



HAL
open science

BNPdensity: Bayesian nonparametric mixture modelling in R

Julyan Arbel, Guillaume Kon Kam King, Antonio Lijoi, Luis E.
Nieto-Barajas, Igor Prünster

► **To cite this version:**

Julyan Arbel, Guillaume Kon Kam King, Antonio Lijoi, Luis E. Nieto-Barajas, Igor Prünster. BNPdensity: Bayesian nonparametric mixture modelling in R. Australian and New Zealand Journal of Statistics, 2021, 63 (3), pp.542-564. 10.1111/anzs.12342 . hal-03433254

HAL Id: hal-03433254

<https://hal.inrae.fr/hal-03433254v1>

Submitted on 17 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

BNPdensity: Bayesian nonparametric mixture modeling in R

J. Arbel¹, G. Kon Kam King^{2*}, A. Lijoi³, L. Nieto-Barajas⁴ and I. Prünster³

Univ. Grenoble Alpes, Inria, Univ. Paris-Saclay, INRAE, Bocconi University, ITAM

Summary

Robust statistical data modelling under potential model mis-specification often requires leaving the parametric world for the nonparametric. In the latter, parameters are infinite dimensional objects such as functions, probability distributions or infinite vectors. In the Bayesian nonparametric approach, prior distributions are designed for these parameters, which provide a handle to manage the complexity of nonparametric models in practice. However, most modern Bayesian nonparametric models seem often out of reach to practitioners, as inference algorithms need careful design to deal with the infinite number of parameters. The aim of this work is to facilitate the journey by providing computational tools for Bayesian nonparametric inference. The article describes a set of functions available in the R package *BNPdensity* in order to carry out density estimation with an infinite mixture model, including all types of censored data. The package provides access to a large class of such models based on normalized random measures, which represent a generalization of the popular Dirichlet process mixture. One striking advantage of this generalization is that it offers much more robust priors on the number of clusters than the Dirichlet. Another crucial advantage is the complete flexibility in specifying the prior for the scale and location parameters of the clusters, because conjugacy is not required. Inference is performed using a theoretically grounded approximate sampling methodology known as the Ferguson & Klass algorithm. The package also offers several goodness of fit diagnostics such as QQ-plots, including a cross-validation criterion, the conditional predictive ordinate. The proposed methodology is illustrated on a classical ecological risk assessment method called the Species Sensitivity Distribution (SSD) problem, showcasing the benefits of the Bayesian nonparametric framework.

Key words: Bayesian nonparametric inference; density estimation; Ferguson and Klass algorithm; infinite mixture models; random probability measures; R

* Author to whom correspondence should be addressed.

¹ Univ. Grenoble Alpes, Inria, CNRS LJK, 38000 Grenoble, France

² Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

³ Department of Decision Sciences and BIDSa, Bocconi University, Italy

⁴ Department of Statistics, ITAM, Mexico

Email: guillaume.kon-kam-king@inrae.fr

Acknowledgment. We would like to thank a Referee for several interesting suggestions and Matti Vihola for useful advice on adaptive MCMC. J. Arbel is partially supported by Grenoble Alpes Data Institute (ANR-15-IDEX-02). A. Lijoi and I. Prünster are partially supported by MIUR, PRIN Project 2015SSNS29B.

1. Introduction

R (RCoreTeam 2019) is often cited by Bayesian statisticians as their favorite programming language due to the many packages that provide tools for Bayesian inference. The general program for Bayesian inference BUGS (Gilks, Thomas & Spiegelhalter 1993) has been available for a couple of decades, with interfaces in R. Since then, additional software has been developed to make that language more accessible to the users, for instance OpenBUGS (Thomas et al. 2006), JAGS (Plummer 2003), and Stan (Stan Development Team & Stan Development Team 2019). All three can be accessed directly from R by respectively using R2OpenBUGS/R2WinBUGS (Sturtz, Ligges & Gelman 2005), rjags (Plummer 2019), runjags (Denwood 2016), and rstan (Stan Development Team 2018). Programs for specific fields of Bayesian statistics have appeared in recent years, for instance bspmma (Burr 2012) for meta-analysis using Dirichlet Process Mixture (DPM) models, DPpackage (Jara 2007; Jara et al. 2011), a bundle of functions for Bayesian nonparametric models, BNPmix (Canale, Corradin & Nipoti 2019), a set of functions for density estimation with Dirichlet process and Pitman–Yor mixing measures via marginal algorithms, PReMiuM (Liverani et al. 2015) for profile regression using the Dirichlet process, Biips (Todeschini, Caron & Fuentes 2014) for Bayesian inference via particle filtering, Bayesian Regression (Karabatsos 2017) for Bayesian nonparametric regression. Packages mcclust (Scrucca et al. 2016), mcclust.ext (Wade & Ghahramani 2018) and GreedyEPL (Rastelli & Friel 2018) provide point estimation and credible sets for Bayesian cluster analysis. The interested reader may refer to the CRAN Task View on Bayesian Inference for an extensive list of R packages dedicated to Bayesian statistics (see Section 4 for a more detailed discussion of R packages for Bayesian density estimation).

Robust statistical data modeling under potential model mis-specification often requires relaxing parametric assumptions for nonparametric assumptions. In Bayesian Nonparametrics (BNP), parameters are infinite dimensional objects such as functions, probability distributions or infinite vectors. Prior distributions are designed for these parameters, which provide a handle to manage the complexity of nonparametric models in practice. However, the applicability of BNP models, for data analysis, depends on the availability of user-friendly software. This is because BNP models typically require complex representations, which may not be immediately accessible to non-experts. This work focuses on inference of densities with mixture models (Frühwirth-Schnatter, Celeux & Robert 2018). The purpose of the present paper is to introduce and describe an extensive revamping of the BNPdensity package, originally presented in Barrios et al. (2013). The package is programmed in R, and is available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=BNPdensity>. To the best of our

43 knowledge, `BNPdensity` is the first R package which implements BNP density models
44 including all types of censored data (left-, right- and interval-censored data), under a general
45 specification of BNP priors called normalised generalised gamma processes (Lijoi, Mena
46 & Prünster 2007b; Barrios et al. 2013). The improvements to the package cover various
47 aspects. Notably, careful profiling and re-writing of some critical parts of the code, along
48 with the use of the R bytecode compiler, yielded a 4-fold decrease of the running time
49 of the algorithm. Drawing on the flexibility of the algorithm to use non-conjugate prior,
50 we also implemented a range of popular new priors on the scale parameter of the clusters
51 such as the half-Cauchy (Gelman 2006; Chung et al. 2015), the truncated Gaussian and the
52 uniform distributions. We also revised the truncation method in the algorithm, intended to
53 deal with the infinite dimensional random measures in the BNP model, to include recent
54 contributions by Arbel & Prünster (2017). These provide a better and principled control of
55 the truncation approximation. Moreover, we extended `BNPdensity` to include all types of
56 censored data (right-, left- or interval-censored data). To leverage on the clustering properties
57 of BNP mixture models, we interfaced `BNPdensity` with other packages to estimate the
58 optimal clustering from posterior samples and provided cluster visualisation tools. We also
59 implemented functions to compute prior distributions on the number of mixture components,
60 for various processes, to better inform prior specification. Finally, we added several new
61 functions for graphical model checking, assessing Markov chain Monte Carlo (MCMC)
62 convergence and parallel computation.

63 The paper is organised as follows. We start with a concise overview of Bayesian
64 nonparametric mixture models for density estimation in Section 2, along with our strategy
65 for posterior inference and a description of the recent improvements to `BNPdensity`. We then
66 describe the package and its general syntax in Section 3, including some simple examples, and
67 provide in Section 4 a comprehensive comparison of the features and functionalities offered
68 in three R packages dedicated to BNP density estimation, namely: `BNPdensity`, `BNPmix`,
69 and `DPpackage`. We then conclude with a case study in Section 5.

70

2. Bayesian nonparametric density estimation

71 This section aims at providing a concise review of the statistical model used in the
72 `BNPdensity` package. As the name suggests, the focus of the package is density estimation
73 based on BNP priors, including all types of censored data. The density model used is a
74 mixture model (Frühwirth-Schnatter, Celeux & Robert 2018), where the mixing measure is a
75 BNP prior, thus leading to an infinite mixture model.

76 The most widely used BNP mixture model for density estimation is the Dirichlet Process
 77 Mixture (DPM) model due to [Lo \(1984\)](#). Generalisations of the DPM correspond to allowing
 78 the mixing distribution to be any discrete nonparametric prior. A large class of such prior
 79 distributions is obtained by normalising increasing additive processes ([Sato 1999](#)). The
 80 normalisation step, under suitable conditions, gives rise to so-called Normalised Random
 81 Measures with Independent Increments (NRMI) as introduced in [Regazzini, Lijoi & Prünster](#)
 82 [\(2003\)](#). See also [Barrios et al. \(2013\)](#).

83 We focus on a class of NRMI that are obtained by normalising the increments of a
 84 generalised gamma process ([Brix 1999](#)) proposed in [Lijoi, Mena & Prünster \(2007a\)](#), which
 85 enjoy analytical tractability and include many well-known priors as special cases. Generalised
 86 gamma processes are discrete random measures $\tilde{\rho}$ of the form

$$\tilde{\rho} = \sum_{i=1}^{\infty} J_i \delta_{\theta_i}, \quad (1)$$

87 where the weights J_i do not sum to one, while the location parameters θ_i are sampled iid
 88 from a measure P_0 , a probability distribution on the parameter space Θ . In what follows, P_0
 89 is considered as diffuse. (J_i, θ_i) are the points of a Poisson process with mean intensity:

$$\nu(dv, d\theta) = \frac{e^{-\kappa v}}{\Gamma(1-\gamma)v^{1+\gamma}} dv \alpha P_0(d\theta), \quad (2)$$

90 which depends on parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$ such that $(\kappa, \gamma) \neq (0, 0)$. The measure
 91 ν in (2) characterises $\tilde{\rho}$ and is often referred to as the Lévy intensity. The base
 92 measure is αP_0 , where $\alpha > 0$. The corresponding generalised gamma NRMI, obtained by
 93 normalising the generalised gamma process as $\tilde{P}(\cdot) := \tilde{\rho}(\cdot)/\tilde{\rho}(\mathbb{X})$ will be denoted as $\tilde{P} \sim$
 94 $\text{NGG}(\alpha, \kappa, \gamma; P_0)$. This class of priors contains as special cases the Dirichlet process which is
 95 a $\text{NGG}(\alpha, 1, 0; P_0)$ process, the normalised inverse Gaussian (N-IG) process ([Lijoi, Mena &](#)
 96 [Prünster 2005](#)), which corresponds to a $\text{NGG}(1, \kappa, 1/2; P_0)$ process, and the N-stable process
 97 ([Kingman 1975](#)) which arises as $\text{NGG}(1, 0, \gamma; P_0)$.

