BNPdensity: Bayesian nonparametric mixture modelling in R

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

BNPdensity: Bayesian nonparametric mixture modeling in R

J. Arbel[1], G. Kon Kam King[2*], A. Lijoi[3], L. Nieto-Barajas[4] and I. Prünster[3]

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 2015SNS29B.

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

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

2. Bayesian nonparametric density estimation

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

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

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

$$\tilde{\rho} = \sum_{i=1}^{\infty} J_i \delta_{\boldsymbol{\theta}_i}, \tag{1}$$

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

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

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

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

$$\tilde{f}(x) = \int_{\Theta} k(x \mid \boldsymbol{\theta}) \tilde{P}(\mathrm{d}\boldsymbol{\theta}). \tag{3}$$

This can equivalently be written in a hierarchical form as

$$\begin{aligned} X_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} k(\cdot \mid \boldsymbol{\theta}_i), \quad i = 1, \ldots, n, \\ \boldsymbol{\theta}_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P}, \quad i = 1, \ldots, n, \\ \tilde{P} &\sim \mathrm{NGG}(\alpha, \kappa, \gamma; P_0). \end{aligned} \tag{4}$$

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

We denote by f_0 the density with respect to the Lebesgue measure of the NGG base measure P_0 on Θ. When P_0 depends on a further hyperparameter $\boldsymbol{\phi}$, we use the notation $f_0(\cdot \mid \boldsymbol{\phi})$. Using the `MixNRMI2` function corresponds to the specification of a nonparametric model for the location and scale parameters of the mixture where the mixture parameter $\boldsymbol{\theta}$ takes the form of the vector (μ, σ). In order to distinguish the hyperparameters for location and scale, we will use the notation $f_0(\mu, \sigma \mid \boldsymbol{\phi}) = f_0^1(\mu \mid \sigma, \varphi) f_0^2(\sigma \mid \varsigma)$. In applications a priori independence between μ and σ is commonly assumed, and this is indeed a natural assumption for the illustration in Section 5.

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

$$\hat{f}_n(x) = \mathrm{E}\big(\tilde{f}(x) \mid \mathbf{X}\big) \tag{5}$$

for any x in $\mathbb{X}$. As for the latter, if R_n is the number of distinct latent values $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{R_n}^*$ out of a sample of size n, one can deduce a partition of the observations such that any two X_i and X_j belong to the same cluster if the corresponding latent variables $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ coincide. Then, it is interesting to determine an estimate $\hat{R}_n$ of the number of clusters into which the data are grouped, along with the clustering structure. For details on clustering estimation in our setting, see Section 2.3.

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

2.1. Posterior sampling via a conditional Gibbs sampler

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

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

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

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

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

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

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

   Sampling U is performed via a Metropolis–Hastings (M-H) step with a gamma proposal distribution $\mathrm{ga}\left(\delta, \delta/u^{[t]}\right)$ centered at the previous U value $u^{[t]}$ with a tuning parameter δ controlling the coefficient of variation. An adaptive version of the M-H algorithm (Roberts & Rosenthal 2009) without the tuning parameter is also implemented in the package, and proposed with the option `adaptive=TRUE`. It uses a log-transformation of the random variable U. Note that the target density (6) not being log-concave, ergodicity cannot be proven as in Roberts & Rosenthal (2009). Nevertheless, the adaptive version appears to offer superior performance in practice.

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

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

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

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

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 jumps J_j^* at the fixed locations $\boldsymbol{\theta}_j^*$ are gamma distributed:

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

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

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

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

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

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

We also include a resampling of the unique values of the cluster parameters via a M-H step to avoid the 'sticky clusters effect', as suggested in Bush & MacEachern (1996).

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

2.2. Moment-matching criterion

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

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

In order to make this methodology applicable, one needs to propose the truncation level $Q(\ell)$ required to achieve a given approximation ℓ. Such map $Q(\ell)$ only depends on the NGG parameters and can be computed once-for-all and distributed with the package. For reference, see the moment matching error $\ell(Q)$ and the map $Q(\ell)$ respectively displayed in Figures 1 and 2 of Arbel & Prünster (2017). Ferguson and Klass posterior sampling based on such a prescribed number of jumps $Q(\ell)$ is computationally more efficient than having to iteratively compute the relative error e_Q as done in the previous package version.

