Bayesian Analysis (2021) **TBA**, Number TBA, pp. 1

Contaminated Gibbs-type priors

Federico Camerlenghi[*,†] and Riccardo Corradin[‡] and Andrea Ongaro[*]

Abstract. Gibbs-type priors are combinatorial processes widely used as key components in several Bayesian nonparametric models. By virtue of their flexibility and mathematical tractability, they turn out to be predominant priors in species sampling problems and mixture modeling. We introduce a new family of processes which extends the Gibbs-type one, by including a contaminant component in the model to account for an excess of observations with frequency one. We first investigate the induced random partition, the associated predictive distribution, the asymptotic behavior of the total number of blocks and the number of blocks with a given frequency: all the results we obtain are in closed form and easily interpretable. A remarkable aspect of contaminated Gibbs-type priors relies on their predictive structure, compared to the one of the standard Gibbs-type family: it depends on the additional sampling information on the number of observations with frequency one out of the observed sample. As a noteworthy example we focus on the contaminated version of the Pitman-Yor process, which turns out to be analytically tractable and computationally feasible. Finally we pinpoint the advantage of our construction in different applications: we show how it helps to improve predictive inference in a species-related dataset exhibiting a high number of species with frequency one; we also discuss the use of the proposed construction in mixture models to perform density estimation and outlier detection.

Keywords: Bayesian nonparametrics, Gibbs-type priors, species sampling models, random partitions, mixture models.

MSC2020 subject classifications: Primary 62G05, 62F15, 60G25.

1 Introduction

The great success of the Dirichlet process within the Bayesian nonparametric framework has paved the way for the definition and investigation of a large variety of random probability measures. Indeed, since the seminal paper by Ferguson (1973), several discrete nonparametric priors have been proposed to accommodate for exchangeable observations, among these we mention: the Pitman-Yor process or two parameter Poisson-Dirichlet process (Perman et al., 1992; Pitman, 1996); species sampling models (Pitman, 1996); priors based on normalization of completely random measures (Regazzini et al., 2003; Lijoi and Prünster, 2010). Gibbs-type priors are another important class of Bayesian nonparametric models early introduced by (Gnedin and Pitman, 2005) and recently investigated in (De Blasi et al., 2015). The Gibbs-type family has the advantage to balance modeling flexibility and mathematical tractability. These processes have

[*]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy federico.camerlenghi@unimib.it, andrea.ongaro@unimib.it

[†]Also affiliated to BIDSA, Bocconi University, Milan, Italy.

[‡]School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom riccardo.corradin@nottingham.ac.uk

© 2021 International Society for Bayesian Analysis

been successfully used in several areas, just to mention a few examples: to face prediction within species sampling framework (e.g. Lijoi et al., 2007a), as mixing measures in mixture models (e.g. Ishwaran and James, 2001; Lijoi et al., 2007b), for survival analysis (e.g. Jara et al., 2010), and for applications in linguistic and information retrieval (e.g. Teh, 2006; Teh and Jordan, 2010). Heaukulani and Roy (2020) have recently discussed a class of feature allocation models parametrized by Gibbs-type random probability measures. Although these priors have been widely employed in the Bayesian nonparametric literature, sometimes a pure Gibbs-type prior may not be flexible enough in presence of a large number of unique values or data contaminated by anomalous quantities. Such a behaviour may occur in several applied fields: in taxonomic data (see Section 5.2), where the species names can be miss-reported; in linguistics, where the author of a manuscript can include many neologisms, and the words' distribution may include a large number of singletons (see, e.g., Harald, 2001); in genomics, where the technologies are subject to sequencing errors (see, e.g., Stoler and Nekrutenko, 2021); in mixture models, when data are contaminated by outilers (see Section 6.2).

In the present paper we introduce a new family of Bayesian nonparametric models where a Gibbs-type prior is contaminated with an exogenous diffuse probability measure, called *contaminant measure*. More precisely, we define a new random probability measure as a convex linear combination of a Gibbs-type prior $\tilde{q}$ and a diffuse probability P_0, i.e. we deal with $\tilde{p} = \beta\tilde{q} + (1-\beta)P_0$, where $\beta \in [0,1]$ is a weight which tunes the impact of the contaminant measure. We refer to $\tilde{p}$ as a contaminated Gibbs-type prior (see Definition 1) and its distribution is then used as a nonparametric prior in a Bayesian context. Although this may seem a simple modification of an existing prior, the proposed construction has a profound impact on the predictive structure of the model and thus on posterior inference. Indeed $\tilde{p}$ is a random probability measure outside the Gibbs-type family, whose predictive distribution has a remarkable advantage with respect to the one induced by a standard Gibbs-type prior: it depends on the additional sampling information on the number of observations with frequency one out of the observed sample. As a consequence our construction allows to enrich the predictive structure of Gibbs-type priors, still maintaining analytical tractability which is a peculiar aspect to develop efficient sampling schemes to address posterior inference. Secondly the use of a contaminant measure P_0 accounts naturally for the presence of anomalies in the data (observations which are under some respects singular). This is crucial when the main inferential interest concerns modeling and predicting observations with frequency one, but it is also of great relevance in other inferential contexts, as modeling such observations incorrectly may lead to wrong inferential conclusions. As a remarkable example, in Section 5.1 we show that this induces a severe bias in estimating critical parameters (such as the reinforcement parameter σ), which have a strong impact on predictive inference, such as the number of new distinct observations in an additional sample. We also point out how contaminated Gibbs-type priors can be exploited for modeling discrete data, when one needs to inflate the observations with frequency one. As an example, in Section 5.2, we consider species detection data from the Global Biodiversity Information Facility project (GBIF.org, 2021) with a high number of species detected only once. Within this framework, we show the advantage of our model with respect to the traditional Pitman-Yor process, and we empirically prove

how they lead to different inferential conclusions. In Section 6 we also discuss the use of a contaminated random probability measure in mixture models, when outliers are present in the data. The use of a contaminated Gibbs type prior, instead of a standard model, results in smoother density estimates and an efficient detection of outliers, which are captured by the contaminant measure.

In order to describe the theoretical properties of the proposed model and develop efficient sampling procedures, we first introduce and deeply investigate the random partition structure induced by contaminated Gibbs-type priors. Moreover we also derive predictive distributions and asymptotic results for the total number of clusters and the number of clusters with a certain frequency. All the stated results are available in closed form, they are simple and with a natural interpretation. The induced prediction rule can be easily explained in terms of a new Chinese restaurant with a room for social persons and one for loners. As a concrete example, throughout the paper we focus on the contaminated version of the Pitman-Yor process, which exhibits more tractable expressions for all the quantities of interest and to face prediction. With regard to this last issue, we determine Bayesian estimators for functionals which depend on an additional unobserved sample of arbitrary size, thus extending some of the results in (Lijoi et al., 2007a; Favaro et al., 2009). As for the contaminated Pitman-Yor process, we finally determine a Pólya urn representation of the updating mechanism.

To the best of our knowledge, the Bayesian nonparametric literature has never focused on Gibbs type priors contaminated with a diffuse measure to model the excess of singular observations. In the past literature, early studies on contamination of random probability measures have been faced by Quintana (2006), in a partially exchangeable setting. The authors used a convex combination of a common discrete component, shared among different groups, with group-specific random probability measures. Later, convex linear combinations of a random probability measure with an atomic component have been used in numerous Bayesian nonparametric models to define spike and slab priors. See, e.g., Scarpa and Dunson (2009); Canale et al. (2017) and references therein. Here we focus on a different convex combination in which we substitute the spike with a diffuse probability measure P_0, with a completely different goal. We also mention that Beraha et al. (2021) have recently used a contamination of the Dirichlet process with a diffuse measure to define the common base measure of a vector of hierarchical Dirichlet processes. Thus, the model proposed by Beraha et al. (2021) is completely different with respect to our proposal, and the aim is different as well: their goal is to define a vector of Dirichlet processes having common, but also specific (i.e., not shared), random atoms. On the contrary, in the present paper we focus on a different prior distribution, designed for one group of observations, with the aim to obtain a richer predictive structure to model singular observations. Models which include contamination have been considered from a probabilistic viewpoint for discrete random structures, for example Kingman's paintbox representation (Kingman, 1978) with dust, coalescent with dust (see, e.g., Freund and Möhle, 2017), and trait allocations with dust (Campbell et al., 2018). However, to the best of our knowledge, these models have never been used in statistical applications. Finally, we mention that priors with a contaminant term have been analysed in Bayesian robustness (ϵ-contaminated priors, Berger and Berliner, 1986). While such a construction is connected with our specification, where a model is contaminated by an

exogenous term, it has been used to study the robustness of Bayesian procedures when the prior is perturbed.

The rest of the article is organized as follows. In Section 2 we introduce the family of contaminated Gibbs-type priors. Section 3 presents the main results on the random partition induced by a contaminated Gibbs-type prior, the predictive structure and asymptotic results on the number of clusters. We further specialize such findings to the contaminated Pitman-Yor process case (Section 4). Section 5.1 shows the flexibility and the predictive ability of contaminated Gibbs-type priors through an extensive simulation study. In Section 5.2 we apply the contaminated Pitman-Yor process to analyse the GBIF dataset, showing the benefits of including a contaminant measure in the model specification. In Section 6, contaminated Gibbs-type priors are employed as mixing measures in mixture models for clustering problems and density estimation in presence of outliers. Section 6.1 describes a detailed algorithm to perform posterior inference in a mixture context, while an astronomical data example (Ibata et al., 2011) is presented in Section 6.2. The paper ends with a discussion on the use of contaminated Gibbs-type priors in other contexts (Section 7). Proofs and additional details are deferred to the Supplementary Material.

2 Definition of contaminated Gibbs-type priors

Let $\{X_i\}_{i\geq 1}$ be a sequence of observations taking values in a Polish space $\mathbb{X}$, equipped with its Borel σ-field $\mathscr{X}$. In the Bayesian nonparametric setting $X_1, X_2, \ldots$ are typically supposed to be exchangeable (de Finetti, 1937), which is tantamount to saying that there exists a random probability measure $\tilde{p}$ such that $X_i|\tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, where the distribution of $\tilde{p}$ works as a prior in the Bayesian nonparametric framework. The distribution of $\tilde{p}$, indicated by $\mathscr{Q}$, is called the de Finetti measure of the sequence $X_1, X_2, \ldots$, and several prior specifications $\mathscr{Q}$ are available in the Bayesian nonparametric literature. Among these we mention the remarkable class of species sampling models (Pitman, 1996). We recall that an exchangeable sequence of observations $\{X_i\}_{i\geq 1}$ is called a *species sampling sequence* if and only if it is governed by a distribution of the following type

$$\tilde{p} = \sum_{j\geq 1} p_j \delta_{Z_j} + \left(1 - \sum_{j\geq 1} p_j\right) P_0, \tag{1}$$

for a sequence of random weights $\{p_j\}_{j\geq 1}$ with $p_j \geq 0$ and $\sum_{j\geq 1} p_j \leq 1$ almost surely, and a sequence of random atoms $\{Z_j\}_{j\geq 1}$ i.i.d. from P_0 independent of $\{p_j\}_{j\geq 1}$, where P_0 is assumed to be a diffuse probability measure on $(\mathbb{X}, \mathscr{X})$. A random distribution $\tilde{p}$ of the form in (1) is called a *species sampling model*. Further, a species sampling model is termed *proper* if and only if $\sum_{j\geq 1} p_j = 1$ almost surely, and most of the current Bayesian nonparametric literature focuses on the proper species sampling models. In this paper we discuss the case in which $\sum_{j\geq 1} p_j < 1$ with positive probability, and we show how non-proper models are particularly suited to take into account contaminated observations or more generally observations with frequency one. It is worth mentioning

that the diffuse component P_0 in (1) generates singleton blocks, which are called dust in the probabilistic literature on random partitions.

