributions of the CRM μ given the points $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, and more generally conditional distributions of a large class functionals of Poisson random measures which includes CRMs. Therefore the goal is to derive a disintegration formula for the joint distribution of the random variables $\{\theta_1, \dots, \theta_n, \mu\}$:

$$\mathcal{P}(d\mu \,|\, \nu) \prod_{i=1}^{n} \mu(d\theta_i) = \mathcal{P}(d\mu \,|\, \boldsymbol{\theta}) M(d\theta_1, \dots, d\theta_n)$$

where $\mathcal{P}(d\mu \,|\, \boldsymbol{\theta})$ can be seen as the posterior distirbution of μ given $\boldsymbol{\theta}$, and $M(d\theta_1, \dots, d\theta_n)$ as the marginal joint measure of $\boldsymbol{\theta}$.
In particular, Proposition 3.1 in [James, 2002] provides the following disintegration formula which can be used to derive posterior distributions in models built on CRMs:

$$e^{-\int f dN} \prod_{i=1}^{n} \mu(d\theta_i) \mathcal{P}(d\mu \,|\, \nu) = P(d\mu \,|\, e^{-f}\nu, \boldsymbol{\theta}) \mathcal{L}_N(f \,|\, \nu) \prod_{j=1}^{k_n} \kappa(m_{n,j} \,|\, \rho) \alpha(d\theta_j^*)$$

where

- k_n is the number of unique elements θ_i^* in $\boldsymbol{\theta}$ and $m_{n,j} := \#\{\theta_i = \theta_j^*\}$
- $\kappa(m \,|\, \rho) := \int \omega^m \rho(d\omega)$ is the m-th moment of the measure ρ
- $P(d\mu \,|\, \nu, \boldsymbol{\theta})$ is the law of $\mu + \sum_{j=1}^{k_n} \omega_j^* \delta_{\theta_j^*}$ with fixed atoms having independent weights whose distribution $\mathbb{P}(\omega_j^* \in d\omega \,|\, \boldsymbol{\theta})$ is proportional to $\omega^{m_{n,j}} \rho(d\omega)$.

Therefore the posterior of the CRM μ given the n points $(\theta_1, \dots, \theta_n)$ is distirbuted as the CRM $\mu^* + \sum_{j=1}^{k_n} \omega_j^* \delta_{\theta_j^*}$ with fixed atoms $\theta_1^*, \dots, \theta_{k_n}^*$, where the

component μ^* is a CRM with mean measure $e^{-f}\nu$ and it is independent of the fixed atoms compoment.

1.1.2 Chinese Restaurant Process

In clustering problems, the unknown partition of the data is the object of interest, and in the Bayesian paradigm this needs to be endowed with a prior. Discrete probability measures with infinite support are useful to induce random partitions of $\mathbb{N}$, and CRMs can be used for this purpose. If μ is a CRM on a measurable space Θ with $\nu(d\omega, d\theta) = \rho(d\omega)\alpha(d\theta)$ such that $\mu(\Theta) < \infty$ almost surely and $\rho(\mathbb{R}_+) = \infty$, then it can be normalised giving a distribution $\bar{\mu} = \mu/\mu(\Theta)$ on the space of discrete probability measures on Θ with countable many atoms. Random discrete probability measures can induce random partitions on integers: taking a sequence of conditionally iid random variables

$$\begin{aligned} P &\stackrel{d}{=} \bar{\mu} \\ \theta_1, \theta_2, \dots \mid P &\overset{iid}{\sim} P \end{aligned}$$

the integers i and j are in the same cluster if and only if $\theta_i = \theta_j$ (see [Aldous, 1985], [Pitman, 1995]). The normalised Gamma process CRM with parameter $\zeta = 1$, called *Dirichlet Process* and first introduced by [Ferguson, 1973], induces the most known random partition model in Bayesian Nonparametrics: the *Chinese Restaurant Process* (CRP). The latter can be defined in numerous different ways, the simplest being through the sequence of predictive distributions. Discarding the actual values of the observations $(\theta_i)_i$, the random partition on the integers can be defined in a sequential fashion. Denoting the cluster label of the i-th observation by c_i, we have

$$\mathbb{P}(c_{n+1} = j \mid c_1, \dots, c_n) = \begin{cases} \frac{m_{n,j}}{\eta+n} & j = 1, \dots, k_n \\ \frac{\eta}{\eta+n} & j = k_n + 1 \end{cases}$$

where k_n is the number of clusters formed by the first n observations and $m_{n,j}$ their sizes for $j = 1, \dots, k_n$. Unfortunately, this simple model, containing

only one parameter, has rather strict asymptotic properties. As shown in [Korwar et al., 1973], the number of clusters grows logarithmically with the number of observations $k_n \sim \alpha(\Theta) \log n$ almost surely, and the number of clusters of size $r \geq 1$, denoted by $k_{n,r}$, is such that $k_{n,r} = o(k_n)$ as n goes to infinity ([Gnedin et al., 2007]). However, the introduction of an additional parameter $\sigma \in (0, 1)$ can provide a model with more flexible properties. The *Pitman-Yor process* ([Pitman, 2003]), also called *two-parameter CRP*, is a generalisation of the previous model and can be defined as follows

$$\mathbb{P}(c_{n+1} = j \mid c_1, \ldots, c_n) = \begin{cases} \frac{m_{n,j} - \sigma}{\eta + n} & j = 1, \ldots, k_n \\ \frac{\eta + \sigma k_n}{\eta + n} & j = k_n + 1 \end{cases}.$$

For this model, the asymptotic regime of the number of clusters is polynomial in this case and can be tuned by the parameter σ: k_n / n^σ converges almost surely to a finite random variable as n tends to infinity. Moreover, the power-law property is gained for the proportion of clusters of given sizes: $k_{n,r}/k_n \sim Cr^{-1-\sigma}$ for $r \gg 1$, $n \to \infty$ and C a positive constant, which is encountered in many real world datasets. Figure 1.1 shows two examples where the power-law behaviour for the proportion of clusters of given sizes appears. The first dataset is a collection of movies' reviews from the Amazon website, which are clustered according to the movies they correspond to. The second dataset contains the answers in the Mathoverflow blog, clustered according to the questions they refer to.

1.1.3 Indian Buffet Process

As discussed above, nonparametric feature allocation models can be described by a sequence of Bernoulli parameters $(\omega_j)_j \in [0, 1]^\infty$, and the features's allocation $Z_i = (z_{ij})_j$ for the i-th observation by a sequence of independent Bernoulli random variables with parameters $(\omega_j)_j$. By definition, the Stable-Beta Process $\mu = \sum_j \omega_j \delta_{\theta_j}$ has random weights in $\omega_j \in [0, 1]$, therefore it can be adopted as a prior on the sequence of Bernoulli parameters and on the features' values $\theta_j \in \Theta$. Therefore the features of the i-th observation can be

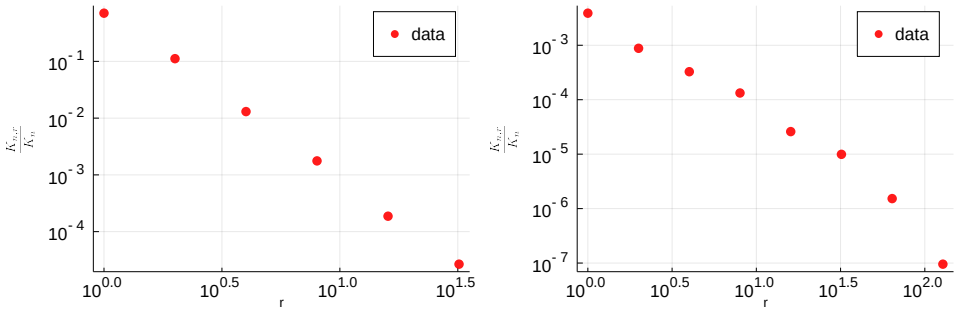


Figure 1.1: Left: Log-log plot of the proportion of movies with a given number of reviews in the Amazon dataset. Right: Log-log plot of the proportion of questions with a given number of answers in the Mathoverflow dataset.

conveniently described by conditionally iid random measures $Z_i = \sum_j z_{ij}\delta_{\theta_j}$ over the features' space Θ. Hence the Bayesian features' allocation model can be written as follows:

$$\begin{aligned} \mu = \sum_j \omega_j \delta_{\theta_j} &\sim \text{SBP}(\alpha, \sigma, c) \\ z_{ij} \,|\, \mu &\overset{ind}{\sim} \text{Bernoulli}(\omega_j) \\ Z_i &\overset{d}{=} \sum_j z_{ij}\delta_{\theta_j}. \end{aligned}$$

Posterior distribution can be derived using the Poisson Partition Calculus ([James, 2017]) and the predictive distribution allows to define the process in a sequential manner: the first object displays a Poisson number of features: $Z_1(\Theta) = \sum_j z_{1j} \sim \text{Poi}(\eta)$, while for the subsequent observations the following holds:

$$Z_{n+1} \,|\, Z_1, \ldots, Z_n \overset{d}{=} Z^*_{n+1} + \sum_{j=1}^{k_n} z_{n+1,j}\delta_{\theta_j}$$

with Z^*_{n+1} and $z_{n+1,j}$'s independent and such that

$$Z^*_{n+1}(\Theta) \sim \text{Poi}\left(\alpha(\Theta)\frac{\Gamma(1+c)\Gamma(n+c+\sigma)}{\Gamma(n+1+c)\Gamma(c+\sigma)}\right)$$
$$z_{n+1,j} \sim \text{Bernoulli}\left(\frac{m_{n,j}-\sigma}{n+c}\right)$$

where k_n is the number of features appeared in the first n observations, and $m_{n,j} = \sum_{i=1}^{n} z_{ij}$ is the number of objects showing the j-th feature. This feature model is referred to as *three-parameter Indian Buffet Process* ([Teh and Gorur, 2009]), which is a generalisation of the *Indian Buffet Process* (IBP) ([Griffiths and Ghahramani, 2011], [Ghahramani and Griffiths, 2006], [Thibaux and Jordan, 2007]) which is recovered for $\sigma = 0$. As in the CRP for clustering, the introduction of the parameter σ adds interesting asymptotic properties to model, namely the number of features k_n grows as n^σ, and the proportions of features shared by $r \geq 1$ items is asymptotically equivalent to $r^{-1-\sigma}$, which corresponds to a power-law behaviour. In Figure 1.2 we can see it appearing in two datasets. The first dataset lists the actors who played in the movies present in the IMDB website, while the second one links the authors to the corresponding papers in the ArXiv website.

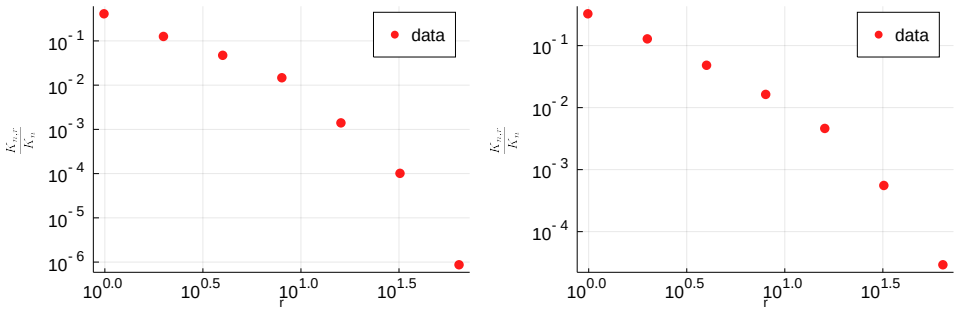


Figure 1.2: Left: Log-log plot of the proportion of actors who played in a given number of movies in the IMDB dataset. Right: Log-log plot of the proportion of authors who wrote a given number of papers in the ArXiv dataset.

1.1.4 Beyond Exchangeability

All the models reviewed in the previous sections are constructed by modelling the observations as conditionally independent. It is easy to show that this implies exchangeability, in other words, the distribution of the data is invariant under permutations. Remarkably, the converse also holds, as stated in the renowed de Finetti's Theorem ([De Finetti, 1937]).

Theorem 2 (de Finetti)**.** *The sequence or random variables $(\theta_i)_i$ is exchangeable if and only if there exists a probability measure Q on the space $\mathcal{P}_\Theta$ of the probability measures on $(\Theta, \mathcal{E})$, called* de Finetti *measure, such that for any $n \geq 1$ and any measurable sets $\{A_i\}_{i=1:n} \subset \mathcal{E}$*

$$\mathbb{P}(\theta_1 \in A_1, \ldots, \theta_n \in A_n) = \int_{\mathcal{P}_\Theta} \prod_{i=1}^{n} \mu(A_i)\, Q(d\mu).$$

For example, in the CRP random partition model the Dirichlet Process is the de Finetti measure, while in the IBP feature allocation model the Beta process is the mixing measure in the display above.
However, if on the one hand the exchangeability justifies the Bayesian approach thanks to the de Finetti Theorem and proves to be convenient when it comes to inference, on the other hand it represents a strong modelling assumption, which limits the flexibility of the models. In particular, in two models of interest, the exchangeability assumption limits the asymptotic properties that the model can capture.
About random partition models, Kingman's representation Theorem ([Kingman, 1978], [Pitman, 1995]) states that any exchangeable random partition model can be described by an exchangeable sequence of random variables, in the same way the CRP has been introduced above. By the strong law of large number, the conditionally independent structure implies that the size of each cluster grows linearly with the number of observations: conditionally on the random measure

$$m_{n,j} \sim \omega_j\, n \qquad a.s.$$

as n tends to infinity, with ω_j being the j-th cluster's frequency. Such be-

haviour is not suitable for modelling datasets where the number of clusters is assumed to grow but with each cluster growing slowly or remaining relatively small. As pointed out in [Miller et al., 2015], linear growth of clusters' sizes would not be appropriate for applications like entity resolution, where the goal is to cluster corrupted copies of the same record —or more in general observations relating to the same entity—, and usually the size of the clusters grow very slowly.

Using the same datasets described in Section 1.1.2, we can illustrate how the linearity in the clusters' sizes growth is not appropriate. In Figure 1.3 two examples of clusters' sizes trajectories are shown for the two real-world datasets. In the first one movies' reviews from Amazon website are clustered according to the movie they refer to, while in the second one answers to questions in the Mathoverflow blog are clustered according to the question they answer to. Datapoints are provided with a temporal ordering, and it is evident that the growth of the clusters is sublinear.

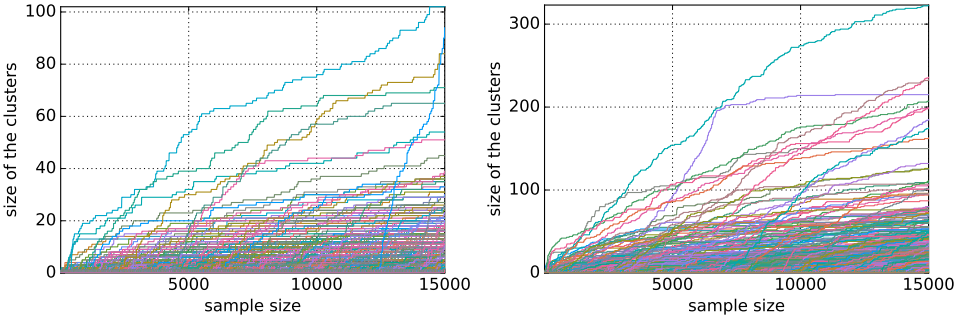


Figure 1.3: Left: number of reviews per movie with respect to the number total number of reviews in the Amazon dataset. Right: number of answer to different questions with respect to the total number of answers in the Mathoverflow dataset.

Analogously, the same phenomenon occurs for exchangeable feature allocation models, where the number of items sharing a particular feature, also referred to as feature's size, grows linearly with the number of items observed. However, such property might not fit some applications, where the growth is instead sublinear. In Fig 1.4 the features' sizes in the two datasets presented in Section 1.1.3 are considered. In the first one, data are collected from the

IMDB movies' website consisting of lists of actors who played in movies. In the feature allocation paradigm, movies are considered as observations and actors as features. Analogously, in the second dataset from the ArXiv website, the publications are considered observations displaying their authors as features. It is reasonable to assume, and as also evident from Fig 1.4, that the features' sizes grow sublinearly with the size of the data.

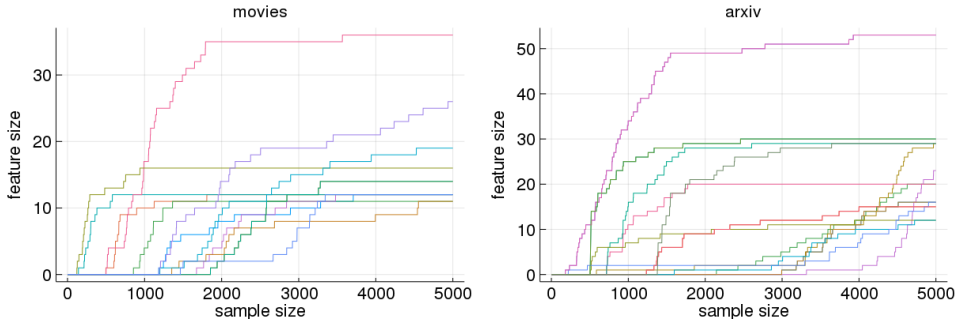


Figure 1.4: Left: Number of movies per actor in the IMDB dataset. Right: Number of articles per author in the ArXiv dataset.

In order to gain more flexibility in these discrete random structures, the exchangeability assumption needs to be dropped. This motivates the works presented in Chapters 2 and 3, where non-exchangeable models are designed to achieve sublinearity in clusters' and features' sizes, without giving up the nice asymptotic properties captured by exchangeable models, such as the power-law properties.

1.2 Frequentist analysis of Bayesian methods

Despite the Bayesian approach makes it possible to incorporate prior knowledge into the model, it could be the case that the statistician lacks of such prior information, but still being willing to adopt a Bayesian procedure, which conveniently provides a posterior probability distribution over the parameter as the result of the inference scheme. In this situations one would rather opt for a default prior that can be used in the model. However, especially in the nonparametric setting, specifying a prior could lead to some undesired

consequences ([Diaconis and Freedman, 1986]). As more and more observations are collected, it is desirable that the prior influence vanishes, so that two Bayesians, starting with different priors, will obtain the same result as the number of observations grows —this is usually referred to as "merging of opinions" property. One way to validate the Bayesian approach is to perform a frequentist analysis: assuming there is a true value of the parameter of interest $\theta_0 \in \Theta$, we can ask if the posterior distribution is able to recover it. This property is referred to as *posterior consistency*. Let $(X^n, P_\theta : \theta \in \Theta)$ be a statistical model with measurable parameter space $(\Theta, \mathcal{E})$, assuming that the distribution of the observations $X^n = (X_1, \ldots, X_n)$ is dominated with density $p_\theta = dP_\theta/d\lambda$ for some measure λ on the observations' space. Denoting by Π the prior distribution over Θ, the posterior mass on a measurable set $B \in \mathcal{E}$ can be written as

$$\Pi(B \,|\, X^n) = \frac{\int_B p_\theta(X^n)\Pi(d\theta)}{\int_\Theta p_\theta(X^n)\Pi(d\theta)} \tag{1.1}$$

and denoting by d a loss function, the posterior distribution is said to be consistent with respect to d if for every $\epsilon > 0$

$$\mathbb{E}_0^{(n)}\Pi(d(\theta, \theta_0) > \epsilon \,|\, X^n) \overset{n\to\infty}{\longrightarrow} 0$$

where $\mathbb{E}_0^{(n)}$ denotes the expectation with respect to the true distribution of the data. Doob's Theorem ([Doob, 1949]) proves under very mild assumptions, that posterior distributions are consistent at θ_0 for Π-almost every θ_0. This result might sound reassuring, and indeed it is if the parameter space Θ is finite or countable. However, in the infinite-dimensional setting, null-sets under the prior can be quite large in a topological sense, and ensuring consistency at every point apart from a prior null-set is not good enough. As a matter of fact, it is possible to find examples where nonparametric priors lead to non-consistent posterior distributions (see for example [Freedman, 1963] and [Diaconis and Freedman, 1986]).

1.2.1 Posterior contraction rates

In 1965 Schwartz ([Schwartz, 1965]) paved the way for the frequentist analysis of nonparametric Bayesian models, providing sufficient conditions for posterior consistency. In the last two decades, the seminal papers [Ghosal et al., 2000] and [Ghosal et al., 2007] strenghtened Schwartz's conditions in order to study at which speed posterior distributions converge to the true parameter value. The general posterior contraction rates result is recalled here, together with a sketch of the proof:

Theorem 3 (Posterior constraction rates)**.** *Let $(\epsilon_n)_n$ be a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. If there exists a constant $c > 0$ such that:*

- *Kullback-Leibler condition: $\Pi(S_n) \geq e^{-cn\epsilon_n^2}$ with*

$$S_n = \{\theta : KL_n(\theta_0, \theta) \leq n\epsilon_n^2,\ V_n(\theta_0, \theta) \leq n\epsilon_n^2\}$$

  with $KL_n(\theta_0, \theta)$ and $V_n(\theta_0, \theta)$ being the first and second moments of the loglikelihood ratio $\log(p_{\theta_0}/p_\theta)(X^n)$ with respect to P_{θ_0}.

- *Test condition: there exists a sequence of test functions $(\phi_n)_n$ such that*

$$\mathbb{E}_0^{(n)}[\phi_n] = o(1), \qquad \sup_{\theta : d(\theta,\theta_0) > \epsilon_n} \mathbb{E}_\theta^{(n)}[1 - \phi_n] = o(e^{-(c+2)n\epsilon_n^2})$$

Then $\mathbb{E}_{\theta_0}\Pi(d(\theta, \theta_0) > \epsilon_n \,|\, X^n) \xrightarrow{n} 0$.

Proof. By multiplying the numerator in Equation 1.1 by $(\phi_n + 1 - \phi_n)$, the expectation of the posterior mass outside the ϵ_n ball around θ_0 can be bounded as follows

$$\mathbb{E}_0^{(n)}\Pi(d(\theta, \theta_0) > \epsilon_n \,|\, X^n) = \mathbb{E}_0^{(n)}[\phi_n(X^n)] + \mathbb{E}_0^{(n)}[(1 - \phi_n(X^n))\Pi(d(\theta, \theta_0) > \epsilon_n \,|\, X^n)]$$

where the first term in the bound goes to zero by assumption. Using equa-

tion 1.1, the second term can be written as

$$\mathbb{E}_0^{(n)}\left[\frac{\int_{\{d(\theta,\theta_0)>\epsilon_n\}}(1-\phi_n(X^n)p_\theta(X^n)/p_{\theta_0}(X^n)\Pi(d\theta)}{\int_\Theta p_\theta(X^n)/p_{\theta_0}(X^n)\Pi(d\theta)}\right].$$

By splitting the integral in the numerator over the set $A_n := \{\int_\Theta p_\theta(X^n)/p_{\theta_0}(X^n)\Pi(d\theta) > e^{-2cn\epsilon_n^2}\}$ and its complement, and applying Fubini's Theorem, we can upper bound the previous quantity by

$$e^{2cn\epsilon_n^2}\sup_{\theta:d(\theta,\theta_0)>\epsilon_n}\mathbb{E}_0^{(n)}[1-\phi_n]+P_0^{(n)}(A_n^c).$$

$P_0^{(n)}(A_n^c)\overset{n}{\to}0$ by applying Markov's inequaliy (see Lemma 10 in [Ghosal et al., 2007]), while the remaining term goes to zero by assumption. □

The first condition ensures that the prior assigns enough mass around a Kullback-Leibler neighbourhood of the true parameter θ_0, while the second condition requires the constructions of test functions testing $H_0:\theta=\theta_0$ versus the alternative $H_1:d(\theta,\theta_0)>M\epsilon_n$. Usually the test condition is unlikely to be satisfied on the whole parameter space, therefore it can be replaced by two requirements. The first one consists in defining test functions testing $H_0:\theta=\theta_0$ versus balls of alternatives $H_1:d(\theta,\theta_1)<\xi\epsilon_n$ with $\xi\in(0,1)$ and θ_1 such that $d(\theta_0,\theta_1)\geq\epsilon_n$. The second one is the definition of an increasing subset of the parameter space $\Theta_n\subset\Theta$, usually referred to as sieve, such that the prior assigns most of the probability mass to it, namely $\Pi(\Theta_n^c)\leq e^{-3cn\epsilon_n^2}$, and such that its complexity can be controlled in terms of the covering number, in particular the number $N(\xi\epsilon_n,d,\Theta_n)$ of d-balls of radius $\xi\epsilon_n$ needed to cover Θ_n can be bounded by $e^{2cn\epsilon_n^2}$.
This version of the Theorem will be recalled in Appendix A, and based on this result, in Chapter 4 posterior contraction rates will be derived for different nonparametric priors for Functional Linear Regression (FLR) and Functional Single-Index Model (F-SIM), both with functional predictor and scalar respose.

1.3 Contributions and Thesis Outline

The thesis is organised into three main Chapters: Chapters 2 and 3 contain two papers which represent the contribution to Bayesian Nonparametric modelling. Chapter 4 contains the results obtained on Bayesian Asymptotics for Functional Data Analysis. All the papers included in the chapters contain supplementary material and bibliography, and for all of them I am the main contributor on both modeling, theoretical and application aspects. In Chapter 5 we summarise the results and discuss further development left for future work. As follows, the methodology adopted and the contribution to each topic are briefly summarised.

1.3.1 Bayesian Nonparametric Methods for Microclustering

The paper [Di Benedetto et al., 2017] included in Chapter 2 defines a novel class of non-exchangeable random partition models. This work is motivated by the microclustering problem, first introduced in [Miller et al., 2015], where slow growth of clusters' sizes are required for specific applications such as entity resolution. In order to satisfy this requirement, the exchangeability assumption has to be dropped, since it would imply a linear growth of the clusters' sizes. In the proposed non-exchangeable model, the distribution of the random partition is induced by a latent Cox process on the space $\mathbb{R}_+ \times \mathbb{R}_+$: the first n points sampled from the point process, ordered according to their first coordinate, form a continuous time embedding of the partition of the first n observations. This construction is inspired by the poissonisation technique used in Probability (see [Gnedin et al., 2007]): defining the model in continuous time provides a convenient framework to derive its asymptotic properties. As a matter of fact, asymptotic results for the number of clusters, the proportions of clusters of given size, and the clusters' sizes are derived for the proposed model. The techniques used to prove these results rely on the properties of the CRMs —which are used to define the mean measure of the Cox process and give the nonparametric nature to the model—, and results

about regularly varying functions.
The main contribution of this work consists in designing a class of random partition models which simultaneously satisfy the projectivity property —which is desirable to make prediction on the evoultion of the partition—, the microclustering property, and the power-law property. Moreover, two interpretable parameters of the model allow to independently tune the asymptotic regimes of the quantities characterising the last two properties. A secondary contribution consists in a novel representation of the Chinese Restaurant Process, which can be derived from the point process defined in our work.
The inference scheme follows an empirical Bayes approach. The parameters of the model are estimated by the MLEs, approximated by Sequential Monte Carlo algorithm which provides an approximation of the distribution of latent point process as well. For clustering application, when the partition of the data is unknown, a likelihood for the observations can be added to the model, and the cluster labels' distribution can be estimated by the Sequential Monte Carlo sampler. Experiments on both synthetic and real data show that the non-exchangeable model can provide a better prediction performance when compared to the exchangeable two-parameter Chinese Restaurant Process.

The paper is under major revision to the Annals of Statistics journal. The ArXiv preprint is referenced as follows.

- **Di Benedetto, G.**, Caron, F. and Teh, Y.W., 2017. Non-exchangeable random partition models for microclustering. arXiv preprint arXiv:1711.07287.

1.3.2 Bayesian Feature Allocation Models

Chapter 3 includes the paper [Benedetto et al., 2020], which presents a Bayesian non-exchangeable feature allocation model. A feature allocation is a collection of possibly overlapping clusters which generalises the definition of partition, allowing the same object to belong to multiple clusters. Feature allocation models are therefore useful to describe a collection of objects sharing a certain number of features. In the nonparametric setting the number of features is assumed to grow with the size of the dataset. It is therefore interesting to

study the asymptotic properties of the quantities that characterise the model, such as the total number of features, the number of objects sharing a particular feature —also called feature's size—, and the proportions of features shared by a given number of objects. The goal of this work is to extend the currently known spectrum of asymptotic regimes for these statistics. In particular, the model we propose satisfies the power-law property for the proportions of features shared by a given number of objects, which is very frequently observed in real data and was already satisfied by the three-parameter Indian Buffet Process introduced by [Teh and Gorur, 2009]. Besides, sublinear rates for the features's sizes can be captured by the non-exchangeable model and tuned by one of its parameters. This property represents the main contribution of this work, and it provides flexibility for modeling data such as bipartite graphs with temporal ordering. In such cases, the non-exchangeable nature of the model provides a better prediction of the evolution of the graph compared to exchangeable alternatives. This is empirically shown in the paper with experiments on both synthetic and real data.
The proposed feature allocation model builds on a latent point process construction which is similar to the one introduced in [Di Benedetto et al., 2017], and it includes the three-parameter Indian Buffet Process as a special case. CRMs have a central role in the definition of the model, with their parameters tuning the asymptotic properties of interest. Results on CRMs and regularly varying functions are used to prove the results presented in the paper. Consistent estimators are derived for the parameters of the model that tune the asymptotic properties, while a Gibbs algorithm is adopted to sample from the posterior distribution of the CRM weights, which are used to make prediction.

The paper has been accepted for publication in the Proceedings of the International Conference on Artificial Intelligence and Statistics 2020. The ArXiv preprint is referenced as follows.

- **Di Benedetto, G.**, Caron, F. and Teh, Y.W., 2020.Non-exchangeable feature allocation models with sublinear growth of the feature sizes. arXiv preprint arXiv:2003.13491.

1.3.3 Asymptotic results in Bayesian Functional Data Analysis

In Chapter 4 we present theoretical results on posterior contraction rates in Bayesian Functional Data Analysis (FDA). Frequentist analysis of the Bayesian procedures, especially in the nonparametric setting, has become extremely popular in the recent years, and it has provided a framework to validate and compare different priors. This work presents two main contributions to the field of Bayesian Asymptotics. Firstly, we study the Bayesian Functional Linear Regression problem with functional predictor and scalar response, which represents the simplest and plausibly the most popular model in applications related to FDA. Posterior contraction rates are derived for two well-known classes of nonparametric priors on the functional slope parameter, namely the sieve priors and the Gaussian process priors. The regression model is studied as an inverse problem, and from this perspective it is interesting to find posterior contraction rates for the loss functions corresponding to the direct and the inverse problems. Up to our knowledge, the only result about posterior contraction rates in Bayesian Functional Linear Regression can be found in [Lian et al., 2016], where the authors solve the direct problem for the Gaussian process prior. We extend their result by showing that minimax posterior contraction rates can be achieved in the inverse problem as well, and for both the two priors mentioned above.

The second contribution resides in asymptotic results for the Bayesian Functional Single-Index model with functional predictor and scalar response introduced by [Müller et al., 2005]. This model could be viewed as the first step towards nonlinearity in FDA, and it is extremely popular in the parametric setting, thanks to the fact that it does not suffer from the so called *curse of dimensionality*, typical of the general nonlinear regression models. While this and several other functional models have been studied in depth in the frequentist literature, asymptotic theory of the Bayesian approaches to FDA has been left rather unexplored. Our work therefore provides the first results about posterior contraction rates for Functional Single-Index models. We shown that the one of the priors studied for Functional Linear Regression

can be used here as priors for the functional index parameter, while a hybrid location-scale mixture of normals prior ([Naulet and Rousseau, 2017]) can be chosen for the link function.
The method used to prove the results contained in this Chapter relies on the standard Theorem developed in [Ghosal et al., 2007], while the inverse problem in the Functional Linear Regression model is solved by using the sieve strategy described by [Knapik and Salomond, 2014].

Chapter 2

Non-exchangeable random partition models for microclustering

Under Major Revision in the Annals of Statistics journal

Submitted to the Annals of Statistics

NON-EXCHANGEABLE RANDOM PARTITION MODELS FOR MICROCLUSTERING

By Giuseppe Di Benedetto*, François Caron* and Yee Whye Teh*,†

*University of Oxford** *and Google Deepmind*†

Many popular random partition models, such as the Chinese restaurant process and its two-parameter extension, fall in the class of exchangeable random partitions, and have found wide applicability in various fields. While the exchangeability assumption is sensible in many cases, it implies that the size of the clusters necessarily grows linearly with the sample size, and such feature may be undesirable for some applications. We present here a flexible class of non-exchangeable random partition models which are able to generate partitions whose cluster sizes grow sublinearly with the sample size, and where the growth rate is controlled by one parameter. Along with this result, we provide the asymptotic behaviour of the number of clusters of a given size, and show that the model can exhibit a power-law behavior, controlled by another parameter. The construction is based on completely random measures and a Poisson embedding of the random partition, and inference is performed using a Sequential Monte Carlo algorithm. Experiments on real datasets emphasize the usefulness of the approach compared to a two-parameter Chinese restaurant process.

1. Introduction. Random partitions arise in a wide range of different applications such as Bayesian model-based clustering [43, 56], population genetics [40], ecology [47] or network modelling [7]. A partition of a set $[n] = \{1, \dots, n\}$ is a set of disjoint non-empty subsets $A_{n,j} \subseteq [n]$, $j = 1, \dots, K_n$ with $\cup_j A_{n,j} = [n]$ where $K_n \leq n$ is the number of clusters and $A_{n,j}$ denotes the set of integers in cluster j. A random partition Π_n of $[n]$ is a random variable taking values in the finite set of partitions of $[n]$. A random partition of $\mathbb{N}$ is a sequence $\Pi = (\Pi_n)_{n\geq 1}$ of random partitions of $[n]$, defined on a common probability space, that satisfy the Kolmogorov consistency condition: for every $1 \leq m < n$, Π_n restricted to $[m]$ is Π_m [39, 1, 61, 59]. For many applications, it is important to characterize the properties of the random partition model as the number of items n grows. Of particular importance are the asymptotic behavior of (i) the number of clusters, (ii) the proportion of clusters of a given size, and (iii) the cluster sizes.

In some contexts a natural and useful assumption is the *exchangeability* of the random partition: Π is said to be exchangeable if for every $n \geq 1$ the

distribution of Π_n is invariant to the group of permutations of $[n]$. Arguably the best known exchangeable random partition model is the Chinese Restaurant Process (CRP) [1]. This model has a single parameter, a very simple generative process and well established asymptotic properties; the number of clusters K_n grows logarithmically with n [42], while the proportion of clusters of any given size goes to zero. Such behaviour is not appropriate for some applications such as natural language processing or image segmentation [67, 66], where these proportions typically exhibit a power-law behavior. This asymptotic property can be achieved by considering the two-parameter CRP [62], another exchangeable random partition model which generalizes the one-parameter CRP. Beyond these two popular models, the class of exchangeable random partitions offers a rich, flexible and tractable framework, including models based on normalised random measures [64, 33, 48, 35], Poisson-Kingman processes [60] or Gibbs-type priors [31, 25, 20, 3].

Although exchangeability is a sensible assumption in many applications, it has strong implications regarding the growth rate of the cluster's sizes: Kingman's representation theorem indeed implies that the size of each cluster grows linearly with the sample size n. As recently noted by Miller et al. [55] and Betancourt et al. [68], this assumption may be unrealistic for some applications, such as entity resolution, which require the construction of random partition models where the cluster sizes grow sublinearly with the sample size; Miller et al. [55] call it the *microclustering* property.

The objective of this article is to present a general class of models for non-exchangeable random partitions of $\mathbb{N}$ which retains the wide range of asymptotic properties of exchangeable partition models, while capturing the microclustering property. The model allows:

- Flexibility in the asymptotic growth rates of (i) the number of cluster, and (ii) the proportion of clusters of a given size, tuned by interpretable parameters; in particular, it is possible to obtain the same growth rates as with the two-parameter CRP, including the power-law regime.
- Flexibility in the asymptotic sublinear growth rates of the cluster sizes, tuned by interpretable parameters.

The paper is organized as follows. In Section 2 we provide background on completely random measures (CRM), exchangeable random partitions, and give a derivation of the partition associated to a normalised completely random measure via a Poissonization technique. In Section 3 we present our novel class of non-exchangeable random partition models that builds on the same Poissonization idea. Section 4 develops the properties of this class of models and posterior inference. In Section 5, we describe how our

model can also be used to build sparse random multigraph models with an asymptotic power-law degree distribution and sublinear degree growth. Section 6 discusses related approaches in the literature. Section 7 provides comparisons between the proposed non-exchangeable model and the two-parameter CRP on two datasets. Most proofs and some definitions can be found in the Appendix.

