Unsupervised ML: Análise de Correspondência Simples e Múltipla I

12 January 2023 10:46

Análise de Correspondência (AC or CA)

É uma técnica **não supervisionada** aplicada quando deseja-se investigar a **associação entre as variáveis e entre suas categorias**. Ou seja, quando há variáveis ***Qualitativas***, podendo ser aplicadas a variáveis ***Quantitativa***s desde que estas sejam ***devidamente transformadas em Qualitativas***. É aplicada ao estudo da relação de interdependência entre duas variáveis categóricas.

As técnicas de análise de correspondência são métodos de representação de linhas e colunas de tabelas cruzadas de dados como coordenadas em um gráfico (mapa perceptual), a partir do qual se pode interpretar as similaridades e diferenças de comportamento entre variáveis e entre categorias.

Portanto a AC tem como **principais objetivos, avaliar a significância dessas similaridades, determinar coordenadas das categorias com base na distribuição dos dados em tabelas cruzadas** e, a partir dessas coordenadas, **construir mapas perceptuais.**

Mapas perceptuais nada mais são que **diagramas de dispersão** que representam as categorias das variáveis na forma de pontos em relação a eixos de coordenadas ortogonais. Mapas perceptuais são, portanto, **mapas de categorias**.

As técnicas de análise de correspondência simples e múltipla permitem considerar todo e qualquer tipo de categoria de variáveis, sem que o pesquisador precise fazer uso de ponderação arbitrária, o que incorreria em erro. Sendo assim, esta técnica pode ser usada para análise de variáveis geradas por escala Likert, evitando o problema da ponderação arbitrária. Para tal, **cada ponto da escala Likert deverá ser transformada em uma categoria da variável na análise.**

LEMBRETE:

A Escala Likert será sempre uma ponderação arbitrária?
Não, há exceções. Tais exceções ocorrem quando, os valores adotados são baseados na literatura científica prévia sendo já bem estabelecidos e fundamentados em ampla base teórica.

Exemplos de aplicação:

A associação entre:

- Faixa de renda e status na aprovação de credito;
- Escolaridade e grupo ocupado;
- Tipo de solo e cultura implementada;
- Gravidade dos sintomas da doença e comorbidades.
- Aplicação de Análise de Correspondência Simples: avaliação do comportamento de consumo, descrito pela preferência por determinados tipos de estabelecimento varejista, e faixa de idade dos consumidores.
- Análise de Correspondência Múltipla: investigar a relação entre o país de origem, o setor de atuação e a faixa de lucratividade de empresas de capital aberto.

Resumo

1. *É um modelo não supervisionado, ou seja :*
   i. *Avalia a relação conjunta entre as variáveis*
   ii. *Não há modelos do tipo "yi = x1i + x2i + ... + ui "*
   iii. *Não é adequada para fins de inferência*
   iv. *Se novas observações forem adicionadas ao banco de dados, é adequado refazer a análise*
2. *Quando aplicar a AC (or CA)?*
   i. *Quando quero avaliar variáveis Qualitativas.*
   ii. *Quando meu objetivo é verificar se existe associação estatísticamente significativa entre essas variáveis categóricas criando o mapa perceptual.*
3. *Se houver variável Quantitativa é necessário que ela passe por um processo de categorização prévia.*
4. *Seu uso para escala Likert demanda que, cada um dos valores da escala seja convertido em uma categoria (variável qualitativa).*

Análise de Correspondência Simples (ANACOR)

Também conhecida por ***ANACOR***, é uma **técnica de análise bivariada** por meio da qual é estudada a **associação entre duas variáveis categóricas e entre suas categoria**s, bem como a **intensidade dessa associação**, a partir de uma tabela cruzada de dados (Tabela de Contingência), em que são dispostas em cada célula as frequências absolutas observadas para cada par de categorias das duas variáveis.

Esta anállise pode ser dividida em 2 grandes partes:

i. Análise de significância estatística da assosciação entre as variáveis e suas categorias por meio do teste Qui-Quadrado.
ii. Elaboração do Mapa Perceptual.

