Explorando a variabilidade interna do Escalonamento Multidimensional – MDSvarint

MOACYR MACHADO CARDOSO JUNIOR
Instituto Tecnológico de Aeronáutica – ITA
Praça Mal. Eduardo Gomes, 50 – São José dos Campos
moacyr@ita.br

RODRIGO ARNALDO SCARPEL
Instituto Tecnológico de Aeronáutica – ITA
Praça Mal. Eduardo Gomes, 50 – São José dos Campos
rodrigo@ita.br

RESUMO

O presente trabalho tem por objetivo propor um novo modelo para análise do mapa perceptual gerado via escalonamento multidimensional não métrico aplicado a matrizes de similaridades (correlação). O modelo explora a variabilidade interna do algoritmo SMACOF (*Scaling by MAjorizing a COmplicated Function*). O SMACOF necessita uma matriz de entrada para obtenção da solução final e no caso deste trabalho diferentes soluções iniciais foram geradas utilizando distribuição Normal das coordenadas. Como resposta o SMACOF produz diferentes mapas perceptuais, que após o alinhamento das configurações via análise de Procrustes Generalizada é obtido o mapa de consenso e também as regiões de confiança dos objetos no espaço bidimensional utilizando técnicas não paramétricas de reamostragem. Como resultado da aplicação de método obteve-se o mapa perceptual de consenso que e incorpora a variabilidade da solução em função da configuração inicial adotada, gerando regiões de confiança.

PALAVRAS CHAVE. Escalonamento Multidimensional. Procrustes. Reamostragem.

Área de classificação principal (Outras aplicações em PO (OA))

ABSTRACT

The main issue of this paper is to propose a new model for evaluation of the perceptual map obtained by nonmetric multidimensional scaling applied to similarities matrix (correlations). The model explores the internal variability of SMACOF algorithm (Scaling by MAjorizing a COmplicated Function). SMACOF needs an initial matrix to obtain the final solution and in this paper different initial solutions were provided using Normal distribution of the coordinates. SMACOF produces different perceptual maps, which are aligned by General Procrustes Analysis and a consensus map is obtained and also objects confidence regions in a bi-dimensional space using non parametric techniques, known as randomization tests. As a result the method generated the consensus perceptual map and the variability introduced by the initial configuration adopted, producing confidence regions.

KEYWORDS. Multidimensional Scaling. Procrustes. Randomization tests.

Main area (Other applications in OR (OA)).

1. Introdução

O Escalonamento Multidimensional – MDS é um método que toma por base a proximidade de objetos, sujeitos ou estímulos utilizados para produzir uma representação espacial dos mesmos (HÄRDLE; SIMAR, 2007). A proximidade expressa a similaridade ou dissimilaridade entre objetos. O MDS é uma técnica de redução de dimensão, uma vez que seu objetivo é encontrar um conjunto de pontos em baixa dimensão (usualmente duas dimensões) que reflitam a configuração dos dados em alta dimensão.

Cox e Cox (2001) apresentam uma definição de MDS que chamam de "estreita" parecida com a apresentada acima, e uma "larga", ou seja, MDS pode conter várias técnicas de análise multivariada de dados, e no extremo, ela se refere a qualquer técnica que produza uma representação gráfica dos objetos a partir de dados multivariados.

Borg e Groenen (2005) definem MDS como um método que representa medidas de similaridade (ou dissimilaridade) entre pares de objetos como distâncias entre pontos em um espaço de baixa dimensão. Os autores descrevem quatro propósitos do MDS: a) MDS como um método para representar (dis)similaridade como distâncias em um espaço dimensional com poucas dimensões, afim de tornar os dados acessíveis à inspeção visual e análise exploratória; b) Como uma técnica que permita que se teste de que forma certos critérios pelos quais uma pessoa pode distinguir diferentes objetos de interesse são espelhados em uma correspondente diferença empírica desses objetos; c) Uma aproximação analítica que permita descobrir as dimensões que estão sob os julgamentos de (dis)similaridade; d) Como um modelo psicológico que explica julgamentos de (dis)similaridade em termos de regras que se aproximam de um tipo particular de função de distância.

Páez (2009) apresenta os modelos MDS divididos em dois grandes grupos: Métrico e Não-Métrico. Os modelos métricos consideram transformações que relacionam dissimilaridades à distâncias considerando além da ordem, os valores de dissimilaridades intrínsecos. O MDS métrico está usualmente associado a transformações de razão, ou seja, $\hat{d}_{ij} = b\delta_{ij}$ onde b é um escalar, $\hat{d}_{ij}$ é a pseudodistância e δ_{ij} é a dissimilaridade. Para transformações intervalares tem-se $\hat{d}_{ij} = a + b\delta_{ij}$, a representa a generalização do modelo de razão, sendo a e b valores desconhecidos. Desta forma o problema do MDS é estimar simultaneamente os valores de $\hat{d}_{ij}$ e a configuração de pontos X, de tal forma que $d_{ij}(x) \approx \hat{d}_{ij}$. Nos modelos não métricos, doravante designados por NMDS as disparidades estão restritas somente a manutenção da ordem das proximidades. Os valores das disparidades são arbitrários e devem obedecer somente a monotonicidade.