98 We now describe the mixture model in more detail. We consider a density kernel $k(\cdot | \theta)$
 99 mixed with respect to $\tilde{P} \sim \text{NGG}(\alpha, \kappa, \gamma; P_0)$ thus obtaining the random mixture density

$$\tilde{f}(x) = \int_{\Theta} k(x | \theta) \tilde{P}(d\theta). \quad (3)$$

100 This can equivalently be written in a hierarchical form as

$$\begin{aligned}
 X_i &| \theta_i \stackrel{\text{iid}}{\sim} k(\cdot | \theta_i), \quad i = 1, \dots, n, \\
 \theta_i &| \tilde{P} \stackrel{\text{iid}}{\sim} \tilde{P}, \quad i = 1, \dots, n, \\
 \tilde{P} &\sim \text{NGG}(\alpha, \kappa, \gamma; P_0).
 \end{aligned}
 \tag{4}$$

101 Details on possible choices for the kernel k and the base measure P_0 are provided in Section 3,
 102 while in Section 4 we argue that conjugacy is not required in this setting.

103 We denote by f_0 the density with respect to the Lebesgue measure of the NGG base
 104 measure P_0 on Θ . When P_0 depends on a further hyperparameter ϕ , we use the notation
 105 $f_0(\cdot | \phi)$. Using the `MixNRM12` function corresponds to the specification of a nonparametric
 106 model for the location and scale parameters of the mixture where the mixture parameter θ
 107 takes the form of the vector (μ, σ) . In order to distinguish the hyperparameters for location
 108 and scale, we will use the notation $f_0(\mu, \sigma | \phi) = f_0^1(\mu | \sigma, \phi) f_0^2(\sigma | \phi)$. In applications a
 109 priori independence between μ and σ is commonly assumed, and this is indeed a natural
 110 assumption for the illustration in Section 5.

111 The most popular uses of mixtures with discrete random probability measures, such as
 112 the one displayed in (4), relate to density estimation and data clustering. The former can be
 113 addressed by evaluating the posterior expectation of the random density \tilde{f} defined in (3),
 114 given a sample $\mathbf{X} = (X_1, \dots, X_n)^\top$,

$$\hat{f}_n(x) = \text{E}(\tilde{f}(x) | \mathbf{X})
 \tag{5}$$

115 for any x in \mathbb{X} . As for the latter, if R_n is the number of distinct latent values $\theta_1^*, \dots, \theta_{R_n}^*$ out
 116 of a sample of size n , one can deduce a partition of the observations such that any two X_i
 117 and X_j belong to the same cluster if the corresponding latent variables θ_i and θ_j coincide.
 118 Then, it is interesting to determine an estimate \hat{R}_n of the number of clusters into which the
 119 data are grouped, along with the clustering structure. For details on clustering estimation in
 120 our setting, see Section 2.3.

121 In the next subsection, we show how to solve all estimation problems with a posterior
 122 sampling algorithm.

123 2.1. Posterior sampling via a conditional Gibbs sampler

124 According to the terminology of Papaspiliopoulos & Roberts (2008), posterior sampling
 125 methods for BNP mixture models can be divided into two classes: marginal and conditional
 126 methods. Marginal methods, such as Escobar & West (1995); MacEachern & Müller (1998);

127 Neal (2000), integrate out the the infinite-dimensional component (1) of the hierarchical
 128 model and sample from the marginal distribution of the remaining variables. Conditional
 129 methods work directly on (4) and must solve the problem of sampling the trajectories of an
 130 infinite-dimensional random element. However, they allow inference on the latent random
 131 measure \tilde{P} , for instance on the jump sizes. An example of conditional method, which nicely
 132 fits our framework, can be the Ferguson and Klass algorithm. Unlike marginal samplers, it
 133 allows for estimating non-linear functionals of the underlying posterior distribution, such as
 134 credible intervals. Here we sketch the conditional algorithm implemented in BNPdensity
 135 which allows to draw posterior simulations from mixtures based on a general NRM (a very
 136 thorough description of the algorithm can be found in Barrios et al. 2013). It works equally
 137 well regardless of whether the kernel k and P_0 form a conjugate pair and readily yields
 138 credible intervals. The algorithm is an implementation of the posterior characterisation of
 139 NRM provided in James, Lijoi & Prünster (2009).

For n observations $\mathbf{X} = (X_1, \dots, X_n)^\top$ in $\mathbb{X} = \mathbb{R}$, we consider the random distribution
 function induced by $\tilde{\rho}$,

$$\tilde{M} := \left\{ \tilde{M}(s) = (\tilde{\rho}((-\infty, s_1]), \dots, \tilde{\rho}((-\infty, s_n]))^\top, \quad s = (s_1, \dots, s_n)^\top \in \mathbb{R}^n \right\}.$$

140 For the implementation of the Gibbs sampling scheme, we use the distributions of $[\tilde{M} \mid \mathbf{X}, \boldsymbol{\theta}]$
 141 and $[\boldsymbol{\theta} \mid \mathbf{X}, \tilde{M}]$. Due to conditional independence properties, the conditional distribution
 142 of \tilde{M} , given \mathbf{X} and $\boldsymbol{\theta}$, does not depend on \mathbf{X} , that is, $[\tilde{M} \mid \mathbf{X}, \boldsymbol{\theta}] = [\tilde{M} \mid \boldsymbol{\theta}]$. Thanks to
 143 Theorem 1 in Barrios et al. (2013) (originating in James, Lijoi & Prünster 2009), the posterior
 144 distribution function $[\tilde{M} \mid \boldsymbol{\theta}]$ can be characterised as a mixture in terms of a latent variable
 145 U , that is through the distributions $[\tilde{M} \mid U, \boldsymbol{\theta}]$ and $[U \mid \boldsymbol{\theta}]$. Thus, the Gibbs sampler uses the
 146 following conditional distributions:

- 147 1. $[U \mid \boldsymbol{\theta}]$: sampling the latent variable U conditionally on the latent parameters $\boldsymbol{\theta}$, where
 148 U follows the distribution:

$$f_{U \mid \mathbf{X}}(u) \propto u^{n-1} (u + \kappa)^{r\gamma - n} \exp \left\{ -\frac{a}{\gamma} (u + \kappa)^\gamma \right\}. \quad (6)$$

149 Sampling U is performed via a Metropolis–Hastings (M-H) step with a gamma
 150 proposal distribution $\text{ga}(\delta, \delta/u^{[t]})$ centered at the previous U value $u^{[t]}$ with a
 151 tuning parameter δ controlling the coefficient of variation. An adaptive version of
 152 the M-H algorithm (Roberts & Rosenthal 2009) without the tuning parameter is also
 153 implemented in the package, and proposed with the option `adaptive=TRUE`. It uses
 154 a log-transformation of the random variable U . Note that the target density (6) not

155 being log-concave, ergodicity cannot be proven as in [Roberts & Rosenthal \(2009\)](#).
 156 Nevertheless, the adaptive version appears to offer superior performance in practice.

157 2. $[\tilde{M} \mid U, \boldsymbol{\theta}]$: simulating the infinite dimensional process conditionally on the parameters
 158 and the latent variable U . This is performed using the [Ferguson & Klass \(1972\)](#)
 159 algorithm. According to to Theorem 1 in [Barrios et al. \(2013\)](#), the conditional
 160 distribution of \tilde{M} is composed of two parts, a part without fixed points of discontinuity
 161 \tilde{M}^* which can be expressed as an infinite sum of random jumps occurring at
 162 random locations and a part with fixed points of discontinuity, or in other words:
 163 $\tilde{M}(\mathbf{s}) = \tilde{M}^*(\mathbf{s}) + \sum_{j=1}^{R_n} J_j \mathbb{I}_{(-\infty, \mathbf{s}]}(\boldsymbol{\theta}_j^*)$ where the $\boldsymbol{\theta}_j^*$, $j = 1, \dots, R_n$ denote the R_n
 164 distinct parameters among $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ and where $(-\infty, \mathbf{s}] = \{\mathbf{x} \in \mathbb{R}^n : x_i \leq s_i, i =$
 165 $1, \dots, n\}$. In the infinite sum:

$$\tilde{M}^*(\mathbf{s}) = \sum_{j=1}^{\infty} J_j \mathbb{I}_{(-\infty, \mathbf{s}]}(\boldsymbol{\vartheta}_j), \quad (7)$$

166 the J_j s are obtained by inverting the relation $\xi_j = N(J_j)$, where ξ_1, ξ_2, \dots are jump
 167 times of a standard Poisson process of unit rate, that is $\xi_1, \xi_2 - \xi_1, \dots \stackrel{\text{iid}}{\sim} \text{ga}(1, 1)$, with
 168

$$N(v) = \frac{a}{\Gamma(1 - \gamma)} \int_v^{\infty} e^{-(\kappa+u)x} x^{-(1+\gamma)} dx, \quad (8)$$

169 while the jumps $\boldsymbol{\vartheta}_j = (\boldsymbol{\vartheta}_j^{(1)}, \dots, \boldsymbol{\vartheta}_j^{(n)})^\top$ are sampled from the base measure P_0 . The
 170 jumps J_j^* at the fixed locations $\boldsymbol{\theta}_j^*$ are gamma distributed:

$$f_j^*(v) = \frac{(\kappa + u)^{n_j - \gamma}}{\Gamma(n_j - \gamma)} v^{n_j - \gamma - 1} e^{-(\kappa+u)v}, \quad (9)$$

171 where n_j are the multiplicities, i.e. the number of $\boldsymbol{\theta}_j$ equal to $\boldsymbol{\theta}_j^*$. A fundamental
 172 merit of Ferguson and Klass' representation, compared to similar algorithms, is the
 173 fact that the random heights J_i are obtained in a descending order. Therefore, one can
 174 truncate the series in (7) at a certain finite index Q to be decided via a moment-matching
 175 criterion (see Section 2.2). This also guarantees that the highest jumps are not left out.