Among the very general class of species sampling models we recover special subclasses of priors, which have been duly investigated in the literature, e.g., homogeneous normalized random measures with independent increments (Regazzini et al., 2003) and Gibbs-type priors (Gnedin and Pitman, 2005; De Blasi et al., 2015). Here we focus on a contaminated version of Gibbs-type priors. For this reason, it is worth recalling that Gibbs-type random probability measures are typically characterized in terms of the exchangeable random partition (Pitman, 2006) induced by the data. More precisely, given a sample $X_{1:n} := (X_1, \ldots, X_n)$ from a species sampling model governed by a random probability measure $\tilde{p}$, the n observations are naturally partitioned into $K_n = k$ groups of distinct values, denoted here as $X_1^*, \ldots, X_k^*$, with corresponding frequencies $(N_{n,1}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)$. The exchangeable partition probability function (EPPF) corresponds to the probability of observing a specific partition of the data into clusters of distinct values, and it can be formalized as

$$\Pi_k^{(n)}(n_1, \ldots, n_k) := \int_{\mathbb{X}^k} \mathrm{E}\left[\prod_{j=1}^{k} \tilde{p}^{n_j}(\mathrm{d}x_j^*)\right]. \tag{2}$$

The EPPF is essential for computational purposes: indeed, it is the basic building block to derive suitable sampling schemes for posterior inference, also in the mixture model case (see, e.g., Section 6).

Gibbs-type priors are proper species sampling models $\tilde{p}$ characterized by means of their sequence of EPPFs $\{\Pi_k^{(n)} : n \geq 1, 1 \leq k \leq n\}$, which can be expressed in the following form

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} (1-\sigma)_{n_i - 1}, \tag{3}$$

for all $n \geq 1$, $k \leq n$ and positive integers $n_1, \ldots, n_k$ with $\sum_{i=1}^{k} n_i = n$, where $(a)_b = \Gamma(a+b)/\Gamma(a)$ in (3), for $a, b > 0$, denotes the Pochhammer symbol. The discount parameter $\sigma < 1$ and the non-negative weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ must satisfy the recurrence relation $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$ for all $k = 1, \ldots, n$, $n \geq 1$, with the proviso $V_{1,1} = 1$ and $V_{0,0} = 1$. The sequence of weights $V_{n,k}$s can be specified to recover prior processes commonly used in literature, such as the Dirichlet process (Ferguson, 1973), the Pitman-Yor process (Pitman and Yor, 1997), the normalized inverse Gaussian process (Lijoi et al., 2005) and the normalized generalized gamma process (see e.g. Lijoi et al., 2007b, and references therein). It is also apparent from (3) that the discount parameter σ affects the distribution of a random partition arising from a Gibbs-type prior: when σ grows, such a distribution favours partitions with few large clusters and a considerable number of small clusters. On the other side, when σ decreases, the EPPF (3) favours partitions with a large number of clusters with substantial sizes. However, if we observe a sample where the frequencies of the common species are properly modeled with a small value of σ, but the frequencies of the rarest species with a large value of σ (see, e.g., Section 5.2), a pure Gibbs-type prior as in (3)

does not have enough flexibility to fully describe the whole observed species. Building upon Gibbs-type priors, we now introduce a new family of prior processes which account for the possibility of contaminated observations.

Definition 1. *Let $\tilde{q}$ be a Gibbs-type prior, specified by the sequence of weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ and $\sigma < 1$. A contaminated Gibbs-type prior is a random probability measure on $(\mathbb{X}, \mathscr{X})$ defined as*

$$\tilde{p} = \beta\tilde{q} + (1-\beta)P_0, \qquad \beta \in (0,1), \tag{4}$$

where $Q_0(\,\cdot\,) = \mathrm{E}[\tilde{q}(\,\cdot\,)]$ is the base measure of $\tilde{q}$, and Q_0, P_0 are diffuse probability measures.

The prior $\tilde{p}$ in (4) is a convex linear combination of two components: an almost surely discrete component $\tilde{q}$ which generates the data, and a diffuse probability measure P_0 which accounts for contaminated observations. In the sequel we refer to P_0 as the *contaminant measure*. Sampling from $\tilde{p}$ can be interpreted as sampling from a population formed by two parts: the first one, representing a β fraction of the entire population, is composed by a countable number of species each appearing with positive probability. The second part ($1-\beta$ fraction) can be thought of as composed by a continuum of individuals each belonging to a different species. Therefore any time we sample from this second part a new species is obtained that cannot be re-observed. As stated above, for simplicity we shall call *contaminant* this second part and *contaminated* the relative observations. However, the diffuse part can be used more generally to account for any population which displays unique elements (see Section 7) and/or to model a high number of generic singletons in the observations. Finally, in Definition 1, the contaminant measure P_0 may be different from the base measure Q_0, thus $\tilde{p}$ in (4) may not be a species sampling model. We finally note that in species problems, specific choices of P_0 and Q_0 are irrelevant, since one is typically interested in the frequencies of the species rather than their labels (see e.g. Section 5.2). On the contrary, the additional flexibility introduced by choosing $P_0 \neq Q_0$ may be useful in model based clustering to deal with outliers (see e.g. Section 6).

We first derive the expectation and the covariance structure of a contaminated Gibbs-type prior in order to understand how the contaminant measure affects the distribution of $\tilde{p}$.

Proposition 1. *Let $\tilde{p}$ be a contaminated Gibbs-type prior as in Definition 1. Let $A, B \in \mathscr{X}$, then*

$$\begin{aligned} \mathrm{E}[\tilde{p}(A)] &= \beta Q_0(A) + (1-\beta)P_0(A), \\ \mathrm{cov}\big(\tilde{p}(A), \tilde{p}(B)\big) &= \beta^2(1-\sigma)\frac{V_{2,1}}{V_{1,1}}\big[Q_0(A\cap B) - Q_0(A)Q_0(B)\big] = \beta^2\mathrm{cov}\big(\tilde{q}(A), \tilde{q}(B)\big). \end{aligned}$$

As consequence of Proposition 1, one has $\mathrm{var}\big(\tilde{p}(A)\big) = \beta^2\mathrm{var}\big(\tilde{q}(A)\big)$, therefore the diffuse probability measure P_0 in (4) has the effect to shrink $\tilde{q}(A)$ towards its expected value. From Proposition 1, one can observe that when $\beta \to 1$ the covariance of the

contaminated model equals the covariance of the non-contaminated one, while as $\beta \to 0$, $\tilde{p}$ degenerated to P_0 and the covariance vanishes. Further, the correlation is invariant under contamination, with $\text{cor}\big(\tilde{p}(A), \tilde{p}(B)\big) = \text{cor}\big(\tilde{q}(A), \tilde{q}(B)\big)$. See Section A.1 of the Supplementary Material for a proof of Proposition 1.

3 Random partition, prediction and asymptotic properties

Having introduced all the modeling assumptions in Section 2, we now study the partition structure induced by a sample of observations from the random probability measure in (4), we further derive a closed form expression for predictive distributions and asymptotic properties for the number of clusters. We first focus on the random partition induced by a sequence of exchangeable observations governed by a contaminated Gibbs-type prior, deriving the EPPF.

Theorem 1. *Let $\tilde{p}$ be a contaminated Gibbs-type prior as in (4), with P_0 and Q_0 two diffuse probability measures on $(\mathbb{X}, \mathscr{X})$. Suppose that $X_i|\tilde{p} \overset{iid}{\sim} \tilde{p}$, as $i \geq 1$, then the probability that n observations $X_{1:n}$ are partitioned into $K_n = k$ clusters of distinct values $X_1^*, \ldots, X_k^*$ with corresponding frequencies $(N_{n,1}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)$ equals*

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1}, k-\bar{M}_{m_1}}\right] \beta^{n-m_1} \prod_{i=1}^{k} (1-\sigma)_{n_i-1} \tag{5}$$

where $\bar{M}_{m_1} \sim \text{Binom}(m_1, 1-\beta)$ and $m_1 = \#\{i : \ n_i = 1\}$ denotes the number of singletons (i.e. observations with frequency one) out of the sample of size n.

See Section A.2 of the Supplementary Material for a proof of Theorem 1. From the expected value in (5), it is apparent that the use of the contaminant measure P_0 in (4) acts on observations with frequency one and, as expected, they play a central role in the expression of the EPPF. In order to fix the terminology we call *singletons* the observations with frequency one, while the *structural singletons* are those values generated from the contaminant measure P_0, whose number equals the latent quantity $\bar{M}_{m_1}$. Note that the term *structural* refers to the fact that these values cannot be observed twice and this statistic could be of potential interest in certain applied problems, as it will be discussed in Section 7.

We now get a glimpse of the probabilistic implications of the random partition (5) induced by contaminated Gibbs-type priors as compared to pure Gibbs-type priors. In order to do this, we denote by $(n_1, \ldots, n_k)$ and $(n_1', \ldots, n_k')$ two distinct compositions having the same number of distinct values k and corresponding to two samples with the same size n; the probability ratio between the EPPFs corresponding to the two compositions will be denoted by $\mathrm{R}(n_1, \ldots n_k; n_1', \ldots n_k'; n, k) := \Pi_k^{(n)}(n_1, \ldots, n_k)/\Pi_k^{(n)}(n_1', \ldots, n_k')$. In Proposition 1 (Section A.3 of the Supplementary Material) we compare the probability ratio R when $\Pi_k^{(n)}$ is a Gibbs-type EPPF (3) and when it equals the EPPF of a contaminated Gibbs-type prior (5). Proposition 1 of the Supplementary Material clarifies

that if the two compositions have the same number of singletons, the ratio is the same for the contaminated and non-contaminated model. On the other side, if the number of singletons out of the composition $(n_1, \ldots, n_k)$ is bigger w.r.t. the number of singletons out of $(n'_1, \ldots, n'_k)$, the relative ratio increases in the contaminated model. Thus, in relative terms and given the number k of distinct values, the contaminated Gibbs model modifies the probabilities of compositions only when a different number of singletons is involved, favouring compositions with a higher number of these elements.

