Uma aplicação de correlação canônica não linear em dados de tuberculose

An application of nonlinear canonical correlation in tuberculosis data

Dalila Camêlo Aguiar
Universidade de Granada - UGR, Granada, Espanha
dalilacamelo@correo.ugr.es

Edwirde Luiz Silva Camêlo
Universidade Estadual da Paraíba - UEPB, Campina Grande, PB
edwirde@uepb.edu.br

Ramón Gutiérrez Sánchez
Universidade de Granada - UGR, Granada, Espanha
ramongs@ugr.es

Resumo: O trabalho apresenta uma aplicação de correlação canônica não linear em dados de tuberculose.O objetivo desta técnica é determinar a semelhança entre dois ou mais conjuntos de variáveis explicando ao máximo a variância das relações entre os conjuntos em um espaço de poucas dimensões. O caso estudado, avalia a relação linear existente entre o desfecho do tratamento da tuberculose, influenciada pela faixa etária e condição de infectado por TB/HIV nas microrregiões de residência do Estado da Paraíba no período de 2009 a 2015.A análise dos resultados determina que existe relação entre os dois conjuntos de variáveis. A importância do estudo resulta na compreensão da variabilidade das notificações de TB nas microrregiões dada a faixa etária do paciente conforme sua condição de infectado por TB/HIV e o desfecho do tratamento.

Palavras-chave: CCNL; estatística multivariada; dados epidemiológicos; microrregiões da Paraíba.

Abstract: The paper presents an application of canonical correlation nonlinear in tuberculosis data. The objective of this technique is to determine the similarity between two or more sets of variables explaining to the maximum the variance of the relations between the sets in a space of few dimensions. The case studied evaluated the linear relationship between the TB treatment outcome, influenced by the age group and condition of infected for TB/HIV in the micro-regions of the State of Paraíba in the period from 2009 to 2015. The analysis of the results determines that there is a relationship between the two sets of variables. The importance of the study results in the understanding of the variability of TB notifications in the microregions given the age group of the patient according to their condition of infected for TB/HIV and the treatment outcome.

Key words: OVERALS; multivariate statistics; epidemiological data; microregions of

Recebido em 23/11/2017 - Aceito em 21/02/2019

Paraíba.

1 Introdução

Sabe-se que a análise de correlação canônica padrão é uma extensão da regressão linear múltipla no caso de várias variáveis dependentes (explicativas ou resposta). Conhecida também por *Nonlinear Canonical Correlation Analysis* (OVERALS), a análise de correlação canônica não linear (CCNL) coincide com a análise de correlação canônica categórica por escalonamento ótimo. O objetivo deste procedimento é determinar a similaridade entre os conjuntos de variáveis categóricas.

A análise de correlação canônica padrão é uma extensão da regressão múltipla, na qual o segundo conjunto não contém uma única variável de resposta, mas várias. O objetivo é explicar o máximo possível da variância sobre as relações existentes entre dois conjuntos de variáveis numéricas em um espaço de poucas dimensões. Inicialmente, as variáveis de cada conjunto são combinadas linearmente de modo que as combinações lineares tenham uma correlação máxima entre si. Uma vez que essas combinações são dadas, é estabelecido que as combinações lineares subsequentes não estão correlacionadas com as combinações anteriores e que elas também tenham a maior correlação possível.

O propósito da análise de CCNL é determinar a semelhança entre dois ou mais conjuntos de variáveis. Como na análise de correlação canônica não linear, o objetivo é explicar ao máximo a variância das relações entre os conjuntos em um espaço de poucas dimensões [1]. No entanto, ao contrário da análise de correlação canônica linear, a análise de correlação canônica não linear não supõe que haja um nível de intervalo de medida ou que as relações sejam lineares. Outra diferença importante é que a análise de correlação canônica não-linear estabelece a semelhança entre os conjuntos mediante a comparação simultânea das combinações lineares das variáveis em cada conjunto com um conjunto desconhecido, isto é, os escores do objeto.

O tratamento de dados epidemiológicos requer da exploração e avaliação de diferentes modelos matemáticos, bem como técnicas estatísticas que permitem estudar e interpretar o desfecho do tratamento da tuberculose (TB) e variáveis influenciadoras.