O NMDS ordinal é um caso especial de MDS, e possivelmente o mais importante na prática (COX; COX, 2001). Ele é usado normalmente quando, por exemplo, desejamos obter o julgamento, colocando os objetos em ordem crescente ou decrescente de importância sob a ótica de um avaliador.

O objetivo deste trabalho é o de propor um novo modelo de NMDS, que incorpore a variabilidade interna do mesmo na solução final, isto é, a solução do NMDS parte, via de regra, da solução do MDS (métrico). Nesta proposta a solução do NMDS se dará a partir de várias soluções iniciais geradas aleatoriamente por uma distribuição ~N(0,1), e com isso se pode obter não somente as coordenadas dos objetos no espaço em baixa dimensão, mas também as regiões de confiança permitindo a análise do resultado obtido considerando a incerteza de posicionamento dos objetos no novo espaço.

Tal solução permite avaliar melhor as diferenças entre os objetos, bem como a distribuição dos dados bivariados (para o caso de 2 dimensões).

Este trabalho encontra-se organizado da seguinte forma: na seção de introdução, foi abordada a motivação e os objetivos para desenvolvimento deste trabalho; a Seção 2 uma breve revisão sobre inferência estatística no MDS. A Seção 3 apresenta o método proposto MDSvarint; a Seção 4 apresenta um experimento realizado e resultados obtidos com a utilização do método; e finalmente a Seção 5 com as considerações finais.

2. Inferência Estatística no MDS

A inferência estatística para problemas MDS foi muito debatida no passado, conforme citado por Cox e Cox (2001), sendo que alguns pesquisadores sugerem que o MDS permaneça com uma característica de técnica exploratória ou representação dos dados, enquanto outros afirmem que alguns esforços deveriam ser envidados para dotar os modelos MDS da teoria do erro e alguns aspectos probabilísticos.

Cox e Cox (2001) citam que desde 1982 algumas pesquisas sobre inferência no MDS foram realizadas, mas que ainda não causaram grande impacto na utilização da ferramenta.

A idéia é que exista uma configuração de pontos no espaço Euclidiano que representa os objetos, e como usual, d_{rs} é a distância Euclidiana entre os objetos r e s. Considerando que d_{rs} tem função de densidade de probabilidade $p(\delta_{rs}\backslash d_{rs})$. Assume-se que as observações são independentes e identicamente distribuídas e, portanto a função de verossimilhança é

$$lnl = \sum_r \sum_s \ln p(\delta_{rs}\backslash d_{rs}) \tag{1}$$

As distâncias podem ser escritas em termos das coordenadas dos pontos, $d_{rs}^2 = (x_r - x_s)^T(x_r - x_s)$, e, portanto a função de verossimilhança pode ser minimizada em relação a x_r e todos os parâmetros da função de densidade de probabilidade p. Obtêm-se dessa forma os estimadores de máxima verossimilhança das coordenadas, $\hat{x}_r$.

Duas distribuições possíveis para $\delta_{rs}\backslash d_{rs}$ são a Normal e a Log-Normal. Para a Normal tem-se:

$$\delta_{rs} \sim N(d_{rs}, d_{rs}^2\sigma^2) \tag{2}$$

Neste modelo existe probabilidade diferente de zero para valores negativos de δ_{rs}.

Para o modelo log-Normal, tem-se:

$$Ln\ \delta_{rs} \sim N(Ln\ d_{rs}, \sigma^2) \tag{3}$$

Ramsay (1982) citado por Cox e Cox (2001) sugere uma transformação das dissimilaridades baseada em regressões *splines*, o que torna o modelo muito complicado, e esta é a principal razão de críticas aos modelos de inferência para o MDS.

O modelo MULTISCALE (*Maximum likelihood Scaling*) foi proposto por Ramsay (1991) para resolver o modelo MDS de máxima verossimilhança, considerando a distribuição Normal e Log-Normal e transformações *splines* das dissimilaridades. O modelo permite inclusive obter as regiões de confiança, pela estimativa da matriz de covariância. Um problema ocorre porém devido ao fato dos modelos MDS serem invariantes à rotação, translação e reflexão, e desta forma a matriz de covariância é afetada. Ramsay resolveu este problema restringindo a configuração de máxima verossimilhança de forma que o centróide da configuração esteja na origem e a matriz X^TX transformada seja diagonal. Observa-se desta forma que a configuração de máxima verossimilhança é dependente da restrição imposta ao modelo.

Weinberg, Carroll e Cohen (1984) utilizaram as técnicas de reamostragem *Bootstrap* e *Jackknife* para verificar se as mesmas podem ser generalizadas para os modelos INDSCAL (*Individual Differences Scaling*) para obtenção das regiões de confiança. As duas abordagens foram comparadas ao modelo MULTISCALE. A conclusão é que as técnicas podem ser

utilizadas para testar a precisão da localização das coordenadas dos objetos no espaço. A técnica *Jackknife* que é uma aproximação linear da estatística em comparação com o *Bootstrap*, que não é linear, e apresentou resultados satisfatórios e muito próximos.

