UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Giovanni Bezerra Barbosa

Técnica de Análise de Componentes Principais: Redução de variáveis econômicas de empresas

Campina Grande - Brasil

Junho de 2018

Giovanni Bezerra Barbosa

Técnica de Análise de Componentes Principais: Redução de variáveis econômicas de empresas

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Edwirde Luiz Silva Camêlo

Campina Grande - Brasil

Junho de 2018

B238t Barbosa, Giovanni Bezerra.
Técnica de análise de componentes principais [manuscrito] : redução de variáveis econômicas de empresas / Giovanni Bezerra Barbosa. - 2018.
40 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2018.
"Orientação : Prof. Dr. Edwirde Luiz Silva Camêlo , Coordenação do Curso de Estatística - CCT."

1. Estatística. 2. Estatística descritiva. 3. Análise de Componentes Principais.

21. ed. CDD 519.53

Elaborada por Giulianne Monteiro Pereira - CRB - 15/714 BC/UEPB

Giovanni Bezerra Barbosa

Técnica de Análise de Componentes Principais: redução de variáveis econômicas de empresas

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 20 de Junho de 2018.

BANCA EXAMINADORA

Dr.Edwirde Luiz Silva Camêlo
Universidade Estadual da Paraíba

Dr.Mácio Augusto Albuquerque
Universidade Estadual da Paraíba

Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba

Dedico este trabalho a minha família, especialmente meus pais, José Aquilino Barbosa e Maria de Lourdes Bezerra Barbosa que sempre me deram força, coragem e constante apoio para seguir em busca de meus objetivos.

Agradecimentos

O meu sincero agradecimento ao meu Deus, autor da minha vida, por ter aberto as portas da Universidade para mim, por sua graça e misericórdia que me fortalecem a cada dia da minha existência. A ele toda honra e louvor por mais essa etapa da minha vida que tem se cumprido.

Aos meus pais, por terem dispensado a mim o amor necessário desde a minha infância e por serem os melhores pais que eu poderia ter.

Ao meu orientador Edwirde Luiz Silva Camêlo pela paciência que teve para comigo e por sua competência profissional que tanto me auxiliou durante esse processo.

Aos meus companheiros de curso em todo o período de convivência acadêmica, em especial a Reginaldo Ferreira Neves e a Washington Luís Pereira, que me ajudaram tantas vezes nos trabalhos acadêmicos.

Tudo tem o seu tempo determinado e há tempo para todo propósito debaixo do céu.

“Não há nada que dominemos inteiramente a não ser os nossos pensamentos...”
(René Descartes)

RESUMO

A Análise de Componentes Principais (ACP) é uma técnica estatística descritiva cujo ponto de partida é uma matriz de dados com uma série de indivíduos de n variáveis. Por isso, geralmente classifica-se como uma técnica multivariante. Aplicou-se a análise de componentes principais em um banco dados que consistia em 42 observações de empresas espanholas com 8 variáveis a ser estudada, mediante análise orientada de interesse, com o intuito de reduzir a dimensionalidade, filtrando aqueles descritores ou variáveis de maior correlação com as unidades de estudo. Assim, discute-se neste trabalho a aplicabilidade e a interpretação da análise de componentes principais e com utilização do gráfico biplot. Observando a distribuição dos produtos destas empresas num gráfico bidiomensional das duas primieras componentes principais, a primeira componente, explicada por 78.75% da variação, foi melhor explicada pelas variáveis actol, recprop, capital e Bimobil; enquanto que as variáveis: plantil1, ventas y benefico contribuem melhor para denominar a segunda componente principal, que explicou aproximadamente 10% da variação.

Palavras-chave: Componentes, Empresas, Análise.

Abstract

Principal Component Analysis (PCA) is a descriptive statistical technique whose starting point is a data matrix with a series of users making different models. The data set considered consists of 42 observations of Spanish companies with 9 variables. The purpose of reducing the dimensionality, filtering the descriptors or variables correlated with as units of study. Thus, it is discussed the applicability and an interpretation of the analysis of main components and using the biplot graph. A first component, explained by 78.75 % of the variation, was better explained by the variables actol, recprop, capital and Bimobil; while they are variable: plantil1, ventas y benefico contribute best to denominate the second principal component, which explained approximately 10% of the variation.

Palavras-chave: Components, Companies, Analysis.

Lista de ilustrações

Lista de tabelas

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

1 Introdução

A análise estatística multivariada ou simplesmente análise multivariada é o ramo da estatística direcionada ao estudo das amostras e distribuição multidimensionais, ou seja, são métodos estatísticos apropriados para estudos em que várias variáveis são consideradas simultaneamente. Entretanto, apesar de as técnicas multivariadas terem eficiência comprovada e proporcionarem enriquecimento das informações extraídas nos dados, é necessário para seu uso a disponibilidade de recursos computacionais, motivo pelo qual a referida técnica ficou limitada no seu uso e do repasse entre os pesquisadores das diversas áreas da ciência no Brasil. Porém, com a incrementação dos recursos da informática nos últimos anos, a técnica atraiu a atenção dos pesquisadores das diversas áreas, tornando o seu emprego potencialmente grande e, consequentemente, o seu conhecimento indispensável. Trabalhos visando descrever, discutir e recomendar o uso de técnicas estatísticas multivariadas na análise de dados florestais são encontrados em literatura (teses, artigos de periódicos, livros, anais de congressos, etc.). Princípios básicos da referida técnica podem ser encontrados em (MORRISON, 1990); (KENDALL, 1975); (JOHNSON; WICHERN, 1988) entre outros.

Atualmente, os pesquisadores defrontam-se com dezenas ou mesmo centenas de variáveis diferentes em suas análises. Com tantas dimensões diferentes, é difícil abarcar ou sequer visualizar os padrões de associação entre elas. O processo é, além do mais, complicado pelo fato de frequentemente haver redundância substancial entre dimensões, o que leva a altos níveis de correlação e multicolinearidade.

