PAULO CÉSAR OSSANI

PROPOSIÇÃO DE UM NOVO ÍNDICE PARA *PROJECTION PURSUIT* NA ANÁLISE DE MÚLTIPLOS FATORES

LAVRAS - MG

2019

PAULO CÉSAR OSSANI

PROPOSIÇÃO DE UM NOVO ÍNDICE PARA *PROJECTION PURSUIT* NA ANÁLISE DE MÚLTIPLOS FATORES

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística Multivariada, para a obtenção do título de Doutor.

Prof. Dr. Marcelo Ângelo Cirillo (UFLA)
Orientador

LAVRAS - MG
2019

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Ossani, Paulo César.
Proposição de um novo índice para *Projection Pursuit* na análise de múltiplos fatores / Paulo César Ossani. - 2019.
75 p. : il.

Orientador(a): Marcelo Ângelo Cirillo.

Tese (doutorado) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. projection pursuit. 2. índice MF. 3. análise de múltiplos fatores. I. Cirillo, Marcelo Ângelo. II. Título.

PAULO CÉSAR OSSANI

PROPOSIÇÃO DE UM NOVO ÍNDICE PARA *PROJECTION PURSUIT* NA ANÁLISE DE MÚLTIPLOS FATORES

PROPOSITION OF A NEW INDEX FOR PROJECTION PURSUIT IN THE MULTIPLE FACTOR ANALYSIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística Multivariada, para a obtenção do título de Doutor.

APROVADA em 28 de fevereiro de 2019

Eliandro Rodrigues Cirilo	UEL
Evelise Roman Corbalan Góis Freire	UFLA
Júlio Sílvio de Sousa Bueno Filho	UFLA
Izabela Regina Cardoso de Oliveira	UFLA

Prof. Dr. Marcelo Ângelo Cirillo (UFLA)
Orientador

LAVRAS - MG
2019

À minha amada esposa, Brígida, pela paciência e carinho dispensados nas horas difíceis que passei estudando para a apresentação deste trabalho.
Aos meus pais, Paulo e Sebastiana, que sempre me apoiaram nas realizações dos meus sonhos.
E aos meus filhos, Gabriel e Rafael, que sempre afagam o meu coração com muito amor e ternura.

DEDICO

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por estar sempre ao meu lado, me intuindo nas escolhas certas e guiando-me no caminho correto.

Aos meus pais, Paulo e Sebastiana, que sempre me apoiaram e incentivaram na jornada em busca de novas realizações.

À Universidade Federal de Lavras (UFLA), ao Departamento de Estatística e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, pela oportunidade de cursar o doutorado.

Ao orientador, professor Dr. Marcelo Ângelo Cirillo, pela paciência ao orientar-me e, acima de tudo, pelo conhecimento que compartilhou comigo.

A todos os meus amigos de Lavras, em especial aos do Departamento de Estatística que me auxiliaram durante o curso.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (Capes) - Código de financiamento 001, pelo qual agradeço.

A certeza repousa no íntimo daqueles com a consciência tranquila por seus deveres cumpridos.

Ossani

RESUMO

Neste trabalho propõe-se um novo índice para a *projection pursuit*, utilizada na redução da dimensão de grupos de variáveis que são analisados pela técnica análise de múltiplos fatores (MFA). A principal vantagem em relação aos outros índices, está no fato de que, o procedimento metodológico preserva a estrutura de variâncias e covariâncias para a realização da decomposição dos valores singulares, quando é utilizado na comparação dos grupos de variáveis. Entre outras contribuições, o estudo apresenta uma modificação no algoritmo *grand tour simulated annealing*, adaptando-o para lidar com grupos de variáveis. A metodologia envolvida na avaliação da proposta do índice foi feita por simulações Monte Carlo em vários cenários com configurações nos seguintes fatores: graus de correlação entre as variáveis, número de grupos e grau de heterogeneidade entre os grupos de variáveis. Comparações foram feitas do índice proposto com 13 índices conhecidos na literatura. Concluiu-se que o novo índice proposto mostrou-se viável na redução dos dados para aplicação na técnica MFA, sendo recomendado nas situações em que os grupos apresentam baixa ou alta heterogeneidade, e forte grau de correlação entre as variáveis $(\rho = 0{,}9)$. De modo geral, os índices são afetados pelo aumento do número de grupos, em função dos cenários avaliados.

Palavras-Chave: Projection pursuit, índice MF, análise de múltiplos fatores, grupos heterogêneos, simulated annealing.

ABSTRACT

This study proposes a new index for projection pursuit, used to reduce the dimensions of groups of variables using multiple factor analysis. The main advantage with respect to other indexes is that the methodological procedure preserves the variance and covariance structures to perform singular value decomposition, when the index is used to compare groups of variables. Among other contributions, the study presents a modification in the grand tour algorithm with simulated annealing, adapting it to deal with groups of variables. The methodology used to assess the proposed index was based on Monte Carlo simulations, in several scenarios and with configurations of the following factors: degrees of correlation between the variables; and number of groups and degrees of heterogeneity among groups of variables. The proposed index was compared with thirteen indexes known in the literature. It was concluded that the proposed index was efficient in the reduction of data to use multiple factor analysis. This index is recommended for situations in which the groups exhibit low or high heterogeneity and a strong degree of correlation between the variables (ρ = 0.9). In general terms, indexes are affected by the increase in the number of groups, depending on the scenarios assessed.

Keywords: Projection pursuit, MF index, multiple factor analysis, heterogeneous groups, simulated annealing.

.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIMBOLOS

$X_{n\times p}$	Conjunto de observações ou matriz de ordem $n \times p$
n	Tamanho amostral
p	Número de colunas ou variáveis em X
k	Número de grupos de variáveis
d	Número de colunas dos dados com dimensão reduzida
$X_{n\times m}$	Conjunto com k grupos $X_{n\times p}$, sendo $m = kp$
$\tilde{X}_{n\times d}$	Conjunto de dados com dimensões reduzidas sendo $d < p$
$\tilde{X}_{n\times s}$	Conjunto com k grupos $\tilde{X}_{n\times d}$, sendo $s = dk$
$A_{p\times d}$	Matriz de projeção ortonormal
$PI(A)$	Função índice de projeção em relação à matriz de projeção $A_{p\times d}$
Z	Matriz de dados esféricos
Y	Matriz dos dados projetados
$S_{n\times m}$	Matriz com os dados balanceados conforme técnica MFA
$A \cdot B$	Produto entre as matrizes A e B

SUMÁRIO

1 INTRODUÇÃO

Em espaços com muitas dimensões, as amostras se tornam esparsas e pouco similares. Tal fato torna-se um problema quando são utilizados métodos de redução de dimensionalidade fundamentados na decomposição de valores singulares, como, por exemplo, a análise de componentes principais, uma vez que as informações das primeiras componentes são diluídas nas demais.

Diante dessa questão, torna-se recomendável uma redução de dimensões, sem a perda das informações contidas na dimensão de origem. Nesse contexto, surge a *projection pursuit* (PP), ou *busca de projeção*, sugerida por Kruskal (1969), implementada por Friedman e Tukey (1974).

Em síntese, a *projection pursuit* pesquisa projeções lineares de baixa dimensão em estruturas de dados de altas dimensões. Para isso, define-se um índice de projeção, $I(u, v)$, entendido como uma função objetivo, que quantifica o grau de interesse de uma projeção sobre o plano pelos vetores (ortogonais) u e v, e, então, utiliza-se um procedimento de otimização numérica para encontrar o plano que maximize esse índice. Neste contexto, o problema consiste na escolha do índice que melhor represente o grau de interesse da projeção.

O método PP torna-se mal adaptado para lidar com estruturas altamente não lineares e também tem a desvantagem de demandar muito tempo computacional (HUBBER, 1985), mas na atualidade esse tempo vem sendo reduzido consideravelmente com as novas ferramentas de processamento. Contudo, com o crescimento da mineração de dados, ele é cada vez mais empregado na classificação e em agrupamentos, para escapar da maldição da dimensionalidade (LEE et al., 2005).

Em dados bivariados por meio de um diagrama de dispersão, o método PP é uma ferramenta que permite a inspeção visual dos dados que estão em alta dimensão, possibilitando detectar concentrações próximas a curvas ou a linhas, estruturas, *outliers*, *skewness* e agrupamentos, utilizando a capacidade de percepção humana para a descoberta instantânea de padrões; em dados em alta dimensão isso seria muito difícil ou quase impossível, devido à limitação humana.

Friedman e Tukey (1974) citam a análise de componentes principais, a análise de fator e outras técnicas usuais multivariadas que lidam com redução de dimensão como casos particulares do método PP, as quais são tidas como métodos lineares, apresentando as vantagens de interpretabilidade simples e economia computacional. Citam, ainda, que a

desvantagem de muitos métodos lineares clássicos é que a única propriedade do enxame pontual que é utilizada para determinar o mapeamento global é a variância ao longo de várias direções no espaço multidimensional. Mas, o método PP combina propriedades globais e locais de enxames pontuais multivariados para obter mapeamentos lineares úteis, utilizando medidas globais aperfeiçoadas, com a vantagem adicional de robustez contra *outliers*.

É notório que a *projection pursuit* tem sido implementada em diversas aplicações, como a classificação exploratória supervisionada de dados (LEE et al., 2005), a análise de componentes principais robusta (CROUX; FILZMOSER; OLIVEIRA, 2007) e a análise de componentes independentes (HYVARINEN; OJA, 2000). Em se tratando da análise de múltiplos fatores, não há relatos, na literatura, da viabilidade da aplicação dessa projeção e da proposição de novos índices que demandem menor esforço computacional com resultados promissores aos índices existentes na literatura.

Nesse contexto, neste trabalho apresenta-se, como proposta, um novo índice a ser utilizado na *projection pursuit* aplicada à técnica análise de múltiplos fatores (MFA), considerando grupos de variáveis quantitativas.

2 OBJETIVOS

Propor um novo índice e algoritmo computacional a serem utilizados pela técnica *projection pursuit* em grupos de variáveis quantitativas, a fim de reduzir a dimensionalidade dos dados, de modo aplicar a técnica MFA na comparação das similaridades entre os grupos.

3 REFERENCIAL TEÓRICO

3.1 Interpolação entre planos

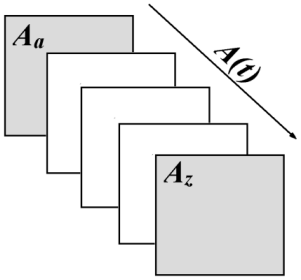
A rotação tridimensional, naturalmente, é feita simplesmente girando-se o objeto de interesse, possibilitando a visualização em todos os ângulos. As visualizações de dados *p*-dimensionais são realizadas de forma semelhante, ou seja, girando todos os dados *p*-dimensionais, ou utilizando projeções em um espaço reduzido $d < p$ dos dados, sendo p a dimensão dos dados originais e d a dimensão dos dados com dimensão reduzida. A exibição de dados *p*-dimensionais corresponde à rotação do plano de projeção *d*-dimensional em vez dos dados (COOK et al., 2008), tendo como referência os dados representados em uma nuvem[1] de pontos fixada.

A rotação do plano de projeção pode ser calculada interpolando-se um plano inicial e um plano alvo. A interpolação entre planos pertence à classe de métodos de visualização de dados multivariados chamada de *grand tours*, os quais são utilizados nas animações das projeções, movendo-se em um plano de projeção bidimensional através do *n*-espaço. A matemática envolvida nesta interpolação pode ser vista em Asimov e Buja (1994) e Hurley e Buja (1990). Rotações planares são discutidas em detalhes em Asimov (1985) e Asimov e Buja (1994), e tecnicamente em Buja et al. (2005).

Segundo Martinez, Martinez e Solka (2010), a ideia por trás da interpolação é um processo que começa em um plano inicial A_a, prosseguindo até o plano alvo A_z através de caminhos geodésicos entre os dois planos.

O quadro (Figura 1) que descreve o plano de partida realiza sequência de planos intermediários. Evitando que os dados girem dentro do plano de visão, esse tipo de rotação é uma distração para quem visualiza, pois se assemelha à visualização de uma cena[2] enquanto está em uma plataforma oscilante. Assim, toda rotação ocorre fora do plano de visualização (COOK et al., 2008).

Figura 1 - Interpolação geodésica entre planos.



Fonte: Do autor (2019).

[1] Refere-se a um conjunto de pontos expresso em um mesmo sistema de coordenadas.
[2] Refere-se a um conjunto de imagens.

Conforme Hurley e Buja (1990), a interpolação geodésica entre pares de planos apresenta a propriedade de suavidade, na qual o caminho geodésico é gerado por rotações no subespaço, abrangidas pelos dois planos. No caso mais simples, dados dois planos A_a e A_z subespaços unidimensionais, caracterizados, respectivamente, pelos vetores unitários a_0 e a_z, a interpolação geodésica é obtida movendo-se um vetor unitário a ao longo do grande círculo que liga a_0 e a_z. Mais precisamente, seja α um ângulo entre a_0 e a_z, e seja a_z^* o vetor unitário obtido por ortogonalização de a_z em relação a a_0 pelo processo de Gram-Schmidt. Dessa forma, o caminho geodésico $A(t)$ de A_a para A_z é dado por

$$a(t) = a_0 cos(t) + a_z^* sen(t), \text{ para } 0 \leq t \leq \alpha.$$

A interpolação geodésica entre dois planos A_a e A_z, é descrita utilizando-se vetores principais. Assim, sejam $a_0 \in A_a$ e $a_z \in A_z$ vetores unitários atingindo o menor ângulo entre A_a e A_z, tendo, ainda, os vetores $b_0 \in A_a$ e $b_z \in A_z$, tal que (a_0, b_0), (a_z, b_z) formam, respectivamente, bases ortonormais para A_a e A_z. Esses quatro vetores são chamados de vetores principais para A_a e A_z, e os ângulos α (entre a_0 e a_z) e β (entre b_0 e b_z) correspondem aos ângulos principais, respectivamente. Pode-se mostrar que a_0 e b_z são ortogonais, e o mesmo para a_z e b_0.

A interpolação geodésica entre A_a e A_z pode ser descrita como uma família de pares de vetores ortonormais $(a(t), \mathrm{b}(t))$, em que $a(t)$ se move de a_0 para a_z ao longo de uma região circular e $\mathrm{b}(t)$ similarmente de b_0 ao b_z. Ambos os vetores se movem em regiões circulares a velocidades constantes (mas, geralmente desiguais), chegando simultaneamente aos seus respectivos alvos, a_z e b_z (HURLEY; BUJA, 1990).

Para um melhor entendimento, a seguir é apresentado um algoritmo para interpolar dois planos, proposto por Cook et al. (2008). Para isso segue que

1: dada uma projeção inicial A_a de ordem $p \times d$, descrevendo o plano inicial, crie uma nova projeção alvo A_z, descrevendo o plano alvo. A projeção pode ser chamada de quadro ortonormal. Um plano pode ser descrito por um número infinito de quadros. Para encontrar a rotação ideal do plano de partida para o plano de destino, é necessário encontrar os quadros em cada plano que são os mais próximos;
2: determine o caminho mais curto entre os quadros utilizando a decomposição em valores singulares $A_a' A_z = V_a \Lambda V_a'$, $\Lambda = \mathrm{diag}(\lambda_1 \geq \cdots \geq \lambda_\mathrm{d})$ e as principais direções em cada plano são $B_a = A_a V_a$, $B_z = A_z V_z$. As direções principais são os quadros que descrevem os planos

de partida e de destino que têm a menor distância entre eles. A rotação é definida com respeito a estas direções principais. Os valores singulares $\lambda_i, i = 1, \cdots, d$, definem os menores ângulos entre as direções principais

3: ortonormalize B_z em B_a, dando B_*, para criar uma estrutura de rotação;

4: calcule os ângulos principais, $\tau_i = \text{acos}(\lambda_i)\,, i = 1, \cdots, d$;

5: girar os quadros dividindo-se os ângulos em incrementos, $\tau_i(t)$, para $t \in (0,1]$, e crie a *i*-ésima coluna do novo quadro, b_i, a partir das *i*-ésimas colunas de B_a e B_*, por $b_i(t) = cos\big(\tau_i(t)\big)b_{ai} + sen\big(\tau_i(t)\big)b_{*i}$. Quando $t = 1$, a moldura será B_z;

6: projete os dados $A(t) = B(t)V_a$;

7: continue a rotação até $t = 1$. Defina a projeção atual como A_a e volte para a etapa 1.

3.2 *Projection pursuit*

A *projection pursuit* (PP), também chamada de *busca de projeção*, é uma técnica para análise exploratória de dados multivariados que pesquisa projeções lineares interessantes de baixa dimensão em dados de alta dimensão. Tais projeções são alcançadas por meio da otimização de uma função objetivo, chamada de *índice de projeção*. É útil para uma análise inicial de dados, especialmente quando os dados estão em um espaço de alta dimensão.

A busca de projeção tem sido utilizada em diversas aplicações, tais como classificação exploratória supervisionada de dados (LEE et al., 2005), análise de componentes principais robusta (CROUX; FILZMOSER; OLIVEIRA, 2007) e análise de componentes independentes (HYVARINEN; OJA, 2000), entre outras.

A ideia básica do método PP, sugerida por Kruskal (1969) e implementada pela primeira vez por Friedman e Tukey (1974), quando o termo *projection pursuit* foi proposto, consiste em definir um índice de projeção $I(u, v)$ que mensura o grau de interesse[3] da projeção sobre o plano, pelos vetores (ortogonais) u e v, e, então, utilizar a otimização numérica para encontrar um plano maximizando o índice.

A questão-chave está na escolha do índice de projeção que melhor represente o grau de interesse da projeção. Assim, em termos práticos, o método PP requer

i) uma função objetivo (índice de projeção), que quantifica o grau de interesse de uma projeção a ser pesquisada;

[3] Refere-se como os dados são projetados em dimensões mais baixas, preservando as características de interesse do pesquisador.

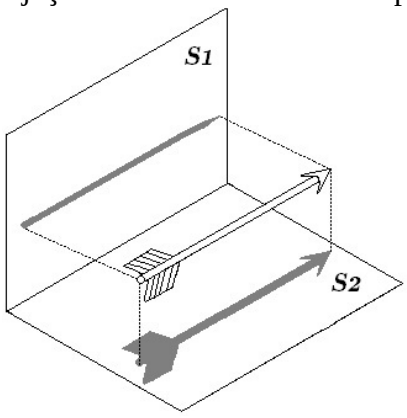
ii) utilizar algum método numérico para otimizar a função objetivo, encontrando bases do espaço de projeção que maximizam a função.

Segundo Hubber (1985), uma característica marcante do método PP é pertencer a um dos poucos métodos multivariados capazes de contornar a maldição da dimensionalidade. Este termo se refere a vários fenômenos que surgem na análise de dados em espaços com muitas dimensões; à medida que o número de variáveis (atributos) cresce, o espaço entre elas vai se tornando vazio e os pontos dos dados tendem a ser equidistantes.