Krzanowski (2006), afirma que muitas análises de MDS findam com a produção de uma representação pictórica em duas ou três dimensões, mas na verdade este é apenas o ponto de partida e assim propôs um método para análise de sensibilidade do MDS métrico que consiste em realizar n+1 análises de coordenadas principais (PCO), uma vez na matriz completa e uma vez para cada omissão de uma linha/coluna desta matriz. As coordenadas dos pontos são calculadas pela decomposição espectral. Cada versão reduzida da matriz é aumentada calculando-se as coordenadas do ponto omitido, usando uma fórmula que requer somente a multiplicação de matrizes já calculadas. O autor sugere novos estudos para avaliar qual modelo probabilístico poderia ser utilizado para inferência, e sugere que estes estudos sejam realizados com *Bootstrap*.

Jacoby (2009) coloca muito bem o problema do MDS utilizado como ferramenta exploratória de dados ao invés de ferramenta para inferência, afirmando que o MDS ao reduzir dimensões e estabelecer as coordenadas dos objetos nesse espaço de baixa dimensão não proporciona nenhuma informação sobre a variabilidade dos dados, de forma que é praticamente impossível generalizar as conclusões para uma população. O autor utilizou o WMDS (modelo INDSCAL) juntamente com o *Bootstrap* para gerar regiões de confiança das coordenadas dos objetos e também para analisar os pesos atribuídos a cada dimensão.

Abdi, Dunlop e Williams (2009) apresentam um método para calcular intervalos de confiança e de tolerância para classificadores representados em mapa tipo MDS utilizando *Bootstrap* com a justificativa de incorporar a variabilidade na análise da representação espacial dos objetos e também comparar a discriminação entre as classes.

Saburi e Chino (2008) desenvolveram um modelo de máxima verossimilhança para o MDS assimétrico, que incorpora no seu resultado a confiabilidade das coordenadas dos objetos na forma de região de confiança. Os autores citam o método como superior aos obtidos via técnicas de reamostragem, tais como *Bootstrap*, pois segundo os mesmos não é simples a implementação para os problemas de MDS.

3. Método proposto - MDSvarint

O método proposto neste trabalho parte da premissa de que a configuração final das coordenadas dos objetos no espaço de baixa dimensão é função da configuração inicial adotada para solução do MDS não métrico, doravante designado por NMDS. A cada configuração inicial adotada, uma solução final é alcançada. Desta forma, pode-se obter regiões de confiança no espaço das possíveis localizações das coordenadas dos objetos e compará-las entre si. Esta variabilidade das coordenadas é obtida internamente ao próprio algoritmo MDS-SMACOF (*Scaling by MAjorizing a COmplicated Function*), e por esta razão passamos a denominá-la de MDS variabilidade interna – MDSvarint.

3.1 Redução de Dimensão

O método inicia-se com uma matriz de dissimilaridade – **D** e à mesma é aplicado o algoritmo de solução SMACOF, proposto por De Leeuw (1977) e recentemente implementado por De Leeuw e Mair (2009) em um pacote do software estatístico R. O algoritmo parte de uma solução inicial, que pode ser gerada por uma distribuição de Poisson, Normal ou Uniforme.

Cox e Cox (2001) citam a utilização de uma configuração arbitrária como uma possível solução inicial para o NMDS. Os pontos podem ser colocados nos vértices de uma grade regular p-dimensional, ou podem ser geradas por um processo de Poisson p-dimensional. Os autores

recomendam que várias configurações iniciais sejam testadas a fim de evitar ótimos locais. A utilização da solução do MDS métrico também é recomendada pelos autores.

Como se busca a variabilidade das configurações finais gera-se uma série de configurações iniciais – **Z** com base nessas distribuições de probabilidade e então se submete cada uma dessas configurações ao algoritmo SMACOF.

O algoritmo SMACOF, para solução do MDS é também denominado de majoração iterativa - MI.

O princípio da MI é que ela gera uma sequência monotônica não crescente de valores da função de Stress (função de perda).

A idéia central do MI é substituir iterativamente a função original complicada $f(x)$ por uma função auxiliar $g(x,z)$, onde z é um valor fixo. A função g deve obedecer aos seguintes requisitos para que $g(x,z)$ seja uma função de majoração de $f(x)$:

- A função auxiliar $g(x,z)$ deve ser mais simples de obter a minimização do que $f(x)$. Por exemplo se $g(x,z)$ é uma função quadrática em x pode ser resolvida em 1 passo;

- A função original deve ser menor ou no máximo igual à função auxiliar; $f(x) \leq g(x,z)$;

- A função auxiliar deve tocar a superfície no ponto de suporte z; $f(z) \leq g(z,z)$.