For computational convenience, we can equivalently describe the EPPF (5) introducing a set of suitable latent variables on an augmented probability space. Indeed, we can denote by $J_1, \ldots, J_n$ Bernoulli random variables, where the generic J_i indicates if the ith observation is generated from the contaminant measure P_0 ($J_i = 0$), or from the a.s. discrete component $\tilde{q}$ ($J_i = 1$). Thus, the introduction of latent elements $J_1, \ldots, J_n$ leads us to deal with the following augmented model

$$\begin{aligned} X_i|\tilde{p}, J_i &\overset{\text{iid}}{\sim} J_i\tilde{q} + (1 - J_i)P_0 \\ J_i &\overset{\text{ind}}{\sim} \text{Bern}(\beta), \end{aligned} \tag{6}$$

from which we may recover the marginal model by integrating (6) with respect to J_i. Furthermore, $J_i = 1$ in (6) if the corresponding observation X_i has been recorded at least twice in the sample: indeed if an observation X_i is generated from P_0, it does not appear again in the sample with probability 1. Thus, the non-degenerate J_is are those values referring to singletons out of the sample $X_{1:n}$. Without loss of generality we can assume that these singletons are the first m_1 observations $X_1, \ldots, X_{m_1}$. Based upon this augmentation, the random variable $\bar{M}_{m_1}$ in (5) equals $\sum_{i=1}^{m_1}(1 - J_i)$ which represents the number of structural singletons and it could be of potential interest in many application areas, as discussed in Section 7. It is worth mentioning that we can perform inference on $\bar{M}_{m_1}$ by introducing the latent variables $J_1, \ldots, J_n$ in the computational strategy, used to face posterior inference. The goal of inferring the value of $\bar{M}_{m_1}$ can be achieved by providing a point estimate of the subset of observations associated with the contaminant measure using a decisional strategy based on information loss function (Wade and Ghahramani, 2018; Rastelli and Friel, 2018). We now describe the predictive distribution of the next observation X_{n+1}, conditionally given $X_{1:n}$ and the latent variables $J_{1:m_1} = (J_1, \ldots, J_{m_1})$.

Proposition 2. *Let $\tilde{p}$ be a contaminated Gibbs-type prior as in (4), with P_0 and Q_0 two diffuse probability measures on $(\mathbb{X}, \mathscr{X})$. Assume that $X_i|\tilde{p} \overset{iid}{\sim} \tilde{p}$, as $i \geq 1$, and consider a sample $X_{1:n}$ which displays $K_n = k$ distinct values, denoted as $X^*_1, \ldots, X^*_k$, with respective frequencies $(N_{n,1}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)$, and the first m_1 values $X^*_1, \ldots, X^*_{m_1}$ are singletons. Then*

$$\begin{aligned} \mathrm{P}(X_{n+1} \in \mathrm{d}x^*|X_{1:n}, J_{1:m_1}) &= (1 - \beta)P_0(\mathrm{d}x^*) + \beta\frac{V_{n-\bar{M}_{m_1}+1,k-\bar{M}_{m_1}+1}}{V_{n-\bar{M}_{m_1},k-\bar{M}_{m_1}}}Q_0(\mathrm{d}x^*) \\ &+ \beta\frac{V_{n-\bar{M}_{m_1}+1,k-\bar{M}_{m_1}}}{V_{n-\bar{M}_{m_1},k-\bar{M}_{m_1}}}\left(\sum_{i=1}^{m_1} J_i(1 - \sigma)\delta_{X^*_i}(\mathrm{d}x^*) + \sum_{i=m_1+1}^{k}(n_i - \sigma)\delta_{X^*_i}(\mathrm{d}x^*)\right), \end{aligned} \tag{7}$$

where $\bar{M}_{m_1} = \sum_{i=1}^{m_1}(1 - J_i)$ represents the latent number of structural singletons.

From the sampling mechanism dictated by the predictive distribution (7), it is apparent that those values sampled from the contaminant measure P_0 cannot be observed twice; moreover, at each sampling step, the probability of sampling a contaminated observation equals $1-\beta$ and does not depend on n. Note also that the sample without the $\bar{M}_{m_1}$ structural singletons is characterized by the usual predictive mechanism of Gibbs-type priors. Finally, it is worth mentioning that the prediction rule has a nice interpretation in terms of a modified Chinese restaurant metaphor. Consider a restaurant with two rooms, one for social people and one for loners. The first customer arrives and chooses a table either in the social room with probability β or in the loners' room with probability $(1-\beta)$, the customer also chooses a dish which is shared by all the customers that will join the same table. The nth customer arrives and first selects either the social room with probability β or the loners' room with probability $(1-\beta)$. In the former case the customer can either sits at a new table or at an occupied table according to the traditional Chinese restaurant metaphor, while in the latter case the customer sits alone at a new table eating a new dish.

If we further assume that $P_0 = Q_0$, which corresponds to a proper species sampling model, we can derive an explicit form of the predictive distribution integrating over $J_{1:m_1}$ as shown in the following result.

Proposition 3. *Under the setting of Proposition 2, with $P_0 = Q_0$, we have*

$$\begin{aligned}\mathrm{P}(X_{n+1} \in \mathrm{d}x^* | X_{1:n}) = &\left((1-\beta) + \beta \frac{\mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1}+1,k-\bar{M}_{m_1}+1}\right]}{\mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1},k-\bar{M}_{m_1}}\right]} \right) P_0(\mathrm{d}x^*) \\ &+ \frac{1}{m_1}\sum_{i=1}^{m_1} \beta(1-\sigma) \frac{\mathrm{E}_{\bar{M}_{m_1}}\left[(m_1-\bar{M}_{m_1})V_{n-\bar{M}_{m_1}+1,k-\bar{M}_{m_1}}\right]}{\mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1},k-\bar{M}_{m_1}}\right]} \delta_{X_i^*}(\mathrm{d}x^*) \\ &+ \sum_{i=m_1+1}^{k} \beta(n_i-\sigma) \frac{\mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1}+1,k-\bar{M}_{m_1}}\right]}{\mathrm{E}_{\bar{M}_{m_1}}\left[V_{n-\bar{M}_{m_1},k-\bar{M}_{m_1}}\right]} \delta_{X_i^*}(\mathrm{d}x^*),\end{aligned} \tag{8}$$

where $\bar{M}_{m_1} \sim \mathrm{Binom}(m_1, 1-\beta)$.

We refer to Section A.5 of the Supplementary Material for a proof of Proposition 3. The predictive distribution (8) clearly shows that the probability that X_{n+1} does not belong to $\{X_1^*, \ldots, X_k^*\}$ depends on the initial sample through the sample size n, the number of distinct values k and the number of singletons m_1. This is a remarkable addition w.r.t. the Gibbs-type family, in which such a probability does not depend on m_1 (Bacallado et al., 2017). Moreover the probability that X_{n+1} equals a previously observed value X_i^*, with $i = 1, \ldots, k$, not only depends on n, n_i and k, as in the Gibbs-type framework, but also on m_1. As a consequence contaminated models allows to enrich the predictive structure of an exchangeable model, though the inclusion of the additional sampling information on the number of singletons out of the observable sample, but they maintain the analytic tractability that characterizes Gibbs-type priors. In Section A.6 we study the re-sampling mechanism induced by contaminated Gibbs-type prior in comparison with standard Gibbs-type priors. More precisely we show that the contaminant measure mainly acts on singletons by decreasing their re-sampling probabilities

w.r.t. observations with higher frequencies. On the other side, for observations with frequency larger than one, we preserve the same reinforcement as the discrete term of the model, and the parameter σ exhibits the same interpretation as in the Gibbs-type case.

We conclude this section with some considerations on distributional properties of the number of clusters with a given frequency in a sample of size n: this helps us to better understand the advantage of contaminated Gibbs-type priors. To fix the notation, we consider a sample $X_{1:n}$ from a contaminated Gibbs-type prior, and we denote by $M_{n,r}$ the random number of elements observed r times out of the sample. In the sequel, if V is a statistic depending on the sample $X_{1:n}$, we write $V(\beta)$ to make explicit the dependence on the parameter β of the contaminated prior (4). The following proposition clarifies the effect of the contaminant component with respect to the Gibbs-type model in terms of stochastic dominance and asymptotic properties.

Proposition 4. *If $\beta_1 < \beta_2$, then $K_n(\beta_1)$ (resp. $M_{n,1}(\beta_1)$) stochastically dominates $K_n(\beta_2)$ (resp. $M_{n,1}(\beta_2)$). Moreover, as $n \to +\infty$, we have*

$$\frac{K_n}{n} \overset{a.s.}{\to} 1-\beta, \quad \frac{M_{n,1}}{n} \overset{a.s.}{\to} 1-\beta \quad \textit{and} \quad \frac{M_{n,r}}{n^\sigma} \overset{a.s.}{\to} \frac{\sigma(1-\sigma)_{r-1}}{r!} S_\sigma \beta^\sigma,$$

where S_σ denotes the σ-diversity random variable (Pitman, 2006).

The first part of Proposition 4 is a result of first order stochastic dominance and it clarifies the effect of the contaminant measure in the model (4). As β decreases, the number of distinct values and the number of singletons out of $X_{1:n}$ increases. By noticing that the case $\beta = 1$ corresponds to a Gibbs-type prior, it is now apparent that our model has the advantage to increase (in mean) the number of distinct values and the number of singletons: the smaller beta, the higher $\mathrm{E}\big[K_n(\beta)\big]$ and $\mathrm{E}\big[M_{n,1}(\beta)\big]$. The second part of Proposition 4 tells us that the number of distinct values K_n and the number of unique values scale linearly with n: this is a remarkable difference with respect to Gibbs-type priors. Indeed, as $n \to +\infty$, for Gibbs-type priors both K_n and $M_{n,1}$ grows as n^σ (Pitman, 2006). This linear behavior may be a realistic assumption in some applications, such as the number of mutation in genome sequencing with a constant rate of mutations or the number of outliers (but further examples will be discussed in Section 7). Also, the asymptotic behavior of $M_{n,r}$ remains unchanged with respect to Gibbs-type priors, apart for the presence of the factor β^σ. This asymptotic behavior clarifies the role of the contaminant measure P_0 in (4), which produces an inflation of the number of singletons, and consequentially of the number of unique elements, but it is not acting on higher frequencies values.

4 The contaminated Pitman-Yor process case

We now focus on a noteworthy example: the contaminated Pitman-Yor process. We specialize all the results of Section 3 for this choice and we detail its peculiar features. The contaminated Pitman-Yor process can be recovered by selecting $\tilde{q}$ in (4) to be a

Pitman-Yor process. In such a case we recall that

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\vartheta + i\sigma)}{(\vartheta+1)_{n-1}}, \tag{9}$$

with $\sigma \in [0,1)$ and $\vartheta > -\sigma$. In particular, by setting $\sigma = 0$ we recover the Dirichlet process. We may find an explicit expression for the EPPF starting from (5) and by substituting the weights $V_{n,k}$s with the expression in (9). Thus, we get

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \prod_{i=1}^{k}(1-\sigma)_{n_i-1}\sum_{\bar{m}_1=0}^{m_1}\binom{m_1}{\bar{m}_1}\beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1}\frac{\sigma^{k-\bar{m}_1}(\vartheta/\sigma)_{k-\bar{m}_1}}{(\vartheta)_{n-\bar{m}_1}}. \tag{10}$$

See Section B.1 of the Supplementary Material for details. The expression of the EPPF (10) plays a central role to carry out posterior inference in applications, indeed all the algorithms we have developed (see Section D of the Supplementary Material) are based on this expression.

