Unsupervised clustering of multivariate circular data

Christophe Abraham[a], Nicolas Molinari[a,b], Rémi Servien[a,*]

[a] *Montpellier SupAgro-INRA, UMR MISTEA 729, Bâtiment 29, 2 place Pierre Viala, 34060 Montpellier Cedex 2, France.*
[b] *CHU de Montpellier, Service DIM, Hôpital Lapeyronie, 371 avenue du Doyen Gaston Giraud, 34295 Montpellier Cedex 5, France.*

Abstract

An unsupervised clustering problem is studied in this paper. The originality of this problem lies in the data, which consist of the positions of five separate x-ray beams on a circle. The five x-ray beam "projectors" are positioned around each patient on a predefined circle. However, similarities exist in positioning for certain groups of patients, and we aim to describe these similarities with the goal of creating pre-adjustment settings that could help save time during x-ray positioning. We therefore performed unsupervised clustering of observed x-ray positions. Because the data for each patient consists of five angle measurements, Euclidean distances are not appropriated. Furthermore, k-means algorithm, usually used for minimising corresponding distortion can not be computed because centers of clusters are not calculables. We present here solutions to these problems. First, we define a suitable distance on the circle. Then, we adapt an algorithm based on simulated annealing to minimize distortion. This algorithm is shown to be theoretically convergent. Finally, simulations on simulated and real data are presented.

Keywords: Unsupervised clustering, Circular data, Radiotherapy machine data.

1. Introduction

Over the past few years, cancer treatment via intensity-modulated radiation therapy (IMRT) has improved. The patient's radiation oncologist prescribes the appropriate treatment volume and dosage. The medical radiation physicist and the dosimetrist determine how to deliver the prescribed dose and calculate the amount of time it will take the accelerator to deliver that dose. Radiation therapists operate the linear accelerator and give patients their daily radiation treatments. When a ray is projected onto a target area using the linear accelerator, such as a tumor, the healthy tissue it crosses is also irradiated. The

*Corresponding author.
Email address: remi.servien@supagro.inra.fr (Rémi Servien)

previous generation of radiotherapy machines helped solve this problem by using two rays that intersected on the target area; thus, the target area received the appropriate therapeutic amount of radiation, while non-target tissues received only half that dose. The latest generation of machines projects five rays, and following the same principle, non-target tissues receive only weak doses of radiation. Multiplying beams concentrates radiation on the tumor while avoiding the massive irradiation of healthy areas, and thus reducing side effects. Recent clinical studies have demonstrated the superiority of these techniques particularly as concerns cardiac side effects associated with lung tumors. However, the five beams are fixed on a circle in the transverse plane around the patient, and their re-positioning on the circle, i.e. their angles in degrees, is required for each patient. Repositioning for the five beam machine takes sufficiently longer as compared to the two beam machine that fewer patients are treated with the new machines. Several algorithms have been developed to make an exhaustive search and determine the best beams compositions (see for instance [1-4]) which are different for each patients. But the pratical implementation of these methods is hindered by the excessive computing time associated with the calculation. There is no others tools to assist the selection of beam orientations other than the therapist's experience and intuition whereas it could be very helpful [5] and accelerate previous algorithms. These algorithms could be sped up by using appropriate initial presets. The purpose of this work is to provide such effective presets. For this, we performed unsupervised clustering of observed x-rays positions. Then, each center of a cluster could be an efficient preset and each new patient should be affected to a cluster using a prior probability which is proportionnal to the number of patients in each cluster.

