Nonparametric bayesian model to cluster individual co-exposure to pesticides found in the French diet.

Amelie Crepet, Jessica Tressou

Nonparametric bayesian model to cluster individual co-exposure to pesticides found in the French diet.

Amélie Crépet & Jessica Tressou

October 15, 2009

Keywords

Dirichlet process; nonparametric bayesian modeling; multivariate normal mixtures; clustering; multivariate exposure; food risk analysis.

Abstract

This work introduces a specific application of bayesian nonparametric statistics in the food risk analysis framework. The goal is to determine cocktails of pesticide residues to which the French population is simultaneously exposed, so as to give directions for future toxicological experiments for studying possible combined effects of those cocktails. For that, the joint distribution of exposures to a large number of pesticides, called the co-exposure distribution, is assessed from the available consumption data and food contamination analyses. We propose to model the co-exposure by a Dirichlet process mixture based on a multivariate Gaussian kernel so as to determine groups of individuals with similar co-exposure patterns. The study of the correlation matrix of these sub-populations will permit to define the cocktails of pesticides to which they are jointly exposed at high doses. The posterior distributions and the optimal partition are computed through a Gibbs sampler based on stick-breaking priors. To reduce computational time due to the high dimension of the data, a random block sampling is used. As an extension, we propose to account for the uncertainty of food contamination through the introduction of an additional level of hierarchy in the model. The results of both specifications are exposed and compared.

1 Introduction

Each food product may contain several residues of pesticides, consequently meals daily ingested may include a large range of pesticides. Therefore, all consumers are expected to be exposed to complex cocktails of pesticides, for which combined effects on health are still unknown. This work proposes a novel methodology to respond to the following question "what are the cocktails of pesticides to which the French population is simultaneously and the most exposed ?" Cocktails of pesticides are

selected based on their joint probability to occur at high doses in the French diet. The population exposure to the P different pesticides found in the diet, called the co-exposure, is firstly estimated considering both the food residue level patterns obtained from the monitoring programmes and the dietary habits of n sounded individuals of the French consumption study INCA 2 (AFSSA, 2009). Secondly, we have developed a method to cluster the population co-exposures to define groups of individuals with similar design of exposures to the P pesticides. Thirdly, the correlations between exposures to the P pesticides of the most exposed individuals have been studied to characterize the relevant cocktails of pesticides. To define homogeneous group of individuals, the population co-exposure to the P pesticides was modeled with a bayesian nonparametric model relying on the use of Dirichlet process, Ferguson (1973). This fully bayesian approach consists of building an infinite mixture model, in which the number of mixture components is potentially unlimited, and is itself a random variable that is part of the overall model. In our pesticide study, a multivariate Normal distribution was chosen as the kernel density of the mixture. In that way, the correlations of the P pesticides were modelled and individuals were clustered both on their co-exposure levels and on their co-exposure correlations. The mixing distribution G was modelled with a Dirichlet process which is the most popular modeling tool as a prior distribution for infinite mixture models in a nonparametric bayesian context, Lo (1984). Indeed, the DP can be viewed as a probabilistic measure on the space of probability measures and fulfills required proprieties defined by Ferguson (1973), Antoniak (1974) to be used as a prior distribution. To integrate the uncertainty of individual exposure to each pesticide, we modified the base model with a hierarchical DP approach, similar to the one proposed by Teh et al. (2006).

As it is often the case in bayesian statistics, inference was conducted via Monte Carlo Markov Chain techniques. A Gibbs sampling method based on the stick-breaking (SB) representation of the DP was retained to account for the complexity of the hierarchy within an effective algorithm, Ishwaran and James (2001). Indeed, the stick-breaking priors can be simply constructed using a sequence of independent Beta random variables. However, due to the high dimension of the datasets and to the iteration principle of the MCMC process, computational time is heavy and must be reduced. In that respect, the random block sampling proposed by Cabrera et al. (2009) was applied to the SB algorithm. This procedure consists of subsampling d variates among the available P dimensions at each iteration.

In the first section, the data of residue levels, consumed quantities and the co-exposure estimation are described. In the second section, infinite mixture and Dirichlet process are outlined. Then, the chosen prior distributions and the model in its hierarchical form are presented. To give practical purposes, the SB and the random block-SB algorithms are detailed. Finally, the models are applied to a set of simulated data and to the French population co-exposures to pesticides.

2 Co-exposure to pesticides application

The pesticide food exposure was estimated from the available datasets on individual food consumption from national dietary survey and on residue levels obtained from national pesticide residues monitoring programmes. Concerning dietary exposure assessment to pesticides residues, it is necessary to identify and take into account all foodstuffs in which significant residues might occur, as well as all pesticides that may be present in the food. Therefore, a first step consists in the identification of food/pesticide combinations to include in the exposure assessment. Most of the chemical analyses of pesticide residues are said "left-censored" when the concentrations levels are lower than the limit of quantification (abbreviated LOQ) of the laboratories. To deal with a large number of quantified data, a selection criteria is used: only the pesticides that have been quantified (residue level over the LOQ) in at least 10% of the analyses realized in one commodity are retained. Because their quantified residues levels were of the same order than the corresponding LOQ, some pesticides for which the percentage of quantified data was lower than 10% have been however included in the study. In such a case, it was considered that the pesticide is really present in the food but has not been quantified due to analytical restrictions. Therefore, even if the quantification level is rather low, the pesticide is of concern in term of risk of exposure.