We can further derive an explicit form of the predictive distribution (7) for the contaminated Pitman-Yor model, conditionally on the latent variables:

$$\begin{aligned}\mathrm{P}(X_{n+1} \in \mathrm{d}x | X_{1:n}, J_{1:m_1}) &= (1-\beta)P_0(\mathrm{d}x) + \beta\frac{\vartheta + (k-\bar{M}_{m_1})\sigma}{\vartheta + n - \bar{M}_{m_1}}Q_0(\mathrm{d}x) \\ &+ \sum_{i=1}^{m_1} J_i\beta\frac{1-\sigma}{\vartheta+n-\bar{M}_{m_1}}\delta_{X_i^*}(\mathrm{d}x) + \sum_{i=m_1+1}^{k}\beta\frac{n_i-\sigma}{\vartheta+n-\bar{M}_{m_1}}\delta_{X_i^*}(\mathrm{d}x).\end{aligned} \tag{11}$$

An important appealing property of the predictive distribution, when the latent variables are integrated out, is that the probability of sampling a new value has monotone behavior as a function of the number of distinct values m_1, which results in a richer predictive structure w.r.t. the Pitman-Yor case, where m_1 does not appear in the probability of sampling a new value. See Section B.2 of the Supplementary Material for a detailed proof. We only mention that the dependence on m_1 is always increasing in the Dirichlet process case ($\sigma = 0$), whereas it is always decreasing in the stable process one ($\vartheta = 0$).

4.1 Pólya urn representation

From the predictive distribution (11) we can describe the sampling structure of a contaminated Pitman-Yor process in terms of a generalization of the urn scheme by Zabell (1997), defining a *strip-and-solid generalized Pólya urn*. To do this, we now assume that the prior distribution for the parameter β is a beta with parameters ϑ and α. Thus, it is easy to check that the distribution of β, conditionally on $X_{1:n}, J_{1:m_1}$ is again a beta with parameters $(\vartheta + n - \bar{M}_{m_1}, \alpha + \bar{M}_{m_1})$, as one can realize from the augmented version of the EPPF (10) with the inclusion of the latent element $\bar{M}_{m_1}$. By integrating

(11) with respect to the conditional distribution of β, we obtain

$$\begin{aligned} \mathrm{P}(X_{n+1} \in \mathrm{d}x | X_{1:n}, J_{1:m_1}) = & \frac{\bar{M}_{m_1} + \alpha}{\alpha + \vartheta + n} P_0(\mathrm{d}x) + \frac{\vartheta + (k - \bar{M}_{m_1})\sigma}{\alpha + \vartheta + n} Q_0(\mathrm{d}x) \\ & + \sum_{i=1}^{m_1} J_i \frac{1-\sigma}{\vartheta + \alpha + n} \delta_{X_i^*}(\mathrm{d}x) + \sum_{i=m_1+1}^{k} \frac{n_i - \sigma}{\vartheta + \alpha + n} \delta_{X_i^*}(\mathrm{d}x). \end{aligned} \tag{12}$$

The predictive distribution (12) can be now described through a Pólya-Eggenberger urn scheme (Eggenberger and Pólya, 1923). The main difference from usual urn schemes is that here we assume the urn is composed by two types of balls: strip and solid balls, where strip balls correspond to elements associated with the contaminant measure while solid balls can be interpreted as elements associated with the discrete term of the model. Initially the urn is composed by a weight α of strip colored balls and a weight ϑ of black solid balls. We want to sample an exchangeable sequence from the urn in such a way that the updating rule is (12), and the balls are sampled proportionally to their weight. At the first sampling step, if a strip colored ball is drawn from the urn, then we return the ball in the urn with an additional strip colored ball of a new color. On the other side if we draw a black solid ball, then we return a black ball in the urn with an additional weight σ and a solid ball of a new color with weight $1 - \sigma$. At the generic ith step, one can sample a strip ball of an arbitrary color, a black solid ball or a colored solid ball. Thus, the updating mechanism of the urn works as follows: i) if we sample a strip ball, we return the strip ball in the urn with another strip ball of a new color having unitary weight; ii) if we sample a black solid ball, we return the solid ball in the urn with a new black solid ball of weight σ and a solid ball of a new color having weight $1 - \sigma$; iii) if we sample a colored solid ball, we return the ball in the urn with an additional new solid ball of the same color having weight 1.
We finally underline that, by introducing a suitable sequence of random variables $J_1, J_2, \ldots$, where the generic $J_i = 1$ if the ith sampled ball is solid, and $J_i = 0$ otherwise, as in the standard theory of Pólya urn schemes we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} J_i = Z \qquad \text{with} \qquad Z \sim \mathrm{Beta}(\vartheta, \alpha), \tag{13}$$

where the limit has to be intended in the almost sure sense. Note that if we initialize the urn without strip balls, we recover the urn of Zabell (1997), and the limiting distribution in (13) degenerates to a point mass at 1. On the other side, if we initialize the urn without solid ball, the urn is producing a sequence of strip balls of different colors, and the limiting distribution in (13) degenerates to a point mass 0.

4.2 Probabilistic investigation of the random partition

We conclude this section with a probabilistic investigation of the random partition induced by a contaminated Pitman-Yor process, both a priori and a posteriori. All the details and additional formulas are deferred to the Supplementary Material. First, we evaluate the expected value of $M_{n,r}$ and K_n. In particular one obtains:

$$\mathrm{E}[M_{n,1}] = n(1-\beta) + n\beta \mathrm{E}\big[(\beta B_1 + (1-\beta))^{n-1}\big],$$

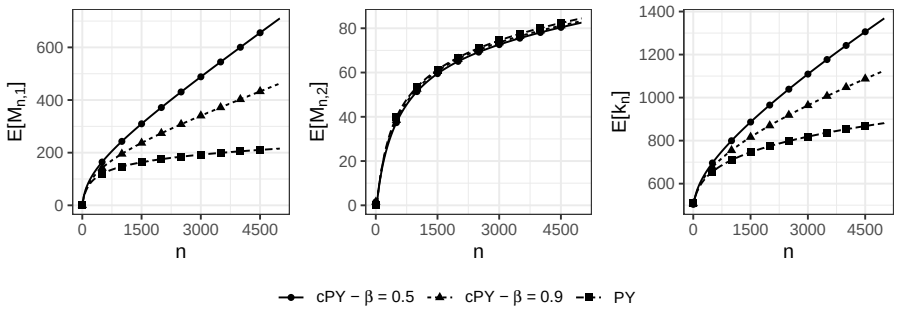

Figure 1: Curves of $\mathrm{E}[K_n]$, $\mathrm{E}[M_{n,1}]$ and $\mathrm{E}[M_{n,2}]$ as n increases, for the contaminated Pitman-Yor model with $\beta = 0.9$ (filled line with circles), the contaminated Pitman-Yor model with $\beta = 0.95$ (dashed line with triangles), and the Pitman-Yor model (dashed line with squares). The parameters are $\vartheta = 100$ and $\sigma = 0.2$.

$$\mathrm{E}[M_{n,r}] = \frac{(1-\sigma)_{r-1}}{(\vartheta+1)_{r-1}}\binom{n}{r}\beta^r \mathrm{E}\big[(B_r\beta + 1 - \beta)^{n-r}\big], \quad \text{if } r \geq 2,$$
$$\mathrm{E}[K_n] = \frac{\vartheta}{\sigma}\mathrm{E}\big[(B_1\beta + 1 - \beta)^n\big] + \frac{n\beta}{\sigma}\mathrm{E}\big[B_1(B_1\beta + 1 - \beta)^{n-1}\big] - \frac{\vartheta}{\sigma} + n(1-\beta),$$

where B_r is a Beta random variable with parameters $(\vartheta + \sigma, r - \sigma)$, as $r \geq 1$. See Section B.3 of the Supplementary Material for further details. In Figure 1, we compare the behavior of the expected values of the statistics K_n, $M_{n,1}$ and $M_{n,2}$ in the Pitman-Yor case with the same quantities for the contaminated model. It is apparent that for the latter model the two curves of $\mathrm{E}[K_n]$ and $\mathrm{E}[M_{n,1}]$ grow faster as function of n, with respect to the Pitman-Yor model. We finally underline that, resorting to the results by Favaro et al. (2013), one may face prediction for a large number of statistics arising in species sampling models. Indeed, in Section B.3 of the Supplementary Material, we evaluated the posterior expected value of the following meaningful statistics: i) $K_m^{(n)}$, which denotes the number of distinct observations out a future sample $X_{n+1:n+m} = (X_{n+1}, \ldots, X_{n+m})$ not yet observed in the initial sample $X_{1:n}$; ii) $N_{m,r}^{(n)}$, which denotes the number of new and distinct observations recorded with frequency r out of the additional sample $X_{n+1:n+m}$, hitherto unobserved in the initial sample of size n. All these posterior expected values display closed form expressions (see Section B.3 for details), which depend not only on n and k, as for all the class of Gibbs-type priors (as a consequence of the characterizations by Bacallado et al. (2017)), but also on the number of singletons m_1. Thus it is now apparent how the contaminated model improves the flexibility of the Pitman-Yor process with the inclusion of the additional sampling information on m_1. We refer to Section C of the Supplementary Material for an illustration of the predictive properties of the contaminated Pitman-Yor process, compared to a Pitman-Yor process without contamination.

5 Illustrations

In this section we illustrate the use of the contaminated Pitman-Yor process through a set of simulation studies and on a real dataset, exhibiting a high number of observations with frequency one. Posterior inference for the urn scheme induced by a contaminated Pitman-Yor process can be done exploiting the representation of the EPPF provided in (10), as described in Section D of the Supplementary Material.

