Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica

Análise Multivariada

Predição de diagnósticos de tumores via características detectadas em exames de imagem

Bernardo Abib de Almeida
RA: 236053

Campinas, São Paulo

ÍNDICE

RESUMO

Neste trabalho, serão aplicadas técnicas de análise multivariada em um conjunto de dados contendo uma variável categórica e 30 variáveis numéricas altamente correlacionadas.

O conjunto de dados em questão são observações de 568 tumores, contendo informações sobre simetria, perímetro, área, concavidade e outras medidas físicas, juntamente com a variável categórica "diagnóstico", que indica se o tumor é benigno ou maligno.

A hipótese inicial do trabalho, era de que os tumores malignos poderiam ser diferenciados dos benignos baseando-se somente em suas características físicas trazidas nos dados obtidos em exames de imagem, antes mesmo do exame de biopsia.

Por se tratar de um conjunto de dados com muitas variáveis numéricas correlacionadas, foi realizado uma redução de dimensão por Análise de Componentes Principais via matriz de correlação. As 30 variáveis numéricas foram transformadas em componentes principais, e através do critério de *Kaisser*, os 6 componentes que mais explicavam a variância dos dados foram selecionados e mantidos.

Após a redução de dimensão via ACP, foi realizado uma Análise de Variância Multivariada para testar se o vetor de médias dos componentes principais era o mesmo para ambos os diagnósticos.

Rejeitou-se essa hipótese, e então concluímos que há diferença no vetor de médias dos componentes principais de acordo com o diagnóstico e, portanto, dado que os componentes principais são combinações lineares das variáveis originais, há evidência estatisticamente significante de que o diagnóstico pode ser predito apenas pelas características físicas do tumor.

DESCRIÇÃO DO BANCO DE DADOS

São 568 observações de tumores, contendo a variável "diagnóstico" (maligno ou benigno) e dez características físicas dadas por variáveis numéricas:

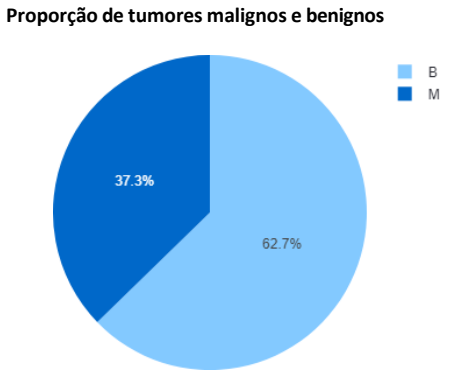
a) raio (média das distâncias do centro aos pontos no perímetro)
b) textura (desvio padrão dos valores de escala de cinza)
c) perímetro
d) área
e) suavidade (variação local nos comprimentos do raio)
f) compacidade (perímetro^2 / área - 1,0)
g) concavidade (severidade das partes côncavas do contorno)
h) pontos côncavos (número de partes côncavas do contorno)
i) simetria

j) dimensão fractal ("aproximação da linha costeira" - 1)

