Acta Scientiarum 20(4):505-509, 1998.
ISSN 1415-6814.

Análise *procrustes* aplicada à seleção de variáveis

Terezinha Aparecida Guedes* e Ivan Ludgero Ivanqui

Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá-Paraná, Brazil.
**Author for correspondence. e-mail: taguedes@maringa.com.br*

RESUMO. Nos estudos exploratórios multivariados, cujo objetivo é a redução da dimensão do conjunto de variáveis, o principal método utilizado é a análise de componentes principais. Neste método, todas as variáveis originais são, em geral, necessárias para definir os subconjuntos de variáveis. Krzanowski (1987) apresentou uma metodologia que combina a análise de componentes principais e a análise *procrustes* para determinar o quanto o novo subconjunto de variáveis representa a estrutura dos dados originais. Steiner (1995) utilizou vários métodos para separar grupos e selecionar variáveis em um problema médico. Neste trabalho, foi aplicada a análise *procrustes* para um conjunto de dados gerado aleatoriamente segundo as distribuições das variáveis definidas por Steiner. O objetivo foi verificar se o subconjunto de variáveis resultantes da análise reproduzem a estrutura original dos dados. Através das análises realizadas, concluiu-se que o método *procrustes* é uma ferramenta indispensável na seleção de variáveis.

Palavras-chave: componentes principais, análise *procrustes*, análise multivariada.

ABSTRACT. Procrustes analysis applied to variables selection. In exploratory multivariate research aiming at the reduction of the, dimension of the variables set, the most frequently used method is the analysis of the principal components. All original variables are generally necessary to define the subset of variables. Krzanowski (1987) has provided a methodology which combines the principal component analysis and the procrustes analysis to determine how much the new subset of variables reproduces the structure of original variables. Steiner (1995) used several methods to separate groups and select the variables in a medical case study. In the present work, the procrustes analysis was applied to a set of data randomly generated according to the variables distributions defined by Steiner. The objective was to verify if the subset of variables resultant from the analysis reproduces the original structure of data. The: results led to the conclusion that the procrustes method is a necessary tool for variables selection in multivariate analysis.

Key words: principal components, procrustes analysis, multivariate analysis.

Nos estudos exploratórios, em diversas áreas, os pesquisadores lidam com análise de dados provenientes de um grande conjunto de variáveis. Isso ocorre pela dificuldade na identificação, à priori, das variáveis que são importantes para o estudo. Geralmente, o objetivo inicial do pesquisador é reduzir a dimensão de seu conjunto de variáveis sem que a estrutura dos dados seja alterada.

A principal técnica exploratória utilizada, nesses casos, é a análise de componentes principais. Esta técnica tem sido abordada por Mardia *et al.,* (1992), Johnson e Wichern (1982), Jolliffe (1972, 1973) e outros. A principal deficiência desse método é que, enquanto a dimensão pode ser reduzida, todas as variáveis originais são, em geral, ainda necessárias para definir as novas variáveis. Isso pode ser insatisfatório não só pelo fato de que, em muitas situações, muito tempo e recursos terão de ser despendidos, mas também porque não é fácil interpretar uma combinação linear. A seleção e a identificação de variáveis redundantes é, há muito tempo, assunto de interesse nas áreas de aplicação de regressão múltipla e de análise discriminante. Jolliffe (1989) discutiu alguns métodos com base em análise de componentes principais para selecionar um subconjunto de variáveis as quais preservam muito da estrutura das variáveis originais. Krzanowski (1987) apresentou uma metodologia que combina a análise de componentes principais e a análise *procrustes* para determinar o quanto o novo subconjunto de variáveis representa a estrutura dos dados originais. Ele demonstrou que esta metodologia pode ser uma ferramenta utilizada na caracterização inicial da aplicação de técnicas como

análise discriminante e análise de agrupamento (*cluster analysis*). Segundo Krzanowski, na análise discriminante, procura-se selecionar variáveis com o objetivo de maximizar a probabilidade de uma alocação correta dos futuros indivíduos no grupo, enquanto o método proposto seleciona aquelas variáveis que melhor indicam a existência do grupo. Na análise de agrupamento, a amostra é particionada em grupos, o pesquisador geralmente tem interesse em identificar um pequeno número de variáveis que podem ser usadas para descrever os grupos e distingui-los. Steiner (1995) utilizou os métodos: componentes principais, função discriminante linear de Fisher e o de k-vizinhos mais próximos (k-médias) para separar grupos e selecionar variáveis, em um problema médico do qual foram coletados dados clínicos de 118 pacientes, destes 35 possuíam câncer e 83 cálculo, ambos de vesícula biliar. Foram examinadas 14 variáveis e após esse estudo ficou estabelecida a distribuição de probabilidade para cada uma das variáveis.

