THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Modèles de mélange bayésiens non-paramétriques et clustering

Bayesian nonparametric mixture models and clustering

Présentée par :

Louise ALAMICHEL

Direction de thèse :

Julyan ARBEL CHARGE DE RECHERCHE HDR, CENTRE INRIA UNIVERSITE GRENOBLE ALPES	Directeur de thèse
Guillaume KON KAM KING CR, INRAE	Co-encadrant de thèse

Rapporteurs :

SYLVIA FRÜHWIRTH-SCHNATTER
FULL PROFESSOR, WIRTSCHAFTSUNIVERSITÄT WIEN
PIERPAOLO DE BLASI
FULL PROFESSOR, UNIVERSITA DEGLI STUDI DI TORINO

Thèse soutenue publiquement le **17 septembre 2024**, devant le jury composé de :

JEAN-FRANÇOIS COEURJOLLY, PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Président
JULYAN ARBEL, CHARGE DE RECHERCHE HDR, CENTRE INRIA UNIVERSITE GRENOBLE ALPES	Directeur de thèse
SYLVIA FRÜHWIRTH-SCHNATTER, FULL PROFESSOR, WIRTSCHAFTSUNIVERSITÄT WIEN	Rapporteure
PIERPAOLO DE BLASI, FULL PROFESSOR, UNIVERSITA DEGLI STUDI DI TORINO	Rapporteur
TOMMASO RIGON, ASSISTANT PROFESSOR, UNIVERSITA DEGLI STUDI DI MILANO-BICOCCA	Examinateur
FLORENCE FORBES, DIRECTRICE DE RECHERCHE, CENTRE INRIA UNIVERSITE GRENOBLE ALPES	Examinatrice

Invités :

GUILLAUME KON KAM KING
CHARGE DE RECHERCHE, CENTRE INRAE ILE-DE-FRANCE - JOUY-EN-JOSAS - ANTONY

Remerciements

Cette thèse n'aurait pas été la même sans la présence de nombreuses personnes, je profite donc ici de l'occasion qui m'est donnée pour les remercier.

En premier lieu et logiquement, mes remerciements vont à Julyan et Guillaume. Merci de m'avoir accompagné durant un peu plus de trois années. Travailler avec vous a toujours été un plaisir, merci aussi pour les moments moins sérieux qu'on a pu partager. Je pense que tout cela a grandement contribué à me laisser de ces trois années de thèse un très bon souvenir.

Now, I will switch to English to thank the members of my jury. I would like to express my gratitude to my thesis reviewers, Pierpaolo De Blasi and Sylvia Frühwirth-Schnatter, for agreeing to review my work. I would also like to thank Jean-François Coeurjolly, Florence Forbes, and Tommaso Rigon for agreeing to be part of my jury. A special thanks to Tommaso for the insightful discussions and feedback throughout my thesis.

Before returning to French, I would like to thank the other people I worked with: Caroline and Dasha. Working with you has been great, and I hope you have enjoyed it too. I would especially like to thank Dasha, who welcomed me on my first day at Inria and was there whenever I needed help or support. Thank you very much for that and for all the discussions and time we shared.

Maintenant, je voudrais remercier l'équipe Statify (je n'ai pas trouvé d'adjectif suffisant pour la qualifier). D'abord merci à tous les permanents d'avoir su créer et maintenir cette ambiance : en premier la cheffe Florence, mais aussi Stéphane, Julyan, Sophie, Pedro, Jonathan et Julien. Merci à ceux avec qui j'ai partagé le bureau G205, aussi connu sous le nom du bureau pause, pendant de plus ou moins long moments : Dasha, Masha, Pierre-Louis, Julien, Matthieu, Kevin-Lâm, Jhouben, Paul et Kostas. Merci à tous les autres doctorants : Alexandre, Benoît, Meryem, Hana, Lucrezia, Minh Tri, Théo, Yuchen, Jacopo, Geoffroy, Razan ; post-docs : Argheesh, Cambyse, Henrique, Paul-Gauthier, Pierre, Tâm, Tin, Yiye ; et stagiaires que j'ai côtoyés pendant mon passage dans l'équipe. Merci à tous pour en vrac les randonnées, les soirées et restos, les discussions, l'escalade, le baby-foot, le five, les soirées jeux de sociétés, Mario kart et tout le reste. En bref, merci d'avoir fait que je suis triste de quitter Grenoble.

J'ai eu l'occasion de faire plusieurs visites à l'INRAe de Jouy-en-Josas. Merci à ceux du 233 de m'avoir toujours si bien accueillie. J'ai notamment pu profiter de deux magnifiques chasses aux oeufs.

Switching back to English, during my PhD, I had the opportunity to spend two months at Duke University. It was a great experience and I would like to thank Filippo for making it possible. I enjoyed the time I spent working with you. I would also like to thank all the PhD students I met there and special thanks to Amy Herring for the bike. I would especially like to thank the wonderful office that welcomed me: Federica, Joe and Youngsoo; but also by extension Davide and Raphaël. Thank you for welcoming me to the climbing team, and for all our Saturday evening meals. My only regret about this stay is that it only lasted two months!

J'ai aussi pu assister à plusieurs conférences. Cela a été à chaque fois l'occasion de belles rencontres, merci de me rappeler à chaque fois à quel point la communauté bayésienne (non paramétrique en particulier) est sympathique. En particulier merci à Antoine et Emma pour ce long mais mémorable trajet calanques CIRM !

On arrive à la fin de mes remerciements mais ce sont sans doute les plus importants. Je tiens à remercier tous mes amis (ils se reconnaîtront). Merci pour les visites à Grenoble, les bivouacs, les vacances, les coups de tête et surtout de me supporter ! Merci aussi à mes co-chefs qui ont partagés mes aventures scoutes pendant cette période, elles m'étaient souvent nécessaires bien que fatigantes, en particulier merci à Nathalie, Seb et Alice. Pour finir, je voudrais remercier ma famille, mes parents et mes frère et sœurs : Claire, Émile et Bénédicte. Je n'ai pas les mots pour qualifier ce que vous m'apportez, juste merci d'être là pour moi. Je glisse aussi un remerciement pour tous les moments partagés à ma famille plus large : mes grands-parents, Anita, mes oncles, mes tantes (avec un remerciement particulier à Domi pour toutes ses visites à Grenoble) ainsi que tous mes cousins et cousines.

Je conclus sur ces mots soufflés par un sage :

"Un remerciement sincère vaut mieux que mille mots, donc merci !"
KL.

Abstract

Abstract

Model-based clustering is a complex problem, addressed for instance using mixture models. In this thesis, we focus on Bayesian nonparametric mixture models. These models are well-known for being consistent when used for density estimation. However, the consistency of the posterior distribution does not provide asymptotic guarantees in the context of clustering problems.

In the first two contributions, we study the consistency of the number of clusters using Bayesian nonparametric mixture models applied to finite mixtures. Results demonstrate posterior inconsistency for the number of clusters in this framework for specific nonparametric priors, such as the Dirichlet process and the Pitman-Yor process. We prove that those results apply to a general class of Bayesian nonparametric priors, the Gibbs-type processes, and some finite-dimensional representations thereof. Next, we discuss possible solutions proposed in the literature and show the application of these solutions to some of the studied priors. Second, we focus on a particular Gibbs-type process, the Pitman–Yor process with a hyperprior on its concentration parameter. Although placing a prior on the concentration hyperparameter, particularly in Dirichlet process mixture models, has been a common strategy to address the inconsistency issue, we provide a rigorous proof that Pitman–Yor process mixture models still suffer from inconsistency in the number of clusters in this framework.

In the final contribution, we apply these models to a real-world problem in ecotoxicology. We propose a Bayesian nonparametric mixture model to assess the ecological risks of water contaminants. The choice of a Bayesian nonparametric approach offers several advantages, including its efficiency in handling small datasets typical of environmental risk assessments, its ability to provide uncertainty quantification, and its capacity for simultaneous density and clustering estimation. We utilize a specific nonparametric prior from the class of normalized random measures with independent increments as the mixing measure, chosen for its robust clustering properties. Following the theoretical results from the first part of the thesis, we do not consider the raw posterior distribution on the number of clusters but follow a decision-theoretic framework to estimate data clustering.

Résumé

La classification, ou clustering, des données est un problème complexe, souvent traité à l'aide de modèles de mélange. Dans cette thèse, nous nous concentrons sur les modèles de mélange bayésiens non paramétriques. Ces modèles sont bien connus pour être consistants lorsqu'ils sont utilisés pour l'estimation de densité. Cependant, la consistance de la distribution a posteriori ne garantit pas asymptotiquement la résolution des problèmes de classification.

Dans les deux premières contributions, nous étudions la consistance du nombre de clusters en utilisant des modèles de mélange bayésiens non paramétriques appliqués à des mélanges finis. Premièrement, nous prouvons que des résultats d'inconsistance s'appliquent à une classe générale de priors bayésiens non paramétriques, les processus de type Gibbs, et à certaines de leurs représentations de dimension finie. Ensuite, nous discutons des solutions possibles proposées dans la littérature et montrons l'application de ces solutions à certains des priors étudiés. Deuxièmement, nous nous concentrons sur un processus de type Gibbs particulier, le processus de Pitman–Yor avec un hyperprior sur son paramètre de concentration. Bien que la mise en place d'un prior sur le paramètre de concentration, notamment dans les modèles de mélange de processus de Dirichlet, soit une stratégie courante pour résoudre le problème d'inconsistance, nous montrons que le nombre de clusters avec un modèle de mélange de processus de Pitman–Yor est encore inconsistant dans ce cas.

Dans la dernière contribution, nous appliquons ces modèles à un problème réel en écotoxicologie. Nous proposons un modèle de mélange bayésien non paramétrique pour évaluer les risques écologiques de contaminants de l'eau. Le choix d'une approche bayésienne non paramétrique offre plusieurs avantages, notamment son efficacité à gérer de petits ensembles de données typiques de l'évaluation des risques environnementaux, sa capacité à fournir une quantification de l'incertitude, ainsi qu'une estimation simultanée de la densité et du clustering. Nous utilisons un prior non paramétrique spécifique de la classe des mesures aléatoires normalisées à incréments indépendants comme mesure de mélange, choisi pour ses propriétés robustes en matière de classification. À cause des résultats théoriques de la première partie de la thèse, nous ne considérons pas la distribution a posteriori du nombre de clusters mais suivons un cadre décisionnel pour estimer le clustering des données.

Contents

List of Figures

List of Tables

List of Acronyms

$\mathcal{VI}$ variation of information.

EC$_x$ Effect Concentration at x%.

EC$_{50}$ Effect Concentration 50%.

HC$_5$ Hazardous Concentration for 5% of the Species.

LC$_{50}$ Lethal Concentration 50%.

i.i.d. identically and independently distributed.

AIC Akaike information criterion.

BIC Bayes information criterion.

BNP Bayesian nonparametric.

CEC Critical Effect Concentration.

CPO conditional predictive ordinate.

CRM completely random measures.

DMP Dirichlet multinomial process.

DP Dirichlet process.

DPM Dirichlet process mixtures.

EPPF exchangeable partition probability function.

IG Inverse-Gamma.

KDE Kernel Density Estimate.

LOO Leave-One-Out.

MAP Maximum a posteriori.

MCMC Markov chain Monte Carlo.

MFM mixture of finite mixture.

MTM Merge-Truncate-Merge.

NGG normalized generalized Gamma process.

NGGM normalized generalized Gamma multinomial process.

NIDM normalized infinitely divisible multinomial process.

NIG normalized inverse Gaussian process.

NRMI normalized random measures with independent increments.

NS normalized stable process.

NTF Nonnegative Tensor Factorization.

PY Pitman–Yor process.

PYM Pitman–Yor multinomial process.

RIVM National Institute for Public Health and the Environment.

SLC Sodium-Lithium Countertranspor.

SSD Species Sensitivity Distribution.

Chapter 1

Introduction

Contents

1.1 The Bayesian non-parametric framework

The origins of Bayesian statistics date back to the 18th century with the introduction of Bayes' rule or Bayes' theorem in 1763 by Thomas Bayes. This result was later generalized by Laplace, who also contributed to some computational aspects of Bayesian statistics. This approach to statistics was later overlooked in favor of frequentist statistics, partly because of the need for numerical calculation to compute the quantities of interest to Bayesian statistics. In the middle of the 20th century, significant advancements in Bayesian computation (Metropolis et al. 1953; Hastings 1970) as well as in Bayesian theory (Finetti 1937; Jeffreys 1939; Savage 1954) contributed to a rise in interest in this area.

Bayesian statistics is often described in terms of its interpretation of probability. Roughly speaking, it aims to give meaning to the notion of subjectivity within statistics. More precisely, the probability of an event is the degree of belief in this very event. Assume that the data X belongs to a general space $\mathcal{X}$ and follows a distribution parametrized by θ in the space of parameters Θ. The way to express this belief is to consider the parameters as random variables sampled from a so-called *prior* distribution, which characterizes this belief. Then, a Bayesian statistical model may be written as

$$X \mid \theta \sim p(x \mid \theta), \quad \theta \sim p(\theta),$$

where $p(\theta)$ is the prior, which is a probability measure on the parameter set Θ, and $p(x \mid \theta)$ is the *likelihood*. Using Bayes' theorem, one may define a *posterior* distribution as follows

$$p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{p(x)}.$$

Here, $p(x)$ is called the marginal distribution of X and acts as a normalizing constant. The posterior distribution is a conditional distribution on the sample space $\mathcal{X}$, it represents the conditional probability of θ given the observed data X. and reflects the updated knowledge of θ after incorporating the data. The analysis of this posterior distribution is the aim of Bayesian statistics.

1.1.1 Bayesian nonparametric inference

One of the main problems in Bayesian statistics is the choice of the prior. The prior needs to summarize prior knowledge and accommodate all uncertainties. A natural way to provide such flexibility on the prior is to consider a so-called Bayesian nonparametric (BNP) framework where the model may involve an arbitrary number of parameters, possibly in infinite dimensions. This allows the number of parameters to adapt to the complexity of the data.

However, defining a prior is more complex and requires more care in a nonparametric framework. In this case, the prior is a probability measure on an infinite dimensional space, hence challenging to define. A way to construct such a prior is to use de Finetti's representation of infinite exchangeable sequences (Finetti 1937; Hewitt and Savage 1955). The infinite sequence $(X_1, \ldots, X_n, \ldots)$ is exchangeable if for any finite permutation σ of $\{1, \ldots, n\}$, $n \geq 1$, $(X_1, \ldots, X_n) \overset{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ with the equality being understood in distribution. More precisely, de Finetti's Theorem states that an infinite sequence $(X_1, \ldots, X_n, \ldots)$ is exchangeable if and only if it is a mixture ofsequences of identically and independently distributed (*i.i.d.*) random variables (see Hjort et al. 2010, for more details). Consequently, from this theorems it follows that $(X_1, \ldots, X_n, \ldots)$ is exchangeable if there exists a probability measure $\mathcal{Q}$ such that

$$X_i \mid G \overset{\text{iid}}{\sim} G, \quad i \in \mathbb{N}^{\star}, \quad G \sim \mathcal{Q}, \tag{1.1}$$

where $\mathcal{Q}$ is the distribution of a random measure G. The distribution $\mathcal{Q}$ can be interpreted as a prior distribution. We recover a parametric setting if $\mathcal{Q}$ degenerates on a finite-dimensional space. For example, $\mathcal{Q}\left(\{G : G(\mathrm{d}x) = \mathcal{N}(\mu, \sigma^2)\mathrm{d}x,\ (\mu, \sigma) \in \mathbb{R}^2\}\right) = 1$ corresponds to the model with a Gaussian likelihood and a prior on the location and scale parameters (μ, σ). Conversely, if $\mathcal{Q}$ is supported on an infinite dimensional space, then $\mathcal{Q}$ is a nonparametric prior.

As mentioned previously, a Bayesian nonparametric model has at least one infinitely dimensional parameter. Typically, this parameter could be a function or a probability measure, for which different kinds of priors are considered. Gaussian processes (see e.g. Williams and Rasmussen 2006) are random functions and a common prior for function space. Another prior for continuous functions is the Pólya tree (Lavine 1992). On the other side, different stochastic processes based on de Finetti's representation theorem are commonly used as prior on probability measures. In both cases, the prior is the law governing the stochastic process. The most famous prior in BNP is the Dirichlet process introduced in Ferguson (1973), more details on this particular prior and generalizations are provided in the following section.

The BNP framework proposes some flexible models, typically used in density estimation, where the unknown distributions are sampled from a prior (Lo 1984). BNP models are also used in model-based clustering (see Section 1.3) or latent factor models (Ghahramani and Griffiths 2005). The theory behind BNP is complex, especially compared to the parametric framework. However, these models perform well in real applications, such as ecology (Zito et al. 2022), genetics (Masoero et al. 2022), and medicine (Albughdadi et al. 2017), among many others.

The Bayesian nonparametric framework is relatively recent, with significant development occurring after the introduction of the Dirichlet process (Ferguson 1973), as well as advancements in computational methods. The community has nonetheless provided useful overall reviews on the field. The interested reader may look at Hjort et al. (2010); Ghosal and van der Vaart (2017).

In what follows, we introduce mixture models which describe heterogeneous data. We focus in particular on BNP mixture models using priors defined below in Section 1.1.2. Then, we consider these mixture models to face clustering problems. We describe model-based clustering with BNP mixture models and a Bayesian decision-theoretic framework to estimate data clustering. Once the framework is established, one may derive asymptotic results on consistency. We finally present such consistency results for BNP mixture models.

1.1.2 Bayesian nonparametric priors

As introduced previously, several different priors based on stochastic processes exist within Bayesian nonparametrics. In this section, we present some priors based on the exchangeability assumption. The class of priors we focus on are called *species sampling processes* (Ghosal and van der Vaart 2017, see Chapter 14) and is constituted of discrete random probability measures.

Before describing the different priors, we introduce some useful notions. A *partition* A of $\{1, \dots, n\}$ in k elements, is a decomposition in k non-empty and disjoints sets of $\{1, \dots, n\}$: $A = (A_1, \dots, A_k)$, with $\cup_{i=1}^k A_i = \{1, \dots, n\}$ and for $i, j \in \{1, \dots, k\}$ such that $i \neq j$, $A_i \cap A_j = \emptyset$. We denote by $n_i = |A_i|$ the cardinality of each set in the partition so that by definition $\sum_{i=1}^k n_i = n$.

Definition 1.1 (Exchangeable partition Ghosal and van der Vaart 2017)**.** *A random partition $\mathbf{A}_n$ of $\{1, \dots, n\}$ is called* exchangeable *if its distribution is invariant under the action of any permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$. In other words, for all partitions $A = (A_1, \dots, A_k)$ of $\{1, \dots, n\}$ the probability $\mathrm{P}\left(\mathbf{A}_n = (\sigma(A_1), \dots, \sigma(A_k)\right)$ is the same for any σ. Equivalently, $\mathbf{A}_n$ is* exchangeable *if there exist a symmetric function p, such that for all partitions $A = (A_1, \dots, A_k)$ of $\{1, \dots, n\}$,*

$$\mathrm{P}\left(\mathbf{A}_n = (\sigma(A_1), \dots, \sigma(A_k))\right) = p(n_1, \dots, n_k). \tag{1.2}$$

The function p is called the exchangeable partition probability function (EPPF) *of $\mathbf{A}_n$.*

With this definition, we can define an *infinite exchangeable random partition* as a sequence of exchangeable random partitions of $\{1, \dots, n\}$, $(\mathbf{A}_n)_{n \in \mathbb{N}}$, such that $\mathbf{A}_{n-1}$ is equal to the partition obtained from $\mathbf{A}_n$ by leaving out the element n. An equivalent EPPF can be defined for the infinite exchangeable random partition (see e.g. Pitman 1995).

In what follows, we describe different nonparametric priors, which are examples of $\mathcal{Q}$ in Equation (1.1). The relation between all these priors and where each is used in the thesis is summarized at the end of the section in Figure 1.3.

Dirichlet process (DP). The Dirichlet process, introduced in Ferguson (1973), is arguably the most used BNP prior. A Dirichlet process is parameterized by a concentration parameter $\alpha > 0$ and a base measure H. A random measure G on the sample space $\mathcal{X}$ has a Dirichlet process distribution if for every finite measurable partition $A = (A_1, \dots, A_k)$ of $\mathcal{X}$

$$(G(A_1), \dots, G(A_k)) \sim \mathrm{Dir}(\alpha H(A_1), \dots, \alpha H(A_k)),$$

and is denoted by $G \sim \mathrm{DP}(\alpha, H)$. Dir denotes the Dirichlet distribution. The random measure G defined in this way turns out to be discrete and can be written

$$G = \sum_{j>1} w_j \delta_{\theta_j},$$

where $\theta_1, \dots, \theta_k, \dots$ is a sequence of random variables such that $\theta_j \overset{iid}{\sim} H$, δ_{θ_j} represents a discrete measure concentrated at θ_j, and $w_1, \dots, w_k, \dots$ is a sequence of random variables representing the weights, such that $\sum_{j>1} w_j = 1$ almost surely.

The construction of these weights follows the stick-breaking representation introduced in Sethuraman (1994). The idea is that considering a stick of unit size, one sequentially breaks the stick into two parts with random sizes. The size of these parts, or equivalently the breaking location, follows a Beta distribution. Figure 1.1 is an illustration of this procedure. Following this representation, a discrete random measure G with a weight distribution defined by $w_1 = v_1$ and $w_j = v_j \prod_{i<j}(1 - v_i)$ where $v_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ and $\theta_j \overset{\text{iid}}{\sim} H$ is such that $G \sim \mathrm{DP}(\alpha, H)$.

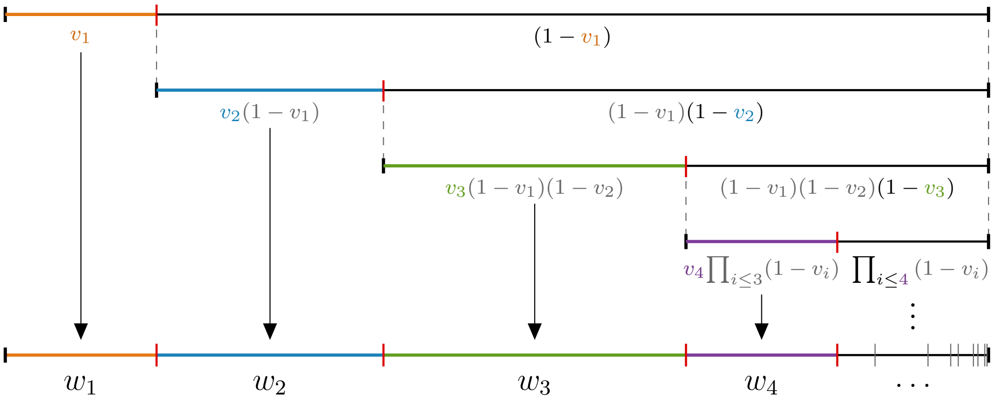


Figure 1.1: Illustration of the stick-breaking representation.

The Dirichlet process distribution can also be constructed through the distribution of the induced partition. As the Dirichlet process is a discrete random measure G, a finite sample $X_{1:n} = (X_1, \dots, X_n)$ from G would have some ties meaning that it would have $K_n \le n$ distinct values: $X_1^\star, \dots, X_{K_n}^\star$, each with a respective frequency: $n_1, \dots, n_{K_n}$ such that $\sum_{j=1}^{K_n} n_i = n$. A partition of $\{1, \dots, n\}$ with K_n clusters, where the clusters are defined by the equivalence relation $i \sim j$ if and only if $X_i = X_j$, is induced by G. The infinite exchangeable random partition generated by a sample of the Dirichlet process $\mathrm{DP}(\alpha, H)$ is called the Chinese Restaurant Process. The EPPF associated is of the form

$$p(n_1, \dots, n_k) = \frac{\alpha^k}{(\alpha + 1)_{n-1}} \prod_{i=1}^{k} (n_i - 1)!,$$

where $(x)_n = x(x + 1) \cdots (x + n - 1)$ is the ascending factorial and $(x)_0 = 1$ by convention. The metaphor associated with the Chinese restaurant process goes as follows: suppose a Chinese restaurant possesses an infinite number of tables, each with infinite seating, and customers arrive sequentially. The first customer sits at an arbitrary empty table with a probability of 1. Then, the second customer can sit at the same table or open a new one. The process repeats for all the customers who can choose between opening a new table or sitting at an already-opened one. The probability of sitting at different tables describes the Dirichlet process distribution. The second customer chooses with probability $1/(\alpha+1)$ the table opened by the first customer and with probability $\alpha/(\alpha+1)$ a new table. More generally, the $(n+1)$th customer finds n customers already seated at k different tables on $n_1,\ldots,n_k$ groups. Then, this customer chooses with probability $n_j/(\alpha+n)$ the table j, and with probability $\alpha/(\alpha+n)$ a new table. Figure 1.2 illustrates this procedure.

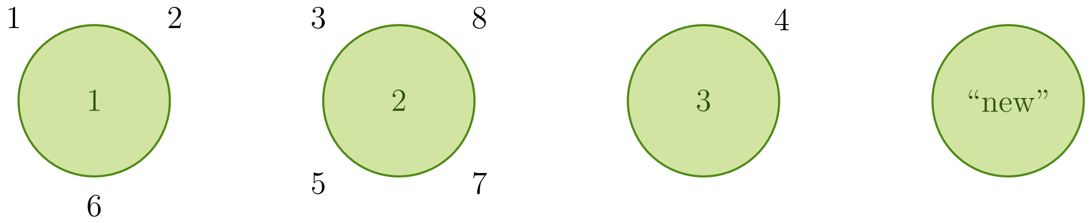


Figure 1.2: Illustration of the Chinese Restaurant Process. In this example, the 9^{th} customer chooses to sit at tables from left to right with probability $3/(\alpha+8)$, $4/(\alpha+8)$, $1/(\alpha+8)$, and $\alpha/(\alpha+8)$.

Pitman–Yor process (PY). Many extensions exist for the Dirichlet process, the most famous being the Pitman–Yor process introduced in Perman et al. (1992) and further investigated in Pitman and Yor (1997). This process is also known as the two-parameter Poisson–Dirichlet process. It is a natural extension of the Dirichlet process parametrized by an extra parameter, increasing its flexibility. The Pitman–Yor process is parametrized by a precision parameter α, a discount parameter σ, and a base measure H. The parameters α and σ are such that $\sigma<0$ and $\alpha\in\{-2\sigma,-3\sigma,\ldots\}$, or $\sigma\in[0,1)$ and $\alpha>-\sigma$.

Similarly to the Dirichlet process, the Pitman–Yor process can be defined through the distribution of the underlying partition. A Pitman–Yor partition is an infinite exchangeable partition with EPPF given by

$$p(n_1,\ldots,n_k)=\frac{\prod_{j=1}^{k-1}(\alpha+j\sigma)}{(\alpha+1)_{n-1}}\prod_{i=1}^{k}(1-\sigma)_{n_i-1}.$$

It is also possible to use the stick-breaking representation: a discrete random measure G has a Pitman–Yor process distribution if the weights are distributed such that $w_1=v_1$ and $w_j=v_j\prod_{i<j}(1-v_i)$ where $v_j\overset{\text{iid}}{\sim}\text{Beta}(1-\sigma,\alpha+j\sigma)$, and $\theta_j\overset{\text{iid}}{\sim}H$.

We denote $G \sim \mathrm{PY}(\alpha, \sigma, H)$.

Note that if $\sigma = 0$ then $G \sim \mathrm{PY}(\alpha, 0, H)$ is equivalent to $G \sim \mathrm{DP}(\alpha, H)$.

Gibbs-type process. Another natural extension of the Dirichlet process is the Gibbs-type process class introduced in Gnedin and Pitman (2006) (see e.g. Pitman 2006; De Blasi et al. 2015, for more details). This class is more general than the Pitman–Yor process class. In particular, the DP and PY are subclasses of the Gibbs-type process class.

Gibbs-type processes are parameterized by $\sigma \in (-\infty, 1)$, which determines the type of the process. In particular, the Gibbs-type process with $\sigma = 0$ is the Dirichlet process.

The common way to define Gibbs-type processes is through random partition. A *Gibbs partition* is an infinite exchangeable random partition with EPPF of the following form

$$p(n_1, \dots, n_k) = V_{n,k} \prod_{i=1}^{k} (1 - \sigma)_{n_i - 1}, \tag{1.3}$$

where the $V_{n,k}$ are nonnegative numbers for $n \in \mathbb{N}^\star$, $k \in \{1, \dots, n\}$, satisfying the recurrence relation $V_{1,1} = 1$ and $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$. The Gibbs-type process is the process characterized by the distribution of the Gibbs partition. Then, Gibbs-type processes are characterized by their EPPF and hence by the $V_{n,k}$ numbers. For example, for a PY the $V_{n,k}$ numbers are as follow

$$V_{n,k} = \frac{\prod_{j=1}^{k-1}(\alpha + j\sigma)}{(\alpha + 1)_{n-1}},$$

and for a DP, $V_{n,k} = \alpha^k/(\alpha)_n$.

Another way to characterize species sampling processes is through the predictive distribution closely related to the EPPF. The predictive distribution is the distribution of a new observation, knowing the previous ones. Given observations $X_{1:n} = (X_1, \dots, X_n)$ sampled from $X_i \mid G \overset{\text{iid}}{\sim} G,\ G \sim \mathcal{Q}$, the predictive distribution is the following,

$$\mathrm{P}(X_{n+1} = x_{n+1} \mid X_{1:n}) = \int p(x_{n+1})\, \mathcal{Q}(\mathrm{d}p \mid X_{1:n}),$$

where $\mathcal{Q}(\cdot \mid X_{1:n})$ denotes the posterior distribution of G. For example, in the DP case, the predictive distribution is

$$\mathrm{P}(X_{n+1} \mid X_{1:n}) = \frac{\alpha}{\alpha + n} H + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}.$$

This quantity and the idea of prediction are used in the Chinese restaurant process.

In the PY case, the predictive distribution when $X_{1:n}$ are grouped in K_n clusters is

$$\mathrm{P}(X_{n+1} \mid X_{1:n}) = \frac{\alpha + \sigma K_n}{\alpha + n} H + \frac{1}{\alpha + n} \sum_{i=1}^{n} (n_j - \sigma) \delta_{X_i}.$$

De Blasi et al. (2015) show that the predictive distribution of a Gibbs-type process depends only on the sample size n and the number of distinct values K_n. This is even an equivalence; if the predictive distribution of a species sampling process only depends on K_n and n, then it is a Gibbs-type process. The predictive distribution of a Gibbs-type prior is the following

$$\mathrm{P}(X_{n+1} \mid X_{1:n}) = \frac{V_{n+1,k+1}}{V_{n,k}} H + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{n} (n_j - \sigma) \delta_{X_i}.$$

We have already introduced two special cases of Gibbs-type processes; however, those are not the only notable processes in this class. Some other examples are normalized generalized Gamma processes, normalized inverse Gaussian processes, and normalized-stable processes. Now, we give further details on these three processes.

Normalized generalized Gamma process (NGG). As said previously, the NGG (see for more details Lijoi et al. 2007), is a particular case of Gibbs-type processes for which $\sigma \in (0, 1)$. The NGG is characterized by the following $V_{n,k}$ numbers,

$$V_{n,k} = \frac{\mathrm{e}^{\beta} \sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right),$$

where $\beta > 0$ is another parameter, and $\Gamma(\cdot;\cdot)$ is the incomplete gamma function: $\Gamma(x; a) = \int_x^{\infty} s^{a-1} \mathrm{e}^{-s} \mathrm{d}s$. Let G be a random measure that follows the normalized generalized Gamma process distribution, we denote $G \sim \mathrm{NGG}(\beta, \sigma)$. NGG is a large class of priors; different interesting priors are part of the class. As a first example, if $\sigma \to 0$, then we recover the Dirichlet process.

If $\sigma = 1/2$, then we obtain the normalized inverse Gaussian process (NIG) (Lijoi et al. 2005a). Similarly to the Dirichlet process, the marginals of this process admit an explicit distribution. The NIG can then be defined in the same way as the DP. A random measure G on $\mathcal{X}$ follows a normalized inverse-Gaussian process with a base measure H if for every measurable partitions $A = (A_1, \ldots, A_k)$ of $\mathcal{X}$

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{NIG}\left(H(A_1), \ldots, H(A_k)\right),$$

where NIG stands for the finite-dimensional normalized inverse-Gaussian distribution.

Finally, if $\beta = 0$, we obtain the normalized stable process (NS). NS introduced in

Kingman (1975) is also a special case of the Pitman–Yor process, obtained by taking $\alpha = 0$ and $\sigma \in (0, 1)$. As the whole class of NGG, this process is also a normalized random measures with independent increments (NRMI) (Lijoi et al. 2008). This prior is used in Chapter 4. NGG is the only Gibbs-type process, which is also a particular case of NRMI, which we briefly describe now.

Normalized random measures with independent increments (NRMI). Before defining NRMI, we will introduce completely random measures (CRM) (Kingman 1967). A completely random measure on the space $\mathcal{X}$ is a random element $\tilde{\mu}$ such that for any measurable partition $(A_1, \dots, A_k)$ of $\mathcal{X}$, the random variables $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_k)$ are independent. CRM are (almost surely) discrete measures and can be represented as

$$\tilde{\mu} = \sum_{i \geq 1} J_i \delta_{Z_i},$$

where J_i are positive jumps and Z_i are locations, both are random and independent. A characterization of a CRM is given by its Laplace transform. The Laplace transform of a measure $\tilde{\mu}(A)$ for $A \in \mathcal{X}$ is defined for $\lambda > 0$ as,

$$L_A(\lambda) = \mathbb{E}[\mathrm{e}^{-\lambda \tilde{\mu}(A)}] = \exp\left[- \int_{\mathbb{R}_+ \times A} (1 - \mathrm{e}^{-\lambda s}) \nu(\mathrm{d}s, \mathrm{d}x) \right].$$

We only consider homogeneous CRM (see Chapter 3 in Hjort et al. 2010) which are characterized by a Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}x) := \rho(s)\mathrm{d}s\, \alpha(\mathrm{d}x)$ where α is a parameter measure on $\mathcal{X}$.

A normalized CRM $\tilde{\mu}$, $\tilde{\mu}/\tilde{\mu}(\mathcal{X})$, defines a normalized random measure with independent increments. NRMI are also characterized by their Lévy intensity. For example, the normalized stable process is a NRMI with Lévy intensity characterized by the following function,

$$\rho(s) = \frac{\sigma}{\Gamma(1 - \sigma) s^{1+\sigma}}.$$

For a NRMI with intensity measure ν, the Laplace exponent is defined (for $\lambda > 0$) by $\psi(\lambda) := -\log(L_{\mathcal{X}}(\lambda)) = \int_{\mathbb{R}_+ \times \mathcal{X}} (1 - \mathrm{e}^{-\lambda s}) \nu(\mathrm{d}s, \mathrm{d}x)$. Then, the EPPF for this NRMI is of the following form,

$$p(n_1, \dots, n_k) = \frac{(-1)^{n-k}}{\Gamma(n)} \int \lambda^{n-1} \mathrm{e}^{-\psi(\lambda)} \prod_{j=1}^{k} \psi^{(n_j)}(\lambda) \mathrm{d}\lambda,$$

where $\psi^{(n_j)}(\lambda) = \int_{\mathbb{R}_+ \times \mathcal{X}} s^{n_j} \mathrm{e}^{-\lambda s} \nu(\mathrm{d}s, \mathrm{d}x)$.
For more details on NRMI see e.g. Regazzini et al. (2003); James et al. (2009).

Figure 1.3 summarizes the relationships between the different priors introduced and outlines this thesis by indicating which priors are considered in each chapter.

All the Gibbs-type processes are introduced again in Section 2.2. As illustrated in Figure 1.3, the priors used in this thesis are closely related. However, due to differences in their definition, some of these priors are preferred in certain applications (Ayed et al. 2019). In addition, the more general these priors are, the more flexible they are, but also the less tractable and easy to use.

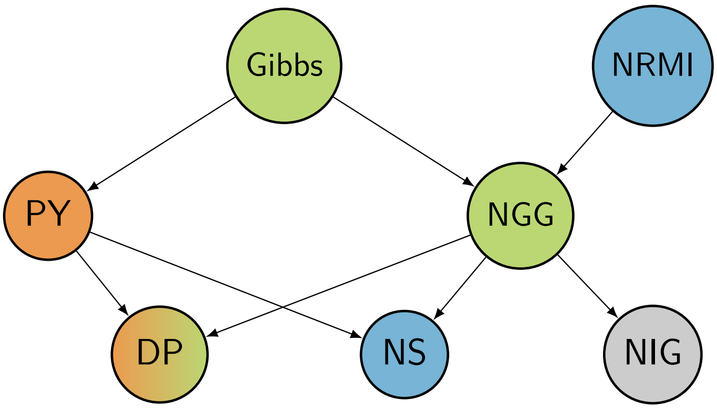


Figure 1.3: Graphical representation of the relationship between the different BNP priors. An arrow indicates that the target is a special case of the origin. In **green** are the priors considered in **Chapter 2**, in **orange** the ones in **Chapter 3** and in **blue** those in **Chapter 4**.

The processes introduced in this section are all associated with a random partition. Then, these processes are naturally used as priors in Bayesian nonparametric mixture models, where the induced infinite random partition yields clusters in the observations. This mechanism is reviewed in the next two sections.

1.2 Mixture models

Mixture models are useful tools for representing complex data, especially that coming from heterogeneous populations. They have been used for over 150 years in both Bayesian and frequentist statistics. We begin by presenting the classic formulation of a mixture model, then we adopt the Bayesian point of view and describe the problem of choosing the number of subpopulations in the data.

1.2.1 Basic formulation

We consider data such that each observation belongs to one subpopulation or group, and each group is characterized by a density, making it homogeneous within a given group. All the different groups are supposed to be unknown and are some latent parameters of the model. Hence, the membership of one observation to one of the different groups is also unknown. These groups are commonly called *components*.

More formally, an observation x comes from a K-components mixture distribution if it is drawn from a mixture density of the following form

$$X \sim f^X(\cdot) = \sum_{k=1}^{K} w_k f(\cdot \mid \theta_k), \tag{1.4}$$

where $w_{1:K} = (w_1, \dots, w_K)$ are positive weights such that $\sum_{k=1}^{K} w_k = 1$, and $f(\cdot \mid \theta_k)$ represents a component-specific kernel density parametrized by $\theta_{1:K} = (\theta_1, \dots, \theta_K)$. Statement (1.4) may equivalently be expressed as

$$X \sim \int f(\cdot \mid \theta) G(\mathrm{d}\theta), \tag{1.5}$$

where G is a discrete mixing measure $G := \sum_{k=1}^{K} w_k \delta_{\theta_k}$ with positive weights $w_{1:K}$ such that $\sum_{k=1}^{K} w_k = 1$ and atoms $\theta_{1:K}$.

Mixture models can also be defined using latent variables describing the component allocation for each observation. More precisely, let K be the number of components, the latter being in proportion w_k, $k = 1, \dots, K$, in the total population. We denote by $z_{1:n} = (z_1, \dots, z_n)$ the allocation variables associated to each data point $(x_1, \dots, x_n)$. The variables z_i are such that $z_i = k$ if x_i belongs to group k. The model is now described by

$$X_i \mid z_i \overset{\text{ind}}{\sim} f(\cdot \mid \theta_{z_i}), \text{ with } \mathrm{P}(z_i = k) = w_k, \ i = 1, \dots, n.$$

This representation is used in Chapter 4, where the allocation variables are inferred.

A commonly used density family for kernel densities is the Gaussian family, defining Gaussian mixture models, an example of which is given in Figure 1.4. This parametric family is the most frequently encountered throughout this thesis.

Because of their definition, mixture models are flexible and can handle a variety of data and problems. Mixture models typically address density estimation (Escobar and West 1995; Ferguson 1983). They can also perform model-based clustering (Fraley and Raftery 2002), which is the main focus of this thesis. They are used in many applications for example in healthcare (Ramírez et al. 2019; Ullah and Mengersen 2019), image analysis (Gerogiannis et al. 2009), ecology (Attorre et al. 2020), econometrics (Frühwirth-Schnatter et al. 2012), networks (Durante and Dunson 2018), genomics (Allison et al. 2002), and many others. For more details on mixture models, the interested reader could refer to the handbook Frühwirth-Schnatter et al. (2019).

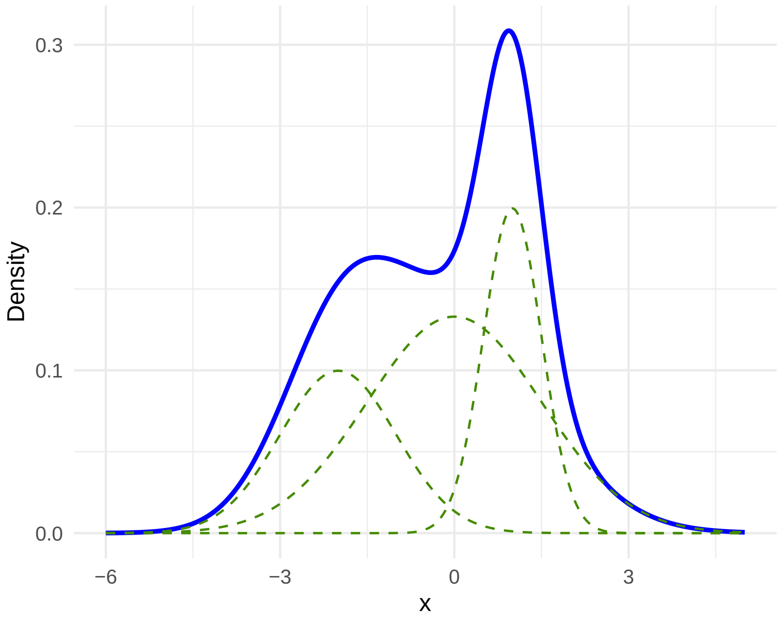


Figure 1.4: Example of a mixture of three univariate normal densities. The model is $^1\!/_4 \times \mathcal{N}(-2,1) + {}^1\!/_2 \times \mathcal{N}(0,1.5^2) + {}^1\!/_4 \times \mathcal{N}(2,0.5^2)$.

1.2.2 Bayesian mixture models

In the Bayesian setting, a common way to deal with mixture models is to consider the model described in Equation (1.5) and put a prior on the mixing measure G. The mixing measure G depends on the number of components K, the weights $w_{1:K}$, and the locations $\theta_{1:K}$. An important and difficult task is to choose the number of components K beforehand. From a frequentist point of view, this problem is usually treated as a model selection problem, using criteria such as the Bayesian Information Criteria (BIC) or the Integrated Completed Likelihood (ICL). In our Bayesian framework, the number of components K can be considered as finite, infinite, or random. In this section, we provide some insights into these three points of view.

1.2.2.1 Finite mixture models

When the number of components K is considered fixed and finite, we deal with finite mixture models where K could be known or unknown. In both cases, in a Bayesian mixture model, the parameters of the model, the mixing distribution, and the component parameters are drawn from a prior distribution. We denote K_0 the real number of components of the mixture distribution from which the data is sampled. The idea of a true value for a parameter is linked with the study of frequentist asymptotic properties for statistical models. We develop this perspective in the upcoming Section 1.4.

If K is known, so that $K = K_0$, we are in an exact-fitted setting. In this case, we only need a prior on $(w_{1:K}, \theta_{1:K})$. Commonly, we choose a prior such that under this prior, the θ_k are independently and identically distributed and independent from $w_{1:K}$. A typical prior for the weights $w_{1:K}$ is the Dirichlet distribution.

If K is unknown, one way to handle the choice of K is to choose K great enough to ensure $K \geq K_0$. In this case, we speak of *overfitted mixture* models. A classic prior for the weights is the Dirichlet distribution, $w_{1:K} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$. Rousseau and Mengersen (2011) study this prior and the properties of the associated overfitted mixture model. Taking $\alpha_1 = \ldots = \alpha_K = {}^{\alpha}/_{K}$ leads to the Dirichlet multinomial process. This process is a finite approximation of the Dirichlet process. In Chapter 2, we also consider different types of parametric priors, which we will describe now.

We introduce some finite-dimensional representations for the BNP priors described in Section 1.1.2 in the sense that $K < \infty$. Using them yields overfitted mixture models. Those finite-dimensional priors provide convenient and tractable models. Considering the limit $K \to \infty$, one retrieves their corresponding nonparametric priors. In the following, we give more details on the Dirichlet multinomial process and recently proposed finite-dimensional versions of the Pitman–Yor process and normalized random measures with independent increments (Lijoi et al. 2024; Lijoi et al. 2020).

Dirichlet multinomial process (DMP). The most commonly used example of such a finite-dimensional representation, introduced previously, is the Dirichlet multinomial distribution (see for instance Muliere and Secchi 1995; Ishwaran and Zarepour 2000; Ishwaran and Zarepour 2002). The DMP is parametrized by a concentration parameter $\alpha > 0$, the number of components K, and a base measure H. It is a random discrete measure $G = \sum_{k=1}^{K} w_k \delta_{\theta_k}$ characterized by a Dirichlet distribution on the weights with parameter ${}^{\alpha}/_{K}$: $w_{1:K} \sim \mathrm{Dir}({}^{\alpha}/_{K}, \ldots, {}^{\alpha}/_{K})$. The location parameters θ_k are distributed according to the base measure H.

Like the other finite-dimensional priors considered below, the DMP can also be defined hierarchically. The DMP is a discrete random probability measure G_K such that

$$G_K \mid G_{0,K} \sim \mathrm{DP}(\alpha; G_{0,K}), \quad G_{0,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\tilde{\theta}_k},$$

where $\tilde{\theta}_k \overset{\text{iid}}{\sim} H$.

The DMP with parameters α, K, and H approximates the Dirichlet process with parameters α and H and converges to it weakly, when $K \to \infty$ (Muliere and Secchi 2003).

Pitman–Yor multinomial process (PYM). The Pitman–Yor multinomial process introduced in Lijoi et al. (2020) is based on the Pitman–Yor process. The PYM is parametrized by a base measure H, parameters α, and σ, which have a similar role as in the PY case. The PYM is defined as a discrete random probability measure G_K such that

$$G_K \mid G_{0,K} \sim \mathrm{PY}(\sigma, \alpha; G_{0,K}), \quad G_{0,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\tilde{\theta}_k},$$

where $\tilde{\theta}_k \overset{\text{iid}}{\sim} H$. As for DMP, the marginal distribution of the PYM is available. Let G_K be a random measure sampled from a PYM, the weights of G_K have a ratio-stable distribution (Carlton 2002), $w_{1:K} = (w_1, \dots, w_K) \sim \mathrm{RS}(\sigma, \alpha; {}^1/_K, \dots, {}^1/_K)$.

The PYM generalises the Dirichlet multinomial process. As with the PY, the random probability measure p_K of the PYM reduces to the Dirichlet multinomial process when $\sigma = 0$. Lijoi et al. (2020) prove that PYM approximates PY, in the sense that PY is obtained as a limiting case when $K \to \infty$ (see Theorem 5 in Lijoi et al. 2020). In addition, PYM is more flexible than DMP (Lijoi et al. 2020).

Normalized infinitely divisible multinomial process (NIDM). NIDM processes are introduced by Lijoi et al. (2024) and can be seen as a finite approximation for NRMI. NIDM processes can be described as NRMI measures using a hierarchical structure

$$G_K \mid G_{0,K} \sim \mathrm{NRMI}(c, \rho; G_{0,K}), \quad G_{0,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\tilde{\theta}_k},$$

where $\tilde{\theta}_k \overset{\text{iid}}{\sim} H$, H is a base measure. In this expression, ρ is a function, defined in Section 1.1.2, that characterizes the NRMI process used. The choice $\rho(s) = s^{-1}e^{-s}$ corresponds to the Dirichlet process, and the NIDM process associated is the Dirichlet multinomial process. Similarly, choosing $\rho(s) = \frac{1}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-\beta s}$, $0 \leq \sigma < 1$ and $\beta \geq 0$ corresponds to NGG. We then get the normalized generalized Gamma multinomial process (NGGM).

Figure 1.5 summarizes the relationships between the different priors introduced. These priors are again introduced with details on their associated EPPF in Section 2.2.

1.2.2.2 Bayesian nonparametric mixture models

Aiming to solve the same problem as overfitted mixture models, infinite mixture models assume an infinite number of components, $K = \infty$, to avoid committing to a choice of K. In this case, the mixing measure is an infinite random measure, so Bayesian nonparametric priors are needed. In this framework, the mixing measure

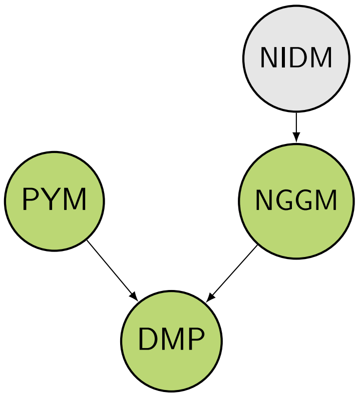


Figure 1.5: Graphical representation of the relationship between the different finite-dimensional priors described in Section 1.2.2.1. An arrow indicates that the target is a special case of the origin. In **green** are the priors considered in **Chapter 2**.

is of this form

$$G = \sum_{i \geq 1} w_i \delta_{\theta_i},$$

where $\theta_1, \theta_2, \ldots \in \mathcal{X}$ is a sequence of location random variables such that $\theta_i \overset{\text{iid}}{\sim} H$, $w_1, w_2, \ldots$ are random variables representing the weights, such that $\sum_{i \geq 1} w_i = 1$.

The Dirichlet process mixture model is the most common BNP mixture model; it was first introduced Lo (1984). The Dirichlet process mixture model is defined as

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} f(\cdot \mid \theta_i), \\ \theta_i \mid G &\overset{\text{iid}}{\sim} G, \\ G &\sim \mathrm{DP}(\alpha, H). \end{aligned} \tag{1.6}$$

It is also possible to opt for other BNP priors on the mixing measure G, such as the extensions to the Dirichlet process mentioned in Section 1.1.2. These priors allow the number of clusters to increase with the number of observations. Using these models, a common way to estimate the true number of components K_0 is to infer the number of clusters, denoted K_n, through its posterior distribution. Chapter 2 and Chapter 3 study the estimation of K_0 using those models. Such models can also be used to perform model-based clustering. This is developed in Section 1.3.

Although an infinite number of components may seem unrealistic, another way of looking at it is that the number of components increases with sample size. Different data behave this way, one of the most known examples being species discovery. In species discovery, the more observations available, the more likely a new species will be discovered. In practice, scientists regularly discover new species. Other examples, such as finding communities in networks, are discussed in Broderick (2016). BNP mixture models are typically used to model these types of data, Frühwirth-Schnatter et al. 2019, Chapter 17 also provide applications of BNP mixture models in finance.

