UMA PROPOSTA DE TRANSFORMAÇÃO DE DADOS PARA ANÁLISE DE COMPONENTES PRINCIPAIS

Ana Carolina Mota CAMPANA[1]
José Ivo RIBEIRO JÚNIOR[2]
Moysés NASCIMENTO[2]

- RESUMO: A análise de Componentes Principais (CPs) não é invariante à escala dos dados, isto é, as estimativas dos CPs com base nos dados originais e padronizados não fornecem os mesmos resultados. Deste modo, definir a escala a ser adotada torna-se uma tarefa de grande importância, uma vez que os resultados do estudo estão diretamente relacionados a esta escolha. A transformação dos dados por meio do coeficiente de variação, faz com que todas as variáveis estejam numa mesma escala de medida, porém com variâncias diferentes. Esta transformação contorna os inconvenientes proporcionados pela padronização de impor a mesma média (zero) e variância (um) para todas as variáveis e, dos dados originais, de sofrer os efeitos das escalas de medida.

- PALAVRAS-CHAVE: Análise multivariada; escala de medida; variabilidade; simulação de dados.

1 Introdução

Os métodos de análise multivariada são aplicados quando várias variáveis são medidas em cada elemento amostral, cujo objetivo é analisá-las simultaneamente sem perda significativa das informações contidas em todas elas (MINGOTI, 2007).

Dentre os métodos multivariados utilizados para reduzir a dimensionalidade dos dados, destaca-se o dos componentes principais (CPs), que foi introduzido por Karl Pearson em 1901 e fundamentado no artigo de Hotelling de 1933. Seu principal objetivo é de explicar a estrutura de variâncias e covariâncias de uma matriz composta por p variáveis (JOHNSON e WICHERN, 2002).

Uma propriedade dos CPs é que eles não são invariantes às diferentes escalas, originais ou transformadas, que os dados podem assumir. Na literatura, encontram-se principalmente trabalhos que utilizam a matriz de covariâncias (S) aplicada aos dados originais ou padronizados. Martins et al. (2007) avaliaram a resistência à ferrugem em 68 genótipos de soja, que foram agrupados pela matriz S aplicada aos dados originais de três características. Lédo et al. (2008) utilizaram a matriz S aos dados originais e padronizados de caracteres forrageiros de *Panicum.* Santos et al. (2004) e Nascimento et al. (2009)

[1] Universidade Federal de Viçosa - UFV, Departamento de Economia Aplicada, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *campanaac@gmail.com*
[2] Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *jivo@ufv.br* / *moysesnascim@ufv.br*

utilizaram a análise por CPs baseada na matriz *S* dos dados padronizados para distinguir grupos ecológicos de 37 espécies arbóreas e para a avaliação da adaptabilidade e estabilidade fenotípica de genótipos de alfafa, respectivamente. Em ambos os trabalhos, os autores utilizaram os dados padronizados para corrigir o problema das diferentes escalas de medidas das variáveis em estudo.

A padronização, expressa a variável em termos de desvios-padrão da média, e é apenas uma das diferentes transformações que podem ser feitas sobre o conjunto de dados. A matriz *S* aplicada aos dados padronizados tem o mesmo efeito que a matriz de correlações (*R*) aplicada aos dados originais. Quando se usa a matriz S aos dados originais, as variâncias das variáveis são influenciadas por suas escalas. Já a utilização da matriz *R* proporciona a mesma variância para todas elas. Porém, não há orientações claras sobre a escolha mais apropriada das matrizes a serem utilizadas nos diferentes estudos multivariados. O que se observa é a utilização da padronização sempre que há diferenças nas unidades de medida das variáveis utilizadas na análise.

Este problema foi ilustrado no trabalho de Naik e Kathree (1996), que questionaram a classificação final encontrada por Dawkins (1989). Pelo fato de existirem discrepâncias entre as variâncias das variáveis analisadas, Dawkins (1989) utilizou a matriz de correlações (*R*). Por outro lado, Naik e Kathree (1996) usaram uma transformação nos dados originais para a estabilização da variância, diferente daquela realizada por Dawkins (1989) e encontraram um novo resultado. Consequentemente, um deles estará associado a interpretações e conclusões erradas sobre o estudo realizado.

