

Análise de correlação canônica na descrição de potenciais de desenvolvimento nos municípios de Minas Gerais

Naje Clécio Nunes da Silva¹

Wederson Leandro Ferreira¹

Gilberto Rodrigues Liska¹

João Domingos Scalon²

Marcelo Ângelo Cirillo²

1 Introdução

Conforme Johnson & Wichern (2007), a análise de correlação canônica é uma técnica estatística multivariada que mede a existência e a intensidade da associação entre dois conjuntos ou grupos de variáveis aleatórias (X e Y), cujo objetivo principal é determinar os pares de combinações lineares $U = a'X$ e $V = b'Y$ tais que tenham a maior correlação possível, ou seja, sendo o primeiro grupo, de p variáveis, representado pelo vetor aleatório $X(p \times 1)$ e o segundo grupo, de q variáveis, representado pelo vetor aleatório $Y(q \times 1)$, então deve-se encontrar os vetores $a'(1 \times p)$ e $b'(1 \times q)$ para os quais a correlação entre $U = a'X$ e $V = b'Y$ seja máxima.

As combinações lineares U e V que podem ser construídas são chamadas de variáveis canônicas e a correlação entre elas ($Corr(U, V)$) é chamada de correlação canônica.

Neste trabalho utilizou-se da análise de correlação canônica para os potenciais de desenvolvimento na área econômica e agropecuária, por meio de um conjunto de dados oriundos do Zoneamento Ecológico Econômico (ZEE) de todos os municípios do estado de Minas Gerais referentes ao ano de 2004 (Scolforo et al., 2008b).

2 Material e métodos

Na análise dos potenciais, foi utilizado dados de todos os 853 municípios do estado de Minas Gerais, e destes foram analisadas 6 variáveis: 1 - Índice do valor adicionado industrial (IVAI); 2 - Índice do valor adicionado serviços (IVAS); 3 - Índice do produto interno bruto (PIB); 4 - Índice do valor adicionado agropecuário (IVAA); 5 - Índice de agricultores familiares (IAF); 6 - Índice de emprego formal (IEF).

¹ Doutorando do Programa de Pós-Graduação em Estatística e Exp. Agropecuária – DEX - UFPA.
Email: naje.silva@ifac.edu.br, wedelean@hotmail.com e gilbertoliska@dex.ufpa.br.

² Docente do Programa de Pós-Graduação em Estatística e Exp. Agropecuária – DEX - UFPA.
Email: scalon@dex.ufpa.br e macufla@dex.ufpa.br.

De acordo com Johnson & Wichern (2007), para validação da análise de correlação canônica se faz necessário uma análise da matriz de covariâncias (Σ) ou de correlações (R) a fim de determinar se elas são próximas ou não da matriz nula.

As hipóteses para validar a análise de correlação canônica podem ser dadas por:

$$H_0 : \sum_{12} = 0 \left(\rho_{U_1, V_1} = \rho_{U_2, V_2} = \dots = \rho_{U_p, V_p} = 0 \right) \text{ contra } H_a : \sum_{12} \neq 0.$$

A estatística do teste é dada por:

$$-2\ln(\Lambda) = - \left[n - 1 - \frac{1}{2}(p + q + 1) \right] \ln \left[\prod_{i=1}^p (1 - \lambda_i) \right], \quad (1)$$

em que n é o tamanho da amostra. A estatística do teste está associada a uma distribuição Qui-quadrado $\chi^2_{(v, \alpha)}$ com $v = pq$ graus de liberdade. Se $-2\ln(\Lambda) > \chi^2_{(v, \alpha)}$ então rejeita-se H_0 , ou seja, a matriz de covariâncias é diferente de zero para algum nível de significância α considerado, indicando que se pode fazer análise de correlação canônica.

