Bayesian Nonparametric Density Estimation in Ecotoxicology

Guillaume Kon Kam King[1], Julyan Arbel[1,2] & Igor Prünster[2]

guillaume.konkamking@gmail.com, julyan.arbel;igor@unibocconi.it

[1] *Collegio Carlo Alberto, Moncalieri, Italy*
[2] *Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Milan, Italy*

Résumé. On revisite une méthode classique d'évaluation de risque écologique via une approche bayésienne non paramétrique. Les modèles bayésiens de mélanges infinis permettent de s'affranchir d'une hypothèse paramétrique classique mais controversée, tout en conservant la possibilité d'utiliser de petits jeux de données, typiques de l'écotoxicologie.

Mots-clés. Distribution de sensibilité d'espèce, Écotoxicologie, Mesures aléatoires normalisées, Modèles de mélange, Statistique bayésienne non paramétrique.

Abstract. We revisit a classical method for ecological risk assessment using a Bayesian nonparametric approach. By resorting to nonparametric mixture models it is possible to overcome a historically debated parametric assumption while retaining the ability to deal with small datasets that are typical of ecotoxicology.

Keywords. Bayesian Nonparametrics, Ecotoxicology, Mixture models, Normalized random measures, Species Sensitivity Distribution.

1 Introduction

Assessing the response of a community of species to an environmental stress is critically important for ecological risk assessment. Species Sensitivity Distribution (SSD) is one of the tools routinely used by environmental managers and regulators in most countries (Australia, China, EU, USA, . . .). The SSD approach characterises, for a given contaminant, the tolerance of all species possibly exposed using information collected on a sample of species. This information consists of Critical Effect Concentrations (CECs), a species-specific concentration marking a limit over which the species suffers a critical level of effect. Examples include: the concentration at which 50% of the tested organisms died (Lethal Concentration 50% (LC_{50})), or the concentration which inhibited growth or reproduction by 50% compared to the control experiment (Effect Concentration 50% (EC_{50})). Each CEC is the summary of long and costly bioassay experiments for a single species, so they are rarely available in large number. Minimal required sample size in Europe is 10 (ECHA, 2008), and there is currently a strong push to reduce animal testing.

To describe the tolerance of all species to be protected, the distribution of the CECs is estimated from the sample. In practice, a parametric distributional assumption is often adopted (Forbes and Calow, 2002): the CECs are assumed to follow a log-normal, log-logistic, triangular or BurrIII distribution.

Once the response of the community is characterised by the distribution, the goal of risk assessment is to define a safe concentration which will protect all or most of the species. To avoid infinitely small concentrations, a cut-off value is often chosen as the safe concentration, typically the Hazardous Concentration for 5% of the Species (HC_5), the 5th percentile of the distribution. The lower bound of the confidence interval on the HC_5 may also be used as the safe concentration, and a safety factor is typically applied a posteriori.

The lack of justification for the choice of any given parametric distribution sparked several research directions. First, authors have sought the best parametric distribution using goodness-of-fit measures for model comparison. The general consensus is that the best distribution depends on the dataset (Forbes and Calow, 2002). Nonetheless, the log-normal distribution has become the customary choice, notably because it readily provides confidence intervals on the HC_5. Moreover, model comparison and goodness of fit tests have low power on small datasets, precluding the emergence of a definite answer. Then, another research direction consisted in using distribution-free approaches. Those efforts included using the empirical distribution function, methods based on ranks, and bootstrap resampling. Wang et al. (2015) proposed a nonparametric kernel density estimation. All these approaches have in common that they require large sample sizes to function well. Finally, authors have considered the possibility that the distribution of the CECs might rather be a mixture of distributions, datasets being an assemblage of several log-normally distributed subgroups (Craig, 2013). This is more realistic from an ecological point of view because several factors influence the tolerance of a species to a contaminant such as the taxonomic group or the mode of action.

