Análise de Componentes Principais*:

Um curioso caso de Pizzas

Ricardo N. Pires, 2001645, Mestrado em Estatística, Matemática e Computação, Universidade Aberta

***Sumário* — Análise de Componentes Principais (ACP) é o método usado perante uma base de dados de pizzas. Pretende-se através desta técnica multivariada reduzir as variáveis ao menor número possível, por um novo conjunto de variáveis ortogonais (componentes principais) e não correlacionada entre si, relativamente à informação nutricional das pizzas. Na elaboração da ACP existem nuances e critérios que ficam ao cargo do analista. Isto é, critérios sobre a retenção ou não do número de componentes e nuances nos pesos das interpretações dos vetores próprios, devido a diferentes métodos rotação. É feito uma passagem breve pelos diversos passos a seguir sendo que estes são uma de muitas formas de proceder, podendo a mesma análise dos respetivos dados através da ACP chegar a valores disparos no momento da escolha de critérios e método diferentes.**

I. Introdução

A Análise de componentes principais (ACP) é certamente dos métodos mais utilizados na estatística multivariada e é praticamente transversal a sua aplicação nas mais diversas disciplinas científicas. Inicialmente introduzido por Pearson [1] e subsequentemente desenvolvido por Hotelling [2]. Tem vindo ao longo das décadas a ser discutida e difundida por diversos autores, tais como Jolliffe [3] e Jackson [4].

Na sua essência, a ACP simplifica os dados ou observações descritas por várias variáveis dependentes que por norma são correlacionadas. O objetivo cinge-se em extrair a informação com maior importância ou relevância dos dados e expressar essa informação num conjunto reduzido de novas variáveis ortogonais, intituladas de componentes principais. De forma grosseira, podemos dizer que se resume a informação existente, "compactando" esta, num menor número de variáveis (componentes). Ou dito de outra forma, procede-se à redução da dimensão dos dados preservando a maior parte da variabilidade desses dados. Não deixar de salientar, que a análise em estudo, não se tratar de uma Análise Fatorial (AF), estamos interessados em resumir ou condensar sem prejuízo de deterioração de informação, as variáveis existentes em número inferior. Isto é, componentes ortogonais à custa de combinações lineares das variáveis observadas. A intenção não é a de explicar as intercorrelações das variáveis observadas pela criação de variáveis latentes ou fatores.

II. Métodos

Um conjunto de pressupostos e validações são necessários para que a ACP possa ser elaborada. As variáveis originais devem estar correlacionadas. A estatística de Kaiser-Meyer-Oklin (KMO), o teste de esfericidade de Bartlett e a Matriz anti imagem, permitem validar a aplicação da ACP aos dados.

As componentes principais são a combinação linear das variáveis X_1, X_2, ..., X_9, tal que:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \tag{1}$$

em que as constantes a_{i1}, a_{i2}, …, a_{ip} são os elementos do vetor próprio.

Inicia-se com o teste de esfericidade de Bartlett com um nível de significância a 5%. Considera-se as hipóteses, H_0: P = I e $H_1: \lambda_1 = \lambda_2 = \cdots = \lambda_p$,com estatística de teste:

$$-\left[n - 1 - \frac{1}{6}(2p + 5)\right] ln|R| \sim \chi^2\left(\frac{1}{2}.p.(p-1)\right) \tag{2}$$

em que R é a matriz de correlações amostrais. De notar que este teste se torna sensível à dimensão da amostra. Podendo induzir a um erro tipo I.

A estatística KMO é definida de seguinte modo:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2} \tag{3}$$

Caso a Matriz anti imagem apresentar valores baixos em número significativo, estamos em condições de aplicar a análise em estudo. Existe assim, evidencia da adequação dos dados da amostra e correlação dos mesmos, para elaborar uma ACP se os testes supramencionados assim o indicarem.