However, BNP mixture models are sometimes used to describe data from a finite mixture with an unknown number of components K_0. In this case, the model is somewhat *misspecified*, as an infinite number of components is used to model K_0. Because of the good properties of BNP mixture models e.g. for density estimation, this practice is quite common but raises some questions about the estimation of K_0. This issue is addressed briefly in Section 1.4 and in more detail in Chapter 2 and Chapter 3.

1.2.2.3 Mixture of finite mixtures

Following the Bayesian approach, a natural way to bypass the choice of the number of components K is to consider K as a parameter of the model and put a prior on it. This defines the Mixture of Finite Mixtures model (MFM), a natural extension of the finite mixture model where the number of components K is an unknown parameter with a prior p_K. Using the latent allocation variable formulation, the model is then defined as

$$\begin{aligned}
&K \sim p_K, \\
&w_1, \dots, w_K \mid K \sim \mathrm{Dir}(\alpha, \dots, \alpha) \\
&z_1, \dots, z_n \mid w_{1:K} \overset{\text{iid}}{\sim} w_{1:K} \\
&\theta_1, \dots, \theta_K \mid (H, K) \overset{\text{iid}}{\sim} H, \\
&X_i \mid \theta_{1:K}, z_{1:n} \overset{\text{ind}}{\sim} p(\cdot \mid \theta_{z_i}), \quad i = 1, \dots, n,
\end{aligned} \tag{1.7}$$

where the third line is a way to describe a variable following a Categorical distribution with parameter $w_{1:K}$, H is a prior or base measure on the parameter space Θ and Dir is the finite Dirichlet distribution. The prior on the mixing measure, $G = \sum_k w_k \delta_{\theta_k}$, is a Gibbs-type prior (see Section 1.1.2) indexed by a negative discount parameter $\sigma = -\alpha$.

This method allows the posterior distribution of the number of components to be examined directly, in addition to the posterior distribution of the number of clusters. This class of models is studied, for instance, in Nobile (1994), Richardson and Green (1997), Miller and Harrison (2018), and Frühwirth-Schnatter et al. (2021).

1.3 Clustering

Clustering is the task of categorising data points with high similarity into groups. This is the central task in unsupervised learning used in various applications, such as genes analysis (McLachlan et al. 2005), marketing (Wedel and Kamakura 2000) or medicine (Rosen and Tanner 1999); these examples and others are detailed in Chapter 8 of Frühwirth-Schnatter et al. 2019. Clustering problems can be addressed using algorithmic approaches such as hierarchical clustering or k-means algorithm. These approaches are based on a definition of the similarities between the observations. Another way to deal with clustering is through *model-based clustering*, which we describe below.

1.3.1 Model-based clustering

Model-based clustering requires formulating a probabilistic model to fit the data and estimate the clusters (Wade 2023; Frühwirth-Schnatter et al. 2019, Chapter 8). The model is commonly considered as a mixture model (Fraley and Raftery 2002). Using a mixture model, each cluster corresponds to a filled (or non-empty) mixture component. Note that we make a difference here, and in this thesis in general, between the notion of *components* and *clusters*. A component represents a group in the population and is part of the model, while a cluster is the estimation of a group. In a Bayesian setting, the components can be seen as the groups a priori and the clusters as the estimation a posteriori.

In a Bayesian framework, one may use a Bayesian mixture model as defined in Section 1.2.2. An important task is to select the number of components. Given our distinction between clusters and components, we denote K the number of components, and K_n the number of clusters. The number of components K is an upper bound on the number of clusters K_n as some components in the model can be empty. As presented in the previous section, different models are considered, such as overfitted mixture models ($K < \infty$), BNP mixture models ($K = \infty$), or mixture of finite mixtures models (K random), see Section 1.2.2.

We can also define a mixture model using latent variables. This model is described in the previous section and recalled here

$$X_i \mid z_i \overset{\text{ind}}{\sim} f(\cdot \mid \theta_{z_i}), \text{ with } \mathrm{P}(z_i = k) = w_k,\ i = 1, \dots, n. \tag{1.8}$$

This definition of a mixture model provides latent variables $z_{1:n} = (z_1, \dots, z_n)$ called the *allocation variables*. For observed data $(x_1, \dots, x_n)$, the allocation variable z_i is such that $z_i = k$ if x_i belongs to group k. By definition, the allocation variable describes the clustering of the data. Following the Bayesian approach, we construct a posterior distribution on this variable

$$p(z_{1:n} \mid X_{1:n}) \propto p(X_{1:n} \mid z_{1:n}) p(z_{1:n}),$$

where $p(z_{1:n})$ is a prior over the space of clusterings, while $p(X_{1:n} \mid z)$ is sometimes called the kernel likelihood and is specified in model-based clustering by Equation (1.8). Alternatively, the quantity $p(X_{1:n} \mid z_{1:n})$ can also be defined using a loss-based approach as proposed in Rigon et al. (2023).

Related to the loss-based approach, using the Bayesian decision theory framework to estimate a point estimate of clustering is also interesting. This method is described below.

1.3.2 Clustering point estimate

Using model-based clustering, we can estimate the clustering posterior distribution. However, one problem with Bayesian clustering is summarising this posterior distribution in a point estimate (Dahl 2006; Lau and Green 2007). In practice, the posterior distribution is available through Markov chain Monte Carlo (MCMC) techniques, which produce a large number of approximate samples from the posterior distribution. Using a point estimate such as the posterior mode, posterior mean, or posterior median is common. However, in clustering analysis, where each sample represents a partition and the partition space is unordered, computing the posterior median is not feasible. Furthermore, defining a sum of clustering to compute a posterior mean is ambiguous. Hence, the posterior mode is the only estimate with a natural definition in this context.

The common method, based on decision theory, proposes choosing the cluster that minimizes a loss function as the point estimate. More formally, given a loss function $\mathcal{L}$ over the clustering space, and $z_{1:n}$ the true clustering, the point estimate $z^\star_{1:n}$ is the estimate minimizing the posterior expected loss

$$z^\star_{1:n} = \arg\min_{\hat{z}_{1:n}} \mathbb{E}\left[\mathcal{L}(z_{1:n}, \hat{z}_{1:n}) \mid X_{1:n}\right] = \arg\min_{\hat{z}_{1:n}} \sum_{z_{1:n}} \mathcal{L}(z_{1:n}, \hat{z}_{1:n}) p(z_{1:n} \mid X_{1:n}). \quad (1.9)$$

It is clear in this formulation that the specification loss $\mathcal{L}$ is significant in the determination of $z^\star_{1:n}$.

Defining a loss over the clustering space is not trivial. However, in the literature, different losses have been studied and proposed. The most classic one is the $0-1$ loss denoted $\mathcal{L}_{0-1}$, and defined as $\mathcal{L}_{0-1}(z_{1:n}, \hat{z}_{1:n}) = \mathbb{I}_{z_{1:n} \neq \hat{z}_{1:n}}$. This loss leads to the following point estimate $z^\star$:

$$\begin{aligned} z^\star_{1:n} &\in \arg\min_{\hat{z}_{1:n}} \sum_{z_{1:n}} \mathbb{I}_{z_{1:n} \neq \hat{z}_{1:n}} p(z_{1:n} \mid X_{1:n}) \\ &\in \arg\min_{\hat{z}_{1:n}} \Big(1 - p(\hat{z}_{1:n} \mid X_{1:n})\Big) \\ &\in \arg\max_{\hat{z}_{1:n}} p(\hat{z}_{1:n} \mid X_{1:n}), \end{aligned}$$

which corresponds to choosing a Maximum a posteriori (MAP). Another common loss is the Binder loss (Binder 1978), which we denote $\mathcal{B}$ and is defined as

$$\mathcal{B}(z_{1:n}, \hat{z}_{1:n}) = \sum_{i<j} \ell_1 \mathbb{I}_{z_i=z_j} \mathbb{I}_{\hat{z}_i \neq \hat{z}_j} + \ell_2 \mathbb{I}_{z_i \neq z_j} \mathbb{I}_{\hat{z}_i = \hat{z}_j},$$

where ℓ_1 and ℓ_2 are penalizing both types of misclassification error, usually $\ell_1 = \ell_2 = 1$. The Binder loss is studied in Bayesian nonparametrics in Lau and Green (2007). Meilă (2007) proposes an alternative loss named variation of information ($\mathcal{VI}$). For $z_{1:n}$ a clustering in k clusters and $\hat{z}_{1:n}$ a clustering in $\hat{k}$, the $\mathcal{VI}$ is defined as

$$\mathcal{VI}(z_{1:n}, \hat{z}_{1:n}) = \sum_{i=1}^{k} \frac{n_{i\bullet}}{n} \log\left(\frac{n_{i\bullet}}{n}\right) + \sum_{j=1}^{\hat{k}} \frac{n_{\bullet j}}{n} \log\left(\frac{n_{\bullet j}}{n}\right) - \sum_{i=1}^{k}\sum_{j=1}^{\hat{k}} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{n}\right),$$

where n_{ij} is the count of data both in cluster z_i and $\hat{z}_j$, $n_{i\bullet} = \sum_j n_{ij}$ and $n_{\bullet j} = \sum_i n_{ij}$. This is studied in Wade and Ghahramani (2018), where its performance is compared with the Binder loss. Other loss functions are considered in e.g. Quintana and Iglesias (2003); Fritsch and Ickstadt (2009); Dombowsky and Dunson (2023).

The different losses described above have various benefits. The MAP clustering based on a Dirichlet process mixture model is studied in Rajkowski (2019). Theoretical results are obtained, such as the unicity of the MAP estimator and good asymptotic properties in specific cases. Lawless (2023), in Chapter 4 shows the consistency of the MAP clustering for a Dirichlet process mixture model when the parameter α decays with the sample size. Chaumeny et al. (2022) study the performance of different losses in practice by conducting a simulation study. Wade (2023) reviewed the performance of the MAP, the Binder loss, and the $\mathcal{VI}$ on two different examples proposed in Miller and Harrison (2013) and Rajkowski (2019).

In practice, the point estimation of the clustering is challenging and computationally expensive. Indeed, the minimization problem in (1.9) is hard to solve due to the large dimension of the clustering space. The cardinality of this space is growing exponentially fast with the sample size and is characterized by the Bell number. Then, it is impossible to enumerate all the possible clustering and explore the whole clustering space. A greedy algorithm to solve this optimization problem is available (see e.g. Wade and Ghahramani 2018; Rastelli and Friel 2018). Another recent algorithm is described in **dahl2022search**.

The results and methods presented here have been discussed in more detail in Wade (2023).

1.4 Asymptotic properties

In statistics, a common way to assess the theoretical quality of a model is to study its asymptotic properties. Ideally, a statistician would like to study the properties of the model for a finite sample of size n. As this is a complex problem with no solution for most models, an asymptotic approach is used as the best possible solution. 1 This section presents two different properties in a Bayesian framework: consistency and its refinement, contraction rate. We also provide some existing results for Bayesian nonparametric mixture models

1.4.1 Consistency

In a frequentist framework, an estimator $\hat{\theta}_n$ is used to estimate the parameter θ. This parameter is supposed to have a fixed and unknown value denoted θ_0. The model is *consistent* at θ_0 if the estimator $\hat{\theta}_n$ converges in probability to θ_0 when the sample size n goes to infinity.

Moving to a Bayesian framework, the posterior distribution characterizes the parameter θ. We are now looking for *posterior consistency*, an asymptotic property of the posterior. The hypothesis of a true value for the parameter is counter-intuitive in the Bayesian framework, where we are traditionally interested in a parameter distribution. On the contrary, we adopt here a frequentist point of view of Bayes procedures. Following Diaconis and Freedman (1986), we assume that the data are distributed according to a true θ_0. Then, the posterior is considered consistent if it converges in any neighborhood of θ_0 when the sample size increases to infinity.

Definition. More formally, given a prior distribution p on the parameter space Θ, where Θ is assumed to be a metric space with a metric d, we denote by $p(\cdot \mid X_{1:n})$ the posterior distribution with $X_{1:n}$ a given sample of the data. The posterior distribution is said to be *consistent* at $\theta_0 \in \Theta$ if

$$p(U^c \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 0,$$

in P_{θ_0}-probability for all neighborhoods U of θ_0.

For instance, considering Bayesian nonparametric (BNP) mixture models, we denote by f_0^X the true density of the data. Then, the posterior density f^X is said to be consistent at f_0^X if, for a distance d on Θ, $p(d(f^X, f_0^X) \geq \varepsilon \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 0$, in $P_{f_0^X}$-probability for all $\varepsilon > 0$. It is also possible to define posterior consistency for the number of clusters or the mixing measure. The posterior number of clusters K_n is said to be consistent at K_0 if $p(K_n = K_0 \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 1$ in $P_{f_0^X}$-probability. There is consistency for the mixing measure G at G_0 if $p(d(G, G_0) \geq \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{} 0$ in $P_{f_0^X}$-probability.

Existing results. An early result on posterior consistency in the nonparametric framework is Doob's Theorem (Doob 1948), which says that for a fixed prior, the posterior distribution is consistent at every value except those in sets that are of null prior measure. This is an interesting result but not enough in BNP because it is a prior-dependent result and the prior null set could be very large for a nonparametric prior (Ghosal and van der Vaart 2017, Section 6.2). Another known result is Schwartz's Theorem (Schwartz 1965). We consider a probability density f_0, and define the Kullback–Leibler support of the prior p as those densities f_0 such that $p(K_\varepsilon(f_0)) > 0$, where K_ε is the Kullback–Leibler neighborhood, $K_\varepsilon(f_0) = \{f, \int f_0 \log(f_0/f) < \varepsilon\}$. Roughly speaking, this theorem states that if the true distribution or parameter is in the K-L support of the prior, then the posterior is consistent at f_0.

Posterior consistency has been studied for BNP mixture models. These models are notably consistent for density estimation (Ghosal et al. 1999; Lijoi et al. 2005b; Ghosal and van der Vaart 2017). Consistency of the mixing measure has also been proven for Dirichlet process (DP) mixture models (Nguyen 2013). The asymptotic behavior of the Gibbs-type process class is studied in De Blasi et al. (2013). In Chapter 2 and Chapter 3, we study the posterior consistency for the number of clusters in BNP mixture models. Existing results in Miller and Harrison (2014) state posterior inconsistency for the number of clusters of a DP and a Pitman–Yor process (PY) mixture model. The posterior consistency of other Bayesian mixture models is also studied. For example, mixture of finite mixture (MFM) are consistent for density estimation (Kruijer et al. 2010), mixing measure estimation Nobile (1994), and the number of components estimation (Guha et al. 2021; Miller 2023).

Misspecification. All these consistency results assume that the kernel of the mixture model is well-specified, meaning that the data are generated from a mixture of distributions that belong to the same family as the kernel used in the model. The study of the misspecified case is also important, as we could expect to face some kernel or mixing measure misspecification in practice. Kleijn and van der Vaart (2006) states general consistency results in the case of prior misspecification and provides examples where these results are valid, e.g. for the mixing measure of a Dirichlet mixture model with a Gaussian location kernel. However, in the MFM case, despite the consistency results for well-specified models, Cai et al. (2021) provide inconsistent results for the number of clusters when the kernel is misspecified or even slightly misspecified. The well-specified assumption is made throughout the rest of the thesis.

1.4.2 Contraction rate

An improvement of the posterior consistency property is the evaluation of the speed at which a posterior distribution concentrates around the true parameter. The quantity that measures this speed is called a *posterior contraction rate*. As before, the parameter space Θ is supposed to be a metric space with a metric d. A sequence ε_n is a posterior contraction rate at the parameter θ_0 with respect to the metric d if for every $M_n \to \infty$, $p(d(\theta, \theta_0) \geq M_n \varepsilon_n \mid X_{1:n}) \xrightarrow[n\to\infty]{} 0$ in P_{θ_0}-probability. As in the case of posterior consistency, it is possible to study the contraction rate at a parameter value but also at a probability density or at a random measure such as the mixing measure in a mixture model. The literature on contraction rates in Bayesian nonparametric has been developed in recent decades (see e.g. Ghosal et al. 2000; Ghosal and van der Vaart 2001; van der Vaart 2004; van der Vaart and van Zanten 2008).

For mixture models, the contraction rates of the posterior distribution of a Dirichlet process mixture model are studied in Ghosal and van der Vaart (2007), and the mixing measure rates of convergence for the same model are studied in Nguyen (2013). The Pitman–Yor process mixture model is also considered in Scricciolo (2014). Some results on contraction rates of finite mixture models (Section 1.2.2.1) are provided in Ho and Nguyen (2016).

For more details and results on posterior consistency or contraction rates, the interested reader could refer to Ghosal and van der Vaart 2017, Chapters 6 to 9.

1.5 Thesis outline

This thesis is separated into two parts. The first part gives theoretical insights into Bayesian mixture models, while the second part presents an application of one of these models to real data.

In the first part, we study the asymptotic properties, as described in Section 1.4, for estimating the number of components K_0 in a finite mixture model. The first part comprises Chapter 2 and Chapter 3. In Chapter 2, we study the consistency when using Gibbs-type process (Section 1.1.2) and finite-dimension representations thereof (Section 1.2.2.1) mixture models to estimate K_0. We prove inconsistency results in both cases. Then, we investigate proposed solutions in the literature, such as the Merge-Truncate-Merge post-processing procedure introduced in Guha et al. (2021). This chapter is based on the following paper:

L. Alamichel, D. Bystrova, J. Arbel, and G. Kon Kam King (2024). "Bayesian mixture models (in)consistency for the number of clusters". In: *Scandinavian Journal of Statistics*.

In Chapter 3, we keep the same framework and study the particular case of Pitman–Yor process mixture models with a hyperprior on the concentration parameter (see Section 1.1.2 for PY). This work is motivated by a solution proposed to the inconsistency issue in Ascolani et al. (2022) for the Dirichlet process mixture model. This solution is to place a prior on the concentration parameter. Contrary to the Dirichlet process case, we prove inconsistency in the PY case. Chapter 3 is an extension of:

C. Lawless, L. Alamichel, J. Arbel, and G. Kon Kam King (2023). "Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter". In: *Fifth Symposium on Advances in Approximate Bayesian Inference*.

In the second part, we propose a Bayesian nonparametric mixture model to assess ecological risk. This model is used to perform density and clustering estimation simultaneously. In the first part, we proved inconsistency results for Bayesian nonparametric mixture models when the true number of components is finite. The framework in this second part is different, as the true number of components is most probably infinite. Still, we do not focus on the posterior number of clusters but choose to use a loss-based approach, as described in Section 1.3, to estimate the clustering; this approach seems to behave well in practice (see Chaumeny et al. 2022). We also use a prior in the class of normalized random measures with independent increments (NRMI), chosen for its robust clustering properties (Barrios et al. 2013). The second part is composed of Chapter 4, which is based on the following recently submitted paper:

L. Alamichel, J. Arbel, G. Kon Kam King, and I. Prünster (2024+). *Species Sensitivity Distribution revisited: a Bayesian nonparametric approach*. Submitted

References

Alamichel, L., J. Arbel, G. Kon Kam King, and I. Prünster (2024+). *Species Sensitivity Distribution revisited: a Bayesian nonparametric approach.* Submitted (cit. on p. 23).

Alamichel, L., D. Bystrova, J. Arbel, and G. Kon Kam King (2024). "Bayesian mixture models (in)consistency for the number of clusters". In: *Scandinavian Journal of Statistics* (cit. on p. 23).

Albughdadi, M., L. Chaari, J.-Y. Tourneret, F. Forbes, and P. Ciuciu (2017). "A Bayesian non-parametric hidden Markov random model for hemodynamic brain parcellation". In: *Signal Processing* 135, pp. 132–146 (cit. on p. 3).

Allison, D. B., G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch (2002). "A mixture model approach for the analysis of microarray gene expression data". In: *Computational Statistics & Data Analysis* 39.1, pp. 1–20 (cit. on p. 11).

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). "Clustering consistency with Dirichlet process mixtures". In: *Biometrika. In press* (cit. on p. 23).

Attorre, F., V. E. Cambria, E. Agrillo, N. Alessi, M. Alfò, M. De Sanctis, L. Malatesta, T. Sitzia, R. Guarino, C. Marcenò, et al. (2020). "Finite Mixture Model-based classification of a complex vegetation system". In: *Vegetation Classification and Survey* 1, p. 77 (cit. on p. 11).

Ayed, F., J. Lee, and F. Caron (2019). "Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior". In: *Proceedings of the 36th International Conference on Machine Learning.* Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 395–404 (cit. on p. 10).

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). "Modeling with Normalized Random Measure Mixture Models". In: *Statistical Science* 28.3, pp. 313–334 (cit. on p. 23).

Binder, D. A. (1978). "Bayesian cluster analysis". In: *Biometrika* 65.1, pp. 31–38 (cit. on p. 19).

Broderick, T. (2016). "Nonparametric Bayesian methods. Tutorial". In: *Machine Learning Summer School, Cádiz* (cit. on p. 15).

Cai, D., T. Campbell, and T. Broderick (2021). "Finite mixture models do not reliably learn the number of components". In: *International Conference on Machine Learning.* PMLR, pp. 1158–1169 (cit. on p. 21).

Carlton, M. A. (2002). "A family of densities derived from the three-parameter Dirichlet process". In: *Journal of applied probability* 39.4, pp. 764–774 (cit. on p. 14).

Chaumeny, Y., J. van der Molen Moris, A. C. Davison, and P. D. W. Kirk (2022). *Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice?* arXiv: 2207.14717 (cit. on pp. 19, 23).

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model". In: *Bayesian inference for gene expression and proteomics*, pp. 201–218 (cit. on p. 18).

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Pruenster, and M. Ruggiero (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 212–229 (cit. on pp. 7, 8).

De Blasi, P., A. Lijoi, and I. Prünster (2013). "An Asymptotic Analysis of a Class of Discrete Nonparametric Priors". In: *Statistica Sinica* 23.3, pp. 1299–1321 (cit. on p. 21).

Diaconis, P. and D. Freedman (1986). "On the Consistency of Bayes Estimates". In: *The Annals of Statistics* 14.1, pp. 63–67 (cit. on p. 20).

Dombowsky, A. and D. B. Dunson (2023). "Bayesian Clustering via Fusing of Localized Densities". In: *arXiv preprint arXiv:2304.00074* (cit. on p. 19).

Doob, J. L. (1948). "Applications of the theory of martingales". In: *Colloques Internationaux du C.N.R.S*, pp. 22–28 (cit. on p. 21).

Durante, D. and D. B. Dunson (2018). "Bayesian Inference and Testing of Group Differences in Brain Networks". In: *Bayesian Analysis* 13.1, pp. 29–58 (cit. on p. 11).

Escobar, M. D. and M. West (1995). "Bayesian Density Estimation and Inference Using Mixtures". In: *Journal of the American Statistical Association* 90.430, pp. 577–588 (cit. on p. 11).

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* 1.2, pp. 209–230 (cit. on pp. 3, 4).

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions". In: *Recent Advances in Statistics*. Elsevier, pp. 287–302 (cit. on p. 11).

Finetti, B. de (1937). "La prévision : ses lois logiques, ses sources subjectives". In: (cit. on pp. 1, 2).

Fraley, C. and A. E. Raftery (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation". In: *Journal of the American Statistical Association* 97.458, pp. 611–631 (cit. on pp. 11, 17).

Fritsch, A. and K. Ickstadt (2009). "Improved criteria for clustering based on the posterior similarity matrix". In: *Bayesian Analysis* (cit. on p. 19).

Frühwirth-Schnatter, S., G. Celeux, and C. P. Robert, eds. (2019). *Handbook of Mixture Analysis*. CRC Press, Taylor & Francis Group (cit. on pp. 11, 15, 17).

Frühwirth-Schnatter, S., G. Malsiner-Walli, and B. Grün (2021). “Generalized Mixtures of Finite Mixtures and Telescoping Sampling”. In: *Bayesian Analysis* 16.4, pp. 1279–1307 (cit. on p. 16).

Frühwirth-Schnatter, S., C. Pamminger, A. Weber, and R. Winter-Ebmer (2012). “Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering”. In: *Journal of Applied Econometrics* 27.7, pp. 1116–1137 (cit. on p. 11).

Gerogiannis, D., C. Nikou, and A. Likas (2009). “The mixtures of Student’s t-distributions as a robust framework for rigid registration”. In: *Image and Vision Computing* 27.9, pp. 1285–1294 (cit. on p. 11).

Ghahramani, Z. and T. Griffiths (2005). “Infinite latent feature models and the Indian buffet process”. In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Vol. 18. MIT Press (cit. on p. 3).

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). “Posterior consistency of Dirichlet mixtures in density estimation”. In: *The Annals of Statistics* 27.1, pp. 143–158 (cit. on p. 21).

Ghosal, S., J. K. Ghosh, and A. van der Vaart (2000). “Convergence rates of posterior distributions”. In: *The Annals of Statistics* 28.2. Number: 2 (cit. on p. 22).

Ghosal, S. and A. van der Vaart (2001). “Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities”. In: *Annals of Statistics* 29.5, pp. 1233–1263 (cit. on p. 22).

Ghosal, S. and A. van der Vaart (2007). “Posterior convergence rates of Dirichlet mixtures at smooth densities”. In: *The Annals of Statistics* 35.2, pp. 697–723 (cit. on p. 22).

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press (cit. on pp. 3, 4, 21, 22).

Gnedin, A. and J. Pitman (2006). “Exchangeable Gibbs partitions and Stirling triangles”. In: *Journal of Mathematical Sciences* 138.3, pp. 5674–5685 (cit. on p. 7).

Guha, A., N. Ho, and X. Nguyen (2021). “On posterior contraction of parameters and interpretability in Bayesian mixture modeling”. In: *Bernoulli* 27.4, pp. 2159–2188 (cit. on pp. 21, 22).

Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1, pp. 97–109 (cit. on p. 1).

Hewitt, E. and L. J. Savage (1955). “Symmetric measures on Cartesian products”. In: *Transactions of the American Mathematical Society* 80.2, pp. 470–501 (cit. on p. 2).

Hjort, N. L., C. Holmes, P. Muller, and S. G. Walker (2010). “Bayesian Nonparametrics”. In: p. 309 (cit. on pp. 2, 3, 9).

Ho, N. and X. Nguyen (2016). “On strong identifiability and convergence rates of parameter estimation in finite mixtures”. In: *Electronic Journal of Statistics* 10.1, pp. 271–307 (cit. on p. 22).

Ishwaran, H. and M. Zarepour (2000). “Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models”. In: *Biometrika* 87.2, pp. 371–390 (cit. on p. 13).

Ishwaran, H. and M. Zarepour (2002). “Exact and approximate sum representations for the Dirichlet process”. In: *Canadian Journal of Statistics* 30.2, pp. 269–283 (cit. on p. 13).

James, L. F., A. Lijoi, and I. Prünster (2009). “Posterior analysis for normalized random measures with independent increments”. In: *Scandinavian Journal of Statistics* 36.1, pp. 76–97 (cit. on p. 9).

Jeffreys, H. (1939). *The theory of probability.* Oxford University Press (cit. on p. 1).

Kingman, J. F. C. (1967). “Completely random measures”. In: *Pacific Journal of Mathematics* 21.1, pp. 59–78 (cit. on p. 9).

Kingman, J. F. C. (1975). “Random discrete distributions”. In: *Journal of the Royal Statistical Society. Series B* 37.1, pp. 1–22 (cit. on p. 9).

Kleijn, B. J. K. and A. van der Vaart (2006). “Misspecification in infinite-dimensional Bayesian statistics”. In: *The Annals of Statistics* 34.2, pp. 837–877 (cit. on p. 21).

Kruijer, W., J. Rousseau, and A. van der Vaart (2010). “Adaptive Bayesian density estimation with location-scale mixtures”. In: *Electronic Journal of Statistics* 4.none, pp. 1225–1257 (cit. on p. 21).

Lau, J. W. and P. J. Green (2007). “Bayesian model-based clustering procedures”. In: *Journal of Computational and Graphical Statistics* 16.3, pp. 526–558 (cit. on pp. 18, 19).

Lavine, M. (1992). “Some aspects of Polya tree distributions for statistical modelling”. In: *The Annals of Statistics*, pp. 1222–1235 (cit. on p. 3).

Lawless, C. (2023). “Advances in Bayesian asymptotics and Bayesian nonparametrics”. PhD thesis. University of Oxford (cit. on p. 19).

Lawless, C., L. Alamichel, J. Arbel, and G. Kon Kam King (2023). “Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter”. In: *Fifth Symposium on Advances in Approximate Bayesian Inference* (cit. on p. 23).

Lijoi, A., R. H. Mena, and I. Prünster (2007). “Controlling the reinforcement in Bayesian non-parametric mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4, pp. 715–740 (cit. on p. 8).

Lijoi, A., R. H. Mena, and I. Prünster (2005a). “Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors”. In: *Journal of the American Statistical Association* 100.472, pp. 1278–1291 (cit. on p. 8).

Lijoi, A., I. Prünster, and T. Rigon (2020). "The Pitman–Yor multinomial process for mixture modelling". In: *Biometrika* 107.4, pp. 891–906 (cit. on pp. 13, 14).

Lijoi, A., I. Prünster, and T. Rigon (2024). "Finite-dimensional Discrete Random Structures and Bayesian Clustering". In: *Journal of the American Statistical Association* 119.546, pp. 929–941 (cit. on pp. 13, 14).

Lijoi, A., I. Prünster, and S. G. Walker (2005b). "On consistency of nonparametric normal mixtures for Bayesian density estimation". In: *Journal of the American Statistical Association* 100.472, pp. 1292–1296 (cit. on p. 21).

Lijoi, A., I. Prünster, and S. G. Walker (2008). "Investigating nonparametric priors with Gibbs structure". In: *Statistica Sinica* 18.4, pp. 1653–1668 (cit. on p. 9).

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates". In: *The Annals of Statistics*, pp. 351–357 (cit. on pp. 3, 15).

Masoero, L., F. Camerlenghi, S. Favaro, and T. Broderick (2022). "More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics". In: *Biometrika* 109.1. Number: 1, pp. 17–32 (cit. on p. 3).

McLachlan, G. J., K.-A. Do, and C. Ambroise (2005). "Analyzing microarray gene expression data". In: (cit. on p. 16).

Meilă, M. (2007). "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5, pp. 873–895 (cit. on p. 19).

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092 (cit. on p. 1).

Miller, J. W. and M. T. Harrison (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components". In: *Advances in Neural Information Processing Systems*, pp. 199–206 (cit. on p. 19).

Miller, J. W. (2023). "Consistency of mixture models with a prior on the number of components". In: *Dependence Modeling* 11.1 (cit. on p. 21).

Miller, J. W. and M. T. Harrison (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components". In: *The Journal of Machine Learning Research* 15.1, pp. 3333–3370 (cit. on p. 21).

Miller, J. W. and M. T. Harrison (2018). "Mixture Models With a Prior on the Number of Components". In: *Journal of the American Statistical Association* 113.521, pp. 340–356 (cit. on p. 16).

Muliere, P. and P. Secchi (1995). "A note on a proper Bayesian bootstrap". In: (cit. on p. 13).

Muliere, P. and P. Secchi (2003). "Weak Convergence of a Dirichlet-Multinomial Process". In: *Georgian Mathematical Journal* 10.2, pp. 319–324 (cit. on p. 13).

Nguyen, X. (2013). “Convergence of latent mixing measures in finite and infinite mixture models”. In: *The Annals of Statistics* 41.1, pp. 370–400 (cit. on pp. 21, 22).

Nobile, A. (1994). “Bayesian Analysis of Finite Mixture Distributions”. PhD thesis. Pittsburgh, PA: Department of Statistics, Carnegie Mellon University (cit. on pp. 16, 21).

Perman, M., J. Pitman, and M. Yor (1992). “Size-biased sampling of Poisson point processes and excursions”. In: *Probability Theory and Related Fields* 92.1, pp. 21–39 (cit. on p. 6).

Pitman, J. (1995). “Exchangeable and partially exchangeable random partitions”. In: *Probability Theory and Related Fields* 102.2, pp. 145–158 (cit. on p. 4).

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d’Eté de Probabilités de Saint-Flour XXXXII-2002*. Springer (cit. on p. 7).

Pitman, J. and M. Yor (1997). “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator”. In: *The Annals of Probability*, pp. 855–900 (cit. on p. 6).

Quintana, F. A. and P. L. Iglesias (2003). “Bayesian clustering and product partition models”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65.2, pp. 557–574 (cit. on p. 19).

Rajkowski, Ł. (2019). “Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model”. In: *Bayesian Analysis* 14.2 (cit. on p. 19).

Ramírez, V. M., F. Forbes, J. Arbel, A. Arnaud, and M. Dojat (2019). “Quantitative MRI Characterization of Brain Abnormalities in DE NOVO Parkinsonian Patients”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1572–1575 (cit. on p. 11).

Rastelli, R. and N. Friel (2018). “Optimal Bayesian estimators for latent variable cluster models”. In: *Statistics and Computing* 28.6, pp. 1169–1186 (cit. on p. 19).

Regazzini, E., A. Lijoi, and I. Prünster (2003). “Distributional results for means of normalized random measures with independent increments”. In: *The Annals of Statistics* 31.2, pp. 560–585 (cit. on p. 9).

Richardson, S. and P. J. Green (1997). “On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4, pp. 731–792 (cit. on p. 16).

Rigon, T., A. H. Herring, and D. B. Dunson (2023). “A generalized Bayes framework for probabilistic clustering”. In: *Biometrika* 110.3, pp. 559–578 (cit. on p. 18).

Rosen, O. and M. Tanner (1999). “Mixtures of proportional hazards regression models”. In: *Statistics in Medicine* 18.9, pp. 1119–1131 (cit. on p. 17).

Rousseau, J. and K. Mengersen (2011). “Asymptotic behaviour of the posterior distribution in overfitted mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, pp. 689–710 (cit. on p. 13).

Savage, L. J. (1954). *The foundations of statistics.* Courier Corporation (cit. on p. 1).

Schwartz, L. (1965). “On Bayes procedures”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 4.1, pp. 10–26 (cit. on p. 21).

Scricciolo, C. (2014). “Adaptive Bayesian Density Estimation in L_p-metrics with Pitman-Yor or Normalized Inverse-Gaussian Process Kernel Mixtures”. In: *Bayesian Analysis* 9.2 (cit. on p. 22).

Sethuraman, J. (1994). “A Constructive Definition of Dirichlet Priors”. In: *Statistica Sinica* 4.2. Publisher: Institute of Statistical Science, Academia Sinica, pp. 639–650 (cit. on p. 5).

Ullah, I. and K. Mengersen (2019). “Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data”. In: *Journal of Big Data* 6.1, pp. 1–25 (cit. on p. 11).

van der Vaart, A. and J. H. van Zanten (2008). “Rates of contraction of posterior distributions based on Gaussian process priors”. In: *Annals of Statistics* 36.3, pp. 1435–1463 (cit. on p. 22).

van der Vaart, A. (2004). “Vitesses de convergence de mesures a posteriori”. In: *Journal de la Société française de statistique* 145.1, pp. 7–30 (cit. on p. 22).

Wade, S. (2023). “Bayesian cluster analysis”. In: *Philosophical Transactions of the Royal Society A* 381.2247, p. 20220149 (cit. on pp. 17, 19).

Wade, S. and Z. Ghahramani (2018). “Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)”. In: *Bayesian Analysis* 13.2. Publisher: International Society for Bayesian Analysis, pp. 559–626 (cit. on p. 19).

Wedel, M. and W. A. Kamakura (2000). *Market segmentation: Conceptual and methodological foundations.* Springer Science & Business Media (cit. on p. 16).

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning.* Vol. 2. 3. MIT press Cambridge, MA (cit. on p. 3).

Zito, A., T. Rigon, O. Ovaskainen, and D. B. Dunson (2022). “Bayesian Modeling of Sequential Discoveries”. In: *Journal of the American Statistical Association*, pp. 1–12 (cit. on p. 3).

Part I
Consistency and properties of Bayesian nonparametric mixture models

Chapter 2

Bayesian mixture models (in)consistency for the number of clusters

The first part of this thesis is devoted to the asymptotic properties of the posterior number of clusters in a mixture model. This first Chapter is a joint work with Daria Bystrova, Julyan Arbel, and Guillaume Kon Kam King. This work began during my internship and continued after the start of my PhD. We extended the inconsistency results for Pitman–Yor and Dirichlet process mixtures provided by Miller and Harrison (2014). We also studied a number of solutions proposed in the literature and proved their applicability.

Daria Bystrova and I contributed equally to this work. All authors contributed to the theoretical development of the paper. Daria Bystrova and I conducted the simulation and real-data study, we wrote the first draft, while all authors contributed to the writing of the final version. This Chapter is based on the following paper, available on arXiv and accepted in the Scandinavian Journal of Statistics:

> L. Alamichel, D. Bystrova, J. Arbel, and G. Kon Kam King (2024). "Bayesian mixture models (in)consistency for the number of clusters". In: *Scandinavian Journal of Statistics*

Contents

Abstract Bayesian nonparametric mixture models are common for modeling complex data. While these models are well-suited for density estimation, recent results proved posterior inconsistency of the number of clusters when the true number of components is finite, for the Dirichlet process and Pitman–Yor process mixture models. We extend these results to additional Bayesian nonparametric priors such as Gibbs-type processes and finite-dimensional representations thereof. The latter include the Dirichlet multinomial process, the recently proposed Pitman–Yor, and normalized generalized gamma multinomial processes. We show that mixture models based on these processes are also inconsistent in the number of clusters and discuss possible solutions. Notably, we show that a post-processing algorithm introduced for the Dirichlet process can be extended to more general models and provides a consistent method to estimate the number of components.

Keywords — Clustering; Finite mixtures; Gibbs-type process; Finite-dimensional BNP representations.

2.1 Introduction

Motivation. Mixture models appeared as a natural way to model heterogeneous data, where observations may come from different populations. Complex probability distributions can be broken down into a combination of simpler models for each population. Mixture models are used for density estimation, model-based clustering (Fraley and Raftery 2002) and regression (Müller et al. 1996). Due to their flexibility and simplicity, they are widely used in many applications such as healthcare (Ramírez et al. 2019; Ullah and Mengersen 2019), econometrics (Frühwirth-Schnatter et al. 2012), ecology (Attorre et al. 2020) and many others (further examples in Frühwirth-Schnatter et al. 2019).

In a mixture model, data $X_{1:n} = (X_1, \dots, X_n)$, $X_i \in \mathcal{X} \subset \mathbb{R}^p$ are modeled as coming from a K-components mixture distribution. If the *mixing measure* G is discrete, i.e. $G = \sum_{i=1}^{K} w_i \delta_{\theta_i}$ with positive weights w_i summing to one and atoms θ_i, then the *mixture density* is

$$f^X(x) = \int f(x \mid \theta) G(\mathrm{d}\theta) = \sum_{k=1}^{K} w_k f(x \mid \theta_k), \tag{2.1}$$

where $f(\cdot \mid \theta)$ represents a component-specific kernel density parameterized by θ. We denote the set of parameters by $\theta_{1:K} = (\theta_1, \dots, \theta_K)$, where each $\theta_k \in \mathbb{R}^d$, $k = 1, \dots, K$. Model (2.1) can be equivalently represented through latent allocation variables $z_{1:n} = (z_1, \dots, z_n)$, $z_i \in \{1, \dots, K\}$. Each z_i denotes the component from which observation X_i comes: $p(X_i \mid \theta_k) = p(X_i \mid z_i = k)$ with $w_k = P(z_i = k)$. Allocation variables z_i define a clustering such that X_i and X_j belong to the same cluster if $z_i = z_j$. Moreover, $z_1, \dots, z_n$ define a partition $A = (A_1, \dots, A_{K_n})$ of $\{1, \dots, n\}$, where K_n denotes the number of clusters.

It is important to distinguish between the number of components K, which is a model parameter, and the number of clusters K_n, which is the number of components from which we observed at least one data point in a dataset of size n (Frühwirth-Schnatter et al. 2021; Argiento and De Iorio 2022; Greve et al. 2022). For a data-generating process with K_0 components, inference on K is typically done by considering the number of clusters K_n and the present article investigates to what extent this is warranted.

Although mixture models are widely used in practice, they remain the focus of active theoretical investigations, owing to multiple challenges related to the estimation of mixture model parameters. These challenges stem from identifiability problems (Frühwirth-Schnatter 2006), label switching (Celeux et al. 2000), and computation complexity due to the large dimension of parameter space.

Another critical question, which is the main focus of this article, regards the number of components and clusters, and whether it is possible to infer them from the data. This question is even more crucial when the aim of inference is clustering. The typical approach to estimating the number of components in a mixture is to fit models of varying complexity and perform model selection using a classic criterion such as the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), etc. This approach is not entirely satisfactory in general, because of the need to fit many separate models and the difficulty of performing a reliable model selection. Therefore, several methods that bypass the need to fit multiple models have been proposed. They define a single flexible model accommodating various possibilities for the number of components: mixtures of finite mixtures, Bayesian nonparametric mixtures, and overfitted mixtures. These methods have been prominently proposed in the Bayesian framework, where the specification of prior information is a powerful and versatile method to avoid overfitting by unduly complex mixture models.

Three types of discrete mixtures. Although we consider discrete mixing measures, G could be any probability distribution (for continuous mixing measures, see for instance Chapter 10 in Frühwirth-Schnatter et al. 2019). Depending on the specification of the mixing measure, there exist three main types of discrete mixture models: *finite mixture* models where the number of components K is considered fixed (known, equal to K_0, or unknown), *mixture of finite mixtures* (MFM) where K is random and follows some specific distribution, and *infinite mixtures* where K is infinite. Under a Bayesian approach, the latter category is often referred to as Bayesian nonparametric (BNP) mixtures.

Specification of the number of components K is different for the three types of mixtures. When K is unknown, the Bayesian approach provides a natural way to define the number of components by considering it random and adding a prior for K to the model, as is done for mixtures of finite mixtures. Inference methods for MFM were introduced by Nobile (1994); Richardson and Green (1997).

Using Bayesian nonparametric (BNP) priors for mixture modeling is another way to bypass the choice of the number of components K. This is achieved by assuming an infinite number of components, which adapts the number of clusters found in a dataset to the structure of the data. The most commonly used BNP prior is the Dirichlet process introduced by Ferguson (1973) and the corresponding Dirichlet process mixture was first introduced by Lo (1984). The success of the Dirichlet process mixture is based on its ease of implementation and computational tractability. However, in some cases the Dirichlet process prior may be restrictive, so more flexible priors such as the Pitman–Yor process can be used. Gibbs-type processes, introduced by Gnedin and Pitman (2006), form an important general class of priors, which contain Dirichlet and Pitman–Yor processes and have flexible clustering properties while maintaining mathematical tractability, see Lijoi and Prünster (2010); De Blasi et al. (2015) for a review. Compared to the Dirichlet process, Gibbs-type priors exhibit a predictive distribution which involves more information, that is, sample size and number of clusters (refer to the sufficientness postulates for Gibbs-type priors of Bacallado et al. 2017). The class of Gibbs-type priors encompasses BNP processes which are widely used, for instance in species sampling problems

(Favaro et al. 2009; Favaro et al. 2012; Cesari et al. 2014; Arbel et al. 2017; Lijoi et al. 2007b), survival analysis (Jara et al. 2010), network inference (Caron and Fox 2017; Legramanti et al. 2022), linguistics (Teh and Jordan 2010) and mixture modeling (Ishwaran and James 2001; Lijoi et al. 2005a; Lijoi et al. 2007a). Miller and Harrison (2018); Frühwirth-Schnatter et al. (2021); Argiento and De Iorio (2022) study the connection between the mixtures of finite mixtures and BNP mixtures with Gibbs-type priors. A common approach to inferring the number of clusters in Bayesian nonparametric models is through the posterior distribution of the number of clusters.

Finally, finite mixture models are considered when K is assumed to be finite. We distinguish two cases, depending on whether the number of components is known or unknown. The case when the number of components is known, say $K = K_0$, is referred to as the exact-fitted setting. An appealing way to handle the other case (K_0 unknown) is to use a chosen upper bound on K_0, i.e. to take the number of components K such that $K \geq K_0$, yielding the so-called overfitted mixture models. A classic overfitted mixture model is based on the Dirichlet multinomial process, which is a finite approximation of the Dirichlet process (see Ishwaran and Zarepour 2002, for instance). Generalizations of the Dirichlet multinomial process were recently introduced by Lijoi et al. (2024); Lijoi et al. (2020), which lead to more flexible overfitted mixture models.

Asymptotic properties of Bayesian mixtures. A minimal requirement for the reliability of a statistical procedure is that it should have reasonable asymptotic properties, such as consistency. This consideration also plays a role in the Bayesian framework, where asymptotic properties of the posterior distribution may be studied. In Table 2.1, we provide a summary of existing results of posterior consistency for the three types of mixture models, when it is assumed that data come from a finite mixture and that the kernel $f(\cdot \mid \theta)$ correctly describes the data generation process (i.e. the so-called *well-specified setting*). We denote by K_0 the true number of components, G_0 the true mixing measure, and f_0^X the true density written in the form of (2.1). For finite-dimensional mixtures, Doob's theorem provides posterior consistency in density estimation (Nobile 1994). However, this is a more delicate question for BNP mixtures. Extensive research in this area provides consistency results for density estimation under different assumptions for Bayesian nonparametric mixtures, such as for Dirichlet process mixtures (Ghosal et al. 1999; Ghosal and van der Vaart 2007; Kruijer et al. 2010) and other types of BNP priors (Lijoi et al. 2005b). In the case of MFM, posterior consistency in the number of clusters as well as in the mixing measure follows from Doob's theorem and was proved by Nobile (1994). Recently, Miller (2023) provided a new proof with simplified assumptions.

For finite mixtures and Bayesian nonparametric mixtures, under some conditions of identifiability, kernel continuity, and uniformity of the prior, Nguyen (2013) proves consistency for mixing measures and provides corresponding contraction rates. These results only guarantee consistency for the mixing measure and do not imply consistency of the posterior distribution of the number of clusters. In contrast, posterior inconsistency of the number of clusters for Dirichlet process mixtures and Pitman–Yor process mixtures is proved by Miller and Harrison (2014). To the best of our knowledge, this result was not shown to hold for other classes of priors. We fill this gap and provide an extension of Miller and Harrison (2014) results for Gibbs-type process mixtures and some of their finite-dimensional representations.

Inconsistency results for mixture models do not impede real-world applications but suggest that inference about the number of clusters must be taken carefully. On the positive side, and in the case of overfitted mixtures, Rousseau and Mengersen (2011) establish that the weights of extra components vanish asymptotically under certain conditions. Additional results by Chambaz and Rousseau (2008) establish posterior consistency for the mode of the number of clusters. Guha et al. (2021) propose a post-processing procedure that allows consistent inference of the number of clusters in mixture models. They focus on Dirichlet process mixtures and we provide an extension for Pitman–Yor process mixtures and overfitted mixtures in this article. Another possibility to solve the problem of inconsistency is to add flexibility for the prior distribution on a mixing measure through a prior on its hyperparameters. For Dirichlet multinomial process mixtures, Malsiner-Walli et al. (2016) observe empirically that adding a prior on the α parameter helps with centering the posterior distribution of the number of clusters on the true value (see their Tables 1 and 2). A similar result is proved theoretically by Ascolani et al. (2022) for Dirichlet process mixtures under mild assumptions.

As a last remark, although we focus on the well-specified case, an important research line in mixture models revolves around misspecified-kernel mixture models, when data are generated from a finite mixture of distributions that do not belong to the kernel family $f(\cdot \mid \theta)$. Miller and Dunson (2019) shows how so-called coarsened posteriors allow performing inference on the number of components in MFMs with Gaussian kernels when data come from skew-normal mixtures. Cai et al. (2021) provide theoretical results for MFMs, when the mixture component family is misspecified, showing that the posterior distribution of the number of components diverges. Misspecification is of course a topic of critical importance in practice, however, the well-specified case is challenging enough to warrant its own extensive investigation.

Contributions and outline. In this rather technical landscape, it can be difficult for the non-specialist to keep track of theoretical advances in Bayesian mixture models. This article aims to provide an accessible review of existing results, as well as the following novel contributions (see Table 2.1):

- We extend Miller and Harrison (2014) results to additional Bayesian nonparametric priors such as Gibbs-type processes (Proposition 2.1) and finite-dimensional representations of them (including the Dirichlet multinomial process and Pitman–Yor and normalized generalized gamma multinomial processes, Proposition 2.2);

- We discuss possible solutions. In particular, we show that the Rousseau and Mengersen (2011) result regarding emptying of extra clusters holds for the Dirichlet multinomial process (Proposition 2.3). Second, we establish that the post-processing algorithm introduced by Guha et al. (2021) for the Dirichlet process extends to more general models and provides a consistent method to estimate the number of components (Propositions 2.4 and 2.5).

- We also provide insight into the non-asymptotic efficiency and practical application of these solutions through an extensive simulation study, and investigate alternative approaches which add flexibility to the prior distribution of the number of clusters.

Quantity of interest	Finite		Infinite	MFM
	$K = K_0$	$K \geq K_0$	$K = \infty$	K random
Density f_0^X	✓ [RGL19]	✓ [RGL19]	✓ [GvdV17]	✓ [KRV10]
Mixing measure G_0	✓ [HN16]	✓ [HN16]	✓ [Ngu13]	✓ [Nob94]
Nb of components K_0	N/A	✗ [ours] / ✓	✗ [MH14, ours] / ✓	✓ [GHN21]

Table 2.1: Results on consistency for different mixture models and quantities of interest in the case where kernel densities are well-specified and data comes from a finite mixture. Consistency is indicated with ✓ and inconsistency with ✗. Our contributions regard the shaded cells. The references cited are [RGL19] Rousseau et al. 2019, Theorem 4.1; [GvdV17] Ghosal and van der Vaart 2017, Theorem 7.15; [KRV10] Kruijer et al. 2010; [HN16] Ho and Nguyen 2016; [Ngu13] Nguyen 2013; [Nob94] Nobile 1994; [MH14] Miller and Harrison (2014); [GHN21] Guha et al. (2021).

The structure of the article is as follows: we start by introducing the notion of a partition-based mixture model and by presenting Gibbs-type processes and finite-dimensional representations of BNP processes in Section 2.2. We then recall in Section 2.3 the inconsistency results of Miller and Harrison (2014) on Dirichlet process mixtures and Pitman–Yor process mixtures and present our generalization.

We discuss some consistency results and a post-processing procedure in Section 2.4. We conclude with a simulation study illustrating some of our results in Section 2.5 and a real data analysis in Section 2.6, while the appendix contains proofs and additional details on the simulation and real data study.

2.2 Bayesian mixture models and mixing measures

We introduce or recall some notions useful for the rest of the paper. We start by defining the mixture model considered. It is based on a partition, whose distribution determines important aspects of the mixture. We introduce different types of priors on the partition, the Gibbs-type process, and some finite-dimensional representations of nonparametric processes such as the Pitman–Yor multinomial process. We conclude this section by recalling the notions of posterior consistency and contraction rate.