Devido ao fato das matrizes *S* e *R* aplicadas aos dados originais, influenciarem diferentemente as estimativas dos CPs, este trabalho teve como objetivo propor uma nova transformação aos dados, de modo que todas as variáveis tenham a mesma média, mas que não percam suas diferenças de variâncias. Esta transformação visa contornar as limitações inerentes às matrizes rotineiramente utilizadas e, consequentemente, melhorar a qualidade da classificação dos elementos amostrais.

2 Material e Métodos

A transformação proposta, para a análise por CPs, baseia-se no coeficiente de variação (CV) das variáveis em estudo. A escolha do CV se deve ao fato deste ser uma medida mais apropriada da variabilidade relativa dos dados, quando as variáveis estão sob diferentes escalas de medida. Assim, definiu-se a seguinte transformação:

$$z_{ij}^{*} = z_{ij} \times CV_j \text{, para } i = 1, 2,..., n \text{ observações e } j = 1, 2,..., p \text{ variáveis} \quad (1)$$

em que:

z_{ij}^{*}: é o valor da *i*-ésima observação da variável X_j transformada;

$z_{ij} = \dfrac{x_{ij} - \mu_j}{\sigma_j}$ é o valor da *i*-ésima observação da variável X_j padronizada; (2)

x_{ij} : é o valor da i-ésima observação da variável X_j;

μ_j : é a média da variável X_j; σ_j : é o desvio-padrão da variável X_j;

$$CV_j = \frac{\sigma_j}{\mu_j} \text{ é o coeficiente de variação da variável } X_j; \quad (3)$$

Substituindo (2) e (3) em (1), pode-se reescrever a transformação proposta como:

$$z_{ij}^{*} = \frac{x_{ij} - \mu_j}{\mu_j} \ . \qquad (4)$$

Por (2), tem-se que: $E(z_{ij}) = 0$ e $V(z_{ij}) = 1$ e, portanto, $E(z_{ij}^{*}) = 0$ e $V(z_{ij}^{*}) = CV_j^2$. Portanto, as variáveis transformadas passam a estar numa mesma escala de medida, com médias iguais a zero, porém com variâncias diferentes de um e proporcionais aos seus respectivos CVs.

Com o objetivo de avaliar a transformação proposta e comparar os seus efeitos e particularidades com as análises em que utilizam as matrizes *S* e *R* aos dados originais, foram simulados diferentes conjuntos de dados com 200 observações de dua variáveis sob diferentes estruturas de correlações. Para tanto, utilizou-se a distribuição normal com os parâmetros descritos na Tabela 1.

Tabela 1 - Parâmetros de média, variância e coeficiente de variação das variáveis X_1, X_2 simuladas sob diferentes casos

Caso	Parâmetro	X_1	X_2
C1	Média	100	100
	Variância	100	100
	CV (%)	10	10
C2	Média	300	100
	Variância	900	100
	CV (%)	10	10
C3	Média	10	100
	Variância	36	100
	CV (%)	60	10
C4	Média	100	100
	Variância	36	100
	CV (%)	6	10
C5	Média	10	100
	Variância	100	100
	CV (%)	100	10

Em cada caso, foram simulados valores para as variáveis X_1 e X_2 considerando três diferentes situações:

a) X_1 e X_2 não correlacionadas ($\rho_{12} = 0$);

b) X_1 e X_2 moderadamente correlacionadas ($\rho_{12} = 0{,}5$);

c) X_1 e X_2 totalmente correlacionadas ($\rho_{12} = 1$).

Formou-se assim um universo de 15 conjuntos de dados para a aplicação dos CPs por meio das três estratégias de análises.

No caso C1, as variáveis X_1 e X_2 foram simuladas com a mesma média e variância e, consequentemente, com o mesmo CV. Este caso serviu de testemunha para o efeito da escala e possibilitou estudar, de forma pura, o efeito da correlação sobre as estimativas

dos CPs. No C2, buscou-se verificar o efeito da escala causadora de média e variância altas em X_1, mas com a mesma variação relativa (CV) de X_2. Em C3, a variável X_1 de menores média e variância foi aquela que teve a maior variação relativa (CV). No caso C4, a variável X_1 apresentou média igual, porém variância e CV menores do que X_2. Já em C5, a média de X_1 foi menor, mas a variância a mesma de X_2. Consquentemente, o CV da primeira foi maior.

A utilização de apenas duas variáveis para o estudo da matriz *S* aplicada aos dados originais, padronizados e transformados, se deve à maior facilidade de entendimento dos resultados. Porém, o caso bi-dimensional, como caso particular, pode ser extendido para o caso multidimensional sem perda de generalização.

