Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica

Departamento de Estatística

ME731 - Análise Multivariada

Análise Multivariada - PCA

Leonardo da Silva Araújo

RA: 220044

CAMPINAS

2024

Introdução

Iris é um gênero de plantas com flor, muito apreciado pelas suas diversas espécies, que ostentam flores de cores muito vivas. São, vulgarmente, designadas como lírios, embora o termo se aplique com mais propriedade a outro tipo de flor. É uma flor muito frequente em diversos jardins. O termo íris é compartilhado, contudo, com outros gêneros botânicos relacionados, da família *Iridaceae*.

Além do seu papel ecológico, as flores possuem alta importância cultural para os seres humanos. Através da história e das diferentes culturas, a flor sempre teve um lugar nas sociedades humanas, seja pela sua beleza intrínseca ou pelo seu simbolismo.[1]. O conjunto de dados que foi analisado é um conjunto de dados que consiste em 50 amostras de cada uma das três espécies de *Iris* (*Iris setosa, Iris virginica e Iris versicolor*). Quatro variáveis foram medidas em cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros.

O objetivo aqui é analisar de forma adequada como podemos reduzir sua dimensionalidade, mantendo características importantes, identificar padrões/agrupamentos entres as diferentes espécies e explorar suas relações, como correlações, visualização de gráfico de nos auxilie na nossa interpretação e obter *insights* e informações valiosas, obtendo conclusões pertinentes acerca da análise multivariada PCA.

Metodologia

Primeiramente foi utlizado o Software Rstudio na importação dos dados foi verificado se há observações faltantes, como também organização dos dados e renomeação de alguma variável que fosse importante para simplificar nossa análise. Em seguida, criamos a matriz de correlação e sua visualização através de uma gráfico. Posteriormente, análise de PCA e suas variâncias tal como a geração do gráfico de PCA para cada componente. Por fim, a visualização do gráfico de boxplot para identificar possíveis *outliers* que possa influenciar em nossa análise. Todos os gráficos como dados estatísticos da análise realizada, foi escrita no LaTeX online (*overleaf*) o que permitiu gerar este relatório em PDF.

[1] https://en.wikipedia.org/wiki/Flower

Resultados

Análise Descritiva

Primeiramente começamos dando uma espiada nos dados, verificando que não há dados faltantes. Em seguida, utilizando apenas dados numéricos com os dados das informações das 50 amostras de cada espécie, sendo o comprimento e o tamanho das pétalas e sépalas em centímetros foi obtida a matriz de correlação que nos permite entender como as variáveis estão realacionadas entre si, como ilustrado na tabela1 .

Tabela 1: Informações do Dataframe

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Em seguida, para faciliar a interpretação , a figura1 nos auxilia na interpretação das nossas correlações. Note que a correlação entre **Sepal.Length**(comprimento da sépala) e **Sepal.Width**(largula da sépala) ou vice-versa representa baixa correlação, como nota-se na figura1 representado por um círculo pequeno(baixa correlação) o que corrobora os dados mostrados na tabela1. Em contrapartida, quanto maior for **Petal.Length**(comprimento da pétala) e **Petal.Width**(largura da pétala) maior será sua correlação entre **Petal.Width**(largura da pétala) e **Petal.Length**(comprimento da pátala) respectivamente. É importante observar que essas correlações são válidas para as 3 espécies, *Iris Setosa, Iris Virginica e Iris Versicolor*.

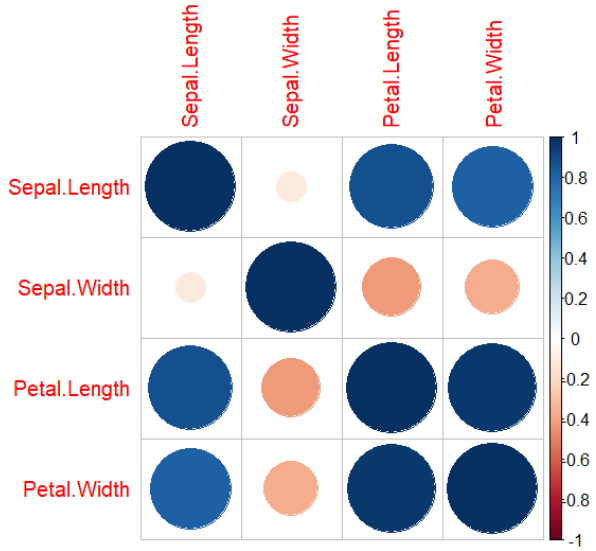

Figura 1: Matriz de correlação

Análise de componentes principais

Em seguida foi realizado PCA para nos ajudar a reduzir a complexidade dos dados, mantendo as características importantes e informações sobre quanto a variância é explicada por cada componente principal, permitindo uma análise mais aprofundada das relações que pode existir entre as variáveis e permitindo visualizar padrões até então ocultos, a priori.