1. Analise de significância estatística Qui quad

i. Criação da Tabela de Contingência: é uma tabela de contagem cruzada (cross-tabulation). Com as frequências absolutas observadas para cada par de categorias das variáveis. It's a 2 dimensional table with groups of variables on the rows and columns. Each cell in the table represents the counts associated to that attribute. Example of contingency table:

Contingency Table

Brands	Attributes		
	Tasty	Aesthetic	Economic
Butterbeer	5	7	2
Squishee	18	46	20
Slurm	19	29	39
Fizzy Lifting Drink	12	40	49
Brawndo	3	7	**16**

ii. Cálculo da Frequência absoluta esperada: also known as expected proportions(E) would be what we expect to see in each cell's proportion, assuming that there is no relationship between rows and columns. Our expected value for a cell would be the row mass of that cell multiplied by the column mass of that cell.

iii. Cálculo da tabela de resíduos: quantifies the difference between the observed data and the data we would expect - assuming there is no relationship between the row and column categories. A positive residual shows us that the count for that object attribute pairing is much higher than expected, suggesting a strong relationship; correspondingly, a negative residual shows a lower value than expected, suggesting a weaker relationship.

iv. Tabela com valores Qui-Quadrado e comparação com o p-valor: a estatística Qui-Quadrado corresponde à somatória, para todas as células, dos valores correspondentes à razão entre o resíduo ao quadrado e a frequência esperada em cada célula. Sendo assim, para dado número de graus de liberdade e determinado nível de significância, se o valor total da estatística Qui-Quadrado for maior que seu valor crítico, poderemos afirmar que existe associação estatisticamente significante entre as duas variáveis categóricas, ou seja, a distribuição das frequências das categorias de uma variável segundo as categorias da outra não será aleatória, e, portanto, haverá um padrão de dependência entre essas variáveis. Podemos, portanto, definir as hipóteses nula e alternativa como:

H0: as duas variáveis categóricas se associam de forma aleatória.
H1: a associação entre as duas variáveis categóricas não se dá de forma aleatória.

ATENÇÃO:
Se nesta etapa H0 for aceita a análise termina aqui. Lembre-se de perguntar o por quê do não resultado.

É importante mencionar que a estatística Qui-Quadrado aumenta à medida que cresce o tamanho da amostra (N), o que pode prejudicar a análise da associação existente em tabelas de contingência. Para que tal problema seja superado, segundo Beh (2004), a

Conceitos

1. **Teste Qui-Quad**
   Fórmula para geração da tabela:
   - $\chi^2 = \frac{(resíduo)^2}{(freq.\ absoluta\ esperada)}$
2. **Tabela de Contingência**
   A tabela de contingência também é chamada de:
   - Tabela de correspondência,
   - Tabela de classificação cruzada;
   - Cross-tabulation

   Quando se refere a ACM:
   A **correspondência múltipla** pode ser utilizada, por se tratar de uma técnica multivariada que possibilita a investigação da existência de associação entre mais de duas variáveis categóricas.
3. **Row and column masses**
   A row or column mass is the proportion of values for that row/column.
4. **Nota**
   A AC se encaixa junto com Cluster e PCA. Rever uso pca e cluster

análise de correspondência faz uso da inércia principal total de uma tabela de contingência para descrever o nível de associação entre duas variáveis categóricas, expressa por:

$$I_T = \frac{\chi^2}{N}$$

Ainda segundo Beh (2004), a decomposição da inércia principal total de uma tabela de contingência pode auxiliar o pesquisador na identificação de fontes importantes de informação que possam ajudar a descrever a associação entre duas variáveis categóricas e, como consequência, propiciar a construção de mapas perceptuais. O **tipo mais comum de decomposição inercial** corresponde à **determinação de autovalores**.

