Statis e Metabiplot - Um estudo comparativo

Sara Morgado Nunes
Escola Superior de Gestão, Instituto Politécnico de Castelo Branco, Portugal
M. Purificación Galindo
Departamento de Estadística, Universidad de Salamanca, España

Resumo: O Statis e o Metabiplot são técnicas factoriais que possibititam a exploração de conjuntos de dados múltiplos com base na procura de uma estrutura comum às matrizes de dados em análise. Neste trabalho aplicam-se ambas as metodologias a um conjunto de dados reais e comparam-se as duas técnicas.

Palavras–chave: Análise em Componentes Principais, Biplot, Statis e Metabiplot

Abstract: Statis and Metabiplot are factorial techniques that allow the exploration of multiple data sets by searching a common structure to the data matrices in analysis. In this work we apply both methodologies to a real data set and compare these tecniques.

Keywords: Principal Component Analysis, Biplot, Statis and Metabiplot

1 Introdução

Desde que em 1936 Hotelling introduziu a Análise Canónica com o objectivo de comparar e analisar duas matrizes de dados, têm vindo a surgir numerosas técnicas destinadas ao tratamento e integração de informação de várias matrizes de dados, surgindo assim muitos trabalhos que abordam o tratamento de dados múltiplos. Uma das situações mais comuns no trabalho com conjuntos de dados múltiplos é a que diz respeito a um conjunto de matrizes que resultam do estudo de um conjunto de variáveis sobre um conjunto de indivíduos em diversas ocasiões (dados temporais) ou correspondentes a situações experimentais distintas (diferentes estudos). Ao trabalhar com este tipo de dados pretende-se, em geral, proceder a uma análise simultânea que possibilite a obtenção de uma estrutura consenso ou compromisso.

Neste trabalho, procede-se a uma breve revisão dos métodos Statis e Metabiplot. Em ambos os procedimentos é possível estudar matrizes que contêm informação sobre um conjunto de variáveis medidas sobre um conjunto de indivíduos. Assim, o objectivo é explorar em simultâneo várias matrizes de dados onde cada uma delas recolhe informação sobre J variáveis em I indivíduos em T ocasiões ou situações experimentais, procurando-se uma estrutura comum a todos os estudos.

2 Statis

O método STATIS ("Structuration des Tableaux A Trois Índices de la Statistique") foi proposto por L´Hermier des Plantes (1976) e aperfeiçoado por Lavit (1988) e Lavit el al (1994). O STATIS utiliza o coeficiente RV proposto por Escoufier (1973) e tem como principal objectivo a extracção de informação relevante contida em conjuntos de dados múltiplos. O Statis possibilita assim a exploração simultânea de T matrizes X_t de dimensão $I \times J$, $t = 1, \ldots, T$, de dados quantitativos obtidos em diferentes ocasiões sobre os mesmos indivíduos (as variáveis podem ser diferentes). Os dados devem ser centrados e podem, eventualmente, ser reduzidos. O método Statis enfatiza pois as posições relativas dos indivíduos (e não as relações entre as variáveis), desenvolvendo-se em quatro etapas essenciais que se descrevem em seguida:

1) Análise da Interestrutura

Com o estudo da interestrutura pretende-se proceder a uma comparação global da estrutura das T matrizes de dados. A estrutura de cada matriz X_t é captada através da configuração

$$W_t = X_t M_t X_t'$$

que é a matriz de produtos escalares entre indivíduos, sendo M_t a métrica do espaço de indivíduos. Cada matriz W_t define portanto as distâncias entre indivíduos. M_t é habitualmente a matriz identidade de ordem J, porém, se o número de variáveis nas matrizes X_t é diferente, é necessário introduzir um termo de compensação tomando os elementos da diagonal iguais a $\frac{1}{J_t}$.