2.3. Clustering estimation

We focus here on the problem of estimating a data clustering from the Bayesian posterior inference conducted so far. This is a long standing problem in Bayesian statistics (see for instance Dahl 2006; Lau & Green 2007). Enumerating all partitions is practically not feasible, 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 $\mathbf{X}$

$$\hat{c} = \arg\min_{c'} \sum_{c} L(c', c)\pi(c \mid \mathbf{X}), \tag{11}$$

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

An asset of the approach presented in Wade & Ghahramani (2018) is that it rests on a greedy search algorithm to determine the minimum loss clustering of (11). Starting from the MCMC output, this greedy approach explores the space of partitions and is not restricted to those visited by the MCMC chain to find the optimum. We include the possibility to estimate the optimal clustering using both the $\mathcal{VI}$ loss and Binder's loss, along with other loss functions, within BNPdensity by adding an optional dependence to GreedyEPL. Note that clustering estimation is also available for censored data, although graphical representation is more tricky (see also the legend to Figure 8).

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

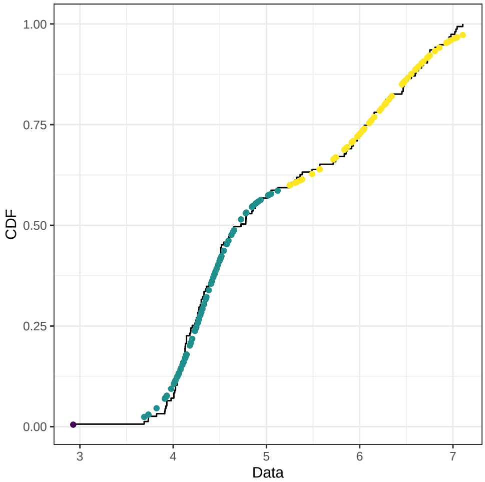

Figure 1. Visualisation of the clustering induced by the BNP mixture model, for the `acidity` dataset. The solid line represents the empirical Cumulative Distribution Function (CDF), dots represent data points. The abscissa of each point is its value, the ordinate is the value of the estimated CDF at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

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*:

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

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

- `distr.k`: Integer number identifying the **mixture kernel** k. Five kernels parameterised by their location and scale are implemented: a Gaussian or double exponential kernel for real data, a gamma or lognormal kernel for positive data and a beta kernel for data on the unit interval. The flexibility of this choice is afforded by the specific algorithm used in BNPdensity.
- `distr.py0`: Integer number identifying the base measure P_0 on the location parameters. Three choices are available, which are constrained by the conjugate prior we place on the hyperparameters of P_0: Gaussian, gamma and beta. Additional arguments can be used to tune the shape of the base measure.
- `distr.py0`, `distr.pz0`: Integer number identifying the base measure P_0 on scale parameters. For the semiparametric model (`MixNRMI1`), this argument is not provided and the base measure is a gamma distribution on the common scale parameter. Traditionally, there is sufficient information in the data to estimate the common scale parameter and inference is not very sensitive to the shape of the base measure. For the fully nonparametric model, the base measure on the scale parameters can be a gamma, lognormal, half Cauchy, half normal, half Student-t, uniform or truncated normal distribution. Additional arguments can be used to tune the shape of the base measure.
- `(Alpha, Kappa, Gama)`: Mixing measure parameters identifying a **Normalised generalised gamma** process, see the Lévy intensity (2) with parameters (α, κ, γ) for more details.
- The rest of the parameters provide handles to tune the MCMC algorithm.

