UFSCar - UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA

Cluster com Modelos de Mistura

Beatriz Luri Yasuda Ikeda
Douglas de Paula Nestlehner
Hélio Mota Ezequiel
Raquel Malheiro de Carvalho

São Carlos - SP

Conteúdo

1 Introdução

Em muitos estudos estatísticos somos confrontados com problemas que pretendem estudar um determinado fenômeno com o objetivo de descrever, explicar e prever o seu comportamento. Porém, na resolução destes problemas topamos com diversas incertezas e isso trás como consequência a impossibilidade de conhecer o fenômeno de forma completa.

Nestas circunstâncias, inicializa-se por realizar uma organização das observações em grupos de acordo com a similaridade das propriedades observadas, sendo útil para identificar padrões que não foram observados anteriormente, exceções e anomalias sobre os dados, sugerir novas hipóteses sobre os grupos de dados. De forma geral, essa organização é fundamental para entender e aprender sobre os dados. Cada grupo (cluster) formado pode ser representado por um ou mais protótipos que são os elementos que irão melhor representar as características do cluster. De seguida, utiliza-se um modelo que constitui uma melhor representação deste fenômeno e que pretende dar resposta aos objetivos inicialmente fixados.

Os métodos de agrupamentos são utilizados em diferentes áreas, desde análise de dados médicos até dados financeiros. Isso se deve a sua grande flexibilidade e disponibilidade atualmente, além de que os agrupamentos diferem de modelos preditivos supervisionados, pois não precisam de uma variável resposta, tornando-o mais abrangente.

Modelos de Mistura de Gaussianas pode ser usado como método de agrupamento baseados em distribuições estatísticas específicas. Qualquer distribuição pode ser modelada por um modelo de mistura. Ao modelar a função densidade de probabilidade de um conjunto de dados, o modelo automaticamente faz agrupamentos do conjunto ao discriminar qual componente da mistura que gerou cada elemento. O modelo de mistura se baseia pela suposição de que cada elemento do conjunto se origina a partir de um componente da mistura com uma determinada probabilidade. Portanto, quando inferimos os parâmetros da mistura, esta probabilidade pode ser utilizada para associar cada elemento ao componente com maior probabilidade de o ter gerado.

Os parâmetros que compõe cada componente de uma mistura podem ser caracterizados como o protótipo do grupo formado pelos dados com maior probabilidade de pertencer à aquele grupo. Quando se referimos à Modelo de Mistura de Gaussianas, o protótipo de cada grupo consiste no conjunto de parâmetros que contém no vetor de médias, matriz de covariâncias e os pesos da mistura. Já as estimativas dos parâmetros da distribuição de cada componente da mistura são obtidas a partir dos cálculos ao maximizar a verossimilhança

utilizando o algoritmo "expectation maximization"(EM), que é um método interativo de encontrar a estimativa de Máxima Verossimilhança dos parâmetros de uma distribuição dado um conjunto de dados.

Este trabalho está organizado em cinco capítulos. Neste capítulo, é realizado toda a contextualização do tema proposto e da motivação do trabalho.

No capítulo 2, são apresentados os principais conceitos necessários para a compreensão e avaliação do método que utilizaremos neste estudo. Sâo descritos os conceitos de Cluster, Modelos de Mistura, Modelos de Mistura Gaussiana e Algoritmo EM.

No capítulo 3, está a aplicação dos conceitos apresentados em um banco de dados.

No capítulo 4, abordamos a conclusão do trabalho.

No capítulo 5, apresentamos os códigos que foram escritos para aplicação dos conceitos.

E por fim, no índice apresentamos toda a bibliografia utilizada para a escrita deste trabalho.

2 Materiais e Métodos

2.1 Cluster

Diferente dos problemas de regressão e classificação, *cluster* está associado á ideia de aprendizado não-supervisionado, ou seja, não já uma saída específica para os padrões de entrada e, busca-se a inferência de uma função que descreve a estrutura dos dados.