2.1 Food consumption data

Consumption data are provided by the second "Individual and National Study of Food Consumption", INCA2 survey, carried out by the French Food Safety Agency, AFSSA (2009). The study was conducted into three fieldwork waves between late 2005 and April 2007 in order to cover seasonal variation. Two independent populations were included in the study: $2,624$ adults aged 18-79 years and $1,455$ children aged 3-17 years. Participants were selected using a three-stage random probability design stratified by region of residence, size of urban area and population category (adults or children). Each subject was asked to complete a seven-day food diary as well as other questionnaires on anthropometrical and socio-economical factors. Food were subsequently coded into $1,305$ "as consumed" food items (INCA2 classification). In order to match the consumption data to pesticide residues data, which are measured on raw agricultural commodities (RAC), the food items defined in the INCA2 survey were decomposed into 181 RAC. For that, 763 standardized recipes, which have been defined by the French Food Safety Agency taking account of industrial processes, home cooking habits and edible portions for the INCA2 survey (AFSSA, 2009) , were used.

In order to consider the heterogeneity of inclusion probabilities, we have created two samples of $n = 2,624$ adults and $n = 1,455$ children from the original sample, by carrying out random trials with replacement and respecting the provided sampling weights of each individual.
In the context of an acute risk, the time window is the 24 hour day, so among the available 7 days of

consumption of each individual, one was randomly selected. Only normo-reporters i.e. individuals whose energetic needs are covered by the declared consumptions, were considered for this study. Therefore, two samples of normo-reporters of $1,898$ adults and of $1,439$ children were used for the analysis.

2.2 Pesticide residues levels data

The data source on pesticide residues in food and drinking water corresponds to the annual monitoring programmes implemented in 2006 by the French administrations (Ministry of Economy, Ministry of Agriculture, Ministry of Health). These surveys provide sample distributions of residues for up to 300 pesticides for about 150 RAC. The number of samples collected varies from about 10 for minor commodities up to 480 for staples (apple, lettuce, etc.). Following the selection criteria described at the begining of this section, 79 pesticides have been retained for the analysis. Residues of the selected pesticides were analyzed in 120 RAC and in drinking water consumed by the INCA2 population. A total of $306,899$ analytical results corresponding to $8,364$ combinations of pesticide/commodity were used in this work.

2.3 Dietary co-exposure assessment

For each commodity a treated with the pesticide p, the daily consumption c_{ia} was multiplied by one residue level q_{pa} and adjusted by the body weight w_i of the consumer i. For the acute exposure, the daily consumption c_{ia} corresponds to the sum of all the quantities of commodity a consumed during the selected day. The intakes calculated for each commodity were summed to obtain a total exposure in milligrams of the chemical per kilogram of body weight of the consumer (mg/kg bw). This process was performed for $m = 1, \ldots, M$ values randomly selected in the contamination distribution of each pesticide/commodity combination to account for the residue level uncertainty. The final data set comprised of a serie of M possible daily exposures $x_{pim} = \sum_{a=1}^{A_p} (c_{ia} \times q_{pam})/w_i$ to each pesticide $p = 1, \ldots, P$, for each individual $i = 1, \ldots, n$. In order to deal with quantitative values, each censored data was uniformly selected between 0 and its censoring value (LOD). Similarly, for each pesticide/commodity combination, random contamination values were uniformly selected between the different observed residues levels, in respect to the probability of being into the interval. Scaling problems between pesticide exposure levels were ruled out by passing to log scale and normalizing the data. Two datasets were created, one considering the 95^{th} percentile of the distribution of the M exposures to the pesticide p of each individual i (one high exposure per individual), the other one considering the entire distribution empirically described by the M exposure values. Therefore, computations were realized, in the first case with a co-exposure matrix of size $n \times P$ and in the second case with a matrix of size $n \times M \times P$.

3 Methodology

3.1 Nonparametric bayesian model based clustering

A common approach to assign data to clusters is to construct a model in which data are generated from a mixture of probability distributions. In that way, the co-exposure of the n individuals to the P pesticides arise from a distribution composed of different sub-distributions, namely the mixture components. Therefore, the groups of individuals with similar patterns of pesticide co-exposure are identified as the ones sharing the same sub-distributions. Let the observed co-exposures $x = (x_1, \ldots, x_i, \ldots, x_n)$ with x_i a P dimensional vector $x_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})$, distributed with a density of probability

$$f(x_i) = \int_{\Theta} k(x_i|\theta) G(d\theta) \tag{1}$$

where $k(.|\theta)$ is the known density of the mixture components called the kernel density, with parameter $\theta \in \Theta$ and G the unknown mixing distribution. Under a nonparametric perspective, the unknown density G is one of an infinite-dimensional function space. In a bayesian approach, the challenge is to place an appropriate prior $P(G)$ on the distribution G. Note that the equation (1) can be broken by introducing the latent variables θ_i

$$\begin{aligned} x_i|\theta_i &\sim k(dx|\theta_i) \\ \theta_i|G &\sim G(d\theta) \\ G &\sim P(G) \end{aligned} \tag{2}$$

Individuals with similar pattern of pesticide co-exposure are the ones with similar values of θ_i.

Now consider a partition, noted as $\mathbf{p}$, that separates the n vectors x_i into $n(\mathbf{p})$ groups of individuals. The partition of size $n(\mathbf{p}) \in \{1, \ldots, n\}$, can be represented as $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$ where C_j denotes the j^{th} cluster for $j = 1, \ldots, n(\mathbf{p})$. The equation (1) can be expressed conditionnaly on the partition $\mathbf{p}$, as a classification likelihood

$$f(x|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} k(x_i, i \in C_j)$$

where $k(x_i, i \in C_j)$, is the normalization constant of the posterior distribution of θ given the measurements of the cluster C_j, that is $k(x_i, i \in C_j) = \int_{\Theta} \prod_{i \in C_j} k(x_i|\theta) G(d\theta)$.

