Durham University

An Exploration into Bayesian Nonparametrics using BNPmix

Jack Lloyd

Department of Mathematical Sciences

Michaelmas 2023

Lecturer: Konstantinos Perrakis

1 Introduction

1.1 The Wine Data

In this brief report we will be analysing the Wine dataset. These data are the results of a chemical analysis of wines grown in Italy but are from three different cultivars[1]. The dataset that we will explore contains 13 chemical variables found in each of the three types of wine. This dataset has been included because we can apply our models to investigate hidden patterns or clusters in the datasets among the three cultivars based on certain variables.

We will perform analysis using the R package BNPmix. We will perform univariate and multivariate analysis. More specifically we will estimate the posterior density of variables, finding the maximum a priori number of clusters in our variables and performing posterior inference for the parameters. This analysis should aim to give us a good estimate of the structure of the posterior density and inference for each of the clusters. We will also perform Bayesian non-parametric regression analysis, where we aim to model a response variable as a linear function of one or more predictors.

In particular, we will perform univariate analysis on the Flavanoids[2] variable. This refers to the flavanoid content in the wine and we will examine its posterior distribution and explore how the values are clustered. We will consider the posterior inference for the predictors to obtain an idea about the mean and variance of each cluster.

2 Data Description

In this section we will get to know our Wine dataset much better. We will start by introducing our variables. Our dataset consists of 174 wines from 3 different cultivars measured across 13 different variables. Many of the variables are measures of different chemical content which is important to consider in the production of wine. We see the histograms of the variables in **Figure 1**.

As we can see, most of our variables are unimodal however there are some exceptions. For example, we can see that Phenols and Flavanoids have two separate spikes so we can argue that the distributions of these variables may be bimodal as they have two distinct separate peaks. Due to this, Flavanoids will be the focus of our density estimation in further sections. We can also see that Alcohol and Alcalinity have rather symmetric distributions since their histograms are univariate with the peak sitting in the middle of the range of values and the amount of observations seeming to decrease at a fairly similar rate as they move towards the smaller quantiles.

[1] A wine cultivar refers to a specific grapevine that has been intentionally selected and cultivated by vineyards for desirable characteristics.

[2] Flavanoids are a class of polyphenolic secondary metabolites found in plants. They are well known for their beneficial effects on health.

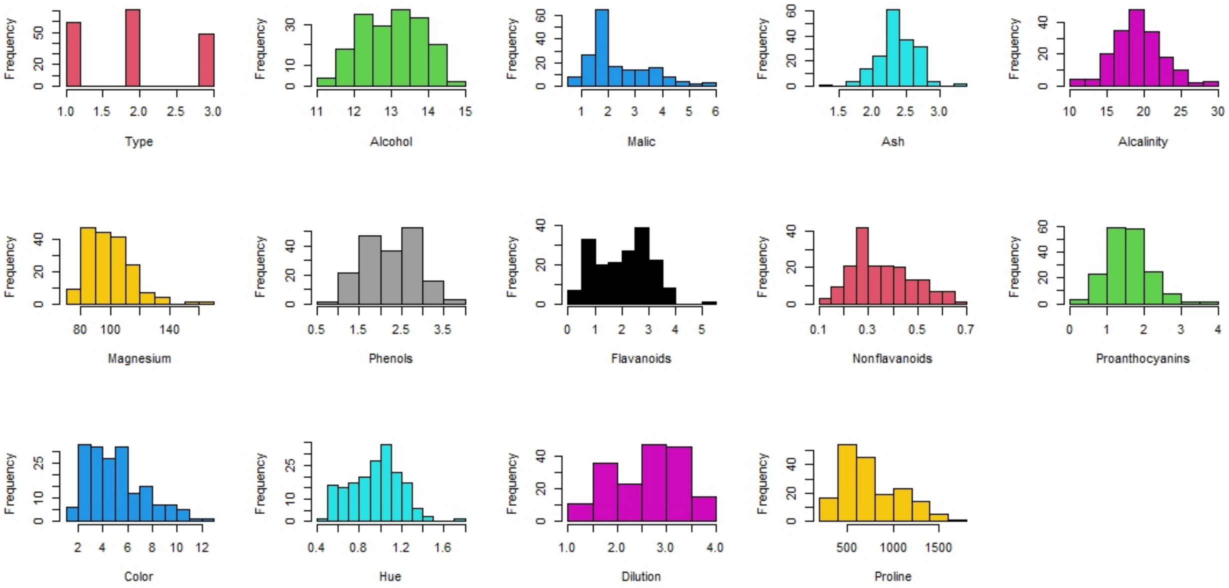


Figure 1: Histograms of all variables in `Wine`

3 Methodology

3.1 Univariate and Multivariate Density Estimation

We will start by performing univariate density estimation on the `Flavanoids` variable. `BNPmix` uses the Pitman-Yor process which is a more general form of the Dirichlet Process. In the DP case we would say

