A clustering Bayesian approach for radiotherapy x-ray beam bouquets

Christophe Abraham, Nicolas Molinari, Rémi Servien

A clustering Bayesian approach for radiotherapy x-ray beam bouquets

Christophe Abraham * Nicolas Molinari † Rémi Servien ‡

Abstract

This paper presents a Bayesian framework for the clustering of multivariate directional or circular data. We introduce a hierarchical model that combines Projected Normal distributions and a Dirichlet Process. The parameters of the model are then inferred using a Metropolis-Hastings within Gibbs algorithm. Simulated datasets are analyzed to study the influence of the parameters of the model. Then, the benefits of our approach are illustrated by clustering real data from 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.

†Université de Montpellier, IMAG, place Eugène Bataillon, 34095 Montpellier cedex 5, France.

‡Toxalim, Université de Toulouse, INRA, Toulouse, France; remi.servien@toulouse.inra.fr

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, chemotherapy and intensity-modulated radiation therapy (IMRT) have demonstrated their effectiveness for cancer treatment. New molecules and new generation of radiotherapy machines are developed by pharmaceutical firms. Latest generation of radiotherapy machines projects multiple rays. Multiplying beams allows concentrating 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 machines takes long time 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. By consequence, an observation is composed of the k beams of a patient, that is k circular measurements. The multivariate trait is due to the number of points k on the unit circle of $\mathbb{R}^2$. One actual observation consists of a (non-ordered) set of k angles rather than of a (ordered) vector of length k.

Abraham et al. (2013) proposed a first tool to assist the selection of beam orientations in addition to 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. A procedure using a criterion of model choice (AIC, BIC, RIC, silhouette index, ...) can be used to choose between two models but an appropriate methodology that automatically finds the optimal number of clusters would be very useful. Second, the final result is only a 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. To our knowledge, such a clustering Bayesian model has never been applied for multivariate circular data in the literature.

The 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 on the sphere of the normal distribution. 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) and 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).

These models are not adapted to treat our general case : multiple points on a sphere that give a multivariate set of non-ordered angles. We propose an extension to multivariate data of the projected normal distribution. We also study an inference method on the variance-covariance matrix of the projected normal distribution. Then, this modelization is associated with a Dirichlet process to perform clustering on this data. Our 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 empirical results first on simulated data and then on the real data set that motivated the present work.

2 Model

A simple way of generating distributions on the p-dimensional unit sphere $\mathcal{S}^p$ is to radially projecting 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 the present paper, the ith of the n observations is given by k angles $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})' \in [0, 2\pi[^k$. 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 $N_{2k}(\mu_i^{\tau_i}, I_{2k})$ where τ_i will be defined later and θ_{ij} is defined as the 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. By abuse of notation, we will denote by $PN_{2k}(\mu_i^{\tau_i}, I_{2k})$ the joint distribution of (θ_i, r_i). Clustering analysis will be based on a Dirichlet process mixture (DPM) model described as follows :

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

where $P_0 = N_{2k}(0, \Sigma_0)$ and n_0 are, respectively, the center and the precision

parameter of the Dirichlet process (DP) introduced by Ferguson (1973). 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 :

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

In the sequel, we will denote by Pólya($n_0 P_0$) the distribution of μ given by (2). It is clear from (2) that $\mu_1, \ldots, \mu_n$ may share values in common; hence the clusters. 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 to choose 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).

At this point, it is important to recall that the actual ith observation consists of a (not ordered) set of the form $\{\theta_{i1}, \ldots, \theta_{ik}\}$ rather than of a (ordered) vector $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})'$. We treat the observations as vectors for convenience but this can have important consequences on the clustering due to the possible permutation between the vector components. To be concrete, consider two observations $\theta_1 = (0, \pi/5, 2\pi/5, 3\pi/5, 4\pi/5)' + \varepsilon_1$ and $\theta_2 = (4\pi/5, 0, 3\pi/5, \pi/5, 2\pi/5)' + \varepsilon_2$ where ε_1 and ε_2 are composed with small values of $[0, 2\pi[$. From a practical point of view, θ_1 and θ_2 should be put together in a same cluster since $\{\theta_{11}, \ldots, \theta_{15}\}$ and $\{\theta_{21}, \ldots, \theta_{25}\}$ are very similar. In this case, the parameters μ_1 and μ_2 are likely to have the same coordinates but in a different order. In

