Métodos em Aprendizado Não Supervisionado de Máquina - Trabalho II

Daniel Lopes Toso

Maio 2024

1 Introdução

Este trabalho tem como objetivo utilizar técnicas de aprendizado de máquina não-supervisionado para classificar 54 linguagens de vários lugares do mundo.

Os dados foram retirados dos dicionários das línguas utilizadas, que, por sua vez, foram obtidos através do webscraping de uma coleção de dicionários disponibilizada por Eymen Efe Altun. O webscraping foi utilizado para obter o conjunto de palavras de cada língua, bem como filtrar os idiomas compostos majoritariamente pelo alfabeto latino (a-z), desconsiderando acentos.

Os códigos para o webscraping em Python e para os gráficos, clusters, manipulação de dados e análises feitas com o R estão disponíveis no Github.

2 Materiais e Métodos

2.1 Métrica

Para realizar a clusterização, as palavras foram extraídas dos dicionários e, para cada linguagem, foram contadas as quantidades de pares de duas letras seguidas, por exemplo "AA", "AB", "AC", etc.

Esse processo foi feito internamente para cada par de letras contido em uma palavra do dicionário com mais de uma letra e os resultados finais foram somados. Abaixo, um exemplo da contagem de pares para a palavra "Banana".

BA	AN	NA
1	2	2

A título de exemplo, para a contagem geral, os primeiros pares de letras para o Holandês são:

AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	...
15777	1818	4490	3941	415	2939	4479	300	1953	184	4804	9801	5428	18155	192	...

Portanto, para cada dicionário, foram obtidas 676 (26×26) covariáveis de contagem.

2.2 Misturas Gaussianas

Devido a natureza de contagem dos dados, assim como a irrestrição no limite de contagem das sílabas pois dicionários estão constantemente em expansão/transformação, pode-se assumir que as células seguem uma distribuição Poisson.

Pela quantidade de palavras encontradas nos dicionários, os parâmetros das Poisson's são de ordem muito grande, o que significa que as distribuições podem ser aproximadas razoavelmente por distribuições gaussianas, após a centralização.

Assim, após a contagem dos pares de letras, as células de cada linha foram divididas pelo número total de palavras disponíveis para cada linguagem, objetivando padronizar todas para a mesma escala.

Assim, devido à natureza de normalidade dos dados após o processamento, o principal modelo a ser avaliado para a clusterização neste trabalho será o de mistura gaussianas (GMM).

2.3 Análise de Componentes Principais

Após o processamento inicial dos dados, temos um *dataframe* com 54 linhas e 676 colunas. Isso é prejudicial não só ao método de GMM mas à clusterização em altas dimensões no geral, devido aos problemas relacionados a "*curse of dimensionality*". Uma maneira apropriada de reduzir as dimensões, partindo do pressuposto que os dados seguem uma distribuição aproximadamente normal multivariada, é a análise de componentes principais.

Devido ao número de observações, só existem 54 colunas linearmente independentes, então é possível reduzir as dimensões dos dados para uma tabela 54×54 sem perda de informação com as componentes principais. Entretanto, é possível realizar a análise com um número ainda menor de componentes e isso foi estudado neste trabalho.