A fim de tornar possível a comparação das T matrizes de dados e averiguar a existência de uma estrutura comum aos indivíduos, define-se o produto interno de Hilbert-Schmidt entre configurações:

$$\langle W_t | W_{t'} \rangle = Tr(W_t D W_{t'} D)$$

onde D é uma matriz diagonal de ordem I, que define a métrica no espaço das variáveis e cujos elementos da diagonal são iguais a $\frac{1}{I}$ (pesos atribuídos aos indivíduos). $\langle W_t | W_{t'} \rangle$ traduz pois a similaridade entre as matrizes W_t e $W_{t'}$.

Este produto induz a uma norma que é definida por $\|W_t\|^2_{HS} = \langle W_t | W_t \rangle$ e portanto a uma distância:

$$d_{HS}(W_t | W_{t'}) = \|W_t - W_{t'}\|_{HS} = \sqrt{\|W_t\| + \|W_{t'}\| - 2\langle W_t | W_{t'} \rangle_{HS}}.$$

No caso em que as configurações possuem normas distintas, é conveniente trabalhar com as configurações normadas.

Na análise da interestrutura recorre-se habitualmente ao Coeficiente de Correlação Vectorial entre dois estudos t e t', proposto por Robert and Escoufier (1976), que se define da seguinte forma:

$$RV(t,t') = \left\langle \frac{W_t}{\|W_t\|_{HS}} | \frac{W_{t'}}{\|W_{t'}\|_{HS}} \right\rangle_{HS} = \frac{Tr(W_t D W_{t'} D)}{\sqrt{Tr(W_t D)^2}\sqrt{Tr(W_{t'} D)^2}}.$$

Os coeficientes RV surgem organizados numa matriz RV, quadrada de ordem T e possibilitam uma interpretação fácil da interestrutura na medida em que são não negativos e variam entre 0 e 1. Se $RV(t,t') = 1$, então a distância entre os estudos t e t' é nula e as estruturas em causa coincidentes. A distância entre dois estudos normados t e t', é dada por:

$$d_{HS}\left(\frac{W_t}{\|W_t\|_{HS}}|\frac{W_{t'}}{\|W_{t'}\|_{HS}}\right)_{HS} = \sqrt{2(1 - RV(t,t')}.$$

Com o objectivo de comparar globalmente a estrutura das T matrizes de dados, é possível obter uma representação euclidiana das mesmas submetendo a matriz RV a uma Análise em Componentes Principais (ACP). No espaço das componentes principais, a distância entre pontos no plano reflecte o grau de semelhança entre matrizes, possibilitando assim a análise da interestrutura. Os coeficientes RV interpretam-se como sendo os cosenos dos ângulos entre os vectores que representam as matrizes e aproximam a correlação vectorial entre as mesmas, pelo que, se há uma estrutura comum, os ângulos são pequenos e a maior parte da variabilidade é explicada pelo primeiro eixo da representação.

2) Matriz Compromisso

Se a análise da interestrutura permite concluir que efectivamente as matrizes se parecem, procede-se à construção de uma matriz que resuma a informação proveniente das configurações em estudo - a matriz compromisso. A matriz compromisso não é mais que uma média ponderada das configurações W_t sendo, portanto, a mais correlacionada, no sentido do produto escalar de Hilbert-Schmidt, com todas as matrizes e é dada por:

$$W = \sum_{t=1}^{T} \alpha_t W_t$$

sendo os coeficientes α_t as componentes do primeiro vector próprio resultante da diagonalização da matriz RV.

3) Análise da Intraestrutura

Esta etapa consiste em representar a estrutura de cada matriz de dados num espaço de baixa dimensão. Para tal, submete-se a matriz compromisso W a uma ACP, permitindo assim obter uma imagem euclidiana compromisso dos indivíduos em estudo. As coordenadas dos objectos da t-ésima matriz no espaço das p componentes principais (espaço compromisso), são dadas por

$$C_t = W_t LE$$

sendo E uma matriz diagonal $f \times f$ cujos elementos da diagonal são o inverso das raízes quadradas dos valores próprios compromisso e L uma matriz de dimensão $I \times f$ que contém os *scores* da ACP da matriz compromisso.