2. Background material.

2.1. *Completely random measures.* Completely random measures, introduced by Kingman [38], have found wide applicability as priors over functional spaces in Bayesian nonparametrics [64, 50], due to their flexibility and tractability; the reader can refer to [19, Chapter 10.1] or [50] for an extended coverage. A homogeneous CRM on $\mathbb{R}_+$ without fixed atoms nor deterministic component is almost surely discrete and takes the form

$$W = \sum_{j \geq 1} \omega_j \, \delta_{\vartheta_j}$$

where $\{(\omega_j, \vartheta_j)\}_{j \geq 1}$ are the points of a Poisson process on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\nu(d\omega, d\theta)$. The measure decomposes as $\nu(d\omega, d\theta) = \rho(d\omega)\alpha(d\theta)$ where α is a non-atomic Borel measure on $\mathbb{R}_+$, called the base measure, such that $\alpha(A) < \infty$ for any bounded Borel set A, and ρ is a Lévy measure on $(0, \infty)$. We write $W \sim \mathrm{CRM}(\alpha, \rho)$. We will also assume in the following that the base measure $\alpha(d\theta)$ is absolutely continuous with respect to the Lebesgue measure with density $\tilde{\alpha}$ and

$$\int_{(0,\infty)\times\mathbb{R}_+} \rho(dw)\alpha(d\theta) = \infty. \tag{2.1}$$

Let

$$\psi(t) = \int_0^\infty \left\{1 - e^{-wt}\right\} \rho(dw) \tag{2.2}$$

be the Laplace exponent and define, for any integer $m \geq 1$ and any $u > 0$

$$\kappa(m, u) = \int_0^\infty \omega^m e^{-u\omega} \rho(d\omega).$$

A remarkable example of CRM is the generalized gamma process [8] (GG) with mean measure

$$\nu(d\omega, d\theta) = \frac{1}{\Gamma(1 - \sigma_0)} \omega^{-1-\sigma_0} e^{-\zeta_0 \omega} d\omega \, \alpha(d\theta)$$

with $\sigma_0 \in (0,1)$ and $\zeta_0 \geq 0$ or $\sigma_0 \in (-\infty, 0]$ and $\zeta_0 > 0$. We write $W \sim \mathrm{GG}(\alpha, \sigma_0, \zeta_0)$. The GG has been a popular model in Bayesian nonparametrics due to its flexibility and attractive conjugacy properties [33, 49, 48, 14]. It includes several important models as special cases: the gamma process for $\sigma_0 = 0$, $\zeta_0 > 0$; the inverse gaussian process for $\sigma_0 = 1/2$, $\zeta_0 > 0$ and the stable process for $\sigma_0 \in (0,1)$ and $\zeta_0 = 0$.

2.2. *Exchangeable random partitions.* For an exchangeable partition $\Pi = (\Pi_n)_{n\geq 1}$ of $\mathbb{N}$ we have, for every $n \geq 1$

$$\mathbb{P}(\Pi_n = \{A_{n,1}, \ldots, A_{n,k_n}\},\, K_n = k_n) = p(|A_{n,1}|, \ldots, |A_{n,k_n}|)$$

where the sets $A_{n,j}$ are considered in order of appearance and p is a symmetric function of its arguments called *exchangeable partition probability function* (EPPF). Therefore, by definition, the ordering in which we observe the data is not taken into account and the only information that affects the distribution of the random partition is the size of the clusters. For an infinite sequence of random variables $(\theta_{(1)}, \theta_{(2)}, \ldots)$ taking values in $\mathbb{R}_+$, let $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ be the random partition of $\mathbb{N}$ defined by the equivalence relation "i and j are in the same cluster" if and only if $\theta_{(i)} = \theta_{(j)}$ [59]. By Kingman's representation theorem [39], every exchangeable random partition has the same distribution as $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$, where the random variables $\theta_{(1)}, \theta_{(2)}, \ldots$ are conditionally independent and identically distributed from some random probability distribution $\mathbb{P}$.

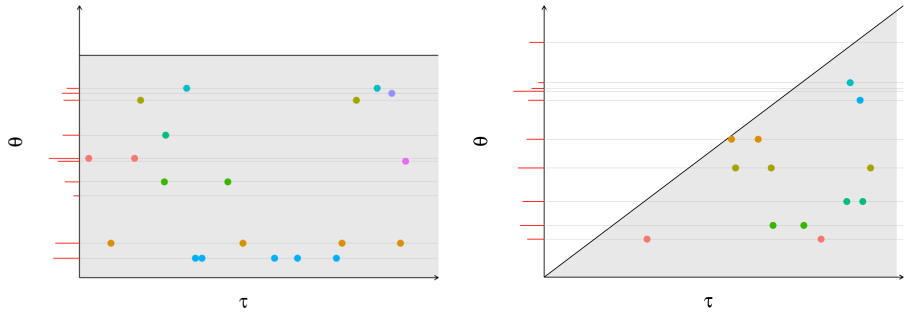
A popular model for this random probability distribution is a normalised completely random measure [64, 34, 35], defined as $P = W/W(\mathbb{R}_+)$, where $W \sim \mathrm{CRM}(\rho, \alpha)$ and $\alpha(\mathbb{R}_+) < \infty$. This condition, together with the condition (2.1), ensures that $0 < W(\mathbb{R}_+) < \infty$ almost surely, and the model is thus properly defined.

2.3. *Continuous-time embedding of exchangeable random partitions via Poissonization.* Let $(\theta_{(1)}, \theta_{(2)}, \ldots)$ be an infinite sequence of random variables taking values in $\mathbb{R}_+$ and $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ be the random partition of $\mathbb{N}$ defined by the equivalence relation "i and j are in the same cluster" if and only if $\theta_{(i)} = \theta_{(j)}$. Let $0 < \tau_{(1)} < \tau_{(2)} < \ldots$ be an infinite sequence of arrival times. Define the continuous-time partition-valued process $(\Pi(t))_{t\geq 0}$ as

$$\Pi(t) := \Pi_{N(t)} = \Pi(\theta_{(1)}, \ldots, \theta_{(N(t))}), \quad t \geq 0$$

where $N(t) = \sum_i 1_{\tau_{(i)} \leq t}$ with $1_{\tau \leq t} = 1$ if $\tau \leq t$ and 0 otherwise. $(\Pi(t))_{t\geq 0}$ defines a continuous-time embedding of the partition Π. Note that we have

$$\Pi_n = \Pi(\tau_{(n)}) \quad n \geq 1.$$



(a) Poisson embedding of an exchangeable random partition

(b) Poisson embedding of the non-exchangeable random partition

FIG 1. *(a) The Chinese restaurant process obtained via a Poisson embedding. Points* (τ_i, θ_i) *are drawn from a Poisson point process on* $\mathbb{R}_+ \times [0,1]$ *with mean measure* $\mu(d\tau, d\theta) = d\tau\, W(d\theta) 1_{\theta \leq 1}$ *where* W *is a gamma random measure with base measure* $\alpha(d\theta) = \alpha_0 d\theta$ *and* $\zeta = 1$*. The red sticks on the* θ*-axis represent the jumps of the Gamma random measure* W*. Points on the same horizontal line are in the same cluster. The random partition* $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ *of* $\mathbb{N}$ *induced by the sequence of points* $\theta_{(1)}, \theta_{(2)}, \ldots$ *ordered by their arrival times* $\tau_{(1)} < \tau_{(2)} < \ldots$ *is the CRP. (b) Non-exchangeable random partition model via a Poisson embedding. Points* (τ_i, θ_i) *are drawn from a Poisson point process on* $\mathbb{R}_+ \times \mathbb{R}_+$ *with mean measure* $\mu(d\tau, d\theta) = d\tau\, W(d\theta) 1_{\theta \leq \tau}$ *where* W *is a CRM. Note that the CRM is not normalised. The red sticks on the* θ*-axis represent the jumps of the CRM* W*. Points on the same horizontal line are in the same cluster.*

A remarkable feature of exchangeable random partitions is that they admit a continuous-time embedding via a Poisson process. Poissonization is a classical technique used in combinatorial problems in order to derive analytical properties of exchangeable partitions and urn schemes [37, 30, 9].

We focus on the important case where the partition is obtained from a normalised CRM [64, 57, 45, 46, 48, 35], which includes as a special case the one-parameter CRP. Let $Q = \{(\tau_i, \theta_i)\}_{i \geq 1}$ be a Poisson (Cox) process on $\mathbb{R}_+ \times \mathbb{R}_+$ with random mean measure

$$\mu(d\tau, d\theta) = d\tau\, W(d\theta) 1_{\theta \leq 1}$$

where $W \sim \mathrm{CRM}(\alpha, \rho)$ with base measure $\alpha(d\theta) = \alpha_0\, d\theta$ with $\alpha_0 > 0$ and Lévy measure ρ satisfying $\int_0^\infty \rho(d\omega) = \infty$. The point process Q has support on $\mathbb{R}_+ \times [0,1]$ and we can write it as follows

$$\begin{aligned} W &\sim \mathrm{CRM}\,(\alpha, \rho) \\ Q \,|\, W &\sim \mathrm{Poisson}(d\tau\, W(d\theta) 1_{\theta \leq 1}) \end{aligned}$$

where Poisson(μ) denotes a Poisson point process with mean μ. This is illustrated in Figure 1(a). Note, importantly, that the CRM is not normalised in the above construction. The distribution of the first n points given W is

$$\mathbb{P}(\theta_{(i)} \in dx_i, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n \mid W) = \left[\prod_{i=1}^{n} W\{dx_i\}\right] e^{-t_n \overline{W}(1)} 1_{t_1<t_2<\dots<t_n} \, dt_1 \dots dt_n \tag{2.3}$$

where $\overline{W}(t) = \int_0^t W(d\theta) = \sum_{i\geq 1} \omega_i 1_{\vartheta_i \leq t}$. Let us denote by $m_{n,j}$ the number of points in the j-th cluster $A_{n,j}$ after having observed n points, and by $(\theta_j^*)_{j=1,\dots,K_n}$ the unique values in $(\theta_{(1)}, \dots, \theta_{(n)})$, ordered by arrival times. Using the results in [33, Proposition 3.1 page 18], we can obtain the expectation of (2.3) with respect to the CRM W

$$\begin{aligned}\mathbb{P}(\Pi_n = \{A_{n,1}, \dots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_n, K_n = k_n, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n) \\ = e^{-\alpha_0 \psi(t_n)} \alpha_0^{k_n} \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n)\, dx_j^*\right] 1_{t_1<\dots<t_n}, dt_1 \dots dt_n.\end{aligned}$$

Integrating over the arrival times $\tau_{(i)}$ and the cluster locations θ_j^* gives

$$\mathbb{P}(\Pi_n) = \int_0^\infty e^{-\alpha_0 \psi(u)} \alpha_0^{k_n} \left[\prod_{j=1}^{K_n} \kappa(m_{n,j}, u)\right] \frac{u^{n-1}}{\Gamma(n)} du,$$

and one recovers the EPPF of the exchangeable random partition associated to a normalised completely random measure [60, Corollary 6], [35, Proposition 3]. In the gamma process case, $\kappa(m, u) = \Gamma(m)/(1+u)^m$ and $\psi(u) = \log(1+u)$, yielding

$$\mathbb{P}(\Pi_n) = \frac{\alpha_0^{K_n}}{\Gamma(n)} \left[\prod_{j=1}^{K_n} \Gamma(m_{n,j})\right] \int_0^\infty \frac{u^{n-1}}{(1+u)^{n+\alpha_0}} du = \frac{\alpha_0^{K_n} \Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \left[\prod_{j=1}^{K_n} \Gamma(m_{n,j})\right]$$

which is the EPPF of the Chinese Restaurant process.

3. Non-exchangeable random partitions. In this section, we build on the Poissonization idea in order to derive a class of non-exchangeable random partitions. This class is shown to have the microclustering property in the next section. The Cox Process $Q = \{(\tau_i, \theta_i)\}_{i\geq 1}$ that defines our non-exchangeable random partition model has the following random mean measure

$$\mu(d\tau, d\theta) = 1_{\theta \leq \tau} W(d\theta) d\tau \tag{3.1}$$

therefore the points will lie under the bisector as shown in Figure 1(b). The overall model is therefore defined as

$$W \sim \mathrm{CRM}\,(\alpha, \rho)$$
$$Q \,|\, W \sim \mathrm{Poisson}(1_{\theta \le \tau}\, d\tau\, W(d\theta))$$

The random partition $\Pi = (\Pi_n)_{n\ge 1}$ of $\mathbb{N}$ is obtained by considering the points $((\tau_{(i)}, \theta_{(i)}))_{i\ge 1}$ of the point process Q ordered by their arrival time, and let $\Pi_n = \Pi(\theta_{(1)}, \ldots, \theta_{(n)})$ be the partition induced by the first n points for any $n \ge 1$. The random partition model is completely specified by the base measure α and the Lévy measure ρ.

The crucial difference with the previous construction is the support of the point process. In the continuous time version of the CRP, every atom of W in $[0, 1]$ was allowed to be chosen at any time, hence the set of potential cluster labels was constant over time. Now, for every fixed $t > 0$, all the clusters whose θ are greater than t cannot be chosen before that time, therefore the set of potential cluster labels increases with t, if for instance the base measure α has unbounded support on $\mathbb{R}_+$. This property intuitively leads to both the non-exchangeability of the random partition induced on $\mathbb{N}$, but also to the microclustering property.

Samples from the process are represented in Figure 2 when $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ with base measure $\alpha(d\theta) = \xi\theta^{\xi-1}d\theta$, for different values of ξ and σ.

PROPOSITION 1. *Let W be a homogeneous CRM(α, ρ) and a point process $Q = \{(\tau_i, \theta_i)\}_{i\ge 1}$ on $\mathbb{R}_+^2$ with mean measure $\mu(d\tau\, d\theta) = 1_{\theta\le\tau} d\tau\, W(d\theta)$. Let $((\tau_{(i)}, \theta_{(i)}))_{i\ge 1}$ be the sequence of points ordered in time, that is such that $\tau_{(1)} < \tau_{(2)} < \ldots$. For any $n \ge 1$, let Π_n be the random partition of $\{1, \ldots, n\}$ defined by the equivalence relation "i and j are in the same cluster" if and only if $\theta_{(i)} = \theta_{(j)}$. For any $n \ge 1$, we have*

$$\mathbb{P}(\Pi_n = \{A_{n,1}, \ldots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \ldots, k_n, K_n = k_n, \tau_{(i)} \in dt_i \text{ for } i = 1, \ldots, n) \tag{3.2}$$
$$= \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n - x_j^*)\alpha(dx_j^*)\right] \times e^{-\int_0^{t_n} \psi(t_n-\theta)\alpha(d\theta)} \left[\prod_{i=1}^{n} 1_{x_{c_i}^* \le t_i}\right] 1_{t_1<t_2<\ldots<t_n}\, dt_1 \ldots dt_n$$

where θ_j^*, $j = 1, \ldots, k_n$, are the unique values of $(\theta_{(1)}, \ldots, \theta_{(n)})$, $m_{n,j} = |A_{n,j}|$ their multiplicities and $c_i \in \{1, \ldots, k_n\}$ is such that $i \in A_{n,c_i}$.

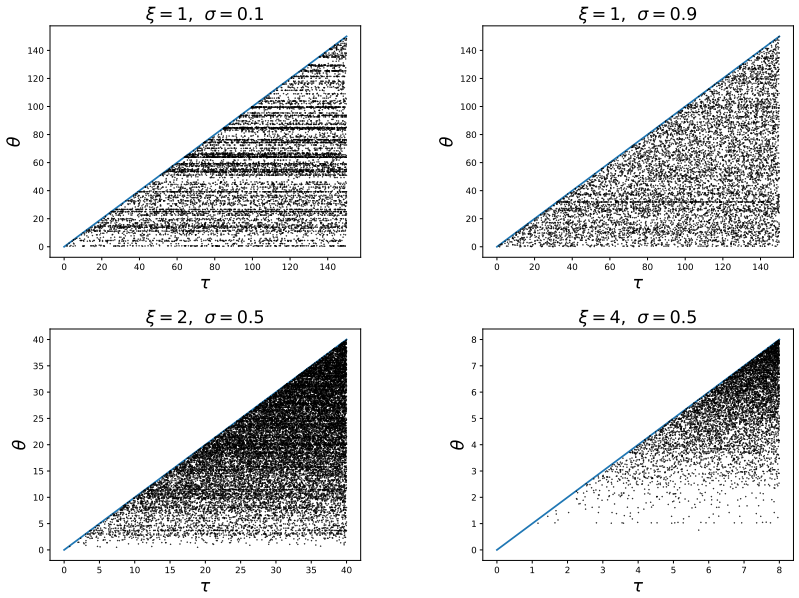


FIG 2. *Samples from the Cox Process Q where $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ with base measure $\alpha(d\theta) = \xi\, \theta^{\xi-1}\, d\theta$, for different values of σ and ξ.*

PROOF. The derivation is similar to the derivation for the exchangeable case described in the previous section. Given W, the set of points $(\tau_i)_{i\geq 1}$ is an inhomogeneous Poisson point process on $\mathbb{R}_+$ with rate $\overline{W}(t)$, hence

$$\mathbb{P}(\tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n \mid W) = e^{-\int_0^{t_n} \overline{W}(t)dt} \left[\prod_{i=1}^{n} \overline{W}(t_i)\right] 1_{t_1<\dots<t_n} dt_1 \dots dt_n.$$

Given the n time variables, the $\theta_{(i)}$'s are distributed as follows

$$\mathbb{P}(\theta_{(i)} \in dx_i \text{ for } i = 1, \dots, n \mid \tau_{(1:n)}, W) = \prod_{i=1}^{n} \frac{W(dx_i)}{\overline{W}(\tau_{(i)})} 1_{x_i<\tau_{(i)}}.$$

It follows that

$$\begin{aligned} &\mathbb{P}(\tau_{(i)} \in dt_i, \theta_{(i)} \in dx_i \text{ for } i = 1, \dots, n \mid W) \\ &\qquad = \left[\prod_{i=1}^{n} W(dx_i) 1_{x_i \leq t_i}\right] e^{-\int_0^{t_n} \overline{W}(t)dt} 1_{t_1<\dots<t_n}\, dt_1 \dots dt_n \end{aligned} \tag{3.3}$$

where $\int_0^{\tau_{(n)}} \overline{W}(t)dt = \sum_j \omega_j(\tau_{(n)} - \vartheta_j)_+ = W(g_{\tau_{(n)}})$ with $g_t(x) = (t - x)_+ = \max(0, t - x)$. Using [33, Proposition 3.1], we can integrate over W to obtain the final result. □

Integrating Equation (3.2) over the cluster allocations $(\theta^*_j)_{j=1,\dots,k_n}$ and the arrival times $\tau_{(1:n)}$, we would obtain the distribution of the random partition Π_n. To the best of our knowledge, there is however no analytical expression for this distribution. We can nonetheless simulate random partitions by using the cluster allocations and arrival times as latent variables. In particular, for the generalized gamma process, we have the following result.

PROPOSITION 2. *Let $W \sim \mathrm{GG}(\alpha, \sigma_0, \zeta_0)$, and $Q = \{(\tau_i, \theta_i)\}_{i\geq 1}$ be the points of a Cox process with mean measure $\mu(d\tau, d\theta) = 1_{\theta\leq\tau} W(d\theta)d\tau$. Then the predictive distribution of $\tau_{(n)}$ is given by*

$$\mathbb{P}(\tau_{(n)} \in dt_n \,|\, (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}) \propto \left[\prod_{j=1}^{K_{n-1}} \frac{1}{(t_n - \theta^*_j + \zeta_0)^{m_{n-1,j}-\sigma_0}}\right] e^{-\int_0^{t_n} \psi(t_n-\theta)\alpha(d\theta)} \times \left(\sum_{j=1}^{K_{n-1}} \frac{m_{n-1,j} - \sigma_0}{t_n - \theta^*_j + \zeta_0} + \int_0^{t_n} \frac{\alpha(\theta)}{(t_n - \theta + \zeta_0)^{1-\sigma_0}}\, d\theta\right) 1_{t_n > \tau_{(n-1)}} dt_n$$

where $\psi(t) = \log(1 + t/\zeta_0)$ for $\sigma_0 = 0$, while $\psi(t) = \left((t + \zeta_0)^{\sigma_0} - \zeta_0^{\sigma_0}\right)/\sigma_0$ for $\sigma_0 \in (0, 1)$. The conditional distribution for $\theta_{(n)}$ is a convex combination of a discrete distribution and a diffuse one,

$$\mathbb{P}(\theta_{(n)} \in dx_n \,|\, (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) \propto H_{\tau_{(n)}}(dx_n) + \sum_{i=1}^{K_{n-1}} \frac{m_{n-1,i} - \sigma_0}{\tau_{(n)} - \theta^*_i + \zeta_0} \delta_{\theta^*_i}(dx_n)$$

where H_t is a diffuse distribution defined as

$$H_t(A) = \int_A \frac{1_{\theta\leq t}}{(t - \theta + \zeta_0)^{1-\sigma_0}}\, \alpha(d\theta) \tag{3.4}$$

for every Borel set $A \subset \mathbb{R}_+$.

For example, if W is a gamma process ($\sigma_0 = 0$) and $\alpha(d\theta) = d\theta$, we obtain

$$\mathbb{P}(\tau_{(n)} \in dt_n \mid (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}) \propto \left[\prod_{j=1}^{K_{n-1}} \frac{1}{(t_n - \theta^*_j + \zeta_0)^{m_{n-1,j}}}\right] e^{-t_n} \left(1 + \frac{t_n}{\xi_0}\right)^{-t_n-\xi_0} \times \left(\sum_{j=1}^{K_{n-1}} \frac{m_{n-1,j}}{t_n - \theta^*_j + \zeta_0} + \log(1 + t_n/\zeta_0)\right) 1_{t_n > \tau_{(n-1)}} dt_n,$$

and

$$\begin{cases} \mathbb{P}(\theta_{(n)} = \theta_j^* \quad | \, (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) = C_n \frac{m_{n-1,j}}{\tau_{(n)} - \theta_j^* + \zeta_0} & \text{for } j = 1, \dots, K_{n-1} \\ \mathbb{P}(\theta_{(n)} \text{ is new} \mid (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) = C_n \log(1 + \tau_{(n)}/\zeta_0) \end{cases}$$

where C_n is the appropriate normalising constant.

4. Properties and inference.

4.1. *Asymptotic properties.* In this section, denote $X_t \sim Y_t$, $X_t = o(Y_t)$ and $X_t = O(Y_t)$ respectively for $X_t/Y_t \to 1$, $X_t/Y_t \to 0$ and $\limsup_t X_t/Y_t < \infty$. The notation $X_t \asymp Y_t$ means both $X_t = O(Y_t)$ and $Y_t = O(X_t)$ hold. When X_t and/or Y_t are random variables the asymptotic relation is meant to hold almost surely.

The properties we are most interested in are the asymptotic behaviour of the cluster sizes $m_{n,j}$, of the number K_n of clusters and the number $K_{n,r}$ of clusters of size r in the random partition. We show in this section that our non-exchangeable model allows for a sublinear growth of the clusters' sizes while retaining desirable properties for the other quantities. Let us list the assumptions on the CRM W to derive the asymptotic results.

(A1) W has finite first two moments, that is

$$\kappa(1,0) = \int_0^\infty \omega \rho(d\omega) < \infty \text{ and } \kappa(2,0) = \int_0^\infty \omega^2 \rho(d\omega) < \infty.$$

(A2) The Lévy tail intensity $\bar{\rho}(x) = \int_x^\infty \rho(d\omega)$ is a *regularly varying* function at 0, that is

$$\overline{\rho}(x) \sim \ell\,(1/x)\, x^{-\sigma}$$

as $x \to 0^+$, where ℓ is a slowly varying function at infinity and $\sigma \in [0,1]$.

(A3) The improper cumulative distribution $\overline{\alpha}(t) = \int_0^t \alpha(dx)$ of the base measure α is a regularly varying function at infinity, that is

$$\overline{\alpha}(t) \sim L(t)\, t^{\xi}$$

as $t \to \infty$, where $\xi > 0$ and L is a slowly varying function. Assume additionally that the base measure α is dominated by the Lebesgue measure, and admits a continuous and monotone density denoted $\tilde{\alpha}(\theta)$.

The moment assumption (A1) excludes the stable process that has infinite first moment. (A2) controls, through the parameter σ, the power-law behaviour of the proportion of clusters of a given size, while condition (A3)

is used to prove the microclustering property and control the sublinear rate of the clusters' size. It is worth noting that the last condition is very mild and allows to pick the density of the base measure from a very large class of functions. Assumptions (A1-A2) are satisfied for the GG with parameters $\sigma_0 \in (-\infty, 1)$ and $\zeta_0 > 0$. In this case, we have $\sigma = \max(\sigma_0, 0)$ and $\ell(t) \propto \log t$ for $\sigma = 0$ and $\ell(t)$ is constant otherwise.

Recall that

$$N(t) = \sum_{i \geq 1} 1_{\tau_i \leq t}$$

denotes the number of points of Q such that $\tau_i \leq t$. For each atom ϑ_j, $j \geq 1$, of the CRM W, let

$$X_j(t) = \sum_{i \geq 1} 1_{\tau_i \leq t} 1_{\theta_i = \vartheta_j}.$$

For $j \geq 1$, let

$$M_j(t) = \sum_{i \geq 1} 1_{\tau_i \leq t} 1_{\theta_i = \theta_j^*}$$

the size of cluster j, ordered by appearance, at time t. Note that $N(\tau_{(n)}) = n$ and $M_j(\tau_{(n)}) = m_{n,j}$.

PROPOSITION 3. *Let $W = \sum_{j \geq 1} \omega_j \delta_{\vartheta_j}$ be a CRM with mean measure $\alpha(d\theta)\rho(d\omega)$ satisfying Assumptions (A1-A3). Let $\{(\tau_i, \theta_i)\}_{i \geq 1}$ be a Poisson point process with mean measure $\mu(d\tau\, d\theta) = 1_{\theta \leq \tau} d\tau\, W(d\theta)$. We have, almost surely as t tends to infinity,*

$$N(t) \sim \frac{\kappa(1,0)}{\xi + 1} t^{\xi+1} L(t)$$

and, for $j \geq 1$

$$X_j(t) \sim W(\{\vartheta_j\})t$$
$$M_j(t) \sim W(\{\theta_j^*\})t.$$

Proposition 3 implies the microclustering property for the random partition Π_n: almost surely, $M_j(t)/N(t) \to 0$ as $t \to \infty$, hence $m_{n,j}/n \to 0$ as $n \to \infty$. In the following theorem, which follows from properties of inverse of regularly varying functions [5, Proposition 1.5.15] or [30, Lemma 22], we obtain exact rates of growth for the cluster sizes.

THEOREM 4 (Microclustering property). *We have*

$$t \sim \left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} L^*_{\xi+1}(N(t))\, N(t)^{1/(\xi+1)} \tag{4.1}$$

almost surely as $t \to \infty$, *where* $L^*_{\xi+1}$ *is a slowly varying function defined in equation* (B.1) *in the Appendix. It follows that the cluster sizes* $m_{n,j} = M_j(\tau_{(n)})$ *verify, for any* $j \geq 1$,

$$m_{n,j} \sim W(\{\theta^*_j\}) \left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} L^*_{\xi+1}(n)\, n^{1/(\xi+1)} \tag{4.2}$$

almost surely as n *tends to infinity. The random weights* $\omega^*_j = W(\{\theta^*_j\})$ *admit the following construction. Denoting by* τ^*_j *the arrival time of the* j*-th cluster, we have that*

$$\begin{aligned}
\tau^*_j &= \Phi^{-1}(\gamma_j) \\
\mathbb{P}(\theta^*_j \in dx \mid \tau^*_j) &= \frac{H_{\tau^*_j}(dx)}{H_{\tau^*_j}(\mathbb{R}_+)} \\
\mathbb{P}(\omega^*_j \in dw_j \mid \theta^*_j, \tau^*_j) &= \frac{w_j e^{-(\tau^*_j - \theta^*_j)w_j} \rho(dw_j)}{\kappa(1, \tau^*_j - \theta^*_j)}
\end{aligned}$$

where $H_\tau(d\theta) = \kappa(1, \tau-\theta)1_{\theta<\tau}\alpha(d\theta)$, $\gamma_j = \sum_{k=1}^j e_k$ *where* $e_1, e_2, \ldots$ *are iid* Exp(1) *and* $\Phi(t) = \int_0^\infty \int_0^\infty \left(1 - e^{-\omega(t-\theta)_+}\right) \rho(d\omega)\alpha(d\theta) = \int_0^\infty \psi(t-\theta)\alpha(d\theta)$. *In the specific case of the GG with parameters* α, σ_0 *and* ζ_0, H_t *takes the form* (3.4) *and*

$$\omega^*_j \,|\, \theta^*_j, \tau^*_j \sim \text{Gamma}(1-\sigma_0, \tau^*_j - \theta^*_j + \zeta_0)$$

Note that the growth rate of the cluster sizes only depends on the parameters ξ and L of the base measure α, and not on the properties of the Lévy measure ρ. For example, taking $\overline{\alpha}(t) = \gamma t^\xi$, with $\xi, \gamma > 0$, we have $L(t) = \gamma$ and $m_{n,j} \asymp n^{1/(\xi+1)}$ and the cluster sizes grow at a rate of n^a where $0 < a < 1$.

We now provide results on the asymptotic rates of the number of clusters and number of clusters of a given size, showing that we can have the same range of behaviour as for exchangeable random partitions. Let

$$K(t) = \sum_{j \geq 1} 1_{X_j(t)>0}$$

be the number of different clusters in $\Pi(t)$ at time t and

$$K_r(t) = \sum_{j\geq 1} 1_{X_j(t)=r}$$

the number of clusters of size r at time t.

PROPOSITION 5. *Let $W = \sum_{j\geq 1}\omega_j\delta_{\vartheta_j}$ be a CRM with mean measure $\alpha(d\theta)\rho(d\omega)$ satisfying Assumptions (A1-A3). Define*

$$\begin{cases} \ell_\sigma(t) = \Gamma(1-\sigma)\ell(t) & \text{if } \sigma \in [0,1) \\ \ell_1(t) = \int_t^\infty y^{-1}\ell(y)dy & \text{if } \sigma = 1. \end{cases}$$

Let $\{(\tau_i,\theta_i)\}_{i\geq 1}$ be a Poisson point process with mean measure $\mu(d\tau\, d\theta) = 1_{\theta\leq\tau} d\tau\, W(d\theta)$. We have, almost surely at t tends to infinity,

$$K(t) \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)}\, L(t)\ell_\sigma(t)\, t^{\sigma+\xi}.$$

For $r \geq 1$, if $\sigma = 0$ then $K_r(t) = o(K(t))$, if $\sigma \in (0,1)$,

$$K_r(t) \sim \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} K(t)\,.$$

If $\sigma = 1$, $K_1(t) \sim K(t)$ and $K_r(t) = o(K(t))$ for all $r \geq 2$.

By noting that $K_n = K(\tau_{(n)})$ and $K_{n,r} = K_r(\tau_{(n)})$, we can combine the results of Proposition 5 and Equation (4.1) to obtain asymptotic expressions for the number K_n of clusters and the number $K_{n,j}$ of clusters of size j in Π_n.

COROLLARY 6. *We have, almost surely as n tends to infinity,*

$$K_n \sim \widetilde{\ell}(n)\, n^{(\sigma+\xi)/(\xi+1)}$$

where $\widetilde{\ell}$ is a slowly varying function defined in equation (B.2) in the Appendix.

For $r \geq 1$, if $\sigma = 0$ then $K_{n,r} = o(K_n)$; if $\sigma \in (0,1)$,

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}\,. \tag{4.3}$$

This corresponds to a power-law behaviour for the proportion of clusters of size r, as

$$\frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \asymp \frac{1}{r^{1+\sigma}}$$

for large r. If $\sigma = 1$, $K_{n,1} \sim K_n$ and $K_{n,r} = o(K_n)$ for all $r \geq 2$. In this case, the proportion of clusters of size 1 tends to one almost surely.

EXAMPLE 7. *If* $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ *with* $\sigma \in (0,1)$ *and base measure* $\alpha(d\theta) = \gamma\xi\theta^{\xi-1}d\theta$ *with* $\xi, \gamma > 0$ *we have* $\ell(t) = \frac{1}{\sigma\,\Gamma(1-\sigma)}$ *and* $L(t) = \gamma$, *therefore*

$$K_n \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\sigma\Gamma(\sigma+\xi+1)}(\xi+1)^{\frac{\sigma+\xi}{1+\xi}}\,\gamma^{1-\frac{\sigma+\xi}{1+\xi}}\,n^{\frac{\sigma+\xi}{1+\xi}}$$

and for all $r \geq 1$

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}$$

almost surely as n *tends to infinity. This power-law behavior is illustrated on Figure 3.*

It is worth noting that although the asymptotic behaviour of the number of clusters and the number of clusters of a given size depend also on the base measure α, the power-law exponent in the proportion of clusters of a given size is solely tuned by the Lévy measure ρ through the parameter σ. The asymptotic proportion of clusters of a given size, given by Equation (4.3), is indeed the same as that obtained with a two-parameter Chinese restaurant process [61] with strictly positive discounting parameter, with a normalised generalized gamma process [48] with parameter $\sigma_0 > 0$ and more generally with a Poisson-Kingman partitions exhibiting α-diversity [60].

REMARK 8. *The number of points sampled by the point process up to time* t *is distributed as the number of points sampled up to time* t *by a Cox process with mean measure* $1_{\theta \leq L(\tau)^{\tau^\xi}} W(d\theta)d\tau$, *where the CRM* W *has base measure* $\alpha(d\theta) = d\theta$. *The same asymptotic result would hold and the proof are analogous and follow by a change of measure.*

REMARK 9. *The counts* K_n *and* $K_{n,r}$ *are asymptotically equivalent to deterministic quantities, contrary to what happens in normalised completely random measures, and more generally in Poisson-Kingman random partitions, where the asymptotics include a random component (see section 6.1 about* α-diversity *in [60]). The reason lies in the fact that our construction involves a CRM with a growing support as the number* n *(or* t*) grows, whose mass* $W([0,t)) \to \infty$ *almost surely as* $t \to \infty$. *Informally, a law of large numbers is therefore at play, which implies* $W([0,t))/\mathbb{E}(W([0,t)) \to 1$ *almost surely as* $t \to \infty$, *resulting in the appearance of the term* $\kappa(1,0) = \mathbb{E}[W([0,t))]/\overline{\alpha}(t)$ *in the depoissonization step (see Equations (4.1) and (B.2)). We refer to the Appendix for a formal proof. Our results hold in the case where* $\overline{\alpha}$ *is regularly varying with index* $\xi > 0$. *It would be interesting to investigate the slowly varying case (*$\xi = 0$*) as a random component may arise in this situation.*

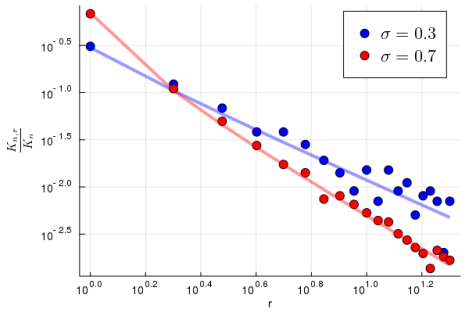


FIG 3. *Log-log plot of the proportions of clusters of given size r for the GG with* $\alpha(d\theta) = d\theta$, $\zeta_0 = 0.1$, $\sigma = 0.3, 0.7$ *and sample size 100k. Dots represent samples from the model (n=100k) and lines represent the asymptotic proportions as* $n \to \infty$, *as given by Equation* (4.3).

4.2. *Inference.*

4.2.1. *Posterior characterization.* Assuming we observe the first n time-ordered points $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ from the Cox process Q, we want to characterize the conditional distribution of the CRM W given the time-ordered observations. The following posterior characterization follows from [33, Proposition 3.1 page 18].