A simulação foi realizada através do software *R* versão 2.9.2 (R Development Core Team, 2007).

Dado os 200 elementos amostrais medidos por duas variáveis (X_1 e X_2), o propósito principal foi estimar as duas novas variáveis (CP_1, CP_2), tal que CP_j (j = 1, 2) é a combinação linear das duas variáveis *Xs*.

Os dois *CPs* estimados a partir da matriz *S* das variáveis *Xs* originais são iguais a:

$$\hat{C}P_1 = \hat{a}_{11}^* X_1 + \hat{a}_{12}^* X_2$$

$$\hat{C}P_2 = \hat{a}_{21}^* X_1 + \hat{a}_{22}^* X_2$$

em que: $\hat{a}_j^* = [\hat{a}_{1j}^* \quad \hat{a}_{2j}^*]$ é a estimativa do autovetor normalizado do CP_j, para j =1, 2.

Geralmente, quando as variáveis estudadas apresentam diferentes escalas de medida, utiliza-se a padronização (2) para que as mesmas fiquem com médias e variâncias iguais a zero e um, respectivamente.

Assim, os dois CPs estimados a partir da matriz *S* das variáveis *Zs* padronizadas ou a partir da matriz *R* das variáveis *Xs* originais são iguais a:

$$\hat{C}P_1 = \hat{a}_{11}^* Z_1 + \hat{a}_{12}^* Z_2$$

$$\hat{C}P_2 = \hat{a}_{21}^* Z_1 + \hat{a}_{22}^* Z_2$$

No caso da transformação proposta, os dois CPs estimados a partir da matriz *S* das variáveis *Z*s* transformadas, denominada de matriz S^*, são iguais a:

$$\hat{C}P_1 = \hat{a}_{11}^* Z_1^* + \hat{a}_{12}^* Z_2^*$$

$$\hat{C}P_2 = \hat{a}_{21}^* Z_1^* + \hat{a}_{22}^* Z_2^*$$

Para os dois *CPs*, foram estimados dois autovalores ($\hat{\lambda}_j$) por meio do determinante da expressão baseada nas matrizes *S, R* e S^*:

$|S - \hat{\lambda}_j I| = 0$, em que: $S = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix} = \begin{bmatrix} s_1^2 & r_{12}s_1s_2 \\ r_{12}s_1s_2 & s_2^2 \end{bmatrix}$; I = matriz identidade de ordem 2; s_1^2 = variância amostral da variável X_1; s_2^2 = variância amostral da variável X_2; s_{12} = covariância amostral entre as variáveis X_1 e X_2; e r_{12} = correlação amostral entre as variáveis X_1 e X_2.

$|R - \hat{\lambda}_j I| = 0$, em que: $R = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$; I = matriz identidade de ordem 2; r_{12} =

coeficiente de correlação amostral entre as variáveis X_1 e X_2.

$|S^* - \hat{\lambda}_j I| = 0$, em que: $S^* = \begin{bmatrix} s_1^{*2} & s_{12}^* \\ s_{12}^* & s_2^{*2} \end{bmatrix} = \begin{bmatrix} CV_1^2 & r_{12}CV_1CV_2 \\ r_{12}CV_1CV_2 & CV_2^2 \end{bmatrix}$; I = matriz identidade de ordem 2; CV_1^2 = coeficiente de variação amostral ao quadrado da variável X_1; CV_2^2 = coeficiente de variação amostral ao quadrado da variável X_2; s_{12}^* = covariância amostral entre as variáveis Z_1^* e Z_2^*.

Para as estimativas dos CPs baseadas em S, R e S^*, têm-se:

$\hat{V}(\hat{CP}_j) = \hat{\lambda}_j$ (autovalor de ordem j da matriz S, R e S^*); em que: $\hat{V}(\hat{CP}_1) \geq \hat{V}(\hat{CP}_2)$ e $Cov(\hat{CP}_1, \hat{CP}_2) = 0$.

Portanto, o CP_1 contém mais informações sobre os dados do que o CP_2, que não contém informações do CP_1.

Para cada autovalor estimado $\hat{\lambda}_j$ (j = 1, 2) tem-se um autovetor não normalizado estimado $\hat{a}_j$, a partir de uma das soluções dos sistemas de equações dados a seguir:

$$[S - \hat{\lambda}_j I]\hat{a}_j = \phi;$$

$$[R - \hat{\lambda}_j I]\hat{a}_j = \phi,$$

$$[S^* - \hat{\lambda}_j I]\hat{a}_j = \phi;$$

em que: $\hat{a}_j = [\hat{a}_{1j} \quad \hat{a}_{2j}]$ = estimativa do autovetor $\hat{a}_j$ não normalizado do CP_j; ϕ = vetor nulo de dimensão 2x1.