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

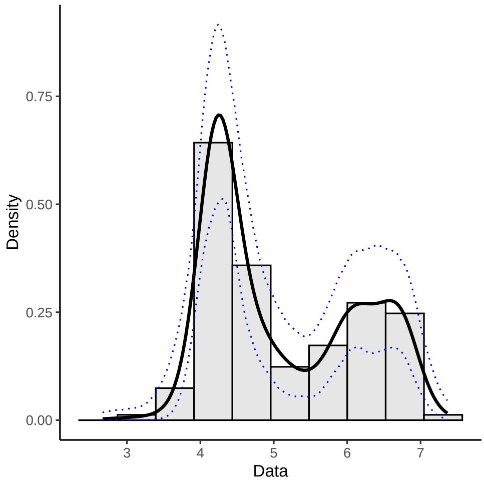

Figure 2. Density estimate (solid black line), 95% credible interval (blue dotted line) and histogram of the `acidity` data fitted with a semiparametric model. Figure obtained using the command `plot(out)`.

```
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).
```

4. Package comparison

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

4.1. Inference algorithm

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

4.2. Mixing measure

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

BNPmix provides a set of functions for density estimation with Dirichlet process and Pitman–Yor mixing measures via marginal algorithms. DPpackage is a more general purpose package than both BNPdensity and BNPmix, including functions for regression models, generalised linear mixed models, and generalised additive models, on top of the density model. However, the implementation is primarily tailored to the Dirichlet process mixing measure. A natural extension to the Dirichlet and Pitman–Yor processes are Gibbs-type priors (De Blasi et al. 2015). NRMI are a larger class of priors than Gibbs-type priors, and their intersection is the NGG priors considered in BNPdensity, as established in Lijoi, Prünster & Walker (2008). Being an extremely general class of priors, Gibbs-type processes are beyond reach for a general treatment in a software, however both BNPdensity and BNPmix packages cover its most commonly used sub-classes. Pitman–Yor process is not implemented in BNPdensity as it is not an NRMI; yet, a dependence to BNPmix is made in BNPdensity, in such a way that users interested in comparing their results with Pitman–Yor can also use the dedicated functions MixPY1 (semiparametric) and MixPY2 (fully nonparametric) that call BNPmix `PYdensity` function. The mixing measures covered by the three packages and their mutual relationships are illustrated in Figure 3.

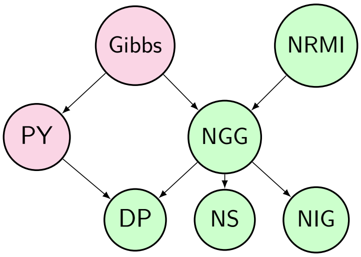

Figure 3. BNP priors mentioned in this section. An arrow indicates that the target is a special case or a limit case of its origin. Gibbs: Gibbs-type process. NRMI: normalised random measures with independent increments. NGG: normalised generalised gamma process. PY: Pitman–Yor process. NIG: normalised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