Retomando a função de Stress do MDS, apresentam-se algumas definições:

a) n representa o número de objetos empíricos (estímulos, variável, item, questão, dentre outros dependendo do contexto);

b) Se uma observação foi realizada para o par de objetos i e j, um valor de proximidade P_{ij} é dado. Se P_{ij} for indefinido, tem-se um valor faltante. O termo proximidade é usado genericamente tanto para similaridade quanto para dissimilaridade. Para similaridades valores altos de P_{ij} indicam que o objeto i e j ao similares;

c) A dissimilaridade é uma proximidade que indica quão diferentes são os objetos. Valores pequenos indicam que os objetos são similares e valores altos, o contrário. A dissimilaridade é representada por δ_{ij}.

d) **X** representa 1) uma configuração de pontos (ie. Um conjunto de n pontos em um espaço m dimensional); 2) uma matriz nxm de coordenadas de n pontos relativos a m eixos de coordenadas cartesianas. Um sistema de coordenadas cartesianas é um conjunto de pares de linhas e retas perpendiculares (eixos coordenados). Todos eixos tem interseção na origem, O. A coordenada de um ponto no eixo **a** é a distância direcionada (sinal) da projeção perpendicular do ponto na direção do eixo **a**, a partir da origem. A m-tupla (x_{i1},...x_{im}) denota as coordenadas do ponto i com respeito aos eixos a=1,...,m, sendo a origem (0,...,0).

e) A distância Euclidiana entre dois pontos quaisquer i e j em **X** é o comprimento de uma linha reta entre os pontos i e j em **X**. É calculado pela Equação $d_{ij} = \left[\sum_{a=1}^{m}(x_{ia} - x_{ja})^2\right]^{\frac{1}{2}}$, onde x_{ia} é a coordenada do ponto i relativo ao eixo a do sistema cartesiano de coordenadas. Utiliza-se também $d_{ij}(X)$ para demonstrar que a distância é função de **X.**

f) O termo $f(P_{ij})$ denota o mapeamento de P_{ij}, ou seja, o valor atribuído a P_{ij} de acordo com a regra f. Algumas vezes pode ser escrito na forma $f: P_{ij} \rightarrow f(P_{ij})$. Também denomina-se $f(P_{ij})$ como uma transformação de P_{ij}. Outra simbologia que pode ser utilizada ao invés de $f(P_{ij})$ é $\hat{d}_{ij}$.

O erro total de representação no MDS é definido por:

$$\sigma_r(X) = \sum_{i<j}^{n} \left(d_{ij} - \delta_{ij}\right)^2 \quad (4)$$

A relação i<j em (4) mostra que é suficiente somar somente metade dos dados, pois as matrizes de dissimilaridade e distâncias são simétricas. Com relação ao termo δ_{ij}, ressalta-se que experimentalmente o mesmo pode não estar disponível, de forma que o termo δ_{ij} é não definido. Valores faltantes não impõem nenhuma restrição em qualquer distância **X**, entretanto define-se um peso fixo w_{ij} com valor de 1 se δ_{ij} é conhecido e 0 se for faltante. Outros valores de w_{ij} também são permitidos, uma vez que $w_{ij} \geq 0$. Isto define a versão final do Stress proposto por Kruskal (1964).

$$\sigma_r(X) = \sum_{i<j}^{n} w_{ij}\left(d_{ij}(X) - \delta_{ij}\right)^2 \quad (5)$$

Para todo conjunto de coordenadas **X**, um valor de Stress pode ser calculado, mas o objetivo é obter o valor que minimiza o erro (4), ou seja, deseja-se minimizar $\sigma_r(X)$ em **X**.

A função de Stress em (5) pode ser desmembrada da seguinte forma:

$$\sigma_r(X) = \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij}d_{ij}^2(X) - 2\sum_{i<j} w_{ij}\delta_{ij}d_{ij}(X) \quad (6)$$

$$\sigma_r(X) = \eta_\delta^2 + \eta^2(X) - 2\rho(X) \quad (7)$$

A primeira parte η_δ^2 é dependente somente dos valores fixos dos pesos w_{ij} e nas dissimilaridades fixas e não depende de **X**, portanto o termo η_δ^2 é constante.

A segunda parte, $\eta^2(X)$ é uma soma ponderada da distância quadrática $d_{ij}^2(X)$. A parte final $-2\rho(X)$ é a soma ponderada da distância plana $d_{ij}(X)$.

No algoritmo SMACOF para solução do MDS, tem-se que:

$$\sigma_r(X) = \eta_\delta^2 + trX'VX - 2trX'B(X)X \leq \eta_\delta^2 + trX'VX - 2trX'B(Z)Z = \tau(X,Z)$$

Assim a função $\tau(X,Z)$ é a função simplificada de majoração, que é quadrática em **X**. O mínimo pode ser obtido analiticamente fazendo a 1ª derivada $\tau' = 0$.

$$\nabla\tau(X,Z) = 2VX - 2B(Z)Z = 0 \quad (8)$$

De forma que $VX = B(Z)Z$

Para resolver o sistema de equações lineares para **X**, pré-multiplica-se por V^{-1}, entretanto a inversa V^{-1} não existe, pois V não tem posto completo. Utiliza-se então a inversa de Moore-Penrose. A inversa de Moore-Penrose de V é dada por $V^+ = \left(V + 11'\right)^{-1} - n^{-2}11'$. O último termo $-n^{-2}11'$ é irrelevante no contexto do SMACOF, uma vez que V^+ é multiplicado por uma matriz ortogonal a 1, pois B(Z) tem autovetor 1 com autovalor = 0, assim atualizamos a Equação.

$$X^u = V^+B(Z)Z \quad (9)$$

Se todos $w_{ij} = 1$, então $V^+ = n^{-1}J$, com $J = I - n^{-1}11'$, e então:

$$X^u = n^{-1}B(Z)Z \quad (10)$$

A equação (9) foi batizada por De Leeuw (1977) de transformação de Guttman.