PROPOSITION 10. *Given the first n time-ordered observations* $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ *from the Cox process Q, with unique cluster labels* $\theta^*_1, \dots, \theta^*_{k_n}$, *the conditional distribution of the CRM W is given by*

$$W' + \sum_{j=1}^{k_n} \omega^*_j \delta_{\theta^*_j}$$

where the random positive weights $(\omega^*_1, \dots, \omega^*_{k_n})$ *are independent of the random measure* W'. W' *is an inhomogeneous CRM with mean measure* $\nu'(d\omega, d\theta) = e^{-\omega(\tau_{(n)}-\theta)_+}\rho(d\omega)\alpha(d\theta)$. *The masses of the fixed atoms are conditionally independent given the data with density*

$$\mathbb{P}(\omega^*_j \in dw_j \mid (\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}) = \frac{\rho(dw_j)\, w_j^{m_{n,j}} e^{-w_j(\tau_{(n)}-\theta^*_j)}}{\kappa(m_{n,j}, \tau_{(n)} - \theta^*_j)}.$$

In particular, when W is a generalized gamma process the masses are conditionally gamma distributed

$$\omega^*_j \mid (\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n} \sim \text{Gamma}(m_{n,j} - \sigma_0, \zeta_0 + \tau_{(n)} - \theta^*_j).$$

4.2.2. *Parameter estimation and prediction.* We consider the CRM with base measure $\alpha(d\theta) = \xi\,\theta^{\xi-1}d\theta$ and generalized gamma Lévy measure with parameters σ_0 and ζ_0. The set of parameters is therefore $\eta = (\xi, \sigma_0, \zeta_0)$. Having observed a partition Π_n, we aim at estimating the parameters η and predict Π_{n+m} for $m \geq 1$. However, unlike in the CRP, the marginal likelihood $\mathbb{P}(\Pi_n|\eta)$ is intractable. We use a sequential Monte Carlo algorithm [23, 22] with target distribution $\mathbb{P}(d\theta_{(1:n)}, d\tau_{(1:n)}|\Pi_n, \eta)$ in order to get unbiased estimators of the marginal likelihoods $\mathbb{P}(\Pi_n|\eta)$ for a grid of values of η, and compute the maximum likelihood estimate $\widehat{\eta}$ (see Appendix C for details). Since the model is not Markovian, the time complexity of the SMC algorithm is of order $\mathcal{O}(nK_n)$. The proposal distribution for the arrival times $\tau_{(n)}$ is a truncated normal on $[\tau_{(n-1)}, \infty)$, while the proposal for the cluster location of a new cluster is uniform on $[0, \tau_{(n)}]$. We also use a sequential Monte Carlo algorithm in order to sample from the predictive $\mathbb{P}(\Pi_{n+m}|\Pi_n, \widehat{\eta})$ using Proposition 10.

We also consider experiments where the partition is not observed. In this setting, we consider a classical Bayesian hierarchical construction, see e.g. [43]. We assume that the observations $(y_1, \ldots, y_n)$ are conditionally independent, with

$$y_i|\Pi_n, U_1, \ldots, U_{k_n} \sim F_{U_{c_i}} \tag{4.4}$$

where F_U is a known probability distribution parameterized by U, with pdf f_U, and the $(U_1, \ldots, U_{k_n})$ are the cluster parameters, iid from some distribution G_0 on Θ. This leads to the following factorization of the likelihood

$$p(y_1, \ldots, y_n|\Pi_n) = \prod_{j=1}^{k_n} g(y_{A_{n,j}}) \tag{4.5}$$

where $y_A = \{y_i | i \in A\}$ and

$$g(y_A) = \left[\int_\Theta \prod_{i\in A} f_U(y_i)\right] dG_0(U) \tag{4.6}$$

We will assume that G_0 is a conjugate prior for f_U, so that g, and therefore the likelihood (4.5) can be evaluated analytically. We use a sequential Monte Carlo algorithm with target distribution $\mathbb{P}(d\theta_{(1:n)}, d\tau_{(1:n)}, \Pi_n|\eta, y_{1:n})$ in order to get unbiased estimators of the marginal likelihoods $\mathbb{P}(y_{1:n}|\eta)$ for a grid of values of η, and compute the maximum likelihood estimate $\widehat{\eta}$. More details about the algorithm are contained in the Appendix C.

5. Random partitions and random multigraphs. The non-exchangeable random partition model proposed can be used to derive models for random multigraphs, see [7]. Recall that $\Pi_n = (A_{n,1}, \ldots, A_{n,K_n})$, where the blocks are sorted in order of appearance. For each $i = 1, 2, \ldots$, let c_i be the index of the cluster to which item i belongs, that is $i \in A_{n,c_i}$ for all $n \geq i$. An undirected multigraph $G = \Phi(\Pi)$, possibly with self-loops and with a countably infinite number of edges, is derived from the random partition Π by

$$G = ((c_1, c_2), (c_3, c_4), \ldots)$$

where each pair (c_{2n-1}, c_{2n}) represents an undirected edge between the vertex c_{2n-1} and the vertex c_{2n}. The set of vertices is either $\{1, \ldots, K\}$ if the partition has a finite number of blocks, or the set $\mathbb{N}$. Let G_n be the restriction of G to the first n edges that is, to the first $2n$ items of Π. Then K_{2n}, the number of clusters in Π_{2n}, is also the number of vertices of G_n, $m_{2n,j}$ is the degree of vertex j, $j = 1, \ldots, K_{2n}$ and $K_{2n,j}/K_{2n}$ is the proportion of vertices of degree j.

The multigraphs G obtained by transformation of an exchangeable random partition form a subclass of the edge-exchangeable graphs [18, 10]. This subclass is called rank one edge-exchangeable graphs by Janson [36]. Of particular interest is the so-called Hollywood model [18], obtained from a two-parameter CRP random partition. In this case, inherited from the properties of the associated random partition [61], one can obtain sparse multigraphs with power-law degree distribution. A consequence of the exchangeability assumption is the fact that the degree sequence grows linearly with the number of edges: for any vertex j, its degree $m_{2n,j} \asymp n$ almost surely as the number of edges n tends to infinity. As shown in the following corollary of the results of Section 4, our construction allows to obtain sparse multigraphs with power-law degree distribution and sublinear growth rate for degree sequences.

COROLLARY 11. *Let Π be a non-exchangeable partition with parameters α and ρ verifying assumptions (A1-A3). Let $G = \Phi(\Pi)$ the associated random multigraph. For a subgraph G_n corresponding to the first n edges, let K_{2n} be the number of vertices, $m_{2n,j}$ the degree of vertex j and $K_{2n,r}$ the number of vertices of degree $r \geq 1$. Then, almost surely as the number of*

edges n tends to infinity

$$m_{2n,j} \asymp L^*_{\xi+1}(n)\, n^{1/(1+\xi)}, \quad j \geq 1$$
$$K_{2n} \sim \widetilde{\ell}(n)(2n)^{(\sigma+\xi)/(1+\sigma)}$$
$$\frac{K_{2n,r}}{K_{2n}} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}, \quad r \geq 1$$

where the slowly varying functions $L^*_{\xi+1}$ and $\widetilde{\ell}$ are defined in Equations (B.1) *and* (B.2).

6. Discussion. To obtain random partitions with the microclustering property, one option is to give up the exchangeability assumption, as we did in this paper. An alternative approach is to drop the Kolmogorov consistency assumption discussed in the introduction. Miller et al. [55] and Betancourt et al. [68], who derived random partition models with the microclustering property, take this option, and consider a collection $(\Pi_n)_{n\geq 1}$ of finitely exchangeable random partitions of $[n]$ that do not define a (Kolmogorov-consistent) random partition of $\mathbb{N}$. Another related contribution is the work of [69] where the authors also define a collection $(\Pi_n)_{n\geq 1}$ of finitely exchangeable random partitions of $[n]$ that do not satisfy Kolmogorov consistency property; the authors emphasize that it is indeed a desirable feature for modeling frequencies of frequencies, which motivates their work. Their model is also based on some Poissonization idea. Other random partition models that do not satisfy Kolmogorov consistency are also described in [4].
In [24] the authors present a multitype Yule process that can be seen as a non-exchangeable generalization of the CRP, where a new customer sits at a new table with constant probability $r \in (0,1)$ and otherwise randomly picks one of the previous customers and sits at his table. This model has a very simple urn construction, and has the microclustering property, as $m_{n,j} \asymp n^{1-r}$. The asymptotic properties of K_n and $K_{n,j}$ are however very different from those obtained with our model. The number of clusters K_n increases linearly with n, and the model exhibits a power-law behavior for the proportion of clusters of a given size within an intermediate range. That is $\sum_{r\geq S} K_{nr} \asymp nS^{-\frac{1}{1-r}}$ valid for $1 \ll S \ll n^{1-r}$. In contrast, our model induces a sublinear growth of the number of clusters K_n, as for exchangeable random partitions, and the same asymptotic power-law behavior as the two-parameter Chinese restaurant process for the proportion of clusters of a given size.

There has been a lot of interest over the past years in the development of non-exchangeable partitions based on dependent Dirichlet processes and

more generally dependent random measures [53, 32, 13, 29, 6, 16, 51, 11, 12]. The focus of these works is rather different though, as they do not aim to capture/characterize the microclustering property. The model presented here builds on a Poisson construction on an augmented space, and is therefore somewhat reminiscent of the work of [63, 52, 17].

The authors of [7] considered a general class of exchangeable and non-exchangeable random partitions of $\mathbb{N}$, motivated by preferential attachments models for random multigraphs. For certain values of the parameters, it can generate partitions with the microclustering property, but with a somewhat different asymptotic behavior for the number of clusters. The microclustering property is obtained whenever the number of clusters grows linearly with the dataset [7, Theorem 7]. In our approach, the number of clusters always grows sublinearly, and the rate can be controlled by the properties of the Lévy measure ρ.

7. Experiments. In what follows we compare our non-exchangeable model to the two-parameter Chinese restaurant process [61]. For the non-exchangeable model, we consider a GG with mean measure $\rho(dw)\alpha(d\theta) = 1/\Gamma(1-\sigma)w^{-1-\sigma}e^{-\zeta w}dw\xi\theta^{\xi-1}d\theta$ where the parameters $\xi \in \{1,2,3\}$, $\sigma \in [0,1)$ and $\zeta > 0$ are unknown. For the two-parameter CRP, the two parameters $\sigma_2 \in [0,1)$ and $\kappa_2 > 0$ are considered unknown. We present two sets of experiments. In the first set of experiments, the partition is observed for a training set, and we aim at predicting the size of the clusters on a test set. In the second set of experiments, we consider an application to (semi-supervised) clustering, where the partition is not completely observed and has to be inferred from data. In each set of experiments the data are partitioned into a training set of size n_{train} and a test set of size n_{test} where $n_{\text{train}} + n_{\text{test}} = n$.

7.1. *Observed partition.* In order to best assess the performance of the two random partition models we first assume that the partition is observed. The parameters of each model are estimated on the training data using maximum likelihood with a grid of values for the parameters: 25 equidistant points in $[0,1)$ for σ, $\xi \in \{1,2,3\}$, and a grid obtained by dichotomic search on the interval $[0,100]$ for ζ, and similarly for the two-parameter CRP. The EPPF $\mathbb{P}(\Pi_{n_{\text{train}}}|\sigma_2,\kappa_2)$ of the two-parameter CRP has an analytic form and is calculated directly. For our method, we approximate the likelihood $\mathbb{P}(\Pi_{n_{\text{train}}}|\xi,\sigma,\zeta)$ using sequential Monte Carlo methods with 10000 particles, as described in Section 4.2.

For each cluster $j = 1,\dots,K_{n_{\text{train}}}$ in the training set, we then aim at predicting its size $m_{k,j}$ for $k = n_{\text{train}}+1,\dots,n$. Let $m_{k,j}^{(\text{true})}$ be the true size

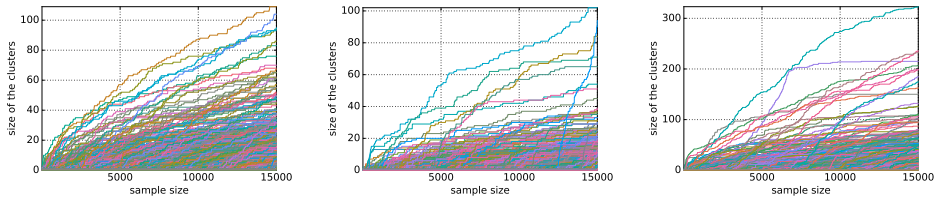
of cluster j in the partition of size k and consider the L2 error

$$E = \frac{1}{K_{n_{\text{train}}}} \sum_{j=1}^{K_{n_{\text{train}}}} \frac{1}{n_{\text{test}}} \sum_{k=n_{\text{train}}+1}^{n} \left(m_{k,j} - m_{k,j}^{(\text{true})} \right)^2 \geq 0.$$

We are interested in the distribution of the predictive error

$$\mathbb{P}\left(E \in dE \mid \Pi_{n_{\text{train}}}, \widehat{\eta}\right) \tag{7.1}$$

where $\widehat{\eta}$ are the fitted parameters, under the two-parameter CRP or our model. Additionally, we want to check that the model can still capture the distribution of the cluster sizes adequately. To this aim, we also report 95% predictive credible intervals for the proportion of clusters of a given size in the test set, and compare this to the empirical distribution.



(a) Synthetic (b) Amazon (c) Math Overflow

FIG 4. *Evolution of the clusters' sizes $m_{j,n}$ with respect to the sample size n for the (a) synthetic, (b) Amazon and (c) Math Overflow datasets.*

Synthetic data. In order to validate the inference procedure, we first run experiments on a simulated dataset, where the data are simulated from our model with parameters set to $(\xi, \sigma, \zeta) = (1, 0.41, 10)$. In this model, the cluster size grows at a rate of $\sqrt{n}$, as can be seen from Figure 4(a) that shows the growth of the cluster sizes with respect to the sample size n. Additionally, the proportion of clusters of a given size has an asymptotic power-law distribution, see the top row of Figure 7. As shown in Figure 5, the SMC estimate of the log-likelihood is rather accurate, and we recover the true parameters. The mean and quantiles of the predictive error under our model and the two-parameter CRP are reported in Table 1. As expected, the predictive under our model outperforms the two-parameter CRP, which is misspecified in that case. Posterior predictive of the proportion of clusters of a given size is reported in the first row of Figure 7.

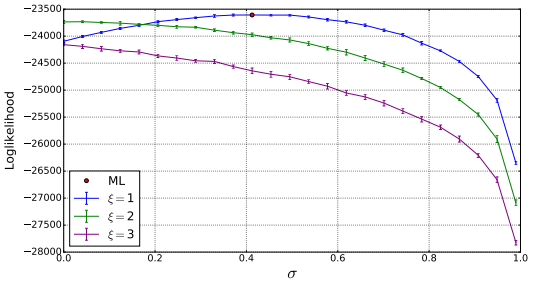


FIG 5. *Loglikelihood estimates for $\zeta = 10$ and different values of ξ and σ. For every grid point, 10 SMC estimates are obtained, and the mean and ± 1 standard deviation error bars are reported.*

Real data. We consider two datasets from the SNAP database [44] of the same size $n = 15000$ with $n_{train} = 5000$. The first one is the Amazon dataset of movies' reviews [54] where each movie represents a cluster containing its reviews, which are ordered. The second dataset is a time-ordered collection of answers to questions in the Math Overflow website[1] where the clusters contain answers to the same question. Evolutions of the cluster sizes are reported in Figure 4(b-c) for these datasets. We aim at predicting, based on the training set, the number of reviews to a given movie for the Amazon dataset, and the number of questions answered to a given question for the Math Overflow dataset. In both cases our non-exchangeable model provides better predictions of the cluster sizes (see Table 1 and Figure 6). Estimates and credible intervals for the parameter σ and σ_2 are reported in Table 2. Figure 7 shows that both models give reasonable predictive fit to the proportion of clusters of a given size.

TABLE 1
Mean and quantiles of the predictive error using 100 samples from the predictive distributions, when the partition is observed.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	6.92	[6.37, 7.42]	14.9	[13.4, 16.2]
Amazon	6.14	[5.58, 6.96]	8.43	[7.35, 9.42]
Math Overflow	1.08×10^2	$[1.01, 1.22] \times 10^2$	1.67×10^2	$[1.58, 1.77] \times 10^2$

[1] https://mathoverflow.net/

TABLE 2
MLE for the the parameter σ of the non-exchangeable model and the discount parameter σ_2 of the two-parameter CRP model, and 0.025 and 0.975 quantiles of their posterior distributions, when the partition is observed.

	Non-exchangeable		Two-parameter CRP	
	MLE of σ	$[q_{0.025}, q_{0.975}]$	MLE of σ_2	$[q_{0.025}, q_{0.975}]$
Synthetic	0.41	[0.38, 0.45]	0.46	[0.41, 0.51]
Amazon	0.63	[0.57, 0.66]	0.65	[0.60, 0.69]
Math Overflow	0.37	[0.34, 0.40]	0.30	[0.24, 0.36]

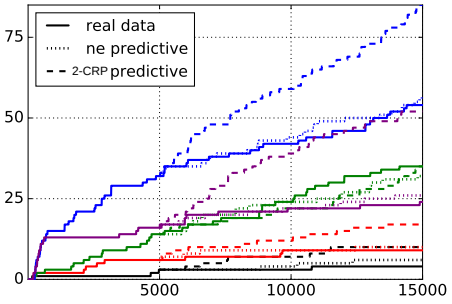


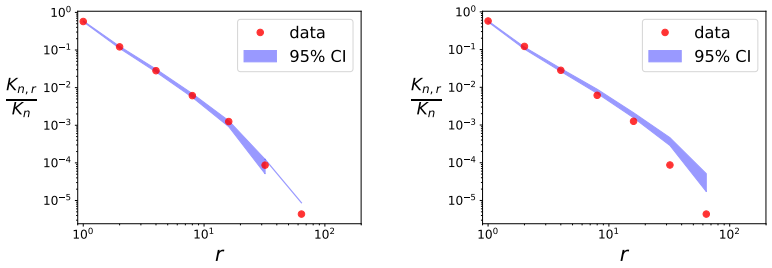
FIG 6. *Amazon dataset. Observed (plain line) and predicted sizes of some clusters (in different colours) from the non-exchangeable (dotted line) and the two-parameter CRP models (dashed line).*

7.2. *Partially observed partition.* In this subsection we consider that observations are word counts in documents, which are modeled using a hierarchical Dirichlet-Multinomial model, and we aim at finding a partition of the documents into topics, and predicting the growth of these topics. Let V denote the size of the vocabulary, and let y_i, $i = 1, \dots, n$, be a V-dimensional vector of word counts in document i. Assume that Π_n follows our non-exchangeable random partition model with $\mathrm{GG}(\alpha, \sigma, \zeta)$ and base measure $\alpha(d\theta) = \xi\theta^{\xi-1}d\theta$, and

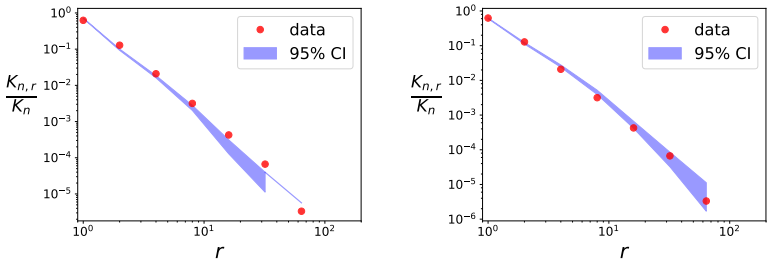
$$U_j \overset{iid}{\sim} \mathrm{Dirichlet}(h, \dots, h) \qquad j = 1, \dots, k_n$$
$$y_i \mid \Pi_n, U_1, \dots, U_{k_n} \overset{ind}{\sim} \mathrm{Multinomial}(N_i, U_{c_i}) \qquad i = 1, \dots, n$$

where U_j is a V-dimensional probability vector defining the word frequencies for the j's cluster/topic and $h > 0$ is the scale parameter of the Dirichlet distribution.

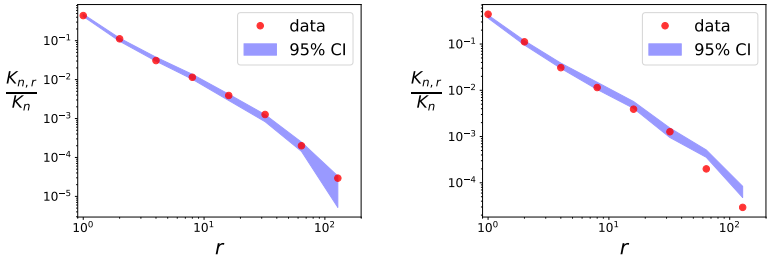
As before, we are interested in predicting the growth of the clusters' sizes and consider the following error which generalizes the error measure defined



(a) Synthetic



(b) Amazon



(c) Math Overflow

FIG 7. *Empirical proportions of clusters of given size (red dots) and* 95% *posterior predictive credible intervals (blue) for our non-exchangeable model (left) and the two-parameter CRP (right).*

in the previous subsection to the case of unobserved partitions:

$$\widetilde{E} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{K_{n_{\text{train}}}} ((m_{k,n} - m_{k,n_{\text{train}}}) - (\hat{m}_{k,n} - \hat{m}_{k,n_{\text{train}}}))^2$$

where $\hat{m}_{k,n}$ is the size of the k-th cluster at time n inferred by the model. In the following experiments we consider two datasets of size $n = 4000$, and split each dataset into three subsets. The first $n_{\text{pre-train}} = 1000$ observations are provided with labels and are used in the pre-training stage to tune the parameters of the random partition models, which are estimated with MLEs for both the non-exchangeable model and the two-parameter CRP, as done in the previous experiments. Given the estimates of the parameter, the objective is to infer and predict the partition of the remaining unlabelled observations. For this purpose, the following $n_{\text{train}} = 1000$ observations are used to infer the partition, and the last $n_{\text{test}} = 2000$ points are used to compare prediction performances.

For the non-exchangeable model the partition is inferred using a SMC algorithm which gives an approximation of the distribution of the cluster labels, as described in Section 4.2.2, while a Gibbs sampler is used for the two-parameter CRP model.

For each dataset we consider two scenarios, corresponding to the clustering task in the semi-supervised and unsupervised settings. In the first one, the first 200 observations in the training set of size $n_{\text{train}} = 1000$ are provided with labels, while in the second case the partition of the training set is fully unobserved.

Synthetic data. A dataset of $n = 4000$ observations $(y_1, \dots, y_n)$ is sampled with parameters $(\xi, \sigma, \zeta) = (1, 0.5, 1)$, $V = 15$, $N_i = 1000$, $h = 1$. Tables 5 reports the estimates of the power-law parameters, and both models obtain rather precise estimates. Tables 3 and 4 shows that the non-exchangeable model outperforms the two-parameter CRP in predicting the partition of the observations in the test set. It is worth noting that providing the labels for the first 200 observations in the training set does not improve the predictive performances significantly. This is due to the fact that the number of counts N_i per observation is much bigger than the dimensionality V of the Dirichlet parameters, which facilitates the clustering. This in not the case in the real data example, where predictive performances will differ between the semi-supervised and unsupervised cases.

Real data. We use again the Amazon dataset of movies' reviews, and consider the actual reviews as observations in a bag-of-words representation. We preprocess the dataset using the text-mining R package *tm* [26, 27] retaining the first time-ordered $n = 4000$ reviews. Estimates of the parameters σ and σ_2 can be found in Tables 5 and the prediction errors in Tables 3 and 4, showing that the proposed model provides more accurate predictions than the two-parameter CRP for this dataset.

TABLE 3
Mean and quantiles of the predictive error using 100 samples from the predictive distributions when the partition of the data is partially observed. Lower is better.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	**3.83**	[2.99, 4.46]	4.00	[3.47, 4, 64]
Amazon	**17.06**	[9.66, 30.29]	336.02	[303.96 366.31]

TABLE 4
Mean and quantiles of the predictive error using 100 samples from the predictive distributions when the partition of the data is unknown. Lower is better.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	**3.85**	[3.12, 4.62]	4.07	[3.30 4.90]
Amazon	**117.70**	[96.85, 137.51]	605.46	[556.84 654.93]

TABLE 5
MLE for the the parameter σ of the non-exchangeable model and posterior mean of the discount parameter σ_2 of the two-parameter CRP model, and 0.025 and 0.975 quantiles of their posterior distributions.

	Non-exchangeable		Two-parameter CRP	
	MLE of σ	$[q_{0.025}, q_{0.975}]$	Posterior mean of σ_2	$[q_{0.025}, q_{0.975}]$
Synthetic	0.54	[0.50, 0.58]	0.43	[0.37, 0.49]
Amazon	0.58	[0.54, 0.62]	0.65	[0.60, 0.70]

Supplementary material. A demo of the simulation and inference for the non-exchangeable random partition model can be found at https://github.com/giuseppedib/microclustering.

Acknowledgements. GDB acknowledges support from EPSRC under grant EP/L016710/1. FC and YWT acknowledge funding from the ERC under the European Union's 7th Frame-work programme (FP7/2007-2013) ERC grant agreement no. 617071. FC acknowledges support from EPSRC under grant EP/P026753/1.

APPENDIX A: PROOF OF THE MAIN THEOREMS

PROOF OF PROPOSITION 2. From Eq. (3.2), the joint distribution of $((\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n})$ is given by

$$\begin{aligned}
&\mathbb{P}(\Pi_n = \{A_{n,1}, \dots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_n, K_n = k_n, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n) \\
&= \left\{ \sum_{i=1}^{k_{n-1}} \left[\prod_{\substack{j=1 \\ j \neq i}}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(m_{n-1,i} + 1, t_n - x_i^*) \tilde{\alpha}(x_i^*)\, \delta_{x_i^*}(dx_{k_n}^*) \right. \\
&\left. + \left[\prod_{j=1}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(1, t_n - x_n^*) \alpha(dx_{k_n}^*) \right\} \\
&\times e^{-\int_0^{t_n} \psi(t_n - x) \alpha(dx)} \left[\prod_{i=1}^{n} 1_{x_{c_i}^* < t_i} \right] 1_{t_1 < t_2 < \dots < t_n} dt_1 \dots dt_n.
\end{aligned}$$

Integrating over the location of the n-th point, we obtain

$$\begin{aligned}
&\mathbb{P}(\Pi_{n-1} = \{A_{n,1}, \dots, A_{n-1,k_{n-1}}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_{n-1}, \\
&\quad K_{n-1} = k_{n-1}, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n) \\
&= \left\{ \sum_{i=1}^{k_{n-1}} \left[\prod_{\substack{j=1 \\ j \neq i}}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(m_{n-1,i} + 1, t_n - x_i^*) \tilde{\alpha}(x_i^*) \right. \\
&\left. + \left[\prod_{j=1}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \int_0^{t_n} \kappa(1, t_n - x) \alpha(dx) \right\} \\
&\times e^{-\int_0^{t_n} \psi(t_n - x) \alpha(dx)} \left[\prod_{i=1}^{n-1} 1_{x_{c_i}^* < t_i} \right] 1_{t_1 < t_2 < \dots < t_n} dt_1 \dots dt_n.
\end{aligned}$$

In the Generalised Gamma Process case we have

$$\kappa(m, u) = \frac{1}{\Gamma(1 - \sigma)} \frac{\Gamma(m - \sigma)}{(\zeta + u)^{m - \sigma}}.$$

Therefore $\kappa(m+1,u)=\kappa(m,u)\frac{m-\sigma}{\zeta+u}$ and

$$\sum_{i=1}^{k_{n-1}}\left[\prod_{\substack{j=1\\ j\neq i}}^{k_{n-1}}\kappa(m_{n-1,j},t_n-x_j^*)\alpha(dx_j^*)\right]\kappa(m_{n-1,i}+1,t_n-x_i^*)\alpha(dx_i^*)$$
$$=\left[\prod_{j=1}^{k_{n-1}}\kappa(m_{n-1,j},t_n-x_j^*)\alpha(dx_j^*)\right]\sum_{i=1}^{k_{n-1}}\frac{m_{n-1,i}-\sigma}{t_n-x_i^*+\zeta}$$

hence

$$\begin{aligned}
&\mathbb{P}(\Pi_{n-1}=\{A_{n,1},\dots,A_{n-1,k_{n-1}}\},\theta_j^*\in dx_j^*\text{ for }j=1,\dots,k_{n-1},\\
&\quad K_{n-1}=k_{n-1},\tau_{(i)}\in dt_i\text{ for }i=1,\dots,n)\\
&=\frac{1}{\Gamma(1-\sigma)^{k_{n-1}}}\left[\prod_{j=1}^{k_{n-1}}\frac{\Gamma(m_{n-1,j}-\sigma)\,\alpha(dx_j^*)}{(t_n-x_j^*+\zeta)^{m_{n-1,j}-\sigma}}\right]\\
&\times\left(\sum_{j=1}^{k_{n-1}}\frac{m_{n-1,j}-\sigma}{t-x_j^*+\zeta}+\int_0^{t_n}\frac{\alpha(dx)}{(t_n-x+\zeta)^{1-\sigma}}\right)\\
&\times e^{-\int_0^{t_n}\psi(t_n-x)\alpha(dx)}\left[\prod_{i=1}^{n-1}1_{x^*_{c_i}<t_i}\right]1_{t_1<\dots<t_n}dt_1\dots dt_n.
\end{aligned}$$

from which we obtain the results of the theorem. □

Proof of proposition 3 and theorem 4. Given W, $N(t)$ is a non-homogeneous Poisson process with rate $\overline{W}(t)=\sum_{j\geq 1}\omega_j 1_{\vartheta_j\leq t}$. Hence, using Fubini's and Campbell's theorems,

$$\mathbb{E}[N(t)]=\mathbb{E}\left[\int_0^t\overline{W}(x)dx\right]=\overline{\overline{\alpha}}(t)\kappa(1,0)$$

where $\overline{\overline{\alpha}}(t)=\int_0^t\overline{\alpha}(t)$. Similarly,

$$\begin{aligned}
\operatorname{var}(N(t))&=\operatorname{var}(\mathbb{E}[N(t)\,|\,W])+\mathbb{E}[\operatorname{var}(N(t)\,|\,W)]=\operatorname{var}(\overline{\overline{W}}(t))+\mathbb{E}[\overline{\overline{W}}(t)]\\
&=\kappa(2,0)\int_0^t(t-\theta)^2\alpha(d\theta)+\kappa(1,0)\overline{\overline{\alpha}}(t)\\
&=\kappa(2,0)\int_0^t\overline{\overline{\alpha}}(x)dx+\kappa(1,0)\overline{\overline{\alpha}}(t)
\end{aligned}$$

Using Karamata's theorem [5, Proposition 1.5.8] and Assumption (A3), we obtain that

$$\mathbb{E}[N(t)] \sim \frac{\kappa(0,1)}{\xi+1} t^{\xi+1} L(t) \text{ and } \operatorname{var}(N(t)) \sim \frac{\kappa(2,0)}{(\xi+1)(\xi+2)} t^{\xi+2} L(t).$$

Therefore, for any $0 < a < \xi$ we have $\operatorname{var}(N(t)) = O(t^{-a}\mathbb{E}[N(t)]^2)$. Using [15, Lemma B.1], we conclude that, almost surely as t tends to infinity

$$N(t) \sim \mathbb{E}[N(t)] \sim \frac{\kappa(1,0)}{\xi+1} t^{\xi+1} L(\xi).$$

Conditional on W, $X_j(t)$ is a non-homogeneous Poisson process with rate $\omega_j 1_{\vartheta_j > t}$ hence

$$X_j(t) \sim \omega_j t$$

almost surely as t tends to infinity. It follows similarly that $M_j(t)$, the size, at time t, of the jth cluster to appear, satisfies

$$M_j(t) \sim \omega_j^* t$$

where $\omega_j^* = W(\{\theta_j^*\})$. Additionally, in the same fashion as in [58, Theorem 4.5], Proposition 1 implies that, denoting by τ_j^* the arrival time of the j-th cluster and by ω_j^*, θ_j^* its weight and location respectively,

$$\{\tau_j^*, \theta_j^*, \omega_j^*\}_{j \geq 1} \sim \text{Poisson}\left(\omega\, e^{-\omega(\tau-\theta)_+} 1_{\theta<\tau} d\tau\, \alpha(d\theta)\rho(d\omega)\right)$$

from which it follows that, given a sequence of iid $e_k \sim \text{Exp}(1)$, the clusters' arrival times are defined as $\tau_j^* = \Phi^{-1}(\gamma_j)$, where $\gamma_j = \sum_{k=1}^{j} e_k$. Proposition 2 then implies the result.

□

Proof of proposition 5. Observe that $\mathbb{P}(X_j(t) > 0 \mid W) = 1 - e^{-\omega_j(t-\vartheta_j)_+}$. By the marking theorem [41, Chapter 5], for each t, $\{(\omega_j, \vartheta_j) \mid j \geq 1, X_j(t) > 0\}$ is a Poisson point process with mean measure $\rho(d\omega)\alpha(d\theta)(1 - e^{-\omega(t-\theta)_+})$. It follows that

$$\mathbb{E}[K(t)] = \operatorname{var}[K(t)] = \int_0^\infty \int_0^t \left(1 - e^{-(t-\theta)\omega}\right) \alpha(d\theta)\rho(d\omega) = \int_0^t \psi(t-\theta)\alpha(d\theta).$$

Similarly to [30, Proposition 2], it follows from the monotonicity of $K(t)$ and the Borel-Cantelli lemma that $K(t) \sim \mathbb{E}[K(t)]$ almost surely as $t \to$

∞. Using the Tauberian theorems [28, Chapter XIII, Section 5] recalled in Lemma 13, Lemma 14 and $\alpha(t) \sim \xi t^{\xi-1} L(t)$, we obtain

$$\mathbb{E}[K(t)] \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}$$

as t tends to infinity.

We proceed similarly for $K_r(t)$. For each $t > 0$ and $r \geq 1$, $\{(\omega_j, \vartheta_j) \mid j \geq 1, X_j(t) = r\}$ is a Poisson point process with mean measure $\rho(d\omega)\alpha(d\theta)\frac{\omega^r (t-\theta)_+^r}{r!} e^{-\omega(t-\theta)_+}$. It follows that

$$\begin{aligned}
\mathbb{E}[K_r(t)] &= \mathrm{var}[K_r(t)] \\
&= \int_0^t \int_0^\infty \frac{(t-\theta)^r}{r!}\, \omega^r e^{-(t-\theta)\omega} \rho(d\omega)\alpha(d\theta) \\
&= \int_0^t \frac{(t-\theta)^r}{r!} \kappa(r, t-\theta)\alpha(d\theta).
\end{aligned}$$

Using Lemma 13 and 14, we obtain: if $\sigma = 0$, $\mathbb{E}[K_r(t)] = o(L(t)\ell(t))t^{\sigma+\xi}$; if $\sigma \in (0,1)$,

$$\mathbb{E}[K_r(t)] \sim \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}.$$

If $\sigma = 1$, $\mathbb{E}[K_1(t)] \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}$ and $\mathbb{E}[K_r(t)] = o(L(t)\ell(t))t^{\sigma+\xi}$ for all $r \geq 2$. For the almost sure result, we proceed as for $K(t)$, using the monotonicity of $\sum_{r\geq s} K_r(t)$ [30, Corollary 21], the equality $\mathrm{var}\left[\sum_{r\geq s} K_r(t)\right] = \mathbb{E}\left[\sum_{r\geq s} K_r(t)\right]$ and the fact that $\mathbb{E}[K_r(t)] \asymp K(t)$ for $\sigma \in (0,1)$, $\mathbb{E}[K_r(t)] = o(K(t))$ for $\sigma = 0$ and $\mathbb{E}[K_1(t)] \sim K(t)$ for $\sigma = 1$. □

Proof of proposition 10. The result follows from known results in Poisson partition calculus. Let us define $f(\omega, \theta) = (\tau_{(n)} - \theta)_+ \omega$ and denote by $\mathcal{P}(dW \mid \nu)$ the distribution of the CRM W with Lévy measure ν, then Proposition 3.1 part (ii) in [33] provides the Bayes formula for our model:

the joint distribution of the CRM W and $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ can be written as

$$
\begin{aligned}
&e^{-\int_0^{\tau_{(n)}} \overline{W}(t)dt} \left[\prod_{i=1}^{n} 1_{\theta_i<\tau_{(i)}}\right] 1_{\tau_{(1)}<\cdots<\tau_{(n)}} d\tau_{(1)} \dots d\tau_{(n)} \mathcal{P}(dW \mid \nu) \\
&= P(dW \mid e^{-f}\nu) \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n - x_j^*)\alpha(dx_j^*)\right] e^{-\int_0^{t_n} \psi(t_n-\theta)\alpha(d\theta)} \\
&\quad \times \left[\prod_{i=1}^{n} 1_{x_{c_i}^*\le t_i}\right] 1_{t_1<t_2<\ldots<t_n}\, dt_1 \dots dt_n
\end{aligned}
$$

where the first term $P(dW \mid e^{-f}\nu)$ is the posterior distribution of W and the remaining terms describe the marginal distribution as in Proposition 1. The result then follows from formulas (33) and (34) in [33]. □

APPENDIX B: BACKGROUND ON REGULAR VARIATION AND TECHNICAL LEMMA

We recall the following definitions which can be found in [5] and [65].

DEFINITION 12 (Regularly varying function). *A measurable function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *is* regularly varying *at* ∞ *with index* $\xi \in \mathbb{R}$ *if for every* $x > 0$

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{\xi}.$$

If $\xi = 0$ we say that the function is *slowly varying*. An important property of the regularly varying function is that they can be written as $f(x) = \ell(x)x^{\xi}$ where ξ is the exponent of variation and ℓ is a slowly varying function.