A estimativa do autovetor normalizado $\hat{a}_j^*$ é dada por:

$$\hat{a}_j^* = \begin{bmatrix} \hat{a}_{1j}^* \\ \hat{a}_{2j}^* \end{bmatrix} = \frac{1}{\sqrt{\hat{a}_{1j}^2 + \hat{a}_{2j}^2}} \begin{bmatrix} \hat{a}_{1j} \\ \hat{a}_{2j} \end{bmatrix}, \text{ em que } \hat{a}_{1j}^{*2} + \hat{a}_{2j}^{*2} = 1.$$

3.1 Três variáveis

Para verificar de forma mais clara o efeito da escala de medida sobre as estimativas dos CPs, foram simuladas três variáveis, com correlação um, com os parâmetros apresentados na Tabela 2.

Tabela 2 - Parâmetros de média, variância e coeficiente de variação de três variáveis (X_1, X_2 e X_3) para o Caso 6 (C6).

Parâmetro	X_1	X_2	X_3
Média	1000	1	1
Variância	400	0,0004	0,14
CV (%)	2	2	37,4

Neste caso, as variáveis X_1 e X_2 são idênticas, exceto pela unidade de medida. Por outro lado, a variável X_3 está na mesma escala de X_2, diferindo da mesma apenas por

possuir maior variância. Desta forma, retirando o efeito da escala, X_1 e X_2 possuem a mesma variabilidade, menor que a de X_3, o que pode ser visto também pelo coeficiente de variação (Tabela 2).

A correlação unitária ($\rho_{ij} = 1,\ i, j = 1,2,3$) foi utilizada na simulação com o objetivo de se obter apenas um CP, suficiente para explicar a totalidade da variação dos dados, facilitando assim a identificação da variável mais importante na análise.

3 Resultados e discussão

Na Tabela 3 são apresentadas algumas estatísticas descritivas das variáveis simuladas (X_1 e X_2). Observa-se que o processo de simulação das variáveis foi eficaz, uma vez que as estimativas estão bastante próximas dos valores paramétricos especificados na Tabela 1.

Tabela 3 - Estimativas das médias, variâncias, coeficientes de variação e de correlação baseadas em 200 observações de duas variáveis (X_1 e X_2) simuladas nos diferentes casos estudados

Caso	Estatística	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
		X_1	X_2	X_1	X_2	X_1	X_2
C1	Média	99,04	100,35	99,77	100,09	100,17	100,17
	Variância	88,90	78,44	85,21	96,83	98,73	98,73
	CV (%)	9,52	8,83	9,25	9,83	9,92	9,92
C2	Média	297,13	100,35	299,32	100,09	300,51	100,17
	Variância	800,14	78,44	766,90	96,83	888,60	98,73
	CV (%)	9,52	8,83	9,25	9,83	9,92	9,92
C3	Média	9,43	100,35	9,86	100,09	10,10	100,17
	Variância	32,01	78,44	30,68	96,83	35,54	98,73
	CV (%)	60,02	8,83	56,15	9,83	59,02	9,92
C4	Média	99,43	100,35	99,86	100,09	100,10	100,17
	Variância	32,01	78,44	30,68	96,83	35,54	98,73
	CV (%)	5,69	8,83	5,55	9,83	5,96	9,92
C5	Média	9,04	100,35	9,77	100,09	10,17	100,17
	Variância	88,90	78,44	85,21	96,83	98,73	98,73
	CV (%)	104,28	8,83	94,44	9,83	97,71	9,92

3.1 Matriz S

Na Tabela 4 são apresentadas as estimativas dos autovalores dos *CPs* obtidas a partir dos dados originais pela decomposição da matriz de covariâncias (*S*) para as duas variáveis X_1 e X_2 avaliadas nos diferentes casos estudados.

Para $r_{12} \cong 0$, as estimativas dos autovalores ($\hat{\lambda}_1$ e $\hat{\lambda}_2$) foram aproximadamente iguais à maior e menor variâncias das variáveis X_1 e X_2, respectivamente (Tabelas 3 e 4). Isso mostra que o grau de explicação de cada componente foi diretamente proporcional à quantidade da variância de cada variável, independentemente se ela esteve ou não relacionada à escala.