5.1 Simulation studies

We investigate the use of contaminated models in discrete scenarios, where the observations are simulated from the urn scheme induced by the Dirichlet process (DP), the Pitman-Yor (PY) Process, the contaminated Pitman-Yor (cPY) process of Section 4 and the contaminated Zipf (cZIPF) distribution with ζ species (see Section E of the Supplementary Material). The true parameters for each model are selected as follows: i) $\vartheta = 100$ for the Dirichlet process; ii) $\vartheta = 100$ and $\sigma = 0.2$ for the Pitman-Yor process; iii) $\vartheta = 100$, $\sigma = 0.6$ and $\beta = 0.9$ for the contaminated Pitman-Yor process; iv) $\zeta = 1000$, $\alpha = 0.25$ and $\beta = 0.9$ for the contaminated Zipf case. For each modelling choice, we have generated a sample of size $n = 50\,000$. We have then estimated the model parameters using three different prior specifications: the urn scheme induced by a contaminated Pitman-Yor, by a Pitman-Yor and by a Dirichlet process. The posterior estimates are reported in Table 1, averaged over 100 replications. We can appreciate how the Dirichlet model is an adequate choice when the reinforcement of the data generating process is also driven by a Dirichlet process. However, when the true generating process is the urn scheme associated to a Pitman-Yor, a contaminated Pitman-Yor or a contaminated Zipf model, the Dirichlet model is not flexible enough to capture the behavior of the data. When the data generating process is not contaminated, the Pitman-Yor model provides good posterior estimates of the parameters, but when we inflate the number of singletons, the Pitman-Yor model overestimates the discount parameter σ. By recalling that the ordered p_js in (1) are asymptotically equivalent to $Zj^{-1/\sigma}$ for a positive random variable Z as $j \to +\infty$ (Mano, 2018, Chapter 2), the overestimation of σ is probably due to fact that the Pitman-Yor tries to capture the singletons by making heavier the tail of the p_js. Finally, we can appreciate how the contaminated Pitman-Yor model provides reasonable posterior estimates in all the scenarios where it is covering the data generating process, but it is also providing good estimates of β in the contaminated Zipf case.

Under the same scenarios as above, we perform predictive inference, in order to do this we consider a sample $(X_1, \ldots, X_n)$ of size $n = 50\,000$ and we try to estimate different statistics which depend on an additional unobserved sample of size $m = 25\,000$. For a statistic $V_m^{(n)}$ depending on both the two samples, we define the prediction error as the square root of the mean square error, i.e.

$$\sqrt{\mathrm{E}[(V_m^{(n)} - \hat{V}_m^{(n)})^2 | X_1, \ldots, X_n]},$$

where $\hat{V}_m^{(n)}$ represents the Bayes estimator under a squared loss function, i.e., the posterior mean. Such a prediction error, in practice, may be estimated by a Monte Carlo

Data	Model	Data parameters	Posterior estimates $\hat{\vartheta}$	$\hat{\sigma}$	$\hat{\beta}$
DP	cPY	$(\vartheta = 100)$	97.33	0.00	1.0
DP	PY	$(\vartheta = 100)$	98.06	0.00	-
DP	DP	$(\vartheta = 100)$	99.82	-	-
PY	cPY	$(\vartheta = 100, \sigma = 0.2)$	107.31	0.18	1.0
PY	PY	$(\vartheta = 100, \sigma = 0.2)$	101.57	0.20	-
PY	DP	$(\vartheta = 100, \sigma = 0.2)$	240.48	-	-
cPY	cPY	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	100.60	0.60	0.9
cPY	PY	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	28.35	0.77	-
cPY	DP	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	4510.80	-	-
cZIPF	cPY	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	233.64	0.00	0.9
cZIPF	PY	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	6.95	0.75	-
cZIPF	DP	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	1766.37	-	-

Table 1: Summaries of the simulation study. First column: data generating process. Second column: model used to analyze the data. Third column: parameters of the data generating process. Fourth to sixth columns: posterior mean of the main parameters of the models. The results are averaged over 100 replications.

integration over replicates of additional samples generated from the true model. For each replicate, we obtained an estimate of the true value $V_m^{(n)}$ by sampling 1 000 distinct additional samples of size m. Table 2 shows the estimated prediction errors for three predictive quantities calculated over the replicates, namely: the number of new distinct observations with frequency one out of the observed sample $N_{m,1}^{(n)}$; the number of new distinct observations with frequency two $N_{m,2}^{(n)}$; the number of new distinct observations $K_m^{(n)}$ detected in an additional sample of size m. Note that the Bayes estimators of these statistics, particularly relevant in species sampling problems, are available in closed form (see Section B of the Supplementary Material). From the results reported in Table 2, it is apparent that the contaminated Pitman-Yor model produces smaller or comparable prediction errors, with respect to the ones of the urn scheme induced by a Dirichlet process for all the scenarios considered in the study. Moreover the performance of the contaminated Pitman-Yor model is always better or similar to the one of the Pitman-Yor model. In particular, when the data generating process has a contaminant component, predictive inference based on the contaminated Pitman-Yor model outperforms the competitors, for both $N_{m,1}^{(n)}$ and $K_m^{(n)}$.

5.2 The North America Ranidae dataset

As a real data example, we consider a set of species detection data, from the Global Biodiversity Information Facility project (GBIF.org, 2021), where the inclusion of additional information in the predictive structure of the model plays a crucial role. The project is an extensive database consisting in record of species found across the world, where for each individual is reported the taxonomy, location and possibly other relevant

Data	Model	Data parameters	Prediction errors $N_{m,1}^{(n)}$	$N_{m,2}^{(n)}$	$K_m^{(n)}$
DP	cPY	$(\vartheta = 100)$	61.70	44.49	508.78
DP	PY	$(\vartheta = 100)$	66.08	44.32	513.33
DP	DP	$(\vartheta = 100)$	66.81	44.42	514.16
PY	cPY	$(\vartheta = 100, \sigma = 0.2)$	169.35	103.83	857.85
PY	PY	$(\vartheta = 100, \sigma = 0.2)$	176.17	103.01	863.74
PY	DP	$(\vartheta = 100, \sigma = 0.2)$	225.06	107.04	917.19
cPY	cPY	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	197.91	309.45	1437.38
cPY	PY	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	344.55	303.39	1541.89
cPY	DP	$(\vartheta = 100, \sigma = 0.6, \beta = 0.9)$	2922.29	271.95	4014.26
cZIPF	cPY	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	208.98	12.88	678.00
cZIPF	PY	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	219.38	85.11	1019.58
cZIPF	DP	$(\zeta = 1000, \alpha = 0.25, \beta = 0.9)$	1674.75	93.67	2453.99

Table 2: Prediction errors of the simulation study. First column: data generating process. Second column: model used to analyze the data. Third column: parameters of the data generating process. Fourth to sixth columns: prediction errors for $N_{m,1}^{(n)}$, $N_{m,2}^{(n)}$ and $K_m^{(n)}$. The results are averaged over 100 replications.

information. Our sample consists of $n = 131\,204$ observations belonging to $k = 619$ distinct species of the Ranidae family observed in North America, and identified by their scientific name. Among the $k = 619$ species, $m_1 = 296$ species were observed only once in the sample, creating a possible inflation of the number of elements with frequency one. Such inflation might be caused by miss reported scientific name of the observed animals. We aim to investigate the benefit of including a contaminant measure in the prior model specification by comparing posterior inference when we use the urn scheme induced by a contaminated and a pure Pitman-Yor process. We choose non-informative prior specifications for the parameters, namely $\vartheta \sim \text{Gamma}(2, 0.02)$ and $\sigma, \beta \sim \text{Unif}(0, 1)$. We carried out posterior inference by exploiting Algorithm 1 described in Section D of the Supplementary Material, and similarly for the standard model. Refer to Section G for diagnostic summaries and algorithmic details.

Figure 2 clarifies how the presence of a large number of species observed only once leverages the estimation of the parameters in the Pitman-Yor model, while the use of a contamination component helps to obtain a much more suitable modeling of the data. Indeed, in the latter case, some of the observations with frequency 1 are assigned to the diffuse component. As consequence of the excessive number of singletons, the estimated posterior distribution of the frequency spectrum is remarkably different on small values of the support, as emphasized in the left panel of Figure 2. Furthermore, both the probability of sampling a new species and the posterior distribution of σ in the Pitman-Yor case are translated with respect to the contaminated model. Additional posteriors summaries are reported in the Supplementary Material.

We finally consider the task of predicting the number of new species and the number of new species observed with a given frequency in a follow-up sample, given an initial training sample. We have retained the 80% of the n data for purposes of training, and

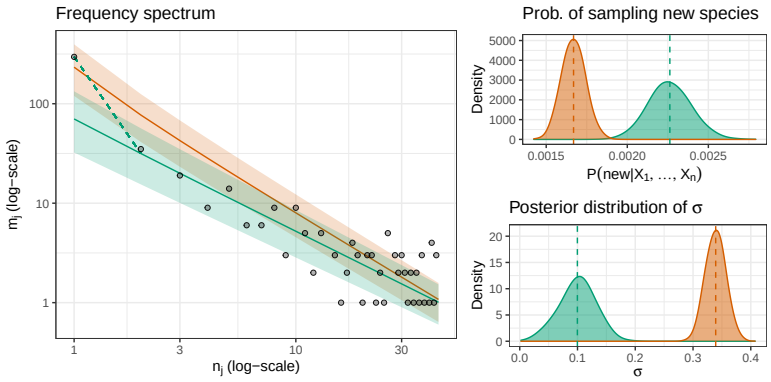

Figure 2: Posterior summaries for contaminated Pitman-Yor model (green) and Pitman-Yor model (orange). Left panel: frequency spectrum of the first non-empty frequencies, with the posterior expectation of the discrete part of the two models, shaded bands represent 90% posterior credible intervals; the dashed line corresponds to the inflation of the diffuse component. Right-top panel: posterior probability of sampling a new species. Right-bottom: posterior distribution of σ.

the remaining m data points are used as a test set. We focused on estimation of: i) $K_m^{(n-m)}$, the distinct number of new species in a follow-up sample hitherto unobserved in the initial training sample of size $n-m$; ii) $N_{m,1}^{(n-m)}$, the number of new species observed with frequency one in an additional sample of size m, hitherto unobserved in the training dataset. The posterior expectations of $K_m^{(n-m)}$ and $N_{m,1}^{(n-m)}$ are evaluated using the corresponding closed-form expressions, reported in Equations (S25) and (S20) respectively, for the contaminated Pitman-Yor model. The predicted values are compared with the true ones, obtained by extrapolating to the remaining m data. We repeated the experiment 1 000 times in order to asses variability. Figure 3 shows the cross-validated distributions of the posterior expectation of $K_m^{(n-m)}$ and $N_{m,1}^{(n-m)}$ when we exploit the contaminated Pitman-Yor model in comparison with the predicted values obtained by using the Pitman-Yor model. The average true value is represented with a dashed black line. From Figure 3, it is apparent how the contaminant measure in the model specification can be crucial also for its predictive properties. Indeed the cross-validated distributions of the posterior expectations of $N_{m,1}^{(n-m)}$ and $K_m^{(n-m)}$ for the contaminated model, conditionally on an observed sample, shrink to the corresponding average of the observed values (black dashed line), while the distributions for the model without a contaminant term provides a systematic error in prediction. Such behavior is also confirmed by the prediction errors, as shown in Table 3. From Table 3 we observe that the inclusion of a contaminant measure in the model specification is helpful to decrease the prediction errors for both $N_{m,1}^{(n-m)}$ and $K_m^{(n-m)}$. Note that the posterior expected values of $N_{m,1}^{(n-m)}$ and $K_m^{(n-m)}$ under the contaminated Pitman-Yor model, found in Section B of the Supplementary Material, depend on the additional sampling information on the number of distinct values with frequency one out of the

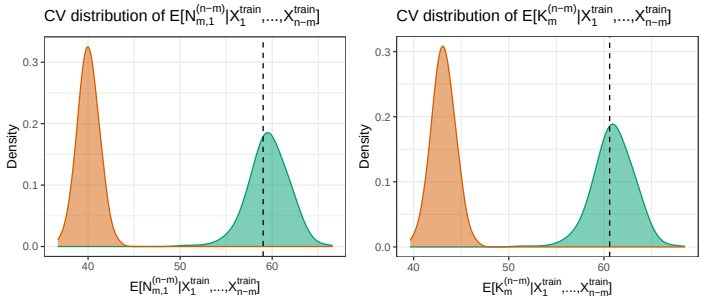

Figure 3: Cross-validated distributions of the posterior expectation of $K_m^{(n-m)}$ and $N_{m,1}^{(n-m)}$, for the contaminated Pitman-Yor model (green) and the Pitman-Yor model (orange). The black dashed lines correspond to the true values.

observed sample. This is a remarkable difference with respect to the posterior expectations of the same quantities under a Pitman-Yor prior, which do not depend on this sampling information. Thus our results underline how the use of this additional sampling information is crucial to decrease the prediction errors. We refer to Section G.1 of the Supplementary Material for further details on the cross-validation study. Finally we stress that this example highlights the strong lack of robustness of the pure Pitman-Yor model. Indeed, drastically erroneous inferential conclusions are caused by relatively few singletons ($m_1 = 296$) compared to the total number of observations ($n = 131\,204$).

