Análise de Correspondência para Dados Longitudinais sobre Atitudes

Laura Leal Nunes[1], Bruno Caetano Vidigal[1], Márcio Luís Moreira de Souza[1] e Ronaldo Rocha Bastos[1]
[1](Departamento de Estatística/ICE/UFJF)

1. Introdução

A Análise de Correspondência (AC) é uma das diversas técnicas de análise multivariada desenvolvida para o estudo da relação entre variáveis qualitativas em tabelas de contingência (BEH, 2004). Ela permite a visualização gráfica das linhas e colunas como pontos em espaços vetoriais de dimensões reduzidas em um novo sistema de eixos ortogonais.

A AC tem se tornado muito comum devido a sua fácil implementação. Necessita apenas de uma tabela com números positivos que representam freqüências observadas de objetos ou indivíduos classificados por uma categoria de linha e uma categoria de coluna; tais categorias devem ser mutuamente exclusivas e exaustivas, ou seja, um indivíduo ou objeto não pode ser classificado em mais de uma categoria de uma mesma variável e devem existir categorias suficientes para esse ser classificado. (GREENACRE, 1984)

2. Metodologia

A AC teve sua origem atribuída a trabalhos desenvolvidos por Karl Pearson, com destaque para a estatística Qui-Quadrado, que permite avaliar se as variáveis dispostas numa tabela bidimensional se afastam do pressuposto de independência a ponto de fornecer evidência estatística suficiente para indicar associação.

A estatística Qui-Quadrado se limita apenas a verificar se há ou não independência entre as variáveis, mas não mostra quais categorias de linha *i* e coluna *j* estão associadas, caso se rejeite a hipótese nula de independência com determinado nível de significância (α).

Assim, a AC surgiu para complementar a análise através da estatística Qui-Quadrado de Pearson, buscando revelar a estrutura de relação entre variáveis categóricas (nominais ou ordinais) principalmente através de representação gráfica. As categorias das variáveis em estudo assumem posições no gráfico que possibilitam a comparação das associações entre elas e similaridade entre categorias de uma mesma variável.

Além da saída gráfica, a aplicação da AC indica qual a porcentagem de associação entre as linhas e colunas da tabela de contingência por fatores ortogonais que explicam a variabilidade total, sendo que o primeiro eixo (fator 1) é responsável por discriminar a maior parte da associação existente.

A AC trabalha com perfis de linha ou coluna, vetor formado por todas as freqüências relativas dos valores observados em cada linha ou coluna em relação aos totais marginais, e não com as freqüências absolutas na análise, uma vez que linhas proporcionais apesar de apresentarem o mesmo perfil podem possuir valores absolutos diferentes; logo a análise feita a partir de valores absolutos levará a uma diferenciação de elementos que essencialmente são iguais.

O perfil de uma linha *i* pode ser interpretado como a descrição da relação entre esta linha com todos os elementos de *J* (coluna). O perfil de uma coluna *j* tem interpretação análoga.

A AC permite incluir variáveis suplementares, ou seja, aquelas que estão na tabela original, mas não participam da solução, apesar de figurarem no gráfico resultante. Tais variáveis podem ser aquelas que possuem maior massa, quociente do total da *i*-ésima linha ou *j*-ésima coluna pelo total geral (Benzécri, 1992), pois o centro de gravidade dos dados no espaço multidimensional está sujeito a maior influência desses pontos.

A ideia de centro de gravidade é a generalização espacial da noção de média ponderada em que cada ponto é representado de forma proporcional à sua massa.

Uma nuvem de pontos é a representação dos perfis em espaços vetoriais de n dimensões, sendo n=min (*I-1, J-1*). (BENZÉCRI, 1992)

A nuvem dos perfis de *I* estará representada em um espaço de dimensão igual ao número de elementos de *J-1*, já que há uma dependência linear entre as coordenadas, ou seja, os elementos de cada perfil somam 1. Raciocínio análogo para a nuvem dos perfis de *J*.

A inércia de uma nuvem de pontos em relação ao seu centro de gravidade é uma medida de variação total, que objetiva captar a dispersão dos pontos da nuvem em torno do centro de gravidade. A inércia total pode ser calculada como o quociente da estatística Qui-Quadrado de Pearson em relação ao total da tabela.