2.2.1 Partition-based mixture model

We consider partition-based mixture models as in Miller and Harrison (2014). Let $\mathcal{A}_k(n)$ be the set of ordered partitions of $\{1,\dots,n\}$ into $k \in \{1,\dots,n\}$ nonempty sets:

$$\mathcal{A}_k(n) := \left\{(A_1,\dots,A_k) : A_1,\dots,A_k \text{ disjoint}, \bigcup_{i=1}^{k} A_i = \{1,\dots,n\},\ |A_i| \geq 1\ \forall i\right\}.$$

We denote by $n_i := |A_i|$ the cardinality of set A_i. We consider a partition distribution $p(A)$ on $\bigcup_{k=1}^{n} \mathcal{A}_k(n)$, which induces a distribution $p(k)$ on $\{1,\dots,n\}$. We denote by p a prior density on the parameters $\theta \in \Theta \subset \mathbb{R}^d$ and $f(\cdot \mid \theta)$ a parametrized component density. The hierarchical structure of a *partition-based mixture model* is:

$$\begin{aligned} p(\theta_{1:k} \,|\, A, k) &= \prod_{i=1}^{k} p(\theta_i), \\ p(X_{1:n} \,|\, A, k, \theta_{1:k}) &= \prod_{i=1}^{k} \prod_{j \in A_i} f(X_j \mid \theta_i), \end{aligned}$$

where $X_{1:n} = (X_1,\dots,X_n)$ with $X_i \in \mathcal{X}$, $\theta_{1:k} = (\theta_1,\dots,\theta_k)$ with $\theta_i \in \Theta$, and $A \in \mathcal{A}_k(n)$. In the rest of the article, we denote by K_n the number of clusters in a dataset of size n, which is denoted k in this section for ease of presentation. K_n highlights this quantity's random nature and dependence on n.

The distribution p on the set of ordered partitions determines the type of the mixture model. Here, we consider two types of prior distributions on the partition: nonparametric ones as a Dirichlet process or a Gibbs-type process, and finite-dimensional ones as a Pitman–Yor multinomial process or a normalized infinitely divisible multinomial process.

2.2.2 Gibbs-type processes

Gibbs-type processes are a natural generalization of the Dirichlet process and Pitman–Yor process (see for example De Blasi et al. 2015). Gibbs-type processes of type $\sigma \in (-\infty, 1)$ can be characterized through the probability distribution of the induced random ordered partition $A \in \mathcal{A}_k(n)$, which has the following form:

$$p(A) = p(n_1, \ldots, n_k) = \frac{V_{n,k}}{k!} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{2.2}$$

where $(x)_n = x(x+1)\cdots(x+n-1)$ is the ascending factorial and $(x)_0 = 1$ by convention. $V_{n,k}$ are nonnegative numbers that satisfy the recurrence relation:

$$V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \quad V_{1,1} = 1. \tag{2.3}$$

The probability distribution for the unordered partition $\tilde{A}$ can be deduced from (2.2) multiplying by $k!$ to adjust for order: $p(\tilde{A}) = V_{n,k} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}$. Parameters $V_{n,k}$ admit the following form (see Pitman 2003; Gnedin and Pitman 2006):

$$V_{n,k} = \frac{\sigma^k}{\Gamma(n - k\sigma)} \int_0^{+\infty} \int_0^1 t^{-k\sigma} p^{n-k\sigma-1} h(t) f_\sigma((1-p)t) \mathrm{d}t\mathrm{d}p, \tag{2.4}$$

with Γ the gamma function, f_σ the density of a positive σ-stable random variable and h a non-negative function. We limit ourselves to the case $0 < \sigma < 1$.

Gibbs-type processes are a general class including the Dirichlet and Pitman–Yor processes and some stable processes. The Pitman–Yor family can be defined by the probability p in (2.2) with parameters

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\alpha + i\sigma)}{(\alpha+1)_{n-1}},$$

where $\sigma \in [0, 1)$ and $\alpha \in (-\sigma, \infty)$. If $\sigma = 0$, we obtain the Dirichlet process for which $V_{n,k} = \alpha^k/(\alpha)_n$.

The normalized generalized gamma process (NGG, Lijoi et al. 2007a) is another particular case of Gibbs-type processes, with parameters

$$V_{n,k} = \frac{\mathrm{e}^{\beta}\sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right), \tag{2.5}$$

where $\sigma \in (0,1)$, $\beta > 0$ and $\Gamma(\cdot;\cdot)$ is the following incomplete gamma function: $\Gamma(x;a) = \int_x^\infty s^{a-1}\mathrm{e}^{-s}\mathrm{d}s$. If $\beta = 0$ we obtain the normalized σ-stable process. Furthermore, if $\sigma \to 0$, then we also recover the Dirichlet process (see Figure 2.1(a) for a graphical representation of the relations between these BNP processes).

2.2.3 Finite-dimensional representations

Finite-dimensional representations for BNP priors have been developed to deal with situations where the increase of the number of clusters with the sample size is unrealistic, such as when an upper bound on the number of clusters is known. They are convenient and tractable models that share many properties of their infinite-dimensional counterparts, such as a clear interpretation of their parameters and efficient sampling algorithms. They naturally approximate their associated nonparametric priors as their dimension increases. See Figure 2.1(b) for a graphical representation of the relations between these multinomial mixing measures.

Dirichlet multinomial process. The simplest example of such a finite-dimensional representation is the Dirichlet multinomial distribution (see for instance Muliere and Secchi 1995; Ishwaran and Zarepour 2000). A Dirichlet multinomial process with concentration parameter $\alpha > 0$, number of components K, and base measure H, is a random discrete measure $G = \sum_{k=1}^K w_k \delta_{\theta_k}$ characterized by a Dirichlet distribution on the weights with parameter α/K: $w_{1:K} = (w_1, \ldots, w_K) \sim \mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$ and, as usual, location parameters θ_k are distributed according to the base measure H. Muliere and Secchi (2003) proves that the Dirichlet multinomial process with parameters α, K, and H approximates the Dirichlet process with parameters α and H, in the sense of the weak convergence, when $K \to \infty$. Recent works by Lijoi et al. (2024); Lijoi et al. (2020) develop finite-dimensional versions of the Pitman–Yor process and normalized random measures with independent increments (Regazzini et al. 2003). The latter include the Dirichlet and normalized generalized gamma multinomial processes as special cases.

Pitman–Yor multinomial process. The Pitman–Yor multinomial process is based on the Pitman–Yor process. Fix some integer $K \geq 1$, base measure H, and parameters α, σ as in the Pitman–Yor process case above. The Pitman–Yor multinomial process is defined by Lijoi et al. (2020) as a discrete random probability measure G_K such that

$$G_K \mid G_{0,K} \sim \mathrm{PY}(\sigma, \alpha; G_{0,K}), \quad G_{0,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\tilde{\theta}_k},$$

where $\tilde{\theta}_k \overset{\text{iid}}{\sim} H$. For all $A \in \mathcal{A}_k(n)$, the partition distribution for the Pitman–Yor multinomial process is:

$$p(A) = \binom{K}{k} \frac{1}{(\alpha+1)_{n-1}} \sum_{(\ell_1,\ldots,\ell_k)} \frac{\Gamma(\alpha/\sigma + |\ell^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{K^{\ell_i}}, \tag{2.6}$$

where $k = |A|$ and the sum runs over the vectors $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ such that $\ell_i \in \{1, \ldots, n_i\}$ and $|\ell^{(k)}| = \ell_1 + \cdots + \ell_k$. Coefficients $C(n, k; \sigma)$ are the generalized factorial coefficients defined as

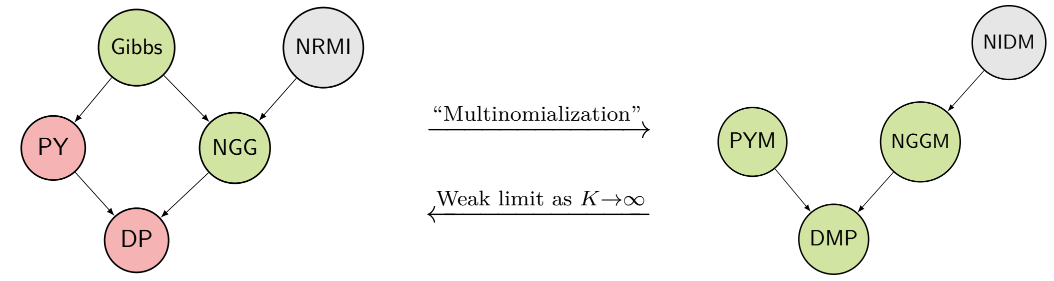
$$C(n, k; \sigma) = \frac{1}{k!}\sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n \tag{2.7}$$

As with the Pitman–Yor process, the random probability measure G_K of the Pitman–Yor multinomial process reduces to the Dirichlet multinomial process when $\sigma = 0$. The Pitman–Yor multinomial process is thus a generalization of the Dirichlet multinomial process. As the latter, the Pitman–Yor multinomial process approximates the Pitman–Yor process, as the Pitman–Yor process is obtained as a limiting case when $K \to \infty$ (see Theorem 5 in Lijoi et al. 2020). In addition, it is also more flexible than the Dirichlet multinomial process. It can be used as an effective computational tool in a nonparametric setting by replacing the stick-breaking construction in the classic Gibbs sampler (see more details in Lijoi et al. 2020).

Normalized infinitely divisible multinomial process. Normalized infinitely divisible multinomial (normalized infinitely divisible multinomial process (NIDM)) processes are introduced by Lijoi et al. (2024) and can be seen as a finite approximation for normalized random measures with independent increments (NRMI), see for instance Regazzini et al. (2003); James et al. (2009). NIDM processes can be described as NRMI measures using a hierarchical structure similar to the previous section

$$G_K \mid G_{0,K} \sim \mathrm{NRMI}(c, \rho; G_{0,K}), \quad G_{0,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\tilde{\theta}_k},$$

where $\tilde{\theta}_k \overset{\text{iid}}{\sim} H$ a base measure. In this expression, ρ is a function that characterizes the NRMI process used. The choice $\rho(s) = s^{-1}e^{-s}$ corresponds to the Dirichlet process. It yields the Dirichlet multinomial process whose distribution for all $A \in$



(a) BNP processes

(b) Multinomial processes

Figure 2.1: Graphical representation of the relationships between the discrete mixing measures considered in this article. An arrow indicates that the target is a special case of the origin. (a) BNP processes: Gibbs-type priors (Gibbs), normalized random measures with independent increments (NRMI), Pitman–Yor process (PY), normalized generalized Gamma process (NGG), and Dirichlet process (DP). (b) Multinomial processes (finite-dimensional approximations of their respective BNP counterparts in the left panel): normalized infinitely divisible multinomial process (NIDM), Pitman–Yor multinomial process (PYM), normalized generalized Gamma multinomial process (NGGM), Dirichlet multinomial process (DMP). Going from left to right can be done according to a “multinomialization” of the BNP processes as described in Section 2.2.3, while the reverse direction is achieved by taking a weak limit as $K \to \infty$. Our contributions generalize results known for mixing measures in red to mixing measures in green. The case of mixing measures in gray remains an open problem.

$\mathcal{A}_k(n)$ is defined as

$$p(A) = \binom{K}{k} \frac{1}{(\alpha)_n} \prod_{j=1}^{k} (\alpha/K)_{n_j}, \tag{2.8}$$

where $k = |A|$. Similarly, choosing $\rho(s) = \frac{1}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-\beta s}$, $0 \leq \sigma < 1$ and $\beta \geq 0$ amounts to considering a normalized generalized Gamma process (NGG) characterized by (2.5). We then get the normalized generalized Gamma multinomial process (NGGM). In this case, for all $A \in \mathcal{A}_k(n)$ the probability is

$$p(A) = \binom{K}{k} \sum_{(\ell_1,\ldots,\ell_k)} \frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}}, \tag{2.9}$$

where $k = |A|$ and $C(n, k; \sigma)$ are defined in (2.7) and the sum over $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ is as in the PY case. Parameters $V_{n,k}$ are defined in (2.5) for the particular case of NGG processes, which depend on β and σ.

2.2.4 Posterior consistency

Posterior consistency is an asymptotic property of the posterior. As in frequentist inference, we can consider that there exists a true value for the parameter of the distribution of the data. Then the posterior is said to be consistent if it converges to the true parameter as the sample size increases to infinity.

More formally, given a prior distribution p on the parameter space Θ, we denote by $p(\cdot \mid X_{1:n})$ the posterior distribution with $X_{1:n}$ a given sample of the data. The posterior distribution is said to be consistent at $\theta_0 \in \Theta$ if $p(U^c \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 0$ in P_{θ_0}-probability for all neighborhoods U of θ_0. For instance, in our case, we consider mixture models for densities. In this type of model, the posterior density is said to be consistent at f_0^X if, for a distance d on the parameter space, $p(d(f, f_0^X) \geq \varepsilon \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 0$ in $P_{f_0^X}$-probability for all $\varepsilon > 0$. It is also possible to define posterior consistency for quantities of interest such as the number of clusters. The posterior number of clusters K_n is said to be consistent at K_0 if $p(K_n = K_0 \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 1$ in $P_{f_0^X}$-probability.

A refinement in the study of posterior consistency is to evaluate the speed at which a posterior distribution concentrates around the true parameter. The quantity that measures this speed is called a posterior contraction rate. More formally, the parameter space Θ is supposed to be a metric space with a metric d. A sequence ε_n is a posterior contraction rate at the parameter θ_0 with respect to the metric d if for every $M_n \to \infty$, $p(d(\theta, \theta_0) \geq M_n\varepsilon_n \mid X_{1:n}) \underset{n\to\infty}{\longrightarrow} 0$ in P_{θ_0}-probability.

For more details on posterior consistency or contraction rates, the reader could refer to Ghosal and van der Vaart 2017, Chapters 6 to 9.

2.3 Inconsistency results

In this section, we generalize the inconsistency results by Miller and Harrison (2014). Under the context defined previously, Miller and Harrison (2014) states sufficient conditions that imply posterior inconsistency of the number of clusters and also proves that these conditions are satisfied for the Dirichlet process and Pitman–Yor process mixture models. For completeness, we first recall here this inconsistency result and then prove that it also applies to the different models introduced in Section 2.2.

2.3.1 Inconsistency theorem of Miller and Harrison (2014)

The central result of Miller and Harrison 2014, Theorem 6 is reproduced below as Theorem 2.1. This result depends on two conditions which are discussed thereafter.

We start with some notations. For $A \in \mathcal{A}_k(n)$, we define $R_A = \bigcup_{i:|A_i|\geq 2} A_i$, the union of all clusters except singletons. For index $j \in R_A$, we define $B(A, j)$ as the ordered partition $B \in \mathcal{A}_{k+1}(n)$ obtained by removing j from its cluster A_ℓ and creating a new singleton for it. Then $B_\ell = A_\ell \setminus \{j\}$, and $B_{k+1} = \{j\}$. Let

$\mathcal{Z}_A := \{B(A, j) : j \in R_A\}$, for $n > k \geq 1$, we define

$$c_n(k) := \frac{1}{n} \max_{A \in \mathcal{A}_k(n)} \max_{B \in \mathcal{Z}_A} \frac{p(A)}{p(B)},$$

with the convention that $0/0 = 0$ and $y/0 = \infty$ for $y > 0$.

Condition 2.1. *Assume* $\limsup_{n\to\infty} c_n(k) < \infty$*, given some particular* $k \in \{1, 2, \ldots\}$*.*

Miller and Harrison (2014) show that this condition holds for any $k \in \{1, 2, \ldots\}$ for the Pitman–Yor process, and thus for the Dirichlet process.

The second condition, named Condition 4 in Miller and Harrison (2014), controls the likelihood through the control of single-cluster marginals. The single-cluster marginal for cluster i is $m(X_{A_i}) = \int_\Theta \left(\prod_{j\in A_i} f(X_j \mid \theta)\right) \pi(\theta)\mathrm{d}\theta$. This condition induces, for example, that as $n \to \infty$, there is always a non-vanishing proportion of the observations for which creating a singleton cluster increases its cluster marginal. This condition only involves the data distribution and is shown to hold for several discrete and continuous distributions, such as the exponential family (see Theorem 7 in Miller and Harrison 2014). In the following, we assume that this condition is satisfied and mainly focus on Condition 2.1.

Theorem 2.1 (Miller and Harrison (2014))**.** *Let* $X_1, X_2, \ldots \in \mathcal{X}$ *be a sequence of random variables and consider a partition-based model. Then, if Condition 4 from Miller and Harrison (2014) holds, and Condition 2.1 above holds for any* $k \geq 1$*, we have for any* $k \geq 1$

$$\limsup_{n\to\infty} p(K_n = k \mid X_{1:n}) < 1 \quad \textit{with probability } 1.$$

As said previously, Condition 2.1 is only related to partition distribution, while Condition 4 from Miller and Harrison (2014) only involves the data distribution and single-cluster marginals. Hence, to generalize this inconsistency result to other processes, it is enough to show that Condition 2.1 also holds for these different processes. This is the focus of the next section, for Gibbs-type processes and finite-dimensional discrete priors.

2.3.2 Inconsistency of Gibbs-type and multinomial processes

We extend the inconsistency result for all the processes introduced in Section 2.2 by proving that Condition 2.1 holds.

Proposition 2.1 (Gibbs-type processes)**.** *Consider a Gibbs-type process with* $0 \leq \sigma < 1$*, then Condition 2.1 holds for any* $k \in \{1, 2, \ldots\}$*, and so does the inconsistency of Theorem 2.1.*

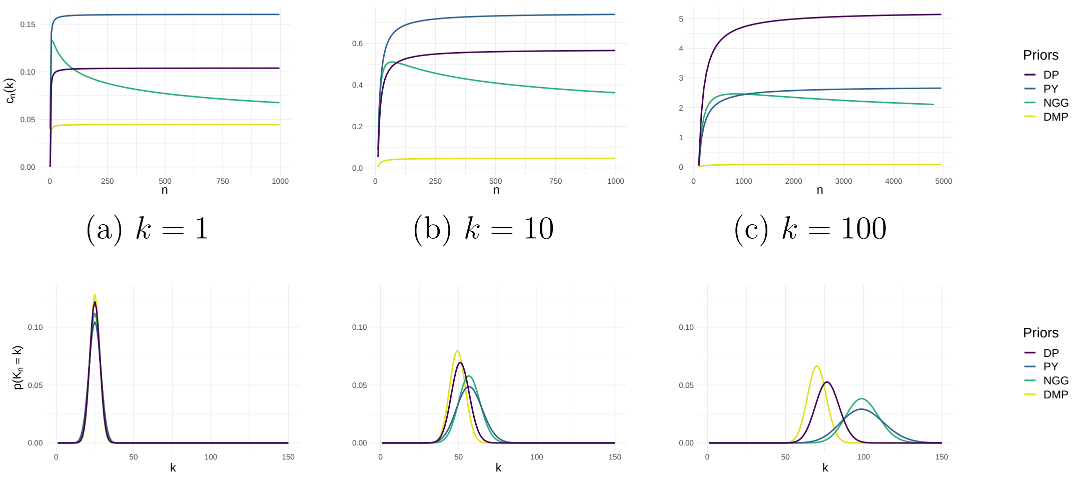
Proposition 2.2 (Multinomial processes). *Consider any of the following priors: Dirichlet multinomial process, Pitman–Yor multinomial process and normalized generalized gamma multinomial process, with K components. Then Condition 2.1 holds for any $k < \min(n, K)$, and so does the inconsistency of Theorem 2.1.*

The proofs of Propositions 2.1 and 2.2 are provided in Appendix 2.A. Note that although the Dirichlet multinomial process is a particular case of the Pitman–Yor multinomial process and the normalized generalized gamma multinomial process, we include it as a separate case in the statement as the proof for this case differs from the proofs for its generalizations.

More precisely, the proof as in Miller and Harrison (2014), Proposition 5 consists in controlling the ratio of probability $\frac{1}{n}\, p(A)/p(B)$, where $B = B(A, j)$ is defined in Section 2.3.1. For the Gibbs-type process, as the ratio of probability is raised by $(V_{n,k}/V_{n,k+1})$, it is enough to show that the sequence $(V_{n,k}/V_{n,k+1})_{n\geq 1}$ is bounded. Since there is no simple formula for $V_{n,k}$ in the general case of the Gibbs-type process, we prove this using a Laplace approximation. The idea of the original proof of Miller and Harrison (2014) is the same but this ratio simplifies as they consider Pitman–Yor process.

For the Pitman–Yor multinomial process and the NGG multinomial process, the partition distribution depends on a sum over the vectors $\ell^{(k)} = (\ell_1, \dots, \ell_k)$ such that $\ell_i \in \{1, \dots, n_i\}$ and $|\ell^{(k)}| = \ell_1 + \dots + \ell_k$. We write this sum as k different sums over each ℓ_i. As in the nonparametric case, we consider the ratio of probability $\frac{1}{n}\, p(A)/p(B)$. By definition of partition B, if $j \in A_k$ then the sum over ℓ_k is different for $p(A)$ and $p(B)$, one is of n_k elements and the other of $n_k - 1$ elements. We separate the sum of n_k elements into two sums, the first one of $n_k - 1$ elements and the second one of one element. In this way, we can use some known properties of the generalized factorial coefficients and some specific properties of each process to conclude.

The top row of Figure 2.2 illustrates Condition 2.1 for different partition distributions, such as the Dirichlet multinomial process (DMP), the Dirichlet process (DP), the Gibbs-type process for the normalized generalized gamma process (NGG) special case and the Pitman–Yor process (PY). In all these cases, we represent the function $c_n(k)$ defined in Section 2.3 for different values of k, $k \in \{1, 10, 100\}$, with $n \in \{1, \dots, 5000\}$ and for some fixed parameters chosen such that $\mathbb{E}[K_{50}] = 25$. We draw all the priors we considered for this choice of the parameters in Figure 2.2 bottom row. We also illustrate how the priors vary depending on n, fixing the priors parameters such that $\mathbb{E}[K_{50}] = 25$ then we made n varying, $n \in \{50, 250, 1000\}$. In Figure 2.2 top row, we can see $n \mapsto c_n(k)$ function reaches a plateau, thus indicating its boundedness for every process and values of k.



(a) $k = 1$ (b) $k = 10$ (c) $k = 100$

(d) $n = 50$ (e) $n = 250$ (f) $n = 1000$

Figure 2.2: (*Top row*) Illustrations of Condition 2.1, for $k \in \{1, 10, 100\}$: the function $n \mapsto c_n(k)$ reaches a plateau for large values of n for a range of priors (see infra).
(*Bottom row*) Prior probability on the number of clusters for different processes and different values of n. In both rows, parameters are fixed such that $\mathbb{E}[K_{50}] = 25$: for Dirichlet process $\mathrm{DP}(\alpha = 19.2)$, for Pitman–Yor process $\mathrm{PY}(\sigma = 0.25, \alpha = 12.2)$, for NGG process $\mathrm{NGG}(\sigma = 0.25, \beta = 48.4)$ and for Dirichlet multinomial process $\mathrm{DMP}(\alpha = 22.5, K = 200)$. Illustrations are made using the package `GibbsTypePriors`.

2.4 Consistency results

The previous results imply that the posterior distribution of the number of clusters for some Bayesian nonparametric mixture models and some overfitted mixture models is inconsistent and thus do not provide good estimates for the number of components in a finite mixture. In these cases, the posterior distribution of the number of clusters is not the relevant summary to consider. Instead, results by Rousseau and Mengersen (2011); Nguyen (2013); Scricciolo (2014) suggest that it might be better to focus on the latent mixing measure. In particular, recent works on consistency can be extended to the models we consider. In this section, we consider the framework of Rousseau and Mengersen (2011) for overfitted mixtures and investigate to which extent it might apply to some models we have been considering, the Dirichlet multinomial process and Pitman–Yor multinomial process mixture models. Moreover Guha et al. (2021) introduce a post-processing procedure, the Merge-Truncate-Merge (MTM) algorithm, for which the output, the number of clusters, is consistent. Guha et al. (2021) proved that this algorithm can be applied to the Dirichlet process mixture model so that there is consistency for the number of clusters after applying this algorithm. We extend this result and prove that we can apply the algorithm to overfitted mixture models and the Pitman–Yor process mixture model.

2.4.1 Emptying extra clusters

Overfitted mixtures can be constructed based on the Dirichlet multinomial process or the Pitman–Yor multinomial process. Rousseau and Mengersen (2011) show in their Theorem 1 that overfitted mixtures, under some conditions on the kernel and the mixture model, have the desirable property that in the mixing measure the weights of extra components tend to zero as the sample size grows. This result only concerns the weights and not the number of clusters, but a near-optimal posterior contraction rate for the mixing measure can be deduced from it (see section 3.1 in Guha et al. 2021). To be more precise, Rousseau and Mengersen (2011) consider a prior p on the mixture weights $w_{1:k}$ written as follows

$$p(w_{1:k}) = C(w_{1:k})w_1^{\alpha_1-1}\cdots w_k^{\alpha_k-1},$$

with specific properties for the function $C(w_{1:k})$ recalled in Condition 2.3. Two types of prior hyper-parameter configurations are studied, which lead to opposite conclusions: merging or emptying of extra components. Let d be the dimension of the component-specific parameter θ. If $\alpha_{\max} = \max_j(\alpha_j)$ is such that $\alpha_{\max} < d/2$, then the posterior expectation for the weights of the extra components tends to zero. This is the case where extra components are emptied. The other case corresponds to $\alpha_{\min} = \min_j(\alpha_j) > d/2$. In this case, the extra components are merged with non-negligible weight, which means that they become identical to an existing component and inadvertently borrow some of its weight. This case is less stable as there are different merging possibilities. It is therefore preferable to choose parameters of the prior that belong to the first case. The result stated in Theorem 1 in Rousseau and Mengersen (2011), depends on five conditions. The first one, Condition 2.2 below, is a posterior contraction condition on the mixture density. The following three conditions, Condition 2.4, Condition 2.5, and Condition 2.6 in Appendix 2.B, are standard conditions on the kernel density, respectively on regularity, integrability, and strong identifiability. Finally, Condition 2.3 below represents a classic continuity property for the prior density. More details on this result are provided in Appendix 2.B where the assumptions on the kernel are recalled and Theorem 1 of Rousseau and Mengersen (2011) is stated.

To apply Theorem 1 in Rousseau and Mengersen (2011) to our case, as the kernel is not the focus of this article, we only need to check the conditions on the mixture model. We recall here these two conditions, Condition 2.2 and Condition 2.3, which correspond to conditions respectively on the posterior contraction of the mixing measure and the prior density.

Condition 2.2 (Rousseau and Mengersen (2011), Assumption 1)**.** *There exists $\varepsilon_n \leq \log(n)^q/\sqrt{n}$, for some $q \geq 0$, such that*

$$\lim_{M\to\infty} \limsup_{n} \left\{ \mathbb{E}_0^n \left[p(\|f^X - f_0^X\|_1 \geq M\varepsilon_n \mid X_{1:n}) \right] \right\} = 0,$$

where f_0^X is the true mixture density.

Condition 2.3 (Rousseau and Mengersen (2011), Assumption 5)**.** *The prior density with respect to Lebesgue measure on the cluster-specific parameter θ is continuous and positive on Θ, and the prior p on $w_{1:K} = (w_1, \dots, w_K)$ satisfies*

$$p(w_{1:K}) = C(w_{1:K}) w_1^{\alpha_1 - 1} \cdots w_K^{\alpha_K - 1},$$

where $C(w_{1:K})$ is a continuous function on the simplex bounded from above and from below by positive constants.

Proposition 2.3. *Assume that the kernel considered satisfies Conditions 2.4, 2.5, and 2.6 (see Appendix 2.B). Let G be a Dirichlet multinomial process. Then, Conditions 2.2 and 2.3 are satisfied, and Theorem 1 of Rousseau and Mengersen (2011) holds.*

The proof of this proposition can be found in Appendix 2.C. It relies on Theorem 4.1 from Rousseau et al. (2019) through which Condition 2.2 holds for mixture models based on the Dirichlet multinomial process.This theorem gives a result on density consistency for finite mixture models in the exact setting, which remains true in the overfitted mixture case. The proof in Appendix 2.C consists mainly of proving that Condition 2.3 holds for the different priors we consider.

We have also studied the Pitman–Yor multinomial process, which is an interesting prior and a natural extension of the Dirichlet multinomial process. As an overfitted mixture, it could be expected that the result in Rousseau and Mengersen (2011) would also apply to this prior. However, even in the special case $\sigma = \frac{1}{2}$ where a prior density for the weights is available in closed form, it can be proven that Condition 2.3 will never be satisfied. More precisely, there exists no $\alpha_1, \dots, \alpha_K$ such that the function $C(w_{1:K})$ defined in Condition 2.3 is bounded from above and below by positive constants. Hence, the Rousseau and Mengersen (2011) framework cannot provide any guarantee for the Pitman–Yor multinomial process. Refer to Appendix 2.C for a detailed description.

2.4.2 Merge-Truncate-Merge algorithm

We assume throughout this section as in Guha et al. (2021) that the parameter space Θ is compact. We denote by $W_r(\cdot,\cdot)$ the Wasserstein distance of order r, $r \geq 1$. We recall in Theorem 2.2 the following result by Guha et al. (2021).

Theorem 2.2 (Guha et al. (2021), Theorem 3.2.)**.** *Let G be a posterior sample from the posterior distribution of any Bayesian procedure, namely $p(\cdot \mid X_{1:n})$ such that for all $\delta > 0$*

$$p\left(G : W_r(G, G_0) \leq \delta\omega_n \mid X_{1:n}\right) \xrightarrow{p_{G_0}} 1,$$

with $\omega_n = o(1)$ a vanishing rate, $r \geq 1$. Let $\tilde{G}$ and $\tilde{K}$ be the outcome of the Merge-Truncate-Merge algorithm (Guha et al. 2021) applied to G. Then the following holds as $n \to \infty$:

(a) $p(\tilde{K} = K_0 \mid X_{1:n}) \longrightarrow 1$ in P_{G_0}-probability.

(b) For all $\delta > 0$, $p(G : W_r(\tilde{G}, G_0) \leq \delta\omega_n \mid X_{1:n}) \longrightarrow 1$ in P_{G_0}-probability.

The Merge-Truncate-Merge algorithm is described in Appendix 2.B.

Proposition 2.4 (Pitman–Yor process)**.** *Let G be a posterior sample from the posterior distribution of a Pitman–Yor process mixture. Under conditions of Lemma 2.1, Theorem 2.2 applies to G.*

Proposition 2.5 (Overfitted mixtures)**.** *Let G be a posterior sample from the posterior distribution of an overfitted mixture. Under conditions of second-order identifiability and uniform Lipschitz continuity of the kernel (Nguyen 2013; Ho and Nguyen 2016), Theorem 2.2 applies to G with $r \leq 2$.*

To prove Proposition 2.4, we introduce a lemma which derives from Theorem 1 in Scricciolo (2014). We assume a location mixture with a known scale parameter τ_0, as stated in Equation (2.14) in Appendix 2.B. The location parameter is univariate, $\Theta \subset \mathbb{R}$. There are three standard conditions, described in Appendix 2.B as Condition 2.9, Condition 2.10 and Condition 2.11, for the theorem. Condition 2.9 is a condition on the kernel density, Condition 2.10 is a tail condition on the true mixing distribution, and Condition 2.11 is a condition on the base measure. Theorem 1 from Scricciolo (2014) is also recalled in Appendix 2.B. To state the lemma, we also need a condition on the kernel $f(\cdot \mid \theta)$. We suppose that for some constants $0 < \rho < \infty$ and $0 < \eta \leq 2$, the Fourier transform $\hat{f}$ of $f(\cdot \mid \theta)$ satisfies $|\hat{f}(t)| \sim e^{-(\rho|t|)^\eta}$.

Lemma 2.1. *Assuming the model is a location mixture as in Equation (2.14), the scale parameter τ_0 is known and $\Theta \subset \mathbb{R}$ is bounded. Under Conditions 2.9, 2.10,* *and 2.11, with G the posterior mixing measure of a Pitman–Yor process mixture model, with $\sigma \in [0, 1)$, then for every $1 \leq r < \infty$, there exists a sufficiently large constant M and some $0 < \eta \leq 2$ such that*

$$p(G : W_r(G, G_0) \geq M \log(n)^{-1/\eta} \mid X^{(n)}) \to 0 \textit{ in } P_{G_0}\textit{-probability.}$$

The proof of this lemma can be found in Appendix 2.C. This lemma is similar to Corollary 2 from Scricciolo (2014) which applies to the special case of the Dirichlet process. With this lemma, we can now prove Proposition 2.4.

Proof of Proposition 2.4. Theorem 2.2 holds when the posterior G is such that for all $\delta > 0$, there exists a vanishing rate ω_n such that

$$p(G : W_r(G, G_0) \geq \delta\omega_n \mid X_{1:n}) \longrightarrow 0 \text{ in } P_{G_0}\text{-probability.}$$

Under the conditions of Lemma 2.1, we have

$$p(G : W_r(G, G_0) \geq M \log(n)^{-1/\eta} \mid X_{1:n}) \to 0 \text{ in } P_{G_0}\text{-probability,}$$

so that $\delta\omega_n = M\,(\log(n))^{-1/\eta}$.

Hence, the consistency results of Theorem 2.2 hold for a Pitman–Yor process mixture model satisfying the conditions of Lemma 2.1. □

In the case of Proposition 2.5, we also need a contraction rate for the mixing measure of overfitted mixture models. To ensure the existence of a contraction rate, two conditions on the kernel are required. These conditions are described in Appendix 2.B as Condition 2.7 and Condition 2.8. Let G be the mixing measure of any overfitted mixture model. It is known that under some conditions on the kernel, there exists a rate of contraction for G (see Equation (5) Guha et al. 2021),

$$p(G : W_2(G, G_0) \gtrsim (\log(n)/n)^{1/4} \mid X_{1:n}) \longrightarrow 0 \text{ in } P_{G_0}\text{-probability.} \tag{2.10}$$

This rate can be suboptimal for some overfitted mixture models but is sufficient to prove Proposition 2.5.

Proof of Proposition 2.5. The proof of Theorem 2.2 is the same in the case of overfitted mixtures as in the Bayesian nonparametric case. This theorem holds when the posterior G is such that for all $\delta > 0$, there exists a vanishing rate ω_n such that

$$p(G : W_r(G, G_0) \geq \delta\omega_n \mid X_{1:n}) \longrightarrow 0 \text{ in } P_{G_0}\text{-probability.}$$

We use Equation (2.10) to conclude with $\delta\omega_n \leq (\log(n)/n)^{1/4}$ and $r = 2$.

Hence, the consistency results of Theorem 2.2 hold for a Pitman–Yor process mixture model satisfying the conditions of Lemma 2.1. □

The work of Guha et al. (2021) can be applied to different Bayesian procedures. The only condition is to have a contraction rate for the mixing measure under the Wasserstein distance. However, this condition is not easy to prove, here we prove it for the Pitman–Yor process but there is no direct generalization for Gibbs-type processes. In the overfitted mixtures case, there is a general contraction rate for the mixing measure under the Wasserstein distance (see Nguyen 2013; Ho and Nguyen 2016). This rate could be suboptimal for some procedures as it is an upper bound but it guarantees the consistency of the Merge-Truncate-Merge algorithm.

2.5 Simulation study

We consider a simulation study to illustrate the three parts of our theoretical results pertaining to (i) inconsistency of the posterior distribution of $\tilde{K}_n$ (Section 2.3.2), (ii) emptying of extra clusters (Section 2.4.1), and (iii) the Merge-Truncate-Merge algorithm (Section 2.4.2). We study the familiar case of a Dirichlet multinomial mixture of multivariate normals. The simulated data was generated using a Gaussian location mixture, with a parameter setting similar to the one of Guha et al. (2021) for the Dirichlet Process. More precisely, we have $K_0 = 3$ clusters and Gaussian kernels such that:

$$f_0^X(x) = \sum_{i=1}^{3} w_i \mathcal{N}(x \mid \mu_i, \Sigma),$$

where $w_{1:3} = (w_1, w_2, w_3)$ are the weights, which we fix as $w_{1:3} = (0.5, 0.3, 0.2)$, and $N(x \mid \mu_i, \Sigma)$ is a multivariate Gaussian distribution with mean μ_i and covariance matrix Σ. We considered the following parameters for the mean and the covariance matrix:

$$\mu_1 = (0.8, 0.8), \mu_2 = (0.8, -0.8), \mu_3 = (-0.8, 0.8) \text{ and } \Sigma = 0.05 I_2.$$

Here, the dimension of the kernel parameter $\theta = (\mu, \Sigma)$ is $d = 5$ (2 for μ and 3 for Σ). In this setting, we generated a sequence of datasets with $n = \{20, 200, 2000, 20000\}$, such that the smaller datasets are subsets of the larger ones. The number of components of the Dirichlet multinomial process is set to $K = 10$, thus satisfying the overfitted condition $K \geq K_0$. We chose the maximum parameter of the Dirichlet distribution, $\bar{\alpha} = \alpha/K$, according to the intuition of Rousseau and Mengersen (2011) results. To obtain vanishing weights for extra components, the parameter $\bar{\alpha}$ should be less than $d/2 = 2.5$. We consider the following values: $\bar{\alpha} \in \{0.01, 1, 2.5, 3\}$. We used the Markov chain Monte Carlo (MCMC) sampler proposed by Malsiner-Walli et al. (2016)*. Although the proposed algorithm allows us to use a hyperprior on the parameter α and shrinkage priors on the component means, we have used the basic version with standard priors on parameters. See details on the number of iterations and simulation practical information in Appendix 2.D. Two situations are considered. In the first case, the prior expected number of clusters is fixed, which leads to decreasing parameter α at a rate asymptotically equivalent to $\log(n)^{-1}$. In the second case, we introduce a prior distribution on $\bar{\alpha}$.

Posterior inconsistency on K_n. In Figure 2.3, we present the posterior distribution of the number of clusters for different values of parameter $\bar{\alpha}$ and different sizes of the dataset n. In addition, we present the prior distribution on the number of clusters for the corresponding $\bar{\alpha}$ and n. Table 2.2 summarizes the values of the parameters $\bar{\alpha}$ and sample sizes n used in the simulation study and displays the associated prior and posterior expected number of clusters K_n. As proved in Proposition 2.2, the posterior distribution diverges with n. This lack of concentration is visible for three of the considered values $\bar{\alpha} \in \{1, 2.5, 3\}$ in our experiments. For $\bar{\alpha} = 0.01$, the posterior distribution stays concentrated around the true value $K_0 = 3$ for the range of sample sizes n. Interestingly, Figure 2.3 makes it clear that the prior with fixed $\bar{\alpha}$ puts increasing mass towards $K_n = 10$ as the sample size grows, which is probably one of the root causes for posterior inconsistency. Allowing $\bar{\alpha}$ to vary, as investigated on Figure 2.8, induces a much less informative prior on the number of clusters and the posterior deterioration as the sample size grows appears much less striking.

n	Prior $\mathbb{E}[K_n]$				Posterior $\mathbb{E}[K_n \mid X_{1:n}]$			
	$\bar{\alpha} = 0.01$	$\bar{\alpha} = 1$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3$	$\bar{\alpha} = 0.01$	$\bar{\alpha} = 1$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3$
20	1.3	6.9	7.9	8	2.8	4.9	5.9	6.0
200	1.5	9.6	9.9	9.98	3.04	6.9	9.5	9.7
2000	1.7	9.9	≈ 10	≈ 10	3.07	8.1	9.98	9.99
20000	1.9	9.99	≈ 10	≈ 10	3.01	8.7	≈ 10	≈ 10

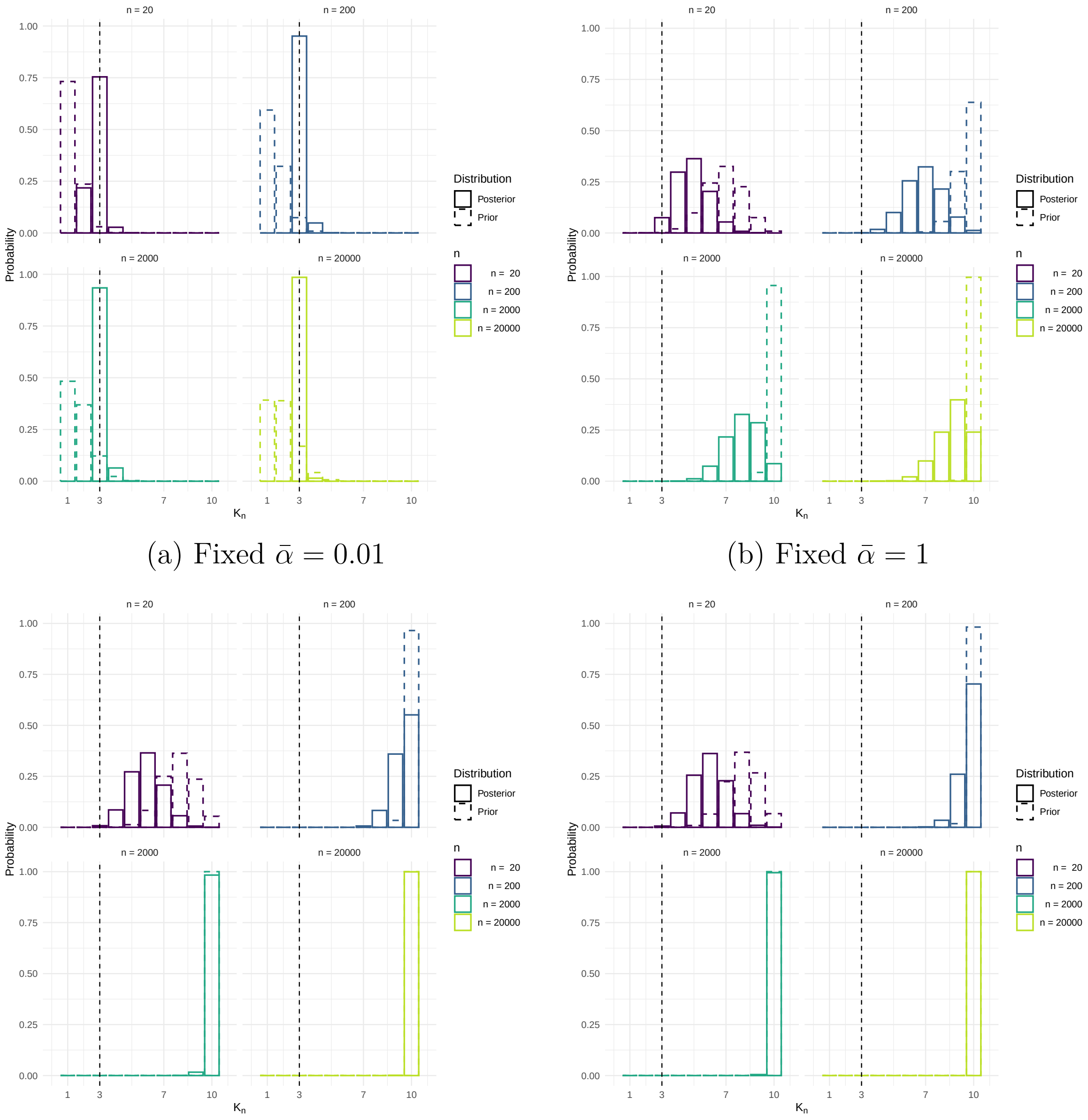
Table 2.2: Prior and posterior expected number of clusters K_n for the various values of $\bar{\alpha}$ considered in our experiments.

Emptying of extra clusters. We are also interested to see how the posterior distribution of the component weights behaves in our simulation setting. Figure 2.4 illustrates the posterior distribution of the weights of the components for different specifications of the parameter $\bar{\alpha}$ and n, and is similar to Figure 1 and Figure 2 in Rousseau and Mengersen (2011). In our case, we sort the weights in decreasing

*The code is available at https://github.com/dbystrova/BNPconsistency.

order to alleviate the label-switching problem. For the minimal values of $\bar{\alpha} = 0.01$, we can see that the posterior weights with growing n are concentrated at the true values of mixture weights, except the largest n. When $\bar{\alpha} = 1$, we can observe the concentration trend, but convergence is slower than in the first case. For $\bar{\alpha} = 2.5$ there are no clear dynamics. And for $\bar{\alpha} = 3$ we can see that the weights become more uniformly distributed, which can be related to the merging weights regime. Specification of our simulation study does not allow to apply the Rousseau and Mengersen (2011) theory directly, as in our case the support of θ is not bounded. However, we can see that the simulation results are still consistent with the theory, suggesting broader applicability.

Merge-Truncate-Merge. We applied the Merge-Truncate-Merge algorithm proposed by Guha et al. (2021) to the posterior distribution of the mixing measure in our simulation setting and illustrate the posterior distribution of the number of clusters $\tilde{K}$ on Figure 2.5. To use the Merge-Truncate-Merge algorithm, we need to know the Wasserstein convergence rates of the corresponding mixing measure. We use the convergence rate for overfitted mixtures $\omega_n = (\log(n)/n)^{1/4}$ (Guha et al. 2021). Note that for this convergence rate the prior on the kernel parameters should be bounded, which is not the case in our simulation (see details in Appendix 2.D), so as in the previous section, we apply Merge-Truncate-Merge out of its theoretically proven domain. The Merge-Truncate-Merge algorithm depends on the specification of a positive scalar c. As there is no explicit guideline for computing c, we tested a range of values $c \in \{0.1, 0.5, 1, 2\}$, see Figure 2.5. We can note that for each value of n, there exists some value of c such that the posterior distribution of the number of clusters remains concentrated around the true number of components $K_0 = 3$. At the same time, some values of c are too restrictive or do not eliminate extra clusters. For example, $c = 0.01$ for $\bar{\alpha} = 1$ does not allow the number of components to be correctly estimated. Conversely, too large a value of c makes the Merge-Truncate-Merge algorithm also fail in the sense that it outputs zero values for $\tilde{K}$. This is because the second step in the algorithm truncates all clusters at once, which corresponds to the case where the set $\mathcal{A}$ of the MTM algorithm recalled in Appendix 2.B is empty and the set $\mathcal{N}$ contains everything. This suggests interpreting c as a regularization parameter, with the estimated number of clusters decreasing with increasing c. Following this intuition, we can draw (Figure 2.6) so-called "regularization paths" plots for c. More specifically, they represent the posterior mean and maximum a posteriori (MAP) for the posterior distribution of $\tilde{K}$ for a range of values $[0, c_{\max}]$ for the parameter c, where $c_{\max}$ is defined as the value of c for which the number of clusters given by the MTM algorithm $\tilde{K}$ is equal to 1. In other words, $c_{\max}$ coincides with the value where all the clusters have been merged or truncated by the MTM post-procedure into a single cluster. We can see that for all specifications of parameter $\bar{\alpha}$ for large $n \geq 2000$, there always exists a region where the posterior mean and the MAP remain approximately constant (exactly constant for the MAP). This suggests a heuristic to use the Merge-Truncate-Merge algorithm: explore regularly spaced values in $[0, c_{\max}]$ and look for a plateau. In the absence of a plateau, the sample size should be increased.



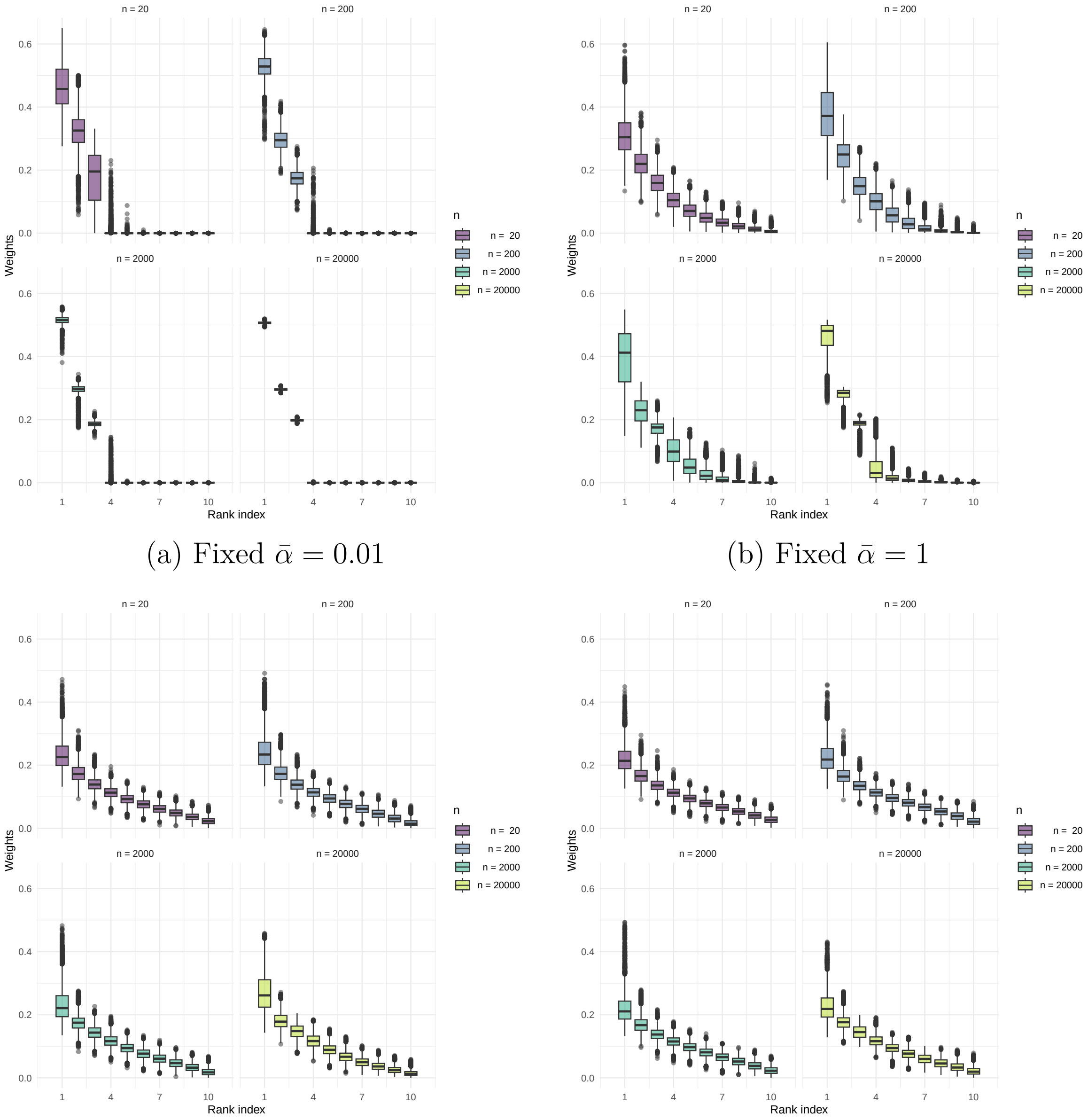
(a) Fixed $\bar{\alpha} = 0.01$

(b) Fixed $\bar{\alpha} = 1$

(c) Fixed $\bar{\alpha} = 2.5$

(d) Fixed $\bar{\alpha} = 3$

Figure 2.3: Prior and posterior distributions of the number of clusters K_n under a Dirichlet multinomial process mixture with fixed parameter $K = 10$, and various choices of $\bar{\alpha} = \alpha/K$ and n. The value $\bar{\alpha} = 2.5$ corresponds to Rousseau and Mengersen (2011)'s threshold.