Ignorance of the group structure is a strong motivation for a nonparametric approach. However, the method must remain applicable to small datasets, which suggests trying to improve on the existing frequentist nonparametric methods. Bayesian nonparametric (BNP) mixture models offer an interesting solution for both large and small datasets, because the complexity of the mixture model adapts to the size of the dataset. Moreover, the low amount of information available in small datasets to estimate the groups parameters can be complemented via the prior distribution, as some a priori degree of information is generally available from other species and contaminants (Craig, 2013). Here, we summarize some of the findings of Kon Kam King et al. (2016). The rest of the paper is organised as follows. In Section 2 we present the BNP model and existing frequentist models for SSD and explain how to obtain a density estimate. Then in Section 3 we compare the different methods on a real dataset, illustrating the benefits of the BNP SSD.

2 Models for SSD

Given that concentrations vary on a wide range, it is common practice to work on log transformed concentrations. Consider a sample of n log-concentrations denoted by $\boldsymbol{X} = (X_1, \ldots, X_n)$. We propose to carry out density estimation for the SSD based on sample $\boldsymbol{X}$ by use of nonparametric mixtures. Bayesian nonparametric mixtures were introduced with Dirichlet process mixtures (DPM) which can be generalized by allowing the mixing distribution to be any discrete nonparametric prior. A large class of such prior distributions is obtained by normalizing random measures known as *completely random measures*. The normalization step, under suitable conditions, gives rise to so-called normalized measures with independent increments (NRMI) as defined by Regazzini et al. (2003), see also Barrios et al. (2013) for a recent review. An NRMI mixture model is defined hierarchically as:

$$\begin{aligned} X_i|\mu_i,\sigma &\overset{\text{ind}}{\sim} k(\cdot|\mu_i,\sigma), \quad \mu_i|\tilde{P} \overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad i=1,\ldots,n, \\ \tilde{P} &\sim \text{NRMI}, \quad \sigma \sim \text{Ga}(a_\sigma, b_\sigma). \end{aligned} \tag{1}$$

where k is a kernel, which we assume parametrized by some $\theta = (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+$, and $\tilde{P}$ is a random probability on $\mathbb{R}$ whose distribution is an NRMI. In our model, all clusters have a common variance. This is easier to fit on a small dataset, because information about the variance is pooled across clusters. Similar mixture SSD models described in Craig (2013) also assume common variance. As described in the Introduction, concentrations are commonly fitted with a log-normal distribution. Our aim is to move from this parametric model to the nonparametric one in (1). In order to allow comparisons to be made, we stick to the normal specification for k on the log-concentrations $\boldsymbol{X}$ by letting: $k(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma)$. Under this framework, density estimation is carried out by evaluating the posterior predictive density along the lines of Barrios et al. (2013). To specify the prior, we choose as mixing random measure the normalized stable process with (i) a stability parameter $\gamma = 0.4$ (ii) a base measure (which corresponds to the mean of the random measure) $P_0(\cdot) = \mathcal{N}(\cdot\,|\,\varphi_1, \varphi_2)$ with mean φ_1 and standard deviation φ_2, hyperparameters fixed a priori to specify a certain knowledge in the degree of smoothness (iii) a common variance for all the clusters with a vaguely informative prior distribution $Ga(0.5, 0.5)$. For posterior sampling, we use the R package BNPdensity and the function MixNRMI1 (see Barrios et al., 2013).

To illustrate the interest of the Bayesian nonparametric SSD, we compare our proposed BNP model to two commonly used frequentist models: the normal distribution (Aldenberg and Jaworska, 2000) and the nonparametric Kernel Density Estimate (KDE) recently proposed by Wang et al. (2015). For both frequentist approaches, the data is assumed to be iid.

For the purpose of comparing the predictive performance of the model, we resort to Leave-One-Out (LOO) cross-validation. We compute the LOO for each of the methods as

$\mathrm{LOO}_i = \hat{f}(X_i \,|\, \boldsymbol{X}_{-i})$ where $\hat{f}(x \,|\, \boldsymbol{X}_{-i})$ is the density for one of the three methods estimated from $\boldsymbol{X}$ with X_i left out. The LOOs for the BNP model correspond to the conditional predictive ordinates (CPOs) (see Barrios et al., 2013).