Enquanto o teste **Qui-Quadrado** permite avaliar se a distribuição das frequências das categorias de uma variável segundo as categorias da outra é aleatória ou se há um padrão de dependência entre as duas, a **análise dos resíduos padronizados ajustados**, revela os padrões característicos de cada categoria de uma variável segundo o excesso ou a falta de ocorrências de sua combinação com cada categoria da outra variável.

IMPORTANTE:

Para a célula referente às categorias 1 das duas variáveis, o valor da estatística χ2 é

$$\chi 2 = \frac{(\text{resíduo}_{11})^2}{(\text{freq. absoluta esperada}_{11})}$$

- *O mesmo cálculo é realizado para cada par de categorias e, em seguida, os valores de todas as células são somados*
- *O objetivo é verificar se há associação estatisticamente significante entre as variáveis (utilizando a soma do χ2)*

H0 : as variáveis se associam de forma aleatória.
H1 : a associação entre as variáveis não se dá de forma aleatória.

- *Dados o nível de significância e os graus de liberdade, se o valor da estatística χ 2 for maior do que seu valor crítico, há associação significante entre as duas variáveis (H1)*
- *Graus de liberdade = (I – 1) x (J – 1)*

v. Cálculo dos resíduos padronizados e dos resíduos padronizados ajustados:
   a) Resíduo padronizado: ajuda a responder como as categorias de uma variável se relacionam com as categorias da outra variável, devido ao excesso ou falta de ocorrência de casos nas categorias das duas variáveis. Seu cálculo é feito por célula e para linha 1 x coluna 1 sua fórmula é:

$$\text{Resíduo}_{padronizado} = \frac{\text{resíduo}_{11}}{\sqrt{(\text{freq. absoluta esperada}_{11})}}$$

   b) Resíduo padronizado ajustado: se o valor do resíduo padronizado ajustado for maior do que 1,96, existe associação significativa, ao nível de significância de 5%, entre as duas categorias que interagem na célula; se for menor do que 1,96, não há associação estatisticamente significativa. Note que 1,96 é o valor crítico tabelado da normal padrão para o nível de significância de 5%. Se adotada outra significância esse valor se modifica de acordo com o tabelado.

$$\text{Resíduo}_{padronizado\ ajustado} = \frac{\text{resíduo padronizado}_{11}}{\sqrt{[(1-\frac{\Sigma_{C1}}{N}) \times (1-\frac{\Sigma_{L1}}{N})]}}$$

Exemplo (comentários sobre o exercicio feito em aula):

No exercicio em aula(ver no Livro) perfil conservador está associado significativamente com o investimento em poupança. Ou seja pessoas com perfil investidor conservador tendem a investir em poupança.

Mesmo que os totais sejam diferentes entre si, em termos absolutos, existem sim mais investidores agressivos mas porém ao olharmos nas colunas investidor percentualmente o maior valor investe em ações

2. Elaboração do mapa perceptual

O mapa perceptual é um grafico de dispersão onde as categorias são os pontos e pontos proximos sao mais correlacionados (alerta ver ao final interpretação label row vs label column).

i. Determinar os autovalores (λ^2): a determinação de autovalores é chamada de Decomposição inercial. A quantidade (m) de autovalores depende da quantidade de categorias nas variáveis, sendo esta o mínimo entre mín(I – 1, J – 1). Na Anacor, os **autovalores** referem-se às **inércias principais parciais e são base para determinar a inércia principal total e o percentual da inércia principal total** em cada dimensão do mapa perceptual.

$$\frac{(\text{Resíduo}_{padronizado})}{\sqrt{N}}$$

Fazendo o calculo acima para cada resíduo obtem-se a matriz A. E com base na matriz A obtem-se W = A'A. Da qual podem ser calculados os autovalores (λ^2) da decomposição inercial, por meio da solução da seguinte equação:

$$\det(\lambda^2 \cdot \mathbf{I} - \mathbf{W}) = \begin{vmatrix} \lambda^2 - w_{11} & -w_{12} & -w_{13} \\ -w_{21} & \lambda^2 - w_{22} & -w_{23} \\ -w_{31} & -w_{32} & \lambda^2 - w_{33} \end{vmatrix} = 0$$

Em outras palavras, a decomposição inercial em determinada tabela de contingência, representada pelas diferenças entre as frequências absolutas observadas e esperadas, pode ser decomposta em m componentes, que se referem aos valores das inércias principais parciais de cada dimensão e que nada mais são que o quadrado dos valores singulares λk de cada dimensão.