Tabela 4 - Estimativas dos autovalores obtidas pela matriz *S* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$
C1	88,96	78,39	53,20	137,82	44,22	75,70	197,47	0	100
C2	800,14	78,44	91,10	794,71	69,02	92,00	987,35	0	100
C3	78,45	32,00	71,00	107,00	20,50	83,90	134,28	0	100
C4	78,45	32,00	71,00	107,00	20,50	83,90	134,28	0	100
C5	88,95	78,39	53,20	137,82	44,20	75,70	197,47	0	100

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Essa situação pode ser observada no caso C3, onde a maior variância de X_2 (78,44) não foi relacionada à maior variabilidade relativa (CV=8,83%). Portanto, simplesmente uma mudança na escala de X_1 trará alterações significativas nas estimativas dos autovalores obtidos em uma outra estratégia de análise de *CPs.*

Pode-se observar também para $r_{12} \cong 0$ (Tabelas 3 e 4), que a soma das estimativas dos autovalores ($\hat{\lambda}_1 + \hat{\lambda}_2$) foi igual à soma das estimativas das variâncias ($s_1^2 + s_2^2$).

Percebe-se nos casos em que se utilizou a matriz S aos dados originais (não transformados), que a variável com maior variância foi a de maior importância no primeiro *CP*, ou seja, aquele que explicou a maior parte da variação total. Essa influência poderá ser drástica, quando houver uma discrepância muito acentuada entre as variâncias das diferentes variáveis. Isto pode ser observado em todos os casos de $r_{12} \cong 0$, dado que a estimativa de λ_1 foi muito parecida com a da maior variância (Tabelas 3 e 4).

O aumento da correlação, em módulo, aumentou a importância do primeiro *CP*, até 100%, para $r_{12} = 1$. No entanto, quando a diferença entre as variâncias foi grande (C2), esse aumento foi desprezível (Tabela 5).

Quando se utilizou a matriz *S*, a magnitude dos autovalores dos *CPs* esteve diretamente relacionada à magnitude das variâncias e covariâncias entre as variáveis estudadas. Portanto, as estimativas dos autovetores normalizados (Tabela 5), que ponderam os valores das variáveis X_1 e X_2 nos respectivos *CPs*, também foram proporcionais às estimativas dos respectivos autovalores.

A existência de variâncias discrepantes entre as variáveis, devido à escala de medida ou à própria variação, proporcionou maior importância relativa ao primeiro autovalor, o que será um problema quando esta variância for influenciada pela escala e não acarretar aumento da variabilidade relativa. Este fato pode ser observado nos casos C3 e C5, onde a variável X_2, de maior importância nos *CPs* por apresentar a maior variância, foi a de menor variação relativa, ou seja, de menor CV (Tabelas 3 e 5).

Já o aumento da correlação, em módulo, aumentou a importância da variável com menor variância. No entanto, quando a diferença entre as variâncias foi grande (C2), esse aumento foi menor (Tabela 5).

Tabela 5 - Estimativas dos autovetores normalizados, associados ao primeiro autovalor, obtidas pela matriz *S* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$
C1	0,998	0,069	0,662	0,750	0,707	0,707
C2	1,000	0,003	0,981	0,196	0,949	0,316
C3	0,009	1,000	0,343	0,939	0,514	0,857
C4	0,009	1,000	0,343	0,939	0,514	0,857
C5	0,998	0,069	0,662	0,750	0,707	0,707

3.2 Matriz R

Na Tabela 6 são apresentadas as estimativas dos autovalores dos *CPs* obtidas a partir dos dados originais pela decomposição da matriz de correlações (*R)*, para as duas variáveis X_1 e X_2 avaliadas nos diferentes casos estudados.

Tabela 6 - Estimativas dos autovalores obtidas pela matriz *R* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$
C1	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C2	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C3	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C4	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C5	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Através da Tabela 6 pode-se observar que todos os cinco casos apresentaram o mesmo resultado, para todos os três coeficientes de correlação, separadamente. Portanto, a média e a variância de cada variável original não interferiram nas estimativas dos CPs.

Observou-se que a análise dos *CPs* utilizando os dados padronizados, resolveu o problema da discrepância na escala de medida das variáveis. Porém, todas passaram a ser igualmente importantes (Tabela 7). E se o objetivo do estudo for o de encontrar as variáveis que mais discriminam os elementos amostrais, essa estratégia não ajudará.