A clusterização consistem em um agrupamento de observações de um conjunto de dados de acordo com as suas similaridades. Isto é, separar em grupos as observações de um conjunto de dados de tal forma que dentro do grupo as observações sejam mais homogêneas possíveis e que os grupos sejam o mais heterogêneos possíveis. O agrupamento é feito através da análise de agrupamento e os grupos encontrados são chamados de *clusters*. Portanto, podemos definir clusters como, sendo $X = \{x_1, x_2, ..., x_N\}$ os nosso dados e x_i representando os atributos da i-ésima observação. Alguns autores definem *m-clustering* de X como partição de X em *m* conjuntos $(C_1, C_2, ..., C_m)$, respeitando as seguintes condições:

- $C_i \neq \emptyset, i = 1, ..., m$;
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j; i, j = 1, ...m$

Isto é, espera-se que as observações contidas em um grupo C_i sejam similares entre si e não tão similares às observações em outros grupos.

Existem dois métodos principais métodos para a realização da análise de agrupamento são o *hierárquico* e o *não-hierárquico* (Frei, 2006).

O método hierárquico consiste em encontrar a matriz de similaridade das observações e agrupar 2 observações (ou agrupamentos) que sejam mais similares em cada etapa até possuir somente um único grupo e escolher o número de grupos que deseja a partir do dendograma. Também é possível agrupar usando medidas de dissimilaridade, isto é, agrupar observações (ou grupos) que são menos diferentes.

Já o método não-hierárquico é um método que exige um esforço computacional para ser realizado e existem alguns algoritmos como o *K-Means*, *Fuzzy C-Means* e muitos outros. Esse método simula diversos agrupamentos possíveis dentro do conjunto de observação e

escolhe o agrupamento com o melhor desempenho. O foco deste trabalho é estudar um método de agrupamento não-hierárquico de modelo de mistura.

2.2 Modelos de Mistura

Geralmente trabalhamos com modelos probabilísticos que fazem o uso de formas de distribuições mais simples (bernoulli, exponencial, poisson, etc.), entretanto podemos nos deparar com situações em que os dados que desejamos modelar são mais complexos, ou seja, não conseguimos caracterizar a distribuição de uma variável de forma clara, como por exemplo na Figura 1, em que é possível notar uma distribuição multimodal.

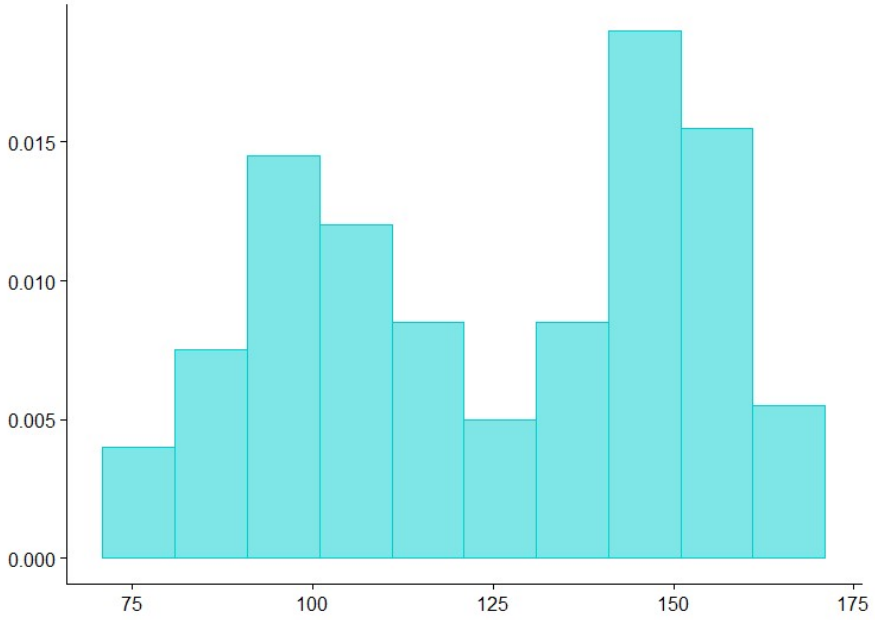

Figura 1: Exemplo de uma distribuição multimodal

Nessas situações, podemos modelar os dados em termos de uma mistura de vários componentes, em que cada componente possui uma forma paramétrica mais simples, como por exemplo a gaussiana, ou seja, declaramos que cada ponto de dados pertence a um dos componentes e a partir disso tentamos inferir a distribuição para cada componente separadamente, denotamos esses tipos de modelos como modelos de mistura.

