

Case study - NGC 2419 globular cluster

Lecturer: Riccardo Corradin

We consider a set of measurements related to stars potentially belonging to the NGC 2419 globular cluster, available in the *NGC2419.csv* file at the module page. For each star, we observe

- The line of sight velocity V .
- The metallicity $[Fe/H]$, a measure of the abundance of iron relative to hydrogen.
- The two-dimensional projection on the plane of the sky of the star position (X, Y) .

We want to identify a cluster of homogeneous star, potentially being part of the globular cluster, while also identifying stars that are noisy observations in this dataset. To this end, we consider a mixture model of the form

$$\begin{aligned} \mathbf{Y}_i &\sim \sum_{j=1}^k w_j \phi(\mathbf{y}_i \mid \boldsymbol{\mu}_j, \Sigma_j), & i = 1, \dots, n, \\ \mathbf{w} &\sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k), \\ (\boldsymbol{\mu}_j, \Sigma_j) &\sim \text{NIW}(\mathbf{m}_0, k_0, \nu_0, \Lambda_0), & j = 1, \dots, k. \end{aligned}$$

Specifically, prior to analyze the data, we marginally standardize each observed variable. Further, we consider $k = 10$, $\alpha_1 = \dots = \alpha_k = \frac{1}{k}$, $\mathbf{m}_0 = \mathbf{0}$, $k_0 = 0.1$, $\nu_0 = 7$ and $\Lambda_0 = \text{diag}_4(1)$.

- Produce a code that produce a sample of latent partition from the posterior distribution of interest. Check with a synthetic dataset that the function works.
- Sample 2000 realizations from the posterior distribution of interest, after 500 burnin iterations. Check the sampled chain using the entropy of the visited partitions.
- Write a function that, given the sampled partitions, produces a point estimate under the Binder loss function.
- Plot the observed data along with the point estimate of the latent partition.