$$G \sim DP(\alpha G_0)$$

where α is our precision and G_0 is our centering measure. In the `BNPmix` package, we assume that

$$G \sim PY(\beta, \alpha, G_0)$$

where $\beta = [0, 1)$ is our discount parameter, α us the strength parameter and G_0 is centering measure. Note when $\beta = 0$ we just have our well known $DP(\alpha, G_0)$.

We must understand the stick-breaking representation to understand how the package approaches DPM models. Simply, `BNPmix` allows us to fit infinite mixtures of normal distributions with the stick breaking representation.

$$p(y_i) = \sum_{k=1}^{\infty} \pi_k N(y_i | \mu_k, {\sigma_k}^2)$$

We have prior specifications for μ_k and σ^2 based on a conditional normal/inverse-gamma centering measure $G_0(\mu_k, {\sigma^2}_k) = G_0(\mu_k|{\sigma^2}_k)G_0({\sigma^2}_k)$ as defined below

$$\mu_k|{\sigma_k}^2 \sim G_0(\mu_k|{\sigma_k}^2) \equiv N(m_0, \frac{{\sigma_k}^2}{k_0}) \quad \text{and} \quad {\sigma_k}^2 \sim G_0({\sigma_k}^2) \equiv \text{InvGamma}(a_0, b_0)$$

In the `PYcalibrate` function, we must specify the prior expected number of clusters. After observing the `Flavanoids` scatter plot (see Appendix - **Figure 6**), we set this number to 3 as there appear to be 3 clusters. If this skews our model output, we may change this.

In the multivariate case, we assume that we have a data matrix $\mathbf{Y}$ with dimensions $n \times q$. In this case, the stick breaking representation is

$$p(\mathbf{y}_i) = \sum_{k=1}^{\infty} \pi_k N_q(\mathbf{y}_i|\mu_k, \Sigma_k)$$

for $\mathbf{y}_i \in \mathbf{R}^q$. The prior distributions for μ_k and Σ_k are based on the multivariate conditional conjugate normal/inverse-Wishart design as a centering measure (see appendix for further details). In our method, we will perform multivariate analysis using the variables `Alcohol`, `Flavanoids` and `Alcalinity`. We only choose 3 variables as `R` cannot perform MV density estimation using all the predictors due to a lack of computational power. We set our grid support (list containing certain values) for each variable to have length equal to 20 so our data frame containing all possible combinations of the variable vectors had dimensions 8000×3. We specified our grid supports to have equally spaced values in a range from the minimum to the maximum value of each variable.

3.2 DPM Regression Models

In the regression setting we have the response $y = (y_1, ..., y_n)^T$ and the $n \times (p+1)$ matrix of predictors x_0. We use the stick breaking process again with

$$p(y_i|x_i) = \sum_{k=1}^{\infty} \pi_k N(y_i|x_{0i}^T\beta_k, {\sigma_k}^2)$$

where we have prior distributions for β_k and σ_k^2 based on a conditional multivariate normal and inverse gamma centering measure, with