So, our data are a circular data set. Directional or circular data arise quite frequently in many natural and physical sciences. For instance, biologists studying bird-migrations record the flight directions of just-released birds as they disappear over the horizon. The first experiment was made by Schmidt-Koenig [6] over homing pigeons. He collects the data on the vanishing angles of birds released singly. Batschelet [7] gives an account of some applications on Biology. From another part, Jammalamadaka and al. [8] discuss a medical application where the angle of knee flexion was measured to assess the recovery of orthopaedic patients. Geologists also analyze paleomagnetic directions of the earth's magnetic pole to investigate the phenomenon of pole-reversal [9]. This brief review is not exhaustive and there is a lot of others applications of directionnal, circular or spherical data. It has also been studied from a theoretical point of view in Mardia and Jupp [10]. Even if our problem has similarities with some previously treated, the specificity of our data requires a specific method. Data are defined by the ballistic of the five angles $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}\}$. To define sets of recurrent angles used by radiotherapy technicians, and so predefine settings, we used an unsupervised clustering method to obtain patient groups with homogeneous ballistics.

The k-means algorithm [11, 12] is habitually used for this kind of problem in-

volving Euclidean distances. We can see in Section 2 that Euclidean distances are not appropriate for this case. So, we defined a suitable distance d for our problem, and a criterion was derived from this distance. However, there is no explicit solution for optimizing this criterion, again because the distances involved are non-Euclidean. The k-medoids clustering methods, like PAM [13] or CLARANS [14], can solve this problem using the most central data of the cluster as centroids (for a large review of clustering methods, we refer the reader to Xu and Wunsch [15]). But, because our real data set is small, we fear that few of the data will be next to their centroids. This can produce bad clustering. For these reasons, and also because these methods only identify local optima, we chose not to use k-medoids clustering methods. Instead we use a simulated annealing type algorithm described below, which can find a better approximation of the cluster centers.

In the present paper, we present a general method for solving this problem. Distance and distortion choices are described in the next section. Section 3 gives the clustering algorithm whose proof of convergence is joined in the Appendix. Section 4 includes empirical results first on simulated data and then on the real data set which motivates the present work.

2. Distance choice

In this section, we must consider two problems: the importance of the modulo 2π in the distance between two points on the circle and the permutations between two subsets, which is a novel feature, and is detailed below.

Data can be viewed as subset of five points on the circle. First, we define a distance δ between two points on the circle as follows :

$$\delta(a,b) = \min_{k\in\mathbb{Z}} |a - b + k2\pi| \text{ for all } a, b \in \mathbb{R}$$

where a and b denotes the angle in radians with respect to an arbitrary origin. Note that δ can be viewed as a L^1-distance on the circle. Also note that the fact that points are angles is immaterial in the rest of this paper and only affects metric δ. So, the following method could be used for any configurations of points lying in any space that has a distance defined between points.
Then, we define a distance between two subsets of five points on the circle. The chosen distance has to test all the permutations between the two subsets. For example, the distance between $x_1 = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$ and $x_2 = \{x_{12}, x_{13}, x_{14}, x_{15}, x_{11}\}$ must be zero. Taking into account these specificities, we propose the following function between two items x_1 and x_2 :

$$d(x_1, x_2) = \inf_{\sigma\in\mathcal{F}} \sum_{l=1}^{5} \delta\left(x_{1\sigma(l)}, x_{2l}\right),$$

where $\mathcal{F}$ is the set of permutations.
The function d is shown to be a distance in Lemma Appendix A.1 joined in the

Appendix. This definition allows us to test all permutations between two angle sets and retain that which corresponds to the smallest distance.

If $x_1, x_2, \ldots, x_n$ denote the n observations to be classified in J clusters, the problem consists in determining the set of cluster centers $\Omega = \{c_1, c_2, \ldots, c_J\}$ which minimizes the distortion D defined by :

$$D(\Omega) = \sum_{i=1}^{n} \min_{c \in \Omega} d(x_i, c).$$

If we set

$$C_j = \{x_i : d(x_i, c_j) = \min_{c \in \Omega} d(x_i, c)\},$$

note that

$$D(\Omega) = \sum_{j=1}^{J} \sum_{x_i \in C_j} d(x_i, c_j)$$

and that $C_1, \ldots, C_J$ defines a partitition of $\{x_1, x_2, \ldots, x_n\}$.
As discussed in the introduction, using a classical algorithm like k-means or k-medoids is not appropriate. So, the clustering algorithm has to approach optimal centers of clusters without the possibility of computing effective centers at each iteration.