A TB é considerada a primeira causa de morte em pacientes com AIDS no Brasil e doentes com coinfecção TB/HIV apresentam maior probabilidade de terem desfecho desfavorável ao tratamento da TB.

Este trabalho mostra uma aplicação da técnica multivariada denominada análise de correlação canônica não linear com a pretensão de estudar a relação existente entre o desfecho do tratamento da tuberculose (TB) e faixa etária segundo a condição de infectado por TB/HIV nas microrregiões de residência do Estado da Paraíba no período de 2009 à 2015.

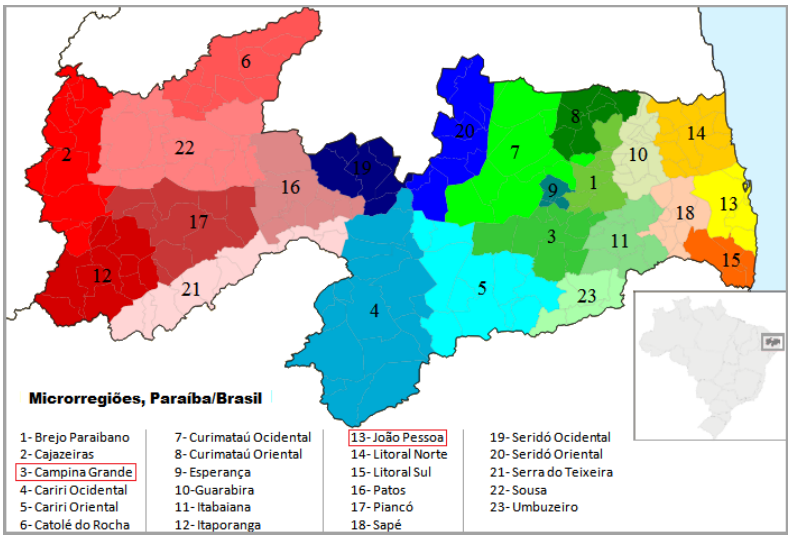

Figura 1. Microrregiões da Paraíba, Brasil.

Foram considerados desfechos clínicos do tratamento de TB, os desfavoráveis (óbito, abandono e falência de tratamento)e alta por cura como desfecho favorável; faixa etária com duas categorias possíveis: menor igual a 39 anos (≤ 39) e idade igual ou superior a 40 anos (≥ 40) e condição de TB/HIV (sim ou não), para as 21 microrregiões com exceção de Campina Grande e João Pessoa por serem *outliers*. O estudo considera dados da Vigilância Epidemiológica do Ministério da Saúde, fornecido pelo Sistema de Informação de Agravos de Notificação (SINAN) [2]. Por se tratar de uma pesquisa de dados secundários e não envolver diretamente seres humanos, este estudo não foi submetido à avaliação do Comitê de Ética em Pesquisa. Os dados obtidos foram adicionados a planilha do programa Excel® 2010, em seguida analisados pelo programa estatístico SPSS, versão 18.0 para Windows [3].

2 Desenvolvimento teórico

Análise de correlação não linear foi primeiramente introduzida por Gifi [4] e Van der Burg [5]. O objetivo da função de otimização de uma análise de CCNL é encontrar o escore X e um conjunto de Y_j para $j = 1, \ldots, m$, o subíndice j indica que pode ser restringido de várias formas, de tal forma que a função:

$$\sigma(X;Y) = \frac{1}{k}\Sigma_k tr\left[(x - \Sigma_{j\in j(k)} G_j Y_j)' M_k (x - \Sigma_{j\in j(k)} G_j Y)\right]$$

seja mínima sobre a normalização da restrição $X^{'} M_c X = knI$, em que $M_c = \Sigma_k M_k$ e I é a matriz identidade $p \times p$. A inclusão de M_k em $\sigma(X;Y)$ prover o seguinte mecanismo para perda de peso: sempre que algum valor dos dados do objeto i em k cai fora do intervalo $[1, k_j]$, a circunstância indicaria outra genuíno valor perdido ou similar a outro valor perdido para causa da análise, e todos outros valores do objeto i em k são desconsiderados.

2.1 Nível de escalonamento

Em análise de correlação não linear as variáveis podem-se ser classificadas em dois ou mais grupos (neste trabalho elas foram classificadas em dois conjuntos). As variáveis da análise se escalam como nominais simples, ordinais ou numéricas. O máximo número de dimensões utilizados no procedimento depende do nível de escalonamento ótimo das variáveis [1].