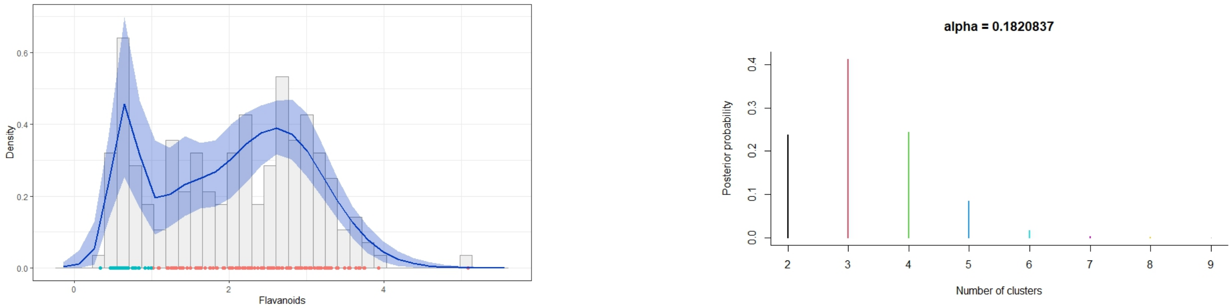
$$\beta_k \sim G_0(\beta_k) \equiv N(m_0, S_0) \quad \text{and} \quad \sigma_k^2 \sim G_0(\sigma_k^2) \equiv \text{InvGamma}(a_0, b_0)$$

In our regression models we are fitting a different distribution to each y_i conditional on x_{0i}. In DPM regression models the conditional distribution can be any distribution and they can take different forms across the data, meaning they are extremely flexible.

4 Results

4.1 Density Estimation using BNPmix

We will first examine the results from performing univariate density estimation on the `Flavanoids` variable. We altered our prior expectation of clusters to 2 as this gave us a more reliable output for average number of clusters.



(a) Flavanoid Density Estimation.

(b) Posterior Probabilities of Clusters

Figure 2: Univariate Density Estimation of `Flavanoids`

After fitting our initial model, we see the density estimator output in **Figure 2a**. Our univariate density estimator found there were a minimum of 2 groups and a maximum of 9 groups, on average outputting a model with 3.2464 clusters. We see in **Figure 2b** that our MAP (Maximum a Posteriori) estimate of the number of clusters in `Flavanoids` is 3 clusters, with the probability of visiting a model with 3 clusters being equal to 0.4121. We also performed posterior inference for the parameters, where we worked conditionally on the MAP estimate for the number of clusters being equal to $K = 3$. **Figure 3** highlights our posterior densities for the means of each cluster. We see that our model identifies Cluster 1 to have $\mu_1 \approx 2.417$, Cluster 2 to have $\mu_2 \approx 0.649$. We get Cluster 3 to have $\mu_3 \approx 1.360$ but it is clear that our model struggles to identify a single point.

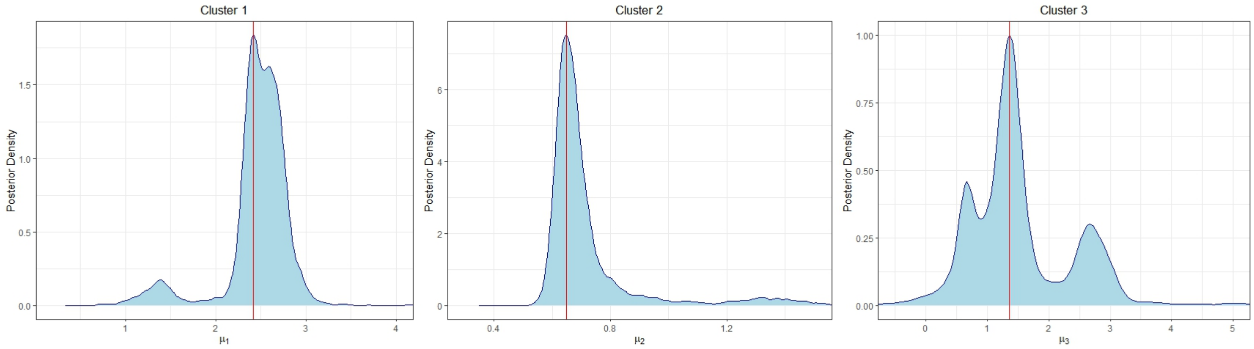


Figure 3: Posterior Means for Each Cluster

4.2 Multivariate Density Estimation using `BNPmix`

In this section we will discuss the results of multivariate density estimation using the variables `Alcohol`, `Flavanoids` and `Alcalinity`. In the initial multivariate model we fitted, it found a minimum of 2 groups, a maximum of 4 groups and an average of 2.902 clusters. we identified the MAP number of clusters to be equal to 3 with $\mathbb{P}[K = 3|\boldsymbol{Y}] = 0.884$. Here we will include some of the plots and comment on the results. The full set of plots is available in the appendix (see **Figure 10**).