2.4 Clusterização Hierárquica

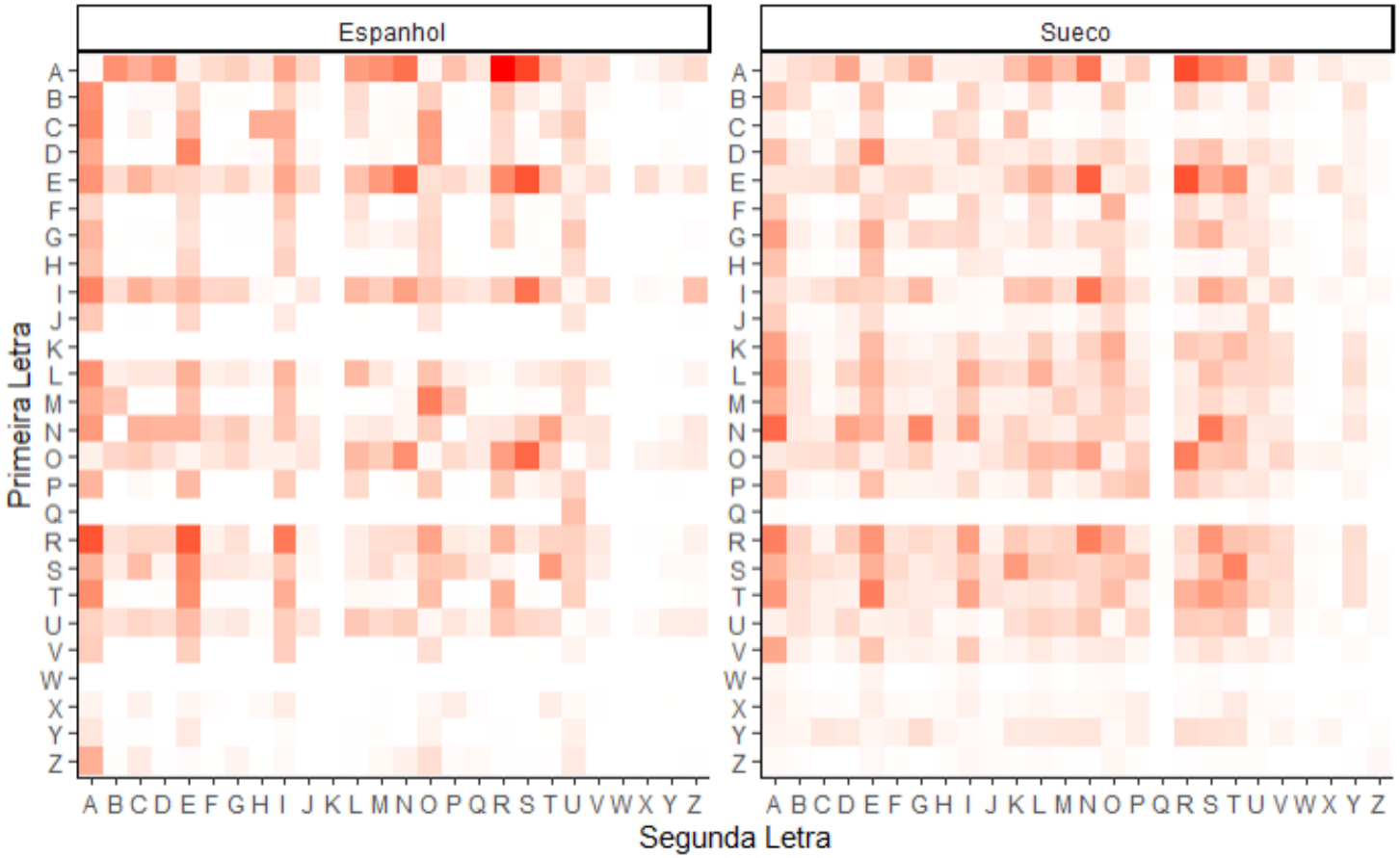
A clusterização hierárquica é uma técnica particularmente interessante para esse problema para notar não só a classificação dentro de cada cluster mas também a distância entre os clusters. Seu uso teve como objetivo comparar a clusterização por GMM às técnicas vistas anteriormente na matéria. Ela também foi utilizada para validar o uso da análise de componentes principais, realizando-se a clusterização hierárquica antes da redução de dimensionalidade e também após a reconstrução dos dados de volta a forma 54×676 a partir das componentes principais, para ilustrar que a perda de informação do processo é mínima.

3 Resultados

3.1 Análise dos Dados

A partir dos dicionários, é possível construir uma relação entre todos dois pares de letras para cada língua. A ideia para a clusterização é que essas relações identificam unicamente cada linguagem e são parecidas em linguagens próximas.