3.2.1 Procedimento para a busca de projeção

De forma lúdica, entende-se que uma projeção pode ser representada por uma sombra[4] de um objeto. Se é uma projeção bidimensional, então, a projeção é a sombra que o objeto lança sob uma luz brilhante, como visto na Figura 2. Se o objeto gira na luz, veem-se muitas sombras bidimensionais diferentes e pode-se inferir a forma do objeto em si (COOK et al., 2008). Na Figura 2 a projeção S_2 mostra mais aspectos relevantes dos dados do que a projeção S_1, o que facilita a identificação do objeto.

Figura 2 - Projeção de uma flecha em dois planos.



Fonte: Do autor (2019).

Matematicamente, a projeção das observações $X_{n\times p}$ é uma transformação linear $Y: \mathbb{R}^p \rightarrow \mathbb{R}^d$, com $p > d$, ou seja, $Y_{n\times d} = X_{n\times p} \cdot A_{p\times d}$, sendo $A_{p\times d}$ uma matriz de projeção ortonormal e, assim, a dimensão que será projetada será d. Por exemplo, para projetar um objeto tridimensional (três colunas ou variáveis) em um plano bidimensional (a sombra do objeto) utiliza-se uma matriz ortonormal $A_{3\times 2}$ (COOK et al., 2008).

O método PP segue o mesmo princípio matemático citado, em que se busca uma transformação linear $Y: \mathbb{R}^p \rightarrow \mathbb{R}^d$, com $p > d$, sendo $Y = XA$, com $A_{p\times d}$. Assim se diz que Y

[4] Refere-se às projeções dos dados em dimensões mais baixas.

é a projeção linear de X, e A é a matriz de projeção, tendo as colunas representando as bases do espaço de projeção.

Convém ressaltar que o interesse é pela busca de projeções lineares com a restrição de ortonormalidade nas bases de projeção, ou seja, $A \cdot A^T = I_p$, sendo I_p a matriz identidade. Essa restrição garante que cada dimensão do espaço de projeção apresente diferentes aspectos dos dados (ESPEZUA et al., 2015; HUBER, 1985; JONES; SIBSON, 1987).

Logo, ao considerar a função índice PI que mede o grau de interesse da projeção em Y, o método PP torna-se um problema de otimização da equação (1) que procura a matriz A que maximize PI (ESPEZUA et al., 2015; JONES; SIBSON, 1987).

$$\tilde{A} = arg\,\max_{A}\{PI(XA)\}, com\ A \cdot A^T = I_p. \quad (1)$$

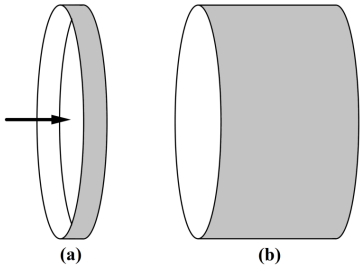
A otimização da equação (1) é feita por meio de métodos numéricos que, tradicionalmente, são fundamentados no gradiente (HUBER, 1985; JONES; SIBSON, 1987) ou na técnica de Newton-Raphson (FRIEDMAN, 1987; FRIEDMAN; TUKEY, 1974), mas não são apropriados para cenários acima de três dimensões (ESPEZUA et al., 2015). Geralmente esses métodos ficam presos a uma bacia de atração, dependente do chute inicial, e isso prejudica encontrar o ponto ótimo.

3.3 Dados esféricos

Segundo Cook et al. (1995), é usual que os dados sejam esféricos antes de iniciar a busca de projeção, removendo a influência de localização e a escala para a busca de projeções estruturadas, sendo necessário para índices que medem a saída da densidade de dados projetados de uma densidade Normal padrão, isso se deve ao fato de que as diferenças de localização e de escala podem dominar as outras diferenças estruturais.

Tornar os dados esféricos muda visivelmente a percepção e a interpretabilidade dos dados. Cook et al. (1995) citam o seguinte exemplo: considere pontos uniformemente distribuídos em um cilindro que tem uma pequena razão entre comprimento e raio, como na Figura 3a. Tornar os dados esféricos é análogo ao aumento do comprimento do tubo, como visto na Figura 3b, resultando em uma cavidade menos visível. Logo, tornar os dados esféricos é alterar graficamente a distribuição dos dados, o que pode, em alguns casos, ocultar recursos que eram vistos anteriormente.

Figura 3 - Efeito visual dos dados esféricos. (a) Antes de os dados tornarem-se esféricos, a cavidade é fácil de ver. (b) Depois de os dados tornarem-se esféricos, a cavidade não tão fácil de ver.



Fonte: Adaptada de Cook et al. (1995).

A transformação necessária para a obtenção de dados esféricos é feita por meio da decomposição espectral na matriz de covariância Σ. Os dados esféricos são obtidos utilizando-se a seguinte equação:

$$Z_i = \Lambda^{-1/2} P^T (X_i - \bar{X}_i), \;\; i = 1, \cdots, p, \qquad (2)$$

sendo Λ a matriz diagonal dos autovalores e P a matriz de autovetores. Observe que, para o cálculo do *i*-ésimo vetor esférico, os dados são centrados nas médias das respectivas variáveis (POSSE, 1995b).

3.4 Principais índices de projeção

Muitas funções índice foram desenvolvidas para definir projeções relevantes. Nesta seção são citadas algumas, cada qual com características que se distinguem nas projeções.

A maioria das projeções de baixa dimensionalidade é aproximadamente Normal (DIACONIS; FREEDMAN, 1984; HUBER, 1985). Muitos índices de projeção que medem o desvio da Normal foram inventados. Lee et al. (2005) citam que a maioria dos índices de projeção é focada na não normalidade. Citam-se, como exemplos, o índice de Entropia e o índice de Momentos (JONES; SIBSON, 1987), o índice de Legendre (FRIEDMAN, 1987), o índice de Hermite (HALL, 1989) e o índice Natural Hermite (COOK; BUJA; CABRERA, 1993).

Para o entendimento da descrição dos índices, consideremos a seguinte notação:

Índices: Holes, Massa Central, Curtose e PCA

- $Y = XA = [y_1, y_2, \cdots, y_d]^T$ é a matriz $n \times d$ dos dados projetados de $X_{n \times p}$.
- $\bar{y}$ é a média de Y.

Índices: LDA e PDA

- X_{ij} é o vetor *p*-dimensional da *j*-ésima observação na *i*-ésima classe, com $\bar{X}_{..} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} X_{ij}$ a média de cada variável em X, em que $i = 1, \cdots, g$, $j = 1, \cdots, n_i$, sendo g o número de classes, n_i é o número de observações na *i*-ésima classe, $n = \sum_{i=1}^{g} n_i$ e $\bar{X}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ é a média da *i*-ésima classe.
- $B = \sum_{i=1}^{g} n_i(\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})^T$ a matriz de distâncias entre classes.
- $H = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_{i.}\right)\left(X_{ij} - \bar{X}_{i.}\right)^T$ a matriz de dispersão intraclasses.

Índices: Friedman-Turkey, entropia, momentos, distâncias L^2 e qui-quadrado

- n é o tamanho amostral.
- Z é a matriz dos dados esféricos.
- z_i é a *i*-ésima observação dos dados esféricos.
- $A = [\alpha, \beta\,]$, em que α e β são vetores ortonormais *n*-dimensionais $(\alpha^T\alpha = 1 = \beta^T\beta \; e \; \alpha^T\beta = 0)$ do plano de projeção.
- $\left(z^\alpha, z^\beta\right)$ são as observações esféricas projetadas sobre os vetores α e β, ou seja, $z^\alpha = z^T\alpha$ e $z^\beta = z^\beta\beta$.
- ϕ_1 é a densidade Normal padrão.
- ϕ_2 é a densidade Normal padrão bivariada.

Mantendo-se essas especificações, enunciam-se os índices a seguir.

3.4.1 Holes

O índice de Holes, definido pela equação (3), formalizado por Cook, Buja e Cabrera (1993), é derivado da função de densidade Normal, sendo um índice sensível a projeções com poucos pontos no centro. Esse índice é definido para uma ou mais dimensões (COOK; SWAYNE, 2007).

$$PI_{Holes}(A) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} exp\left(-\frac{1}{2}y_i^T y_i\right)}{1 - exp\left(-\frac{d}{2}\right)}. \qquad (3)$$

3.4.2 Massa central

O índice de massa central, definido pela equação (4), foi proposto por Cook, Buja e Cabrera (1993), sendo um complemento do índice Holes, mantendo a característica de ser

sensível a projeções com muitos pontos no centro. Esse índice é definido para uma ou mais dimensões (COOK; SWAYNE, 2007).

$$PI_{MC}(A) = 1 - PI_{Holes}(A) = \frac{\frac{1}{n}\sum_{i=1}^{n} exp\left(-\frac{1}{2}y_i^T y_i\right) - exp\left(-\frac{d}{2}\right)}{1 - exp\left(-\frac{d}{2}\right)} \quad (4)$$

3.4.3 PCA

O índice PCA tem como base a análise de componentes principais. Sua principal característica é encontrar projeções em que os dados estão mais dispersos, sendo a variância total dos dados projetados o aspecto de interesse. Este índice é definido somente para uma dimensão (COOK; SWAYNE, 2007) e a sua expressão é dada pela equação (5).

$$PI_{PCA}(A) = \frac{1}{n} Y^T Y = \frac{1}{n}\sum_{i=1}^{n} y_i^2. \quad (5)$$

3.4.4 Curtose

O índice de Curtose, definido por Pena e Prieto (2001), foi proposto para a identificação de agrupamentos em dados multivariados, utilizando informações das projeções univariadas dos dados da amostra em certas direções. As direções escolhidas são aquelas que minimizam ou maximizam o coeficiente de curtose. A maximização desse índice favorece a detecção de *outliers*, enquanto a minimização favorece a detecção de agrupamentos. O índice é definido pela equação (6) para uma dimensão.

$$PI_{Curtose}(A) = \frac{(n-1)^2 \sum_{i=1}^{n}(y_i - \bar{y})^4}{n(\sum_{i=1}^{n}(y_i - \bar{y})^2)^2}. \quad (6)$$

3.4.5 LDA

O índice LDA, mencionado por Espezua et al. (2015) e Lee et al. (2005), e tem como fundamento a análise discriminante linear (*Linear Discriminant Analysis*) e é empregado na classificação exploratória supervisionada. Esse índice é definido para uma ou mais dimensões, e utiliza as ideias do discriminante linear de Fisher na busca por projeções para classificação, favorecendo as projeções lineares com a maior separação entre classes e a menor dispersão intraclasse, sendo definido pela equação (7).

$$PI_{lda}(A) = \begin{cases} 1 - \dfrac{|A^T HA|}{|A^T(H+B)A|} & , para\ |A^T(H+B)A| \neq 0 \\ 0 & , para\ |A^T(H+B)A| = 0. \end{cases} \quad (7)$$

3.4.6 PDA

O índice *Penalized Discriminant Analysis*, ou PDA, baseia-se na penalização do índice LDA e foi desenvolvido para resolver os problemas com muitos preditores altamente correlacionados (HASTIE; BUJA; TIBSHIRANI, 1995; LEE; COOK, 2010). Esse índice é definido para uma ou mais dimensões, e sua expressão é dada pela equação (8).

$$PI_{pda}(A) = 1 - \frac{\left|A^T\{(1-\lambda)H + n\lambda I_p\}A\right|}{\left|A^T\{(1-\lambda)(H+B) + n\lambda I_p\}A\right|}, \quad (8)$$

sendo I_p a matriz identidade de ordem p e $\lambda \in [0,1)$ um parâmetro a ser estimado. Se $\lambda = 0$, tem-se que o índice PDA é o mesmo do LDA (LEE; COOK, 2010).

3.4.7 Lr-norm

O índice *Lr*-norm, mencionado por Lee et al. (2005), é empregado para classificação exploratória supervisionada e pode ser utilizado na detecção de *outliers*, pois existe uma relação linear crescente entre a medida r e a medida leverage[5] do *outlier*.

Dada a projeção A *d*-dimensional, então $Y = XA$ é a matriz $n \times d$ dos dados projetados. Assim, tem-se que y_{ijl} é a *j*-ésima observação na *i*-ésima classe da *l*-ésima variável em Y, sendo $\bar{y}_{..l} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ijl}$ a média geral de cada variável em Y, com $i = 1, \cdots, g$, $j = 1, \cdots, n_i$, g é o número de classes, n_i é o número de observações na classe i, e $n = \sum_{i=1}^{g} n_i$, $\bar{y}_{i.l} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ijl}$ é a média geral de cada variável em Y da *i*-ésima classe. Este índice é definido pela equação (9) para uma ou mais dimensões.

$$PI_{Lr}(A) = \left(\frac{\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{y}_{i.l} - \bar{y}_{..l})^r}{\sum_{l=1}^{k}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(y_{ijl} - \bar{y}_{i.l}\right)^r}\right)^{1/r}. \quad (9)$$

[5] Refere-se à uma medida de alavancagem, tornando mais favorável a detecção dos *outlier.*

3.4.8 Friedman-Tukey

O índice de projeção proposto por Friedman e Tukey (1974) é baseado em distâncias interpontos, bem como na variância da nuvem de pontos para buscar a projeção ótima. É utilizado para a identificação de agrupamentos em dados multivariados, empregando-se informações das projeções dos dados da amostra em certas direções. As direções escolhidas são aquelas que maximizam o coeficiente, o que proporciona a maior separação para os diferentes agrupamentos. Aplicando-se novamente o índice a cada agrupamento separadamente, podem-se encontrar novas projeções que revelam mais agrupamentos dentro de cada conjunto de dados isolados. Estes subgrupos podem ser isolados e o processo repetido. Este índice foi revisto do original para ser um invariante afim e é definido pela equação (10) para duas dimensões (MARTINEZ; MARTINEZ, 2007; POSSE, 1995b).

$$PI_{FT}(A) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(R^2 - r_{ij}^2\right)^3 \mathbf{1}(x)\left(R^2 - r_{ij}^2\right) \tag{10}$$

sendo $R = 2{,}29n^{-1/5}$, $r_{ij}^2 = \left(z_i^{\alpha} - z_j^{\alpha}\right)^2 + \left(z_i^{\beta} - z_j^{\beta}\right)^2$, e $\mathbf{1}(x)$ a função indicadora para valores positivos, sendo $x = R^2 - r_{ij}^2$, e

$$\mathbf{1}(x) = \begin{cases} 1; x > 0 \\ 0; x \leq 0. \end{cases}$$

3.4.9 Entropia

O índice de entropia, mencionado por Huber (1985) e Jones e Sibson (1987), é uma extensão do índice de Friedman-Tukey, construído utilizando-se a entropia negativa de uma estimativa do núcleo da densidade. Ele é definido pela equação (11) para duas dimensões (MARTINEZ; MARTINEZ, 2007; POSSE, 1995b).

$$PI_{Entropy}(A) = \frac{1}{n}\sum_{i=1}^{n} log\left[\frac{1}{nh_\alpha h_\beta}\sum_{j=1}^{n}\phi_2\left(\frac{\left(z_i^{\alpha} - z_j^{\alpha}\right)}{h_\alpha}, \frac{\left(z_i^{\beta} - z_j^{\beta}\right)}{h_\beta}\right)\right] + log(2\pi e), \tag{11}$$

sendo ϕ_2 a densidade Normal padrão bivariada. Os valores de h_γ, $\gamma = \alpha, \beta$ são obtidos a partir de

$$h_\gamma = 1{,}06n^{-1/5}\left(\sum_{i=1}^{n}\left\{z_i^\gamma - \sum_{j=1}^{n}\frac{z_j^\gamma}{n}\right\}^2 /(n-1)\right)^{1/2}.$$

3.4.10 Baseado em momentos

Esse índice foi desenvolvido por Jones e Sibson (1987) e baseia-se no terceiro e no quarto momentos bivariados, o que o torna muito rápido para ser calculado e útil para grandes conjuntos de dados, aliviando o cálculo das integrais do índice baseado na entropia. No entanto, um problema com este índice é que ele tende a localizar estrutura nas caudas da distribuição. O índice é definido pela equação (12) para duas dimensões (MARTINEZ; MARTINEZ; SOLKA, 2010; POSSE, 1995b).

$$PI_M(A) = \frac{1}{12}\left\{k_{30}^2 + 3k_{21}^2 + 3k_{12}^2 + k_{03}^2 + \frac{1}{4}(k_{40}^2 + 4k_{31}^2 + 6k_{22}^2 + 4k_{13}^2 + k_{04}^2)\right\}, \quad (12)$$

sendo

$k_{12} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(z_i^\beta\right)^2 z_i^\alpha$, $k_{21} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^\alpha)^2 z_i^\beta$

$k_{22} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}(z_i^\alpha)^2\left(z_i^\beta\right)^2 - \frac{(n-1)^3}{n}\right\}$,

$k_{03} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(z_i^\beta\right)^3$, $k_{30} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^\alpha)^3$

$k_{31} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(z_i^\alpha)^3 z_i^\beta$, $k_{13} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(z_i^\beta\right)^3 z_i^\alpha$

$k_{40} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}(z_i^\alpha)^4 - \frac{3(n-1)^3}{n}\right\}$ e

$k_{04} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}\left(z_i^\beta\right)^4 - \frac{3(n-1)^3}{n}\right\}$.

3.4.11 Distâncias L^2

Vários índices estimam a distância L^2 entre a densidade dos dados projetados e uma densidade Normal padrão bivariada. Todos os índices de projeção desta classe utilizam várias expansões polinomiais ortonormais truncadas para estimar a densidade marginal dos dados projetados.

Uma proposta foi feita por Friedman (1987), construída pela inversão da densidade por meio de uma função de distribuição cumulativa Normal com as transformações $y^\alpha = 2\phi(z^\alpha) - 1$ e $y^\beta = 2\phi\left(z^\beta\right) - 1$, em que ϕ é a distribuição Normal padrão, e

utilizando J termos de polinômios de Legendre para a expansão. Esta expansão é, então, comparada com uma densidade uniforme no intervalo $[-1,1]$ por uma distância L^2. Esse é um bom índice geral, e rápido para ser calculado. O índice de Legendre é obtido pela equação (13) para duas dimensões (MARTINEZ; MARTINEZ, 2007; POSSE, 1995b).

$$PI_{Leg}(A) = \frac{1}{4}\left\{\sum_{j=1}^{J}(2j+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_j(y_i^{\alpha})\right)^2 + \sum_{k=1}^{J}(2k+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_k\left(y_i^{\beta}\right)\right)^2 + \sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_j(y_i^{\alpha})\,P_k\left(y_i^{\beta}\right)\right)^2\right\} \tag{13}$$

sendo $P_a(\cdot)$ o polinômio de Legendre de ordem a.