O objetivo da AC é encontrar o subespaço ótimo de menor dimensão que melhor se ajusta aos pontos da nuvem. (BENZÉCRI, 1992) Daí, esse subespaço ótimo é formado pelos eixos fatoriais que passam pelo centro de gravidade e que minimizam as distâncias dos pontos até ele. Primeiramente encontra-se um eixo que passa pelo centro de gravidade e que minimiza a distância perpendicular dos pontos a reta (primeira dimensão). Logo após, traça-se uma segunda reta ortogonal a primeira que também passa pelo centro de gravidade que minimize as distâncias dos pontos em relação a esta reta. Esse processo ocorrerá até se encontrar min (*I-1, J-1*) dimensões.

Uma forma de se obter as coordenadas dos perfis de linha e coluna no novo sistema de eixo fatorial é através da decomposição do valor singular da matriz $A = UD_\mu V^T$ onde U contém os vetores singulares a esquerda de A, V os valores singulares a direita de A e D_μ é uma matriz diagonal de números positivos, valores singulares, em ordem decrescente.

Dessa forma as coordenadas principais dos perfis de linha e coluna são dadas por $F = D_r^{-1/2} U D_\mu$ e $G = D_c^{-1/2} V D_\mu$, respectivamente, onde D_r é a matriz diagonal do perfil da coluna marginal e D_c é a matriz diagonal do perfil da linha marginal.

A projeção conjunta da nuvem de pontos dos perfis de linha e coluna, onde as coordenadas utilizadas são as principais (simétrica), pode ser interpretada como se as soluções gráficas de cada uma fossem sobrepostas. Isto significa que as distâncias entre as projeções dos perfis não são interpretáveis. Apenas as direções, ou tendências, podem ser analisadas. (NYFJALL, 2002)

Como o objetivo da AC é a redução de dimensionalidade em um subespaço ótimo, faz-se necessário calcular e interpretar a contribuição de cada dimensão (fator) para a inércia total já comentada acima.

A contribuição relativa do eixo α para a inércia do ponto i, calculada por $COR_\alpha(i) = \frac{f_i(F_\alpha(i))^2}{I_{f_J}(f_J^i,\ f_i)}$, onde $F_\alpha(i)$ é a coordenada de i no eixo α, permite conhecer os fatores que explicam a posição de f_J^i relativa ao centro da nuvem.

Através da soma das contribuições relativas, se obtém o que é conhecido como a qualidade de representação de um ponto em um subespaço, denominado QLT, onde quanto mais próximo de 1 o valor estiver, melhor será a representação do ponto.

$$QLT = COR_{\alpha_1}(i) + COR_{\alpha_2}(i)$$

A contribuição absoluta de cada elemento *i* para a inércia de certa dimensão α (λ_α) é dada por $CTA_\alpha(i) = f_i(F_\alpha(i))^2$ e a partir dela tem-se a contribuição relativa do ponto *i* para a inércia de um eixo fatorial, definida por $CTR_\alpha(i) = \frac{CTA_\alpha(i)}{\lambda_\alpha}$. (BENZÉCRI, 1992)

Assim como a AC, a Análise de Correspondência Múltipla (ACM) é também uma técnica de análise exploratória de dados, só que aplicada a tabelas multidimensionais.

Existem duas maneiras de se desenvolver a ACM. Uma delas é a partir da matriz indicadora, onde as linhas correspondem a cada indivíduo ou objeto de análise e as colunas são as variáveis com suas categorias. Essa matriz possui os elementos na forma de variáveis *dummy*, ou seja, quando for 1, o objeto *i* é classificado na categoria *j* da variável *q*, e 0, caso contrário. O objeto *i* só pode pertencer a uma categoria para cada variável. A outra possível abordagem da ACM é utilizando a matriz de Burt, uma matriz simétrica que contém todos os cruzamentos possíveis 2x2 entre as *Q* variáveis envolvidas, sendo que a diagonal principal é formada de matrizes diagonais de freqüências marginais de coluna.

A matriz de Burt é calculada a partir do produto da transposta da matriz indicadora por ela mesma. Uma vantagem do uso da matriz de Burt é reduzir o problema de células vazias quando se tem muitas categorias e variáveis, justamente pelo fato desta trabalhar com freqüências marginais.

3. Aplicação

Nesse trabalho foi utilizado uma base de dados da BHPS (*British Household Panel Survey*), que tem o objetivo de fazer pesquisas do tipo painel (mesmas e/ou diferentes variáveis são medidas para os mesmos indivíduos, em pelo menos dois pontos do tempo) com indivíduos em famílias da Grã-Bretanha com perguntas sócio-demográficas, econômicas e de atitudes, entre outras. Foi aplicado questionário a mulheres que tinham entre 16 e 39 anos no ano de 1991 e este foi replicado nos anos subseqüente de 1993, 1995, 1997 e 1999, perfazendo um total de 1340 mulheres. A técnica de amostragem aplicada foi a estratificada em múltiplos estágios, onde todas as residências tinham aproximadamente a mesma probabilidade de inclusão na amostra.