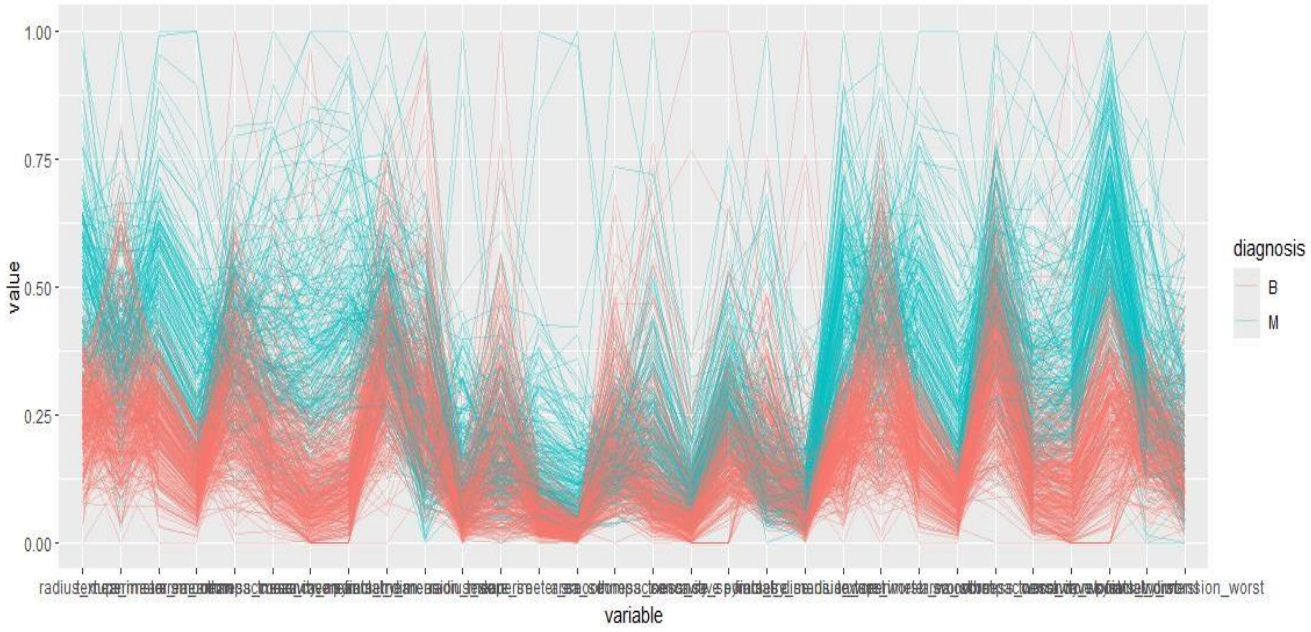
Para cada uma dessas características, há as medidas de média (*variável_mean*), desvio padrão (*variável_se*) e maior/pior medida (*variável_worst*), totalizando 30 variáveis numéricas e uma categórica. Não há valores faltantes.

ANÁLISE EXPLORATÓRIA

O conjunto de observações tem a seguinte proporção de diagnósticos:



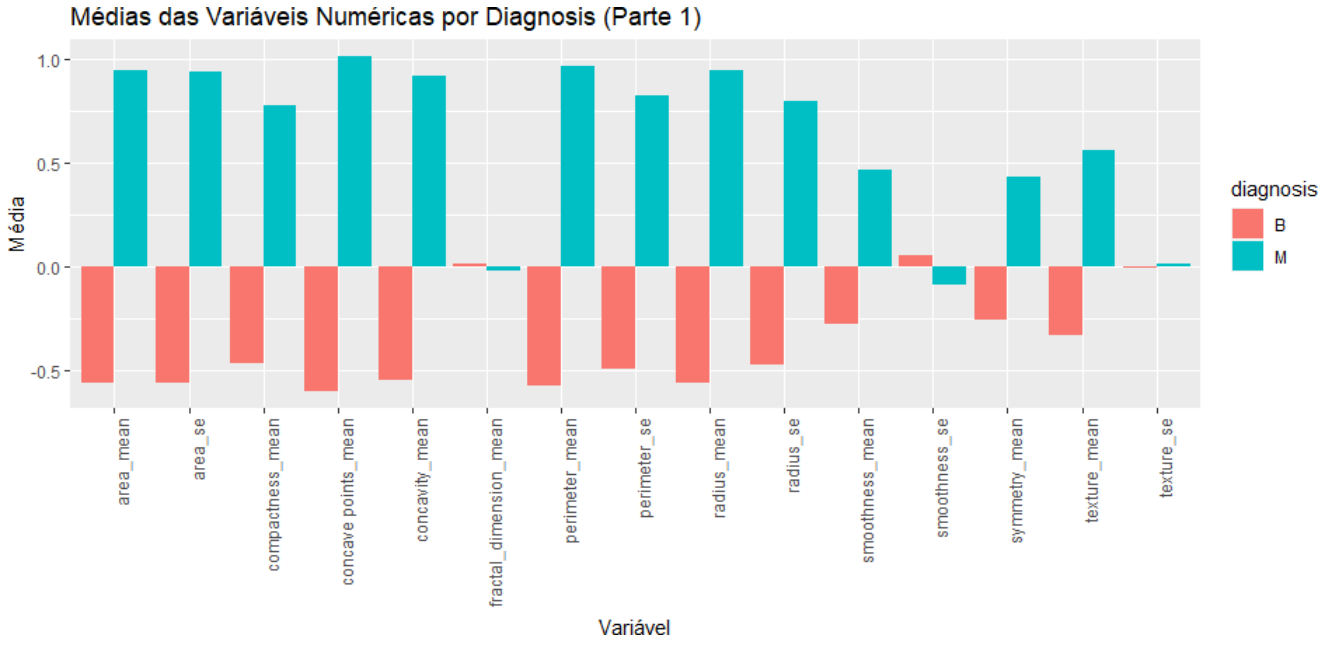
Para visualizar padrões de comportamento das variáveis no âmbito multivariado, podemos analisar o gráfico de coordenadas paralelas:

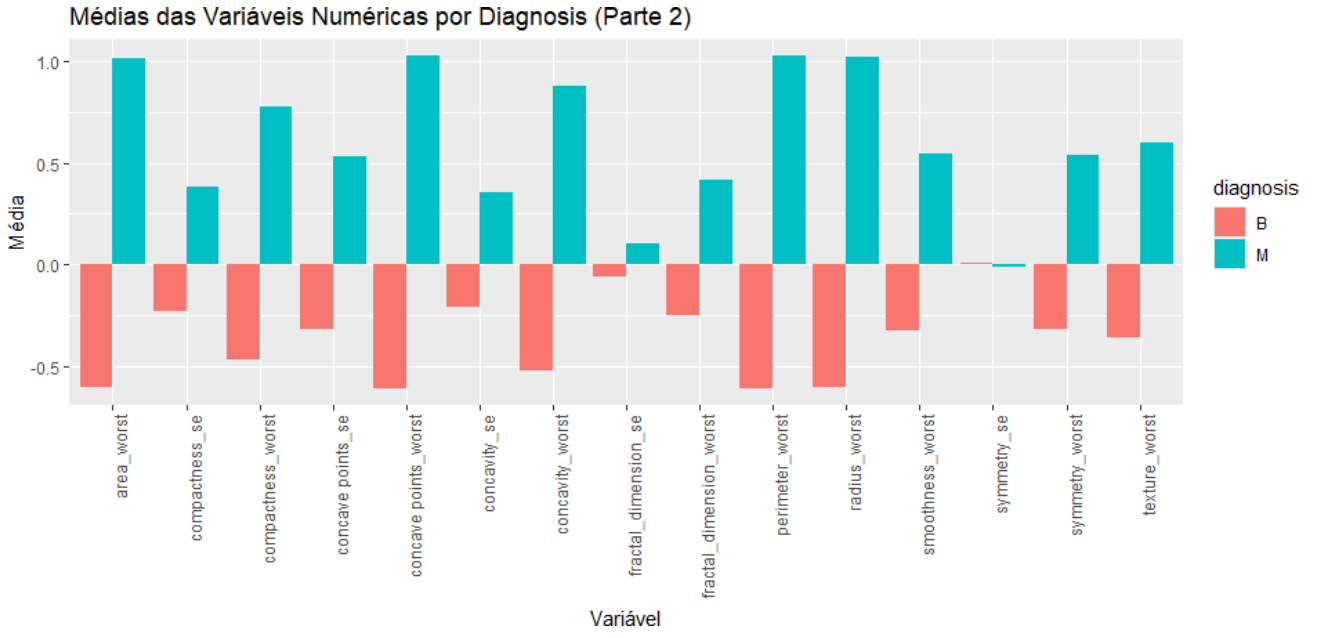


Apesar de poluído devido a alta dimensão dos dados, ainda é possível identificar os padrões das linhas vermelhas (observações benignas) e azuis (observações malignas), que claramente possuem comportamentos diferentes.

A maior parte das observações cujo tumor era benigno parece se acumular no fundo do gráfico, indicando características físicas com medidas menores. Já as observações cujo tumor era maligno, parecem se concentrar em valores mais altos na maior parte das variáveis.