Os modelos de mistura são úteis para descrever a heterogeneidade em dados que não podem ser capturados adequadamente com uma única distribuição. Isso é feito agrupando conjuntos de observações que maximizam tanto a homogeneidade da classe latente quanto a heterogeneidade da classe latente, onde as classes latentes são padrões de respostas que

são usados para identificar subpopulações dentro de um conjunto de dados (McLachlan and Peel, 2000). Em outras palavras, os modelos de mistura tem como objetivo identificar grupos de respostas semelhantes entre si para colocar em classes separadas e identificar classes diferentes umas das outras.

No caso da Figura 1 podemos notar com clareza a presença de duas classes latentes (padrões de respostas), desse modo, a mistura de duas distribuições será o suficiente caracterizar os dados. Entretanto, na maioria das situações não conseguimos notar de inicio o numero de classes latentes (grupos), caberá ao modelo identificar as possíveis classes latentes e como interpretação desse resultado podemos atribuir significados para cada classe (nem sempre isso sera possível, dependendo da complexidade do modelo).

Em geral, os modelos de mistura assumem que os dados são gerados pelo seguinte procedimento:

- Inicia-se amostrando as variáveis latentes, denotamos essas variáveis como L;
- Em seguida amostramos as respostas observadas (Y) de uma distribuição que depende de L, ou seja:

$$p(L, Y) = p(L)p(Y|L)$$

Em que, $p(L)$ é uma distribuição multimodal, e $p(Y|L)$ assume qualquer distribuição mais simples.

Desse modo, os modelos de mistura produzem uma variável categórica latente que agrupa padrões de respostas em um número discreto de grupos não observados (McLachlan and Peel, 2000).

2.2.1 Funcionamento da mistura de distribuições

No caso da Figura 1 podemos supor como sendo uma mistura de duas distribuições gaussianas. Como exemplo, componente 1 assume distribuição N(100, 20) e componente 2 assume distribuição N(150, 10), e que a probabilidade do componente 1 é 0.8, desse modo, o processo gerador da mistura das distribuição é dada por:

- Escolhemos o componente com base na probabilidade;
- Se escolhemos o componente 1, então retiramos uma amostra y da distribuição gaussiana com média 100 e variância 20;

- Caso seja o componente 2, então retiramos uma amostra y da distribuição gaussiana com média 150 e variância 10;

2.3 Modelos de Mistura Gaussiana

Sabendo as definições/informações do que são cluster e modelos de mistura, iremos abordar a clusterização utilizando modelos de mistura gaussiana.

Supondo que em um banco de dados, nosso interesse seja agrupar os pontos de dados em várias partes com base na sua similaridade, no aprendizado de máquina, isso é denominado como *clustering*. Para isso, temos alguns métodos disponíveis:

- K-means;
- Agrupamento hierárquico;
- Modelos de mistura;
- etc.

Como dito anteriormente, em modelos de mistura podemos combinar distribuições mais simples no intuito de estimar a distribuição dos dados. Nesse estudo iremos abordar com mais detalhes os modelos de misturas como mistura de distribuições gaussianas.

Iremos utilizar o modelo de mistura gaussiana para exemplificar como funciona e uma aplicação. Os modelos de misturam lidam com a limitação de não construir um modelo estimador de densidade limitado apenas a função de densidade de probabilidade, sendo que estes combinam um conjunto de distribuições e se baseiam na **Estimativa de Máxima Verossimilhança(EMV)**.

