A Clustering Bayesian Approach for Multivariate Non-Ordered Circular Data

Christophe Abraham, Rémi Servien, Nicolas Molinari

A Clustering Bayesian Approach for Multivariate Non-Ordered Circular Data

Christophe Abraham [*] Rémi Servien [†]
Nicolas Molinari [‡]

Abstract

This paper presents a Bayesian model for the clustering of non-ordered multivariate directional or circular data. The particular trait of our data is that each single observation is made up of $k \geq 2$ non-ordered points on the circle. We introduce a hierarchical model that combines a symmetrization technique, Projected Normal distributions and a Dirichlet Process. One parameter is introduced to model the non-ordered trait and another one to control the variability of the angles on the circle. An informative prior on the relative locations of the k angles is also provided. The gain of the symmetrization is highlighted by a theoretical study. The parameters of the model are then inferred using a Metropolis-Hastings within Gibbs algorithm. Simulated datasets are analyzed to study the sensitivity to hyperparameters. Then, the benefits of our approach are illustrated by clustering real data made up of the positions of five separate radiotherapy x-ray beams on a circle.

Keywords : Circular data; Dirichlet process; Non-ordered multivariate data; Projected Normal Distribution; Radiotherapy machine data; Unsupervised clustering.

[*]Montpellier SupAgro-INRA, UMR MISTEA 729, Bâtiment 29, 2 place Pierre Viala, 34060 Montpellier Cedex 2, France.

[†]Toxalim, Université de Toulouse, INRA, Toulouse, France; remi.servien@inra.fr

[‡]Université de Montpellier, IMAG, place Eugène Bataillon, 34095 Montpellier cedex 5, France.

1 Introduction

Circular and directional data arise in a number of different fields such as oceanography (wave direction), meteorology (wind direction), biology (animal movement direction). The present paper is motivated by circular data in medicine. Nowadays, intensity-modulated radiation therapy (IMRT) has demonstrated its effectiveness for cancer treatment. The latest generation of radiotherapy machines projects multiple rays. Multiplying beams allows to concentrate radiation on the tumor while avoiding the massive irradiation of healthy areas. However, the selection of the incident angles of the treatment beams may be a crucial component of IMRT planning. Due to variations in tumor locations, size and patient anatomy, repositioning for the multiple beams takes long time and is based on the planner's experience to find an optimal set of beams. So, establishing a small set of standardized beam bouquets for planning could be of valuable help. The set of beam bouquets could be determined by learning the beam configuration features from previous IMRT datasets. The multiple beams are fixed on a circle in the transverse plane around the patient. Consequently, an observation is composed of the k beams of a patient, that is k circular measurements. A real data set from post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France, is represented in Figure 1. One actual observation consists of a (non-ordered) set of k angles rather than of a vector (ordered) of length k but to cope with the technical difficulty of dealing with sets, it is convenient to store the angles of each patient in a vector in increasing order (or in any other given order). Of course, the derived vectors may be very different even for similar sets of angles. This is easily seen by considering a simple case of two patients with angles $\{1°, 60°, 100°, 150°, 180°\}$ and $\{60°, 100°, 150°, 180°, 359°\}$: the two patients should share the same cluster as the sets of angles are very similar (modulo 360) although the derived vectors are very different and, if any classical clustering method was applied, are not likely to share the same cluster.

Abraham et al. (2013) proposed a first tool to assist the selection of beam orientations to enhance the therapist's experience. A suitable distance on the circle was defined and, for a fixed number of clusters, an algorithm based on simulated annealing was proposed. Yuan et al. (2015) generalized the precedent approach using k-medoids to cluster beam configuration features with different numbers of beams. These methods suffer from some major flaws. First, the number of clusters has to be supplied by the user. An additional

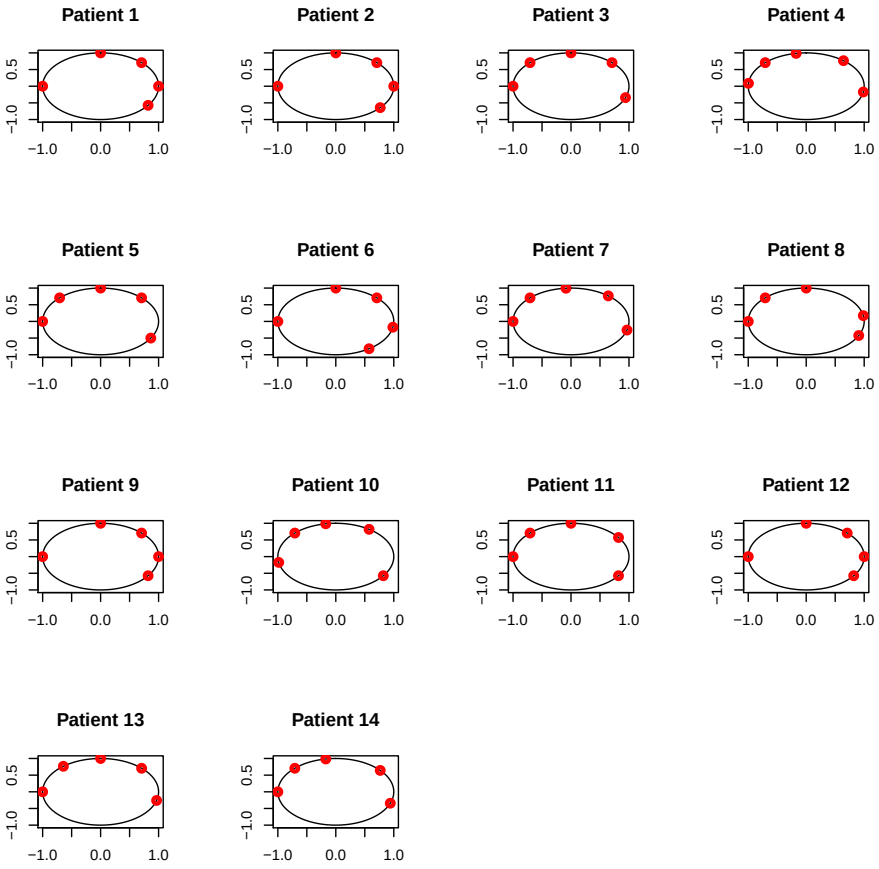

Figure 1: Real data set of 14 patients with $k = 5$ angles. A point on the circle represents the location of a treatment beam.

procedure of model selection (AIC, BIC, RIC, silhouette index, ...) can be used to select the number of clusters but an appropriate methodology that automatically finds this number would be very useful. Second, the final result is only one unique clustering whereas there are probably other clusterings that could be acceptable. A final result with all possible clusterings and a probability of appearance for each could be of great help for the practitioner. These problems can naturally be solved with a Bayesian clustering method based on Dirichlet Process as it does not require a preselected number of clusters and provides different clusterings (possibly with different numbers of clusters) with their posterior probabilities. Also note that the Bayesian framework is well adapted to our application as the sample size is low and can be compensated to some extent by prior information.To our knowledge, such a clustering Bayesian model has never been applied for multivariate circular data in the literature.

Circular data have first been studied using classical non-Bayesian approaches. Three main models for circular data can be found in the litterature: the von-Mises distributions, the wrapped distributions and the projected normal distributions. The von-Mises distributions, first introduced by Von Mises (1918) and extended by Singh et al. (2002) and Mardia et al. (2008), are the natural analogues of the normal distribution on the sphere. The wrapped distributions (Mardia and Jupp, 2009) are based on a simple fact that a probability distribution on a circle can be obtained by wrapping a probability distribution defined on the real line. Projected normal distributions are obtained by projecting multivariate normal random variables radially onto the sphere (Presnell et al., 1998). These latter distributions allow for asymmetric and possible bimodal models. We refer the reader to Mardia and Jupp (2009) for a complete review on probability distributions of circular data.

Bayesian litterature on circular data is more recent. Von Mises distributions are used in the univariate case in Damien and Walker (1999) and are applied to a change-point problem in SenGupta and Laha (2008). Wrapped distributions appear in Ravidran and Ghosh (2011), with a data augmentation algorithm to overcome some computational difficulties, and in Jona-Lasinio et al. (2012), to handle structured dependences between spatial measurements. Nuñez-Antonio and Gutiérrez-Peña (2005), Wang and Gelfand (2013) adapted the projected normal distributions in a Bayesian framework. A more sophisticated model was considered in Wang and Gelfand (2014) to capture structured spatial dependence for modeling directional data at

different spatial locations. This model was then upgraded to capture joint structured spatial and temporal dependence (Wang et al., 2015).