Tem-se também, se a hipótese nula $H_0 : \sum_{12} = 0$ for rejeitada, é natural buscar um número de correlações canônicas r que diferem significativamente de zero. Em que ρ_k^* é a anotação abreviada de ρ_{U_k, V_k}^* . A princípio, testa-se a hipótese de que a primeira correlação é não nula e as demais $(p - 1)$ são nulas; em seguida, testa-se que as duas primeiras correlações são não nulas e as demais $(p - 2)$; e assim por diante. Para o k -ésimo passo desse processo testa-se a hipótese $H_0^{(k)}$ dada por:

$$\begin{cases} H_0^{(k)} : \rho_1^* \neq 0, \rho_2^* \neq 0, \dots, \rho_k^* \neq 0, \rho_{k+1}^* = \rho_{k+2}^* = \dots = \rho_p^* = 0 \\ H_a^{(k)} : \rho_i^* \neq 0 \text{ para algum } i \geq k + 1 \end{cases}$$

A estatística do teste é dada por:

$$-2\ln(\Lambda) = - \left[n - 1 - \frac{1}{2}(p + q + 1) \right] \ln \left[\prod_{i=k+1}^p (1 - \lambda_i) \right], \quad (2)$$

em que a estatística do teste está também associada a uma distribuição Qui-quadrado $\chi^2_{(v, \alpha)}$, porém com $v = (p - k)(q - k)$ graus de liberdade.

De acordo com Mingoti (2007), a proporção da variância total explicada por cada variável é dada por:

$$PVTE_{U_k} = \frac{\sum_{i=1}^p \text{corr}(U_k, X_i)^2}{p} \times 100 \text{ e } PVTE_{V_k} = \frac{\sum_{i=1}^q \text{corr}(V_k, Y_i)^2}{q} \times 100. \quad (3)$$

Todas as análises foram feitas no software R (R Development Core Team, 2012).

3 Resultados e discussões

Na Figura 1, é apresentado o gráfico biplot com o intuito de agrupar as variáveis com correlações próximas, preliminar à análise de correlação canônica.

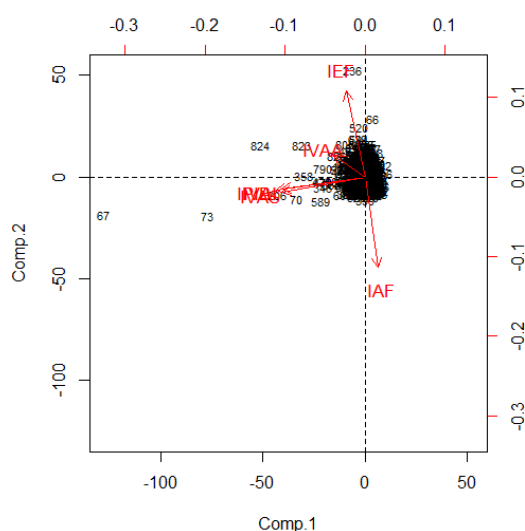


Figura 1: Biplot considerando a matriz de correlação das variáveis em estudo.

Pela Figura 1, observa-se que as variáveis IVAI, IVAS e IPIB podem formar o grupo X (indicador econômico), já que essas variáveis são correlacionadas, pois os vetores por elas formados possuem normas semelhantes e estão no mesmo quadrante. As variáveis IVAA, IAF e IEF não são independentes das demais variáveis e podem formar o grupo Y (indicador agropecuário). Na tabela 1, tem-se a matriz de correlações entre as variáveis de cada grupo.

Tabela 1: Matriz de correlação entre as 6 variáveis referentes aos 853 municípios.

Variável	IVAI	IVAS	IPIB	IVAA	IAF	IEF
IVAI	1,0000	0,7388	0,8899	0,1379	-0,0646	0,1057
IVAS	0,7388	1,0000	0,8814	0,2144	-0,0164	0,0930
IPIB	0,8899	0,8813	1,0000	0,2605	-0,0527	0,1031
IVAA	0,1379	0,2144	0,2605	1,0000	-0,1240	0,0718
IAF	-0,0647	-0,0164	-0,0527	-0,1240	1,0000	-0,4164
IEF	0,1057	0,0930	0,1030	0,0718	-0,4164	1,0000

De acordo com a matriz de correlação, apresentadas na Tabela 1, existem correlações fortes entre as variáveis IPIB e IVAI, IPIB e IVAS, IVAI e IVAS (variáveis do grupo X), localizados na parte superior esquerda (3x3), e correlações fraca e moderada entre as variáveis do grupo Y, localizados na parte inferior direita (3x3).

Para validação da análise de correlação canônica, o valor da estatística do teste foi de $-2\ln(\Lambda) = 4,87$, e o valor tabelado da Qui-quadrado foi de $\chi^2_{(9,5\%)} = 3,33$. Como $-2\ln(\Lambda) > \chi^2_{(9,5\%)}$, então rejeita-se a hipótese nula, ou seja, a matriz das covariância dos grupos X e Y é estatisticamente diferente de zero ao nível de 5% de significância, indicando que se pode fazer análise de correlação canônica.