Além disso, é importante ressaltar que os coeficientes dos *CPs* obtidos pela matriz *S* não foram numericamente iguais aos da matriz *R*. Em geral, o percentual de variação explicado pela matriz *S* foi mais concentrado no primeiro *CP*. Pela *R* a concentração foi menor. Portanto, de acordo com a última matriz, houve melhor distribuição da variabilidade e desta forma, será necessário um maior número de componentes para explicar a mesma quantidade da variância total.

Tabela 7 - Estimativas dos auto-vetores normalizados, associados ao primeiro autovalor, obtidas pela matriz *R* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$
C1	0,707	0,707	0,707	0,707	0,707	0,707
C2	0,707	0,707	0,707	0,707	0,707	0,707
C3	0,707	0,707	0,707	0,707	0,707	0,707
C4	0,707	0,707	0,707	0,707	0,707	0,707
C5	0,707	0,707	0,707	0,707	0,707	0,707

A matriz *S* deu maior importância às variáveis com maiores variâncias, o que pode levar ao domínio de uma variável no primeiro CP. Já a matriz *R* deu importâncias iguais a todas as variáveis, independentemente da escala.

3.3 Matriz S^*

Na Tabela 8 são apresentadas as estimativas dos autovalores dos CPs obtidas a partir dos dados transformados (Z^*) pela decomposição da matriz S^*, para as duas variáveis X_1 e X_2, nos diferentes casos estudados.

Tabela 8 - Estimativas dos autovalores obtidas pela matriz S^* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$
C1	0,008	0,006	56,90	0,013	0,004	76,10	0,019	0	100
C2	0,008	0,006	56,90	0,013	0,004	76,10	0,019	0	100
C3	0,320	0,006	98,10	0,271	0,007	97,50	0,354	0	100
C4	0,006	0,003	68,00	0,010	0,002	85,20	0,013	0	100
C5	0,967	0,006	99,40	0,763	0,007	99,10	0,952	0	100

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Observou-se no caso C2, para $r_{12} \cong 0$, que foram necessários dois componentes para explicarem a variação dos dados, uma vez que o primeiro CP explicou somente 56,90% desta variação (Tabela 8). Este resultado foi diferente do observado quando se utilizaram os dados originais decompostos pela matriz *S*, onde somente o primeiro CP explicou mais de 91% da variação total. Nesse último caso, a variável mais importante foi a X_1, cuja variância foi muito superior à de X_2. Porém, esta variância foi devida somente à escala e não a uma variação real, dado que ambas apresentaram CVs semelhantes (Tabela 3). Portanto, o grau de explicação do CP_1 estimado pela matriz S^*, foi mais próximo daquele obtido pela aplicação da matriz *R* aos dados originais, que considerou a mesma variabilidade para as duas variáveis padronizadas. Isso implicou que a transformação das

variáveis em função das suas variabilidades relativas medidas pelos CVs, foi eficiente em diminuir a importância da variável que apresenta alta variância em função da escala de medição.

Os resultados apresentados para o caso C4 foram similares, quando aplicados às matrizes S e S^*. Este resultado era esperado, pois ambas as variáveis estavam numa mesma escala de medida e, portanto, a maior variância associada a X_2 não foi proveniente da diferença entre as mesmas (Tabelas 4 e 8).

Quando $r_{12} = 1$, ou seja, quando as variáveis foram totalmente correlacionadas, foi necessário apenas um componente para explicar toda a variação contida nos dados, sendo que o coeficiente de maior grandeza nestes componentes esteve associado à variável de maior variação relativa (CV), que foi considerada a mais importante na discriminação dos elementos amostrais (Tabelas 8 e 9).

Tabela 9 - Estimativas dos autovetores normalizados, associados ao primeiro autovalor, obtidas pela matriz S^* nos diferentes casos estudados

Caso	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$
C1	1,000	0,031	0,617	0,787	0,707	0,707
C2	1,000	0,031	0,617	0,787	0,707	0,707
C3	1,000	0,001	0,995	0,097	0,986	0,166
C4	0,011	1,000	0,317	0,949	0,515	0,857
C5	1,000	0,001	0,998	0,057	0,995	0,101

Nos casos C3 e C5, o CP_1 explicou quase que totalmente a variação total, sendo este componente dominado pelas variáveis X_1 no C3 e X_2 no C5, que possuíram os maiores coeficientes de variação (Tabelas 3 e 9). Este resultado foi bastante interessante quando comparado aos casos C3 e C5 da análise baseada na matriz S, onde as variáveis X_2 e X_1 foram consideradas as mais importantes em função das suas maiores variâncias, respectivamente.

