BAYESIAN STATISTICAL MODELS - University of Milano-Bicocca

Case study - Clustering stars

Lecturer: Riccardo Corradin

We consider a set of synthetic data, sampled from the following distribution

$$f(y) = 0.25\phi(y; -3\gamma, 1) + 0.25\phi(y; 0, 1) + 0.25\phi(y; 3\gamma, 1)$$

We first consider a sample of size $n = 150$. A priori we consider a model of the form

$$\begin{aligned} y_i \mid \mu^*_{1:k}, \sigma^{2*}_{1:k}, s_i &\sim N(\mu^*_{s_i}, \sigma^{2*}_{s_i}), && i = 1, \dots, n \\ \boldsymbol{w} &\sim Dir(\alpha, \dots, \alpha) \\ \mu^*_j, \sigma^{2*} &\sim NIG(m_0, k_0, a_0, b_0), && j = 1, \dots, k. \end{aligned}$$

with $m_0 = 0$, $k_0 = 0.1$, $a_0 = 2$, $b_0 = 1$, $\alpha = 1$.

- Implement a function that produce a sample of latent partitions from the posterior distribution of interest, updating the cluster weights and the cluster-specific parameters.
- Sample $2\,000$ realizations from the posterior distribution, after 500 burnin iterations, using the previous function. 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 and the produced mixture model.
- Repeat the estimates with $k_0 = 10$.
- Study the functions in terms of efficiency (computational time) and precision (point estimate close to the true one), varying the sample size $n \in \{150, 250, 1000\}$ and the dispersion of the data generating process γ.

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

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

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

- Implement a function that produce a sample of latent partition from the posterior distribution of interest. Check with a synthetic dataset that the function works.
- Sample $2\,000$ 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.