Neste trabalho, será aplicada a análise *procrustes* para um conjunto de dados gerado aleatoriamente segundo as distribuições das variáveis definidas por Steiner. O objetivo é verificar se o conjunto de variáveis resultantes de análises, no qual a escolha de variáveis é realizada, reproduz a estrutura original dos dados.

Análise *procrustes*

Suponha que p variáveis $X_1, X_2, ..., X_p$ foram observadas sobre n indivíduos e que o vetor de observações para o i-ésimo indivíduo seja denotado por $X^i = (X_{i1}, X_{i2}, ..., X_{ip})'$, então a matriz dos dados será denotada por $X(nxp)$. Na análise de componentes principais, as variáveis $X_1, X_2, ..., X_p$ são transformadas linearmente em novas variáveis $Y_1, Y_2, ..., Y_p$, denominadas componentes principais. Dessa forma, os dados observados X^i são também transformados em escores dos componentes principais correspondentes, $Y^i = (Y_{i1}, Y_{i2}, ..., Y_{ip})'$. Assim sendo, a matriz original X de ordem nxp é transformada em uma matriz Y de ordem nxp. A redução da dimensão será importante quando o número de componentes Y_i que conservam muito da informação amostral, inerente as variáveis X_i, $i=1, ..., p$, for menor que p. Nesse caso, as observações originais X^i podem ser trocadas pelos primeiros q ($q < p$) elementos dos escores dos componentes principais correspondentes e pode-se escrever $Y^i = (Y_{i1}, Y_{i2}, ..., Y_{iq})'$. Um gráfico de dimensão q dos escores exibirá uma aproximação da configuração original dos dados. A principal deficiência desse método é que, enquanto a dimensão pode ser reduzida de p para q, todas as p variáveis originais são, em geral, ainda necessárias para definir as q novas variáveis Y_i.

Suponha que k variáveis foram selecionadas como sendo as essenciais para a análise através de algum método de seleção, tendo sido levado em consideração se a variabilidade dos dados está sendo suficientemente avaliada nesta escolha.

Seja Y a matriz dos escores das componentes principais obtida das k variáveis selecionadas, que produz a melhor aproximação da configuração dos dados originais.

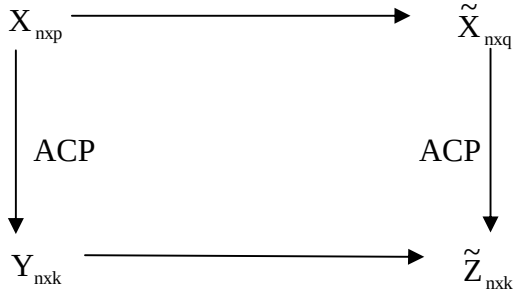
O objetivo é selecionar q das variáveis originais, onde $q < p$ e $p \geq k$ esperando que as variáveis selecionadas representem a verdadeira estrutura dos dados.

Suponha que $\tilde{X}$ denota a matriz de dados que contém somente as k variáveis selecionadas e que $\tilde{Z}$ é a matriz dos escores dos componentes principais do conjunto reduzido de dados. Isto é, a melhor aproximação de dimensão k para a configuração de dimensão q definida pelo subconjunto de dados.

Se a verdadeira dimensão dos dados é de fato k, então Y pode ser vista como a verdadeira configuração, enquanto $\tilde{Z}$ é a aproximação correspondente da configuração baseada em somente q variáveis. Para medir a discrepância entre a verdadeira configuração e a obtida pelo subconjunto de dados, uma análise *procrustes* pode ser realizada.