Para todos os casos estudados, as variáveis de maior variabilidade relativa (CV), foram aquelas que apresentaram os maiores coeficientes no primeiro CP, ou seja, aquele que explicou a maior parte da variabilidade contida nos dados.

Os autovetores normalizados estimados a partir da matriz S baseada nos dados originais (Tabela 5) foram diferentes daqueles obtidos a partir da matriz S^* baseada nos dados transformados (Tabela 8). Como pode-se observar, um coeficiente é alto e o outro baixo, em valores absolutos. A diferença é que o coeficiente alto pondera as variáveis de maiores variância e CV, respectivamente. Mais especificamente, verifica-se na Tabela 2 que X_1 tem menor variância que X_2, o que faz com que X_2 tenha maior estimativa absoluta de autovetor em S (Tabela 5). No entanto X_1 tem maior variância relativa (Tabela 2), o que faz com que X_1 tenha maior estimativa absoluta de autovetor em S^* (Tabela 9).

Considerando-se que a variância relativa é mais importante que a variância absoluta (σ^2), a escolha da matriz S^* mostra-se mais apropriada que a matriz S. Outra diferença foi o grau de explicação de cada autovetor, medido pela magnitude do seu respectivo autovalor. A transformação dos dados não proporcionou estimativas dos coeficientes mais

equilibradas, como aquelas oriundas da matriz *R* (Tabela 7). Assim, conclui-se que qualquer diferença de variabilidade, por menor que seja, entre as variáveis, promoverá uma desproporcionalidade, maior ou menor, entre os valores absolutos dos coeficientes normalizados dos autovetores.

3.4 Três variáveis

A Tabela 10 apresenta as estatísticas descritivas das três variáveis simuladas para o caso 6. Como já verificado nos casos C1 a C5, espera-se que as duas primeiras variáveis (X_1 e X_2) sejam semelhantes na discriminação dos elementos amostrais, dado que ambas são idênticas, exceto pela unidade de medida. Por outro lado, espera-se que X_3, a de maior variância relativa, CV=34,8%, seja mais importante na discriminação dos mesmos elementos.

Tabela 10 - Estimativas das médias, variâncias e coeficientes de correlação baseadas em 200 observações das variáveis (X_1, X_2 *e* X_3) simuladas para o Caso 6

Caso	Estatística	$r_{ij} = 0{,}98$		
		X_1	X_2	X_3
C6	Média	1000,27	1,00	1,00
	Variância	384,7725	0,0004	0,1213
	CV (%)	1,96	1,93	34,83

As estimativas dos autovalores dos *CPs* obtidas pela decomposição das matrizes de covariâncias (S e S^*) e de correlação (*R*) para as três variáveis X_1, X_2 e X_3 do caso 6 são apresentadas na Tabela 11.

Tabela 11 - Estimativas dos autovalores obtidas pelas matrizes *S, R* e S^* para o Caso 6

Matriz	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$IR_{\lambda_1}(\%)$
R	2,969327	0,022987	0,007686	98,98
S	384,888730	0,005496	0,000004	99,99
S^*	0,114235	0,000021	0,000003	99,98

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Como esperado, dada a alta correlação entre as variáveis simuladas, tem-se que apenas o primeiro componente é suficiente para explicar quase 100% da variabilidade total dos dados (Tabela 11).

Considerando as estimativas dos autovetores normalizados associados ao primeiro autovalor (Tabela 12), tem-se que, quando da utilização da matriz *R*, todas as variáveis são igualmente importantes em explicar a variabilidade total dos dados. Enquanto que, no caso da matriz *S*, a variável X_1 apresenta-se como a mais importante na discriminação dos elementos amostrais. Porém, como já foi dito anteriormente, esta variável sofre o efeito da escala, o que não é levado em consideração pela análise de CPs. Este inconveniente é

corrigido com a utilização da matriz S^*, uma vez que as variáveis X_1 e X_2 são consideradas igualmente importantes na análise e a variável X_3, de maior variabilidade relativa, foi a de maior correlação com o primeiro CP, isto é, aquele que mais explica a variabilidade total dos dados e, portanto, é a mais importante em discriminar os elementos amostrais.