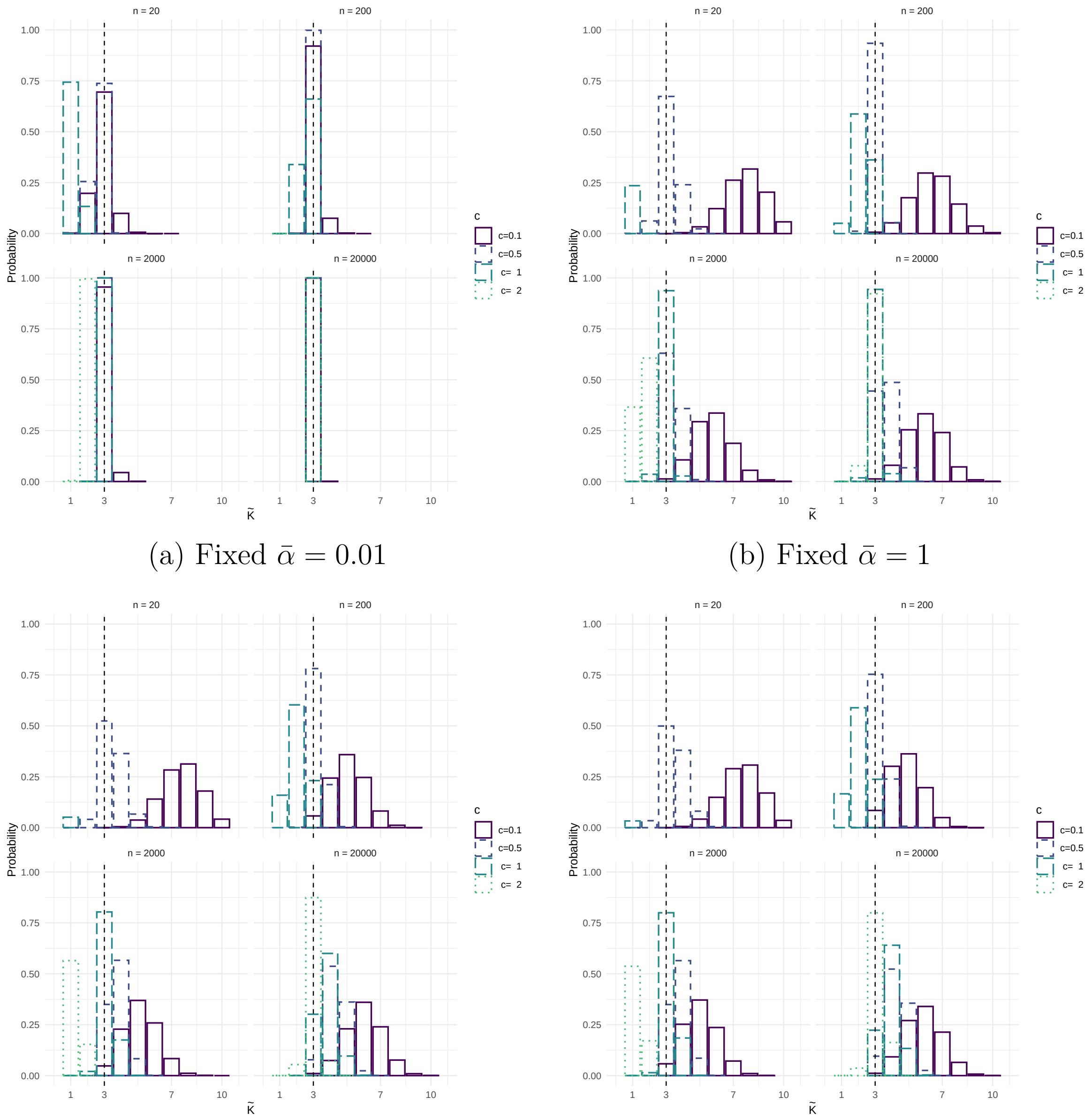
(a) Fixed $\bar{\alpha} = 0.01$

(b) Fixed $\bar{\alpha} = 1$

(c) Fixed $\bar{\alpha} = 2.5$

(d) Fixed $\bar{\alpha} = 3$

Figure 2.4: Mixture weights under a Dirichlet multinomial process mixture with fixed parameter $K = 10$, and various choices of $\bar{\alpha} = \alpha/K$ and n.



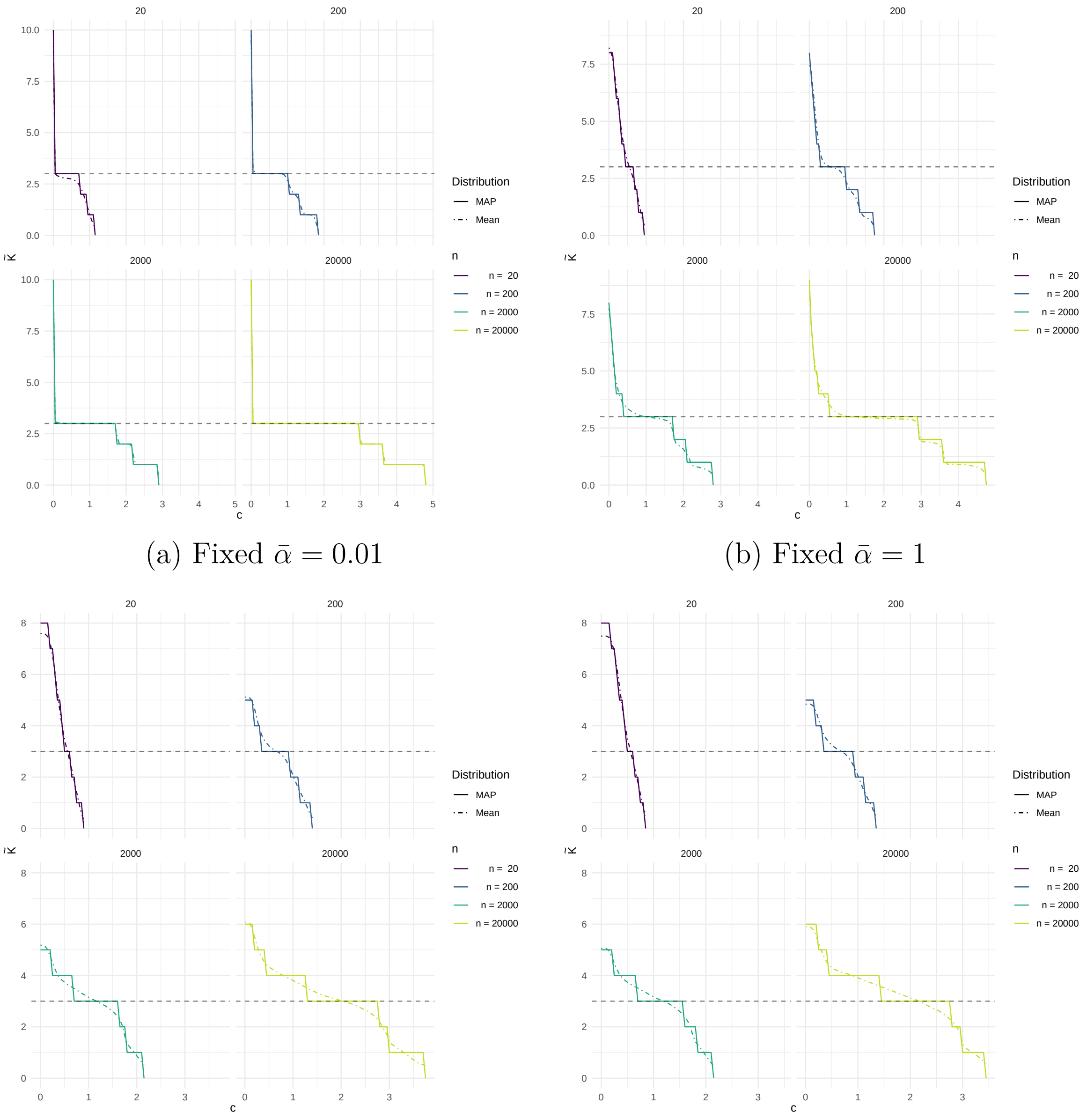
(a) Fixed $\bar{\alpha} = 0.01$

(b) Fixed $\bar{\alpha} = 1$

(c) Fixed $\bar{\alpha} = 2.5$

(d) Fixed $\bar{\alpha} = 3$

Figure 2.5: Distribution of $\tilde{K}$, that is the posterior number of clusters after applying the Merge-Truncate-Merge algorithm of Guha et al. (2021), with c parameter in $\{0.1, 0.5, 1, 2\}$, under a Dirichlet multinomial process mixture with fixed parameter $K = 10$, and various choices of $\bar{\alpha} = \alpha/K$ and n.



(a) Fixed $\bar{\alpha} = 0.01$

(b) Fixed $\bar{\alpha} = 1$

(c) Fixed $\bar{\alpha} = 2.5$

(d) Fixed $\bar{\alpha} = 3$

Figure 2.6: "Regularization path" for $\tilde{K}$, that is the posterior number of clusters after applying the Merge-Truncate-Merge algorithm of Guha et al. (2021), with parameter c in $[0, c_{\max}]$, under a Dirichlet multinomial process mixture with fixed parameter $K = 10$, and various choices of $\bar{\alpha} = \alpha/K$ and n. The dotted dashed curves represent the posterior mean while the solid curves represent the maximum a posteriori (MAP) and the dotted horizontal line represents $K_0 = 3$.

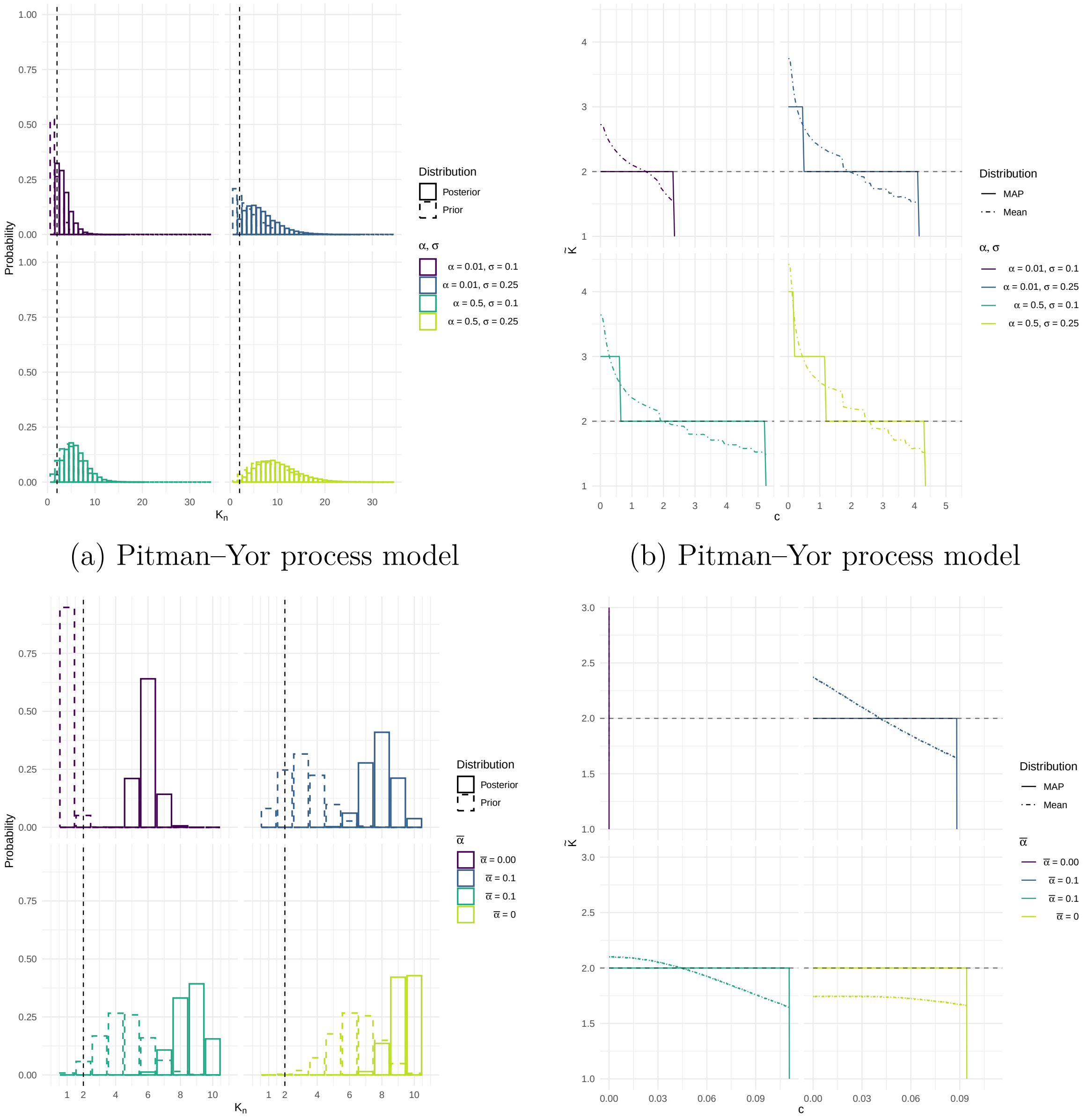
2.6 Real-data Analysis

We now consider the Sodium-Lithium Countertransport (SLC) dataset introduced by Dudley et al. (1991). This dataset is composed of 190 individual measurements of SLC level. This dataset was studied by Miller (2014) with a location-scale Gaussian mixture of finite mixture (MFM) model. In the following, we consider the maximum a posteriori of the number of clusters found in Miller (2014) using MFM as ground truth, hence the true number of components is assumed to be $K_0 = 2$.

We used two different models to study this dataset: a Pitman–Yor process mixture model with various values of $\alpha \in \{0.01, 0.5\}$ and $\sigma \in \{0.1, 0.25\}$, and a Dirichlet multinomial process mixture model with $K = 10$ components and various choices of parameter $\bar{\alpha} = \alpha/K \in \{0.01, 0.5, 1, 2\}$. In both cases, we used a location-scale mixture model described in detail in Appendix 2.E.

In Figure 2.7, we present the posterior distribution of the number of clusters and the so-called "regularization paths" for the two models. More precisely, Figure 2.7 (a) presents the posterior distribution of the number of clusters for a Pitman–Yor process mixture model. We used the marginal sampler from `BNPmix` package proposed in Corradin et al. (2021). In Figure 2.7 (b), we illustrate the application of the Merge-Truncate-Merge algorithm (MTM, Guha et al. 2021) to the posterior distribution of the mixing measure for the Pitman–Yor process mixture model with the "regularization paths" plots for the parameter c from the algorithm. It is worth noticing that even if the contraction rate given in Lemma 2.1 is only valid for a location mixture with the scaling parameter known, here we used this rate for a location-scale mixture model as the scaling parameter is unknown. More precisely, we used the rate of Lemma 2.1 with $\eta = 2$ as the kernel is Gaussian. In the same way, in Figure 2.7 (c), we present the posterior distribution of the number of clusters for a Dirichlet multinomial process mixture model with the number of components $K = 10$ and various choices of parameter $\bar{\alpha}$. In Figure 2.7 (d), we illustrate the application of the Merge-Truncate-Merge algorithm (Guha et al. 2021) to the posterior distribution of the mixing measure for the Dirichlet multinomial process mixture model with the "regularization paths" plots for the parameter c from the algorithm.

For both models, in Figure 2.7 (a) and (c), the posterior distribution of the number of clusters is not centered around the ground truth $K_0 = 2$. This aligns with the inconsistency results in Section 2.3. In Figure 2.7 (b), for each value of α and σ a plateau can be observed on the true value $K_0 = 2$ for the maximum a posteriori as a function of the parameter c of the MTM algorithm. In Figure 2.7 (b) and (d), the range of values for the parameter c is $[0, c_{\max}]$ where $c_{\max}$ is defined as in Section 2.5, such that for the greater value of c $\tilde{K}$ is equal to 1. On the other hand, $c = 0$ is such that only the first stage of MTM algorithm is performed, see Appendix 2.B for more details. In Figure 2.7 (d), the ranges of values for c are very small, illustrating the fact that the first stage of the MTM algorithm alone already has a very strong merging effect. Plateaus on the true value $K_0 = 2$ for the maximum a posteriori as a function of c can still be observed for $\bar{\alpha} > 0.01$. For $\bar{\alpha} = 0.01$, the first stathe ge of MTM algorithm is not strong enough to find $\tilde{K} = 2$ (finds $\tilde{K} = 3$)



(a) Pitman–Yor process model

(b) Pitman–Yor process model

(c) Dirichlet multinomial process model

(d) Dirichlet multinomial process model

Figure 2.7: (a) Prior and posterior distributions of the number of clusters K_n under a Pitman–Yor process model with various choices of α and σ applied on the SLC dataset. (b) Corresponding "regularization path" for $\tilde{K}$, that is the posterior number of clusters after applying the Merge-Truncate-Merge algorithm of Guha et al. (2021), with c parameter in $[0, 6]$.
(c) Prior and posterior distributions of the number of clusters K_n under a Dirichlet multinomial process mixture with fixed parameter $K = 10$, and various choices of $\bar{\alpha} = \alpha/K$ applied on the SLC dataset. (d) Corresponding "regularization path" for $\tilde{K}$, that is the posterior number of clusters after applying the Merge-Truncate-Merge algorithm of Guha et al. (2021), with c parameter in $[0, 0.12]$.
For (b) and (d) the dotted dashed curves represent the posterior mean while the solid curves represent the maximum a posteriori (MAP). For (a) and (c) the solid curves represent the posterior distribution of K_n while the dotted curves represent the prior distribution. The dotted line represents $K_0 = 2$.

and the second stage of the algorithm, which applies when $c > 0$, directly jumps to one cluster $\tilde{K} = 1$.

2.7 Discussion

We studied the finite and infinite mixture models with well-specified kernels applied to data generated from a mixture with a finite number of components. In this setting, we have proved that Gibbs-type process mixtures are inconsistent a posteriori for the number of clusters. It is also the case for some finite-dimensional representations of Gibbs-type priors such as the Dirichlet multinomial, Pitman–Yor multinomial and normalized generalized gamma multinomial processes. However, we did not prove inconsistency in general for NIDM (Lijoi et al. 2024). Further, we discussed the different approaches to solving inconsistency problems for both finite and infinite mixtures.

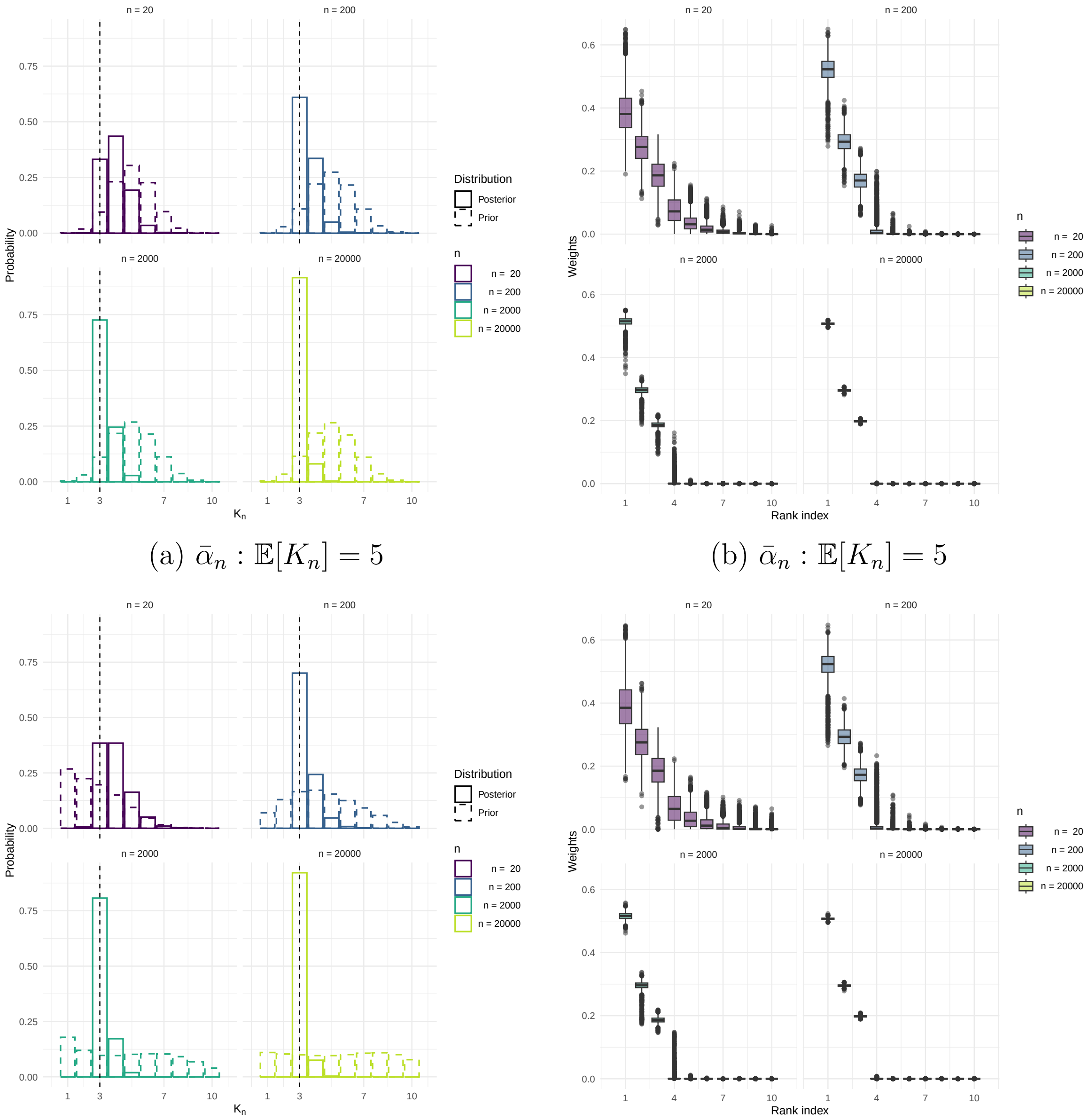
For overfitted mixtures, Rousseau and Mengersen (2011) prove that for some parameter specifications, the weights for extra components vanish, but it does not guarantee the posterior consistency of the number of clusters. We show that this guides prior specification for some of the models that are inconsistent a posteriori, such as overfitted mixtures based on the Dirichlet multinomial process. On the other hand, we also proved that the Pitman–Yor multinomial process does not satisfied the conditions of Theorem 1 in Rousseau and Mengersen (2011). When the Wasserstein convergence rate of the mixing measure is known, the Merge-Truncate-Merge (MTM) algorithm proposed by Guha et al. (2021) allows obtaining a consistent estimate of the number of components in Bayesian nonparametric and overfitted mixtures. In particular, we showed that in contrast to the results of Rousseau and Mengersen (2011), the Merge-Truncate-Merge algorithm can be applied to the Dirichlet multinomial and Pitman–Yor multinomial processes without parameter constraints. Moreover, we also proved that Merge-Truncate-Merge can be applied to the Pitman–Yor process in the case of location mixtures.

Even if it seems possible to recover some consistency with, for example, the Merge-Truncate-Merge procedure, our inconsistency results suggest that Gibbs-type process mixture models face challenges when employed to estimate a finite number of components. This can be related to the fact that this usage corresponds to model misspecification, as these models assume an infinite number of components or a number of clusters growing with the sample size. When it is known that the number of components is finite, we can also use a Mixture of Finite Mixtures which is better specified for this case. MFM are consistent for the number of components as proved in Guha et al. (2021). However, MFM are notoriously more computationally challenging than Dirichlet process mixtures, for instance, when the number of components is large (see remark in Section 3.2 Guha et al. 2021). This might be a motivation to favour using misspecified Gibbs-type process mixture models in conjunction with the Merge-Truncate-Merge algorithm for instance in place of MFM. However, recent works introduced new samplers for MFM which appear more computationally efficient than the usual ones (Miller and Harrison 2018; Frühwirth-Schnatter et al. 2021; De Blasi and Gil–Leyva 2023).

It is known that the Dirichlet process mixture model tends to create some extra little clusters which are linked to the inconsistency result (see e.g. Miller and Harrison 2014, and references therein). To avoid these clusters, some authors propose to use repulsive mixture models (see e.g. Petralia et al. 2012). Such models introduce a dependence on the components to better spread them out in the parameter space. Xie and Xu (2020) prove consistency for the density and the mixing measure for repulsive mixture models with Gaussian kernel. As for the number of components, no consistency is proven, but it is shown that some shrinkage effect occurs.

Another way to solve the inconsistency problem of the posterior number of clusters in the Dirichlet process mixture is introduced by Ohn and Lin (2023). Their solution is to make the concentration parameter α decrease when the sample size increases. With this assumption, they obtain a nearly tight upper bound on the true number of components through the posterior number of clusters. They also present a simulation study showing posterior consistency for the number of components. We can wonder if control over the concentration parameter α when the sample size increases can allow posterior consistency for the number of components. Indeed, Ascolani et al. (2022) proposes a way to control this parameter through a prior, which gives consistency for the number of clusters for a Dirichlet process mixture.

We investigate empirically these two directions, with a simulation study for Dirichlet multinomial mixtures where (i) we fix the expected number of clusters a priori when the sample size increases, implying that α decreases (Figure 2.8 (a) and (b)) and (ii) we use a Gamma prior on the concentration parameter (Figure 2.8 (c) and (d)) (see details in Appendix 2.D). As illustrated in Figure 2.8, the posterior number of clusters in both cases seems to estimate the true number of components well even for large sample sizes, and the posterior seems to be consistent. This observation is corroborated by the posterior distribution of the weights shown in Figure 2.8 (b) and (d). However, there are no theoretical guarantees for consistency or inconsistency since the results of respectively Ohn and Lin (2023) and Ascolani et al. (2022) do not apply in both cases. Although we cannot directly compare our experimental results with results obtained by Ohn and Lin (2023) due to different theoretical assumptions, we can note that the theoretical results obtained by Ohn and Lin (2023) require that $\bar{\alpha}$ decreases as n^{-a_0}, where $a_0 > 0$, which is faster than the $1/\log(n)$ decrease induced by fixing the expectation. So, the obtained results suggest that a slower decrease rate for α might be enough to obtain consistency.



(a) $\bar{\alpha}_n : \mathbb{E}[K_n] = 5$

(b) $\bar{\alpha}_n : \mathbb{E}[K_n] = 5$

(c) $\bar{\alpha} \sim \text{Ga}(a, bK)$

(d) $\bar{\alpha} \sim \text{Ga}(a, bK)$

Figure 2.8: Dirichlet multinomial process mixtures varying concentration parameter $\bar{\alpha}$. (a) and (b): $\bar{\alpha}$ chosen such that $\mathbb{E}[K_n] = 5$ for various choices of n. (c) and (d): a $\text{Ga}(a, bK)$ prior is used on $\bar{\alpha}$. (a) and (c): Prior and posterior distributions of the number of clusters K_n. (b) and (d): Boxplots of mixture weights.

Another way to estimate the number of components is to use the approach of Wade and Ghahramani (2018). This approach consists of a point estimation of the partition of the data and is commonly used in practice. As it is widely used in practice, it would be interesting to investigate the consistency in this case. Chaumeny et al. (2022) investigates this question from a practical point of view, using a simulation study, some positive results are found, but no theory is provided.

Bayesian nonparametric or overfitted mixtures are often used in practical applications. In our work, we showed that the number of clusters estimated using these models is inconsistent for some of these models. We have also discussed possible ways to obtain consistent estimates in practice, using either prior- or post-processing procedures. However, throughout the paper, we considered a well-specified kernel case, where the data is generated from the finite mixture of distributions that belong to the considered kernel family. In practice, this condition can be easily violated, and an interesting avenue of research would be to investigate misspecified settings.

Supporting Information

Additional information for this article is available online, consisting of the code used for the simulations and the figures.

2.A Proofs of the results of Section 2.3

Proof of Proposition 2.1. For all $k \in \{1, 2, \ldots\}$, we want to prove that

$$\limsup_{n\to\infty} c_n(k) = \limsup_{n\to\infty} \frac{1}{n} \max_{A\in\mathcal{A}_k(n)} \max_{B\in\mathcal{Z}_A} \frac{p(A)}{p(B)} < \infty,$$

where $\mathcal{Z}_A$ and $\mathcal{A}_k(n)$ are defined in Section 2.2.1.
So, it is sufficient to prove that for any fixed k, there exists a constant C such that for any n, for all $A \in \mathcal{A}_k(n)$ and $B = B(A, j)$ with $j \in A_\ell$, $\frac{1}{n}\frac{p(A)}{p(B)} \leq C$.

We consider the Gibbs-type prior case with $\sigma > 0$, as case, $\sigma = 0$ is a Dirichlet process and is already proven in Miller and Harrison (2014). As we are in the Gibbs-type prior case, we have, for $A \in \mathcal{A}_k(n)$, $p(A) = \frac{V_{n,k}}{k!} \prod_{i=1}^{k} (1-\sigma)_{n_i-1}$, and so

$$\begin{aligned}\frac{1}{n}\frac{p(A)}{p(B)} &= \frac{1}{n}\frac{V_{n,k}}{k!}\prod_{i=1}^{k}(1-\sigma)_{|A_i|-1}\frac{(k+1)!}{V_{n,k+1}}\left(\prod_{i=1}^{k+1}(1-\sigma)_{|B_i|-1}\right)^{-1}\\ &= \frac{k+1}{n}\frac{V_{n,k}}{V_{n,k+1}}\underbrace{(1-\sigma+|A_\ell|-2)}_{\leq n}\\ &\leq \frac{V_{n,k}}{V_{n,k+1}}(k+1).\end{aligned}$$

Therefore, we just have to prove that the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 1}$ is bounded.

Using the recurrence relation (2.3), we have

$$\begin{aligned}V_{n,k} = V_{n+1,k+1} + (n-\sigma k)V_{n+1,k} \iff \frac{V_{n,k}}{V_{n+1,k+1}} &= \frac{V_{n+1,k+1}}{V_{n+1,k+1}} + (n-\sigma k)\frac{V_{n+1,k}}{V_{n+1,k+1}}\\ \iff \frac{V_{n+1,k}}{V_{n+1,k+1}} &= \left(\frac{V_{n,k}}{V_{n+1,k+1}} - 1\right)\frac{1}{n-\sigma k}. \quad (2.11)\end{aligned}$$

We denote by $f_n(p,t) = t^{-\sigma k} p^{n-1-k\sigma} h(t) f_\sigma(t(1-p))$ the integrand function of Equation (2.4). From the definition of the $V_{n,k}$ in (2.4), we can write

$$\frac{V_{n+1,k}}{V_{n,k}} = \frac{1}{n-\sigma k}\frac{\iint pf_n}{\iint f_n}.$$

Using again the recurrence relation (2.3), we have

$$\frac{V_{n+1,k+1}}{V_{n,k}} = 1 - (n-\sigma k)\frac{V_{n+1,k}}{V_{n,k}}.$$

Then, applying the Laplace approximation method twice and by setting (t_n, p_n) the mode of f_n, we obtain as in Arbel and Favaro (2021)

$$\frac{V_{n+1,k+1}}{V_{n,k}} = g(t_n, p_n) + o\left(\frac{1}{n}\right), \tag{2.12}$$

with $g(t_n, p_n) = 1 - p_n$. Indeed, to use the Laplace approximation, we write the integrand as $f_n = e^{n\ell_n}$, then

$$\frac{V_{n+1,k+1}}{V_{n,k}} = \frac{\iint g e^{n\ell_n}}{\iint e^{n\ell_n}}.$$

As the exponential term is the same in both integrands of this ratio, by applying the Laplace approximation method to both integrals, we obtain

$$\frac{V_{n+1,k+1}}{V_{n,k}} = \frac{g(t_n, p_n) + a(t_n, p_n)/n + \mathcal{O}\left(\frac{1}{n^2}\right)}{1 + \mathcal{O}\left(\frac{1}{n}\right)},$$

where $a(t_n, p_n)$ is a second order term such that $a(t_n, p_n) = o(1/n)$. Hence, the previous ratio finally simplifies to (2.12).

Let $\varphi_h(t) = -th'(t)/h(t)$, we can finally write using the partial derivatives above

$$\frac{V_{n+1,k+1}}{V_{n,k}} = \frac{\sigma k + \varphi_h(t_n)}{n + \varphi_h(t_n) - 1} + o\left(\frac{1}{n}\right). \tag{2.13}$$

Thus, if $\varphi_h(t_n)$ converges as n tends to infinity, we have that $\frac{V_{n+1,k+1}}{V_{n,k}} \times \frac{n}{\sigma k} \to 1$ as $n \to \infty$, so with the relation (2.11), $\frac{V_{n+1,k}}{V_{n+1,k+1}} \xrightarrow[n\to\infty]{} \frac{1}{\sigma k}$. If $\varphi_h(t_n)$ diverges as n tends to infinity, we have that

$$\lim_{n\to\infty} \frac{V_{n+1,k+1}}{V_{n,k}} = \begin{cases} \frac{1}{c+1} & \text{if } \frac{n}{\varphi_h(t_n)} \xrightarrow[n\to\infty]{} c,\ c \in \mathbb{R}, \\ 0 & \text{if } \frac{n}{\varphi_h(t_n)} \xrightarrow[n\to\infty]{} \pm\infty. \end{cases}$$

And then, using again (2.11), $\frac{V_{n+1,k}}{V_{n+1,k+1}} \xrightarrow[n\to\infty]{} 0$. Hence,

$$\lim_{n\to\infty} \frac{V_{n+1,k}}{V_{n+1,k+1}} = \begin{cases} \frac{1}{\sigma k} & \text{if } \varphi_h(t_n) \text{ converges,} \\ 0 & \text{if } \varphi_h(t_n) \text{ diverges.} \end{cases}$$

Thus, the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 1}$ is bounded and Condition 2.1 is satisfied.

□

Proof of Proposition 2.2. We consider $A \in \mathcal{A}_k(n)$ and $B = B(A, j)$, and we assume for simplicity, and without loss of generality, that the cluster in A which contains the element j is the k-th cluster A_k. As in the previous proof, we want to bound the ratio $\frac{p(A)}{p(B)}$ for the three different partition probabilities considered in the proposition.

First, we consider the Dirichlet multinomial process, which is a special case of the Pitman–Yor multinomial process and normalized generalized gamma when $\sigma = 0$. Then we consider the Pitman–Yor multinomial process and the normalized generalized gamma process with $\sigma > 0$.

(a) Dirichlet multinomial process: using (2.8), we have

$$\frac{1}{n}\frac{p(A)}{p(B)} = \frac{1}{n}\frac{p(n_1,\dots,n_k)}{p(n_1,\dots,n_k-1,1)}.$$

So,

$$\begin{aligned}\frac{1}{n}\frac{p(A)}{p(B)} &= \frac{1}{n}\frac{(k+1)!(K-k-1)!\prod_{j=1}^{k}(c/K)_{n_j}(c)_n}{k!(K-k)!\prod_{i=1}^{k+1}(c/K)_{n_i}(c)_n}\\ &= \frac{(k+1)(c/K+n_k-1)}{n(K-k)c/K} \leq \frac{K(k+1)}{c(K-k)}.\end{aligned}$$

Thus, Condition 2.1 is satisfied for the Dirichlet multinomial process.

(b) Pitman–Yor multinomial process with $\sigma > 0$: we denote by $q_{\ell^{(k)}} = \prod_{i=1}^{k} C(n_i,\ell_i;\sigma)/K^{\ell_i}$. Using (2.6), we have

$$\begin{aligned}\frac{1}{n}\frac{p(A)}{p(B)} &= \frac{1}{n}\frac{p(n_1,\dots,n_k)}{p(n_1,\dots,n_k-1,1)}\\ &= \frac{(k+1)!(K-(k+1))!}{nk!(K-k)!}\frac{\sum_{(\ell_1,\dots,\ell_k)}\frac{\Gamma(\alpha/\sigma+|\ell^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)}\;q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k+1})}\frac{\Gamma(\alpha/\sigma+|\ell^{(k+1)}|)}{\sigma\Gamma(\alpha/\sigma+1)}\;q_{\ell^{(k+1)}}}\\ &= \frac{k+1}{n(K-k)}\frac{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k}\Gamma(\alpha/\sigma+|\ell^{(k)}|)\;q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\sum_{n_{k+1}=1}^{1}\Gamma(\alpha/\sigma+|\ell^{(k+1)}|)\;q_{\ell^{(k+1)}}}\\ &= \frac{k+1}{n(K-k)}\frac{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k}\Gamma(\alpha/\sigma+|\ell^{(k)}|)\;q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\sum_{n_{k+1}=1}^{1}\Gamma(\alpha/\sigma+|\ell^{(k+1)}|)\;q_{\ell^{(k)}}\;\frac{C(1,1;\sigma)}{K^{\ell_{k+1}}}}\\ &= \frac{K(k+1)}{n\sigma(K-k)}\frac{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k}\Gamma(\alpha/\sigma+|\ell^{(k)}|)\;q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\Gamma(\alpha/\sigma+|\ell^{(k)}|+1)\;q_{\ell^{(k)}}}\\ &=: \frac{K(k+1)}{n\sigma(K-k)}(R_1+R_2).\end{aligned}$$

We separate the sum over ℓ_k in the numerator in two, R_1 corresponds to the first

$n_k - 1$ terms and R_2 to the last one. We compute separately R_1 and R_2.

$$\begin{aligned}
R_1 &= \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}| + 1)\ q_{\ell^{(k)}}} \\
&= \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} (\alpha/\sigma + |\ell^{(k)}|)\Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}} \\
&\leq \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}}{(\alpha/\sigma + k)\ \ \sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}} \\
&\leq \frac{1}{\alpha/\sigma + k}.
\end{aligned}$$

Using twice the fact that $k \mapsto C(n, k; \sigma)$ is non increasing for $k \in \{1, \dots, n\}$ (see Bystrova et al. 2021), so $C(n_k, 1; \sigma) \geq C(n_k, \ell_k; \sigma) \geq C(n_k, n_k; \sigma)$, and that $\Gamma(\alpha/\sigma + |\ell^{(k-1)}| + n_k) \leq \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k-1)}| + \ell_k + 1)$, we obtain

$$\begin{aligned}
R_2 &= \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=n_k}^{n_k} \Gamma(\alpha/\sigma + |\ell^{(k)}|)\ q_{\ell^{(k)}}}{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k)}| + 1)\ q_{\ell^{(k)}}} \\
&= \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} \Gamma(\alpha/\sigma + |\ell^{(k-1)}| + n_k)\ q_{\ell^{(k-1)}} \frac{C(n_k,n_k;\sigma)}{K^{n_k}}}{\sum_{(\ell_1,\dots,\ell_{k-1})} \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k-1)}| + \ell_k + 1)\ q_{\ell^{(k-1)}}\ \frac{C(n_k,\ell_k;\sigma)}{K^{\ell_k}}} \\
&\leq \frac{C(n_k, n_k; \sigma)}{K^{n_k}} \frac{K^{n_k-1}}{C(n_k, 1; \sigma)} \frac{\sum_{(\ell_1,\dots,\ell_{k-1})} q_{\ell^{(k-1)}} \Gamma(\alpha/\sigma + |\ell^{(k-1)}| + n_k)}{\sum_{(\ell_1,\dots,\ell_{k-1})} q_{\ell^{(k-1)}}\ \sum_{\ell_k=1}^{n_k-1} \Gamma(\alpha/\sigma + |\ell^{(k-1)}| + \ell_k + 1)} \\
&\leq \frac{C(n_k, n_k; \sigma)}{K\ C(n_k, 1; \sigma)} \leq \frac{1}{K}.
\end{aligned}$$

Finally, we have that

$$\frac{1}{n}\frac{p(A)}{p(B)} = \frac{K(k+1)}{n\sigma(K-k)}(R_1 + R_2) \leq \frac{K(k+1)}{n\sigma(K-k)} \left(\frac{1}{\alpha/\sigma + k} + \frac{1}{K} \right).$$

So Condition 2.1 is satisfied for the Pitman–Yor multinomial process.

(c) Normalized generalized gamma multinomial process: using (2.9) and following the same way as for the Pitman–Yor case, we have

$$
\begin{aligned}
\frac{1}{n}\frac{p(A)}{p(B)} &= \frac{1}{n}\frac{p(n_1,\dots,n_k)}{p(n_1,\dots,n_k-1,1)} \\
&= \frac{k+1}{n(K-k)}\left(\sum_{(\ell_1,\dots,\ell_k)}\frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)\left(\sum_{(\ell_1,\dots,\ell_{k+1})}\frac{V_{n,|\ell^{(k+1)}|}}{K^{|\ell^{(k+1)}|}}\prod_{i=1}^{k+1}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)^{-1} \\
&= \frac{k+1}{n(K-k)}\left(\sum_{(\ell_1,\dots,\ell_k)}\frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)\left(\sum_{(\ell_1,\dots,\ell_k)}\frac{V_{n,|\ell^{(k)}|+1}}{K^{|\ell^{(k)}|+1}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)^{-1} \\
&=: \frac{K(k+1)}{n(K-k)}(R_1+R_2).
\end{aligned}
$$

As in PYM (b) proof, we separate the sum over ℓ_k in the numerator in two, R_1 corresponds to the first n_k-1 terms and R_2 to the last one.

In the proof of Proposition 2.1, we have shown that the ratio $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 1}$ is bounded. Let $B \in \mathbb{R}_+^\star$ denote an upper bound of this sequence. Then

$$
\begin{aligned}
R_1 &= \left(\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)\left(\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\frac{V_{n,|\ell^{(k)}|+1}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)^{-1} \\
&\leq B\left(\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)\left(\sum_{(\ell_1,\dots,\ell_{k-1})}\sum_{\ell_k=1}^{n_k-1}\frac{V_{n,|\ell^{(k)}|}}{K^{|\ell^{(k)}|}}\prod_{i=1}^{k}\frac{C(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}\right)^{-1} \\
&\leq B.
\end{aligned}
$$

Combining $\frac{V_{n,|\ell^{(k-1)}|+n_k}}{K^{|\ell^{(k-1)}|+n_k}} \leq \sum_{\ell_k=1}^{n_k-1}\frac{V_{n,|\ell^{(k)}|+1}}{K^{|\ell^{(k)}|}}$ with similar arguments to the bounding of R_2 term in PYM (b) above yield $R_2 \leq \frac{1}{\sigma}$ Finally, we obtain

$$
\frac{1}{n}\frac{p(A)}{p(B)} \leq \frac{K(k+1)(\sigma B+1)}{n\sigma(K-k)},
$$

so Condition 2.1 is satisfied for the normalized generalized gamma multinomial processes.

Hence, there is inconsistency in the sense of Theorem 2.1 for the Pitman–Yor multinomial process, the Dirichlet multinomial process, and the NGGM process.

□

2.B Details on the results of Section 2.4

2.B.1 Theorem 1 of Rousseau and Mengersen (2011)

We recall the main result of Rousseau and Mengersen 2011, Theorem 1. This result holds under some conditions on the mixture density, the kernel and the prior of the mixture model. Conditions 2.2 and 2.3, stated previously, are conditions on the mixture density and on the prior density. Under the notations used here, we stated the conditions on the kernel density, which need to have some regularity, integrability and strong identifiability properties. As a reminder, $\theta_{1:K} = (\theta_1, \dots, \theta_K)$ denotes the set of component parameters, $w_{1:K} = (w_1, \dots, w_K)$ denotes the weights of the mixing measure, $f(\cdot \mid \theta_i)$ denotes a component specific kernel density and $G = \sum_i w_i \delta_{\theta_i}$ denotes the mixing measure. We have data observations $X_1, \dots, X_n$ assumed to be independent and identically distributed from a mixture model with K_0 components, where $K_0 < K$:

$$f_0^X(x) = \sum_{k=1}^{K_0} w_k^0 f(x \mid \theta_k^0), \qquad \theta_k^0 \in \Theta.$$

Condition 2.4 (Rousseau and Mengersen (2011), Assumption 2)**.** *The kernel function $\tilde{\theta} \in \Theta \to f(\cdot \mid \tilde{\theta})$ is three time differentiable and regular in the sense that for all $\tilde{\theta} \in \Theta$ the Fisher information matrix that is associated with $f(\cdot \mid \tilde{\theta})$ is positive definite at $\tilde{\theta}$. Denote by $\nabla f(x \mid \theta)$ and $\mathrm{D}^2 f(x \mid \theta)$ respectively the vector of the first derivatives and the matrix of second derivatives of $f(x \mid \theta)$ with respect to θ. Denote also by $\mathrm{D}^{(3)} f(x \mid \theta)$ the array whose components are $\frac{\partial^3 f(x|\theta)}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}}$.*

For all $i \leq K_0$, there exists $\delta > 0$ such that

$$\int f_0^X(x) \frac{\sup_{|\theta_i^0 - \theta| \leq \delta} f(x \mid \theta)^3}{\inf_{|\theta_i^0 - \theta| \leq \delta} f(x \mid \theta)^3} \mathrm{d}x < \infty, \quad \int f_0^X(x) \frac{\sup_{|\theta - \theta_i^0| \leq \delta} |\nabla f(x \mid \theta)|^3}{\inf_{|\theta_i^0 - \theta| \leq \delta} f(x \mid \theta)^3} \mathrm{d}x < \infty,$$

$$\int f_0^X(x) \frac{|\nabla f(x \mid \theta_i^0)|^4}{f_0^X(x)^4} \mathrm{d}x < \infty,$$

$$\int f_0^X(x) \frac{\sup_{|\theta - \theta_i^0| \leq \delta} |\mathrm{D}^2 f(x \mid \theta)|^2}{\inf_{|\theta_i^0 - \theta| \leq \delta} f(x \mid \theta)^2} \mathrm{d}x < \infty, \quad \int f_0^X(x) \frac{\sup_{|\theta - \theta_i^0| \leq \delta} |\mathrm{D}^{(3)} f(x \mid \theta)|^2}{\inf_{|\theta_i^0 - \theta| \leq \delta} f(x \mid \theta)} \mathrm{d}x < \infty.$$

Assume also that for all $i = 1, \dots, K_0$, $\theta_i^0 \in \mathrm{int}(\Theta)$ the interior of Θ.

Condition 2.5 (Rousseau and Mengersen (2011), Assumption 3)**.** *There exists $\Theta_0 \subset \Theta$ satisfying $\lambda(\Theta_0) > 0$, where $\lambda(A)$ denotes the Lebesgue measure of A, and for all $i \leq K_0$,*

$$d(\theta_i^0, \Theta_0) = \inf_{\theta \in \Theta_0} |\theta - \theta_i^0| > 0,$$

and such that for $\theta \in \Theta_0$ there exists a $\delta > 0$,

$$\int f_0^X(x) \frac{f(x \mid \theta)^4}{f_0(x)^4} \mathrm{d}x < \infty, \quad \int f_0^X(x) \frac{f(x \mid \theta)^3}{\sup_{|\theta' - \theta_i^0| \leq \delta} f(x \mid \theta')^3} \mathrm{d}x < \infty, \ \forall i \leq K_0.$$

Condition 2.6 (Rousseau and Mengersen (2011), Assumption 4)**.** *For all ordered partitions $\mathbf{t}$ of $\{1, \dots, K\}$ in $K_0{+}1$ clusters defined by the cardinality of each cluster,*

$\mathbf{t} = (t_i)_{i=0}^{K_0}$ *with* $0 = t_0 < t_1 < \cdots < t_{K_0} \leq K$*, let* $(w, \theta) = (w_1, \ldots, w_K, \theta_1, \ldots, \theta_K)$ *and write* (w, θ) *as* $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$*, where*

$$\phi_{\mathbf{t}} = ((\theta_j)_{j=1,\ldots,t_{K_0}}, (s_i)_{i=1,\ldots,K_0-1}, (w_j)_{j=t_{K_0}+1,\ldots,K}) \in \mathbb{R}^{dt_{K_0}+K_0+K+t_{K_0}-1},$$

$s_i = \sum_{j=t_{i-1}+1}^{t_i} w_j - w_i^0$*,* $i = 1, \ldots, K_0$*, and*

$$\psi_{\mathbf{t}} = ((q_j)_{j=1,\ldots,t_{K_0}}, \theta_{t_{K_0}+1}, \ldots, \theta_{t_K}), \quad q_i = w_i / \sum_{j=t_{i-1}+1}^{t_i} w_j, \text{ where } i \in \{t_{i-1}+1, \ldots, t_i\}.$$

We denote by $g^X(x)$ *the density associated to the parameterization* $(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})$ *of* (w, θ)*. Then*

$$(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T g^{X\prime} + \frac{1}{2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T g^{X\prime\prime}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) = 0 \quad \Leftrightarrow$$
$$\forall i \leq K_0,\ s_i = 0 \text{ and } \forall j \in \{t_{i-1}+1, \ldots, t_i\},\ q_j(\theta_j - \theta_j^0) = 0, \quad \forall i \geq t_{K_0}+1,\ w_i = 0.$$

Assuming also that if $\theta \notin \{\theta_1, \ldots \theta_k\}$ *then for all functions* $h(\cdot \mid \theta)$ *which are linear combinations of derivatives of* $f(\cdot \mid \theta)$ *of order less than or equal to 2 with respect to* θ*, and all functions* h_1 *which are also linear combinations of derivatives of* $f(\cdot \mid \theta_j)$*,* $j = 1, \ldots, K$*, and its derivatives of order less than or equal to 2, then* $ah(\cdot \mid \theta) + bh_1(\cdot) = 0$ *if and only if* $ah(\cdot \mid \theta) = bh_1(\cdot) = 0$*.*

This last condition can be extended to the non-compact cases if Θ is not compact as explained in Rousseau and Mengersen (2011).

We recall that d denotes the dimension of θ. Under the three conditions detailed above, Condition 2.2 and Condition 2.3, we can state the main result in Rousseau and Mengersen (2011) in the following theorem.

