UFRRJ UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PÓS-GRADUAÇÃO EM AGRONOMIA – CPGA-CS

Análise Multivariada Aplicada as Ciências Agrárias

Análise de Componentes Principais

Carlos Alberto Alves Varella

Seropédica - RJ
11/12/2008

Conteúdo

ANÁLISE DE COMPONENTES PRINCIPAIS

Carlos Alberto Alves Varella[1]

Introdução

A análise de componentes principais é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais. Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados. A análise de componentes principais é associada à idéia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados. Esta técnica pode ser utilizada para geração de índices e agrupamento de indivíduos. A análise agrupa os indivíduos de acordo com sua variação, isto é, os indivíduos são agrupados segundo suas variâncias, ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de características que define o indivíduo, ou seja, a técnica agrupa os indivíduos de uma população segundo a variação de suas características. Segundo REGAZZI (2000), apesar das técnicas de análise multivariada terem sido desenvolvidas para resolver problemas específicos, principalmente de Biologia e Psicologia, podem ser também utilizadas para resolver outros tipos de problemas em diversas áreas do conhecimento. A análise de componentes principais é a técnica mais conhecida, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas da estatística multivariada para resolver a maioria dos problema práticos.

[1] Professor. Universidade Federal Rural do Rio de Janeiro, IT-Departamento de Engenharia, BR 465 km 7 - CEP 23890-000 – Seropédica – RJ. E-mail: varella@ufrrj.br.

Matriz de dados X

Considere a situação em que observamos ‘p’ características de ‘n’ indivíduos de uma população π. As características observadas são representadas pelas variáveis X_1, X_2, X_3, ..., X_p. A matriz de dados é de ordem ‘n x p’ e normalmente denominada de matriz ‘X’.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

A estrutura de interdependência entre as variáveis da matriz de dados é representada pela matriz de covariância ‘S’ ou pela matriz de correlação ‘R’. O entendimento dessa estrutura através das variáveis X_1, X_2, X_3, ..., X_p, pode ser na prática uma coisa complicada. Assim, o objetivo da análise de componentes principais é transformar essa estrutura complicada, representada pelas variáveis X_1, X_2, X_3, ..., X_p, em uma outra estrutura representada pelas variáveis Y_1, Y_2, Y_3, ..., Y_p não correlacionadas e com variâncias ordenadas, para que seja possível comparar os indivíduos usando apenas as variáveis Y_{is} que apresentam maior variância. A solução é dada a partir da matriz de covariância S ou da matriz de correlação R.

Matriz de covariância S

A partir da matriz X de dados de ordem ‘n x p’ podemos fazer uma estimativa da matriz de covariância Σ da população π que representaremos por S. A matriz S é simétrica e de ordem ‘p x p’.

$$S = \begin{bmatrix} \hat{V}ar(x_1) & \hat{C}ov(x_1x_2) & \hat{C}ov(x_1x_3) & \cdots & \hat{C}ov(x_1x_p) \\ \hat{C}ov(x_2x_1) & \hat{V}ar(x_2) & \hat{C}ov(x_2x_3) & \cdots & \hat{C}ov(x_2x_p) \\ \hat{C}ov(x_3x_1) & \hat{C}ov(x_3x_2) & \hat{V}ar(x_3) & \cdots & \hat{C}ov(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{C}ov(x_px_1) & \hat{C}ov(x_px_2) & \hat{C}ov(x_px_3) & \cdots & \hat{V}ar(x_p) \end{bmatrix}$$

Normalmente as características são obervadas em unidades de medidas diferentes entre si, e neste caso, segundo REGAZZI (2000) é conveniente padronizar as variáveis X_j (i=1, 2, 3,

..., p). A padronização pode ser feita com média zero e variância 1, ou com variância 1 e média qualquer.

Padronização com média zero e variância 1

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s(x_j)}, \quad i = 1, 2, \cdots, n \quad \text{e} \quad j = 1, 2, \cdots, p$$

Padronização com variância 1e média qualquer

$$z_{ij} = \frac{x_{ij}}{s(x_j)}, \quad i = 1, 2, \cdots, n \quad \text{e} \quad j = 1, 2, \cdots, p$$

em que, $\overline{X}_j$ e $S(x_j)$ são, respectivamente, a estimativa da média e o desvio padrão da característica j:

$$\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

$$\text{e} \qquad s(x_j) = \sqrt{\hat{V}ar(x_j)} \quad , \quad j = 1, 2, \cdots p$$

$$\sqrt{\hat{V}ar(x_j)} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2}{n-1} \quad \text{ou} \quad \sqrt{\hat{V}ar(x_j)} = \frac{\sum_{i=1}^{n} x_{ij}^2 - \frac{\left(\sum_{i=1}^{n} x_{ij}\right)^2}{n}}{n-1}$$