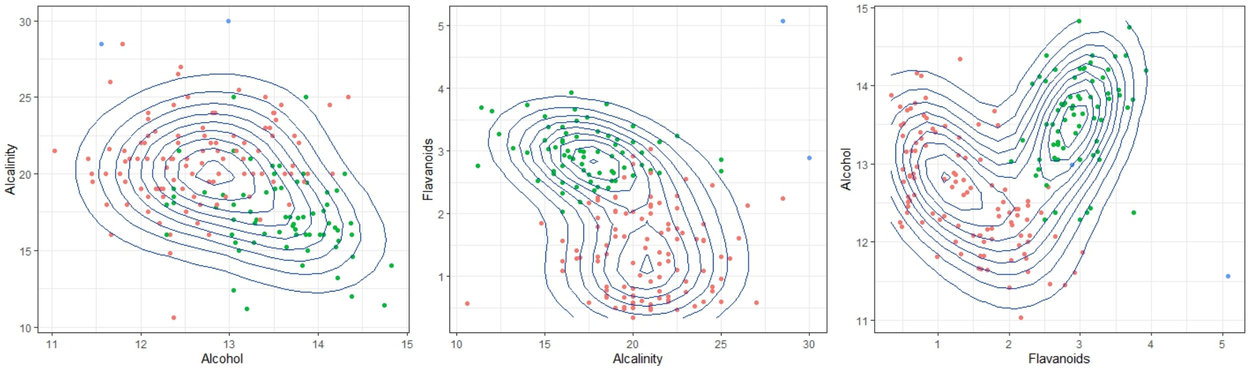


Figure 4: Bivariate Density Estimators

From left to right, **Figure 4** show the Bivariate Density Estimators of *(Alcohol, Alcalinity)*, *(Alcalinity, Flavanoids* and *(Flavanoids, Alcohol)*. One can see that our density estimators are picking up the majority of the data. The marginal univariate distributions (**Figure 10**) show that our model has correctly identified the shape of the distribution for each variable when we compare to the histograms.

4.3 DPM Regression Models using `BNPmix`

In this section, we considered the level of alcohol in the wine as the response variable. We started by observing the histogram of alcohol to give us an idea of the distribution and plotted `Alcohol` against the other variables in our dataset.

We can say that there seems to be some form of relationship between `Alcohol` and `Flavanoids`. We then used `Flavanoids` as our single predictor in our DPM regression. Since it would be too much to ask for `R` to provide distributions of Alcohol$_i$ given every Flavanoid$_i$, we chose five values from Flavanoids, specifically the 1% quantile, 1^{st} quartile, median, 3^{rd} quartile and the 99% quantile.

We see in **Figure 5** that our model expresses uncertainty in the posterior conditional densities of Alcohol. As we consider larger values of `Flavanoids`, our posterior conditional densities start to become bimodal which shows our model struggles to identify the value of `Alcohol` which corresponds with the respective value of `Flavanoids`. This may be due to there being few large values of flavanoids so more uncertainty is expressed (as we see in the large credible bands). The MAP estimate for the number of clusters is $K = 2$, which we then used when performing further posterior analysis (see **Section 6.2.2** in Appendix). We obtain that our two clusters have the linear regression equations $a_1 = 12.923 + 0.225 f_1$ for Cluster 1 and $a_2 = 12.551 - 0.166 f_2$ for Cluster 2 where a_i and f_i are the `Alcohol` and `Flavanoids` values in Cluster i.