4.3. Prior characteristics

4.3.1. Non-conjugacy

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

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

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

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

4.3.2. Prior distribution on number of components

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

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

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
```

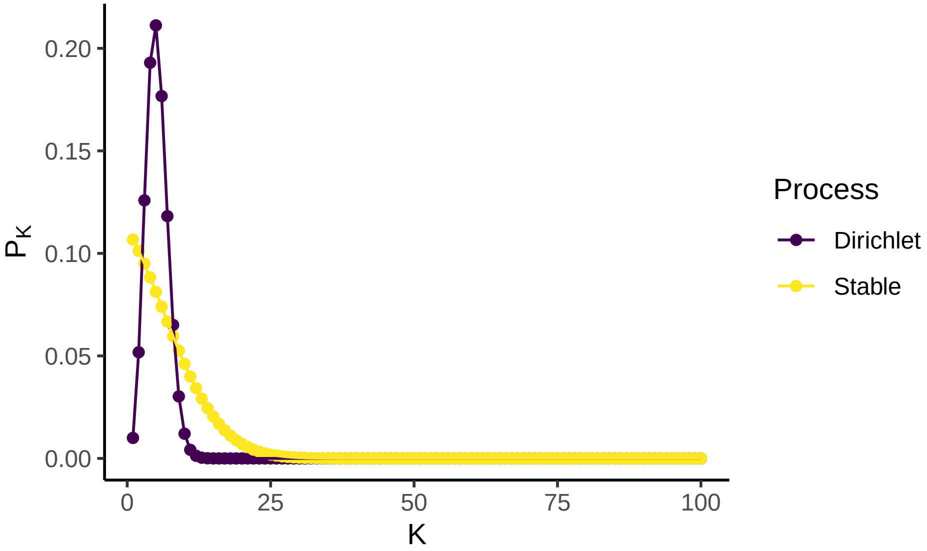

Figure 4. Prior distribution on the number of clusters with 100 data points, for the stable process with $\gamma = 0.4$ and for the Dirichlet process with $\alpha = 1$.

4.4. Censored data

BNPdensity can deal with left, right and interval-censored data by using the functions `MixNRMI1cens` and `MixNRMI2cens`. The same holds true for DPpackage, while BNPmix does not handle censored data at all.

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

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

4.5. Visualisation and programming

4.5.1. Convergence checking and model evaluation

BNPdensity offers several tools for assessing MCMC convergence and performing model checking and comparison. Notably, we provide a conversion function `as.mcmc` to interface the package with the coda package for analysing output and carrying out diagnostics on MCMC. We are not aware of such tools for BNPmix or DPpackage. This is done by running multiple chains starting from different initial conditions, potentially in parallel, and converting them into an `mcmc` object that can be processed by coda. A simple solution for running multiple chains does not seem available for BNPmix and DPpackage. One conceptual detail for assessing convergence is that, due to the nonparametric nature of the model, the number of parameters which could potentially be monitored to measure auto-correlation of the chains or effective sample size varies. The location parameters of the clusters, for instance, vary at each iteration, and even the labels of the clusters vary, which makes it tricky to follow. However, it is possible to monitor the log-likelihood of the data along the iterations, the value of the latent variable u, the number of components and for the semi-parametric model, the value of the common scale parameter. The following code shows how to compute the potential scale reduction factor (Gelman & Rubin 1992):