3. Clustering algorithm and convergence result

Given the chosen distance and its characteristics mentioned above, we use, with a fixed number of clusters J, a clustering algorithm based on simulated annealing [16]. The $\nu - 1^{th}$ iteration of the algorithm ends giving us a set of J centers Ω^a. We describe below the ν^{th} iteration :

1. Each data is assigned to its nearest center according to distance d. This provides us with a distortion D_ν^a defined by

$$D_\nu^a(\Omega^a) = \sum_{i=1}^{n} \min_{c \in \Omega^a} d(x_i, c).$$

2. A cluster j with center $c_j = \{c_{j1}, c_{j2}, c_{j3}, c_{j4}, c_{j5}\}$ is randomly chosen according to a discrete Uniform distribution. Then, a new center c'_j is proposed for this cluster, with coordinates $c'_{js} \sim N^w(c_{js}, \sigma_a^2)$ for $1 \leq s \leq 5$.
3. The new distortion

$$D_\nu^b(\Omega^b) = \sum_{i=1}^{n} \min_{c \in \Omega^b} d(x_i, c)$$

is computed with $\Omega^b = \{c_1, \ldots, c_{j-1}, c'_j, c_{j+1}, \ldots, c_J\}$.

(a) The new center is accepted with probability $1 \wedge \exp\left(-(D_\nu^b - D_\nu^a)/(t_\nu)\right)$, where t_ν is the so-called temperature, and we return to step 1.
(b) If rejected, we return to step 2 and another center is taken.

The distribution $N^w(c_{js}, \sigma_a^2)$ is the wrapped normal distribution on the circle [10]. It is obtained by wrapping a common normal distribution $N(c_{js}, \sigma_a^2)$ onto the circle. Its probability density function is

$$f(x; c_{js}, \sigma_a^2) = \frac{1}{\sqrt{2\pi}\sigma_a} \sum_{l=-\infty}^{\infty} \exp\left\{-\frac{(x - c_{js} + 2l\pi)^2}{2\sigma_a^2}\right\}.$$

This distibution is unimodal and symmetric about its mode c_{js}.
The set of centers $\{c_1, \ldots, c_J\}$ which provides the lowest distortion D over all the chain is retained. This algorithm requires that the user sets in advance the number of clusters J, the shape of the temperature t_ν and the variance of normal distributions σ_a^2.
We are now in position to study the convergence of the algorithm from a theoretical point of view. Let K be the transition kernel associated with the described algorithm. We provide this transition kernel in the Appendix. And let us define $\mathrm{osc}_K(D)$ as follows

$$\underset{K}{\mathrm{osc}}(D) = \sup\{|D(x) - D(y)|, x \in E, y \in \mathrm{supp}\, K(x, .)\}$$

where $\mathrm{supp}\, K(x, .)$ denotes the support of $K(x, .)$.
We state the following Proposition 3.1 whose proof is joined in the Appendix.

Proposition 3.1. *Taking* $t_\nu = \frac{C_0}{\log(\nu+e)}$ *with* $C_0 > J\,\mathrm{osc}_K(D)$*, then, for all* $\varepsilon > 0$*,* $Pr(x_\nu \in D^\varepsilon) \to 1$ *as* $\nu \to \infty$ *where*

$$D^\varepsilon = \{x \in E, D(x) \leq \underset{\lambda}{\mathrm{essinf}}(D) + \varepsilon\} \text{ and } \underset{\lambda}{\mathrm{essinf}}(\mathrm{D}) = \sup\{\mathrm{a} \geq 0, \lambda(\mathrm{a} \leq \mathrm{D}) = 1\}.$$