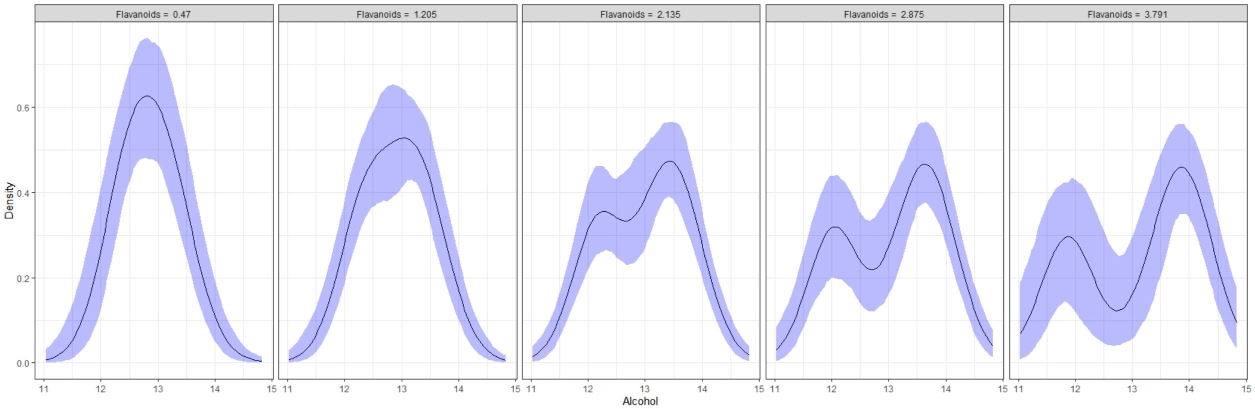


Figure 5: Posterior Conditional Densities of Alcohol for specified values of Flavanoids

5 Final Thoughts

In this report we displayed how we can use the `R` package `BNPmix` to analyse the `Wine` dataset. In our univariate density estimation, our model managed to identify the MAP estimate for the number of clusters in `Flavanoids` to be $K = 3$. This was in line with what was expected after considering the scatter plot (**Figure 6** in appendix). We see that despite getting the overall shape of the distribution correct, the model still had a considerable amount of uncertainty, displayed in the credible bands of **Figure 2b**. We see that, when performing posterior inference for the parameters (using the MAP estimate of $K = 3$ clusters) our posterior densities were unimodal apart from the posterior mean in Cluster 3. This factor and the incredibly low posterior probablity of an observation belonging to Cluster 3 lead us to question whether our model had made a mistake in identifying a third cluster. When considering further improvements, one could perform further analysis to explore how our model identified this third cluster.

In our multivariate density estimation, where we used variables `Alcohol`, `Alcalinity` and `Flavanoids`, we saw consistent results. Our bivariate density estimation plots (**Figure 4 and 9**) showed that our model seemed to identify the patterns between clusters. In **Figure 9**, our marginal density estimators identified `Alcohol` and `Alcalinity` as unimodal (which we know to be true from the initial histograms in **Figure 1**) and suggested that `Flavanoids` had a more bimodal distribution, which we also expected.

In our regression, we saw our model struggle in identifying the posterior densities of `Alcohol` given our `Flavanoids` values. One further improvement that could be made to perhaps reduce this uncertainty would be to increase the MCMC sample amount as this gives our regression model more samples to work with. We saw the credible bands the largest at our largest value of `Flavanoids` so another improvement could be trying to obtain a larger wine dataset with more wines measured, since then one would find more large values of `Flavanoids`. In practise this would be difficult.

Overall, we can say that our analysis using `BNPmix` had mixed success. The models we created faced two main issues: large uncertainty in places and potential misidentification of clusters. Both of these issues can be solved by using computers with greater power, so we can greatly increase our MCMC samples to give us a better idea of the posterior distribution.

6 Appendix

This appendix will mainly contain further plots which have been referenced in **Sections 2,3,4 and 5**. We will provide small comments on these results.

6.1 Methodology: Further Plots and Theory

6.1.1 Univariate and Multivariate Density Estimation

We chose to set our prior expected number of clusters in the `Flavanoids` variable after observing its scatter plot. One can argue there are 3 main clusters in the data from observing this plot.

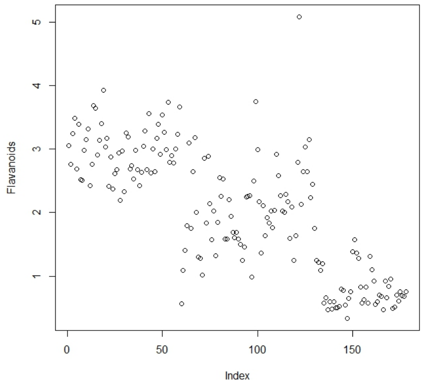


Figure 6: Scatter Plot for the `Flavanoids`

6.2 Results: Additional Plots and Comments

6.2.1 Univariate Density Estimation

We include the posterior densities for the variances of each cluster in **Figure 7**. We observe that the model outputs the variances with the highest posterior densities are $\sigma_1^2 = 0.5604$, $\sigma_2^2 = 0.0262$ and $\sigma_3^2 = 0.0500$. We can see that our model struggled to identify a variance for Cluster 1 due to the width of the peak.