Finally, the quantity of interest for ecological risk assessment is the HC_5, which corresponds to the 5th percentile of the SSD 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 KDE is obtained by numerical inversion of the cumulative distribution function, and the confidence intervals using nonparametric bootstrap. The 5th percentile of the normal SSD and its confidence intervals are obtained following the classical method of Aldenberg and Jaworska (2000).

3 Application to real data

We apply this model to a selection of contaminants extracted from a large database collected by National Institute for Public Health and the Environment (RIVM). This database was prepared, studied and published by Hickey et al. (2012). We only consider non censored data, censored data being either discarded or transformed. Kon Kam King et al. (2016) will describe how the method can be adapted to include censored data. Using a continuous distribution for the CECs implies that the model does not support ties (or, in other words, observing ties has zero probability). However, ties may appear in the dataset due to the rounding of concentrations. Hence, we use a small jittering of the data.

We selected two example datasets: a medium-sized `temephos` dataset (CAS: 3383-96-8, mosquito larvicide, 21 species), and a small `captan` dataset (CAS: 133-06-2, fungicide, 13 species). Datasets for new contaminants are always small, the minimum requirement set by the European Chemical Agency being 10 species. The datasets can be visualised on the histograms of Figure 1. These datasets illustrate different features of the three approaches: when there is a clear multimodality in the data, the BNP SSD is more flexible than the fixed bandwidth KDE SSD (Figure 1, `captan`). When the data do not exhibit strong multimodality, as for `temephos`, the BNP reduces to the normal SSD model, whereas the KDE is by construction a mixture of 21 normal components.

One might think to increase the flexibility of the KDE by simply decreasing the bandwidth. However, that would also decrease the robustness of the method. On the second column of Figure 1, the LOO give an indication of the robustness to over-fitting of the three methods. For `captan`, they show that the superior flexibility of the BNP SSD compared to the KDE SSD does not come at the expense of robustness, because the median CPO of the BNP SSD is higher than the other two. In the case of `temephos`, the median LOO likelihood estimate of the normal model is very similar to the median CPO for the BNP SSD, sign that there is little over-fitting. This generally illustrates the fact that model complexity in a BNP model scales with the amount and structure of the data. On the right hand side of Figure 1, the credible intervals of the HC_5s for the BNP SSD are generally larger than the confidence interval of the normal SSD.