```
library(coda)
data(acidity)
fit = multMixNRMI1(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
```

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

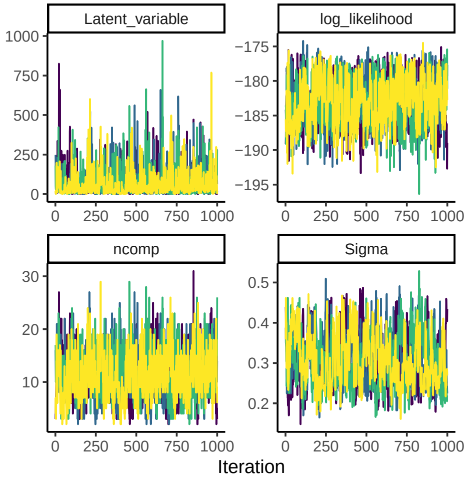

Figure 5. Trace plot of four chains in the MCMC for a semi-parametric model.

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

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

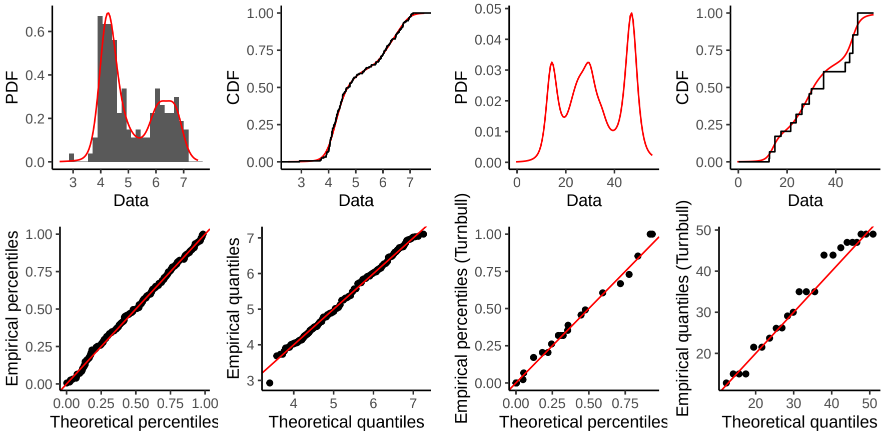

Figure 6. Graphical goodness of fit plots for censored (right) and non censored data (left). The top row is the mean density estimate with a histogram for the non censored data. The middle row is the estimated CDF with the empirical CDF for non censored data, and with the Turnbull estimate of the CDF 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.

We also provide tools for model comparison based on expected predictive density. The conditional predictive ordinate (CPO) is the expected predictive density of a data point given the prior and all other data points, so it is the leave-one-out expected predictive density of the model (Gelman et al. 2014), a typical cross-validation criterion. As such, it is a measure of predictive power with a penalisation for over-fitting. A Monte Carlo approximation of the CPO is easily available and can be used to compare a semi-parametric model to the fully 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 customary choice, notably because it readily provides confidence intervals on the HC_5, and because model comparison and goodness of fit tests have relatively low power on small datasets, precluding the emergence of a definite answer to the question.

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

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

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

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

In keeping with the traditional assumption of normality of the species sensitivity distribution, we choose to use a normal kernel for the mixture ($\texttt{distr.k} = 1$).

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

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

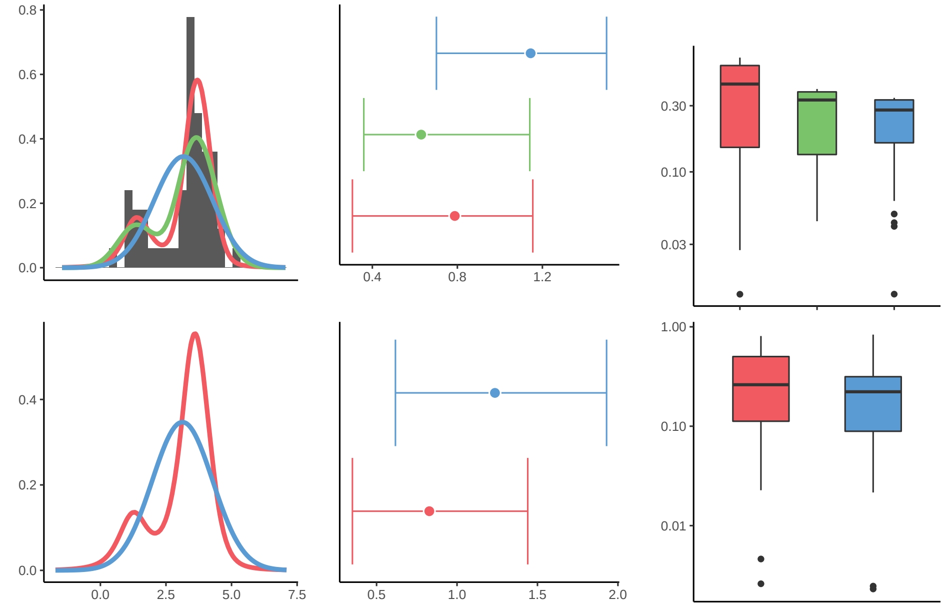

Figure 7. Top panel: non censored data. Bottom panel: censored data. The normal model is represented in blue, the KDE in green and the BNP in red. Left: density plot and histogram for the Carbaryl data using several SSD methods. The histogram is not available for censored data. Center: 5th percentile 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.

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

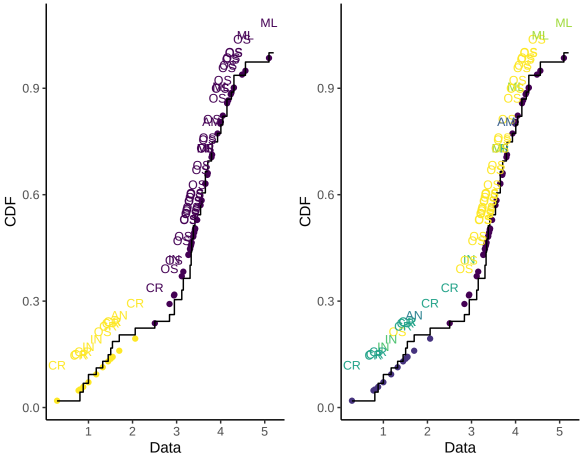

Figure 8. Graphical representation of the clustering induced by the mixture model for the Carbaryl data. The solid line represents the Turnbull estimate of the CDF, the points loosely represent the data. Interval censored data are represented at the middle of the interval, left and right censored data 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.

Computational details

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

References

Aldenberg, T. & Jaworska, J.S. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety* **46**, 1–18. doi:10.1006/eesa.1999.1869. URL http://www.ncbi.nlm.nih.gov/pubmed/10805987.

Arbel, J. & Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing* **27**, 3–17. doi:10.1007/s11222-016-9676-8.

Barrios, E., Lijoi, A., Nieto-Barajas, L.E. & Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science* **28**, 313–334.