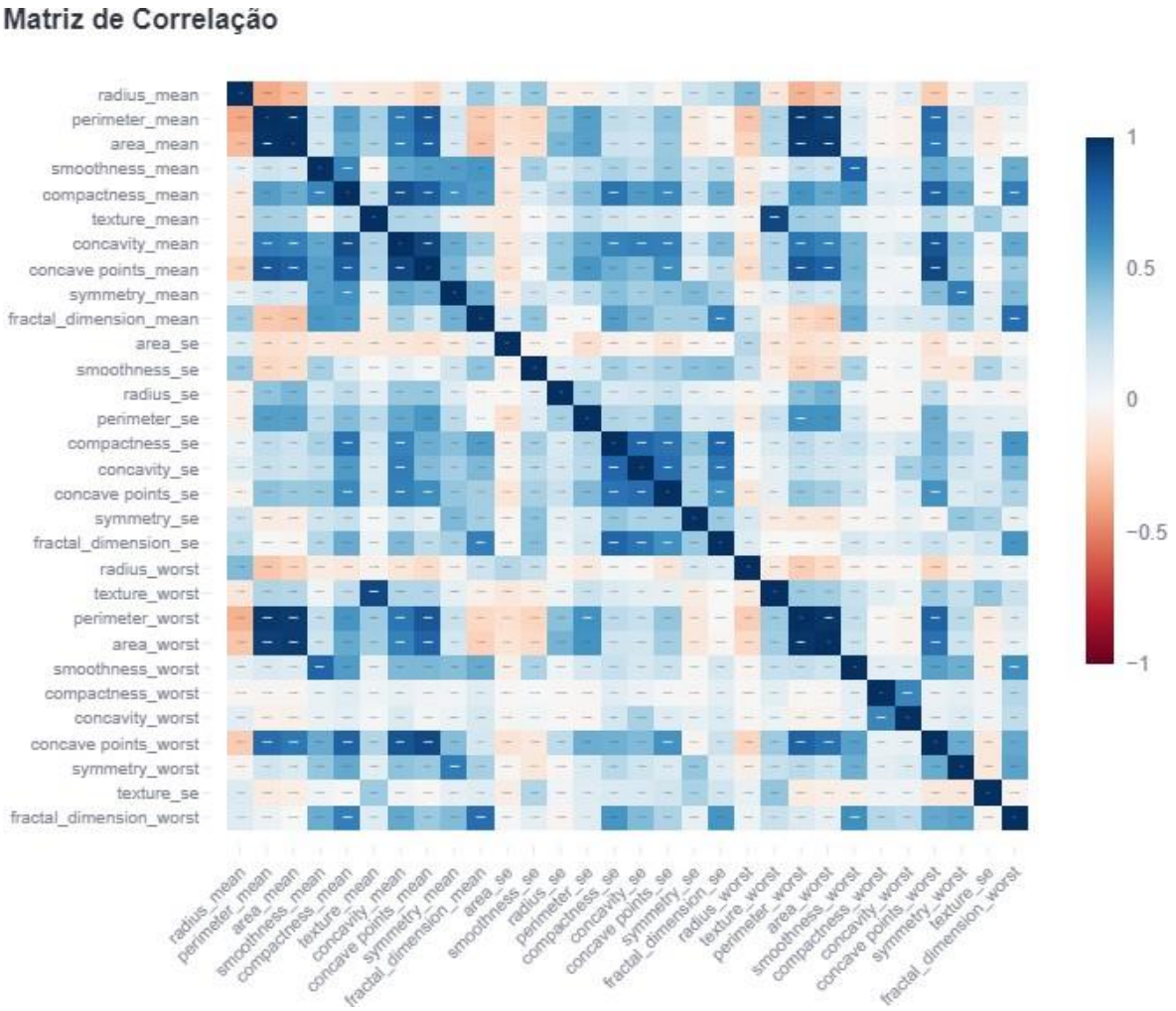
Esse padrão de comportamento fica ainda mais claro ao comparar o vetor de médias das observações com diagnóstico de tumor maligno (M) com o vetor de médias das observações com diagnóstico de tumor benigno (B) das variáveis normalizadas:





No gráfico acima, observa-se imediatamente que a média da maior parte das variáveis tem grandes alterações dependendo do diagnóstico do tumor. Das 30 variáveis numéricas, apenas as variáveis de média e de desvio padrão da ‘dimensão fractal’, desvio padrão da ‘lisura’, desvio padrão da ‘textura’ e desvio padrão da ‘simetria’ aparentam não exibir alterações notáveis em suas médias. O restante das características físicas segue o padrão que começamos a identificar quando analisamos as coordenadas paralelas: tumores malignos parecem ter medidas maiores do que os benignos.