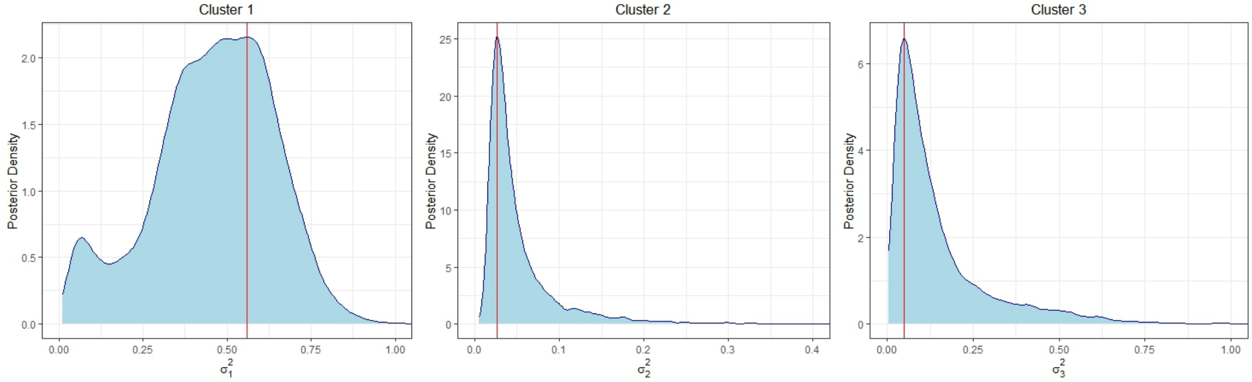


Figure 7: Posterior Variances for each Cluster

We also consider the posterior densities of the probability of each cluster i, π_i in **Figure 8**. Our model concludes that the an observation is most likely to belong to Cluster 1, with $\pi_1 = 0.7808$. We then have $\pi_2 = 0.2014$ and $\pi_3 = 0.0058$, suggesting that we are extremely unlikely to find an observation in Cluster 3. This leads us to question if Cluster 3 is a valid cluster or just a point of overlap between Cluster 1 and 2 which has been misidentified as a separate cluster.