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.

A análise dos componentes principais consiste numa transformação linear de 'p' variáveis originais em 'q' novas variáveis, de tal modo que a primeira nova variável

computada seja responsável pela maior variação possível existente no conjunto de dados, a segunda pela maior variação possível restante e assim por diante até que toda a variação do conjunto tenha sido explicada (JOHNSON; WICHERN, 1988).

Neste trabalho, foi utilizada a técnica do PCA para identificar variáveis dependentes para redução dos parâmetros que caracterizassem determinados variáveis do conjunto de dados da empresa estudada. O objetivo principal da PCA é explicar a estrutura de variâncias e covariâncias de um vetor aleatório composto de p-variáveis aleatórias iniciais, podendo-se resumir sua informação, ou reduzir a dimensão dos dados, eliminando as informações redundantes contida no complexo das variáveis originais, o que torna os resultados mais simples e de interpretação mais clara. O desenvolvimento desta tem as seguintes características:

- Não requer uma suposição de normalidade multivariada. Entretanto, as componentes principais (CP's) derivadas para populações com distribuição normal multivariada têm interpretações uteis em termos de elipsóides de confiança, além da possibilidade de se fazer algumas inferências sobre eles.
- Não impõe qualquer modelo causal, mas também não permite detectar quaisquer relações de causa-efeito entre as variáveis iniciais mesmo se existirem.
- É uma técnica de análise intermediária (exploratória) em muitas investigações (como por exemplo, na regressão múltipla, análise de agrupamentos etc.) e, portanto não constitue um método final e conclusivo (FERREIRA, 2008).

A técnica depende somente da estrutura de covariâncias S ou da matriz de correlações R do conjunto de variáveis observadas.

A técnica de ACP, realizada a partir da matriz de variâncias e covariâncias, consiste em transformar um conjunto de variáveis iniciais $X_1; X_2; \cdots; X_p$, relacionadas entre si, em um novo conjunto de variáveis $C_1; C_2; \cdots; C_p$, não correlacionadas (ortogonais); chamadas de Componentes Principais, arranjadas em ordem decrescente de variâncias (REGAZZI, 2000). Isso é feito construindo combinações lineares das variáveis originais.

Na seção 1 serão apresentados um breve resumo do que será apresentado no trabalho, a metodologia utilizada e os métodos estatísticos serão abordados mostrando a ferramenta de multivariada utilizando a análise de componentes principais e com seus aspectos históricos e o método de coleta das variáveis coletados no banco de dados, no qual a estrutura física e o sistema de informação utilizados na coleta dos dados são também descritos. Na seção 2 é descrita toda fundamentação teórica sobre o tema no qual foi feita a aplicação da análise de componentes principais, detalhando todo o processo e no tratamento dos dados disponíveis no trabalho. Na seção 3 é apresentado os resultados obtidos gerados

após a aplicação da análise de componentes principais e seus resultados obtidos. Finalmente, Na seção 4, será apresentado a conclusão e suas referencias bibliográficas.

1.1 Objetivo

Analisar as variáveis métricas reduzindo em poucas dimensões para melhor visualização num espaço bidimensional.

1.2 Fundamentação Teórica

A análise estatística multivariada ou simplesmente análise multivariada é o ramo da estatística direcionado ao estudo de p-variáveis. No entanto, apesar de as técnicas multivariadas terem eficiência comprovada e proporcionarem enriquecimento das informações extraídas de dados amostrais, é necessária para seu uso a disponibilidade de recursos computacionais, motivo pelo qual a referida técnica ficou limitada no seu uso e do repasse entre os pesquisadores das diversas áreas da ciência, no Brasil. Entretanto, com o desenvolvimento dos recursos da informática nos últimos anos, a técnica atraiu a atenção dos pesquisadores das diversas áreas, tornando o seu emprego potencialmente grande e, consequentemente, o seu conhecimento indispensável. A análise multivariada compreende várias técnicas que, segundo (KENDALL, 1975) distinguem-se em:

a) Técnicas de Avaliação da Interdependência: estuda as relações de um conjunto de variáveis entre si. - "Cluster Analysis" ou Análise de Agrupamento - Componentes Principais - Correlações Canônicas - Análise Fatorial - Escala

b) Técnicas de Avaliação da Dependência: estuda a dependência de uma ou mais variáveis em relação às outras. Como exemplo: Regressão e Análise Discriminante

A análise de componentes principais é uma técnica multivariada, que segundo (KENDALL, 1975), é uma técnica de avaliação da interdependência, ou seja, estuda as relações de um conjunto de variáveis entre si.

1.3 Marco Histórico

A técnica de componentes principais foi originalmente descrita por (PEARSON, 1901) , em um artigo onde deu ênfase à sua utilização no ajustamento de um subespaço a uma nuvem de pontos. Posteriormente, a técnica foi consolidada por (HOTELLING, 1933) ,para o propósito particular de analisar estruturas de correlações ((MORRISON et al., 1976),(MARDIA, 1976)).A técnica de componentes principais procura explicar a estrutura de variâncias e covariâncias através de poucas combinações lineares das variáveis originais, com os objetivos de reduzir os dados, colocá-los numa forma mais adequada

para análise, evidenciar as tendências e facilitar sua interpretação. A utilização da análise de componentes principais tem por finalidade proporcionar simplificação estrutural dos dados,de modo que a diversidade, influenciada a princípio por um conjunto p-dimensional (p = números de caracter considerados no estudo), possa ser avaliada por um complexo bi ou tridimensional de fácil interpretação geométrica. Ou ainda, a análise por componentes principais, consiste em transformar um conjunto original de variáveis em outro conjunto, de dimensões equivalentes, mas com propriedades importantes de grande interesse em certos estudos.