O questionário consistia de perguntas em escala tipo Likert, onde 1 significava concordo totalmente, 2, concordo, 3, nem concordo nem discordo, 4, discordo e 5, discordo totalmente, sobre nove afirmações a respeito da família, o papel das mulheres e o trabalho fora do lar. Os escores das variáveis (C), (D) e (E) foram invertidos para garantir que escores mais elevados caracterizassem as mulheres liberais.

Para efeito de análise, foram retiradas do estudo as variáveis (G), (H) e (I) (ver Berrington, 2001).

As variáveis foram descritas abaixo.

- A: Criança em idade pré-escolar pode sofrer se sua mãe trabalha fora de casa.
- B: Todos na família sofrem quando a mãe trabalha em tempo integral.
- C: A mulher e a família seriam mais felizes se ela trabalhasse fora.
- D: O marido e a mulher deveriam contribuir para as despesas da casa.
- E: Ter emprego integral é a melhor forma para a mulher ser independente.
- F: A função do homem é ganhar dinheiro, e a da mulher cuidar da casa e da família.
- G: As crianças precisam que os pais estejam tão envolvidos quanto as mães na questão da educação.
- H: Os empregadores deveriam fazer um horário especial para que as mulheres conseguissem conciliar o trabalho e o cuidado com as crianças.
- I: Solteiros podem ser tão bons em cuidar das crianças quanto os casados.

Esse estudo contempla uma série de técnicas desenvolvidas no software livre R, versão 2.11.0, disponível gratuitamente em www.r-project.org (R Development Core Team, 2010), através da análise da matriz super indicadora Z^q, tridimensional com os elementos z_{ijt}^q, onde *q* indica a variável, *i* os objetos, *j* as categorias da variável *q* e *t* o momento da observação (SAPORTA e Nyiang, 2006). O problema é reduzir essa matriz para duas dimensões de forma a aplicar a AC, técnica já comentada acima.

A ACM capta a estrutura através da concatenação das matrizes, formando duas formas básicas, BROAD (as linhas representam os indivíduos e nas colunas estão as categorias das variáveis, sendo repetidas nos momentos) e LONG (as colunas representam as categorias das variáveis e as linhas são os indivíduos, sendo estes repetidos em cada momento). Nesse estudo será utilizado apenas a forma BROAD.

As possíveis soluções do problema são agregar por uma dimensão e obter matrizes marginais, observar apenas as sub-matrizes $Z^{i(q)}$, $Z^{j(q)}$, $Z^{t(q)}$, concatenar estas em cada momento *t* na forma BROAD ou LONG e transformar a matriz Z^q na matriz de Burt.

Agregar por uma dimensão é obter as matrizes $Z^{IJ(q)}$, com os elementos z_{ij+}^q (agregação no tempo), $Z^{JT(q)}$, com os elementos z_{+jt}^q (agregação dos objetos ou indivíduos) e $Z^{IT(q)}$, com os elementos z_{i+t}^q (agregação de categorias).

Aplicou-se a AC na matriz super indicadora de 1340 linhas por 150 colunas, já que como foi dito anteriormente, são 1340 indivíduos e 6 variáveis com 5 categorias cada. A análise desta matriz mostra uma mudança de estrutura entre as variáveis, ou seja, indicam como a relação entre as variáveis no momento *t* difere daquela no momento *t'* e indicam como a relação entre variáveis no momento *t* com outras variáveis no momento *t'* pode diferir da relação entre estas mesmas variáveis entre os momentos *t''* e *t'''*. Essa mudança de estrutura é revelada através dos *component scores* (produto da matriz diagonal de valores singulares pela matriz à direita da decomposição) e não pelos *object scores*, já que os escores para cada indivíduo ou objeto serão os mesmos.

Agregando pelo tempo tem-se a matriz indicadora $Z^{IJ(q)}{}_{(6700x30)}$ que foi transformada na matriz de Burt, matriz de freqüências marginais, com 30 linhas e 30 colunas, com cada categoria somada no tempo.

A matriz $Z^{JT(q)}{}_{(5x30)}$ mostra as freqüências marginais de cada categoria nos cinco momentos analisados (*waves*), ou seja, ela agrega por indivíduos.