No espaço compromisso, a distância entre pontos interpreta-se como sendo a distância compromisso entre os indivíduos por eles representados. Por outro lado, os ângulos definidos entre as variáveis e os eixos, interpretam-se em termos de correlação, o que permite, por sua vez, interpretar as posições dos indivíduos relativamente aos eixos compromisso.

4) Interpretação das Trajectórias

Projectando as linhas das matrizes originais no espaço compromisso como elementos suplementares, é possível analisar a evolução das trajectórias ao longo dos distintos grupos. As trajectórias obtidas definem a mudança na posição de um indivíduo e, eventualmente, das variáveis na configuração consenso ao longo do tempo, permitindo portanto analisar a forma como os indivíduos contribuem para as distâncias entre configurações. Assim, trajectórias de grande amplitude reflectem mudanças na estrutura dos indivíduos, enquanto trajectórias envolventes descrevem evoluções médias.

3 Metabiplot

A Análise Biplot foi proposta por Gabriel (1971) e é uma técnica multivariada que pode ser usada sobre qualquer matriz de dados X sobre I indivíduos e J variáveis. Um Biplot é uma representação gráfica da aproximação de uma matriz X $I \times J$ mediante marcadores $g_1, g_2, \ldots, g_I$ para as suas linhas e marcadores $h_1, h_2, \ldots, h_J$ para as suas colunas, de tal forma que o produto interno aproxime o elemento x_{ij}. Para tal, a matriz X decompõe-se em valores e vectores singulares:

$$X = UDV$$

onde U é a matriz de vectores próprios da matriz XX', D a matriz diagonal de valores singulares de X e V a matriz de vectores próprios da matriz $X'X$. Uma escolha adequada dos marcadores para as linhas e para as colunas de X proporciona uma representação Biplot num espaço de baixa dimensão. Duas factorizações possíveis são $X_{(2)} = A^o B^{*'} = A^* B^{o'}$ com os factores definidos, respectivamente, por

$$RMP: A^* \approx (UD)_{C2} \quad \text{e} \quad B^o \approx V_{C2} \quad \text{ou} \quad CMP: A^o \approx U_{C2} \quad \text{e} \quad B^* \approx (VD)_{C2}$$

onde o índice $C2$ indica que apenas as primeiras duas colunas da matriz são retidas. A notação RMP (*Row Metric Preserving*) indica que se preserva a métrica para as linhas, enquanto CMP (*Column Metric Preserving*) indica que a factorização em causa preserva a métrica para as colunas. Cada factorização tem um factor principal que contém os valores singulares e nota-se com " $*$ " e um factor standard notado por "o". Dependendo do Biplot escolhido, a métrica no espaço de linhas e colunas é diferente. Por exemplo, num CMP-Biplot, os produtos escalares das colunas de X coincidem com os produtos escalares dos

marcadores B^* e a distancia de Mahalanobis entre as linhas de X aproxima-se pela distancia euclidiana entre os marcadores A^o.

De facto, é possível utilizar representações Biplot para integrar informação de várias matrizes de dados. O problema a resolver é a obtenção de uma configuração consenso na qual seja possível comparar as configurações resultantes de análises Biplot. Neste contexto, Martín-Rodriguez et al (2001) propõem a Análise Metabiplot.

Sejam X_1 e X_2 duas matrizes de dados que contêm informação sobre J variáveis e I indivíduos em duas ocasiões ou situações distintas. Suponha-se que ambas as matrizes foram previamente submetidas à mesma análise Biplot após ter sido aplicado o mesmo tipo de transformação. Assuma-se que os Biplots k-dimensionais foram obtidos a partir dos k primeiros vectores singulares:

$$X_1 = U_1 D_1 V_1' \quad \text{e} \quad X_2 = U_2 D_2 V_2',$$

sendo U_1, D_1, V_1, U_2, D_2 e V_2 as respectivas matrizes de valores e vectores singulares.