O Modelo de Mistura Gaussianas (MMG) é um modelo o qual a combinação linear de todos os $p_j(x)$ forma uma combinação finita de distribuições gaussiana da seguinte forma:

$$p(\overrightarrow{x}\,|\,\overrightarrow{\theta}) = \sum_{k=1}^{K} \pi_k p(\overrightarrow{x}\,|\overrightarrow{\mu_k}, \Sigma_k)$$

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

Onde θ é a coleção de todos os parâmetros do modelo, ou seja, pesos, médias e matrizes de covariância de cada componetente. Sendo expresso por:

$$\theta := \{\pi_1, ...\pi_K, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_k\}$$

Como temos que estimar estes parâmetros e não temos os rótulos de cada classe, um método para auxiliar nisso é o algoritmo EM como veremos a seguir.

2.3.1 Algoritmo EM

O algoritmo EM, busca encontrar o EMV da verossimilhança marginal iterativamente aplicando os dois passos a seguir:

Expectation (passo E): Este passo visa o cálculo do valor esperado da log-verossimilhança com respeito a distribuição condicional de Z dado X, utilizando a estimativa atual dos parâmetros $\overrightarrow{\theta}^{(t)}$. Fazemos isso por meio da seguinte expressão:

$$Q(\overrightarrow{\theta} \mid \overrightarrow{\theta}^{(t)}) = E_{z|X,\overrightarrow{\theta}^{(t)}}[logL(\overrightarrow{\theta}; X, Z)]$$

Maximization (passo M): Este passo serve para encontrar os parâmetros que maximizam essa quantidade, dada pela expressão:

$$\overrightarrow{\theta}^{(t+1)} = \arg\max_{\overrightarrow{\theta}} Q(\overrightarrow{\theta} \mid \overrightarrow{\theta}^{(t)})$$

Geralmente, os modelos nos quais o algoritmo EM é aplicado usam Z como variável latente, ou seja, variáveis não diretamente observadas, , indicando a pertinência das amostras X em um conjunto de clusters. Observações importantes:

- As observações X podem ser tanto discretas quanto contínuas (Gaussianas por exemplo);
- As variáveis latentes Z são discretas, podendo assumir um número fixo de valores, de modo que podemos associar uma variável latente por observação;
- Os parâmetros são contínuos e podem ser de dois tipos: parâmetros associados a todas as amostras, e parâmetros associados com um valor específico de variável latente;

3 Aplicação

Nessa seção, iremos apresar um exemplo de uso do modelo de mistura gaussiana, aplicado na base de dados Iris. Sabemos a priori que na base exitem três tipos de espécies de flores (“Setosa”, “Versicolor” e “Virginica”), as quais são medidas por quatro covariáveis quantitativas (“Sepal.Length”, “Sepal.Width”, “Petal.Length” e “Petal.Width”).

Para esse exemplo, iremos supor que não temos a informação de qual flor cada observação corresponde, ou seja, supor que não conhecemos a variável “Species”. Desse modo, o nosso objetivo é identificar cada grupo de flor, com base em suas covariáveis.

Para atingir o objetivo proposto, poderíamos utilizar de diversas técnicas de agrupamento, entretanto o modelo de mistura gaussiana é uma alternativa que apresenta bons resultados, e sua aplicação é relativamente fácil, pois temos o algoritmo EM já implementado em diversos pacotes do R.

A seguir temos representado como os dados foram definidos e separados para a aplicação.

```
# bibliotecas
library(mclust)

# dados
dados = iris
head(dados)

# separando a variavel categorica dos dados

# variavel categorica
cat = dados[,5]

# variaveis continuas
dados = dados[,-5]
```

Em seguida, utilizamos o pacote “mclust”, o qual proporciona a função “Mclust()”, responsável pelo ajuste do modelo gaussiano utilizando o algoritmo EM. Na chamada da função Mclust(), apenas a matriz de dados é fornecida, e o número de componentes de

mistura e a parametrização de covariância são selecionados usando o Critério de Informação Bayesiano (BIC).

```
M1 = Mclust(dados)
summary(M1) # Numero de componentes de mistura indicado
plot(M1, what = "BIC") # Gráfico BIC
```