Também foi observado que as variáveis de medidas são altamente correlacionadas entre si, o que é intuitivo, pois faz sentido que a área do tumor esteja correlacionada a medidas como o perímetro e o raio. Essas relações podem ser visualizadas na matriz abaixo.



A independência entre as observações, o alto número de variáveis numéricas e as fortes correlações entre elas foram os motivadores para a redução de dimensão via ACP.

ANÁLISE DE COMPONENTES PRINCIPAIS

A partir daqui, vamos pressupor que as variáveis possuem relações lineares entre si (averiguado por gráficos de dispersão omitidos da análise exploratória) e que seguem a distribuição Normal (pelo Teorema Central do Limite). Esses pressupostos serão discutidos na sessão de "Discussões e Conclusão".

Os componentes principais são combinações lineares das variáveis originais, cujos coeficientes são os elementos dos autovetores associados à matriz de covariância ou de correlação.

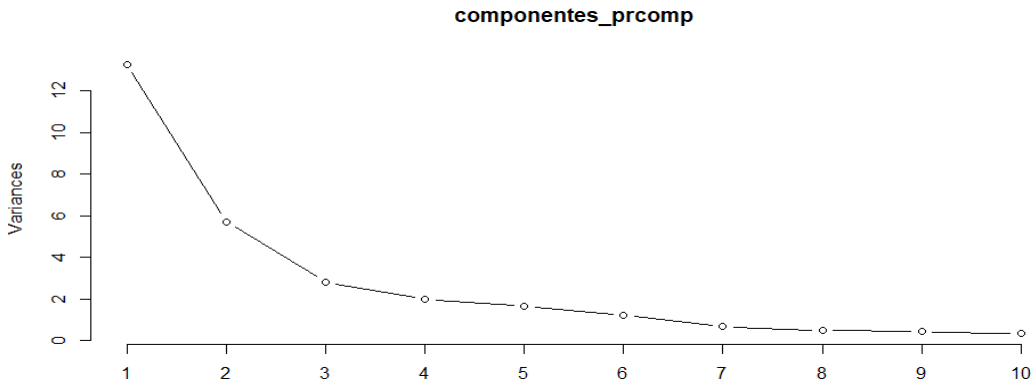
Na prática, para encontrar os componentes de um conjunto de dados **X**:

1. Calcula-se a matriz de covariância ou de correlação dos dados (chamamos de **S**)
2. Extrai-se os autovetores associados à essa matriz (obtendo uma nova matriz cujas colunas são os autovetores de **S**; chamamos de **M**)
3. Calcula-se o produto de **X** e **M**, obtendo as combinações lineares descritas acima.

É intuitivo que a quantidade de componentes seja igual ao número de variáveis originais. Porém, os componentes são não-correlacionados (o que posteriormente será relevante para a MANOVA, que é sensível a correlação entre preditores).

Devido à diferença de unidades e de escala entre as variáveis numéricas, neste trabalho, foi decidido realizar a análise dos componentes principais através da decomposição espectral da matriz de correlação, ao invés da matriz de covariância.