In that classification likelihood context, the partition $\mathbf{p}$ is the parameter for which a prior-posterior analysis is required. The posterior distribution of $\mathbf{p}$ is the product between the prior distribution $P(G) \propto \prod_{j=1}^{n(\mathbf{p})} g(C_j)$, where g is a function of the cluster only, e.g. its size, and the joint distribution k of the subset of x:

$$\pi(\mathbf{p}|x) \propto \prod_{j=1}^{n(\mathbf{p})} g(C_j)k(x_i, i \in C_j). \tag{3}$$

An estimation of the optimal partition is the one that maximizes the posterior distribution (3), which is approximated in this paper with a Gibbs sampler described in the section 3.3. The number of clusters in the optimal partition represents the number of sub-populations of individuals with similar design of pesticide co-exposure.

Dirichlet process

A collection of distribution functions called random probability measures (RPM) can be assigned to the density G (Walker et al., 1999; Muller and Quintana, 2004). Ferguson (1973) stated properties of this class of measures and introduced the Dirichlet Process (DP) as one of the RPM. The DP is defined by two parameters, a scaling parameter γ and a base probability measure H. The distribution of probability G is drawn from a DP, noted $G \sim DP(\gamma, H)$, if and only if for any partition $(A_1, \ldots, A_k)$ of Ω, the vector of random probabilities $(G(A_1), \ldots, G(A_k))$ is drawn from a Dirichlet distribution

$$(G(A_1), \ldots, G(A_k)) \sim Dir(\gamma H(A_1), \ldots, \gamma H(A_k)). \tag{4}$$

From the equation (4), it is easy to show that for $A \in \Omega$

$$E[G(A)] = H(A) \text{ et } V[G(A)] = \frac{H(A)(1 - H(A))}{1 + \gamma}.$$

Stick-breaking representation of the Dirichlet process

Many representation of the DP, as Polya urn scheme closed to the Chinese restaurant process (Blackwell and MacQueen, 1973; Pitman and Yor, 1996) are relevant for computational purposes. Sethuraman (1994) introduced the stick-breaking (SB) representation of the DP. In that way, $G \sim DP(\gamma, H)$ can be represented as an infinite mixture of point masses

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where ϕ_k are random variables sampled from H, δ_{ϕ_k} refers to a point masss concentrated at atom ϕ_k and β_k are the "stick-breaking weights" depending on γ. Drawing θ_i from G in equation 2 means that θ_i is equal to one of ϕ_k with the associated probability β_k.

As infinite mixture is impossible to construct in practice, Ishwaran and James (2001) have shown that for a reasonable N ($N < \infty$) the quality of approximation of the G's is good. The random weights β_k are built from auxiliary weight $\beta_k^* \sim Beta(1, \gamma)$ through the stick-breaking

procedure given by

$$\beta_1 = \beta_1^*, \qquad \beta_k = \beta_k^* \prod_{l=1}^{k-1}(1-\beta_l^*) \text{ for } k = 2, \ldots, N-1, \text{ and } \beta_N = 1 - \sum_{k=1}^{N-1} \beta_k.$$

In our study, clusters of individuals with similar design of co-exposure are identified as the ones sharing the same atoms ϕ_k.

3.2 Specific models

Multivariate normal mixture model

A P dimensional multivariate Normal distribution $N_P(\mu, \tau^{-1})$ with mean vector $\mu \in R^P$ and random covariance matrix $\tau^{-1} \in R^{P \times P}$ is assigned to the kernel density k.

The conjugate Wishart-Normal $WN(\alpha, \Psi, m, t)$ density is used as the prior distribution of the parameters (μ, τ). Conditionally of the random precision matrix τ, the random vector μ is assigned a P-dimensional Normal distribution $\mu|\tau \sim N_P\left(m, (t\tau)^{-1}\right)$. A Wishart distribution is used for the symmetric and positive finite precision matrix τ as $\tau \sim W(\alpha, \Psi)$, where α is a scalar degree of freedom and Ψ a $P \times P$ scale matrix. Therefore, the base probability measure H is the combination of the Wishart and the Normal distributions

$$\begin{aligned} H(d\mu, d\tau) &= \left\{ 2^{-\alpha P/2} |\Psi|^{\alpha/2} \left(\Gamma_P(\alpha/2)\right)^{-1} \times |\tau|^{(\alpha-P-1)/2} \exp\left[-\frac{1}{2} Tr\left(\Psi\tau\right)\right]\right\} \\ &\quad \times \left\{ (2\pi)^{-P/2} |t\tau|^{1/2} \exp\left[-\frac{t}{2}(\mu-m)'\tau(\mu-m)\right]\right\} d\mu d\tau \end{aligned}$$

where Γ_P is the multivariate Gamma function $\Gamma_P(\alpha/2) = \pi^{P(P-1)/4} \prod_{r=1}^{d} \Gamma(\frac{\alpha+1-r}{2})$ and $Tr(A)$ is the trace of the matrix A.

The marginal density $k(x_i, i \in C_j)$ is obtained by integrating over the parameters μ and τ the product of the probability measure H and the product of the kernel density k of the $x_i, i \in C_j$, that is

$$k(x_i, i \in C_j) = \int\int \prod_{i \in C_j} k(x_i|\mu, \tau) H(d\mu, d\tau).$$