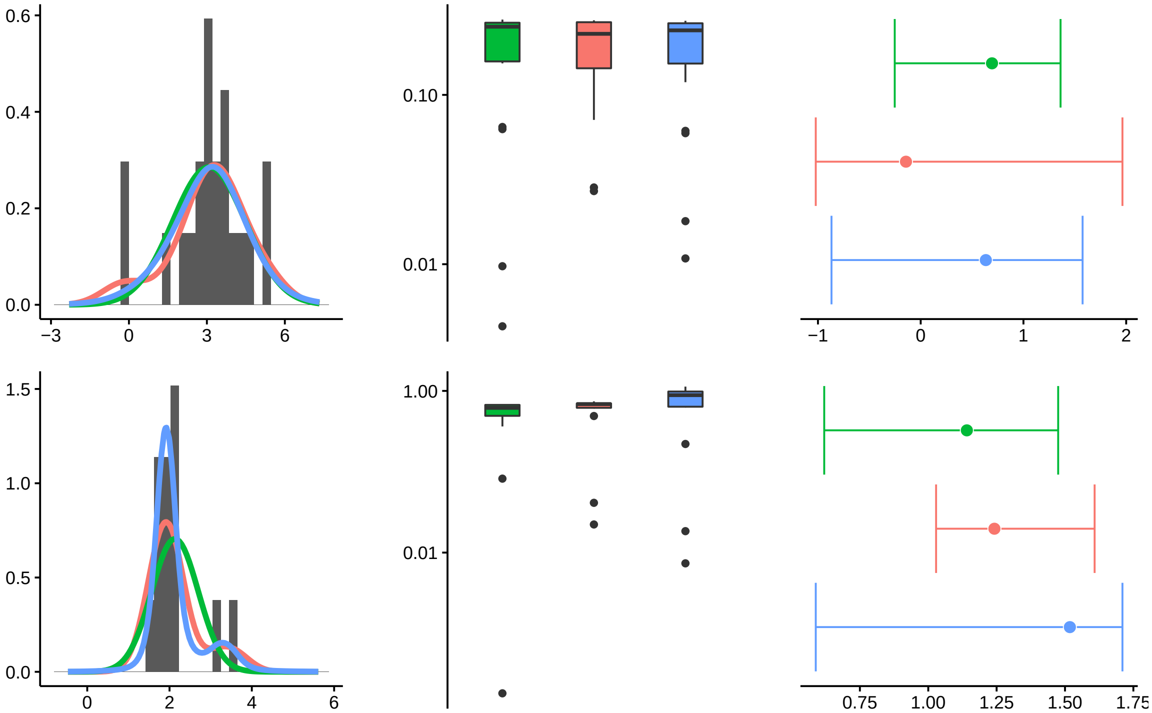

Figure 1: The top panel represents the medium-sized `temephos` dataset, the bottom panel represents small-sized `captan` dataset. Fits of the Normal, KDE and BNP models. Concentrations are log transformed. *Left*: Histogram and density estimates. *Centre*: Boxplot for the LOO (for Normal and KDE) and the CPO (for BNP) on logarithmic scale. *Right*: HC_5 and associated confidence/credible intervals.

In conclusion, the BNP SSD seems to perform well when the dataset deviates from a normal distribution. Its great flexibility is an asset to describe the data, while it does not seem prone to over-fitting. It can be thought of as an intermediate model between the single component normal SSD and the KDE with as many components as there are species. We chose to base the BNP SSD on NRMI rather than on the more common Dirichlet Process, because it is more robust in case of misspecification in the number of clusters (Barrios et al., 2013). The BNP SSD provides several benefits for risk assessment: it is an effective and robust standard model which adapts to many datasets. Moreover, it readily provides credible intervals. While it is always possible to obtain confidence intervals for a frequentist method using bootstrap, it can be difficult to stabilise the interval for small datasets even with a large number of bootstrap samples. As such, the BNP SSD represents a safe tool to remove one of the arbitrary parametric assumptions of SSD (Forbes and Calow, 2002).

Future work to support the BNP SSD will include a comparison of methods on simulated data, an extension to the case of censored data and an emphasis on the potential benefits of the approach from a biological point of view.

Acknowledgement

J. Arbel and I. Prünster are supported by the European Research Council (ERC) through StG “N-BNP” 306406.

References

Aldenberg, T. and Jaworska, J. S. (2000). Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and environmental safety*, 46(1):1–18.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science*, 28(3):313–334.

Craig, P. S. (2013). Exploring novel ways of using species sensitivity distributions to establish PNECs for industrial chemicals: Final report to Project Steering Group. Technical report.

ECHA (2008). Characterisation of dose concentration-response for environment. In *Guidance on information requirements and chemical safety assessment*, chapter R.10. European Chemicals Agency, Helsinki.

Forbes, V. E. and Calow, P. (2002). Species Sensitivity Distributions Revisited: A Critical Appraisal. *Human and Ecological Risk Assessment*, 8(3):473–492.

Hickey, G. L., Craig, P. S., Luttik, R., and de Zwart, D. (2012). On the quantification of intertest variability in ecotoxicity data with application to species sensitivity distributions. *Environmental Toxicology and Chemistry*, 31(8):1903–1910.

Kon Kam King, G., Arbel, J., and Prünster, I. (2016). Species Sensitivity Distribution revisited: a Bayesian nonparametric approach. *In preparation.*

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585.

Wang, Y., Wu, F., Giesy, J. P., Feng, C., Liu, Y., Qin, N., and Zhao, Y. (2015). Nonparametric kernel density estimation of species sensitivity distributions in developing water quality criteria of metals. *Environmental Science and Pollution Research*, 22(18):13980–13989.