Note that, for all the models cited above, each observation is simply a point on a circle or on a sphere while in our case, a single observation is made up of k ($k \geq 2$) non-ordered points on the circle. For this reason these models cannot straightforwardly be adapted to our dataset. We propose an extension of the projected normal distribution to our data. This extension does not reduce to a simple projection of a multivariate normal distribution but enables us to model the multivariate and the non-ordered features of our data. We also provide an informative prior distribution on the relative locations of the k angles on the circle. This prior distribution expresses that the k angles are a priori regularly spaced on the circle. A new parameter is also introduced to control the variability of the angles on the circle. Inference on the variabiliy of the angles is of particular interest for a clustering purpose as an inadequate value of this parameter can alter the final results. The projected normal distribution is then associated with a Dirichlet process to perform clustering. Therefore, the proposed method includes an automated selection for the number of clusters.

In the present paper, the Bayesian model is described in the next section. Section 3 is devoted to the inference of the parameters of the model. Section 4 provides a theoretical study to highlight the adaptability of our model to the multivariate and the non-ordered features of the data. Section 5 and 6 provide empirical results first on simulated data and then on the real data set that motivated the present work. A short conclusion is given in Section 7.

2 Model

A simple way of generating distributions on the p-dimensional unit sphere $\mathcal{S}^p$ is to radially project probability distributions originally defined on the p-dimensional space $\mathbb{R}^p$ (Presnell et al., 1998). Let x be a random p-dimensional vector, then $x/||x||$ is a random point on $\mathcal{S}^p$. If x has a p-variate Normal distribution $N_p(\mu, \Sigma)$ then $x/||x||$ is said to have a projected normal distribution, denoted by $PN_p(\mu, \Sigma)$. The literature has been first confined to the special case where $p = 2$ and $\Sigma = \mathbf{I}$ (Presnell et al., 1998; Nuñez-Antonio and Gutiérrez-Peña, 2005; Nuñez-Antonio et al., 2011). Then, Wang and Gelfand (2013) studied the projected normal family with a general covariance matrix

Σ and refer to this richer class $PN_p(\mu, \Sigma)$ as the general projected normal distribution. This general version allows asymmetry and bimodality (see Figure 2. in Wang and Gelfand, 2014). The general projected normal distribution is not identifiable because $x/||x||$ is invariant to scale transformation. To overcome this problem Wang and Gelfand (2013) fixed some variance parameters in Σ to provide identifiability.

In a first step of simplification, we assume that the ith of the n observations is given by a vector of k angles $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})' \in [0, 2\pi[^k$ instead of a non-ordered set $\{\theta_{i1}, \ldots, \theta_{ik}\}$. Using a projected normal distribution, we denote by $x_i = (x_{i1}, \ldots, x_{ik})' \in (\mathbb{R}^2)^k$ a random vector with distribution $N_{2k}(\mu_i, I_{2k})$ where θ_{ij} is defined as the radial projection of x_{ij} on the unit circle of $\mathbb{R}^2$. In other words, we have $x_{ij} = (x_{ij1}, x_{ij2})' = (r_{ij}\cos\theta_{ij}, r_{ij}\sin\theta_{ij})'$ for all $i \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, k\}$ where r_{ij} denotes the Euclidean norm of x_{ij}. Note that θ_i is observed while $r_i = (r_{i1}, \ldots, r_{ik})'$ is not and is treated as an unknown parameter. We denote by $PN_{2k}(\mu_i, I_{2k})$ the joint distribution of (θ_i, r_i). Clustering analysis will be based on a Dirichlet process mixture (DPM) model described as follows:

$$\begin{array}{rcl} \theta_i, r_i|\mu & \sim & PN_{2k}(\mu_i, I_{2k}) \\ \mu_i|P & \sim & P \\ P & \sim & DP(n_0 P_0), \end{array} \tag{1}$$

where $\mu = (\mu_1, \ldots, \mu_n)$ and where $DP(n_0P_0)$ denotes the Dirichlet process (DP) introduced by Ferguson (1973) with center $P_0 = N_{2k}(0, \Sigma_0)$ and precision parameter n_0. The clustering properties of the DP are well known and date back to Blackwell and MacQueen (1973). It is shown that the parameter $\mu = (\mu_1, \ldots, \mu_n)$ follows the Pólya urn scheme:

$$\begin{array}{rcl} \mu_1 & \sim & P_0 \\ \mu_{i+1}|\mu_1, \ldots, \mu_i & \sim & \frac{1}{n_0+i}\sum_{j=1}^{i}\delta_{\mu_i} + \frac{n_0}{n_0+i}P_0, \text{ for } i \geq 2. \end{array} \tag{2}$$

with δ_{μ_i} indicating the point measure on μ_i. So, μ_{i+1} may be equal to one of the previous μ_i's or may be drawn from P_0. This results in a positive probability of sharing the parameter value with previous observations; hence the clusters. In the sequel, we will denote by Pólya(n_0P_0) the distribution of μ given by (2). Although the DPM is very popular for Bayesian clustering, other model-based cluster methods exist. For a review of these methods, we refer the reader to Quintana (2006); Lau and Green (2007); Fritsch and

Ickstadt (2009) and references therein. Note that the DPM model does not require choosing the number of clusters. On the other hand, it is well known that the number of clusters can be controlled by n_0. Learning about n_0 from the data may be addressed by assuming a Gamma prior distribution $n_0 \sim G(a_{n_0}, b_{n_0})$ (Escobar and West, 1995).

Now recall that the actual ith observation consists of a (non ordered) set of the form $\{\theta_{i1}, \ldots, \theta_{ik}\}$ rather than of a vector (ordered) $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})'$. The impact of this simplification is quite easy to understand. Using model (1), two observations i_1 and i_2 with the same angles but in different orders would have a very low posterior probability of sharing the same cluster, that is $\mu_{i1} = \mu_{i2}$. We treat the observations as vectors for convenience but we have to introduce a permutation parameter τ_i to compensate this simplification. More precisely, for all $\mu_i = (\mu'_{i1}, \ldots, \mu'_{ik})'$ and all permutation τ_i of $\{1, \ldots, k\}$, we set $\mu_i^{\tau_i} = (\mu'_{i\tau_i(1)}, \ldots, \mu'_{i\tau_i(k)})'$; $\mu_i^{\tau_i}$ can be viewed as a random permutation of the coordinates of μ_i. Therefore, the clustering model becomes:

$$\begin{array}{rcl} \theta_i, r_i|\mu, \tau & \sim & PN_{2k}(\mu_i^{\tau_i}, I_{2k}) \\ \mu_i|P & \sim & P \\ P & \sim & DP(n_0 P_0), \end{array} \tag{3}$$

where $\tau = (\tau_1, \ldots, \tau_n)$ and $\mu = (\mu_1, \ldots, \mu_n)$. The permutations τ_i are assumed to be a priori independent with a uniform distribution $\mathcal{U}_{\mathcal{P}}$ on the set $\mathcal{P}$ of permutations of $\{1, \ldots, k\}$. The posterior probability that two observations i_1 and i_2 with the same angles but in different orders would share the same cluster is increased with model (3) as there exist some values of τ_{i1} and τ_{i2} such that $\mu_{i1}^{\tau_{i1}} = \mu_{i2}^{\tau_{i2}}$. A theoretical study of the impact of the symmetry introduced by τ_i is given in Section 4.

Prior information It is natural to assume that the k angles $\theta_{i1}, \ldots, \theta_{ik}$ are a priori roughly equally spaced on the unit circle. This prior information can be incorporated into the covariance matrix Σ_0 of P_0 as follows. From (3), it is well known that the marginal distribution of μ_i is $P_0 = N_{2k}(0, \Sigma_0)$. Denote by R the 2×2-matrix of the rotation in $\mathbb{R}^2$ with angle $2\pi/k$ and center 0. Set $\mu_{i1} \sim N_2(0, \rho I_2)$ where ρ is a positive number and $\mu_{ij}|\mu_{i,j-1} \sim N_2(R\mu_{i,j-1}, I_2)$ for $j \in \{2, \ldots, k\}$. Then, roughly, $\mu_{i1}, \ldots, \mu_{ik}$ are approximately equally spaced on the circle with center 0 and radius $\sqrt{\rho}$. Note that the variance parameter ρ has an important impact on the prior: a large value of ρ enables to generate $\mu_{i1}, \ldots, \mu_{ik}$ approximately situated on a circle with a large radius.

For such a large radius, the angles θ_{ij} of the projections on the unit circle have small variances. Hence, ρ can also be viewed as a precision parameter for θ_i (see Subsection 5.1 and Figure 2). It is shown in the Appendix that the derived matrix Σ_0, also denoted by $\Sigma_0(\rho)$ in the sequel to highlight the dependence on ρ, can be expressed as a closed-form expression as well as the inverse Σ_0^{-1} and the determinant $|\Sigma_0|$. Inference on ρ can then be performed using an inverse gamma prior $\rho \sim IG(a_\rho, b_\rho)$ for which the full posterior conditional distribution will be calculated in the following section.