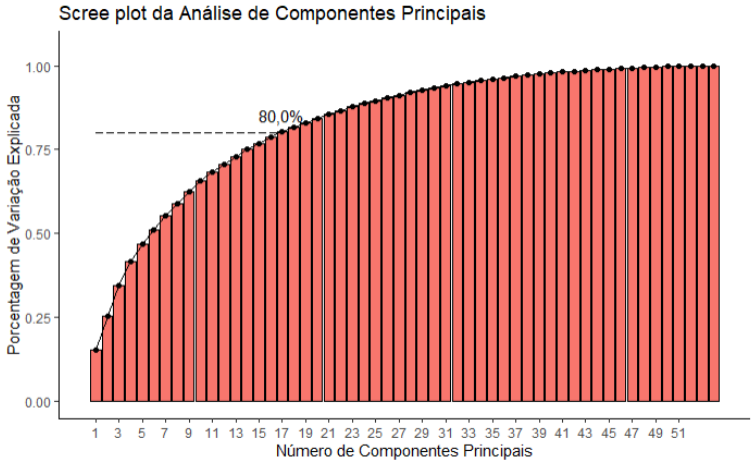
Abaixo, a comparação com um *heatmap* entre as relações encontradas para o Espanhol e o Sueco. Os valores utilizados para a produção do gráfico foram transformados com a raiz quadrada, para facilitar a visualização dos pontos de calor.



É possivel perceber que essa técnica evidencia a maior estruturação do espanhol em torno das vogais (no sentindo de que cada par de letras envolve uma vogal) devido a sua origem latina, evidenciado pelas linhas notáveis em torno dessas letras. Por outro lado, o sueco apresenta distribuição mais homogênea de relações de pares com as consoantes, o que corrobora com a ideia de que é possível perceber características intrínsecas aos idiomas por meio dessa métrica.

3.2 Clusterização Hierárquica e Validação das Componentes Principais

Utilizando-se o Scree Plot abaixo, foram escolhidas as primeiras 17 componentes principais para representar os dados, o que oferece em torno de $80,0\%$ da variação total explicada.



Para justificar o uso de componentes principais para reduzir a dimensionalidade de 676 para 54 sem perda de informação, foram aplicadas técnicas de clusterização hierárquica nos dados antes da aplicação de componentes principais e após sua reconstrução, os resultados foram:

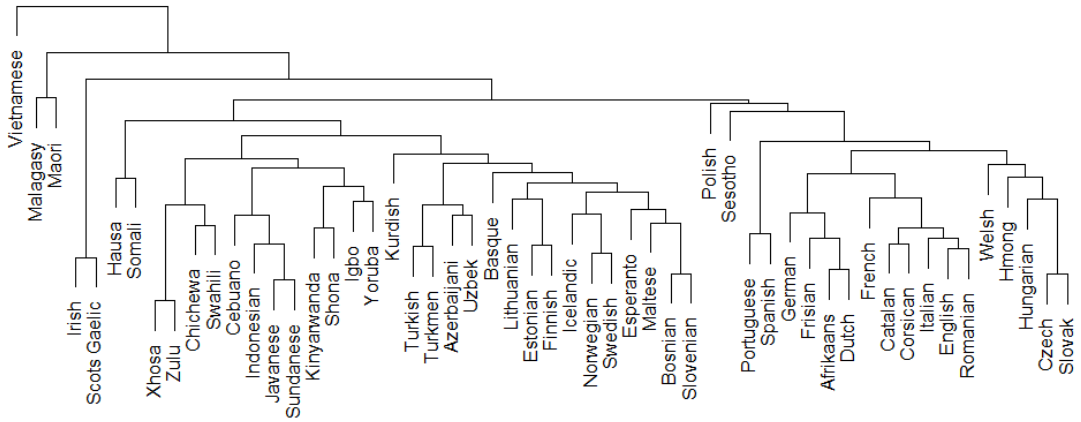


Figura 1: Clusterização com 676 dimensões antes da PCA

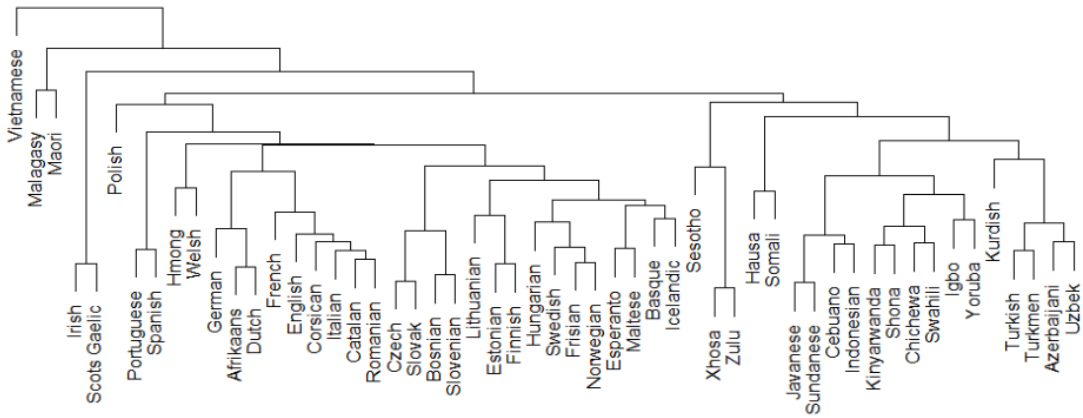


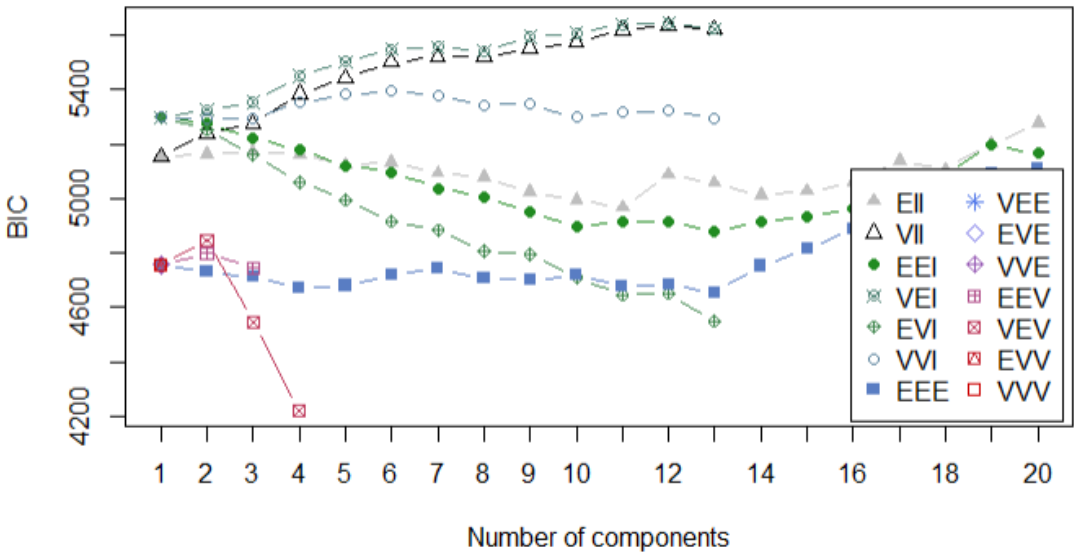
Figura 2: Clusterização com 676 dimensões reconstruído das primeiras 17 PCA's

A reconstrução a partir das 54 PCA's resultou em uma clusterização idêntica à figura 1, confirmando que toda a informação foi mantida. No caso da reconstrução a partir das 26 primeiras PCA's, é possível ver que a ordem foi levemente alterada mas os grupos de linguagem se mantiveram praticamente os mesmos: as línguas românicas como português, espanhol, catalão, córsico se mantiveram unidas, bem como tcheco, eslovaco, bósnio e esloveno. Irlandês e escocês se mantêm na mesma posição, assim como magalasi e maori, entre outros.