In respect with our distribution choice, the marginal density is written as

$$k(x_i, i \in C_j) = \frac{\prod_{r=1}^{d} \Gamma(\frac{\alpha_j^*+1-r}{2})}{\prod_{r=1}^{d} \Gamma(\frac{\alpha+1-r}{2})} \frac{t^{d/2}}{\pi^{de_j/2}\left(t_j^*\right)^{d/2}} \frac{|\Psi|^{\alpha/2}}{\left|\Psi_j^*\right|^{\alpha_j^*/2}} = \frac{\Gamma_d(\alpha_j^*/2)}{\Gamma_d(\alpha/2)} \frac{t^{d/2}}{\pi^{de_j/2}\left(t_j^*\right)^{d/2}} \frac{|\Psi|^{\alpha/2}}{\left|\Psi_j^*\right|^{\alpha_j^*/2}},$$

where $(^*_j)$ are the updated values of the parameters of the Wishart-Normal noted $WN(\alpha_j^*, \Psi_j^*, m_j^*, t_j^*)$,

and equal to

$$\alpha_j^* = \alpha + e_j,\ m_j^* = \frac{tm + e_j\overline{x_j}}{t_j^*},\ t = t + e_j,\ \Psi_j^* = \Psi + S_j + \frac{e_j t}{t_j^*}(m - \overline{x_j})(m - \overline{x_j})',$$

where e_j is the number of observations classified in the cluster C_j, $\overline{x_j} = \frac{1}{e_j}\sum_{i \in C_j} x_i$ is the mean of the cluster C_j and $S_j = \sum_{i \in C_j} (x_i - \overline{x_j})(x_i - \overline{x_j})$ is the corresponding covariance matrix.

The optimum number of mixture components $n(\mathbf{p})$ and the posterior distribution of the component mixture parameters are obtained by maximizing the following criteria corresponding up to a constant to the posterior empirical log likelihood (i.e. the log of the equation (3))

$$Q(p) = n(\mathbf{p}) \times \ln(\gamma) + \sum_{j=1}^{n(\mathbf{p})} \ln \Gamma(e_j) + \sum_{i=1}^{n} \ln k(x_i|\theta_i). \tag{5}$$

Hierarchical model to account for the uncertainty of the exposure

An additional Dirichlet process is used to account for the uncertainty of the exposure given by the set of data $x_{im} = \{x_{pim}, p = 1, \ldots, P\}$ for each individual $i = 1, \ldots, n$ and the contamination value $m = 1, \ldots, M$

$$\begin{aligned} x_{im}|\theta_{im} &\sim k(.|\theta_{im}) \\ \theta_{im}|G_i &\sim G_i \\ G_i &\sim DP(\alpha_i, G_0) \\ G_0 &\sim DP(\gamma, H). \end{aligned} \tag{6}$$

With such model, the co-exposure to the P pesticides of each individual i is composed of several sub-distributions identified by the ones sharing the same θ_{im}. Therefore, the shape of the individual exposures to each pesticide given by the M values is considered into the clustering process.

3.3 Algorithm

3.3.1 Stick-breaking algorithm of the base model

The algorithm of the base model described in equation (2) is presented below in three steps. Considering $G \sim DP(\gamma, H)$, we use the stick-breaking representation of the Dirichlet process $G = \sum_{k=1}^{N} \beta_k \delta_{\phi_k}(.)$, where ϕ_k denotes the hidden parameters of the multivariate gaussian distribution (μ_k, τ_k), $\beta = (\beta_1, \ldots, \beta_N)$ are the stick-breaking weights, and N is the maximum number of atom of G. A vector $K = (K_i)$ is introduced to store the affectation of each data $x_i = (x_{ip}, p = 1, \ldots, P)$ for $i = 1, \ldots, n$ to an atom $(\phi_k)_{k=1,\ldots,N}$ so that K_i is an integer from 1 to N. Only N^* of the N

available atoms are distinct values which set is denoted by K^*.

1. Sampling $(\phi|K, \beta, X)$: for those k in K^*, sample ϕ_k with respect to the "updated" base measure H_k^* (a Wishart-Normal with parameters $\alpha_k^*, \Psi_k^*, m_k^*, t_k^*$ obtained from the posterior distribution given by $\{x_i, K_i = k\}$) and for the remaining $(N - N^*)$ atoms, get ϕ_k from the base measure H (the prior Wishart-Normal(α, Ψ, m, t)).

2. Sampling $(K|\beta, \phi, X)$ for $k = 1, \ldots, N$, and $i = 1, \ldots, n$

$$\Pr(K_i = k) \propto \beta_k \times k(x_i|\phi_k). \tag{7}$$

3. Sampling $(\beta|\phi, K, X)$ based on the $\beta_k^* \sim Beta(1 + e_k, \gamma + \sum_{l=k+1}^{N} e_l)$ with e_k corresponding to $\#\{x_i, K_i = k\} \leq n$ for $k = 1, \ldots, N$, then

$$\beta_1 = \beta_1^*, \qquad \beta_k = \beta_k^* \prod_{l=1}^{k-1}(1 - \beta_l^*), \quad \text{for} \quad k = 2, \ldots, N-1, \qquad \beta_N = 1 - \sum_{l=1}^{N-1} \beta_l$$

Hierarchical model

The algorithm of the stick-breaking representation of the hierarchical Dirichlet process presented in equation (6) requires the sampling of additional intermediary weights $\pi = (\pi_{ik})$ and the definition of a matrix $K = (K_{im})$ describing the affectation of each data $x_{im} = (x_{pim}, p = 1, \ldots, P)$ to one of the N atoms, for $i = 1, \ldots, n$ and $m = 1, \ldots, M$. Steps 2 and 3 are replaced with steps 2' and 3' below.

2'. Sampling $(K|\pi, \beta, \phi, X)$ for $k = 1, \ldots, N$, $i = 1, \ldots, n$ and $m = 1, \ldots, M$

$$\Pr(K_{im} = k) \propto \pi_{ik} \times k(x_{im}|\phi_k)$$

3'. Sampling $(\pi|\beta, \phi, K, X)$, independently on the fixed i's and based on

$$\pi_{ik}^* \sim Beta(\alpha_0\beta_i + e_{ik}, \alpha_i \left(1 - \sum_{l=1}^{k} \beta_l\right) + \sum_{l=k+1}^{N} e_{il}),$$

with e_{ik} corresponding to $\#\{x_{im}, K_{im} = k\} \leq M$ for $i = 1, \ldots, n$ and $k = 1, \ldots, N$, then

$$\pi_{i1} = \pi_{i1}^*, \qquad \pi_{ik} = \pi_{ik}^* \prod_{l=1}^{k-1}(1 - \pi_{il}^*), \quad \text{for} \quad k = 2, \ldots, N-1, \qquad \pi_{iN} = 1 - \sum_{l=1}^{N-1} \pi_{il}.$$

Note that $\pi_{iN}^* \sim Beta(\alpha_i\beta_N + e_{iN}, 0)$ is a properly defined Beta distribution.