Let $L^{\#}$ be the de Brujin conjugate [5] of a slowly varying function L. Regularly varying functions $f(x) = L(x)x^{\xi}$ of index $\xi > 0$ admit asymptotic inverse $g(x) = L_{\xi}^*(x)x^{1/\xi}$ which are regularly varying of index ξ^{-1} (see [5, Proposition 1.5.15] or [30, Lemma 22]) with slowly varying part

$$L_{\xi}^*(x) = \{L^{1/\xi}(x^{1/\xi})\}^{\#}. \tag{B.1}$$

Note that if $L(t) = c$, then $L_{\xi}^*(t) = c^{1/\xi}$. From Equation (4.1) and Proposition 5, it follows that the slowly varying function appearing in Corollary 6

is

$$\widetilde{\ell}(n) = \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} \left(\frac{\xi+1}{\kappa(1,0)}\right)^{(\sigma+\xi)/(\xi+1)} L^*_{\xi+1}(n)^{\sigma+\xi}$$

(B.2)

$$\times L\left\{\left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} n^{1/(\xi+1)} L^*_{\xi+1}(n)\right\} \ell_\sigma\left\{\left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} n^{1/(\xi+1)} L^*_{\xi+1}(n)\right\}.$$

The following lemma is a compilation of Tauberian results in Propositions 17, 18 and 19 in [30]. See also [28, Chapter XIII].

LEMMA 13. *Let ρ be a Lévy measure on $(0,\infty)$ with tail Lévy intensity $\overline{\rho}(x) = \int_x^\infty \rho(d\omega)$. Assume*

$$\overline{\rho}(x) \sim x^{-\sigma}\ell(1/x)$$

as x tends to 0, where $\sigma \in [0,1]$ and ℓ is a slowly varying function at infinity. For any $\sigma \in [0,1)$,

$$\psi(t) \sim \Gamma(1-\sigma)t^\sigma \ell(t)$$

and for $r = 1, 2, \dots$

$$\begin{cases} \kappa(r,t) \sim t^{\sigma-r}\ell(t)\Gamma(r-\sigma) & \text{if } \sigma \in (0,1) \\ \kappa(r,t) = o(t^{\sigma-r}\ell(t)) & \text{if } \sigma = 0 \end{cases}$$

as t tends to infinity. For $\sigma = 1$,

$$\psi(t) \sim t\ell_1(t)$$
$$\kappa(1,t) \sim \ell_1(t)$$

and

$$\kappa(r,t) \sim t^{1-r}\ell(t)\Gamma(r-1)$$

for all $r \geq 2$ as t tends to infinity, where $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx$.

LEMMA 14. *Let f and g be locally bounded, regularly varying functions with $f(x) = \ell_f(x)x^a$ and $g(x) = \ell_g(x)x^b$ where $a, b > -1$ and ℓ_f, ℓ_g are slowly varying functions. Then as t tends to infinity*

$$\int_0^t f(x)g(t-x)dx \sim \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} tf(t)g(t)$$
$$\sim \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} \ell_f(t)\ell_g(t)t^{a+b+1}.$$

PROOF. Let us split the integral in the following way

$$\int_0^t f(x)g(t-x)dx = \int_0^{\frac{t}{2}} f(x)g(t-x)dx + \int_0^{\frac{t}{2}} f(t-x)g(x)dx.$$

Let $\delta \in (0, \min(a,b)+1)$. From Potter's Theorem [5, Theorem 1.5.6], there is X such that for all $t > 2X$, $u \in [X/t, 1/2]$,

$$\frac{f(tu)}{f(t)} \leq 2u^{a-\delta}, \quad \frac{g(t(1-u))}{g(t)} \leq 2(1-u)^{b-\delta}.$$

Take $t > 2X$. We have

$$\int_X^{\frac{t}{2}} \frac{f(x)g(t-x)}{tf(t)g(t)}\,dx = \int_0^{\frac{1}{2}} \frac{f(ut)}{f(t)} \frac{g((1-u)t)}{g(t)} 1_{u\in[X/t,1/2]}\,du$$

where the integrand function is bounded by $4u^{a-\delta}(1-u)^{b-\delta}$ which is integrable, hence we have convergence to $\int_0^{\frac{1}{2}} u^a(1-u)^b du \in (0,\infty)$ by the dominated convergence theorem. We proceed analogously for the second integral. Since $\int_0^1 u^a(1-u)^b du = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$, we have the result. □

APPENDIX C: ALGORITHM'S DETAILS

This section presents details about the parameter estimation in the non-exchangeable model in the case the partition in not observed and the observations follow the model described in Equations 4.4-4.6. The random partition model's parameters $\eta = (\xi, \sigma_0, \zeta_0)$ are estimated by the MLE $\hat{\eta}$, and marginal likelihood estimates are obtained with a Sequantial Monte Carlo algorithm. Let us denote the unobserved variables as $x_i = (c_i, \theta_{(i)}, \tau_{(i)})$ for $i = 1, \ldots, n$. Defining for each $n \in \mathbb{N}$

$$f(n) := \left[\prod_{j=1}^{k_{n-1}} \frac{\kappa(m_{n-1,j}, \tau_{(n)} - \theta_j^*)}{\kappa(m_{n-1,j}, \tau_{(n-1)} - \theta_j^*)}\right] \times e^{-\int_0^{\tau_n} \psi(\tau_{(n)}-\theta)\alpha(d\theta) + \int_0^{\tau_{n-1}} \psi(\tau_{(n-1)}-\theta)\alpha(d\theta)} 1_{\theta_{(n)}<\tau_{(n)}} 1_{\tau_{(n-1)}<\tau_{(n)}}$$

the predictive distribution can be written as

$$p(x_n \mid x_{1:n-1}, \eta) = \begin{cases} f(n)\frac{m_{n-1,j}-\sigma}{\tau_{(n)}+\theta_j^*-\zeta} & c_n = j \in \{1, \ldots, k_{n-1}\} \\ f(n)\frac{\alpha(\theta_{(n)})}{(\zeta+\tau_{(n)}-\theta_{(n)})^{1-\sigma}} & c_n = k_{n-1}+1 \end{cases}$$

Denoting by N_p the number of particles in the SMC sampler, the marginal likelihood estimate can be computed using the importance weights of the

particles (see [21],[2]): $\hat{p}(y_{1:n} \mid \eta) = \hat{p}(y_1 \mid \eta) \prod_{t=2}^{n} \hat{p}(y_t \mid y_{1:t-1}, \eta)$ with $\hat{p}(y_t \mid y_{1:t-1}, \eta) = \frac{1}{N_p} \sum_{j=1}^{N_p} w_t^{(j)}$ and

$$w_t^{(j)} = \frac{p(x_t^{(j)}, y_t \mid x_{1:t-1}^{(j)}, y_{1:t-1}, \eta)}{q(x_t^{(j)} \mid x_{1:t-1}^{(j)}, y_{1:t}, \eta)} \qquad \text{for all } j = 1, \ldots, N_p$$

where q denotes the proposal distribution which we define as

$$\begin{aligned} q(x_n | x_{1:n-1}, y_{1:n}, \eta) = {} & q(\tau_{(n)} \mid \tau_{(n-1)})\, q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_n, \eta) 1_{c_n \in \{1,\ldots,k_{n-1}\}} \\ & + q(\tau_{(n)} \mid \tau_{(n-1)})\, q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_{1:n}, \eta)\, q(\theta_n \mid \tau_{(n)}) 1_{c_n = k_{n-1}+1} \end{aligned}$$

with $q(\tau_{(n)} \mid \tau_{(n-1)})$ being the density of truncate Gaussian distributions centred at $\tau_{(n-1)}$, $q(\theta_n \mid \tau_{(n)}) = 1/\tau_{(n)}$, and

$$q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_{1:n}, \eta) \propto \begin{cases} \frac{m_{n-1,j} - \sigma}{\tau_{(n)} - \theta_j^* + \zeta} \frac{g(y_{A_{n-1,j} \cup \{n\}})}{g(y_{A_{n-1,j}})} & c_n = j \in \{1, \ldots, k_{n-1}\} \\ H_{\tau_{(n)}}(\mathbb{R}_+)\, g(y_{\{n\}}) & c_n = k_{n-1} + 1 \end{cases}$$

where g is defined in Equation4.5 and $H_{\tau_{(n)}}$ in Equation 3.4. Therefore we have that the importance weight of the j-th particle at time t can be written as

$$w_t^{(j)} = \begin{cases} \frac{Sf(t)}{q(\tau_{(t)}^{(j)} \mid \tau_{(t-1)}^{(j)})} & c_t = j \in \{1, \ldots, k_{t-1}\} \\ \frac{Sf(t)}{q(\tau_{(t)}^{(j)} \mid \tau_{(t-1)}^{(j)}) q(\theta_{(t)}^{(j)} \mid \tau_{(t)}^{(j)}) H_{\tau_{(t)}^{(j)}}(\mathbb{R}_+)} & c_t = k_{t-1} + 1 \end{cases}$$

where S is the normalising constant.

REFERENCES

[1] Aldous, D. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983* 1–198.

[2] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 269–342.

[3] Bacallado, S., Favaro, S. and Trippa, L. (2015). Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli* **21** 1–37.

[4] Betz, V., Ueltschi, D. and Velenik, Y. (2011). Random permutations with cycle weights. *The Annals of Applied Probability* **21** 312–331.

[5] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular variation* **27**. Cambridge university press.

[6] Blei, D. and Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* **12** 2461–2488.

[7] Bloem-Reddy, B. and Orbanz, P. (2017). Preferential Attachment and Vertex Arrival Times. *arXiv preprint arXiv:1710.02159*.

[8] Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* 929–953.
[9] Broderick, T., Jordan, M. I. and Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian analysis* **7** 439–476.
[10] Cai, D., Campbell, T. and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4249–4257. Curran Associates, Inc.
[11] Camerlenghi, F., Lijoi, A., Orbanz, P., Prünster, I. et al. (2019a). Distribution theory for hierarchical processes. *The Annals of Statistics* **47** 67–92.
[12] Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., Rodríguez, A. et al. (2019b). Latent nested nonparametric priors (with discussion). *Bayesian Analysis* **14** 1303–1356.
[13] Caron, F., Davy, M. and Doucet, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007* 33–40.
[14] Caron, F. and Fox, E. (2017). Sparse Graphs using Exchangeable Random Measures. *Journal of the Royal Statistical Society B* **79** 1295-1366. Part 5.
[15] Caron, F. and Rousseau, J. (2017). On sparsity and power-law properties of graphs based on exchangeable point processes. *arXiv preprint arXiv:1708.03120*.
[16] Caron, F., Neiswanger, W., Wood, F., Doucet, A. and Davy, M. (2017). Generalized Pólya urn for time-varying Pitman–Yor processes. *Journal of Machine Learning Research* **18** 1-32.
[17] Chen, C., Rao, V., Buntime, W. and Teh, Y. W. (2013). Dependent Normalized Random Measures. In *International Conference on Machine Learning*.
[18] Crane, H. and Dempsey, W. (2017). Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*.
[19] Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, second ed. Springer.
[20] De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence* **37** 212–229.
[21] Del Moral, P. (2004). Feynman-kac formulae. In *Feynman-Kac Formulae* 47–93. Springer.
[22] Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 411–436.
[23] Doucet, A., de Freitas, N. and Gordon, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.
[24] Durrett, R., Schweinsberg, J. et al. (2005). Power laws for family sizes in a duplication model. *The Annals of Probability* **33** 2094–2126.
[25] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability* **23** 1721–1754.
[26] Feinerer, I., Hornik, K. and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* **25** 1–54.
[27] Feinerer, I. and Hornik, K. (2018). tm: Text Mining Package R package version 0.7-6.
[28] Feller, W. (1971). *An introduction to probability theory and its applications* **2**. John Wiley & Sons.
[29] Foti, N. J. and Williamson, S. A. (2015). A survey of non-exchangeable priors for

Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence* **37** 359–371.
[30] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv* **4** 88.
[31] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical sciences* **138** 5674–5685.
[32] Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association* **101** 179–194.
[33] James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*.
[34] James, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics* **33** 1771–1799.
[35] James, L. F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36** 76–97.
[36] Janson, S. (2017). On edge exchangeable random graphs. *arXiv preprint arXiv:1702.06396*.
[37] Karlin, S. (1967). Central Limit Theorems for Certain Infinite Urn Schemes. *Journal of Mathematics and Mechanics* **17** 373-401.
[38] Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78.
[39] Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–22.
[40] Kingman, J. F. C. (1978). Random partitions in population genetics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **361** 1–20. The Royal Society.
[41] Kingman, J. F. C. (1993). *Poisson processes* **3**. Oxford University Press, USA.
[42] Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability* 705–711.
[43] Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16** 526–558.
[44] Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[45] Lijoi, A., Mena, R. H. and Prünster, I. (2005a). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Statistical Inference for Stochastic Processes* **8** 283–309.
[46] Lijoi, A., Mena, R. H. and Prünster, I. (2005b). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100** 1278–1291.
[47] Lijoi, A., Mena, R. and Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786.
[48] Lijoi, A., Mena, R. H. and Prünster, I. (2007b). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 715–740.
[49] Lijoi, A. and Prünster., I. (2003). On a normalized random measure with independent increments relevant to Bayesian nonparametric inference. In *Proceedings of the 13th European Young Statisticians Meeting* 123-134. Bernoulli Society.
[50] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) Cambridge University Press.

[51] Lijoi, A., Nipoti, B., Prünster, I. et al. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291.
[52] Lin, D., Grimson, E. and Fisher, J. W. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in neural information processing systems* 1396–1404.
[53] MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.
[54] McAuley, J., Targett, C., Shi, Q. and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* 43–52. ACM.
[55] Miller, J., Betancourt, B., Zaidi, A., Wallach, H. and Steorts, R. (2015). Microclustering: When the Cluster Sizes Grow Sublinearly with the Size of the Data Set. *arXiv:1512.00792v1*.
[56] Müller, P. and Rodriguez, A. (2013). *Chapter 8: Random Partition Models*. In *Nonparametric Bayesian Inference. NSF-CBMS Regional Conference Series in Probability and Statistics* **Volume 9** 87–92. Institute of Mathematical Statistics and American Statistical Assocation, Beachwood, Ohio, USA; and Alexandria, Virginia, USA.
[57] Nieto-Barajas, L. E., Prünster, I. and Walker, S. G. (2004). Normalized random measures driven by increasing additive processes. *The Annals of Statistics* **32** 2343–2360.
[58] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92** 21–39.
[59] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102** 145–158.
[60] Pitman, J. (2003). *Poisson-Kingman partitions*. In *Statistics and science: a Festschrift for Terry Speed. Lecture Notes–Monograph Series* **Volume 40** 1–34. Institute of Mathematical Statistics, Beachwood, OH.
[61] Pitman, J. (2006). *Combinatorial stochastic processes. Ecole d'été de Probabilité de Saint-Flour - 2002*. Springer.
[62] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900.
[63] Rao, V. and Teh, Y. W. (2009). Spatial normalized gamma processes. In *Advances in neural information processing systems* 1554–1562.
[64] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31** 560–585.
[65] Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.
[66] Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems* 1585–1592.
[67] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* 985–992. Association for Computational Linguistics.
[68] Zanella, G., Betancourt, B., Wallach, H., Miller, J. W., Zaidi, A. and Steorts, R. C. (2016). Flexible Models for Microclustering with Application to Entity Resolution. In *Advances in Neural Information Processing Systems* 1417–1425.

[69] Zhou, M., Favaro, S. and Walker, S. G. (2016). Frequency of Frequencies Distributions and Size Dependent Exchangeable Random Partitions. *Journal of the American Statistical Association*.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Non-exchangeable random partition models for microclustering
Publication Status	☐Published ☐ Accepted for Publication ☑Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style
Publication Details	Under major revision for the Annals of Statistics journal

Student Confirmation

Student Name:	Giuseppe Di Benedetto		
Contribution to the Paper	I am the first author on this paper and I have contributed in deriving the theoretical asymptotic results and implementing inference scheme to apply the model to different scenarios, namely in the unsupervised and supervised settings. About the applications, I have run the experiments on both synthetic and real datasets to compare the proposed model against a well-known alternative model.		
Signature		Date	20/04/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor François Caron			
Supervisor comments			
Signature		Date	20/04/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Non-exchangeable feature allocation models with sublinear growth of feature sizes

Accepted for publication in the Proceedings of the International Conference on Artificial Intelligence and Statistics 2020

Non-exchangeable feature allocation models with sublinear growth of the feature sizes

Giuseppe Di Benedetto
University of Oxford

François Caron
University of Oxford

Yee Whye Teh
University of Oxford
DeepMind

Abstract

Feature allocation models are popular models used in different applications such as unsupervised learning or network modeling. In particular, the Indian buffet process is a flexible and simple one-parameter feature allocation model where the number of features grows unboundedly with the number of objects. The Indian buffet process, like most feature allocation models, satisfies a symmetry property of exchangeability: the distribution is invariant under permutation of the objects. While this property is desirable in some cases, it has some strong implications. Importantly, the number of objects sharing a particular feature grows linearly with the number of objects. In this article, we describe a class of non-exchangeable feature allocation models where the number of objects sharing a given feature grows sublinearly, where the rate can be controlled by a tuning parameter. We derive the asymptotic properties of the model, and show that such model provides a better fit and better predictive performances on various datasets.

1 Introduction

Feature allocation models are probabilistic models over multisets (Broderick et al., 2013), which represent the allocation of a set of objects to a (potentially unbounded) set of features. Contrary to models on partitions, where each object is assigned to a single group, these models allow each object to be allocated more than one feature. For example, the objects may be movies, and the features correspond to the actors performing in that movie. Movies have different number of actors, and actors participate to a different number of movies. Informally, feature allocations models can be interpreted as distributions over sparse binary matrices whose rows represent the objects and whose columns represent the features; non-zero entries indicate the allocation of features to objects. Feature allocation models have been used in various applications, including topic modeling (Williamson et al., 2010b), image analysis (Zhou et al., 2011a), network modeling (Palla et al., 2012; Cai et al., 2016) or inference in tumor heterogeneity (Lee et al., 2015; Xu et al., 2015).

A classical assumption is that of exchangeability: the feature allocation model is invariant over permutations of the objects. Taking the interpretation as a binary matrix, the matrix is invariant over permutations of its rows. The most remarkable example of an exchangeable feature allocation is the Indian buffet process (IBP) (Ghahramani and Griffiths, 2006; Thibaux and Jordan, 2007; Griffiths and Ghahramani, 2011) which has a simple and intuitive generative model. The model allows the number of features K_n to grow unboundedly with the number of objects n at a logarithmic rate. The IBP admits a three-parameter generalisation, the stable IBP (Teh and Gorur, 2009), which is also exchangeable. For some values of its parameters, the stable IBP can capture a power-law behavior, where the number of features grows at a rate of n^σ for some $\sigma \in (0, 1)$.

While the exchangeability assumption is reasonable for many applications and has computational advantages, it may not be adequate in some cases. In particular, assuming exchangeability implies that, out of n objects, the number $m_{n,j} \leq n$ of objects having a particular feature j (called feature's size) scales linearly with the number of objects n (see Figure 4(b) for an illustration). Such assumption may be undesirable. For instance, even if she's a prolific actress, one does not expect the filmography of Meryl Streep to scale linearly with the overall number of movies released.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

The objective of this article is to present a model that can have a sublinear growth of the size of the features, while retaining the properties of the (stable) Indian buffet in terms of overall growth of the number of features and power-law properties. The article is organised as follows. Section 2 provides some background on feature allocation models and the IBP. Our non-exchangeable model is presented in Section 3 and its asymptotic properties given in Section 4. In Section 5 we derive a Gibbs sampler for posterior inference. Experimental results are presented in Section 6. Related approaches are discussed in Section 7.

2 Background

2.1 Feature allocation models and the Indian buffet process

A feature allocation (Broderick et al., 2012) $f_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$ is a multiset of a set of objects $\{1, \ldots, n\}$ such that $A_{n,k}$, $k = 1, \ldots, K_n$ are (possibly overlapping) non-empty subsets of $\{1, \ldots, n\}$; $A_{n,k}$ represents the set of objects having feature k and K_n is the number of different features shared by the n objects. For example, $f_4 = \{\{1,3\}, \{1,3\}, \{3,4\}, \{4\}, \{4\}\}$ indicates that object 1 has features 1 and 2, object 2 has no feature, object 3 has features 1, 2, 3 and object 4 has features 3, 4 and 5. Note that the labelling of the features as 1 to 5 is arbitrary.

A feature allocation model is a distribution over a growing family of random feature allocations $(f_n)_{n=1,2,\ldots}$. The most popular feature allocation model is the Indian buffet process, where f_n has distribution proportional to

$$\Pr(f_n) \propto \eta^{K_n} \prod_{k=1}^{K_n} \frac{(\widetilde{m}_{n,k} - 1)!(n - \widetilde{m}_{n,k})!}{n!} \qquad (1)$$

where $\widetilde{m}_{n,k}$ is the number of objects having feature k, for $k = 1, \ldots, K_n$ and $\eta > 0$ is a tuning parameter. The IBP admits a three-parameter generalisation, called stable IBP (Teh and Gorur, 2009), where[1]

$$\Pr(f_n) \propto \frac{\eta^{K_n}}{\Gamma(1-\sigma)^{K_n}} \prod_{k=1}^{K_n} \frac{\Gamma(\widetilde{m}_{n,k} - \sigma)\Gamma(n - \widetilde{m}_{n,k} + \zeta)}{\Gamma(n + \zeta - \sigma)} \qquad (2)$$

with $\eta > 0$, $\sigma \in (-\infty, 1)$ and $\zeta > 0$. It reduces to the one-parameter IBP when $\zeta = 1$ and $\sigma = 0$. When $\sigma > 0$, the model exhibits power-law properties.

A convenient way to encode a feature allocation model is via a collection of atomic random measures $(Z_1, \ldots, Z_n)$ on some space (here $\mathbb{R}_+$) where for $i = 1, \ldots, n$,

$$Z_i = \sum_{j \geq 1} z_{ij} \delta_{\theta_j} \qquad (3)$$

with $z_{ij} = 1$ if object i has feature j and $(\theta_j)_{j \geq 1}$ are continuous random variables on $\mathbb{R}_+$ whose distribution is irrelevant here. Note that in this notation there is no particular ordering of the features, and we now use the index j instead of k to emphasize this difference.

2.2 Completely random measures

A homogeneous completely random measure (CRM) (Kingman, 1967; Lijoi and Prünster, 2010) on $\mathbb{R}_+ = [0, \infty)$ is an almost surely discrete random measure

$$B = \sum_{j \geq 1} \omega_j \delta_{\theta_j} \qquad (4)$$

where $\{(\omega_j, \theta_j)_{j \geq 1}\}$ are points of a Poisson point process on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\nu(d\omega, d\theta) = \rho(\omega) d\omega \alpha(\theta) d\theta$ where ρ and α satisfy

$$\int_0^\infty (1 - e^{-\omega}) \rho(\omega) d\omega < \infty \quad \text{and} \quad \int_A \alpha(\theta) d\theta < \infty$$

for any bounded set $A \subset \mathbb{R}_+$. The above condition ensures that $B(A) < \infty$ almost surely. The CRM is said to be infinite-activity if $\int_0^\infty \rho(\omega) d\omega = \infty$. In this case, the measure B has a countably infinite support on any non-empty interval $A \subset \mathbb{R}_+$.

The (stable) Indian buffet process admits the following hierarchical construction via CRMs (Thibaux and Jordan, 2007; Teh and Gorur, 2009). Let $\widetilde{B} = \sum_j \pi_j \delta_{\theta_j}$ be a homogeneous CRM with

$$\rho(\pi) = \frac{\eta}{\Gamma(1-\sigma)} \pi^{-1-\sigma} (1 - \pi)^{\zeta - 1} \, 1_{\pi \in (0,1)} \qquad (5)$$

and $\alpha(\theta) = 1_{\theta \leq 1}$. The parameter π_j can be interpreted as the popularity of feature j. For $i = 1, \ldots, n$ define the atomic measure Z_i as in Equation (3) with

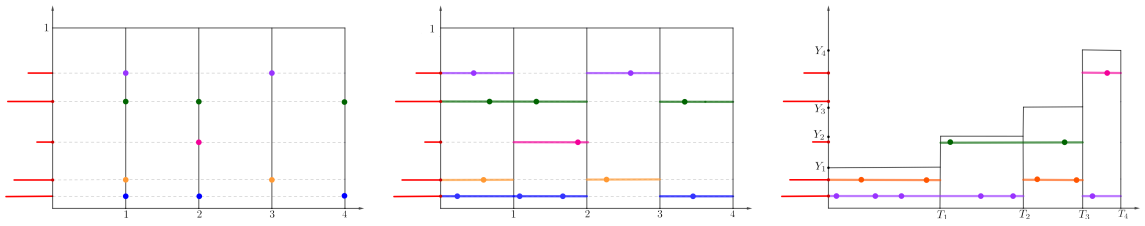
$$z_{ij} \mid \pi_j \sim \text{Ber}(\pi_j)$$

for $j \geq 1$, where $\text{Ber}(\pi)$ denotes the Bernoulli distribution with parameter $\pi \in [0, 1]$.

2.3 An alternative construction for the Indian buffet process

We present here an equivalent construction for the IBP. It relies on the introduction of latent Poisson processes, similar to the construction proposed by Di Benedetto et al. (2017) for non-exchangeable random partitions, adapted to feature models. This approach will give the intuition for the generalisation of the IBP introduced in the next section.

[1] Note that we use a slightly different parameterisation compared to that of Teh and Gorur (2009).



(a) Beta-Bernoulli process (b) Continuous time IBP (c) Non-exchangeable point process

Figure 1: Point process defining the feature models: red sticks represent weights of the CRM, points and bars of the same colour refer to the same feature. (a) Beta-Bernoulli process; (b) Point process construction of the IBP: each rectangle of basis $[T_{i-1}, T_i]$ represents the $i-th$ object whose features are depicted by the coloured bars, which are present if at least one of the corresponding points falls in the rectangle; (c) Non-exchangeable point process.

Using the change of variable $\omega_j = -\log(1-\pi_j) \in (0,\infty)$, the random measure $B = \sum_j \omega_j \delta_{\theta_j}$ is itself a CRM with $\alpha(\theta) = 1_{\theta \leq 1}$ and

$$\rho(\omega) = \frac{\eta}{\Gamma(1-\sigma)} e^{-\zeta\omega}(1-e^{-\omega})^{-1-\sigma}. \tag{6}$$

We call (6) a transformed stable beta (TSB) Lévy measure, and the associated random measure B a TSB process (TSBP). Consider for each j a homogeneous Poisson process $N_j(t)$ on $\mathbb{R}_+$ with rate ω_j. Let $z_{ij} = 1_{N_j(i)-N_j(i-1)>0}$ be a binary variable indicating if there is any event in the time interval $[i-1, i)$. Then

$$\begin{aligned}\Pr(z_{ij} = 1 \mid \omega_j) &= \Pr(N_j(i) - N_j(i-1) > 0 \mid \omega_j) \\ &= 1 - e^{-\omega_j} = \pi_j.\end{aligned}$$

This construction is illustrated in Figure 1.

3 Non-exchangeable feature allocation model

Let $B = \sum_{j\geq 1} \omega_j \delta_{\theta_j}$ be a homogeneous completely random measure on $\mathbb{R}_+$. While the model can be defined for a general CRM, we focus here on the case where $\alpha(\theta) = 1$ and ρ is either the transformed stable beta measure (6), or the generalised gamma (GG) measure (Hougaard, 1986), given by

$$\rho(\omega) = \frac{\eta}{\Gamma(1-\sigma)} \omega^{-1-\sigma} e^{-\zeta\omega} 1_{\omega>0} \tag{7}$$

where $\eta > 0$, $\sigma \in (-\infty, 1)$, $\zeta > 0$. As we will show in Section 4, both models lead to the same asymptotic behavior. The GG measure has however a conjugate form that makes it more amenable to posterior inference, as detailed in Section 5. A CRM with mean measure (7) will be called a Generalised Gamma Process (GGP).

Inspired by the work on random partition models by Di Benedetto et al. (2017), let $(Y_n)_{n\geq 1}$ and $(T_n)_{n\geq 1}$ be two increasing sequences of positive reals defined as

$$T_n = n^{\frac{1}{\xi+1}}, \qquad Y_n = n^{\frac{\xi}{\xi+1}} \tag{8}$$

where $\xi \geq 0$ is a tuning parameter. Define the sequence $(\Delta_n)_{n\geq 1}$ by $\Delta_n := T_n - T_{n-1}$. The feature allocation of an object i is represented by a random measure Z_i on $\mathbb{R}_+$ as in Equation (3) where

$$z_{ij} \mid \omega_j, \theta_j \sim \text{Ber}\left(1 - e^{-\omega_j \Delta_i 1_{\theta_j \leq Y_i}}\right). \tag{9}$$

Note that $z_{ij} = 0$ a.s. if $\theta_j > Y_i$.

The model admits the following construction using a latent Poisson process. For each j, let $(N_j(t))_{t>0}$ be a homogeneous Poisson process with rate ω_j. Then the binary variable $z_{ij} = 1_{N_j(T_i)-N_j(T_{i-1})>0} 1_{\theta_i \leq Y_i}$ has distribution (9). See the illustration in Figure 1(c).

Note that if one sets $\xi = 0$, we have $\Delta_i = Y_i = 1$. The distribution in the right-handside of Equation (9) does not depend on i and the associated feature allocation model is therefore exchangeable. If additionally we use the mean measure ρ as in Equation (6), we recover the three-parameter IBP as a special case.

The model is parameterised by the four parameters η, σ, ζ and ξ. In the next section we show how these parameters tune the asymptotic properties of the model. The critical parameter is the parameter ξ and we show that for $\xi > 0$, the features' sizes grow sublinearly with n, at a rate controlled by this parameter.

4 Asymptotic Properties

4.1 Notations

In this section we use the following notations for asymptotics. $a_n \overset{n\to\infty}{\sim} b_n$ means that $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ and $a_n \overset{n\to\infty}{\asymp} b_n$ means $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$ and $\limsup_{n\to\infty} \frac{b_n}{a_n} < \infty$ (that is, a_n is of the same order as b_n). Let

$$Z_i(\mathbb{R}_+) = \sum_{j\geq 1} z_{ij}$$

be the number of features for object i. For each feature $j = 1, 2, \ldots$, let us define its size as

$$m_{n,j} = \sum_{i=1}^{n} z_{ij}$$

i.e. the number of objects having that feature, and let

$$m_n = \sum_{j\geq 1} m_{n,j} = \sum_{i=1}^{n} \sum_{j\geq 1} z_{ij}$$

be the total number of features allocated to the n objects. Denote

$$K_n = \sum_j 1_{m_{n,j}>0}$$

the number of unique features, and

$$K_{n,r} = \sum_j 1_{m_{n,j}=r}$$

the number of unique features allocated to r objects, where $r \geq 1$. Note that $\sum_r K_{n,r} = K_n$.

We will use the following notation for the Laplace exponent and the tilted moments

$$\psi(t) = \int_0^\infty \left\{1 - e^{-\omega t}\right\} \rho(\omega) d\omega$$
$$\kappa(m, u) = \int_0^\infty \omega^m e^{-u\omega} \rho(\omega) d\omega.$$

for any integer $m \geq 1$ and any $u > 0$. $\kappa(m, 0)$ correspond to the m-th moment of the measure ρ. For the GG measure, we have $\psi(t) = \frac{\eta}{\sigma}((t+\zeta)^\sigma - \zeta^\sigma)$ and $\kappa(m, u) = \eta \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}(u+\zeta)^{\sigma-m}$. For the TSB, we have $\psi(1) = \frac{\eta\Gamma(\zeta)}{\Gamma(\zeta+1-\sigma)}$, but there is no analytical expression for $\psi(t)$ and $\kappa(m, u)$.

4.2 Asymptotic properties when $\xi = 0$

We first recall here, for comparison, the asymptotic properties of our model for $\xi = 0$ when B is a GGP or a TSBP. As mentioned in Section 3, the model is exchangeable in that case and if B is the TSBP, it reduces to the three parameter IBP, whose asymptotic properties are well known (Ghahramani and Griffiths, 2006; Thibaux and Jordan, 2007; Teh and Gorur, 2009; Broderick et al., 2012). Asymptotics when B is the GGP model are similar. First, for the number of features of object i,

$$Z_i(\mathbb{R}_+) \sim \text{Poisson}\left(\psi(1)\right) \tag{10}$$

where $\psi(1) = \frac{\eta\Gamma(\zeta)}{\Gamma(\zeta+1-\sigma)}$ for the TSBP and $\psi(1) = \frac{\eta}{\sigma}((\zeta+1)^\sigma - \zeta^\sigma)$ for the GGP.

Given the CRM B, we have almost surely

$$m_{n,j} \overset{n\to\infty}{\sim} n(1 - e^{-\omega_j}) \tag{11}$$
$$m_n \overset{n\to\infty}{\sim} n \sum_{j\geq 1} (1 - e^{-\omega_j}). \tag{12}$$

Hence both the features' sizes and the total number of features grow linearly with n, as a consequence of the exchangeability assumptions. For the number of unique features, we have almost surely

$$K_n \overset{n\to\infty}{\sim} \begin{cases} \eta \log(n) & \sigma = 0 \\ \frac{\eta}{\sigma} n^\sigma & \sigma \in (0, 1) \end{cases}. \tag{13}$$

Finally we have, for the proportions of features allocated to r objects for $\sigma \in (0, 1)$

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \tag{14}$$

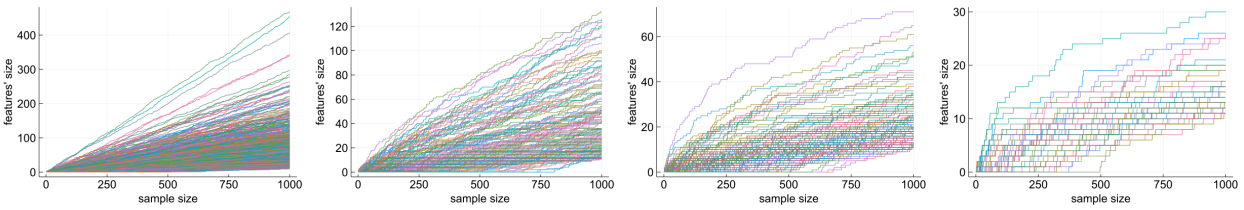
almost surely as n tends to infinity, and $K_{n,r}/K_n \to 0$ almost surely otherwise for all $r \geq 1$. Equation (14) corresponds to a power-law behaviour as $\Gamma(r-\sigma)/r! \simeq r^{-(1+\sigma)}$ for large r.

4.3 Asymptotic properties when $\xi > 0$

When $\xi > 0$, the model is non-exchangeable. In this case, most of the properties of the exchangeable case are retained, such as the Poisson number of features per object (see Proposition 1 below), the linear growth of the total number of features (Proposition 2) and the power-law behaviour, solely controlled by the parameter σ (Proposition 5). The number of unique features K_n still grows sublinearly, but at a rate now controlled by both σ and ξ (Proposition 4). The key difference is that, as shown in Proposition 3, the features' sizes grow at a rate $n^{1/(1+\xi)}$, which is sublinear for $\xi > 0$.

Propositions 1, 2 and 3 are valid for any choice of Lévy measure ρ. Propositions 4 and 5 hold for any Lévy measure ρ such that

$$\psi(t) \overset{t\to\infty}{\sim} \begin{cases} \eta \log(t) & \sigma = 0 \\ \frac{\eta}{\sigma} t^\sigma & \sigma \in (0, 1) \end{cases}. \tag{15}$$



(a) IBP (b) $\xi = 0.25$ (c) $\xi = 1$ (d) $\xi = 2$

Figure 2: Evolution of the features' sizes with the sample size for different synthetic datasets: (a) IBP; (b-d) Non-exchangeable model for $\xi = 0.25, 1, 2$. The feature size growth is (a) n, (b) $n^{0.8}$, (c) $n^{1/2}$ and (d) $n^{1/3}$.

Equation (15) holds in particular for the GGP and the TSBP. The proofs are given in the Supplementary Material.

The first proposition shows that the number of features per object is Poisson distributed, with a rate converging to a constant.

Proposition 1 (number of features per object)**.** *Consider the model of Section 3 with a generic ρ and $\xi > 0$. For each object n, we have*

$$Z_n(\mathbb{R}_+) \sim \text{Poisson}\left(\mathbb{E}\left[Z_n(\mathbb{R}_+)\right]\right) \tag{16}$$

where

$$\mathbb{E}\left[Z_n(\mathbb{R}_+)\right] = Y_n \psi(\Delta_n) \to (1+\xi)^{-1}\kappa(1,0),$$

with $\kappa(1,0) = \eta\zeta^{\sigma-1}$ for the GGP.