Finally, the complete Bayesian model can be expressed as follows:

$$\begin{array}{rcl} \theta_i, r_i|\mu, \tau & \sim & PN_{2k}(\mu_i^{\tau_i}, I_{2k}) \\ \mu|n_0, \rho & \sim & \text{Pólya}(n_0 P_0(\rho)) \\ \tau_i & \sim & \mathcal{U}_{\mathcal{P}} \\ \rho & \sim & IG(a_\rho, b_\rho) \\ n_0 & \sim & G(a_{n_0}, b_{n_0}). \end{array} \tag{4}$$

where $P_0(\rho) = N_{2k}(0, \Sigma_0(\rho))$. By convention, it is assumed that the random variables at a stage of the hierarchy are independent.

3 Inference

We set $\theta = (\theta_1, \ldots, \theta_n)$, $r = (r_1, \ldots, r_n)$, $\mu = (\mu_1, \ldots, \mu_n)$, $\tau = (\tau_1, \ldots, \tau_n)$ and $\xi = (r, \mu, \tau, \rho, n_0)$. Thus, the parameter is ξ and the observation is θ. We sample from the posterior distribution of ξ with a Metropolis-Hastings-Within-Gibbs algorithm. In what follows, p stands for a generic notation for a density distribution.

Simulations of μ We can restrict our attention to model (3) instead of the full model (4) for the simulations of μ as every component of ξ except μ remains fixed. An alternative parameter setting of μ, θ and ρ will prove useful. Denote $x = (x_1, \ldots, x_n)$ where $x_i = (x'_{i1}, \ldots, x'_{ik})'$. Firstly, note that the full conditional distribution of μ reduces to the conditional distribution of μ given (x, n_0, ρ, τ) as there is a natural bijection between x_{ij} and (θ_{ij}, r_{ij}). Secondly, if we denote by $N_{2k}(x_i; \mu_i, I_{2k})$ the value of the density of $N_{2k}(\mu_i, I_{2k})$ at x_i, it is easy to check that:

$$N_{2k}(x_i; \mu_i^{\tau_i}, I_{2k}) = N_{2k}(x_i^{\tau_i^{-1}}; \mu_i, I_{2k}). \tag{5}$$

Consequently, if we set $y_i = x_i^{\tau_i^{-1}}$, sampling from the posterior distribution of μ in the DPM model (3) reduces to sampling from the posterior distribution of μ in the following conjugate DPM model:

$$\begin{array}{rcl} y_i|\mu & \sim & N_{2k}(\mu_i, I_{2k}) \\ \mu_i|P & \sim & P \\ P & \sim & DP(n_0 P_0). \end{array} \tag{6}$$

There are several samplers for conjugate DPM models; for a review, we refer the reader to MacEachern (1998); Neal (2000); Griffin and Holmes (2010). Following the notations of Dahl (2003), we use a parameter setting of μ in terms of:

- a set partition $\eta = \{S_1, \ldots, S_q\}$ for $\{1, \ldots, n\}$ where each S_j represents a cluster, i.e., $\mu_i = \mu_j$ if there exists $j_1 \in \{1, \ldots, q\}$ such that $i, j \in S_{j_1}$ and $\mu_i \neq \mu_j$ if there exist $j_1, i_1 \in \{1, \ldots, q\}$, $i_1 \neq j_1$ such that $i \in S_{i_1}$, $j \in S_{j_1}$,
- a vector $\phi = (\phi_1, \ldots, \phi_q)$ composed of the distinct values of μ, i.e., $\phi_j = \mu_i$ for all $i \in S_j$.

Then, the conjugate DPM model (6) may be expressed as:

$$\begin{array}{rcl} y_i|\eta, \phi & \sim & N_{2k}(\sum_{j=1}^q \phi_j \mathbf{1}_{\{i \in S_j\}}, I_{2k}) \\ \phi_j|\eta & \sim & P_0 \\ \eta & \sim & p(\eta) \propto \prod_{i=1}^q n_0 \Gamma(|S_j|), \end{array} \tag{7}$$

where $|S_j|$ is the cardinal of S_j, $\mathbf{1}_A$ is the indicator function for the event A, Γ denotes the gamma function and p stands for the generic notation for any density. We can integrate over the cluster location parameter ϕ analytically in (7) as P_0 is conjugate to the normal distribution of y_i given η and ϕ. Then, we run the SAMS sampler of Dahl (2003) for simulating η. This sampler may improve the merge-split sampler initially proposed by Jain and Neal (2004). Once a simulation of η is obtained, it is easy to simulate the cluster location parameter ϕ from its full conditional which reduces to sample independently each ϕ_j from a $N_{2k}(\Sigma_j \sum_{i \in S_j} y_i/|S_j|, \Sigma_j)$ distribution with $\Sigma_j^{-1} = |S_j|^{-1} I_{2k} + \Sigma_0^{-1}(\rho)$. As recommended by the previous authors, we combine three runs of the Metropolis-Hastings update of the SAMS sampler with a full scan of Gibbs sampling for μ (see MacEachern, 1994, for a presentation of this particular Gibbs sampler). Some details of the SAMS and the Gibbs samplers used in this article are given in the Appendix.

Simulations of r It is shown in the Appendix that the r_{ij} are independent given $(\theta, \tau, \mu, \rho, n_0)$ with density:

$$p(r_{ij}|\theta, \tau, \mu, \rho, n_0) \propto r_{ij} e^{-\frac{1}{2}\left(r_{ij} - u'_{ij}\mu_{i\tau_i(j)}\right)^2}, \tag{8}$$

with $u'_{ij} = (\cos\theta_{ij}, \sin\theta_{ij})$. If we denote by $N_1^+(m, v)$ the univariate normal distribution truncated to $[0, \infty)$, we remark that (8) is close to the value of the density of $N_1^+(u'_{ij}\mu_{i\tau_i(j)}, 1)$ at r_{ij}. It is then natural to simulate from (8) by a Metropolis-Hastings step with a $N_1^+(u'_{ij}\mu_{i\tau_i(j)}, 1)$ as the proposal distribution. Clearly, the probability of acceptance reduces to the ratio $\min\{r_{ij}^{new}/r_{ij}^{old}, 1\}$ where r_{ij}^{old} and r_{ij}^{new} are, respectively, the current and the proposed values of r_{ij} in the algorithm.

Simulations of τ As the prior distribution of τ is uniform, we have:

$$\begin{aligned} p(\tau|\theta, r, \mu, \rho, n_0) &= p(\tau|x, \mu) \\ &\propto p(x|\tau, \mu) \\ &\propto \prod_{i=1}^{n} N_{2k}(x_i; \mu_i^{\tau_i}, I_{2k}). \end{aligned}$$

Thus, given $(\theta, r, \mu, \rho, n_0)$, the τ_i are independent with density (with respect to the counting measure on the set T of permutations of $\{1, \ldots, k\}$):

$$p(\tau_i|x, \mu) = \frac{N_{2k}(x_i; \mu_i^{\tau_i}, I_{2k})}{\sum_{t\in T} N_{2k}(x_i; \mu_i^{t}, I_{2k})}. \tag{9}$$

Simulations of ρ From (4), it is clear that the full conditional distribution of ρ reduces to the conditional distribution of ρ given μ. Then, using the parametrization of μ in terms of (η, ϕ) and (7), we note that η and ρ are independent and that:

$$\begin{aligned} p(\rho|\theta, r, \mu, \tau, n_0) &= p(\rho|\eta, \phi) \\ &\propto p(\phi|\eta, \rho)p(\rho|\eta) \\ &\propto \left(\prod_{j=1}^{q} p(\phi_j|\rho)\right) p(\rho). \end{aligned} \tag{10}$$

We show in the Appendix that $|\Sigma_0^{-1}(\rho)| = \rho^{-2}$ and that the components of the matrix $\Sigma_0^{-1}(\rho)$ are independant (constant) of ρ except the components

of the first 2 by 2 diagonal submatrix (lines and columns 1 and 2). As this submatrix is equal to $(\rho^{-1}+(k-1))I_2$, it is easily seen that

$$\begin{aligned}\phi_i'\Sigma_0^{-1}(\rho)\phi_i &= (\rho^{-1}+(k-1))\phi_{i1}'\phi_{i1}+constant\\ &= \rho^{-1}\phi_{i1}'\phi_{i1}+constant.\end{aligned}$$

where *constant* stands for a generic notation for an expression independent of ρ. Since $\phi_j|\rho \sim P_0(\rho) = N_{2k}(0,\Sigma_0(\rho))$ and $\rho \sim IG(a_\rho, b\rho)$, we have:

$$\prod_{j=1}^{q} p(\phi_j|\rho) \propto \rho^{-q} e^{-\frac{1}{2}\rho^{-1}\sum_{j=1}^{q}\phi_{i1}'\phi_{i1}},$$

and it is easy to conclude from (10) that the full conditional of ρ is

$$IG\left(a_\rho+q, b_\rho+\frac{1}{2}\sum_{i=1}^{q}\phi_{i1}'\phi_{i1}\right). \tag{11}$$