other words, if we denote by $\mu_{i1}, \ldots, \mu_{ik} \in \mathbb{R}^2$ the components of μ_i, it is likely that $\{\mu_{11}, \ldots, \mu_{1k}\} = \{\mu_{21}, \ldots, \mu_{2k}\}$ and $(\mu_{11}, \ldots, \mu_{1k}) \neq (\mu_{21}, \ldots, \mu_{2k})$. If it was simply assumed that θ_i had a projected normal distribution $PN_{2k}(\mu_i, I_{2k})$ then, θ_1 and θ_2 would belong to the same cluster only when $\mu_1 = \mu_2$; and this is a posteriori unlikely. To cope with this problem, we introduce the permutation parameter τ_i. Denote by $\mathcal{U}_{\mathcal{P}}$ the uniform distribution on the set of permutations of $\{1, \ldots, k\}$. Assume that the parameters τ_i are a priori independent with distribution $\mathcal{U}_{\mathcal{P}}$ and, for all $\mu_i = (\mu'_{i1}, \ldots, \mu'_{ik})' \in (\mathbb{R}^2)^k$, 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. Clearly, the introduction of τ_i in (1) enables our model to put θ_1 and θ_2 in a same cluster although $\mu_1 \neq \mu_2$ since there exist some values of τ_1 and τ_2 for which $\mu_1^{\tau_1} = \mu_2^{\tau_2}$.

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 (1), 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 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 4.1 and Figure 1). 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 conditionnal distribution will be calculated in the following section.

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

$$\begin{aligned} \theta_i, r_i|\mu, \tau &\sim PN_{2k}(\mu_i^{\tau_i}, I_{2k}), \\ \mu|n_0, \rho &\sim \text{Pólya}(n_0 P_0(\rho)), \\ \tau_i &\sim \mathcal{U}_{\mathcal{P}}, \\ \rho &\sim IG(a_\rho, b_\rho), \\ n_0 &\sim G(a_{n_0}, b_{n_0}). \end{aligned} \tag{3}$$

where $P_0(\rho) = N_{2k}(0, \Sigma_0(\rho))$ and $\mu = (\mu_1', \ldots, \mu_n')'$. By a usual convention, it is assumed that the random variables at a stage of the hierarchy are independent.

3 Inference

We set $\theta = (\theta_1', \ldots, \theta_n')'$, $r = (r_1', \ldots, r_n')'$, $\tau = (\tau_1, \ldots, \tau_n)'$ 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.

Simulations of μ We can restrict our attention to model (1) instead of the full model (3) for the simulations of μ as every component of ξ except μ remains fixed. An alternative reparametrization of μ, θ and ρ will prove useful. Denote $x = (x_1', \ldots, x_n')'$ where $x_i = (x_{i1}, \ldots, x_{ik})'$. First, note that the full conditional distribution of μ reduces to the conditional distribution of μ given (x, n_0, ρ, τ) as there exists a natural bijection between x_{ij} and (θ_{ij}, r_{ij}). Second, 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

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

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

$$\begin{aligned} y_i|\mu_i &\sim N_{2k}(\mu_i, I_{2k}), \\ \mu_i|P &\sim P, \\ P &\sim DP(n_0 P_0). \end{aligned} \tag{4}$$

There exist 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 parametrization 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 (4) may be expressed as:

$$\begin{aligned} y_i|\eta, \phi &\sim N_{2k}(\textstyle\sum_{j=1}^{q} \phi_j \mathbf{1}_{\{i \in S_j\}}, I_{2k}), \\ \phi_j|\eta &\sim P_0, \\ \eta &\sim p(\eta) \propto \textstyle\prod_{i=1}^{q} n_0 \Gamma(|S_j|), \end{aligned} \tag{5}$$

