APLICAÇÃO DE ANÁLISE PROCRUSTES EM DADOS ARQUEOMÉTRICOS

Paulo Tadeu Meira e Silva de OLIVEIRA[1], Lúcia Pereira BARROSO[2], Casimiro Sepúlveda MUNITA[3]

1. INTRODUÇÃO

Há muitos anos tem sido grande a preocupação dos estatísticos quando dispõem de um grande número de variáveis para análise. Muitas dessas variáveis podem ser redundantes e nesse caso torna-se útil a eliminação de algumas delas, mantendo a estrutura dos dados sem perda de muita informação e tornando menor a dimensionalidade do problema, facilitando assim a interpretação das inter-relações envolvidas.

Na literatura, muitos trabalhos têm sido publicados sobre métodos de seleção de variáveis. Entretanto, existem poucos trabalhos que utilizam um método objetivo para a eliminação das que não são significativas.

Neste trabalho, aplicou-se o método de Análise Procrustes (Krzanowski, 1996) como regra de parada na eliminação de variáveis, na determinação, por ativação com nêutrons, dos elementos químicos Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Ce, Eu, Tb, Hf e Th em 75 amostras de fragmentos cerâmicos.

2. MATERIAL E MÉTODOS

O método apresentado neste trabalho foi aplicado à concentração de quinze elementos químicos (Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Ce, Eu, Tb, Hf e Th) em um conjunto de 75 amostras de fragmentos cerâmicos coletados em um sítio arqueológico. Os detalhes sobre preparação das amostras e o método analítico foram publicados por Munita et al. (2000).

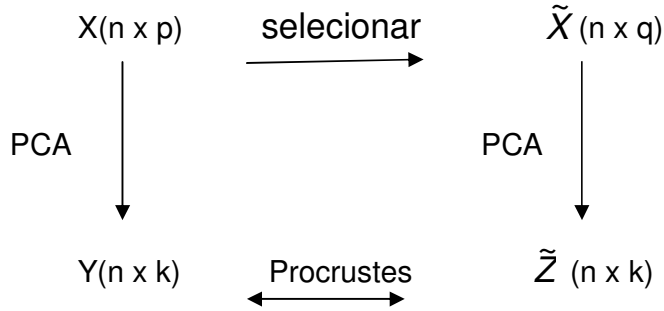
Inicialmente, considera-se uma matriz **X** (matriz de dados com dimensão $\mathbf{n} \times \mathbf{p}$, sendo **n** o número de fragmentos cerâmicos e **p** o número de elementos químicos observados); suponha que a dimensão essencial dos dados, a ser usada em alguma comparação, é **k** e que essa dimensão assegure que a variabilidade suficiente dos dados seja explicada na escolha de **k**. A seguir, considera-se **Y** a matriz ($\mathbf{n} \times \mathbf{k}$) dos escores das componentes principais que

[1] Aluno do programa de doutorado em estatística. Depto de Estatística, Instituto de Matemática e Estatística – USP, C. P.: 66281, CEP. 05315-970, São Paulo/SP, e-mail: poliveir@ime.usp.br, Tel. 0 XX (11) 3106.7760.
[2] Pesquisadora. Depto de Estatística, Instituto de Matemática e Estatística – USP, C. P.: 66281, CEP. 05315-970, São Paulo/SP, e-mail: lbarroso@ime.usp.br, Tel. 0 XX (11) 3091-6130.

produz a melhor aproximação **k**-dimensional da configuração dos dados originais **X.** Suponha que **q** das **p** variáveis originais (**q** < **p** e **q** ≥ **k**) sejam suficientes para representar a mesma estrutura apresentada em **Y**. Considera-se $\tilde{X}$ de dimensão **n** × **q**, a matriz de dados com as variáveis selecionadas e $\tilde{Z}$, a matriz de dimensão (**n** × **k**) dos escores das componentes principais dos dados reduzidos que produz a melhor aproximação **k**-dimensional da configuração **q**-dimensional definida no subconjunto dos dados.

Se a verdadeira dimensão dos dados é **k**, então **Y** pode ser vista como a verdadeira configuração, e $\tilde{Z}$ como a correspondente configuração aproximada baseada somente nas **q** variáveis.

O diagrama esquemático abaixo mostra os passos do procedimento.



Para medir a discrepância entre as configurações Y e $\tilde{Z}$ foi utilizada a Análise Procrustes (Sibson, 1978), avaliando o ajuste entre as duas configurações pela soma de quadrados residual (M^2), que mede a perda de informação sobre a estrutura dos dados quando somente as **q** variáveis selecionadas são usadas em vez das **p** variáveis originais.