Simulations of n_0 Using the arguments of Escobar and West (1995), under the $G(a_{n_0}, b_{n_0})$ prior, n_0 is updated at each Gibbs iteration by sampling first an additional variable ζ from a Beta distribution and then a new value of n_0 from a mixture of Gamma distributions as follows:

$$\begin{aligned}\zeta|n_0 &\sim B(n_0+1, n)\\ n_0|\zeta, q &\sim \pi_n G(a_{n_0}+q, b_{n_0}-\log\zeta)+(1-\pi_n)G(a_{n_0}+q-1, b_{n_0}-\log\zeta),\end{aligned} \tag{12}$$

with weights π_n defined by $\pi_n/(1-\pi_n) = (a_{n_0}+q-1)/[n(b_{n_0}-\log\zeta)]$.

The whole procedure is summarized in Algorithm 1.

4 Theoretical study of the symmetrized model

To investigate the impact of the symetrization induced by the variables τ_i, we consider a simple model of the following form:

$$\begin{aligned}x_i|\eta,\phi &\sim N_{2k}(\textstyle\sum_{j=1}^{q}\phi_j\mathbf{1}_{\{i\in S_j\}}, I_{2k})\\ \phi_j|\eta &\sim P_0\\ \eta &\sim G\end{aligned} \tag{I}$$

Algorithm 1

Require: Data set $\theta = (\theta_1, \ldots, \theta_n)$.
Require: Hyperparamaters $a_\rho, b_\rho, a_{n_0}, b_{n_0}$.
Repeat :

1. Simulate η.
   (a) Run the SAMS sampler three times.
   (b) Run the Gibbs sampler.
2. Simulate $\phi_j \sim N_{2k}(\Sigma_j \sum_{i \in S_j} y_i/|\Sigma_j|, \Sigma_j)$ for each cluster j.
3. Propose $r_{ij}^{new} \sim N_1^+(u_{ij}' \mu_{i\tau_i(j)}, 1)$, accept with probability $\min(r_{ij}^{new}/r_{ij}^{old}, 1)$.
4. Simulate new τ_i from 9.
5. Simulate new ρ from 11.
6. Simulate n_0 from 12.

and its symmetrized version :

$$\begin{array}{rcl} x_i|\eta,\phi & \sim & N_{2k}(\sum_{j=1}^{q}\phi_j^{\tau_i}\mathbf{1}_{\{i\in S_j\}}, I_{2k}) \\ \phi_j|\eta & \sim & P_0 \\ \eta & \sim & G \\ \tau_i & \sim & \mathcal{U}_{\mathcal{P}}, \end{array} \qquad \text{(II)}$$

where $\phi_j^{\tau_i} = (\phi'_{j\tau_i(1)}, \ldots, \phi'_{j\tau_i(k)})'$ is obtained by random permutation of the coordinates of $\phi_j = (\phi'_{j1}, \ldots, \phi'_{jk})' \in (\mathbb{R}^2)^k$. In both models, $P_0 = N_{2k}(0, \Sigma_0)$ and G is any distribution of the partition $\eta = \{S_1, \ldots, S_q\}$ of $\{1, \ldots, n\}$. Such distributions include the distribution derived from the Dirichlet process given by (7). Model (II) can be viewed as a simplified and reparametrized version of (4). Now consider an idealized sample $x_1, \ldots, x_n$ for which every observation x_i is simply a random permutation of one unique observation $x_0 = (x'_{01}, \ldots, x'_{0k})' \in (\mathbb{R}^2)^k$; in other words, for every i, there exists a permutation α_i such that $x_i = (x'_{0\alpha_i(1)}, \ldots, x'_{0\alpha_i(k)})'$. As the coordinates x_{ij} of all the x_i are the same but in a different order, it is expected that all the observations are put together in one unique cluster. The aim of this section is to study whether model (II) is more appropriate than model (I) for this purpose.

Let p_0 and $p_{\mathrm{I}}(x|\eta)$ denote respectively the density of P_0 and the conditional density of $x = (x_1, \ldots, x_n)$ given η for model (I). We have:

$$\begin{array}{rcl} p_{\mathrm{I}}(x|\eta) & = & \displaystyle\int \prod_{j=1}^{q} \prod_{i\in S_j} N_{2k}(x_i; \phi_j, I_{2k}) p_0(\phi_j) d\phi_j \\ & = & \displaystyle\prod_{j=1}^{q} m(x_{S_j}), \end{array}$$

where $x_{S_j} = (x_i, i \in S_j)$ and

$$m(x_{S_j}) = \int \prod_{i\in S_j} N_{2k}(x_i; \phi_j, I_{2k}) p_0(\phi_j) d\phi_j.$$

Denote by $p_{\mathrm{II}}(x|\eta)$ the conditional density of x given η for model (II). By

(5) and noting that $\{\tau_i^{-1}, \tau_i \in \mathcal{P}\} = \mathcal{P}$, we have:

$$\begin{aligned} p_{\mathrm{II}}(x|\eta) &= \sum_\tau \frac{1}{(k!)^n} \int \prod_{j=1}^q \prod_{i \in S_j} N_{2k}(x_i; \phi_j^{\tau_i}, I_{2k}) p_0(\phi_j) d\phi_j \\ &= \sum_\tau \frac{1}{(k!)^n} \int \prod_{j=1}^q \prod_{i \in S_j} N_{2k}(x_i^{\tau_i}; \phi_j, I_{2k}) p_0(\phi_j) d\phi_j \\ &= \frac{1}{(k!)^n} \sum_\tau \prod_{j=1}^q m(x_{S_j}^\tau), \end{aligned}$$

where the sum above is taken for all the values of $\tau = (\tau_1, \ldots, \tau_n)$ in $\mathcal{P}^n$, $x_{S_j}^\tau = (x_i^{\tau_i}, i \in S_j)$ and $x_i^{\tau_i} = (x'_{i\tau_i(1)}, \ldots, x'_{i\tau_i(k)})'$. Therefore, models (I) and (II) reduce to

$$\begin{aligned} x|\eta &\sim \textstyle\prod_{j=1}^q m(x_{S_j}) \\ \eta &\sim G, \end{aligned} \tag{I'}$$

and

$$\begin{aligned} x|\eta &\sim \tfrac{1}{(k!)^n} \textstyle\sum_\tau \prod_{j=1}^q m(x_{S_j}^\tau). \\ \eta &\sim G. \end{aligned} \tag{II'}$$

For all partition $\eta = \{S_1, \ldots, S_q\}$ and all observation x, we set

$$f(x, \eta) = \frac{1}{(k!)^n} \sum_{\tau \in \mathcal{P}^n} \exp \frac{1}{2} \sum_{j=1}^q \left(\| \sum_{i \in S_j} x_i^{\tau_i} \|_{S_j}^2 - \| \sum_{i \in S_j} x_i \|_{S_j}^2 \right) \tag{13}$$

where $\Sigma_S = \left(\Sigma_0^{-1} + |S| I_{2k}\right)^{-1}$ for all subset $S \subset \{1, \ldots, n\}$ and $\|t\|_S^2 = t' \Sigma_S t$ for all $t \in (\mathbb{R}^2)^k$.

Proposition 1. *a) For all partition $\eta = \{S_1, \ldots, S_q\}$ and all observation $x = (x_1, \ldots, x_n)$, we have:*

$$\frac{p_{\mathrm{II}}(x|\eta)}{p_{\mathrm{I}}(x|\eta)} = f(x, \eta).$$

b) For all distribution G, there exists a positive number B_G such that:

$$\frac{p_{\mathrm{II}}(\eta|x)}{p_{\mathrm{I}}(\eta|x)} = B_G \ f(x, \eta),$$

for all partition η and all observation x.

c) For all distribution G, all partition η and all observation x, we have:

$$\frac{p_{\mathrm{II}}(\eta|x)}{p_{\mathrm{I}}(\eta|x)} \geq f(x,\eta)\ \frac{1}{\max_\eta f(x,\eta)} \tag{14}$$

where the maximum is taken over all partitions of $\{1,\ldots,n\}$.