Theorem 2.3 (Rousseau and Mengersen (2011), Theorem 1)**.** *Under all the five conditions recalled previously that the prior distribution satisfies, let* $\mathcal{S}_K$ *be the set of permutations of* $\{1, \ldots, K\}$*,* $\alpha_{\max} = \max(\alpha_j,\ j \leq K)$ *and* $\alpha_{\min} = \min(\alpha_j,\ j \leq K)$*.*

(i) If $\alpha_{\max} < d/2$*, set* $\rho = [dK_0 + K_0 - 1 + \alpha_{\max}(K - K_0)] / (d/2 - \alpha_{\max})$*, then*

$$\lim_{M\to\infty} \limsup_n \left(\mathbb{E}_0^n \left[p \left\{ \min_{\sigma \in \mathcal{S}_K} \sum_{i=K_0+1}^{K} w_{\sigma(i)} > Mn^{-1/2} \log(n)^{q(1+\rho)} \,\middle|\, X_{1:n} \right\} \right] \right) = 0.$$

(ii) If $\alpha_{\min} > d/2$*, set* $\rho' = [dK_0 + K_0 - 1 + d(d - K_0)/2] / (\alpha_{\min} - d/2)(K - K_0)$*, then*

$$\lim_{\epsilon\to 0} \limsup_n \left(\mathbb{E}_0^n \left[p \left\{ \min_{\sigma \in \mathcal{S}_K} \sum_{i=K_0+1}^{K} w_{\sigma(i)} < \epsilon \log(n)^{-q(1+\rho')} \,\middle|\, X_{1:n} \right\} \right] \right) = 0.$$

2.B.2 Merge-Truncate-Merge Algorithm of Guha et al. (2021)

We recall the Merge-Truncate-Merge algorithm in Guha et al. (2021) used in Section 2.4.2. This algorithm is a post-processing procedure applied on a posterior sample of the mixing measure G. Applying this algorithm, a posterior contraction rate for the mixing measure under the Wasserstein metric, denoting ω_n, is mandatory. More precisely, we need G such that

$$p\left(G : W_r(G, G_0) \le \delta\omega_n \mid X_{1:n}\right) \xrightarrow{p_{G_0}} 1,$$

with $\omega_n = o(1)$ a vanishing rate, $r \ge 1$. We also need to choose a constant c used in the second stage of the algorithm. There is no explicit way of choosing this constant in Guha et al. (2021), we describe it as a regularisation parameter, which we illustrate in figure 2.6.

Algorithm 1 Recall of Merge-Truncate-Merge Algorithm (MTM) (Guha et al. 2021)

```
Input: Posterior sample $G = \sum_i w_i\delta_{\theta_i}$, rate $\omega_n$, constant $c$.
Output: Discrete measure $\tilde{G}$ and its number of supporting atoms $\tilde{k}$.
    {Stage 1: Merge procedure}
    Reorder atoms $\{\theta_1, \theta_2, \dots\}$ by simple random sampling without replacement with
    corresponding weights $\{w_1, w_2, \dots\}$, let $\tau_1, \tau_2, \dots$ denote the new indices and set
    $\mathcal{E} = \{\tau_j\}_j$ as the existing set of atoms.
    Sequentially for each index $\tau_j \in \mathcal{E}$, if there exists an index $\tau_i < \tau_j$ such that
    $\|\theta_{\tau_i} - \theta_{\tau_j}\| \le \omega_n$, then update $w_{\tau_i} = w_{\tau_i} + w_{\tau_j}$, and remove $\tau_j$ from $\mathcal{E}$.
    Collect $G' = \sum_{j:\,\tau_j\in\mathcal{E}} w_{\tau_j}\delta_{\theta_{\tau_j}}$, write $G'$ as $\sum_{i>1} q_i\delta_{\gamma_i}$ so that $q_1 \ge q_2 \ge \cdots$.
    {Stage 2: Truncate-Merge procedure}
    Set $\mathcal{A} = \{i : q_i > (c\omega_n)^r\}$ and $\mathcal{N} = \{i : q_i \le (c\omega_n)^r\}$.
    For each index $i \in \mathcal{A}$, if there is $j \in \mathcal{A}$ such that $j < i$ and $q_i\|\gamma_i - \gamma_j\|^r \le (c\omega_n)^r$,
    then remove $i$ from $\mathcal{A}$ and add it to $\mathcal{N}$.
    For each $i \in \mathcal{N}$, find the atom $\gamma_j$ among $j \in \mathcal{A}$ that is nearest to $\gamma_i$, update
    $q_j = q_j + q_i$.
    Return $\tilde{G} = \sum_{j\in\mathcal{A}} q_j\delta_{\gamma_j}$ and $\tilde{K} = |\mathcal{A}|$.
```

As recalled in Theorem 2.2, Guha et al. (2021) prove that the output $\tilde{K}$ of the MTM algorithm consistently estimates the number of clusters for any $c > 0$. This result holds under the assumption that there exists a contraction rate for the mixing measure. In order to have a contraction rate the kernel $f(\cdot \mid \theta)$ needs to satisfy some assumptions presented below.

Condition 2.7 (Second-order identifiability)**.** *The family of densities* $\{f(\cdot \mid \theta),\ \theta \in \Theta\}$ *is identifiable in the second order if* $f(x \mid \theta)$ *is twice differentiable in* θ *and for* *any finite k different $\theta_1, \dots, \theta_k \in \Theta$, the equality*

$$\sup_x \left| \sum_{j=1}^{k} (\alpha_j f(x \mid \theta_j) + \beta_j^T \frac{\partial f}{\partial \theta}(x \mid \theta_j) + \gamma_j^T \frac{\partial^2 f}{\partial \theta^2}(x \mid \theta_j) \gamma_j \right| = 0$$

implies that $\alpha_j = 0 \in \mathbb{R}$, $\beta_j = \gamma_j = 0 \in \mathbb{R}^d$ for $j = 1, \dots, k$.

Condition 2.8 (Uniform Lipschitz-continuity). *The family of densities $\{f(\cdot \mid \theta),\ \theta \in \Theta\}$ is uniformly Lipschitz continuous up to the second order if there exists a positive constant δ such that for any $R > 0$ $\|\theta\| \le R$, $\gamma \in \mathbb{R}^d$, $\theta_1, \theta_2 \in \Theta$, there is a positive constant $C > 0$ depending on R such that for all $x \in \mathcal{X}$*

$$\left| \gamma^T \left(\frac{\partial^2 f}{\partial \theta^2}(x \mid \theta_1) - \frac{\partial^2 f}{\partial \theta^2}(x \mid \theta_2) \right) \gamma \right| \le C \|\theta_1 - \theta_2\|_1^\delta \|\gamma\|_2^2.$$

For more details on these conditions and on contraction rate see Chen (1995); Ho and Nguyen (2016).

2.B.3 Theorem 1 of Scricciolo (2014)

In Section 2.4, we introduce Lemma 2.1 which is a corollary of Theorem 1 in Scricciolo (2014). This theorem gives a posterior contraction rate for the mixing measure of a Pitman–Yor mixture model relative to the L^p-metric. We detailed below the conditions for this theorem and then we recalled the result.

Here, we assume that $\Theta \subset \mathbb{R}$. The model is less general than in the rest of the paper, we assume that the model is a location mixture defined as:

$$f^X(x) = \int f(x \mid \theta, \tau) G(\mathrm{d}\theta) = \int \tau^{-1} f((x - \theta_k)/\tau) G(\mathrm{d}\theta), \qquad (2.14)$$

where τ is a scale parameter and $f(\cdot \mid \theta) = f(\cdot)$ denotes the kernel density. In this Section, we will assume that the scale parameter τ_0 is known as the true scale parameter τ_0. This can also be seen as τ following the prior distribution δ_{τ_0}.

The theorem holds under three conditions on the kernel density, the true mixing measure G_0 and the base measure of the Pitman–Yor process.

Condition 2.9 (Scricciolo (2014), Assumption A1). *The kernel probability density $f(\cdot \mid \theta) : \mathbb{R} \to \mathbb{R}^+$ is symmetric around 0, monotone decreasing in $|x|$ and satisfies the tail condition $f(x \mid \theta) \gtrsim e^{-c|x|^\kappa}$ as $|x| \to \infty$, for some constants $0 < c$, $\kappa < \infty$.*

Condition 2.10 (Scricciolo (2014), Assumption A2). *The true mixing measure G_0 satisfies the tail condition $G_0(\theta : |\theta| > t) \gtrsim e^{-c_0 t^\varpi}$ as $t \to \infty$, for some constants $0 < c_0 < \infty$ and $0 < \varpi \le \infty$.*

Condition 2.11 (Scricciolo (2014), Assumption A3)**.** *The base measure P has a continuous and positive density p' on $\mathbb{R}$ such that $p'(\theta) \propto e^{-b|\theta|^{\delta}}$ as $|\theta| \to \infty$, for some constants $0 < b$, $\delta < \infty$.*

We also introduce the following set,

$$\mathcal{A}^{\rho,\eta,L} := \left\{ f : \mathbb{R} \to \mathbb{R}^{+} \mid \|f\|_1 = 1, \int e^{2(\rho|t|)^{\eta}} |\hat{f}(t)|^2 \mathrm{d}t \leq 2pL^2 \right\},$$

where $\hat{f}$ denotes the Fourier transform of f and ρ, L, η are some positive constants.

We can now recall Theorem 1 from Scricciolo (2014). Here, we state a simplified version of the theorem as we assumed the scale parameter to be known. The general statement requires additional conditions on the scale parameter.

Theorem 2.4 (Scricciolo (2014), Theorem 1)**.** *Let $f(\cdot \mid \theta) \in \mathcal{A}^{\rho,\eta,L}(\mathbb{R})$, $0 < \rho, \eta, L < \infty$, be as in Condition 2.9. Suppose that the true mixture density f_0^X, with*

(i) G_0 satisfying Condition 2.10 for some constants $0 < c_0 < \infty$ and given numbers $0 < \kappa$, $\eta < \infty$ ϖ be such that $0 < \max\left\{\kappa,\ [1 + \nVdash_{(1,\infty)}(\eta)/(\eta - 1)]\right\} \leq \varpi \leq \infty$.

Let $G \sim \mathrm{PY}(\alpha, \sigma; P)$, with $0 \leq \sigma < 1$, $-\sigma < \alpha < \infty$ and a base measure P. Assume that

(ii) P satisfies Condition 2.11 for constants $0 < b$, $\delta < \infty$, with $\delta \leq \varpi$ when $\varpi < \infty$;

Then, the posterior contraction rate relative to the L^p-metric, $1 \leq p \leq \infty$, denoted by $\omega_{n,p}$, is $n^{-1/2} \log(n)^{\mu}$, with a constant $0 < \mu\infty$ possibly depending on p.

2.C Proofs of the results of Section 2.4

Proof of Proposition 2.3. In the Dirichlet multinomial process case, the prior on the weights $w_{1:K} = (w_1, \ldots, w_K)$ is a finite-dimensional Dirichlet distribution which is of the form

$$p(w_{1:K}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} w_1^{\alpha/K-1} w_2^{\alpha/K-1} \cdots w_K^{\alpha/K-1} \mathbb{I}(w_{1:K} \in \Delta_K),$$

where Δ_K denotes the K-dimensional simplex. So, the prior is of the same form as in Condition 2.3 with $C(w_{1:K}) = \Gamma(\alpha)/\Gamma(\alpha/K)^K\, \mathbb{I}(w_{1:K} \in \Delta_K)$ which is a constant on the simplex. Condition 2.2 is verified using Theorem 4.1 from Rousseau et al. (2019) which can also be applied to overfitted mixtures. Hence, the result Rousseau and Mengersen (2011) applies in this case.

In the Pitman–Yor multinomial case, the prior on the weights is a ratio-stable distribution defined in Carlton (2002) and denoted by $w_{1:K} \sim RS(\sigma, \tilde{\alpha}; 1/K, \dots, 1/K)$. In the general case, no density is available so it is not possible to satisfy 2.3 and the Rousseau et al. (2019) results cannot give us any guarantee. In the interesting $\sigma = 1/2$ case, the density is

$$p(w_{1:K}) = \frac{(1/K)^K}{p^{\frac{K-1}{2}}} \frac{\Gamma(\tilde{\alpha} + K/2)}{\Gamma(\tilde{\alpha} + 1/2)} \frac{w_1^{-3/2} \cdots w_K^{-3/2}}{\left(\frac{1}{w_1 K^2} + \cdots + \frac{1}{w_K K^2}\right)^{\tilde{\alpha}+K/2}} \mathbb{I}(w_{1:K} \in \Delta_K).$$

To write this density in the form $p(w_{1:K}) = C(w_{1:K}) \prod_{i=1}^K w_i^{\alpha_i - 1}$ we must set:

$$C(w) \propto \frac{\prod_{i=1}^K w_i^{-\alpha_i - 1/2}}{\left(\sum_{i=1}^K w_i^{-1}\right)^{\tilde{\alpha}+K/2}}$$

Condition 2.3 from requires $C(w_{1:K})$ to be bounded above and below. A necessary condition is obtained by studying the limit of $C(w_{1:K})$ for any $w_i \to 0$ with the others remaining bounded away from 0.

$$\text{As } w_i \to 0;\ C(w_{1:K}) = \mathcal{O}\left(w_i^{\tilde{\alpha} - \alpha_i + \frac{K-1}{2}}\right)$$

For $C(w_{1:K})$ to remain bounded above and below in this limit requires $\alpha_i = \tilde{\alpha} + \frac{K-1}{2}$. Enforcing this necessary condition for each α_i independently requires rewriting $C(w)$ as:

$$\begin{aligned} C(w_{1:K}) &\propto \frac{\prod_{i=1}^K w_i^{-(\alpha+K/2)}}{\left(\sum_{i=1}^K w_i^{-1}\right)^{\tilde{\alpha}+K/2}} \\ &\propto \frac{1}{(w_1 \cdots w_K)^{\tilde{\alpha}+K/2}} \times \left(\frac{w_1 \cdots w_K}{w_2 \cdots w_K + w_1 w_3 \cdots w_K + \cdots + w_1 \cdots w_{K-1}}\right)^{\tilde{\alpha}+K/2} \\ &\propto \left(\frac{1}{w_2 \cdots w_K + w_1 w_3 \cdots w_K + \cdots + w_1 \cdots w_{K-1}}\right)^{\tilde{\alpha}+K/2} \end{aligned}$$

This quantity is bounded from below for all $\tilde{\alpha} > 0$, for any $w_i \to 0$ independently, but not for two $w_i, w_j \to 0$ at different rates, in which case all terms at the denominator vanish and $C(w_{1:K})$ diverges. We have found a set of necessary conditions which are incompatible, hence there is no choice of α_i such that Condition 2.3 can be satisfied, and in this case too the Rousseau et al. (2019) results cannot give us any guarantee. □

Proof of Lemma 2.1. This is a direct application of Corollary 1 from Scricciolo (2014). To apply this corollary, we must check that the kernel $f(\cdot \mid \theta)$ associated with the mixing measure G is a symmetric probability density such that, for some constants $0 < \rho < \infty$ and $0 < \eta \leq 2$, the Fourier transform $\hat{f}$ of $f(\cdot \mid \theta)$ satisfies:

$$|\hat{f}(t)| \sim e^{-(\rho|t|)^\eta} \text{ as } |t| \to \infty.$$

This is satisfied by assumption. In Condition 2.9, the kernel $f(\cdot \mid \theta)$ is assumed to be symmetric, monotone decreasing in $|x|$ and to satisfy a tail condition. The kernel $f(\cdot \mid \theta)$ also belongs to the set

$$\mathcal{A}^{\rho,\eta,L} := \left\{ f : \mathbb{R} \to \mathbb{R}^+ \mid \|f\|_1 = 1, \int e^{2(\rho|t|)^\eta} |\hat{f}(t)|^2 \mathrm{d}t \leq 2pL^2 \right\},$$

where $\hat{f}$ denotes the Fourier transform of f and ρ, L, η are some positive constants.

We also need to check that for a sequence $\tilde{\varepsilon}_n > 0$ such that $\tilde{\varepsilon}_n \to 0$ as $n \to \infty$ and $n\tilde{\varepsilon}_n^2 \gtrsim \log(n)^{1/\eta}$, we have

$$p(B_{\mathrm{KL}}(f_0^X; \tilde{\varepsilon}_n^2)) \gtrsim \exp(-Cn\tilde{\varepsilon}_n^2) \text{ for some constant } 0 < C < \infty,$$

where $B_{\mathrm{KL}}(f_0^X; \varepsilon^2) := \left\{ f : \int f_0^X \log(f_0^X/f) \leq \varepsilon^2, \int f_0^X (\log(f_0^X/f))^2 \leq \varepsilon^2 \right\}$ denotes Kullback–Leibler neighbourhoods of the true density f_0^X. This condition is verified in the second part of the proof of Theorem 1 in Scricciolo (2014). □

2.D Details on the simulation study of Section 2.5

We consider the mixture model, with $K = 10$:

$$f^X(x) = \sum_{k=1}^{K} w_k f(x \mid \mu_k, \Sigma_k).$$

Parameters have the following prior distributions:

$$\begin{aligned} (w_1, \ldots, w_K) &\sim \mathrm{Dir}_K(\bar{\alpha}, \ldots, \bar{\alpha}), \quad \bar{\alpha} = \alpha/K, \\ \mu_k &\sim \mathcal{N}(b_0, B_0), \quad k = 1, \ldots, K, \\ \Sigma_k^{-1} &\sim \mathcal{W}(c_0, C_0), \quad C_0 \sim \mathcal{W}(d_0, D_0). \end{aligned}$$

Parameters for Wishart distribution are defined as in Malsiner-Walli et al. (2016): $c_0 = 2.5 + \frac{r-1}{2}$, $d_0 = 0.5 + \frac{r-1}{2}$, $D_0 = \frac{100\, d_0}{c_0} \mathrm{diag}(1/R_1^2, \ldots, 1/R_r^2)$, and $B_0 = \mathrm{diag}(R_1^2, \ldots, R_r^2)$, where r is dimension of Σ matrix, and R_j is the range of the data in each dimension. Parameter b_0 is set to the median of the data.

We run two MCMC chains of 20 000 iterations each, with 10 000 burn-in iterations. Convergence assessment was done through the calculation of Gelman-Rubin diagnostics (Gelman and Rubin 1992) and visual inspection of the trace plots.

We provide here some more details on Figure 2.8 in Discussion. Figure 2.8 illustrates two different cases where the parameter $\bar{\alpha}$ is not fixed. First, we consider the fixed prior expected number of clusters, such as $\mathbb{E}[K_n] = 5$, which leads to decreasing of the parameter $\bar{\alpha}$ with n. Posterior distribution of the number of clusters is presented in Figure 2.8 (a). The second approach consists in using the hyperprior for parameter $\bar{\alpha}$. We consider $\bar{\alpha} \sim \text{Ga}(a, bK)$, where parameters $a = 1$, $b = 0.1$ and $K = 10$ is the number of components, which leads to less informative prior distribution of the number of clusters. This simulation setting is also different from theoretical assumptions required by Ascolani et al. (2022).

2.E Details on the real-data analysis of Section 2.6

We consider two different mixture models in Section 2.6. The first one is of the form:

$$f^X(x) = \int_\Theta f(x \mid \theta) G(\mathrm{d}\theta),$$

where $\theta_k = (\mu_k, \tilde{\sigma}_k^2)$ and $f(x \mid \theta) = \mathcal{N}(x \mid \mu, \tilde{\sigma}^2)$. Here, the mixing measure is distributed as a Pitman–Yor process, $G \sim \text{PY}(\alpha, \sigma; P)$, where P is the base measure defined hierarchically as the following:

$$\begin{aligned} \tilde{\sigma}_k^2 &\overset{\text{iid}}{\sim} \text{IG}(a_0, b_0), \qquad k > 1, \\ \mu_k \mid \tilde{\sigma}_k^2 &\overset{\text{ind}}{\sim} \mathcal{N}(m_0, \tilde{\sigma}_k^2), \qquad k > 1. \end{aligned}$$

IG denotes the Inverse-Gamma distribution. Parameters for the Inverse-Gamma distribution are defined as the default values of the `BNPmix` package (see Corradin et al. 2021): $a_0 = 2$, b_0 is set as the sample variance of the data and m_0 as the sample mean of the data. We used various values of $\alpha \in \{0.01, 0.5\}$ and $\sigma \in \{0.1, 0.25\}$

We run four MCMC chains of 20 000 iterations each, with 10 000 burn-in iterations using the marginal sampler of the `BNPmix` package.

The second model is of the following form:

$$f^X(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x \mid \mu_k, \tilde{\sigma}_k^2),$$

with $K = 10$. In this case, the mixing measure is distributed as a Dirichlet multinomial process. The parameters have the following prior distributions:

$$
\begin{aligned}
(w_1, \dots, w_K) &\sim \mathrm{Dir}_K(\bar{\alpha}, \dots, \bar{\alpha}), \quad \bar{\alpha} = \alpha/K, \\
\mu_k &\sim \mathcal{N}(b_0, B_0), \quad k = 1, \dots, K, \\
\tilde{\sigma}_k^2 &\sim \mathrm{IG}(c_0, C_0), \quad C_0 \sim \mathrm{Ga}(h_0, H_0).
\end{aligned}
$$

The parameters for the Gamma distribution are defined as previously but in a univariate form: $c_0 = 2.5$, $d_0 = 0.5$, $D_0 = \frac{100\, d_0}{c_0\, R^2}$, and $B_0 = R^2$ where R is the range of the data. Parameter b_0 is set to the median of the data. We used various values of $\bar{\alpha} = \alpha/K \in \{0.01, 0.5, 1, 2\}$.

We run two MCMC chains of 80 000 iterations each, with 30 000 burn-in iterations. For both models, convergence assessment was done through the calculation of Gelman–Rubin diagnostics (Gelman and Rubin 1992) and visual inspection of the trace plots.

References

Alamichel, L., D. Bystrova, J. Arbel, and G. Kon Kam King (2024). “Bayesian mixture models (in)consistency for the number of clusters”. In: *Scandinavian Journal of Statistics* (cit. on p. 32).

Arbel, J. and S. Favaro (2021). “Approximating predictive probabilities of Gibbs-type priors”. In: *Sankhya A* 83.1, pp. 496–519 (cit. on p. 66).

Arbel, J., S. Favaro, B. Nipoti, and Y. W. Teh (2017). “Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptiotics”. In: *Statistica Sinica*, pp. 839–858 (cit. on p. 36).

Argiento, R. and M. De Iorio (2022). “Is infinity that far? A Bayesian nonparametric perspective of finite mixture models”. In: *The Annals of Statistics* (cit. on pp. 34, 36).

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). “Clustering consistency with Dirichlet process mixtures”. In: *Biometrika. In press* (cit. on pp. 37, 62, 77).

Attorre, F., V. E. Cambria, E. Agrillo, N. Alessi, M. Alfò, M. De Sanctis, L. Malatesta, T. Sitzia, R. Guarino, C. Marcenò, et al. (2020). “Finite Mixture Model-based classification of a complex vegetation system”. In: *Vegetation Classification and Survey* 1, p. 77 (cit. on p. 34).

Bacallado, S., M. Battiston, S. Favaro, and L. Trippa (2017). “Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations”. In: *Statistical Science* 32.4, pp. 487–500 (cit. on p. 35).

Bystrova, D., J. Arbel, G. Kon Kam King, and F. Deslandes (2021). “Approximating the clusters’ prior distribution in Bayesian nonparametric models”. In: *Third Symposium on Advances in Approximate Bayesian Inference* (cit. on p. 68).

Cai, D., T. Campbell, and T. Broderick (2021). “Finite mixture models do not reliably learn the number of components”. In: *International Conference on Machine Learning.* PMLR, pp. 1158–1169 (cit. on p. 37).

Carlton, M. A. (2002). “A family of densities derived from the three-parameter Dirichlet process”. In: *Journal of applied probability* 39.4, pp. 764–774 (cit. on p. 75).

Caron, F. and E. B. Fox (2017). “Sparse graphs using exchangeable random measures”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.5, pp. 1295–1366 (cit. on p. 36).

Celeux, G., M. Hurn, and C. P. Robert (2000). “Computational and inferential difficulties with mixture posterior distributions”. In: *Journal of the American Statistical Association* 95.451, pp. 957–970 (cit. on p. 34).

Cesari, O., S. Favaro, and B. Nipoti (2014). “Posterior analysis of rare variants in Gibbs-type species sampling models”. In: *Journal of Multivariate Analysis* 131, pp. 79–98 (cit. on p. 36).

Chambaz, A. and J. Rousseau (2008). “Bounds for Bayesian order identification with application to mixtures”. In: *The Annals of Statistics* 36.2, pp. 938–962 (cit. on p. 37).

Chaumeny, Y., J. van der Molen Moris, A. C. Davison, and P. D. W. Kirk (2022). *Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice?* arXiv: 2207.14717 (cit. on p. 63).

Chen, J. (1995). “Optimal Rate of Convergence for Finite Mixture Models”. In: *The Annals of Statistics* 23.1. Publisher: Institute of Mathematical Statistics, pp. 221–233 (cit. on p. 73).

Corradin, R., A. Canale, and B. Nipoti (2021). “BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures”. In: *Journal of Statistical Software* 100, pp. 1–33 (cit. on pp. 59, 77).

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Pruenster, and M. Ruggiero (2015). “Are Gibbs-type priors the most natural generalization of the Dirichlet process?” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 212–229 (cit. on pp. 35, 40).

De Blasi, P. and M. F. Gil–Leyva (2023). “Gibbs Sampling for Mixtures in Order of Appearance: The Ordered Allocation Sampler”. In: *Journal of Computational and Graphical Statistics* 32.4, pp. 1416–1424 (cit. on p. 62).

Dudley, C. R., L. A. Giuffra, A. E. Raine, and S. T. Reeders (1991). “Assessing the role of APNH, a gene encoding for a human amiloride-sensitive Na+/H+ antiporter, on the interindividual variation in red cell Na+/Li+ countertransport.” In: *Journal of the American Society of Nephrology* 2.4, p. 937 (cit. on p. 58).

Favaro, S., A. Lijoi, R. H. Mena, and I. Prünster (2009). “Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 993–1008 (cit. on p. 36).

Favaro, S., A. Lijoi, and I. Prünster (2012). “A new estimator of the discovery probability”. In: *Biometrics* 68.4, pp. 1188–1196 (cit. on p. 36).

Ferguson, T. S. (1973). “A Bayesian analysis of some nonparametric problems”. In: *The Annals of Statistics* 1.2, pp. 209–230 (cit. on p. 35).

Fraley, C. and A. E. Raftery (2002). “Model-Based Clustering, Discriminant Analysis, and Density Estimation”. In: *Journal of the American Statistical Association* 97.458, pp. 611–631 (cit. on p. 33).

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Vol. 425. Springer (cit. on p. 34).

Frühwirth-Schnatter, S., G. Celeux, and C. P. Robert, eds. (2019). *Handbook of Mixture Analysis*. CRC Press, Taylor & Francis Group (cit. on pp. 34, 35).

Frühwirth-Schnatter, S., G. Malsiner-Walli, and B. Grün (2021). "Generalized Mixtures of Finite Mixtures and Telescoping Sampling". In: *Bayesian Analysis* 16.4, pp. 1279–1307 (cit. on pp. 34, 36, 62).

Frühwirth-Schnatter, S., C. Pamminger, A. Weber, and R. Winter-Ebmer (2012). "Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering". In: *Journal of Applied Econometrics* 27.7, pp. 1116–1137 (cit. on p. 34).

Gelman, A. and D. B. Rubin (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical Science* 7.4, pp. 457–472 (cit. on pp. 76, 78).

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999). "Posterior consistency of Dirichlet mixtures in density estimation". In: *The Annals of Statistics* 27.1, pp. 143–158 (cit. on p. 36).

Ghosal, S. and A. van der Vaart (2007). "Posterior convergence rates of Dirichlet mixtures at smooth densities". In: *The Annals of Statistics* 35.2, pp. 697–723 (cit. on p. 36).

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press (cit. on pp. 38, 44).

Gnedin, A. and J. Pitman (2006). "Exchangeable Gibbs partitions and Stirling triangles". In: *Journal of Mathematical Sciences* 138.3, pp. 5674–5685 (cit. on pp. 35, 40).

Greve, J., B. Grün, G. Malsiner-Walli, and S. Frühwirth-Schnatter (2022). "Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis". In: *Australian & New Zealand Journal of Statistics* 64.2, pp. 205–229 (cit. on p. 34).

Guha, A., N. Ho, and X. Nguyen (2021). "On posterior contraction of parameters and interpretability in Bayesian mixture modeling". In: *Bernoulli* 27.4, pp. 2159–2188 (cit. on pp. 37, 38, 47, 48, 50–52, 54, 57–62, 72).

Ho, N. and X. Nguyen (2016). "On strong identifiability and convergence rates of parameter estimation in finite mixtures". In: *Electronic Journal of Statistics* 10.1, pp. 271–307 (cit. on pp. 38, 50, 52, 73).

Ishwaran, H. and L. F. James (2001). "Gibbs Sampling Methods for Stick-Breaking Priors". In: *Journal of the American Statistical Association* 96.453, pp. 161–173 (cit. on p. 36).

Ishwaran, H. and M. Zarepour (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models". In: *Biometrika* 87.2, pp. 371–390 (cit. on p. 41).

Ishwaran, H. and M. Zarepour (2002). “Exact and approximate sum representations for the Dirichlet process”. In: *Canadian Journal of Statistics* 30.2, pp. 269–283 (cit. on p. 36).

James, L. F., A. Lijoi, and I. Prünster (2009). “Posterior analysis for normalized random measures with independent increments”. In: *Scandinavian Journal of Statistics* 36.1, pp. 76–97 (cit. on p. 42).

Jara, A., E. Lesaffre, M. D. Iorio, and F. Quintana (2010). “Bayesian semiparametric inference for multivariate doubly-interval-censored data”. In: *The Annals of Applied Statistics* 4.4, pp. 2126–2149 (cit. on p. 36).

Kruijer, W., J. Rousseau, and A. van der Vaart (2010). “Adaptive Bayesian density estimation with location-scale mixtures”. In: *Electronic Journal of Statistics* 4.none, pp. 1225–1257 (cit. on pp. 36, 38).

Legramanti, S., T. Rigon, D. Durante, and D. B. Dunson (2022). “Extended stochastic block models with application to criminal networks”. In: *The Annals of Applied Statistics* 16.4, pp. 2369–2395 (cit. on p. 36).

Lijoi, A., R. H. Mena, and I. Prünster (2005a). “Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors”. In: *Journal of the American Statistical Association* 100.472, pp. 1278–1291 (cit. on p. 36).

Lijoi, A., R. H. Mena, and I. Prünster (2007a). “Controlling the reinforcement in Bayesian non-parametric mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4, pp. 715–740 (cit. on pp. 36, 40).

Lijoi, A., R. H. Mena, and I. Prünster (2007b). “Bayesian Nonparametric Estimation of the Probability of Discovering New Species”. In: *Biometrika* 94.4, pp. 769–786 (cit. on p. 36).

Lijoi, A. and I. Prünster (2010). “Models beyond the Dirichlet process”. In: *Bayesian Nonparametrics*. Ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, pp. 80–136 (cit. on p. 35).

Lijoi, A., I. Prünster, and T. Rigon (2020). “The Pitman–Yor multinomial process for mixture modelling”. In: *Biometrika* 107.4, pp. 891–906 (cit. on pp. 36, 41, 42).

Lijoi, A., I. Prünster, and T. Rigon (2024). “Finite-dimensional Discrete Random Structures and Bayesian Clustering”. In: *Journal of the American Statistical Association* 119.546, pp. 929–941 (cit. on pp. 36, 41, 42, 61).

Lijoi, A., I. Prünster, and S. G. Walker (2005b). “On consistency of nonparametric normal mixtures for Bayesian density estimation”. In: *Journal of the American Statistical Association* 100.472, pp. 1292–1296 (cit. on p. 36).

Lo, A. Y. (1984). “On a class of Bayesian nonparametric estimates: I. Density estimates”. In: *The Annals of Statistics*, pp. 351–357 (cit. on p. 35).

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). "Model-based clustering based on sparse finite Gaussian mixtures". In: *Statistics and Computing* 26.1-2, pp. 303–324 (cit. on pp. 37, 52, 76).

Miller, J. W. (2014). "Nonparametric and Variable-Dimension Bayesian Mixture Models: Analysis, Comparison, and New Methods". PhD thesis. Brown University, Division of Applied Mathematics (cit. on pp. 58, 59).

Miller, J. W. (2023). "Consistency of mixture models with a prior on the number of components". In: *Dependence Modeling* 11.1 (cit. on p. 36).

Miller, J. W. and D. B. Dunson (2019). "Robust Bayesian Inference via Coarsening". In: *Journal of the American Statistical Association* 114.527, pp. 1113–1125 (cit. on p. 37).

Miller, J. W. and M. T. Harrison (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components". In: *The Journal of Machine Learning Research* 15.1, pp. 3333–3370 (cit. on pp. 32, 37–39, 44–46, 62, 65).

Miller, J. W. and M. T. Harrison (2018). "Mixture Models With a Prior on the Number of Components". In: *Journal of the American Statistical Association* 113.521, pp. 340–356 (cit. on pp. 36, 62).

Muliere, P. and P. Secchi (1995). "A note on a proper Bayesian bootstrap". In: (cit. on p. 41).

Muliere, P. and P. Secchi (2003). "Weak Convergence of a Dirichlet-Multinomial Process". In: *Georgian Mathematical Journal* 10.2, pp. 319–324 (cit. on p. 41).

Müller, P., A. Erkanli, and M. West (1996). "Bayesian curve fitting using multivariate normal mixtures". In: *Biometrika* 83.1, pp. 67–79 (cit. on p. 33).

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models". In: *The Annals of Statistics* 41.1, pp. 370–400 (cit. on pp. 37, 38, 47, 50, 52).

Nobile, A. (1994). "Bayesian Analysis of Finite Mixture Distributions". PhD thesis. Pittsburgh, PA: Department of Statistics, Carnegie Mellon University (cit. on pp. 35, 36, 38).

Ohn, I. and L. Lin (2023). "Optimal Bayesian estimation of Gaussian mixtures with growing number of components". In: *Bernoulli* 29.2, pp. 1195–1218 (cit. on p. 62).

Petralia, F., V. Rao, and D. Dunson (2012). "Repulsive mixtures". In: *Advances in Neural Information Processing Systems* 25 (cit. on p. 62).

Pitman, J. (2003). "Poisson-Kingman partitions". In: *Statistics and science: a Festschrift for Terry Speed* 40. Publisher: Institute of Mathematical Statistics, pp. 1–35 (cit. on p. 40).

Ramírez, V. M., F. Forbes, J. Arbel, A. Arnaud, and M. Dojat (2019). "Quantitative MRI Characterization of Brain Abnormalities in DE NOVO Parkinsonian

Patients". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1572–1575 (cit. on p. 34).

Regazzini, E., A. Lijoi, and I. Prünster (2003). "Distributional results for means of normalized random measures with independent increments". In: *The Annals of Statistics* 31.2, pp. 560–585 (cit. on pp. 41, 42).

Richardson, S. and P. J. Green (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4, pp. 731–792 (cit. on p. 35).

Rousseau, J., C. Grazian, and J. E. Lee (2019). "Bayesian mixture models: Theory and methods". In: *Handbook of Mixture Analysis*. Ed. by S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. Chapman and Hall/CRC, pp. 53–72 (cit. on pp. 38, 49, 74, 75).

Rousseau, J. and K. Mengersen (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, pp. 689–710 (cit. on pp. 37, 38, 47–49, 52–55, 61, 69–71, 74).

Scricciolo, C. (2014). "Adaptive Bayesian Density Estimation in L_p-metrics with Pitman-Yor or Normalized Inverse-Gaussian Process Kernel Mixtures". In: *Bayesian Analysis* 9.2 (cit. on pp. 47, 50, 51, 73–76).

Teh, Y. W. and M. I. Jordan (2010). "Hierarchical Bayesian nonparametric models with applications". In: *Bayesian nonparametrics* 1, pp. 158–207 (cit. on p. 36).

Ullah, I. and K. Mengersen (2019). "Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data". In: *Journal of Big Data* 6.1, pp. 1–25 (cit. on p. 34).

Wade, S. and Z. Ghahramani (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)". In: *Bayesian Analysis* 13.2. Publisher: International Society for Bayesian Analysis, pp. 559–626 (cit. on p. 63).

Xie, F. and Y. Xu (2020). "Bayesian Repulsive Gaussian Mixture Model". In: *Journal of the American Statistical Association* 115.529, pp. 187–203 (cit. on p. 62).

Chapter 3

Pitman–Yor mixture models with a prior on the precision inconsistency in the number of clusters

In Chapter 2, we first proved some inconsistency results for the number of clusters of mixture models, and then studied some solutions to the inconsistency. In the same context, we studied the consistency of the number of clusters of mixtures of the Pitman-Yor process with an a priori on the precision parameter, which is motivated by results in Ascolani et al. (2022). This Chapter is a joint work with Caroline Lawless, Julyan Arbel, and Guillaume Kon Kam King. Caroline Lawless and I contributed equally to this work. All authors contributed to the theoretical development of the paper. I conducted the simulation study. Caroline Lawless wrote the first draft, while all authors contributed to the writing of the final version. This Chapter present a generalisation of the following paper:

> C. Lawless, L. Alamichel, J. Arbel, and G. Kon Kam King (2023). “Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter”. In: *Fifth Symposium on Advances in Approximate Bayesian Inference*

Contents

Abstract Bayesian nonparametric (BNP) mixture models are widely used for handling complex data. While considerable research has focused on establishing the convergence of their posterior distributions to the true data-generating distribution at the optimal minimax rate, it is important to note that the consistency of the posterior distribution does not guarantee the consistency of the inferred number of clusters. Until recently, there has been a lack of asymptotic guarantees regarding the posterior number of clusters for these models.

Recent investigations have revealed that these models may exhibit inconsistency in estimating the number of clusters. Although placing a prior on the concentration hyperparameter α, particularly in Dirichlet process mixture models, has been a common strategy to address this issue, it has been found that Pitman–Yor process mixture models can still suffer from inconsistency in the number of clusters, particularly in scenarios where the discount parameter σ is a fixed constant within the range $(0, 1)$. This work provides a rigorous proof of this inconsistency in Pitman–Yor process mixture models under the condition of a fixed constant discount parameter σ in the range $(0, 1)$, despite the incorporation of a prior on α.

Keywords— mixture models, Bayesian nonparametric, number of clusters, Pitman–Yor process.

3.1 Introduction

Mixture models, popular for their flexibility and simplicity, are commonly used in the statistical analysis of heterogeneous data where observations are assumed to come from an unknown number of different populations. Since in a mixture, each observation is assumed to come from one population, such models naturally induce a clustering: two data points belong to the same cluster if they come from the same population. We focus on the problem of inferring the number of clusters in the data.

One solution is to fit mixture models with an increasing number of components and select the best model using the Akaike information criterion (AIC), the Bayes information criterion (BIC), etc. This method, however, may be computationally expensive since many models must be fitted. A Bayesian approach could alternatively be taken by putting a parametric prior (such as a Poisson) on the number of components, but inference can be challenging when the dimensionality or the amount of data becomes large (although new strategies have been proposed recently Miller and Harrison (2018)).

In this work, we consider infinite mixture models where the mixing measure is modelled with a nonparametric prior. In such models, the number of components possible has no upper bound. Inference may be performed in a unified way without the need for strong assumptions on the number of components and with no need to fit multiple models.

While the most standard nonparametric prior remains the Dirichlet process (DP) introduced by Ferguson (1973), many extensions now exist. In this work, we focus on the Pitman–Yor process (PY) (Pitman and Yor 1997), a natural extension of the DP with an extra parameter increasing model flexibility. Compared with DP mixtures, PY mixtures are better suited when the sizes of clusters are more evenly distributed. Due to the interpretability of their hyperparameters, ease of implementation, and nice mathematical properties, DP and PY priors are widely used in practice, and in the last two decades a huge amount of research has focused on their properties (see for example Ghosal and van der Vaart 2017; Müller et al. 2018). The use of the DP as a mixing measure was first introduced by Lo (1984). Thanks to the wide variety of efficient computational methods which have been introduced for their inference (Escobar and West 1998; MacEachern and Müller 1998; Neal 2000; Blei and Jordan 2006), nonparametric mixture models have become common in a wide range of modeling applications.

In the context of density estimation, under certain conditions the posterior distribution of DP mixture models concentrates at the true data-generating density at the minimax-optimal rate (Ghosal and van der Vaart 2017; Ghosal et al. 1999). This holds for other types of Bayesian nonparametric priors, such as PY priors (Lijoi et al. 2005). Nguyen (2013) further proved posterior consistency of the mixing distribution in the Wasserstein metric DP mixture models.

It is important to realize that consistency of the posterior distribution for the data-generating density and even for the mixing measure does not imply consistency of the inferred number of clusters. Empirically, many researchers have observed that DP mixture posteriors tend to overestimate the number of clusters (West and Escobar 1993; Lartillot and Philippe 2004; Onogi et al. 2011). More recently, Miller and Harrison (2013); Miller and Harrison (2014) proved non-consistency for the number of components in DP and PY mixtures. Alamichel et al. (2024) extended this result to the case of Gibbs-type processes and finite-dimensional representations of Bayesian nonparametric (BNP) priors. A possible explanation for this inconsistency result can be found in a result proved by Rousseau and Mengersen (2011), that in overfitted finite or infinite mixture models, the weight attributed to extra cluster goes to zero as the number of observations grows. Provided that the weights for the extra components are infinitesimally small, any mixture can be approximated arbitrarily well by a mixture with a larger number of components.

Despite the above inconsistency results, it can be possible to achieve posterior consistency for the number of clusters in the mixture models we consider. Guha et al. (2021) introduce a fast and simple post-processing procedure for DP mixtures which provides clustering consistency. Zeng et al. (2023) introduce a quasi-Bernoulli stick-breaking process and prove posterior consistency for the number of clusters in the associated mixture model. Consistency in this class of BNP priors requires the prior to be calibrated based on the sample size, hence the model is no longer projective. Ascolani et al. (2022) show that posterior consistency for the number of clusters can be achieved for a projective model by putting a prior on the DP concentration parameter α. DP mixtures modeled in this way can be considered as mixtures of DP mixtures (Antoniak 1974) and are commonly used in practice.

We show that Ascolani et al. (2022)'s result cannot be directly extended to PY mixtures: we prove clustering inconsistency for Pitman–Yor process mixture models with a prior on the concentration parameter when the discount parameter $\sigma \in (0,1)$ is a fixed constant.

3.2 Notations

We assume that data $X_{1:n} \in \mathcal{X}^n$ is generated by a mechanism of the following form:

$$X_i \overset{\text{iid}}{\sim} f^X(\cdot) = \sum_{j=1}^{t} w_j f(\cdot \mid \theta_j^\star), \tag{3.1}$$

where the w_j are probability weights in $(0,1)$ summing to one, and where the $f(\cdot \mid \theta_j^\star)$ are probability kernels, each depending on some parameter $\theta_j^\star$. The above may alternatively be expressed as a convolution of the component-specific kernel $f(\cdot \mid \theta)$ with the discrete mixing measure $G = \sum_{j=1}^t w_j \delta_{\theta_j^\star}$:

$$f^X(x) = \int f(x \mid \theta) G(\mathrm{d}\theta).$$

We consider the well-specified case where the kernel density $f(\cdot \mid \theta)$ is known, but where the integer t, the weights w_j, and the latent variables $\theta_j^\star$ in Equation (3.1) are all unknown. To allow for an unbounded number of components t in the mixture, we consider nonparametric mixture models with nonparametric priors on the mixing measure G.

Ascolani et al. (2022) consider Dirichlet process mixture models with a prior on the concentration parameter α:

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} f(\cdot \mid \theta_i), \quad \theta_i \mid G \overset{\text{iid}}{\sim} G, \quad G \mid \alpha \sim \mathrm{DP}(\alpha, H), \quad \alpha \sim \pi, \tag{3.2}$$

where π is a prior distribution on α, and H is the DP base measure with a density h.

We consider an extension of Ascolani et al. (2022)'s model, which are Pitman–Yor mixture models with a prior π on the concentration parameter $\alpha > 0$ and with a fixed discount parameter $\sigma \in (0, 1)$:

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} f(\cdot \mid \theta_i), \quad \theta_i \mid G \overset{\text{iid}}{\sim} G, \quad G \mid \alpha, \sigma \sim \mathrm{PY}(\alpha, \sigma, H), \quad \alpha \sim \pi. \tag{3.3}$$

For every pair of numbers $(n, s) \in \mathbb{N}^2$ with $s \leq n$, we let $\mathcal{A}_s(n)$ denote the set of partitions of $\{1, \dots, n\}$ into s non empty subsets. Conditional on parameters α and σ, a Pitman–Yor mixture model induces the following prior distribution on the space of partitions on n, for any $n \in \mathbb{N}$, and any $A = \{A_1, \dots, A_s\} \in \mathcal{A}_s(n), s \leq n$,

$$p(A \mid \alpha, \sigma) = \frac{\sigma^{s-1}(1 + \frac{\alpha}{\sigma})_{(s-1)}}{(1+\alpha)_{(n-1)}} \prod_{j=1}^{s} (1-\sigma)_{(n_j - 1)}, \tag{3.4}$$

where $\alpha_{(n)} = \alpha \cdots (\alpha + n - 1)$ is the ascending factorial and $n_j = |A_j|$ stands for the cardinality of the set A_j. Conditionally on the partition A, the probability distributions of the data $X_{1:n} = (X_1, \dots, X_n)$ and of the cluster-specific parameters $\hat{\theta}_{1:s} = (\hat{\theta}_1, \dots, \hat{\theta}_s)$ are

$$p(X_{1:n} \mid \hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} f(X_i \mid \hat{\theta}_j), \quad p(\hat{\theta}_{1:s} \mid A, \theta) = p(\hat{\theta}_{1:s} \mid A) = \prod_{j=1}^{s} h(\hat{\theta}_j).$$

We use the standard notation K_n to denote the number of clusters in a sample of size n. The concentration parameter α essentially controls the prior mean of K_n, while the discount parameter σ has more impact on the variance (Bystrova et al. 2021). More specifically, the prior number of clusters is known to grow asymptotically with n as a power-law, e.g. in expectation we have $\mathbb{E}[K_n] \sim \frac{\Gamma(\alpha+1)}{\sigma\Gamma(\alpha+\sigma)} n^{\sigma}$ when $n \to \infty$ (see Section 3.3 of Pitman 2006). Under our model (3.3), K_n has the following prior distribution

$$p(K_n = s \mid \sigma) = \int \sum_{A \in \mathcal{A}_s(n)} p(A \mid \alpha, \sigma) \pi(\mathrm{d}\alpha)$$

where $p(A \mid \alpha, \sigma)$ is as above.

To study the asymptotic behaviour of the number of clusters, we consider $p(K_n = s \mid X_{1:n}, \sigma)$. We start with the joint distribution $(X_{1:n}, K_n \mid \sigma)$ which, for every $x_{1:n} = (x_1, \dots, x_n) \in \mathcal{X}^n$, is given by:

$$p(X_{1:n} = x_{1:n}, K_n = s \mid \sigma) = \sum_{A \in \mathcal{A}_s(n)} p(A \mid \sigma) \prod_{j=1}^{s} m(x_{A_j})$$

where $p(A \mid \sigma) = \int p(A \mid \alpha, \sigma)\pi(\mathrm{d}\alpha)$ and $m(x_{A_j}) = \int \prod_{i \in A_j} f(x_i \mid \theta)h(\theta)\mathrm{d}\theta$ is the marginal likelihood for the subset of observations identified by A_j, given that they are clustered together.

3.3 Theoretical result

Condition 4 from Miller and Harrison (2014) is required. This assumption controls the likelihood through the single-cluster marginals. Introducing

$$\varphi_s(x_{1:n}, c) := \min_{A \in \mathcal{A}_s(n)} \frac{1}{n} \left| S_A(x_{1:n}, c) \right|,$$

where $S_A(x_{1:n}, c)$ is the set of indices $j \in \{1, \ldots, n\}$ such that the part A_ℓ containing j satisfies $m(x_{A_\ell}) \leq c m(x_{A_\ell \setminus j}) m(x_j)$, i.e. the set of observations for which the marginals of the new clusters obtained after taking out that observation and creating a new singleton cluster dominates the marginal of the original cluster up to a constant c.

Assumption 3.1 (Condition 4 of Miller and Harrison (2014))**.** *Given a sequence of random variables $X_1, X_2, \ldots \in \mathcal{X}$, and $s \geq 1$, assume*

$$\sup_{c \in [0,\infty)} \liminf_{n \to \infty} \varphi_s(X_{1:n}, c) > 0 \quad a.s.$$

This condition induces, for example, that as $n \to \infty$, there is always a non-vanishing proportion of the observations for which creating a singleton cluster increases its cluster marginal. This condition only involves the data distribution and is shown to hold for several discrete and continuous distributions, such as the exponential family (see Theorems 7 and 8 in Miller and Harrison (2014)).

Theorem 3.1. *Suppose that the prior π over the concentration parameter α is proper. If Condition 4 of Miller and Harrison (2014) recalled above holds, for every G as in* (3.1)*, we have for any $t \in \mathbb{N}$,*

$$p(K_n = t \mid X_{1:n}) \not\to 1 \ as \ n \to \infty.$$

The proof of Theorem 3.1 rests on analysing the ratio $\frac{p(K_n = s \mid X_{1:n})}{p(K_n = t \mid X_{1:n})}$ for a fixed $t \in \mathbb{N}$, as consistency cannot hold if it does not converge to 0 as $n \to \infty$. Following the strategy of Ascolani et al. (2022), this ratio can be split into the product of two quantities, one capturing the impact of the prior distribution on the concentration parameter α, and the other independent of the prior on α. In the Dirichlet process case with a prior on α, the first quantity goes to 0 and the second remains bounded. We show that in the Pitman–Yor case, the σ parameter enters the first quantity and prevents it from vanishing as $n \to \infty$, destroying consistency and highlighting a fundamental difference between the DP and PY processes.

3.4 Proof of Theorem 3.1

The proof of our result relies on the following simple lemma, used by and proved by Ascolani et al. (2022). It justifies working with ratios, which allows us to avoid calculations of marginal likelihoods of the observed data.

Lemma 3.1. *The convergence $p(K_n = t \mid X_{1:n}) \to 1$ as $n \to \infty$ holds if and only if one has*

$$\sum_{s \neq t} \frac{p(K_n = s \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \to 0 \quad \textit{as } n \to \infty.$$

Proof of Theorem 3.1. We fix $t \in \mathbb{N}$. By Lemma 3.1, it will be sufficient to prove that $\frac{p(K_n=s|X_{1:n})}{p(K_n=t|X_{1:n})} \not\to 0$ as $n \to \infty$, for some s. We will prove this using $s = t+1$.

In order to prove our result, we make use of similar notations as in Ascolani et al. (2022) for the Dirichlet process mixture model. Throughout this proof we will use the subscript PY to indicate that a quantity is related to the Pitman–Yor model.