Os princípios básicos desta técnica são descritos por vários autores, tais como (MORRISON et al., 1976); (MARDIA K. V.; KENT, 1979); (KENDALL, 1980); (JOHNSON; WICHERN, 1988), entre outros. Segundo estes autores, cada componente principal é uma combinação linear das variáveis originais, que são independentes entre si e estimadas com o propósito de reter, em ordem de estimação, o máximo da informação, em termos de variação total, contida nos dados originais. Assim, entre todos os componentes principais, o primeiro tem a maior variância, o segundo tem a segunda maior e assim sucessivamente.

A grande importância do conhecimento da técnica dos componentes principais, reside no fato de constituir um procedimento básico do qual derivam vários outros métodos de análise de dados multivariados, como por exemplo, análise de agrupamento (cluster analysis).

Obtenção dos Componentes Principais algebricamente, componentes principais são combinações lineares particulares das p variáveis aleatórias $(X_1, X_2, \cdots, X_p)$. Geometricamente, estas combinações lineares representam a seleção de um novo sistema de coordenadas obtidas pela rotação do sistema original como $(X_1, X_2, \cdots, X_p)$ como eixos. Os novos eixos representam as direções com variabilidade máxima e fornece uma descrição mais simples e mais parcimoniosa da estrutura de covariâncias.

Os componentes principais dependem somente da matriz de covariâncias (S) ou da matriz de correlação (R) de $(X_1, X_2, \cdots, X_p)$. Assim, a técnica de componentes principais caracteriza-se por trabalhar com a média amostral ou ser usada nas situações em que não há repetições de dados. O seu desenvolvimento não necessita de normalidade. No entanto, a análise de componentes derivada de populações normais multivariadas têm suas interpretações usuais em termos de elipsoides de densidade constante (JOHNSON; WICHERN, 1988). Entretanto, embora a análise, formalmente não requeira a distribuição normal multivariada, ela é mais apropriada para variáveis quantitativas contínuas. Quando os dados são constituídos de contagem, razões, proporções ou percentagens, a transformação é recomendada para tornar sua distribuição mais apropriada, previamente à análise de componentes principais.

a) Examinar as correlações entre caracteres estudados;

b) Resumir um grande conjunto de caracteres em outro menor e de sentido biológico;

c) Avaliar a importância de cada caracter e promover a eliminação daqueles que contribuem pouco, em termos de variação, no grupo de indivíduos avaliados;

d) Construir índices que possibilitem o agrupamento de indivíduos; e

e) Permitir o agrupamento de indivíduos com o mais alto grau de similaridade, mediante exames visuais em dispersão gráficas no espaço bi ou tridimensional.

A PCA tem como principais vantagens: retirar a multicolinearidade das variáveis, pois permite transformar um conjunto de variáveis originais intercorrelacionadas em um novo conjunto de variáveis não correlacionadas (componentes principais). Além disso, reduz muitas variáveis a eixos que representam algumas variáveis, sendo estes eixos perpendiculares (ortogonais) explicando a variação dos dados de forma decrescente e independente, (REGAZZI, 2000). Pode-se identicar algumas das desvantagens da PCA que são: a sensibilidade a outliers, não recomendada quando se tem duplas ausências (muitos zeros na matriz) e dados ausentes.

1.4 Material e método

O conjunto de dados utilizado neste trabalho foi retirado do anuário El País de 1994 em que foi publicado informações econômicas financeiras das 229 maiores empresas espanholas. Para este trabalho, considerou-se apenas as 40 maiores empresas em relação ao volume de vendas em euro em 1992 (URIEL; MANZANO, 2002). A tabela 1 representa o ganho brutos, das empresas em medidas de unidades monetárias (euro)referentes as variáveis em estudo.

Para realizar a PCA, faz-se necessário o auxílio de um software versão R 3.0.0, pois a amostra em estudo possui a dimensão $p = 8$, ou seja, tem-se 8 variáveis (considerados 7 variáveis para facilitar a interpretação). Essas variáveis suplementares são utilizadas quando o pesquisador busca identificar o comportamento em relação às demais variáveis. As variáveis envolvidas neste estudo foram:

1. Cifra de vendas (ventas)
2. Capital social (Capital)
3. Recursos próprios (Recpprop)
4. Ativos totais (acttot)
5. Benefícios netos (Benfico)

Tabela 1 – Informação econômica financeira de 40 (do total de 229 maiores) empresas em quanto a seu volume de vendas em 1992

empresa	ventas	capital	recprop	acttot	benefico	inmovil	cashflow
E1	1271510	463480	1488378	3917167	83899	3568851	425802
E2	775104	46207	248813	463335	23795	173572	37005
E3	775218	437564	1028297	3133889	58778	2911741	164567
E4	700963	3247	4869	150236	1531	61898	1990
E5	674063	84000	91056	472004	-12756	381012	7934
E6	631003	18442	84646	220062	14729	101784	21725
E7	537744	22781	129448	245723	9059	95735	30795
E8	489155	36154	101465	332985	12541	194837	26196
E9	448465	18925	94010	258782	13495	130790	22804
E10	445853	213281	138922	544617	-34824	424760	-8757
E11	430878	208004	503065	1114274	95118	1015846	164761
E12	427333	21070	126718	228674	10700	46909	27250
E13	425909	14012	87373	296022	34329	113939	44955
E14	414472	41547	95963	216214	21201	77787	36055
E15	371566	44596	143544	320857	11248	194183	17686
E16	341596	145069	309904	1116935	13420	970791	66167
E17	304852	8574	28179	145642	5670	73005	11106
E18	299375	7506	45339	297389	11351	71137	16674
E19	291305	29724	100864	254492	10325	48424	13073
E20	288109	5344	76759	132017	5687	40226	14664
E21	258331	148201	298108	792174	12772	709445	40021
E22	242303	14012	43450	78374	207	38580	1486
E23	234929	19500	30757	159328	523	47070	12636
E24	206811	208593	273371	762166	15247	693523	47212
E25	206259	4971	40655	164872	4757	27089	7359
E26	200380	10500	16878	64514	-3194	22002	-448
E28	173067	55766	54844	198097	-14389	146813	-5788
E30	158106	125145	33663	339125	-67644	234579	-53188
E32	151688	9772	16601	81701	6603	50300	11170
E33	148822	2638	5881	58299	1850	40342	6753
E34	147634	46692	31700	139679	841	60862	5716
E35	132294	20020	50179	92498	8369	69612	13577
E36	129850	18000	41690	91828	6342	57876	10214
E37	127241	46893	135955	221399	4521	192907	12882
E38	125099	17515	22688	55192	1487	18084	4023
E39	121076	12639	21345	99751	800	20331	2864
E40	119624	5625	32180	91594	2270	30208	4747