Após a padronização obtemos uma nova matriz de dados Z:

$$Z = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1p} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2p} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & z_{np} \end{bmatrix}$$

A matriz Z das variáveis padronizadas z_j é igual a matriz de correlação da matriz de dados X. Para determinar os componentes principais normalmente partimos da matriz de correlação R. É importante observar que o resultado encontrado para a análise a partir da matriz S pode ser diferente do resultado encontrado a partir da matriz R. A recomendação é que a

padronização só dever ser feita quando as unidades de medidas das características observadas não forem as mesmas.

Determinação dos componentes principais

Os componentes principais são determinados resolvendo-se a equação característica da matriz S ou R, isto é:

$$\det\left[\mathrm{R}-\lambda\mathrm{I}\right]=0 \quad \text{ou} \quad \left|\mathrm{R}-\lambda\mathrm{I}\right|=0$$

$$R=\begin{bmatrix} 1 & r(x_1x_2) & r(x_1x_3) & \cdots & r(x_1x_p) \\ r(x_2x_1) & 1 & r(x_2x_3) & \cdots & r(x_2x_p) \\ r(x_3x_1) & r(x_3x_2) & 1 & \cdots & r(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(x_px_1) & r(x_px_2) & r(x_px_3) & \cdots & 1 \end{bmatrix}$$

Se a matriz R for de posto completo igual a 'p', isto é, não apresentar nenhuma coluna que seja combinação linear de outra, a equação $\left|\mathrm{R}-\lambda\mathrm{I}\right|=0$ terá 'p' raízes chamadas de autovalores ou raízes características da matriz R. Na montagem da matriz de dados X é importante observar que o valor de 'n' (indivíduos, tratamentos, genótipos, etc.) dever ser pelo menos igual a 'p+1', isto é, se queremos montar um experimento para analisar o comportamento de 'p' características de indivíduos de uma população é recomendado que o delineamento estatístico apresente pelo menos 'p+1' tratamentos.

Sejam $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_p$ as raízes da equação característica da matriz R ou S, então:

$$\lambda_1 > \lambda_2 > \lambda_3 \cdots, \lambda_p .$$

Para cada autovalor λ_i existe um autovetor $\tilde{a}_i$:

$$\tilde{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix}$$

Os autovetores $\tilde{a}_i$ são normalizados, isto é, a soma dos quadrados dos coeficientes é igual a 1, e ainda são ortogonais entre si. Devido a isso apresentam as seguintes propriedades:

$$\sum_{j=1}^{p} a_{ij}^2 = 1 \quad \left(\tilde{a}_i' \cdot \tilde{a}_i = 1\right)$$

$$\text{e} \sum_{j=1}^{p} a_{ij} \cdot a_{kj} = 0 \quad \left(\tilde{a}_i' \cdot \tilde{a}_k = 0 \text{ para } i \neq k\right)$$

Sendo $\tilde{a}_i$ o autovetor correspondente ao autovalor λ_i , então o i-ésimo componente principal é dado por:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

Os componentes principais apresentam as seguintes propriedades:

1) A variância do componente principal Y_i é igual ao valor do autovalor λ_i.

$$\hat{\text{V}}\text{ar}\left(Y_i\right) = \lambda_i$$

2) O primeiro componente é o que apresenta maior variância e assim por diante:

$$\hat{\text{V}}\text{ar}(Y_1) > \hat{\text{V}}\text{ar}(Y_2) > \cdots > \hat{\text{V}}\text{ar}(Y_p)$$

3) O total de variância das variáveis originais é igual ao somatório dos autovalores que é igual ao total de variância dos componentes principais:

$$\sum \hat{\text{V}}\text{ar}(X_i) = \sum \lambda_i = \sum \hat{\text{V}}\text{ar}(Y_i)$$

4) Os componentes principais não são correlacionados entre si:

$$\hat{\text{C}}\text{ov}\left(Y_i, Y_j\right) = 0$$

Contribuição de cada componente principal

A contribuição C_i de cada componenete principal Y_i é expressa em porcentagem. É calculada dividindo-se a variância de Y_i pela varância total. Representa a proporção de variância total explicada pelo componenete principal Y_i.