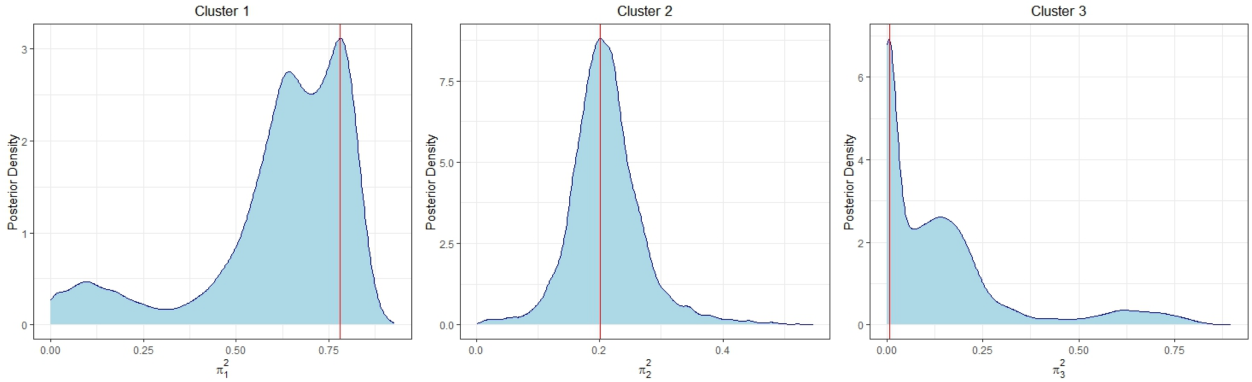
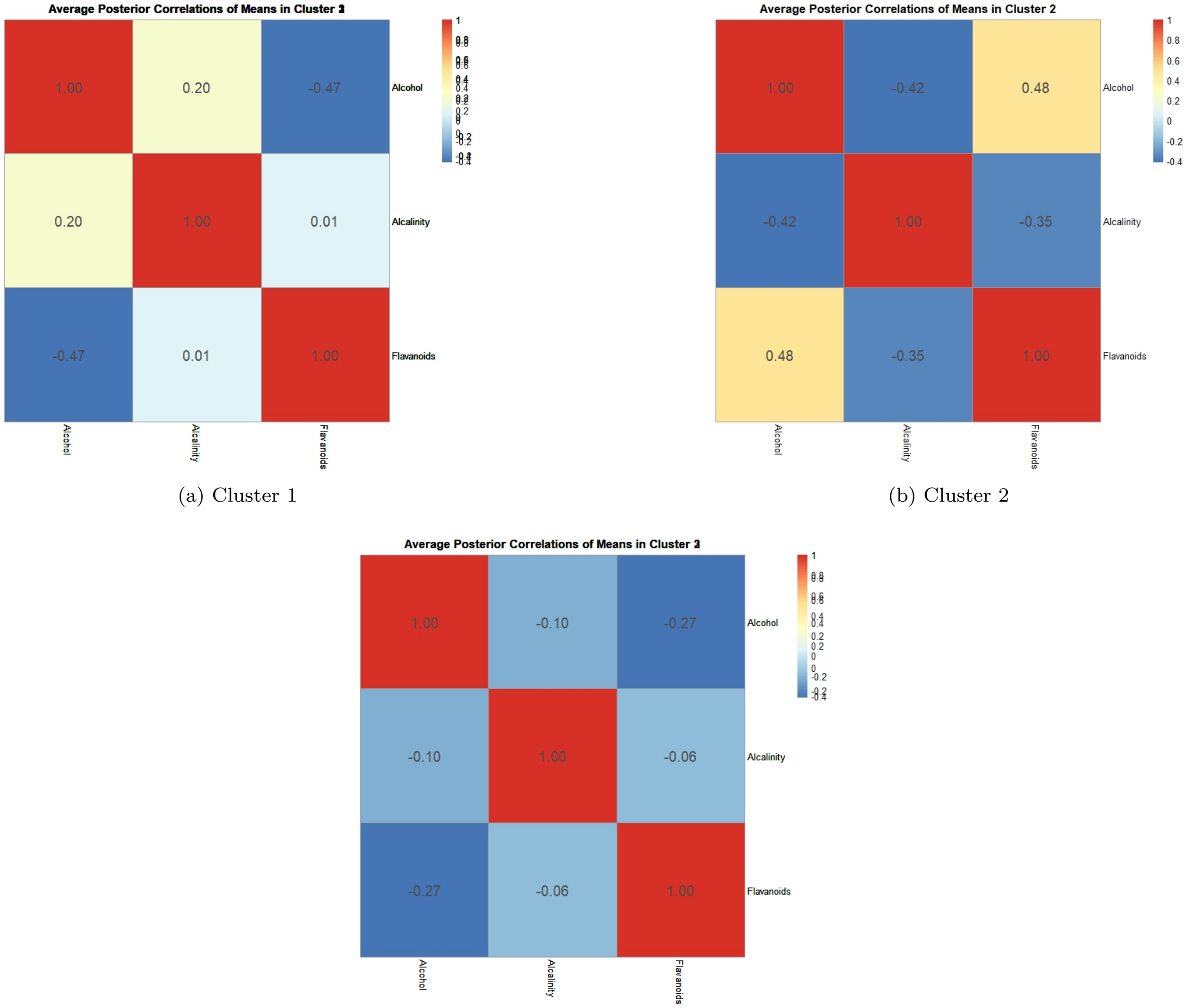


Figure 8: Posterior Probabilities of Each Cluster

6.2.2 Multivariate Density Estimation

Here we will provide some further results from our multivariate density estimation. In the figures on the diagonal of **Figure 10** we see that our model has identified the distribution of `Alcohol` and `Alcalinity` to be unimodal and the distribution of `Flavanoids` to be bimodal. We see that from the bivariate density estimators that our model has managed to identify the majority of patterns and clusters in the data. We also consider the posterior correlations between variables in each cluster (see **Figure 9**), where interestingly we saw that in Cluster 3 (**Figure 9c**) there was very little correlation between variables compared to Cluster 1 and Cluster 2. This could suggest that we have a cluster within the dataset in which the variables do not follow any form of pattern, or that our clustering algorithm has failed to identify the pattern correctly.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 9: Posterior Correlation Between Variables in each Cluster