Após a obtenção dos componentes principais, analisou-se o quanto da variância dos dados cada componente explica:



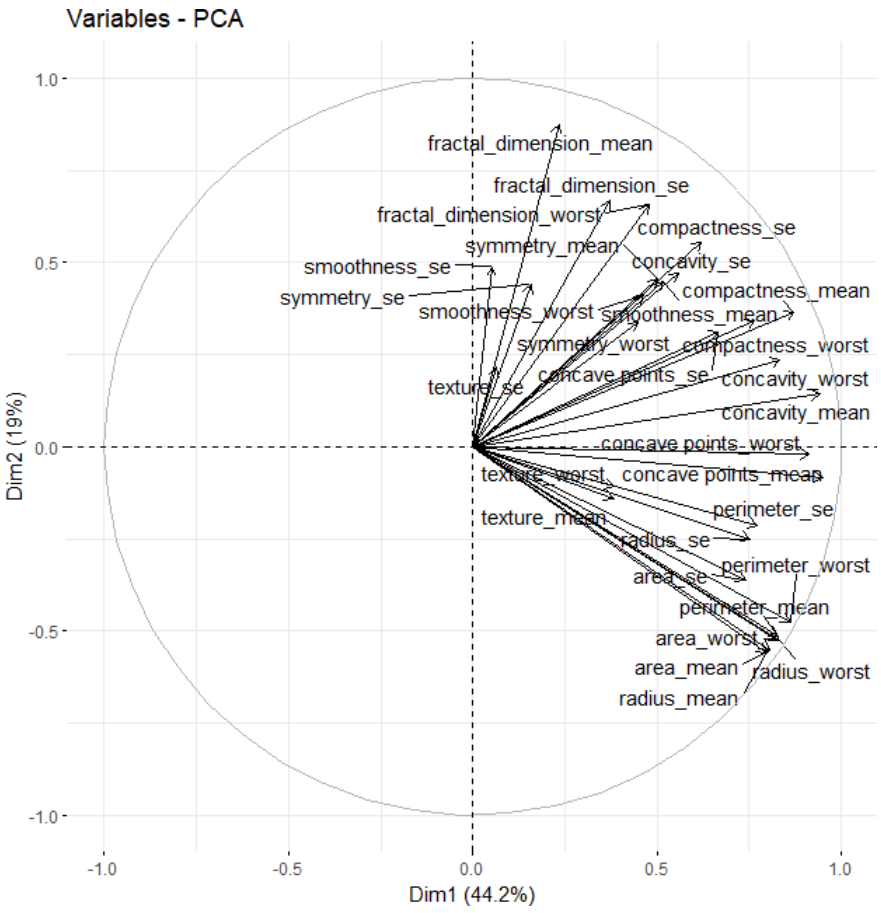
Os componentes estão em ordem de "maior responsabilidade pela variância total". Ou seja: o primeiro componente é o que mais explica a variabilidade dos dados, seguido pelo segundo, e assim por diante.

Pelo gráfico, observa-se que grande parte da variabilidade dos dados é explicada pelos seis primeiros componentes, e que a variância explicada a partir do décimo componente é insignificante. Para decidir quantos componentes seriam utilizados para representar a variância dos dados, foi utilizado o critério de autovalores que satisfazem

$\lambda_i > Média\ de\ \lambda$ (conhecido como *critério de Kaisser*). Com isso, foi decidido manter os 6 primeiros componentes principais, que explicam 88,77% da variabilidade total dos dados. Essa variabilidade é calculada pela proporção dos 6 primeiros autovalores em relação a soma total. Ou seja:

$$var\ total = \frac{\lambda 1 + \lambda 2 + \lambda 3 + \lambda 4 + \lambda 5 + \lambda 6}{\sum_{i=1}^{p} \lambda_i}$$

Na nuvem de variáveis abaixo, é possível visualizar as cargas e quais variáveis originais tem mais impacto em cada um dos dois componentes principais (que juntos explicam 63,3% da variabilidade total).



Observa-se que todas as variáveis são positivamente correlacionadas com o primeiro componente principal. Analisando os vetores mais próximos, também é possível visualizar as variáveis que possuem correlação mais alta entre si.

As variáveis de média e 'worst' do rádio e de média e 'worst' da área, são as variáveis que mais se correlacionam negativamente com o segundo componente. Em contra partida, a média da dimensão fractal é a variável que possui maior correlação positiva.

ANÁLISE DE VARIÂNCIA MULTIVARIADA (ONE-WAY)

Após a redução de dimensão realizada na Análise de Componentes Principais, passamos de 30 variáveis numéricas altamente correlacionadas para 6 componentes principais não-correlacionados.