Como exemplo se todas as variáveis estão especificadas como ordinais, nominais simples ou métricas, o número máximo de dimensões é o mesmo do número de observações menos 1 e o número total de variáveis. Entretanto, se apenas define-se dois conjuntos de variáveis, o número máximo de dimensões é o número de variáveis no conjunto menor. Se algumas variáveis são nominais múltiplas, o número máximo de dimensões é o número total de categorias nominais múltiplas mais o número de variáveis nominais não múltiplas menos o número de variáveis nominais múltiplas [6].

Os possíveis níveis de escalonamentos utilizados para quantificar cada variável são os seguintes:

- Ordinal: a ordem das categorias da variável observada conserva-se na variável escalar ótima. Os pontos das categorias estarão sobre uma reta (vetor) que passam pela origem.

- Nominal: a única informação da variável observada que se conserva na variável escalar ótima é a agrupação dos objetos nas categorias. Não se conserva a ordem das categorias das variáveis observadas. Os pontos de categorias estarão sobre uma reta (vetor) que passa pela origem.

O nível de escalonamento ótimo é distinguido na análise de correlação não linear através das seguintes categorias:

a) Múltipla nominal: $Y_j = Y_j$ (apenas restrito igualmente);

b) Nominal simples: $Y_j = Y_j a'$ (igualdade e grau – uma restrição);

em que: a_j variável peso para simples variável, de ordem p, Y_j quantificação da categoria para múltiplas variáveis, de ordem $k_j \times p$.

c) Ordinal simples: $Y_j = Y_j a_j^{'}$ e $Y_j \in C_j$ (igualdade, grau 1 e monotonicamente restrito). A restrição monotônica $Y_j \in C_j$ significa que Y_j deve ser locado no cone convexo de todos k_j - vetores com elementos não decrescentes.

d) Numeral simples: $Y_j = Y_j a_j^{'}$ e $Y_j \in L_j$ (igualdade, grau 1 e monotonicamente restrito). A restrição monotônica $Y_j \in L_j$ significa que Y_j deve ser locado no subespaço de todo k_j - vetores que são uma linear transformação do vetor de k_j sucessivamente interagido.

Para cada variável, estes níveis podem ser escolhidos independentemente. A geral necessidade para todas as opções é que igual categoria indica receber igual quantificação. A maior necessidade para a não múltipla opção é $Y_j = Y_j a_j^{'}$, isto é, Y_j é de grau 1; para o propósito da identificação, Y_j é sempre normalizado tal que $y_j^{'} D_j y_j = n_w$.

3 Resultados e discussão

Os valores ajustados e perdidos indicam a eficácia do ajuste da solução da análise de correlação canónica não linear para os dados ótimos quantificados em relação à associação entre os conjuntos. O resumo da Tabela 1 mostra o valor ajustado, os valores perdidos e os autovalores.

Tabela 1. *Resumo da perda dos conjuntos por dimensão.*

		Dimensão		
		1	2	**Soma**
Perda	Conjunto 1	0,276	0,505	0,781
	Conjunto 2	0,276	0,452	0,728
	Média	0,276	0,478	0,755
Autovalores		0,724	0,522	
Ajuste				1,245

A perda é dividida pelas dimensões e conjuntos. Para cada dimensão e conjunto, a perda representa a proporção de variação nos escores do objeto que não pode ser explicada através da combinação ponderada das variáveis definidas. Na segunda dimensão ocorreu a maior perda (0,478). A perda média sobre os conjuntos informa a diferença entre o máximo ajuste e o real (2-1,245 = 0,755), e portanto, o valor do ajuste mais a perda média é igual ao número de dimensões.

O autovalor indica a parte da relação mostrada por cada dimensão. O autovalor para cada dimensão é igual a 1 menos a perda média para a dimensão. Os autovalores são somados até o ajuste total. A primeira dimensão explica, 0,724/1,245 = 58% do ajuste real.

O valor de ajuste é bom (1,245 sobre 2) o que indica uma boa qualidade do ajuste por correlação canônica não linear.