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 den-

sity. We can integrate over the clusters locations parameter ϕ analytically in (5) 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 clusters locations 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 :

$$p(r_{ij}|\theta, \tau, \mu, \rho, n_0) \propto r_{ij} e^{-\frac{1}{2}\left(r_{ij} - u'_{ij}\mu_{i\tau_i(j)}\right)^2}, \tag{6}$$

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 (6) 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 (6) 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 that:

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

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 $\{1, \ldots, k\}$) :

$$p(\tau_i|x, \mu) = \frac{N_{2k}(x_i; \mu_i^{\tau_i}, I_{2k})}{\sum_{t \in T} N_{2k}(x_i; \mu_i^{t}, I_{2k})}. \tag{7}$$

Simulations of ρ From (3), 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 (5), we note that η and ρ are independent and that:

$$\begin{aligned} p(\rho|\theta, r, \mu, \tau, n_0) &= p(\rho|\eta, \phi) \\ &\propto p(\phi|\eta, \rho)p(\rho|\eta) \\ &\propto \left(\prod_{j=1}^{q} p(\phi_j|\rho) \right) p(\rho). \end{aligned} \tag{8}$$

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 that:

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

and it is easy to conclude from (8) that the full conditional of ρ is

$$IG\left(a_\rho + q, b_\rho + \frac{1}{2}\sum_{i=1}^{q}\phi'_{i1}\phi_{i1}\right). \tag{9}$$

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:

$$\begin{aligned} \zeta|n_0 &\sim B\left(n_0 + 1, n\right) \\ n_0|\zeta, q &\sim \pi_n G(a_{n_0} + q, b_{n_0} - \log\zeta) + (1 - \pi_n)G(a_{n_0} + q - 1, b_{n_0} - \log\zeta), \end{aligned} \tag{10}$$

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 the Algorithm 1.

4 Simulations

Before using our algorithm on real data, we test it on two simulation studies. 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} and N_{11} the numbers of pairs that are in the same cluster in both partitions and the number of pairs that are in different clusters in both partitions respectively,

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 three times the SAMS sampler.
   (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 7.
5. Simulate new ρ from 9.
6. Simulate n_0 from 10.

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 clusterings comparison and for a study on the usefulness of the adjusted measures, we refer the reader to Fritsch and Ickstadt (2009) and Nguyen et al. (2009).

4.1 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}, $j =$

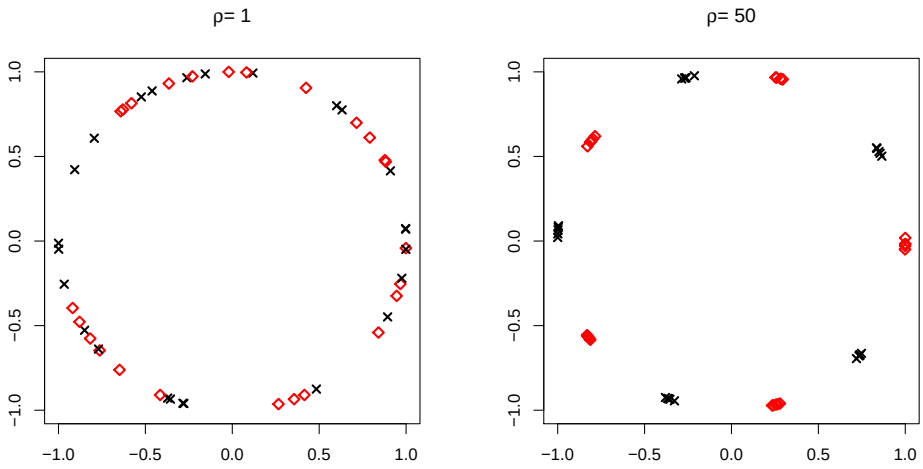

Figure 1: Two data sets are generated with two different values for the parameter ρ to highlight the influence of this parameter. Two clusters of five data are represented on each plot. Each data is composed of $k = 5$ angles on the circle. One cluster is represented by the black cross, the other by the red square.