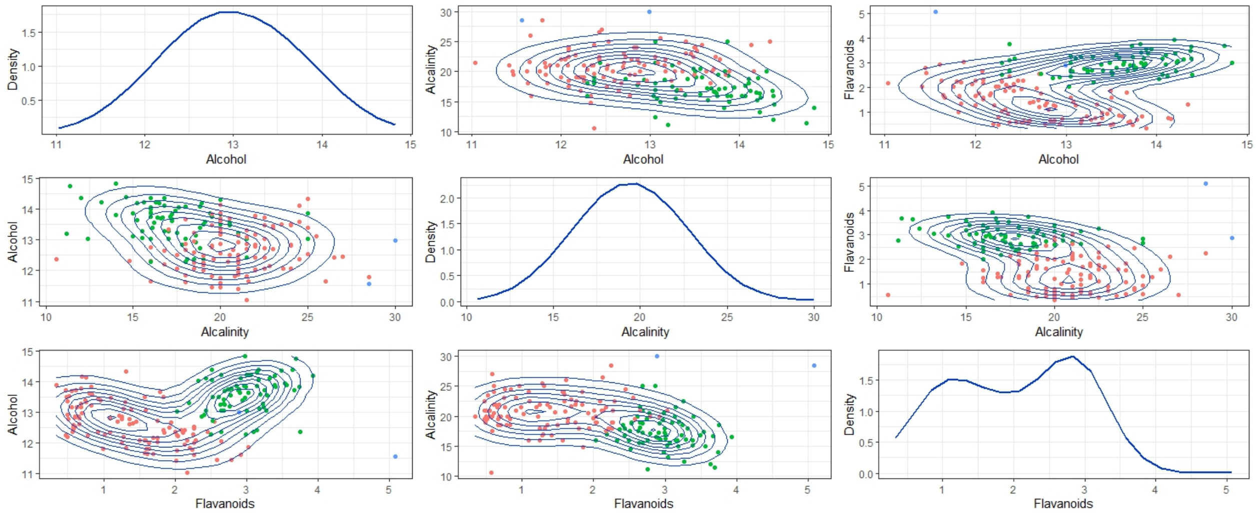


Figure 10: All Plots for Multivariate Density Estimation

6.2.3 Regression Analysis

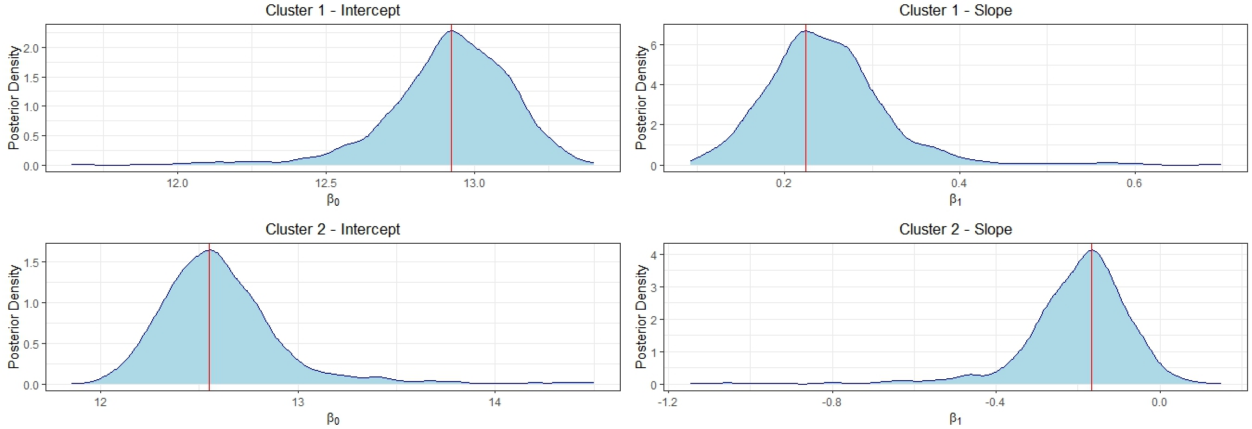


Figure 11: Posterior Densities for Regression Coefficients

We can see that our posterior densities are all unimodal, which shows that our model has coped with performing regression with one predictor. However the peaks are still rather wide, which shows that the model was still unsure of where in the region the point with the highest posterior density sat. We could look to reduce the width of these peaks by using a larger MCMC sample size.