176 3. $[\boldsymbol{\theta} \mid \mathbf{X}, \tilde{M}]$: resampling the latent cluster parameters given the data and the random
 177 measure. The support of the conditional distribution of $\boldsymbol{\theta}_i$ are the locations of the
 178 jumps of \tilde{M} , $\{\bar{J}_j\}_{j=1}^{\infty} = \{J_1^*, \dots, J_{R_n}^*, J_1, \dots\}$ with associated jumps $\{\bar{\boldsymbol{\vartheta}}_j\}_{j=1}^{\infty} =$
 179 $\{\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{R_n}^*, \boldsymbol{\vartheta}_1, \dots\}$,

$$f_{\boldsymbol{\theta}_i \mid X_i, \tilde{M}}(\mathbf{s}) \propto \sum_j k(X_i \mid \mathbf{s}) \bar{J}_j \delta_{\bar{\boldsymbol{\vartheta}}_j}(\mathbf{s}). \quad (10)$$

180 Simulating from this conditional distribution when an approximation with a finite
 181 number of jumps has been determined is straightforward: one just needs to evaluate
 182 the right-hand side of the expression above and normalise.

183 4. Updating the hyperparameters of P_0 . We only put a prior on the hyperparameters for the
 184 location parameters, found to have a higher impact. Assuming a priori independence
 185 between location and scale parameters of the clusters, the conditional posterior
 186 distribution on the hyperparameters given the data and the rest of the parameters only
 187 depends on the distinct location parameters. A simple way to proceed is thus to consider
 188 a prior conjugate to the base measure.

189 We also include a resampling of the unique values of the cluster parameters via a M-H
 190 step to avoid the ‘sticky clusters effect’, as suggested in [Bush & MacEachern \(1996\)](#).

191 We devote the next section to explaining the moment-matching criterion used for
 192 truncation in the second conditional, which is a recent addition to the package `BNPdensity`.

193 2.2. Moment-matching criterion

194 Normalised Generalised Gamma (NGG) priors are infinite dimensional objects that are
 195 obtained by normalising a generalised gamma process. Concrete implementation of NGG
 196 priors requires to truncate the random series (1) at some level denoted Q , which results in
 197 some truncation error. Previous implementation of the package used to appeal to a relative
 198 error index, that we will denote $e_Q = \sum_{i>Q} J_i \delta_{\theta_i}$, based on the jumps themselves. We
 199 improve on this approach, by implementing the methodology proposed by [Arbel & Prünster](#)
 200 (2017) which relies on a moment-based evaluation of the error, denoted by ℓ_M . One of the
 201 main contributions of [Arbel & Prünster \(2017\)](#) is to warn that relying on the relative error
 202 index e_M can lead to overly optimistic conclusions in terms of approximation, especially for
 203 large values of the discount parameter γ .

204 To be more specific, consider K moments of the total mass of the CRM $\tilde{\rho}(\mathbb{X}) =$
 205 $\sum_{i=1}^{\infty} J_i$, denoted by $\mathbf{m}_K = (m_1, \dots, m_K)^\top$. Such moments have a simple expression in
 206 terms of the cumulants, which are themselves available in closed form, see for instance Table
 207 1 in [Arbel & Prünster \(2017\)](#). Thus, these exact moments can be computed and compared
 208 with their empirical counterparts obtained with the Ferguson & Klass algorithm ([Ferguson &](#)
 209 [Klass 1972](#)).

210 In order to make this methodology applicable, one needs to propose the truncation level
 211 $Q(\ell)$ required to achieve a given approximation ℓ . Such map $Q(\ell)$ only depends on the NGG
 212 parameters and can be computed once-for-all and distributed with the package. For reference,
 213 see the moment matching error $\ell(Q)$ and the map $Q(\ell)$ respectively displayed in Figures 1

214 and 2 of [Arbel & Prünster \(2017\)](#). Ferguson and Klass posterior sampling based on such a
 215 prescribed number of jumps $Q(\ell)$ is computationally more efficient than having to iteratively
 216 compute the relative error e_Q as done in the previous package version.

217 2.3. Clustering estimation

218 We focus here on the problem of estimating a data clustering from the Bayesian posterior
 219 inference conducted so far. This is a long standing problem in Bayesian statistics (see for
 220 instance [Dahl 2006](#); [Lau & Green 2007](#)). Enumerating all partitions is practically not feasible,
 221 which typically requires resorting to approximations.

Many ad-hoc procedures have been devised in the literature. However, as noted by [Dahl \(2006\)](#), it seems counter-intuitive to apply an ad-hoc clustering method on top of a model which itself produces clusterings. We adopt instead a fully Bayesian route by undertaking clustering on decision-theoretic grounds. We consider a loss function L and propose a Bayesian point estimator \hat{c} for a clustering obtained as an argument which minimises the posterior expected loss given data \mathbf{X}

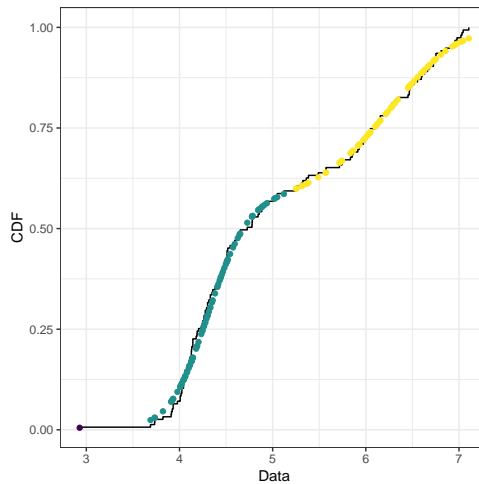
$$\hat{c} = \arg \min_{c'} \sum_c L(c', c) \pi(c | \mathbf{X}), \quad (11)$$

222 where $\pi(c | \mathbf{X})$ is the posterior distribution of clustering c . Often considered in the literature,
 223 the posterior mode is an example of such a Bayesian estimator, based on the very crude 0-1
 224 loss function. When n is large, an MCMC sample from the posterior generally hardly visits
 225 twice the same clustering, thus rendering the empirical mode of the MCMC output very
 226 sensitive to the initialisation of the chain and of very limited validity in practice. Manifestly,
 227 many other loss functions can be considered and expected to perform better than the 0-1
 228 loss. One particular choice of a loss function stands out from these in best estimating the
 229 number of groups in a clustering. It is called the variation of information, denoted by \mathcal{VI} ,
 230 which is a loss function firmly established in information theory ([Meila 2007](#); [Wade &](#)
 231 [Ghahramani 2018](#)). The variation of information between two clusterings is defined as the
 232 sum of their information (their Shannon entropies) minus twice the information they share.
 233 Simulations indicate that the variation of information is a sensible choice: when other losses
 234 such as the Binder loss ([Binder 1978](#)) typically tend to overestimate the number of clusters,
 235 the variation of information instead seems to consistently recover it (see for instance the
 236 simulated examples, and more specifically Figures 6 to 8, of [Wade & Ghahramani 2018](#)).

237 An asset of the approach presented in [Wade & Ghahramani \(2018\)](#) is that it rests on a
 238 greedy search algorithm to determine the minimum loss clustering of (11). Starting from the
 239 MCMC output, this greedy approach explores the space of partitions and is not restricted

240 to those visited by the MCMC chain to find the optimum. We include the possibility to
 241 estimate the optimal clustering using both the \mathcal{VZ} loss and Binder's loss, along with other loss
 242 functions, within `BNPdensity` by adding an optional dependence to `GreedyEPL`. Note that
 243 clustering estimation is also available for censored data, although graphical representation is
 244 more tricky (see also the legend to Figure 8).

```
245 data(acidity)
    out <- MixNRMI2(acidity)
    clustering = compute_optimal_clustering(out)
    plot_clustering_and_CDF(out, clustering)
```



246

Figure 1. Visualisation of the clustering induced by the BNP mixture model, for the `acidity` dataset. The solid line represents the empirical Cumulative Distribution Function (CDF), dots represent data points. The abscissa of each point is its value, the ordinate is the value of the estimated CDF at that point. Each colour denotes the cluster estimated by minimising the \mathcal{VZ} loss function.

247

248

3. Package description

The implementation of `BNPdensity` package is available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=BNPdensity>. Fitting a model with `BNPdensity` starts with calling one of the two functions, `MixNRMI1` or `MixNRMI2`, or their versions for censored data. The function `MixNRMI1` fits a semiparametric mixture model where all components have a common scale parameter σ with an independent parametric prior, $\sigma \sim P_\sigma$, while `MixNRMI2` is devoted to fully

nonparametric mixtures of *location and scale parameters*:

$$\begin{aligned} X_i \mid \theta_i, \sigma_i &\stackrel{\text{ind}}{\sim} k(\cdot \mid \theta_i, \sigma_i), \quad i = 1, \dots, n, \\ (\theta_i, \sigma_i) \mid \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P}, \quad i = 1, \dots, n, \\ \tilde{P} &\sim \text{NGG}(\alpha, \kappa, \gamma; P_0). \end{aligned}$$

249 Data and prior parameters are passed to the model function as arguments. The
 250 `MixNRMIX` functions also take a number of arguments to choose the BNP model, the mixture
 251 kernels, a variety of priors and tuning parameters for the Markov chain Monte Carlo sampling
 252 algorithm. The main arguments of the model functions are presented below.

- 253 • `distr.k`: Integer number identifying the **mixture kernel** k . Five kernels
 254 parameterised by their location and scale are implemented: a Gaussian or double
 255 exponential kernel for real data, a gamma or lognormal kernel for positive data and
 256 a beta kernel for data on the unit interval. The flexibility of this choice is afforded by
 257 the specific algorithm used in `BNPdensity`.
- 258 • `distr.py0`: Integer number identifying the base measure P_0 on the location
 259 parameters. Three choices are available, which are constrained by the conjugate
 260 prior we place on the hyperparameters of P_0 : Gaussian, gamma and beta. Additional
 261 arguments can be used to tune the shape of the base measure.
- 262 • `distr.py0, distr.pz0`: Integer number identifying the base measure P_0 on
 263 scale parameters. For the semiparametric model (`MixNRMIX1`), this argument is not
 264 provided and the base measure is a gamma distribution on the common scale parameter.
 265 Traditionally, there is sufficient information in the data to estimate the common scale
 266 parameter and inference is not very sensitive to the shape of the base measure. For
 267 the fully nonparametric model, the base measure on the scale parameters can be a
 268 gamma, lognormal, half Cauchy, half normal, half Student-t, uniform or truncated
 269 normal distribution. Additional arguments can be used to tune the shape of the base
 270 measure.
- 271 • `(Alpha, Kappa, Gama)`: Mixing measure parameters identifying a **Normalised**
 272 **generalised gamma** process, see the Lévy intensity (2) with parameters (α, κ, γ) for
 273 more details.
- 274 • The rest of the parameters provide handles to tune the MCMC algorithm.