The next result is on the asymptotic behaviour of the total number of features observed in the first n objects.

Proposition 2 (total number of features)**.** *Consider the model of Section 3 with a generic ρ and $\xi > 0$. The total number of features observed in the first n objects satisfies*

$$m_n \overset{n\to\infty}{\sim} (1+\xi)^{-1}\kappa(1,0)\, n.$$

The following proposition describes the growth of the features' sizes $m_{n,j}$ with respect to n.

Proposition 3 (number of objects per feature)**.** *Consider the model of Section 3 with a generic Lévy measure ρ and $\xi \geq 0$. For each $j \geq 1$ and conditionally on the CRM B, we have that, almost surely,*

$$m_{n,j} \overset{n\to\infty}{\sim} \omega_j T_n = \omega_j n^{1/(1+\xi)}.$$

The features' sizes therefore grow linearly if $\xi = 0$, and sublinearly when $\xi > 0$, with a rate decreasing at ξ increases. This is illustrated in Figure 2.

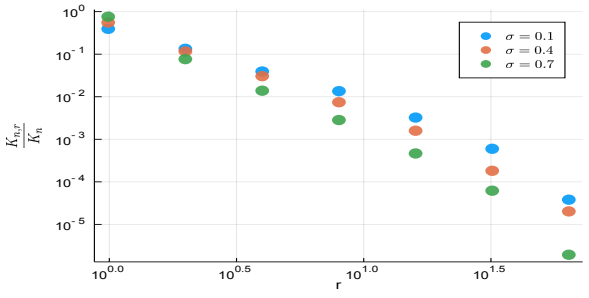


Figure 3: Log-log plot of the proportions of features shared by a fixed number of objects. Simulated examples for $\sigma \in \{0.1, 0.4, 0.7\}$, with $\xi = 1$ and $\eta = 30$.

Proposition 4 (number of unique features)**.** *Consider the model of Section 3 with $\xi \geq 0$ and a Lévy measure ρ whose Laplace exponent ψ satisfies Equation* (15)*. The number of unique features K_n is such that*

$$K_n \overset{n\to\infty}{\sim} \begin{cases} \eta\, n^{\frac{\xi}{\xi+1}} \log(n) & \sigma = 0 \\ \eta\, \frac{\Gamma(\xi+1)\Gamma(\sigma)}{\Gamma(\sigma+\xi+1)}\, n^{\frac{\xi+\sigma}{\xi+1}} & \sigma \in (0,1) \end{cases}. \tag{17}$$

Analogously to the three-parameter IBP model, the proposed non-exchangeable feature allocation model exhibits the power-law property for the number of features shared by a given number of objects (see Figure 3 for an illustration of this property).

Proposition 5 (power-law properties)**.** *Consider the model of Section 3 with $\xi > 1 - \sigma$ and a Lévy measure ρ whose Laplace exponent ψ satisfies Equation* (15)*. We have, almost surely as n tends to infinity*

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}.$$

We conjecture that the condition $\xi > 1 - \sigma$ introduced in the above proposition, which is needed for the proof, is not a necessary condition, and that the above proposition actually holds for any $\xi \geq 0$.

5 Inference

5.1 Data augmentation and conditional distribution

For posterior inference, we introduce a latent process, similar to that of Caron (2012). For $i = 1, \dots, n$, let $U_i = \sum_{j \geq 1} u_{ij} \delta_{\theta_j}$ where $u_{ij} = 1$ if $z_{ij} = 0$, and

$$u_{ij} \mid z_{ij} = 1, \omega_j \sim \text{tExp}(\Delta_i \omega_j, 1) \tag{18}$$

$\text{tExp}(\lambda, T)$ denotes the right-truncated exponential distribution with rate $\lambda > 0$ and truncation $T > 0$ with probability density function $\lambda e^{-\lambda x}(1 - e^{-T\lambda})^{-1} 1_{x<T}$. Note that by construction, $z_{ij} = 1_{u_{ij}<1}$ is a deterministic function of u_{ij}.

Denote $\widetilde{\theta}_1, \dots, \widetilde{\theta}_{K_n}$ the set of θ_j's such that $m_{n,j} > 0$, $\widetilde{\omega}_1, \dots, \widetilde{\omega}_{K_n}$ the corresponding weights, $(\widetilde{u}_{i,k})_{i=1,\dots,n;k=1,\dots,K_n}$ the corresponding latent variables, $(\widetilde{z}_{i,k})_{i=1,\dots,n;k=1,\dots,K_n}$ the corresponding binary variables and $\widetilde{m}_{n,1}, \dots, \widetilde{m}_{n,K_n}$ the associated features' sizes. For $k = 1, \dots, K_n$, let $Y_k^* = \inf_i\{Y_i : \widetilde{z}_{ik} = 1\}$. With analogous calculations to Caron (2012) it is possible to write down the conditional distributions

$$\begin{aligned}
&\mathbb{P}(U_1, \dots, U_n \,|\, B) \\
&= e^{-\sum_{i=1}^n \Delta_i \left(\bar{B}(Y_i) - \sum_{k=1}^{K_n} \widetilde{\omega}_k 1_{\widetilde{\theta}_k < Y_i}\right)} \left(\prod_{i=1}^n \Delta_i^{\sum_{k=1}^{K_n} \widetilde{z}_{ik}}\right) \\
&\left(\prod_{k=1}^{K_n} \widetilde{\omega}_k^{\widetilde{m}_{n,k}} \, e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} 1_{\widetilde{\theta}_k < Y_k^*}\right)
\end{aligned}$$

where $\bar{B}(Y_i) = B((0, Y_i)) = \sum_j \omega_j 1_{\theta_j < Y_i}$. Using the results on Poisson partition calculus (James, 2002), we can marginalize out the CRM

$$\begin{aligned}
\mathbb{P}(U_1, \dots, U_n) = e^{-\int_0^{Y_n} \psi(f_n(\theta)) d\theta} &\left(\prod_{i=1}^n \Delta_i^{\sum_{k=1}^{K_n} \widetilde{z}_{ik}}\right) \\
&\times \prod_{k=1}^{K_n} \kappa\left(\widetilde{m}_{n,k}, \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}\right) 1_{\widetilde{\theta}_k < Y_k^*}
\end{aligned} \tag{19}$$

with $f_n(\theta) := \sum_{i=1}^n \Delta_i 1_{\theta \leq Y_i}$. The conditional distribution is

$$B \,|\, (\widetilde{u}_{ik}) \stackrel{d}{=} B^* + \sum_{k=1}^{K_n} \widetilde{\omega}_k \delta_{\widetilde{\theta}_k} \tag{20}$$

where B^* is an inhomogeneous CRM with Lévy measure $\nu^*(d\omega, d\theta) = e^{-\omega \sum_{i=1}^n \Delta_i 1_{\theta < Y_i}} \rho(\omega) d\omega d\theta$, and it is independent of $(\widetilde{\omega}_k, \widetilde{\theta}_k)_{k=1,\dots,K_n}$ which are independent (in k), with joint posterior probability density

$$p(\widetilde{\omega}_k, \widetilde{\theta}_k | \, (\widetilde{u}_{ik})) \propto \widetilde{\omega}_k^{\widetilde{m}_{n,k}} e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} \rho(\widetilde{\omega}_k) 1_{\widetilde{\theta}_k < Y_k^*}.$$

5.2 Gibbs sampler

Assume that we have observed the binary feature allocations $(\widetilde{z}_{i,k})_{i=1,\dots,n,k=1,\dots,K_n}$. We take an empirical Bayes approach, and wish to approximate the posterior distribution

$$p((\widetilde{u}_{i,k}), (\widetilde{\omega}_k), (\widetilde{\theta}_k) \mid (\widetilde{z}_{i,k}), \widehat{\phi}) \tag{21}$$

where $\widehat{\phi}$ is a point estimate of the hyperparameters $\phi = (\xi, \sigma, \eta, \zeta)$. We explain in the next section how to obtain consistent estimators of the hyperparameters using our asymptotic results. For the GGP, a Gibbs sampler with target distribution (21) can be derived as follows

- For $i = 1, \dots, n$, $k = 1, \dots, K_n$, such that $\widetilde{z}_{i,k} = 1$ sample $\widetilde{u}_{ik} \mid \text{rest} \sim \text{tExp}(\Delta_i \widetilde{\omega}_k, 1)$.
- For $k = 1, \dots, K_n$, sample

$$\widetilde{\omega}_k \mid \text{rest} \sim \text{Gamma}\left(\widetilde{m}_{n,k} - \sigma, \zeta + \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}\right).$$

- For $k = 1, \dots, K_n$, sample

$$\widetilde{\theta}_k \mid \text{rest} \sim e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ik} 1_{\widetilde{\theta}_k < Y_i}} 1_{\widetilde{\theta}_k < Y_k^*}.$$

The last distribution is piecewise constant, and one can therefore straightforwardly sample from it.

5.3 Estimation of the Hyperparameters

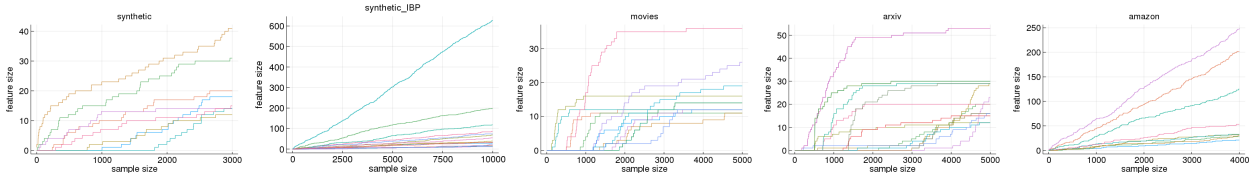
We use here the asymptotic results of Section 4 to derive consistent estimators of the hyperparameters. The parameter ξ controls the features' growth and is consistently estimated using Proposition 3, while Proposition 5 can be used to consistently estimate the parameter σ from the proportion of features of size one:

$$\hat{\xi} = \frac{\log n}{\log \max_{k=1,\dots,K_n} \widetilde{m}_{n,k}} - 1, \qquad \hat{\sigma} = \frac{K_{n,1}}{K_n}.$$

Using Propositions 2 and 4 we have the following consistent estimators for the parameters η and ζ (GGP):

$$\hat{\eta} = \frac{\Gamma(\hat{\sigma} + \hat{\xi} + 1)}{\Gamma(\hat{\sigma})\Gamma(\hat{\xi} + 1)} \frac{K_n}{n^{\frac{\hat{\xi}+\hat{\sigma}}{1+\hat{\sigma}}}}, \qquad \hat{\zeta} = \left(\frac{(\hat{\xi} + 1)\, m_n}{\hat{\eta}\, n}\right)^{\frac{1}{\hat{\sigma}+1}}.$$

Although these estimators are consistent, it should be noted that $\hat{\xi}$ depends logarithmically on the sample size, leading potentially to a slow convergence. However, as shown in the simulated experiments (see section 6 and Table 1), the estimated values of the hyperparamenters ξ and σ are close to the true values.



(a) Synthetic - NE (b) Synthetic - IBP (c) Movies (d) ArXiv (e) Amazon

Figure 4: Evolution of some of features' sizes with respect to the sample size for different datasets.

	Non-exchangeable			Indian Buffet process		
	$\hat{\sigma}_{\text{NE}}$	L2 error	90% CI	$\hat{\sigma}_{\text{IBP}}$	L2 error	90% CI
Synthetic	0.61	15.41	$[15.32, 15.50]$	0.51	19.93	$[19.72, 20.13]$
Synthetic IBP	0.41	30.47	$[30.37, 30.54]$	0.39	31.19	$[31.11, 31.28]$
IMDb movies	0.31	21.27	$[21.20, 21.35]$	0.07	36.78	$[36.59, 37.04]$
Arxiv	0.48	7.96	$[7.89, 7.96]$	0.31	13.95	$[13.81, 14.08]$
Amazon	0.57	142.06	$[141.74, 142.38]$	0.61	142.62	$[142.25, 143.02]$

Table 1: Estimated values for the parameter σ obtained by the Indian Buffet process ($\hat{\sigma}_{\text{IBP}}$) and our non-exchangeable model ($\hat{\sigma}_{\text{NE}}$). Mean and quantiles of the L_2 predictive error.

6 Experiments

In this section we present some experiments to compare the proposed model with GG measure against the three-parameter IBP. We estimate the parameters on a training set, and compute the empirical normalized L_2 error between the true and predicted features' sizes on a test set

$$\text{Err} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{K_{n_{\text{train}}}} \sum_{i=n_{\text{train}}+1}^{n_{\text{train}}+n_{\text{test}}} (\widetilde{m}_{i,k}^{\text{true}} - \widetilde{m}_{i,k})^2$$

where $\widetilde{m}_{i,k}^{\text{true}}$ is the observed size of feature k, n_{train} and n_{test} are the numbers of objects in the training set and test set respectively and $K_{n_{\text{train}}}$ is the number of features observed in the training set. We aim at reporting the posterior mean L_2 error $\mathbb{E}[\text{Err} \mid (\widetilde{z}_{ik})_{i=1,\ldots,n_{\text{train}},k=1,\ldots,K_{n_{\text{train}}}}]$ and the 90% credible interval of the posterior distribution of the L_2 error. Details on the posterior predictive under our non-exchangeable model are given in Section A of the Supplementary Material. In what follows we denote the total number of objects in the dataset as n. Given the closed form expression (2), the hyperparameters (η, σ, ζ) of the three-parameter IBP are estimated by maximum likelihood.

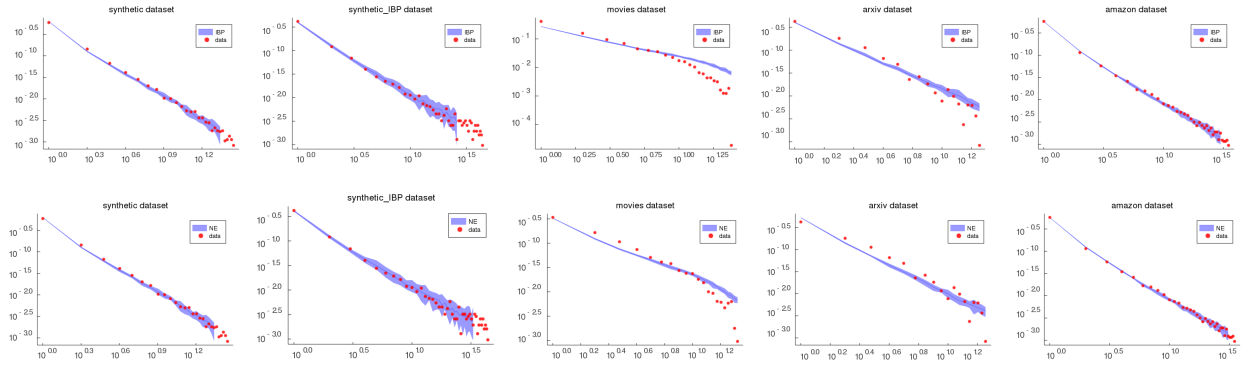
Synthetic data. The models are first tested on synthetic datasets. We generate a binary matrix with $n = 3000$ rows from our non-exchangeable model with parameters $(\eta, \sigma, \xi, \zeta) = (30, 0.5, 1, 1)$, with a training set of size $n_{\text{train}} = 1000$. Figure 4 shows that the feature' sizes are clearly sublinear, and as expected the prediction error is higher under the IBP model, as shown in Table 1. The second synthetic dataset of size $n = 10000$ is generated from the IBP model with parameters $(\eta, \sigma, \zeta) = (20, 0.4, 2.4)$, with $n_{\text{train}} = 3000$. The estimated value for the parameter ξ in the non-exchangeable model is close to zero ($\hat{\xi} = 0.04$). This shows that our model is able to capture the linearity in the feature size, typical of the exchangeable datasets. Both models are able to recover the parameter σ correctly and the prediction errors are about the same (see Table 1 and Figure 4).

Amazon data. This real dataset contains time-ordered reviews of movies from Amazon[2]. We consider the first $n = 4000$ reviews and use $n_{\text{train}} = 1000$ reviews for training. In this context the binary matrix represents the presence of words in the reviews. Figure 4 shows that the presence of words in the reviews increases linearly with the number of reviews, which is expected to be observed in most of text data. As a consequence, the estimated parameter $\hat{\xi} = 0.01$ is very close to zero and the prediction errors of the two models are the same (see Table 1 and Figure 4).

ArXiv data. The ArXiv dataset[3] contains time-ordered articles uploaded on the Arxiv website. Each article is represented as a binary array that encodes its authors. Features size then represents the number of

[2]https://snap.stanford.edu

[3]https://www.kaggle.com/neelshah18/arxivdataset



(a) Synthetic - NE (b) Synthetic - IBP (c) Movies (d) ArXiv (e) Amazon

Figure 5: Log-log plots of the empirical proportions $(K_{n,r}/K_n)_{r\geq 1}$ (red dots) and 95% posterior predictive intervals (blue). Top: Indian Buffet Process (IBP); bottom: non-exchangeable model (NE).

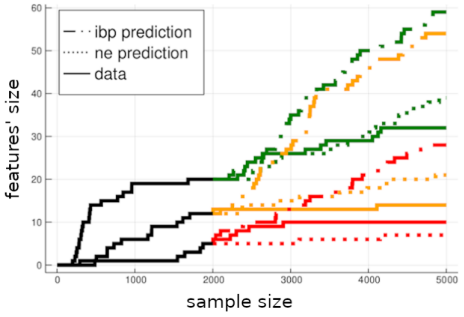


Figure 6: Movies dataset: sizes of some of the features (solid lines, training set in black), predicted features' sizes by the non-exchangeable model (dot lines), predictions by the IBP (dash-dot lines).

articles published by an author as a function of all the articles published on ArXiv. It is therefore expected a sublinear trend of the features' sizes, which is shown in Figure 4. We considered the first $n = 5000$ movies and used the first $n_{train} = 1000$ for training. The non-exchangeable model captures this asymptotic property ($\hat{\xi} = 0.96$) resulting in a better predictive performance compared to the IBP (see Table 1 and Figure 4).

Movies data. This dataset[4] contains time-ordered movies listed in Wikipedia and IMDB. Here the features are represented by the actors who have performed in the movie. The number of movies in which a specific actor has performed is obviously sublinear. We consider a subset of $n = 5000$ movies with $n_{\text{train}} = 2000$. The sublinear trend of the features' sizes is correctly modelled by our model ($\hat{\xi} = 0.45$) which outperforms the IBP in the prediction error (see Table 1 and Figure 4 and 6).

[4]http://udbms.cs.helsinki.fi/?datasets/film_dataset

Finally, Figure 5 shows the posterior predictive of the proportions of features shared by a given number of objects. Both models have the same asymptotic power-law property for this statistic and their fit is similar across all the datasets as expected.

7 Discussion

Various feature allocation models have been proposed in the literature that relax the exchangeability assumption. In particular dependent Indian buffet processes (Caron and Doucet, 2009; Williamson et al., 2010a; Zhou et al., 2011b; Ren et al., 2011; Miller et al., 2012; Gershman et al., 2014; Perrone et al., 2017) include models with covariates (time, space, graph, etc.), so that objects with similar covariates have similar feature weights; see (Foti and Williamson, 2013) for a review. Contrary to this line of work, the aim here is to derive feature allocation models with provable sublinear growth of the features' sizes, retaining the good asymptotic properties of the exchangeable models, namely the power-law behaviour and the control on the number of unique features. The proposed class of models allows to control these quantities by interpretable and tunable parameters, and inference is carried out by a simple Gibbs algorithm. Finally, the problem addressed in this paper is closely related to the problem of microclustering (Miller et al., 2015), which aims at finding models for random partitions where the size of the clusters grows sublinearly.

Acknowledgments. FC acknowledges support from EPSRC under grant EP/P026753/1 and from the Alan Turing Institute under EPSRC grant EP/N510129/1. GDB is funded by EPSRC under grant EP/L016710/1.

References

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

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

Cai, D., Campbell, T., and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems*, pages 4249–4257.

Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 2051–2059.

Caron, F. and Doucet, A. (2009). Bayesian nonparametric models on decomposable graphs. In *Advances in Neural Information Processing Systems*, pages 225–233.

Di Benedetto, G., Caron, F., and Teh, Y. W. (2017). Non-exchangeable random partition models for microclustering. *arXiv preprint arXiv:1711.07287*.

Foti, N. J. and Williamson, S. A. (2013). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):359–371.

Gershman, S. J., Frazier, P. I., and Blei, D. M. (2014). Distance dependent infinite latent feature models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):334–345.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

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

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396.

James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*.

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

Lee, J., Müller, P., Gulukota, K., and Ji, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics*, 9(2):621–639.

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

Miller, J., Betancourt, B., Zaidi, A., Wallach, H., and Steorts, R. (2015). Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *arXiv:1512.00792v1*.

Miller, K. T., Griffiths, T., and Jordan, M. I. (2012). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. *arXiv preprint arXiv:1206.3279*.

Palla, K., Knowles, D. A., and Ghahramani, Z. (2012). An infinite latent attribute model for network data. In *In Proceedings of the International Conference on Machine Learning (ICML)*. Citeseer.

Perrone, V., Jenkins, P. A., Spano, D., and Teh, Y. W. (2017). Poisson random fields for dynamic feature models. *The Journal of Machine Learning Research*, 18(1):4626–4670.

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

Teh, Y. W. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1838–1846. Curran Associates, Inc.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *AISTATS*.

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010a). Dependent Indian buffet processes. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 924–931.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010b). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 1151–1158.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514.

Zhou, M., Chen, H., Paisley, J., Ren, L., Li, L., Xing, Z., Dunson, D., Sapiro, G., and Carin, L. (2011a). Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21(1):130–144.

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

Supplementary material of Non-exchangeable feature allocation models with sublinear growth of the feature sizes

Giuseppe Di Benedetto
University of Oxford

François Caron
University of Oxford

Yee Whye Teh
University of Oxford
DeepMind

In this document we provide the proofs of the Propositions stated in the main paper, together with the posterior predictive distribution of the feature allocations.

A Posterior predictive distribution of feature allocations

The data augmentation described in Section 5 in the main paper allows us to compute the posterior predictive distribution of the $n+1$-th object given the first n ones.

$$Z_{n+1} \,|\, U_1, \ldots, U_n \overset{d}{=} Z^*_{n+1} + \sum_{k=1}^{K_n} \widetilde{z}_{n+1,k} \delta_{\widetilde{\theta}_k}$$

where Z^*_{n+1} is independent of the $\widetilde{z}_{n+1,k}$ which are distributed as follows

$$\widetilde{z}_{n+1,k} \,|\, U_1, \ldots, U_n \sim \text{Ber}\left(1 - \frac{\kappa\left(\widetilde{m}_{nj}\,,\, \Delta_{n+1} + \sum_{i=1}^n \Delta_i \widetilde{u}_{ij} 1_{\widetilde{\theta}_k \leq Y_i}\right)}{\kappa\left(\widetilde{m}_{nj}\,,\, \sum_{i=1}^n \Delta_i \widetilde{u}_{ij} 1_{\widetilde{\theta}_k \leq Y_i}\right)}\right).$$

By the marking theorem for Poisson point processes and Equation (20), $Z^*_{n+1} = \sum_{k=K_n+1}^{K_n+K^*_{n+1}} \delta_{\widetilde{\theta}_k}$ is a Poisson random measure on $\mathbb{R}_+$ with mean measure $\mu^*_{n+1}(d\theta) = \int_0^\infty \left(1 - e^{-\omega \Delta_{n+1} 1_{\theta \leq Y_{n+1}}}\right) d\nu^*(d\omega, d\theta)$, where we recall that

$$\begin{aligned}\nu^*(d\omega, d\theta) &= e^{-\omega \sum_{i=1}^n \Delta_i 1_{\theta \leq Y_i}} \rho(\omega) d\omega d\theta \\ &= \left(1_{\theta > Y_n} + \sum_{i=1}^n e^{-\omega(T_n - T_{i-1})} 1_{Y_{i-1} < \theta \leq Y_i}\right) \\ &\quad \times \rho(\omega) d\omega d\theta\end{aligned}$$

where $T_0 = Y_0 = 0$. Therefore, the number $K^*_{n+1} = Z^*_{n+1}(\mathbb{R}_+)$ of new features of the $n+1$ object is Poisson distributed with mean

$$\begin{aligned}&\mathbb{E}[Z^*_{n+1}(\mathbb{R}_+)] \\ &= \int_0^\infty \int_0^\infty \left(1 - e^{-\omega \Delta_{n+1} 1_{\theta \leq Y_{n+1}}}\right) d\nu^*(d\omega, d\theta)\end{aligned}$$

If follows that

$$\begin{aligned}&\mathbb{E}[Z^*_{n+1}(\mathbb{R}_+)] \\ &= \sum_{i=1}^{n+1} (Y_i - Y_{i-1}) \int_0^\infty \left(1 - e^{-\omega \Delta_{n+1}}\right) e^{-\omega(T_n - T_{i-1})} \rho(\omega) d\omega\end{aligned} \tag{A.1}$$

The locations $\widetilde{\theta}_{K_n+1}, \ldots, \widetilde{\theta}_{K_n+K^*_{n+1}}$ are sampled iid from the piecewise constant distribution on $[0, Y_{n+1}]$ with pdf proportional to

$$\sum_{i=1}^{n+1} 1_{Y_{i-1} < \theta \leq Y_i} \int_0^\infty \left(1 - e^{-\omega \Delta_{n+1}}\right) e^{-\omega(T_n - T_{i-1})} \rho(\omega) d\omega \tag{A.2}$$

If B is a GGP, the integral in Equations (A.1) and (A.2) is tractable and we have

$$\begin{aligned}&\int_0^\infty \left(1 - e^{-\omega \Delta_{n+1}}\right) e^{-\omega(T_n - T_{i-1})} \rho(\omega) d\omega \\ &= \begin{cases} \frac{\eta}{\sigma}\left[(T_{n+1} - T_{i-1} + \zeta)^\sigma - (T_n - T_{i-1} + \zeta)^\sigma\right] & \sigma > 0 \\ \eta \log\left(1 + \frac{\Delta_{n+1}}{T_n - T_{i-1} + \zeta}\right) & \sigma = 0 \end{cases}\end{aligned}$$

Proof. We have

$$\Pr(\widetilde{z}_{n+1,k} = 1 \,|\widetilde{\omega}_k, U_1, \ldots, U_n) = 1 - e^{-\widetilde{\omega}_k \Delta_{n+1}}$$

and

$$p(\widetilde{\omega}_k \mid U_1, \ldots, U_n) = \frac{\widetilde{\omega}_k^{\widetilde{m}_{n,k}} e^{-\widetilde{\omega}_k \sum_{i=1}^n \Delta_i \widetilde{u}_{ij} 1_{\widetilde{\theta}_k \leq Y_i}} \rho(\widetilde{\omega}_k)}{\kappa\left(\widetilde{m}_{n,k}, \sum_{i=1}^n \Delta_i \widetilde{u}_{ij} 1_{\widetilde{\theta}_k \leq Y_i}\right)}$$

Hence

$$\begin{aligned}&\Pr(\widetilde{z}_{n+1,k}=1 \mid U_1,\dots,U_n)\\&=1-\frac{\kappa\left(\widetilde{m}_{n,k},\Delta_{n+1}+\sum_{i=1}^n\Delta_i\widetilde{u}_{ij}1_{\widetilde{\theta}_k\le Y_i}\right)}{\kappa\left(\widetilde{m}_{n,k},\sum_{i=1}^n\Delta_i\widetilde{u}_{ij}1_{\widetilde{\theta}_k\le Y_i}\right)}\end{aligned}$$

□

The conditional distribution of the latent point process U_{n+1} can be written as follows:

$$\begin{aligned}&p(\widetilde{u}_{n+1,k}\,|\,\widetilde{z}_{n+1,k}=1,\,\widetilde{u}_{1:n,k})\\&\propto\kappa\left(\widetilde{m}_{n,k}+1,\Delta_{n+1}\widetilde{u}_{n+1,k}+\sum_{i=1}^n\Delta_i\widetilde{u}_{ik}\right)1_{u_{n+1,j}<1}\end{aligned}$$

Proof. We have

$$\begin{aligned}&p(\widetilde{u}_{n+1,k}\,|\,\widetilde{\omega}_k,\widetilde{z}_{n+1,k}=1,\,\widetilde{u}_{1:n,k})\\&=\frac{\Delta_{n+1}\widetilde{\omega}_k e^{-\widetilde{u}_{n+1,k}\Delta_{n+1}\widetilde{\omega}_k}1_{\widetilde{u}_{n+1,k}<1}}{1-e^{-\Delta_{n+1}\widetilde{\omega}_k}}\end{aligned}$$

and

$$\begin{aligned}&p(\widetilde{\omega}_k\mid\widetilde{z}_{n+1,k}=1,U_1,\dots,U_n)\\&\propto(1-e^{-\Delta_{n+1}\widetilde{\omega}_k})\widetilde{\omega}_k^{\widetilde{m}_{n,k}}e^{-\widetilde{\omega}_k\sum_{i=1}^n\Delta_i\widetilde{u}_{ij}1_{\widetilde{\theta}_k\le Y_i}}\rho(\widetilde{\omega}_k)\end{aligned}$$

□

B Proofs

B.1 Proof of Proposition 1

By the marking theorem for Poisson point processes, the set of points $\{(\omega_j)_{j\ge1}\mid z_{ij}=1\}$ is drawn from a Poisson point process with mean measure $Y_i(1-e^{-\Delta_i\omega})\rho(\omega)d\omega$. The total number of such points $Z_i(\mathbb{R}_+)$ is therefore Poisson distributed with mean $\mathbb{E}\left[Z_i(\mathbb{R}_+)\right]=Y_i\psi(\Delta_i)$. Using integration by part, we have

$$\psi(t)=t\int_0^\infty e^{-wt}\overline{\rho}(w)dw$$

where

$$\overline{\rho}(x)=\int_x^\infty\rho(w)dw. \tag{B.1}$$

Hence, by monotone convergence,

$$\lim_{t\to\infty}\frac{\psi(t)}{t}=\int_0^\infty\overline{\rho}(w)dw=\kappa(1,0)$$

and it follows that

$$Y_n\psi(\Delta_n)\overset{n\to\infty}{\sim}Y_n\Delta_n\kappa(1,0)$$

Finally note that $\Delta_n\overset{n\to\infty}{\sim}(1+\xi)^{-1}n^{-\xi/(\xi+1)}$, hence $Y_n\Delta_n\to(1+\xi)^{-1}$.

B.2 Proof of Proposition 2

The number of features observed in the first n objects can be written as

$$m_n:=\sum_{i=1}^n\sum_{j\ge1}z_{ij}=\sum_{i=1}^nZ_i(\mathbb{R}_+)$$

Since $\mathbb{E}[Z_n(\mathbb{R}_+)]=\mathbb{E}[m_n]-\mathbb{E}[m_{n-1}]$, it follows by Stolz-Cesàro theorem that

$$\mathbb{E}[m_n]\overset{n\to\infty}{\sim}(1+\xi)^{-1}\kappa(1,0)n.$$

In order to get the almost sure convergence of m_n to its expectation we can use the Kolmogorov strong law of large numbers which, under the assumption $\sum_{n\ge1}\mathrm{Var}\left(\frac{Z_n(\mathbb{R}_+)}{n}\right)<\infty$, gives

$$\frac{m_n-\mathbb{E}[m_n]}{n}\to0\text{ almost surely.}$$

Recall that $\mathrm{Var}(Z_n(\mathbb{R}_+))=\mathbb{E}[Z_n(\mathbb{R}_+)]$. Therefore the summability condition on the variance boils down to the convergence of the sum $\sum_{n\ge1}\frac{1}{n^2}Y_n\,\psi(\Delta_n)$, which holds true since the elements of the sum are of order n^{-2}.

B.3 Proof of Proposition 3

Since $\mathbb{E}[m_{n,j}|B]=\sum_{i=1}^n\left(1-e^{-\omega_j\Delta_i}\right)1_{\theta_j<Y_i}$, we have $\frac{\mathbb{E}[m_{nj}|B]-\mathbb{E}[m_{n-1\,j}|B]}{T_n-T_{n-1}}=\frac{1-e^{-\omega_j\Delta_n}}{\Delta_n}1_{\theta_j<Y_n}\to\omega_j$, then by Stolz-Cesàro theorem we have that

$$\mathbb{E}[m_{nj}|B]\overset{n\to\infty}{\sim}\omega_jT_n.$$

We have

$$\begin{aligned}\mathrm{Var}(m_{n,j}\mid B)&=\sum_{i=1}^n\left(1-e^{-\omega_j\Delta_i}\right)e^{-\omega_j\Delta_i}1_{\theta_j<Y_i}\\&\le\mathbb{E}[m_{n,j}|B].\end{aligned}$$

Using the sandwiching argument in Proposition 2 of Gnedin et al. (2007) it follows that, conditionally on B, $\frac{m_{nj}}{\mathbb{E}[m_{nj}\mid B]}\to1$ almost surely.

B.4 Proof of Proposition 4

Applying Campbell's theorem

$$\begin{aligned}\mathbb{E}[K_n]&=\mathbb{E}[\mathbb{E}[K_n\mid B]]\\&=\mathbb{E}\left[\sum_j\Pr(m_{n,j}>0\mid B)\right]\\&=\int_0^\infty\int_0^\infty\left(1-e^{-\omega f_n(\theta)}\right)\rho(\omega)d\omega d\theta\\&=\int_0^{Y_n}\psi(T_n-g(\theta))d\theta\end{aligned}$$

where

$$f_n(\theta) := \sum_{i=1}^{n} \Delta_i 1_{\theta \leq Y_i} = (T_n - g(\theta)) 1_{\theta \leq Y_n} \qquad \text{(B.2)}$$

with $g(\theta) := \sum_{i=1}^{\infty} \Delta_i 1_{\theta > Y_i}$ is a monotone increasing step function satisfying, for all $\theta \geq 0$

$$\max(0, g_1(\theta)) \leq g(\theta) \leq g_2(\theta) \qquad \text{(B.3)}$$

where $g_1(\theta) = (\theta^{(\xi+1)/\xi} - 1)^{1/(\xi+1)}$ and $g_2(\theta) = \theta^{1/\xi}$. Note that $g_1(\theta) \overset{\theta \to \infty}{\sim} g_2(\theta) \overset{\theta \to \infty}{\sim} \theta^{1/\xi}$. As ψ is an increasing function, it follows

$$\int_0^{Y_n} \psi(T_n - g_2(\theta)) d\theta \leq \mathbb{E}[K_n] \leq \int_1^{Y_n} \psi(T_n - g_1(\theta)) d\theta + \psi(T_n).$$

Using a change of variable, we obtain

$$\int_0^{Y_n} \psi(T_n - g_2(\theta)) d\theta = \xi \int_0^{T_n} \psi(T_n - \theta)\, \theta^{\xi - 1} d\theta$$

Finally, noting that

$$\psi(t) \overset{t \to \infty}{\sim} t^{\sigma} \ell(t)$$

where

$$\ell(t) = \begin{cases} \eta \log(t) & \sigma = 0 \\ \frac{\eta}{\sigma} & \sigma \in (0, 1) \end{cases}$$

and using (Di Benedetto et al., 2017, Lemma 14), we obtain

$$\int_0^{Y_n} \psi(T_n - g_2(\theta)) d\theta \overset{n \to \infty}{\sim} \frac{\Gamma(\xi+1)\Gamma(\sigma+1)}{\Gamma(\sigma+\xi+1)} n^{\frac{\xi+\sigma}{\xi+1}} \ell(n)$$

Similarly, we have

$$\int_1^{Y_n} \psi(T_n - g_1(\theta)) d\theta \overset{n \to \infty}{\sim} \frac{\Gamma(\xi+1)\Gamma(\sigma+1)}{\Gamma(\sigma+\xi+1)} n^{\frac{\xi+\sigma}{\xi+1}} \ell(n).$$

It follows by sandwiching that

$$\mathbb{E}[K_n] \overset{n \to \infty}{\sim} \frac{\Gamma(\xi+1)\Gamma(\sigma+1)}{\Gamma(\sigma+\xi+1)} n^{\frac{\xi+\sigma}{\xi+1}} \ell(n).$$

Using Campbell's theorem again,

$$\begin{aligned} \mathrm{Var}[K_n] &= \mathrm{Var}[\mathbb{E}[K_n \mid B]] + \mathbb{E}[\mathrm{Var}[K_n \mid B]] \\ &= \int \left(1 - e^{-\omega f_n(\theta)}\right) e^{-\omega f_n(\theta)} \rho(d\omega) d\omega d\theta \\ &\quad + \int \left(1 - e^{-\omega f_n(\theta)}\right)^2 \rho(d\omega) d\omega d\theta \\ &= \int \left(1 - e^{-\omega f_n(\theta)}\right) \rho(d\omega) d\omega d\theta \end{aligned}$$

therefore the almost sure asymptotic equivalence follows by Chebyshev inequality and the strong law of large numbers for K_n (see (Gnedin et al., 2007, Proposition 2)).