$2, \ldots, 5$ (so $k = 5$) 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 1 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 1 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 $a_\rho = b_\rho = 0.01$.

4.2 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 $\rho = 20$. The mean values for the ARI over 100 simulated data sets are given in Table 1.

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)

As expected, the choice of these hyperparameters has an impact on the results. Let us remind that the a priori distribution of n_0 is a gamma distribution with parameter a_{n_0} and b_{n_0} with expected value equal to a_{n_0}/b_{n_0} and variance equal to $a_{n_0}/b_{n_0}^2$. The results of Table 1 suggests that a choice of $a_{n_0}/b_{n_0} = 10$ provides good results. According to the results of this section, we fix $a_{n_0} = 10$ and $b_{n_0} = 1$ as the real data set is similar to these simulated data sets.

5 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 (Table 2). This data set was previously analyzed in Abraham et al. (2013) in which the number of clusters was preselected to $q = 2$. As already explained, we do not preselect the number of clusters but we estimate it from the data. Recall that we fix the hyperparameters $a_{n_0} = 10$, $b_{n_0} = 1$ and $a_\rho = b_\rho = 0.01$ in Section 4.

The majority clustering is the same as in Abraham et al. (2013) 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 enlights the fact that all the data share some common traits and the main difference in the two clusters of the majority clustering only stands for one angle. All the clusterings are included in Figure 2 sorted by their posterior probabilities. Remark that a credible region with a posterior probability of 71% is composed of the 4 previous clusterings.

We give in Figure 3 the posterior distribution of the number of clusters. The posterior probabilities of 1, 2 or 3 clusters are respectively 65%, 21% and 12%. 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) but also add some information. Indeed, the choice of 2 clusters is not made a priori. We have different clusterings associated with different prob-

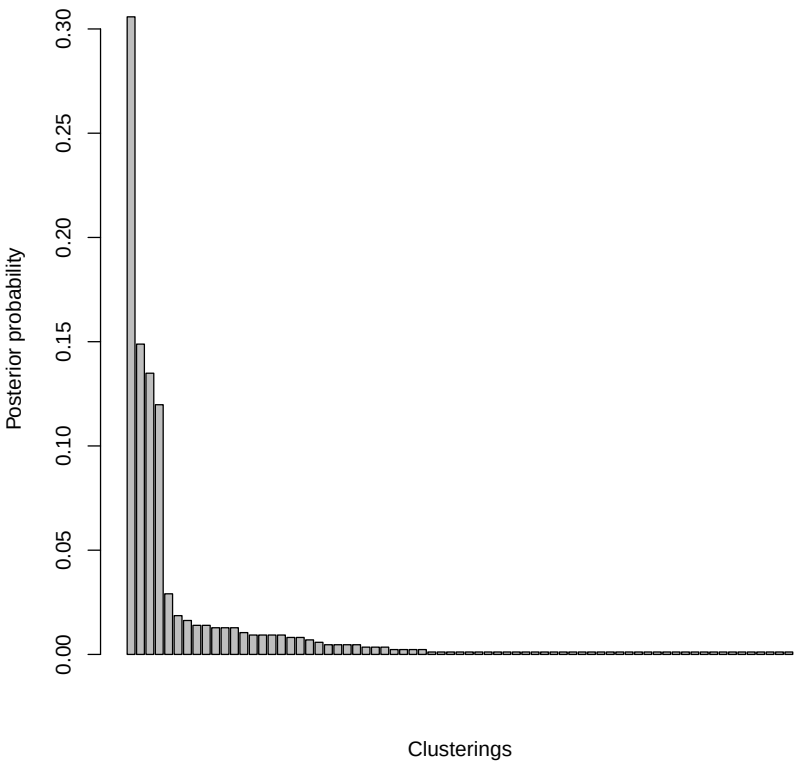

Figure 2: Barplot of the proportion of the different clusterings.