Inércia principal total:

$$\%\ da\ Inércia\ Principal\ Total\ (para\ cada\ dimensão) = \frac{\lambda^2_{dimensão}}{\lambda^2_{total}}$$

$$Inércia\ Principal\ Total = \frac{\chi^2}{N}$$

Como a análise de correspondência tem, como um de seus principais objetivos, propiciar ao pesquisador a construção de mapas perceptuais que mostram a relação entre as categorias das variáveis dispostas em linha e em coluna na tabela de contingência, cada componente da **inércia principal total será utilizado** para que se **identifique como determinada linha ou coluna contribui para a construção de cada eixo** (dimensão) do referido mapa.

ii. Determinação as massas em linha e coluna

As **massas** representam a **influência ou preponderância, que cada categoria exerce sobre as demais** categorias de sua variável, seja na coluna (column profiles) ou linha (row profiles).

As massas são os percentuais das margens da tabela de contingencia. Ou seja o percentual de cada uma das categorias em relação ao todo. São geradas duas matrizes (tabelas) uma para cálculo de linha e uma para cálculo de coluna. Com essas massas faz-se a massa total de coluna e linha.

Com base nos valores das massas médias em linha e em coluna, podemos definir duas matrizes diagonais, DL e DC, que contêm, respectivamente, esses valores em suas diagonais principais e 0 nos outros valores. DL e DC sao posicionadas em tabelas diferentes.

iii. Determinação dos autovetores:

Para a matriz W, é possível encontrar os autovetores a partir dos autovalores (λ^2) , substituindo os autovalores de cada dimensão na matriz definida como $\det(\lambda^2 . I - W) = 0$ e resolvendo o sistema de equações que parte dela, é possível encontrar os autovetores da coluna (V) e, com base neles, encontrar os autovetores da linha (U)

$$u'_k = \underbrace{[\,D_l^{-1/2} \,.\, (P - lc') \,.\, D_c^{-1/2}]}_{\text{Trata-se da matriz A}} \,.\, v'_k \,.\, \lambda_k^{-1}$$

Cálculo do autovetor resulta em V1 = eixo x e V2= eixo y, autovetor U1 = variavel em linha para dimensao 1 (eixo x), U2 = variavel em linha para dimensao 2 (eixo y).

iv. Definição das coordenadas (scores) das categorias no mapa perceptual

- **Variável em linha na tabela de contingência:**
  - Coordenadas da primeira dimensão (abcissas):

$$\mathbf{X}_l = \begin{pmatrix} \mathbf{x}_{l1} \\ \vdots \\ \mathbf{x}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_1} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_1 \qquad (11.27)$$

  - Coordenadas da segunda dimensão (ordenadas):

$$\mathbf{Y}_l = \begin{pmatrix} \mathbf{y}_{l1} \\ \vdots \\ \mathbf{y}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_2} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_2 \qquad (11.28)$$

  - Coordenadas da k-ésima dimensão:

$$\mathbf{Z}_l = \begin{pmatrix} \mathbf{z}_{l1} \\ \vdots \\ \mathbf{z}_{lI} \end{pmatrix} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_k} \cdot \mathbf{D}_l^{-1/2} \cdot \mathbf{u}_k \qquad (11.29)$$

- **Variável em coluna na tabela de contingência:**
  - Coordenadas da primeira dimensão (abcissas):

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{x}_{c1} \\ \vdots \\ \mathbf{x}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_1} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_1 \qquad (11.30)$$

  - Coordenadas da segunda dimensão (ordenadas):

$$\mathbf{Y}_c = \begin{pmatrix} \mathbf{y}_{c1} \\ \vdots \\ \mathbf{y}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_2} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_2 \qquad (11.31)$$