275 Functions to fit a model return an object with `print`, `summary` and `plot` methods, as
 276 follows (the latter plot is represented in Figure 2):

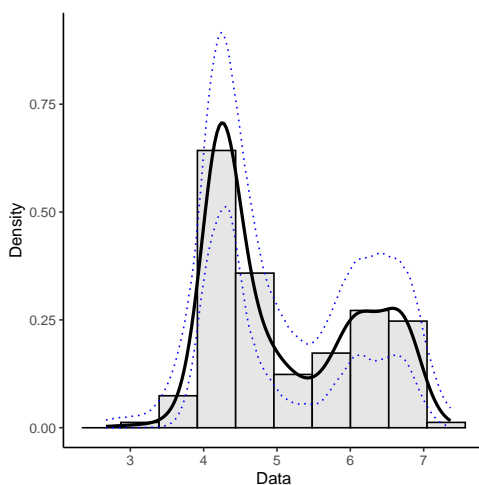


Figure 2. Density estimate (solid black line), 95% credible interval (blue dotted line) and histogram of the `acidity` data fitted with a semiparametric model. Figure obtained using the command `plot(out)`.

```

data(acidity)
out <- MixNRMI1(acidity)
## MCMC iteration 500 of 1500
## MCMC iteration 1000 of 1500
## MCMC iteration 1500 of 1500
## >>> Total processing time (sec.):
##   user  system elapsed
##  49.166   0.083   49.255

summary(out)
## Density estimation using a Normalized stable process,
## with stability parameter Gamma = 0.4
##
## A semiparametric normal mixture model was used.
##
## There were 155 data points.
##
## The MCMC algorithm was run for 1500 iterations with 10%
##
## To obtain information on the estimated number of clusters,
## please use summary(object, number_of_clusters = TRUE).

```

277

278

4. Package comparison

279

280

281

282

In this section, we discuss in detail the features and functionalities offered in three R packages addressing BNP density estimation, namely: `BNPdensity`, `BNPmix` (Canale, Corradin & Nipoti 2019), and `DPpackage` (Jara et al. 2011) (`DPpackage` was removed from the CRAN repository, but former versions are available at <https://cran.r-project.org/web/packages/BNPdensity/index.html>).

283 r-project.org/src/contrib/Archive/DPpackage/). Since the focus of the
284 present paper is mixture modeling and density estimation, note that other packages relying on
285 BNP approaches but tackling other questions such as regression (PReMiuM, [Liverani et al.](#)
286 [2015](#), Bayesian Regression, [Karabatsos 2017](#)), or meta-analysis (bspmma, [Burr 2012](#))
287 are not discussed here. Likewise, non Bayesian approaches are deliberately set aside. Table 1
288 summarises the comparative study of this section.

289 4.1. Inference algorithm

290 Efficient posterior computation for BNP mixture models relies on two types of
291 approaches: marginal or conditional. Marginal methods incorporate analytic integration of
292 infinite dimensional parts of the parameter, which is the case of DPpackage and BNPmix.
293 Instead, BNPdensity relies on a conditional sampler that directly samples trajectories of the
294 processes. More specifically, the Ferguson & Klass algorithm is employed (see Section 2.1),
295 with the crucial merit of ensuring that largest weights in the series representation are not left
296 out. This is to be compared to the stick-breaking representation where the weights sequence
297 is decreasing only stochastically (that is, in expectation).

298 4.2. Mixing measure

299 As described in Section 2, BNP mixture modeling and density estimation require to
300 specify some mixing measure. We start here by comparing the mixing measures available in
301 the three packages.

302 BNPmix provides a set of functions for density estimation with Dirichlet process
303 and Pitman–Yor mixing measures via marginal algorithms. DPpackage is a more general
304 purpose package than both BNPdensity and BNPmix, including functions for regression
305 models, generalised linear mixed models, and generalised additive models, on top of the
306 density model. However, the implementation is primarily tailored to the Dirichlet process
307 mixing measure. A natural extension to the Dirichlet and Pitman–Yor processes are Gibbs-
308 type priors ([De Blasi et al. 2015](#)). NRMI are a larger class of priors than Gibbs-type priors,
309 and their intersection is the NGG priors considered in BNPdensity, as established in [Lijoi,](#)
310 [Prünster & Walker \(2008\)](#). Being an extremely general class of priors, Gibbs-type processes
311 are beyond reach for a general treatment in a software, however both BNPdensity and
312 BNPmix packages cover its most commonly used sub-classes. Pitman–Yor process is not
313 implemented in BNPdensity as it is not an NRMI; yet, a dependence to BNPmix is made in
314 BNPdensity, in such a way that users interested in comparing their results with Pitman–
315 Yor can also use the dedicated functions `MixPY1` (semiparametric) and `MixPY2` (fully

316 nonparametric) that call `BNPmix` `PYdensity` function. The mixing measures covered by
 317 the three packages and their mutual relationships are illustrated in Figure 3.

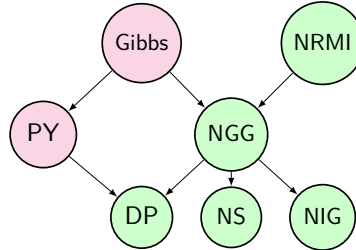


Figure 3. BNP priors mentioned in this section. An arrow indicates that the target is a special case or a limit case of its origin. Gibbs: Gibbs-type process. NRMI: normalised random measures with independent increments. NGG: normalised generalised gamma process. PY: Pitman–Yor process. NIG: normalised inverse Gaussian process. NS: normalised stable process. DP: Dirichlet process. In green: covered by `BNPdensity` package.

318

319 4.3. Prior characteristics

320 4.3.1. Non-conjugacy

321 Mixture models present the difficulty that the likelihood goes to infinity for infinitely
 322 small clusters located exactly on one observed data point. This may induce numerical
 323 problems and instabilities, and such tiny clusters are almost invariably undesirable in practical
 324 applications. A reasonable solution in the Bayesian framework is to use a prior distribution
 325 on scale parameters with little mass on very small values, i.e. a gamma distribution with
 326 shape parameter larger than 1 or a truncated distribution. We might also want to provide a
 327 different kind of information on cluster scales: for instance, for a dataset whose variance has
 328 been scaled to 1, there is no reason to find clusters with a variance much larger than one. This
 329 would suggest using a prior with an upper bound, or with light tails for large values. Finally,
 330 flexibility in the choice of the kernel k is a clear asset when modelling real data, to choose a
 331 reasonable error model. These three examples suggest that we might need a certain flexibility
 332 in the specification of the prior distribution on scale parameters or in the choice of the kernel.

333 The inference algorithm used in `BNPdensity` and presented in Section 2.1 does not
 334 rely on conjugacy between the base measure and the kernel of the mixture, as do standard
 335 algorithms for sampling from a Dirichlet mixture process such as that presented in Escobar

336 & West (1995). In contrast, `DPpackage` and `BNPmix` are limited to using conjugate couples
 337 of base measure and the mixture kernel.

338 Not being bounded to conjugacy allows us first to use any relevant kernel for the mixture.
 339 Moreover, even in the case of the normal kernel, this removes the dependence imposed in
 340 the conjugate case between the location of the clusters and their variances. More precisely,
 341 this allows a full flexibility on specifying priors based on external knowledge, and proves
 342 particularly useful concerning the scale parameters of the kernels. Indeed, half-Cauchy or
 343 half-Gaussian priors for hierarchical variance parameters have recently become popular
 344 Gelman (2006); Chung et al. (2015). The illustration on Species Sensitivity Distribution
 345 (SSD) (Section 5), where the data is scaled, offers such an example where both an upper
 346 bound and lower bound on the cluster variances are useful.

347 4.3.2. Prior distribution on number of components

348 Prior elicitation is a delicate task in Bayesian modeling. `BNPdensity` provides some
 349 guidelines on how to choose parameters (`Alpha`, `Kappa`, `Gama`) with two functions,
 350 one for computing the prior expected number of components, and one for plotting this prior
 351 distribution. Comparable functionalities are offered in `BNPmix` and `DPpackage`.

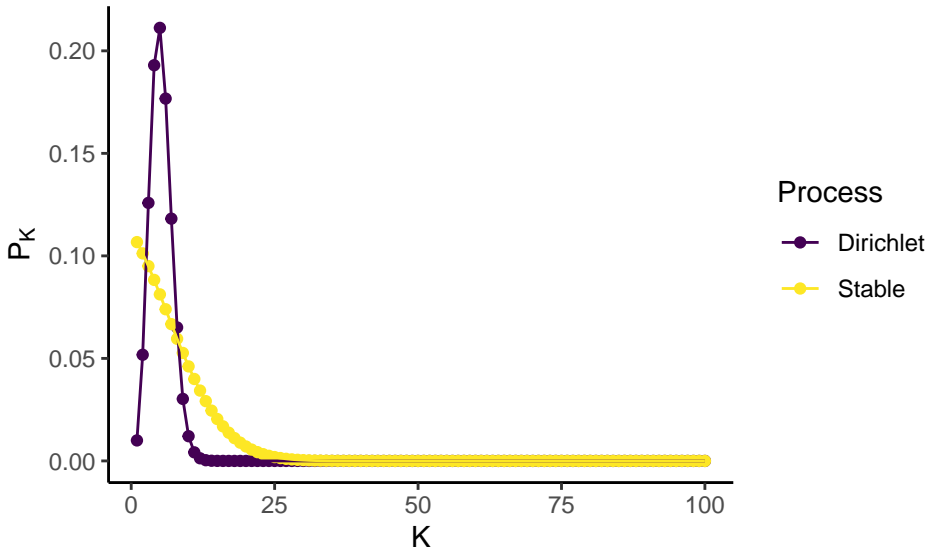
352 The (`Alpha`, `Kappa`, `Gama`) parametrisation allows to easily compare several well
 353 known priors. We already mentioned that the Dirichlet process can be obtained by setting
 354 `Gama = 0`, the normalised inverse Gaussian process by setting `Alpha = 1`, `Gama =`
 355 `1/2` and the normalised stable process by setting `Alpha = 1`, `Kappa = 0`. The stable
 356 process is a convenient model because its parameter γ has a simple interpretation: it can be
 357 used to tune how informative the prior on the number of components is. Small values of `Gama`
 358 bring the process closer to a Dirichlet process, where the prior on the number of components
 359 is a relatively peaked distribution around $\alpha \log n$. In contrast, the larger the value of `Gama`
 360 is, the flatter the distribution is. More guidelines on how to choose the parameters may be
 361 found in Lijoi, Mena & Prünster (2007b), notably by considering the expected prior number
 362 of components. The expected prior number of components for normalised generalised gamma
 363 processes is not trivial to compute due to numerical instabilities, but we provide functions to
 364 compute prior distribution on the number of clusters for the normalised stable process and
 365 for the Dirichlet process. These functions require installing the packages `gmp` and `Rmpfr` for
 366 Multiple Precision Arithmetic, both available on CRAN.

```
Rmpfr::asNumeric(expected_number_of_components_stable(n = 100, Gama = 0.4))
## [1] 7.102731

expected_number_of_components_Dirichlet(n = 100, Alpha = 1.)
## [1] 5.187378
```