Desvio Padrão:

- **PC1:** 1.7084
- **PC2:** 0.9560
- **PC3:** 0.38309
- **PC4:** 0.14393

Os resultados do desvio padrão acima, indica a quantidade de variabilidade que cada componente principal captura. PC1, sendo o primeiro componente, captura a maior quantidade de variância, seguido por PC2, e assim por diante. Isso significa que a maioria das informações nos dados está concentrada em PC1.

Proporção de variância:

- **PC1:** 0.7296 (ou 72.96%)
- **PC2:** 0.2285 (ou 22.85%)
- **PC3:** 0.03669 (ou 3.67%)
- **PC4:** 0.00518 (ou 0.518%)

Acima podemos ver o quanto da variância total dos dados é explicado por cada componente. O PC1 explica cerca de 73% da variância total, o que é bastante significativo. O PC2 explica cerca de 23%. Já os PCs 3 e 4 explicam muito pouco da variância (3.67% e 0.518%, respectivamente). Isso indica que, após os dois primeiros componentes, a adição de mais componentes não traz informações relevantes.

Proporção acumulada:

- **PC1:** 0.7296 (72.96%)
- **PC2:** 0.9581 (95.81%)
- **PC3:** 0.99482 (99.48%)
- **PC4:** 1.00000 (100%)

A proporção acumulada da variância visualiuzada acima, podemos notar que, ao considerar apenas os dois primeiros componentes (PC1 e PC2), explica cerca de 95.81% da variância total dos dados. Com três componentes, você atinge cerca 99.48%. Isso sugere que, para muitos propósitos práticos, pode ser suficiente utilizar apenas os dois primeiros componentes.

Podemos ver na figura2 o comportamento dos componentes principais e suas variâncias são exibidos, corroborando as nossas informações relatadas acima. Em suma, **PC1** é o mais importante, explicando quase 73% da variância. O **PC2** também, com 22.85% porém nos componentes 3 e 4 há pouca relevância e perceba que sua variância começa a se estabilizar, sugerindo que a mioria das informações pode ser capturada apenas com os componentes **PC1** e **PC2**. Por fim, a combinação dos dois primeiros componentes captura mais de 95% da variância, o que é eficaz para reduzir a dimensionalidade dos dados sem perder muita informação.

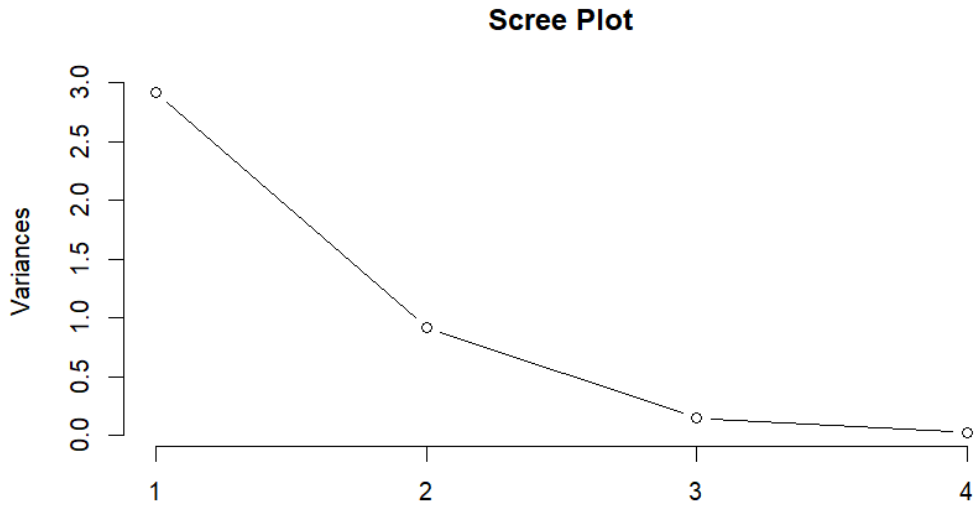

Figura 2: Variâncias PCA

Agrupamento PCA entre as espécies

Em seguida, optamos por gerar um gráfico de dispersão com as diferentes espécies agrupadas.