Morton (1989) notou que este índice não é invariante afim e, assim, propôs o seguinte índice, expressando a distância L^2 em coordenadas polares, usando

$$\rho = (z^{\alpha})^2 + \left(z^{\beta}\right)^2 \text{ e } \theta = arctan\left(\frac{z^{\beta}}{z^{\alpha}}\right).$$

Em seguida, a parte angular da densidade marginal dos dados projetados foi expandida em K termos de uma série de Fourier e a parte radial, em L termos de polinômios de Laguerre. O índice de Laguerre-Fourier é definido pela equação (14) para duas dimensões (MARTINEZ; MARTINEZ, 2007; POSSE, 1995b).

$$PI_{LF}(A) = \frac{1}{\pi}\sum_{l=0}^{L}\sum_{k=1}^{K}\left[\left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)exp\left(-\frac{\rho_i}{2}\right)cos(k\theta_i)\right)^2 + \left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)exp\left(-\frac{\rho_i}{2}\right)sin(k\theta_i)\right)^2\right] + \frac{1}{2\pi}\sum_{l=0}^{L}\left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)exp\left(-\frac{\rho_i}{2}\right)\right)^2 - \frac{1}{2\pi n}\sum_{i=1}^{n}exp\left(-\frac{\rho_i}{2}\right) + \frac{1}{8\pi}, \tag{14}$$

sendo L_a o polinômio de Laguerre de ordem a.

Hall (1989) estimou a mesma distância que Morton (1989), sem transformar os dados, expandindo $f_{\alpha\beta}$ (densidade marginal de Z no plano $P(\alpha,\beta)$) em H termos de polinômios de

Hermite, sendo ortogonal em relação a ϕ_1. O índice de Hermite é definido pela equação (15) para duas dimensões (POSSE, 1995b).

$$PI_{Her}(A) = \sum_{j=0}^{H}\sum_{k=0}^{H-j}\frac{2^{-(j+k)}}{j!\,k!}\left(\frac{1}{n}\sum_{i=1}^{n}H_j(z_i^{\alpha})\phi_1(z_i^{\alpha})\right)^2\left(\frac{1}{n}\sum_{i=1}^{n}H_k\left(z_i^{\beta}\right)\phi_1\left(z_i^{\beta}\right)\right)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\phi_1(z_i^{\alpha})\sum_{i=1}^{n}\phi_1\left(z_i^{\beta}\right)+\frac{1}{4\pi}, \tag{15}$$

sendo $H_a(\cdot)$ o polinômio Hermite de ordem a.

Devido ao peso ϕ_2, Cook, Buja e Cabrera (1993) expandiram $f_{\alpha\beta}$ em N termos de polinômios Naturais de Hermite, sendo ortogonal em relação a ϕ_2. O índice Natural Hermite é definido pela equação (16) para duas dimensões (POSSE, 1995b).

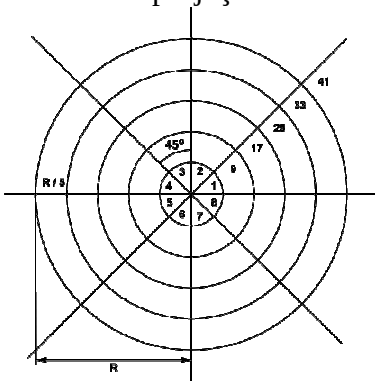
$$PI_{Nat}(A) = \sum_{j=0}^{N}\sum_{k=0}^{N-j}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{j!\,k!}}N_j(z_i^{\alpha})N_k\left(z_i^{\beta}\right)\phi_2\left(z_i^{\alpha},z_i^{\beta}\right)-b_jb_k\right)^2, \tag{16}$$

sendo que $N_a(\cdot)$ denota o polinômio Natural Hermite de ordem a e $b_{2i+1}=0$ e $b_{2i}=(-1)^i\sqrt{(2i)!}/\left(\sqrt{\pi}i!\,2^{2i+1}\right)$, para $i=0,1,2,\cdots,N$.

3.4.12 Qui-quadrado

Posse (1995a, 1995b) desenvolveu um índice busca de projeção baseado na distância Qui-quadrado, dividindo o plano de projeção (α,β) em 48 regiões ou caixas B_r $(r=1,\cdots,48)$ distribuídas em anéis (Figura 4), tendo em conta a simetria radial da distribuição Normal bivariada.

Figura 4 - Plano para avaliar o índice de projeção com base na distância Qui-quadrado.



Fonte: Adaptada de Posse (1995a).

As regiões apresentam a mesma largura angular de 45 graus, as regiões internas têm a mesma largura radial de $R = (2\log 6)^{1/2}$, e as regiões no anel externo têm probabilidade de $1/48$. Essa escolha para a largura radial fornece regiões com, aproximadamente, a mesma probabilidade para a distribuição Normal bivariada. O índice é definido pela equação (17) para duas dimensões (MARTINEZ; MARTINEZ; SOLKA, 2010).

$$PI_{\chi^2}(A) = \frac{1}{9}\sum_{j=0}^{8}\sum_{r=1}^{48}\frac{1}{c_r}\left[\frac{1}{n}\sum_{i=1}^{n} I_{B_r}\left(z_i^T\alpha(\eta_j), z_i^T\beta(\eta_j)\right) - c_r\right]^2, \tag{17}$$

sendo

- c_r é a probabilidade avaliada sobre a r-ésima região utilizando a Normal padrão bivariada

$$c_r = \iint_{B_r} \phi_2 dx_1 dx_2.$$

- B_r é uma caixa no plano de projeção.
- I_{B_r} é a função indicadora para a região B_r.
- $\eta_j = \pi j/36, j = 0, \cdots, 8$ é o ângulo pelo qual os dados são girados no plano antes de serem atribuídos a regiões B_r.

Segue ainda que

$$\alpha(\eta_j) = \alpha\cos\eta_j - \beta\sin\eta_j,$$
$$\beta(\eta_j) = \alpha\sin\eta_j + \beta\cos\eta_j.$$

O índice Qui-quadrado não é afetado pela presença de *outliers* e é sensível a distribuições com um furo no núcleo e a projeções que contenham *clusters*.

3.5 Algoritmo de otimização

Segundo Friedman (1987), métodos que são fundamentados no gradiente ou na técnica de Newton-Raphson são suscetíveis de cair em pseudomáximos, quando aplicados a uma função objetivo, a menos que o ponto de partida esteja dentro do domínio de atração de um máximo local ou global. Isso ocorre devido ao fenômeno de ondulação causado, principalmente, por flutuações de amostragem, o que pode distrair o algoritmo de busca de projeção de encontrar pontos importantes (máximos locais ou globais). Esses pseudomáximos podem ser visualizados como uma ondulação de alta frequência sobreposta à principal

estrutura variacional da função objetivo. A amplitude destas ondulações aumenta com o aumento da dimensão e a diminuição do tamanho da amostra.

Para Lee et al. (2005), um bom procedimento de otimização é uma parte muito importante na busca de projeções. O algoritmo de otimização de busca de projeção precisa ser flexível o suficiente para localizar máximos locais e globais, pois o objetivo da otimização é encontrar todas as projeções interessantes, e não só encontrar um máximo global, pois, às vezes, os máximos locais podem revelar estrutura de dados inesperadamente interessante. Assim, o método de otimização *simulated annealing* produz resultados tão bons quanto os obtidos por métodos convencionais de otimização, além de funcionar bem em grandes conjuntos de dados.

Existem outros otimizadores, como, por exemplo, *tribes* (COOREN; CLERC; SIARRY, 2009), *random scan sampling algorithm* (WEBB-ROBERTSON et al., 2005), *genetic algorithm* (GUO et al., 2000) e *particle swarm optimization* (KENNEDY; EBERHART, 1995). Nesse trabalho será usado o método de otimização *grand tour simulated annealing*, ver seção 3.5.2, nada impedindo que outros métodos fossem usados.

3.5.1 Algoritmo *simulated annealing* de otimização

Lee et al. (2005) demonstraram o seguinte algoritmo baseado em *simulated annealing*, que pode ser utilizado em muitas aplicações em que se busca encontrar projeções ótimas:

1: defina uma projeção inicial, A_0 e calcule o valor do índice de busca da projeção inicial $PI_0 = PI$;.

Para a *i*-ésima iteração,

2: gere uma projeção A_i de $N_{D_i}(A_0)$, em que $D_i = c^i$, c é o parâmetro de resfriamento predeterminado na faixa $(0,1)$, comumente chamado de *cooling*; $N_{D_i}(A_0) = \{A: A$ é uma projeção ortonormal com direção $A_0 + D_i B$ e B é uma projeção aleatória que é extraída de uma distribuição uniforme em uma unidade esférica $(p-1)$-dimensional$\}$;

3: calcule $PI_i = PI(A_i), \Delta PI_i = PI_i - PI_0, T_i = \frac{T_0}{log(i+1)}$;

4: defina $A_0 = A_i$ e $PI_0 = PI_i$, com probabilidade $\rho = min\left(exp\left(\frac{\Delta PI_i}{T_i}\right), 1\right)$ e incremente i para $i + 1$.

Repita 2-4 até que ΔPI_i seja suficientemente pequeno.

3.5.2 Algoritmo *grand tour simulated annealing* de otimização

Cook et al. (2008) utilizaram *grand tour* alternada, com uma otimização *simulated annealing* de um índice de busca de projeção, para criar um fluxo contínuo de projeções que são exibidas para visualização exploratória de dados multivariados. Segue o algoritmo.

1: da projeção inicial, A_a, calcule o valor do índice de busca da projeção inicial, PI_0.

Para a *i*-ésima iteração,

2: gerar um número fixo de novas projeções, $A_i^* = A_a + cA_i$, a partir de uma vizinhança da projeção atual, na qual o tamanho da vizinhança é especificado pelo parâmetro de resfriamento c na faixa $(0,1)$, e A_i é uma projeção aleatória;

3: calcule $PI_i = PI(A_i)$, $\Delta PI_i = PI_i - PI_0$;

4: defina a nova projeção com o valor de índice mais alto para ser o alvo, A_z e interpole a partir da projeção atual, A_a para a projeção de destino, A_z.

Repita 2-4 até que ΔPI_i seja suficientemente pequeno.

Este último algoritmo, essencialmente, difere do anterior apenas no processo de interpolação conforme apresentado na seção 3.1.

3.6 Análise de múltiplos fatores

Dentre as inúmeras técnicas propostas para analisar dados, a análise de múltiplos fatores (MFA) se caracteriza por permitir analisar grupos de variáveis com tamanhos diferentes e de naturezas distintas, que podem ser quantitativas ou categóricas, definidas no mesmo conjunto de observações (ESCOFIER; PAGÈS, 1982, 1990, 2008).

Este método pode ser muito útil para a análise de estudos em que se podem identificar vários grupos de variáveis, ou para os estudos em que as mesmas perguntas são feitas em intervalos de tempo diversos (ABDI; WILLIAMS; VALENTIN, 2013).

A vantagem proporcionada por esse método consiste em poder trabalhar com variáveis categóricas com escalas diferentes, gerando resultados com análise similar à análise de componentes principais (*Principal Components Analysis*, PCA), possibilitando também a visualização num espaço de duas ou três dimensões, os grupos de variáveis (sendo cada grupo representado por um ponto), as variáveis, os eixos principais e, ainda, as observações.

O cerne da técnica MFA é uma análise de fator aplicada ao conjunto todo de variáveis, no qual cada grupo de variáveis é balanceado, conduzindo a uma representação das

observações e variáveis, como em qualquer análise de fator. Devido ao balanceamento, esta análise de fator pode ser interpretada como uma análise canônica (ESCOFIER; PAGÈS, 1994).

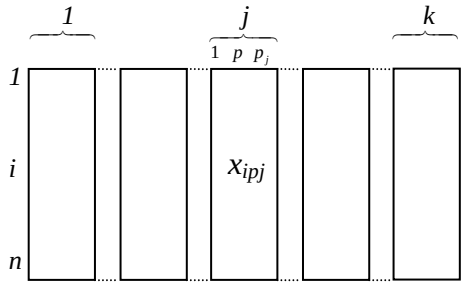
Convém ressaltar que, na análise de múltiplos fatores, o número de variáveis em cada grupo pode diferir e a natureza delas (quantitativa ou categórica) pode variar de um grupo para o outro, mas as variáveis devem ser da mesma natureza no grupo dado (ABDI; VALENTIN, 2007; ESCOFIER; PAGÈS, 2008).

Dados perdidos são permitidos no caso de variáveis categóricas e, assim, as observações nas quais não há nenhuma categoria p assumirão 0 (zero) para todas as variáveis indicadoras associadas à categoria p (ESCOFIER; PAGÈS, 1994).

3.6.1 Notação

Consideremos vários grupos de variáveis, representados por $X_1, \dots, X_j, \dots, X_k$, porém, justapostos em um único conjunto, conforme ilustrado na Figura 5.

Figura 5 - *Layout* dos conjuntos de dados.



Fonte: Adaptada de Bécue-Bertaut e Pagès (2008).

Na Figura 5, as n observações são descritas por k conjuntos de variáveis. A letra j refere-se a um conjunto, a letra p refere-se a uma coluna, p_j é o número de colunas no conjunto j e $m = \sum_{k \in j} p_j$ é o número de colunas em todos os conjuntos.

Uma tabela $n \times p_j$ está associada a cada conjunto j. As k tabelas juntas compõem uma tabela múltipla ou global $n \times m$. Para um conjunto j quantitativo, p_j é tanto o número de colunas quanto o número de variáveis. Já para um conjunto categórico j, p_j é tanto o número de colunas quanto o número de categorias. Este tipo de conjunto é representado por uma tabela de variáveis indicadoras na qual a coluna p está associada com a categoria p (BÉCUE-BERTAUT; PAGÈS, 2008; PAGÈS, 2002).

No cruzamento da linha i e coluna p (que pertencem a tabela j), tem-se

- se j é um conjunto quantitativo, x_{ipj} é o valor da variável p para a observação i.
- se j é um conjunto categórico, $x_{ipj} = 1$ se i pertence à categoria p e 0, caso contrário.

3.6.2 Procedimento para análise de múltiplos fatores

Primeiramente, centralizam-se os dados por coluna, subtraindo-se de cada elemento da coluna a sua respectiva média. Com isso, a média de cada coluna será zero, ou seja,

$$C_{ip} = x_{ip} - \overline{x}_p. \tag{18}$$

Em seguida, normalizam-se as colunas, dividindo-se cada elemento da coluna pela raiz quadrada da soma do quadrado da respectiva coluna, sendo

$$x_{ip}^{*} = \frac{C_{ip}}{\sqrt{\sum_{i=1}^{n} (C_{ip})^2}}, \tag{19}$$

e, desse modo, cada coluna irá se comportar como um vetor de módulo 1.

Mas se os dados forem quantitativos, aplica-se um PCA (Análise de Correspondência Múltipla, se forem de variáveis categóricas), em cada grupo X_j de variáveis para encontrar os primeiros autovalores λ_{j1}, fazendo-se $\varphi_j = a_p/\sqrt{\lambda_{j1}}$, em que a_p é o peso da variável p. Para variáveis quantitativas, $a_p = 1$ para todo p. Para variáveis categóricas, a_p é a proporção das observações que não apresentam categoria p. Assim segue que

$$S_j = X_j \times \varphi_j. \tag{20}$$

A matriz global será uma matriz $n \times m$ denotada por

$$S = [S_1 \ldots S_j]. \tag{21}$$

Com os dados balanceados segue a análise global em S.

3.6.3 MFA como um PCA

MFA consiste de um PCA na matriz global S, cujo termo geral é dado por (21). Do PCA, também chamado de análise global, os componentes principais são obtidos de modo

usual e os escores dos fatores globais são dados pela equação (23) (ABDI; VALENTIN, 2007; ABDI; WILLIAMS, 2010; PAGÈS, 2002, 2004).

Utilizando-se a decomposição em valores singulares, sabe-se que

$$S = U\Lambda V^T, \text{ com } U^T U = V^T V = \boldsymbol{I}, \tag{22}$$

sendo U e V as matrizes de autovetores e $\Lambda = diag\left(\sqrt{\lambda_i}\right)$ em que $\lambda_i > 0$ são os autovalores e $\boldsymbol{I}$ é a matriz identidade. Assim, extraindo-se os autovetores e autovalores da matriz S obtêm-se os escores dos fatores globais.

3.6.4 Os escores dos fatores globais

Os escores dos fatores globais são dados por

$$F = M^{\frac{1}{2}} U\Lambda, \tag{23}$$

sendo M a matriz diagonal das massas das observações, com $m_i = 1/n$, e n o número de linhas da matriz S. Em F, cada linha representa as observações e cada coluna, as variáveis.

3.6.5 Análise parcial

A análise global revela a estrutura comum das observações. Para observar como cada variável interpreta este espaço, projeta-se o conjunto de dados de cada variável sobre a análise global. Isso é alcançado obtendo-se a matriz de projeção, reescrevendo-se a equação (23), podendo os escores dos fatores globais serem computados como

$$F = M^{\frac{1}{2}} U\Lambda = SS^T M^{\frac{1}{2}} U\Lambda^{-1}. \tag{24}$$

Isto mostra que

$$P = M^{\frac{1}{2}} U\Lambda^{-1}, \tag{25}$$

é a matriz de projeção que transforma a matriz SS^T em escores dos fatores.

A matriz de projeção é utilizada para projetar os grupos sobre o espaço global, fazendo

$$F_j = k \times S_j S_j^T P, \tag{26}$$

sendo k o número de conjuntos de variáveis.

3.6.6 Inércia parcial entre os grupos de variáveis

As relações entre as variáveis e a solução global são analisadas calculando-se a inércia parcial de cada variável por cada dimensão da análise global. Isto é calculado, para cada variável, como a soma das projeções ao quadrado das variáveis do vetor singular V de S multiplicado pelo autovalor correspondente. Os vetores singulares são normalizados. A soma das inércias parciais para todos os grupos para uma determinada dimensão é igual ao seu autovalor (ABDI; VALENTIN, 2007; ABDI; WILLIAMS, 2010; PAGÈS, 2002, 2004), ou seja,

$$W_{jr} = \lambda_r \times \sum_{1}^{p} v_{prj}^2. \tag{27}$$

sendo W_{jr} o índice de similaridade do *j*-ésimo grupo na *r*-ésima componente dos dados com p observações, λ_r é o *r*-ésimo autovalor da decomposição em valores singulares e v_{prj} os valores do autovetor V.

As representações das linhas e colunas são analisadas conjuntamente, dispostas em um mapa perceptual. Dessa forma, seguem as interpretações de forma semelhante à análise de componentes principais.

Veja que a técnica MFA baseia-se em buscar variáveis latentes ($L_i = V_i\lambda_i$) que representam combinações lineares das variáveis relacionadas na matriz S, como no PCA. Assim, por meio da matriz V da decomposição em valores singulares ($S = U\Lambda V^T$), busca-se estabelecer a similaridade dos grupos por meio da proporção de explicação da variância representada por λ_i dentro dos grupos de variáveis em V. Logo, os índices de similaridade são as proporções explicadas por cada grupo de variáveis nessa combinação linear em V.

Com intuito didático, na seção 6.1 encontra-se um exemplo do uso da técnica MFA.

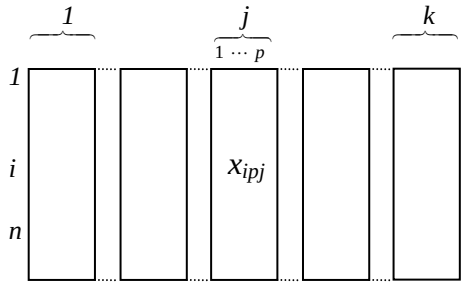
4 METODOLOGIA

Em consonância com as definições mencionadas nas seções anteriores, a metodologia estatística a ser desenvolvida neste trabalho é fundamentada nas etapas descritas a seguir.