Under our Pitman–Yor mixture model, by applying Equation (3.4), we have

$$\begin{aligned}\frac{p(K_n = s \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} &= \frac{\int \sigma^{s-1} \left(1 + \frac{\alpha}{\sigma}\right)_{(s-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}} \mathrm{d}\alpha}{\int \sigma^{t-1} \left(1 + \frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}} \mathrm{d}\alpha} \frac{\sum_{A \in \mathcal{A}_s(n)} \prod_{j=1}^{s} (1-\sigma)_{(a_j-1)} m(X_{A_j})}{\sum_{A \in \tau_t(n)} \prod_{j=1}^{t} (1-\sigma)_{(a_j-1)} m(X_{A_j})} \\ &=: C_{\mathrm{PY}}(n,t,s) R_{\mathrm{PY}}(n,t,s)\end{aligned}$$

where

$$C_{\mathrm{PY}}(n,t,s) = \frac{\int \sigma^{s-1} \left(1 + \frac{\alpha}{\sigma}\right)_{(s-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}} \mathrm{d}\alpha}{\int \sigma^{t-1} \left(1 + \frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}} \mathrm{d}\alpha}$$

and

$$R_{\mathrm{PY}}(n,t,s) = \frac{\sum_{A \in \mathcal{A}_s(n)} \prod_{j=1}^{s} (1-\sigma)_{(a_j-1)} m(X_{A_j})}{\sum_{A \in \tau_t(n)} \prod_{j=1}^{t} (1-\sigma)_{(a_j-1)} m(X_{A_j})}.$$

Since our expression $R_{\mathrm{PY}}(n,t,s)$ above does not depend on α, it is identical to the corresponding expression in the setup of Miller and Harrison (2014), who prove that it does not converge to zero as $n \to \infty$. Finer bounds can be found in Yang et al. (2020). What is left to show is that our expression $C_{\mathrm{PY}}(n,t,s)$ above does not converge to zero as $n \to \infty$.

Taking $s = t+1$, we then have for a fixed value of n,

$$\begin{aligned} C_{\mathrm{PY}}(n,t,t+1) &= \frac{\int \sigma^t \left(1+\frac{\alpha}{\sigma}\right)_{(t)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha}{\int \sigma^{t-1} \left(1+\frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha} \\ &= \sigma \frac{\int \left(t+\frac{\alpha}{\sigma}\right)\left(1+\frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha}{\int \left(1+\frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha} \\ &= t\sigma + \frac{\int \alpha \left(1+\frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha}{\int \left(1+\frac{\alpha}{\sigma}\right)_{(t-1)} \frac{\pi(\alpha)}{(1+\alpha)_{(n-1)}}\mathrm{d}\alpha} \\ &\geq t\sigma, \end{aligned}$$

as the second term of the sum is the integral of a product of positive terms.

Finally, as $C_{\mathrm{PY}}(n,t,t+1) \geq t\sigma > 0$, $\lim_{n\to\infty} C_{\mathrm{PY}}(n,t,t+1) > 0$, hence the result. □

3.5 Simulation study

We illustrate our results through a simulation study. Data is generated using a Gaussian location mixture with $K_0 = 3$ components: $P(x) = \sum_{i=1}^{3} p_i \mathcal{N}(x \mid \mu_i, \Sigma)$, where $p = (p_1, p_2, p_3) = (0.5, 0.3, 0.2)$ and $\mathcal{N}(x \mid \mu_i, \Sigma)$ is a multivariate Gaussian with mean μ_i and covariance matrix Σ with $\mu_1 = (0.8, 0.8)$, $\mu_2 = (0.8, -0.8)$, $\mu_3 = (-0.8, 0.8)$ and $\Sigma = 0.05\, I_2$. We adapt the Importance Conditional Sampler for PY mixtures of Canale et al. (2022), with the following prior specification:

$$\begin{aligned} G &\sim \mathrm{PY}(\alpha, \sigma, H), \quad \mu_i \sim \mathcal{N}(b_0, B_0), \quad i = 1, \ldots, t, \\ \Sigma^{-1} &\sim \mathcal{W}(c_0, C_0), \quad C_0 \sim \mathcal{W}(d_0, D_0). \end{aligned}$$

The Wishart prior on Σ^{-1} and the prior on μ_i are the same as in Malsiner-Walli et al. (2016). We run four Markov chain Monte Carlo (MCMC) chains of 22 000 iterations each, with 20 000 burn-in iterations.

We have proved inconsistency for the number of clusters when fitting Pitman–Yor mixture models with a prior on the concentration parameter α and fixed discount parameter σ. Figure 3.1 illustrates this result for varying values of σ and different parameters on the prior on α. In Figure 3.1, we can observe that as n increases in each scenario, the posterior distribution of the number of cluster K_n does not concentrate on the true value $K_0 = 3$.

While our result is limited to the setting where the discount parameter σ is kept fixed, it is common in practice to put a prior on both PY parameters α and σ in

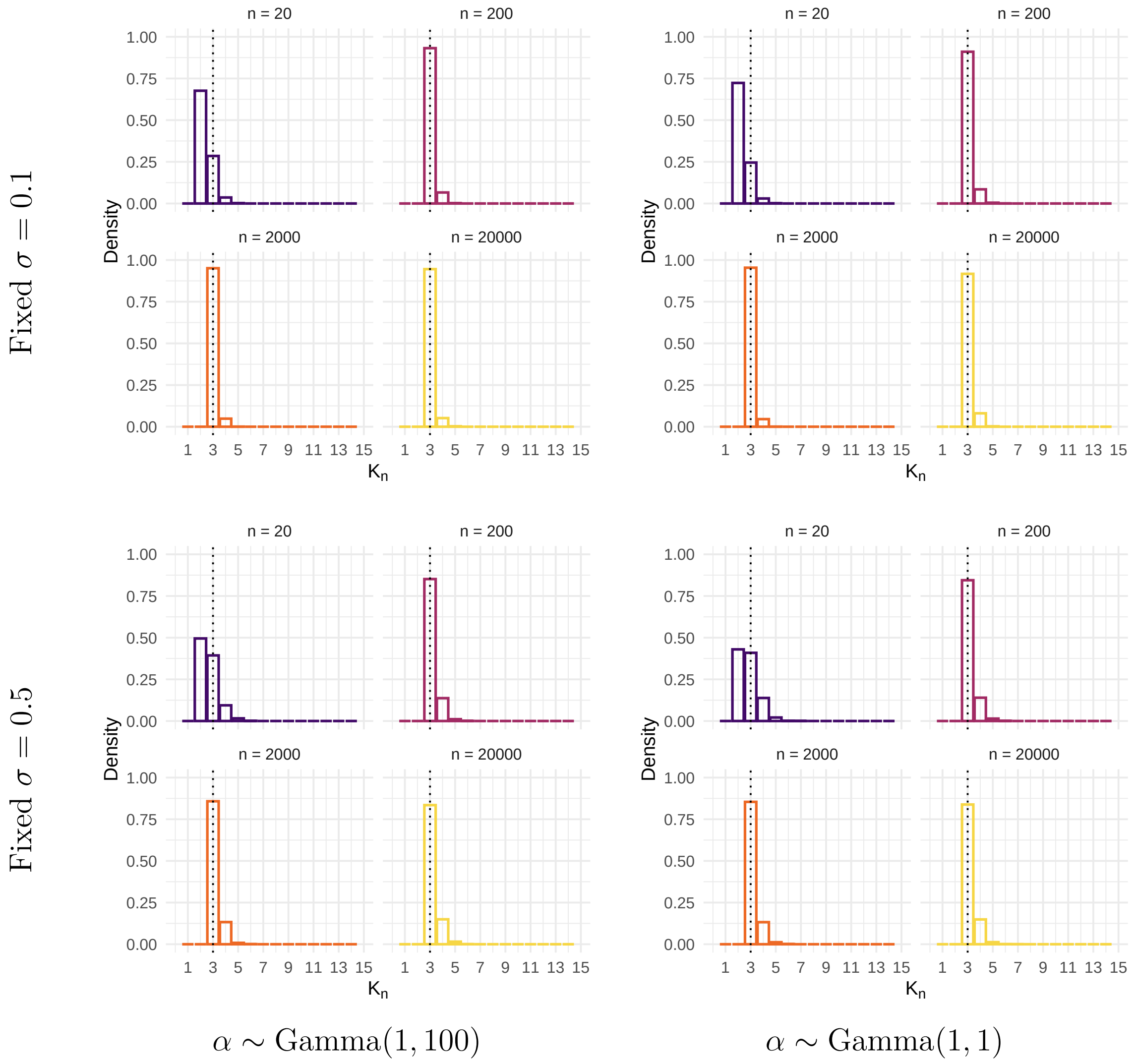


Figure 3.1: Posterior distribution of the number of clusters K_n under a Pitman–Yor process mixture for various choices of $n \in \{20, 200, 2000, 20000\}$ and with different values of the fixed parameter $\sigma \in \{0.1, 0.5\}$ and parameters of the Gamma prior on α. The dotted line represents the true number of components $K_0 = 3$.

PY mixture models. This situation is the subject of current investigations.

3.6 Discussion

We have proved inconsistency for the number of clusters when fitting single-component mixtures with Pitman–Yor mixture models with a prior on the concentration parameter α and fixed discount parameter σ. Our result holds when the true number of clusters in the data-generating mechanism is one. While hinting at what to expect, further study would be needed to fully understand clustering consistency for a data-generating mechanism with an arbitrary number of components.

While our result is limited to the setting where the discount parameter σ is kept fixed, it is common in practice to put a prior on both PY parameters α and σ in PY mixture models. Normalized-stable priors are the limiting case of Pitman–Yor priors as the parameter α goes to 0. Consequently, the study of the normalized-stable process mixture model with a prior on the parameter σ could constitute a first step in the study of the PY mixture model with a prior on σ.

Another interesting extension of this work would be to look at the Dirichlet multinomial process mixture model with a prior on the parameter α. Dirichlet multinomial process priors are the finite version of DP priors, useful in situations where there is some known upper bound K on the number of clusters (see Ishwaran and Zarepour 2000; Muliere and Secchi 1995). The Dirichlet process is recovered in the limit as the parameter K goes to infinity. We can then expect a similar behavior as the one proved by Ascolani et al. (2022) in the Dirichlet process case. While no theoretical results have yet been proven, simulation studies of Alamichel et al. (2024) (see Chapter 2 Section 2.7) suggest that clustering consistency is indeed achieved in the case of Dirichlet multinomial mixture models.

References

Alamichel, L., D. Bystrova, J. Arbel, and G. Kon Kam King (2024). "Bayesian mixture models (in)consistency for the number of clusters". In: *Scandinavian Journal of Statistics* (cit. on pp. 87, 94).

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems". In: *Source: The Annals of Statistics The Annals of Statistics* 2.6, pp. 1152–1174 (cit. on p. 88).

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). "Clustering consistency with Dirichlet process mixtures". In: *Biometrika. In press* (cit. on pp. 85, 88–91, 94).

Blei, D. M. and M. I. Jordan (2006). "Variational inference for Dirichlet process mixtures". In: *Bayesian analysis* 1.1, pp. 121–144 (cit. on p. 87).

Bystrova, D., J. Arbel, G. Kon Kam King, and F. Deslandes (2021). "Approximating the clusters' prior distribution in Bayesian nonparametric models". In: *Third Symposium on Advances in Approximate Bayesian Inference* (cit. on p. 89).

Canale, A., R. Corradin, and B. Nipoti (2022). "Importance conditional sampling for Pitman–Yor mixtures". In: *Statistics and Computing* 32.3, p. 40 (cit. on p. 92).

Escobar, M. D. and M. West (1998). "Computing nonparametric hierarchical models". In: *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 1–22 (cit. on p. 87).

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* 1.2, pp. 209–230 (cit. on p. 87).

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). "Posterior consistency of Dirichlet mixtures in density estimation". In: *The Annals of Statistics* 27.1, pp. 143–158 (cit. on p. 87).

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press (cit. on p. 87).

Guha, A., N. Ho, and X. Nguyen (2021). "On posterior contraction of parameters and interpretability in Bayesian mixture modeling". In: *Bernoulli* 27.4, pp. 2159–2188 (cit. on p. 88).

Ishwaran, H. and M. Zarepour (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models". In: *Biometrika* 87.2, pp. 371–390 (cit. on p. 94).

Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process". In: *Molecular biology and evolution* 21.6, pp. 1095–1109 (cit. on p. 87).

Lawless, C., L. Alamichel, J. Arbel, and G. Kon Kam King (2023). "Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter”. In: *Fifth Symposium on Advances in Approximate Bayesian Inference* (cit. on p. 85).

Lijoi, A., I. Prünster, and S. G. Walker (2005). “On consistency of nonparametric normal mixtures for Bayesian density estimation”. In: *Journal of the American Statistical Association* 100.472, pp. 1292–1296 (cit. on p. 87).

Lo, A. Y. (1984). “On a class of Bayesian nonparametric estimates: I. Density estimates”. In: *The Annals of Statistics*, pp. 351–357 (cit. on p. 87).

MacEachern, S. N. and P. Müller (1998). “Estimating mixture of Dirichlet process models”. In: *Journal of Computational and Graphical Statistics* 7.2, pp. 223–238 (cit. on p. 87).

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). “Model-based clustering based on sparse finite Gaussian mixtures”. In: *Statistics and Computing* 26.1-2, pp. 303–324 (cit. on p. 92).

Miller, J. W. and M. T. Harrison (2013). “A simple example of Dirichlet process mixture inconsistency for the number of components”. In: *Advances in Neural Information Processing Systems*, pp. 199–206 (cit. on p. 87).

Miller, J. W. and M. T. Harrison (2014). “Inconsistency of Pitman-Yor process mixtures for the number of components”. In: *The Journal of Machine Learning Research* 15.1, pp. 3333–3370 (cit. on pp. 87, 90, 91).

Miller, J. W. and M. T. Harrison (2018). “Mixture Models With a Prior on the Number of Components”. In: *Journal of the American Statistical Association* 113.521, pp. 340–356 (cit. on p. 86).

Muliere, P. and P. Secchi (1995). “A note on a proper Bayesian bootstrap”. In: (cit. on p. 94).

Müller, P., F. A. Quintana, and G. Page (2018). “Nonparametric Bayesian inference in applications”. In: *Statistical Methods & Applications* 27, pp. 175–206 (cit. on p. 87).

Neal, R. M. (2000). “Markov chain sampling methods for Dirichlet process mixture models”. In: *Journal of Computational and Graphical Statistics* 9.2, pp. 249–265 (cit. on p. 87).

Nguyen, X. (2013). “Convergence of latent mixing measures in finite and infinite mixture models”. In: *The Annals of Statistics* 41.1, pp. 370–400 (cit. on p. 87).

Onogi, A., M. Nurimoto, and M. Morita (2011). “Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods”. In: *BMC Bioinformatics* 12.1, pp. 1–16 (cit. on p. 87).

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d’Eté de Probabilités de Saint-Flour XXXXII-2002*. Springer (cit. on p. 89).

Pitman, J. and M. Yor (1997). “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator”. In: *The Annals of Probability*, pp. 855–900 (cit. on p. 87).

Rousseau, J. and K. Mengersen (2011). “Asymptotic behaviour of the posterior distribution in overfitted mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, pp. 689–710 (cit. on p. 87).

West, M. and M. D. Escobar (1993). *Hierarchical priors and mixture models, with application in regression and density estimation.* Institute of Statistics and Decision Sciences, Duke University (cit. on p. 87).

Yang, C.-Y., E. Xia, N. Ho, and M. I. Jordan (2020). *Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models.* arXiv: 1905.09959 (cit. on p. 91).

Zeng, C., J. W. Miller, and L. L. Duan (2023). “Consistent Model-based Clustering using the Quasi-Bernoulli Stick-breaking Process.” In: *J. Mach. Learn. Res.* 24, pp. 153–1 (cit. on p. 88).

Part II
Bayesian nonparametric mixture models application

Chapter 4

Species Sensitivity Distribution revisited: a Bayesian nonparametric approach

The first part focused on some theoretical aspects of clustering with Bayesian nonparametric mixture models. In this second part, we used a similar model to assess some ecological risk.
This Chapter is a joint work with Julyan Arbel, Guillaume Kon Kam King and Igor Prünster. Julyan Arbel, Guillaume Kon Kam King and Igor Prünster proposed the model, Guillaume Kon Kam King conducted the simulation study and real-data analysis on the model. I joined the project later and conducted the clustering analysis. I also created the Shiny applicaation provided with the paper. All authors contributed to the writing of the final version. This Chapter is based on the following recently submitted paper:

L. Alamichel, J. Arbel, G. Kon Kam King, and I. Prünster (2024+). *Species Sensitivity Distribution revisited: a Bayesian nonparametric approach.* Submitted

Contents

Abstract We present a novel approach to ecological risk assessment by reexamining the Species Sensitivity Distribution (SSD) method within a Bayesian nonparametric framework. Widely mandated by environmental regulatory bodies globally, SSD has faced criticism due to its historical reliance on parametric assumptions when modeling species variability. By adopting nonparametric mixture models, we address this limitation, establishing a more statistically robust foundation for SSD.

Our Bayesian nonparametric approach offers several advantages, including its efficacy in handling small datasets typical of ecological risk assessment and its ability to provide principled uncertainty quantification alongside simultaneous density estimation and clustering. We utilize a specific nonparametric prior from the class of normalized random measures with independent increments as the mixing measure, chosen for its robust clustering properties—a crucial consideration given the lack of strong prior beliefs about the number of components among SSD practitioners.

Notably, we extend our mixture model to accommodate censored data, which are common in ecotoxicology studies. Through systematic simulation studies and analysis of real datasets, we demonstrate the superiority of our Bayesian nonparametric SSD over classical normal SSD and kernel density estimate SSD methods.

Moreover, we exploit the inherent clustering structure of the mixture model to explore patterns in species sensitivity. Our findings underscore the effectiveness of our approach in improving ecological risk assessment methodologies.

Keywords— Bayesian Nonparametrics; Critical Effect Concentration; Ecological Risk Assessment; Ecotoxicology; Hazardous Concentration; Mixture Models.

4.1 Introduction

4.1.1 Background

Assessing the response of a community of species to environmental stress is critical for ecological risk assessment. Methods developed for this purpose vary greatly in levels of complexity and realism. Species Sensitivity Distribution (SSD) represents an intermediate tier method, more refined than rudimentary assessment factors (Posthuma et al. 2002) but practical enough to be used routinely by environmental managers and regulators in most developed countries (Australia and New Zealand, ANZECC (2000), Canada, CCME (2007), China, Liu et al. (2014), EU, ECHA (2008), South Africa, Roux et al. (1996), USA, USEPA (1998)). The SSD approach is intended to provide, for a given contaminant, a description of the tolerance of all species possibly exposed using information collected on a sample of species. This information consists of Critical Effect Concentrations (CECs), a concentration specific to a species that marks a limit over which the species suffers a critical level of effect. Such levels of effect are for instance the concentration at which 50% of the tested organisms died, referred to as Lethal Concentration 50% (LC_{50}), or the concentration which inhibited growth or reproduction by 50% compared to the control experiment, referred to as Effect Concentration 50% (EC_{50}). Each CEC is the summary of costly bioassay experiments for a single species, so data is usually in short supply. The European Chemical Agency (ECHA) sets the minimal data requirement to a sample size of 10, preferably 15 (ECHA 2008).

To describe the tolerance of all species to be protected, the distribution of the CECs is then estimated from the sample of tested species. In practice, a parametric distributional assumption is often adopted (Forbes and Calow 2002): the CECs are assumed to follow a log-normal (Wagner and Lokke 1991; Aldenberg et al. 2002), log-logistic (Aldenberg and Slob 1993; Kooijman 1987), triangular (Van Straalen 2002; Zhao and B. Chen 2016) or BurrIII (Shao 2000) distribution.

Once the response of the community is characterized by the distribution, the goal of the risk assessment is to define a safe concentration, which will protect all or most of the species. In the case of distributions without a lower bound above 0, no concentration would protect all species potentially exposed, so a cut-off value is often chosen as the safe concentration. This is typically the Hazardous Concentration for 5% of the Species (HC_5), which is the 5th percentile of the distribution. Instead of the estimate of 5th percentile, the lower extreme of a confidence interval around the 5th percentile is often used in practice, and, on top of that, a safety factor is to be subsequently applied to the resulting value.

The difficulty to justify the choice of any given parametric distribution for the

SSD has sparked various research directions. Several contributions (F.-L. Xu et al. 2015; He et al. 2014; Jagoe and Newman 1997; Van Straalen 2002; Xing et al. 2014; Zhao and B. Chen 2016) have sought to find the best parametric distribution using goodness-of-fit measures for model comparison. The consensus on this issue is that no single distribution seems to provide a uniformly superior fit and that the answer is essentially dataset dependent (Forbes and Calow 2002). In fact, model comparison and goodness of fit tests have relatively low power on small datasets, precluding the emergence of a clear-cut answer. Therefore, the log-normal distribution has emerged as the customary choice, notably because it readily provides confidence intervals on the HC_5.

Another research direction has aimed at avoiding any reference to a distribution using so-called distribution-free approaches. These efforts include using the empirical distribution function (Suter II et al. 1999; Jones et al. 1999), rank-based methods (Van Der Hoeven 2001; L. Chen 2004), bootstrap resampling (Jagoe and Newman 1997; Grist et al. 2002; B. Wang et al. 2008) or nonparametric kernel density estimation (Y. Wang et al. 2015). All these approaches have in common that they require large sample sizes to be applicable, which clashes with the low amount of data usually available for SSD as well as with the general goal of reducing animal testing. Finally, some contributions have considered the possibility that the distribution of the CECs might not be a single distribution but rather a mixture of distributions (Zajdlik et al. 2009), datasets being an assemblage of several log-normally distributed subgroups (Kefford et al. 2012; Craig 2013). This is more realistic from an ecological point of view: several factors influence the tolerance of a species to a contaminant such as the taxonomy or habitat, and contaminants such as pesticides might even have target species groups, which therefore react very differently from non-target species. Hence, scientific knowledge provides support to the assumption that an SSD might be composed of several subgroups, although the CECs within a group might still be well-approximated by a log-normal distribution.

4.1.2 Objectives and outline

Given the determinants of species sensitivity to contaminants are still an open research question, little knowledge is available a priori on the group structure: this represents a strong motivation for a nonparametric approach. However, any SSD method needs to be accurate for small datasets, which suggests trying to improve on the existing frequentist nonparametric methods, which are mostly based on asymptotic guarantees. Bayesian nonparametric (BNP) mixture models offer an effective solution for both large and small datasets because the complexity of the mixture model adapts to the size of the dataset. It offers a good compromise between simplistic parametric models and kernel density methods which might exhibit drawbacks such as lack of flexibility (Barrios et al. 2013), problematic uncertainty quantification, or overfitting. Indeed, while it is always possible to obtain confidence intervals for a frequentist method using bootstrap, it can be difficult to stabilize the interval for small datasets even with a large number of bootstrap samples. Importantly, the low amount of information available in small datasets to estimate the groups' parameters can be complemented via the prior, as some a priori degree of information is generally available from other species or contaminants (Awkerman et al. 2008; Craig 2013; Craig et al. 2012). Finally, note that the current official recommendation in the case of apparent multimodality of the toxicity dataset is to use only the most sensitive group (ECHA 2008; Zajdlik et al. 2009). As a result, a multimodal approach has the potential to greatly improve the current methodology by making better use of existing data. Indeed, Fox et al. (2021) recognizes that observing multimodal data is common and one proposed solution is to use a mixture model.

In Section 4.2 we present a Bayesian nonparametric approach to SSD based on a nonparametric mixture model. We show that our BNP-SSD approach can include censored data, which are common in ecotoxicology (Kon Kam King et al. 2014), and that it provides a rigorous description of the uncertainty on the variable of interest, the HC_5. We showcase the value of our approach by comparing the BNP-SSD with the most standard normal-SSD approach (Aldenberg and Jaworska 2000) and with a nonparametric SSD method based on Kernel Density Estimate (KDE) (Y. Wang et al. 2015). The comparison is performed on simulated data in Section 4.3, to demonstrate the higher accuracy of the BNP-SSD, and we study real censored and noncensored datasets in Section 4.4 highlighting the robustness of our method. Additionally, we perform an exploratory analysis to describe what biological insight we can gain with BNP-SSD by studying patterns of species or contaminants induced by the clustering in Section 4.5. Finally, we conclude and describe further research directions in Section 4.6.

The code used in this paper is available on GitHub at `https://github.com/alamichL/BNP_SSD/`.

4.2 Methods

Due to the wide spectrum of variation of SSD concentrations and to their positivity, it is common practice to work on log transformed concentrations. We consider a sample of n log-concentrations that we denote by $X_{1:n} = (X_1, \dots, X_n)$, that typically represents the CEC for a collection of n species tested with a given contaminant. Moreover, the data are standardized: observations are centered and rescaled to be of variance one. After the inference, all estimations are transformed back to the original scales.

We carry out density estimation for the SSD based on a sample $X_{1:n}$ using Bayesian nonparametric mixtures. The method of mixtures of probability density kernels with a nonparametric prior as mixing measure is due to Lo (1984), where Dirichlet process mixtures (DPM) is introduced. Generalizations of the DPM correspond to allowing the mixing distribution to be any discrete nonparametric prior. A large class of such prior distributions is obtained by normalizing random measures known as *completely random measures* (Kingman 1967). The normalization step gives rise to so-called normalized random measures with independent increments (NRMI) as defined by Regazzini et al. (2003). See Lijoi and Prünster (2010); Jordan (2010); Barrios et al. (2013) for reviews. More details on specific choices of the NRMI prior and their inferential impact are provided in the sequel. An NRMI mixture model is defined as

$$\begin{aligned} X_i \mid G &\overset{\text{i.i.d.}}{\sim} \tilde{f}(x) = \int f(x \mid \theta) G(\mathrm{d}\theta), \\ G &\sim \text{NRMI} \end{aligned} \tag{4.1}$$

where k is a probability density kernel parametrized by some $\theta \in \Theta$ and G is a random probability on Θ whose distribution is an NRMI. Alternatively, the mixture model can also be formulated hierarchically as

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} f(x \mid \theta_i), \quad i = 1, \dots, n, \\ \theta_i \mid G &\overset{\text{i.i.d.}}{\sim} G, \quad i = 1, \dots, n, \\ G &\sim \text{NRMI}. \end{aligned} \tag{4.2}$$

Specifically, we consider location-scale mixtures and denote by $\theta_i = (\mu_i, \sigma_i)$ the vectors of individual means and standard deviations, $\theta_i \in \mathbb{R} \times \mathbb{R}_+$. As discussed in the Introduction, log-concentrations are commonly fitted with a normal distribution, or with mixtures of such distributions. Our aim is to move from these parametric models to the nonparametric specification in (4.2), and in order to allow for comparisons with competing approaches, we stick to the normal specification for k on the log-concentrations $X_{1:n}$, $f(x \mid \mu, \sigma) = \mathcal{N}(x \mid \mu, \sigma^2)$.

Within this framework, density estimation is carried out by evaluating the posterior mean

$$\hat{f}(x \mid X_{1:n}) = \mathbb{E}\Big(\tilde{f}(x) \mid X_1, \dots, X_n\Big) \tag{4.3}$$

for any x in $\mathbb{R}$.

We compare the proposed BNP-SSD to two competitors. First, the most commonly used model for the SSD, the normal distribution (Aldenberg and Jaworska

2000), with estimated density $\hat{f}_{\mathcal{N}}(x) = \mathcal{N}(x \mid \hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu}$ and $\hat{\sigma}$ are the data empirical mean and standard deviation. Second, the frequentist nonparametric kernel density method recently applied to the SSD by Y. Wang et al. (2015), with estimated density $\hat{f}_{\mathrm{KDE}}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(x \mid X_i, h_n^2)$ where $h_n = 1.06\hat{\sigma}n^{-\frac{1}{5}}$ is the asymptotically optimal default bandwidth recommendation of Silverman (1986), also used by Y. Wang et al. (2015).

4.2.1 Censored data

Kon Kam King et al. (2014) explained how to use censored data with the normal SSD, and showcased the drawbacks of the common approach, which consists of transforming or discarding censored data in SSD. It is similarly possible to incorporate censored data into the BNP-SSD.

Indeed, only the first line of the hierarchical model defined in (4.2) needs to be suitably adapted, while the other levels in the hierarchy remain unchanged. More specifically, denote by F the cumulative density function of the kernel k, then: $f(x \mid \theta)$ has to be replaced by $F(x \mid \theta)$ for a left-censored observation, by $1 - F(x \mid \theta)$ for a right-censored observation, and by $F(x_r \mid \theta) - F(x_l \mid \theta)$ for an interval-censored observation $[x_l, x_r]$. This approach can be used for the standard normal SSD and any likelihood-based inference, but there is no widely available tool to perform KDE on all types of censored data: the R package `ICE` does not handle left/right censored data (not maintained anymore on CRAN); R packages `muhaz` or `Kernelheaping` can deal with right or interval-censored data respectively, but there does not seem to be an available implementation for all three types of censored data.

In this paper, we illustrate the differences among the various approaches on a censored dataset and, for comparison purposes, we study censored and non-censored versions of the dataset. To obtain a non-censored dataset from a censored dataset, we follow the customary approach, which consists of discarding left and right censored data and replacing interval-censored data with the central value of the interval.

4.2.2 Prior specification

The class of NRMI priors is very broad and we refer the reader to Lijoi and Prünster (2010) for an extensive review. Here we focus on a specific member of the class known as the normalized stable process (Kingman 1975), which as argued in Barrios et al. (2013) represents a natural default choice. It is specified in terms of a stability parameter $\gamma \in (0, 1)$ and a base measure H, which corresponds to the expectation of the random probability measure. The stability parameter γ controls the variability of the prior distribution on the number of clusters: heuristically, a small γ corresponds to an informative prior, while a large γ indicates a vague prior. Large values of γ can require expensive computations to preserve the quality of the posterior sampling. We chose $\gamma = 0.4$ as a compromise between the flexibility of the model and the computational requirements. See Lijoi et al. (2007); Barrios et al. (2013) for details.

As mentioned previously, we shall consider location-scale mixtures, meaning that the NRMI prior should be defined on $\mathbb{R}\times\mathbb{R}_+$, the space of locations and scales. Thus the base measure H is defined on $\mathbb{R}\times\mathbb{R}_+$, and we denote by f_0 its density with respect to the Lebesgue measure on $\mathbb{R}\times\mathbb{R}_+$. We assume that the locations μ and scales σ are a priori independent. Thus, we use the notation $f_0(\mu, \sigma) = f_0^1(\mu) f_0^2(\sigma)$ with possible hyperparameters for f_0^1 and f_0^2. The possibility to specify independent priors for μ and σ is a beneficial feature of NRMIs which do not require any conjugacy structure in the prior. Therefore, the prior specification can be derived in a straightforward way.

The specific choice of distribution on the location parameters of the clusters μ is a normal distribution $f_0^1(\mu) = \mathcal{N}(\mu \mid \varphi_1, \varphi_2^{-1})$ where mean φ_1 and precision φ_2 are assigned a normal-Gamma conjugate hyperprior. This corresponds to a vague prior for the location of the clusters, which can just as well be at the center of the dataset or at the borders. Regarding the scale parameter σ of the clusters, given the standardization of the data during the pre-processing where the variance is set to one, σ should be smaller than one, approaching one only in cases of unimodality. Moreover, σ should also be a priori bounded from below since numerous extremely small clusters do not make biological sense regarding species sensitivity distributions. Therefore, we choose a uniform prior $f_0^2(\sigma) = \text{Unif}_{[0.1,1.5]}(\sigma)$, leaving room around the upper bound of one to allow for potential unimodality. We studied the sensitivity with respect to this prior specification by varying its extreme points and also with respect to a left-truncated normal prior of mean 0.5, variance 1 and lower bound 0.1. Note that the latter has approximately 3/4 of its mass within the support of the $\text{Unif}_{[0.1,1.5]}$ distribution. The sensitivity analyses showed little variation to moderate changes in the parameters of these two prior distributions.

4.2.3 Posterior sampling

Several software packages devoted to the implementation of Bayesian nonparametric models have appeared in recent years: `DPpackage` (Jara et al. 2011) is an R package containing a rather comprehensive bundle of functions for Bayesian nonparametric models, but is not anymore maintained on CRAN; `Bayesian Regression` (Karabatsos 2016) is a software for Bayesian nonparametric regression; `BNPmix` (Corradin et al. 2021) is an R package for Bayesian nonparametric multivariate density estimation, clustering, and regression, using Pitman-Yor mixture models; `BayesMix` (Beraha et al. 2022) is a C++ library for MCMC posterior simulation for general

Bayesian mixture models. Here, we use BNPdensity[1], an R package which performs BNP density estimation and clustering under a general specification of NRMI prior based on generalized Gamma processes see Barrios et al. (2013); Arbel et al. (2021). BNPdensity leverages the popular Ferguson and Klass algorithm (Ferguson and Klass 1972), extended with a Metropolis–Hastings within Gibbs scheme.

4.2.4 Estimation of the HC_5

The main quantity of interest for ecological risk assessment is the HC_5, which corresponds to the 5th percentile of the SSD distribution. In our BNP framework, we rely on the posterior expectation, which corresponds to the Bayes estimator under a quadratic loss function, as our density estimator. Moreover, the 95% credible bands are formed by the 2.5% and 97.5% quantiles of the HC_5 posterior distribution.

The 5th percentile of the KDE can be obtained by numerical inversion of the cumulative distribution function, and the confidence intervals using nonparametric bootstrap. The 5th percentile of the normal SSD and its confidence intervals were obtained following the classical method by Aldenberg and Jaworska (2000).

4.2.5 Robustness comparison using Leave-One-Out cross validation

We compare the predictive performance of the three SSD models using Leave-One-Out (LOO) cross-validation. We compute the LOO for each method as:

$$\mathrm{LOO}_i = \hat{f}(X_i \mid X_{-i}), \tag{4.4}$$

where $\hat{f}(\cdot \mid X_{-i})$ is the density estimate based on $X_{1:n}$ with X_i left out for each of the three methods. For the BNP models, LOOs are referred to as conditional predictive ordinates (CPOs) statistics. They are commonly used in applications, see for instance Gelfand (1996).

A CPO statistic is defined for each data point X_i as

$$\mathrm{CPO}_i = \hat{f}(X_i \mid X_{-i}) = \int f(X_i \mid \theta) p(\mathrm{d}\theta \mid X_{-i}),$$

where $p(\mathrm{d}\theta \mid X_{-i})$ is the posterior distribution associated to X_{-i} and $\hat{f}(\cdot \mid X_{-i})$ is the (cross-validated) posterior predictive distribution of (4.3). CPOs can be easily

[1] Available at https://CRAN.R-project.org/package=BNPdensity.

approximated by Monte Carlo as

$$\widehat{\mathrm{CPO}_i} = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(X_i \mid \theta^{(t)})}\right)^{-1},$$

where $\{\theta^{(t)}, t = 1, 2, \ldots, T\}$ is an MCMC sample obtained as detailed in Section 4.2.3. For the two frequentist models, the LOOs can be computed by fitting them directly on the leave-one-out data X_{-i} for each i.

4.2.6 Clustering estimation

The question of how to estimate data clustering based on an MCMC posterior sample is a long-standing problem in Bayesian statistics (see Dahl 2006; Lau and Green 2007). Estimating a clustering structure is computationally expensive, owing to the extremely rapid growth in the cardinality of the partition space with the sample size n, known as the Bell number of order n. Enumeration of all partitions is infeasible in practice, thus one typically resorts to approximations. Many ad-hoc procedures have been devised in the literature. However, as noted by Dahl (2006), it seems counter-intuitive to apply an ad-hoc clustering method on top of a model that itself produces clusterings.

We adopt instead a fully Bayesian route by undertaking clustering on decision-theoretic grounds. We consider a loss function $\mathcal{L}$ and propose a Bayesian point estimator $\hat{z}_{1:n}$ for a clustering structure obtained as an argument that minimizes the posterior expected loss given the data $X_{1:n}$

$$\hat{z}_{1:n} \in \arg\min_{z'_{1:n}} \sum_{z_{1:n}} \mathcal{L}(z'_{1:n}, z_{1:n}) p(z_{1:n} \mid X_{1:n}), \tag{4.5}$$

where $p(z_{1:n} \mid X_{1:n})$ is the posterior distribution of clustering $z_{1:n}$.

The maximum a posteriori (MAP), often considered in the literature, is an example of such a Bayesian estimator, based on the very crude $0-1$ loss function $\mathcal{L}_{0-1}$. However when n is large, a posterior sample generally hardly visits twice the same clustering, thus making the empirical MAP of the MCMC output very sensitive to the initialization of the chain and of very limited validity in practice.

Manifestly, many other loss functions can be considered and expected to perform better than $\mathcal{L}_{0-1}$. One particular choice of a loss function stands out from these in best estimating the number of groups in a clustering. It is known as the variation of information, denoted by $\mathcal{VI}$, which is a loss function firmly established in information theory (Meilă 2007; Wade and Ghahramani 2018). The variation of information between two clusterings is defined as the sum of their information (their Shannon entropies) minus twice the information they share. Simulations indicate that the variation of information is a sensible choice: when other losses such as the Binder loss (Binder 1978) typically tend to overestimate the number of clusters, the variation of information instead seems to effectively recover it (see for instance the simulated examples, and more specifically Figures 6–8, of Wade and Ghahramani 2018).

A merit of the approach presented in Wade and Ghahramani (2018) is that it rests on a greedy search algorithm to determine the minimum loss clustering of (4.5). Starting from the MCMC output, this greedy approach explores the space of partitions not restricted to those visited by the MCMC chain to find the optimum. The algorithm is available as an R package called **mcclust.ext** that we used in this study. See also **dahl2022search** for another recent algorithm.

4.3 Simulation study

In order to compare the performance of BNP-SSD, the normal-SSD (Aldenberg and Jaworska 2000) and the nonparametric KDE-SSD (Y. Wang et al. 2015), we perform a simulation study with synthetic datasets corresponding to different scenarios.

4.3.1 Simulated data

We consider three distinct simulated data scenarios corresponding to various situations:

(a) a standard normal distribution: this is a situation where the normal assumption made for SSD is justified;

(b) a t-distribution with three degrees of freedom and noncentrality parameter equal to -2: this is a situation where some species are relatively more sensitive, creating a heavier tail on the left of the distribution;

(c) a bimodal distribution corresponding to a mixture of normals $^1/_3\mathcal{N}(-2,1) + {}^2/_3\mathcal{N}(5,1)$: this is a situation where a group of species is much more sensitive than all the others, typical of some pesticides which disproportionately affect the target species.

These three scenarios represent the diversity of empirical distributions found in real data such as the National Institute for Public Health and the Environment (RIVM) database (Zwart 2001).

For all settings, we sampled independently $S = 40$ datasets of sizes 10, 20, 50, 100. These sizes are representative of the dataset sizes in the field, as described in the Introduction. Figure 4.1 depicts the data generating densities and the different estimates obtained from the three different approaches with datasets of size 20.

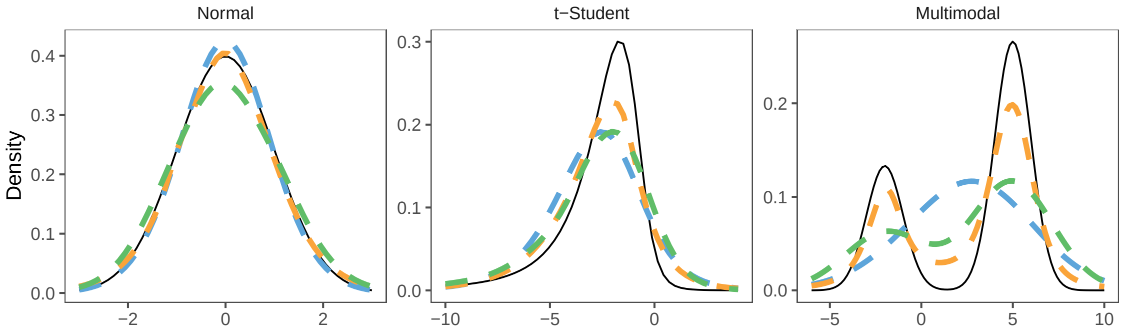


Figure 4.1: Three simulation scenarios: data generating density (solid line) and density estimates for each model based on datasets of size $n = 20$ (dashed lines). Orange (- -) for the BNP model, blue (- -) for the normal model, and green (- -) for the KDE model.

4.3.2 Performance comparison of the three approaches

For each sampled dataset $i \in \{1, \dots, S\}$ and each model considered, we estimate the data generating density f by a density denoted $\hat{f}_i$, and compute the associated 5th percentile $\hat{q}_i$, which is used as an estimate of the true HC_5 denoted by q_0. We denote by $(\hat{l}_i, \hat{u}_i)$ a 95% confidence/credible interval for $\hat{q}_i$, and by $\hat{\ell}_i = \hat{u}_i - \hat{l}_i$ its length. To account for sampling variation, we compute averaged summaries (using the notation $\left\langle \cdot \right\rangle_S$ to denote averaging over the S independent samples).

We compute two performance indicators, the mean absolute error $\mathrm{MAE} = \left\langle |\hat{q}_i - q_0| \right\rangle_S$, and the mean integrated squared error $\mathrm{MISE} = \left\langle \int (\hat{f}_i - f)^2 \right\rangle_S$. Moreover, we compute the mean confidence/credible interval length $\mathrm{MCIL} = \left\langle \hat{\ell}_i \right\rangle_S$ as a measure of uncertainty: as the BNP model captures model uncertainty, we expect it to give a more conservative estimate of uncertainty than the other models. However, we would not want to be conservative to the point that the estimates are useless for the practical purpose of estimating an HC_5.

The density estimates give a first intuition of the superiority of the BNP-SSD over the other two models in recovering the true density (Figure 4.1). The results from the simulation study are presented in Figure 4.2, which we describe from top to bottom and left to right.

On the normal simulated data, the well-specified normal model obviously performs best. However, the mean absolute error on the HC_5 of the BNP is very similar to that of the normal. For small sample sizes, the MISE of the BNP is almost the same as that of the normal. This illustrates the fact that the BNP model complexity scales with the amount of data and that in data-poor contexts, it essentially reduces to a normal model. For small dataset sizes, the mean CI length is larger for the BNP model than for the normal, reflecting the model uncertainty built into the BNP

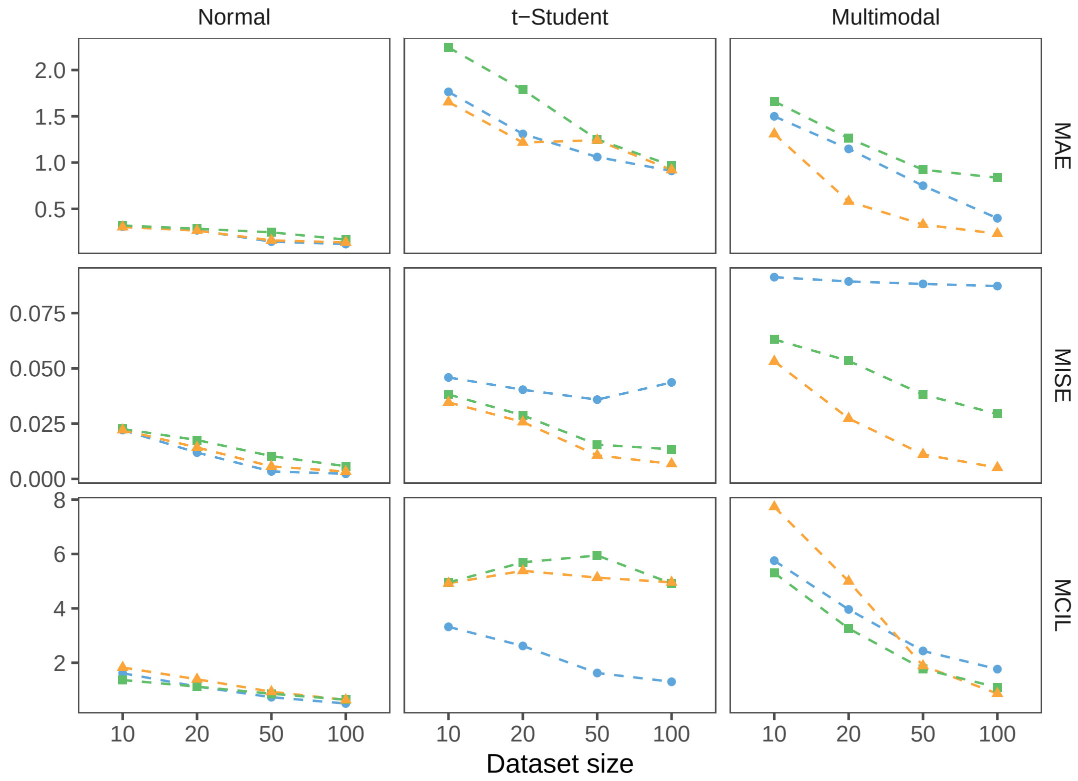


Figure 4.2: Normal, t-student, and normal mixture simulation scenarios (from left to right); mean absolute error (MAE), mean integrated squared error (MISE), and mean confidence/credible interval length (MCIL) as a function of the dataset size (from top to bottom). Orange (-▲-) for the BNP model, blue (-●-) for the normal model, and green (-■-) for the KDE model.

model, which is a sign of its potential to flexibly adapt in case of deviations as more data become available. For larger sizes, the model uncertainty decreases and the BNP and normal model coincide.

For the t-student simulated data, the BNP model outperforms the other two. The normal and BNP have a smaller mean absolute error than the KDE, the BNP and KDE have a smaller MISE and the mean CI length for the normal model is misleadingly small.

Finally, for the multimodal simulated data, the BNP model is clearly superior to the other two models in terms of mean absolute error and MISE. The mean CI length is relatively larger than the other two models for small dataset sizes and smaller for larger dataset sizes.

4.4 Analysis of contaminant-wise clustering

4.4.1 Real data description

We illustrate the advantages of the proposed Bayesian nonparametric approach by means of a selection of contaminants extracted from a large database collected by the RIVM and first presented in Zwart (2001).

We study the dataset already curated by Hickey et al. (2012) with the same restrictions concerning data quality and homogeneity. We consider both the censored and the non-censored versions of the dataset, the non-censored version being obtained by following the traditional approach of discarding left and right-censored data and taking the central value of interval-censored data. The dataset is an aquatic ecotoxicity research database, with 1,557 species and 3,448 distinct chemicals. Specifically, the dataset records the following covariates: species, chemical, and concentration.

To deal with the presence of multiple CECs values for one species, we used the classical approach to replace these values by the geometric mean of the values as a surrogate (ECHA 2008) for non-censored data, and followed Kon Kam King et al. (2014) in the case of censored data.

4.4.2 Density, quantiles and HC_5 estimation

For illustration purposes, we present three categories of contaminants: contaminants with large datasets, consisting of more than 60 values, medium datasets, with around 25 values, and small datasets, with a little over 10 values. For each of these categories, we select a roughly unimodal, a skewed and a bimodal dataset, as in the simulation study. This selection was performed for non-censored datasets. The three models (BNP, KDE and normal) were fitted on each dataset and we studied the estimate of the HC_5 and its credible interval, the LOO error and the shape of the estimated density compared to the histogram. The censored version of the same datasets was also studied with the BNP and normal model, while there does not seem to exist any implementation of the KDE model for censored data (see Section 4.2.1). The results are displayed in Appendix 4.A, see Figure 4.7 to Figure 4.12.

The BNP model is both more flexible than the KDE model, as apparent from the density estimates for the bimodal datasets, and comparably, or even more, robust. The length of the confidence/credible intervals does not exhibit substantial differences among the three methods. This represents strong evidence in favor of the claim that being less restrictive (in terms of distributional assumptions) than the normal model does not result in over-conservative estimates of the HC_5. This is of great importance since over-conservative estimates would seriously compromise a wide adoption of the BNP SSD approach. More precisely, in the case of the roughly normal datasets, the BNP method results in an estimate for the HC_5 comparable to that of the normal model. When the datasets strongly deviate from the normal model, the HC_5 estimates from the normal and BNP model differ substantially and strongly support the use of the more flexible BNP model over the normal.

4.4.3 Contaminant-wise clustering

We have so far demonstrated the advantages of the BNP method over the existing approaches to SSD for density estimation and for the determination of the HC_5. There is an additional benefit connected to the BNP-SSD: the mixture model induces a clustering of the species, which conveys interesting information from the biological point of view. Indeed, one of the long-standing questions around SSD, and ecotoxicology in general, is to understand what drives the sensitivity of species to a contaminant. Craig (2013) assumes that taxonomy is a driving factor and effectively imposes a clustering based on taxonomic units. Zwart (2001) investigates the influence of habitat by comparing freshwater and saltwater species, while Kefford et al. (2012) study the variations in sensitivity in different regions of the world.

All these approaches start from a possible clustering structure and test for a significant difference among cluster units. The BNP-SSD takes the opposite path by endogenizing the clustering in a probabilistically principled way. Indeed, it allows the clustering structure to emerge from the data, and, by using meta-data about the species, this structure can be examined a posteriori to verify whether it matches certain scientific hypotheses about the driving forces behind species sensitivity.

The clustering induced by the BNP model may or may not coincide with particular information about the species, which can challenge or support existing theories about the determinants of species sensitivity. Figure 4.3 compares the estimated clustering structure for Carbaryl and a quasi-taxonomic grouping expected to be relevant for species sensitivity in Zwart (2001). Two clusters emerge, one predominantly composed of crustaceans and containing all crustaceans but one, and another predominantly composed of fishes, containing all fishes but one, and all the molluscs. The three insects are scattered over the two clusters, the only annelid is grouped with the crustaceans while the only amphibian is grouped with the fishes. Thus, for Carbaryl, the estimated clustering structure seems strongly associated with the quasi-taxonomic grouping and supports the theory that species sensitivity is dependent on taxonomy, with fish forming a cluster relatively resilient to Carbaryl while crustaceans form a more sensitive cluster. However, this parallel between taxonomy and sensitivity is not observed for every contaminant; indeed, it is possible to

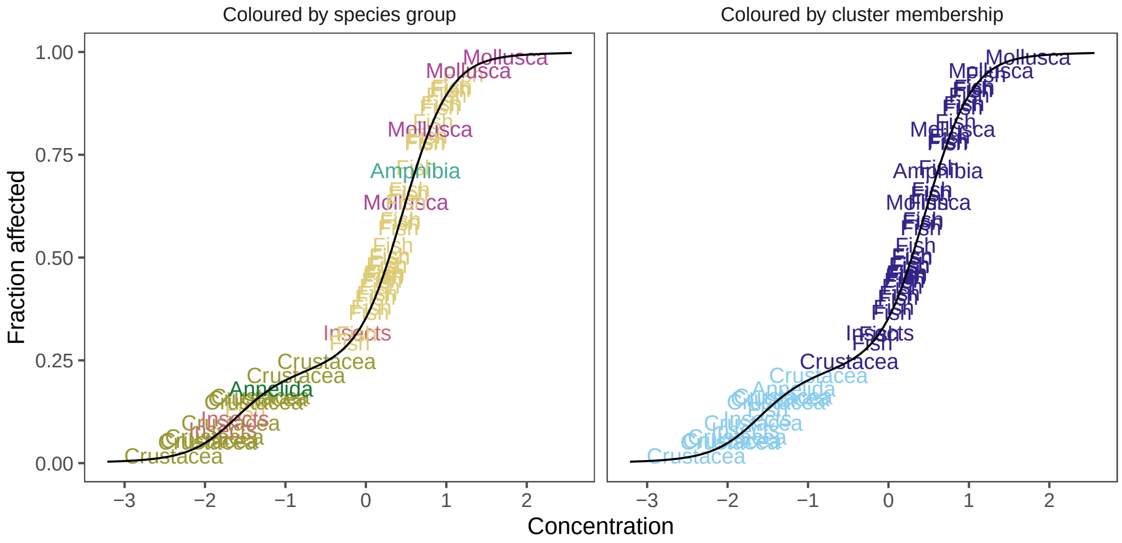


Figure 4.3: SSD for Carbaryl (CAS: 63-25-2) with the quasi-taxonomic group of each species overlaid on the curve. Left: Species coloured by quasi-taxonomic group. Right: Species coloured by cluster membership in the BNP model.

identify contaminants for which the estimated clustering does not match the quasi-taxonomic grouping. A general tendency observed over many contaminants is that fishes tend to group in a single cluster. This insight could be used to argue for reducing the number of fish species to be tested, as their contribution to the complete SSD could be emulated by giving a higher weight to a few representative species.