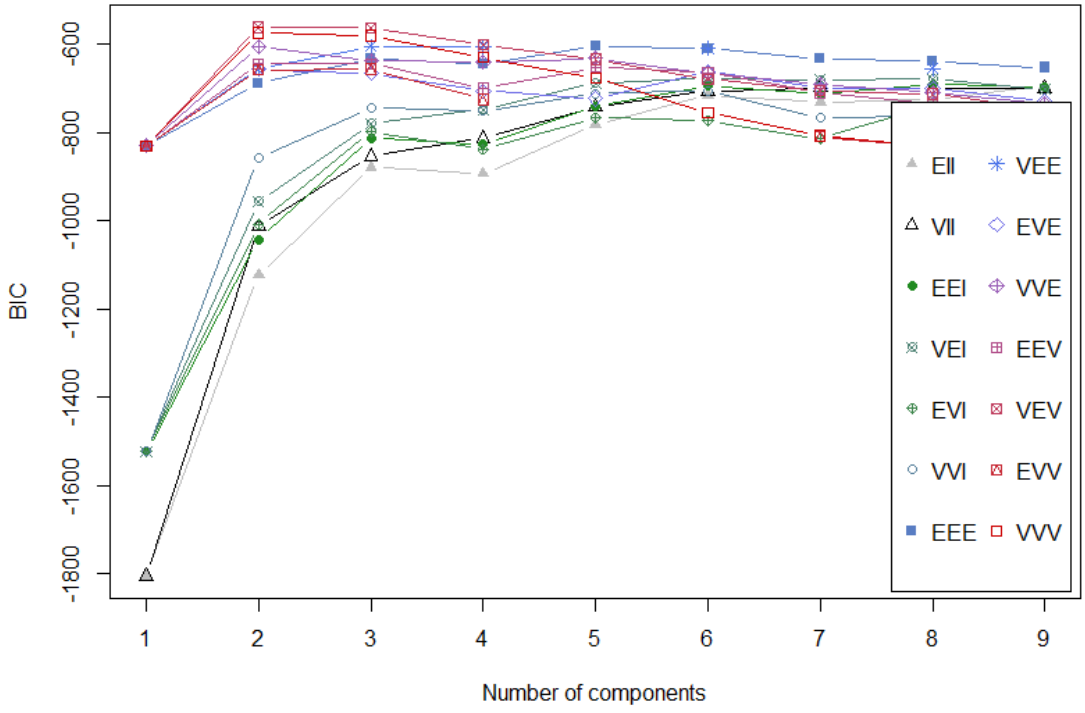

Figura 2: BIC estimado para diferentes número de componentes e parametrizações de covariância

Observando as estimativas obtidas no modelo, notamos que os maiores BIC são para os modelos com numero de componentes igual a 2, entretanto, os valores são bem próximos aos modelos com números de componentes igual a 3.

Realizamos então, um novo ajuste considerando 3 componentes, por meio da mesma função “Mclust() só que agora definindo o número de componentes.

```
M2 = Mclust(dados, 3) # DEFININDO O NUMERO DE COMPONENTES
summary(M2, parameters = T)
plot(M2, what = "classification")
```