4.1 Índice proposto para uso com MFA

Dada a amostra multivariada $X_{n\times m} = \left[X^1_{n\times p}\middle|\cdots\middle|X^j_{n\times p}\middle|\cdots|X^k_{n\times p}\right]$ composta por k grupos de variáveis quantitativas e n observações em cada grupo, sendo $m = pk$. Dessa forma, a observação x_{ipj}, sendo $i = 1,\cdots,n$ e $j = 1,\cdots,k$, é identificada como a *i*-ésima observação da *p*-ésima variável no *j*-ésimo grupo, conforme sugere o *layout* da Figura 6.

Figura 6 - *Layout* da estrutura de dados de uma análise de múltiplos fatores.



Fonte: Adaptada de Bécue-Bertaut e Pagès (2008).

A aplicação da *projection pursuit* na análise de múltiplos fatores consistiu em reduzir a dimensão de cada grupo $X^j_{n\times p}$ para uma dimensão $d < p$, podendo d ser diferente em cada grupo. Seguindo essas especificações, a amostra multivariada é representada pelo vetor $\tilde{X}_{n\times s} = [\tilde{X}^1_{n\times d}\middle|\cdots\middle|\tilde{X}^j_{n\times d}\middle|\cdots|\tilde{X}^k_{n\times d}]$, sendo $\tilde{x}_{idj}$ a *i*-ésima observação na *d*-ésima coluna do *j*-ésimo grupo, com $s = dk$.

O que se deseja é que, ao aplicar a técnica MFA nas matrizes $X_{n\times m}$ e $\tilde{X}_{n\times s}$, os resultados sejam análogos em relação às similaridades entre os grupos, conforme sugere a equação (28).

$$W_{jr} \cong \widetilde{W}_{j\tilde{r}} \Rightarrow \lambda_r \times \sum_1^p v^2_{prj} \cong \tilde{\lambda}_{\tilde{r}} \times \sum_1^d \tilde{v}^2_{d\tilde{r}j}. \tag{28}$$

sendo W_{jr} o índice de similaridade do *j*-ésimo grupo na *r*-ésima componente dos dados com as dimensões não reduzidas, e $\widetilde{W}_{j\tilde{r}}$ o índice de similaridade do *j*-ésimo grupo na $\tilde{r}$-ésima componente dos grupos com dimensões reduzidas, com $r > \tilde{r}$, sendo r e $\tilde{r}$ o posto das matrizes X e $\tilde{X}$, respectivamente. Reportando que $X = U\Lambda V^T$, convém ressaltar que os índices

de similaridades descritos na equação (28) foram definidos inicialmente utilizando o quadrado dos elementos da matriz V de autovetores e o quadrado dos valores singulares $\Lambda^2 = diag(\lambda_r)$. Assim, v_{prj}^2 representa o quadrado da *p*-ésima variável, na *r*-ésima componente no *j*-ésimo grupo em V, de modo semelhante a $\tilde{v}_{d\tilde{r}j}^2$ em $\tilde{V}$.

Note que a igualdade em (28) se resume em comparar os resultados com dimensões diferentes. Para exemplificar, consideremos as matrizes $X_{n\times m}$ e $\tilde{X}_{n\times s}$. Ao se aplicar a técnica MFA, têm-se as respectivas matrizes de autovetores V e $\tilde{V}$, e as similaridades entre os grupos podem ser expressas, respectivamente, na Tabela 1 e na Tabela 2.

Tabela 1 - Similaridade dos grupos de variáveis em $X_{n\times m}$ pela técnica MFA.

Comp.	Grupo 1	⋯	Grupo j	⋯	Grupo k
1	$\lambda_1 \times \left(v_{111}^2 + \cdots + v_{p11}^2\right)$	⋯	$\lambda_1 \times \left(v_{11j}^2 + \cdots + v_{p1j}^2\right)$	⋯	⋯
2	$\lambda_2 \times \left(v_{121}^2 + \cdots + v_{p21}^2\right)$	⋯	$\lambda_2 \times \left(v_{12j}^2 + \cdots + v_{p2j}^2\right)$	⋯	⋯
⋮	⋮	⋮	⋮	⋮	⋮
r	$\lambda_r \times \left(v_{1r1}^2 + \cdots + v_{pr1}^2\right)$	⋯	$\lambda_r \times \left(v_{1rj}^2 + \cdots + v_{prj}^2\right)$	⋯	⋯

Fonte: Do autor (2019).

Tabela 2 - Similaridade dos grupos de variáveis em $\tilde{X}_{n\times s}$ pela técnica MFA.

Comp.	Grupo 1	⋯	Grupo j	⋯	Grupo k
1	$\tilde{\lambda}_1 \times (\tilde{v}_{111}^2 + \cdots + \tilde{v}_{d11}^2)$	⋯	$\tilde{\lambda}_1 \times \left(\tilde{v}_{11j}^2 + \cdots + \tilde{v}_{d1j}^2\right)$	⋯	⋯
2	$\tilde{\lambda}_2 \times (\tilde{v}_{121}^2 + \cdots + \tilde{v}_{d21}^2)$	⋯	$\tilde{\lambda}_2 \times \left(\tilde{v}_{12j}^2 + \cdots + \tilde{v}_{d2j}^2\right)$	⋯	⋯
⋮	⋮	⋮	⋮	⋮	⋮
$\tilde{r}$	$\tilde{\lambda}_{\tilde{r}} \times (\tilde{v}_{1\tilde{r}1}^2 + \cdots + \tilde{v}_{d\tilde{r}1}^2)$	⋯	$\tilde{\lambda}_{\tilde{r}} \times \left(\tilde{v}_{1\tilde{r}j}^2 + \cdots + \tilde{v}_{d\tilde{r}j}^2\right)$	⋯	⋯

Fonte: Do autor (2019).

Seguindo essas especificações, enunciam-se os resultados (1) e (2) que podem ser generalizados para qualquer matriz. Nesse contexto, a justificativa da formalização do índice proposto é verificada.

(1) Assumindo $X_{n\times m}$ uma matriz de posto r. Pela decomposição do valor singular $X = U\Lambda V^T$. Então, para $\left[x_{ij}^2\right]$, representando o quadrado de cada elemento em X, tem-se que

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}^2 &= \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{q=1}^{r} u_{iq}^2 \times \lambda_{qq} \times v_{qj}^2 = \sum_{q=1}^{r}\sum_{i=1}^{n} u_{iq}^2 \times \lambda_{qq} \\ &= \sum_{j=1}^{m}\sum_{q=1}^{r} \lambda_{qq} \times v_{qj}^2 = \sum_{q=1}^{r} \lambda_q.\end{aligned} \tag{29}$$

sendo $[u_{ir}^2]$, $[\lambda_{rr}]$ e $\left[v_{rj}^2\right]$, respectivamente, o quadrado dos elementos de U, Λ e V^T.

(2) Assumindo a matriz $X_{n\times m}$ composta por k grupos de variáveis, sendo cada grupo composto por $s > 1$ variáveis, então, $X_{n\times m} = [X_{n\times s}^1|\cdots|X_{n\times s}^j|\cdots|X_{n\times s}^k]$, com x_{nsj} representando a *n*-ésima observação na *s*-ésima coluna do *j*-ésimo grupo, sendo $m = s \times k$. Ao considerar $[x_{il}^2]$ representando o quadrado de cada elemento da matriz X, então, para cada grupo j em X tem-se que

$$\sum_{l=1}^{s}\sum_{i=1}^{n} x_{ilj}^2 = \sum_{l=1}^{s}\sum_{q=1}^{r} \lambda_{qq} \times v_{qlj}^2, \tag{30}$$

sendo $\left[\lambda_{qq}\right]$ o quadrado dos elementos de Λ e $\left[v_{rsj}^2\right]$ o quadrado dos elementos das projeções do *j*-ésimo grupo de X em V^T.

Com o propósito de padronizar as variáveis com as mesmas unidades de medidas, necessariamente aplicou-se uma transformação linear, o que é usualmente feito na técnica MFA, conforme o procedimento sugerido por Bécue-Bertaut e Pagès (2008), considerando os dados centrados na média, descritos matricialmente por (31)

$$C_{n\times p} = X_{n\times p} - \boldsymbol{J}_{n\times p} \cdot diag\left(\bar{X}_{1\times p}\right), \tag{31}$$

sendo $\boldsymbol{J}_{n\times p}$ a matriz unitária e $\bar{X}_{1\times p}$ o vetor linha das médias das colunas de $X_{n\times p}$. Normaliza-se $C_{n\times p}$ dividindo-se cada elemento da coluna pela raiz quadrada da soma do quadrado da respectiva coluna, conforme (32)

$$N_{n\times p} = C_{n\times p} \times diag\left(1/\sqrt{\mathbf{1}_{1\times n} \cdot \left(C_{n\times p} \odot C_{n\times p}\right)}\right), \tag{32}$$

sendo $\mathbf{1}_{1\times n}$ o vetor unitário, lembrando que $C_{n\times p} \odot C_{n\times p}$ é produto de Hadamard. Logo, dado o primeiro autovalor λ_1 obtido de $N_{n\times p}$, para os dados transformados, resulta em (33)

$$S_{n\times p} = \frac{1}{\sqrt{\lambda_1}} \times N_{n\times p,} \tag{33}$$

Portanto, o índice proposto chamado de MF (acrônimo de *Multiple Factorial*) é definido pela expressão (34) para duas ou mais dimensões.

$$PI_{MF}(A) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(s_{ij}\right)^2, \tag{34}$$

sendo s_{ij} a *í*-ésima linha da *j*-ésima coluna de $S = (S_1, S_2, \ldots, S_p)$ e p o número de variáveis aleatórias (p-dimensional), na qual cada componente de S é representado por $S_p = (s_{1p}, s_{2p}, \ldots, s_{np})$.

A otimização numérica para encontrar um plano que maximize o índice (34) foi dada por uma modificação no algoritmo *grand tour simulated annealing* (seção 4.2), proposto por Cook et al. (2008), de modo que grupos de variáveis possam ser considerados em consonância com a técnica MFA.

Conjecturou-se a validade do índice proposto (34), comparando-se o grau de concordância ao aplicar a *projection pursuit* em duas dimensões, para cada grupo de variáveis. O mesmo foi feito a todos os índices da seção 3.4, exceto aos índices PCA e Curtose por serem avaliados apenas em uma dimensão. E posteriormente comparou-se os resultados do índice proposto em relação aos outros índices analisados.

Com esse propósito, procedeu-se à realização de 1.000 simulações Monte Carlo. Assim, códigos (Apêndice) foram desenvolvidos no software R (R CORE TEAM, 2018) para uso público, por meio do pacote MVar.pt versão 2.0.2 (OSSANI; CIRILLO, 2018), incorporando uma modificação no algoritmo *grand tour simulated annealing*, adaptado para a análise de múltiplos fatores (seção 4.2), com as especificações dadas em $5e^3$ iterações, *cooling* $= 0{,}95$, *eps* $= 1e^{-4}$ e *half* $= 30$ para computar os índices mencionados na seção 3.4, em função dos cenários descritos na Tabela 4, assumindo dados não esféricos nas simulações com redução de duas dimensões nos dados em cada grupo.

4.2 Algoritmo de busca de projeção para a técnica MFA

Aqui propõe-se um algoritmo de otimização na busca de projeção para ser usado com a técnica MFA. Esse algoritmo tem como base o algoritmo apresentado por Cook et al. (2008), ver seção 3.5.2, adaptado para lidar com grupos de variáveis. Assim segue

Input: $X_{n\times pk} = \left[X^1_{n\times p}\middle|\cdots\middle|X^j_{n\times p}\middle|\cdots|X^k_{n\times p}\right]$ matriz com *k* grupos de variáveis

p: número de variáveis no grupo *j*

n: número de observações

Maxiter: número máximo de iterações

Cooling: valor inicial do parâmetro de resfriamento na faixa $(0,1)$

Half : número de etapas sem alterar *Cooling*

Eps: precisão de aproximação para *Cooling*

for $j = 1 : k$ **do**

Gere a projeção inicial aleatória A_a, com $d < p$.

Execute a transformação linear $T_0 = X^j A_a$, para o *j*-ésimo grupo.

Calcule o índice de busca da projeção inicial, $PI_{MF}^0(T_0)$.

Faça $h = 0$, $Cooling = 0.95$, $Half = 30$, $Eps = 1e^{-4}$ e $i = 1$.

while ($i <$ *Maxiter* **and** *Cooling* > *Eps*) **do**

Gere uma nova projeção aleatória A_i.

Em seguida, gere nova projeção $A_i^* = A_a + Cooling \times A_i$.

Faça $A_z = interpolação(A_a, A_i^*)$, interpolação a partir da projeção A_a até a projeção A_i^*, conforme apresentado na seção 3.1.

Execute a transformação linear $T_i = X^j A_z$, para o *j*-ésimo grupo.

Calcule $PI_{MF}^i(T_i)$.

if $PI_{MF}^i < PI_{MF}^0$ **do**

Faça $A_a = A_z$, $PI_{MF}^0 = PI_{MF}^i$

else

Faça $h \mathrel{+}= 1$

end if

if $h = Half$ **then**

Faça $Cooling = Cooling * 0.9$ e $h = 0$

end if

Faça $i \mathrel{+}= 1$

end while

Faça $\tilde{X}_{n\times d}^j = T_i$, o que representa a matriz com a dimensão reduzida do *j*-ésimo grupo

end for

A nova matriz com *k* grupos de variáveis com as dimensões reduzidas será representada por $\tilde{X}_{n\times dk} = [\tilde{X}_{n\times d}^1 | \cdots | \tilde{X}_{n\times d}^j | \cdots | \tilde{X}_{n\times d}^k]$.

Output: Aplica-se a técnica MFA aos grupos formados em $\tilde{X}_{n\times dk}$.

4.3 Procedimento Monte Carlo para validação do novo índice proposto

A fim de validar o novo índice proposto na seção 4.1, considerando variáveis e grupos correlacionados em diversos cenários, a seguir descrevem-se os passos.

4.3.1 Geração das amostras para aplicação da técnica de análise de múltiplos fatores

Com o propósito de avaliar o desempenho do índice proposto na seção 4.1, consideraram-se diversos cenários, que evidenciam diferentes graus de correlação entre as variáveis pertencentes a cada grupo, e a heterogeneidade entre os grupos. Para isso, assumiu-se o procedimento proposto por Cirillo et al. (2010), no qual as amostras foram geradas de populações Normais dependentes e heterogêneas.

A observação multivariada foi especificada pelo vetor $\check{X}$, em que cada componente foi dado por $\check{X}_j = \left(X_{j1}, \cdots, X_{jp}\right)^T$ para $j = 1, \cdots k$, sendo k o número total de grupos e p o número de variáveis, considerando a estrutura de correlação global autorregressiva de ordem 1, $AR(1)$, definida em (35). Cada bloco, delimitado pelas linhas tracejadas, corresponde à estrutura de correlação das amostras geradas por uma Normal multivariada com estrutura de correlação global R_b.

$$R_b = \begin{bmatrix} \boxed{\begin{matrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \cdots & \cdots & 1 \end{matrix}} & \begin{matrix} \rho^{p+1} & \rho^{p+2} & \rho^{p+3} & \cdots & \rho^{pk-1} \\ \ddots & \ddots & \ddots & \cdots & \rho^{pk-2} \\ \ddots & \ddots & \ddots & \cdots & \rho^{pk-3} \\ \ddots & \ddots & \ddots & \cdots & \rho^{pk-4} \\ \ddots & \ddots & \ddots & \ddots & \vdots \end{matrix} \\ \begin{matrix} \rho^{p+1} & \ddots & \ddots & \ddots \\ \rho^{p+2} & \ddots & \ddots & \ddots \\ \rho^{p+3} & \ddots & \ddots & \ddots \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{pk-1} & \rho^{pk-2} & \rho^{pk-3} & \rho^{pk-4} \end{matrix} & \begin{matrix} \ddots \\ \ddots \\ \ddots \\ \ddots \\ \cdots \end{matrix} & \boxed{\begin{matrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \cdots & \cdots & 1 \end{matrix}} \end{bmatrix}, \quad (35)$$

A matriz de covariância global foi obtida por meio da equação (36)

$$\sum^{*} = D^{\frac{1}{2}} R_b D^{\frac{1}{2}}, \quad (36)$$

em que $D^{\frac{1}{2}}$ é uma matriz diagonal com os desvios padrões das variáveis. Com estas especificações, as amostras multivariadas provenientes de populações Normais dependentes foram geradas por $\check{X} \sim N_{pk}\left(\check{\mu}_{pk}, \sum^{*}\right)$, assumindo os valores paramétricos definidos em (37) e (38)

$$\breve{\mu}_{pk\times 1} = \begin{bmatrix} \breve{\mu}_1 = 0 \\ \vdots \\ \breve{\mu}_1 = 0 \end{bmatrix}, \tag{37}$$

$$\sum\nolimits^{*}_{(pk\times pk)} = \begin{bmatrix} \sum_{11} & \cdots & \sum_{k1} \\ \vdots & \ddots & \vdots \\ \sum_{k1} & \cdots & \sum_{kk} \end{bmatrix}. \tag{38}$$

As covariâncias representadas em (36) não são nulas, uma vez que cada elemento na diagonal representa a *j*-ésima matriz de covariância do grupo de variáveis indexadas por $j = 1, \cdots, k$.

A heterogeneidade entre cada matriz de covariância no processo de simulação foi determinada por meio de um grau de heterogeneidade δ especificado em (40), seguindo o algoritmo proposto por Cirillo et al. (2010), enunciado nos seguintes passos:

(1) simula-se uma amostra da distribuição Normal multivariada $N_{pk}\left(\breve{\mu}_{pk}, \sum^{*}\right)$ obtendo-se uma matriz $Y_{n\times pk}$, representada na Tabela 3. Cada bloco de p colunas (variáveis) corresponde ao *j*-ésimo grupo. Dessa forma, a unidade amostral multivariada é disposta nas n linhas.

Tabela 3 - *Layout* da matriz $Y_{n\times pk}$ utilizada na determinação do parâmetro da matriz de covariância sob a situação de heterogeneidade $(\delta > 1)$.

	Grupos									
	1			2			j	k		
1	y_{111}	$\cdots$	y_{1p1}	y_{112}	$\cdots$	y_{1p2}	$\cdots$	y_{11k}	$\cdots$	y_{1pk}
2	y_{211}	$\cdots$	y_{2p1}	y_{212}	$\cdots$	y_{2p2}	$\cdots$	y_{21k}	$\cdots$	y_{2pk}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_{n11}	$\cdots$	y_{np1}	y_{n12}	$\cdots$	y_{np2}	$\cdots$	y_{n1k}	$\cdots$	y_{npk}

Fonte: Do autor (2019).

