Análise de Componentes Principais, Clustering em K-Médias, Clustering Hierárquico e Clustering por Misturas

Métodos Estatísticos em Data Mining

Mestrado em Engenharia Matemática
Faculdade de Ciências da Universidade do Porto

Ana Beatriz Carvalho (202107558)
Ana Rita Chamusca (202106828)

1 O Dataset e Tratamento de Dados

O Dataset detalha dados sobre cirrose e a sua evolução em pacientes provenientes da Mayo Clinic, entre 1974 e 1984.

Originalmente, era constituído por 418 instâncias - pacientes - e 20 variáveis, 8 categóricas e 12 numéricas.

Sendo que uma das variáveis numéricas era o ID dos pacientes, removemos do Dataset. Removemos também uma variável categórica, pois não seria útil para a análise dos métodos a aplicar.

Omitimos ainda todos os missing values, o que reduziu significativamente as instâncias para estudo, para um total de 276 instâncias e 18 variáveis.

Para além disso transformamos as restantes variáveis categóricas em variáveis binárias, para ser possível aplicar os métodos estudados.

2 PCA - Análise de Componentes Principais

Aplicando o PCA em R, obtivemos os seguintes resultados:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
Standard deviation     2.0497758 1.4284851 1.17196900 1.10404673 1.05078412 1.01275135
