MODEL-BASED CLUSTERING FOR TORUS DATA

Luca Greco [1], Antonio Lucadamo [2] and Claudio Agostinelli [3],

[1] University G. Fortunato, Benevento, Italy, (e-mail: l.greco@unifortunato.eu)

[2] Department DEMM, University of Sannio, Italy (e-mail: antonio.lucadamo@unisannio.it)

[3] Department of Mathematics, University of Trento, Italy, (e-mail: claudio.agostinelli@unitn.it)

ABSTRACT: Torus data are multivariate circular observations that arise as measurements on a periodic scale and are often recorded as angles. In this paper, we focus on parsimonious model based clustering for torus data by building on the `mclust` methodology. Therefore, covariance constraints are imposed on the completely general heterogeneous clustering model allowing a flexible and general framework to clustering torus data.

KEYWORDS: torus data, model-based clustering, wrapped distribution

1 Introduction

Torus data are multivariate circular observations. Many applications involve torus data in several fields: protein bioinformatics, wind directions, animal movements, people orientation, human motor resonance, robotics, astronomy, meteorology, geology, medicine, oceanography. Actually, multivariate circular data can be thought of as points on a p-torus $\mathbb{T}^p$, $p > 1$, whose surface is obtained by revolving the unit circle in a p-dimensional manifold. The multivariate wrapped normal (WN) distribution is a very attractive model for torus data (Mardia & Jupp, 2000). In Greco *et al.*, 2022, the WN distribution has been proved to be very useful in modeling mixtures of torus data and providing an effective tool for model based clustering and classification, but only under a completely general heterogeneous clustering model. In this paper, by paralleling a widely used methodology for *linear* data on $\mathbb{R}^p$, we focus on parsimonious model based clustering for torus data by building on the `mclust` methodology (Scrucca *et al.*, 2016).

2 Parsimonious model based clustering

Let us consider a sample of size n of torus data $y = (y_1, y_2, \dots, y_n)$, from the finite mixture model with density function

$$f^\circ(y;\tau) = \sum_{g=1}^{G} \delta_g m^\circ(y;\theta_g), \tag{1}$$

where we set $\tau = (\delta_1, \dots, \delta_G, \theta_1, \dots, \theta_G)$, G denotes the number of groups, δ_g are membership probabilities, $\delta_g > 0$, $\sum_{g=1}^{G} \delta_g = 1$, $\theta_g = (\mu_g, \Sigma_g)$ are component specific location and scatter and $m^\circ(y;\theta_g) = \sum_{j \in \mathbb{Z}^p} m(y + 2\pi j;\theta)$ is the wrapped density function, where j is the vector of wrapping coefficients and $m(\cdot)$ the corresponding unwrapped density. Let $m^\circ(y;\theta_g)$ be the density of a WN distribution (being $m(\cdot)$ the normal density). Building on `mclust`, we enforce constraints on the scatter matrices Σ_g using the parsimonious models of Celeux & Govaert, 1995 that can be obtained by means of the eigenvalue decomposition of the covariance matrices of the form $\Sigma_g = \lambda_g D_g A_g D_g^\top$, where $\lambda_g = [\det(\Sigma_g)]^{1/d}$, $d = 1, 2, \dots, p$, is a measure of the volume of the g^{th} cluster, A_g is a diagonal matrix with the eigenvalues of Σ_g, with $\det(A_g) = 1$, specifying the shape and D_g is an orthogonal matrix whose columns are given by the eigenvectors of Σ_g which determines the orientation.

In order to make estimation of wrapped models feasible, the infinite sum over $\mathbb{Z}^p$ is replaced by a sum over the Cartesian product $\mathcal{C}_J = \otimes \mathcal{J}^p$, $\mathcal{J} = (-J., -J+1, \dots, 0, \dots, J-1, J)$, for some J providing a good approximation. Then, maximum likelihood estimation of the model in (1) follows from the maximization of the mixture log-likelihood function.

$$\ell(\tau) = \sum_{i=1}^{n} \log f^\circ(y_i;\tau) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \delta_g \sum_{j \in \mathcal{C}_J} m(y + 2\pi j;\theta_g) \right]. \tag{2}$$

The operations of mixing and wrapping commute, and (2) can be rewritten as

$$\ell(\tau) = \sum_{i=1}^{n} \log \left[\sum_{j \in \mathcal{C}_J} \sum_{g=1}^{G} \delta_g m(y + 2\pi j;\theta_g) \right] = \sum_{i=1}^{n} \log f(y_i + 2\pi j;\tau)$$

where $f(y + 2\pi j;\tau)$ is a mixture density for linear data.