(2) Para o 1º grupo $(j = 1)$, as observações não são alteradas. As p variáveis do *j*-ésimo grupo foram multiplicadas por d_j^*, $(j > 1)$, definido por

$$d_j^* = \left(1 + (j-1)\frac{(\delta - 1)}{(k-1)}\right)^{\frac{1}{p}}, \tag{39}$$

sendo δ o grau de heterogeneidade entre as matrizes de covariâncias especificadas. Finalizando este procedimento, adotou-se como parâmetro da matriz de covariância global $\sum_n$ definida por

$$\sum_n = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})(Y_j - \bar{Y}). \quad (40)$$

Após a definição dos parâmetros da matriz de covariância $\sum^*$ e $\sum_n$, mediante as situações de heterogeneidade entre as matrizes de covariâncias, especificadas em $\delta = 2$ e 8, foram feitas as simulações Monte Carlo, nas quais as observações amostrais multivariadas foram geradas. A matriz dos dados amostrais utilizados na análise de múltiplos fatores, foi, então, composta dos n vetores gerados por (41)

$$X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}. \quad (41)$$

Logo, fixando $p = 10$ e $n = 150$, os cenários considerados no procedimento de simulação foram definidos na Tabela 4.

Tabela 4 - Cenários para geração das amostras Normais multivariadas a serem utilizadas na análise de múltiplos fatores.

Heterogeneidade entre as matrizes de covariâncias (δ)	Número de grupos (k)	Grau de correlação entre as variáveis (ρ)
2	7	0,2
		0,5
	10	0,9
8	7	0,2
		0,5
	10	0,9

Fonte: Do autor (2019).

5 RESULTADOS E DISCUSSÕES

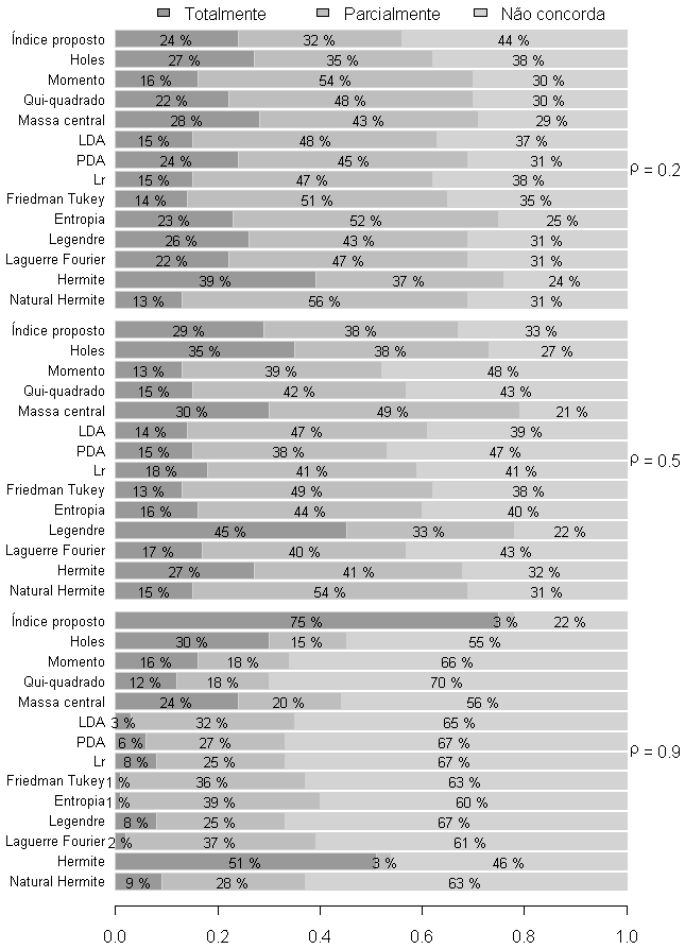
Em função dos objetivos propostos, os resultados ilustrados na Figura 7 à Figura 10 correspondem às dimensões reduzidas dadas em relação à primeira componente principal, justamente por apresentarem sempre a maior explicação na técnica MFA. Dessa forma, são descritos a seguir, a avaliação dos índices em função da heterogeneidade, número de grupos e das correlações entre as variáveis, conforme procedimento apresentado na seção 4.3.

Nas figuras, a palavra totalmente significa que os resultados dos dados em dimensão reduzida concordam na sua totalidade com os dados em dimensões originais ao aplicar a técnica MFA, parcialmente os resultados concordam em parte, e não concordam os resultados são todos diferentes.

5.1 Avaliação do índice proposto em função da heterogeneidade entre os grupos

Os resultados ilustrados na Figura 7 indicaram que o índice proposto apresentou, inicialmente, elevada taxa de discordância, em relação aos demais índices, quando se considerou um fraco grau de correlação entre as variáveis $(\rho = 0{,}2)$. Contudo, à medida que o grau de correlação foi incrementando, essa taxa foi reduzida de tal forma que, mantendo um forte grau de correlação entre as variáveis $(\rho = 0{,}9)$, o índice proposto resultou em baixa discordância e resultados promissores em relação à concordância com os demais. Vale ressaltar que os dados simulados foram não esféricos, situação esta que, em geral, desfavorece a aplicabilidade dos índices propostos.

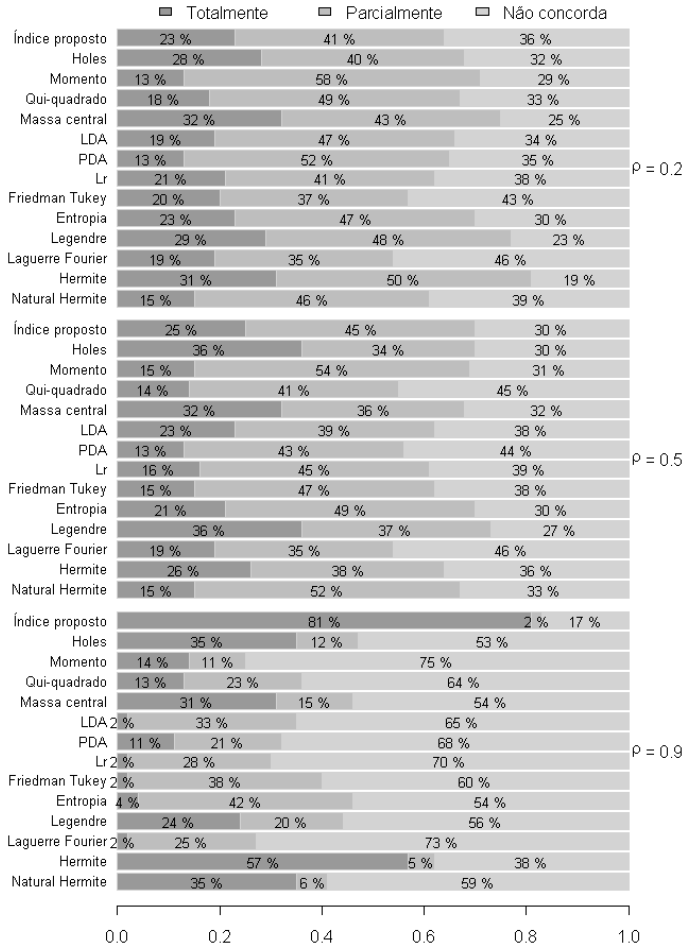
Figura 7 - Resultados das simulações com grau $\delta = 2$ de heterogeneidade entre $k = 7$ grupos e grau de correlação $\rho = 0{,}2;\ 0{,}5$ e $0{,}9$ entre as variáveis dentro dos grupos.



Fonte: Do autor (2019).

Diante do exposto, com o propósito de comparar a validação do índice para os mesmos cenários, porém, aumentando o grau de heterogeneidade entre os grupos ($\delta = 8$), os resultados ilustrados na Figura 8 foram comparados em relação à situação de baixa heterogeneidade ($\delta = 2$) ilustrada na Figura 7.

Figura 8 - Resultados das simulações com grau $\delta = 8$ de heterogeneidade entre $k = 7$ grupos e grau de correlação $\rho = 0{,}2;\ 0{,}5$ e $0{,}9$ entre as variáveis dentro dos grupos.



Fonte: Do autor (2019).

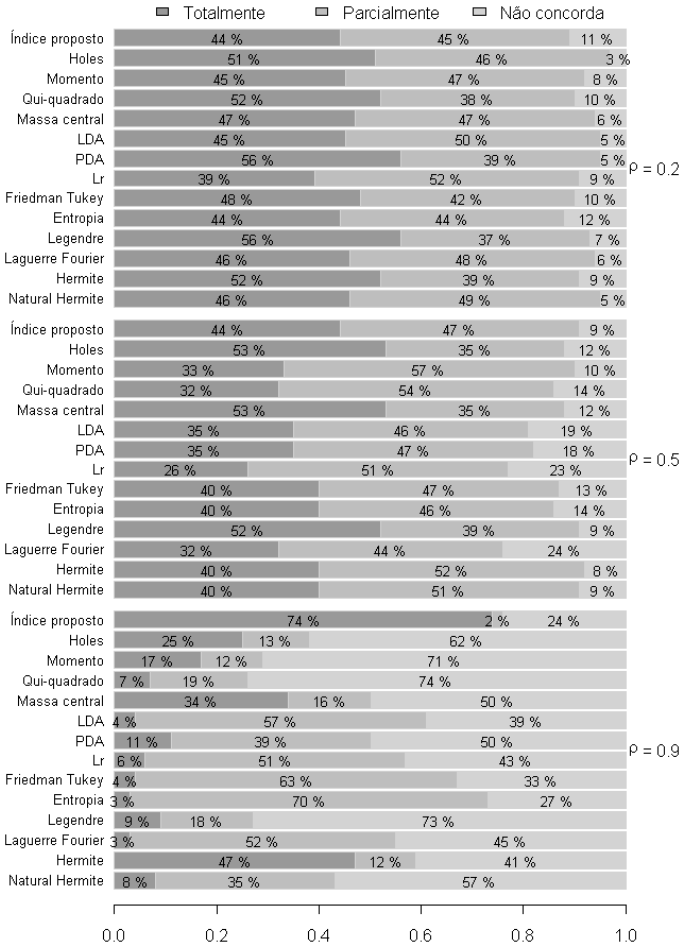
Os resultados ilustrados na Figura 8 praticamente confirmaram o mesmo comportamento dos índices apresentados na Figura 7, em relação ao grau de correlação entre as variáveis e os graus de concordância e discordância entre os índices. Praticamente qualquer diferença é dada devido à oscilação do erro Monte Carlo. Entretanto, um resultado que convém destacar é verificado para o grau forte de correlação ($\rho = 0{,}9$). Neste contexto, nota-se que o índice apresentou elevado grau de concordância, quando as amostras foram simuladas para um baixo grau de heterogeneidade ($\delta = 2$), com porcentagem próxima a 75%. Aumentando para ($\delta = 8$), o índice proposto apresentou uma melhoria considerável, com porcentagem de concordância próxima a 81%.

5.2 Avaliação do índice proposto em função do aumento do número de grupos

Em comparação com os resultados ilustrados na Figura 7, na qual foi considerada baixa heterogeneidade entre os grupos e $k = 7$ grupos, nota-se na Figura 9 que, nas mesmas

configurações, o incremento no número de grupos ($k = 10$) confirmou que os graus de concordância e discordância do índice proposto foram aproximadamente iguais quando $\rho = 0{,}9$. Portanto, há evidências estatísticas de que o índice apresenta resultados promissores para a situação em que as variáveis são fortemente correlacionadas, independente do tamanho do grupo.

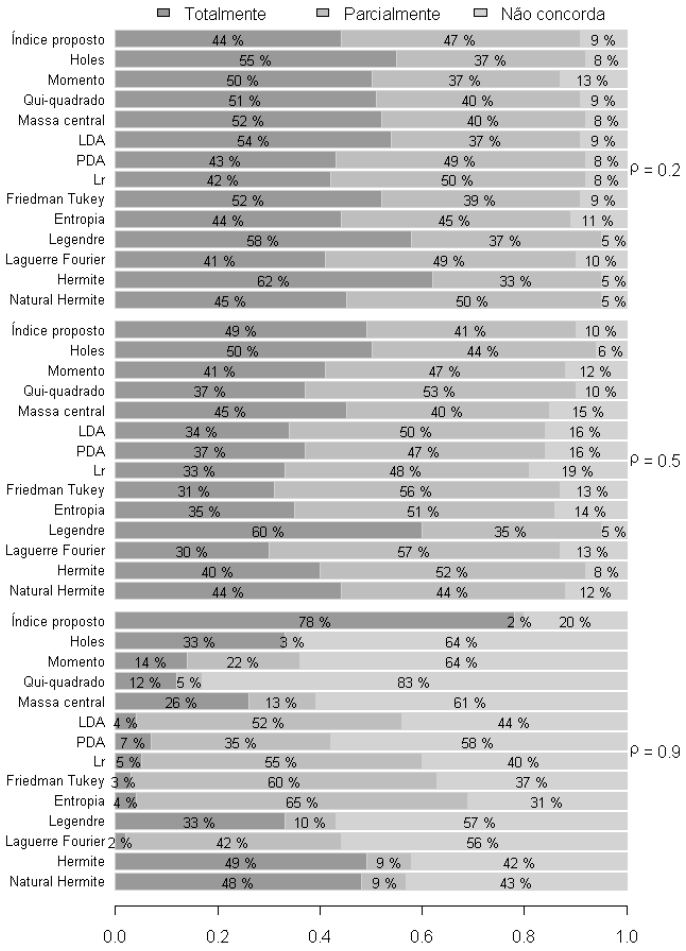
Figura 9 - Resultados das simulações com grau $\delta = 2$ de heterogeneidade entre $k = 10$ grupos e graus de correlação $\rho = 0{,}2$; $0{,}5$ e $0{,}9$ entre as variáveis dentro dos grupos.



Fonte: Do autor (2019).

Em se tratando da alta heterogeneidade entre os grupos ($\delta = 8$), o aumento do número de grupos ($k = 10$), de modo geral, não afetou o desempenho dos índices. Nota-se, por meio da Figura 10, que ocorreu o mesmo comportamento dos índices em relação ao grau de correlação e à sua taxa de concordância, em comparação com os índices competidores.

Figura 10 - Resultados das simulações com grau $\delta = 8$ de heterogeneidade entre $k = 10$ grupos e graus de correlação $\rho = 0{,}2,\ 0{,}5$ e $0{,}9$ entre as variáveis dentro dos grupos.



Fonte: Do autor (2019).

Ressalta-se que, em todos os cenários simulados, nas situações de fraca correlação ($\rho = 0{,}2$) e moderada correlação ($\rho = 0{,}5$), o índice proposto apresentou redução da taxa de discordância entre os seus competidores. Neste contexto, há evidências para afirmar que, em média, todos os índices são afetados pelo aumento do número de grupos. Dessa forma, pode-se afirmar que a ocorrência desses resultados não inviabiliza a aplicação dos índices avaliados neste trabalho, em se tratando da *projection pursuit* aplicada na análise de múltiplos fatores.

Com intuito didático, na seção 6.2 encontra-se um exemplo do uso da técnica MFA com redução de dimensão utilizando a *projection pursuit* por meio do índice proposto.

6 EXEMPLOS APLICADOS

Com finalidade didática, encontram-se, nesta seção, um exemplo do uso da técnica MFA e um exemplo do uso da técnica MFA com a redução de dimensão usando a *projection pursuit* por meio do índice proposto.

6.1 Uso da técnica MFA

Foram simulados três grupos de variáveis, conforme apresentado na Tabela 5. Os dados de cada grupo foram gerados conforme metodologia proposta por Cirillo et al. (2010), visto na seção 4.3.1, utilizando-se os códigos do Apêndice, com os parâmetros $\rho = 0{,}9$ e $\delta = 1$. A ideia aqui é usar alto grau de correlação e baixa heterogeneidade, esses valores foram escolhidos por conta dos resultados da seção 5.

Tabela 5 - Matriz $X_{6\times10}$ de dados simulados dos grupos.

Obs.	Grupo 1				Grupo 2			Grupo 3		
	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
1	1,39	2,44	3,06	4,35	1,75	2,13	2,88	1,90	3,08	3,58
2	0,98	1,85	3,02	3,76	-0,42	0,19	1,38	1,58	2,05	3,76
3	2,38	3,56	4,37	5,40	0,03	1,27	1,76	-0,06	1,21	2,21
4	1,51	2,41	3,68	5,31	0,33	1,02	2,36	0,62	2,40	3,58
5	0,53	1,69	1,88	4,24	1,01	2,41	2,41	1,73	3,10	3,66
6	1,40	2,02	3,40	3,58	1,03	2,56	3,16	1,24	1,89	2,77

Fonte: Do autor (2019).

Aplicando-se a metodologia da técnica MFA, descrita na seção 3.6, nos grupos de variáveis na Tabela 5, inicialmente equilibra-se a influência das variáveis V_p em cada grupo X^j, para executar, posteriormente, uma análise global. Centralizam-se os dados por coluna, subtraindo-se de cada elemento da coluna a sua respectiva média, conforme descrito pela equação (18). Logo, tem-se a Tabela 6.

Tabela 6 - Matriz $C_{6\times10}$ dos dados centrados nas médias das respectivas colunas.

Grupo 1				Grupo 2			Grupo 3		
V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
0,025	0,112	-0,175	-0,090	1,128	0,533	0,555	0,732	0,792	0,320
-0,385	-0,478	-0,215	-0,680	-1,042	-1,407	-0,945	0,412	-0,238	0,500
1,015	1,232	1,135	0,960	-0,592	-0,327	-0,565	-1,228	-1,078	-1,050
0,145	0,082	0,445	0,870	-0,292	-0,577	0,035	-0,548	0,112	0,320
-0,835	-0,638	-1,355	-0,200	0,388	0,813	0,085	0,562	0,812	0,400
0,035	-0,308	0,165	-0,860	0,408	0,963	0,835	0,072	-0,398	-0,490

Fonte: Do autor (2019).

Em seguida, ao aplicar a equação (19) na Tabela 6, normalizam-se as colunas, dividindo-se cada elemento da coluna pela raiz quadrada da soma do quadrado dos elementos da respectiva coluna, o que gera a Tabela 7.

Tabela 7 - Matriz $x^*_{6\times10}$ dos dados normalizados por colunas.

Grupo 1				Grupo 2			Grupo 3		
V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
0,018	0,074	-0,095	-0,053	0,640	0,257	0,372	0,435	0,484	0,229
-0,279	-0,318	-0,116	-0,397	-0,591	-0,679	-0,633	0,244	-0,146	0,357
0,737	0,818	0,613	0,561	-0,335	-0,158	-0,379	-0,730	-0,659	-0,750
0,105	0,054	0,240	0,508	-0,165	-0,278	0,023	-0,326	0,068	0,229
-0,606	-0,424	-0,732	-0,117	0,220	0,393	0,057	0,334	0,496	0,286
0,025	-0,205	0,089	-0,503	0,232	0,465	0,560	0,043	-0,243	-0,350

Fonte: Do autor (2019).