Binder, D.A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* **31**, 929–953.

Burr, D. (2012). bspmma: An R package for Bayesian semi-parametric models for metaanalysis. *Journal of Statistical Software* **50**, 1–23.

Bush, C.A. & MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Canale, A., Corradin, R. & Nipoti, B. (2019). BNPmix: an R package for Bayesian nonparametric modelling via Pitman–Yor mixtures. *Journal of Statistical Software* , to appear.

Chung, Y., Gelman, A.G., Rabe-Hesketh, S., Liu, J. & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and 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.

Dahl, D.B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics* **4**, 201–218.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229. doi:10.1109/TPAMI.2013.217. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6654160. arXiv:1503.00163v1.

Denwood, M.J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software* **71**, 1–25.

Dowse, R., Tang, D., Palmer, C.G. & Kefford, B.J. (2013). Risk assessment using the species sensitivity distribution method: Data quality versus data quantity. *Environmental Toxicology and Chemistry* **32**, 1360–1369. doi:10.1002/etc.2190. URL http://www.ncbi.nlm.nih.gov/pubmed/23440771.

Escobar, M.D. & West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**, 577–588. doi:10.1080/01621459.1995.10476550. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550.

Ferguson, T.S. & Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Stat.* **43**, 1634–1643.

Forbes, V.E. & Calow, P. (2002). Species Sensitivity Distributions Revisited: A Critical Appraisal. *Human and Ecological Risk Assessment* **8**, 473–492. doi:10.1080/20028091057033. URL http://www.tandfonline.com/doi/abs/10.1080/10807030290879781.

Frühwirth-Schnatter, S., Celeux, G. & Robert, C.P. (2018). *Handbook of Mixture Analysis*. Chapman & Hall/CRC.

Gelman, A.G. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* **1**, 515–534. doi:10.1214/06-BA117A.

Gelman, A.G., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2014). *Bayesian Data Analysis*. Boca Raton, FL: CRC press, 3rd edn.

Gelman, A.G. & Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**, 457–511. doi:10.1214/ss/1177011136.

Gilks, W.R., Thomas, A. & Spiegelhalter, D.J. (1993). A Language and program for complex bayesian modelling. *Journal of the Royal Statistical Society. Series D (The Statistician)* **43**, 169–177. doi:Doi10.2307/2348941.

He, W., Qin, N., Kong, X., Liu, W., Wu, W., He, Q., Yang, C., Jiang, Y., Wang, Q., Yang, B. & Xu, F. (2014). Ecological risk assessment and priority setting for typical toxic pollutants in the water from Beijing-Tianjin-Bohai area using Bayesian matbugs calculator (BMC). *Ecological Indicators* **45**, 209–218. doi:10.1016/j.ecolind.2014.04.008. URL http://dx.doi.org/10.1016/j.ecolind.2014.04.008.

Helsel, D.R. (2005). *Nondetects and data analysis. Statistics for censored environmental data.* Wiley-Interscience. doi:10.2136/vzj2005.0106br.

Helsel, D.R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **65**, 2434–2439.

Hickey, G.L., Craig, P.S., Luttik, R. & de Zwart, D. (2012). On the quantification of intertest variability in ecotoxicity data with application to species sensitivity distributions. *Environmental Toxicology and Chemistry* **31**, 1903–1910. doi:10.1002/etc.1891. URL http://www.ncbi.nlm.nih.gov/pubmed/22619109.

Jagoe, R.H. & Newman, M.C. (1997). Bootstrap estimation of community NOEC values. *Ecotoxicology* **6**, 293–306. doi:10.1023/A:1018639113818. URL http://dx.doi.org/10.1023/A:1018639113818.