From a) of Proposition 1, we see that $f(x,\eta)$ is the likelihood ratio of models (II') and (I'). From b), we know that the posterior odds ratio is large when $f(x,\eta)$ is large. It would be of interest to know whether this ratio is greater than one. Unfortunately, this is not an easy task except for a few particular cases given below. Indeed, although the factor B_G is actually known (see the proof of Proposition 1 in the Appendix), it is rather intractable. From c), we deduce that the posterior odds is actually greater or equal to one at least for the partition η_x that maximizes $f(x,\eta)$. This partition does exist for any observation x and is independent of G. In other words, for any x, there exists a partition η_x such that $p_{\mathrm{II}}(\eta_x|x) \geq p_{\mathrm{I}}(\eta_x|x)$ for all prior G. Finally, we can remark from the proof of the theorem that the equality in (14) is obtained when G is a Dirac distribution; a meaningless prior.

Consider the partition $\bar{\eta}$ with a single cluster: $q=1$ and $S_1=\{1,\ldots,n\}$. From (13), the posterior odds ratio when $\eta=\bar{\eta}$ is likely to be large when $\sum_{i=1}^n x_i \approx 0$ and small when all the $x_i \approx x_0$ for all $i \in \{1,\ldots,n\}$. Assume from now that $\sum_{i=1}^n x_i = 0$ and that $\Sigma_0 = I_{2k}$. Remenber that Σ_0 models the prior information about the mutual positions of the angles on the circle. Therefore $\Sigma_0 = I_{2k}$ can be viewed as a non informative prior. In this case, $\|t\|_{S_j}^2 = (1+|S_j|)^{-1}t't = (1+|S_j|)^{-1}\|t\|$ for all $t \in (\mathbb{R}^2)^k$ and we have:

$$f(x,\bar{\eta}) = \frac{1}{(k!)^n}\sum_{\tau\in\mathcal{P}^n} \exp\frac{1}{2(n+1)}\left(\|\sum_{i=1}^n x_i^{\tau_i}\|^2\right). \tag{15}$$

Example 1 below provides a typical sample $x=(x_1,\ldots,x_n)$ for which the posterior probabilty of a unique cluster is greater with model (II) than with model (I) independently of the prior distribution G.

Example 1. *First, by noting that:*

$$\sum_{l=0}^{n} e^{il\theta} = \frac{\sin\frac{\theta(n+1)}{2}}{\sin\frac{\theta}{2}} e^{i\frac{\theta}{2}n},$$

for all $\theta \in \mathbb{R}$, we deduce that:

$$\sum_{l=0}^{k-1} \cos\left(\frac{2\pi l}{k}\right) = 0 \text{ and } \sum_{l=0}^{k-1} \sin\left(\frac{2\pi l}{k}\right) = 0. \tag{16}$$

Assume $n = k$ and set $x_{ij} = (\cos(i+j-2)2\pi/k, \sin(i+j-2)2\pi/k)'$ for $i \in \{1,\ldots,k\}$ and $j \in \{1,\ldots,k\}$. In other words, $x_1 = (x'_{11},\ldots,x'_{1k})' \in (\mathbb{R}^2)^k$ is made up of k consecutive points on the unit circle separated from an angle of $2\pi/k$, x_2 is obtained by a rotation with angle $2\pi/k$ of each point of x_1 and so on. Therefore, it is easy to see from (16) that $\sum_{i=1}^n x_i = 0$. Our conjecture is that $\max_\eta f(x,\eta) = f(x,\bar{\eta})$ for all integer k which implies, from c) of Proposition 1, that the probability of a unique cluster is greater for model (II) than for model (I) for any distribution G. For $n = k = 2$ the conjecture reduces to $f(x,\eta) \leq f(x,\bar{\eta})$ for a single partition $\eta = \{\{x_1\},\{x_2\}\}$. As $\|x_i\|_{S_j} = \|x_i^{\tau_i}\|_{S_j}$ for all i and τ_i, it is easily seen from (13) that $f(x,\eta) = 1$. On the other hand, as $\|x_1\|^2 = k$ and $x_1 = -x_2$, we see from (15) that

$$\begin{aligned} f(x,\bar{\eta}) &= \frac{1}{4}\left(2\exp\frac{1}{6}\|x_1+x_2\|^2 + 2\exp\frac{1}{6}\|2x_1\|^2\right) \\ &= \frac{1}{2}\left(1+2\exp\frac{4}{3}\right), \end{aligned}$$

hence the proof of the conjecture for $n = k = 2$. We also proved the conjecture for $n = k = 3$ with a rather large amount of calculations (not given here) to take into account all the partitions η and all the permutation $\tau = (\tau_1,\tau_2,\tau_3)$. We are not in a position to provide general proof of the conjecture for $n = k \geq 4$.

5 Simulations

With small data sets, a misspecification of the prior could have a strong negative impact on the final results. Therefore, special attention has to be paid to the prior and the hyperparameter specifications. Consequently, we test our algorithm on two simulation studies to evaluate the influence of some hyperparameters.

The performances of our method are investigated using the Adjusted Rand Index (ARI), proposed by Hubert and Arabie (1985), to compare our

obtained partition to the actual one. The Rand Index (Rand, 1971) is a well known measure of the similarity between two partitions. If we denote by N_{00} the numbers of pairs that are in the same cluster in both partitions and by N_{11} the number of pairs that are in different clusters in both partitions, then the Rand Index is defined by the ratio $(N_{00} + N_{11})/\binom{n}{2}$. The ARI is a corrected-for-chance version of the Rand index. Its expected value (under the generalized hypergeometric model) is equal to 0 and its maximum is 1 while the expected value of the Rand Index depends on the number of clusters. For a presentation of the different criteria for clustering comparison and for a study investigating the usefulness of the adjusted measures, we refer the reader to Fritsch and Ickstadt (2009) and Nguyen et al. (2009).

5.1 Influence of the Precision Parameter ρ

First we choose to simulate data using a procedure which is close to our model in order to investigate the influence of the precision parameter ρ. We set $q = 3$ clusters of 10 data. We simulate the coordinates μ_{ij} of each center μ_i on the circle with fixed radius ρ. The first coordinate μ_{i1} is simulated according to a uniform distribution on the circle with radius ρ. The other coordinates μ_{ij}, $j = 2, \ldots, 5$ $(k = 5)$ are generated according to a noisy rotation with angle $2\pi j/5$ of μ_{i1}. For each cluster i, we generate 10 data according to $PN_{10}(\mu_i, \mathbf{I}_{10})$. A comparison of the generated data is provided in Figure 2 with different values for ρ; for the clarity of the picture we choose to represent only $q = 2$ clusters of 5 observations. It is clear from Figure 2 that large values of ρ provide small variability for the projected observations. According to this remark we choose a noninformative prior for ρ by setting $a_\rho = b_\rho = 0.01$.

5.2 Robustness to the Hyperparameters a_{n_0} and b_{n_0}

It is well-known that the number of clusters does depend on n_0 whose prior distribution is fixed by the hyperparameters a_{n_0} and b_{n_0}. In this subsection we investigate the sensitivity of the ARI with respect to these hyperparameters. We apply the same simulation strategy as in the previous subsection with a fixed $\rho = 20$. Note that the parameters a_{n_0} and b_{n_0} are not at all involved in the simulation of the dataset. The mean values for the ARI over 100 simulated data sets are given in Table 1.

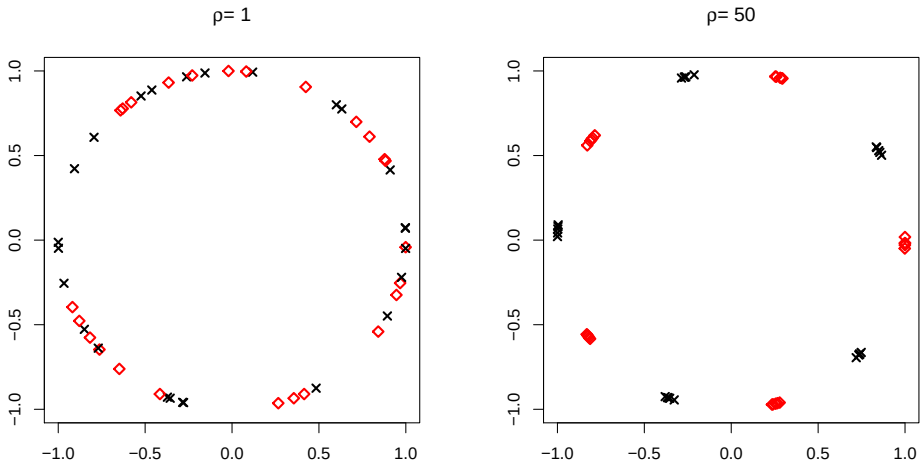

Figure 2: Two data sets are generated with two different values for the parameter ρ to highlight the influence of this parameter. Two clusters of five data are represented on each plot. Each data is composed of $k = 5$ angles on the circle. One cluster is represented by the black cross, the other by the red square.