Para isso é necessário encontrar a soma de quadrados da diferença entre os escores correspondentes às configurações. Esta soma de quadrados residual mede a perda de informação sobre a estrutura de dados quando somente q variáveis são utilizadas ao invés das p variáveis originais.

Krzanowski (1987) apresentou o diagrama abaixo para mostrar os passos do procedimento:



ACP = Análise de componentes principais

O objetivo é detectar a perda de informação causada pela retirada de algumas variáveis.

Y → configuração natural

$\tilde{Z}$ → configuração transformada.

A análise *procrustes* minimiza o traço da matriz de quadrados da diferença entre as duas configurações:

Min traço $\{(Y - \tilde{Z})(Y - \tilde{Z})^T\}$,

sob translação rotação e reflexão de $\tilde{Z}$.

A solução pode ser obtida rotacionando $\tilde{Z}$ para $\tilde{Z}Q$. A matriz de rotação é dada por $Q = UV^T$, onde VDU^T é a decomposição em valores singulares de $Y^T\tilde{Z}$ e então:

$$M^2 = \text{traço}\{YY^T + \tilde{Z}\tilde{Z}^T - 2YQ^T\tilde{Z}^T\}$$
$$= \text{traço}(YY^T) + \text{traço}(\tilde{Z}\tilde{Z}^T) - 2\,\text{traço}(D). \qquad (1)$$

A quantidade dada pela equação acima pode ser calculada facilmente para qualquer subconjunto selecionado e representa a proximidade com que a configuração desse subconjunto selecionado representa a configuração do conjunto original. O melhor subconjunto de variáveis apresentará o menor valor de M^2 entre todos os subconjuntos possíveis. A escolha das variáveis, que irão compor o subconjunto, pode ser realizada através de qualquer método de seleção de variáveis.

Na prática, pode-se utilizar o algoritmo *procrustes* abaixo:

1. Inicialmente, tome $q = p$, e para k fixo, calcule a matriz Y de escores dos componentes principais.
2. Obtenha as matrizes $\tilde{Z}$ de escores dos componentes principais de cada um dos subconjuntos de variáveis selecionadas.
3. Calcule M^2, através da equação (1), para a matriz *Y* e para cada uma das outras matrizes $\tilde{Z}$ de escores e identifique o subconjunto de variáveis que resulta o menor M^2.

Obs. No passo (2) pode-se retirar uma variável de cada vez.
No passo (3) escolher o subconjunto que resultou o menor M^2 e voltar ao passo (2).
Não há regra de parada.

Aplicação. Os dados para a aplicação da análise *procrustes* foram gerados aleatoriamente, através do programa estatístico SAS, segundo as distribuições de probabilidades das variáveis, conforme obtido por Steiner (1995). As variáveis e suas distribuições estão apresentadas na tabela a seguir.

O método de *procrustes* foi utilizado através do programa SAS/IML e o procedimento PRINCOM.

Tabela 1. Variáveis relacionadas ao câncer de fígado e suas distribuições de probabilidades

Variável	Distribuição	Média	desvio-padrão	Teste de aderência KS - p-value
X1 - Sexo	Binomial	p=0,568	σ = 0,495	0,657
X2 - Idade	Normal	μ=57,18	σ = 13,06	0,995
X3 - Albumina	Normal	μ=3,053	σ=0,489	0,903
X4 - Vg	Normal	μ=36,484	σ=6,033	0,844
X5 - Bilirrubina total	Gama	μ=22,060 α= 7,993	σ =7,802 β=0,362	0,885
X6 - Bilirrubina direta	Gama	μ=12,895 α=8,195	σ =4,504 β=0,635	0,993
X7 - Bilirrubina indireta	Gama	μ=9,352 α=4,610	σ =4,355 β=0,493	0,932
X8 - Sgpt	Gama	μ=81,981 α=1,867	σ=4,355 β=0,023	0,680
X9 - Sgot	Gama	μ=97,311 α=3,910	σ=49,21 β=0,040	0,736
X10 - Fosfatose Alcalinas	Gama	μ=383,19 α=2,184	σ=259,27 β=0,006	0,384
X11 - Amilase	Gama	μ=103,00 α=5,589	σ=43,568 β=0,054	0,587
X12 - Tap	Gama	μ=14,228 α=79,619	σ=1,594 β=5,596	0,820
X13 - Creatinina	Gama	μ=0,858 α=10,306	σ=0,267 β=12,009	0,981
X14 - Leucócitos	Gama	μ=9,872 α=9,933	σ=3,132 β=1,006	0,631