4. Resultados

Após as análises das matrizes observou-se a consistência das perguntas e a escala ordinal adotada confirmada pela disposição dos pontos no gráfico representado na Figura 1.

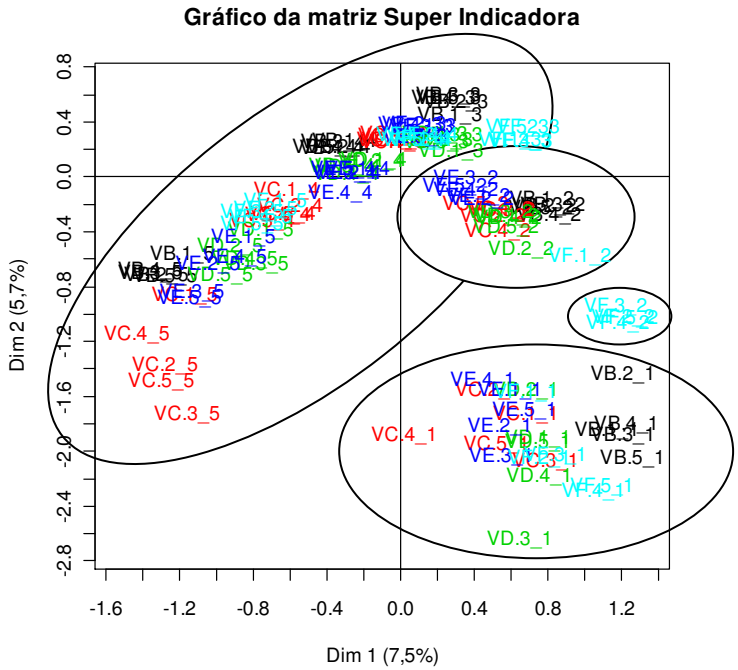

Figura 1. Mapa das quantificações das categorias das variáveis relacionadas à atitude do grupo de mulheres analisada ao longo dos 5 anos não consecutivos.

A partir da análise do gráfico apresentado na figura acima, podemos perceber que, independente do tempo (momentos) e das variáveis que focamos, não houve mudanças drásticas na atitude das mulheres estudadas: aquelas de opiniões mais conservadores se aglutinam no quadrante inferior direito do gráfico (Figura 1), bem como as mais liberais se aglutinam no lado oposto em relação à primeira dimensão (que conserva a maior inércia) do gráfico bidimensional. O primeiro índice refere-se a *wave* e o segundo representa a categoria da variável, por exemplo, a VD.3_1 representa a variável D no momento 3, categoria 1.

Utilizamos os círculos pontilhados a fim de chamar a atenção à formação de pequenos *clusters* que nos possibilitam ter uma análise um tanto quanto substantiva do perfil médio das atitudes femininas ao longo dos cinco momentos de análises. As mulheres de opiniões muito conservadoras (categoria 1) têm perfis um pouco isolados dos demais, ao

passo que as muito liberais e liberais, junto com as aquelas que não tem opinião quanto as afirmativas, estão mais relacionadas. Há também o grupo das conservadoras (categoria 2), que foi dividido em dois grupos,as da variável F, que estão graficamente mais próximas das muito conservadoras, e o grupo das demais mulheres conservadoras, mais próximas do centróide, representada pelas demais variáveis em questão.

5. Agradecimentos

Os autores agradecem a PROPESQ/UFJF pela Bolsa de Iniciação Científica de Laura Leal Nunes e Bruno Caetano Vidigal, e a FAPEMIG, pela Bolsa de apoio técnico de Márcio Luís Moreira de Souza.

6. Referências Bibliográficas

BEH, E. J. Simple correspondence analysis: a bibliographic review. International Statistical . Review, v. 72, n.2, p. 257-284, aug. 2004.

BENZÉCRI, J.P. (1992) *Correspondence Analysis Handbook.* New York: Marcel Dekker.

BERRINGTON, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study . *In* Lesthaeghe R. ed. *Meaning and Choice: Value Orientations and Life Course Decisions*. Brussels: NIDI.

GREENACRE,M. (2007).Correspondence Analysis in Practice, second edition: Boca Raton: Chapman & Hall/CRC.

HEIJDEN, P. G. M. van der. (1987) Correspondence analysis of longitudinal categorical data by. Tese de Doutorado: Universidade de Leiden, Holanda.

SAPORTA, G e Nyiang. N. (2006) *Correspondence Analysis and Classification.* In: Michael Greenacre & Jörg Blasius (eds.) Multiple Correspondence Analysis and Related Methods.

Boca Raton: Chapman & Hall/(CRC).