É possível comparar dois subespaços definidos pelos marcadores num CMP-Biplot com base nos seguintes teoremas:

Teorema 1 O ângulo mínimo entre um vector arbitrário do espaço definido pelos marcadores linha (coluna) de X_1 resultante de um CMP-Biplot e um quase paralelo a este no espaço dos marcadores linha (coluna) de X_2 é dado por $cos^{-1}\sqrt{\lambda_1}$, onde λ_1 é o maior valor próprio de $M = U_1' U_2 U_2' U_1$ ($N = D_1 V_1' V_2 V_2' V_1 D_1$).

Teorema 2 Seja λ_i o i-ésimo maior valor próprio de $M(N)$, e a_i o vector próprio que lhe está associado. Seja $e_i = U_1 a_i$ e $c_i = U_2 U_2' e_i$ ($e_i = V_1 D_1 a_i$ e $c_i = V_2 V_2' e_i$) para $i = 1, \ldots, k$. Então, os vectores $e_1, \ldots, e_k$ constituem um conjunto ortogonal (não ortogonal para colunas) no subespaço definido pelos marcadores linha (coluna) de X_1. Os vectores $c_1, \ldots, c_k$ são um conjunto ortogonal no subespaço dos marcadores linha (coluna) de X_2. O ângulo entre o par (e_i, c_i) é dado por $cos^{-1}\sqrt{\lambda_1}$.

Estendendo este raciocínio a um conjunto de T matrizes de dados, é possível proceder à comparação de subespaços definidos pelos marcadores de vários grupos num CMP-Biplot com base no seguinte teorema:

Teorema 3 Seja b um vector no subespaço I-dimensional original (J-dimensional) e seja ∂_t o ângulo entre o vector b e o vector mais próximo paralelo a este no espaço gerado pelos marcadores linha (coluna) do grupo t ($t = 1, \ldots, g$). Então, o valor de b que minimiza o ângulo ∂_t e logo que maximiza $V = \sum_{t=1}^{g} cos^2 \partial_t$ é o vector próprio b_1, correspondente ao maior valor μ_1 da matriz $G = \sum_{t=1}^{g} U_t U_t'$ ($H = \sum_{t=1}^{g} V_t D_t D_t V_t'$).

Para um RMP-Biplot, os resultados são análogos.

4 Aplicação a Dados Reais

Com o objectivo de comparar as metodologias apresentadas, aplicou-se o Statis e o Metabiplot a um conjunto de dados reais, referentes ao Índice Metropolitano da Qualidade do Ar (IMECA) que se define como um valor representativo dos níveis de contaminação atmosférica e seus efeitos na saúde, dentro de uma região determinada. Os dados em estudo foram obtidos em http://www.sima.org.mx/ e reportam-se à medição dos níveis de Ozono (O3), Dióxido de Azoto (NO2), Dióxido de Enxofre (SO2), Monóxido de Carbono (CO) e Quantidade de Partículas Suspensas (PSup) entre a 1 e as 19h, em intervalos de duas horas, em cinco regiões do Vale do México: Zona Noroeste (NO), Zona Nordeste (NE), Zona Centro (CE), Zona Sudeste (SE) e Zona Sudoeste (SO). Os dados foram previamente standardizados.

Com a aplicação das metodologias Statis e Metabiplot, pretende-se comparar globalmente as regiões em estudo bem como averiguar a existência de uma tipologia comum às matrizes de informação em análise que possibilite a descrição, numa única estrutura, das matrizes referentes às várias horas.

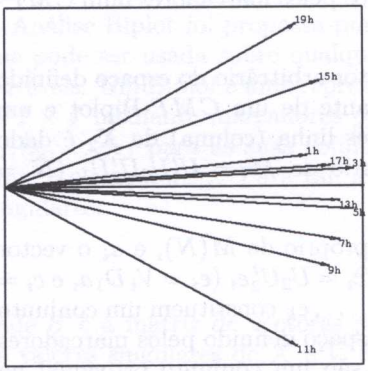

Figura 1: Gráfico da interestrutura.