Após a redução de dimensões via SMACOF, procede-se ao alinhamento das configurações utilizando a Análise Procrustes Generalizada – GPA.

3.2 Análise Procrustes Generalizada - GPA

Como resultados obtêm-se uma série de configurações finais, que pelo fato de suas soluções serem invariantes à rotação, reflexão e translação, as mesmas são então submetidas ao processo de alinhamento utilizando a análise Procrustes generalizada – GPA, que submete as diferentes configurações a um processo de obtenção de uma configuração de consenso por meio das transformações rígidas admissíveis das configurações, obtendo consequentemente configurações finais alinhadas segundo um critério de erro das coordenadas. O termo Procrustes originou-se do trabalho de Hurdley e Cattell (1962) *apud* Gower e Dijksterhuis (2004) que estavam à busca de comparar uma configuração obtida por intermédio da análise fatorial com uma configuração hipotética.

A GPA segundo Brombin e Salmaso (2009) é uma ferramenta da estatística da forma. O termo forma é definido pelos autores associando as propriedades geométricas de uma configuração de pontos que são invariantes a mudanças de translação, rotação e escala. A análise direta de um conjunto de pontos não é conveniente devido à presença de erros sistêmicos tais como posição, orientação e tamanho, e usualmente para que se possa conduzir uma análise estatística da forma confiável normalmente o GPA é utilizado para eliminar os fatores não relativos à forma e para alinhar as configurações para um sistema de coordenadas comum.

A GPA é uma técnica estatística multivariada empírica na qual três dimensões estão envolvidas: os objetos de estudo, as pessoas que avaliam os objetos e os atributos nos quais os objetos são avaliados. A GPA é ideal para analisar dados oriundos de diferentes indivíduos (DIJKSTERHUIS; GOWER, 2010).

As transformações permitidas no GPA são translação, rotação/reflexão e escalonamento isotrópico, de forma que as distâncias relativas entre os objetos permaneçam inalterada. (RODRIGUE, 1999)

Diferentemente da análise Procrustes clássica que visa ajustar uma configuração teste em relação ao uma configuração alvo, o GPA permite o ajuste simultâneo das m configurações a uma configuração comum, referenciada por Rodrigue (1999), Borg e Groenen (2005) como configuração de consenso, ou espaço de consenso. O espaço de consenso é a média de todas as configurações após as transformações. Desta forma, ao invés de examinar todos os $\binom{m}{2}$ pares de configurações, o princípio é generalizar de forma que todas as m configurações sejam simultâneamente transladas, rotacionadas, refletidas e escalonadas usando um processo iterativo, de forma que o critério de ajuste seja minimizado.

A similaridade pode ser expressa segundo Rodrigue (1999) como uma minimização da soma de quadrados das distâncias entre cada um dos mn pontos $Q_i^{(j)}$ (configurações individuais transformadas) e sua configuração centróide (soma de todas as configurações), denotada por Z_i, e o critério de otimização consiste na minimização das distâncias pela aplicação de transformações adequadas as configurações.

A formulação matemática do GPA pode ser descrita da seguinte forma: Seja T_j uma matriz nxp com todas n linhas iguais a t_j (1xp vetor linha), H_j uma matriz ortogonal pxp e seja ρ_j escalares (j=1,...m). A translação para uma nova origem é dada pela adição do mesmo vetor linha (1xp) t_j a toda linha de X_j. O escalonamento, a rotação e a translação podem, portanto, ser expressos pela transformação

$$X_j \longmapsto \rho_j X_j H_j + T_j \qquad (11)$$

O problema do GPA é a determinação dos escalares ρ_j, e das matrizes H_j e T_j (j=1,...,m), de forma que a soma dos quadrados dos resíduos, S_r , seja mínima. Isto é, obter uma transformação que torne as configurações iniciais tão similares quanto possível.

A estimação dos parâmetros desconhecidos usa a minimização da soma dos quadrados das diferenças entre cada par de configuração:

$$S_r = \sum_{i=1}^{n}\sum_{j,j^*}^{m} \Delta^2\left(\rho_j X_{ij} H_j + T_{j,}\rho_{j^*} X_{ij^*} H_{j^*} + T_{j^*}\right) \quad (12)$$

A forma de estimar os parâmetros T_j, ρ_j e H_j consiste na diferenciação de S_r com relação a T_j, ρ_j e H_j. Entretanto alguns aspectos desse problema devem ser investigados antes da realização dessa operação. Inicialmente, sem perda do poder de generalização, todas as configurações X_j podem ser centralizadas em seus centróides. Essa operação simplifica a operação de translação. S_r pode ser minimizado, tomando todos ρ_j como 0, mas esta solução não é interessante, e uma outra aproximação é fixar o fator de escalonamento igual a 1. Gower e Dijksterhuis (2004) propuseram uma forma mais satisfatória para a estimativa do escalonamento impondo a seguinte restrição $\sum_{j=1}^{m} \rho^2 tr\left(X_j X_j^T\right) = \sum_{j=1}^{m} tr\left(X_j X_j^T\right)$, ou seja, a soma final de quadrados na origem das configurações rotacionadas e escalonadas não é alterada em relação a inicial, e este fato garante que a soma de quadrados dos resíduos será a mesma independente do escalonamento realizado.