Análise do conjunto de dados

A análise conduzida por Steiner (1995) sugere que menos de 14 variáveis são suficientes para identificar a estrutura do conjunto de dados. Pelo método dos k-vizinhos mais próximos, as variáveis selecionadas foram: sexo, bilirrubina indireta, sgpt (transminases gultâmico-pirúvicas), vg e leucócitos. No ajuste logístico as variáveis selecionadas foram: idade, bilurrubina total, bilurrubina direta, amilase, sgpt (transminases gultâmico-pirúvicas), sgot (transminases oxalacéticos do soro), fosfatases alcalinas e vg. Além destes dois conjuntos de variáveis, foram selecionados mais dois conjuntos: um pelo método de componentes principais e outro pelo método *procrustes*.

Inicialmente, foi realizada uma análise de componentes principais com todas as variáveis e a matriz de escores desses componentes foi obtida. Observou-se que os sete primeiros componentes principais acumulam 99,92% da variação total, indicando que esta matriz é uma boa aproximação da matriz de dados original.

Com base nesse resultado e utilizando o método de seleção, através de componentes principais, descrito por Mardia, *et al.,* (1992), foram selecionadas as seguintes variáveis: Idade, bilirrubina total, bilirrubina direta, sgpt, sgot, fosfatose alcalina e amilase.

A seguir, uma nova seleção de variáveis foi realizada utilizando o algoritmo *procrustes*, descrito acima, para um conjunto de 50 observações. Vários subconjuntos foram obtidos, eliminando-se as variáveis uma a uma, e os seus respectivos quadrados

médios do resíduo (M^2) foram calculados. A Tabela 2 abaixo apresenta os subconjuntos correspondentes aos menores valores de quadrado médio do resíduo.

Tabela 2. Subconjunto de variáveis selecionadas e respectivos valores dos quadrados médios dos resíduos (M^2)

N° de variáveis eliminadas	Variáveis selecionadas	M^2
1	X2 até X14	2,64E-9
2	X2 até X12 e X14	4,99E-9
3	X2, X4 até X12 e X14	4,83E-8
4	X2,X4 até X11	4,46E-6
5	X2, X5 até X11 e X14	2,25E-5
6	X2, X5 até X11	6,10E-5
7	X2, X5, X7 até X11	1,16E-4
8	X2, X5, X8 até X11	3,35E-4
9	X2, X8 até X11	4,40E-3

O gráfico dos escores, obtidos através de análise de componentes principais, de cada um dos subconjuntos selecionados foi construído e os que apresentaram configuração semelhante a do conjunto original de dados foram os subconjuntos com 1, 2, 3, 4, 5, 6, 7 variáveis eliminadas. Então, para comparação com os outros conjuntos foi escolhido o conjunto com 7 variáveis eliminadas sendo as variáveis selecionadas: idade, belirrubina total, bilirrubina direta, bilirrubina indireta, sgpt, sgot, fosfatose alcalina e amilase.

No presente trabalho, o objetivo foi investigar o quão bem cada um desses quatro subconjuntos selecionados representa a estrutura do conjunto original de dados. Serão realizadas comparações dos resultados obtidos, para os subconjuntos selecionados, através de componentes principais combinada com o método *procrustes*.

Na seqüência, através de análise de componentes principais, foram obtidas as matrizes de escores para cada um dos quatro subconjuntos de variáveis, obtidos pelos métodos estudados e estas foram, então comparadas com a matriz de escores do conjunto completo, através da análise *procrustes* (Tabela 3).

Foram construídos os gráficos (Figuras 1 a 5) dos dois primeiros componentes principais, para o conjunto de 50 observações, para determinar o quanto os subconjuntos de variáveis capturam a estrutura dos dados originais. Esses gráficos acumulam 93,61%, 93,61%, 94,47%, 93,65% e 93,66%, respectivamente, da variabilidade total dos dados.