Finally sampling $(\beta|\pi, \phi, K, X)$ exactly as described in the original step 3.

Learning about the parameter γ

Considering γ as a random parameter leads to an additional last step.

4. Sampling $(\gamma|\phi, K, \pi, \beta, X)$ based on an auxiliary variable $\gamma^* \sim Beta(\gamma + 1, n)$, according to the following mixture distribution

$$\gamma \sim w_{\gamma^*} \times \Gamma(a_\gamma + k, b_\gamma - \ln \gamma^*) + (1 - w_{\gamma^*}) \times \Gamma(a_\gamma + k - 1, b_\gamma - \ln \gamma^*), \tag{8}$$

with weights w_{γ^*} defined by $\frac{w_{\gamma^*}}{1-w_{\gamma^*}} = \frac{a_\gamma+k-1}{b_\gamma-\ln \gamma^*} = c_{\gamma^*}$ that is $w_{\gamma^*} = \frac{c_{\gamma^*}}{1-c_{\gamma^*}}$.

Starting values of hyperparameters

Starting values of the hyperparameters were taken equals to $\alpha = P$, $\Psi = 0_{P\times P}$, $m = 0_P$, $t = 1$, as it is proposed in Cabrera et al. (2009) in order to use vague prior distributions. Some tests were done with the simulated datasets for the starting value related to the parameter γ: no prior distribution but γ is fixed to 1, a prior Gamma distribution with (a_γ, b_γ) the shape and the rate parameters equal to: $(a_\gamma, b_\gamma) = (2, 4)$ for informative prior and equal to $(a_\gamma, b_\gamma) = (1, 1)$ and $(a_\gamma, b_\gamma) = (0.01, 0.01)$ for more vague prior distributions (Escobar and West, 1995). In the case of the hierarchical model, the weigths α_i were fixed to 1 for each $i = 1, \ldots, n$.

3.3.2 Random-block Gibbs Stick-breaking

Cabrera et al. (2009) introduced a novel procedure called the random block Gibbs weighted Chinese restaurant process algorithm to reduce the heavy computational time to estimate the optimal partition, induced by the Gibbs sampler and the high dimensionality of the data. We propose to apply this method to the SB algorithm. The principle is to randomly reduce the dimension of the data by selecting a number $d(d < P)$ of pesticides among the original number P at each gibbs cycle. Therefore, given the sequence of random integers $v_d = \{l_1, \ldots, l_d\}$, a subset of observations $x_i = (x_{il_1}, \ldots, x_{il_d})$ is used instead of the $x_i = (x_{i1}, \ldots, x_{iP})$ for the $i = 1, \ldots, n$ individuals. This procedure will be referred to as RB-SB in the following.

4 Application

4.1 Simulated datasets

To investigate the quality of the clustering estimates under various settings, a simulation study was conducted for the stick-breaking algorithms applied to two datasets created from the one proposed by Cabrera et al. (2009). The datasets were built from three component mixture of $P = 5$ dimensional multivariate Normal distributions noted $N_P(\mu_k, \Sigma_k)_{k=1,\ldots,3}$. The parameters (μ_k, Σ_k) are detailed in Table 1.

Table 1: Parameters of the multivariate Normal distributions $N_P(k, S_k)_{k=1,\ldots,3}$ used to generate the simulated datasets

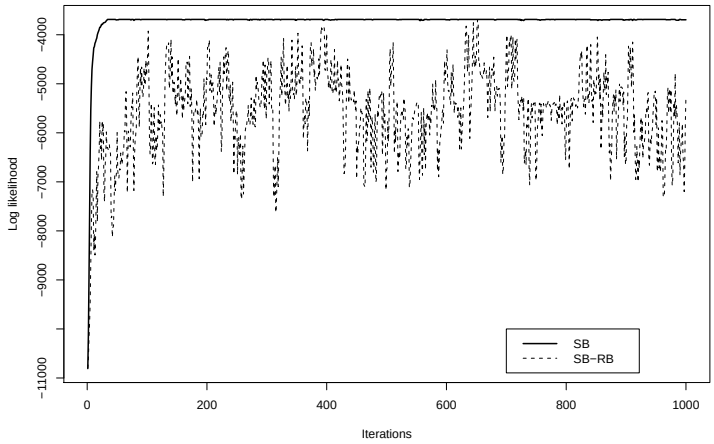
k	μ_k	Σ_k				
	2	1.0	0.5	0.2	0.1	0.1
	2	0.5	2.0	0.5	0.2	0.1
1	4	0.2	0.5	1.0	0.5	0.2
	5	0.1	0.2	0.5	1.0	0.5
	6	0.1	0.1	0.2	0.5	3.0
	-2	1.0	0.5	0.2	0.1	0.1
	-2	0.5	2.0	0.5	0.2	0.1
2	-4	0.2	0.5	3.0	0.5	0.2
	-5	0.1	0.2	0.5	2.0	0.5
	-6	0.1	0.1	0.2	0.5	1.0
	-5	3.0	0.5	0.2	0.1	0.1
	5	0.5	1.0	0.5	0.2	0.1
3	-7	0.2	0.5	2.0	0.5	0.2
	7	0.1	0.2	0.5	2.0	0.5
	-9	0.1	0.1	0.2	0.5	3.0