  - Coordenadas da k-ésima dimensão:

$$\mathbf{Z}_c = \begin{pmatrix} \mathbf{z}_{c1} \\ \vdots \\ \mathbf{z}_{cJ} \end{pmatrix} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda} = \sqrt{\lambda_k} \cdot \mathbf{D}_c^{-1/2} \cdot \mathbf{v}_k \qquad (11.32)$$

É importante ressaltar que as coordenadas da variável em linha também podem ser obtidas por meio das coordenadas da variável em coluna e vice-versa. Assim, caso o pesquisador tenha apenas as coordenadas das categorias de uma das variáveis, porém possua as massas de cada uma das categorias da outra, além dos valores singulares, poderá calcular as coordenadas das categorias desta última variável.

Conforme comentam Fávero et al. (2009), as coordenadas das categorias da variável em linha para uma específica dimensão podem ser obtidas multiplicando-se a matriz de massas (row profiles) pelo vetor de coordenadas das categorias da variável em coluna e dividindo-se os valores obtidos pelo valor singular daquela determinada dimensão. Analogamente, as coordenadas das categorias da variável em coluna, também para dada dimensão, podem ser obtidas multiplicando-se a matriz de massas (column profiles) pelo vetor de coordenadas das categorias da variável em linha e dividindo-se também os valores obtidos pelo valor singular daquela dimensão.

As coordenadas X e Y obtidas por meio das expressões (11.27) a (11.32) são utilizadas para construir um mapa perceptual conhecido como mapa simétrico, em que os pontos que representam as linhas e colunas das categorias das variáveis possuem a mesma escala, também conhecida por normalização simétrica. Caso o pesquisador deseje, por outro lado, privilegiar exclusivamente a visualização das massas em linha ou das massas em coluna de determinada tabela de contingência para a construção do mapa perceptual, poderá abrir mão da normalização simétrica e optar, respectivamente, por aquelas conhecidas como principal linha e principal coluna. Nesses casos, o

cálculo das coordenadas é elaborado por expressões apresentadas no Quadro 11.1.

Quadro 11.1 Expressões para determinação das abcissas e ordenadas em mapas perceptuais.

Normalização	Expressão para as Abcissas	Expressão para as Ordenadas
Simétrica	$\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda}$	$\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda}$
Principal Linha	$\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U}) \cdot \mathbf{\Lambda}$	$\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V})$
Principal Coluna	$\mathbf{X} = \mathbf{D}_l^{-1} \cdot (\mathbf{D}_l^{1/2} \cdot \mathbf{U})$	$\mathbf{Y} = \mathbf{D}_c^{-1} \cdot (\mathbf{D}_c^{1/2} \cdot \mathbf{V}) \cdot \mathbf{\Lambda}$

Enquanto, no perfil linha, apenas o cálculo das abcissas leva em consideração a matriz de valores singulares, no perfil coluna, essa matriz é utilizada apenas para o cálculo das ordenadas. Com base na determinação das coordenadas de cada categoria, pode ser construído um mapa perceptual com m dimensões. Embora essa possibilidade seja matematicamente possível, apenas as duas primeiras dimensões (m = 2) são geralmente utilizadas para a elaboração da análise gráfica, o que gera um mapa perceptual conhecido por biplot.

Step by Step

I. *Criação da Tabela de Contingência:*
II. *Cálculo da frequência absoluta esperada.*
III. *Cálculo da tabela de resíduos freq abs obs- freq abs esperado*
IV. *Tabela com valores Qui-Quadrado e comparação com o p-valor*
V. *Cálculo dos resíduos padronizados e dos resíduos padronizados ajustados*
VI. *Determinar autovalores*
VII. *Determinação as massas em linha e coluna*
VIII. *Definição das coordenadas (scores) das categorias no mapa perceptual*

Análise de Correspondência Múltipla (ACM)

Na ACM só são usadas as variáveis que se correlacionam ao menos com uma outra variável se não houver correlação a variável é excluída. Ou seja, só pode ser usada quando existe associação significativa estatisticamente entre as variáveis. Nesse sentido, é recomendável que seja elaborado um teste Qui-Quarado para cada par de variáveis antes da elaboração de uma análise de correspondência múltipla. Se uma delas não apresentar associação estatisticamente significante a nenhuma das demais variáveis, a determinado nível de significância, recomenda-se que seja excluída da análise de correspondência múltipla.