6. Imobilizado neto (inmovol)

7. Cas-flow (Cashflow)

8. Rentabilidade (Benreccpr)

1.4.1 Análise de componentes principais(ACP)

A PCA também não é recomendada quando se tem mais variáveis do que unidades amostrais. Ao reduzir o número de variáveis, há perda da informação de variabilidade das variáveis originais. Mas que a parte explicada seja o padrão de resposta e a outra parte o ruído, ou seja, erro de medida e redundância.

As PCA associadas a uma vetor de variáveis $X = (X_1, \ldots, X_p)'$, são combinações lineares (CL) de ditas variáveis submetidas a certas variáveis. É uma técnica cujo objetivo básico é a redução da dimensão de um problema de p variáveis a outro de dimensão menor com um possível novas variáveis. Aqui na ACP paramétrico, o vetor aleatório X será considerado padronizado, na hora de considerar a inferência, por uma distribuição normal p-dimensional.

1.4.2 Construção das componentes principais (CP)

A ACP pretende explicar a estrutura de covariância de un vetor aleatório X mediante a procura de um novo sistema de eixos coordenados (as CP) que indicam as direções de maior variabilidade en uma situação teórica dada (com Σ matriz de covariância de X conhecida) ou posteriormente de uma matriz Σ estimada a partir de dados observados (PEÑA, 2002).

Suponha que $X = (X_1, \cdots, X_p)'$ com $Cov(X) = \Sigma$ semidefinida positiva e com $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, as raízes características correspondentes. Sejam as combinações lineares

$$\begin{cases} Y_1 = a_1'X = a_{11}'X_1 + \cdots + a_{1p}'X_p \\ \vdots \\ Y_p = a_p'X = a_{p1}'X_1 + \cdots + a_{pp}'X_p \end{cases}$$

Considere o vetor aleatório $Y = (Y_1, \ldots, Y_p)'$. Considerando duas quaisquer de suas componentes, i e j, é claro que

$$Var(Y_i) = a_i'\Sigma a_i \quad Cov(Y_i, Y_j) = a_i'\Sigma a_j$$

e isto é certo para todo vetor X, qualquer que seja sua distribuição.

Denomina-se *componentes principais (CP)* as combinações lineares $Y_1, \ldots, Y_p$ que são não correlacionadas entre si, tais que serão máximas e as variâncias $a_i'\Sigma a_i, i = 1, \ldots, p$.

Passos para construir as CP

- Considere a combinação linear de variância máxima (chamando-se Y_1) de modo que esta variância será $Var(Y_i) = a_i'\Sigma a_i$. Obviamente isto tem uma indeterminação já que a variância aumentará sem mais que multiplicar a por uma constante positiva.
- Introduz-se portanto a restrição de que os vetores a sejam unitários em todas as CP a obter, portanto $a_i'a_i = 1$.
- Denomina-se primeira componente principal a CL $Y_1 = a_1'X$ ta que faz máxima $Var(Y_1)$ com a restrição $a_1'a_1 = 1$.
- Denomina-se segunda componente principal a CL $Y_2 = a_2'X$ tal que se faz máxima $Var(Y_2)$ com a restrição $a_2'a_2 = 1$ e com a restrição adicional de ser incorrelacionada com Y_1, isto é,
$$Cov(a_1'X, a_2'X) = 0.$$
- O procedimento continua até construir as p combinações lineares $Y_1, \ldots, Y_p$. Tal que uma Y_i qualquer, $i = 1, \ldots, p$, por definição, maximiza $Var(a_i'X)$ sujeita a $a_i'a_i = 1$ e a $Cov(a_i'X, a_k'X) = 0$ para $k < i$.

Introduz-se esta restrição mediante o multiplicador de Lagrange definido por:

$$M = a_1\Sigma a_1 - \lambda(a_1'a_1 - 1)$$

1.4.3 Cálculo da primeira componente principal

Defini-se a primeira componente principal como:

$$Y_1 = e_1'X,\ e_1'e_1 = 1(vetor\ ortonormal)$$

tal que

$$Var(Y_1) = Var(a'X),\ \max_a Var(a'X) = Var(e_1'X) = e_1'\Sigma e_1.$$

Resolvendo por meio do problema de Lagrange de maximização condicionada, tem-se:

$$\left\{ \begin{array}{l} \max\limits_{a}\{a'\Sigma a\} \\ a'a = 1 \end{array} \right\} \Rightarrow \Phi_1(a) = a'\Sigma a - \lambda(a'a - 1) \Rightarrow \frac{\partial \Phi_1(a)}{\partial a} = 2\Sigma a - 2\lambda a = 0 \Rightarrow (\sigma - \lambda I)a = 0$$

Suposto que $\Sigma_{p\times p}$ tem autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$,[1] con autovectores associados $e_1, e_2, \ldots, e_p$ e como $a'\Sigma a = \lambda a'a = 1$, $Var(a'\Sigma a) = \lambda$ e é claro que tomando $a = e_1$, correspondente ao maior autovalor, se resolve o problema, de modo que a primeira CP é $Y_1 = e_1'X$ e tem-se $Var(Y_1) = \lambda_1$.