Sejam as configurações: **Y** matriz de dimensão (**n** × **k**) e $\tilde{Z}$ matriz de dimensão (**n** × **k**), então

$$M^2 = \text{traço}\left\{YY' + \tilde{Z}\tilde{Z}' - 2\tilde{Z}Q'Y'\right\},$$

que pode ser reescrito como

$$M^2 = \text{traço}\left[YY'\right] + \text{traço}\left[\tilde{Z}\tilde{Z}'\right] - 2\text{traço}\left[\Lambda\right],$$

onde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$

$\mathbf{Q} = \mathbf{VU}'$

onde **Q** é matriz ortogonal de dimensão **k** × **k**;

[3] Pesquisador, Instituto de Pesquisas Energéticas e Nucleares – IPEN-CNEN/SP, C. P. 11049 CEP.: 05422-970, São Paulo/SP, e-mail: camunita@net.ipen.br Tel. 0 XX (11) 3816.9289.

U, **Λ** e **V** são obtidos da decomposição por valor singular (Golub, 1970) da matriz $\tilde{Z}'Y$ de dimensão (**k** × **k**), isto é,

$$_{k}\tilde{Z}'Y_{k} = U\Lambda V'.$$

Krzanowski (1996) mostra que, sob determinada suposição, M^2 tem distribuição proporcional a qui-quadrado com $nk - k(k-1)/2$ graus de liberdade, se as variáveis não são estruturadas; a proporção é dada por $\alpha\chi^2$ onde $\alpha = \sigma^2(2p - 2k - i)/(p - k)$ e $\boldsymbol{\sigma^2}$ é a estimativa da variância do resíduo, obtida pela soma de quadrados dos elementos de $U_{p\text{-}i}\, D_{p}i\, V'_{p\text{-}k}$, dividida por (n-k-1) (p-k). As **p-q** variáveis são eliminadas uma a uma, em ordem crescente de importância, e a eliminação é feita enquanto M^2 não excede o valor crítico determinado pela distribuição de referência.

O melhor subconjunto de **q** variáveis é o subconjunto que fornece o menor valor de M^2 entre todos os subconjuntos de **q** variáveis.

Para a eliminação das variáveis pode ser utilizado o seguinte procedimento:

a) Inicialmente considerar **q = p**, e para **k** fixado calcular a matriz dos escores das componentes principais **Y**.

b) Utilizar algoritmos de atualização (Bunch et al., 1978) para obter e armazenar a matriz de escores das componentes principais, excluindo sucessivamente cada variável.

c) Calcular M^2 para cada matriz dos escores e identificar a variável $\mathbf{X_u}$ que fornece o menor M^2. Seja $\tilde{Z}_{(u)}$ a correspondente matriz dos escores.

d) Eliminar a variável $\mathbf{X_u}$. Fazer $Z = \tilde{Z}_{(u)}$ e retornar à etapa b) com **p – 1** variáveis. Continuar este ciclo até que restem somente **q** variáveis.

3 – RESULTADOS E DISCUSSÃO

Neste trabalho, o algoritmo de análise foi elaborado no aplicativo R e empregado nas concentrações de Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Ce, Eu, Tb, Hf e Th determinadas em 75 amostras de fragmentos cerâmicos. A técnica de validação cruzada dos dados sugeriu que a verdadeira dimensionalidade dos dados é dois. Então, o procedimento de seleção com **k** = 2 foi aplicado ao conjunto de dados padronizados. As duas primeiras componentes principais explicaram 57% da variância total. A Tabela 1 mostra o procedimento de seleção de variáveis incluindo a seqüência de eliminação, a distância da bidimensionalidade da configuração das componentes principais usando todos os dados (M^2) e o valor crítico de 5% (stop rule *cv*). Na Tabela 1 a variável Th é o primeiro elemento a ser eliminado, uma vez que o valor para M^2 foi

de 5,5, que mede a proximidade da bidimensionalidade da configuração das componentes principais. Isso mostra que para Th o valor de 226,4 é o valor crítico com nível de significância de 5% e como M^2 é menor que esse valor, então, a eliminação de Th não afeta significativamente a configuração das componentes principais. Uma vez que as variáveis são eliminadas pela distância da configuração das componentes principais, M^2 aumenta e o valor crítico diminui, que por sua vez depende do número de variáveis, que também diminui.

Quando a variável é eliminada, a configuração associada é alterada até o ponto em que chega a ser inaceitável, o que ocorre quando M^2 é maior que o valor crítico. Esse ponto é alcançado quando o valor do M^2 é maior que o *cv.* Este procedimento sugere que Fe, Lu, Yb, Sc, K, Tb e U sejam eliminadas, uma a uma, nessa ordem e que Ce e as outras variáveis restantes devam ser consideradas (Ce, Eu, Hf, Na, La, Cr e Cs).