Finalmente a minimização é submetida à restrição de que as matrizes H_j, matrizes de rotação, sejam ortogonais.

3.3 Obtenção das Regiões de Confiança - reamostragem

Após o GPA obtêm-se as configurações alinhadas, cujas coordenadas são então submetidas a uma estratégia de reamostragem (*randomization tests*), uma vez que não existe garantia de que as coordenadas seguem uma determinada distribuição probabilística (ie.: distribuição Normal), especialmente por se tratar de um caso multivariado. A técnica escolhida para esta implementação foi o *Jackknife*, pois o mesmo apresentou bons resultados no trabalho de Weinberg, Carroll e Cohen (1984), e o mesmo é de fácil implementação. O objetivo do teste é o de estimar a matriz de covariância, permitindo assim obter as regiões de confiança das coordenadas finais.

O *Jackknife*, introduzido por Quenouille (1949, 1956) *apud* Weinberg, Carroll e Cohen (1984) como uma técnica estatística para reduzir o viés na estimação de parâmetros, foi generalizado por Tukey (1958) *apud* Weinberg, Carroll e Cohen (1984) para estimar o desvio padrão dos parâmetros estimados.

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho N de uma distribuição F desconhecida. Se θ é um parâmetro de F, $\hat{\theta}$ é um estimador de θ e $\hat{\theta}_i$ é o estimador obtido pela eliminação do *i-ésimo* membro da amostra. A estimação da variância de $\hat{\theta}$:

$$\widehat{Var} = \frac{n-1}{n}\sum_{i=1}^{N}\left[\hat{\theta}_i - \hat{\theta}_.\right]^2 \quad (13)$$

Sendo $\hat{\theta}_. = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_i$

Tukey (1958) *apud* Weinberg, Carroll e Cohen (1984) definiu ainda os pseudovalores, $\tilde{\theta}_i$, e os mesmos calculados da forma:

$$\tilde{\theta}_i = N\hat{\theta} - (N-1)\hat{\theta}_i \quad (14)$$

Desta forma, a estimativa da variância em função dos pseudovalores é dada por:

$$\widehat{Var}_J = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left[\tilde{\theta}_i - \tilde{\theta}_.\right]^2 \quad (15)$$

Sendo $\tilde{\theta}_. = \frac{1}{N}\sum_{i=1}^{N} \tilde{\theta}_i$

No caso multivariado, θ passa a ser um vetor de parâmetros da população e $S_{\tilde{\theta}}$ a matriz de covariância. Dessa forma tem-se que:

$$S_{\tilde{\theta}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\tilde{\theta}_i - \tilde{\theta}_.\right)\left(\tilde{\theta}_i - \tilde{\theta}_.\right)^T \quad (16)$$

4. Experimento

Com o objetivo de demonstrar o método proposto, MDSvarint foi realizado um experimento utilizando os dados extraídos do livro de Borg e Groenen (2005) a partir de um exemplo apresentado por Wish (1971), referente ao julgamento das similaridades entre 12 nações.

O algoritmo SMACOF necessita na entrada de uma matriz de dissimilaridade, e desta forma a matriz original foi transformada segundo $\delta_{ij} = S_{ij} - 6{,}68$, onde δ_{ij} é a dissimilaridade e S_{ij} representa a similaridade de forma que a maior similaridade seja transformada na menor dissimilaridade. Os elementos diagonais por definição, são igualados a zero.

Na sequência, para gerar uma referência comparativa, o problema foi resolvido utilizando o MDS métrico e também o NMDS, para duas dimensões, utilizando o algoritmo KYST (Kruskal-Young-Shepard-Togerson), com o resultado representado na Figura 1. No caso do NMDS a configuração inicial foi a solução do MDS métrico. O valor de Stress obtido foi de 19,21.

O próximo passo é a geração aleatória de soluções iniciais utilizando uma distribuição ~N(0,1). No caso deste experimento foram utilizadas 100 iterações do algoritmo (C=100). A resolução via SMACOF, gerou uma distribuição dos valores de Stress que não tem aderência a distribuição Normal (Teste KS - D = 0.3812, p-value = 4.766×10^{-13}), com média igual a 25,19, valor mínimo de 21,59 e valor máximo de 47,06.

Comparando com a referência do MDS não métrico, Stress de 19,21, conclui-se pelo teste de Wilcoxon com correção de continuidade que os valores de Stress obtido são superiores ao da referência ao nível de significância de 5%. (p-value < 2.2×10^{-16})

Esta já representa uma constatação importante, pois as soluções geradas aleatoriamente não foram capazes de superar o ajuste obtido a partir da configuração obtida pelo MDS métrico.

A Figura 1 apresenta as 100 soluções finais obtidas e a comparação com a solução do MDS métrico e não métrico NMDS.