367 We also provide a way to visualise the prior distribution on the number of components:

```
plot_prior_number_of_components(100, 0.4)
## Computing the prior probability on the number of clusters for the Dirichlet process
## Computing the prior probability on the number of clusters for the Stable process
```



368

369 Figure 4. Prior distribution on the number of clusters with 100 data points, for the stable process with $\gamma = 0.4$ and for the Dirichlet process with $\alpha = 1$.

370 4.4. Censored data

371 BNPdensity can deal with left, right and interval-censored data by using the functions
 372 `MixNRM11cens` and `MixNRM12cens`. The same holds true for `DPpackage`, while
 373 `BNPmix` does not handle censored data at all.

374 Censored data usually emerge from imperfections of the measurement process, such as
 375 detection limits (high or low) or saturation, low measurement precision, or binning of the
 376 data. Improper treatment of censored data is clearly a source of bias (Helsel 2005): in the
 377 case of right-censored data due to a detection limit for high values, for instance, data are not
 378 censored at random and discarding them or substituting them deteriorates the dataset.

379 We deal with censored data by using a version of the likelihood (Helsel 2005) adapted to
 380 censored data. More specifically, denote by F_k the cumulative distribution function of the
 381 kernel k . The heart of the method is then to replace $k(x | \theta)$ by $F_k(x | \theta)$ for a left-censored
 382 observation, by $1 - F_k(x | \theta)$ for a right-censored observation, and by $F_k(x_r | \theta) - F_k(x_l |$
 383 $\theta)$ for an interval-censored observation $[x_l, x_r]$.

384 4.5. Visualisation and programming

385 4.5.1. Convergence checking and model evaluation

386 BNPdensity offers several tools for assessing MCMC convergence and performing model
 387 checking and comparison. Notably, we provide a conversion function `as.mcmc` to interface
 388 the package with the `coda` package for analysing output and carrying out diagnostics on
 389 MCMC. We are not aware of such tools for `BNPmix` or `DPpackage`.

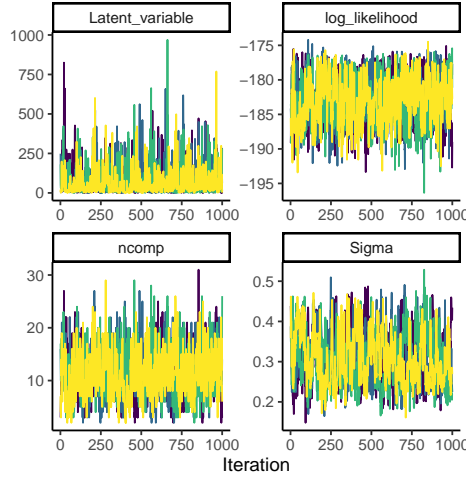
390 This is done by running multiple chains starting from different initial conditions, potentially
 391 in parallel, and converting them into an `mcmc` object that can be processed by `coda`. A simple
 392 solution for running multiple chains does not seem available for `BNPmix` and `DPpackage`.

393 One conceptual detail for assessing convergence is that, due to the nonparametric nature
 394 of the model, the number of parameters which could potentially be monitored to measure
 395 auto-correlation of the chains or effective sample size varies. The location parameters of the
 396 clusters, for instance, vary at each iteration, and even the labels of the clusters vary, which
 397 makes it tricky to follow. However, it is possible to monitor the log-likelihood of the data
 398 along the iterations, the value of the latent variable u , the number of components and for the
 399 semi-parametric model, the value of the common scale parameter. The following code shows
 400 how to compute the potential scale reduction factor (Gelman & Rubin 1992):

```
library(coda)
data(acidity)
fit = multMixNRM1(acidity, extras = TRUE, Nit = 20000)
mcmc_list = as.mcmc(fit)
gelman.diag(mcmc_list)
```

```
## Potential scale reduction factors:
##
##           Point est. Upper C.I.
## ncomp           1.02      1.06
## Sigma           1.02      1.07
## Latent_variable  1.02      1.05
## log_likelihood   1.01      1.04
##
## Multivariate psrf
##
## 1.03
```

401 A trace plot for the chains may also be obtained by calling `traceplot(fit)`; see Figure 5.



402

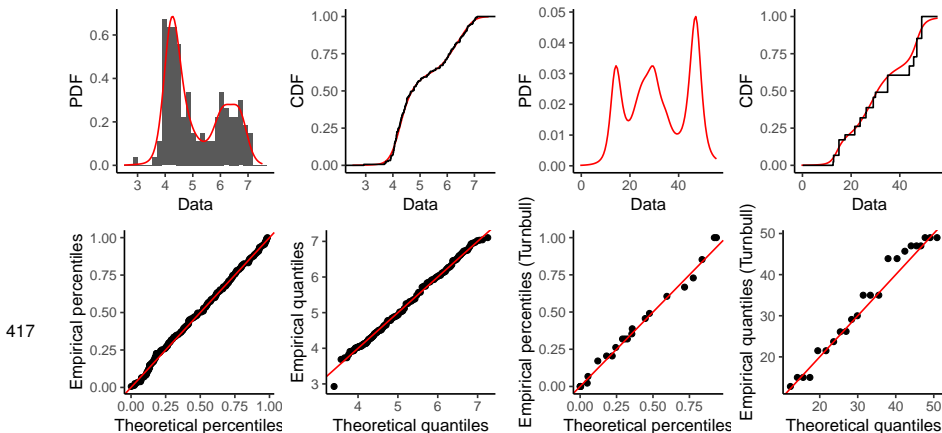
403

Figure 5. Trace plot of four chains in the MCMC for a semi-parametric model.

Table 1. Comparison of R packages performing BNP density estimation: `BNPdensity`, `BNPmix`, and `DPpackage`. (a) See discussion in Section 4.2. (b) The `DPpackage` `LDPDdoublyint` function, for *Linear Dependent Poisson Dirichlet Process Mixture Models for the Analysis of Doubly-Interval-Censored Data* could in principle be used for Pitman–Yor process mixture density estimation, although the interface (and the name) suggests it is not intended for this.

		BNPdensity	BNPmix	DPpackage
4.1 Inference algorithm	Conditional	yes	no	no
	Marginal	no	yes	yes
4.2 Mixing measure	Dirichlet process (DP)	yes	yes	yes
	Norm. inverse Gaussian (NIG)	yes	no	no
	Norm. stable (NS)	yes	no	no
	Norm. gener. gamma (NGG)	yes	no	no
	Pitman–Yor (PY)	no ^(a)	yes	no ^(b)
4.3 Prior characteristics	Non Gaussian kernels allowed	yes	no	no
	Functions for prior elicitation	yes	yes	yes
4.4 Data	All types of censored data	yes	no	yes
4.5 Vis. & Programming	MCMC conv. assessm.	yes	no	no
	Graphical model checking	yes	no	no
	Clustering vis. tools	yes	no	no
	Parallel computing	yes	no	no

404 We also provide tools for assessing goodness of fit. Graphical assessment can be performed
 405 comparing various representations of the estimated distributions against representations
 406 of the empirical distribution (Figure 6). Such plots may be obtained from a fitted
 407 object using the command `GOFplots(fit, qq_plot = TRUE)`. The density plot
 408 provides a familiar representation of the Nonparametric distribution, while the CDF plot
 409 is probably the most classical visualisation of goodness of fit. The percentile-percentile
 410 plot focuses on the goodness of fit in the center of the distribution, while the quantile-
 411 quantile plot focuses on the goodness of fit in the tails of the distribution. The density,
 412 CDF, percentile and quantiles used in the plots are the expected posterior quantities,
 413 computed from the MCMC sample. Computation of the theoretical quantiles is a fairly
 414 expensive operation because it requires numerically inverting the CDF. We choose not to
 415 compute the quantile-quantile plot by default, and when we do, the computation is done
 416 on a thinned MCMC chain with an argument provided to control the level of thinning.



417 Figure 6. Graphical goodness of fit plots for censored (right) and non censored data (left). The top row is the mean density estimate with a histogram for the non censored data. The middle row is the estimated CDF with the empirical CDF for non censored data, and with the Turnbull estimate of the CDF for censored data. The bottom row are percentile-percentile plots where the empirical percentiles are computed from the empirical CDF for the non censored data, and from the Turnbull estimate for the censored data.