A análise de correlação simples se baseia na tabela de contingência, na múltipla devido ao grande número de diferentes tabelas de contingência é necessário agregar, a grande quantidade de tabelas de contingência geradas, por um dos 2 métodos seguintes:

i) Matriz binaria
ii) Matriz de Burt

A análise de correspondência múltipla, também conhecida como ACM, é uma técnica de análise multivariada que representa uma extensão natural da análise de correspondência simples (ANACOR), uma vez que permite que sejam estudadas as associações entre mais de duas variáveis categóricas e entre suas categorias, bem como a intensidade dessas associações.

Ao contrário da ANACOR, técnica de análise bivariada, não é possível verificar a existência de associações entre mais de duas variáveis simultaneamente para a elaboração da análise de correspondência múltipla, visto que a estatística do teste Qui-Quarado é calculada apenas com base em uma tabela de contingência bidimensional. Isso não impede, por outro lado, que, em função das massas das categorias de cada uma das variáveis a serem inseridas na análise de correspondência múltipla, sejam calculados autovalores utilizados para que se definam as coordenadas daquelas categorias em um mapa perceptual. Portanto, a lógica da análise de correspondência múltipla é semelhante à estudada para a análise de correspondência simples.

Matriz binaria: Para que seja elaborada a análise de correspondência múltipla, é necessário apresentar o conceito de matriz binária. Imaginemos um banco de dados com N observações e Q variáveis (Q > 2), e que cada variável q (q = 1, ... , Q) possua Jq categorias. Logo, o número total de categorias envolvidas em uma análise de correspondência múltipla é:

$$J = \sum_{q=1}^{Q} J_q$$

Através da matriz binária pode ser definida a inércia principal total da análise de correspondência múltipla, cujo cálculo é bastante simples e depende apenas da quantidade total de variáveis inseridas na análise e do número de categorias de cada uma delas, não dependendo das frequências absolutas das categorias.

Conforme discute Greenacre (2008), a matriz binária Z é composta por matrizes Zq agrupadas lateralmente, uma para cada variável q. Como cada matriz Zq apresenta somente um valor 1 em cada linha, todos os perfis linha se situam nos vértices de um sistema de coordenadas, e, portanto, estamos diante de um exemplo de matriz em que ocorrem as maiores associações possíveis entre linhas e colunas. Como consequência, para cada matriz Zq, a inércia principal parcial da dimensão principal será sempre igual a 1, e a inércia principal total, igual a Jq - 1. Dessa forma, a inércia principal total de Z corresponde à média das inércias principais totais das matrizes Zq que a compõem, ou seja, pode ser obtida por meio da seguinte expressão:

$$I_T = \frac{\sum_{q=1}^{Q} (J_q - 1)}{Q} = \frac{J - Q}{Q}$$

Por meio do método da codificação binária, pode-se supor que a matriz Z seja uma tabela de contingência de uma análise de correspondência simples, a partir da qual podem ser definidos os valores das inércias principais parciais de cada uma das J - Q dimensões.

Os autovalores e autovetores calculados a partir da matriz binária Z (considerada uma tabela de contingência de uma ANACOR), são usados para a definição das coordenadas de cada uma das categorias das variáveis inseridas na análise de correspondência múltipla, o que permite que seja construído o mapa perceptual. As coordenadas geradas por meio do método da matriz binária são conhecidas como coordenadas-padrão.