A perda de cada conjunto se divide pela análise de correlação canônica não linear de muitas formas. A Tabela 2 apresenta o ajuste múltiplo, o ajuste simples e a perda simples por dimensão para cada variável de cada um dos conjuntos da análise. O ajuste múltiplo menos ajuste simples é igual a perda simples.

0.9

Tabela 2. *Partição do ajuste e da perda dos conjuntos por dimensão.*

		Ajuste múltiplo			**Ajuste simples**			**Perda simples**		
		Dimensão			Dimensão			Dimensão		
	Conjunto	1	2	Soma	1	2	Soma	1	2	Soma
1	Microrregião	0,694	0,033	0,728	0,694	0,021	0,715	0,000	0,012	0,012
	Faixa etária	0,019	0,483	0,502	0,019	0,483	0,502	0,000	0,000	0,000
2	Condição de TB/HIV	0,730	0,000	0,730	0,730	0,000	0,730	0,000	0,000	0,000
	Desfecho	0,004	0,550	0,554	0,004	0,550	0,554	0,000	0,000	0,000

A perda simples indica a perda que é obtida restringindo as variáveis a um conjunto de quantificações (isto é, nominal simples, ordinal ou nominal). Na Tabela 2 percebe-se que o

ajuste simples e o múltiplo são quase iguais, o que significa que as coordenadas múltiplas estão quase em uma linha reta na direção marcada pelos pesos.

O ajuste múltiplo é igual à variância das coordenadas da categoria múltipla para cada variável. Pelo ajuste múltiplo é possível verificar que as variáveis *Condição de TB/HIV* e *Desfecho do tratamento da TB* são as que melhor discriminam, sendo que a primeira é totalmente discriminada na primeira dimensão. Os valores ajustados, somados ao longo das duas dimensões, são 0,730 para *Condição de TB/HIV* e 0,554 para *Desfecho do tratamento da TB*. Ou seja, a condição de TB/HIV de um paciente proporciona maior força discriminatória que o desfecho ao que está inserido.

Os diferentes níveis nos quais cada variável pode ser escalada impõem restrições sobre as quantificações [7]. Os gráficos de transformação ilustram a relação entre as quantificações e as categorias originais resultantes do nível ótimo de escalonamento ótimo selecionado, como se observa na Figura 2.

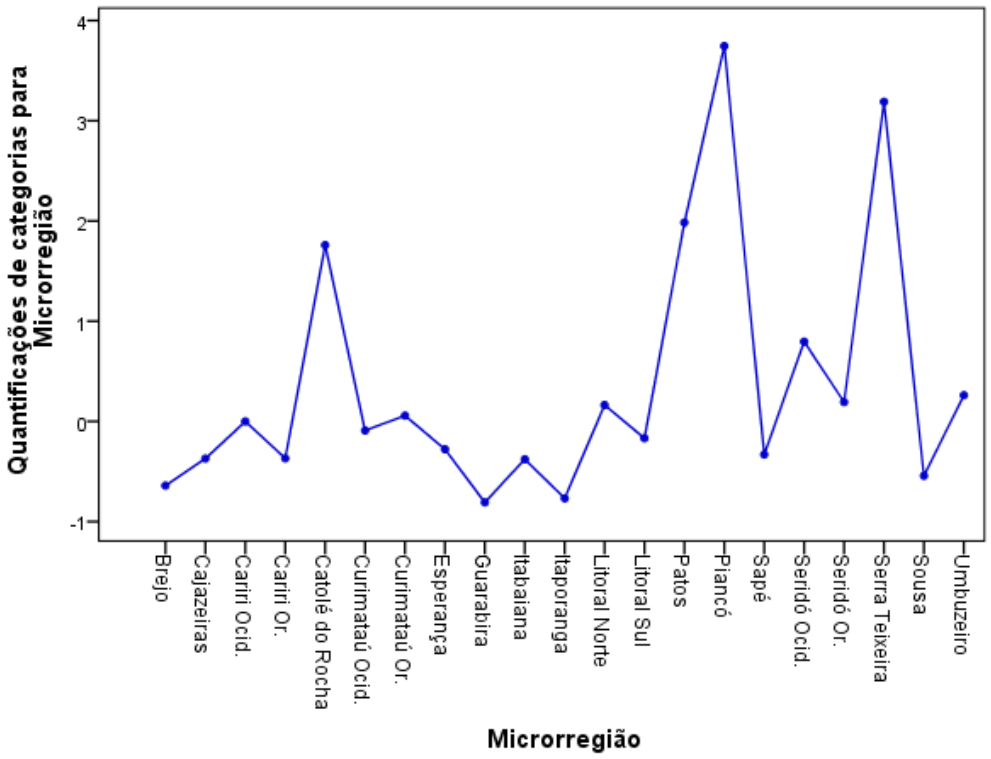

Figura 2. Transformação para a variável desfecho do tratamento de TB por Microrregião, Paraíba, Brasil.