Os pares de variáveis canônicas formados pelos dois grupos e suas respectivas correlações estão expressos a seguir:

$$\begin{cases} U_1 = 1,1178IVAI + 0,4499IVAS - 2,1017IPIB \\ V_1 = -0,0197IVAA - 0,0197IAF - 0,0816IEF \end{cases} \quad e \quad \rho_{U_1, V_1} = 0,3440, \quad (4)$$

$$\begin{cases} U_2 = -1,4038IVAI - 0,0834IVAS + 0,8903IPIB \\ V_2 = 0,1003IVAA + 0,0685IAF - 0,4637IEF \end{cases} \quad e \quad \rho_{U_2, V_2} = 0,0930, \quad (5)$$

$$\begin{cases} U_3 = -0,1028IVAI - 2,1716IVAS + 1,6092IPIB \\ V_3 = 0,0215IVAA - 0,2131IAF - 0,4135IEF \end{cases} \quad e \quad \rho_{U_3, V_3} = 0,0719. \quad (6)$$

Verifica-se que o primeiro par de variáveis canônicas (4), apresentou uma correlação moderada de 0,3440 e os demais pares de variáveis canônicas apresentaram correlações fracas.

Realizando o teste sequencial das correlações canônicas individuais encontrou-se que:

Para $H_0^{(1)} : \rho_{U_1, V_1}^* \neq 0, \rho_{U_2, V_2}^* = \rho_{U_3, V_3}^* = 0$ contra $H_a^{(1)} : \rho_{U_i, V_i}^* \neq 0$ para algum $i \geq 2$, obteve-se que $-2\ln(\Lambda) = 8,06 > \chi^2_{(4,5\%)} = 0,71$, então rejeita-se a hipótese nula, ou seja, uma das duas correlações ($i \geq 2$) é diferente de zero ao nível de 5% de significância.

Para $H_0^{(2)} : \rho_{U_1, V_1}^* \neq 0, \rho_{U_2, V_2}^* \neq 0, \rho_{U_3, V_3}^* = 0$ contra $H_a^{(2)} : \rho_{U_3, V_3}^* \neq 0$, obteve-se que $-2\ln(\Lambda) = 12,93 > \chi^2_{(1,5\%)} = 0,001$, ou seja, a terceira correlação também é diferente de zero ao

nível de 5% de significância. Logo, utilizam-se os três pares de correlações canônicas para representar os dois grupos de variáveis.

A proporção da variação total explicada por cada variável canônica está expressa a seguir:

$$\left\{ \begin{array}{lll} PVTE_{U_1} = 61,87\% & PVTE_{U_2} = 66,39\% & PVTE_{U_3} = 26,00\% \\ PVTE_{V_1} = 35,98\% & PVTE_{V_2} = 7,18\% & PVTE_{V_3} = 30,76\% \end{array} \right. \quad (7)$$

Em (7), observa-se que a proporção da variação total explicada para U_1 é de 61,87% e para V_1 é de 35,98%, ou seja, a variável U_1 representa 61,87% da variação total do grupo X e a variável V_1 representa 35,98% da variação total do grupo Y. Portanto, o primeiro par de variáveis canônicas representa os grupos X e Y, uma vez que a correlação canônica entre esses grupos é de 34,40%, e os demais pares apresentaram correlações canônicas muito baixas.

4 Conclusões

A análise de correlação canônica foi empregada com sucesso para as variáveis do zoneamento ecológico econômico. O grupo formado pelas variáveis IVAI, IVAS e IPIB pode ser explicado pelo grupo de variáveis IVAA, IAF e IEF, uma vez que a proporção da variação total do primeiro par de variáveis canônicas associadas aos grupos é de 61,87% para o primeiro grupo e de 35,98% para o segundo grupo.

5 Referências

- [1] JOHNSON, R. A., WICHERN, D.W.; **Applied Multivariate Statistical Analysis**. 6. ed. New Jersey: Prentice Hall, 2007. 794 p.
- [2] MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Editora UFMG, Belo Horizonte, 2005. 295 p.
- [3] R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing. www.r-project.org. , 2012.
- [4] SCOLFORO, J. R. S.; OLIVEIRA, A. D.; CARVALHO, L. M. T. **Zoneamento ecológico-econômico do Estado de Minas Gerais: componente sócio econômico**. Lavras: UFLA, 2008b. 195 p.