4.1.1 Dataset for the base model

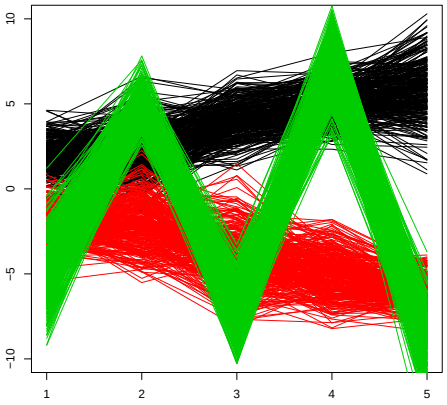
A sample of $1,000$ values (x_i) was built from $f(x) = \sum_{k=1}^{3} \beta_k N_P(\mu_k, \Sigma_k)$ with $(\beta_1, \beta_2, \beta_3) = (0.3, 0.3, 0.4)$. From $30,000$ iterations, and with the parameter γ fixed to 1, the optimal partition was obtained at the 34^{th} iteration for our Q-criteria, given in equation 5, of $Q = -3,687$, cf. Fig. 1(a). The optimal partition was composed of 3 clusters corresponding to the 3 components of the mixture dataset, cf. Fig. 1(b). Performing the RB-SB algorithm with dimension reduced to $d = 2$, the maximum Q-criteria is reached at the 650^{th} iteration. When using different prior distributions for the parameter γ, the same maximum value of $Q = -3,687$ is reached, as with γ fixed to 1. This attests to the robustness of the γ value in so far as the issues of predictive density estimation are concerned. Figure 2 shows the 3 different prior distributions attributed to the parameter γ and their corresponding posterior distributions. Note that 50% of posterior values of γ is below $0.27 < 1$, showing that it is possible, even though maybe not crucial here, to learn about γ.

4.1.2 Dataset for the hierarchical model

To reproduce the co-exposure data structure including uncertainty of exposure, 240 individuals were generated from the $f(x) = \sum_{k=1}^{3} \beta_k N_P(\mu_k, \Sigma_k)$ with $(\beta_1, \beta_2, \beta_3) = (0.33, 0.17, 0.5)$. For each individual i, $M = 100$ values were sampled resulting of a total sample of $4,800$ observations. The convergence of the SB and the RB-SB algorithms to the optimal partition is very slow. After



(a) Log-likelihood of the $1,000$ first iterations



(b) Optimum partition obtained with SB algorithm

Figure 1: The Stick-Breaking (SB) and the Random-Block Stick-Breaking (RB-SB) algorithms for the base model applied to simulated dataset ($N = 30$ atoms and $30,000$ iterations).

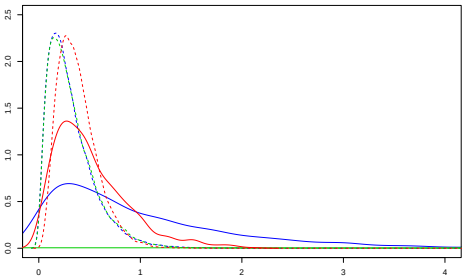


Figure 2: Densities of the $Gamma(a_\gamma, b_\gamma)$ prior (solid) and the posterior (dashed) ditributions of the parameter γ. Red line : $(a_\gamma, b_\gamma) = (2, 4)$, blue line : $(a_\gamma, b_\gamma) = (1, 1)$, green line : $(a_\gamma, b_\gamma) = (0.01, 0.01)$, $\ln(\gamma)$ for the prior.

200,000 iterations, the number of cluster which maximized the Q-criteria was 11. The size of the clusters ranged from 269 to 16,089 observations. However, some clusters had similar values of parameters to the distributions used to generate the dataset. This confirms that the algorithm will eventually converge to an optimal partition corresponding to the 3 components of the mixture dataset.

4.2 Real datasets on co-exposure to pesticides

4.2.1 Base model for the 95^{th} percentile of co-exposure

Different values of d ($d = \{15, 25, 41, 79\}$) have been tested running 100,000 iterations of the algorithm and using a prior distribution Gamma(0.01, 0.01) for γ as a flat prior. The presented results are for the value of d which maximizes the Q-criteria: $d = 41$. To test the convergence of the algorithm to the optimal partition an extra of 200,000 iterations had been performed with this value of d.

For the adults sample, the optimal partition was obtained after 196,445 iterations and is composed of 17 clusters. The adult population was clustered into 3 main sub-populations composed of 582, 412 and 870 individuals. The other 14 clusters were discarded as they jointly only count 34 individuals. For each main cluster, the boxplots of the 79 pesticide exposures are shown in Fig. 3. The sub-populations of the clusters 2 and 3 are the most highly exposed to a large number of pesticides. For these 2 populations, the correlation matrix of the pesticide exposure are drawn from the posterior distribution of the parameter τ and shown in cf. Fig. 4(a) and Fig. 4(c). To define cocktails, we focus on pesticides with at least one correlation with another, greater than 0.95 (cf. Fig. 4(b) and Fig. 4(d)). With this criteria, from the 79 pesticides and the two sub-populations, 34 pesticides have been selected and combined into 20 cocktails.

For the children sample, the optimal partition was obtained after 98,362 iterations and is composed of 16 clusters. As for the adults, 14 clusters totalizing only 45 individuals were discarded to focus on the 2 main ones. The first cluster is composed of 743 children which are highly exposed to a large number of pesticides and the second one is made of 651 children. As for the adult sample, the correlation matrix of the most exposed sub-population was analysed to determine cocktails of pesticides. There are 39 pesticides with correlations over 0.95, divided into 13 different cocktails. From this 39 pesticides, 28 are similar to the ones obtained with the adult population.