As quantificações para as microrregiões de residência apresentam uma tendência ascendente principalmente nas microrregiões de Catolé do Rocha, Patos, Piancó e Serra do Teixeira, por outro lado, Guarabira e Itaporanga recebem as menores quantificações.

Para cada variável tratada como nominal simples, ordinal ou numérica, se determinam as quantificações, as coordenadas das categorias simples e as coordenadas de categoria múltiplas. Na Tabela 3 tem-se estas estatísticas para as microrregiões.

Tabela 3. *Coordenadas por tipo de categorias para microrregiões.*

			Coordenadas de categoria simples		**Coordenadas de categoria múltipla**	
	Frequência		Dimensão		Dimensão	
	marginal	**Quantificações**	1	2	1	2
Brejo	51	-0,640	0,533	-0,094	0,523	-0,156
Cajazeiras	75	-0,371	0,309	-0,054	0,305	-0,076
Cariri Ocid.	24	-0,001	0,001	0,000	-0,002	-0,014
Cariri Or.	26	-0,369	0,307	-0,054	0,290	-0,153
Catolé do Rocha	26	1,759	-1,465	0,257	-1,422	0,502
Curimataú Ocid.	27	-0,091	0,076	-,013	0,087	0,048
Curimataú Or.	27	0,058	-0,048	0,009	-0,048	0,014
Esperança	25	-0,278	0,232	-0,041	0,214	-0,139
Guarabira	163	-0,809	0,674	-0,118	0,648	-0,265
Itabaiana	74	-0,378	0,315	-0,055	0,337	0,069
Itaporanga	30	-0,767	0,639	-0,112	0,653	-0,033
Litoral Norte	127	0,163	-0,136	0,024	-0,123	0,099
Litoral Sul	94	-0,169	0,141	-0,025	0,152	0,039
Patos	64	1,985	-1,653	0,290	-1,683	0,120
Piancó	24	3,745	-3,120	0,548	-3,144	0,409
Sapé	146	-0,330	0,275	-0,048	0,298	0,080
Seridó Ocid.	12	0,796	-0,663	0,116	-0,645	0,220
Seridó Or.	22	0,192	-0,160	0,028	-0,145	0,115
Serra Teixeira	30	3,188	-2,656	0,466	-2,652	0,487
Sousa	133	-0,544	0,453	-0,080	0,438	-0,168
Umbuzeiro	17	0,261	-0,217	0,038	-0,192	0,181

As coordenadas de uma determinada categoria correspondem à quantificação multiplicada pelos pesos da dimensão da variável. As coordenadas das categorias simples para a microrregião de Piancó (-3,120, 0,548) são a quantificação (3,745) multiplicada pelos pesos da dimensão (-0,833, 0,146). As coordenadas das categorias múltiplas para as variáveis que são tratadas como simples, ordinais ou numéricas, representam as coordenadas das categorias no espaço do objeto antes de aplicar as restrições ordinais ou lineares. Esses valores são minimizadores sem restringir a perda. Para as variáveis nominais múltiplas, estas coordenadas representam as quantificações das categorias.

A Figura 3 apresenta o gráfico dos centroides rotulados pelas categorias das variáveis. Este gráfico mostra a eficácia com a que as variáveis separam grupos de objetos (os centroides estão no centro de gravidade dos objetos).