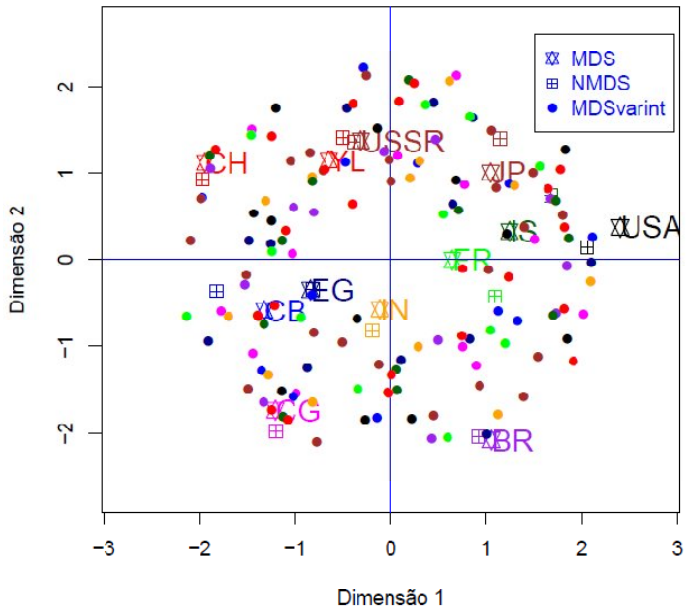

FIGURA 1 – Resultado obtido pela solução do MDS métrico, Não métrico e MDSvarint, sem alinhamento das configurações.

Como era de se esperar, verifica-se que as soluções geradas pelo MDSvarint, são invariantes à rotação e reflexão, e desta forma os pontos homólogos das diferentes iterações assumem uma distribuição aleatória em torno da origem com diferentes raios.

Para eliminar este efeito realiza-se então o alinhamento das configurações por meio do GPA. Na Figura 2 apresenta-se o resultado desse alinhamento.

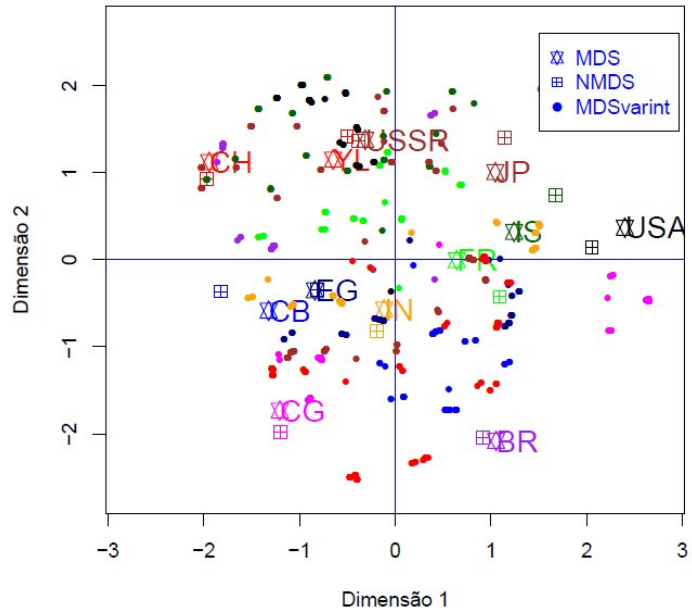

FIGURA 2 – Resultado obtido pela solução do MDS métrico, Não métrico e MDSvarint, após o alinhamento das configurações com GPA.

Os pontos agora mostram a dispersão obtida comparada com as soluções do MDS e NMDS.

Para possibilitar analisar os dados (objetos individualmente) as matrizes originais (k,m,n)=(12,2,100) foram convertidas em (n,m,k)= (100,2,12).

A seguir os pontos foram avaliados quanto à aderência a distribuição normal multivariada pelo teste Shapiro-Wilk, e todos os pontos (12) foram rejeitados (não seguem distribuição normal multivariada)

Finalmente apresenta-se o resultado final do MDSvarint, na Figura 3, com as regiões de confiança construídas a partir da matriz de covariância obtida pelo *Jackknife*.Não foram representados todos os Países pois o mapa ficaria confuso, e o objetivo é o de mostrar a habilidade do mesmo na análise de cada objeto.

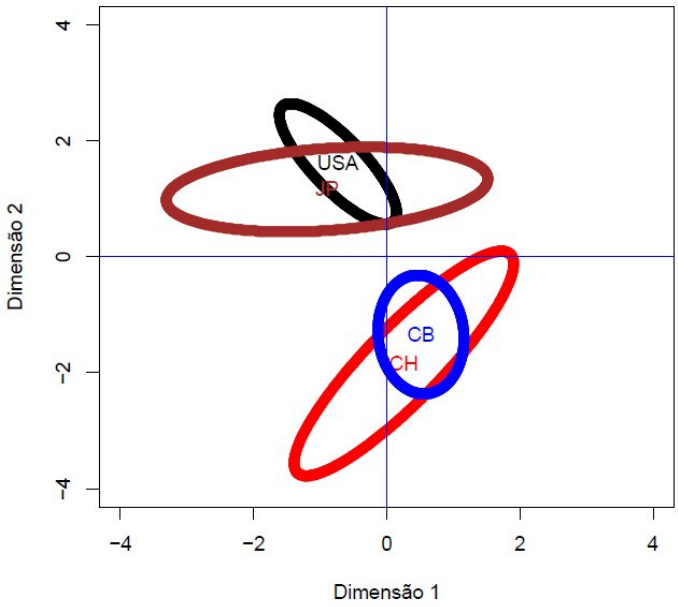

FIGURA 3 – Resultado do MDSvarint, com regiões de confiança gerados pelo *Jackknife* ao nível de significância de 95% para os Países Estados Unidos – USA, Japão – JP, Cuba – CB e China - CH.