	Prediction errors	
Model	$N_{m,1}^{(n-m)}$	$K_m^{(n-m)}$
cPY	8.61	8.69
PY	20.58	19.25

Table 3: Prediction errors of $N_{m,1}^{(n-m)}$ and $K_m^{(n-m)}$ under the contaminated Pitman-Yor (cPY) and Pitman-Yor model (PY). The results are averaged over 1 000 cross-validated samples.

6 Mixtures of contaminated Gibbs-type priors

Contaminated Gibbs-type priors are not restricted only to species sampling models, but they can be also convoluted with a kernel function to build contaminated mixture models. Such a modelling strategy can be exploited to face simultaneously model based clustering and density estimation in presence of outliers, by inflating the number of structural singletons through the contaminant measure. As we will show in Section 6.2, the inclusion of a contaminant measure has the effect to smooth the density estimates, rather than producing spikes in correspondence of outliers. Since there is no gold standard definition of outliers in presence of clusters, here the term outlier generically refers to an observation which markedly deviates from the rest of the data/clusters. As a

consequence we recognize an observation as an outlier if it belongs to a cluster with frequency one.

Mixture models in Bayesian nonparametrics were early introduced by Lo (1984) for the Dirichlet process mixtures of univariate Gaussian distribution case, and later extended in several directions. Remarkable examples that have been studied with different kernel functions are the Dirichlet process mixture of multivariate Gaussian distributions (Müller et al., 1996) for multivariate continuous data, the Dirichlet process mixture of Poisson distributions (Krnjajić et al., 2008) for counting data, and the Dirichlet process mixture of Gaussian processes (Bigelow and Dunson, 2009) for functional data. The seminal work of Lo (1984) has been also extended by considering mixing distributions different from the Dirichlet process, such as the Pitman-Yor process (Ishwaran and James, 2001), the Normalized Generalized Gamma process (Lijoi et al., 2007a), the Normalized Inverse-Gaussian process (Lijoi et al., 2005), and more in general Gibbs-type priors (e.g. De Blasi et al., 2015). See also Frühwirth-Schnatter et al. (2019) for an extensive review on mixture models. The standard general framework can be described as follows. It is assumed that observations are $\mathbb{Y}$-valued random elements generated from a random density $\tilde{f}(y) = \int_{\Xi} \mathcal{K}(y;\xi)p(\mathrm{d}\xi)$, where $\mathcal{K}(y;\xi) : \mathbb{Y} \times \Xi \to \mathbb{R}^+$ is a kernel and $\mathbb{Y}, \Xi$ are general Polish spaces. Furthermore, the mixing measure p is usually assumed to belong to a specific class of discrete random probability measures. If one denotes by $\xi_1, \ldots, \xi_n$ the latent variables corresponding to a sample of size n from p, the standard mixture model may be expressed in the following hierarchical form

$$Y_i|\xi_i \overset{\text{ind}}{\sim} \mathcal{K}(\,\cdot\,;\xi_i), \quad \xi_i|p \overset{\text{iid}}{\sim} p \tag{14}$$

for any $i = 1, \ldots, n$. We remark that the model (14) describes a general formulation of a mixture model. Nowadays it is an established opinion in the applied statistics framework that mixture models are flexible tools for density estimation and model-based clustering analysis (Frühwirth-Schnatter et al., 2019).

Here we propose to extend such framework by choosing as mixing measure p the contaminated Gibbs-type prior $\tilde{p}$. Thanks to the definition of $\tilde{p}$, which is a linear convex combination of two elements, we can decompose the mixture in two terms, a first term corresponding to the discrete part of $\tilde{p}$ and a second term which corresponds to the diffuse component,

$$\tilde{f}(y) = \beta \sum_{j=1}^{\infty} p_j \mathcal{K}(y;Z_j) + (1-\beta) \int_{\Xi} \mathcal{K}(y;\xi) P_0(\mathrm{d}\xi) \tag{15}$$

where the last equality holds in force of the almost sure discreteness of $\tilde{q} = \sum_{j\geq 1} p_j \delta_{Z_j}$. The first term on the r.h.s. of equation (15) describes the standard random mixture components of the model, while the second term corresponds to a different probabilistic mechanism contaminating the mixture, typically over-disperse, which is particularly suited to model outliers in the data. Outlier detection is a crucial problem in Statistics and similar convex constructions are available also in classical setting, see, e.g., Bouveyron et al. (2019) for an account. In the Bayesian framework, some contributions are available and rely on the use of traditional Dirichlet process. Quintana and Iglesias (2003); Quintana (2006) focus on product partition models, and they develop a

decision-theoretic approach that allows selecting a partition with the purpose of outlier detection in regression problems. Shotwell and Slate (2011) identify an outlier detection criterion based on the Bayes factor, where they compare a partition containing outliers against a partition with fewer or no outliers. As a remarkable addition with respect to the current literature, our prior process contemplates a specific component in the model, i.e. the contaminant measure P_0, which has two effects in the model specification: i) it accounts for the presence of outliers, which could follow a different generating process with respect to the other observations; ii) in presence of these outliers, the contaminant component allows to inflate the tails of the mixture model with respect to a model with only local components, thus producing smoother density estimates. In particular, the contaminated Pitman-Yor model accommodates for outliers by increasing the weight of the over-disperse contaminant component, instead of generating specific local and peaked components for them, which is a typical behavior observed for Pitman-Yor mixture models.

We also point out that the choice of P_0 and Q_0 is a fundamental aspect in the mixture model case. By allowing $Q_0 \neq P_0$, further flexibility is added to the model. In this case the contaminant measure P_0 can be determined according to the specific information related to the problem under study. If our prior opinion on contaminated observations is restricted to a particular subset of the sample space $\mathbb{Y}$, we can shrink P_0 to induce a relevant mass on it, thereby forcing the model to expect more structural singletons on a specific part of the support. As an example, we could restrict the support of P_0 on the basis of an observed sample. However, such a specification may turn out detrimental, as the truncated support of P_0 might be unable to account for contaminant observations in an additional sample. On the other hand, in absence of specific information, we may aim to model possible contaminated observations over the entire sample space. In this case, we can specify P_0 to produce an over-dispersed predictive distribution (possibly with heavy tails) with respect to the predictive distribution induced by Q_0. Furthermore, the measure Q_0 can be elicitated by resorting to an empirical Bayes initialization. See, e.g., Section 6.2. Finally, any prior knowledge on the expected rate of contaminant observations, and more generally on their distribution, can be used to specify the prior distribution on β, otherwise it can be elicitated in a vague manner.

6.1 Posterior inference for mixtures of contaminated Pitman-Yor process

Thanks to the predictive distribution (12), posterior inference in mixture models can be performed by exploiting an augmented marginal sampling strategy, in the spirit of Escobar (1988); Escobar and West (1995). We denote by $S_1, \dots, S_n$ the variables describing the latent group allocations in the mixture, with $S_i = j$ if the ith observation belongs to the jth group of the mixture, with the proviso $S_i = 0$ if the ith observation comes from the diffuse component. Further, we denote by $A_{1:n} := (A_1, \dots, A_n)$ a generic vector of n elements, where $A_{(i)}$ is denoting the vector $A_{1:n}$ with the ith element removed. Here we provide the algorithm to face posterior inference by sampling R realizations from an MCMC scheme. To clarify the notation used in Algorithm 1, the quantities reported with a tilde refer to observations allocated to the discrete component of the model.

Algorithm 1 Sampling scheme for contaminated Pitman-Yor mixture model.

[0] At time $r = 0$, set the initial values $S_{1:n}^{(0)}$, $\xi_{1:n}^{(0)}$, $\sigma^{(0)}$, $\vartheta^{(0)}$, $\beta^{(0)}$

for $r = 1, \ldots, R$ **do**

for $i = 1, \ldots, n$ **do**

[1.1] Update the cluster allocation of the ith element at time r, generating $S_i^{(r)}$ from:

$$\mathrm{P}(S_i = j \mid Y_i, S_{(i)}, \xi_{(i)}) \propto \begin{cases} (1-\beta) \int_{\Xi} \mathcal{K}(Y_i; \xi) P_0(\mathrm{d}\xi) & \text{if } j = 0 \\ \beta \frac{\tilde{n}_j - \sigma}{\vartheta + n - \bar{m}_{1(i)} - 1} \mathcal{K}(Y_i; \tilde{\xi}_j) & \text{if } j = 1, \ldots, \tilde{k}_{(i)} \\ \beta \frac{\vartheta + \tilde{k}_{(i)} \sigma}{\vartheta + n - \bar{m}_{1(i)} - 1} \int_{\Xi} \mathcal{K}(Y_i; \xi) Q_0(\mathrm{d}\xi) & \text{if } j = \tilde{k}_{(i)} + 1 \end{cases}$$

where $\tilde{\xi}_1, \ldots, \tilde{\xi}_{\tilde{k}_{(i)}}$ denote the unique values in $\{\xi_{\ell,(i)} : S_{\ell,(i)} \neq 0\}_{\ell=1}^{n-1}$ with frequencies $\tilde{n}_1, \ldots, \tilde{n}_{\tilde{k}_{(i)}}$, where $\tilde{n}_j = \sum_{\ell \neq i} \mathbb{1}_{\{j\}}(S_\ell)$, $\tilde{k}_{(i)}$ represents the number of distinct unique latent parameters associated with the discrete component of the model, and $\bar{m}_{1(i)} = \sum_{\ell \neq i} \mathbb{1}_{\{0\}}(S_\ell)$

Let $\tilde{n}_1, \ldots, \tilde{n}_{\tilde{k}}$ denote the frequencies of the unique values $\tilde{\xi}_1, \ldots, \tilde{\xi}_{\tilde{k}}$ out of $\{\xi_i : S_i \neq 0\}_{i=1}^{n}$, with $\tilde{k}$ denoting the number of distinct unique latent parameters of the discrete component.