Table 1: Adjusted Rand Index (Proportion of clustering with the actual number of clusters) according to a_{n_0} and b_{n_0}.

	$b_{n_0} = 0.1$	$b_{n_0} = 1$	$b_{n_0} = 10$	$b_{n_0} = 100$	$b_{n_0} = 1000$
$a_{n_0} = 0.1$	0.73 (0.80)	0.71 (0.79)	0.62 (0.72)	0.63 (0.75)	0.59 (0.67)
$a_{n_0} = 1$	0.76 (0.91)	0.72 (0.84)	0.65 (0.79)	0.67 (0.76)	0.64 (0.71)
$a_{n_0} = 10$	0.72 (0.76)	0.78 (0.96)	0.69 (0.84)	0.67 (0.80)	0.65 (0.74)
$a_{n_0} = 100$	0.70 (0.70)	0.68 (0.79)	0.79 (0.92)	0.72 (0.82)	0.62 (0.75)
$a_{n_0} = 1000$	0.66 (0.69)	0.62 (0.72)	0.68 (0.79)	0.75 (0.88)	0.65 (0.76)

Table 1 suggests that a choice of a_{n_0}/b_{n_0} approximately between 1 and 10 provides good and similar results.

6 Real Data

We then apply the methodology to a real data set from post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France (see Figure 1 and Table 2). Let us recall that no other competing methods exist for these kind of multivariate circular data except the method described in Abraham et al. (2013) with a fixed number of clusters. Consequently, our results are compared to those of Abraham et al. (2013) in which the number of clusters was preselected to $q = 2$.

Let us remind you that the a priori distribution of n_0 is a gamma distribution with parameter a_{n_0} and b_{n_0} with an expected value equal to a_{n_0}/b_{n_0} (if $a_{n_0} > 1$) and a variance equal to $a_{n_0}/b_{n_0}^2$ (if $a_{n_0} > 2$). Remember that the expected number of clusters given n_0 is approximately equal to $n_0 \log(1 + n/n_0)$ (Teh, 2010). According to the results of Section 5, the results are robust with respect to the choice of the hyperparameters a_{n_0} and b_{n_0} with $1 \leq a_{n_0}/b_{n_0} \leq 10$. We choose a rather non-informative prior by setting $a_{n_0} = 3$ and $b_{n_0} = 0.3$ which leads to a distribution of n_0 centered around 3 with a large variance. Other values for a_{n_0} and b_{n_0} have been tested and give nearly the same results. As in Section 5, we choose a non-informative prior by setting $a_\rho = b_\rho = 0.01$.

MCMC convergence diagnostics was investigated with the clustering en-

tropy

$$-\sum_{i=1}^{q} \frac{|S_i|}{n} \log\left(\frac{|S_i|}{n}\right).$$

Traceplots for this quantity and for other parameters of the model suggest a good mixing and the convergence of our chain.

The majority clustering (mode of the posterior distribution of the clusterings) is the same as in Abraham et al. (2013) (two clusters: one containing data 1,2,6,9 and 12, the second containing data 3,4,5,7,8,10,11,13 and 14) with a posterior probability equal to 30.5%. This result was awaited and is coherent with the choice of 2 clusters in the previous method. But the real gain from our Bayesian approach is to look beyond this majority clustering. Here there are 3 more clusterings that are significant and that could give some information on this real dataset. The second majority clustering is nearly the same as the previous one : the clusters are the same but data 6 is alone in a third cluster. Indeed, this data is very atypical because it is the only one that contains an angle near 1.69π. The posterior probability for this clustering is 14.9%. The third majority clustering gives nearly the same information with a posterior probability of 13.5%. There are two clusters: one with data 6 and a second with all the others. Finally, another clustering with a posterior probability of 12.0% is made up of only one cluster. Even with other choices for the hyperparameters a_{n_0} and b_{n_0}, the posterior probability of this clustering remains high. It highlights the fact that all the data share some common traits and the main difference in the two clusters of the majority clustering only concerns one angle. All the clusterings are included in Figure 3 sorted by their posterior probabilities. It can be noted that a credible region with a posterior probability of 71% is composed of the 4 previous clusterings.

We give in Figure 4 the posterior distribution of the number of clusters. The posterior probabilities of 1, 2 or 3 clusters are respectively 12%, 65% and 21%. Consequently, the number of clusters is certainly (with probability 98%) less than or equal to 3.

As expected, these results are in line with the clusterings obtained in Abraham et al. (2013). The final choice of 2 clusters (that is not made here a priori) could provide, using the centers of the clusters, preset positions for praticians. As explained above, these two centers share only one main difference on one unique angle. This is highlighted by the important posterior

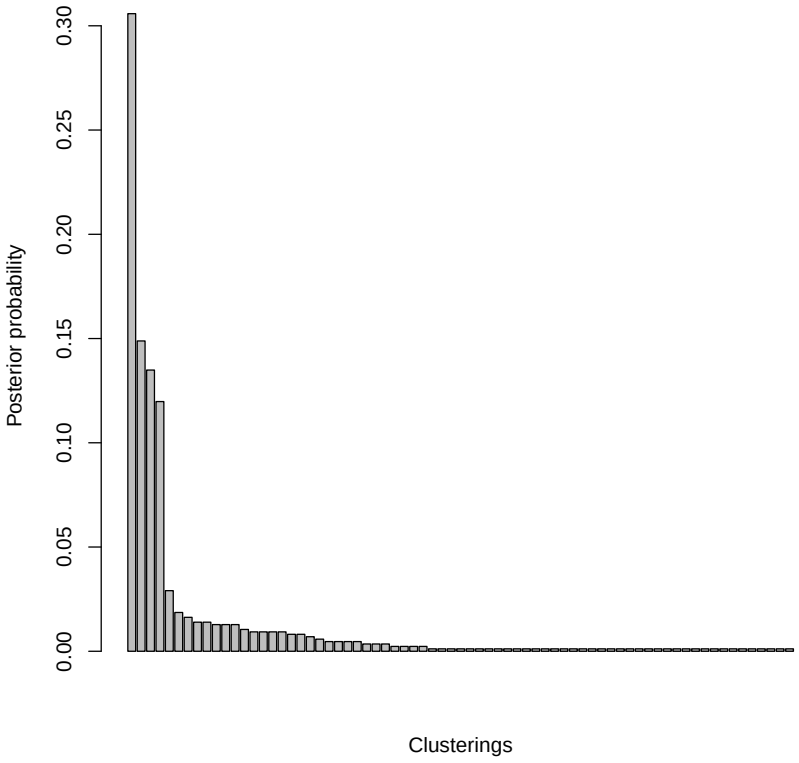

Figure 3: Barplot of the proportion of the different clusterings.

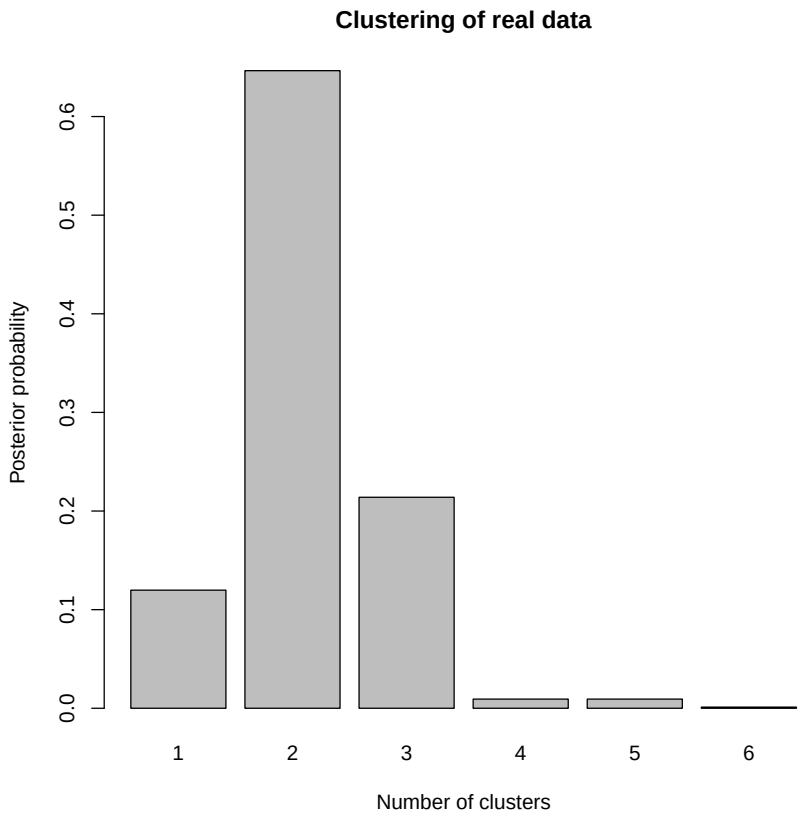

Figure 4: Posterior distribution of the number of clusters.

probability of the clustering with only one cluster. Thus, using these preset positions should be fairly easy for praticians, with four fixed values and only two choices for the last one. Furthermore, the results suggest another preset position that should be added and tested if the two previous one do not fit: the beam angles of data 6.