Como os dados são quantitativos, encontra-se o primeiro autovalor para cada grupo de variáveis apresentadas na Tabela 7. Em seguida, divide-se cada elemento do grupo pela raiz quadrada do respectivo autovalor (desvio padrão). Os primeiros autovalores de cada grupo são dados por λ_{11} =1,834829, λ_{21} =1,652816 e λ_{31} =1,596909. Assim ao aplicar a equação (20), tem-se a matriz mostrada na Tabela 8.

Tabela 8 - Matriz global $S_{6\times10}$ dos dados balanceados.

Grupo 1				Grupo 2			Grupo 3		
V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
0,010	0,040	-0,052	-0,029	0,387	0,156	0,225	0,272	0,303	0,143
-0,152	-0,173	-0,063	-0,217	-0,357	-0,411	-0,383	0,153	-0,091	0,224
0,401	0,446	0,334	0,306	-0,203	-0,095	-0,229	-0,457	-0,413	-0,470
0,057	0,030	0,131	0,277	-0,100	-0,168	0,014	-0,204	0,043	0,143
-0,330	-0,231	-0,399	-0,064	0,133	0,238	0,034	0,209	0,311	0,179
0,014	-0,112	0,049	-0,274	0,140	0,281	0,339	0,027	-0,152	-0,219

Fonte: Do autor (2019).

Pela decomposição em valores singulares conforme apresentado pela equação (22), sabe-se que $S = U\Lambda V^T$, sendo U e V as matrizes de autovetores e $\Lambda = diag\left(\sqrt{\lambda_i}\right)$, em que $\lambda_i > 0$ são os autovalores. Assim, ao extrair os autovetores e os autovalores da matriz S obtêm-se

$$U_{6\times6} = \begin{bmatrix} -0{,}3304 & -0{,}3316 & 0{,}3634 & -0{,}5158 & 0{,}4648 & -0{,}4082 \\ -0{,}0223 & 0{,}8020 & -0{,}3154 & -0{,}2481 & 0{,}1690 & -0{,}4082 \\ 0{,}7813 & -0{,}1690 & 0{,}0236 & 0{,}2648 & 0{,}3517 & -0{,}4082 \\ 0{,}1575 & 0{,}1143 & 0{,}5237 & -0{,}1233 & -0{,}7113 & -0{,}4082 \\ -0{,}4972 & 0{,}0358 & 0{,}1001 & 0{,}7547 & 0{,}0729 & -0{,}4082 \\ -0{,}0889 & -0{,}4516 & -0{,}6954 & -0{,}1323 & -0{,}3470 & -0{,}4082 \end{bmatrix},$$

$$V_{10\times6} = \begin{bmatrix} 0{,}3434 & -0{,}2148 & 0{,}0839 & -0{,}3313 & 0{,}1677 & 0{,}3966 \\ 0{,}3310 & -0{,}1910 & 0{,}2594 & -0{,}0642 & 0{,}4899 & -0{,}1141 \\ 0{,}3494 & -0{,}1169 & 0{,}0070 & -0{,}5363 & -0{,}1872 & -0{,}4873 \\ 0{,}2494 & -0{,}0659 & 0{,}6828 & 0{,}2877 & -0{,}1634 & -0{,}2629 \\ -0{,}2636 & -0{,}4726 & 0{,}1941 & -0{,}1958 & 0{,}2677 & 0{,}1034 \\ -0{,}2025 & -0{,}5274 & -0{,}1319 & 0{,}4424 & 0{,}0296 & -0{,}4487 \\ -0{,}2049 & -0{,}5175 & -0{,}0474 & -0{,}2842 & -0{,}5519 & 0{,}1443 \\ -0{,}4161 & 0{,}0859 & -0{,}1119 & -0{,}3338 & 0{,}4751 & -0{,}3533 \\ -0{,}3922 & -0{,}0200 & 0{,}4999 & 0{,}0178 & 0{,}0846 & 0{,}3211 \\ -0{,}3298 & 0{,}3496 & 0{,}3737 & -0{,}2986 & -0{,}2450 & -0{,}2495 \end{bmatrix}$$

e

$$\Lambda = [1{,}4146 \quad 0{,}9533 \quad 0{,}5775 \quad 0{,}3597 \quad 0{,}3004 \quad 0{,}0000].$$

Logo, os componentes principais apresentam as variâncias iguais a

$$\Lambda^2 = [2{,}0010 \quad 0{,}9087 \quad 0{,}3335 \quad 0{,}1294 \quad 0{,}0902 \quad 0{,}0000].$$

As explicações dos componentes principais podem ser observadas na Tabela 9.

Tabela 9 - Explicação dos autovalores em relação às componentes principais.

Componentes	Autovalor	% da variância	% acumulada da variância
1	2,0010	57,79	57,79
2	0,9087	26,24	84,03
3	0,3335	9,63	93,66
4	0,1294	3,74	97,39
5	0,0902	2,61	100,00

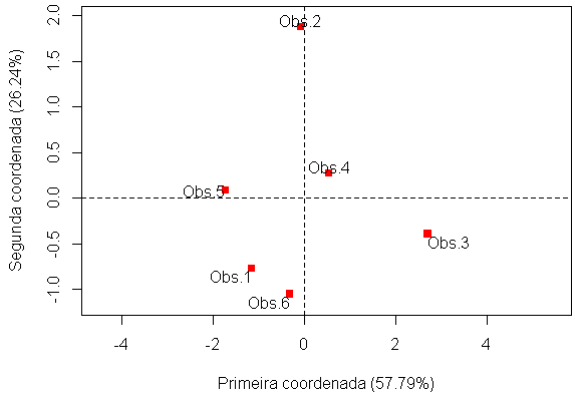
Fonte: Do autor (2019).

Ao aplicar a equação (23), geram-se os escores dos fatores globais, que são dados por

$$F = \begin{bmatrix} -1{,}1449 & -0{,}7743 & 0{,}5140 & -0{,}4545 & 0{,}3420 \\ -0{,}0773 & 1{,}8727 & -0{,}4461 & -0{,}2186 & 0{,}1244 \\ 2{,}7071 & -0{,}3946 & 0{,}0334 & 0{,}2333 & 0{,}2588 \\ 0{,}5458 & 0{,}2668 & 0{,}7408 & -0{,}1086 & -0{,}5234 \\ -1{,}7227 & 0{,}0837 & 0{,}1416 & 0{,}6649 & 0{,}0536 \\ -0{,}3081 & -1{,}0544 & -0{,}9837 & -0{,}1165 & -0{,}2554 \end{bmatrix}.$$

Utilizando-se as duas primeiras colunas de F, gera-se o gráfico da Figura 11, que revela a estrutura comum das observações em duas dimensões.

Figura 11 - Gráfico da análise global das observações nas duas primeiras componentes principais.



Fonte: Do autor (2019).

A análise global revela a estrutura comum das observações, mas, para ver como cada grupo interpreta este espaço, projeta-se o conjunto de dados de cada grupo sobre a análise global. Isto é implementado por meio da equação (26), portanto,

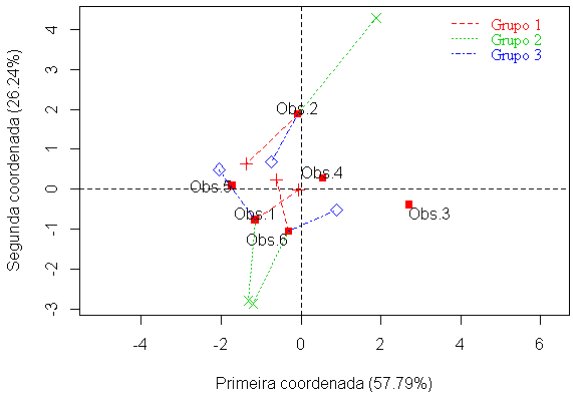
$$F_1 = \begin{bmatrix} -0{,}0616 & -0{,}0142 & -0{,}0633 & 0{,}0993 & 0{,}2630 \\ -1{,}3649 & 0{,}6426 & -1{,}5138 & 0{,}2441 & -0{,}4638 \\ 3{,}5159 & -1{,}6946 & 2{,}6486 & -1{,}8584 & 1{,}2726 \\ 1{,}0609 & -0{,}3787 & 1{,}4887 & -0{,}0842 & -0{,}3359 \\ -2{,}5365 & 1{,}2192 & -0{,}9842 & 2{,}3507 & -0{,}6133 \\ -0{,}6138 & 0{,}2256 & -1{,}5760 & -0{,}7515 & -0{,}1226 \end{bmatrix},$$

$$F_2 = \begin{bmatrix} -1{,}3206 & -2{,}8039 & 0{,}3228 & -0{,}5204 & -0{,}1175 \\ 1{,}8807 & 4{,}2906 & 0{,}0219 & -0{,}0213 & 0{,}7620 \\ 0{,}8802 & 1{,}9460 & -0{,}1173 & 0{,}4603 & 0{,}5093 \\ 0{,}4231 & 0{,}9463 & 0{,}0156 & -0{,}4333 & -0{,}2910 \\ -0{,}6635 & -1{,}5144 & -0{,}0522 & 0{,}5086 & 0{,}1740 \\ -1{,}1999 & -2{,}8646 & -0{,}1907 & 0{,}0060 & -1{,}0366 \end{bmatrix} \text{ e}$$

$$F_3 = \begin{bmatrix} -2{,}0525 & 0{,}4953 & 1{,}2827 & -0{,}9423 & 0{,}8807 \\ -0{,}7475 & 0{,}6849 & 0{,}1535 & -0{,}8785 & 0{,}0749 \\ 3{,}7253 & -1{,}4351 & -2{,}4311 & 2{,}0980 & -1{,}0056 \\ 0{,}1533 & 0{,}2329 & 0{,}7181 & 0{,}1916 & -0{,}9434 \\ -1{,}9680 & 0{,}5462 & 1{,}4611 & -0{,}8647 & 0{,}6002 \\ 0{,}8894 & -0{,}5242 & -1{,}1843 & 0{,}3959 & 0{,}3932 \end{bmatrix}.$$

Utilizando-se as duas primeiras colunas de F_1, F_2 e F_3 gera-se o gráfico da Figura 12, que demonstra como as observações se comportam em duas dimensões, mediante os grupos de variáveis.

Figura 12 - Gráfico da análise global das observações com os grupos.



Fonte: Do autor (2019).

Na Figura 12, os grupos de variáveis orbitam as observações. Assim, por meio da distância dos pontos que correspondem aos grupos, pode-se afirmar que o Grupo 2 se diferencia dos outros grupos, pois os Grupos 1 e 3 apresentam distâncias semelhantes em relação às observações grafadas.

Na Figura 12 foi apresentado apenas a relação das observações 1, 2 e 6 aos Grupos formados, por facilitar a visualização, o que não ocorreria caso fosse feito a todas as observações.

As relações entre os grupos e a solução global são analisadas calculando-se a inércia parcial de cada grupo para cada dimensão da análise global. Isto é calculado para cada grupo, como a soma das projeções ao quadrado das variáveis do vetor singular V de S, multiplicada pelo autovalor correspondente. Como os vetores singulares são normalizados, a soma das inércias parciais para todos os grupos para uma determinada dimensão é igual ao seu autovalor. Assim, ao usar a equação (27), as similaridades dos grupos em relação à primeira componente principal são dadas por

$$W_{11} = \lambda_1 \times \sum_{1}^{4} v_{p11}^2 = 2{,}0010 \times [(0{,}3434)^2 + (0{,}3310)^2 + (0{,}3494)^2 + (0{,}2494)^2]$$

$$= 0{,}8240$$

$$W_{21} = \lambda_1 \times \sum_{1}^{3} v_{p11}^2 = 2{,}0010 \times [(-0{,}2636)^2 + (-0{,}2025)^2 + (-0{,}2049)^2] = 0{,}3052$$

$$W_{31} = \lambda_1 \times \sum_{1}^{3} v_{p11}^2 = 2{,}0010 \times [(-0{,}4161)^2 + (-0{,}3922)^2 + (-0{,}3298)^2] = 0{,}8718$$

Cálculos semelhantes em relação às outras componentes resultam nos valores informados na Tabela 10.

Tabela 10 - Inércias dos grupos em cada componente principal.

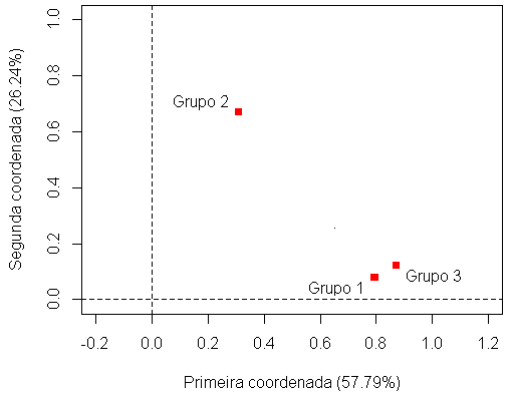
Grupo	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
1	0,8240	0,0915	0,1803	0,0626	0,0298
2	0,3052	0,6991	0,0191	0,0407	0,0340
3	0,8718	0,1181	0,1341	0,0260	0,0264
λ_i	2,0010	0,9087	0,3335	0,1294	0,0902

Fonte: Do autor (2019).

Pelos dados da Tabela 10, observa-se, em relação à primeira componente principal, que os Grupos 1 e 3 apresentam forte similaridade entre si, respectivamente 0,8240 e 0,8718; o Grupo 2 difere dos outros. Em relação à segunda componente, pode-se dizer que os Grupos 1 e 3 são similarmente fracos, e o Grupo 2 continua sendo original. Outras análises podem ser feitas em relação às outras componentes, mas como toda explicação já foi dada pela primeira componente, com a maior explicação (57,79%), nesse caso não se faz necessária outra explicação além da que foi dada pela primeira componente principal.

A partir das inércias obtidas dos grupos na Tabela 10, visando uma melhor interpretação, o gráfico das inércias, Figura 13 é gerado, mostrando que há relação forte entre os Grupos 1 e 3 em relação à primeira componente, e confirma a originalidade do Grupo 2.

Figura 13 - Gráfico das inércias dos grupos



Fonte: Do autor (2019).

6.2 Uso da técnica MFA com redução de dimensão utilizando a *projection pursuit* por meio do índice proposto

Utilizando-se a matriz $X_{6\times10}$ apresentada na Tabela 5 da seção 6.1, ao aplicar a *projection pursuit* com o novo índice proposto apresentado pela equação (34) e novo

algoritmo apresentado na seção 4.2, obtêm-se as matrizes de projeções para cada grupo, com os índices de projeções, ambos mostrados na Tabela 11.

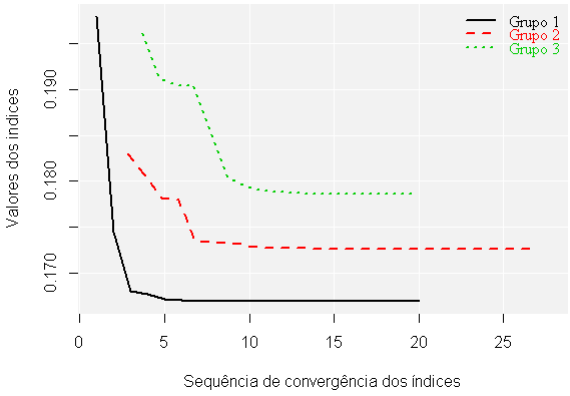
Tabela 11 - Matriz de projeção de cada grupo X^j em $X_{6\times10}$.

Projeção	Grupo 1		Grupo 2		Grupo 3	
	Vetor 1	Vetor 2	Vetor 1	Vetor 2	Vetor 1	Vetor 2
1	-0,243820	0,915174	-0,777026	-0,034119	0,520753	0,365951
2	0,682251	0,001828	-0,574801	-0,368557	-0,063393	0,917547
3	0,654536	0,229692	0,256581	-0,928979	0,851351	-0,155523
4	0,216028	0,331201	-	-	-	-
Índice de projeção	0,1668638		0,1725528		0,1786068	

Fonte: Do autor (2019).

Na Figura 14 é mostrado como foi a convergência do novo índice para cada grupo de variáveis, nas otimizações numéricas.

Figura 14 - Gráfico das convergências do novo índice proposto para cada grupo.



Fonte: Do autor (2019).

Ao aplicar as matrizes de projeções nos grupos em $X_{6\times10}$ obtiveram-se os dados da Tabela 12, que representam os dados com dimensão reduzida em duas dimensões para cada conjunto X^j, de modo que a matriz $X_{6\times10}$ será representada pela nova matriz $\tilde{X}_{6\times6}$.

Tabela 12 - Matriz $\tilde{X}_{6\times6}$ dos dados com as dimensões reduzidas de $X_{6\times10}$.

Obs.	Grupo 1		Grupo 2		Grupo 3	
	Projeção 1	Projeção 2	Projeção 1	Projeção 2	Projeção 1	Projeção 2
1	4,268384	3,420136	-1,845168	-3,520194	3,842017	2,964580
2	3,812184	2,839240	0,571221	-1,337686	3,893913	1,874408
3	5,875395	4,976865	-0,301725	-2,104094	1,773535	0,744569
4	4,831857	3,990264	-0,237184	-2,579577	3,218560	1,872230
5	3,170266	2,324246	-1,551706	-3,161522	3,820329	2,908277
6	4,035601	3,251591	-1,461030	-3,914222	2,884163	1,757145

Fonte: Do autor (2019).

Ao aplicar a metodologia da técnica MFA, descrita na seção 3.6, nos grupos de variáveis na Tabela 12, as explicações dos componentes principais são dadas na Tabela 13.

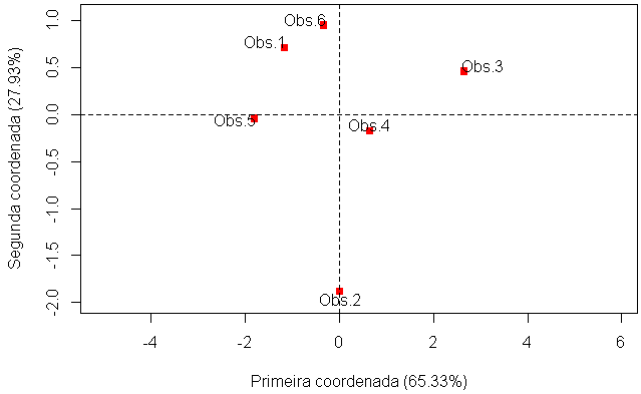
Tabela 13 - Explicação dos autovalores em relação aos componentes principais.

Componentes	Autovalor	% da variância	% acumulada da variância
1	2,0306	65,33	65,33
2	0,8681	27,93	93,26
3	0,1678	5,40	98,66
4	0,0382	1,23	99,89
5	0,0034	0,11	100,00

Fonte: Do autor (2019).

Ao aplicar-se a equação (23), geram-se os escores dos fatores globais, que são utilizados para gerar o gráfico da Figura 15, que revela a estrutura comum das observações em duas dimensões.