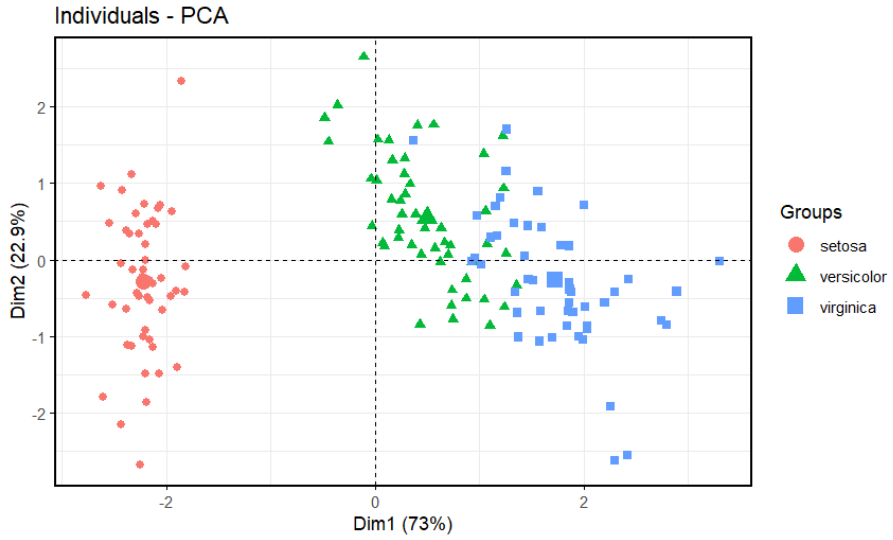

Figura 3: PCA agrupado por espécie

No gráfico da figura3 os dados foram agrupados por espécie. Além disso, a PCA busca maximizar a variância nos dados, daí se as características medem diferenças significativas entre as espécies, isso refletirá em como os pontos se agrupam. Nota-se que as espécies *Setosa, Versicolor* e *Virginica* estão bem agrupadas, principalmente a espécie *Setosa*(em vermelho) que no agrupamento se distância das demais espécies, isso indica que as características, ou seja, a medição das variáveis são eficazes para classificar essa espécie. As espécies *Versicolor*(verde) e *Virginica*(azul) quase não se sobrepõem, porém estão os pontos observados estão muito próximos e isso pode sugerir que essas duas espécies pode ser semelhantes em suas características. Em contrapartida, se essas duas espécies não for semelhante em relação as características medidas, os pontos verdes e azuis deveriam se enmcontrar ageupadas e distantes umas das outras.

Contudo o gráfico da figura3 facilita nossa visualização de como as espécies se agrupam com base nas características medidas. As diferenças e similaridades entre as espécies podem ser observadas em relação a esses dois componentes mais importantes PC1**(Dim1)** e PC2**(Dim2)** que juntos somam 95.9% da variância total.

Interpretando os resultados

O que são cargas dos componentes?

- Cada valor nas cargas representa a contribuição de uma variável para um componente principal. Um valor alto(positivo ou negativo) indica que a variável tem uma influência significativa nesse componente.
- As colunas correspondem aos componentes principais (PC1, PC2, PC3, PC4) e as linhas correspondem às variáveis originais (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width).

Na tabela2 podemos visualizar como cada variável original contribui para cada componente principal PCA, vamos interpretá-las.

Tabela 2: Componente principal em cada variável

	PC1	PC2	PC3	PC4
Sepal.Length	0.5210659	-0.37741762	0.7195664	0.2612863
Sepal.Width	-0.2693474	-0.92329566	-0.2443818	-0.1235096
Petal.Length	0.5804131	-0.02449161	-0.1421264	-0.8014492
Petal.Width	0.5648565	-0.06694199	-0.6342727	0.5235971

PC1 - Primeiro componente principal:

- **Sepal.Length:** 0.5211 — Aumenta a variância em PC1.
- **Sepal.Width:** -0.2693 — Tem uma contribuição negativa, mas menor.

- **Petal.Length:** 0.5804 — Contribuição positiva significativa.
- **Petal.Width** 0.5649 — Também contribui positivamente.

O PC1 é influenciado principalmente por Sepal.Length, Petal.Length e Petal.Width, sugerindo que esses atributos estão inter-relacionados e explicam uma grande parte da variação nos dados.

PC2 - Segundo componente principal:

- **Sepal.Length:** -0.3774 — Contribuição negativa.
- **Sepal.Width:** -0.9233 — Contribuição muito forte e negativa.
- **Petal.Length:** -0.0245 — Contribuição quase nula.
- **Petal.Width:** 0.0669 — Também pequena.

O PC2 é fortemente influenciado por Sepal.Width. Essa direção nos espaços dos componentes pode ser interpretada como uma variação oposta aquela do primeiro componente.