B.5 Proof of Proposition 5

We have the following inequality, for any $x \geq 0$

$$0 \leq x - (1 - e^{-x}) \leq \frac{x^2}{2}. \qquad \text{(B.4)}$$

Let us recall that

$$K_{n,r} = \sum_{j \geq 1} 1_{m_{n,j} = r}.$$

where $m_{n,j} = \sum_{i=1}^{n} z_{ij}$. Conditional on the CRM B we have

$$\mathbb{E}[K_{n,r} \mid B] = \sum_{j \geq 1} \Pr\left(m_{n,j} = r \,\middle|\, B\right).$$

Note that $\Pr\left(m_{n,j} = r \,\middle|\, B\right)$ only depends on (θ_j, ω_j). Write $S_{n,r}(\theta_j, \omega_j) = \Pr\left(m_{n,j} = r \,\middle|\, B\right)$. Let us denote $q_i(\theta_j, \omega_j) := \Pr(z_{ij} = 1 \mid B) = 1 - e^{-\Delta_i \omega_j 1_{\theta_j \leq Y_i}}$; $\lambda_n(\theta_j, \omega_j) := \sum_{i=1}^{n} q_i(\theta_j, \omega_j)$. Conditional on B the random variable $m_{n,j}$ has a Poisson-Binomial distribution with parameters $(q_1(\theta_j, \omega_j), \ldots, q_n(\theta_j, \omega_j))$. For each fixed (ω_j, θ_j) Le Cam's inequality Le Cam (1960) and inequality (B.4) give

$$\begin{aligned} &\sum_{r \geq 0} \left| \Pr(m_{n,j} = r \mid B) - \mathrm{Poisson}\left(r; \lambda_n(\theta_j, \omega_j)\right) \right| \\ &\leq 2 \sum_{i=1}^{n} q_i(\theta_j, \omega_j)^2 \leq 2\omega_j^2 \sum_{i=1}^{n} \Delta_i^2 1_{\theta_j \leq Y_i}. \end{aligned} \qquad \text{(B.5)}$$

where $\mathrm{Poisson}(r; \lambda)$ denote the probability mass function of a Poisson random variable with rate parameter λ evaluated at r. Note that for any $0 < \lambda_1 \leq \lambda_2$, using coupling inequalities (see. e.g. (Roch, 2015, Example 4.10 p. 154))

$$\sum_{r \geq 0} \left| \mathrm{Poisson}(r; \lambda_1) - \mathrm{Poisson}(r; \lambda_2) \right| \leq 2(\lambda_2 - \lambda_1).$$

Noting that $\lambda_n(\theta_j, \omega_j) \leq \omega_j f_n(\theta_j)$, where f_n is defined in Equation (B.2), and using inequality (B.4), we obtain

$$\begin{aligned} &\sum_{r \geq 0} \left| \mathrm{Poisson}(r; \lambda_n(\theta_j, \omega_j)) - \mathrm{Poisson}(r; \omega_j f_n(\theta_j)) \right| \\ &\leq 2 \sum_{i=1}^{n} \left(\omega_j \Delta_i - (1 - e^{-\omega_j \Delta_j})\right) 1_{\theta_j \leq Y_i} \\ &\leq \omega_j^2 \sum_{i=1}^{n} \Delta_i^2 1_{\theta_j \leq Y_i} \end{aligned}$$

Combining the above inequality with the inequality (B.5), we obtain the total variation bound

$$\sum_{r\geq 0}\left|\Pr(m_{n,j}=r\mid B)-\text{Poisson}\left(r;\omega_j f_n(\theta_j)\right)\right| \leq 3\omega_j^2\sum_{i=1}^{n}\Delta_i^2 1_{\theta_j\leq Y_i}. \tag{B.6}$$

Using Campbell's theorem,

$$\mathbb{E}\left[\sum_{j\geq 1}\omega_j^2\sum_{i=1}^{n}\Delta_i^2 1_{\theta_j\leq Y_i}\right]=\kappa(2,0)\sum_{i=1}^{n}Y_i\Delta_i^2 \overset{n\to\infty}{\asymp} n^{1/(1+\xi)}. \tag{B.7}$$

Using Campbell's theorem again,

$$\begin{aligned}&\mathbb{E}\left[\sum_{j\geq 1}\text{Poisson}\left(r;\omega_j f_n(\theta_j)\right)\right]\\ &=\frac{1}{r!}\int_0^\infty\int_0^\infty e^{-\omega f_n(\theta)}\omega^r f_n(\theta)^r\,\rho(\omega)\,d\omega d\theta\\ &=\frac{1}{r!}\int_0^\infty\kappa(r,f_n(\theta))f_n(\theta)^r d\theta\\ &=\frac{1}{r!}\int_0^{Y_n}\kappa(r,T_n-g(\theta))(T_n-g(\theta)))^r d\theta\end{aligned}$$

We use again the inequality (B.3) to bound the above expression. The upper bound is given by

$$\frac{1}{r!}\int_0^{Y_n}\kappa(r,T_n-g_2(\theta))(T_n-g_1(\theta))^r d\theta$$

Using a change of variable, we obtain

$$\begin{aligned}&\frac{\xi}{r!}\int_0^{T_n}\kappa(r,T_n-\theta)(T_n-g_1(\theta^\xi)))^r\theta^{\xi-1}d\theta\\ &=\frac{\xi}{r!}\int_0^{T_n}\kappa(r,\theta)(g_1(\theta^\xi))^r(T_n-\theta)^{\xi-1}d\theta.\end{aligned} \tag{B.8}$$

Noting that $\kappa(r,\theta)\overset{\theta\to\infty}{\sim}\eta\theta^{\sigma-r}\frac{\Gamma(r-\sigma)}{\Gamma(1-\sigma)}$ and $(g_1(\theta^\xi))^r\overset{\theta\to\infty}{\sim}\theta^r$, and using (Di Benedetto et al., 2017, Lemma 14), we obtain that (B.8) is asymptotically equivalent to

$$\eta\frac{\Gamma(\xi+1)\Gamma(r-\sigma)}{r!\Gamma(\sigma+\xi+1)}\,n^{\frac{\xi+\sigma}{\xi+1}}.$$

A similar asymptotic equivalence is obtained for the lower bound, and we conclude by sandwiching that

$$\mathbb{E}\left[\sum_{j\geq 1}\text{Poisson}\left(r;\omega_j f_n(\theta_j)\right)\right]\overset{n\to\infty}{\sim}\eta\frac{\Gamma(\xi+1)\Gamma(r-\sigma)}{r!\Gamma(\sigma+\xi+1)}\,n^{\frac{\xi+\sigma}{\xi+1}}$$

Combining the above asymptotic result with Equations (B.6) and (B.7), and assuming $\xi+\sigma>1$, we conclude

$$\mathbb{E}[K_{n,r}]=\mathbb{E}\left[\sum_{j\geq 1}\Pr\left(m_{n,j}=r\,\middle|\,B\right)\right]\overset{n\to\infty}{\sim}\eta\frac{\Gamma(\xi+1)\Gamma(r-\sigma)}{r!\Gamma(\sigma+\xi+1)}\,n^{\frac{\xi+\sigma}{\xi+1}}.$$

The variance of $K_{n,r}$ can be written as

$$\begin{aligned}\text{Var}[K_{n,r}]&=\text{Var}[\mathbb{E}[K_{n,r}\mid B]]+\mathbb{E}[\text{Var}[K_{n,r}\mid B]]\\ &=\int S_{n,r}(\theta,\omega)\rho(\omega)d\omega d\theta\\ &+\int S_{n,r}(\theta,\omega)(1-S_{n,r}(\theta,\omega))\,\rho(\omega)d\omega d\theta\\ &=\mathbb{E}[K_{n,r}].\end{aligned}$$

Using the result below Proposition 2 in Gnedin et al. (2007), we obtain, almost surely, $\mathbb{E}\left[\sum_{r\geq j}K_{n,r}\right]\overset{n\to\infty}{\sim}\sum_{r\geq j}K_{n,r}$. Using a proof similar to that of Corollary 21 in Gnedin et al. (2007), we obtain

$$K_{n,r}\overset{n\to\infty}{\sim}\eta\frac{\Gamma(\xi+1)\Gamma(r-\sigma)}{r!\Gamma(\sigma+\xi+1)}\,n^{\frac{\xi+\sigma}{\xi+1}}. \tag{B.9}$$

Combining Equation (B.9) with Equation (17) gives the final result.

References

Di Benedetto, G., Caron, F., and Teh, Y. W. (2017). Non-exchangeable random partition models for microclustering. *arXiv preprint arXiv:1711.07287*.

Gnedin, A., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv*, 4(146-171):88.

Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197.

Roch, S. (2015). Modern discrete probability: An essential toolkit. chapter 4. https://www.math.wisc.edu/ roch/mdp/roch-mdp-chap4.pdf.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Non-exchangeable feature allocation models with sublinear growth of the feature sizes
Publication Status	☐Published ☑ Accepted for Publication ☐Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style
Publication Details	Accepted for publication in the Proceedings of the International Conference on Artificial Intelligence and Statistics 2020

Student Confirmation

Student Name:	Giuseppe Di Benedetto		
Contribution to the Paper	I am the first author on this paper and I have contributed in designing the model,deriving the theoretical asymptotic results and proposing the point estimates for the model's parameters. About the applications, I have implemented the Gibbs sampler algorithm and I have run the experiments on both synthetic and real datasets to compare the proposed model against a well-known alternative model.		
Signature		Date	20/04/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor François Caron		
Supervisor comments 		
Signature 	Date	20/04/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Posterior contraction rates in Bayesian Functional Regression models

In this Chapter posterior contraction rates are derived for two Bayesian functional regression models: Functional Linear Regression and Functional Single-Index Model both with random predictor and scalar response.

4.1 Background

Functional Data Analysis (FDA) has gained increasing attention in the last decades, thanks to the rising of data being recorded over very fine grids, making it reasonable to model data as random functions. The models developed in this area have found applications in various fields such as econometrics ([Laurini, 2014]), neuroimaging ([Lila et al., 2016]), linguistics ([Gubian et al., 2015]), meteorology ([Jamaludin and Yusop, 2016]), just to mention a few. We refer to [Ramsay, 2005] and [Ferraty and Vieu, 2006] for a detailed review of Functional Data Analysis. Among the numerous models belonging to FDA, probably the simplest and most popular one is the functional linear regression (FLR) with functional predictor and scalar response. This model have been widely studied in the frequentist literature and asymptotic properties have been de-

rived for different estimators ([Cardot et al., 1999], [Chagny and Roche, 2016], [Cai and Hall, 2006], [James et al., 2009]). Asymptotic results of the Bayesian approaches, on the other hand, are rather scarse, and up to our knowledge, [Choi et al., 2011] is the only work where the posterior contraction rates with respect to the prediction error are provided for the functional slope parameter, endowed with a Gaussian process prior. In this Chapter we will study FLR as an inverse problem, and show that different priors can be used to achieve minimax posterior contraction rates with respect to the prediction loss and the norm of the Hilbert space where the functional predictor and parameter live, which are the loss functions in the direct and inverse problem respectively.

However, the assumption of linear dependence is too strong to model some phenomena, and a nonlinear model is necessary. It is known that nonlinear regression models suffer from the curse of dimensionality: in the finite dimensional case, the minimax rate of convergence is $n^{-\frac{\alpha}{2\alpha+d}}$ where α and d are the regularity of the regression function and the dimensionality of the predictor respectively. In the functional setting, the convergence deteriorates to logarithmic rates. Single-index models are used to tackle this issue, providing interpretability and flexibility, while retaining good asymptotic properties. Several works have studied the functional single-index model with a functional link and a finite dimensional index paramenter ([Roy et al., 2017], [Jiang et al., 2011], [Alquier and Biau, 2013], [Dhara et al., 2018], [Ganti et al., 2017], [Antoniadis et al., 2004]), and frequentist estimators has been proposed when the index parameter is infinite dimensional ([Chen et al., 2011], [Ma, 2016]). In this Chapter, where we will focus on the functional single-index model with functional predictor and scalar response, we derive posterior contraction rates and give an example of prior that can achieve such rates.

This Chapter is organized in two main sections. In section 4.2, posterior contraction rates for Bayesian FLR are derived with respect to the prediction error and the Hilbert norm. Such rates are achieved by the sieve priors and the Gaussian process priors, and the proofs rely on the construction of tests with vanishing errors, following the techniques introduced in [Ghosal et al., 2007]. In section 4.3 posterior contraction rates are derived for the Bayesian F-SIM

with respect to the empirical prediction loss. In that section we show that the sieve prior used for FLR can be adopted for the index function, while a hybrid location-scale mixture of normals prior can be chosen for the link function. We prove that these priors induce a prior on the nonlinear function which achieves contration rate which is the maximum between the rate for FLR and the rate for nonlinear regression.

4.2 Functional Linear Regression model

In functional linear regression models the scalar response $Y \in \mathbb{R}$ is modelled as

$$Y = \langle \beta, X \rangle + \epsilon$$

with functional covariate X and slope parameter β living in a Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ endowed with its norm which will be denoted by $\|\cdot\|$. The responses are observed up to an error $\epsilon \sim \mathcal{N}(0, \sigma^2)$ which is assumed to be centred and uncorrelated with the covariate. In Bayesian FLR a prior is assigned to the unknown slope parameter, and the goal here is to investigate about the convergence of the posterior distribution to the true parameter β_0. The rates will be described for two different norms: the norm of the Hilbert space, and a norm which depends on the covariate distribution. Deriving the rates for these two norms is equivalent, when looking at the regression problem as an inverse problem, to solve both the direct and the inverse problem. The direct problem will be solved using the standard Theorems developed in the seminal papers [Ghosal et al., 2007], which requires the design of a sequence of test functions with vanishing errors. We will provide such tests and show that both sieve priors and a Gaussian process priors can achieve the minimax rates of convergence in the direct problem. The inverse problem is instead tackled using the same approach of [Knapik and Salomond, 2014].
In what follows we will denote the Sobolev norm by $\|\cdot\|_\alpha$, defined as $\|\beta\|_\alpha^2 = \sum_j j^{2\alpha}\beta_j^2$ where $\alpha \in \mathbb{R}$ and $(\beta_j)_j$ is the sequence of coefficients of β with respect to some orthonormal (ON) basis of $\mathbb{H}$. $\mathbb{H}_\alpha = \{\beta : \|\beta\|_\alpha^2 < \infty\}$ will denote the Sobolev space of regularity $\alpha \in \mathbb{R}$, and $B(\beta, r, \|\cdot\|_\alpha)$ a ball of radius r with

respect to some norm $\|\cdot\|_\alpha$ centered in β, while C^α denotes the space of Hölder functions with exponent $\alpha > 0$. Henceforth a sequence of n iid observations will be denoted by $(X^n, Y^n) = \{(X_i, Y_i)\}_{i=1}^n$ consisting of pairs of predictors X_i and responses Y_i. The prior on the parameter of interest and its posterior will be denoted by $\Pi(\cdot)$ and $\Pi(\cdot \mid X^n, Y^n)$ respectively. $\mathbb{P}_0^{(n)}$ will denote the distribution of n observations under the true value of the parameter, and $\mathbb{E}_0^{(n)}$ its expectation. In what follows $\ell_n(\beta)$ will denote the log-likelihoood of n observations given β. $KL(f, g) = \int \log(f/g) f d\mu$ denotes the Kullback-Leibler divergence between two densities f and g, and $V(f, g) = \int |\log(f/g)|^2 f d\mu$. With abuse of notation, $KL_n(\beta_0, \beta)$ and $V_n(\beta_0, \beta)$ will denote the expectation and variance of the log-likelihood ratio of n observations with respect to β_0 and β respectively. Inequalities up to a constant will be denoted by $\lesssim$ and $\gtrsim$, the notation $a_n \asymp b_n$ is used for sequences a_n and b_n satisfying $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

4.2.1 Assumptions on the functional covariate

The functional predictor X is assumed to be a centered stochastic process with trajectories in $\mathbb{H}$ with finite second moment $\mathbb{E}\|X\|^2 < \infty$. This assumption ensures the existence of the covariance operator $\Gamma : \mathbb{H} \to \mathbb{H}$ defined as $\Gamma f = \mathbb{E}[\langle X, f\rangle X]$, through which the functional linear regression problem can be viewed as an inverse problem. As pointed out in [Cardot and Johannes, 2010], by multiplying the linear equation by X and taking the expectation, we get the functional equation $\mathbb{E}[YX] = \Gamma\beta$. Therefore the problem consists in inverting the covariance operator Γ. Another link to inverse problem is described in [Meister, 2011], where it is shown the asymptotic equivalence of FLR and a white noise inverse problem involving $\Gamma^{1/2}$. However, the covariance operator is a trace-class operator and therefore it does not admit a continuous inverse in $\mathbb{H}$, hence the inverse problem is said to be ill-posed. The properties of the covariance operator imply a spectral decomposition of Γ: there exist a nonnegative and summable sequence of eigenvalues $(\lambda_j)_j$ and a sequence of eigenfunctions $(\varphi_j)_j$ that form a ON basis of $\mathbb{H}$. Therefore the stochastic process X admits a series representation, called Karhuen-

Loéve expansion, which is a functional principal components decomposition: $X = \sum_j \sqrt{\lambda_j}\xi_j\varphi_j$ where the $(\xi_j)_j$ is a sequence of centered and uncorrelated random variables such that $\mathbb{E}\,\xi_j^2 = 1$. The operator Γ induces norm: $\|\beta\|_\Gamma^2 = \langle\Gamma\beta,\beta\rangle^2 = \left\|\Gamma^{1/2}\beta\right\|^2 = \sum_j \lambda_j\beta_j^2$, weaker than the Hilbert norm, making the recovery of the true parameter β_0 with respect to the Hilbert norm more difficult. The metric induced by this norm is also called prediction loss and can be written as

$$\|\beta - \beta_0\|_\Gamma^2 = \mathbb{E}[\langle X, \beta - \beta_0\rangle^2] = \sum_{j\geq 1} \lambda_j(\beta_j - \beta_{0j})^2.$$

The degree of ill-posedness of the inverse problem depends on how fast the eigenvalues converge to zero, and two regimes are considered in the literature: the *mildly ill-posed* inverse problem refers to the polynomial decrease $\lambda_j \asymp j^{-2\eta}$, while *severely ill-posed* refers to the exponential descrease $\lambda_j \asymp e^{-j^p}$. In this Chapter we focus on the mildly ill-posed case.
The assumptions on the functional covariate are summarized as follows:

- $\mathbb{E}[X] = 0$, $\mathbb{E}[\|X\|^2] < \infty$, and there exists a positive constant C such that
$$\mathbb{P}\left(\sum_{i=1}^n \|X_i\|^2 \leq C\,n\right) \overset{n}{\to} 1$$
.
- The covariance operator Γ has eigenvalues $\lambda_j = j^{-2\eta}$, with $\eta > 1/2$.
- There exist constants $C_1 > 0$ and $t > 0$ such that for every $\beta \in \{\beta : \|\beta - \beta_0\| \leq t\}$
$$\mathbb{E}[\langle X, \beta - \beta_0\rangle^4] \leq C_1\,\mathbb{E}[\langle X, \beta - \beta_0\rangle^2]$$

The first and third assumption are useful to prove the convergence rates in the direct problem, in particular to show that the sequence of likelihood ratio tests have exponentially decreasing errors.

4.2.2 Prior models and results

Contraction rates results are presented for two different priors: the sieve prior and the Gaussian process prior. Both can be represented as a series with respect to a ON basis of $\mathbb{H}$:

$$\beta \stackrel{d}{=} \sum_{j=1}^{K} \beta_j \varphi_j. \tag{4.1}$$

with the number of components $K \in \mathbb{N} \cup \{\infty\}$. For the sieve prior K is finite but random, while for the Gaussian process prior $K = \infty$ almost surely. Assumptions on the distribution of the coefficients β_j needs to be specified in order for the prior to satisfy the conditions of the Theorem used to solve the direct problem.

Sieve Prior Sieve priors represent a very well known class of non-parametric priors in the Bayesian literature (see for example [Arbel et al., 2013] and [Rousseau and Szabo, 2016]). Such priors are defined over finite dimensional spaces, with the number of dimension being itself random. In the functional setting, the distribution of the unknown slope parameter can be written in a hierarchical fashion:

$$\begin{aligned} \beta &\stackrel{d}{=} \sum_{j=1}^{K} \beta_j \varphi_j \\ K &\sim \pi \\ (\beta_1, \ldots, \beta_K) \,|\, K &\sim \Pi_K \end{aligned} \tag{4.2}$$

where π is a random distribution on $\mathbb{N}$ and Π_k is a distribution on $\mathbb{R}^k$ satisfying some conditions on the tails which are described in [Arbel et al., 2013] and recalled here:

- There exist two positive constants b_1 and b_2 such that $e^{-b_1 k L(k)} \leq \pi(k) \leq e^{-b_2 k L(k)}$ for all $k \in \mathbb{N}$ with L slowly varying function.
- As in [Arbel et al., 2013], we assume that for all $k \geq 1$ the distribution

Π_k is absolute continuous with respect to the Lebesgue measure with $\Pi_k(\theta) = \prod_{j=1}^{k} \frac{1}{\tau_k} g\left(\frac{\theta_j}{\tau_j}\right)$ with scale parameters such that $\overline{\tau} \leq \tau_j \leq \underline{\tau}$ for all $j \geq 1$ and for $\overline{\tau}$ and $\underline{\tau}$ positive constants. The density function g is such that $g(x) \leq G_1 e^{-G_2|x|^\gamma}$ for all $x \in \mathbb{R}$ with G_1, G_2, γ positive constants, and we assume there exists a compact subset $X_0 \subset \mathbb{R}$ such that $g(x) \geq c$ for all $x \in X_0$ and $c > 0$.

Gaussian Process prior An alternative and widely used prior for functions is the Gaussian process:

$$\beta \sim \mathrm{GP}(0, \Lambda) \tag{4.3}$$

characterized by the covariance function $\Lambda(s,t) = \sum_j j^{-2\gamma-1} \varphi_j(s)\varphi_j(t)$, with $(\varphi_j)_j$ being the sequence of eigenfunctions of the covariance operatore Γ of X. The parameter $\gamma > 0$ tunes the smoothness of the prior: the Sobolev space $\mathbb{H}_{\gamma+1/2}$ is the reproducing kernel Hilbert space associated to Λ. The prior on β can be written in a series representation as in 4.1, with an infinite number of components

$$\beta \overset{d}{=} \sum_{j \geq 1} \beta_j \varphi_j$$

where $\beta_j = j^{-\gamma-1/2} Z_j$ and $Z_j \overset{iid}{\sim} \mathcal{N}(0,1)$.

Posterior contraction rates

The direct problem is solved using the Theorem by [Ghosal et al., 2007], recalled in section 4.4. In order to solve the inverse problem we use the Theorem 2.1 from [Knapik and Salomond, 2014], which requires the design of a sieve where the norms corresponding to the direct and inverse problems are equivalent. The proofs of the following posterior contraction results are given in Section 4.4.

Sieve prior

Theorem 4. *Let $\beta_0 \in \mathbb{H}_{\alpha_1}$ and define the sequences $\epsilon_n = n^{-\frac{\alpha_1+\eta}{2(\alpha_1+\eta)+1}}(\log n)^{a_1}$, $u_n \asymp n^{-\frac{\alpha_1}{2(\alpha_1+\eta)+1}}(\log n)^{a_2}$ and $j_n = \frac{n\epsilon_n^2}{\log n}$, with a_1 and a_2 non-negative constants. Let Π be a sieve prior on β described in 4.2 such that*

$$\sum_{j=1}^{\lfloor j_n \rfloor} \frac{|\theta_{0j}|^\gamma}{\tau_j^\gamma} \lesssim j_n \log n$$

then the posterior distribution converges to β_0 with respect to the prediction error and the Hilbert norm with rates ϵ_n and u_n respectively:

$$\mathbb{E}_0^{(n)}\Pi\left(\beta : ||\beta - \beta_0||_\Gamma > \epsilon_n | X^n, Y^n\right) \xrightarrow{n} 0$$
$$\mathbb{E}_0^{(n)}\Pi\left(\beta : ||\beta - \beta_0|| > u_n | X^n, Y^n\right) \xrightarrow{n} 0.$$

The conditions are satisfied for example by taking Gaussian or Laplace densities with constant scaling parameters $\tau_j = 1$. While for the distribution on the number of components a Poisson or a Geometric distribution can be chosen. The condition required in the Theorem is used to prove that the prior assigns enough mass to a Kullback-Leibler neighbourhood around the true value β_0.

Gaussian process prior In order to prove contraction rates in the direct problem, we apply the Theorem in [van der Vaart and van Zanten, 2008], which provides conditions for the Gaussian process prior to satisfy the Kullback-Leibler condition required in the standard Theorem in [Ghosal et al., 2007].

Theorem 5. *Let $\beta_0 \in \mathbb{H}_{\alpha_1}$ and define the sequences $\epsilon_n = n^{-\frac{\alpha_1 \wedge \gamma + \eta}{2(\alpha_1+\eta)+1}}(\log n)^{a_1}$ and $u_n \asymp n^{-\frac{\alpha_1 \wedge \gamma}{2(\alpha_1+\eta)+1}}(\log n)^{a_1}$ with a_1 and a_2 non-negative constants. Let Π be a Gaussian process prior on β described in 4.3 with the regularity parameter $\gamma > \max(0, \alpha_1 - 1/2)$. Then the posterior distribution converges to β_0 with respect to the prediction error and the Hilbert norm with rates ϵ_n and u_n respectively:*

$$\mathbb{E}_0^{(n)}\Pi\left(\beta : ||\beta - \beta_0||_\Gamma > \epsilon_n | X^n, Y^n\right) \xrightarrow{n} 0$$
$$\mathbb{E}_0^{(n)}\Pi\left(\beta : ||\beta - \beta_0|| > u_n | X^n, Y^n\right) \xrightarrow{n} 0.$$

Adaptive methods for Gaussian process prior exist in the literature ([Szabó et al., 2013], [Knapik et al., 2016]), in particular the work of [Knapik et al., 2016] describes two strategies for inverse problems: an empirical Bayes approach and a full hierarchical Bayes one. In the latter, a prior is assigned to the unknown regularity parameter γ, which needs to satisfy mild assumptions (see Assumption 1 in [Knapik et al., 2016]) met for instance by gamma, inverse gamma, and exponential distributions. The hierarchical prior leads to the rates $n^{-\frac{\alpha_1+\eta}{2(\alpha_1+\eta)+1}}$ and $n^{-\frac{\alpha_1}{2(\alpha_1+\eta)+1}}$ for direct and inverse problems respectively, up to logarithmic terms. As expected, the rate for the inverse problem is slower due to compactness of the covariance operator. As a matter of fact, the equivalence of the norms in the sieve involves an increasing term, which represent the deterioration factor of the contraction rate of the direct problem. See Section 4.4 for details. It is worth noting that, up a logarithmic term, the posterior distribution achieves minimax rates for both the direct and inverse problem in the mildly ill-posed case (see for example [Cardot and Johannes, 2010]).

4.3 Functional Single-index Model

The linear assumption can be too strong for many application, hence it is desirable to have more flexibility in modeling the relation between the covariate and the response. However, the very general non-parametric functional regression problem $Y = f(X)$ suffers of the curse of dimensionality, leading to logarithmic rates of convergence for estimators of f (see [Chagny and Roche, 2016]). Therefore one would need to constraint the nonlinearity in order to have a good flexibility in modeling real-world phenomena and retaining good asymptotic properties. This trade-off can be achieved by the single-index model

$$Y = f(\langle X, \beta \rangle) + \epsilon$$

where β is an element in a separable Hilber space $(\mathbb{H}, \langle \cdot, \cdot \rangle)$, $f : \mathbb{R} \to \mathbb{R}$ is a real function and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X which is assumed to be a centered random variable with values in the Hilbert space. For the sake of simplicity, the error variance is assumed to be known, however, as

mentioned in [Arbel et al., 2013], the same results would hold for σ unknown, by assigning an inverse Gamma prior to the unknown σ and studying the three different regimes $\sigma/\sigma_0 < 1/2$, $\sigma/\sigma_0 > 3/2$ and $\sigma/\sigma_0 \in [1/2, 3/2]$. It is worth noting that the domain of the link function f cannot be restricted to a compact. This assumption does not hold in this case, since we cannot bound the product $\langle X, \beta \rangle$. In this section we will show how the sieve prior used to achive minimax rate of convergence for both the direct and inverse problem in functional linear regression can be used to build a prior in nonlinear functional regression models.

4.3.1 Assumptions

In the functional single-index model, the parameter space is

$$\Theta = \{g : \mathbb{H} \to \mathbb{R},\ \exists \beta \in \mathbb{H}, f : \mathbb{R} \to \mathbb{R} \text{ such that } \forall X \in \mathbb{H},\ g(X) = f(\langle \beta, X \rangle)\}$$

and the objective is to recover the generating function $g_0(\cdot) = f_0(\langle \beta_0, \cdot \rangle)$. Analogously to the finite dimensional case, in order to ensure identifiability, the link function has to be non-constant and the index function β_0 is required to satisfy $\|\beta_0\| = 1$ (see [Geenens and Delecroix, 2006]). The prior Π_g on the function $g : \mathbb{H} \to \mathbb{R}$ can be induced by setting a prior Π_β on β and a prior Π_f on f. The recovery of the function g_0 is studied with respect to the empirical prediction error defined as

$$d_n(g, g_0) = \left(\frac{1}{n} \sum_{i=1}^{n} (g(X_i) - g_0(X_i))^2 \right)^{1/2}$$

and the posterior contraction rates will depend on the rates to recover β_0 in the functional linear regression derived in the previous sections, and the rates to recover the link function f_0 in the one-dimensional nonlinear regression. The assumptions we make to prove the posterior convergence result are the following:

- **Covariate**:

- $\mathbb{E}[X] = 0$ and $\mathbb{E}[\|X\|^2] < \infty$
- Defining $\xi_j := \frac{\langle X, \varphi_j \rangle}{\lambda_j}$ and $\lambda_j = \mathrm{Var}(\langle X, \varphi_j \rangle)$ for all $j \geq 1$, there exists $p \geq 2$ such that

$$\sup_{j \geq 1} \mathbb{E}[|\xi_j|^p] \leq C_2$$

- The eigenvalues of the covariance operator Γ of X decrease polynomially: $\lambda_j = j^{-2\eta}$ with $\eta > 1/2$

- **Regularity**: We assume that the true link function to be $f_0 \in C^{\alpha_2} \cap L^1_{\langle \beta_0, X \rangle}$, where $L^1_{\langle \beta_0, X \rangle}$ is the space of integrable functions with respect to the distribution of $\langle \beta_0, X \rangle$, and the index function $\beta_0 \in \mathbb{H}_{\alpha_1}$.

- **Prior**: The prior Π on g, induced by the priors on the index and link functions, satisfies the following conditions

  (P1) $\Pi(\|\beta - \beta_0\|_\Gamma < \epsilon_{\beta,n}) \geq e^{-cn\epsilon^2_{\beta,n}}$ for a sequence $\epsilon_{\beta,n} \xrightarrow{n} 0$

  (P2) Let $\epsilon_{f,n} \xrightarrow{n} 0$ and J_n such that $\epsilon_{\beta,n} = CJ_n^2\, 2^{-\alpha_2 J_n}$ with $C > 0$ constant, there exists $\mathcal{M}_n \subset C^{\alpha_2} \cap L^1_{\langle \beta_0, X \rangle}$ such that $\Pi(\mathcal{M}_n) > e^{-cn\epsilon^2_{f,n}}$, and for all $f \in \mathcal{M}_n$:

$$|f(x) - f_0(x)| \lesssim J_n \qquad \forall x \in \mathbb{R}$$
$$|f(x) - f_0(x)| \lesssim J_n^2 2^{-\alpha_2 j} \qquad \forall j = 0 \ldots, J_n \quad \forall x \in A_j$$

  with $A_j = [-2^{2\alpha_2(J_n - j)/p},\, 2^{2\alpha_2(J_n - j)/p}]$.

  (P3) There exists a sieve $\mathcal{F}_n \in \Theta$ such that $\log N(\xi\epsilon_n, \mathcal{F}_n, d_n) \lesssim n\epsilon_n^2$ for $\xi \in (0,1)$ and $\Pi(\mathcal{F}_n^c) \leq e^{-(c+2)n\epsilon_n^2}$, with $\epsilon_n = \max\left(\epsilon_{\beta,n}^{\min(\alpha_2,1)},\, \epsilon_{f,n}\right)$

The assumptions on the prior ensure that the contraction rate result can be proven using the standard Theorem in [Ghosal et al., 2007]: the first prior assumption requires Π_β to assign enough mass to a Kullback-Leibler neighbourhood which is equivalent to a ball in the norm induced by Γ. The second prior assumption requires Π_f to assign enough mass to a set of functions that can approximate f_0 both uniformly and on a sequence of balls A_j, this is

needed since the domain of the function f is not compact. The third prior assumption requires the existence of a sieve with enough prior mass and with bounded complexity measured with the logarithm of the covering number, this assumption is required to build a sequence of test with exponentially decreasing errors which are needed in the proof of the Theorem. The proof of the following Theorem can be found in Section4.4.2.

Theorem 6 (Contration rates in functional single-index models)**.** *Under the assumptions listed above, the posterior distribution of g converges to g_0 with respect to the d_n distance:*

$$\mathbb{E}_0[\Pi(d_n(g, g_0) \gtrsim \epsilon_n \,|\, Y^n, X^n)] \xrightarrow{n} 0$$

at rate

$$\epsilon_n = \max\left(\epsilon_{\beta,n}^{\min(\alpha_2,1)},\, \epsilon_{f,n}\right)$$

with $\epsilon_{\beta,n}$ and $\epsilon_{\beta,n}$ being the rates of the direct problem in the functional linear regression and the nonlinear regression problem respectively.

The minimax rate for the direct problem in functional linear regression can be achieved by the priors described in the previous sections. Regarding the nonlinear regression, minimax rates are still unknown in the setting of non compact domain, and up to our knowledge, the best rate in the case is achieved by the hybrid location-scale mixture prior introduced in [Naulet and Rousseau, 2017]. In the next Section we provide an example of prior that satisfies the assumptions required by the Theorem above.

4.3.2 Example of prior for the Functional Single-Index Model

Let the prior on the functional index β be the sieve prior described in Section 4.2.2 and satisfying the assumptions of Theorem 4. A class of priors introduced in [Naulet and Rousseau, 2017] can be used for the link function in single-index models. We will show that such priors satisfy the assumptions required for Theorem 6 to hold. Let us recall the hybrid location-scale mixture

of normals prior:

$$f(x) = \int_{\mathbb{R}} e^{-\frac{(x-\mu)^2}{\omega}} dM(\omega, \mu)$$

where $M \sim \Pi_A$ is a random signed measure distributed as a symmetric Gamma process (see for instance [Wolpert et al., 2011], [Naulet and Barat, 2018], and [Naulet and Barat, 2018]) with parameter A being a random measure distributed as follows

$$\begin{aligned} P_\omega &\sim \Pi_\omega \\ A &= \overline{A} P_\omega \otimes G_\mu, \quad \overline{A} > 0 \end{aligned}$$

where Π_ω is a distribution on the space of probability measures on $\mathbb{R}_+$ such that

$$\begin{aligned} &\Pi_\omega \left(P_\omega : \ P_\omega(\omega > x) \geq \exp(-a_1 x^{b_1}/2) \right) \leq C_1 \exp(-a_2 x^{b_1}) \\ &\Pi_\omega \left(P_\omega : \ P_\omega(\omega < 1/x) \geq \exp(-a_3 x^{b_2}/2) \right) \leq C_2 \exp(-a_4 x^{b_2}) \end{aligned}$$

for some positive constants $c_1, c_2, a_1, a_2, a_3, a_4, b_1, b_2$, and for all $r \geq 1$ there exist positive constants a_5, b_3, c_3 such that for J large enough

$$\Pi_\omega \left(\bigcap_{j=0}^{J} \{ P_\omega : \ P_\omega([2^{-j}, 2^{-j}(1 + 2^{-Jr})]) \geq 2^{-J} \} \right) \geq c_3 \exp(-a_5 J^{b_3} 2^J).$$

while G_μ is a probability measure on $\mathbb{R}$ satisfying, for some positive constants c_4, b_4, b_5,

$$G_\mu(|\mu - x| \leq t) \geq c_4 t^{b_4} (1 + |x|)^{-b_5} \qquad \forall t \in (0, 1).$$

As shown in [Naulet and Rousseau, 2017], the Dirichlet process, under some conditions on its base measure, can satisfy the conditions required for Π_ω (see Proposition 1 in [Naulet and Rousseau, 2017]), and it is worth noting that if Π_ω is a Dirichlet process, the hybrid location-scale mixture prior assigns positive probability to mixtures with two or more components sharing the

same scale parameter. Under these assumptions on the prior for the link function, we have the following result, whose proof is given in Section 4.4.