---

for $j = 1, \ldots, \tilde{k}$ **do**

[2.1] Generate $\tilde{\xi}_j^{(r)}$ at time r from the full conditional distribution

$$\mathcal{L}(\tilde{\xi}_j \mid Y_{1:n}, S_{1:n}) \propto Q_0(\tilde{\xi}_j) \prod_{\{i: S_i = j\}} \mathcal{K}(Y_i; \tilde{\xi}_j);$$

[3] Generate the updated values of the parameters $\sigma^{(r)}$, $\theta^{(r)}$ and $\beta^{(r)}$ according to steps [1-3] of Algorithm 1 in Section D of the Supplementary Material.

In Algorithm 1, the acceleration step [2.1] is not mandatory, but it improves the mixing performances of the algorithm. The integral in the predictive distribution of step [1] can be easily solved for suitable choices of the kernel function and the measures P_0 and Q_0, leading to a closed form expression for the predictive distribution. Otherwise such integral can be approximated via Monte Carlo methods, in the spirit of Algorithm 8 in Neal (2000). Furthermore, we have tested the capability of the contaminated Pitman-Yor mixture model with a sampling strategy as in Algorithm 1 to identify outliers by performing an extensive simulation study in Section F of the Supplementary Material. From these numerical experiments, one can observe that the pure Pitman-Yor mixture model fails to detect contaminant observations, while the contaminated Pitman-Yor mixture model is able to identify such observations. Moreover, when the data are not contaminated, we have empirically shown that the contaminated Pitman-Yor mixture model and the pure Pitman-Yor mixture model with $P_0 \neq Q_0$ have similar performances,

consistently with the specification of P_0 and Q_0 that we used in the simulation study. We conclude the present section with an application of the mixture model when the mixing measure is a contaminated Pitman-Yor process.

6.2 Analysis of the NGC 2419 data

We consider a set of data composed by $n = 139$ stars, possibly belonging to the globular cluster NG 2419 and sharing the same galactic center. The data were early introduced and studied by Ibata et al. (2011) and early analyzed by Arbel et al. (2021). For each observation we have measurements of $d = 4$ different variables: the two-dimensional projection on the plane of the position of the star (D_1, D_2), the line of sight velocity V on a logarithmic scale, and the metallicity of the star $[Fe/H]$ on a logarithmic scale, which is a measure of the abundance of iron relative to hydrogen. We denote by $\boldsymbol{Y}_i = (D_{1,i}, D_{2,i}, V_i, [Fe/H]_i)$ the ith observed vector. A crucial problem for the astronomical community is to identify which stars belong to the globular cluster, and which stars are contaminants (or outliers), to properly study the dynamic of a group of stars. To this aims, we consider a contaminated Pitman-Yor mixture model, specified with a multivariate Gaussian kernel function $\mathcal{K}(\cdot\,; (\boldsymbol{\mu}, \boldsymbol{\Sigma}))$, with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We further assume a base measure conjugate to the kernel function, i.e., $Q_0 \equiv \text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{S}_0)$ is a Normal-Inverse-Wishart distribution, where by $\text{NIW}(\boldsymbol{a}, b, c, \boldsymbol{D})$ we mean that $\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \text{N}(\boldsymbol{a}, \boldsymbol{\Sigma}/b)$ and $\boldsymbol{\Sigma} \sim IW(\boldsymbol{D}, c)$. We also assume $P_0 \equiv \text{NIW}(\boldsymbol{\mu}_1, \kappa_1, \nu_1, \boldsymbol{S}_1)$. We specify the base measure of the discrete component Q_0 by setting $\boldsymbol{\mu}_0$ equal to the sample mean of the data, $\kappa_0 = 1$, $\nu_0 = d + 3 = 7$, and $\boldsymbol{S}_0$ equals to the diagonal of the sample variance of the data. We further specify the parameters of the contaminant measure as follows: $\boldsymbol{\mu}_1$ equals the sample mean of the data, $\kappa_1 = 0.1$, $\nu_1 = d + 3 = 7$, and $\boldsymbol{S}_1$ matches the sample variance of the data, in order to force an over-disperse contaminant measure with respect to the base measure. We complete the model specification assuming vague priors for the parameters of the mixing measure, $\vartheta \sim \text{Gamma}(2, 0.02)$ and $\sigma, \beta \sim \text{Unif}(0, 1)$. Posterior inference is carried out by using the sampling scheme described in Section 6.1, see also Section H for diagnostics. We exploit the decisional approach based on the variation of information loss function (Wade and Ghahramani, 2018; Rastelli and Friel, 2018) to provide an optimal posterior point estimate of the latent partition induced by the data. Relying on our outlier definition (see beginning of Section 6) we recognize an observation as an outlier if it belongs to a cluster with frequency one, regardless if this is associated with the contaminant measure or the discrete term of the model.

The results of the point estimate of the latent partition in the data are summarized in Table 4, in comparison with the previous clusters identified by Ibata et al. (2011). Within the 16 stars identified as contaminants, 4 belongs to the globular cluster identified by Ibata et al. (2011), 5 stars to the likely globular cluster group, and 7 to the contaminants. Most of the stars of the main estimated cluster, denoted by A in Table 4, belong to the globular cluster of Ibata et al. (2011). We have also recovered two additional clusters: a cluster of stars mainly belonging to the globular cluster in (Ibata et al., 2011), and a cluster with two contaminant stars. Our findings are coherent with the previous literature, but providing a more conservative detection of the contaminant

			CPY partition			
			Singletons	*A*	*B*	*C*
		total	*16*	*115*	*4*	*4*
	globular cluster	*118*	4	109	3	2
Ibata et al. (2011)	*likely globular cluster*	*12*	5	6	1	0
	contaminants	*9*	7	0	0	2

Table 4: Comparison between the partition described in Ibata et al. (2011) and optimal partition estimated using a contaminated Pitman-Yor mixture model

stars. We can further compare our findings with the latent partition obtained using a Pitman-Yor mixture model, as described in Section H.1 of the Supplementary Material: the optimal latent partition recovered with the contaminated Pitman-Yor prior is characterized by a larger main globular cluster but also a higher number of singletons.

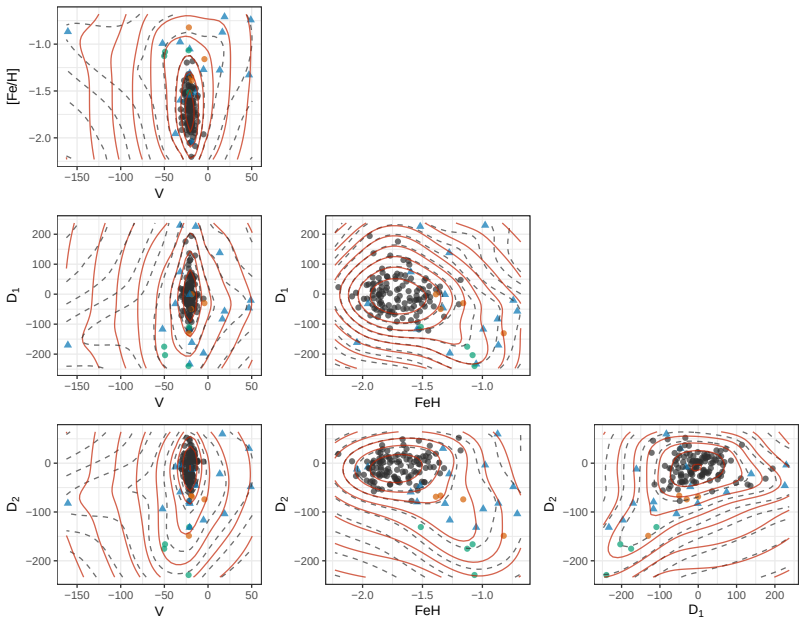

Figure 4: Posterior expectation of the random density and optimal partition of the NGC 2419 data. Red lines: mean of the posterior random density with the contaminated Pitman-Yor mixture model of Gaussian distributions. Black dashed lines: mean of the posterior random density with the Pitman-Yor mixture model of Gaussian distributions. The points are colored according to the point estimate of the latent partition of the data with a contaminated Pitman-Yor mixture model of Gaussian distributions.

Figure 4 shows the point estimate of the latent partition of the data under the contaminated Pitman-Yor mixture model. The red lines represent the contour lines of the mean of the posterior density under the contaminated prior, while the contour

lines estimated under the standard Pitman-Yor mixture model correspond to the black dashed lines. Note that the inclusion of a diffuse component results in smoothed contour lines, thus in smoothed density estimates, indeed the contour lines corresponding to the standard Pitman-Yor mixture model exhibit some peaks in correspondence of the contaminants. We finally analyze the predictive performance of the proposed model by considering the leave-one-out cross-validation criterion (LOO-CV). We further compare the mean of the posterior random densities with respect to a benchmark density f_0. The density f_0 corresponds to the kernel density estimate on the basis of a trimmed version of data, including only the observations identified as non-contaminants, according to Ibata et al. (2011). The estimated density $\hat{f}_n = \mathrm{E}[\tilde{f} \mid Y_1, \ldots, Y_n]$ under the two models (contaminated and standard mixture) is then compared with f_0 by evaluating the symmetrised Kullback-Leibler divergence $\text{S-KL}(\hat{f}_n, f_0) = \mathrm{KL}\big(\hat{f}_n \mid f_0\big) + \mathrm{KL}\big(f_0 \mid \hat{f}_n\big)$, where $\mathrm{KL}(g \mid h) = \int_{\mathbb{X}} g(\boldsymbol{x}) \log\big(g(\boldsymbol{x})/h(\boldsymbol{x})\big)\mathrm{d}\boldsymbol{x}$ for generic densities g, h.

	LOO-CV	S-KL$(\hat{f}_n, f_0)$
cPY	3921.85	0.38
PY	3912.03	0.51

Table 5: Summaries of the posterior random densities for the Pitman-Yor (cPY) and the Pitman-Yor (PY) mixture models. Middle column: LOO-CV criterion. Right column: S-KL divergence for the mean of the posterior random densities $\hat{f}_n$. The reference density f_0 corresponds to the kernel density estimate (KDE) obtained with the trimmed data.

Table 5 shows comparable values of the LOO-CV criterion for the cPY and the PY models. However, we can appreciate that the inclusion of a contaminant term in the mixing measure specification is producing a posterior random density with mean closer to the benchmark density, in terms of S-KL divergence. Such behaviour is due to the contaminant measure which inflates the tails of the distribution in presence of outliers, instead of generating specific components for them.