Proportion of Variance 0.2334212 0.1133650 0.07630619 0.06771773 0.06134151 0.05698141
Cumulative Proportion  0.2334212 0.3467862 0.42309234 0.49081007 0.55215159 0.60913299
                           Comp.7    Comp.8    Comp.9    Comp.10   Comp.11    Comp.12
Standard deviation     0.98496965 0.93867660 0.88015364 0.86497503 0.7896363 0.75628949
Proportion of Variance 0.05389807 0.04895076 0.04303725 0.04156566 0.0346403 0.03177632
Cumulative Proportion  0.66303106 0.71198182 0.75501907 0.79658473 0.8312250 0.86300135
                          Comp.13    Comp.14    Comp.15    Comp.16    Comp.17    Comp.18
Standard deviation     0.71358052 0.69893885 0.68566952 0.64263282 0.56003963 0.52105568
Proportion of Variance 0.02828873 0.02713975 0.02611904 0.02294316 0.01742469 0.01508328
Cumulative Proportion  0.89129008 0.91842983 0.94454887 0.96749203 0.98491672 1.00000000
```

Figure 1: Sumário do PCA

Aplicando o critério de Kaiser, escolhemos utilizar seis componentes principais. Este critério foi utilizado após a standardização dos dados, e foi escolhido pois garantia uma quantidade "manuseável" de componentes principais.

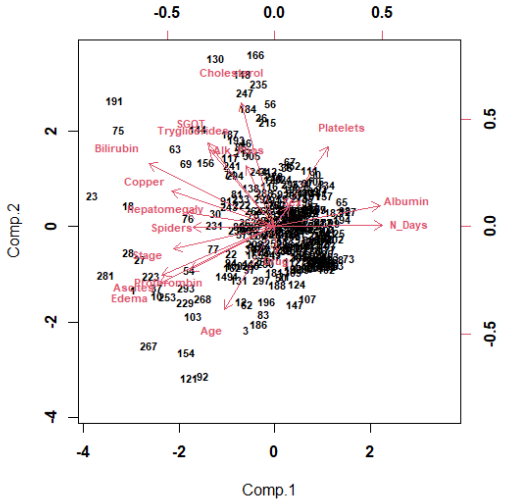