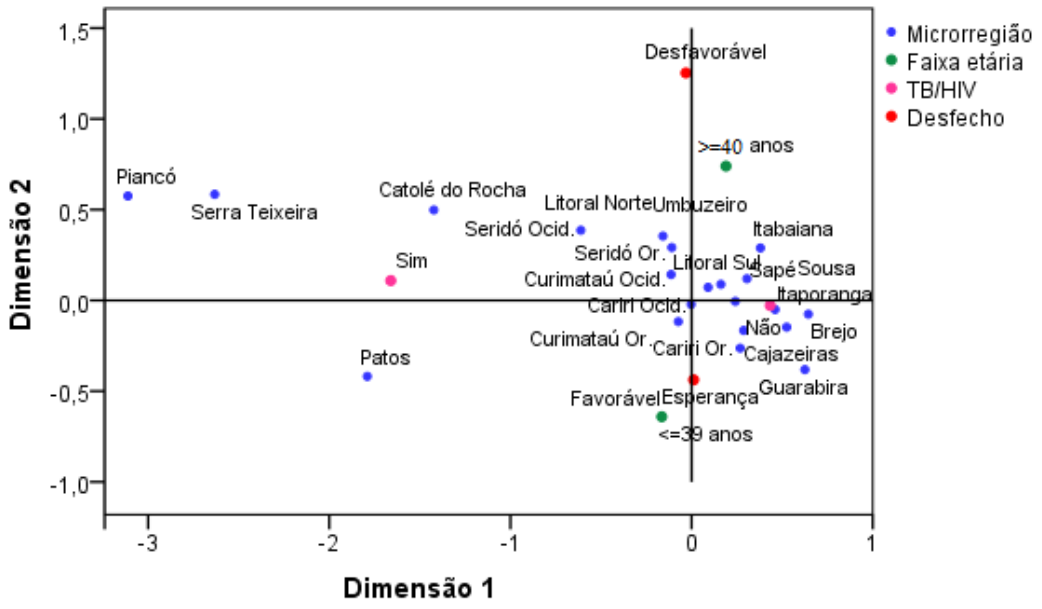

Figura 3. Centróides rotulados por variáveis. Paraíba, Brasil.

Com destaque para as microrregiões de Itaporanga, Brejo, Cajazeiras, Cariri Oriental, Guarabira e Esperança, o desfecho favorável do tratamento da TB (cura) está relacionado a pacientes com idade inferior a 40 anos principalmente aos que não possuem a comorbilidade TB/HIV, enquanto o desfecho desfavorável (óbito, abandono e falência de tratamento) está associado a faixa etária de maior idade ($\geq$ 40 anos).

4 Conclusões

No presente estudo, a eficácia do modelo demonstrou uma boa qualidade do ajuste (62%) por correlação canônica não linear, em que a primeira dimensão explicou 58% do ajuste real.

A semelhança estabelecida entre os conjuntos através da comparação simultânea das combinações lineares das variáveis no conjunto 1 (microrregião e faixa etária) e conjunto 2 (condição de TB/HIV e desfecho clínico do tratamento) com os escores do objeto permitiu identificar associações para algumas microrregiões entre o desfecho do tratamento da TB, faixa etária segundo a condição de TB/HIV. Identificou-se também que a condição de TB/HIV de um paciente proporciona maior força discriminatória que o tipo de desfecho ao qual pertence.

Estas evidências demandam mais investigação sobre o assunto e finalmente é esperado que o conhecimento gerado pelo presente estudo possa contribuir para novos estudos e decisões efetivas por parte dos programas de controle da TB do Estado da Paraíba.

Referências

[1] LÓPEZ, C.P. Métodos estadísticos avanzados con SPSS. Thomson: Madrid. 613-622p. 2005.

[2] BRASIL. Ministério da Saúde. Sistema de Informação de Agravos de Notificação Tuberculose – casos confirmados notificados no Sistema de Informação de Agravos de Notificação – SINAN. Brasília (DF). Disponível em: $< http://www2.datasus.gov.br/http://www2.datasus.gov.br/ >$ Acesso em: 17 de setembro de 2017.

[3] SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.

[4] GIFI, A. Nonlinear multivariate analysis. Chichester: John Wiley and Sons. 1990.

[5] VAN Der Burg, E. Nolinear canonical carrelation and some related techniques. Leiden: DSWO Press, 1988.

[6] KRUSKAL, J.B. Multidimensional Scaling by Optimizing Gooddness of Fit to a Nonmetric Hypothesis, v. 29, n. 1, 1-27p. 1964a.

[7] HO, R. Handbook of univariate and multivariate data analysis and interpretation with SPSS. Boca Raton: Chapman. 2006.