Para verificar se o comportamento das variáveis selecionadas representam a estrutura completa dos dados, uma análise adicional de componentes principais (Jolliffe, 1972 e 1973) foi realizada usando uma matriz de dados de 75 amostras e os 7 elementos selecionados (Ce, Eu Hf, Na, La, Cr e Cs). O gráfico das duas primeiras componentes baseado em todas as variáveis é apresentado na Figura 1 e baseado nas variáveis selecionadas, na Figura 2.

Comparando-se as duas figuras, confirma-se que a análise de componentes principais usando as sete variáveis produz resultados similares aos de análise de componentes principais usando todas as variáveis.

Muitas variáveis foram eliminadas porque a estrutura dos dados é muito forte e o ajuste dos dados é bom. Isto implica em que o erro residual da variância é baixo.

O baixo valor crítico pode resultar em um baixo erro do desvio padrão, que por sua vez implica em que muitas variáveis são necessárias para serem estatisticamente significantes.

Tabela 1. Resultado do procedimento de eliminação para o conjunto de dados

M^2	5,5	12,9	19,7	32,4	51,3	74,2	99,7	126,5	158,6	216,8	
cv	226,4	215,6	204,9	194,1	183,3	172,5	161,7	150,9	140,2	129,4	
Elemento	Th	Fe	Lu	Yb	Sc	K	Tb	U	Ce	Hf	Na, La, Cr, Cs, Eu

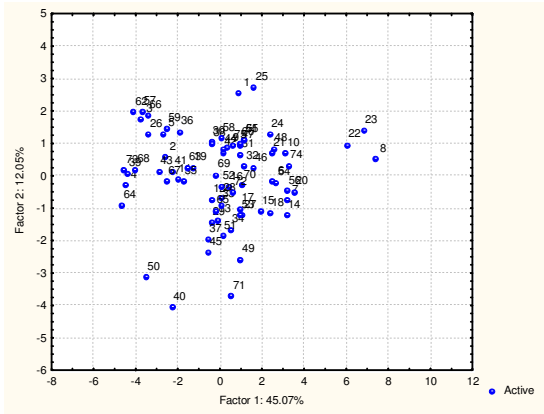


Fig.1. Componente principal 1 vs componente principal 2, considerando todas as variáveis.

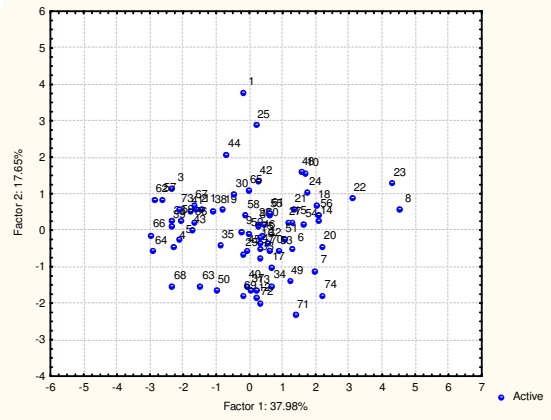


Fig.2. Componente principal 1 vs. componente principal 2, considerando as variáveis Ce, Eu Hf, Na, La, Cr e Cs.

4. CONCLUSÃO

Neste trabalho, foi mostrado, por meio do gráfico das componentes principais, que o subconjunto **q** das variáveis selecionadas (Ce, Eu Hf, Na, La, Cr e Cs) representam o conjunto das variáveis, implicando que ao usar um número menor de variáveis não há grande perda de informação.

5. REFERÊNCIAS BIBLIOGRÁFICAS

1. Bunch, J.R., Nielsen, C.P. and Sorensen, D.C. (1978). Rank one modification of the symmetric eigenproblem. *Numerische Mathematik*, **31**, 31-48.
2. Golub, G.H. and Reinsch, C (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403-420.
3. Jolliffe, I.J. (1972). Discarding variables in principal component analysis. I: artificial data. *Applied Statistics,* **21**, 160-173.
4. Jolliffe, I.J. (1973). Discarding variables in principal component analysis. II: real data. *Applied Statistics*, **22**, 21-31.
5. Krzanowski, W.J. (1996). A stopping rule for structure preserving variable selection. *Statistics and Computing,* **6**, 51-56.
6. Munita, C.S., Paiva, R.P., Alves, M.A., Oliveira, P.M.S. and Momose, E.F. (2000). Contribution of neutron activation analysis to archaeological studies. *J. Trace Microprobe Techniques*, **18** (3), 381-387.
7. Sibson, R. (1978). Studies in the robustness of multidimensional scaling. *Journal of the Royal Statistical Society, B*, **40**, 234-238.