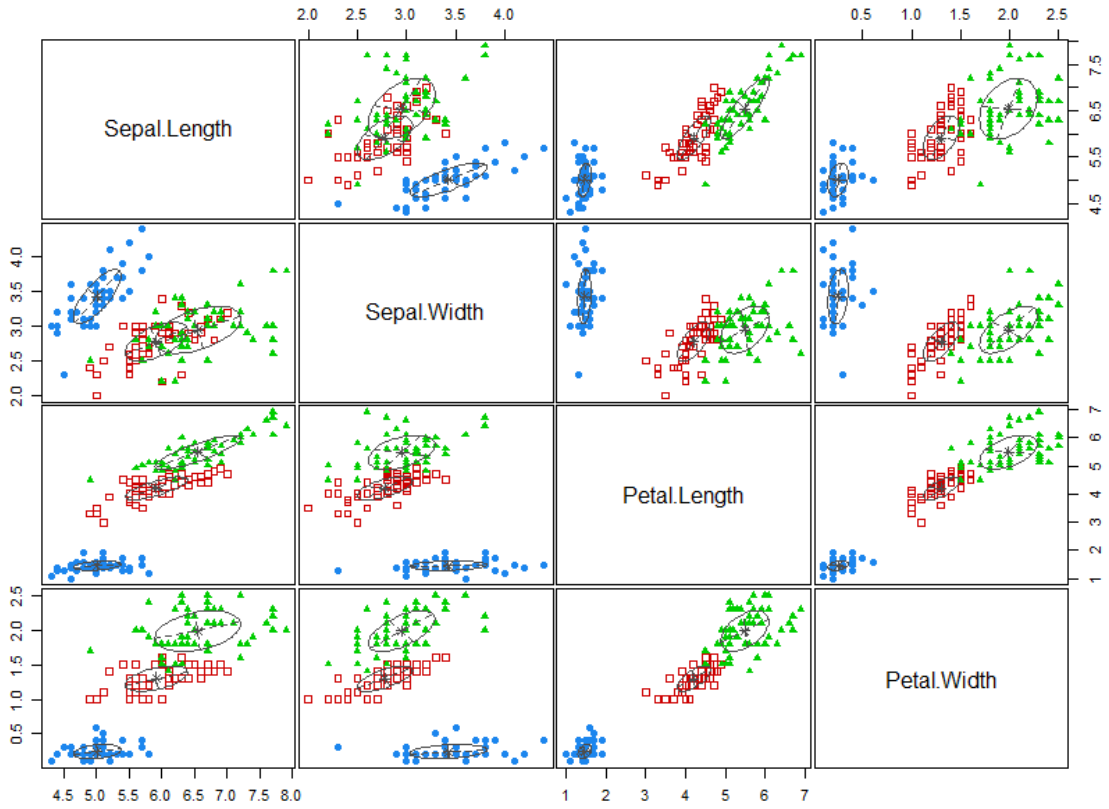

Figura 3: Agrupamentos obtidos pelo modelo de mistura gaussiana.

Podemos notar que o agrupamento feito pelo modelo apresenta bons resultados, na Tabela 4 temos a representados o numero de observações de cada grupo (real e estimado).

Classificações	1	2	3
setosa	50	0	0
versicolor	0	45	5
virginica	0	0	50

Vale ressaltar que o modelo não atribui significados para os agrupamentos, ou seja, no caso da Tabela 4 das 50 observações “setosas” o modelo conseguiu identificar um grupo de 50 observações (que podem caracterizar as “setosas”); modelo encontrou outro grupo contendo 45 observações (que podem caracterizar como “versicolor”) e outro grupo contendo 55 observações.

A função mixmodCluster() nos permite observar quais distribuições foram atribuídas em cada componente, também a mistura de distribuições estimada.

```
# distriuições misturadas
mixmodBIC = mixmodCluster(dados, 3)
summary(mixmodBIC)
hist(mixmodBIC)
```