418 We also provide tools for model comparison based on expected predictive density. The
 419 conditional predictive ordinate (CPO) is the expected predictive density of a data point given
 420 the prior and all other data points, so it is the leave-one-out expected predictive density of
 421 the model (Gelman et al. 2014), a typical cross-validation criterion. As such, it is a measure
 422 of predictive power with a penalisation for over-fitting. A Monte Carlo approximation of the
 423 CPO is easily available and can be used to compare a semi-parametric model to the fully
 424 nonparametric model for instance:

```

set.seed(0)
normal_mixture <- MixNRMI2(acidity, distr.k = 1, Nit = 15000)
dbl_exponential_mixture <- MixNRMI1(acidity, distr.k = 4, Nit = 15000)
c(median(normal_mixture$cpo), median(dbl_exponential_mixture$cpo))

```

```
## [1] 0.279 0.271
```

Model	Mean CPO	Median CPO
Nonparametric normal mixture	0.362	0.279
Semi parametric double exponential mixture	0.357	0.271

4.5.2. Clustering visualisation tools

As described in Section 2.3, BNPdensity provides functions for clustering estimation, `compute_optimal_clustering`, and visual representation, `plot_clustering_and_CDF`. See also Figure 1 and Figure 8 for illustrations. We are not aware of such clustering tools for BNPmix or DPpackage.

5. Case study: Species Sensitivity Distribution

We present an application of nonparametric density estimation for environmental data. Assessing the response of a community of species to an environmental stress is of critical importance for ecological risk assessment. Methods for this purpose vary in levels of complexity and realism. SSD represents an intermediate tier, more refined than rudimentary assessment factors (Posthuma, Suter II & Trass 2002) but practical enough for routine use by environmental managers and regulators in most developed countries (Australia, Canada, China, EU, South Africa, USA, ...). The SSD approach is intended to provide, for a given contaminant, a description of the tolerance of all species possibly exposed using information collected on a sample of those species. This information consists of a single species-specific value, which marks a limit over which the species suffers adverse effects. This value is very often censored (Kon Kam King et al. 2014), because measuring it is both costly and difficult (bioassay experiments). The tolerance of all species possibly exposed is described by a distribution, fitted on the sample of species (Aldenberg & Jaworska 2000). The quantity of interest for ecological risk assessment is the Hazardous Concentration for 5% of the Species (HC_5), which corresponds to the 5th percentile of the SSD distribution. The lack of justification for the choice of any given parametric distribution has sparked several research directions. Some authors (Xu et al. 2015; He et al. 2014; Jagoe & Newman 1997; Van Straalen 2002; Xing et al. 2014; Zhao & Chen 2016) have sought to find the best parametric distribution by model comparison using goodness-of-fit measures. The general understanding is that no single distribution seems to provide a superior fit and that the answer is dataset dependent (Forbes & Calow 2002). Therefore, the log-normal distribution has become the

453 customary choice, notably because it readily provides confidence intervals on the HC_5 , and
454 because model comparison and goodness of fit tests have relatively low power on small
455 datasets, precluding the emergence of a definite answer to the question.

456 The availability of a package such as `BNPdensity` allows to move beyond this customary
457 assumption very easily. NRMIs offer a flexible nonparametric mixture model, which can
458 accommodate distributions very different from a normal distribution. [Barrios et al. \(2013\)](#)
459 and [Kon Kam King, Arbel & Prünster \(2017\)](#) show that NRMIs have better performance than
460 Dirichlet process mixtures, kernel density estimates (the recent approach proposed by [Wang
461 et al. \(2015\)](#)) or simple one-component normal models. Moreover, there are good reasons to
462 believe that the distribution of species sensibility should at least allow for multimodality.
463 Indeed, many stressors target specifically certain species groups, such as insecticides for
464 insects, while they are developed with the aim of leaving other species group unaffected.
465 Therefore, it is expected that there should at the very least be a group of sensitive species and
466 a group of less sensitive species. This is why [Zajdlik, Dixon & Stephenson \(2009\)](#) propose
467 to model the species sensitivity distribution as a finite mixture, with raises customary issues
468 of model choice. Using a BNP approach via `BNPdensity` allows generalising this approach
469 while circumventing the theoretical and technical difficulties of estimating the right number
470 of components in a mixture.

471 It is also important to use a method which may be applied to small datasets. This is another
472 motivation for using a BNP approach, where model complexity adapts to the number of data
473 points, and will tend to suggest simple or even univariate mixtures when few data points
474 are present. On the contrary, many classical nonparametric approaches to modelling species
475 sensitivity distribution ([Wang et al. 2015](#); [Verdonck et al. 2001](#)) only work well on large
476 datasets.

477 To model species sensitivity distribution, we carefully select the parameters in the package
478 `BNPdensity`. Given that concentrations vary on a wide range, it is common practice to
479 work on log-transformed concentrations. We choose a fully nonparametric model using
480 the normalised stable process ([Kingman 1975](#)) as mixing random measure (hence setting
481 $\text{Alpha} = 1$ and $\text{Beta} = 0$). We favor this process over the more classical Dirichlet process
482 because it allows specifying less informative prior on the number of components, which
483 makes it more robust to model misspecification ([Barrios et al. 2013](#)). With this process,
484 the amount of information from the prior is controlled by the stability parameter γ , which
485 we set to 0.4 ($\text{Gama} = 0.4$). This choice reflects a compromise between model flexibility
486 ($\gamma \rightarrow 1$) and computational effort (γ small, see also section 3). As we wish the location
487 parameter of the clusters μ to be estimated freely, we use the default weakly informative
488 prior of a normal base measure $f_0^1(\mu|\varphi) = \mathcal{N}(\mu|\varphi_1, \varphi_2)$ with hyperpriors on φ given by
489 $f(\varphi) = \mathcal{N}(\varphi_1|\psi_1, \psi_2)\text{ga}(\varphi_2|\psi_3, \psi_4)$ (see also [Barrios et al. \(2013\)](#) for more details).

490 For the prior on the scale of the clusters, we want to use two pieces of information: first,
491 since the data has been scaled, scale parameters are likely to be smaller than 1, the extreme
492 case being a mixture with a single component. Second, we want to avoid the possibility of
493 extremely small clusters centred on a data point, because they are not very interesting from
494 an interpretation point of view, and because they cause numerical problems (the likelihood
495 diverges when a cluster scale goes to 0). Therefore, we choose a uniform distribution between
496 0.1 and 1.5 for the prior on the cluster scales.

497 In keeping with the traditional assumption of normality of the species sensitivity distribution,
498 we choose to use a normal kernel for the mixture (`distr.k = 1`).

499 We now compare three approaches to modelling Species Sensitivity Distribution (SSD):
500 the most standard and recommended approach of [Wagner & Lokke \(1991\)](#); [Aldenberg &](#)
501 [Jaworska \(2000\)](#), which is a simple normal model, the most recent proposal by ([Wang et al.](#)
502 [2015](#)) which is a normal kernel density estimate and the BNP normal mixture made available
503 with `BNPdensity` that we presented above. As already stated, a quantity of interest is the
504 5th percentile of the distribution. We choose as an estimator the median of the posterior
505 distribution of the 5th percentile, while the 95% credible bands are formed by the 2.5%
506 and 97.5% quantiles of the posterior distribution of the 5th percentile. The 5th percentile
507 of the Kernel Density Estimate (KDE) is obtained by numerical inversion of the cumulative
508 distribution function, and the confidence intervals using the nonparametric bootstrap. The 5th
509 percentile of the normal SSD and its confidence intervals are obtained following the classical
510 method of [Aldenberg & Jaworska \(2000\)](#).

511 We use data from an ecotoxicity research database as pre-processed in [Hickey et al. \(2012\)](#).
512 We extract data for the insecticide Carbaryl. The dataset contains 57 species, of which
513 approximately 40% have censored data. We obtain a non censored version of this dataset by
514 excluding right or left censored data, and replacing interval censored data by the midpoint of
515 the interval. [Helsel \(2006\)](#); [Dowse et al. \(2013\)](#); [Kon Kam King et al. \(2014\)](#) have shown
516 that transforming censored data risks inducing bias, hence the ability of `BNPdensity` to
517 accommodate censoring is particularly valuable for SSD. There does not appear to be any
518 easily available approach to use KDE methods on all types of censored data. Figure 7 shows
519 a comparison of three approaches to SSD. The left hand side of Figure 7 shows that the BNP
520 model is more flexible than both the KDE and normal model, while the right hand side shows
521 that it is no less robust, according to a leave-one-out cross validation criterion. The middle
522 panel shows that although the BNP model is more flexible and takes into account uncertainty
523 on the number of clusters, the estimation of the 5th percentile is not much more uncertain
524 than with the other methods. Significantly larger uncertainty would have jeopardised the real
525 world applicability of the BNP-SSD.

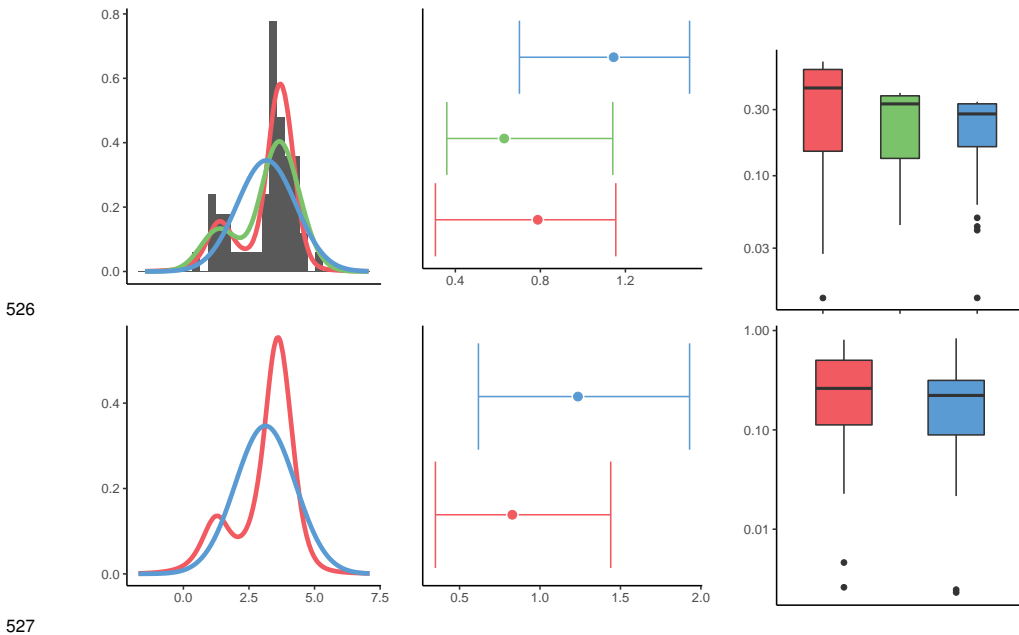
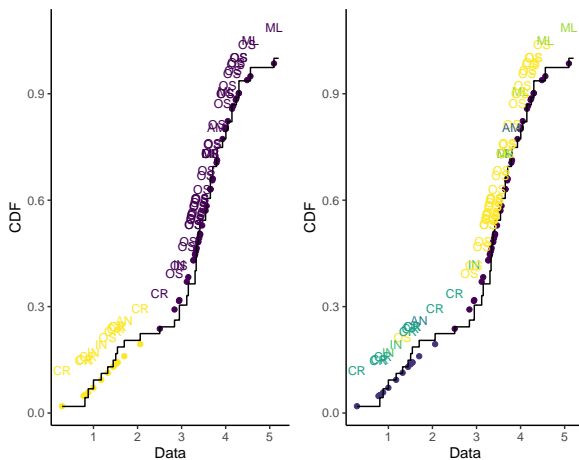


Figure 7. Top panel: non censored data. Bottom panel: censored data. The normal model is represented in blue, the KDE in green and the BNP in red. Left: density plot and histogram for the Carbaryl data using several SSD methods. The histogram is not available for censored data. Center: 5th percentile estimate (not available for KDE with censored data). Right: boxplot of the CPO (for BNP) and Leave-One-Out (LOO) (for normal and KDE, not available for KDE with censored data), one value for each data point.