1.4.4 Cálculo da segunda componente principal

Trata-se de obter, segundo a definição anterior, uma combinação linear $Y_2 = a'X$, incorrelacionada com Y_1 e de variância máxima. Portanto,

$$\max_{a}\{a'Xa\},\ con\ a'a = 1,\ a'\Sigma e_1 = 0$$

(PEÑA, 2002).

Assim, para a segunda componente tem-se a seguinte função objetivo:

$$\phi_2 = a_1'l_1 + a_2'\Sigma a_2 - \lambda_1(a_1'a_1 - 1) - \lambda_2(a_2'a_2 - 1)$$

$$\Phi_2(a) = a'\Sigma a - \lambda(a'a - 1) - 2v(a'\Sigma e_1) \Rightarrow \frac{\partial \Phi_2(a)}{\partial a} = 2\Sigma a - 2\lambda a - 2v\Sigma e_1.$$

De uma forma mais simplificada, tem-se:

$$\begin{aligned} \frac{\delta \phi}{\delta a_1} &= 2\Sigma a_1 - 2\lambda_1 a_1 = 0 \\ \frac{\delta \phi}{\delta a_2} &= 2\Sigma a_2 - 2\lambda_2 a_2 = 0 \end{aligned}$$

A solução do sistema acima será

$$\begin{aligned} \Sigma a_1 &= \lambda_1 a_1 \\ \Sigma a_2 &= \lambda_2 a_2 \end{aligned}$$

[1] Σ, em geral, como matriz de covariancia, é semidefinida positiva

Que indica que l_1 e l_2 devem ser autovetores de Σ. Tomando os autovetores associados de norma unidade e substituindo na Equação 1.1, obtém-se que, no máximo, a função objetivo será

$$\phi \;=\; \lambda_1 + \lambda_2$$

É claro que λ_1 e λ_2 devem ser os dois autovalores maiores da matriz de covariância Σ.

1.4.5 Cálculo da $(r+1)$-ésima componente principal $(1 \leq r+1 \leq p)$.

Neste caso temos:

$$Y_{r+1} = a'X;\ a'a = 1;\ a'\Sigma e_i = 0,\ i = 1, \ldots r$$
$$\Phi_{r+1}(a) = a'\Sigma a - \lambda(a'a - 1) - 2\sum_{i=1}^{r} v_i a'\Sigma e_i.$$

Pode demonstrar que, sendo $\lambda_i \neq 0,\ i = 1, \ldots, r$ o problema conduz a $v_i = 0,\ i = 1, \ldots, r$ de modo que o sistema que resolve o problema de maximização é

$$\{2\Sigma a - 2\lambda a = 0, \Sigma a - \lambda a = 0, (\Sigma - \lambda I)a = 0\}\cdot$$

Se $\lambda_{r+1} \neq 0$, basta tomar $\lambda = \lambda_{r+1}$, $l = e_{r+1}$ e se obtém a $(r+1)$-ésima CP que é

$$Y_{r+1} = e'_{r+1}X,\ Var(Y_{r+1}) = \lambda_{r+1}.$$

No caso em que $\lambda_{r+1} = 0,\ \lambda_i \neq 0, i \neq r+1$, toma-se uma CL de α_{r+1} y α_i para qual $\alpha_i \neq 0$.

Uma vez conseguidos $A = (e_1, \ldots, e_p),\ \Lambda = diag(\lambda_1, \ldots, \lambda_p),\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, como $A'A = I$ y $\Sigma A = A\Lambda$, tem-se que $A'\Sigma A = \Lambda$.

1.4.6 Construção conjunta das p componentes principais.

No lugar de ir obtendo sucessivamente as CP resolvendo os sucessivos problemas de máximo condicionado e ao final considerar globalmente todos. Para resolver cabalmente desde o início. Claro que se obtém os mesmos resultados, mas em lugar de ir aplicando e resolvendo os sucessivos problemas de máximos condicionados de Lagrange, vamos nos basear em um conhecido resultado de maximização.

Seja H uma matriz $p \times p$ definida positiva, com autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ e autovalores normalizados $e_1, \ldots, e_p$, e seja x um vetor $p \times 1$, arbitrário não nulo. Então cumpre-se que:

$$\max_x \frac{x'Hx}{x'x} = \lambda_1, \text{ alcanzado en } x = e_1,$$
$$\min_{x'x \neq 0} \frac{x'Hx}{x'x} = \lambda_p, \text{ alcanzado em } x = e_p,$$

O procedimento de cálculo desta maximização forge do objetivo do trabalho.

1.4.7 Principais propriedades das componentes

As componentes principais são novas variáveis com as seguintes propriedades (PEÑA, 2002).

1. A soma das variâncias das componentes é igual a soma das variâncias das variáveis oroginais, e a variância generalizada das componentes é igual a original

$$tr(S) \quad = \quad Var(x_1) + \cdots + var(x_p) = \lambda_1 + \cdots + \lambda_p$$

Portanto, $\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(z_i)$. As novas variáveis z_i tem conjuntamente a mesma variabilidade que as variáveis originais.

As componentes principais também conserva a variância generalizada(determinante da matriz de covariância das variáveis). Como o determinante é o produto dos autovalores, teremos então:

$$|S| \quad = \quad \lambda_1 \cdots \lambda_p = \prod_{i=1}^{p} Var(z_i) = |S_z|$$

2. A proporção da variabilidade explicada por uma componente é o quociente entre sua variância, o autovalor associado ao autovetor que o define, e a soma dos autovalores da matriz.

Em efeito, a variância dfa componente h é λ_h, e a soma das variâncias das variáveis originais é $\sum_{i=1}^{p} \lambda_i$, igual. como se pode ver, a soma das variâncias dos componentes.
A porporção de variabilidade total explicada pela componente h é $\frac{\lambda_h}{\sum_{i=1}^{p}} \lambda_i$.