Tabela 3. Valores dos quadrados médios dos resíduos (M^2) para cada um dos subconjuntos de variáveis obtidos pelos métodos estudados, segundo o tamanho das amostras e número de variáveis selecionadas

Métodos	Tamanho da amostra			
	20	50	100	200
Componentes Principais (7 variáveis)	9,06E-4	3,10E-4	5,03E-4	4,46E-4
K-vizinhos (5 variáveis)	62,89	235,48	519,40	1123,42
Regressão Logística (8 variáveis)	7,25E-4	2,86E-4	0,76E-4	1,86E-4
Procrustes (8 variáveis)	1,87E-4	6,10E-5	4,73E-4	2,71E-4

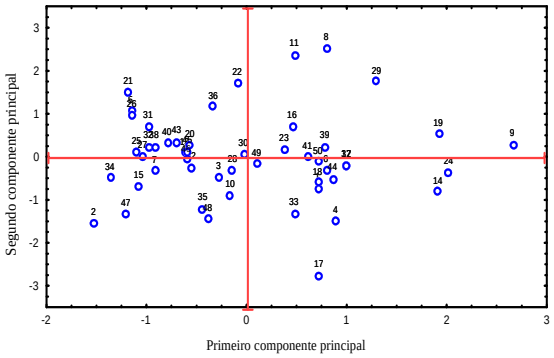


Figura 1. Gráfico das 50 observações geradas segundo as distribuições das variáveis relacionadas ao câncer de fígado para todas as variáveis

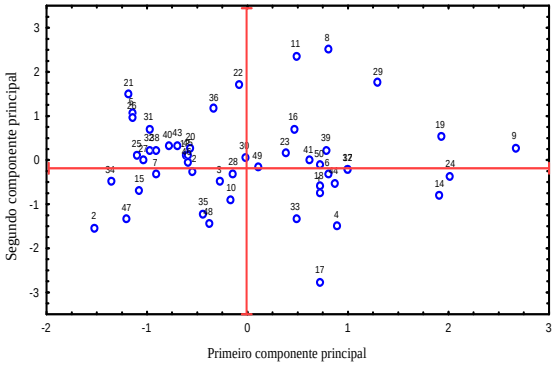


Figura 2. Gráfico das 50 observações geradas segundo as distribuições das variáveis relacionadas ao câncer de fígado para as variáveis selecionadas pelo método de componentes principais

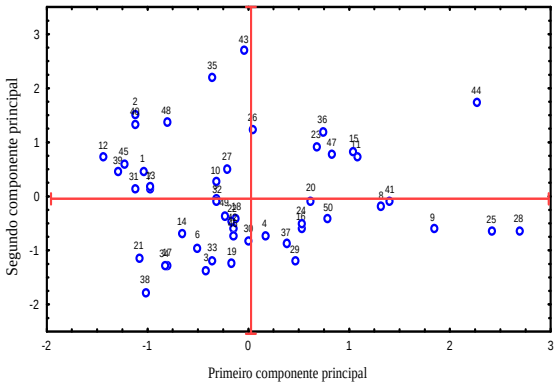


Figura 3. Gráfico das 50 observações geradas segundo as distribuições das variáveis relacionadas ao câncer de fígado para as variáveis selecionadas pelo método dos k-vizinhos

Discussão

As somas de quadrados (M^2) das análises apresentadas na Tabela 3 fornecem uma ordem de importância aos subconjuntos. O subconjunto de variáveis com o menor valor nessa soma e que apresente uma configuração semelhante a do

conjunto original de dados é o que melhor representa o conjunto original.

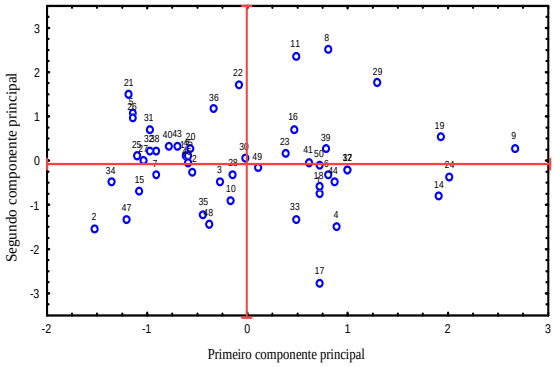


Figura 4. Gráfico das 50 observações geradas segundo as distribuições das variáveis relacionadas ao câncer de fígado para as variáveis selecionadas pelo método da regressão logística