4.4.4 BNP-SSD Shiny application

The method described above can be used directly with the `BNPdensity` package, but this requires a certain level of fluency in the R language. Thus, we developed a Shiny application, named BNP-SSD, tailored to SSD problems and based on the functions of the package `BNPdensity`, available at https://alamichl.shinyapps.io/BNP_SSD/. This application is inspired by the application `shinyssdtools` of Dalgarno (2021).

In this application, the BNP model described in Section 4.2 is fitted to censored or uncensored data. Before fitting the model the concentration data are cleaned, dealing with the possible presence of multiple CECs values for one species, transformed using a log-scale, and centered-scaled. For greater flexibility, some options are left to the user, such as the number of iterations of the MCMC algorithm. Once the model has been fitted, the estimated density is plotted, along with some goodness-of-fit graphs. In another panel, an estimate of HC_5 is made using the posterior distribution over quantiles. It is also possible to estimate a percentile of the distribution other than the 5th percentile. The credible bands of this estimate are also computed. Finally, in the last panel, the induced optimal clustering is computed and plotted.

4.5 Cross-contaminant clustering

It would be highly interesting to establish whether similar patterns occur commonly for contaminants by studying the clustering structure for all contaminants in the dataset. A complete clustering analysis would require a hierarchical model with a contaminant effect, which is beyond the scope of the present paper and will be the object of future work. Here we approach the issue in a simple, yet insightful way by fitting the model independently for all contaminants. Consequently, we present a post-processing of the clustering structure estimated for each contaminant. The general idea is, over all contaminants, to assess how often each pair of species is grouped. This defines sensitivity communities of species, which we compare to the quasi-taxonomic grouping.

We restrict ourselves to contaminants tested by at least eight species, which is a little below the minimum threshold recommended for fitting a Species Sensitivity Distribution (ECHA 2008). We first fit the model on all such contaminants. We then combine the information from the clustering for each contaminant, to understand if some common patterns may be observed. To extract information from the clustering structure for each contaminant, we transform each estimated clustering into an association matrix. Stacking all the association matrices on top of each other forms a three-dimensional array, also called a tensor, each slice corresponding to a contaminant. One difficulty is that contaminants are tested on different sets of species, with potentially little overlap. This results in a large proportion of missing values (pairs of contaminants-species that have not been tested) that need to be dealt with.

4.5.1 Non-negative tensor factorization

We perform non-negative three-way tensor factorization (Cichocki et al. 2009), which is a tensor generalization of principal component analysis. It is a dimension-reduction technique that decomposes the association tensor into a sum of R rank-one tensors. Developed in Chemometrics, this technique has also been employed in Biostatistics, Signal Processing, Linguistics, and Machine Learning (Gauvin et al. 2014). The technique also allows the imputation of missing values.

We use a Parallel Factors Analysis (PARAFAC), also referred to as Canonical Decomposition (CANDECOMP) factorization; Section 4.B in Supplement provides details on this technique and background on tensors properties. Denoting by $\mathbf{Y}$ the tensor of the data described previously, we have that $\mathbf{Y} \in \mathbb{R}^{n_S \times n_S \times n_C}$ is a symmetric tensor in the first two dimensions, where n_S and n_C, respectively, denote the number of species and contaminants. The general PARAFAC factorization for some tensor $\hat{\mathbf{Y}} \in \mathbb{R}^{I \times J \times K}$ is denoted by

$$\hat{\mathbf{Y}} = [\![A, B, C]\!] = \sum_{r=1}^{R} a_r \circ b_r \circ c_r,$$

where $A = [a_1, \ldots, a_R] \in \mathbb{R}^{I \times R}$, $B = [b_1, \ldots, b_R] \in \mathbb{R}^{J \times R}$ and $C = [c_1, \ldots, c_R] \in \mathbb{R}^{K \times R}$ are three components or factors matrices, and $\circ$ stands for the vector outer product. For the considered data, the symmetry of the tensor $\mathbf{Y}$ in the first two dimensions implies that the PARAFAC factorization can be simplified as $\mathbf{Y} = [\![A, A, C]\!]$, where $A = [a_1, \ldots, a_R] \in \mathbb{R}^{n_S \times R}$ and $C = [c_1, \ldots, c_R] \in \mathbb{R}^{n_C \times R}$. Note that the factorization is only an approximation and incurs in some additive error $\mathbf{E}$ in the form of $\mathbf{Y} = [\![A, A, C]\!] + \mathbf{E}$. To give physical meaning to the different components found, we use the non-negative PARAFAC factorization (Y. Xu and Yin 2013) from the `multiway` R package (Leeuw 2011). This adds non-negativity constraints on the component matrices $a_{ir} \in \mathbb{R}_+$ and $c_{jr} \in \mathbb{R}_+$ for all $i \in \{1, \ldots, n_S\}$, $j \in \{1, \ldots, n_C\}$ and $r \in \{1, \ldots, R\}$.

A popular heuristic to determine the number of components R is the core-consistency diagnostic (Bro and Kiers 2003). This diagnostic requires the imputation of the missing values for efficient computation. As the number of missing values in our type of data is large, we used instead a cross-validation method. The cross-validation consists of removing a chosen proportion of the tensor non-missing values, performing the decomposition for different ranks, and then evaluating the reconstruction error on the removed value, a type of K-fold cross-validation. We measure the reconstruction error using the Frobenius distance between this tensor and the original one on the non-missing values (see Figure 4.13).

The decomposition can be performed once the rank of the decomposition is chosen. The result of the three-way decomposition consists of three factor matrices, two of which with dimension $n_S \times R$, and one with dimension $n_C \times R$. The first two encode the degree of membership of each species to each component of the decomposition and are equal by construction. The third matrix encodes, for each contaminant, its degree of membership to each component. To facilitate the interpretation of the results, we threshold the membership degrees and decide whether each species and contaminant belongs to a component or not. To do this, we use K-means clustering on the degree of membership vectors to adaptively threshold the membership degrees. Species and contaminants may be allocated to 0 or several components (see Figure 4.14 and Figure 4.15).

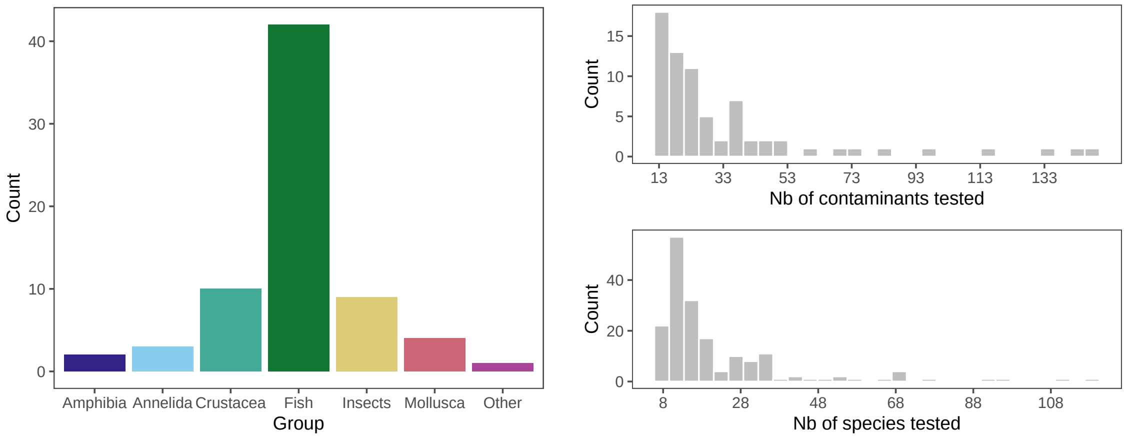


Figure 4.4: Left: Quasi-taxonomic composition of the species in the part of the data considered. The groups are defined according to the classification in Zwart (2001). Right: (Top) Number of contaminants tested for each species in the data considered (at least 13 contaminants by species). (Bottom) Number of species tested for each contaminant.

4.5.2 Results

We used this methodology on a part of the dataset described in Section 4.4.1 (see Figure 4.4 for a description of the restricted dataset). To lower the proportion of missing data, we only considered the species for which at least 13 contaminants have been tested; with this restriction, 95% of the data are missing. Using the cross-validation approach, the rank 41 was selected (see Figure 4.13). We obtained 41 components of the contaminants and the species. We present seven components among these, selecting those with the highest contrast using Figure 4.14 and Figure 4.15, deciding to only select the well-separated components in the sense of K-means clustering. These seven components are analyzed in Figure 4.5 and Figure 4.6.

Figure 4.5 presents the quasi-taxonomic composition of the seven components in count (left) and in relative proportion (right). Figure 4.19 and Figure 4.18 present the same result for other taxonomic ranks. We can not observe a clear difference in composition among the different components. This suggests that quasi-taxonomy does not appear to be the main driver for species to be co-clustered, or in other words, quasi-taxonomy does not appear to strongly determine species sensitivity. This observation should be modulated by the fact that, as we see in Figure 4.4, fish species are over-represented in the dataset, so they form a substantial part of every cluster.

Figure 4.6 presents the weights of each contaminant in the various components. We can observe that only one contaminant (Benzenamine) has a significant weight in component E. In components B and F, the most weighted contaminants are mostly insecticides. More precisely, component B is composed of six insecticides

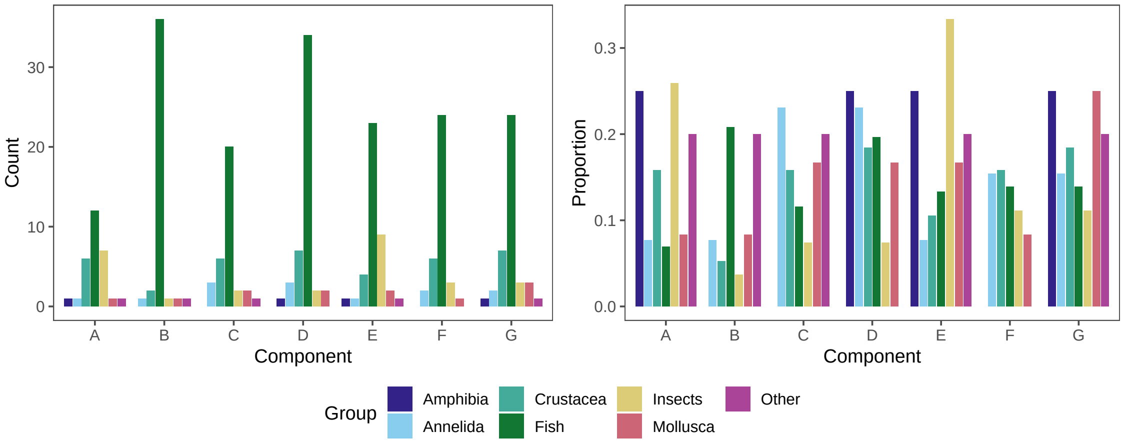


Figure 4.5: Quasi taxonomic composition of the components. The groups are defined according to the classification in Zwart (2001). Left: Number of species in each component. Right: Proportion of species compared to the distribution of species in the whole data in each component.

and one larvicide (Temephos). Component F contains five insecticides including four organophosphate insecticides, and Cadmium nitrate, which is mostly used in glass coloration. Component G is composed of two fungicides mainly used as wood preservatives. Component A mostly contains solvents. Two of the highest-weight contaminants in component C are composed of sulfuric acid, all the contaminants in component C have a wide variety of usages. Finally, component D contains some insecticides but also some contaminants notably used in photography fixators, antiseptics or hygiene products. Together, these observations suggest that components are associated to contaminants of similar type, or, in other words, that species which respond similarly to one type of contaminant could tend to respond similarly to another component of the same type. This could be seen as providing support to the idea that species sensitivity across contaminants may be correlated, which has been used for instance in Awkerman et al. (2008).

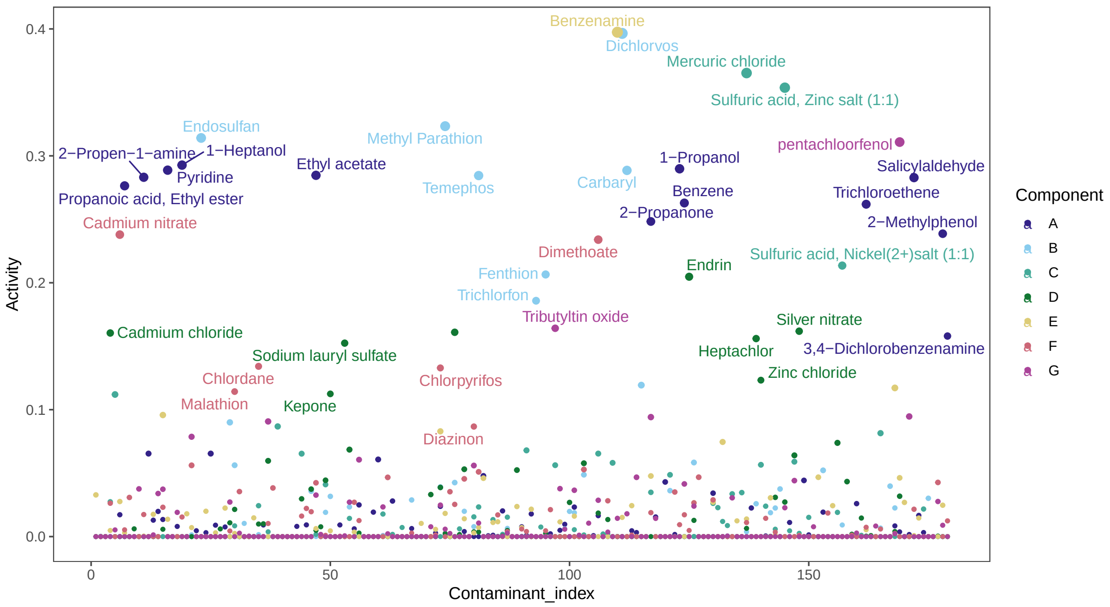


Figure 4.6: Weight of each contaminant in the various components. The size of the point is proportional to the intensity of the activity of the contaminants.

4.6 Discussion and future research

We presented a novel approach to SSD based on a Bayesian nonparametric mixture model. We performed an extensive comparison to the current SSD models both on simulated and on real datasets, to demonstrate the added value of the proposed approach. The BNP-SSD performs particularly well when the dataset deviates from a log-normal distribution, which allows to leverage its great flexibility in describing the data. At the same time, the proposed approach turns out to be relatively robust and does not seem prone to over-fitting. The BNP-SSD can be thought of as an intermediate model between the single component log-normal-SSD, and the KDE with as many components as there are species. Indeed, in all practical cases of the RIVM dataset, the number of clusters necessary to describe the data was no greater than 3.

The BNP-SSD provides several benefits for risk assessment: it is an effective and robust standard model that adapts to many datasets. As such, the BNP-SSD represents a safe tool to remove one of the arbitrary parametric assumptions of SSD (Forbes and Calow 2002). Moreover, as a Bayesian method, it readily provides credible intervals.

The traditional approach to SSD is to consider contaminants independently. In a context of data scarcity, only exacerbated by the drive to reduce animal testing, it would be desirable to leverage the long history of ecotoxicity testing to borrow information from experiments regarding different contaminants. Large databases are already available. We can hope that the ongoing discussion about transparency on the regulation of chemicals will even push towards greater public availability of data. Therefore, it is a timely effort to develop models that harness all the information already available about species' sensitivity to other contaminants. This is, in essence, the proposal brought forward by Awkerman et al. (2008) and Craig (2013) which use taxonomic information to predict the sensitivity of unknown species to a contaminant. An important by-product of the BNP-SSD approach is that it provides interesting opportunities for subsequent cluster analysis. We presented such an example of a post-processing analysis using non-negative tensor factorization, summarizing insights over 179 contaminants, which suggested some regularities in species sensitivity based on contaminant type. However, we did not observe a strong relation between taxonomy and species sensitivity across contaminants. As such, we may hypothesize that there could be structures other than taxonomic which would be relevant to species sensitivity, such as ecological niche.

This idea motivates a natural extension of the present work: the BNP-SSD could be augmented to model inter-contaminant variation in a hierarchical model. This is particularly relevant as some contaminants can have very similar toxicity because they belong to the same class of chemicals. A variety of Bayesian nonparametric hierarchical models already exist (Teh et al. 2006; Camerlenghi et al. 2019), and have demonstrated their usefulness and applicability in various contexts, so the tools necessary for this extension would only need some tailoring. Moreover, as discussed when studying the species clustering, the groups could be specified either based on the known chemical nature of the contaminants or learned in a flexible manner using a Bayesian nonparametric approach. This would open the door to a principled investigation of similarities among potentially very different contaminants. Additionally, one would also like to move forward from summarizing the sensitivity of species by a single value. Ecotoxicological tests are usually analyzed by fitting a dose-response model which describes several aspects of the species' response to a contaminant, such as the time between exposure and effect. Using all parameters from the dose-response model would imply performing the clustering in a higher dimensional space, increasing the discriminating power between groups and ultimately resulting in more meaningful clusters.

Acknowledgements

L. Alamichel is partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir. J. Arbel is partially supported by ANR-21-JSTM-0001 grant. I. Prünster is supported by the European Union – Next Generation EU funds, PRIN-PNRR 2022 (P2022H5WZ9).

This supplementary material is organized as follows: Section 4.A provides additional results on real data, Section 4.B contains details on the non-negative tensor factorization, and Section 4.C displays additional figures related to Section 4.5 of the main document.

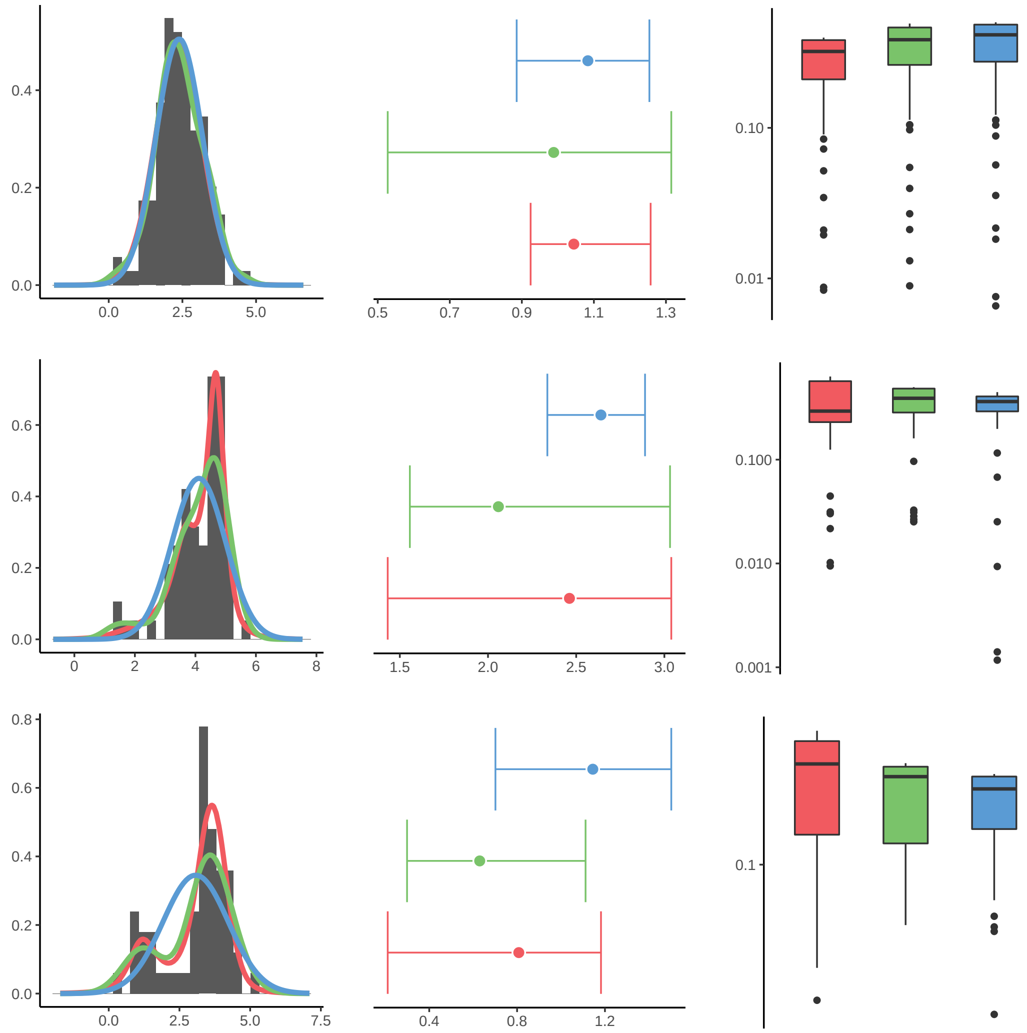
4.A Results on real data

This section illustrates the results stated in Section 4.4.2. We present the comparison of the three models (BNP, KDE and normal) on real data. We consider three categories of contaminants: contaminants with large datasets, consisting of more than 60 values, medium datasets, with around 25 values, and small datasets, with a little over 10 values. The three models were fitted on each dataset and we studied the estimate of the HC_5 and its credible interval, the LOO error and the shape of the estimated density compared to the histogram. The censored version of the same datasets was also studied with the BNP and normal model.

4.A.1 Model comparison on non-censored data

We present the results for the non-censored version of the datasets.

4.A.1.1 Large non-censored datasets



Density estimates HC$_5$ CPO

Figure 4.7: Density estimates, HC$_5$ and CPO for three large non-censored datasets. Red (—) for the BNP model, blue (—) for the normal model, and green (—) for the KDE model.From top to bottom: Cadmium chloride, Potassium Dichromate, and Carbaryl.

4.A.1.2 Medium non-censored datasets

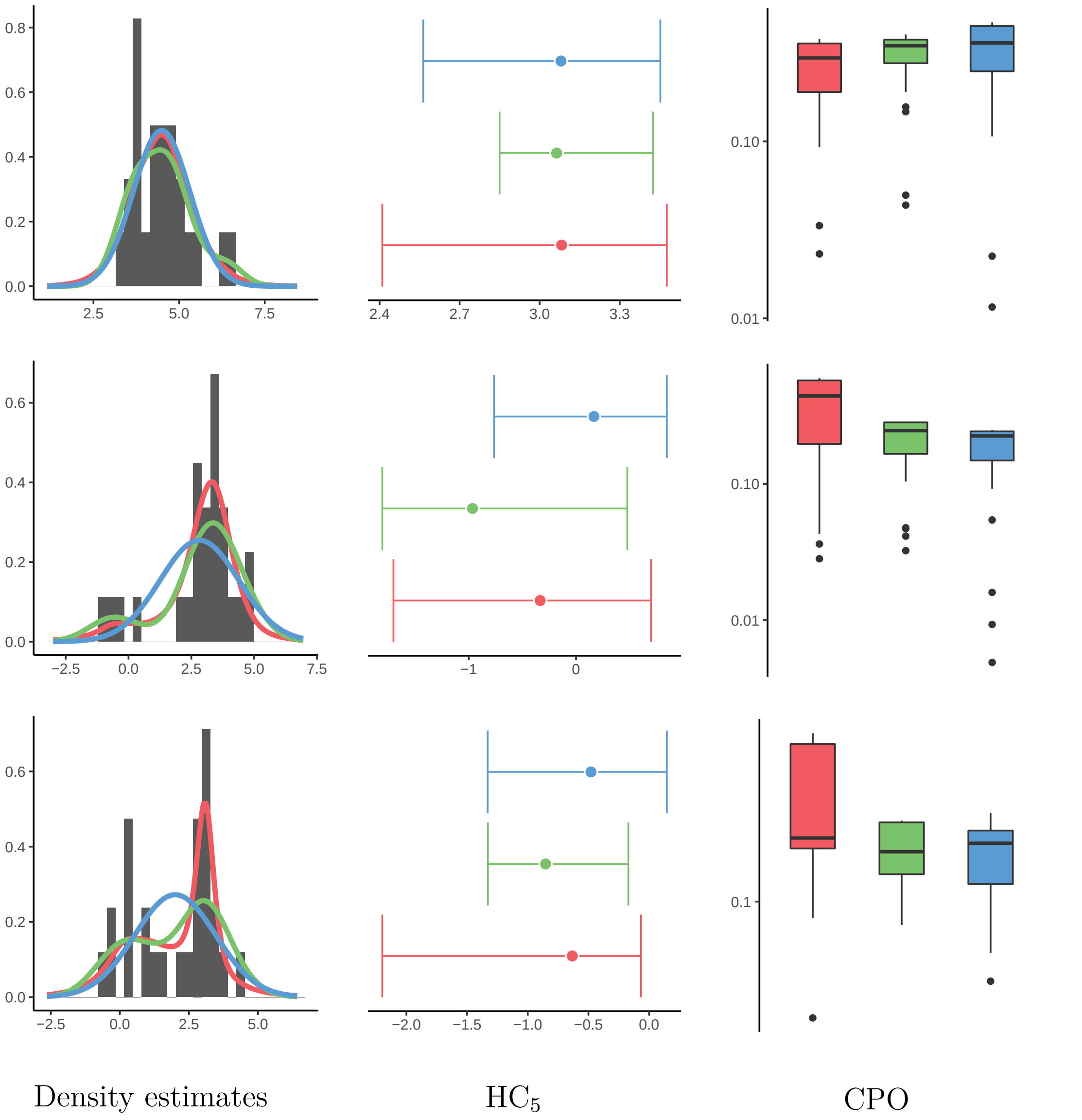
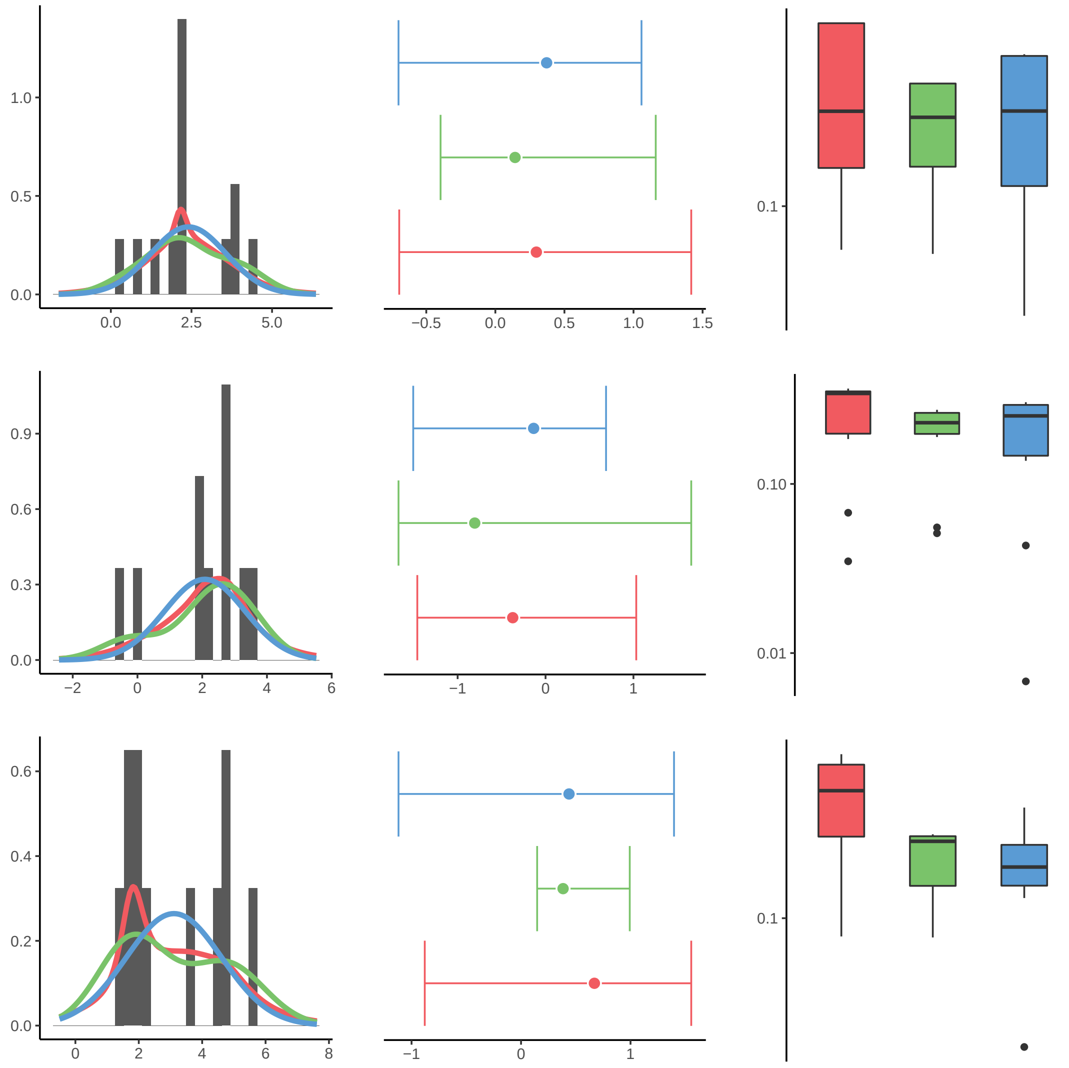


Figure 4.8: Density estimates, HC$_5$ and CPO for three medium non-censored datasets. Red (—) for the BNP model, blue (—) for the normal model, and green (—) for the KDE model.From top to bottom: 2,4-D Acid, Trichlorfon, Parathion.

4.A.1.3 Small non-censored datasets



Density estimates HC$_5$ CPO

Figure 4.9: Density estimates, HC$_5$ and CPO for three small non-censored datasets. Red (—) for the BNP model, blue (—) for the normal model, and green (—) for the KDE model.From top to bottom: Phosmet, Naled, Sodium dichromate.

4.A.2 Model comparison on censored datasets

Here we present the results for the censored version of the datasets. As extensions for kernel density estimators with censored data are not available (see Section 4.2.1), we only compare the normal and BNP models.

4.A.2.1 Large censored datasets

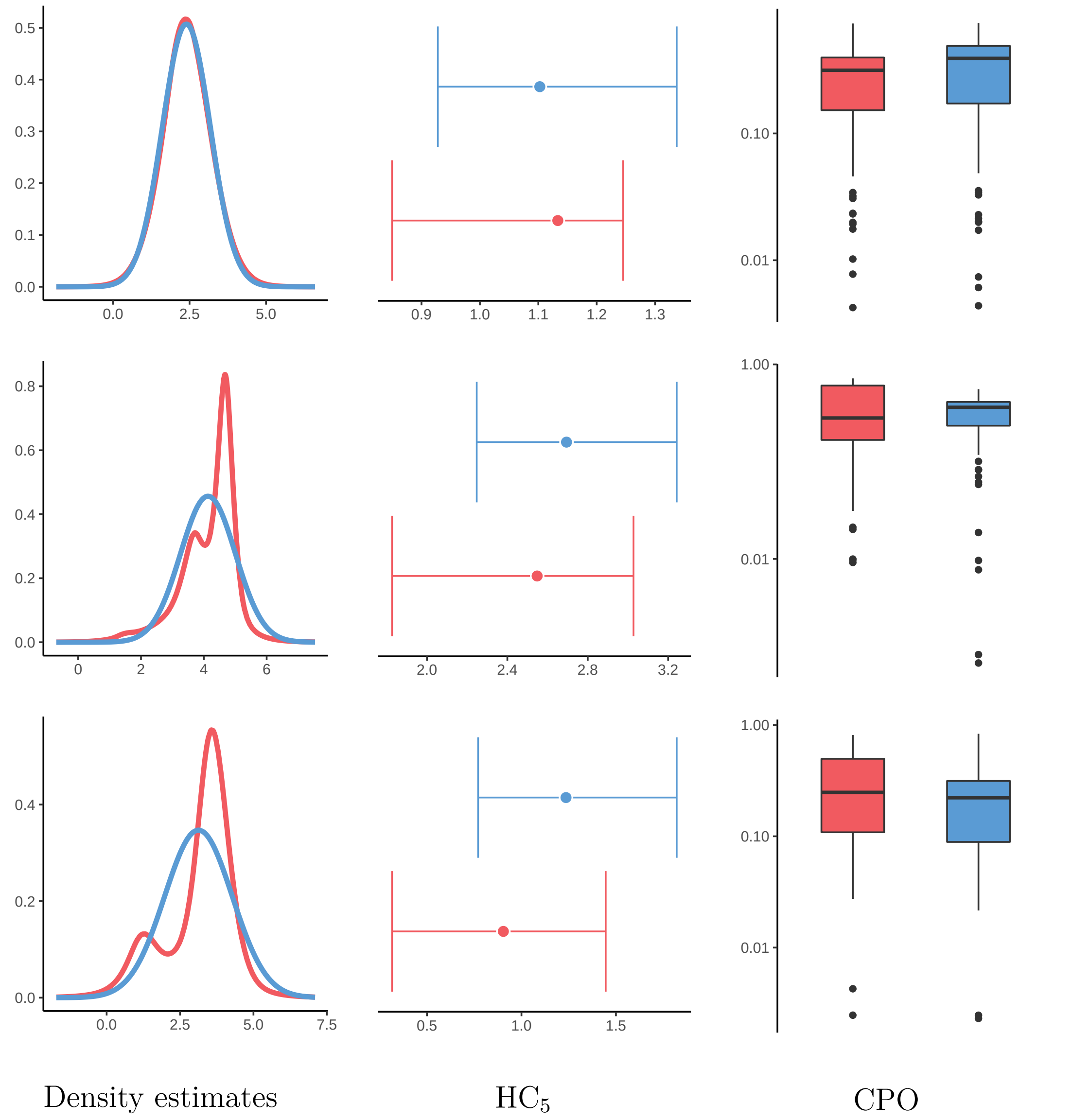
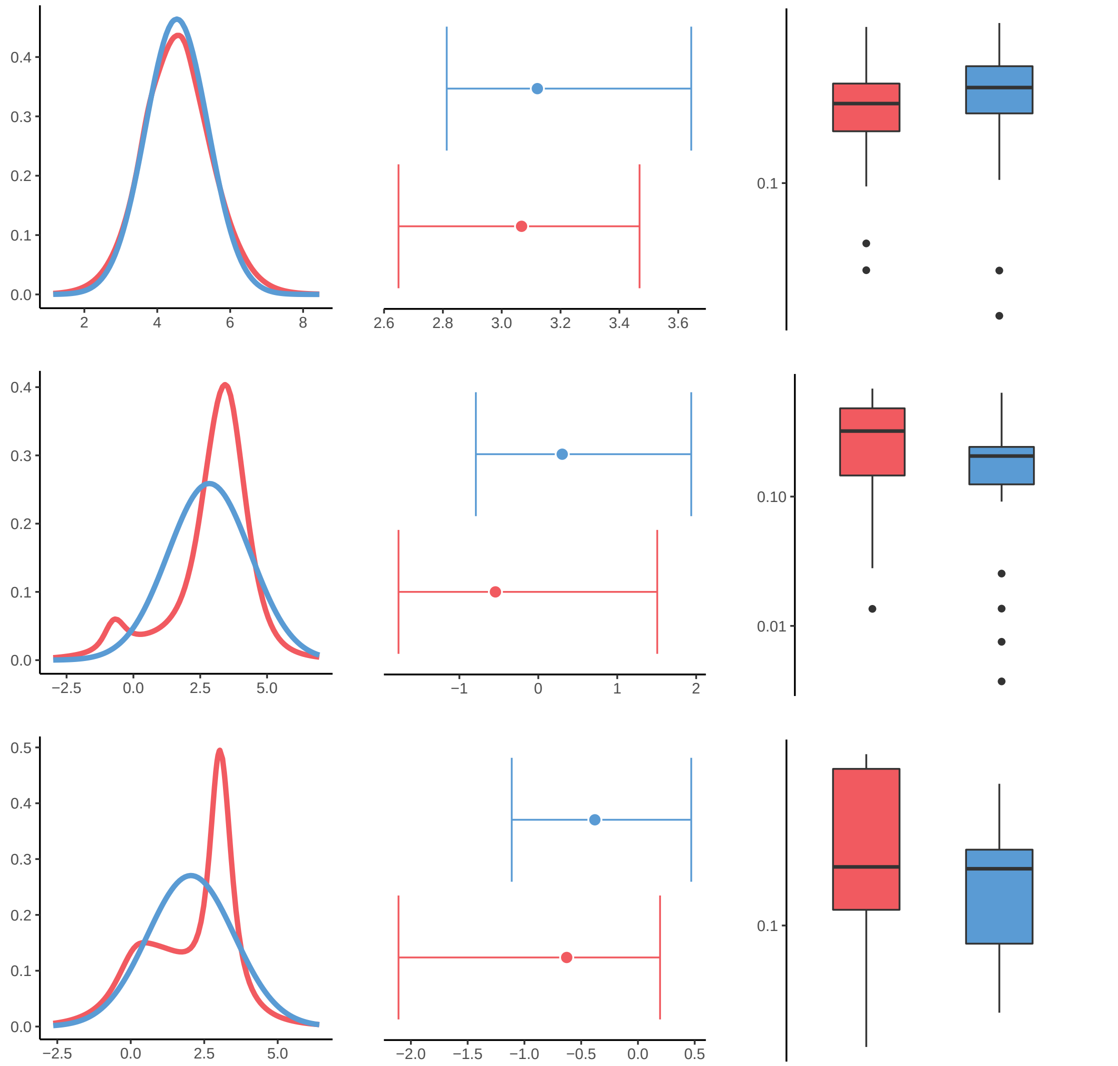


Figure 4.10: Density estimates, HC$_5$ and CPO for three large censored datasets. Red (—) for the BNP model, blue (—) for the normal model(KDE not implemented for censored data). From top to bottom: Cadmium chloride, Potassium Dichromate, and Carbaryl.

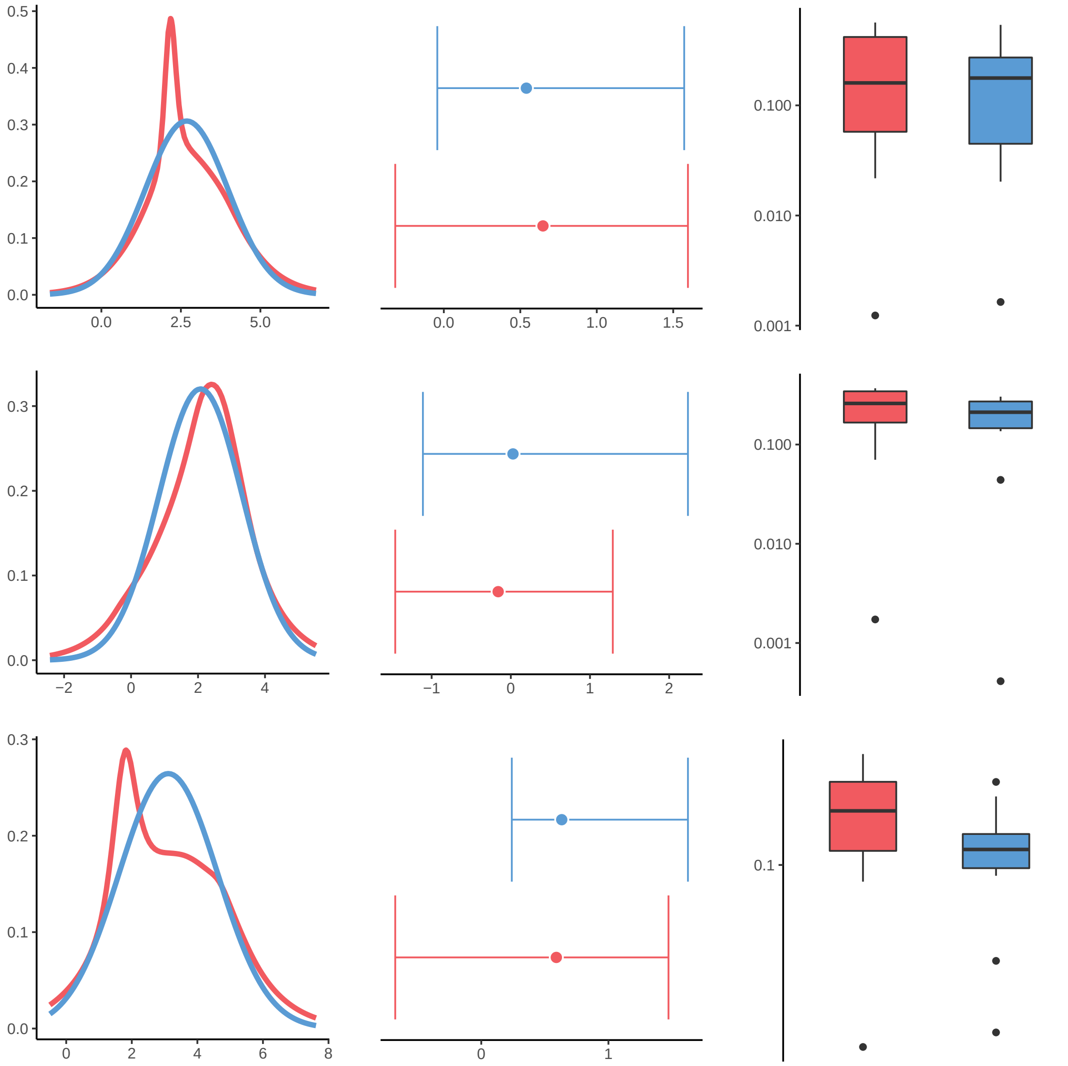
4.A.2.2 Medium censored datasets



Density estimates HC$_5$ CPO

Figure 4.11: Density estimates, HC$_5$ and CPO for three medium censored datasets. Red (—) for the BNP model, blue (—) for the normal model(KDE not implemented for censored data). From top to bottom: 2,4-D Acid, Trichlorfon, Parathion.

4.A.2.3 Small censored datasets



Density estimates HC$_5$ CPO

Figure 4.12: Density estimates, HC$_5$ and CPO for three small censored datasets. Red (—) for the BNP model, blue (—) for the normal model(KDE not implemented for censored data). From top to bottom: Phosmet, Naled, Sodium dichromate.

4.B Details on the non-negative tensor factorization

In this paper, a tensor is a high-dimensional form of matrices. We consider only three-order tensors, it means any tensor with three dimensions: $\mathbf{Y} \in \mathbb{R}^{I \times J \times K}$ where $I, J, K \in \mathbb{N}$. We introduce one useful product for tensor factorization or decomposition, the outer product. The outer product of two vectors $a \in \mathbb{R}^I$ and $b \in \mathbb{R}^J$, denoted by $\circ$, yields a matrix $A \in \mathbb{R}^{I \times J}$,

$$A = a \circ b = ab^T.$$

The outer product of three vectors $a \in \mathbb{R}^I$, $b \in \mathbb{R}^J$ and $c \in \mathbb{R}^K$ yields a third-order tensor $\mathbf{Y} \in \mathbb{R}^{I \times J \times K}$,

$$\mathbf{Y} = a \circ b \circ c, \qquad \text{with } y_{ijk} = a_i b_j c_k.$$

A three-order tensor defined as the outer product of three vectors is called a rank-one tensor. The rank of a tensor $\mathbf{Y}$ if defined as the minimal number R of rank-one tensors $\mathbf{Y}_1, \ldots, \mathbf{Y}_R$ such that $\mathbf{Y} = \sum_{r=1}^{R} \mathbf{Y}_r$.

The tensor factorization generalizes the matrix factorization techniques. The different matrix factorizations are useful notably for feature selection or dimensionality reduction. The tensor or multi-way array factorization allows to consider applications where the data contains high-order structures. One of the most popular models for the factorization of high-order tensors is the PARAFAC model. In the following, we will describe the decomposition for a third-order tensor, but the model can be extended to decompose a higher-order tensor.

4.B.1 PARAFAC factorization

We recall the Parallel Factors Analysis (PARAFAC) described in Section 4.5. Given a tensor $\mathbf{Y} \in \mathbb{R}^{I \times J \times K}$, the PARAFAC factorization is denoted by $\mathbf{Y} = [\![A, B, C]\!]$ where $A = [a_1, \ldots, a_R] \in \mathbb{R}^{I \times R}$, $B = [b_1, \ldots, b_R] \in \mathbb{R}^{J \times R}$ and $C = [c_1, \ldots, c_R] \in \mathbb{R}^{K \times R}$ are three components matrices. More formally,

$$\mathbf{Y} = \sum_{r=1}^{R} a_j \circ b_j \circ c_j + \mathbf{E} = [\![A, B, C]\!] + \mathbf{E},$$

where the tensor $\mathbf{E} \in \mathbb{R}^{I \times J \times K}$ represents the approximation error.

A difficult problem for the PARAFAC model is to choose the appropriate number of components R. This problem is equivalent to determining the rank of a tensor, in the sense that the rank of a tensor is the smallest number of R components in an exact PARAFAC decomposition, i.e. with a null approximation error tensor $\mathbf{E}$. Determining the rank of a given tensor is known to be a NP-hard problem.

In the ideal case, there is no noise in the data, so we can fit the PARAFAC model for different values of R until we have an exact decomposition. This assumes a perfect procedure for fitting the PARAFAC model, which is not the case in practice. Furthermore, in a more realistic case, the data is noisy and the procedure described is no longer applicable.

There are different proposed methods to solve this problem, such as core consistency diagnostic, residual analysis, visual appearance of loadings (also called factors represented by the components matrices), or cross-validation. The method we used in this paper is the cross-validation as described in Section 4.5.

4.B.2 Non-negative tensor factorization

On the PARAFAC model presented previously, it is possible to impose some non-negativity constraints. The non-negativity constraints allow to give physical meaning to the different components found.

In practice, adding constraints on the PARAFAC model yields to solve an optimization problem under constraints. Indeed, finding the PARAFAC factorization means solving the following optimization problem,

$$\min_{A,B,C} \|\mathbf{Y} - [\![A, B, C]\!]\|_F^2 ,$$

where $\|\cdot\|_F$ denotes the tensor Frobenius norm defined by $\|\mathbf{Y}\|_F^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2$. The non-negative PARAFAC factorization is a PARAFAC factorization with the following non-negative constraints

$$\min_{A,B,C \text{ s.t. } a_{ir}, b_{jr}, c_{kr} \in \mathbb{R}_+} \|\mathbf{Y} - [\![A, B, C]\!]\|_F^2 .$$

4.C Additional figures for Section 4.5

We now present some additional figures illustrating the results presented in Section 4.5.

Figure 4.13 illustrates the choice of the rank decomposition using the cross-validation method.

Figure 4.14 and Figure 4.15 illustrate the K-means clustering used on each components resulting from the tensor decomposition.

Figure 4.16 presents the heatmap of the components and the species, and Figure 4.17

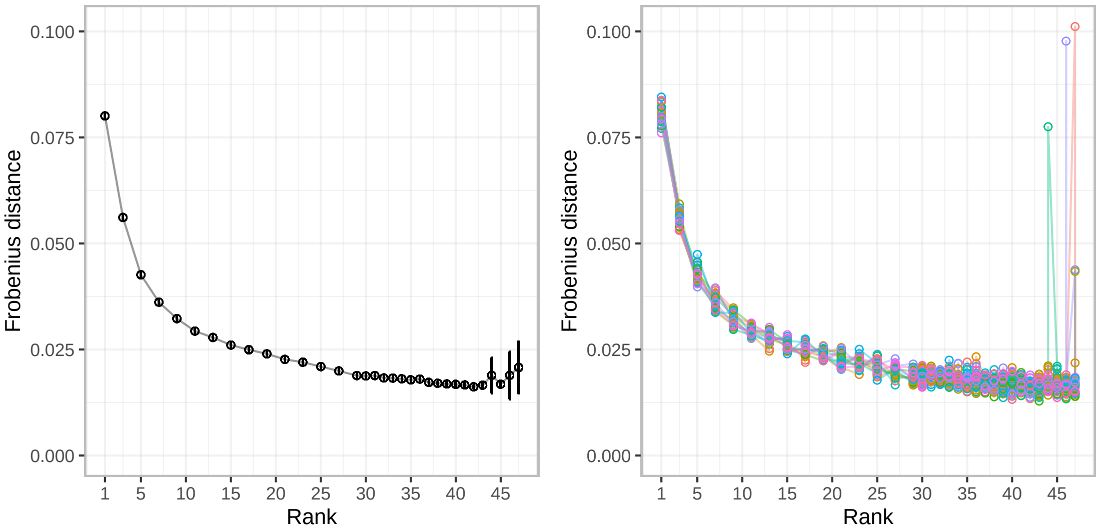


Figure 4.13: Cross-validation results for the tensor decomposition. Left: Frobenius distance mean of the 30-fold cross-validation with a confidence interval at 95% around each point with the rank varying between 1 and 47. Right: Frobenius distance for each of the 30 folds. The rank of the decomposition is chosen by taking the lowest rank (here 41) contained in the confidence interval from the rank minimizing the error (here 43).

presents the heatmap of the components and the contaminants. The heatmaps indicate which contaminants or species have more weights in a component. This is another way to illustrate the composition of the components.

Figure 4.19 and 4.18 present the taxonomic composition of the seven components previously selected at two different taxonomic ranks. These two figures support the idea that the taxonomy does not seem to strongly determine species sensitivity.

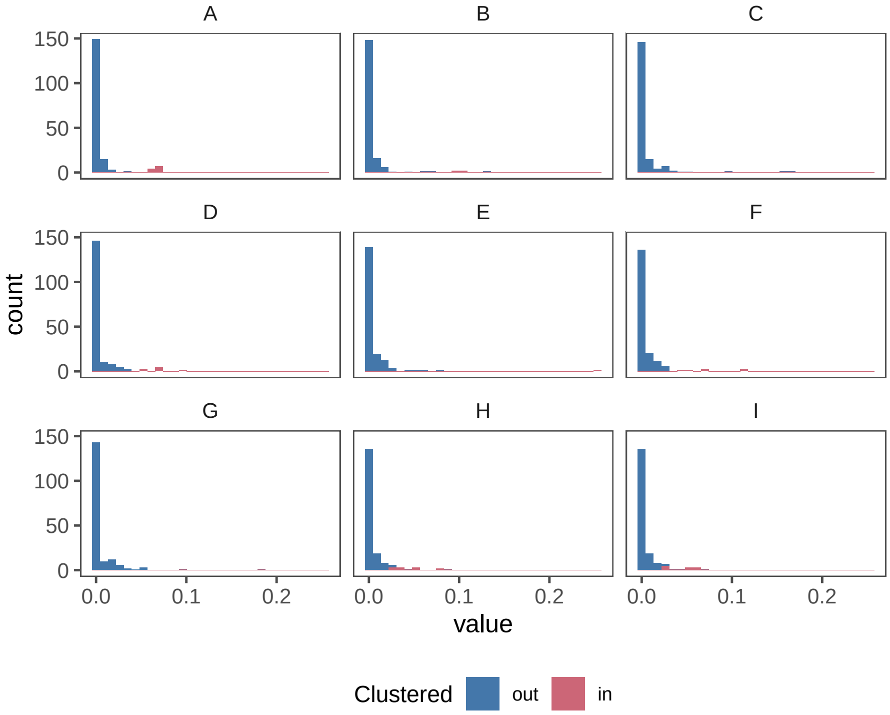


Figure 4.14: Distribution of the contaminant weight in each component. The presence of two well-separated groups in the distribution suggests separating the contaminants that belong to the component and those that do not. The clustering is performed using the K-means method. In Section 4.5 only the components where there is no overlap between the red part and the blue part, components A, B, C, D, E, F, and G are considered while components H and I are not.