Retomando a nossa hipótese inicial do trabalho, o objetivo agora é procurar evidências significativas de que as características físicas dos tumores podem predizer seu diagnóstico com alto nível de confiança, antes do exame de biópsia.

Como os componentes principais são combinações lineares das variáveis originais, se indicarmos que há mudanças significativas nos vetores de médias dos componentes de acordo com o tipo de diagnóstico, isso também será evidência de que há alterações significativas nas variáveis originais.

Seja μ_B o vetor de médias dos componentes principais onde o diagnóstico é "benigno" e μ_M o vetor de médias dos componentes principais onde o diagnóstico é "maligno". Então, formula-se as hipóteses:

H_0: $\mu_B = \mu_M$

H_1: $\mu_B \neq \mu_M$.

Finalmente, foi realizada uma análise de variância multivariada pelo teste de Wilks (construído como a generalização multivariada do teste F univariado, no caso da ANOVA)

O modelo, utilizando os componentes principais como preditores e o diagnóstico como resposta, obteve o seguinte resultado:

	Df	Wilks	approx F	num Df	den Df	Pr(>F)	
diagnosis	1	0.29425	224.26	6	561	< 2.2e-16	***

Com a estatística do teste suficientemente extrema e um baixo p-valor, rejeita-se a hipótese de que os vetores de médias dos componentes são iguais em ambos os diagnósticos. Logo, como os componentes são combinações lineares das características físicas nos dados originais, há evidência de que estas têm variação estatisticamente significativa quando o tumor é benigno ou maligno.

É importante ressaltar, porém, que a realização da 'one-way' MANOVA (que avalia a alteração no vetor de médias em relação a uma única variável categórica) tem como pressuposto a normalidade multivariada e a equivalência das matrizes de covariância. Após realizar um teste M de Box (comparando as variâncias generalizadas), obteve-se um resultado que rejeitou a hipótese das matrizes de covariância serem iguais:

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  componentes_selecionados[, -7]
Chi-Sq (approx.) = 555.79, df = 21, p-value < 2.2e-16
```

Uma possibilidade, é a rejeição da hipótese ter ocorrido devido à falta de normalidade multivariada (pressuposto que havia sido assumido baseado no Teorema Central do Limite), visto que o teste de Box é sensível a desvios da normalidade que, em casos mais extremos, deve ser tratado via transformações.

DISCUSSÕES E CONCLUSÃO

Apesar da análise de componentes principais e da análise de variância multivariada trazerem resultados poderosos, ambas possuem limitações de acordo com os pressupostos estatísticos que as devem preceder.

Por exemplo, a ACP trabalha somente com correlações lineares. Apesar das relações entre as 30 variáveis terem sido visualizadas em gráficos de dispersão (omitidos deste trabalho) e aparentemente tratarem-se de relações lineares, não foi realizado nenhum teste rigoroso para verificá-las.

Apesar dessas limitações, tanto a ACP quanto a MANOVA foram um sucesso. Reduzimos a dimensão das nossas trinta variáveis altamente correlacionadas à 6 componentes não-correlacionados que explicam 88,77% da variabilidade total dos dados, e com eles pudemos testar a variação das características físicas originais em relação ao diagnóstico, confirmando nossa hipótese levantada na análise exploratória de que nossas variáveis originais podem determinar com nível significativo de confiança o diagnóstico de um tumor.

Numa extensão interessante deste trabalho, poderíamos utilizar o modelo para fazer predições de diagnósticos, e testar a hipótese de que sua acurácia é tão alta quanto a da biópsia.

REFERÊNCIAS

Johnson, R. A., & Wichern, D. W. (2007). “Applied multivariate statistical analysis”. 6ª ed. Capítulo 4

Härdle, W. K., & Simar, L. (2019). “Applied Multivariate Statistical Analysis”. 5ª ed. Capítulos 7 e 11

BioTuring Team, (2018). “How to read PCA biplots and scree plots”, Artigo no Medium

Gráficos gerados no site https://basic-stats-automation.streamlit.app/ e no R.