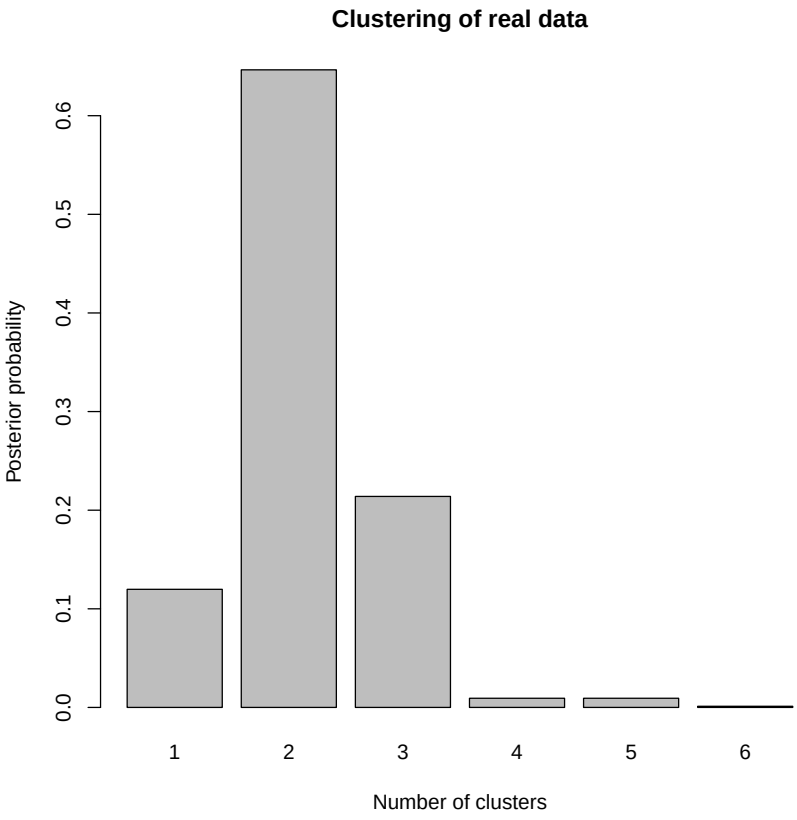

Figure 3: Posterior distribution of the number of clusters.

abilities that help us understanding this data set : all data have similarities and data 6 is the most dissimilar.

MCMC convergence diagnostics was investigated with the clustering entropy

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

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

Table 2: Real data set (radians).

Data	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

6 Conclusion

We present a full Bayesian framework for the clustering of multivariate directional or circular data. It is based on a hierarchical model that combines Projected Normal distributions and a Dirichlet Process. The parameters of the model are then inferred using a Metropolis-Hastings within Gibbs algorithm.

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 contemplated as, for example, incorporating covariates (shape or size of the tumor, stage of the cancer, sex, age, ...) to preselect the beam positions.

7 Appendix

7.1 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

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

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

7.2 SAMS and Gibbs sampler

7.2.1 SAMS Sampler

The SAMS sampler is given in details 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 $S_j = \{i_1, \dots, i_{|S_j|}\}$, $S_{-j} = \{i_1, \dots, i_{j-1}\}$ and $\Sigma_{-j} = \left(|S_{-j}|^{-1} I_{2k} + \Sigma_0^{-1}(\rho)\right)^{-1}$.

7.2.2 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

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

7.3 Full conditional distributions

Full conditional of r Recall 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,

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

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 *Bayesian Nonparametrics* (eds. N. Hjort, C. Holmes, P. Mller and S. G. Walker), 208–222. Cambridge University Press.

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

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

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

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: Pro-*

ceedings of the 26th Annual International Conference on Machine Learning, 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. Von Mises, R. (1918) Über die ganzzahligkeit der atomgewicht und verwandte fragen. *Physikalische Zeitschrift*, **19**, 490–500.

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

30. — (2014) Modeling space and space-time directional data using projected gaussian processes. *Journal of the American Statistical Association*, **109**, 1565–1580.

31. 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.

32. 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.