528

529 An added value of the BNP-SSD is that on top of being more flexible than the classic normal
 530 SSD and more robust than the nonparametric approach of Wang et al. (2015), as a mixture
 531 model it naturally induces a clustering of the data which may contain some biologically
 532 interesting information. We implemented functions to estimate the optimal clustering from
 533 the MCMC sample and visualise it, potentially including a label on each point to reflect
 534 available meta data for interpretation. In the context of SSD, it is interesting to know what
 535 drives species sensitivity: it might be taxonomy, in the sense that taxonomically close species
 536 will tend to respond in the same way and belong to the same cluster, but other drivers have
 537 been suggested such as habitat, feeding behaviour or respiration, which may not coincide with
 538 taxonomy. Figure 8 shows the clustering induced in the case of the insecticide Carbaryl. In
 539 this case, there is a large cluster mostly composed of fish and molluscs, and a cluster mostly
 540 composed of insects and crustaceans, showing that the clustering structure is consistent with
 541 a finer taxonomic structure. This suggests that for Carbaryl, taxonomy may very well be the
 542 main driver for sensitivity.



543

Figure 8. Graphical representation of the clustering induced by the mixture model for the Carbaryl data. The solid line represents the Turnbull estimate of the CDF, the points loosely represent the data. Interval censored data are represented at the middle of the interval, left and right censored data are not represented. A label describing the taxonomic group of each species is written above each point, AM: Amphibians, AN: Annelids (worms), CR: Crustaceans, IN: Insects, ML: Molluscs, OS: Osteichthyes (fish). On the left panel, the points and the labels are coloured according to the estimated cluster index. On the right panel, the labels are coloured according to the taxonomic group and the points are not coloured.

544

545

Computational details

546 The results in this paper were obtained using R 4.1.1 with the BNPdensity package version
 547 2020.3.4. R itself and all packages used are available from the Comprehensive R Archive
 548 Network (CRAN) at <https://CRAN.R-project.org/>.

549

References

- 550 ALDENBERG, T. & JAWORSKA, J.S. (2000). Uncertainty of the hazardous concentration and fraction
551 affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety* **46**, 1–18.
552 doi:10.1006/eesa.1999.1869. URL <http://www.ncbi.nlm.nih.gov/pubmed/10805987>.
- 553 ARBEL, J. & PRÜNSTER, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and*
554 *Computing* **27**, 3–17. doi:10.1007/s11222-016-9676-8.
- 555 BARRIOS, E., LIJOI, A., NIETO-BARAJAS, L.E. & PRÜNSTER, I. (2013). Modeling with normalized
556 random measure mixture models. *Statistical Science* **28**, 313–334.
- 557 BINDER, D.A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.
- 558 BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied*
559 *Probability* **31**, 929–953.
- 560 BURR, D. (2012). bspmma: An R package for Bayesian semi-parametric models for metaanalysis. *Journal*
561 *of Statistical Software* **50**, 1–23.
- 562 BUSH, C.A. & MACEACHERN, S.N. (1996). A semiparametric Bayesian model for randomised block
563 designs. *Biometrika* **83**, 275–285.
- 564 CANALE, A., CORRADIN, R. & NIPOTI, B. (2019). BNPmix: an R package for Bayesian nonparametric
565 modelling via Pitman–Yor mixtures. *Journal of Statistical Software* , to appear.
- 566 CHUNG, Y., GELMAN, A.G., RABE-HESKETH, S., LIU, J. & DORIE, V. (2015). Weakly Informative
567 Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and*
568 *Behavioral Statistics* **40**, 136–157. doi:10.3102/1076998615570945. URL <http://jeb.sagepub.com.ezproxy.lancs.ac.uk/content/40/2/136>. arXiv:1011.1669v3.
- 570 DAHL, D.B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model.
571 *Bayesian inference for gene expression and proteomics* **4**, 201–218.
- 572 DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R.H., PRÜNSTER, I. & RUGGIERO, M. (2015).
573 Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE*
574 *Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229. doi:10.1109/
575 TPAMI.2013.217. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6654160>. arXiv:1503.00163v1.
- 577 DENWOOD, M.J. (2016). runjags: An R package providing interface utilities, model templates, parallel
578 computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical*
579 *Software* **71**, 1–25.
- 580 DOWSE, R., TANG, D., PALMER, C.G. & KEFFORD, B.J. (2013). Risk assessment using the species
581 sensitivity distribution method: Data quality versus data quantity. *Environmental Toxicology and*
582 *Chemistry* **32**, 1360–1369. doi:10.1002/etc.2190. URL <http://www.ncbi.nlm.nih.gov/pubmed/23440771>.
- 584 ESCOBAR, M.D. & WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal*
585 *of the American Statistical Association* **90**, 577–588. doi:10.1080/01621459.1995.10476550. URL
586 <https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550>.
- 587 FERGUSON, T.S. & KLASS, M.J. (1972). A representation of independent increment processes without
588 Gaussian components. *Ann. Math. Stat.* **43**, 1634–1643.
- 589 FORBES, V.E. & CALOW, P. (2002). Species Sensitivity Distributions Revisited: A Critical Appraisal.
590 *Human and Ecological Risk Assessment* **8**, 473–492. doi:10.1080/20028091057033. URL <http://www.tandfonline.com/doi/abs/10.1080/10807030290879781>.
- 592 FRÜHWIRTH-SCHNATTER, S., CELEUX, G. & ROBERT, C.P. (2018). *Handbook of Mixture Analysis*.
593 Chapman & Hall/CRC.
- 594 GELMAN, A.G. (2006). Prior distributions for variance parameters in hierarchical models (Comment on
595 Article by Browne and Draper). *Bayesian Analysis* **1**, 515–534. doi:10.1214/06-BA117A.

- 596 GELMAN, A.G., CARLIN, J.B., STERN, H.S. & RUBIN, D.B. (2014). *Bayesian Data Analysis*. Boca Raton,
597 FL: CRC press, 3rd edn.
- 598 GELMAN, A.G. & RUBIN, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences.
599 *Statistical Science* **7**, 457–511. doi:10.1214/ss/1177011136.
- 600 GILKS, W.R., THOMAS, A. & SPIEGELHALTER, D.J. (1993). A Language and program for complex
601 bayesian modelling. *Journal of the Royal Statistical Society. Series D (The Statistician)* **43**, 169–177.
602 doi:Doi10.2307/2348941.
- 603 HE, W., QIN, N., KONG, X., LIU, W., WU, W., HE, Q., YANG, C., JIANG, Y., WANG, Q., YANG,
604 B. & XU, F. (2014). Ecological risk assessment and priority setting for typical toxic pollutants in
605 the water from Beijing-Tianjin-Bohai area using Bayesian matbugs calculator (BMC). *Ecological*
606 *Indicators* **45**, 209–218. doi:10.1016/j.ecolind.2014.04.008. URL [http://dx.doi.org/10.](http://dx.doi.org/10.1016/j.ecolind.2014.04.008)
607 [1016/j.ecolind.2014.04.008](http://dx.doi.org/10.1016/j.ecolind.2014.04.008).
- 608 HELSEL, D.R. (2005). *Nondetects and data analysis. Statistics for censored environmental data*. Wiley-
609 Interscience. doi:10.2136/vzj2005.0106br.
- 610 HELSEL, D.R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what
611 can be done about it. *Chemosphere* **65**, 2434–2439.
- 612 HICKEY, G.L., CRAIG, P.S., LUTTIK, R. & DE ZWART, D. (2012). On the quantification of interest
613 variability in ecotoxicity data with application to species sensitivity distributions. *Environmental*
614 *Toxicology and Chemistry* **31**, 1903–1910. doi:10.1002/etc.1891. URL [http://www.ncbi.nlm.](http://www.ncbi.nlm.nih.gov/pubmed/22619109)
615 [nih.gov/pubmed/22619109](http://www.ncbi.nlm.nih.gov/pubmed/22619109).
- 616 JAGOE, R.H. & NEWMAN, M.C. (1997). Bootstrap estimation of community NOEC values. *Ecotoxicology*
617 **6**, 293–306. doi:10.1023/A:1018639113818. URL [http://dx.doi.org/10.1023/A:](http://dx.doi.org/10.1023/A:1018639113818)
618 [1018639113818](http://dx.doi.org/10.1023/A:1018639113818).
- 619 JAMES, L.F., LIJOI, A. & PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with
620 independent increments. *Scandinavian Journal of Statistics* **36**, 76–97. doi:10.1111/j.1467-9469.2008.
621 00609.x.
- 622 JARA, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *R News* **7**, 17–26.
623 URL <https://CRAN.R-project.org/doc/Rnews/>.
- 624 JARA, A., HANSON, T.E., QUINTANA, F.A., MÜLLER, P. & ROSNER, G.L. (2011). DPpackage: Bayesian
625 non-and semi-parametric modelling in R. *Journal of statistical software* **40**, 1.
- 626 KARABATSOS, G. (2017). A menu-driven software package of bayesian nonparametric (and parametric)
627 mixed models for regression analysis and density estimation. *Behavior Research Methods* **49**, 335–
628 362. [1506.05435](https://doi.org/10.1037/xap0000114).
- 629 KINGMAN, J. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B* **37**,
630 1–15.
- 631 KON KAM KING, G., ARBEL, J. & PRÜNSTER, I. (2017). A Bayesian Nonparametric Approach to
632 Ecological Risk Assessment. In *Bayesian Statistics in Action: BAYSM 2016, Florence, Italy, June 19-*
633 *21*, eds. R. Argiento, E. Lanzarone, I. Antoniano Villalobos & A. Mattei. Cham: Springer International
634 Publishing, pp. 151–159. doi:10.1007/978-3-319-54084-9_14. URL [http://dx.doi.org/10.](http://dx.doi.org/10.1007/978-3-319-54084-9_14)
635 [1007/978-3-319-54084-9_{_}14](http://dx.doi.org/10.1007/978-3-319-54084-9_14).
- 636 KON KAM KING, G., VEBER, P., CHARLES, S. & DELIGNETTE-MULLER, M.L. (2014). MOSAIC.SSD:
637 A new web tool for species sensitivity distribution to include censored data by maximum likelihood.
638 *Environmental Toxicology and Chemistry* **33**, 2133–2139. doi:10.1002/etc.2644. URL [http:](http://www.ncbi.nlm.nih.gov/pubmed/24863265)
639 [//www.ncbi.nlm.nih.gov/pubmed/24863265](http://www.ncbi.nlm.nih.gov/pubmed/24863265).
- 640 LAU, J.W. & GREEN, P.J. (2007). Bayesian model-based clustering procedures. *Journal of Computational*
641 *and Graphical Statistics* **16**, 526–558.
- 642 LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2005). Hierarchical mixture modeling with normalized inverse-
643 Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291.