The choice of C_0 is a known problem for the convergence of the algorithm. If C_0 is chosen too large, the algorithm will take a long time to converge because the denominator is $\log(\nu + e)$. On the other hand, if C_0 is chosen too small, the algorithm converges too quickly and does not sufficiently explore the space of possible values to find the optimal clustering. In our problem, it is clear that we have $\mathrm{osc}_K(D) \leq 5n\pi$, which leads us to the sufficient condition $C_0 > 25n\pi$. This is a rather crude bound, but we cannot obtain a better one without making strong assumptions about the data distribution. In order to reasonably estimate C_0, we run a chain of ν_0 sets of centers Ω and we calculate the variation of the distortion D at each iteration which leads to the following estimate of $\mathrm{osc}_K(D)$:

$$\hat{\mathrm{osc}}_K(D) = \sup_{1 \leq h \leq \nu_0} |D(\Omega^h) - D(\Omega^{h+1})|$$

where $\Omega^{h+1} \sim K(\Omega^h, .)$. This enables us to estimate C_0.
Note that in our algorithm only one randomly chosen center is updated. This

provides us with an acceptable trade-off between exploration and convergence. Other strategies could be considered like updating all the centers at each iteration. In any case, the variance σ_a^2 of the proposal distribution must be carefully chosen in order to balance between exploration and convergence.

4. Results

4.1. Simulated data

Before using our algorithm on real data, we test it on simulated data. We randomly generate a number of clusters J between 2 and 20, the five coordinates of the J centers c_j on $[0, 2\pi]$ and the number of data of each cluster between 2 and 30. Each coordinate s of the elements x_i of the cluster j are independently generated according to a wrapped normal distribution $N^w(c_{js}, \sigma_g^2)$, with fixed σ_g^2. Note the difference between σ_g^2 and σ_a^2, the variance of the wrapped normal distribution from which new centers in the algorithm are generated. Finally, we run the algorithm on a thousand of these data sets, with beginning centers taken randomly on $[0, 2\pi]$ and we obtain results presented in Table 1.

Table 1: Percentage of correct classification of our algorithm on simulated data according to σ_g and σ_a.

	$\sigma_g = 0.10$	$\sigma_g = 0.17$	$\sigma_g = 0.24$	$\sigma_g = 0.35$	$\sigma_g = 0.52$
$\sigma_a = 0.10$	99	94	88	78	54
$\sigma_a = 0.17$	99	95	90	76	55
$\sigma_a = 0.24$	99	95	90	75	54

Our algorithm performs well. Note that, according to Table 1, the choice of σ_a according to σ_g has little effect on the final results.

4.2. Real data set

We then apply the algorithm to the real data set, which come from post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France (Table A.2).
The clinicians informed us they believe that there are two different groups of patients. Running our algorithm with $J = 2$ we find the following two groups : one containing data 1,2,6,9 and 12, the second containing data 3,4,5,7,8,10,11,13 and 14. These results are relatively independent of the input parameters, such as initial centers or variance of wrapped normal distributions σ_a. Indeed, they have little effect on the clustering. We obtained two presets corresponding to the centers of our two groups:

$$c_1 = \{\pi/4, \pi/2, \pi, 1.81\pi, 1.99\pi\} \text{ and } c_2 = \{\pi/4, 0.51\pi, 3/4\pi, \pi, 1.88\pi\}.$$

We remark that the two centers have three common angles : $\pi/4, \pi/2$ and π and one slightly different from 1.85π. The principal difference resides in only one angle whose presets are $\pi/4$ or 0. Thus, using these preset positions should be fairly easy for praticians, with four fixed values and two choices for the last one. They should only have to make a few minor adjustments around these presets to correctly position beams. Each new patient should be affected to the first cluster with a probability 5/14 and to the second with a probability 9/14. In the first tests, the practitians will realize quickly a possible wrong assignment of a patient and have just a few quick changes to be done to correct this.