3. As covariâncias entre cada componente principal e as variáveis X vem dada pelo produto das coordenadas do autovalor que define a componente pelo seu autovalor

$$Cov(z_i; x_1, \cdots, x_p) = \lambda_i a_i = (\lambda_i a_{i1}, \cdots, \lambda_i a_{ip})$$

Em que a_i é o vetor do coeficiente da componente z_i.

Para justificar este resultado, vemos calcular a matriz $p \times p$ de covariâncias entre os componentes e as variáveis originais. Esta matriz é:

$$Cov(z, x) = \frac{1}{n}\mathbf{Z}'\mathbf{X}$$

E sua primeira linha proporciona as covariâncias entre a primeira componente e as p variáveis originais. Como $\mathbf{Z} = \mathbf{XA}$, substituindo tem-se:

$$Cov(z, x) = \frac{1}{n}\mathbf{A}'\mathbf{X}'\mathbf{X} = \mathbf{A}'\mathbf{S} = \mathbf{DA}'$$

Em que A contém em colunas os autovetores de $\mathbf{S}$ e $\mathbf{D}$ é a matriz diagonal dos autovalores. Em consequência,a covariância entre, por exemplo, o primeiro componente principal e as p variáveis serão dada pela primeira fila de $\mathbf{A}'\mathbf{S}$, ou seja, $a_i S$ ou também $\lambda_1 a'1$, em que $a'1$ é o vetor de coeficiente da primeira componente principal.

1.4.8 A busca de um modelo de análise de componentes principais

A busca de componentes $f_1, \cdots, f_m$ é semelhante à rotação dos eixos $X_1, \cdots, X_p$, uma vez que as seguintes equações são usadas e se relacionam X_i com f_j:

$$\sum_{j=1}^{m} w_i j f_j (i = 1, \cdots, n) \tag{1.1}$$

Em que x_i se denominam variáveis (critérios) w_{ij} é chamado de carga do j-ésimo componente da i-ésima variável.

A equação 1.1 pode ser expressa em forma de matriz da seguinte forma:

$$X = WF \tag{1.2}$$

Na primeira etapa de busca dos componentes principais, em vez dos eixos $f_1, \cdots, f_m$ consideramos os eixos auxiliares $y_1, \cdots, y_m$. assume-se que x_i, y_i, f_i são centrados, isto é,

$$M_{xi} = M_{yi} = M_{fi} = 0 (i = 1, \cdots, n),$$

a rotação dos eixos é expressa na forma.

$$Y = UX, \ X = U'Y, \tag{1.3}$$

onde U é uma matriz ortogonal.

Se a k-ésima linha da matriz U for denotada por u_k, temos;

$$y_i = u_{iX} = \sum_{k=1}^{n} u_{ik} x_k,$$

ou seja,

$$\lambda_i = M y_i^2.$$

Esta é a variância da variável y_i. É natural considerar que y_i não está correlacionado, ou seja, $M(y_i, y_j) = 0 \ (i \neq j)$. Por tanto,

$$\begin{aligned} M y_i y_j = M \sum_{k=1,s=1}^{n} u_{ik} \ x_k \ u_{js} \ x_s &= \sum_{k=1,s=1}^{n} u_{ik} \ M(x_k \ x_s (u)'_{sj} \\ &= \sum_{k=1,s=1}^{n} u_{ik} k_{ks} (u)'_{sj} = u_i k_x u'_j \lambda_i \sigma_{ij} \ , \end{aligned}$$

onde $k_x = \| \ k_{kj} \ \|, k_{kj} = M(x_k x_j)$ é a matriz de covariância das variáveis x.

Então temos a matriz

$$\lambda = U K_x U'$$

que é diagonal dessa forma, acrescentamos à rotação dos eixos, reduzindo assim a matriz de covarianza na forma diagonal. colocado-se os elementos $\lambda_1, \cdots, \lambda_p$ da matriz λ em ordem decrescente. Temos:

$$u_i K x U'_i = \lambda$$

A matriz U é ortogonal, isto é, $u_i u'_i = 1$, podemos escrever;

$$K_x u'_i = \lambda_i u'_i.$$

Isso significa que λ_i são os autovalores da matriz k_x e u_i são seus próprios vetores. Agora podemos encontrar os componentes principais normalizando os componentes y_i;

$$f_i = \lambda_i^{-1/2} yi = \lambda_i^{-1/2} u_i X$$

ou, na forma de matriz,

$$F = \lambda^{-1/2} Y = \lambda^{-1/2} UX.$$

daqui temos

$$X = U' \lambda^{1/2} F,$$

e para a matriz,

$$W = U' \lambda^{1/2.}.$$

Sendo que $\sum_{i=1}^{n} Dx_i = Sp[K_x] = Sp[\lambda] = \sum_{i=1}^{n} y_i$, podemos afirmar que a variância total das variáveis x_i é igual à variância total dos componentes principais não normalizados y_i. Agora é fácil encontrar a porcentagem introduzida por cada componente principal na variância total das variáveis x_i.

Geralmente esta é a principal idéia de análise de componentes principais vários componentes principais $f_1, \cdots f_m$ (m < n) o que introduz uma contribuição suficientemente grande na variância total.

1.4.9 Matriz X de dados de componentes principais

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 $X_1, X_2, X_3, \cdots, X_p$. A matriz de dados é de ordem $(n \times p)$ e normalmente denominada de matriz 'X'.

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

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 $X_1, X_2, X_3, \cdots, X_p$, pode ser na prática uma coisa complicada. Assim, o objetivo da análise de componentes principais é transformar essa estrutura complicada, representada pelas variáveis $X_1, X_2, X_3, \cdots, X_p$, em uma outra estrutura representada pelas variáveis $Y_1, Y_2, Y_3, \cdots, Y_p$ não correlacionadas e com variâncias ordenadas, para que seja possível comparar os indivíduos usando apenas as variáveis Y_i que apresentam maior variância. A solução é dada a partir da matriz de covariância S ou da matriz de correlação R.