Observe that the wrapping coefficients j are unknown. Then, they can be considered as latent variables and the observed torus data y as being incomplete. In the following, maximum likelihood estimation relies on a data augmentation approach and is performed according to a suitable Classification Expectation Maximization algorithm. The point is that there are two sources of incompleteness in (2): one given by the wrapping coefficient vectors, the other from group memberships. The proposed algorithm iterates between an outer Classification Expectation (CE) step, in which the circular data are unwrapped to fitted linear data $\hat{x} = y + 2\pi\hat{j}$ (see Nodehi *et al.*, 2021), and an inner run of a classical EM algorithm for (linear) finite mixtures using the fitted linear data. Actually, the algorithm maximizes the (approximated) classification log-likelihood function based on the complete torus data (y, j):

$$\ell_c(\tau) = \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_J} v_{ij} \log \left[\sum_{g=1}^{G} \delta_g m(y_i + 2\pi j; \theta_g) \right] , \tag{3}$$

where $v_{ij} = 1$ or $v_{ij} = 0$ according to wheter y_i has $j \in \mathcal{C}_J$ as wrapping coefficients vector.

Formal approaches to infer the number of clusters and select the best model among the available parsimonious alternatives can be based on the value of the penalized complete log-likelihood function (3) at convergence or, alternatively, of the incomplete data log-likelihood function (2). Classical model selection criteria are given by the Bayesian Information Criterion (BIC) or the integrated complete-data likelihood criterion (ICL).

3 A numerical example

Let us consider a synthetic data example to illustrate the proposed methodology. The sample size is $n = 500$. Data have been generated according to a bivariate WN mixture model with two components and unbalanced memberships probabilities, imposing an EII covariance structure. Starting values are driven from cluster-wise constrained maximum likelihood estimation under the assumed model from an initial partition obtained using the angular separation distance and the Ward agglomerative method. The BIC selects the right model, in this example. Cluster assignments are plotted in Figure 1. Tolerance ellipses are also given, based on the normal model. Note that the data have been represented on a flat torus, that is the same data structure repeats itself on the Euclidean space to account for the wraparound nature of the data. i.e. data are represented for different js. The procedure has been repeated 500 times. The model EII has been correctly selected in 95.6% of the simulations. The average Adjusted Rand Index (ARI) between the inferred partitions and the true component memberships is 0.963.

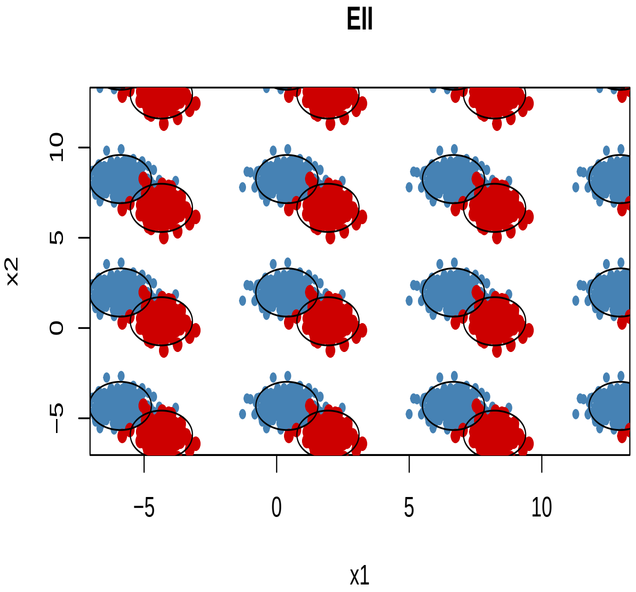

Figure 1. *Cluster assignments and tolerance ellipses under the EII model.*

References

Celeux, G., & Govaert, G. 1995. Gaussian parsimonious clustering models. *Pattern recognition*, **28**(5), 781–793.

Greco, L., Novi Inverardi, P., & Agostinelli, C. 2022. Finite mixtures of multivariate Wrapped Normal distributions for model based clustering of p-torus data. *Journal of Computational and Graphical Statistics*.

Mardia, K. V., & Jupp, P. E. 2000. *Directional statistics*. Wiley Online Library.

Nodehi, A., Golalizadeh, M., Maadooliat, M., & Agostinelli, C. 2021. Estimation of parameters in multivariate wrapped models for data on ap-torus. *Computational Statistics*, **36**, 193–215.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, **8**(1), 289.