Matriz de Burt: é a matriz gerada pela análise de correspondência múltipla elaborada por de tabelas de contingência combinadas, em uma única matriz,com os cruzamentos de todos os pares de variáveis. É uma matriz, quadrada e simétrica.

Considerando a matriz de Burt (B) uma tabela de contingência, podemos também elaborar uma análise de correspondência simples, da qual se pode verificar que as coordenadas das categorias das variáveis corresponderão às coordenadas-padrão geradas por meio do método da matriz binária Z, porém com valores em escala reduzida. Esse fato, segundo discute Greenacre (2008), faz os mapas perceptuais construídos a partir das coordenadas geradas pelo método da matriz de Burt serem mais reduzidos e com pontos mais concentrados em torno da Origem, o que, em alguns casos, pode prejudicar a análise visual das associações entre as categorias, embora isso não afete o estudo da relação entre as variáveis. As coordenadas geradas por meio do método da matriz de Burt são conhecidas por coordenadas principais, e a relação entre essas coordenadas principais e as coordenadas-padrão obtidas pelo método da matriz binária é dada pela seguinte expressão:

$$(\text{coord. principal}_{\text{dim.}k})_{\mathbf{B}} = \hat{\lambda}_k \cdot (\text{coord. padrão}_{\text{dim.}k})_{\mathbf{Z}}$$

ou seja, as coordenadas principais de determinada dimensão são as coordenadas-padrão multiplicadas pela raiz quadrada da inércia principal parcial daquela dimensão. Como as inércias principais parciais são menores que 1, explica-se a redução de escala do mapa perceptual construído a partir do método da matriz de Burt.

Interpreação de resultados

Correspondence analysis places the row labels on the plot such that the closer two rows are to each other, the more similar their residuals. This also applies to the column labels. Most had the ***wrong*** assumption that the greater the proximity between a row label and a column label, then the higher the residual and association.

The way that correspondence analysis works means that we can compare between row labels based on distances. We can also compare between column labels based on distances. However, ***if we want to compare a row label to a column label, we need to follow the rules from i to iii***:

i. Look at the length of the line connecting the row label to the origin. Longer lines indicate that the row label is highly associated with some of the column labels (i.e., it has at least one high residual).

ii. Look at the length of the label connecting the column label to the origin. Longer lines again indicate a high association between the column label and one or more row labels.
iii. Look at the angle formed between these two lines. Really small angles indicate association. 90 degree angles indicate no relationship. Angles near 180 degrees indicate negative associations.

Resumo

I. *Só participam da ACM as variáveis que apresentam associação estatisticamente significativa com pelo menos uma outra variável contida na análise*
II. *Antes de elaborar a ACM, é importante realizar um teste χ 2 para cada par de variáveis*
III. *Se alguma delas não apresentar associação com outras, não é incluída na análise de correspondência Análise de Correspondência Múltipla*
IV. *A matriz binária é obtida pela transformação das variáveis qualitativas em variáveis binárias, ou seja, valores 0 ou 1*
    1) *Com base na matriz binária (Z), pode ser obtida a inércia principal total na ACM*
    2) *Supondo que a matriz binária (Z) seja semelhante a uma tabela de contingência da Anacor, é possível obter a inércia principal parcial das dimensões, autovalores, autovetores e coordenadas dessa matriz*
V. *A matriz de Burt é definida como: $B = Z' \cdot Z$*
    1) *Com o uso da Mateiz de Burt, possível combinar em uma única matriz as tabelas de contingência com o cruzamento de todos os pares variáveis*
    2) *Ao considerar a matriz de Burt uma tabela de contingência, é possível realizar uma ANACOR e obter as coordenadas das categorias*

BIBLIOGRAFIA:

Luiz Paulo Fávero and Belfiore, P. (2017). Manual de Análise de Dados. Elsevier Brasil.

Artigo 1:
https://www.qualtrics.com/eng/correspondence-analysis-what-is-it-and-how-can-i-use-it-to-measure-my-brand-part-1-of-2/>

Artigo 2:
https://www.displayr.com/how-correspondence-analysis-works/