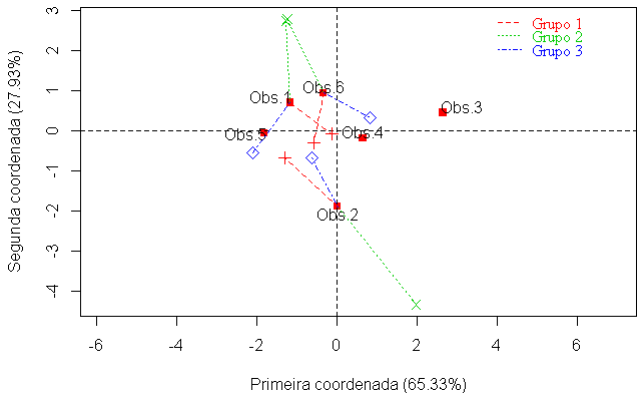
Figura 15 - Gráfico da análise global das observações dos dados com dimensões reduzidas nas duas primeiras componentes principais.



Fonte: Do autor (2019).

Por meio da equação (26), gera-se o gráfico da Figura 16, que demonstra como as observações se comportam em duas dimensões, mediante os grupos de variáveis com dimensões reduzidas.

Figura 16 - Gráfico da análise global das observações com os grupos com dimensões reduzidas nas duas primeiras componentes principais.



Fonte: Do autor (2019).

De modo análogo ao gráfico da Figura 12, o gráfico da Figura 16 sugere que o Grupo 2 difere dos outros Grupos, pois os Grupos 1 e 3 apresentam distâncias semelhantes em relação às observações grafadas.

Aqui, novamente, ao aplicar a equação (27), as similaridades dos Grupos são descritas na Tabela 14, em relação às componentes principais,.

Tabela 14 - Inércias dos grupos em cada componente principal.

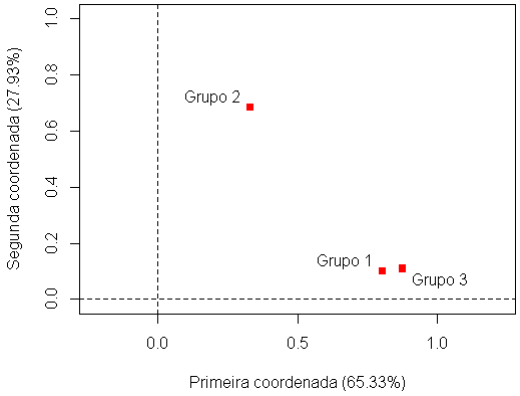
Grupo	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
1	0,8278	0,0956	0,0768	0,0009	0,0001
2	0,3300	0,6705	0,0065	0,0273	0,0010
3	0,8727	0,1021	0,0846	0,0100	0,0023
λ_i	2,0306	0,8681	0,1678	0,0382	0,0034

Fonte: Do autor (2019).

As mesmas observações feitas em relação à Tabela 10, que trata das similaridades dos grupos de variáveis com dimensões originais, se aplicam a Tabela 14, ou seja, em relação à primeira componente principal, os Grupos 1 e 3 apresentam forte similaridade entre si, e o Grupo 2 difere dos outros.

Novamente a partir das inércias obtidas dos grupos na Tabela 14, visando uma melhor interpretação, o gráfico das inércias (Figura 17) é gerado, mostrando que há relação forte entre os Grupos 1 e 3, e confirmando a originalidade do Grupo 2.

Figura 17 - Gráfico das inércias dos grupos com as dimensões reduzidas.



Fonte: Do autor (2019).

Como pode-se perceber, os resultados das similaridades com dados com dimensão reduzida produziram os mesmos resultados dos dados com as dimensões originais, o que mostra a viabilidade do novo índice proposto. E temos ainda que nos dados com dimensão reduzida, as explicações nas duas primeiras componentes principais foi de 93,26% (Tabela 13), bem acima dos dados com dimensões originais com 84,03% (Tabela 9).

Nesse ultimo exemplo embora os resultados sejam praticamente os mesmos do exemplo dado na seção 6.1, ele é computacionalmente mais caro, pois foi preciso utilizar da *projection pursuit* para reduzir as dimensões dos grupos de variáveis para depois aplicar a técnica MFA, sendo que ao aplicar a técnica diretamente surtiu o mesmo resultado. Mas o objetivo desse exemplo foi mostrar a viabilidade do novo índice proposto.

Lembrando que o propósito da *projection pursuit* é reduzir as dimensão em dados em alta dimensão. Ao aplicar a técnica MFA em dados de alta dimensão ocorre o mesmo que no PCA, as explicações nas componentes principais são diluídas, o que prejudica a aplicabilidade da técnica, devido a perda de informações. Mas se ao reduzir as dimensões houver um aumento das explicações nas primeiras componentes, isso assegura maior confiabilidade nas análises, esse aumento nas explicação nas primeiras componentes pode ser visto no último exemplo dado.

7 CONCLUSÕES

Em consonância com objetivo, o novo índice proposto mostrou-se viável na redução dos dados para aplicação na técnica MFA, sendo recomendado nas situações em que os grupos apresentam baixa ou alta heterogeneidade, e forte grau de correlação entre as variáveis ($\rho = 0{,}9$).

Os índices da seção 3.4 que foram testados, obtiveram baixa concordância nos resultados, mostrando que não são apropriados para o propósito de redução de dados na aplicação da técnica MFA.

De modo geral, os índices avaliados são afetados pelo aumento do número de grupos, em função dos cenários avaliados, apresentando melhores resultados nas correlação em que $\rho = 0{,}2$ e $\rho = 0{,}5$.

Como consequência ao reduzir as dimensões dos dados, ocorre uma maior explicação nas primeiras componentes principais ao aplicar à técnica MFA, o que beneficia o uso do novo índice proposto.

A limitação do novo índice de ser recomendado para grupos com forte correlação, mostra que outras pesquisas podem ser feitas a fim de melhorar a aproximação, servindo de base para pesquisas futuras.

Uma outra abordagem em pesquisas futuras seria a utilização da *projection pursuit* na redução das dimensões das observações, o que seria de grande utilidade ao trabalhar com dados desbalanceados, pois até o momento todas as pesquisas com essa técnica visa reduzir as dimensões somente dos atributos.

REFERÊNCIAS

ABDI, H.; VALENTIN, D. Multiple factor analysis (MFA). In: SALKIND, N. J. (Ed.). **Encyclopedia of measurement and statistics**. Thousand Oaks: Sage, 2007. p. 657-663.

ABDI, H.; WILLIAMS, L. Principal component analysis. **WIREs Computational Statistics**, New York, v. 2, p. 433-459, 2010.

ABDI, H.; WILLIAMS, L.; VALENTIN, D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. **WIREs Computational Statistics**, New York, v. 5, n. 2, p. 149-179, Mar./Apr. 2013.

ASIMOV, D. The grand tour: a tool for viewing multidimensional data. **SIAM Journal of Scientific and Statistical Computing**, Philadelphia, v. 6, n. 1, p. 128-143, 1985.

ASIMOV, D.; BUJA, A. The grand tour via geodesic interpolation of 2-frames. In: VISUAL DATA EXPLORATION AND ANALYSIS, SYMPOSIUM ON ELECTRONIC IMAGING SCIENCE AND TECHNOLOGY, 1994, San Jose. **Proceedings…** San Jose: SPIE, 1994. Disponível em: <https://www.nas.nasa.gov/assets/pdf/techreports/1994/rnr-94-004.pdf>. Acesso em: 10 mar. 2018.

BÉCUE-BERTAUT, M.; PAGÈS, J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. **Computational Statistics & Data Analysis**, New York, v. 52, p. 3255-3268, 2008.

BUJA, A. et al. Computational methods for high-dimensional rotations in data visualization. In: RAO, C. R.; WEGMAN, E. J.; SOLKA, J. L. (Ed.). **Handbook of statistics:** data mining and visualization. Amsterdam: Elsevier, 2005. p. 391-413.

CIRILLO, M. A. et al. Generalized variances ratio test for comparing k covariance matrices from dependent normal populations. **Journal of Modern Applied Statistical Methods**, Berlin, v. 9, n. 2, p. 369-378, 2010.

COOK, D.; BUJA, A.; CABRERA, J. Projection pursuit indexes based on orthonormal function expansions. **Journal of Computational and Graphical Statistics**, Alexandria, v. 2, n. 3, p. 225-250, 1993.

COOK, D. et al. Grand tour and projection pursuit. **Journal of Computational and Graphical Statistics**, Alexandria, v. 4, n. 3, p. 155-172, 1995.

COOK, D. et al. Grand tours, projection pursuit guided tours and manual controls. In: CHEN, C. et al. (Ed.). **Handbook of data visualization:** Springer handbooks of computational statistics. 2nd ed. New York: Springer, 2008. chap. 3, p. 295-314.

COOK, D.; SWAYNE, D. F. **Interactive and dynamic graphics for data analysis:** with R and GGobi. New York: Springer, 2007.

COOREN, Y.; CLERC, M.; SIARRY, P. Performance evaluation of TRIBES, an adaptive particle swarm optimization algorithm. **Swarm Intelligence**, Boca Raton, v. 3, p. 149-178, 2009.

CROUX, C.; FILZMOSER, P.; OLIVEIRA, M. Algorithms for projection-pursuit robust principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 87, p. 218-225, 2007.

DIACONIS, P.; FREEDMAN, D. Asymptotics of graphical projection pursuit. **Annals of Statistics**, Hayward, v. 12, p. 793-815, 1984.

ESCOFIER, B.; PAGÈS, J. **Analyse factorielles simples et multiples**. Paris: Dunod, 1990.

ESCOFIER, B.; PAGÈS, J. **Analyses factorielles simples et multiples:** objectifs, méthodes et interprétation. 4th ed. Paris: Dunod, 2008. 318 p.

ESCOFIER, B.; PAGÈS, J. **Comparaison de groupes de variables définies sur le même ensemble d´individus:** un exemple d´applications. Le Chesnay: Institut National de Recherche en Informatique et en Automatique, 1982. (Working Paper, 165).

ESCOFIER, B.; PAGÈS, J. Multiple factor analysis (AFUMULT package). **Computational Statistics & Data Analysis**, New York, v. 18, p. 121-140, 1994.

ESPEZUA, S. et al. A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets. **Neurocomputing**, New York, v. 149, p. 767-776, 2015.

FRIEDMAN, J. H. Exploratory projection pursuit. **American Statistical Association**, New York, v. 82, n. 397, p. 249-266, 1987.

FRIEDMAN, J. H.; TUKEY, J. W. A projection pursuit algorithm for exploratory data analysis. **IEEE Transaction on Computers**, New York, v. 23, n. 9, p. 881-890, 1974.

GUO, Q. et al. Sequential projection pursuit using genetic algorithms for data mining of analytical data. **Analytical Chemistry**, Washington, v. 72, n. 13, p. 2846-2855, 2000.

HALL, P. On polynomial-based projection indexes for exploratory projection pursuit. **Annals of Statistics**, Hayward, v. 17, p. 589-605, 1989.

HASTIE, T.; BUJA, A.; TIBSHIRANI, R. Penalized discriminant analysis. **Annals of Statistics**, Hayward, v. 23, n. 1, p. 73-102, 1995.

HUBER, P. J. Projection pursuit. **Annals of Statistics**, Hayward, v. 13, n. 2, p. 435-475, 1985.

HURLEY, C.; BUJA, A. Analyzing high-dimensional data with motion graphics. **SIAM Journal of Scientific and Statistical Computing**, Philadelphia, v. 11, n. 6, p. 1193-1211, 1990.

HYVARINEN, A.; OJA, E. Independent component analysis: algorithms and applications. **Neural Networks**, New York, v. 13, n. 4/5, p. 411-430, 2000.

JONES, M. C.; SIBSON, R. What is projection pursuit: with discussion. **Journal of the Royal Statistical Society**, Series A, London, v. 150, p. 1-36, 1987.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 4., 1995, Perth. **Proceedings…** Perth: IEEE, 1995. p. 1942-1948.

KRUSKAL, J. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". **Statistical Computation**, New York, p. 427-440, 1969.

LEE, E. K.; COOK, D. A projection pursuit index for large p small n data. **Statistics and Computing**, London, v. 20, n. 3, p. 381-392, 2010.

LEE, E. K. et al. Projection pursuit for exploratory supervised classification. **Journal of Computational and Graphical Statistics**, Alexandria, v. 14, n. 4, p. 831-846, 2005.

MARTINEZ, W. L.; MARTINEZ, A. R. **Computational statistics handbook with MATLAB**. 2nd ed. New York: Chapman & Hall/CRC, 2007. 794 p.

MARTINEZ, W. L.; MARTINEZ, A. R.; SOLKA, J. **Exploratory data analysis with MATLAB**. 2nd ed. New York: Chapman & Hall/CRC, 2010. 499 p.

MORTON, S. **Interpretable projection Pursuit**. Stanford: Stanford University Press, Laboratory for Computational Statistics, 1989. (Technical Report, 106).

OSSANI, P. C.; CIRILLO, M. A. **MVar.pt:** análise multivariada. R Package Version 2.0.2. 2018. Disponível em: <https://CRAN.R-project.org/package=MVar.pt>. Acesso em: 10 jun. 2018.

PAGÈS, J. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. **Revue de Statistique Appliquée**, Paris, v. 50, n. 4, p. 5-37, 2002.

PAGÈS, J. Multiple factor analysis: main features and application to sensory data. **Revista Colombiana de Estadística**, Bogotá, v. 27, n. 1, p. 1-26, 2004.

PENA, D.; PRIETO, F. Cluster identification using projections. **Journal of the American Statistical Association**, New York, v. 96, n. 456, p. 1433-1445, 2001.

POSSE, C. Projection pursuit exploratory data analysis. **Computational Statistics and Data Analysis**, Amsterdam, v. 20, p. 669-687, 1995a.

POSSE, C. Tools for two-dimensional exploratory projection pursuit. **Journal of Computational and Graphical Statistics**, Alexandria, v. 4, p. 83-100, 1995b.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: <http://www.R-project.org>. Acesso em: 10 abr. 2018.

WEBB-ROBERTSON, B. et al. An improved optimization algorithm and bayes factor termination criterion for sequential projection pursuit. **Chemometrics and Intelligent Laboratory Systems**, Berlin, v. 77, n. 1/2, p. 149-160, 2005.

APÊNDICE

Código R usado nas simulações dos dados.