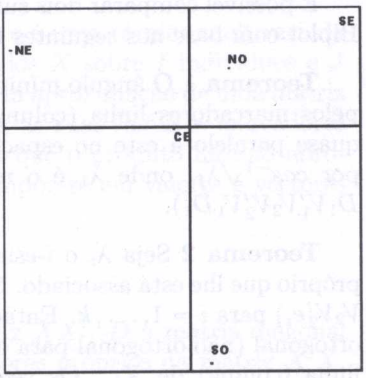

Figura 2: Representação compromisso.

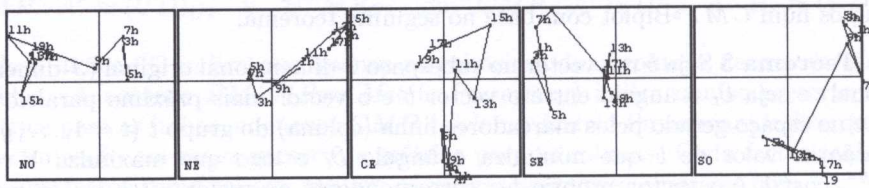

Figura 3: Trajectórias das regiões no espaço compromisso.

Da aplicação da metodologia Statis resultou a matriz dos coeficientes RV que expressa o grau de semelhança entre as estruturas em análise. Assim, submetendo a matriz RV a uma ACP, obteve-se a representação euclidiana (Figura 1), em que os cosenos dos ângulos formados pelos vectores aproximam a correlação vectorial entre os mesmos, ou seja, quanto menores forem os ângulos, mais parecidas serão as estruturas em causa. Os coeficientes RV obtidos são, em geral, elevados, podendo contudo referir-se que as matrizes relativas às 11, 15 e 19h são as que mais se afastam das restantes.

A partir da decomposição em valores e vectores singulares da matriz compromisso, constrói-se a imagem euclidiana compromisso dos indivíduos num espaço de baixa dimensão (Figura 2) que resume a informação proveniente de todas as configurações. A distância entre dois pontos interpreta-se em termos de distância compromisso entre regiões. A região sudoeste (SO) parece ser a que mais se afasta das restantes marcando o eixo 2.

Projectando as matrizes originais no espaço compromisso, obtêm-se as trajectórias das diferentes regiões no espaço compromisso ao longo do tempo (Figura 3). As trajectórias são, em geral, bastante irregulares, indiciando que ocorreram alterações bruscas no comportamento das variáveis ao longo do tempo.

Tabela 1: Quadrados dos cosenos entre os subespaços e o subespaço consenso.

Grupo	1h	3h	5h	7h	9h	11h	13h	15h	17h	19h
c1	0.60	0.93	0.87	0.96	0.85	0.66	0.64	0.93	0.94	0.84
c2	0.82	0.94	1	0.77	0.49	0.05	0.84	0.88	0.43	0.93

O mesmo conjunto de dados foi submetido a uma Análise Metabiplot. Começou por efectuar-se uma Análise Biplot das matrizes de dados correspondentes às diversas horas, procedendo-se posteriormente à comparação e integração dos resultados num subespaço consenso.

O primeiro passo consistiu em procurar o sistema de eixos ortogonais que melhor faz convergir as principais direcções de inércia. A proporção de inércia no plano é de 91%. Os quadrados dos cosenos entre os subespaços correspondentes a cada grupo (Tabela 1) interpretam-se como medida de similaridade de cada subespaço com as componentes comuns. Pode considerar-se que a similaridade entre as componentes de cada grupo e novas as componentes comuns $c1$ e $c2$ é próxima da unidade, indiciando que as novas componentes são bastante similares às de cada grupo. Constata-se que a matriz referente às 3h é a mais semelhante à estrutura consenso por ser aquela que forma ângulos menores com as componentes consenso. A configuração consenso (Figura 4) integra todas as estruturas em análise, é a mais parecida a todos os subespaços e permite observar o posicionamento das variáveis no plano principal. Assim, ao eixo 1 associa-se a variável SO2 enquanto o eixo 2 aparece essencialmente marcado