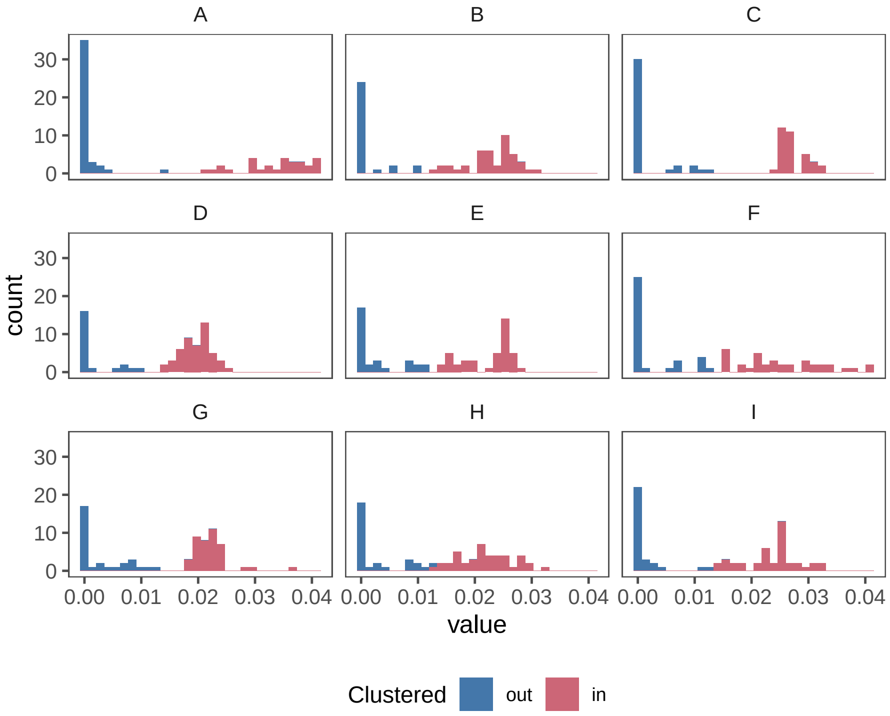


Figure 4.15: Distribution of the species weight in each component. The presence of two well-separated groups in the distribution suggests separating the contaminants that belong to the component and those that do not. The clustering is performed using the K-means method.

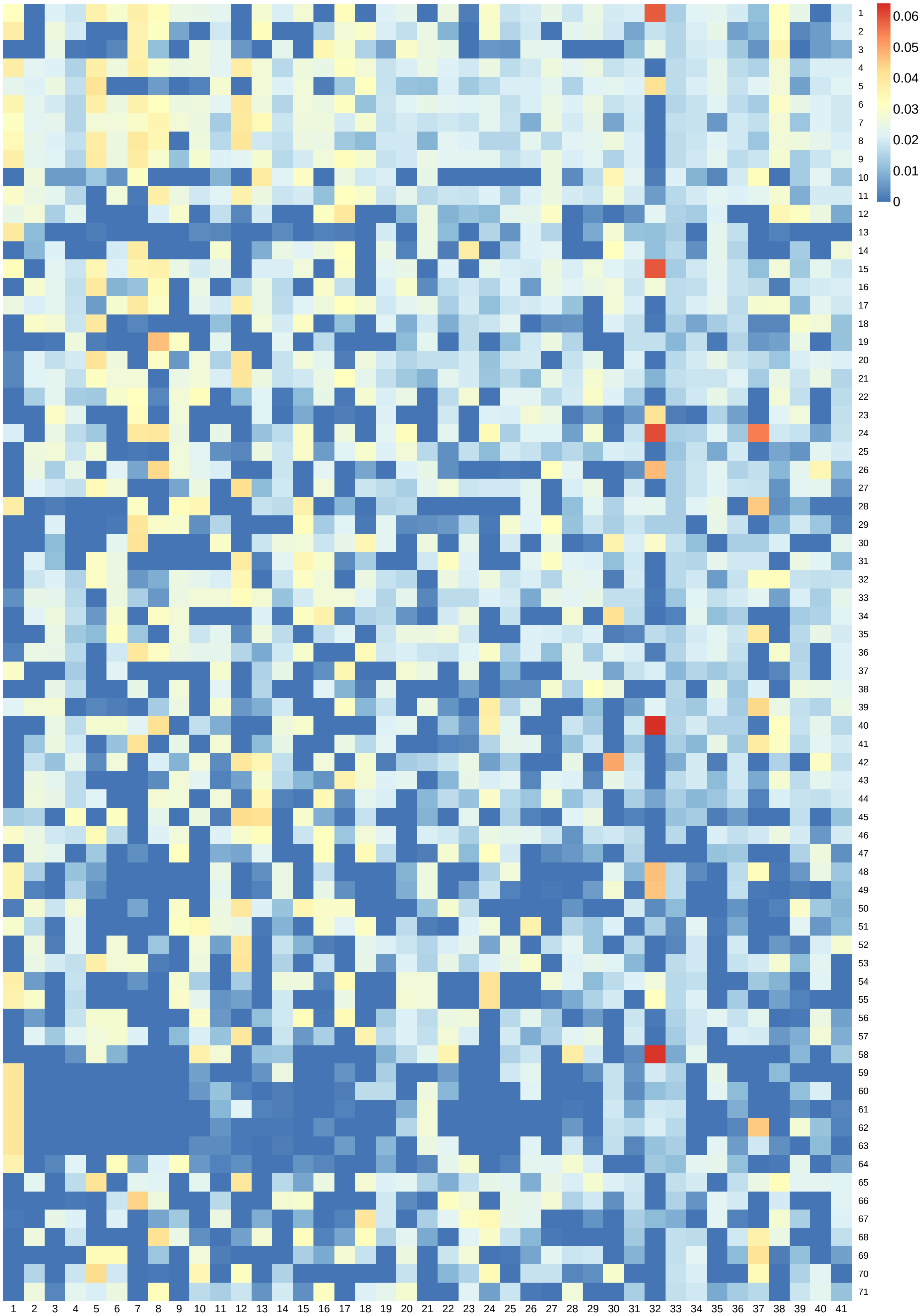


Figure 4.16: Heatmap of all the species and components.

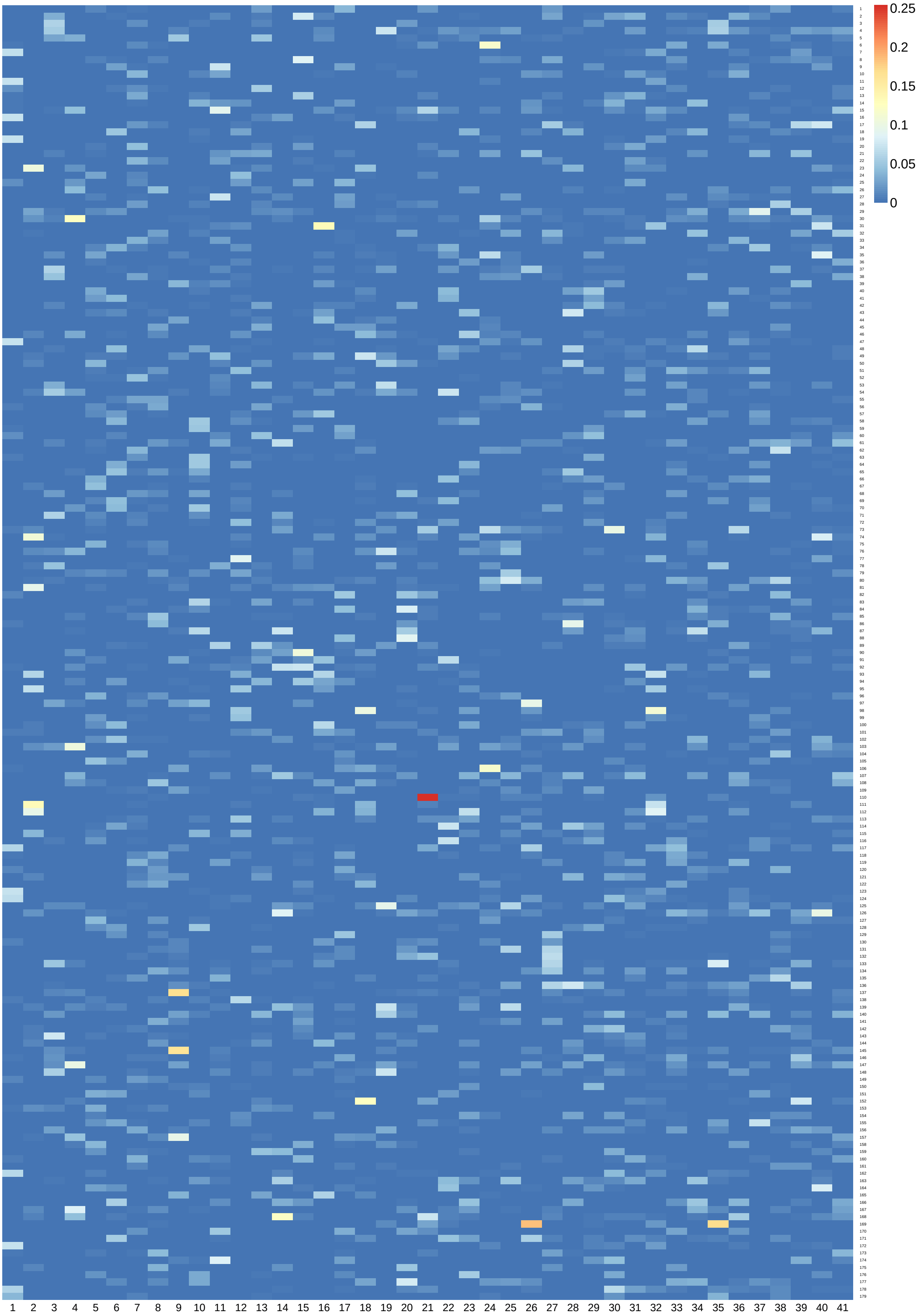


Figure 4.17: Heatmap of all the contaminants and components.

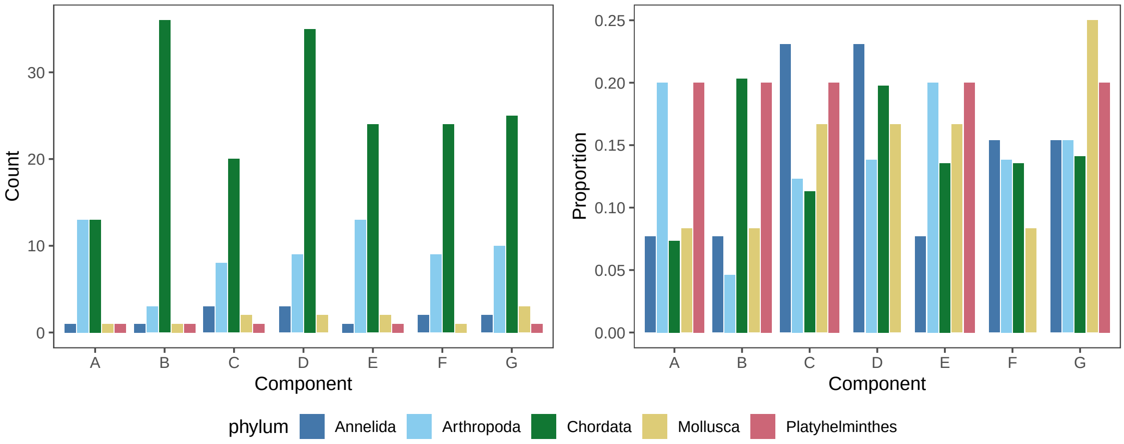


Figure 4.18: Taxonomic composition of the selected components at the Phylum level.

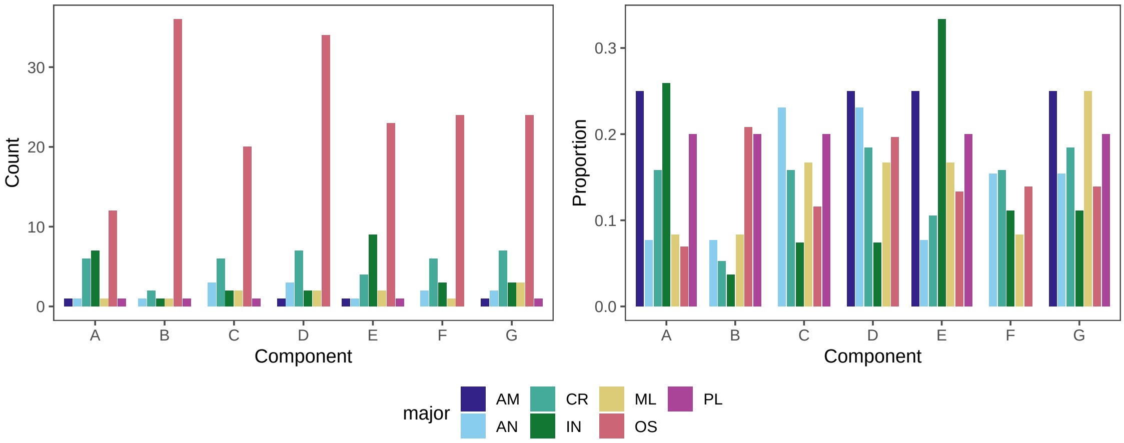


Figure 4.19: Taxonomic composition of the selected components at the major level.

References

Alamichel, L., J. Arbel, G. Kon Kam King, and I. Prünster (2024+). *Species Sensitivity Distribution revisited: a Bayesian nonparametric approach.* Submitted (cit. on p. 99).

Aldenberg, T. and J. S. Jaworska (2000). "Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions." In: *Ecotoxicology and Environmental Safety* 46.1, pp. 1–18 (cit. on pp. 103, 104, 107, 109).

Aldenberg, T., J. S. Jaworska, and T. P. Traas (2002). "Normal Species Sensitivity Distributions and Probalistic Ecological Risk Assessment". In: *Species sensitivity Distribution in Ecotoxicology.* Ed. by L. Posthuma, G. I. Suter, and T. P. Traas. Boca Raton, FL: Lewis Publishers, pp. 49–102 (cit. on p. 101).

Aldenberg, T. and W. Slob (1993). "Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data." In: *Ecotoxicology and Environmental Safety* 25.1, pp. 48–63 (cit. on p. 101).

ANZECC (2000). *Australian and New Zealand guidelines for fresh and marine water quality.* Tech. rep. Canberra, Australia: Australian et al. (cit. on p. 101).

Arbel, J., G. Kon Kam King, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2021). "BNPdensity: Bayesian nonparametric mixture modeling in R". In: *Australian & New Zealand Journal of Statistics* 63 (3), pp. 542–564 (cit. on p. 107).

Awkerman, J. A., S. Raimondo, and M. G. Barron (2008). "Development of species sensitivity distributions for wildlife using interspecies toxicity correlation models". In: *Environmental Science and Technology* 42.9, pp. 3447–3452 (cit. on pp. 103, 118, 120).

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). "Modeling with Normalized Random Measure Mixture Models". In: *Statistical Science* 28.3, pp. 313–334 (cit. on pp. 103–107).

Beraha, M., B. Guindani, M. Gianella, and A. Guglielmi (2022). *BayesMix: Bayesian Mixture Models in C++.* arXiv: 2205.08144 [stat.CO] (cit. on p. 106).

Binder, D. A. (1978). "Bayesian cluster analysis". In: *Biometrika* 65.1, pp. 31–38 (cit. on p. 109).

Bro, R. and H. A. L. Kiers (2003). "A new efficient method for determining the number of components in PARAFAC models". In: *Journal of Chemometrics* 17.5, pp. 274–286 (cit. on p. 116).

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). "Distribution theory for hierarchical processes". In: *Annals of Statistics* 47, pp. 67–92 (cit. on p. 120).

CCME (2007). "A protocol for the derivation of water quality guidelines for the protection of aquatic life". In: *Canadian Environmental Quality Guidelines.* Ccme

1991. Winnipeg: Canadian Council of Ministers of the Environment (cit. on p. 101).

Chen, L. (2004). "A conservative, nonparametric estimator for the 5th percentile of the species sensitivity distributions". In: *Journal of Statistical Planning and Inference* 123.2, pp. 243–258 (cit. on p. 102).

Cichocki, A., R. Zdunek, A. H. Phan, and S.-i. Amari (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons (cit. on p. 115).

Corradin, R., A. Canale, and B. Nipoti (2021). "BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures". In: *Journal of Statistical Software* 100.15, pp. 1–33 (cit. on p. 106).

Craig, P. S. (2013). *Exploring novel ways of using species sensitivity distributions to establish PNECs for industrial chemicals: Final report to Project Steering Group*. Tech. rep. (cit. on pp. 102, 103, 113, 120).

Craig, P. S., G. L. Hickey, R. Luttik, and A. Hart (2012). "Species non-exchangeability in probabilistic ecotoxicological risk assessment". In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 175.1, pp. 243–262. arXiv: arXiv:0811.2183v2 (cit. on p. 103).

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model". In: *Bayesian inference for gene expression and proteomics*, pp. 201–218 (cit. on p. 108).

Dalgarno, S. (2021). "shinyssdtools: A web application for fitting Species Sensitivity Distributions (SSDs)". In: *Journal of Open Source Software* 6.57, p. 2848 (cit. on p. 114).

ECHA (2008). "Characterisation of dose concentration-response for environment". In: *Guidance on information requirements and chemical safety assessment*. May. Helsinki: European Chemicals Agency. Chap. R.10 (cit. on pp. 101, 103, 112, 115).

Ferguson, T. S. and M. J. Klass (1972). "A representation of independent increment processes without Gaussian components". In: *The Annals of Mathematical Statistics* 43.5, pp. 1634–1643 (cit. on p. 107).

Forbes, V. E. and P. Calow (2002). "Species Sensitivity Distributions Revisited: A Critical Appraisal". In: *Human and Ecological Risk Assessment* 8.3, pp. 473–492 (cit. on pp. 101, 102, 119).

Fox, D., R. van Dam, R. Fisher, G. Batley, A. Tillmanns, J. Thorley, C. Schwarz, D. Spry, and K. McTavish (2021). "Recent Developments in Species Sensitivity Distribution Modeling". In: *Environmental Toxicology and Chemistry* 40.2, pp. 293–308 (cit. on p. 103).

Gauvin, L., A. Panisson, and C. Cattuto (2014). “Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach”. In: *PLoS ONE* 9.1. arXiv: arXiv:1308.0723v3 (cit. on p. 115).

Gelfand, A. E. (1996). “Model determination using sampling-based methods”. In: *Markov chain Monte Carlo in practice*, pp. 145–161 (cit. on p. 107).

Grist, E. P. M., K. M. Y. Leung, J. R. Wheeler, and M. Crane (2002). “Better bootstrap estimation of hazardous concentration thresholds for aquatic assemblages”. In: *Environmental Toxicology and Chemistry* 21.7, pp. 1515–1524 (cit. on p. 102).

He, W., N. Qin, X. Kong, W. Liu, W. Wu, Q. He, C. Yang, Y. Jiang, Q. Wang, B. Yang, and F. Xu (2014). “Ecological risk assessment and priority setting for typical toxic pollutants in the water from Beijing-Tianjin-Bohai area using Bayesian matbugs calculator (BMC)”. In: *Ecological Indicators* 45, pp. 209–218 (cit. on p. 102).

Hickey, G. L., P. S. Craig, R. Luttik, and D. de Zwart (2012). “On the quantification of intertest variability in ecotoxicity data with application to species sensitivity distributions”. In: *Environmental Toxicology and Chemistry* 31.8, pp. 1903–1910 (cit. on p. 112).

Jagoe, R. H. and M. C. Newman (1997). “Bootstrap estimation of community NOEC values”. English. In: *Ecotoxicology* 6.5, pp. 293–306 (cit. on p. 102).

Jara, A., T. E. Hanson, F. A. Quintana, P. Müller, and G. L. Rosner (2011). “DPpackage: Bayesian Semi- and Nonparametric Modeling in R”. In: *Journal of Statistical Software* 40.5, p. 1 (cit. on p. 106).

Jones, D. S., L. W. Barnthouse, G. W. Suter II, R. A. Efroymson, J. M. Field, and J. J. Beauchamp (1999). “Ecological risk assessment in a large river-reservoir: 3. Benthic invertebrates”. In: *Environmental Toxicology and Chemistry* 18.4, pp. 599–609 (cit. on p. 102).

Jordan, M. I. (2010). “Hierarchical Models, Nested Models and Completely Random Measures”. In: *Frontiers of statistical decision making and Bayesian analysis: In honor of James O. Berger*. New York: Springer, pp. 207–218. arXiv: arXiv:1312.6184v5 (cit. on p. 104).

Karabatsos, G. (2016). “A Menu-Driven Software Package of Bayesian Nonparametric (and Parametric) Mixed Models for Regression Analysis and Density Estimation”. In: *Behavior Research Methods*. eprint: ArXive-prints1506.05435 (cit. on p. 106).

Kefford, B. J., G. L. Hickey, A. Gasith, E. Ben-David, J. E. Dunlop, C. G. Palmer, K. Allan, S. C. Choy, and C. Piscart (2012). “Global scale variation in the salinity sensitivity of riverine macroinvertebrates: Eastern Australia, France, Israel and South Africa”. In: *PLoS ONE* 7.5, e35224 (cit. on pp. 102, 113).

Kingman, J. F. C. (1967). “Completely random measures”. In: *Pacific Journal of Mathematics* 21.1, pp. 59–78 (cit. on p. 104).

Kingman, J. F. C. (1975). “Random discrete distributions”. In: *Journal of the Royal Statistical Society. Series B* 37.1, pp. 1–22 (cit. on p. 105).

Kon Kam King, G., P. Veber, S. Charles, and M. L. Delignette-Muller (2014). “MOSAIC_SSD: A new web tool for species sensitivity distribution to include censored data by maximum likelihood”. In: *Environmental Toxicology and Chemistry* 33.9, pp. 2133–2139 (cit. on pp. 103, 105, 112).

Kooijman, S. (1987). “A safety factor for LC 50 values allowing for differences in sensitivity among species”. In: *Water Research* 21.3, pp. 269–276 (cit. on p. 101).

Lau, J. W. and P. J. Green (2007). “Bayesian model-based clustering procedures”. In: *Journal of Computational and Graphical Statistics* 16.3, pp. 526–558 (cit. on p. 108).

Leeuw, J. d. (2011). “The Multiway Package”. In: (cit. on p. 116).

Lijoi, A., R. H. Mena, and I. Prünster (2007). “Controlling the reinforcement in Bayesian non-parametric mixture models”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 69.4, pp. 715–740 (cit. on p. 106).

Lijoi, A. and I. Prünster (2010). “Models beyond the Dirichlet process”. In: *Bayesian nonparametrics.* Ed. by N. L. Hjort, C. C. Holmes, P. Müller, and S. G. Walker. Cambridge University Press, Cambridge, pp. 80–136 (cit. on pp. 104, 105).

Liu, Y., F. Wu, Y. Mu, C. Feng, Y. Fang, L. Chen, and J. P. Giesy (2014). “Setting Water Quality Criteria in China: Approaches for Developing Species Sensitivity Distributions for Metals and Metalloids”. In: *Reviews of Environmental Contamination and Toxicology volume.* Springer, pp. 35–57 (cit. on p. 101).

Lo, A. Y. (1984). “On a class of Bayesian nonparametric estimates: I. Density estimates”. In: *The Annals of Statistics*, pp. 351–357 (cit. on p. 104).

Meilă, M. (2007). “Comparing clusterings—an information based distance”. In: *Journal of Multivariate Analysis* 98.5, pp. 873–895 (cit. on p. 108).

Posthuma, L., G. W. Suter II, and P. T. Trass (2002). *Species sensitivity distributions in ecotoxicology.* CRC press, p. 616 (cit. on p. 101).

Regazzini, E., A. Lijoi, and I. Prünster (2003). “Distributional results for means of normalized random measures with independent increments”. In: *Annals of Statistics* 31.2, pp. 560–585 (cit. on p. 104).

Roux, D. J., S. H. J. Jooste, and H. M. MacKay (1996). “Substance-specific water quality criteria for the protection of South African freshwater ecosystems: Methods for derivation and initial results for some inorganic toxic substances”. In: *South African Journal of Science* 92.4, pp. 198–206 (cit. on p. 101).

Shao, Q. (2000). “Estimation for hazardous concentrations based on NOEC toxicity data: An alternative approach”. In: *Environmetrics* 11.5, pp. 583–595 (cit. on p. 101).

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Vol. 26. CRC press (cit. on p. 105).

Suter II, G. W., L. W. Barnthouse, R. A. Efroymson, and H. Jager (1999). “Ecological Risk Assessment in a Large River–Reservoir: 2. Fish Community”. In: *Environmental Toxicology and Chemistry* 18.4, pp. 589–598 (cit. on p. 102).

Teh, Y., M. Jordan, M. J. Beal, and D. M. Blei (2006). “Hierarchical Dirichlet processes”. In: *J. Am. Stat. Assoc.* 101.476, pp. 1566–1581. arXiv: arXiv:1210.6738v2 (cit. on p. 120).

USEPA (1998). *Guidelines for ecological risk assessment.* Tech. rep. Washington, DC.: US Environmental Protection Agency (cit. on p. 101).

Van Der Hoeven, N. (2001). “Estimating the 5-percentile of the species sensitivity distributions without any assumptions about the distribution”. In: *Ecotoxicology* 10.1, pp. 25–34 (cit. on p. 102).

Van Straalen, N. M. (2002). “Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc”. In: *Environmental Toxicology and Pharmacology* 11.3-4, pp. 167–172 (cit. on pp. 101, 102).

Wade, S. and Z. Ghahramani (2018). “Bayesian Cluster Analysis: Point Estimation and Credible Balls”. In: *Bayesian Analysis* 13.2, pp. 559–626 (cit. on pp. 108, 109).

Wagner, C. and H. Lokke (1991). “Estimation of ecotoxicological protection levels from NOEC toxicity data”. In: *Water Research* 25.10, pp. 1237–1242 (cit. on p. 101).

Wang, B., G. Yu, J. Huang, and H. Hu (2008). “Development of species sensitivity distributions and estimation of HC(5) of organochlorine pesticides with five statistical approaches.” In: *Ecotoxicology* 17.8, pp. 716–724 (cit. on p. 102).

Wang, Y., F. Wu, J. P. Giesy, C. Feng, Y. Liu, N. Qin, and Y. Zhao (2015). “Non-parametric kernel density estimation of species sensitivity distributions in developing water quality criteria of metals”. In: *Environmental Science and Pollution Research* 22.18, pp. 13980–13989 (cit. on pp. 102, 103, 105, 109).

Xing, L., H. Liu, X. Zhang, M. Hecker, J. P. Giesy, and H. Yu (2014). “A comparison of statistical methods for deriving freshwater quality criteria for the protection of aquatic organisms”. In: *Environmental Science and Pollution Research* 21.1, pp. 159–167 (cit. on p. 102).

Xu, F.-L., Y.-L. Li, Y. Wang, W. He, X.-Z. Kong, N. Qin, W.-X. Liu, W.-J. Wu, and S. E. Jorgensen (2015). “Key issues for the development and application of the species sensitivity distribution (SSD) model for ecological risk assessment". In: *Ecological Indicators* 54, pp. 227–237 (cit. on p. 102).

Xu, Y. and W. Yin (2013). "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion". In: *SIAM Journal on Imaging Sciences* 6.3, pp. 1758–1789 (cit. on p. 116).

Zajdlik, B. A., D. G. Dixon, and G. Stephenson (2009). "Estimating Water Quality Guidelines for Environmental Contaminants Using Multimodal Species Sensitivity Distributions: A Case Study with Atrazine". In: *Human and Ecological Risk Assessment* 15.3, pp. 554–564 (cit. on pp. 102, 103).

Zhao, J. and B. Chen (2016). "Species sensitivity distribution for chlorpyrifos to aquatic organisms: Model choice and sample size." In: *Ecotoxicology and Environmental Safety* 125, pp. 161–9 (cit. on pp. 101, 102).

Zwart, D. de (2001). "Observed regularities in species sensitivity distributions for aquatic species". In: *Species sensitivity distributions in ecotoxicology.* CRC Press (cit. on pp. 109, 112, 113, 117, 118).

Conclusion & Perspectives

Conclusion

This thesis provides some contributions to Bayesian mixture modeling, with a particular focus on Bayesian nonparametric mixture models. Mixture models are efficient tools for representing complex data. One of the most important tasks when using mixture models is to choose the number of components in the model. As presented in Section 1.2, using a Bayesian nonparametric mixture model is a natural way to deal with this problem by taking an infinite number of components a priori. In the first part of the thesis, some theoretical guarantees for this choice were provided.

More precisely, the first part contains theoretical contributions to studying the consistency of the number of clusters in Bayesian mixture models. In Chapter 2, we studied consistency for a wide class of Bayesian nonparametric priors, the Gibbs-type prior. Following the Miller and Harrison (2014) method, inconsistency for the number of clusters was proven for this class of priors. The same property was studied for overfitted finite mixture model priors such as the Dirichlet multinomial process, the Pitman–Yor multinomial process, and the normalized generalized Gamma multinomial process. We proved inconsistency for the number of clusters for overfitted mixture models based on these priors. We also discussed some approaches to solving inconsistency problems, notably the Merge-Truncate-Merge post-processing procedure from Guha et al. (2021). An approach to solve the inconsistency problem in the Dirichlet process mixture model is to add a prior on the parameter of the Dirichlet process. We presented in Chapter 3 the same approach for Pitman–Yor process mixture models. We proved the inconsistency for the number of clusters of Pitman–Yor mixture models when there is a prior on the concentration parameter.

In the second part, Chapter 4 provides some applied insight into Bayesian nonparametric (BNP) mixture modeling, particularly a practical application of Bayesian nonparametric mixture models. In this chapter, we proposed a new approach to assess ecological risk using Bayesian nonparametric mixture models. The proposed model has several advantages due to its Bayesian nonparametric nature alongside the fact that it is a mixture model. The model notably deals with multimodality and provides some uncertainty quantification and clustering estimation. In Chapter 4, the loss-based clustering approach proposed in Wade and Ghahramani (2018) is used to find a point estimate of the underlying partition. We exploited this clustering using a nonnegative tensor factorization. We obtained groups of species and contaminants, which depend on the type of contaminants but do not reveal any taxonomic link for species.

Perspectives

In what follows, we mention a few perspectives for each work in this thesis. One of the perspectives is an ongoing project and is more detailed.

BNP approach to Species Sensitivity Distribution (SSD)

Chapter 4 discussed a Bayesian nonparametric approach to assessing ecological risk using SSD method. We proposed a Bayesian nonparametric mixture model to estimate the risk distribution for one contaminant. We also study the information between the contaminants using the Nonnegative Tensor Factorization (NTF) as a post-processing procedure. Our results show that this information is significant, as the contaminant clusters found group together contaminants of a similar type. A natural extension of this work would be to incorporate the information from the contaminants in the model. We could consider all contaminants simultaneously by using a hierarchical model. This way, a post-processing procedure such as NTF will no longer be necessary to obtain this information, as the contaminants clustering will be provided by a Bayesian model. The uncertainties associated with this clustering will then be available thanks to the model.

Furthermore, it will be interesting to have theoretical validations of the clustering summarizing method used in Chapter 4. This is discussed in the last perspective of the thesis.

Number of cluster consistency

Chapter 2 contributed to extending some inconsistency results to a class of priors, Gibbs-type priors, and their finite-dimensional representations. A natural perspective to this work is to generalize the inconsistency results to a more general class of priors, the homogeneous normalized random measures with independent increments (NRMI) and its corresponding finite-dimensional representation normalized infinitely divisible multinomial process (NIDM). This extension is not an obvious result as the priors' exchangeable partition probability function (EPPF) are complex.

Simulation. Another natural way to improve the contribution presented in Chapter 2 is to conduct a simulation study also for the Pitman–Yor multinomial process (PYM). Indeed, this prior is more flexible than the Dirichlet multinomial process (DMP) but less used in practice. The PYM does not satisfy the conditions to apply Theorem 2.2 in Rousseau and Mengersen (2011). Hence it would be interesting to study the behavior of the mixing measure weights. A simulation study could show a difference in behaviour compared to the DMP in practice. It is also possible that the simulation study will show similar behavior; in this case, we might assume that the result of Theorem 2.2 holds for PYM but that some conditions could be relaxed.

Repulsive mixtures. At the end of Chapter 2, we discussed another possible class of Bayesian nonparametric mixture models known as *repulsive mixture* models. These models introduce some dependence on the components; they solve a well-known problem of e.g. Dirichlet process mixture model, which tends to create extra small clusters. While this sounds appealing, there is still a lack of theoretical guarantees for the repulsive mixture models. A recent paper Beraha et al. (2023) provides an EPPF for some particular repulsive mixture models. For these special cases, it could be possible, using the same machinery as in Chapter 2, to provide some consistency results for the number of clusters.

Misspecification. In the first part of this thesis, Bayesian nonparametric mixture models are *misspecified* in the sense that an infinite number of components mixture model is used to estimate a finite mixture model. This partially explains the inconsistency results. An interesting remaining open question is to study the consistency of the number of clusters when the true number of components is infinite, $K_0 = \infty$, or when it grows to infinity with the data sample size n. More precisely, we can wonder if the number of clusters of a Bayesian nonparametric mixture model will grow at the same rate as the number of components in the true distribution. Some ideas to solve this problem are discussed in Yang et al. (2020), which provide lower bounds on the ratio of posterior probabilities studied in Proposition 2.1. Providing some upper bounds on these ratios and studying the asymptotic regime of these bounds will be a way to answer this question.

Another type of *misspecification* can also be considered to get more realistic results. In the first part, we made the strong assumption that the kernel distribution was well-specified. A decisive improvement would be to prove some consistency while relaxing this assumption. However, as proved in Cai et al. (2021), adding even a small amount of misspecification leads to strong inconsistency even for mixture of finite mixture (MFM) models which are consistent in the well-specified case. The question of consistency under misspecification is an interesting and tough problem.

Indeed, most of the tools used in this thesis would have to be adapted, modified, or extended to deal with kernel misspecification. Inspiration for this could be taken from Guha et al. (2021), which shows some contraction rates for mild misspecification, or from Kleijn and van der Vaart (2006), which proves rates of convergence for density estimation under misspecification.

Necessity. In Chapter 2, we introduce Theorem 2.1 (Miller and Harrison 2014). This result states inconsistency for the number of clusters of mixture models satisfying Conditions 3.1 and 2.1. An interesting open question about this theorem is whether the conditions are sharp or not. In other words, since these conditions are sufficient, we wonder whether they are necessary. This question is important because the same conditions could be used to prove consistency results. As a preliminary finding, the MFM models, which are known to be consistent, do not satisfy Condition 2.1. This raises the question of whether this is also the case for other mixture models.

Contraction rate. The Merge-Truncate-Merge algorithm introduced in Chapter 2 can be used as a solution to inconsistency. To apply this procedure, all you need to know is the contraction rate in the Wasserstein distance of the model's mixing measure. Some contraction rates for different models are available in literature (see e.g. Ho and Nguyen 2016; Nguyen 2013). In this thesis, we provided a mixing measure contraction rate for the Pitman–Yor process (PY) mixture model in a quite restrictive case. Our result holds for a univariate location mixture. First, extending these results to a location-scale mixture would be interesting as these mixtures are more commonly used in practice. Our result uses a L_p contraction rate proved in Scricciolo (2014). Direct study of the Wasserstein distance might be one way of finding a more general contraction rate, but this is not a simple problem.

Hyperprior. Chapter 3 also contributes to the question of consistency for the number of clusters in Bayesian nonparametric mixture models. In this chapter, the idea to put an *hyperprior* on the parameters of the mixing measure prior as in Ascolani et al. (2022) is studied for the PY prior. In the Dirichlet process (DP) mixture model case, it is possible to deal with the inconsistency by adding a prior on the concentration parameter. In the PY case, the inconsistency remains when adding a prior on the concentration parameter and keeping the discount parameter fixed. A natural extension of this work is to study the case where the discount parameter σ is also considered random and is assigned some prior distribution. The study of the normalized stable process (NS) mixture model with a prior on σ could be a natural first step for this, as NS is a particular case of PY with α equals 0.

Another possible extension is to study other priors, such as the DMP case. In both cases, the difficulty comes from the fact that it will no longer be possible to separate the ratio, studied in Chapter 3, in a product of two distinct terms.

To sum up, the first part of this thesis focused on estimating the number of components in mixture models. However, this number of components is not the only quantity of interest relevant to mixture modeling. An even more informative quantity of interest is the underlying partition of the data. This extension and its advantages are developed in the following section.

Partition point-estimate

In Chapter 4, we used the variation of information ($\mathcal{VI}$) loss-based method to estimate the data clustering. This choice follows some empirical results, see e.g. Wade and Ghahramani (2018); Chaumeny et al. (2022). However, as presented in Section 1.3, theoretical results in the decision-theoretic framework for data clustering estimation are scarce. An ongoing project in collaboration with Filippo Ascolani aims to fill some of these theoretical gaps in the literature. I started this project with Filippo Ascolani during my PhD, while I was visiting the Statistical Department of Duke University.

As presented in Section 1.3, the problem of finding a good estimator for data clustering is complex. A way to face this issue is to choose a point estimate given a loss $\mathcal{L}$. The point estimate $z^\star_{1:n}$ is the estimate of the data partition, which minimizes the posterior expected loss

$$z^\star_{1:n} = \arg\min_{\hat{z}_{1:n}} \mathbb{E}\left[\mathcal{L}(z_{1:n}, \hat{z}_{1:n}) \mid X_{1:n}\right] = \arg\min_{\hat{z}_{1:n}} \sum_{z_{1:n}} \mathcal{L}(z_{1:n}, \hat{z}_{1:n}) p(z_{1:n} \mid X_{1:n}).$$

Rajkowski (2019) proved some results on the Bayesian estimator based on the $0-1$ loss function $\mathcal{L}_{0-1}$, the Maximum a posteriori (MAP). Firstly, it is proved that the minimizer under this loss is a unique *convex partition*. A *convex partition* is defined in Rajkowski (2019) as a partition that divides data into clusters whose convex hulls are disjoints with possibly the exception of one point. This leads to consistency results for the MAP in some special cases (Ascolani and Ghidini 2023). However, the $0-1$ loss function $\mathcal{L}_{0-1}$ is very basic. When n is large, a posterior sample generally rarely visits the same clustering twice, making the empirical MAP of the Markov chain Monte Carlo (MCMC) output very sensitive to the chain's initialization and of limited validity in practice. Other losses have been proposed with some good behavior in practice. The Binder loss (Binder 1978) and the variation of information loss (Meilă 2007) are the most used in practice (Wade 2023). Here, the focus is on the Binder loss because of its simple formulation.

We recall that the Binder loss is of this form

$$\mathcal{B}(z_{1:n}, \hat{z}_{1:n}) = \sum_{i<i'} \ell_1 \mathbb{I}_{z_i = z_{i'}} \mathbb{I}_{\hat{z}_i \neq \hat{z}_{i'}} + \ell_2 \mathbb{I}_{z_i \neq z_{i'}} \mathbb{I}_{\hat{z}_i = \hat{z}_{i'}},$$

where $z_{1:n}$, $\hat{z}_{1:n}$ are the latent clustering allocation variables, see Section 1.3 for more details. In the following, we will consider the case where $\ell_1 = \ell_2 = 1$. Because of its pairwise representation, this loss is easier to study than, for example, the variation of information (Section 1.3). Consequently, the expected Binder loss is a function of pairwise clustering equality probabilities. Therefore, the probability of two observations being in the same cluster is a key quantity.

We study a specific mixture model defined as follows

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} \mathcal{N}(\cdot \mid \theta_i, 1), \quad \theta_i \mid G \overset{\text{iid}}{\sim} G, \quad G \sim \text{Gt},$$

where Gt denotes a Gibbs-type prior with a standard Gaussian distribution base measure. The Gibbs-type prior, described in Section 1.1, is a general class of priors such that the same probability a priori is given on partitions of the same cluster sizes.

In the following, we present our preliminary results from the study of the Binder loss using the model above. Our objective is to prove that the expected Binder risk of a convex partition is lower than the risk of any non-convex partition with the same cluster sizes. We expect this to be the case for Gibbs-type priors since they depend only on the cardinality of the clusters.

Prior. In this case, the a priori expected Binder loss gives interesting information:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(z_{1:n}, \hat{z}_{1:n})\right] &= \sum_{i<i'} P(z_i = z_{i'}) \mathbb{I}_{\hat{z}_i \neq \hat{z}_{i'}} + P(z_i \neq z_{i'}) \mathbb{I}_{\hat{z}_i = \hat{z}_{i'}} \\ &= P(z_1 = z_2) \sum_{i<i'} \mathbb{I}_{\hat{z}_i \neq \hat{z}_{i'}} + [1 - P(z_1 = z_2)] \sum_{i<i'} \mathbb{I}_{\hat{z}_i = \hat{z}_{i'}}.\end{aligned}$$

With a Gibbs-type prior, the z_i are identically and independently distributed (*i.i.d.*) a priori. The a priori point estimate $z^\star_{\text{prior}}$ is the point estimate minimizing the *prior* expected loss:

$$\begin{aligned}z^\star_{\text{prior}} &= \arg\min_{\hat{z}_{1:n}} \mathbb{E}\left[\mathcal{L}(z_{1:n}, \hat{z}_{1:n})\right] \\ &= \arg\min_{\hat{z}_{1:n}} P(z_1 = z_2) \sum_{i<i'} \mathbb{I}_{\hat{z}_i \neq \hat{z}_{i'}} + [1 - P(z_1 = z_2)] \sum_{i<i'} \mathbb{I}_{\hat{z}_i = \hat{z}_{i'}}.\end{aligned}$$

Three scenarios are possible depending on the quantity $P(z_1 = z_2)$.

1. If $P(z_1 = z_2) < 1/2$ then the second term $[1 - P(z_1 = z_2)] \sum_{i<i'} \mathbb{I}_{\hat{z}_i = \hat{z}_{i'}}$ dominates

and $z^\star_{\text{prior}}$ characterizes the partition with n singletons. Each observation is in its own cluster.

2. If $P(z_1 = z_2) > 1/2$, the first term $P(z_1 = z_2) \sum_{i<i'} \mathbb{I}_{\hat{z}_i \neq \hat{z}_{i'}}$ dominates. Then $z^\star_{\text{prior}}$ characterizes the partition with 1 cluster where all the observations are clustered together.

3. Finally, if $P(z_1 = z_2) = 1/2$, neither term dominates and all possible partitions of $\{1, \dots, n\}$ minimize the expected loss.

Interestingly, this result is intuitive, as the prior probability for two observations to be clustered together characterizes the point-estimate clustering obtained.

Posterior. Working on the posterior expected risk of the Binder loss is more challenging. Similarly to the prior case, a key quantity is now the posterior probability for two observations to be in the same cluster. This quantity is of the following form:

$$\begin{aligned} p(z_i = z_j \mid X_{1:n}) &= \sum_{s>1} \sum_{\substack{A \in \mathcal{A}_s(n) \\ \text{s.t.}\, \exists \ell,\, X_i, X_j \in A_\ell}} p(A \mid X_{1:n}) \\ &= \sum_{s>1} \sum_{\substack{A \in \mathcal{A}_s(n) \\ \text{s.t.}\, \exists \ell,\, X_i, X_j \in A_\ell}} \frac{p(A)\, p(X_{1:n} \mid A)}{p(X_{1:n})}. \end{aligned}$$

As a starting point, the focus is on a univariate framework. In this framework, the observations can be ordered, and the *convex partitions* defined above are ordered partitions. More formally, we now consider the ordered dataset $X_{(1):(n)}$ and the associated clustering-allocation variables $z_{(1):(n)}$ where the clusters are relabeled with $z_{(1)} = 1$. A convex partition is characterized by the behavior of all the triplets of $z_{(1):(n)}$. The partition is convex if and only if all triplets of $z_{(1):(n)}$ are ordered. In this context, we study the posterior distribution of the allocation variables, and in particular, we focus on the three-point scenario.

We consider three ordered observations: $X_{(1)}$, $X_{(2)}$ and $X_{(3)}$. We compute the posterior probability that $X_{(1)}$ and $X_{(3)}$ are grouped together while $X_{(2)}$ is in another cluster. With a Gibbs-type process prior, we proved that this probability is smaller than having either $X_{(1)}$ and $X_{(2)}$ or $X_{(2)}$ and $X_{(3)}$ grouped,

$$\begin{aligned} p(z_{(1)} = z_{(3)}, z_{(1)} \neq z_{(2)} \mid X_{1:n}) \leq p(&\{z_{(1)} = z_{(2)}, z_{(1)} \neq z_{(3)}\} \\ &\cup \{z_{(2)} = z_{(3)}, z_{(1)} \neq z_{(3)}\} \mid X_{1:n}). \end{aligned}$$

A similar result also holds in a four points scenario,

$$p(z_{(1)} = z_{(4)}, z_{(2)} = z_{(3)}, z_{(1)} \neq z_{(2)} \mid X_{1:n}) \leq p(z_{(1)} = z_{(2)}, z_{(3)} = z_{(4)}, z_{(1)} \neq z_{(3)} \mid X_{1:n}).$$

We can also show a result for specific partitions. We define two partitions the first one $\tilde{A} \in \mathcal{A}_s(n)$ is convex and the second one $\hat{A} \in \mathcal{A}_s(n)$ is defined as,

$$\hat{A}_1 \neq \tilde{A}_1,\ \hat{A}_2 \neq \tilde{A}_2, \quad \hat{A}_1 \cup \hat{A}_2 = \tilde{A}_1 \cup \tilde{A}_2, \qquad \hat{A}_i = \tilde{A}_i,\ i = 3, \ldots, s,$$

hence elements in $\hat{A}_1$ and $\hat{A}_2$ are not in a convex order, i.e. there exist $(i, j) \in \hat{A}_1$ and $k \in \hat{A}_2$ such that $X_i < X_k < X_j$. We proved, in preliminary findings, that the posterior probability of $\tilde{A}$ is greater than the one of $\hat{A}$ for a case where $\tilde{A}_1$ and $\tilde{A}_2$ are equal-size clusters. The next step is to generalize this finding for any cluster size.

We then study the posterior expected risk associated with a convex partition. Because of previous results on the posterior distribution, we expect that this risk will be smaller than the risk of any non-convex partition with the same cluster sizes. However, the generalization of the previous findings to the posterior expected risk is not straightforward. Finally, we aim to show that similar properties as proved for the $\mathcal{L}_{0\text{-}1}$ loss in Rajkowski (2019) also hold for the Binder loss and then extend it for other losses such as the $\mathcal{VI}$ loss.

Bibliography

Ascolani, F. and V. Ghidini (2023). “Posterior clustering for Dirichlet process mixtures of Gaussians with constant data”. In: *14th Scientific Meeting of the Classification and Data Analysis Group*, p. 42 (cit. on p. 146).

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). “Clustering consistency with Dirichlet process mixtures”. In: *Biometrika. In press* (cit. on p. 145).

Beraha, M., R. Argiento, F. Camerlenghi, and A. Guglielmi (2023). *Normalized Random Meaures with Interacting Atoms for Bayesian Nonparametric Mixtures* (cit. on p. 144).

Binder, D. A. (1978). “Bayesian cluster analysis”. In: *Biometrika* 65.1, pp. 31–38 (cit. on p. 146).

Cai, D., T. Campbell, and T. Broderick (2021). “Finite mixture models do not reliably learn the number of components”. In: *International Conference on Machine Learning*. PMLR, pp. 1158–1169 (cit. on p. 144).

Chaumeny, Y., J. van der Molen Moris, A. C. Davison, and P. D. W. Kirk (2022). *Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice?* arXiv: 2207.14717 (cit. on p. 146).

Guha, A., N. Ho, and X. Nguyen (2021). “On posterior contraction of parameters and interpretability in Bayesian mixture modeling”. In: *Bernoulli* 27.4, pp. 2159–2188 (cit. on pp. 142, 145).

Ho, N. and X. Nguyen (2016). “On strong identifiability and convergence rates of parameter estimation in finite mixtures”. In: *Electronic Journal of Statistics* 10.1, pp. 271–307 (cit. on p. 145).

Kleijn, B. J. K. and A. van der Vaart (2006). “Misspecification in infinite-dimensional Bayesian statistics”. In: *The Annals of Statistics* 34.2, pp. 837–877 (cit. on p. 145).

Meilă, M. (2007). “Comparing clusterings—an information based distance”. In: *Journal of Multivariate Analysis* 98.5, pp. 873–895 (cit. on p. 146).

Miller, J. W. and M. T. Harrison (2014). “Inconsistency of Pitman-Yor process mixtures for the number of components”. In: *The Journal of Machine Learning Research* 15.1, pp. 3333–3370 (cit. on pp. 142, 145).

Nguyen, X. (2013). “Convergence of latent mixing measures in finite and infinite mixture models”. In: *The Annals of Statistics* 41.1, pp. 370–400 (cit. on p. 145).

Rajkowski, Ł. (2019). “Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model”. In: *Bayesian Analysis* 14.2 (cit. on pp. 146, 149).

Rousseau, J. and K. Mengersen (2011). “Asymptotic behaviour of the posterior distribution in overfitted mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, pp. 689–710 (cit. on p. 144).

Scricciolo, C. (2014). “Adaptive Bayesian Density Estimation in L_p-metrics with Pitman-Yor or Normalized Inverse-Gaussian Process Kernel Mixtures”. In: *Bayesian Analysis* 9.2 (cit. on p. 145).

Wade, S. (2023). “Bayesian cluster analysis”. In: *Philosophical Transactions of the Royal Society A* 381.2247, p. 20220149 (cit. on p. 146).

Wade, S. and Z. Ghahramani (2018). “Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)”. In: *Bayesian Analysis* 13.2. Publisher: International Society for Bayesian Analysis, pp. 559–626 (cit. on pp. 143, 146).

Yang, C.-Y., E. Xia, N. Ho, and M. I. Jordan (2020). *Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models*. arXiv: 1905.09959 [stat.ML] (cit. on p. 144).