5. Conclusion

To solve our problem, we first defined a distance between two data on the ray-positioning circle. Secondly, we used the latter distance to establish the distortion we needed to minimize in order to have the best clustering. Then we built a convergent algorithm to find the minimizer of this distortion. This is a simulated annealing like algorithm and it includes an empirical search for cluster centres. Finally, we obtained, using this algorithm, good results both on simulated and real data sets. However, the algorithm can be improved. Including an automated search for the number of clusters would be interesting and useful. Measuring covariables on the patients could also help us to refine the prior probabilities of assignment in each cluster.

Acknowledgment : The authors greatly thank an anonymous referee for his useful comments and suggestions.

Appendix A. Proofs

Proof of Proposition 3.1:
Let K be a transition kernel on $E \times \mathcal{E}$ and denote by (x^ν) the Markov chain of the simulated annealing algorithm with transition kernel K. The proof of Proposition 3.1 is based on the following proposition of Bartoli and Del Moral [16].

Proposition Appendix A.1. *Assume that:*
(1) there exists λ a probability measure on $\mathcal{E}$, such that $\lambda(dx)K(x,dy) = \lambda(dy)K(y,dx)$,
(2) there exists an integer $p > 0$, $\varepsilon > 0$ and γ, a probability measure on $\mathcal{E}$, such that, for all $(x,A) \in E \times \mathcal{E}, K^p(x,A) \geq \varepsilon\gamma(A)$.
Take $\beta_\nu = C_0^{-1}\log(\nu + e)$ with $C_0 > p\,\mathrm{osc}_K(D)$. Then, for all $\varepsilon > 0$, $Pr(x_\nu \in D^\varepsilon) \to 1$ as $\nu \to \infty$ where

$$D^\varepsilon = \{x \in E, D(x) \leq \operatorname*{essinf}_\lambda(D) + \varepsilon\} \text{ and } \operatorname*{essinf}_\lambda(\mathrm{D}) = \sup\{\mathrm{a} \geq 0, \lambda(\mathrm{a} \leq \mathrm{D}) = 1\}.$$

We now proceed with studying the convergence of the algorithm defined above. Denote by $x = \{x_1, \ldots, x_J\}$ the subset of the centers of the clusters where $x_i = (x_{i1}, x_{i2}, \ldots, x_{i5})$.

The Markov chain associated with the algorithm is based on the transition kernel K defined below. For all, $x, y \in E$, let

$$K(x, dy) = \frac{1}{J} \sum_{j=1}^{J} \tilde{K}_j(x, dy)$$

with

$$\tilde{K}_j(x, dy) = \left(\prod_{g \neq j} \mathbf{1}_{\{x_g = dy_g\}} \right) K_j\left(x_j, dy_j\right)$$

and

$$K_j\left(x_j, dy_j\right) = \prod_{s=1}^{5} f\left(y_{js}; x_{js}, \sigma_g^2\right) dy_{js}$$

where $f\left(y_{js}; x_{js}, \sigma_g^2\right)$ is the density of the wrapped normal distribution defined in Section 3. Then, we verify (1) and (2) for our algorithm.
To state (1), it is sufficient to prove that $\tilde{K}_j(x, dy)\lambda(dx) = \tilde{K}_j(y, dx)\lambda(dy)$ for all $j \in \{1, \ldots, J\}$. Note that

$$\tilde{K}_j(x, dy)\lambda(dx) = \kappa^{-1} \left[\prod_{s=1}^{5} f\left(y_{js}; x_{js}, \sigma_g^2\right) \right] \left[\prod_{g \neq j} \mathbf{1}_{\{x_g = dy_g\}} dx_g \right] dy_j dx_j,$$

as

$$\lambda(dx) = \kappa^{-1} \prod_{j=1}^{J} dx_j$$

with $\kappa = \int_E dx_1 dx_2 \ldots dx_J$.
As $f\left(y_{js}; x_{js}, \sigma_g^2\right)$ is symmetric on x and y, (1) is verified.
It is also clear that (2) is checked with $\varepsilon = J!/J^J$, γ the product of J wrapped normal densities on the circle and $p = J$. □