James, L.F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36**, 76–97. doi:10.1111/j.1467-9469.2008.00609.x.

Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *R News* **7**, 17–26. URL https://CRAN.R-project.org/doc/Rnews/.

Jara, A., Hanson, T.E., Quintana, F.A., Müller, P. & Rosner, G.L. (2011). DPpackage: Bayesian non-and semi-parametric modelling in R. *Journal of statistical software* **40**, 1.

Karabatsos, G. (2017). A menu-driven software package of bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods* **49**, 335–362. 1506.05435.

Kingman, J. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B* **37**, 1–15.

Kon Kam King, G., Arbel, J. & Prünster, I. (2017). A Bayesian Nonparametric Approach to Ecological Risk Assessment. In *Bayesian Statistics in Action: BAYSM 2016, Florence, Italy, June 19-21*, eds. R. Argiento, E. Lanzarone, I. Antoniano Villalobos & A. Mattei. Cham: Springer International Publishing, pp. 151–159. doi:10.1007/978-3-319-54084-9_14. URL http://dx.doi.org/10.1007/978-3-319-54084-9{_}14.

Kon Kam King, G., Veber, P., Charles, S. & Delignette-Muller, M.L. (2014). MOSAIC_SSD: A new web tool for species sensitivity distribution to include censored data by maximum likelihood. *Environmental Toxicology and Chemistry* **33**, 2133–2139. doi:10.1002/etc.2644. URL http://www.ncbi.nlm.nih.gov/pubmed/24863265.

Lau, J.W. & Green, P.J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

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

LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2007b). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Stat. Soc. B Met.* **69**, 715–740.

LIJOI, A., PRÜNSTER, I. & WALKER, S.G. (2008). Investigating nonparametric priors with Gibbs structure. *Statistica Sinica* **18**, 1653–1668. URL http://www.ams.org/mathscinet-getitem?mr=MR2469329{%}5Cnpapers2://publication/uuid/1CD2CD58-C0D3-42E1-8129-24261868FAB1.

LIVERANI, S., HASTIE, D.I., AZIZI, L., PAPATHOMAS, M. & RICHARDSON, S. (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software* **64**, 1–30. URL http://www.jstatsoft.org/v64/i07/.

LO, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12**, 351–357.

MACEACHERN, S.N. & MÜLLER, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics* **7**, 223–238. doi:10.1080/10618600.1998.10474772. URL http://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474772.

MEILA, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* **98**, 873–895.

NEAL, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9**, 249–265.

PAPASPILIOPOULOS, O. & ROBERTS, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169.

PLUMMER, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20-22, Vienna, Austria. ISSN 1609-395X.*, vol. 124. p. 125. doi:10.1.1.13.3406.

PLUMMER, M. (2019). *rjags: Bayesian Graphical Models using MCMC.* URL https://CRAN.R-project.org/package=rjags. R package version 4-9.

POSTHUMA, L., SUTER II, G.W. & TRASS, P.T. (2002). *Species sensitivity distributions in ecotoxicology.* CRC press. URL http://www.amazon.com/Species-Sensitivity-Distributions-Ecotoxicology-Posthuma/dp/1566705789.

RASTELLI, R. & FRIEL, N. (2018). Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing* **28**, 1169–1186. doi:10.1007/s11222-017-9786-y. URL https://doi.org/10.1007/s11222-017-9786-y.

RCORETEAM (2019). R: A Language and Environment for Statistical Computing. URL http://www.r-project.org/http://www.r-project.org.

REGAZZINI, E., LIJOI, A. & PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31**, 560–585. doi:10.1214/aos/1051027881.