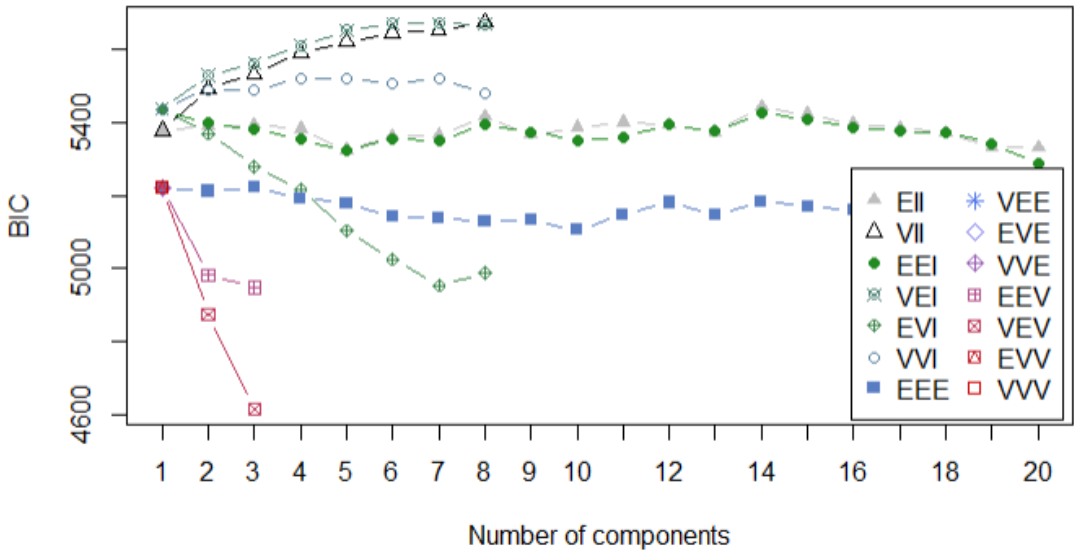
3.3 Misturas Gaussianas

3.3.1 BIC

Utilizando as primeiras 17 componentes principais, foram calculados os BIC's para determinar o número de clusters e tipo do modelo de misturas. Os resultados obtidos pelo ICL foram omitidos por serem idênticos ao BIC.



A partir da análise do BIC, conclui-se que o modelo mais apropriado é de 12 clusters do tipo "VII"ou "VEI".
Também foi feita a análise do BIC após a retirada de observações consideradas como "ruído", os resultados foram:



De onde se percebe que o modelo ideal possui 8 clusters e tipo "VII"ou "VEI".
Com ou sem a retirada de ruído, os modelos do tipo "VII"e "VEI"se mostraram idênticos, portanto, para evitar redundância, as análises que se seguem vão considerar apenas o tipo "VII", por ser teoricamente mais simples (distribuição esférica, de volume variável e formas iguais).

3.3.2 Clusters

Os resultados dos clusters para ambos os métodos foram:

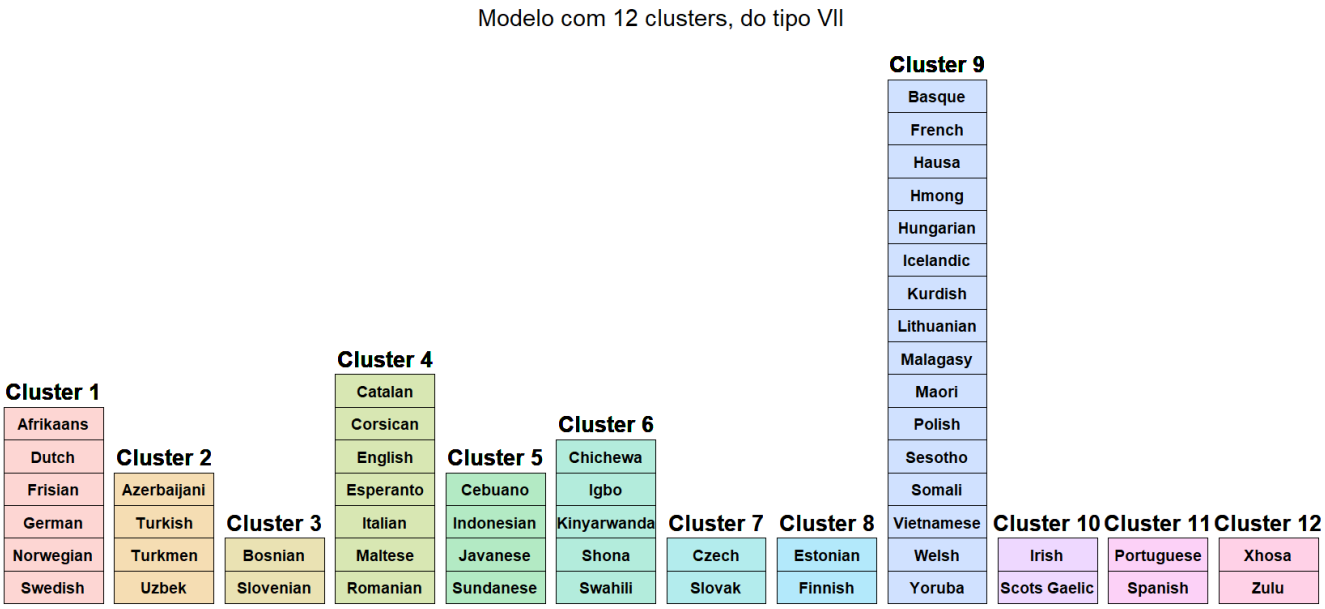


Figura 3: Mistura de normais com 12 clusters, sem remoção de ruído

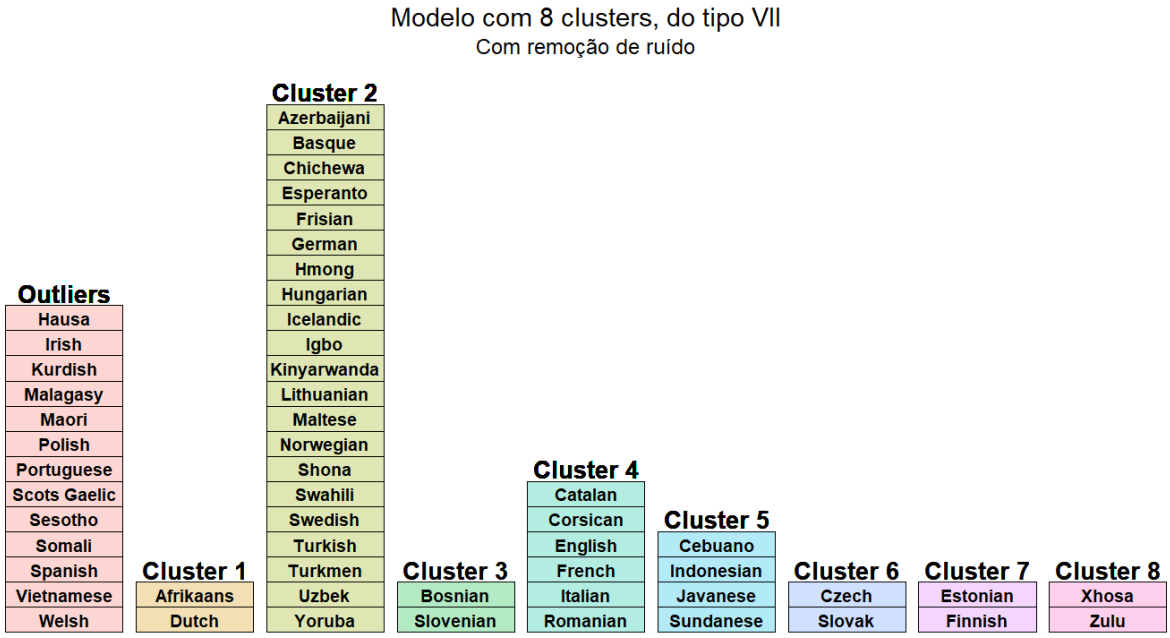


Figura 4: Mistura de normais com 8 clusters, com remoção de ruído

4 Discussão

4.1 A respeito das técnicas

É possível perceber que em ambas as técnicas utilizadas, um dos clusters encontrados parece ter incluído as linguagens que não se adequavam a nenhum outro grupo. No caso da figura 3, isso pode ser visto no cluster 9, enquanto na figura 4 esse comportamento ocorre para o cluster 2.

A remoção de ruído, por se basear na distância em relação a outros pontos, eliminou que menos se adequavam aos outros clusters existentes, como Vietnamita, polonês e Sesotho (figura 1 e 2), porém também removeu grupos inteiros já estabelecidos, como os clusters 11 e 10 da figura 3, sem resolver o problema das observações centrais que não pareciam pertencer claramente a nenhum cluster.