Theorem 7. *Let $f_0 \in C^{\alpha_2} \cap L^1_{\langle \beta_0, X \rangle}$ and $\beta_0 \in \mathbb{H}_{\alpha_1}$. Let Π_f be the Hybrid location-scale mixture of normals prior for the link function described above, and Π_β the sieve prior satisfying the assumptions of Theorem 4. Then the posterior distribution converges to g_0 with respect to the empirical prediction loss at rate $\epsilon_n = \max\left(\epsilon_{\beta,n}^{\min(\alpha_2,1)}, \epsilon_{f,n}\right)$:*

$$\mathbb{E}_0^{(n)}[\Pi(d_n(g, g_0) > \epsilon_n | X^n, Y^n)] \xrightarrow{n} 0$$

where

$$\epsilon_{\beta,n} = n^{-\frac{\alpha_1+\eta}{2(\alpha_1+\eta)+1}}, \qquad \epsilon_{f,n} = \begin{cases} n^{-\frac{\alpha_2}{2\alpha_2+\max(1,\min(\alpha_2+1,2\alpha_2/p))}} & \text{if } 0 < p \leq 2\alpha_2 \\ n^{-\frac{\alpha_2}{2\alpha_2+1}} & \text{if } p > 2\alpha_2 \end{cases}$$

up to logarithmic terms.

4.4 Proofs

Here we recall the conditions required by the standard Theorem of [Ghosal et al., 2007] to obtain posterior contraction rates. $X^n = (X_1, \ldots, X_n)$ denotes a vector dataset of n observations whose distribution is described by the parameter $\theta \in \Theta$, which in this Chapter is the functional slope β in FLR, and the non-linear function $g(\cdot) = f(\langle \beta, \cdot \rangle)$ in F-SIM. Let us denote the loss function by d, and the true value of the parameter of interest by θ_0. Let $(\epsilon_n)_n$ be a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Let us assume that the following conditions hold true:

- Kullback-Leibler condition:

$$\Pi\left(B_n^{KL}(\theta_0, \epsilon_n)\right) \geq e^{-cn\epsilon_n^2}$$

with $B_n^{KL}(\theta_0, \epsilon_n) := \{\theta \in \Theta : KL_n(\theta_0, \theta) \leq n\epsilon_n^2, V_n(\theta_0, \theta) \leq n\epsilon_n^2\}$ and $c > 0$ a constant.

- Sieve condition: there exists $\mathcal{F}_n \subset \Theta$ such that

$$\Pi(\mathcal{F}_n^c \leq e^{-(c+4)n\epsilon_n^2})$$

and the logarithm of the covering number of $\mathcal{F}_n$ with balls of radius $\xi\epsilon_n$ with respect to d, with $\xi \in (0,1)$ constant, satisfies the following:

$$\log N(\xi\epsilon_n, \mathcal{F}_n, d) \leq n\epsilon_n^2.$$

- Tests condition: for $\xi \in (0,1)$ constant and for each $\theta_1 \in \mathcal{F}_n$ with $d(\theta_0, \theta_1) \geq \epsilon_n$, there exist a sequence of test functions $(\phi_n)_n$ such that

$$\mathbb{E}_0^{(n)}\phi_n \leq e^{-Cn\epsilon_n^2}, \qquad \sup_{\theta \in \mathcal{F}_n : d(\theta,\theta_1) < \xi\epsilon_n} \mathbb{E}_\theta^{(n)}[1-\phi_n] \leq e^{-Cn\epsilon_n^2}$$

with $C > 0$ constant.

Then the posterior distribution converges to θ_0 at rate ϵ_n:

$$\mathbb{E}_{\theta_0}^{(n)}\Pi(d(\theta,\theta_0) \gtrsim \epsilon_n \,|\, X^n) \overset{n}{\to} 0$$

For the test condition, standard likelihood ratio tests can be used for both the FLR and the F-SIM models. They can be defined for the general nonlinear regression problem with functional predictor and scalar response $Y = g(X)+\epsilon$, where $\epsilon \sim \mathcal{N}(0,\sigma^2)$ uncorrelated with the predictor, and with respect to the empirical prediction loss, as stated in the following result.

Lemma 8. *In the nonlinear regression problem $Y = g(X) + \epsilon$, for $\xi \in (0,1)$ and every g_1 such that $d_n(g_0, g_1) \geq r\epsilon_n$, there exist a sequence of test functions $(\phi_n)_n$, defined as $\phi_n(g) = 1(\ell_n(g) - \ell_n(g_0) > \delta_0 n d_n(g, g_0)^2)$ with $\delta_0 > 0$ constant, such that for $C > 0$ constant*

$$\mathbb{E}_0^{(n)}[\phi_n | X^n] \leq e^{-Cn\epsilon_n^2} \qquad \sup_{g : d_n(g,g_1) < \xi\epsilon_n} \mathbb{E}_g^{(n)}[1-\phi_n | X^n] \leq e^{-Cn\epsilon_n^2}$$

Proof. See Appendix A. □

4.4.1 Proofs for posterior contraction rates in FLR

The definition of the test functions are independent of the prior used. Therefore Lemma 8 can be adopted for both the sieve priors and the Gaussian process priors.
The next Subsections 4.4.1 and 4.4.1 provide the Kullback-Leibler and sieve conditions to solve the direct problem in FLR. The strategy to solve the inverse problem requires the definition of a sieve $\mathcal{G}_n$ where the norms associated to the direct and inverse problems are equivalent. The rate of convergence is then given by a quantity called *modulus of continuity* defined in this case as

$$\omega(\mathcal{G}_n, \epsilon_n) = \sup_{\beta \in \mathcal{G}_n : ||\beta - \beta_0||_\Gamma < \epsilon_n} ||\beta - \beta_0||.$$

Given the contraction rate result for the direct problem, the only condition to show is posterior consistency in the sieve: $\mathbb{E}_0^{(n)} \Pi(\mathcal{G}_n^c \,|\, Y^n, X^n) \to 0$. The latter is implied by exponentially decreasing prior mass assigned to the complement of the sieve S_n: $\Pi(\mathcal{G}_n^c) \lesssim e^{-2cn\epsilon_n^2}$, with $c > 0$ used in the Kullback-Leibler condition. For both priors the same sieves defined for the direct problem are used to invert the covariance operator as shown below.

Proofs for sieve prior in FLR

Under the assumption $\beta \in \mathbb{H}_{\alpha_1}$, the following lemmas provide the conditions for the contraction rate result in the direct problem with rate $\epsilon_n \asymp n^{-\frac{\alpha_1}{2(\alpha_1+\eta)+1}} (\log n)^{a_1}$. The proof for the next lemma is analogous to the proof of Lemmas 2 in [Arbel et al., 2013] and therefore it is omitted.

Lemma 9. *Under the assumptions of Theorem 4 there exists $c > 0$ such that*

$$\Pi(B_n^{KL}(\beta_0, \epsilon_n)) \gtrsim e^{-cn\epsilon^2}.$$

The second condition allows us to focus on a sequence of increasing subsets $\mathcal{F}_n \subset \mathbb{H}$ when defining the sequence of test functions required by the Theorem in [Ghosal et al., 2007]. For the sieve prior, a finite dimensional sieve $\mathcal{F}_n$ is

defined as

$$\mathcal{F}_n = \left\{ \beta = \sum_{j=1}^{j_n} \beta_j \varphi_j : \|\beta\|^2 \leq n^H \right\}$$

with $H > 0$ and $j_n \asymp \frac{n\epsilon_n^2}{\log n} \to \infty$. The next Lemma shows that the sieve condition is satisfied.

Lemma 10. *Under the assumptions of Theorem 4,*

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-(c+4)n\epsilon_n^2}$$
$$\log N(\xi\epsilon_n, \mathcal{F}_n, d) \lesssim n\epsilon_n^2.$$

Proof. See Appendix A. □

Lemma 11. *For $\xi \in (0,1)$ constant and for every $\beta_1 \in \mathcal{F}_n$ with $\|\beta_1 - \beta_0\| \geq \epsilon_n$, there exist tests $(\phi_n)_n$ such that*

$$\mathbb{E}_0^{(n)}[\phi_n] \leq e^{-Cn\epsilon_n^2} \qquad \sup_{\beta : \|\beta - \beta_1\|_\Gamma < \xi\epsilon_n} \mathbb{E}_\beta^{(n)}[1 - \phi_n] \leq e^{-Cn\epsilon_n^2}$$

Proof. See Appendix A. □

It is worth noting that, in the sieve $\mathcal{F}_n$, the covariance operator is continuously invertible, namely the two norms are equivalent up to a term that vanishes to zero as $n \to \infty$: for all $\beta \in \mathcal{F}_n$

$$\begin{aligned}
\|\beta - \beta_0\|_\Gamma^2 \leq \|\beta - \beta_0\|^2 &\leq \sum_{j=1}^{j_n} (\beta_j - \beta_{0j})^2 + \sum_{j > j_n} \beta_{0j}^2 \\
&\leq j_n^{2\eta} \sum_{j=1}^{j_n} j^{-2\eta} (\beta_j - \beta_{0j})^2 + j_n^{-2\alpha_1} \sum_{j > j_n} j^{2\alpha_1} \beta_{0j}^2 \\
&\leq j_n^{2\eta} \|\beta - \beta_0\|_\Gamma^2 + C j_n^{-2\alpha_1}
\end{aligned}$$

The Kullback-Leibler condition and the sieve condition are sufficient for the consistency in the sieve. Therefore $\omega(\mathcal{F}_n, \epsilon_n) = j_n^\eta \epsilon_n \asymp n^{-\frac{\alpha_1}{2(\alpha_1+\eta)+1}} (\log n)^{a_2}$ is the rate of convergence for the inverse problem.

Proofs for Gaussian process prior in FLR

Solving the direct problem is equivalent to recover $\Gamma^{1/2}\beta_0$ with respect to the Hilbert space norm. In the Gaussian process case, the covariance operator induces a Gaussian process prior on $\Gamma^{1/2}\beta \sim \text{GP}(0, \Gamma^{1/2}\Lambda\Gamma^{1/2})$, whose RKHS is the Sobolev space $\mathbb{H}_{\gamma+\eta+1/2}$. The contraction rate of the posterior distribution depends on the small ball probabilty and on the position of $\Gamma^{1/2}\beta_0$ with respect to the RKHS. Theorem 2.1 in [van der Vaart and van Zanten, 2008] shows that the prior mass condition is satisfied if the concentration function can be upper bounded by $n\epsilon_n^2$, where $\epsilon_n = n^{-\frac{\alpha_1 \wedge \gamma + \eta}{2(\alpha_1+\eta)+1}}(\log n)^{a_1}$ will be the posterior contraction rate. The proof of the following lemma can be found in chapter 10 of [Ghosal and van der Vaart,] and it is therefore omitted.

Lemma 12. *Under the assumptions of Theorem 5, let the concentration function be defined as*

$$\phi_{\Gamma^{1/2}\beta_0}(\epsilon_n) = \inf_{\substack{h \in \mathbb{H}_{\gamma+\eta+1/2} \\ ||h-\Gamma^{1/2}\beta_0|| \leq \epsilon_n}} \frac{1}{2}||h||^2_{\gamma+\eta+1/2} - \log \mathbb{P}(||\Gamma^{1/2}\beta|| < \epsilon_n)$$

If $\phi_{\Gamma^{1/2}\beta_0}(\epsilon_n) \leq n\epsilon_n^2$ *then*

$$\Pi(\beta : \|\beta - \beta_0\|_\Gamma < 2\epsilon_n) \geq e^{-n\epsilon_n^2}$$

Regarding the sieve condition, as done in in the sieve prior case, we will define a sieve such that its complexity, in tems of covering number, can be controlled, and such that the norms relative to the direct and inverse problems are equivalent up to a vanishing term. The former requirement on the complexity is used for bounding the type II error of tests uniformly in the sieve, while the latter is used to apply the Theorem in [Knapik and Salomond, 2014] to solve the inverse problem. Let us define the sieve $\mathcal{F}_n$ as

$$\mathcal{F}_n = \left\{ \beta : \sum_{j > j_n} \beta_j^2 < Q\xi_n^2, \quad \sum_{j=1}^{j_n} \beta_j^2 < v_n^2 \right\}$$

with $Q > 0$ constant, and the sequences $\xi_n \overset{n}{\to} 0$, $j_n \overset{n}{\to} \infty$ and $v_n \overset{n}{\to} \infty$ defined

as

$$j_n \asymp n^{\frac{1}{2(\gamma+\eta)+1}}, \qquad v_n > j_n^{\gamma}, \qquad \xi_n \asymp n^{-\frac{\alpha_1 \wedge \gamma}{2(\gamma+\eta)+1}}.$$

The following Lemma provides upper bounds for the prior mass assigned to the complement of the sieve and for the logarithm of its covering number:

Lemma 13. *Under the assumptions of Theorem 5 we have*

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-cn\epsilon_n^2}$$
$$\log N(\epsilon_n, \mathcal{F}_n, \|\cdot\|) \lesssim cn\epsilon_n^2$$

Proof. See Appendix A. □

The next Lemma provides the test condition for the direct problem. Its proof is analogous to the one given for the sieve prior.

Lemma 14. *For $\xi \in (0,1)$ and for every $\beta_1 \in \mathcal{F}_n$ with $\|\beta_1 - \beta_0\| \geq \epsilon_n$, there exist tests $(\phi_n)_n$ such that*

$$\mathbb{E}_0^{(n)}[\phi_n] \leq e^{-Cn\epsilon_n^2} \qquad \sup_{\beta:\|\beta-\beta_1\|_\Gamma < \xi\epsilon_n} \mathbb{E}_\beta^{(n)}[1-\phi_n] \leq e^{-Cn\epsilon_n^2}$$

The last three Lemmas ensure that the direct problem is solved with rate $\epsilon_n = n^{-\frac{\alpha_1 \wedge \gamma + \eta}{2(\alpha_1+\eta)+1}}$. The rate depends on the degree of ill-posedness η, the regularity of the true parameter α_1 and the smoothness of the prior γ, and the optimal rate is achieved when the prior smoothness matches the smoothness of the truth β_0.

In order to solve the inverse problem we can apply Theorem 2.1 in [Knapik and Salomond, 2014] whose conditions are satisfied by Lemmas 12 and 13, ensuring that the posterior mass vanishes exponentially fast outside the sieve $\mathcal{F}_n$. The change of metric in the sieve gives the rate of convergence in the inverse problem and can be computed as follows

$$\begin{aligned}\|\beta - \beta_0\|^2 &\leq \sum_{j=1}^{j_n} (\beta_j - \beta_{0j})^2 + \sum_{j>j_n} \beta_j^2 + \sum_{j>j_n} \beta_{0j}^2 \\ &\leq j_n^{2\eta} \|\beta - \beta_0\|_\Gamma^2 + M\xi_n^2 + j_n^{-2\alpha_1} \|\beta_0\|^2 \leq \omega(\mathcal{F}_n, \epsilon_n)\end{aligned}$$

with $\omega(\mathcal{F}_n, \epsilon_n) \asymp n^{-\frac{\alpha_1 \wedge \gamma}{2(\gamma+\eta)+1}}$. Therefore we obtain the rate of convergence for the inverse problem $\omega(\mathcal{F}_n, \epsilon_n) = n^{-\frac{\alpha_1 \wedge \gamma}{2(\alpha_1+\eta)+1}}$, where the optimal rate is again achieved when $\gamma = \alpha_1$.

4.4.2 Proofs for posterior constraction rates in F-SIM

We present here the proof of Theorem6.

Proof. The standard Theorem of [Ghosal et al., 2007] requires the existence of a sequence of test functions $(\phi_n)_n$ such that

$$\mathbb{E}_{g_0}^{(n)}[\phi_n | X^n] = o(1)\,, \qquad \sup_{g:\, d_n(g,g_0) > C\epsilon_n} \mathbb{E}_g^{(n)}[1 - \phi_n | X^n] \leq e^{-(c_1+2)n\epsilon_n^2}$$

Likelihood ratio tests can be used to test g_0 against balls of alternatives, as shown in Lemma 8. Tests whose type II errors are uniformly bounded on $\mathcal{F}_n$ can be defined in the same way it is done for the FLR case. Regarding the Kullback-Leibler condition, it is worth noting that the prior mass assigned to the Kullback-Leibler neighbourhood $S_n = \{g \in \Theta : \mathbb{E}_0[(g(X) - g_0(X))^2] \leq \epsilon_n^2\}$ can be lower bounded by the taking $f \in \mathcal{M}_n$. By restricting the parameter space by selecting f in the set $\mathcal{M}_n$, the KL distance can be upper bounded by the sum of two terms:

$$\mathbb{E}_{g_0}[(g(X) - g_0(X))^2] \leq 2\|f_0\|_\alpha^2 \mathbb{E}_{g_0}[\langle \beta - \beta_0, X\rangle^2]^{\min(\alpha_2,1)} + C J_n^4 2^{-2\alpha J_n} \mathbb{E}_{g_0}[\langle \beta, X\rangle^p]. \tag{4.4}$$

This bound can be proven as follows:

$$\mathbb{E}_0[(g_0(X) - g(X))^2] \leq 2\mathbb{E}_0[(f_0(\langle \beta_0, X\rangle) - f_0(\langle \beta, X\rangle))^2] + 2\mathbb{E}_0[(f_0(\langle \beta, X\rangle) - f(\langle \beta, X\rangle))^2].$$

From the regularity assumption on f_0 and using Jensen's inequality, the first term can be bounded by

$$\begin{aligned}\mathbb{E}_0[(f_0(\langle \beta_0, X\rangle) - f_0(\langle \beta, X\rangle))^2] &\leq \|f_0\|_\alpha^2 \mathbb{E}_0[|\langle \beta_0 - \beta, X\rangle^{2\min\{\alpha_2,1\}}] \\ &\leq \|f_0\|_\alpha^2 \mathbb{E}_0[|\langle \beta_0 - \beta, X\rangle^2]^{\min\{\alpha_2,1\}}\end{aligned}$$

The second term can be bounded using the assumptions of the approximating set $\mathcal{M}_n$, in particular the integral can be split over the partition $\{I_0, \dots, I_{J_n}, A_0^c\}$ of $\mathbb{R}$, with $I_j := A_j \setminus A_{j+1}$ for $j = 0, \dots, J_n - 1$ and $I_{J_n} = A_{J_n} = [-1, 1]$. In each I_j the expectation can be bounded as

$$\begin{aligned}\mathbb{E}_0[(f_0(\langle\beta, X\rangle) - f(\langle\beta, X\rangle))^2 1(\langle\beta, X\rangle \in I_j)] &\leq C_1 J_n^3 2^{-2j\alpha} \mathbb{P}_0\left(|\langle\beta, X\rangle| > 2^{2\alpha_2(J_n - j)/p}\right) \\ &\leq C_1 J_n^3 2^{2J_n\alpha} \mathbb{E}_{g_0}[|\langle\beta, X\rangle|^p]\end{aligned}$$

Analogously we have

$$\begin{aligned}\mathbb{E}_0[(f_0(\langle\beta, X\rangle) - f(\langle\beta, X\rangle))^2 1(\langle\beta, X\rangle \in A_0^c)] &\leq C_2 J_n^3 \mathbb{P}_0\left(|\langle\beta, X\rangle| > 2^{2\alpha_2 J_n}\right) \\ &\leq C_2 J_n^3 2^{-2\alpha_2 J_n} \mathbb{E}_0[|\langle\beta, X\rangle|^p]\end{aligned}$$

hence the KL distance upper bound 4.4 is proven. The p-th moment appearing in equation 4.4 can be bounded using Cauchy-Schwarz inequality and the assumption on the covariate:

$$\mathbb{E}_0[|\langle\beta, X\rangle|^p] \leq 2^{p-1}\mathbb{E}_0\left[\|X\|^p \left(\|\beta_0\|^p + \|\beta - \beta_0\|^p\right)\right] \lesssim C$$

since β_0 is in a Sobolev ball and $\|\beta - \beta_0\|$ is bounded with high probability by assumption on the sieve prior. By assumption $\epsilon_{f,n} = J_n^4 2^{-2J_n\alpha_2}$, hence it follows that $\Pi(S_n) \geq e^{-cn\epsilon_n^2}$ with $\epsilon_n = \max\left(\epsilon_{\beta,n}^{\min(\alpha_2,1)}, \epsilon_{f,n}\right)$. □

Proofs for location-scale mixture prior for the link function and sieve prior for the index function

The hybrid location-scale of normals prior satisfies the prior condition (P2) as shown in Propositions 7 and 8 of [Naulet and Rousseau, 2017]. Condition (P1) is satisfied by the sieve prior on the index function presented in the Section 4.2.2. In what follows we show that there exists a sieve satifying condition (P3). The sieve $\mathcal{F}_n$ used for the function non-linear function $g(\cdot) = f(\langle\beta, \cdot\rangle)$ is defined using sieves $\mathcal{F}_{\beta,n}$ and $\mathcal{F}_{f,n}$ for the functions β and f. The sieve for the link function f is analogous to the one introduced in [Naulet and Rousseau, 2017], while the sieve for β is a finite dimensional ball

with respect to Hilbert space norm:

$$\mathcal{F}_n = \left\{ \begin{array}{l} g = f(\langle \beta, X \rangle) : \ f(\cdot) = \int e^{-\frac{(\cdot-\mu)^2}{2\sigma^2}} dM(\sigma,\mu), \ M = \sum_{j\geq 1} u_j \delta_{\sigma_j,\mu_j} \\ \sum_{j\geq 1} |u_j| \leq n, \ \sum_{j\geq 1} |u_j| 1(|u_j| \leq \frac{1}{n}) \leq \epsilon_n \\ \sum_{j\geq 1} |u_j| 1(\sigma_j^2 < n^{-1/b_2}) \leq \epsilon_n, \ \sum_{j\geq 1} |u_j| 1(\sigma_j^2 > n^{1/b_1}) \leq \epsilon_n \\ |\{j : |u_j| > n^{-1}, \ n^{-1/b_2} \leq \sigma_j^2 \leq n^{1/b_1}\}| \leq \frac{Hn\epsilon_n^2}{\log n} \\ \beta = \sum_{j=1}^{K_n} \beta_h \varphi_j : \ \|\beta\|^2 = \sum_{j=1}^{K_n} \beta_j^2 \leq n^{b_3} \end{array} \right\}$$

for positive constants H, b_1, b_2, b_3, and for $K_n = \lfloor n\epsilon_{\beta,n}^2 / \log n \rfloor$. In order to control its complexity, let us define the set

$$\widetilde{\mathcal{F}}_n = \left\{ \begin{array}{l} g = f(\langle \beta, X \rangle) : \ f(\cdot) = \int e^{-\frac{(\cdot-\mu)^2}{2\sigma^2}} dM(\sigma,\mu), \ M = \sum_{j\in I} u_j \delta_{\sigma_j,\mu_j}, \ |I| \leq \frac{Hn\epsilon_n^2}{\log n} \\ \forall j \in I : \ \mu_j \in R_n, \ \sigma_j \in Q_n, \\ \sum_{j\geq 1} |u_j| \leq n, \forall j \in I : \ u_j = \frac{kn^{-3/2}}{H}, k \in \mathbb{Z}, \beta \in B_n^* \end{array} \right\}$$

where $B_n^* = \{\beta_1^*, \ldots, \beta_{N_\beta(n)}^*\}$ is the set of centers of ϵ_n-covering balls of $\{\beta = \sum_{j=1}^{K_n} \beta_j \phi_j \ : \ \|\beta\|^2 \leq n^{1/b_3}\}$ with respect to the Hilbert space norm; $\mathcal{J} = \mathbb{N} \cup \{\infty\}$, $\mathcal{K} = \{i \ : \ |u_i| \ > \ n^{-1}\}$, $\mathcal{L} = \{i \ : \ \mu_i \ \in \ S_n\}$, $\mathcal{I} = \mathcal{J} \cap \mathcal{K} \cap \mathcal{L}$; $R_n = \{\mu \in \mathbb{R} : \ \mu = kn^{-(3/2+1/b_2)}, k \in \mathbb{Z}\} \cap S_n$; $S_n = \cup_{i=1}^n \cup_{j=1}^{N_\beta(n)} \{z \in \mathbb{R} : |z - \langle \beta_j^*, X_i \rangle|\} \leq 4n^{1/b_1} \log n\}$; $Q_n = \{\sigma = n^{-1/b_2}(1 + n^{-(1+1/b_1+1/b_2)}\epsilon_n)^k, \ k \in \mathbb{N}, \ n^{-1/b_2} \leq \sigma \leq n^{1/b_1}\}$.

Lemma 15. *$\widetilde{\mathcal{F}}_n$ is a $C\epsilon_n$-net covering $\mathcal{F}_n$ with respect to the empirical prediction loss d_n.*

Proof. We need to prove that for each $m_1 \in \mathcal{F}_n$ there exist $m_2 \in \widetilde{\mathcal{F}}_n$ such that

$$d_n(m_1, m_2)^2 = \frac{1}{n} \sum_{i=1}^n (m_1(X_i) - m_2(X_i))^2 \lesssim \epsilon_n^2$$

Let β^* be such that $\|\beta - \beta^*\| \leq \epsilon_n$, and define

$$m_1(x) = \sum_{j\geq 1} u_j \phi_{\sigma_j}(\langle \beta, x \rangle - \mu_j), \qquad m_2(x) = \sum_{j\in I} u_j' \phi_{\sigma_j'}(\langle \beta^*, x \rangle - \mu_j')$$

then

$$
\begin{aligned}
d_n(m_1, m_2)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\geq 1} u_j\phi_{\sigma_j}(\langle\beta, X_i\rangle - \mu_j) - \sum_{j\in I} u'_j\phi_{\sigma'_j}(\langle\beta^*, X_i\rangle - \mu'_j)\right)^2 \\
\leq \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\geq 1} u_j\phi_{\sigma_j}(\langle\beta, X_i\rangle - \mu_j) - \sum_{j\in I} u'_j\phi_{\sigma'_j}(\langle\beta, X_i\rangle - \mu'_j)\right)^2 \\
+ \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\in I} u'_j\phi_{\sigma'_j}(\langle\beta, X_i\rangle - \mu'_j) - \sum_{j\in I} u'_j\phi_{\sigma'_j}(\langle\beta^*, X_i\rangle - \mu'_j)\right)^2.
\end{aligned}
$$

The first term in the last display can be bounded by $2\epsilon_n^2$, since for the indices $j \in \mathcal{I}$, u'_j and μ'_j can be chosen such that $|u_j - u'_j| \leq n^{-3/2}/H$ and $\left\|\mu_j - \mu'_j\right\| \leq n^{-(3/2+1/b_2)}$ (see Section 4.1.2 of [Naulet and Rousseau, 2017]). With abuse of notation we denote the vector of the first K_n coordinates of the functions β and β^* in the series representaion by β and β^*. The second term in the display above is bounded by

$$
\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\in I}\frac{u'_j}{\sigma_j^2}\langle\beta - \beta^*, X_i\rangle\right)^2 \leq \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\in I} n^{1+2/b_2}\langle\beta - \beta^*, X_i\rangle\right)^2 \\
&\leq 2n^{2+4/b_2^2}(\beta - \beta^*)^T\left(\frac{1}{n}\sum_{i=1}^{n}\overline{X}_{i,K_n}\overline{X}_{i,K_n}^T\right)(\beta - \beta^*)
\end{aligned}
$$

where $\overline{X}_{i,K_n} = (X_{i1}, \ldots, X_{iK_n})$ is the vector of the first K_n coefficients in the Karhuen-Loéve expansion $X_i = \sum_j X_{ij}\varphi_j$, since β and β^* live in a K_n-dimensional space. The quantity in the last display can be bounded using Cauchy-Schwarz inequality by

$$
2n^{2+4/b_2^2}\|\beta - \beta^*\|^2 \operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{X}_{i,K_n}\overline{X}_{i,K_n}^T\right) \leq 2n^{2+4/b_2^2}\epsilon_{\beta,n}^2\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_n}X_{ij}^2.
$$

By Markov inequality, we can bound the probability of the sum involving the covariates being larger that a small quantity $L_n \asymp n^{H_1}$ with $H_1 > 0$ constant:

$$
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_n}X_{ij}^2 > L_n\right) \leq \frac{1}{nL_n}\sum_{i=1}^{n}\sum_{j=1}^{K_n}\mathbb{E}[X_{ij}^2] = \frac{1}{L_n}\sum_{j=1}^{K_n} j^{-2\eta} \overset{n}{\to} 0
$$

which vanishes to zero for $H_1 > 0$. Therefore with high probability and for appropriate choice of H_1, we can bound the empirical distance

$$d_n(m_1, m_2)^2 \leq 2n^{2+4/b_2^2} \epsilon_n^2 L_n \leq \epsilon_n^2$$

□

Lemma 16. *The prior mass vanishes exponentially fast outside the sieve:*

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-c_2 n\epsilon_n^2}$$

Proof. Denoting by $\mathcal{F}_{\beta,n}$ and $\mathcal{F}_{f,n}$ the sieves for the functions β and f respectively, we have that the prior mass outside the sieve $\mathcal{F}_n$ for g is exponentially small:

$$\Pi(\mathcal{F}_n^c) \leq \Pi(\mathcal{F}_{f,n}^c) + \Pi(\mathcal{F}_{\beta,n}^c) \lesssim e^{-c_2 n\epsilon_n^2}$$

where the first term is bounded as in [Naulet and Rousseau, 2017] and the second term can be bounded as in the functional linear regression model (see 4.2.2). □

Lemma 17. *The complexity of the sieve can be controlled as follows*

$$\log N(\xi\epsilon_n, \mathcal{F}_n, d_n) \lesssim n\epsilon_n^2$$

Proof.

$$\log N(\xi\epsilon_n, \mathcal{F}_n, d_n) \leq \log N_f(n) + \log N_\beta(n) \lesssim n\epsilon_n^2$$

since $\log N_\beta(n) \leq n\epsilon_n^2$ and $\log N_f(n) \leq n\epsilon_n^2$ are the logarithms of the covering numbers of the sieves $\mathcal{F}_{\beta,n}$ and $\mathcal{F}_{\beta,n}$ respectively. □

It follows that the prior on g induced by hybrid location-scale mixture of normals prior for the link function f and by the sieve prior presented in 4.2.2 for the index function β satisfy the conditions of Theorem 6.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Posterior contraction rates in Bayesian Functional Regression models
Publication Status	☐Published ☐ Accepted for Publication ☐Submitted for Publication ☑Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Prof. Judith Rousseau (University of Oxford) and Prof. Angelina Roche (University of Paris Dauphine).

Student Confirmation

Student Name:	Giuseppe Di Benedetto		
Contribution to the Paper	I have derived the posterior contraction rates for the direct and inverse problems in Bayesian functional linear regression and applied the hybrid location-scale mixture of normals prior to the Bayesian functional single-index model.		
Signature		Date	20/04/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François Caron		
Supervisor comments			
Signature		Date	20/04/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Discussion

This Chapter concludes the dissertation by summarizing the results presented and describing possible extensions which have been left for future work.

5.1 Summary

The contributions that this manuscript brings to the Bayesian Nonparametrics' literature are twofold. From the statistical modelling perspective, Chapters 2 and 3 present two nonparametric models for discrete random structures, extending the number of applications Bayesian nonparametric methods can be used for. The first model consists in a class of non-exchangeable random partitions which captures the so called *microclustering* property, allowing the growth of clusters' sizes to be tuned with a parameter of the model. The second model is a non-exchangeable feature allocation model which satisfies a similar property, capturing sublinear trends for the features' sizes. Both models retains the good asymptotic properties of the exchangeable counterparts such as the power-law property, and manage to outperform them on real-world datasets.

From the statistical theory perspective, this thesis presents new asymptotic results for posterior distributions in Functional Data Analysis. In particular, posterior contraction rates are derived for Functional Linear Regression and Functional Single-Index Models for well-know priors used in Bayesian Non-

parametrics.

5.2 Future directions

Some questions regarding the topics presented are still unanswered. In the following we highlight some future directions that could be interesting to investingate.

5.2.1 Efficient inference scheme for Microclustering

Despite the non-exchangeable random partition models presented in Chapter 2 is attractive because of the flexibility it provides, the proposed inference scheme results to be computationally expensive. As described in [Di Benedetto et al., 2017], in order for the model to estimate the unobserved cluster labels, it has to infer the points $(\tau_i, \theta_i)_i$ of a latent point process. Based on the results obtained in the paper, we have studied a similar model where the arrival times $(\tau_i)_i$ are set deterministically to their asymptotic values. We prove the microclustering property for this model, however the alternative inference scheme which is derived for the new model does not seem to benefit from the removal of the sequence of arrival times in the estimation procedure. Details are provided in the Appendix B.

5.2.2 Network modelling

As shown in Section 5 of Chapter 2, the proposed non-exchangeable random partition model can be used to model sparse multigraphs with power-law degree distribution and sublinear growth rate for degree sequences. The ideas and a similar point process construction presented in Chapters 2 and 3 could be adapted to network modeling, in particular to design a generative model for sparse networks with dense communities. Let us recall that a network is said to be *sparse* if the number of edges grows subquadratically with the number of nodes, namely $n^{(e)} = o(n^2)$, where $n^{(e)}$ is the number of edges in a graph of n nodes; otherwise it is said to be *dense*. *Communities* are subgraphs

whose nodes share similar characteristics. Several work have focused on modeling networks with community structures, such as the stochastic block-model ([Holland et al., 1983]). In some applications it is reasonable to assume that such subgraphs contain many connections between their members, implying that communities are dense. At the same time, most of the networks observed are also sparse. These two assumptions lead to a necessary condition of sublinearity of the number of nodes of a community j, denoted by $m_{n,j}$, with respect to the overall number of nodes n in the network. To prove this, let us assume by contradition that $m_{n,j} \asymp n$. The above assumptions can be written as $n_j^{(e)} \asymp m_{n,j}^2$ (dense communities) and $n^{(e)} = o(n^2)$ (sparse graph). It follows that $n^{(e)} = o\left(n_j^2\right)$; since $n_j^{(e)} \leq n^{(e)}$ we have $n_j^{(e)} = o\left(m_{n,j}^2\right)$ which contradicts the assumption of dense communities. Therefore in the case each node belongs to only one community, the non-exchangeable random partition model presented in Chapter 2 might be adapted to design a solution that satisfies the asymptotic requirements mentioned above.

5.2.3 Non-exchangeable trait allocation models

The non-exchangeable model presented in [Di Benedetto et al., 2017], being a feature allocation model, assumes that each observation i shows a collection of features or equivalently belongs to a latent subgroup. This is encoded by a binary sequence $Z_i = (z_{ij})_j$ for each $i = 1, \ldots, n$. However, for some applications it is desirable to describe observations showing different levels of belonging to each subgroup. In this case Z_i can be modelled as a random sequence of integers. This models are referred to as *trait allocations*, and we refer to [Campbell et al., 2018] for more details and for a representation Theorem that applies to exchangeable trait allocations (see also [Zhou, 2014]). In particular, this de Finetti representation result implies that sum of all the levels of belongings to a particular subgroup (trait) grows linearly with the size of the data.

The model presented in Chapter 3 can be generalised to a non-exchangeable trait allocation model. Using the same notation, one could substitute the Bernoulli distribution for the variables z_{ij} in Equation (9) of [Di Benedetto et al., 2017]

with a Poisson distribution:

$$z_{ij} \,|\, \omega_j \theta_j \sim \text{Poi}\ (\omega_j \Delta_i 1_{\theta_j \leq Y_i}).$$

It is easy to show, following the same proof of the paper, that a sublinear trend for the traits' sizes can be captured: conditionally on the CRM defining the model,

$$m_{n,j} = \sum_{i=1}^{n} z_{ij} \overset{n\to\infty}{\sim} \omega_j T_n = \omega_j n^{\frac{1}{1+\xi}}.$$

The asymptotics for the number of traits remains unchanged, since $\mathbb{P}(z_{ij} > 0 \,|\, \omega_j \theta_j) = 1 - e^{-\omega_j \Delta_i 1_{\theta_j \leq Y_i}}$ which is equal to the Bernoulli parameter in the feature allocation model. Same asymptotics also holds for the number of total traits $m_n = \sum_{i=1}^{n} \sum_{j \geq 1} z_{ij}$.