ROBERTS, G.O. & ROSENTHAL, J.S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics* **18**, 349–367.

SATO, K.I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, *Cambridge Studies in Advanced Mathematics*, vol. 68. Cambridge University Press.

SCRUCCA, L., FOP, M., MURPHY, T.B. & RAFTERY, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**, 205–233. URL https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf.

STAN DEVELOPMENT TEAM (2018). RStan: the R interface to Stan. URL http://mc-stan.org/. R package version 2.18.2.

STAN DEVELOPMENT TEAM & STAN DEVELOPEMENT TEAM (2019). Stan: A C++ Library for Probability and Sampling, Version 2.19. URL http://mc-stan.org/.

Sturtz, S., Ligges, U. & Gelman, A.E. (2005). R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. (2006). Making BUGS open. *R News* **6**, 12–17.

Todeschini, A., Caron, F. & Fuentes, M. (2014). Rbiips: Bayesian inference with interacting particle systems. *arXiv* URL http://alea.bordeaux.inria.fr/biips.

Van Straalen, N.M. (2002). Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc. *Environmental Toxicology and Pharmacology* **11**, 167–172. doi:10.1016/S1382-6689(01)00114-4. URL http://www.ncbi.nlm.nih.gov/pubmed/21782599.

Verdonck, F.A.M., Jaworska, J., Thas, O. & Vanrolleghem, P.A. (2001). Determining environmental standards using bootstrapping, Bayesian and maximum likelihood techniques: A comparative study. *Analytica Chimica Acta* **446**, 429–438. doi:10.1016/S0003-2670(01)00938-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0003267001009382.

Wade, S. & Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* **13**, 559–626.

Wagner, C. & Lokke, H. (1991). Estimation of ecotoxicological protection levels from NOEC toxicity data. *Water Research* **25**, 1237–1242. doi:10.1016/0043-1354(91)90062-U. URL http://www.sciencedirect.com/science/article/pii/004313549190062U.

Wang, Y., Wu, F., Giesy, J.P., Feng, C., Liu, Y., Qin, N. & Zhao, Y. (2015). Non-parametric kernel density estimation of species sensitivity distributions in developing water quality criteria of metals. *Environmental Science and Pollution Research* **22**, 13980–13989. doi:10.1007/s11356-015-4602-8. URL http://link.springer.com/10.1007/s11356-015-4602-8.

Xing, L., Liu, H., Zhang, X., Hecker, M., Giesy, J.P. & Yu, H. (2014). A comparison of statistical methods for deriving freshwater quality criteria for the protection of aquatic organisms. *Environmental Science and Pollution Research* **21**, 159–167. doi:10.1007/s11356-013-1462-y.

Xu, F.L., Li, Y.L., Wang, Y., He, W., Kong, X.Z., Qin, N., Liu, W.X., Wu, W.J. & Jorgensen, S.E. (2015). Key issues for the development and application of the species sensitivity distribution (SSD) model for ecological risk assessment. *Ecological Indicators* **54**, 227–237. doi:10.1016/j.ecolind.2015.02.001. URL http://www.sciencedirect.com/science/article/pii/S1470160X15000692.

Zajdlik, B.A., Dixon, D.G. & Stephenson, G. (2009). Estimating Water Quality Guidelines for Environmental Contaminants Using Multimodal Species Sensitivity Distributions: A Case Study with Atrazine. *Human and Ecological Risk Assessment* **15**, 554–564. URL http://www.tandfonline.com/doi/abs/10.1080/10807030902892539.

Zhao, J. & Chen, B. (2016). Species sensitivity distribution for chlorpyrifos to aquatic organisms: Model choice and sample size. *Ecotoxicology and Environmental Safety* **125**, 161–9. doi:10.1016/j.ecoenv.2015.11.039. URL http://www.sciencedirect.com/science/article/pii/S0147651315301883.