PC3 - Terceiro componente principal:

- **Sepal.Length:** 0.7196 — Contribuição positiva significativa.
- **Sepal.Width:** -0.2444 — Contribuição negativa.
- **Petal.Length:** -0.1421 — Contribuição negativa menor.
- **Petal.Width:** -0.6343 — Contribuição negativa significativa.

O PC3 mostra uma relação complexa mas é dominado por Sepal.Length em direção positiva e Petal.Width em direção negativa.

PC4 - Quarto componente principal:

- **Sepal.Length:** 0.2613 — Contribuição positiva.
- **Sepal.Width:** -0.1235 — Contribuição negativa pequena.
- **Petal.Length:** -0.8014 — Contribuição negativa forte.
- **Petal.Width:** 0.5236 — Contribuição positiva.

O PC4 é fortemente influenciado por Petal.Length negativamente e Petal.Width positivamente.

Em seguida, verificamos se há outliers nos dados e como eles podem estar afetando o PCA através da figura4 e notamos que apenas na variável **Sepal.Width** há presença de *outliers*. Isso pode ter algumas implicações diretas, como aumento da variância na variável, o que pode fazer com que a PCA se concentre mais nessa variável. Contudo, **Sepal.Width** é uma variável que impacta principalmente PC2, onde sua carga é mais forte. Isso indica que a largura da sépala é uma variável importante para a variação que está sendo capturada nesse componente.

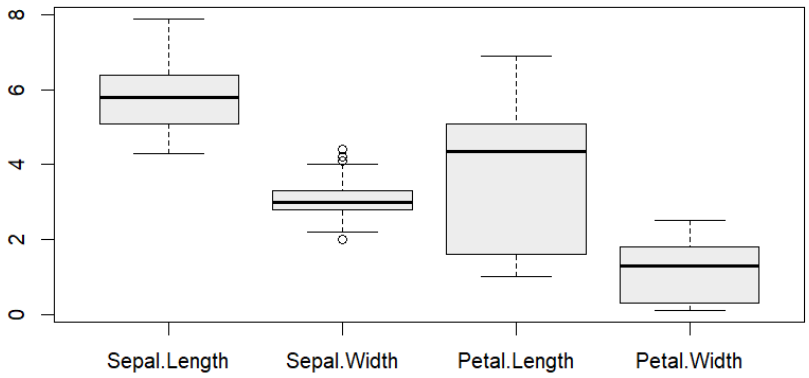

Figura 4: Boxplot das variáveis

Conclusões

A análise de PCA nos revelou alguns padrões significativos nas varáveis do conjunto de dados analisados **iris**, destacando a importância das características das pétalas na separação das espécies. O **PC1** mostrou-se predominantemente influenciado por essas características, elucidando que desempenham um papel crucial na identificação das diferentes espécies de flores e por outro lado, o **PC2** apresentou uma forte relação com a largura da sépala, sugerindo que esta variável pode servir como um discriminador eficaz entre as espécies.

Além disso, os outros dois componentes **PC3** e **PC4** capturam interações mais complexas entre as medidas de comprimento e largura das sépalas e pétalas. Isso nos aponta que, embora as variáveis individuais sejam essenciais, compreender suas interações pode proporcionar uma visão mais profunda as relações entre as características morfológicas das flores.

Por fim, uma possível aplicação prática por exemplo, seria em taxonomia, horticultura e conservação, permitindo uma identificação mais precisa e eficiente das espécies se baseando-se nessas medidas simples.

Referências

[0]Notas de aula.

[1] Johnson, R.A. & Wichern, D.W.. *Applied Multivariate Analysis,* Quarta Edição, Prentice-Hall, Nova Jersey, 1998. * Roteiro do curso.

[1.A] —— , Sexta Edição, Pearson Education Limited, Nova Jersey, 2007.

[2] Artes, R.& Barroso, L.. *Métodos Multivariados de Análise Estatística Estatística.* Blucher, São Paulo, 2023.

[3] Everitt, B..*Cluster Analysis.* Quinta Edição, Wiley & Sons, Nova Iorque, 2011.

[4] Koch, I.. *Analysis of Multivariate and High-Dimensional Data: Theory and Practice.* Cambridge University Press, Nova Iorque, 2014.

[5] Mardia, K.V., Kent, J.T.& Bibby, J.M.. *Multivariate Analysis.* Sétima Reimpressão, Academic Press, Londres, 2000.

[6] Volpato, G.L.. *Guia Prático para Redação Científica.* Best Writing, Botucatu, 2015.