Figure 2: Biplot das duas primeiras componentes principais

Devido ao facto das duas primeiras componentes principais explicarem relativamente pouco da variância total (cerca de 35%), não é possíveis retirar conclusões deste Biplot. De facto, qualquer par de componentes principais que possamos escolher não revelará um gráfico mais esclarecedor, visto que, individualmente, as componentes principais explicam pouco da variância total.

3 Clustering em K-Médias

Em primeiro lugar, verificamos, pelo método do cotovelo, quantos clusters seria adequado utilizar no método de K-Médias. Decidimos aplicar este método, e todos os que seguirão, às primeiras seis componentes principais, de modo a fazer a ligação entre a aplicação do PCA e os restante métodos.

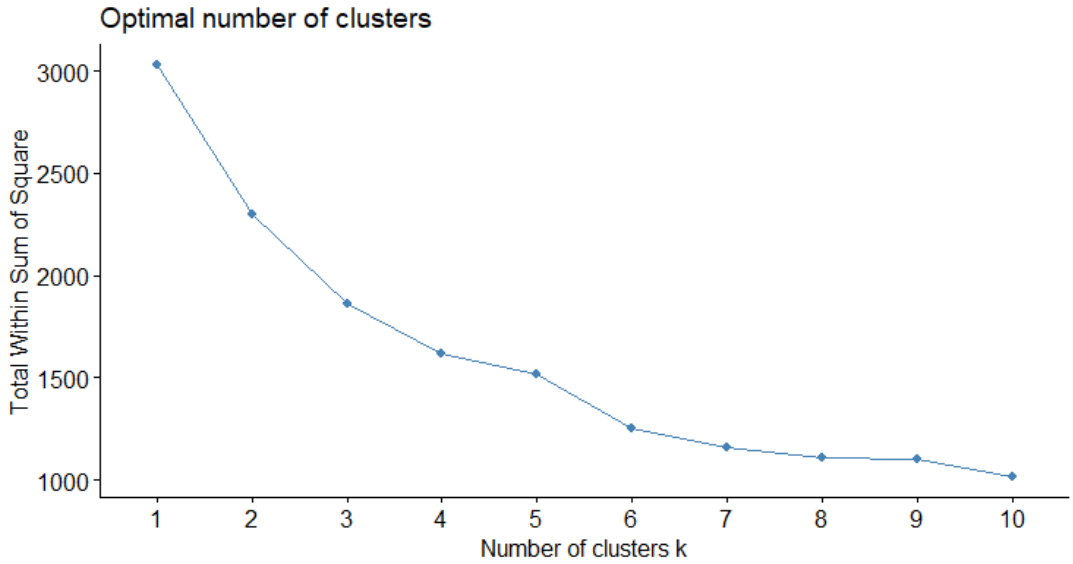

Figure 3: Gráfico do método do Cotovelo

Pela observação do gráfico, concluímos que seria mais adequada a utilização de cerca de cinco clusters.

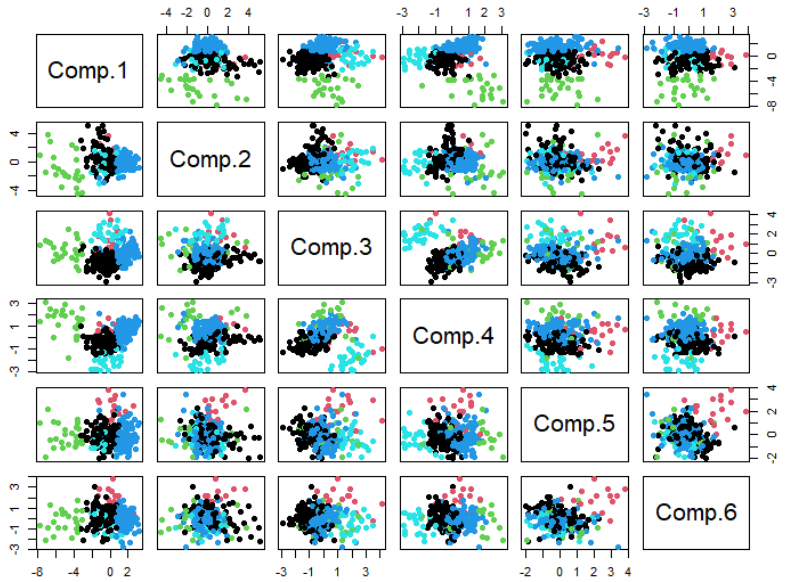

Figure 4: Seis primeiras componentes principais agrupadas em 5 clusters

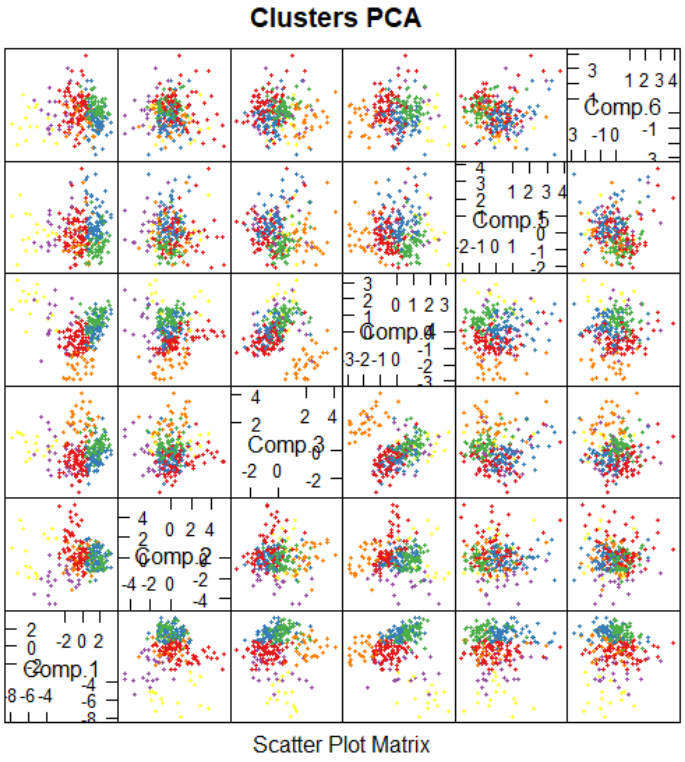

Figure 5: Seis primeiras componentes principais agrupadas em 5 clusters (SPLOM)