7 Discussion

We introduced a new family of priors outside the Gibbs-type one which are still tractable from an analytical and computational viewpoint. According to the characterization by Bacallado et al. (2017), the predictive probability weights of Gibbs-type priors cannot depend on the number of observations recorded with frequency one $M_{n,1}$ in the initial sample. Our prior choice has the advantage to enrich the predictive structure of Gibbs-type priors with the inclusion of the additional sampling information on $M_{n,1}$, retaining interpretability of the reinforcement mechanism. Moreover we discussed the usage of contaminated Gibbs-type priors in two situations: i) for discrete data in presence of an excess of ones; ii) in mixture models to account for outliers, showing the importance of their increased flexibility. In particular, our simulation studies highlights that a correct modeling of singletons is of paramount importance to prevent undesirable inferential conclusions both in terms of prediction of relevant quantities and for estimating densities or parameters. Nevertheless, the use of contaminated Gibbs-type

priors is not restricted to the scenarios presented in this manuscript, but they can be relevant in other applications, where the presence of elements with frequency one is a key inferential interest.

Firstly, contaminated Gibbs-type priors could be of potential interest in the analysis of disclosure risk for microdata. Microdata files typically contains two types of categorical information about individuals: identifying and sensitive information. Before releasing a dataset, statistical agencies estimate different measures of disclosure risk, which are typically based on the number of sample records which have a unique combination of the categorical variables and that are not shared with any element of the entire population. See, e.g., Bethlehem et al. (1990a); Skinner and Elliot (2002); Bethlehem et al. (1990b) for possible definitions and estimators of disclosure indexes. In the disclosure risk framework, the random variable $\bar{M}_{m_1}$ appearing in our model represents a measure of disclosure, i.e., the number of records that contain a unique element both in the sample and in the whole population. Notice that, as mentioned in section 3, the latent variable $\bar{M}_{m_1}$ can be estimated through a suitable posterior computational strategy.

Language modeling constitutes another application area when one is interested to estimate the number of *hapax legomena* in a corpus of documents. An hapax legomena is indeed a word that occurs only once in the entire production of an author. These unique words are particularly important since they have been recognized as peculiar usage of words by specific authors, and they represent an interesting problem to study from a statistical perspective. See e.g. Baayen (2001) for further details on word frequency distributions. In such a framework, one may use a contaminated Gibbs-type prior to estimate the number of hapax legomena on the basis of the latent quantity $\bar{M}_{m_1}$.

Finally, contaminated Gibbs-type models can be extended in several directions. For example, they can be exploited to test the presence of contaminated observations in a set of data by selecting a spike and slab prior (Mitchell and Beauchamp, 1988) for the parameter β in (4). More precisely one may specify a prior for β which assigns positive mass to the point $\beta = 1$. Another interesting direction of research is the generalization of contaminated species sampling models to a contaminated version of feature allocation models (see, e.g., Ghahramani et al., 2007; Broderick et al., 2013). Work on these points is ongoing.

Supplementary Material

Supplementary material for "Contaminated Gibbs-type priors". Section A provides all the proofs of the results presented in the paper. Section B contains results for the contaminated Pitman-Yor case. Section C describes numerical illustrations on the predictive structure of contaminated Gibbs-type priors. The algorithms for the discrete case is presented in Section D. Details on the contaminated Zipf distribution are deferred to Section E. Section F contains the summaries of the simulation studies for the contaminated Pitman-Yor mixture model. Sections G and H provide further insights on the North America Ranidae dataset analysis and the NGC 2419 globular cluster dataset, respectively.

References

Arbel, J., Corradin, R., and Nipoti, B. (2021). "Dirichlet process mixtures under affine transformations of the data." *Comput. Stat.*, 36: 577–601. 22

Baayen, H. R. (2001). *Word Frequency Distributions*. Springer Netherlands. 25

Bacallado, S., Battiston, M., Favaro, S., and Trippa, L. (2017). "Sufficientness postulates for Gibbs-type priors and hierarchical generalizations." *Statist. Sci.*, 32(4): 487–500. 9, 13, 24

Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). "The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions." *Bayesian Anal.*, 1 – 33. 3

Berger, J. O. and Berliner, L. M. (1986). "Robust Bayes and Empirical Bayes Analysis with ϵ-Contaminated Priors." *Annals of Statistics*, 14: 461–486. 3

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990a). "Disclosure Control of Microdata." *J. Amer. Statist. Assoc.*, 85(409): 38–45. 25

— (1990b). "Disclosure Control of Microdata." *Journal of the American Statistical Association*, 85(409): 38–45. 25

Bigelow, J. L. and Dunson, D. B. (2009). "Bayesian Semiparametric Joint Models for Functional Predictors." *Journal of the American Statistical Association*, 104(485): 26–36. 19

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. 19

Broderick, T., Pitman, J., and Jordan, M. I. (2013). "Feature Allocations, Probability Functions, and Paintboxes." *Bayesian Analysis*, 8(4): 801 – 836. 25

Campbell, T., Cai, D., and Broderick, T. (2018). "Exchangeable trait allocations." *Electronic Journal of Statistics*, 12(2): 2290 – 2322. 3

Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2017). "On the Pitman–Yor process with spike and slab base measure." *Biometrika*, 104(3): 681–697. 3

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2): 212–229. 1, 5, 19

de Finetti, B. (1937). "La prévision : ses lois logiques, ses sources subjectives." *Ann. Inst. H. Poincaré*, 7(1): 1–68. 4

Eggenberger, F. and Pólya, G. (1923). "Über die Statistik verketteter Vorgänge." *AMM - Zeitschrift Für Angewandte Mathematik Und Mechanik*, 3(4): 279–289. 12

Escobar, M. D. (1988). "Estimating the means of several normal populations by non-

parametric estimation of the distribution of the means." Ph.D. thesis, Department of Statistics, Yale University. 20

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *J. Amer. Statist. Assoc.*, 90(430): 577–588. 20

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). "Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5): 993–1008. 3

Favaro, S., Lijoi, A., and Prünster, I. (2013). "Conditional formulae for Gibbs-type exchangeable random partitions." *Ann. Appl. Probab.*, 23(5): 1721–1754. 13

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209 – 230. 1, 5

Freund, F. and Möhle, M. (2017). "On the size of the block of 1 for Ξ-coalescents with dust." *Modern Stochastics: Theory and Applications*, 4(4): 407–425. 3

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC. 19

GBIF.org (2021). "GBIF Occurrence Download, https://doi.org/10.15468/dl.cr98vh." 2, 15

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). "Bayesian nonparametric latent feature models." *Bayesian statistics*, 8: 1–25. 25

Gnedin, A. and Pitman, J. (2005). "Exchangeable Gibbs partitions and Stirling triangles." *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12): 83–102, 244–245. 1, 5

Harald, B. R. (2001). *Word Frequency Distributions*. Text, Speech and Language Technology. Springer. 2

Heaukulani, C. and Roy, D. M. (2020). "Gibbs-type Indian buffet processes." *Bayesian Anal.*, 15(3): 683–710. 2

Ibata, R., Sollima, A., Nipoti, C., Bellazzini, M., Chapman, S., and Dalessandro, E. (2011). "The globular cluster ngc 2419: a crucible for theories of gravity." *ApJ*, 738(2): 1–23. 4, 22, 23, 24

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *J. Amer. Statist. Assoc.*, 96(453): 161–173. 2, 19

Jara, A., Lesaffre, E., Iorio, M. D., and Quintana, F. (2010). "Bayesian semiparametric inference for multivariate doubly-interval-censored data." *Ann. Appl. Statist.*, 4(4): 2126 – 2149. 2

Kingman, J. F. C. (1978). "The Representation of Partition Structures." *Journal of the London Mathematical Society*, s2-18(2): 374–380. 3

Krnjajić, M., Kottas, A., and Draper, D. (2008). "Parametric and nonparametric

Bayesian model specification: A case study involving models for count data." *Comput. Stat. Data Anal.*, 52(4): 2110–2128. 19

Lijoi, A., Mena, R. H., and Prünster, I. (2005). "Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors." *J. Amer. Statist. Assoc.*, 100(472): 1278–1291. 5, 19

Lijoi, A., Mena, R. H., and Prünster, I. (2007a). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94(4): 769–786. 2, 3, 19

— (2007b). "Controlling the reinforcement in Bayesian non-parametric mixture models." *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, 69(4): 715–740. 2, 5

Lijoi, A. and Prünster, I. (2010). "Models beyond the Dirichlet process." In *Bayesian nonparametrics*, volume 28 of *Camb. Ser. Stat. Probab. Math.*, 80–136. Cambridge Univ. Press, Cambridge. 1

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *Ann. Statist.*, 12(1): 351 – 357. 19

Mano, S. (2018). *Partitions, hypergeometric systems, and Dirichlet processes in statistics*. SpringerBriefs in Statistics. Springer, Tokyo. JSS Research Series in Statistics. 14

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *J. Amer. Statist. Assoc.*, 83(404): 1023–1032. 25

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83(1): 67–79. 19

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *J. Comput. Graph. Statist.*, 9(2): 249–265. 21

Perman, M., Pitman, J., and Yor, M. (1992). "Size-biased sampling of Poisson point processes and excursions." *Probab. Theory Related Fields*, 92(1): 21–39. 1

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." In *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes Monogr. Ser.*, 245–267. Inst. Math. Statist., Hayward, CA. 1, 4

— (2006). *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. 5, 10

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *Ann. Probab.*, 25(2): 855–900. 5

Quintana, F. A. (2006). "A predictive view of Bayesian clustering." *J. Statist. Plann. Inference*, 136(8): 2407–2429. 3, 19

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(2): 557–574. 19

Rastelli, R. and Friel, N. (2018). "Optimal Bayesian estimators for latent variable cluster models." *Stat. Comput.*, 28(6): 1169–1186. 8, 22

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *Ann. Statist.*, 31(2): 560–585. 1, 5

Scarpa, B. and Dunson, D. B. (2009). "Bayesian Hierarchical Functional Data Analysis via Contaminated Informative Priors." *Biometrics*, 65(3): 772–780. 3

Shotwell, M. S. and Slate, E. H. (2011). "Bayesian outlier detection with Dirichlet process mixtures." *Bayesian Anal.*, 6(4): 665–690. 20

Skinner, C. J. and Elliot, M. J. (2002). "A measure of disclosure risk for microdata." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4): 855–867. 25

Stoler, N. and Nekrutenko, A. (2021). "Sequencing error profiles of Illumina sequencing instruments." *NAR Genomics and Bioinformatics*, 3(1). 2

Teh, Y. W. (2006). "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 985–992. 2

Teh, Y. W. and Jordan, M. I. (2010). *Hierarchical Bayesian nonparametric models with applications*, 158–207. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 2

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27: 1413–1432.

Wade, S. and Ghahramani, Z. (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)." *Bayesian Analysis*, 13(2): 559 – 626. 8, 22

Zabell, S. (1997). *The continuum of inductive methods revisited*, 351–385. University of Pittsburgh Press. 11, 12

Acknowledgments

The authors are grateful to the Associate Editor and two anonymous Referees for their valuable comments and suggestions, which lead to a substantial improvement of the paper. The authors gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022, and the DEMS Data Science Lab for supporting this work through computational resources. Federico Camerlenghi is a member of the *Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni* (GNAMPA) of the *Istituto Nazionale di Alta Matematica* (INdAM).