O MDSvarint proporciona ao analista visualização das regiões do espaço dos possíveis pontos da solução MDS para o espaço bidimensional. Verifica-se, por exemplo, que Estados Unidos e Japão não diferem estatisticamente entre si ao nível de significância de 95%. O mesmo para China e Cuba. No entanto, tanto Estados Unidos quanto Japão difere estatisticamente de Cuba e/ou China.

5. Considerações finais.

O método proposto possibilita a análise exploratória dos dados (ie.: representação pictórica) e proporciona incorporar a variabilidade da solução em função da configuração inicial adotada.

Do ponto de vista teórico, não se está analisando a variabilidade dos dados de entrada, até porque as mesmas não existem para o caso em que se parte de uma matriz de correlações (similaridades) entre julgamentos de vários observadores, mas sim a variabilidade interna do MDS, que incorpora a variabilidade do próprio método, ou seja, a solução do MDS não é única, no sentido que a mesma é dependente da solução inicial. O fato de partirmos sempre da solução do MDS métrico reflete: 1) uma facilidade, pois se garante que não existirão dois pontos iguais (que inviabiliza a solução pelos métodos tradicionais); 2) limita a análise a um padrão estabelecido; 3) Define o NMDS como determinístico.

A adoção do MDSvarint apresenta como vantagens: a) O algoritmo SMACOF permite a solução, mesmo com configurações iniciais com "empate" entre os objetos ("*ties*"); b) Apresenta um novo modelo de análise permitindo explorar as possíveis configurações do espaço final, dotando o NMDS com a teoria do erro. O fato das soluções iniciais aleatórias gerarem valores de Stress superiores aos métodos clássicos pode ser considerado como limitante e merece ser melhor investigado.

Como desenvolvimentos futuros, os autores sugerem a comparação do MDSvarint utilizando métodos não paramétricos e método de máxima verossimilhança, bem como a aplicação de outros métodos não paramétricos tais como *Bootstrap,* e modelos que permitam inferência (teste de hipótese) via Permutação.

Referências

Abdi, H.; Dunlop, J.P.; Williams, L.J. (2009), How to compute estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). *Neuroimage*, 45, 89-95.

Brombin, C. e Salmaso, L. (2009), Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics and Data Analysis*. 53, 3921-3931.

Borg, I.; Groenen, P. J. *Modern Mutidimensional Scaling: Theory and Applications*. New York: Springer, 2005.

Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*. 2ª Ed. ed. London: Chapman & Hall/CRC, 2001.

Dijksterhuis, G.B.; Gower, J.G. *Procrustes Problems*. New York: Oxford University Press, 2004.

Dijksterhuis, G.B.; Gower, J.G. *The interpretations of generalized Procrustes analysis and allied methods*. Urtecht: Oliemans Punter and Partners, 2010.

De Leeuw, J. (1977). Applications of Convex Analysis to Multidimensional Scaling." In JR Barra, F Brodeau, G Romier, B van Cutsem (eds.), Recent Developments in Statistics, pp. 133-145. North Holland Publishing Company, Amsterdam.

De Leeuw, J.; Mair, P. (2009), Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software,* 31(3).

Hardle, W.; Simar, L. *Applied Multivariate Statistical Analysis*, 2ª Ed. Berlin: Springer, 2007.

Jacoby, W.G. (2009), Public opinion during a presidential campaign: Distinguishing the effects of environmental evolution and attitude change. *Electoral Studies*, 28, 422-436.

Krzanovski, W. J. (2006), Sensitivity in metric Scaling and Analysis of Distance. *Biometrics* , 239-244.

Páez, R.M. *Modelos de Classificación y Multidimensional Scaling y su tratamiento computacional*. Tese de Doutoramento em Estadistica e Investigación Operativa. Granada: Universidad de Granada. 2009.

Rodrigue, N. *A Comparison of the Performance of Generalized Procrustes Analysis and the Intraclass Coefficient of Correlation to Estimate Interrater Reliability*. Master of Science in Epidemiology and Biostatistics. Montreal: McGill University, 1999.

Saburi, S.; Chino, N. (2008), A maximum likelihood method for an asymmetric MDS model. *Computational Statistics and Data Analysis*, 52, 4673-7684.

Weinberg, S.L., Carroll, J.D.; Cohen, H.S. (1984), Confidence regions for INDSCAL using the Jackknife and Bootstrap techniques. *Psychometrika*, 46 (4), 475-491.