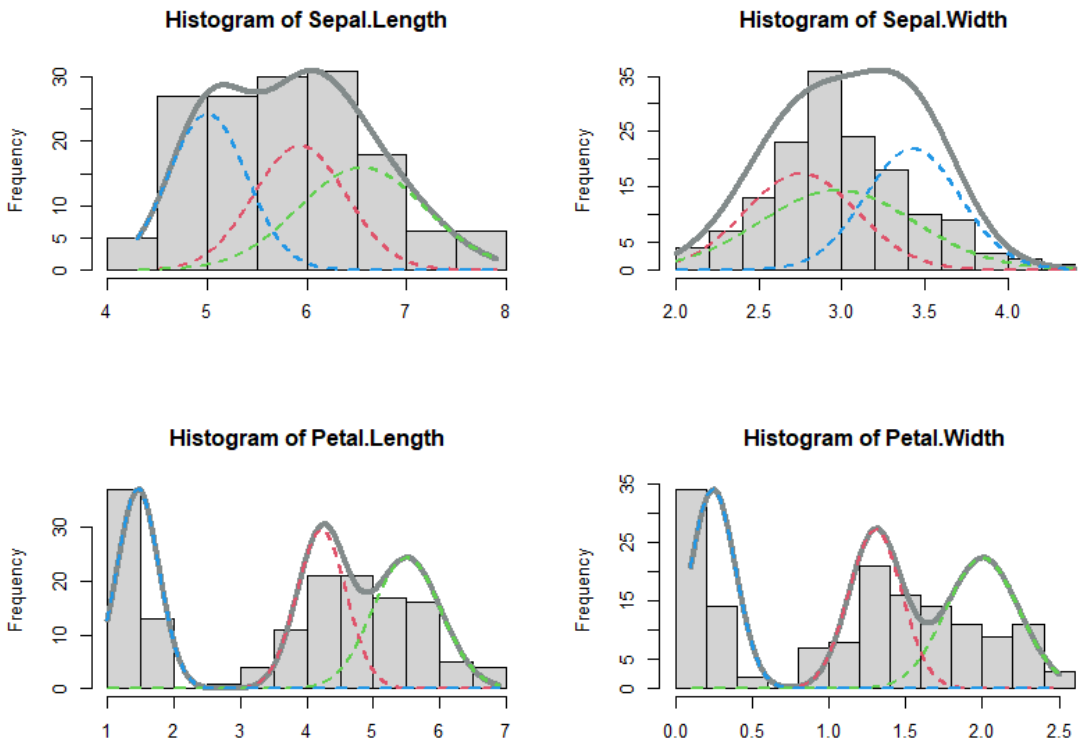

Figura 4: Histograma das misturas de distribuições.

Nota-se que algumas das covariáveis apresentam distribuição “claras”, portanto para esses dados, a aplicação do modelo de mistura não seja tão indicada.

De modo geral, o modelo ajustado conseguiu realizar os respectivos agrupamentos, e apresentou bons resultados

4 Conclusão

Podemos analisar que agrupamento é útil para identificação de padrões, sendo o Modelo de Mistura Gaussiana (GMM), um método de *clustering*, também definido como um modelo probabilístico que assume que todos os pontos de dados são gerados a partir de uma mistura de um número finito de distribuições Gaussianas com parâmetros desconhecidos.

O MMG supera a limitação de agrupamento que outros modelos e algoritmos possuem, como o *K-means* que vimos em aulas da disciplina, sendo ele uma generalização de algoritmos de mistura onde, ao invés de tentar estimar apenas os “centroides"de cada agrupamento, ele tenta estimar além da forma e proporção de cada Gaussiana da mistura observada.

Este modelo de mistura pode ser descrito como uma combinação de modelos os quais se combina a distribuição gaussiana com pesos, tendo assim o modelo de mistura gaussiana.

Vimos também a necessidade do Algoritmo EM (*Expectation-Maximization*) na otimização de alguns parâmetros enquanto outros são fixados para a identificação dos agrupamentos, e fornece , além da informação do centro dos *clusters*, a dispersão dos dados em torno dos centros.

Bibliografia

Fernando Frei. *Introdução à análise de agrupamentos*. Unesp, 2006.

Alexandre Henrique. O raciocinio por tras do algoritmo expectation-maximization, 2021. URL `https://medium.com/b2w-engineering/o-racioc%C3%ADonio-por-tr%C3%A1s-do-algoritmo-expectation-maximization-91d4a8588778`.

G. McLachlan and D. Peel. *Finite Mixture Models*. *Wiley Series in Probability and Statistics, John Wiley Sons, Inc., 2000*.

Surya Pryi. Clustering in machine learning, 2022. URL `https://www.geeksforgeeks.org/clustering-in-machine-learning/`.

R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.*

Carlos Eduardo Martins Relvas. Agrupamento baseado em modelos de mistura de gaussianas com covariáveis. *PhD thesis, Universidade de São Paulo, 2020.*

Julio M Singer, Juvêncio S Nobre, and Francisco MM Rocha. Diagnostic and treatment for linear mixed models. In Proceedings 59th ISI World Statistics Congress Session CPS203*, pages 5486–5491, 2013.*

Daiane Aparecida Zuanetti and Luis Aparecido Milan. Data-driven reversible jump for QTL mapping. Genetics*, 202(1):25–36, 2016.*