Como é possível observar, grande parte dos dados agrupados pelos clusters estão bastante sobrepostos. Isto, aliado à variância relativamente baixa (60%) explicada pelas seis primeiras componentes principais, fará com que qualquer conclusão retirada seja apenas parcialmente confiável, qualquer tipo de analise sobre os dados necessitará de complementação adicional.

Isto poderá ser consequência dos dados não apresentarem uma divisão natural, o que seria expectável, considerando que o tipo de dados em estudo são relativos à saúde humana.

É percetível que, na figura 4, o grupo a verde (amarelo na figura 5) é consistentemente o mais espalhado, o que é sinal de apresenta uma grande variabilidade nos seus dados, Em contraste, o grupo a azul escuro (verde na figura 5) está de maneira constante bastante compacto, o que é indicador de que é um grupo bastante homogéneo.

4 Clustering Hierárquico

Para realizar clustering hierárquico, às seis primeiras componentes principais, foi necessário colocar os dados em questão à mesma escala, para que valores mais altos não tivessem uma maior influência na distância.

Utilizamos o método da distância de Ward, de forma a minimizar a variância dentro dos clusters.

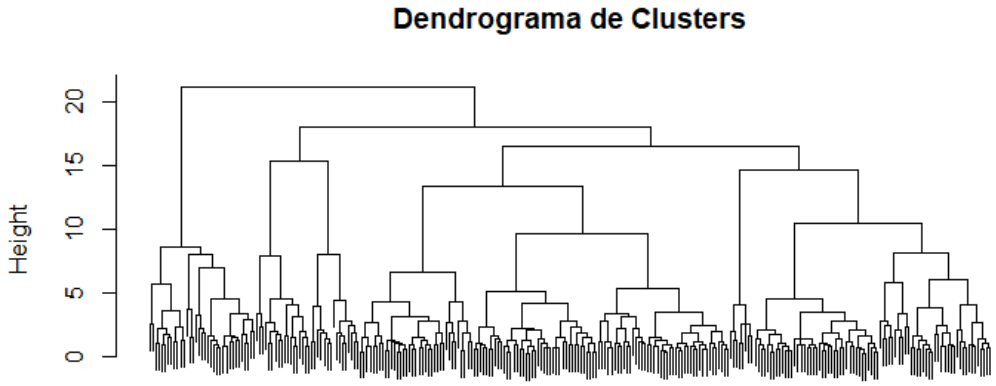