- 644 LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2007a). Bayesian nonparametric estimation of the probability of
645 discovering new species. *Biometrika* **94**, 769–786.
- 646 LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2007b). Controlling the reinforcement in Bayesian non-parametric
647 mixture models. *J. Roy. Stat. Soc. B Met.* **69**, 715–740.
- 648 LIJOI, A., PRÜNSTER, I. & WALKER, S.G. (2008). Investigating nonparametric priors with
649 Gibbs structure. *Statistica Sinica* **18**, 1653–1668. URL [http://www.ams.org/
650 mathscinet-getitem?mr=MR2469329{%}5Cnpapers2://publication/uuid/
651 1CD2CD58-C0D3-42E1-8129-24261868FAB1](http://www.ams.org/mathscinet-getitem?mr=MR2469329{%}5Cnpapers2://publication/uuid/1CD2CD58-C0D3-42E1-8129-24261868FAB1).
- 652 LIVERANI, S., HASTIE, D.I., AZIZI, L., PAPATHOMAS, M. & RICHARDSON, S. (2015). PReMiuM: An R
653 package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*
654 **64**, 1–30. URL <http://www.jstatsoft.org/v64/i07/>.
- 655 LO, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*
656 **12**, 351–357.
- 657 MACEACHERN, S.N. & MÜLLER, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal*
658 *of Computational and Graphical Statistics* **7**, 223–238. doi:10.1080/10618600.1998.10474772. URL
659 <http://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474772>.
- 660 MEILA, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*
661 **98**, 873–895.
- 662 NEAL, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of*
663 *computational and graphical statistics* **9**, 249–265.
- 664 PAPASPILIOPOULOS, O. & ROBERTS, G. (2008). Retrospective Markov chain Monte Carlo methods for
665 Dirichlet process hierarchical models. *Biometrika* **95**, 169.
- 666 PLUMMER, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling.
667 In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003),*
668 *March 20-22, Vienna, Austria. ISSN 1609-395X*, vol. 124. p. 125. doi:10.1.1.13.3406.
- 669 PLUMMER, M. (2019). *rjags: Bayesian Graphical Models using MCMC*. URL [https://CRAN.
670 R-project.org/package=rjags](https://CRAN.R-project.org/package=rjags). R package version 4-9.
- 671 POSTHUMA, L., SUTER II, G.W. & TRASS, P.T. (2002). *Species sensitivity dis-*
672 *tributions in ecotoxicology*. CRC press. URL [http://www.amazon.com/
673 Species-Sensitivity-Distributions-Ecotoxicology-Posthuma/dp/
674 1566705789](http://www.amazon.com/Species-Sensitivity-Distributions-Ecotoxicology-Posthuma/dp/1566705789).
- 675 RASTELLI, R. & FRIEL, N. (2018). Optimal Bayesian estimators for latent variable cluster models. *Statistics*
676 *and Computing* **28**, 1169–1186. doi:10.1007/s11222-017-9786-y. URL [https://doi.org/10.
677 1007/s11222-017-9786-y](https://doi.org/10.1007/s11222-017-9786-y).
- 678 RCORETEAM (2019). R: A Language and Environment for Statistical Computing. URL [http://www.
679 r-project.org/http://www.r-project.org](http://www.r-project.org/http://www.r-project.org).
- 680 REGAZZINI, E., LIJOI, A. & PRÜNSTER, I. (2003). Distributional results for means of normalized random
681 measures with independent increments. *Annals of Statistics* **31**, 560–585. doi:10.1214/aos/1051027881.
- 682 ROBERTS, G.O. & ROSENTHAL, J.S. (2009). Examples of adaptive mcmc. *Journal of Computational and*
683 *Graphical Statistics* **18**, 349–367.
- 684 SATO, K.I. (1999). *Lévy Processes and Infinitely Divisible Distributions, Cambridge Studies in Advanced*
685 *Mathematics*, vol. 68. Cambridge University Press.
- 686 SCRUCCA, L., FOP, M., MURPHY, T.B. & RAFTERY, A.E. (2016). mclust 5: clustering, classification and
687 density estimation using Gaussian finite mixture models. *The R Journal* **8**, 205–233. URL [https:
688 //journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf](https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf).
- 689 STAN DEVELOPMENT TEAM (2018). RStan: the R interface to Stan. URL <http://mc-stan.org/>. R
690 package version 2.18.2.
- 691 STAN DEVELOPMENT TEAM & STAN DEVELOPEMENT TEAM (2019). Stan: A C++ Library for Probability
692 and Sampling, Version 2.19. URL <http://mc-stan.org/>.

- 693 STURTZ, S., LIGGES, U. & GELMAN, A.E. (2005). R2WinBUGS: a package for running WinBUGS from
694 R. *Journal of Statistical Software* **12**, 1–16.
- 695 THOMAS, A., O’HARA, B., LIGGES, U. & STURTZ, S. (2006). Making BUGS open. *R News* **6**, 12–17.
- 696 TODESCHINI, A., CARON, F. & FUENTES, M. (2014). Rbiips: Bayesian inference with interacting particle
697 systems. *arXiv* URL <http://alea.bordeaux.inria.fr/biips>.
- 698 VAN STRAALLEN, N.M. (2002). Threshold models for species sensitivity distributions applied to aquatic
699 risk assessment for zinc. *Environmental Toxicology and Pharmacology* **11**, 167–172. doi:10.1016/
700 S1382-6689(01)00114-4. URL <http://www.ncbi.nlm.nih.gov/pubmed/21782599>.
- 701 VERDONCK, F.A.M., JAWORSKA, J., THAS, O. & VANROLLEGHEM, P.A. (2001). Determining
702 environmental standards using bootstrapping, Bayesian and maximum likelihood techniques: A
703 comparative study. *Analytica Chimica Acta* **446**, 429–438. doi:10.1016/S0003-2670(01)00938-2. URL
704 <http://linkinghub.elsevier.com/retrieve/pii/S0003267001009382>.
- 705 WADE, S. & GHARAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with
706 discussion). *Bayesian Analysis* **13**, 559–626.
- 707 WAGNER, C. & LOKKE, H. (1991). Estimation of ecotoxicological protection levels from NOEC toxicity
708 data. *Water Research* **25**, 1237–1242. doi:10.1016/0043-1354(91)90062-U. URL <http://www.sciencedirect.com/science/article/pii/004313549190062U>.
- 710 WANG, Y., WU, F., GIESY, J.P., FENG, C., LIU, Y., QIN, N. & ZHAO, Y. (2015). Non-parametric kernel
711 density estimation of species sensitivity distributions in developing water quality criteria of metals.
712 *Environmental Science and Pollution Research* **22**, 13980–13989. doi:10.1007/s11356-015-4602-8.
713 URL <http://link.springer.com/10.1007/s11356-015-4602-8>.
- 714 XING, L., LIU, H., ZHANG, X., HECKER, M., GIESY, J.P. & YU, H. (2014). A comparison of statistical
715 methods for deriving freshwater quality criteria for the protection of aquatic organisms. *Environmental
716 Science and Pollution Research* **21**, 159–167. doi:10.1007/s11356-013-1462-y.
- 717 XU, F.L., LI, Y.L., WANG, Y., HE, W., KONG, X.Z., QIN, N., LIU, W.X., WU, W.J. & JORGENSEN,
718 S.E. (2015). Key issues for the development and application of the species sensitivity distribution
719 (SSD) model for ecological risk assessment. *Ecological Indicators* **54**, 227–237. doi:10.1016/j.
720 ecolind.2015.02.001. URL [http://www.sciencedirect.com/science/article/pii/
721 S1470160X15000692](http://www.sciencedirect.com/science/article/pii/S1470160X15000692).
- 722 ZAJDLIK, B.A., DIXON, D.G. & STEPHENSON, G. (2009). Estimating Water Quality Guidelines for
723 Environmental Contaminants Using Multimodal Species Sensitivity Distributions: A Case Study
724 with Atrazine. *Human and Ecological Risk Assessment* **15**, 554–564. URL <http://www.tandfonline.com/doi/abs/10.1080/10807030902892539>.
- 726 ZHAO, J. & CHEN, B. (2016). Species sensitivity distribution for chlorpyrifos to aquatic organisms:
727 Model choice and sample size. *Ecotoxicology and Environmental Safety* **125**, 161–9. doi:10.1016/
728 j.ecoenv.2015.11.039. URL [http://www.sciencedirect.com/science/article/pii/
729 S0147651315301883](http://www.sciencedirect.com/science/article/pii/S0147651315301883).