O pacote MVar.pt usado nas simulações utiliza o algoritmo *grand tour simulated annealing* de otimização (OSSANI; CIRILLO, 2018).

```
# Rotina desenvolvida por Paulo Cesar Ossani para simular dados para o
# metodo MFA e comparar os resultados apos usar a projection pursuit.

rm(list=ls(all=T)) # limpa a memoria do sistema

library(MVar.pt)
library(MASS)
library(mvtnorm)

source("Funcoes_Simulacao.R") # funcoes usadas nas simulacoes

##### INICIO - Cenario para simulacoes #####
## Indices para a busca de projecao usados nas simulacoes
# Findex <- "holes"
# Findex <- "Moment"
# Findex <- "chi"
Findex <- "MF"
# Findex <- "cm"
# Findex <- "LDA"
# Findex <- "PDA"
# Findex <- "Lr"
# Findex <- "FriedmanTukey"
# Findex <- "Entropy"
# Findex <- "Legendre"
# Findex <- "LaguerreFourier"
# Findex <- "Hermite"
# Findex <- "NaturalHermite"

Dim       <- 2      # dimensao de reducao
MetOtimiz <- "GTSA" # metodo de otimizacao
Sphere    <- FALSE  # utiliza dados esfericos
NumComp   <- 1      # numero de componentes comparadas
VarOrGroups <- "V"  # simula G - grupos ou V - variaveis correlacionadas
NumIter   <- 5000   # numero maximo de iteracoes
VrCool    <- 0.95   # valor coolling
VrEps     <- 1e-4   # valor de aproximacao
VrHalf    <- 30     # valor half

NumSimula <- 100    # numero de simulacoes
VarGrupos <- c(3,5,10,15) # numero de variaveis em cada grupo
QGrupos   <- c(3,7,10) # numero de grupos
LinGrupos <- c(20,50,70,100,150) # numero de linhas em cada grupo
NivRho    <- c(0.2,0.5,0.7,0.9) # nivel de correlacao a simular entre as
variaveis
NivDelta  <- c(2,8) # grau de heterogeneidade entre as matrizes
                    # caso simulacao de grupos correlacionados

N.VarGrupos <- length(VarGrupos) # no de elementos em NumVarGrupos
N.Grupos    <- length(QGrupos)   # no de elementos em NumGrupos
N.LinGrupos <- length(LinGrupos) # no de elementos em NumLinGrupos
N.Rho       <- length(NivRho)    # numero de elementos em Rho
```

```
# numero de cenarios das simulacoes
NumCenario  <- N.VarGrupos * N.Grupos * N.LinGrupos * N.Rho
##### FIM - Cenario para simulacoes #####

##### INICIO - Simulacoes #####
ResultMatrix <- NULL # matriz com os resultados

Findex <- toupper(Findex) # transforma em maiusculo

VarOrGroups <- toupper(VarOrGroups) # transforma em maiusculo

if (VarOrGroups == "V") NivDelta <- c(1)

N.Delta <- length(NivDelta) # numero de elementos em Delta

Tot.Simula <- NumSimula*length(VarGrupos)*length(QGrupos)*
              length(LinGrupos)*length(NivRho)*length(NivDelta)

Current.Simula <- 1 # contador das simulacoes

NLGrupos <- 1
while(NLGrupos <= N.LinGrupos) { # numero de linhas em cada grupo
  NVGrupos <- 1
  while(NVGrupos <= N.VarGrupos) { # no. de variaveis em cada grupo
    NGrupos  <- 1
    while(NGrupos <= N.Grupos) { # numero de grupos
      NVRho <- 1
      while(NVRho <= N.Rho) { # nivel correlacao a simular entre variáveis
        NDelta <- 1
        while(NDelta <= N.Delta) {# grau de heterogeneidade entre variáveis
        NumVarGrupos <- VarGrupos[NVGrupos] # numero variaveis cada grupo
        NumGrupos    <- QGrupos[NGrupos] # numero de grupos
        NumLinGrupos <- LinGrupos[NLGrupos] # numero linhas em cada grupo
        Rho   <- NivRho[NVRho] # nivel correlacao a simular entre variáveis
        # grau de heterogeneidade entre as matrizes caso
        # simulacao de grupos correlacionados
        Delta <- NivDelta[NDelta]
        Concorda <- NULL # resultado da simulacao

Sim <- 1
while(Sim <= NumSimula) { # simulacoes
  ### INICIO - cria a base de dados para as analises ###
  ## simulacoes para funcao indice Findex diferente de LDA, PDA ou Lr
  if (!(Findex %in% c("LDA", "PDA", "LR"))) {
    if (VarOrGroups == "V") { # simula variaveis correlacionadas
       Z <- NULL
       i <- 1
       while(i <= NumGrupos) {
         M <- NULL
         j <- 1
         while(j <= NumVarGrupos) {
           M <- cbind(M,rnorm(NumLinGrupos))
           j <- j + 1
         }
         Res <- SimCor(M, Rho) # simula correlação das colunas
                               # nas bases de dados
         Z <- cbind(Z, Res)
         i <- i + 1
       }
    } else {  # simula grupos correlacionados
         Z <- Cirillo(NumVarGrupos,NumGrupos,Rho,Delta,NumLinGrupos,100)
```

```
  }

  Z <- as.data.frame(Z)
  ### FIM - cria a base de dados para as analises ###

  ### INICIO - simulacao com MFA ###
  Grup <- rep(NumVarGrupos, NumGrupos) # parametros p/ numero de colunas
  ### Metodo MFA
  MF  <- MFA(Z,Groups = Grup, TypeGroups = c(rep("n",NumGrupos)))

  ### Metodo MFA com as dimensoes reduzidas - Busca de projecao
  Zpp <- NULL
  Gr  <- rep(NumVarGrupos, NumGrupos)  # numero elementos em cada grupo
  j   <- 1
  k   <- Gr[1]
  i   <- 1
  while(i <= NumGrupos) {
    Res <- PP_Optimizer(Data = as.data.frame(Z[,j:k]), Findex = Findex,
                        OptMethod = MetOtimiz, DimProj =Dim, Half =VrHalf
                        Sphere = Sphere, Weight=TRUE, Lambda = 0.1, r =1,
                        Cooling = VrCool, Eps = VrEps, Maxiter = NumIter)

    Zpp <- cbind(Zpp, Res$Proj.Data)
    j <- j + Gr[i] # coluna inicial do grupo de variaveis
    k <- k + Gr[i+ifelse(i!=NumGrupos,1,0)] #coluna final grupo variaveis
    i <- i + 1
  }

  Grup <- rep(Dim,NumGrupos) # parametros p/ no. de colunas em cada grupo

  MFpp <- MFA(as.data.frame(Zpp), Groups = Grup,
              TypeGroups = c(rep("n", NumGrupos)))
  ### FIM - simulacao com MFA ###
} else { # simulacoes para funcao indice Findex = LDA, PDA ou Lr
  if (VarOrGroups == "V") {  # simula variaveis correlacionadas
     ## Estabelece a correlação para cada grupo em Z
     Z1 <- NULL
     Z2 <- NULL
     i <- 1
     while (i <= NumGrupos) {
       M <- NULL
       j <- 1
       while (j <= NumVarGrupos) {
         M <- cbind(M, rnorm(NumLinGrupos))
         j <- j + 1
       }
       Res <- SimCor(M, Rho) # simula correlação das colunas
                             # das bases de dados
       Z1 <- cbind(Z1,Res)
       AA <- rep(paste("Grupo",i), NumLinGrupos, step = "")
       M <- cbind(as.data.frame(Res), AA)
       Z2 <- rbind(Z2, M)
       i <- i + 1
     }
     ### FIM - cria a base de dados para as analises
   } else { # simula grupos correlacionados
     Z <- Cirillo(NumVarGrupos, NumGrupos, Rho, Delta, NumLinGrupos, 100)
     ##### Estabelece a correlação para cada grupo em Z ######
     Z1 <- NULL
     Z2 <- NULL
     i  <- 1
```

```
    j  <- 1
    while (i <= NumGrupos) {
      Res <- Z[,j:(j+(NumVarGrupos-1))] # grupos formados
      Z1  <- cbind(Z1,Res)
      AA  <- rep(paste("Grupo",i), NumLinGrupos, step="")
      M   <- cbind(as.data.frame(Res), AA)
      Z2  <- rbind(Z2,M)
      j   <- j + NumVarGrupos
      i   <- i + 1
    }
    ### FIM - cria a base de dados para as analises
  }

  ### INICIO - simulacao com MFA ###
  Z1 <- as.data.frame(Z1)

  Grup <- rep(NumVarGrupos, NumGrupos) # parametros para o numero colunas

  ### Metodo MFA
  MF <- MFA(Z1, Groups = Grup, TypeGroups = c(rep("n", NumGrupos)))

  ### Metodo MFA com PP
  Z2    <- as.data.frame(Z2)
  Class <- Z2[,(NumVarGrupos+1)]
  Z2  <- Z2[,1:NumVarGrupos]
  Gr  <- rep(NumVarGrupos, NumGrupos) # numero de elementos em cada grupo
  Res <- PP_Optimizer(Data = as.data.frame(Z2), Class = Class, Eps=VrEps,
                      Findex = Findex, OptMethod=MetOtimiz, DimProj=Dim,
                      Sphere = Sphere, Weight = TRUE, Lambda = 0.1, r=1,
                      Cooling = VrCool, Maxiter = NumIter, Half = VrHalf)
  Zpp <- NULL
  j   <- 1
  Gr  <- rep(NumLinGrupos, NumGrupos)
  k   <- Gr[1]
  i   <- 1
  while (i <= NumGrupos) {
    Zpp <- cbind(Zpp, as.matrix(Res$Proj.Data[j:k, 1:Dim]))
    j <- j + Gr[i] # coluna inicial do grupo de variaveis
    k <- k + Gr[i+ifelse(i!=NumGrupos,1,0)] # col. final grupo variaveis
    i <- i + 1
  }

  Grup <- rep(Dim,NumGrupos) # parametros p/ no. colunas em cada grupo

  MFpp <- MFA(as.data.frame(Zpp), Groups = Grup,
              TypeGroups = c(rep("n", NumGrupos)))
  ### FIM - simulacao com MFA ###
}

#### pode analizar ate a sexta componente
Help.Con <- NULL
Con <- 1
while(Con <= NumComp) {
  # funcao retorna resultados da comparacao dos metodos MFA e o proposto
  ResComp  <- CompResult(MF$MatrixEscVar[,Con], MFpp$MatrixEscVar[,Con])
  Help.Con <- cbind(Help.Con, ResComp$NTrue)
  Con <- Con + 1
}
Concorda <- rbind(Concorda,Help.Con[order(Help.Con, decreasing = T)][1])
Esferico <- ifelse(Sphere, " Dados Esfericos", " Dados NAO Esfericos")
VarGrop  <- ifelse(VarOrGroups == "V"," Variaveis Correlacionadas",
```

```
                    " Grupos Correlacionados")
  Otimizador <-  ifelse(toupper(MetOtimiz) == "SA"," Otimizador: SA",
                        " Otimizador: GTSA")

  print(paste(Help.Con[order(Help.Con,decreasing = T)][1], "- Simulacao:",
        Current.Simula, "de", Tot.Simula, " Metodo:", Findex, Esferico,
        VarGrop,Otimizador," Nº Variaveis Grupos:", NumVarGrupos,
        " Nº Grupos:", NumGrupos, " Nº Linhas:", NumLinGrupos,
        " Nivel Correlacao Variaveis:", Rho, " Grau Heterogeneidade:",
        Delta, " Half:" , VrHalf))

  Current.Simula <- Current.Simula + 1
  Sim <- Sim + 1
}

### INICIO - Resumo dos resultados ###
NumElem  <- length(Concorda)
ConTotal <- length(Concorda[Concorda == "T"]) # total concordam totalmente
ConParci <- length(Concorda[Concorda == "P"]) # tot. concordam parcialmente
NaoConrd <- length(Concorda[Concorda == "N"])# total que nao concordam

Resultados <- cbind((ConTotal / NumElem),(ConParci / NumElem),
                    (NaoConrd / NumElem),ConTotal, ConParci, NaoConrd,
                    sum(c(ConTotal,ConParci, NaoConrd)), NumVarGrupos,
                    NumGrupos, NumLinGrupos, Rho, Delta)

rownames(Resultados) <- c("Resultados")

colnames(Resultados) <- c("Totalmente", "Parcialmente", "Nao Concorda",
                          "Qtd.Sim.Totalmente", "Qtd.Sim.Parcialmente",
                          "Qtd.Sim. Nao Concorda", "Num.Simulacoes",
                          "Num.de Variaveis", "Num.de Grupos",
                          "Num.de Observacoes", "Correlacao Variaveis",
                          "Heterogeneidade")

ResultMatrix <- rbind(ResultMatrix, Resultados)
### FIM - Resumo dos resultados ###

         NDelta <- NDelta + 1  # nivel correlacao a simular entre os grupos
        }
       NVRho <- NVRho + 1  # nivel correlacao a simular entre as variáveis
      }
     NGrupos <- NGrupos + 1 # numero de grupos
    }
  NVGrupos <- NVGrupos + 1  # numero de variaveis em cada grupo
  }

  ### INICIO - salva resultados por grupos de linhas dado por LinGrupos ###
  NiveisLinhas <- paste("- Num.Linhas", LinGrupos[NLGrupos])
  Otimizador <- ifelse(toupper(MetOtimiz) == "SA", "- Otimizador SA",
                       "- Otimizador GTSA")
  File <- paste("Resultados - ", Findex,NiveisLinhas, Otimizador, ".csv")

  write.table(file=File, ResultMatrix, sep=";", dec=",",row.names = FALSE)
  ResultMatrix <- NULL # Matriz com os resultados
  ### FIM - salva os resultados por grupos de linhas dado por LinGrupos ###

 NLGrupos <- NLGrupos + 1  # numero de linhas em cada grupo
}
##### FIM - Simulacoes #####
```

Funções que formam o arquivo “Funcoes_Simulacao.R”

```
SimCor <- function(Z, Rho) {
  # Funcao gera vetores correlacionados a nivel de Rho.

  # Entrada
  # Z   - Matriz referencia.
  # Rho - Nivel de correlacao entre as colunas.

  # Saida:
  # M - Matriz com as colunas correlacionadas a nivel de Rho

  NVar  <- ncol(Z)
  NLin  <- nrow(Z)
  Media <- c(1:NVar) # media diferente em cada coluna na solucao final
  Cov   <- matrix(Rho, nrow = NVar, ncol = NVar) +
                  diag(NVar)*(1-Rho) # matriz correlacao
  M     <- mvrnorm(n = NLin , mu = Media, Sigma = Cov)
  return(M)
}


CompResult <- function(ResMFA, ResRM) {
  # Rotina desenvolvida por Paulo Cesar Ossani para comparar os resultados
  # do metodo proposto com os resultados da tecnica MFA

  # Entrada:
  # ResMFA - Matriz com resultados da componente principal no MFA
  # ResRM  - Matriz com resultados do metodo proposto

  # Saida:
  # PosMFA - Matriz com as posicoes onde os grupos sao similares no MFA
  # PosRM  - Matriz posicoes onde os grupos sao similares metodo proposto
  # MatCompFinal - Matriz com os resultados do MFA e do metodo
  #                proposto e as comparacoes
  # NTrue - Diz se os metodos concordam totalmente (T)
  #         nao concordam (N) ou parcialmente (P)

  ### INICIO - similaridade no MFA ###
  Round <- 4 # numero de casas decimais

  # faixa de comparacao das similaridade para a tecnica MFA
  VrangeMFA <- as.matrix(rbind(c(0, 0.39), c(0.4, 0.69), c(0.7, 1.1)))

  ResMFA <- round(ResMFA, Round) # resultados do MFA

  VetCompMFA <- cbind(ResMFA,0) # vetor com as comparações

  NumLin <- nrow(VetCompMFA)

  h <- 1
  while (h <= NumLin){
    i <- h
    while (i <= NumLin) { # linhas do MFA
      j <- 1
      while(j <= nrow(VrangeMFA)) { # linhas do Vrange
        if (ResMFA[i] >= VrangeMFA[j,1] && ResMFA[i] < VrangeMFA[j,2]) {
            VetCompMFA[i,2] <- j
        }
        j <- j + 1
```

```
    }
    i <- i + 1
  }
  h <- h + 1
}
### FIM - similaridade no MFA ###

### INICIO - similaridade no metodo proposto ###
VrangeRM  <- VrangeMFA # faixa comparacao similaridade metodo proposto
ResRM     <- round(ResRM, Round) # resultados do MFA
VetCompRM <- cbind(ResRM,0) # vetor com as comparações
NumLin    <- nrow(VetCompRM)

h <- 1
while(h <= NumLin) {
  i <- h
  while(i <= NumLin) { # linhas do MFA
    j <- 1
    while(j <= nrow(VrangeRM)) { # linhas do Vrange
      if (ResRM[i] >= VrangeRM[j,1] && ResRM[i] < VrangeRM[j,2]) {
         VetCompRM[i,2] <- j
      }
      j <- j + 1
    }
    i <- i + 1
  }
  h <- h + 1
}
### FIM - similaridade no metodo proposto ###

### INICIO-compara convergencia solucoes entre MFA e metodo proposto ###
## matriz com todas as comparacoes
MatCompFinal <- data.frame(cbind(VetCompMFA, VetCompRM))
colnames(MatCompFinal) <- c("MFA", "Comp.", "MFApp", "Comp")

### INICIO - Iguala os codigos ###
Cod <- c("a", "b", "c", "d", "e", "f", "g", "h", "i", "j", "k","l", "m",
         "n", "o", "p", "q", "r", "s", "t", "u", "v","w", "x", "y", "z")

n <- nrow(MatCompFinal)
a <- MatCompFinal
i <- 1
while(i <= n) {
  vla <- a[i,2]
  vlb <- a[i,4]
  if ( i > 1) {
    if (!is.na(suppressWarnings(as.numeric(a[i,2]))))
       a[a[,2] == vla,2] = Cod[i] # substibui codigos por letras em Cod

    if (!is.na(suppressWarnings(as.numeric(a[i,4]))))
       a[a[,4] == vlb,4] = Cod[i] # substibui codigos por letras em Cod
  } else {
       a[a[,2] == vla,2] = Cod[i] # substibui codigos por letras em Cod
       a[a[,4] == vlb,4] = Cod[i] # substibui codigos por letras em Cod
  }
  i <- i + 1
}

# transforma letras em numeros
i <- 1
while(i <= length(Cod)) {
```

```
  a[a[,2] == Cod[i],2] = i # substibui codigos por letras em Cod
  a[a[,4] == Cod[i],4] = i # substibui codigos por letras em Cod
  i <- i + 1
}
### FIM - Iguala os codigos ###

MatMFA <- MatCompFinal[,2] # comparacoes do MFA
MatRM  <- MatCompFinal[,4] # comparacoes no metodo proposto

# encontra as posicoes onde os grupos sao semelhantes no MFA
NumLinMFA <- nrow(VrangeMFA)
FaixaMFA  <- 1:NumLinMFA
PosMFA <- NULL # vetor posicao onde os grupos sao iguais no MFA

i <- 1
while(i <= NumLinMFA) {
  Equal <- MatMFA == i
  if (!is.na(pmatch(TRUE, Equal))) {
     Pos.MFA <- rep(0, nrow(VetCompMFA))
     j <- 1
     while(j <= nrow(VetCompMFA)) {
       if (Equal[j] == TRUE) Pos.MFA[j] <- j
      j <- j + 1
    }
    if (sum(Pos.MFA) != 0) {
      if (length(Pos.MFA) == 1) Pos.MFA <- 0
      PosMFA <- cbind(PosMFA, Pos.MFA)
    }
  }
  i <- i + 1
}

# encontra as posicoes onde os grupos sao semelhantes no metodo proposto
NumLinRM <- nrow(VrangeRM)
FaixaRM  <- c(1:NumLinRM)
PosRM    <- NULL # vetor posicao onde grupos sao iguais metodo proposto
i <- 1
while(i <= NumLinRM) {
  Equal  <- MatRM == i
  if (!is.na(pmatch(TRUE, Equal))) {
    Pos.RM <- rep(0,nrow(VetCompRM))
    j <- 1
    while(j <= nrow(VetCompRM)) {
      if (Equal[j] == TRUE) Pos.RM[j] <- j
      j <- j + 1
    }
    if (sum(Pos.RM) != 0) {
      if (length(Pos.RM) == 1) Pos.RM <- 0
      PosRM <- cbind(PosRM, Pos.RM)
    }
  }
  i <- i + 1
}

# Compara os resultados do MFA com metodo proposto
NColMFA <- ncol(PosMFA)
NColRM  <- ncol(PosRM)
NTrue   <- "N" # resultado da comparacao
if (!is.null(NColMFA) && !is.null(NColRM))
if (NColMFA == NColRM) {
   k <- 0
```

```
      i <- 1
      while(i <= NColMFA) {
         j <- 1
         while(j <= NColRM) {
            Ajud <- PosMFA[PosMFA[,i]>0,i] %in% PosRM[PosRM[,j]>0,j]
            if (length(Ajud[Ajud == TRUE]) >= 1) { k <- k + 1 }
            j <- j + 1
         }
         i <- i + 1
      }
     if (k == NColMFA || k == 1) NTrue = "T"
  } else {
      i <- 1
      while(i <= NColMFA) { # as tecnicas concordam parcialmente
         j <- 1
         while(j <= NColRM) {
           Ajud <- PosMFA[PosMFA[,i]>0,i] %in% PosRM[PosRM[,j]>0,j]
           if (length(Ajud[Ajud == TRUE]) > 1) {
               NTrue = "P"
               i = NColMFA + 1
               break
           }
           j <- j + 1
         }
         i <- i + 1
      }
  }
  ### FIM - compara convergencia solucoes entre MFA e metodo proposto ###

  return(list(NTrue = NTrue, MatCompFinal = MatCompFinal))
}


Cirillo = function(p, k, pho, delta, n, sim) {
  # Funcao desenvolvida por Marcelo Angelo Cirillo para
  # gerar grupos de variaveis correlacionados

  # Entrada:
  # p - numero de variaveis
  # k - numero de populacoes
  # pho - correlacao
  # delta - grau de heterogeneidade entre as matrizes
  # n - tamanho amostral
  # sim - numeros de simulacoes

  # Saida:
  # amostra - matriz com grupo correlacionados a nivel delta

  ### INICIO - cria matriz de correlacao global ###
  pk = p * k # numero de elementos na diagonal
  ar1 = diag(pk) # matriz de correlacao global

  i <- 1
  while(i <= pk) {
    j <- 1
    while(j <= pk) {
      ## Estrutura AR(1)
      if (i == j) ar1[i,j] = 1
      else ar1[i,j] = pho^(abs(i-j))
      j <- j +1
    }
```

```
    i <- i + 1
  }
  ### FIM - cria matriz de correlacao global ###

  ### INICIO - Simula o parametro da matriz de covariancia ###
  if (delta > 1) {
     mi    = matrix(0, nrow(ar1), 1) # vetor de medias
     gn    = 10000 # tamanho da amostra a simular
     amos  = rmvnorm(gn, mean = mi, sigma = ar1)
     d     = (delta - 1) / (k - 1)
     b1    = amos[1:gn, 1:p] # primeira populacao
     conti = p + 1
     contf = conti + p - 1
     yf = matrix(0, gn, p)
     j <- 2
     while(j <= k) {
       mc    = amos[1:gn, conti:contf]
       dest  = (1 + d*(j - 1))^(1/p)
       yaux  = mc * dest
       conti = conti + p
       contf = contf + p
       yf    = cbind(yf, yaux)
       j <- j + 1
     }
     yf1  = yf[,(p+1):(p*k)]
     yf2  = cbind(b1, yf1)
     mcov = cov(yf2)
  }

  if (delta == 1) mcov = ar1
  ### FIM - Simula o parametro da matriz de covariancia ###

  ### INICIO - Simulacoes ###
  mi = matrix(0, nrow(mcov), 1)

  # amostra com as populacoes correlacionadas
  amostra = rmvnorm(n, mean = mi, sigma = mcov)
  ### FIM - Simulacoes ###

  return(amostra)
}
```