Figure 6: Dendograma de clustering hierárquico

Há um bom ponto de corte natural por volta dos cinco clusters, o que coincide com o que foi concluído na observação do gráfico da figura 3. Este corte natural está ilustrado na figura 7.

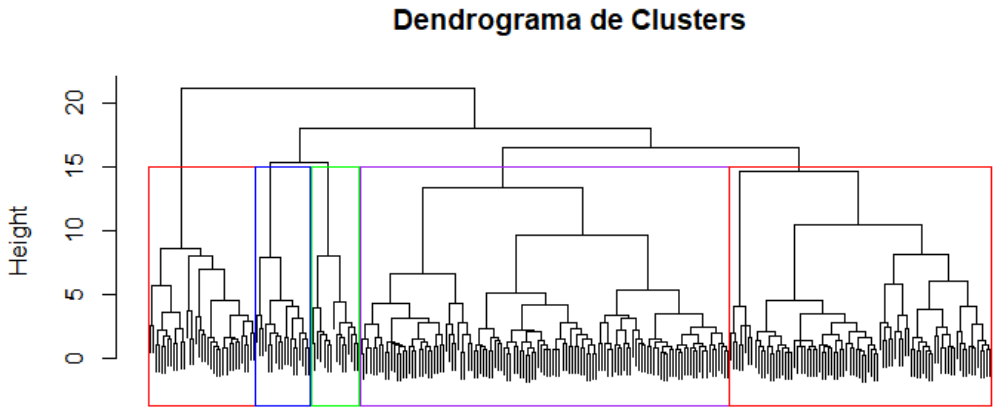

Figure 7: Dendograma de clustering hierárquico - Agrupado

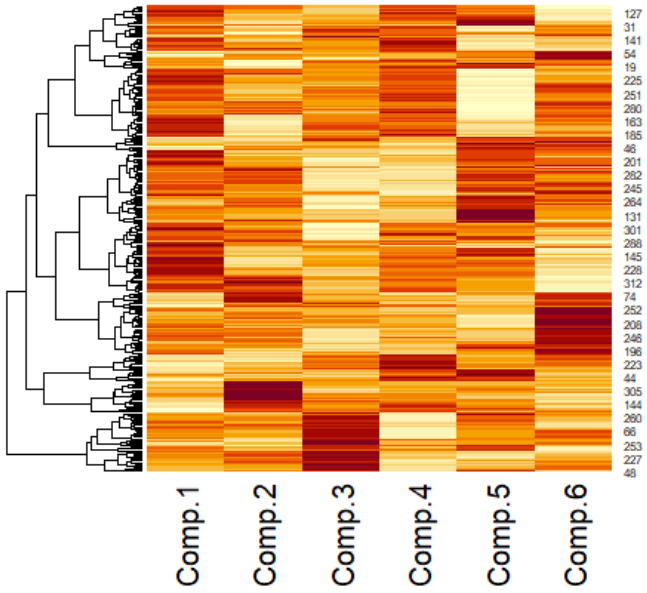

Figure 8: Heatmap de clustering hierárquico

Como podemos observar, existe uma grande variação de cores dentro do mesmo grupo de cada componente, supostamente onde a cor se encontra mais escura significa que a componente reconhece um forte padrão entre os dados, contudo, como foi dito acima, estas componentes não explicam 70 % dos dados, por isso não podemos tirar grandes informações deste heatmap, o que vai de acordo com as conclusões acima referidas.

5 Clustering por Misturas

Aqui aplicamos o mclust aos dados, e o método decidiu que o numero apropriado de clusters a serem usados seriam 9, bem como que o melhor modelo a aplicar seria o VEE. O que podemos também observar no gráfico abaixo, uma vez que por volta das 5 componentes que é quando o modelos começam a estabilizar o maior valor de BIC (mais próximo de 0) é atingido pelo modelo VEE.