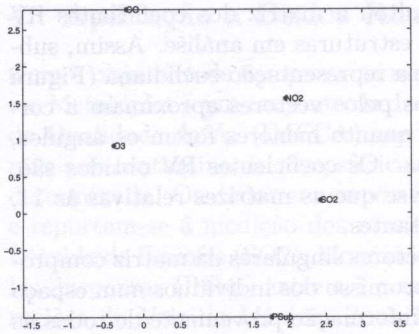

Figura 4: Configuração consenso (variáveis).

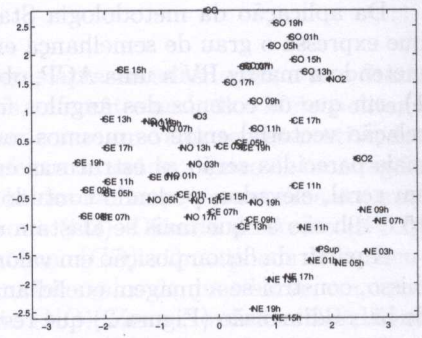

Figura 5: Projecção das regiões na configuração consenso.

pela variável CO. Projectando as regiões em estudo na configuração consenso, é possível observar o seu comportamento durante o período em estudo através da respectiva posição relativamente às variáveis (Figura 5).

5 Considerações finais

As metodologias Statis e Metabiplot constituem ferramentas úteis no tratamento e integração de informação proveniente de várias matrizes de dados e ambas possibilitam a obtenção de uma configuração consenso que integra as configurações resultantes das várias análises.

Para aplicar o Statis é necessário que os coeficientes de correlação vectorial entre configurações estejam próximos de 1, ou seja, as matrizes de covariância devem ser similares para que possam ser integradas, enquanto no caso do Metabiplot, as estruturas de covariância devem ser similares mas as componentes não têm por que coincidir. Enquanto no Statis a matriz compromisso resume a informação proveniente das configurações em estudo, sendo a mais correlacionada com todas as configurações, no Metabiplot a configuração consenso integra as configurações resultantes de várias análises Biplot. Desta forma, pode considerar-se que o Metabiplot proporciona uma interpretação mais rica em termos de estudo de relações indivíduos-variáveis que o Statis já que, ao integrar toda a informação numa única estrutura, permite identificar quais as variáveis responsáveis pela configuração obtida .

É interessante notar que no Statis assumem grande importância na matriz consenso as matrizes altamente correlacionadas com o primeiro eixo. Às matrizes pouco correlacionadas com o primeiro eixo, dá-se uma importância muito baixa, ainda que tais variáveis sejam decisivas na análise parcial. Por sua vez, o Metabiplot permite ultrapassar esta limitação.

Referências

[1] Gabriel, K. R. (1971). The Biplot graphic display os matrices with application to principal components analysis, *Biometrika* Vol.58(3), pp.453-467.

[2] Escoufier, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, pp.750-760.

[3] L´Hermier des Plantes, H. (1976). *Structuration des tableaux a trois indices de la statistique.* Thèse de 3 ème cycle, Université de Montpellier.

[4] Lavit, Ch. (1988). *Analyse conjointe de tableaux quantitatifs.* Masson, Paris.

[5] Lavit, Ch., Escoufier, Y., Sabatier, R. and Traissac, P. (1994). The ACT (Statis Method). *Computacional Statistics and Data Analysis.* Vol.18, pp.97-119.

[6] Martín-Rodríguez, J., Galindo-Villardón, M.P. and Vicente-Villardón, J.L. (2001). Comparison and integration of subspaces from a biplot perspective. *Journal of Statistical Planning and Inference.* Vol.102,(2), pp.1-13