$$C_i = \frac{\hat{\text{V}}\text{ar}\left(Y_i\right)}{\sum_{i=1}^{p} \hat{\text{V}}\text{ar}\left(Y_i\right)} \cdot 100 = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \cdot 100 = \frac{\lambda_i}{\text{traço}\left(S\right)} \cdot 100$$

A importância de um componente principal é avaliada por meio de sua contribuição, isto é, pela proporção de variância total explicada pelo componente. A soma dos primeiros *k* autovalores representa a proporção de informação retida na redução de *p* para *k* dimensões. Com essa informação podemos decidir quantos componente vamos usar na análise, isto é, quantos componentes serão utilizados para diferenciar os indivíduos. Não existe um modelo estatístico que ajude nesta decisão. Segundo REGAZZI (2000) para aplicações em diversas áreas do conhecimento o número de componentes utilizados tem sido aquele que acumula 70% ou mais de proporção da variância total.

$$\frac{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{Y}_1)+\cdots\hat{\mathrm{V}}\mathrm{ar}(\mathrm{Y}_k)}{\sum_{i=1}^{k}\hat{\mathrm{V}}\mathrm{ar}(\mathrm{Y}_i)}\cdot 100 \geq 70\% \quad \text{onde} \quad k < p$$

Interpretação de cada componente

Esta análise é feita verificando-se o grau de influência que cada variável X_j tem sobre o componente Y_i. O grau de influência é dado pela corelação entre cada X_j e o componente Yi que está sendo interpretado. Por exemplo a correlação entre X_j e Y_1 é:

$$\mathrm{Corr}(\mathrm{X}_{j,}\mathrm{Y}_1)=\mathrm{r}_{\mathrm{Xj}\cdot\mathrm{Y1}}=\mathrm{a}_{1j}\cdot\frac{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{Y}_1)}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X}_j)}}= \quad \sqrt{\lambda_1}\cdot\frac{\mathrm{a}_{1j}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X}_j)}}$$

Para comparar a influência de X_1, X_2, ..., Xp sobre Y_1 análisamos o peso ou loading de cada variável sobre o componente Y_1. O peso de cada variável sobre um determinado componente é dado por:

$$\mathrm{w}_1=\frac{\mathrm{a}_{11}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X}_1)}},\ \mathrm{w}_2=\frac{\mathrm{a}_{12}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X}_2)}},\cdots\mathrm{w}_p=\frac{\mathrm{a}_{1p}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X}_p)}},\ \text{sendo } \mathrm{w}_1 \text{ o peso de } \mathrm{X}_1.$$

Se o objetivo da análise for a obtenção de índices, prática muito comum em Economia, a análise termina aqui.

Se o objetivo da análise é comparar ou agrupar indivíduos, a análise continua e é necessário calcular os escores para cada componente principal que será utilizado na análise.

Escores dos componentes principais

Os escores são os valores dos componentes principais. Após a redução de p para k dimensões, os k componentes principais serão os novos índivíduos e toda análise é feita utilizando-se os escores desses componentes. No Quadro 1 é exemplificado a organização de um conjunto de dados composto por **n** tratamentos, **p** variáveis e **k** componentes principais.

Quadro 1. Organização de um conjunto de dados com n tratamentos, p variáveis e k componentes

Tratamentos	Variáveis				Escores dos componentes principais			
(Indivíduos)	X1	X2	...	Xp	Y1	Y2	...	Yk
1	X11	X12	⋮	X1p	Y11	Y12	...	Y1k
2	X21	X22	⋮	X2p	Y21	Y22	...	Y2k
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Xn1	Xn2	...	Xnp	Yn1	Yn2	...	Ynk

Assim temos que os escores do **primeiro** componente para os **n** tratamentos são:

Trat	Primeiro conponente principal
1	$Y_{11} = a_{11}X_{11} + a_{12}X_{12} + \cdots + a_{1p}X_{1p}$
2	$Y_{21} = a_{11}X_{21} + a_{12}X_{22} + \cdots + a_{1p}X_{2p}$
⋮	⋮
N	$Y_{n1} = a_{11}X_{n1} + a_{12}X_{n2} + \cdots + a_{1p}X_{np}$

Exemplo de aplicação

No Quadro 2 estão os valores originais observados (X_1 e X_2) e padronizados (Z_1 e Z_2) de duas variáveis para cinco tratamentos (n=5).