1.4.10 Matriz de covariância S

A partir da matriz X de dados de ordem $(n \times p)$ podemos fazer uma estimativa da matriz de covariância $\sum$ da população Π que representaremos por S. A matriz S é simétrica e de ordem $(p \times p)$ (PEÑA, 2002).

$$S = \begin{bmatrix} Var_{(x_1)} & Cov_{(x_1,x_2)} & Cov_{(x_1,x_3)} & \cdots & Cov_{(x_1,x_p)} \\ Cov_{(x_2,x_1)} & Var_{(x_2)} & Cov_{(x_2,x_3)} & \cdots & Cov_{(x_2,x_p)} \\ Cov_{(x_3,x_1)} & Cov_{(x_3,x_2)} & Var_{(x_3)} & \cdots & Cov_{(x_3,x_p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov_{(x_p,x_1)} & Cov_{(x_p,x_2)} & Cov_{(x_p,x_3)} & \cdots & Var_{(x_p)} \end{bmatrix}$$

Fazendo as variâncias $Var_{(x_i)} = s_{ii}$ e as covariâncias $Cov_{(x_i,x_j)} = s_{ij}$, tem-se:

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{(1,p)} \\ s_{2,1} & s_{22} & s_{2,3} & \cdots & s_{2,p} \\ s_{3,1} & s_{3,2} & s_{33} & \cdots & s_{3,p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p,1} & s_{p,2} & s_{p,3} & \cdots & s_{pp} \end{bmatrix}$$

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

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

Padronização com variância 1 e média qualquer

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

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

$$\bar{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

$$Var(x_j) = \frac{\sum_{i=1}^{n} x_{ij}^2 - \left(\frac{\sum_{i=1}^{n} x_{ij}}{n}\right)^2}{n-1}$$

E o desvio-padrão seria

$$S(x_j) = \sqrt{Var(x_j)},\ j = 1, 2, \cdots, p$$

1.4.11 Cálculo a partir da matriz de dados centrados

A matriz S pode ser obtida diretamente a partir da matriz de dados centrados $\tilde{X}$, que se define como a matriz resultado do resto de sua média (PEÑA, 2002).

$$\tilde{X} = \tilde{X} - \mathbf{1}\bar{x}^{'}$$

Substituindo no vetor de médias tem-se:

$$1\bar{x} = \frac{1}{n}\tilde{X}^{'}\mathbf{1}$$

Em que $\mathbf{1}$ representa sempre um vetor de uns de dimensão adequada ao tamanho da amostra. Assim, o vetor de média pode ser escrito da seguinte forma:

$$\tilde{X} = \tilde{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^{'}\mathbf{X} = \mathbf{P}\mathbf{X}$$

Em que a matriz quadrada $\mathbf{P}$ está definida por:

$$\tilde{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{'}$$

A matriz $\mathbf{P}$ é simétrica e idempotente, ou seja, $\mathbf{PP} = \mathbf{P}$. Esta matriz tem posto $n - 1$ (é ortogonal ao espaço definido pelo vetor 1, já que $\mathbf{P1} = 0$ e projeta os dados ortogonalizados

ao espaço definido pelo vetor constante (com todas as coordenadas iguais). Então, a matriz **S** pode ser escrita da seguinte forma:

$$\mathbf{S} = \frac{1}{n}\tilde{X}'\tilde{X} = \frac{1}{n}\mathbf{X}'\mathbf{P}\mathbf{X}$$

A matriz de covariância é dividida por $n-1$ para obter um estimador justo da matriz da populacional. Assim, chama-se esta matriz de covariância corrigida com dados centrados

$$\hat{S} = \frac{1}{n-1}\tilde{X}'\tilde{X}$$

Após a padronização obtemos uma nova matriz de dados Z definida da seguinte forma:

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

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 covariância, S pode ser diferente do resultado encontrado a partir da matriz de correlação, 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:

$$det[R - \Lambda I] = 0 \; ou \; [R - \Lambda I] = 0$$

$$R = \begin{bmatrix} 1 & r_{(x_1x_2)} & r_{(x_1x_3)} & \cdots & r_{(x_1x_p)} \\ r_{(x_2x_1)} & 1 & r_(x_2x_3) & \cdots & r_{(x_2x_p)} \\ r_{(x_3x_1)} & r_{(x_3x_2)} & 1 & \cdots & r_{(x_3x_p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{(r_px_1)} & r_{(x_px_2)} & r_{(x_px_3)} & \cdots & 1 \end{bmatrix}$$

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 $[R \ - \Lambda I] = 0$ 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 $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda p$ as raízes da equação característica da matriz R ou S, então:

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

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

$$\tilde{a}_i = \left[\begin{array}{cccc} \tilde{a}_{i1} & \tilde{a}_{i2} & \cdots & \tilde{a}_{ip} \end{array} \right]$$

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:

$$\sum_{j=1}^{p} a_{ij}^2 = 1 \ (\tilde{a}_i \ . \ \tilde{a}_i = 1)$$

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

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 varinância do componente principal Y_i é igual ao valor do autovalor λ_i.

$$Var(Y_i) = \lambda_i$$

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

$$Var(Y_1) \ > \ Var(Y_2) \ > \ \cdots Var(Y_p)$$

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

$$\sum Var(X_i) = \sum \lambda_i = \sum Var(Y_i)$$

4. Os componentes principais não sâo correlacionados entre si:

$$Cov(Y_i, Y_j) = 0$$

1.4.12 Contribuição de cada componente principal

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

$$C_i = \frac{Var(Y_i)}{\sum_{n=1}^{p} Var(Y_i)} \dot{1}00 = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \dot{1}00 = \frac{\lambda_i}{S} \dot{1}00$$

