A clustering Bayesian approach for non-ordered multivariate circular data

Christophe Abraham[1] & Nicolas Molinari[2] & Rémi Servien[3]

[1] *Montpellier SupAgro-INRA, UMR MISTEA 729, 2 place Pierre Viala, 34060 Montpellier Cedex 2 ; christophe.abraham@supagro.fr*
[2] *Université de Montpellier, IMAG, place Eugène Bataillon, 34095 Montpellier cedex 5 ; nicolas.molinari@inserm.fr*
[3] *Toxalim, Université de Toulouse, INRA, Toulouse ; remi.servien@inra.fr*

Résumé. On propose une méthode bayésienne de classification de données circulaires multivariées non ordonnées. Les observations consistent en des ensembles (non ordonnés) de k angles. Elles sont modélisées par des distributions normales projetées sur le cercle unité couplées à un processus de Dirichlet. Des paramètres supplémentaires sont introduits pour prendre en compte le caractère non ordonné des observations et pour modéliser leur variance. L'inférence est réalisée par un algorithme de type Metropolis-Hastings within Gibbs. La méthode est d'abord testée sur des simulations puis appliquée à des données de radiothérapie.

Mots-clés. Méthodes bayésiennes, données circulaires, processus de Dirichlet, données multivariées non ordonnées, distribution normale projetée, radiothérapie, classification.

Abstract. This paper presents a new 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 data are made up of (non ordered) sets of k angles. Additional parameters are introduced in order to take into account the non ordered particularity of the data and for modelling their variance. 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. 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. Bayesian Statistics, Circular data, Dirichlet process, Non-ordered multivariate data, Projected Normal Distribution, Radiotherapy machine data, Unsupervised clustering.

1 Introduction

The latest generation of radiotherapy machines projects multiple rays. The selection of the incident angles of the treatment beams may be a crucial component of IMRT planning.

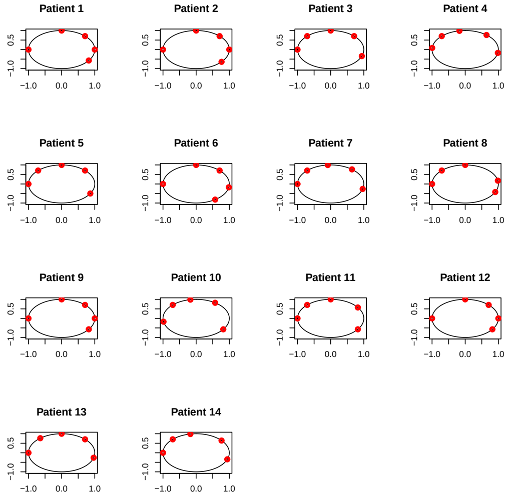

Figure 1: Real data set of 14 patients with $k = 5$ angles. A point on the circle represents the location of a treatment beam.

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. Therefore, 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. In Figure 1, a real data set from post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France, is represented. To understand the specificity of the data, consider a simple case of two patients with angles $\{1°, 60°, 100°, 150°, 180°\}$ and $\{60°, 100°, 150°, 180°, 359°\}$. To cope with technical difficulties, it is convenient to store the angles of each patient in a vector in increasing order (or in any other specific order). Note that the two patients should share the same cluster as the sets of angles are very similar (modulo 360) but that the derived vectors are not similar and are not likely to share the same cluster.

Abraham et al. (2013) defined a suitable distance on the circle and, given the number of clusters, proposed an algorithm based on simulated annealing to cluster the n patients. The number of clusters has to be supplied by the user and the final result reduces to a unique clustering whereas there are probably other clusterings that could be acceptable. To overcome these two drawbacks, we proposed a Bayesian method based on the Dirichlet Process mixture (DPM) for multivariate circular data.

2 Model

For simplicity, first assume that the ith of the n observations is given by a vector of k ordered angles $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})' \in [0, 2\pi[^k$ instead of a set $\{\theta_{i1}, \ldots, \theta_{ik}\}$; the latter case will be addressed later. Using a projected normal distribution (Presnell et al., 1998), 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 define θ_{ij} 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 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 :

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

where $DP(n_0 P_0)$ 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): some μ_i can have the same value; hence the clusters. 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 (ordered) vector $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})'$. We treat the observations as vectors for convenience and introduce the permutation parameter τ_i to compensate this simplification. More precisely, for all $\mu_i = (\mu_{i1}, \ldots, \mu_{ik})$ and all permutation τ_i, 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. The impact of the parameter τ_i can be understood by removing τ_i in (1). In this case, two observations with the same angles but in different orders would have a very low posterior probability of sharing the same cluster. We show that this probability is actually high with (1) thanks to the symmetry introduced by τ_i.

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 will be the case when $\mu_{i1}, \ldots, \mu_{ik}$ will be approximately equally spaced on a circle with center 0 and radius $\sqrt{\rho}$. We incorporate this prior information into the covariance matrix $\Sigma_0(\rho)$ which is given in a closed form as well as its inverse and its determinant. We also note that the variance of the angles θ_{ik} is actually controlled by ρ: the larger the value of ρ, the lower the variance of the angles. Inference on ρ can be handled using an inverse gamma prior $\rho \sim IG(a_\rho, b_\rho)$.

To summarize, the complete Bayesian model can be expressed as follows:

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

where $P_0(\rho) = N_{2k}(0, \Sigma_0(\rho))$, $\mu = (\mu_1', \ldots, \mu_n')'$, $\mathcal{P}$ is the set of permutations of $\{1, \ldots, k\}$ and $\mathcal{U}_{\mathcal{P}}$ denotes the uniform distribution on $\mathcal{P}$.

3 Inference and Application

We integrate over P as usual and we set $\theta = (\theta_1', \ldots, \theta_n')'$, $r = (r_1', \ldots, r_n')'$, $\tau = (\tau_1, \ldots, \tau_n)'$ and $\xi = (r', \mu', \tau', \rho, n_0)'$. Therefore the parameter reduces to ξ and the observation is θ. We sample from the posterior distribution of ξ with a Metropolis-Hastings-Within-Gibbs algorithm. We provide the complete conditional distribution for all the parameters except for r and μ. A Metropolis-Hastings step is needed for r and we use the SAMS sampler of Dahl (2003) for μ. This sampler may improve the merge-split sampler initially proposed by Jain and Neal (2004).

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) (Hubert and Arabie, 1985) to compare our obtained partition to the actual partition. 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. The majority clustering (mode of the posterior distribution of the clusterings) is the same as in Abraham et al. (2013) with a posterior probability equal to 30.5%. It can be noted that a credible region with a posterior probability of 71% is made up of only 4 clusterings.

The interested reader is referred to Abraham et al. (2017) for more details on the whole approach and the different results.

References

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

Abraham, C., Molinari, N., and Servien, R. (2017). A clustering bayesian approach for radiotherapy x-ray beam bouquets. *Submitted. https://hal.archives-ouvertes.fr/hal-01326166 .*

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

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

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

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

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

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.

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.