Quadro 2. Valores originais e padronizados de duas variáveis para cinco tratamentos

Tratamentos	Variáveis originais		Variáveis padronizadas	
	X_1	X_2	Z_1	Z_2
1	102	96	24,3827	6,9554
2	104	87	24,8608	6,3033
3	101	62	24,1436	4,4920
4	93	68	22,2313	4,9268
5	100	77	23,9046	5,5788
Variância	17,50	190,50	1	1
Média	100,00	78,00	23,9046	5,6513

Os dados estão padronizados para variância 1:

$$Z_{ij} = \frac{X_{ij}}{s(X_j)} \quad \Rightarrow \quad Z_{12} = \frac{104}{\sqrt{17,5}} = 24,8608$$

A matriz de correlação é:

$$R = \begin{bmatrix} 1 & 0,5456 \\ 0,5456 & 1 \end{bmatrix}$$

A equação característica é: $|R - \lambda I| = 0$

$$\begin{vmatrix} 1-\lambda & 0,5456 \\ 0,5456 & 1-\lambda \end{vmatrix} = 0$$

$$\lambda^2 - 2\lambda + 0,7023 = 0$$

Os autovalores da matriz de correlação R são:

$$\lambda_1 = 1,5456 \text{ e } \lambda_2 = 0,4544$$

A soma de λ_1 e λ_2 é igual ao traço da matriz R. O traço de uma matriz é a soma dos elementos de sua diagonal principal.

$$\text{traço(R)} = 1+1=2$$

Obtenção dos componentes principais

O autovetor normalizado para o primeiro componente principal é:

$$\tilde{a}_1 = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0{,}7071 \\ 0{,}7070 \end{bmatrix}$$

e o primeiro componente principal é:

$$Y_1 = 0{,}7071Z_1 + 0{,}7071Z_2$$

Da mesma forma para o segundo componente principal temos:

$$\tilde{a}_{21} = \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0{,}7071 \\ 0{,}7070 \end{bmatrix}$$

$$Y_2 = -0{,}7071Z_1 + 0{,}7071Z_2$$

Quadro 3. Informações que podem ser obtidas com a análise de componentes principais

Componente principal	Variância (Autovalor)	Coeficiente de ponderação		Correlação entre Zj eYi		Porcentagem da variância total	Porcentagem acumulada de variância dos Y_i
		Z_1	Z_2	Z_1	Z_2		
Y1	1,5456	0,7071	0,7071	0,879	0,879	77,28	77,28
Y2	0,4544	-0,7071	0,7071	-0,476	0,476	22,72	100,00

Quadro 4. Escores dos dois componentes principais para os cinco tratamentos obtidos a partir da matriz de correlação R.

Tratamentos	Escores dos componentes principais	
	Y_1	Y_2
1	22,16	-12,32
2	22,04	-13,12
3	20,25	-13,90
4	19,20	-12,24
5	20,85	-12,96

Gráfico de dispersão

São utilizados para visualizar a dispersão dos tratamentos em função dos escores dos componentes principais em espaço bi ou tridimensional. A dispersão das médias de tratamentos para este exemplo está ilustrada na Figura 2.

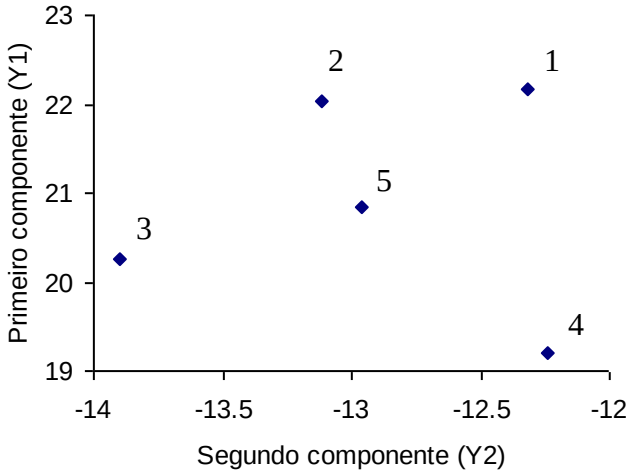

Figura 2. Dispersão dos tratamentos em função dos escores dos componentes principais.

Programa SAS para obtenção dos componentes principais

BIBLIOGRAFIA

REGAZZI, A.J. Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

KHATTREE, R. & NAIK, D.N**. Multivariate data reduction and discrimination with SAS software**. Cary, NC, USA: SAS Institute Inc., 2000. 558 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.