As comunalidades são alvo de análise pois permitem aferir o quão as componentes retidas explicam cada variável original. A Matriz de correlações permite inclusivamente, uma análise detalhada sobre a correlação das variáveis em consonância com os testes globais mencionados acima. O critério para a retenção das componentes, neste caso, aplica-se a regra do valor próprio superior a 1 em combinação com o

*Relatório da Atividade 3 (Trabalho Individual) da Unidade Curricular de Análise de Dados Multivariados e Aplicações (22002) 2020/2021

gráfico scree. Análise da matriz das componentes é prosseguida com a possibilidade da rotação da mesma para uma melhor interpretabilidade, usando o método elaborado por Kaiser [5].

São analisados os diversos itens (variáveis originais) nas componentes principais.

III. Caso em Estudo

Nos dados em questão tem-se informação nutricional de uma amostra com dimensão de trezentas pizzas. Com nove variáveis relativas a diferentes valores nutricionais todos eles por 100 gramas. Temos assim, ***pizza*** (dez tipos diferentes), ***id, água, proteína, gordura, cinza, sódio, açúcar*** e ***calorias***. Tornar-se-á útil para posterior análise ou modelação que esta informação fosse condensada e não correlacionada. A própria possibilidade de ter um número inferior de variáveis sem deteriorar a informação existente, permite uma leitura da informação nutricional simples e eficaz. Assim sendo, usou-se ACP para esse efeito.

IV. Resultados

Com KMO de 0.524, indicando que a adequação da amostra é má, porém aceitável [6] e no teste de esfericidade de Bartlett rejeita-se a hipótese nula existindo evidência que as variáveis estão correlacionadas. Reforça-se o que foi mencionado em II sobre o teste de Bartlett. Sendo que se procedeu a uma análise da Matriz de correlações, a variável ***id*** encontra-se significativamente correlacionada somente com a variável ***pizza*** pelo que decidiu-se retirar a variável da análise. Pela regra do valor próprio superior a 1, escolheu-se duas componentes ortogonais que explicam 89,137% da variância total das variáveis originais. A primeira componente explica 60,346% da variância e a segunda componente explica 28,791% da variância não explicada pela primeira. A *fig.1* ilustra o gráfico scree que apoia na decisão da retenção do número de componentes. Quanto às comunilidades, verifica-se que a variável ***pizza*** (0.719) e ***proteína*** (0.783) são as que mais informação perdem no momento de transformar as oito variáveis de origem em duas componentes principais. As restantes variáveis mantem valores elevados próximo de 0.9 e acima.

A suspeita inicial sobre a exclusão do variável ***id*** torna-se evidente quando analisada a variância total explicada, sendo que existiu uma diminuição redundante de 1,5 pontos percentuais, quando esta não está na análise e as componentes passam de 3 para 2. Sendo a variável em causa meramente informativa e pelas razões supramencionadas, achou-se que não iria ter relevância na análise.

No caso em concreto, o uso da rotação pelo método Varimax de pouco acrescentava na interpretação das componentes, os pesos acabam por não serem acentuados de forma relevante, de tal modo que se acaba por analisar a Matriz de componentes sem rotação. Temos a primeira componente principal,

$$Z_1 = 0.947X_1 + 0.92X_2 + 0.882X_3 - 0.844X_4 - 0.842X_5 + 0.77X_6 + 0.085X_7 + 0.535X_8 \quad (4)$$

A segunda componente principal,

$$Z_2 = -0.211X_1 + 0.314X_2 + 0.262X_3 + 0.522X_4 - 0.105X_5 - 0.436X_6 - 0.958X_7 + 0.837X_8 \quad (5)$$

Sendo que as variáveis ***cinzas*** (X_1), ***gordura*** (X_2), ***sódio*** (X_3) e ***proteína*** (X_6) estão fortemente associadas com a componente principal 1, temos ainda nesta componente ***açúcar*** (X_4) e ***pizza*** (X_5) que estão ambas fortemente associadas mas em sentido opostas às outras quatro variáveis, enquanto as variáveis ***agua*** (X_7) e ***calorias*** (X_8) estão fortemente associadas à componente principal 2 e em sentidos opostos uma da outra. Onde as X_p devem ser variáveis padronizadas.