The base model have been compared with a classical principal component analysis (PCA) for the adult population. The axis 1, which represents 68% of the total variance is mainly determined by the 34 pesticides selected with our base model. Indeed, among the 24 pesticides which mostly contribute to axis 1, 23 have also been selected with our model, and the 34 pesticides selected with our base model are among the 50 first which build the axis 1. The second axis only represents 6% of the total variance. Note that the coordinates of the 79 pesticides are positive on the axis 1, while the individuals are present on the both sides of the axis. This leads to the conclusion that

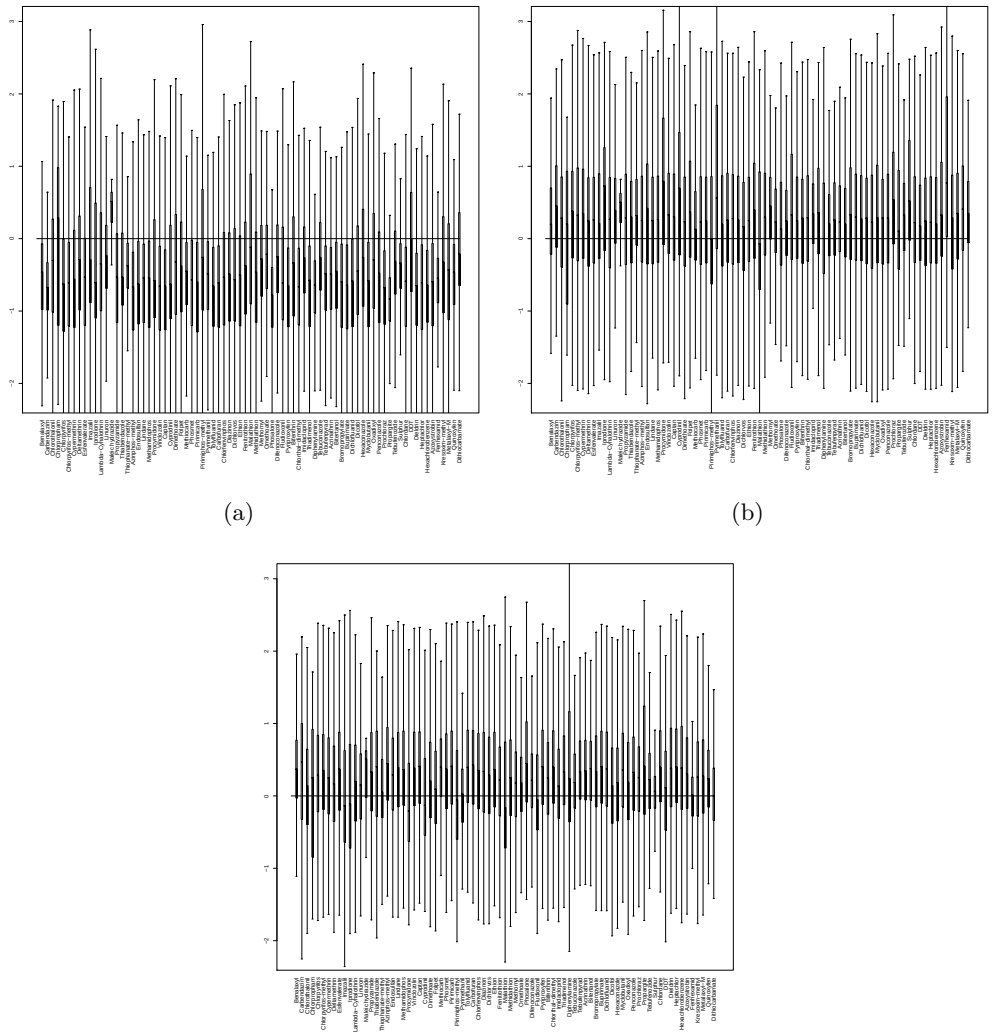


Figure 3: Boxplots of the 79 pesticide exposures for the cluster 1 (a), cluster 2 (b) and cluster 3 (c).