Tabela 12 - Estimativas dos autovetores normalizados, associados ao primeiro autovalor, obtidas pelas matrizes *S, R* e S^* para C6

Matriz	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{13}$
R	-0,9933	-0,9974	-0,9939
S	-19,6156	-0,0190	-0,3403
S^*	0,0188	0,0184	0,3370

Os resultados para o caso com três variáveis (C6) confirma aquele encontrado nos casos bi-dimensionais (C1 a C5), isto é, a análise por CPs dá importâncias iguais para todas as variáveis se a decomposição for da matriz *R*, maior importância para a variável de maior variância se utilizada a matriz *S* e maior importância as variáveis de maior variância relativa no caso da transformação proposta por meio do *CV*.

Conclusões

- A análise de componentes principais baseada na matriz de covariâncias dos dados originais leva em consideração as variâncias das variáveis. Desta forma, as estimativas dos componentes principais podem ser prejudicadas quando essas variâncias forem inerentes à escala.
- Quando se utiliza a matriz de correlações às variáveis originais, as suas variâncias não são importantes para as estimativas dos componentes e, portanto, para a classificação dos elementos amostrais. Desta forma, apesar de contornar o problema da escala dos dados, ela faz com que todas as variáveis tenham a mesma variância, o que pode não ser útil quando se pretende identificar aquelas variáveis com maior grau de discriminação.
- A análise a partir das variáveis transformadas utilizando o coeficiente de variação proporciona a formação dos componentes principais com base nas variáveis de maior variabilidade relativa. Esta transformação contorna as limitações das análises baseadas nas matrizes de covariâncias e de correlações das variáveis originais, dado que as mesmas passam a estar numa mesma escala, porém com variâncias diferentes.
- A transformação proposta deve ser utilizada quando o pesquisador quer corrigir as diferenças entre as escalas de medidas das variáveis e dar importância relativa à variabilidade das variáveis estudadas.

Agradecimentos

À CAPES pelo apoio financeiro.

CAMPANA, A. C. M.; RIBEIRO JÚNIOR., J. I.; NASCIMENTO, M. A proposal of data transformation for principal component analysis. *Rev. Bras. Biom.*, São Paulo, v.28, n.2, p.103-115, 2010.

- *ABSTRACT: The principal component (PC) analysis is not independent of the data scale, ie results of standardized and data-based PC estimates differ. The choice of the scale to be adopted is therefore a task of central importance, directly related to the study results. The data transformation through the coefficient of variation unites all variables on the same measurement scale, but with different variances. This transformation circumvents the disadvantages caused by the standardization, where the same mean (zero) and variance (one) are applied to all variables, and avoids the effects of the measurement scales on the original data.*
- *KEYWORDS: Multivariate analysis; measurement scale; variability; data simulation.*

Referências

DAWKINS, B. Multivariate analysis of national track records. *Am. Stat.*, Hayward, v.43, p.110-115, 1989.

JOHNSON, R. A.; WICHERN, D.W. *Applied multivariate statistical analysis***.** 5. ed. New Jersey: Prentice Hall, 2002. 767p.

LÉDO, F. J. S ET al. Estimativas de repetibilidade para caracteres forrageiros em *Panicum maximum*. *Ciênc. Agrotec***.**, Lavras, v.32, n.4, p.1299-1303, 2008

MARTINS, J. A. S. et al. Período latente e uso da análise de componentes principais para caracterizar a resistência parcial à ferrugem da soja. *Summa Phytopathol.*, Botucatu, v. 33, n.4, p.364-371, 2007

MINGOTI, S.A. *Análise de dados através de métodos de estatística multivariada***:** uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2007. 297p.

NAIK, D.N.; KATTREE, R. Revisiting Olympic track records: some practical considerations in the principal component analysis. *Am. Stat.*, Hayward, v.50, n.2, p.140-144, 1996.

NASCIMENTO, M. et al. Alteração no método centroide de avaliação da adaptabilidade genotípica. *Pesqui. Agropec. Bras.*, Brasília, v.44, n.3, p.263-269, 2009.

SANTOS, J. H. S. et al. Distinção de grupos ecológicos de espécies florestais por meio de técnicas multivariadas. *Rev. Árvore*, Viçosa-MG, v.28, n.3, p. 387-396, 2004.

R DEVELOPMENT CORE TEAM (2007). *R: A language and environment for statistical computing*. *R Fundation for Statistical Computing*. Vienna, Austria. Disponível em: <http://r-project .org>.

Recebido em 12.01.2010.

Aprovado após revisão 11.05.2010.