A existência de um agrupamento com alta concentração se deve principalmente ao pequeno volume de dados conjuntamente com um alto número de variáveis. Como foi visto em aula, o número de parâmetros a serem estimados cresce rapidamente conforme a dimensão dos dados, problema que foi mitigado com a análise de componentes principais mas não completamente solucionado, devido à baixa correlação entre as covariáveis à priori. Assim, os modelos que parecem melhor se adequar ao problema, como "VII"e "VEI"não convergem para um número maior de clusters.

Como a remoção de ruído removeu muitas observações, incluindo alguns dos clusters encontrados com outras técnicas, sem resolver o problema de superlotação de um único grupo, a modelo a ser considerado para a análise de perfil será a mistura de normais com 12 clusters.

4.2 Análise de Perfil

De acordo com a figura 3 contendo as classificações para o modelo de mistura de gaussianas, podemos tirar as seguintes conclusões a respeito dos grupos latentes encontrados pelos modelos, bem como as diferenças das técnicas:

- Cluster 1: grupo de línguas germânicas, incluindo idiomas do oeste alemão como Africâner (que tem origem da colonização holandesa e alemã), Holandês e Frísio, porém também contém linguas do norte germânico (escandinávia) e da própria Alemanha.
- Cluster 2: grupo de línguas túrquicas, localizadas no centro-oeste da ásia, no Azerbaijão, Turquia e Uzbequistão.
- Cluster 3 e Cluster 7: grupos de línguas eslavas-meriodionais, da região da Bósnia e Hezergovina, Eslováquia, Eslovênia e República Tcheca. Entretanto, foi dividido em dois grupos tanto na figura 3 quanto na figura 4.
- Cluster 4: grupo composto por línguas românicas como o Catalão, Córsico, Italiano e Românico, bem como línguas que sofreram influência do latim ao longo da sua história, como o inglês e o maltês. O Esperanto é uma língua artificial criada com elementos latinos e anglo-saxões, o que justifica sua classificação neste grupo.
- Cluster 5: idiomas da oceânia. Cebuano é uma das línguas regionais das Filipinas, Javanês é relativo à ilha de Java na Indonésia, bem como o Sundanês, do povo Sundanês.
- Cluster 6: grupo de línguas originadas na faixa central da África, do grupo das Línguas Bantas. Chichewa e Shona estão presentes em Moçambique e em Malawi, Igbo no Congo, Quiniaruanda em Ruanda e Suaíli na Quênia, Tanzânia e Uganda.
- Cluster 8: grupo formado por duas línguas do Norte Germânico, porém com grande influência histórica do Eslavo, devido sua proximidade com a Rússia.
- Cluster 9: grupo formado por línguas de todos os lugares do mundo, com exemplos da europa como o Francês e o Basco, da África com o Somali e Sesotho, da Oceania com o Maori e Malagasi e da Ásia com o Vietnamita e Hmong. No geral, se porta como um grupo de "ruído".
- Cluster 10: possui dois idiomas gaélicos: o Irlândes e o Escocês Gaélico, são idiomas tradicionais nas Ilhas Britânicas.
- Cluster 11: o Português e o Espanhol são idiomas românicos, assim como os presentes no cluster 4, porém tem como característica marcante o uso estruturado das vogais na maioria das sílabas, como foi ilustrado na análise de dados em 3.1. Por não terem uma contagem alta de pares de letras com duas consoantes, os dois idiomas se separam das demais línguas românicas e se encontram em um cluster próprio.
- Cluster 12: por fim, mais um grupo formado por Línguas Bantu, semelhante ao Cluster 6. Porém, tanto Xhosa como Zulu são encontradas mais ao Sul da África, fator que explica sua proximidade e cluster próprio.

É possível dizer que o modelo de misturas gaussianas foi apropriado para encontrar conexões latentes entre as línguagens a partir da métrica utilizada. Apesar de não permitir uma visualização abrangente de todos os níveis de conexões e suas proximidades, como o clustering hierárquico provê na Figura 2, o GMM trouxe respostas assertivas para questões essenciais no aprendizado de máquinas não-supervisionado, como o número ideal de clusters e até mesmo a detecção de outliers.

O clustering hierárquico, por si só, provê resultados diferentes e não muito claros a depender do ponto de corte escolhido, bem como dificulta o reconhecimento de observações que podem ser tidas como ruídos. Neste caso, a análise dos dados se beneficia do uso conjunto das duas técnicas.