Table 2: Real data set (radians).

Patient	1^{st}angle	2^{nd}angle	3^{rd}angle	4^{th}angle	5^{th}angle
1	1.81π	0	$\pi/4$	$\pi/2$	π
2	1.78π	0	$\pi/4$	$\pi/2$	π
3	1.89π	$\pi/4$	$\pi/2$	$3/4\pi$	π
4	1.94π	0.28π	0.56π	$3/4\pi$	0.97π
5	-0.17π	$\pi/2$	$\pi/4$	$3/4\pi$	π
6	1.69π	-0.06π	$\pi/4$	$\pi/2$	π
7	$3\pi/4$	0.28π	0.53π	$3/4\pi$	π
8	1.86π	0.06π	$\pi/2$	$3/4\pi$	π
9	$\pi/2$	π	1.81π	0	$\pi/4$
10	0.31π	0.56π	$3/4\pi$	$1\pi/2$	-0.19π
11	1.81π	0.1π	$\pi/2$	$3/4\pi$	π
12	$\pi/4$	$\pi/2$	π	1.81π	0
13	0.72π	π	-0.08π	$\pi/4$	$\pi/2$
14	0.22π	0.56π	$3/4\pi$	π	1.89π

7 Conclusion

We present a full Bayesian framework for the clustering of multivariate circular and non-ordered data. It is based on a hierarchical model that combines Projected Normal distributions and the Dirichlet Process. Two original parameters are also introduced in this model: the parameter ρ to infer the variance of the angles and the symmetrization parameter τ to model the non-ordered feature of the data. The parameters of the model are then inferred using a Metropolis-Hastings within Gibbs algorithm and a theoretical study of the impact of the symmetrization parameter is provided. The simulation study and the real data example show the benefits of this approach.

Indeed, the number of clusters is chosen automatically by the method and the final result is much more complete than the majority clustering which is usually provided by classical clustering algorithms. However some improvements could be considered, such as, incorporating covariates (shape or size of the tumor, stage of the cancer, sex, age, ...) to preselect the beam positions.

8 Appendix

8.1 Specification of the prior of P_0

Let us recall the notations of Section 2. We denote by R the 2×2-matrix of the rotation in $\mathbb{R}^2$ with angle $2\pi/k$ and center 0 and set $\mu_{i1} \sim N_2(0, \rho I_2)$ and $\mu_{ij}|\mu_{i,j-1} \sim N_2(R\mu_{i,j-1}, I_2)$ for $j \in \{2, \ldots, k\}$. We denote by P_0 the distribution of $\mu_i = (\mu_{i1}', \ldots, \mu_{ik}')'$. Then, there exist independent random variables $\epsilon_j \sim N_2(0, I_2)$, independent of μ_{i1} such that $\mu_{ij} = R^{j-1}\mu_{i1} + \epsilon_j$ for $j \in \{2, \ldots, k\}$. It is then clear that P_0 is centered, Gaussian with covariance matrix

$$\Sigma_0(\rho) = \begin{pmatrix} \rho I_2 & \rho R' & \rho R^{2'} & \ldots & \rho R^{(k-1)'} \\ \rho R & (\rho+1) I_2 & \rho R' & \ldots & \rho R^{(k-2)'} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho R^{k-2} & \rho R^{k-3} & \ldots & (\rho+1) I_2 & \rho R' \\ \rho R^{k-1} & \rho R^{k-2} & \ldots & \rho R & (\rho+1) I_2 \end{pmatrix},$$

where R' is the transposed matrix of R and that

$$\Sigma_0^{-1}(\rho) = \begin{pmatrix} (\rho^{-1} + (k-1)) I_2 & -R' & -R^{2'} & \ldots & -R^{(k-2)'} & -R^{(k-1)'} \\ -R & I_2 & 0 & \ldots & \ldots & 0 \\ -R^2 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -R^{k-2} & \vdots & \ddots & \ddots & I_2 & 0 \\ -R^{k-1} & 0 & \ldots & \ldots & 0 & I_2 \end{pmatrix}.$$

Furthermore, Theorem 13.3.8 of Harville (1997) leads us to $|\Sigma_0^{-1}(\rho)| = \rho^{-2}$.

8.2 SAMS and Gibbs Samplers

8.2.1 SAMS Sampler

The SAMS sampler is given in detail in Dahl (2003). Formula (12) of Dahl (2003) reduces to

$$P(l \in S_i | S_i, S_j) = \frac{|S_i| N_{2k}\left(y_l; \Sigma_i \sum_{h\in S_i} y_h/|S_i|, I_{2k}+\Sigma_i\right)}{|S_i| N_{2k}\left(y_l; \Sigma_i \sum_{h\in S_i} y_h/|S_i|, I_{2k}+\Sigma_i\right) + |S_j| N_{2k}\left(y_l; \Sigma_j \sum_{h\in S_j} y_h/|S_j|, I_{2k}+\Sigma_j\right)}$$

and formula (14) for the Metropolis-Hastings ratio is obtained with

$$p(y_{S_j}) = \prod_{h=1}^{|S_j|} N_{2k}\left(y_{i_h}; \Sigma_{-j} \sum_{h\in S_{-j}} y_h/|S_{-j}|, I_{2k}+\Sigma_{-j}\right)$$

where $S_j = \{i_1, \ldots, i_{|S_j|}\}$, $S_{-j} = \{i_1, \ldots, i_{j-1}\}$ and $\Sigma_{-j} = \left(|S_{-j}|^{-1} I_{2k} + \Sigma_0^{-1}(\rho)\right)^{-1}$.

8.2.2 Gibbs Sampler

Let us denote by $\eta = \{S_1, \ldots, S_q\}$ the current partition of the algorithm. For $i = 1, \ldots, n$, the observation i is assigned to cluster S_j, $j \in \{1, \ldots, q\}$ with probability proportional to

$$|S_j^-| \times N_{2k}\left(y_i; \Sigma_j \sum_{i\in S_j} y_i/|S_j^-|, I_{2k}+\Sigma_j\right)$$

where $|S_j^-|$ is the cardinal of $S_j \setminus \{i\}$, or to (a new) cluster S_{q+1} with probability proportional to

$$n_0 \times N_{2k}\left(y_i; 0, I_{2k}+\Sigma_0(\rho)\right).$$

8.3 Full Conditional Distributions

Full Conditional of r Remember that $x_i = (x'_{i1}, \ldots, x'_{ik})' \in (\mathbb{R}^2)^k$, $i \in \{1, \ldots, n\}$, are independant with distribution $N_{2k}(\mu_i^{\tau_i}, I_{2k})$ with $\mu_i^{\tau_i} = (\mu'_{i\tau_i(1)}, \ldots, \mu'_{i\tau_i(k)})' \in$

$(\mathbb{R}^2)^k$ and that $x_{ij} = (x_{ij1}, x_{ij2})' = (r_{ij}\cos\theta_{ij}, r_{ij}\sin\theta_{ij})'$. Then, it is easy to see that (θ_{ij}, r_{ij}) are independant given τ, μ, ρ and n_0, with density:

$$p(\theta_{ij}, r_{ij}|\tau, \mu, \rho, n_0) = (2\pi)^{-1}\ e^{-\frac{1}{2}\mu'_{i\tau_i(j)}\mu_{i\tau_i(j)}}\ r_{ij}\ e^{-\frac{1}{2}\left(r_{ij}^2 - 2r_{ij}u'_{ij}\mu_{i\tau_i(j)}\right)},$$

with $u'_{ij} = (\cos\theta_{ij}, \sin\theta_{ij})$. Then,

$$\begin{aligned} p(r|\theta, \tau, \mu, \rho, n_0) &\propto p(\theta, r|\tau, \mu, \rho, n_0) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{k} p(\theta_{ij}, r_{ij}|\tau, \mu, \rho, n_0) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{k} r_{ij} e^{-\frac{1}{2}\left(r_{ij} - u'_{ij}\mu_{i\tau_i(j)}\right)^2}. \end{aligned}$$