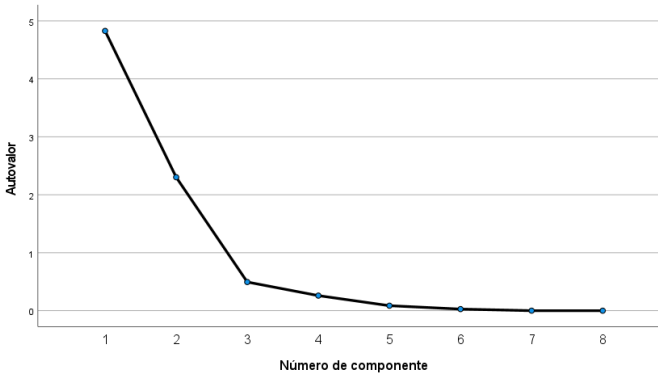

Figure 1. Gráfico Scree

V. Conclusão

Em suma, verifica-se que a componente principal um, fica associada com seis variáveis, enquanto a componente principal dois fica associadas com duas variáveis. A fig. 2 ilustra o mapa bidimensional das duas componentes retidas. Pelo posicionamento e os pesos da variáveis nas componentes não é de todo claro a possibilidade nomear as próprias. Consegue-se através desta ACP reter uma grande parte (89,137%) da informação total das variáveis originais. Porém, é visível pela fig.2 uma certa disparidade entre as variáveis nas componentes.

Simplificou-se a estrutura dos dados, sendo possível doravante prosseguir com análises mais elaboradas, ou mesmo modelar questões nutricionais em prol da perfeita junção de nutrientes, ou talvez com objetivo de maximizar o custo benefício na produção de pizzas. As possibilidades são vastas. O objeto em estudo, foi sim, reduzir variáveis sem comprometer em demasia a informação existente.

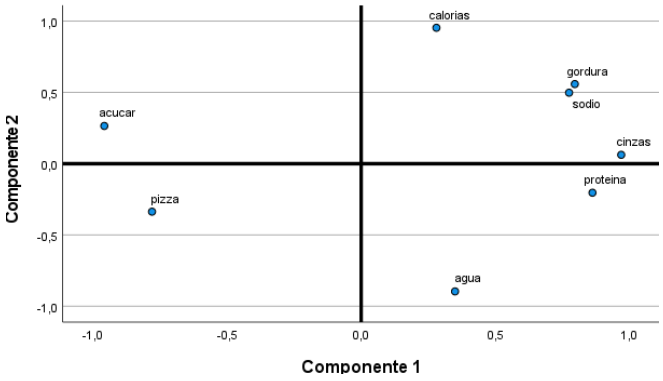

Figure 2. Gráfico de Componentes

REFERÊNCIAS

[1] Pearson K., "On lines and planes of closest fit to systems of points in space" Philo Mag. A 1901, 6 pp 559-572.

[2] Hotelling H., "Analysis of a complex of statistical variables into principal components." J. Educ. Psychol., 1933, 25, pp. 417-441.

[3] Jolliffe I.T. *Principal Component Analysis*, New York Springer, 2002.

[4] Jackson J.E., *A User's Guide to Principal Components*. New York John Wiley & Sons, 1991.

[5] Kaiser, H.F., "The varimax criterion for analytic rotation in factor analysis.", Psychometrika, 1958, 23, pp. 187-200.

[6] Reis, E., *Estatística Multivariada Aplicada*, 2ª edição, edições Sílabo, 2001.

[7] Marôco, J., *Análise Estatistica com o SPSS STATISTICS*, 8ª edição, Report Number, 2021