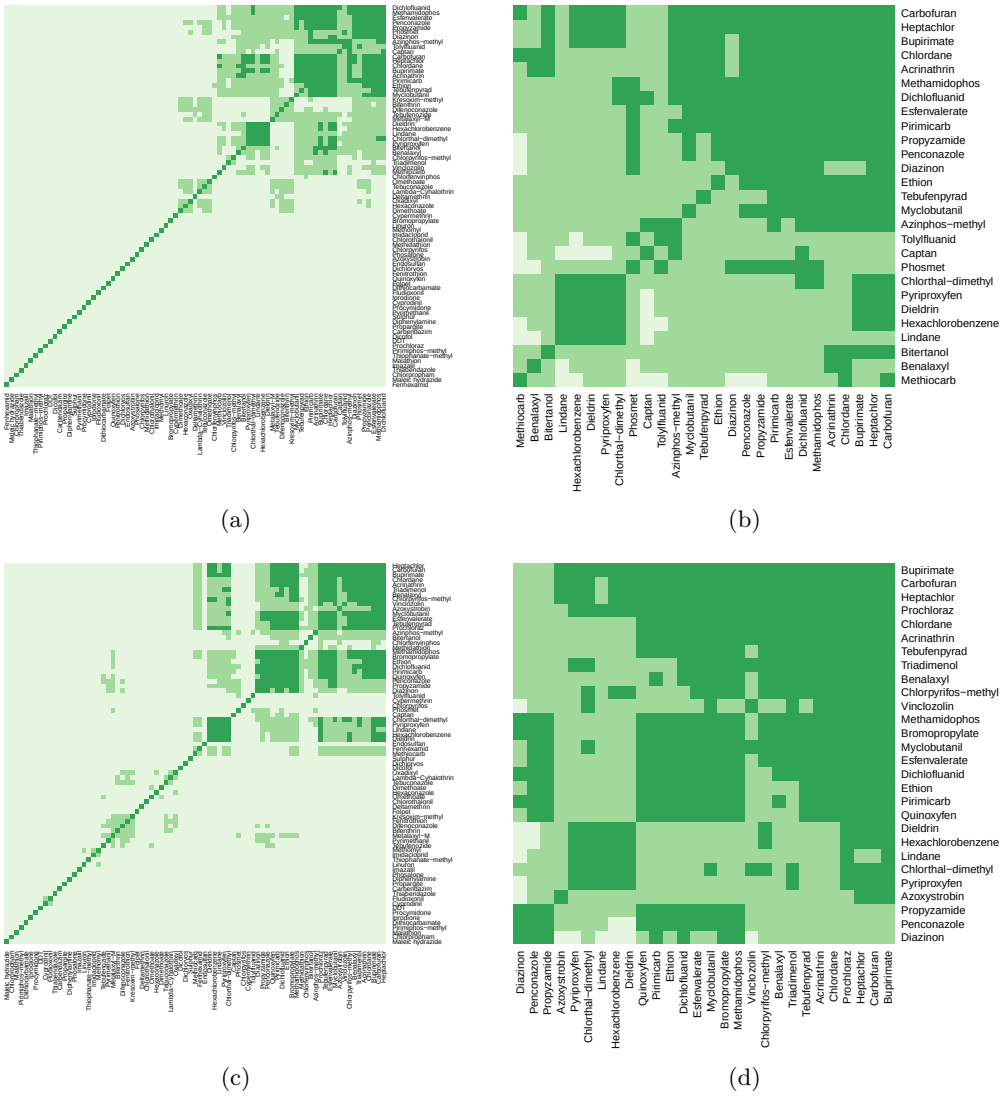


Figure 4: Heatmap of the correlation matrix of the cluster 2 (a,b) and the cluster 3 (c,d). Right maps are focus on pesticides with correlation upper than 0.95

the individuals are either highly exposed to all pesticides or have a low exposure for all pesticides, and hence that it is their consumption behavior that matters.

4.2.2 Hierarchical model to integrate exposure uncertainty

Because of the heavy computational time of the algorithm, the hierarchical model was only applied to the 34 pesticides selected with the base model for the adult population. From $30,000$ iterations, the optimal partition was found after $25,853$ iterations and was composed of 6 clusters. From these clusters, three were composed of large sets of observations according to the clustering obtained with the base model (cf. previous subsection). Moreover, two of these clusters were composed of individuals highly exposed to a large part of the 34 pesticides. The distribution of co-exposure of each individual, was mostly found composed of 2 or 3 component mixtures. The correlation matrix of the two main clusters shows that the correlations between pesticides were very low, ranged between 0.2 and 0.45, compared to the ones found with the base model. These low correlations could be due to the high uncertainty around the exposure to each pesticide. Indeed, for a one random set of contamination selected for the P pesticides, an individual can be exposed to a low level for one pesticide and to a high level for another pesticide, leading to low correlations. The integration of the exposure uncertainty is more realistic in terms of exposure assessment but imply difficulty to define cocktails of pesticides.

5 Conclusion

This paper presents a nonparametric bayesian model based on Dirichlet process mixtures, applied to cluster the co-exposure of the French population to various pesticides in order to define cocktails of pesticides which are relevant to study for toxicological effects. Such nonparametric bayesian model has several advantages i.e. the number of clusters is automatically determined through the estimation process, no parametric assumption on the shape of the co-exposure distribution is required and the structure of the data set may be introduced through a specific hierarchy to account for exposure uncertainty. However, the required hypothesis done to construct the co-exposure dataset in order to deal with low residues levels, lead with homogeneous population: individuals highly exposed to a large part of pesticides and individuals lowly exposed to a large part of pesticides. In that way, defining sub-populations and cocktails of pesticides is a difficult task and the results obtained are preliminary. Therefore, several extensions or changes in the framework can be considered to improve the clustering. For example, to obtain more clusters of individuals with similar co-exposure patterns, the Poisson Dirichlet process could be used as a prior distribution, see Pitman and Yor (1997). Another extension can be to cluster both the individuals and the pesticides. In that way, the Mondrian processes which are multidimensional generalizations of Poisson processes and which have been introduced by Roy and Teh (2009) to model relational

data, could be considered.

Acknowledgment

This work is a part of a project granted by the National Agency of Research (ANR) and the French Agency for Environmental and Occupational Health Safety (AFSSET).

References

AFSSA (2009). "INCA 2 (2006-2007). Rapport de l'étude Individuelle Nationale des Consommations Alimentaires 2."

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2: 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Polya urn schemes." *Annals of Statistics*, 1: 353–355.

Cabrera, J., Lau, J. W., and Lo, A. Y. (2009). "Random Block Sampling for high dimensional clustering (from the Bayesian point of view)." *Working Paper*.

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230.

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173.

Lo, A. Y. (1984). "On a class of bayesian nonparametric estimates: I. Density Estimates." *Annals of Statistics*, 12(1): 351–357.

Muller, P. and Quintana, F. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19(1): 95–110.

Pitman, J. and Yor, M. (1996). "Some developments of the Blackwell-MacQueen Urn scheme." *IMS, Hayward, California*.

— (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *Annals of Probability*, 25: 855–900.

Roy, D. M. and Teh, Y. W. (2009). "The Mondrian Process." *Working Paper*.

Sethuraman, J. (1994). "A constructive definition of Dirichlet prior." *Statistica Sinica*, 4: 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476): 1566–1581.

Walker, S., Damien, P., Laud, P., and Smith, A. (1999). "Bayesian nonparametric inference for random distributions and related functions." *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3): 485–527.