8.4 Proof of Proposition 1

a) If we denote by ϕ_{2k} the density of the $N_{2k}(0, I_{2k})$ distribution, it can be shown after some calculations, that:

$$m(x_S) = \left(\prod_{i\in S}\phi_{2k}(x_i)\right)|\Sigma_0|^{-1/2}|\Sigma_S|^{-1/2}\exp\left(\frac{1}{2}\|\sum_{i\in S}x_i\|_S^2\right).$$

Then, we have:

$$\prod_{j=1}^{q} m(x_{S_j}) = \left(\prod_{i=1}^{n}\phi_{2k}(x_i)\right)|\Sigma_0|^{-q/2}|\Sigma_S|^{-q/2}\exp\left(\frac{1}{2}\sum_{j=1}^{q}\|\sum_{i\in S}x_i\|_S^2\right),$$

and

$$\prod_{j=1}^{q} m(x^\tau_{S_j}) = \left(\prod_{i=1}^{n}\phi_{2k}(x_i^{\tau_i})\right)|\Sigma_0|^{-q/2}|\Sigma_S|^{-q/2}\exp\left(\frac{1}{2}\sum_{j=1}^{q}\|\sum_{i\in S}x_i^{\tau_i}\|_S^2\right).$$

From (5), it can be seen that $\phi_{2k}(x_i) = \phi_{2k}(x_i^{\tau_i})$ and we conclude that:

$$\frac{\prod_{j=1}^{q} m(x^\tau_{S_j})}{\prod_{j=1}^{q} m(x_{S_j})} = \exp\left(\frac{1}{2}\sum_{j=1}^{q}\left(\|\sum_{i\in S_j}x_i^{\tau_i}\|_{S_j}^2 - \|\sum_{i\in S_j}x_i\|_{S_j}^2\right)\right),$$

hence the result.
b) If we denote by g the density of G with respect to the counting measure, we have:

$$p_I(\eta|x) \propto g(\eta)\ p_I(x|\eta),$$

and

$$p_{II}(\eta|x) \propto g(\eta)\ \frac{1}{(k!)^n} \sum_{\tau \in \mathcal{P}^n} p_I(x^\tau|\eta).$$

and we deduce that:

$$\frac{p_{II}(\eta|x)}{p_I(\eta|x)} = B_G f(x,\eta),$$

where

$$B_G = \frac{\sum_\eta g(\eta)\ p_I(x|\eta)}{\sum_\eta g(\eta) \frac{1}{(k!)^n} \sum_{\tau \in \mathcal{P}^n}\ p_I(x|\eta)}.$$

c) From Lemma 1 below, we have:

$$\begin{aligned} \min_G B_G &= \min_\eta \frac{p_I(x|\eta)}{\frac{1}{(k!)^n} \sum_{\tau \in \mathcal{P}^n}\ p_I(x|\eta)}, \\ &= \min_\eta \frac{1}{f(x,\eta)} \\ &= \frac{1}{\max_\eta f(x,\eta)}. \end{aligned}$$

□

Lemma 1. *Let $h \in \mathbb{R}^n$ and $f \in \mathbb{R}^n$ such that $f_i > 0$ and $h_i > 0$ for all $i \in \{1,\dots,n\}$. Write $\mathcal{D} = \{p \in \mathbb{R}^n, \sum_{i=1}^n p_i = 1, p_i \geq 0$ for all $i\}$. We have:*

$$\inf_{p \in \mathcal{D}} \frac{\sum_{i=1}^n p_i f_i}{\sum_{i=1}^n p_i h_i} = \min_{1 \geq i \geq n} \frac{f_i}{g_i}.$$

Proof of Lemma 1 Assume without loss of generality that $\min_i f_i/g_i = f_1/g_1$. Then we have:

$$\begin{aligned}
\frac{f_i}{g_i} \geq \frac{f_1}{g_1} \text{ for all } i \in \{1,\ldots,n\} \quad &\iff \quad f_i g_1 - f_1 g_i \geq 0 \text{ for all } i \in \{1,\ldots,n\} \\
&\iff \quad \sum_{i=1}^{n} p_i (f_i g_1 - f_1 g_i) \geq 0 \text{ for all } p \in \mathcal{D} \\
&\iff \quad g_1 \sum_{i=1}^{n} p_i f_i \geq f_1 \sum_{i=1}^{n} p_i g_i \text{ for all } p \in \mathcal{D} \\
&\iff \quad \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i g_i} \geq \frac{f_1}{g_1} \text{ for all } p \in \mathcal{D}.
\end{aligned}$$

We conclude by noting that

$$\frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i g_i} = \frac{f_1}{g_1}$$

for $p = (1, 0, \ldots, 0)$. □

References

1. Abraham, C., Molinari, N., and Servien, R. (2013). Unsupervised clustering of multivariate circular data. Statistics in Medicine **32,** 1376–1382.

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

3. Dahl, D. B. (2003). An improved merge-split sampler for conjugate dirichlet process mixture models. Technical Report, Univ. of Wisconsin - Madison **1086,** 1–32.

4. Damien, P. and Walker, S. (1999). A full bayesian analysis of circular data using the von mises distribution. The Canadian Journal of Statistics **27,** 291–298.

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

6. Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. The Annals of Statistics **1,** 209–230.

7. Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. Bayesian Analysis **4,** 367–392.

8. Griffin, J. and Holmes, C. (2010). Computational issues arising in bayesian nonparametric hierarchical models. In Hjort, N., Holmes, C., Mller, P., and Walker, S. G., editors, Bayesian Nonparametrics, pages 208–222, Cambridge University Press.

9. Harville, D. A. (1997). Matrix algebra from a statistician's perspective. Springer, New York.

10. Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification **2,** 193–218.

11. Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Journal of Computational and Graphical Statistics **13,** 158–182.

12. Jona-Lasinio, G., Gelfand, A., and Jona-Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped gaussian processes. Ann. Appl. Stat. **6,** 1478–1498.

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

14. MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. Communications in Statistics: Simulation and Computation **23,** 727–741.

15. MacEachern, S. N. (1998). Computational methods for mixture of dirichlet process models. In Dey, D., Mller, P., and Sinha, D., editors, Practical Nonparametric and Semiparametric Bayesian Statistics, pages 23–44, New-York: London. Lecture Notes in Statistics 133.

16. Mardia, K. and Jupp, P. (2009). Directional Statistics. John Wiley & Sons, New-York.

17. Mardia, K. V., Hugues, G., Taylor, C. C., and Singh, H. (2008). A multivariate von mises distribution with applications to bioinformatics. The Canadian Journal of Statistics **36,** 99–109.

18. Neal, R. M. (2000). Markov chain sampling method for dirichlet process mixture models. Journal of Computational and Graphical Statistics **9,** 249–265.

19. Nguyen, X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clustering comparison: Is a correction for chance necessary ? ICML'09: Proceedings of the 26th Annual International Conference on Machine Learning pages 1073–1080.

20. Nuñez-Antonio, G. and Gutiérrez-Peña, E. (2005). A bayesian analysis of directional data using the projected normal distribution. Journal of Applied Statistics **32,** 995–1001.

21. Nuñez-Antonio, G., Gutiérrez-Peña, E., and Escarela, G. (2011). A bayeisan regression model for circular data based on the projected normal distribution. Statistical Modeling **11,** 185–201.

22. Presnell, B., Morrison, S. P., and Littell, R. C. (1998). Projected multivariate linear models for directional data. Journal of the American Statistical Association **93,** 1068–1077.

23. Quintana, F. A. (2006). A predictive view of bayesian clustering. Journal of Statistical Planning and Inference **136,** 2407–2429.

24. Rand, W. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66,** 846–850.

25. Ravidran, P. and Ghosh, S. K. (2011). Bayesian analysis of circular data using wrapped distributions. Journal of Statistical Theory and Practice **5,** 547–560.

26. SenGupta, A. and Laha, A. K. (2008). A bayesian analysis of the change-point problem for directional data. Journal of Applied Statistics **35,** 693–700.

27. Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic model for two dependant circular variables. Biometrika **89,** 719–723.

28. Teh, Y. W. (2010). Dirichlet processes. In Encyclopedia of Machine Learning. Springer.

29. Von Mises, R. (1918). Über die ganzzahligkeit der atomgewicht und verwandte fragen. Physikalische Zeitschrift **19,** 490–500.

30. Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. Statistical Methodology **10,** 113–127.

31. Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected gaussian processes. Journal of the American Statistical Association **109,** 1565–1580.

32. Wang, F., Gelfand, A. E., and Jona-Lasinio, G. (2015). Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the adriatic sea. Statistica Sinica **25,** 25–29.

33. Yuan, L., Wu, Q. J., Yin, F., Li, Y., Sheng, Y., Kelsey, C. R., and Ge, Y. (2015). Standardized beam bouquets for lung IMRT planning. Physics in Medicine & Biology **60,** 1821–1843.