Lemma Appendix A.1. *The function d is a distance.*

Proof of Lemma Appendix A.1:
First, for all subsets of five angles x, y it is clear that $d(x, y) = 0$ implies that $\forall s \in \{1, ..., 5\} : \delta(x_s, y_s) = 0$ for a $\sigma \in \mathcal{F}$ which gives us $x = y \quad [2\pi]$.
It is also clear that for all x, y we have $\delta(x, y) = \delta(y, x)$ because making permutations on x or on y gives the same results.
Finally, it is easy to see that, for all $x, y, z \in \mathbb{R}$, we have $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$. This relation gives us

$$\inf_{\sigma_1 \in \mathcal{F}} \sum_{s=1}^{5} \delta(x_{\sigma_1(l)}, z_l) = \inf_{\sigma_1, \sigma_2 \in \mathcal{F}} \sum_{s=1}^{5} \delta(x_{\sigma_1(l)}, z_{\sigma_2(l)}) \leq \inf_{\sigma_1, \sigma_2 \in \mathcal{F}} \quad \left(\sum_{s=1}^{5} \delta(x_{\sigma_1(l)}, y_l) + \sum_{s=1}^{5} \delta(z_{\sigma_2(l)}, y_l) \right)$$

which leads us to the conclusion. □

References

1. Wang Z, Zhang X, Dong L, Liu H, Wu Q, Mohan R. Development of methods for beam angle optimization for imrt using an accelerated exhaustive search strategy. *International Journal of Radiation Oncology, Biology, Physics* 2004; **60**.

2. Liu H, Jauregui M, Zhang X, Wang Z, Dong L, Mohan R. Beam angle optimization and reduction for intensity-modulated radiation therapy of non-small-cell lung cancers. *International Journal of Radiation Oncology, Biology, Physics* 2006; **65**.

3. Lei J, Li Y. An approaching genetic algorithm for automatic beam angle selection in imrt planning. *Computer Methods and Programs in Biomedicine* 2009; **93**.

4. Lee E, Fox T, Crocker I. Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy. *International Journal of Radiation Oncology, Biology, Physics* 2006; **64**.

5. Pugachev A, Li G, Boyer A, Hancock S, Le QT, Donaldson S, Xing L. Role of beam orientation optimization in intensity-modulated radiation therapy. *International Journal of Radiation Oncology, Biology, Physics* 2001; **50**.

6. Schmidt-Koenig K. *Circular Statistics in Biology*. Springer-Verlag, Berlin, 1975.

7. Batschelet E. *Circular Statistics in Biology*. Academic Press, London, 1981.

8. Jammalamadaka S, Bhadra N, Chaturvedi D, Kutty T, Majumdar P, Poduval G. Functionnal assessment of knee and ankle during level walking. In Krishnan T, ed., *Data Analysis in the Life Sciences*, 21–54. Indian Statistical Institute, Calcutta, 1986; .

9. Fuller S, Butcher S, Cheng R, Baker T. Three-dimensional reconstruction of icosahedral particles-the uncommon line. *Journal of Structural Biology* 1996; **116**.

10. Mardia K, Jupp P. *Directional Statistics*. John Wiley & Sons, New-York, 2000.

11. MacQueen J. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, University of California Press, 1967; .

12. Hartigan J, Wong M. A k-means clustering algorithm. *Journal of the Royal Statistical Society* 1979; **28**.

13. Kaufman L, Rousseeuw P. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, New-York, 1990.

14. Ng R, Han J. Efficient and effective clustering method for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, 1994; .

15. Xu R, Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Network* 2005; **16**.

16. Bartoli N, Del Moral P. *Simulations et algorithmes stochastiques : une introduction avec applications*. CEPADUES, Toulouse, 2001.

Table A.2: Real data set in 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	95	$3/4\pi$	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