5.2.4 Recovery of link and index functions in Functional SIM

While in Chapter 4 we provide we provide posterior contraction rates for FLR models with respect to the norms of the direct and inverse problems, in F-SIM the rates are derived only with respect to the empirical loss function. Therefore it would be interesting to marginalise out the predictor and derive the posterior contraction rates with respect to the prediction loss. Moreover, it would be worth studying the also to separately recover the two functional parameters of the model: the link and the index functions. A second direction that would be worth considering consists in doing an empirical analysis of the Bayesian F-SIM with the prior described in Chapter 4, together with applications on real datasets. A Gibbs sampler that alternates updates for the two functional parameters could be used for sampling from the posterior. If a sieve prior is used on the index function, a Reversible Jump Monte Carlo Markov Chain algorithm can be adopted to make the updates. Analogously, if the hybrid location-scal mixture of normals is used for the link function, then a Reversible Jump Monte Carlo Markov Chain ([Green, 1995]) similar to the one introduced in [Wolpert et al., 2011] might be used for updating the link

function in the Gibbs step.

Appendix A

Technical proofs of Chapter 4

Proof of Lemma 8. Let us consider the likelihood ratio tests defined as

$$\phi_n(\beta) = 1(Z_n(\beta) > \delta_n(\beta))$$

where $Z_n(g) = \ell_n(g) - \ell_n(g_0)$ and $\delta_n(g) = \delta_0 n d_n^2(g, g_0)$, with the empirical distance defined as $d_n^2(\beta, \beta_0) = \frac{1}{n}\sum_{i=1}^n (g(X_i) - g_0(X_i))^2$. The aim is to bound type I and II errors by $\exp(-cn\epsilon_n^2)$, conditionally on X^n. The type I error can be written as

$$\mathbb{E}_0^{(n)}[\phi_n(g)|X^n] = \mathbb{P}_0^{(n)}(Z_n(g) > \delta_n(g)|X^n)$$

Given X^n, $Z_n(\beta)$ is a Gaussian random variable with mean $\mathbb{E}_0^{(n)}[Z_n(\beta)|X^n] = -\frac{1}{2\sigma^2}\sum_{i=1}^n (g(X_i) - g_0(X_i))^2$ and variance $\mathrm{Var}_0^{(n)}[Z_n(\beta)|X^n] = \frac{1}{\sigma^2}\sum_{i=1}^n (g(X_i) - g_0(X_i))^2$, therefore using Gaussian tail inequality we have

$$\mathbb{E}_0^{(n)}[\phi_n(g)|X^n] \leq \exp\left\{ -\frac{1}{2}\frac{(\delta_n(g) - \mathbb{E}_0^{(n)}[Z_n(g)|X^n])^2}{\mathrm{Var}_0^{(n)}(Z_n(g)|X^n)} \right\}$$

$$= \exp\left\{ -\frac{\left(\delta_0 n d_n^2(g, g_0) + \frac{n}{2\sigma^2} d_n^2(g, g_0)\right)^2}{\frac{2n}{\sigma^2} d_n^2(g, g_0)} \right\} \leq e^{-cn d_n^2(g_0, g)}$$

with the positive constant $c = \frac{1}{2\sigma^2}\left(\delta_0\sigma^2 + \frac{1}{2}\right)^2$.
Recalling $\epsilon_i = Y_i - g(X_i)$, we have that the type II error can be written as

follows

$$\begin{aligned}
\mathbb{E}_g^{(n)}[1-\phi_n(\beta_1)|X^n] &= \mathbb{P}_g^{(n)}(\ell_n(g_1)-\ell_n(g_0)<\delta_n(g_1)|X^n) \\
&= \mathbb{P}_g\left(\sum_{i=1}^n \epsilon_i(g_1(X_i)-g_0(X_i)) + g(X_i)(g_1(X_i)-g_0(X_i)) + \frac{g_0(X_i)^2}{2} - \frac{g_1(X_i)^2}{2} < \delta_n(\beta_1)\right) \\
&= \mathbb{P}_g\Bigg(\sum_{i=1}^n \epsilon_i(g_1(X_i)-g_0(X_i)) + \frac{n}{2}d_n^2(g_0,g_1) + \underbrace{\sum_{i=1}^n (g_0(X_i)-g_1(X_i))(g_1(X_i)-g_0(X_i))}_{\Delta_n} < \delta_n(\beta_1)\Bigg) \\
&= \mathbb{P}_g\left(-\sum_{i=1}^n \epsilon_i(g_1(X_i)-g_0(X_i)) > \frac{n}{2}d_n^2(\beta_0,\beta_1) - \delta_n(\beta_1) + \Delta_n\right)
\end{aligned}$$

Since $|\Delta_n| \leq \left(\sum_{i=1}^n (g(X_i)-g_1(X_i))^2\right)^{1/2}\left(\sum_{i=1}^n (g_1(X_i)-g_0(X_i))^2\right)^{1/2} \leq n d_n(g_0,g_1)^2$ it follows that

$$\mathbb{E}_g^{(n)}[1-\phi_n(\beta_1)|X^n] \leq e^{-cnd_n(g_0,g)^2}.$$

□

Proof of Lemma 10. The proof of the first statement about the sieve prior mass is analogous to the proof of Lemma 3 in [Arbel et al., 2013]. Regarding the complexity of the sieve, $\mathcal{F}_n$ is a ball of radius n^H in $\mathbb{R}^{j_n}$, hence if $r_n = \xi\epsilon_n$ is the radius of the covering balls, its entropy number $\log N_n$ can be bounded as follows

$$\log N(r_n, \mathcal{F}_n, \|\cdot\|) \leq j_n \log\left(\frac{n^H\sqrt{j_n}}{2r_n}\right) \leq j_n \log\left(n^H j_n^{\frac{1}{2}+2\eta}\right)$$

$$\leq \left(H + \frac{\frac{1}{2}+2\eta}{2(\alpha_1+\eta)+1}\right) j_n \log n \lesssim n\epsilon_n^2$$

where $j_n \asymp \frac{n\epsilon_n^2}{\log n}$. □

Proof of Lemma 14. From Lemma 8, in the linear case, the unconditional type I error of the likelihood ratio tests can be bounded by taking the expectation with respect to X^n. Therefore we have

$$\mathbb{E}_0^{(n)}[\phi_n(\beta)] \leq \mathbb{E}\left[\prod_{i=1}^n \exp\left\{-c^2\langle X_i, \beta_0-\beta\rangle^2\right\}\right] = \left(\mathbb{E}\left[\exp\left\{-c^2\langle X, \beta_0-\beta\rangle^2\right\}\right]\right)^n$$

By the fourth moment assumption, described in Section 4.2.1, we have that for $t > 0$ small enough $\mathbb{E}[\langle X, \beta - \beta_0\rangle^4] \leq c_1 \mathbb{E}[\langle X, \beta - \beta_0\rangle^2]$ when $\|\beta - \beta_0\|^2 < t$, therefore using the inequality $1 - x \leq e^{-x} \leq 1 - x + x^2/2$ valid for $x > 0$, we have that

$$\mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] = \mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] 1_{\|\beta-\beta_0\|<t} + \mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] 1_{\|\beta-\beta_0\|\geq t})$$

The first term can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] 1_{\|\beta-\beta_0\|<t} &\leq \mathbb{E}\left[1 - \left(c - \frac{c_1 c^2}{2}\right)\langle X, \beta - \beta_0\rangle^2\right] 1_{\|\beta-\beta_0\|<t} \\
&\leq (1 - \tilde{c}\,\|\beta - \beta_0\|_\Gamma^2) 1_{\|\beta-\beta_0\|<t} \\
&\leq e^{-\tilde{c}\|\beta-\beta_0\|_\Gamma^2} 1_{\|\beta-\beta_0\|<t}
\end{aligned}$$

In order to bound the second term let us take a constant $s \in (0, t]$ and denote $\delta := s\frac{\beta-\beta_0}{\|\beta-\beta_0\|}$, then we can write

$$\begin{aligned}
\mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] = \mathbb{E}\left[e^{-\frac{c^2\|\beta-\beta_0\|^2}{s^2}\langle X,\delta\rangle^2}\right] &\leq 1 - \frac{c^2\|\beta-\beta_0\|^2}{s^2}\left(1 - \frac{c^2\|\beta-\beta_0\|^2}{s^2}c_1\right)\mathbb{E}\langle X,\delta\rangle^2 \\
&\leq e^{-\frac{c^2\|\beta-\beta_0\|^2}{s^2}\left(1-\frac{c^2\|\beta-\beta_0\|^2}{s^2}c_1\right)\frac{s^2}{\|\beta-\beta_0\|^2}\mathbb{E}\langle X,\beta-\beta_0\rangle^2}
\end{aligned}$$

Let us further split the space and first consider $\beta \in \mathbb{H}$ such that $\|\beta - \beta_0\|^2 \geq \frac{s^2}{c_1 c^2}$

$$\begin{aligned}
\mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] = \mathbb{E}\left[e^{-\frac{c\|\beta-\beta_0\|^2}{s^2}\langle X,\delta\rangle^2}\right] &\leq \mathbb{E}\left[e^{-\frac{1}{c_1}\langle X,\delta\rangle^2}\right] \\
&\leq e^{-\tilde{c}\mathbb{E}\langle X,\delta\rangle^2} \leq e^{-\tilde{c}t^2\frac{\|\beta-\beta_0\|_\Gamma^2}{\|\beta-\beta_0\|^2}} \leq e^{-\tilde{c}t^2}
\end{aligned}$$

where in the second inequality follows analogously from the previous calculation. For the case $\|\beta - \beta_0\|^2 < \frac{s^2}{c_1 c^2}$ we have

$$\mathbb{E}\left[e^{-c\langle X,\beta-\beta_0\rangle^2}\right] \leq e^{-c^2\|\beta-\beta_0\|_\Gamma^2}$$

Therefore we have that the type I error is exponentially decreasing

$$\mathbb{E}_0^{(n)}[\phi_n(\beta)] \leq e^{-c_5 n\epsilon_n^2}.$$

About the type II error, in the linear case, following the proof of Lemma 14, $|\Delta_n|$ can be bounded by $\|\beta-\beta_1\|\,\|\beta_1-\beta_0\|\,\sum_{i=1}^n \|X_i\|_2^2$ and $\sum_{i=1}^n \frac{\|X_i\|^2}{n} \leq 2$, hence $|\Delta_n| \leq \xi\, n\, \|\beta_1-\beta_0\|_\Gamma^2$ is implied by $\|\beta-\beta_1\| \leq \xi \frac{\|\beta_1-\beta_0\|_\Gamma^2}{2\|\beta_1-\beta_0\|}$, which is true if

$$\|\beta-\beta_1\| \leq \frac{\xi}{\lambda_{j_n}^{-1}+\frac{c_{12}}{\epsilon_n^2}} \leq \xi\epsilon_n$$

since $\|\beta_1-\beta_0\|_\Gamma > \epsilon_n$ and $\|\beta_1-\beta_0\|^2 = \sum_{j=1}^{j_n}(\beta_{1j}-\beta_{0j})^2 + \sum_{j=j_n+1}^{\infty}\beta_{0j}^2 \leq \lambda_{j_n}^{-1}\|\beta_1-\beta_0\|_\Gamma^2 + b_{j_n}$ with b_{j_n} decreasing to zero. Therefore, applying Gaussian tail inequality we get

$$\mathbb{E}_\beta^{(n)}(1-\phi_n(\beta_1)|X^n) \leq \exp\left\{-\frac{\left(\xi n d_\lambda^2(\beta_0,\beta_1)+\frac{n}{4}d_n^2(\beta_0,\beta_1)\right)^2}{2nd_n^2(\beta_0,\beta_1)}\right\} \leq \exp\left\{-\frac{1}{32}nd_n^2(\beta_0,\beta_1)\right\}$$

In order to maginalize out X^n we can use the same technique used to bound the type I error, obtaining

$$\sup_{\beta:\|\beta-\beta_1\|_\Gamma<\xi\epsilon_n} \mathbb{E}_\beta^{(n)}[1-\phi_n(\beta_1)] \leq e^{-c_5 n\epsilon_n^2}.$$

□

Proof of lemma 13. The prior mass of $\mathcal{F}_n^c$ can be bounded from below as follows:

$$\mathbb{P}\left(\sum_{j=1}^{j_n} j^{-2\gamma-1}Z_j^2 > v_n^2\right) \leq e^{-sv_n^2}\prod_{j=1}^{j_n}\mathbb{E}e^{s\,j^{-2\gamma-1}Z_j^2} \leq e^{-sv_n^2+\frac{s}{1-2s}j_n^{2\gamma}} \leq e^{-v_n^2/8} \leq e^{-n\epsilon_n^2/8}$$

by Markov inequality setting $s=\frac{1}{8}$ and for $v_n > j_n^\gamma$.

$$\mathbb{P}\left(\sum_{j>j_n} j^{-2\gamma-1}Z_j^2 > M\xi_n^2\right) \leq e^{-sM\xi_n^2}\prod_{j>j_n}(1-2sj^{-2\gamma-1})^{-1/2}$$

taking the log on both sides and using the inequality $\log x \geq \frac{-x}{1-x}$, we have the

upper bound

$$-sM\xi_n^2 + \sum_{j>j_n} \frac{sj^{-2\gamma-1}}{1-2sj^{-2\gamma-1}} \leq -sM\xi_n^2 + \frac{2\gamma+1}{2\gamma} \frac{sj_n^{-2\gamma}}{1-2sj_n^{-2\gamma-1}} \leq cn\epsilon_n^2$$

by taking $s = \frac{j_n^{1+2\gamma}}{4}$ and $j_n^{2\gamma} \geq 2\frac{2\gamma+1}{\gamma M}\xi_n^2$.
The complexity of the sieve is measured with the number of covering balls of radius $r_n = \xi\epsilon_n$. By assumption on ξ_n and v_n, it is sufficient to cover a j_n-dimensional ball, since for every two points $\beta, \beta' \in \mathcal{F}_n$ we have that $\|\beta - \beta'\|^2 = \sum_{j=1}^{j_n}(\beta_j - \beta_j')^2 + 2M\xi_n^2 \leq c_3 v_n^2$ for $c_3 > 0$ constant. Therefore we can cover the ball of radius $c_3 v_n$ using balls of radii r_n. Hence the logarithm of the covering number can be bounded as follows

$$\log N(r_n, \mathcal{F}_n, \|\cdot\|) \leq j_n \log\left(\frac{v_n\sqrt{j_n}}{2r_n}\right) \lesssim n\epsilon_n^2$$

□

Appendix B

Details about extension of the random partition model for microclustering

Deterministic arrival times model

Given the asymptotic results of the random partition model defined in the paper [Di Benedetto et al., 2017], we ask whether the random partition model still satisfies the microclustering property when the arrival times are defined deterministically following Theorem 4 in [Di Benedetto et al., 2017], where the asymptotic arrival time $\tau_{(n)}$ of the n-th observation is derived. Keeping the same notation of the paper, let us denote $\overline{\overline{\alpha}}(t) = \int_0^t \overline{\alpha}(s)ds$ and the deterministic arrival time of the n-th element as $f(n) := \overline{\overline{\alpha}}^{-1}(n)$, up to a slowly varying factor. Starting from the point process construction described in the paper, we can remove the random arrival times, and sample the clusters' locations $(\theta_{(k)})_k$ as follows

$$\begin{aligned} W &\sim CRM(\alpha, \rho) \\ \theta_{(k)} \,|\, W &\overset{ind}{\sim} \frac{W(d\theta)}{\overline{W}(f(k))} 1_{(\theta_{(k)} < f(k))}, \qquad k \in \mathbb{N} \end{aligned} \tag{B.1}$$

It can be proved that this model satisfies the microclustering property as shown in the Proposition below.

Proposition 18. *The random partition model given by* (B.1) *satisfies the microclustering property, in particular as* $n \to \infty$ *conditionally on* W

$$m_{n,j} \sim_{a.s.} W(\theta_j^*)\, f(n)$$

Proof. Let us denote the number of observations with cluster location parameter ϑ_j by $X_{n,j} = \sum_{k=1}^{n} 1(\theta_{(k)} = \vartheta_j)$, where ϑ_j is an atom of W. Conditionally on W, the indicator random variables $Y_k := 1(\theta_{(k)} = \vartheta_j)$ are independent Bernoulli distributed random variables with parameters $\frac{\omega_j}{\overline{W}(f(k))}$. The objective is to prove $\frac{X_{n,j}}{f(n)} \to C \in (0, \infty)$ as $n \to \infty$ a.s. conditionally on W. The Kolmogorov SLLN can be used, with scaling sequence $f(n)$ and partial sum $X_{n,j}$. The condition to prove is the summability of the variances:

$$\sum_{k \geq 1} \mathrm{Var}\left(\frac{Y_k}{f(k)}\right) = \omega_j \sum_{k \geq 1} \left(\frac{\overline{W}(f(k)) - \omega_j}{\overline{W}(f(k))^2} \frac{1}{f(k)^2}\right).$$

By assumption and from the results in the paper, we have that $f \in RV_{\frac{1}{\xi+1}}$ and $\overline{W}(t) \in RV_{\xi}$, since $\frac{\overline{W}(t)}{\overline{\alpha}(t)} \to \kappa(1,0)$ a.s. as $t \to \infty$. It follows that the addends in the infinite sum are evaluations of a regularly varying function with index $-\frac{\xi+2}{\xi+1} < -1$, therefore the infinite sum is convergent. In fact, for any function $h \in RV_a$ with $a < -1$, we can observe that the function $h_\epsilon(k) := k^{a+\epsilon}$ with $\epsilon > 0$ and $a + \epsilon < -1$ is such that $h_\epsilon(k) \geq h(k)$ for k large and its infinite sum converges. The Kolmogorov SLLN gives

$$\frac{X_{n,j} - \mathbb{E}[X_{n,j}]}{f(n)} \to 0 \qquad a.s.$$

We now need to prove that $\frac{\mathbb{E}[X_{n,j}]}{f(n)} \to c \in (0, \infty)$. We can prove that using Stolz-Cesàro theorem. Let us use the notation $g := \overline{W} \circ f$,

$$\frac{1}{f(n)} \sum_{k=1}^{n} \frac{1}{g(k)} = \frac{\sum_{k=1}^{n} a_k b_k}{\sum_{k=1}^{n} b_k} \tag{B.2}$$

with $a_k = \frac{1}{g(k)(f(k)-f(k-1))}$ and $b_k = f(k) - f(k-1)$. Then, if $\lim_k a_k = c$ and $\sum_{k=1}^{n} b_k$ is strictly increasing, it follows that the ratio in (B.2) converges

to c. Now, denoting by ℓ_f and ℓ_f the slowly varying parts of the f and g respectively, we have

$$\lim_k g(k)(f(k) - f(k-1)) = \lim_k \ell_g(k) k^{\frac{\xi}{\xi+1}} \left(\ell_f(k) k^{\frac{1}{\xi+1}} - \ell_f(k-1)(k-1)^{\frac{1}{\xi+1}} \right)$$

and considering the case $\overline{\alpha}(t) = t^\xi$, the previous limit converges to a non-zero constant

$$\lim_k k^{\frac{\xi}{\xi+1}} \left(\frac{k^{-\frac{\xi}{\xi+1}}}{\xi+1} + o\left(k^{-\frac{\xi}{\xi+1}}\right) \right) = \frac{1}{\xi+1}$$

from which the result follows. □

Inference

An alternative way to sample from the model would consist in a two step procedure. Let us define the random variables $u_j = W([f(j-1), f(j)])$ and for each $n \in \mathbb{N}$, let us introduce n random variables $\pi_1^{(n)}, \ldots, \pi_n^{(n)}$ summing up to 1, defined as $\pi_i^{(n)} = \frac{u_i}{\sum_{k=1}^n u_k}$. Given the variables $u_{1:n}$, we can sample the interval, among $\{[f(j-1), f(j)]\}_{j=1,\ldots,n}$, that each θ_i belongs to. Let us index the interval sampled for the i-th observation by z_i. In particular we have that $z_1 = 1$ a.s.,

$$z_2 = \begin{cases} 1 & \text{wp } \pi_1^{(2)} = \frac{u_1}{u_1+u_2} \\ 2 & \text{wp } \pi_2^{(2)} = \frac{u_2}{u_1+u_2} \end{cases}$$

and in general $z_k = j$ with probability $\pi_j^{(k)} = \frac{u_j}{\sum_{m \leq k} u_m}$ for $k \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$. Henceforth we omit the superscript of the variables π_i's. We can write

$$p(z_{1:n} \,|\, u_{1:n}) = u_1^{s_1 - 1} \prod_{j=2}^{n} u_j^{s_j} \prod_{j=2}^{n-1} \left(\sum_{\ell=1}^{j} u_\ell \right)^{-1}$$

where $s_j = \sum_{\ell=1}^n 1(z_\ell = j)$ is the number of points falling in the j-th segment. In what follows we will marginalize out the variables $u_{1:n}$ in the case the CRM is a Gamma Process.

Gamma Process case

In the case $W \sim \mathrm{GP}(\alpha)$, then $\pi_{1:n} \sim \mathrm{Dir}(\alpha_{1:n})$ with $\alpha_j := \alpha([f(j-1), f(j)])$. Therefore we have that

$$p(\pi_{1:n} \,|\, z_{1:n}) \propto \pi_1^{\alpha_1+s_1-2} \prod_{j=1}^{n} \pi_j^{\alpha_j+s_j-1} \prod_{j=2}^{n-1} \left(\sum_{\ell \le j} \pi_\ell \right)^{-1}$$

which is the Generalized Dirichlet distribution defined by [Connor and Mosimann, 1969]. In order to find the parameters of the distributions, let us introduce n independent random variables $X_{1:n}$ such that $X_1 = 1$ a.s. and $X_i \sim \mathrm{Beta}(a_i, b_i)$ for $i = 2, \ldots, n$; define $Y_{1:n}$ as $Y_j = X_j \prod_{k>j}(1 - X_k)$. It follows that $Y_{1:n}$ is a *neutral to the left* vector since $\frac{Y_i}{1-\sum_{k>j} Y_k} = X_i$ are independent, and we get

$$\mathbb{E}[Y_j] = \frac{a_j}{a_j + b_j} \prod_{k>j} \frac{b_k}{a_k + b_k}$$

In order to get the joint distribution of $Y_{1:n}$ we can just apply a change of variable to

$$p(X_{2:n} = x_{2:n}) \propto \prod_{j=2}^{n} x_j^{a_j-1} (1 - x_j)^{b_j-1}$$

obtaining

$$\begin{aligned} p(Y_{1:n} = y_{1:n}) \propto & \left(\prod_{j=2}^{n-1} y_j^{a_j-1} \left(1 - \sum_{k>j} y_k \right)^{2-(a_j+b_j)} \left(1 - \sum_{k \ge j} y_k \right)^{b_j-1} \right) \\ & \times y_n^{a_n-1} (1 - y_n)^{b_n-1} \prod_{j=2}^{n-1} \left(1 - \sum_{k<j} y_k \right)^{-1} \\ = & \, y_1^{b_2-1} \left(\prod_{j=2}^{n} y_j^{a_j-1} \right) \prod_{j=2}^{n-1} \left(\sum_{k \le j} y_k \right)^{b_{j+1}-(a_j+b_j)} \end{aligned}$$

In the distribution of $\pi_{1:n} \,|\, z_{1:n}$ the parameters are

$$\begin{aligned} a_j^{(n)} &= \alpha_j + s_j^{(n)} \\ b_j^{(n)} &= \sum_{k<j} (\alpha_k + s_k^{(n)}) - (j-2) \end{aligned}$$

for $j = 2, \ldots, n$, therefore

$$\begin{aligned} p(z_k = j \,|\, z_{1:k-1}) &= \mathbb{E}_{\pi_{1:n} \,|\, z_{1:n-1}} \left[p(z_n \,|\, \pi_{1:n}) \right] \\ &= \frac{a_j^{(k-1)}}{a_j^{(k-1)} + b_j^{(k-1)}} \prod_{\ell > j} \frac{b_\ell^{(k-1)}}{a_\ell^{(k-1)} + b_\ell^{(k-1)}}. \end{aligned}$$

After having sampled the segments the $\theta_{(n)}$ belongs to, we have

$$\mathbb{P}(\theta_{(n)} \in dx \,|\theta_{(1:n-1)}, z_n) = \frac{\alpha(dx)}{\alpha_j + s_j} + \frac{1}{\alpha_j + s_j} \sum_{k \,|\, z_k = z_n} \delta_{\theta_{(k)}}(dx)$$

which means that conditioning on the segment variables, the partition of points belonging to the same segment is distributed according to the CRP. Denoting by $c_{1:n}$ the partition of the first n observations, we have

$$\begin{aligned} p(c_k = j, z_k \,|\, c_{1:k-1}, z_{1:k-1}) = &\left(\frac{m_j}{a_{z_k}^{(k)} + b_{z_k}^{(k)}} \prod_{\ell = z_k + 1}^{k} \frac{b_\ell^{(k)}}{a_\ell^{(k)} + b_\ell^{(k)}} \right) \delta(j < K_k) \quad \text{(B.3)} \\ &+ \left(\frac{\alpha_{z_k}}{a_{z_k}^{(k)} + b_{z_k}^{(k)}} \prod_{\ell = z_k + 1}^{k} \frac{b_\ell^{(k)}}{a_\ell^{(k)} + b_\ell^{(k)}} \right) \delta(j = K_k + 1) \end{aligned}$$

where K_n is the number of clusters in $\{1, \ldots, n\}$ and m_j is the number of observations in the j-th cluster.

Assuming we observe the true partition, we now have a single layer of latent variables $z_{1:n}$ to infer. The weights of the particles are updated by computing (B.3) which involves a product with a large number of factors in the case z_k takes a low value. It is therefore worth inverstigating, at least empirically, which values such latent variables takes more often. Let us denote the CMR's weight associated to the j-th cluster by $\omega_j^* := W(\theta_j^*)$, and by

$\overline{\omega}_j^* = \sum_{k=1}^n \omega_k^* 1_{(z_k=j)}$ the sum of masses given by the CRM to the cluster locations falling in the j-th segment. Let us recall $f(n) = \overline{\overline{\alpha}}^{-1}(n)$ and denote $g(n) = \overline{\alpha}(f(n))$. Therefore we have that conditionally on W, as $n \to \infty$

$$a_j(n) \sim g(j) - g(j-1) + \overline{\omega}_j^* f(n)$$
$$b_j(n) \sim g(j-1) + f(n) \sum_{l=1}^{n-1} \overline{\omega}_l^*$$

hence

$$\frac{b_j(n)}{a_j(n) + b_j(n)} \sim \frac{\sum_{l=1}^{n-1} \overline{\omega}_l^*}{\sum_{l=1}^{n} \overline{\omega}_l^*}$$
$$\frac{a_j(n)}{a_j(n) + b_j(n)} \sim \frac{\overline{\omega}_j^*}{\sum_{l=1}^{n-1} \overline{\omega}_n^*}$$

and

$$p(z_n = j \,|\, z_{1:n-1}) \sim \frac{\overline{\omega}_j^*}{\sum_{l=1}^{n-1} \overline{\omega}_n^*} \prod_{i=k+1}^{n} \frac{\sum_{l=1}^{i-1} \overline{\omega}_l^*}{\sum_{l=1}^{i} \overline{\omega}_l^*} = \frac{\overline{\omega}_j^*}{\sum_{i=1}^{n} \overline{\omega}_i^*}$$

The conditional probability $p(z_n = j \,|\, z_{1:n-1})$ scales as the mass proportion of the j-th segment. As shown in B.1, the variables z_i's are more likely to take smaller values, making the computation of the probabilities in (B.3) very expensive.

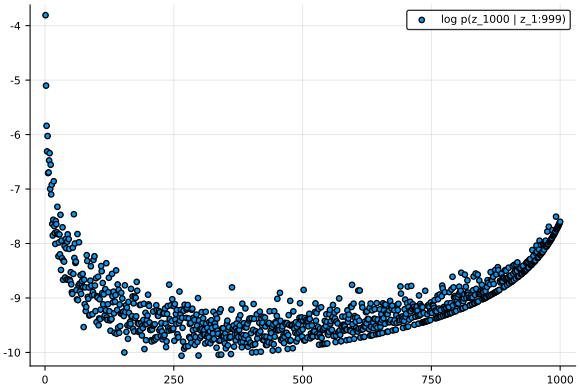


Figure B.1: Mean of the conditional log probabilities of z_{1000} using 30 samples with $\xi = 1$.

Bibliography

[Aldous, 1985] Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

[Alquier and Biau, 2013] Alquier, P. and Biau, G. (2013). Sparse single-index model. *Journal of Machine Learning Research*, 14(Jan):243–280.

[Antoniadis et al., 2004] Antoniadis, A., Grégoire, G., and McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, pages 1147–1164.

[Arbel et al., 2013] Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian journal of statistics*, 40(3):549–570.

[Archambeau et al., 2014] Archambeau, C., Lakshminarayanan, B., and Bouchard, G. (2014). Latent ibp compound dirichlet allocation. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):321–333.

[Benedetto et al., 2020] Benedetto, G. D., Caron, F., and Teh, Y. W. (2020). Non-exchangeable feature allocation models with sublinear growth of the feature sizes.

[Bernardo et al., 2003] Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742.

[Blackwell et al., 1973] Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

[Brunel et al., 2016] Brunel, É., Mas, A., and Roche, A. (2016). Non-asymptotic adaptive prediction in functional linear models. *Journal of Multivariate Analysis*, 143:208–232.

[Cai and Hall, 2006] Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.*, 34(5):2159–2179.

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

[Cardot et al., 1999] Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

[Cardot and Johannes, 2010] Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Analysis*, 101:395–408.

[Chagny and Roche, 2016] Chagny, G. and Roche, A. (2016). Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, 146:105 – 118. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

[Chen et al., 2011] Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.*, 39(3):1720–1747.

[Choi et al., 2011] Choi, T., Shi, J. Q., and Wang, B. (2011). A gaussian process regression approach to a single-index model. *Journal of Nonparametric Statistics*, 23(1):21–36.

[Connor and Mosimann, 1969] Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the

dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

[Daley and Vere-Jones, 2007] Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

[Dang and Chainais, 2017] Dang, H.-P. and Chainais, P. (2017). Indian buffet process dictionary learning: algorithms and applications to image processing. *International Journal of Approximate Reasoning*, 83:1–20.

[De Finetti, 1937] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[Dhara et al., 2018] Dhara, K., Lipsitz, S., Pati, D., Sinha, D., et al. (2018). A new bayesian single index model with or without covariates missing at random. *Bayesian Analysis*.

[Di Benedetto et al., 2017] Di Benedetto, G., Caron, F., and Teh, Y. W. (2017). Non-exchangeable random partition models for microclustering. *arXiv preprint arXiv:1711.07287*.

[Diaconis and Freedman, 1986] Diaconis, P. and Freedman, D. (1986). On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26.

[Doob, 1949] Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27.

[Escobar and West, 1995] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

[Ferguson, 1973] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

[Freedman, 1963] Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[Ganti et al., 2017] Ganti, R., Rao, N., Balzano, L., Willett, R., and Nowak, R. (2017). On learning high dimensional structured single index models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Geenens and Delecroix, 2006] Geenens, G. and Delecroix, M. (2006). A survey about single-index models theory. *International Journal of Statistics and Systems*.

[Ghahramani and Griffiths, 2006] Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

[Ghosal et al., 2000] Ghosal, S., Ghosh, J. K., Van Der Vaart, A. W., et al. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

[Ghosal and van der Vaart,] Ghosal, S. and van der Vaart, A. Fundamentals of nonparametric bayesian inference; 2016. *Cambridge Series on Statistical and Probabilistic Mathematics*, 44.

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

[Ghosal et al., 2007] Ghosal, S., Van Der Vaart, A., et al. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

[Gnedin et al., 2007] Gnedin, A., Hansen, B., Pitman, J., et al. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171.

[Green, 1995] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

[Griffiths and Ghahramani, 2011] Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224.

[Gubian et al., 2015] Gubian, M., Torreira, F., and Boves, L. (2015). Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49:16–40.

[Hall and Horowitz, 2007] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35(1):70–91.

[Holland et al., 1983] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

[Jamaludin and Yusop, 2016] Jamaludin, S. and Yusop, Z. (2016). Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology*.

[James et al., 2009] James, G. M., Wang, J., Zhu, J., et al. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

[James, 2002] James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*.

[James, 2017] James, L. F. (2017). Bayesian poisson calculus for latent feature modeling via generalized indian buffet process priors. *Ann. Statist.*, 45(5):2016–2045.

[James et al., 2005] James, L. F. et al. (2005). Bayesian poisson process partition calculus with an application to bayesian lévy moving averages. *The Annals of Statistics*, 33(4):1771–1799.

[Jiang et al., 2011] Jiang, C.-R., Wang, J.-L., et al. (2011). Functional single index models for longitudinal data. *The Annals of Statistics*, 39(1):362–388.

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

[Kingman, 1992] Kingman, J. (1992). *Poisson Processes*, volume 3. Clarendon Press.

[Kingman, 1978] Kingman, J. F. (1978). The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380.

[Knapik and Salomond, 2014] Knapik, B. and Salomond, J.-B. (2014). A general approach to posterior contraction in nonparametric inverse problems. *arXiv preprint arXiv:1407.0335*.

[Knapik et al., 2016] Knapik, B. T., Szabó, B., Van Der Vaart, A. W., and Van Zanten, J. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3-4):771–813.

[Korwar et al., 1973] Korwar, R. M., Hollander, M., et al. (1973). Contributions to the theory of dirichlet processes. *The Annals of Probability*, 1(4):705–711.

[Laurini, 2014] Laurini, M. P. (2014). Dynamic functional data analysis with non-parametric state space models. *Journal of Applied Statistics*, 41(1):142–163.

[Lian et al., 2016] Lian, H., Choi, T., Meng, J., and Jo, S. (2016). Posterior convergence for bayesian functional linear regression. *Journal of Multivariate Analysis*, 150:27–41.

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

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

[Lijoi et al., 2010] Lijoi, A., Prünster, I., et al. (2010). Models beyond the dirichlet process. *Bayesian nonparametrics*, 28(80):342.

[Lila et al., 2016] Lila, E., Aston, J. A., Sangalli, L. M., et al. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879.

[Ma, 2016] Ma, S. (2016). Estimation and inference in functional single-index models. *Annals of the Institute of Statistical Mathematics*, 68(1):181–208.

[Meister, 2011] Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.*, 39(3):1471–1495.

[Miller et al., 2015] Miller, J., Betancourt, B., Zaidi, A., Wallach, H., and Steorts, R. C. (2015). Microclustering: When the cluster sizes grow sublinearly with the size of the data set. *arXiv preprint arXiv:1512.00792*.

[Müller et al., 2005] Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *the Annals of Statistics*, 33(2):774–805.

[Naulet and Barat, 2018] Naulet, Z. and Barat, E. (2018). Some aspects of symmetric gamma process mixtures. *Bayesian Anal.*, 13(3):703–720.

[Naulet and Rousseau, 2017] Naulet, Z. and Rousseau, J. (2017). Posterior concentration rates for mixtures of normals in random design regression. *Electron. J. Statist.*, 11(2):4065–4102.

[Neal, 2000] Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

[Pitman, 1995] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158.

[Pitman, 2003] Pitman, J. (2003). Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34.

[Ramsay, 2005] Ramsay, J. (2005). Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*.

[Rasmussen, 2000] Rasmussen, C. E. (2000). The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560.

[Ray, 2013] Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549.

[Richardson and Green, 1997] Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

[Robert and Soubiran, 1993] Robert, C. P. and Soubiran, C. (1993). Estimation of a normal mixture model through gibbs sampling and prior feedback. *Test*, 2(1-2):125–146.

[Rousseau and Szabo, 2016] Rousseau, J. and Szabo, B. (2016). Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors. *arXiv preprint arXiv:1609.05067*.

[Roy et al., 2017] Roy, A., Ghosal, S., Choudhury, K. R., et al. (2017). High dimensional single index bayesian modeling of the brain atrophy over time. *arXiv preprint arXiv:1712.06743*.

[Schwartz, 1965] Schwartz, L. (1965). On bayes procedures. *journal for probability theory and related fields*, 4(1):10–26.

[Szabó et al., 2013] Szabó, B., van der Vaart, A., van Zanten, J., et al. (2013). Empirical bayes scaling of gaussian priors in the white noise model. *Electronic Journal of Statistics*, 7:991–1018.

[Teh and Gorur, 2009] Teh, Y. W. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In *Advances in neural information processing systems*, pages 1838–1846.

[Thibaux and Jordan, 2007] Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571.

[van der Vaart and van Zanten, 2008] van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *Ann. Statist.*, 36(3):1435–1463.

[Williamson et al., 2010] Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010). The ibp compound dirichlet process and its application to focused topic modeling.

[Wolpert et al., 2011] Wolpert, R. L., Clyde, M. A., and Tu, C. (2011). Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *Ann. Statist.*, 39(4):1916–1962.

[Zhou, 2014] Zhou, M. (2014). Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling. In *Advances in Neural Information Processing Systems*, pages 3455–3463.