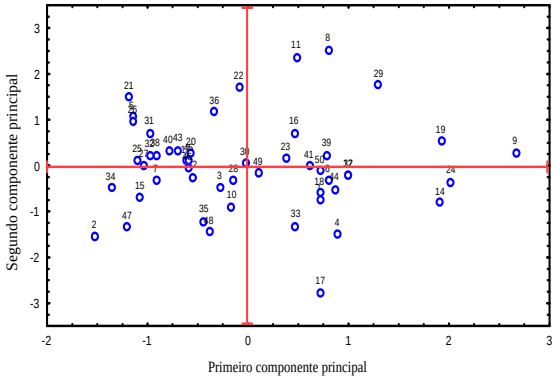


Figura 5. Gráfico das 50 observações geradas segundo as distribuições das variáveis relacionadas ao câncer de fígado para as variáveis selecionadas pelo método *procrustes*

Os subconjuntos de variáveis que apresentaram os mais baixos valores para quadrado médio do resíduo (M^2) foram obtidos pelos métodos de regressão logística e de *procrustes* com 8 variáveis cada um. Sendo que a variável "bilirrubina indireta" aparece no subconjunto selecionado pelo método *procrustes* e "vg" aparece no subconjunto selecionado pelo método de regressão logística.

Da Tabela 3 observa-se que o método de seleção dos k-vizinhos foi o que apresentou o maior valor para o quadrado médio do resíduo (M^2).

Ainda na Tabela 3, pode-se observar que para amostras de tamanho 100 e 200 o menor valor para o quadrado médio do resíduo (M^2) foi o obtido pelo método de regressão logística.

As representações bidimensionais dos escores dos subconjuntos de variáveis (Figuras 1 a 5) revelam quais dos métodos melhor capturam a estrutura do conjunto de dados com todas as variáveis.

Pelo gráfico dos escores do conjunto de dados com todas as variáveis (Figura 1) pode-se observar dois grupos distintos de observações. Tal configuração é mantida pelos gráficos dos escores dos subconjuntos de dados, obtidos pelos métodos de componentes principais, regressão logística e *procrustes*.

Pelo gráfico dos escores do subconjunto de dados obtido pelo método dos k-vizinhos (Figura 3) observa-se que o mesmo não é uma boa aproximação do conjunto original, pois não apresenta a mesma configuração.

Assim sendo, dos quatro métodos de seleção de variáveis estudados, o único método que não capturou adequadamente a estrutura dos dados originais foi o dos k-vizinhos.

Pode-se observar que não foi utilizada nenhuma regra de parada no algoritmo *procrustes,* pois não foi encontrada na literatura nenhuma regra definida.

Com este estudo, ficou evidenciado que se o objetivo é selecionar subconjuntos de variáveis que preservem a estrutura dos dados originais, o método *procrustes* é uma ferramenta auxiliar de grande importância na comparação entre a configuração do conjunto original de dados e o conjunto selecionado; embora para a seleção de variáveis o método *procrustes* não tenha apresentado resultados melhores que os métodos de regressão logística e de componentes principais.

Referências bibliográficas

Johnson, R.A.; Wichern, D. W. *Applied multivariate statistical analysis*. New Jersey : Prentice Hall, 1982.

Jolliffe. I.T. Discarding variables in a principal components analysis. I: Artificial data. *Appl. Statist.*, *21*:160-173, 1972.

Jolliffe. I.T. Discarding variables in a principal components analysis. I: Real data. *Appl. Statist.*, *22*:21-31, 1973.

Jolliffe. I.T. Rotation of III-defined Principal Components. *Appl. Statist.*, *38*:139-147, 1989.

Krzanowski, W.J. Selection of variables to preserve multivariate data structure, using principal components, *Appl. Statist.*, *36*:22-33, 1987.

Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*. New York: Academic Press, 1992.

Steiner, M.T.A. *Reconhecimento de padrão na indústria de papel.* Florianópolis, 1995. (Doctoral Thesis in Operation Research) - Universidade Federal de Santa Catarina.

Received on September 22, 1998.
Accepted on November 09, 1998.