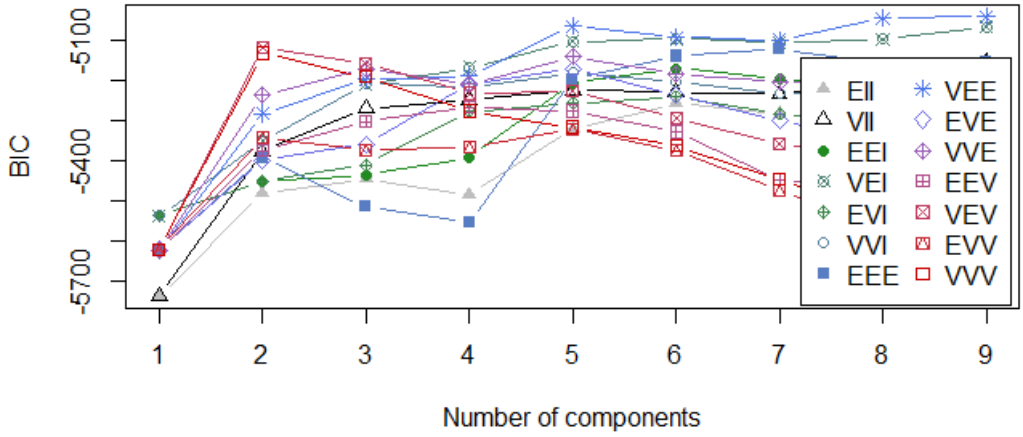

Figure 9: Gráfico BIC

Para além do gráfico acima, o método forneceu o gráfico de classificação dos dados, onde podemos ver os grupos de clusters, como temos vindo a concluir, podemos observar que são visíveis as cores porém existe sobreposição de alguns grupos o que poderá indicar que existe alguma ambiguidade em certos grupos.

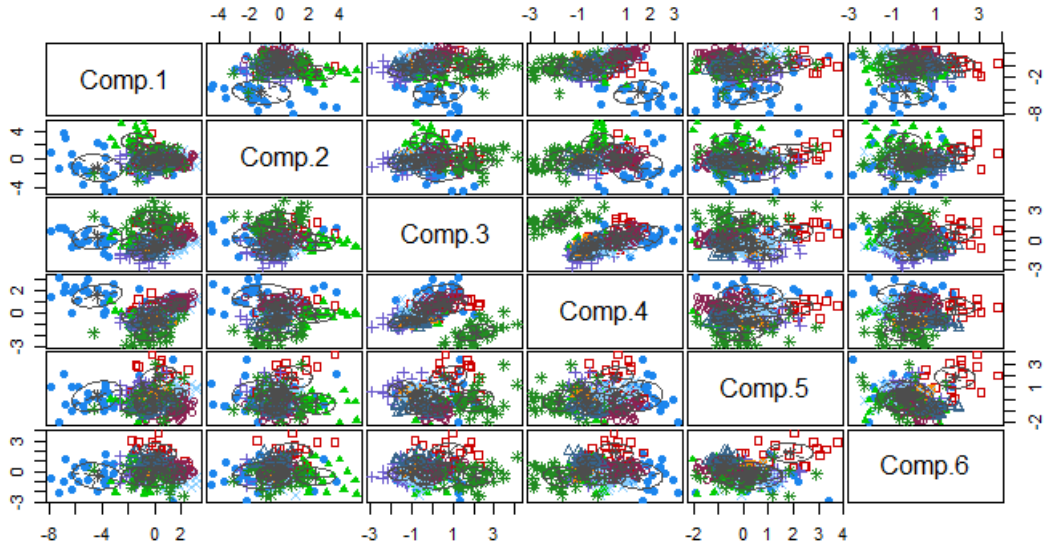

Figure 10: Classificação dos dados

Pelo gráfico seguinte podemos observar as elipses que representam a área de confiança de cada cluster, tendo em conta isto podemos ver que existe uma grande incerteza nos dados uma vez que grande parte se encontra fora da elipse representante do cluster. Para além disso estas elipses estão sobrepostas em maior parte dos casos, o que apoia mais uma vez as conclusões que temos vindo a tirar até agora de que o modelo tem dificuldade a distinguir certos tipos de conjuntos de dados.

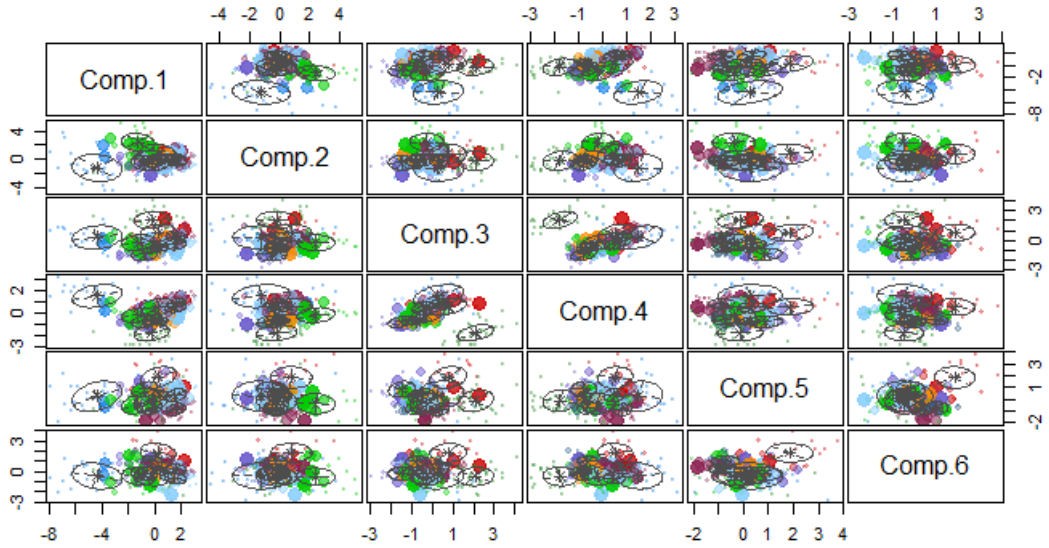

Figure 11: Incerteza

Por último obtivemos o gráfico das densidades e mais uma vez, podemos observar que as curvas estão sobrepostas o que indica sobreposição entre clusters.

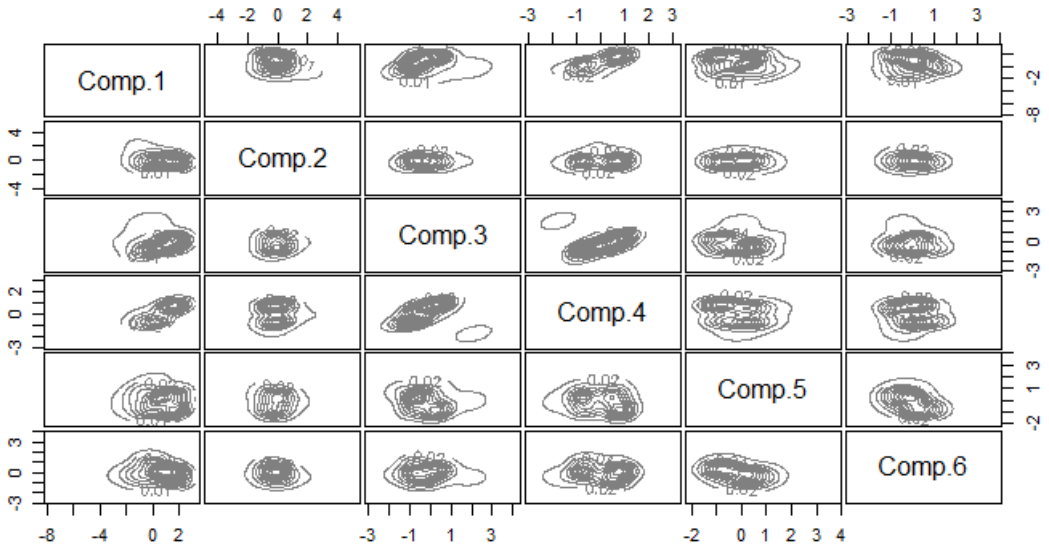

Figure 12: Densidade dos Dados