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 serâo 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.

$$\frac{Var(Y_i) \ + \ldots \ Var(Y_k)}{\sum_{i=1}^{k} Var(Y_i)} 100 \geq \ 70\% \ onde \ k < \ p$$

1.4.13 Interpretação de cada componente

Antes de obter o coeficiente de correlação de cada componente com cada variável, vamos calcular a covariância entre a variável X_i e a componente Y_i. Para isso, considere o vetor amostral da variável X_i.

$$x_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \cdots \\ X_{in} \end{bmatrix}$$

Os vetores amostrais da componente principal Y_i

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{in} \end{bmatrix}$$

A covariância amostral entre X_i e Y_i será dada por:

$$Cov(X_i, Y_i) \quad = \quad \frac{1}{n} x_i^{'} y_i$$

O vetor x_i pode ser escrito em função da matriz X, utilizando o vetor de ordem p, que vamos designar por δ, que tem 1 na posição i-ésima e 0 nas posições restantes. Assim

$$x_i^{'} = \delta^{'} X = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} X_{1,1} & \cdots & X_{1,i} & \cdots & X_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{j,1} & \cdots & X_{j,i} & \cdots & X_{j,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{p,1} & \cdots & X_{p,i} & \cdots & X_{p,n} \end{bmatrix}$$

Tendo em conta que

$$z_h \quad = \quad X u_h$$

E covariância 1.4, pode-se expressar $\delta^{'} X$ da seguinte forma forma:

$$cov(X_j, Z_h) \quad = \quad \frac{1}{n} \delta^{'} X^{'} X u_h = \delta^{'} V u_h = \lambda_h \delta^{'} u_h = \lambda_h u_{hj}$$

Em consequência, pode-se calcular a correlação existente entre a variável X_j e a componente Z_h. 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 correlação entre cada X_j e o componente Y_i que está sendo interpretado. Por exemplo a correlação entre X_j e Y_1 é:

$$Corr(X_j, Y_1) = r_{xj.y1} = a_{1j} \ . \ \frac{\sqrt{Var(Y_1)}}{\sqrt{Var(X_j)}} = \sqrt{\lambda_1} \ . \ \frac{a_{1j}}{\sqrt{Var(X_j)}}$$

Para comparar a influência de $X_1, X_2, \cdots, X_p$ sobre Y_1 analisamos o peso ou *loading* de cada variável sobre o componente Y_1. O peso de cada variável sobre um determinado componente é dado por:

$$W_1 = \frac{a_{11}}{\sqrt{Var(X_1)}}, W_2 = \frac{a_{12}}{\sqrt{Var(X_2)}}, ...W_p = \frac{a_{1p}}{\sqrt{Var(X_p)}},$$

sendo w_1 o peso de X_1.

1.4.14 Gráfico de cotovelo

O Scree-Plot (gráfico de cotovelo) foi proposto por (CATTELL, 1966), é um gráfico dos autovalores , j, em funcão da ordem das CP's, representando graficamente a porcentagem de variância explicada por componente. Quando esta porcentagem se reduz e a curva passa a ser quase paralela ao eixo das abscissas, podemos excluir os componentes correspondentes.

Critério de (KAISER, 1958): Incluir apenas as CP's cujos autovalores são superiores ou iguais na média dos autovalores. Se a for feita a partir de uma matriz de correlações, reter as CP's com variâncias 1.

Para cada uma das CP's escolhidas pode-se calcular os escores de cada elemento amostral. Esses escores podem ser analisados utilizando-se técnicas estatísticas usuais como análise de variância e análise de regressão, dentre outras. No caso da análise multivariada, a ACP serve para auxiliar na escolha das variáveis independentes.

1.4.15 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.

Nas Equações 1.4 é exemplificado a organização de um conjunto de dados composto por p variáveis e k componentes principais.

$$Y_1 \quad = \quad a_{11}X_{11} + a_{12}X_{12} + a_{1p}X_{1p}$$

$$
\begin{aligned}
Y_2 &= a_{11}X_{21} + a_{12}X_{22} + a_{1p}X_{1p} \\
\vdots &= \vdots \\
Y_k &= a_{11}X_{n1} + a_{12}X_{n2} + a_{1p}X_{np}
\end{aligned}
$$

2 Resultados e Discussão

Utilizando dos procedimentos descritos em análise de componentes principais poderemos obter outliers multivariados que serão identificados como individuos e retirados da análise. Removeu-se o efeito de grupos dos dados originais e em seguida a matriz de variâncias e covariâncias amostrais, S, foi calculada.

Tabela 2 – Componentes principais (VCP) e Autovalores da variância explicada (%) e da variância acumulada (%)

Cp	Autovalores	Var. explicada (%)	Var. Explicada acumulada (%)
c_1	6.300020969	78.75026212	78.75026
c_2	0.757563809	9.46954761	88.21981
c_3	0.605591083	7.56988854	95.78970
c_4	0.239616717	2.99520896	98.78491
c_5	0.060918375	0.76147969	99.54639
c_6	0.027645423	0.34556779	99.89195
c_7	0.007326482	0.09158103	99.98354
c_8	0.001317141	0.01646426	100.00000

De acordo com a expressão denotada por cp (1), observa-se na Tabela 2 o percentual de variância explicada na primeira componente principal é $(6.300020969/8).100 = 78.750262\%$ o autovalor foi dividido por 8, pois este número corresponde ao traço da matriz de correlação, onde a diagonal principal é formada por valores iguais a 1. Após a extração dos autovalores e percentual da variância explicada, é necessário decidir-se pelo número de fatores a serem retirados para análise. Para isso, utiliza-se o método gráfico sugerido por (CATTELL, 1966), tal como fora mencionado anteriormente.

Através do exame do gráfico dos autovalores disposto na Figura 1, observa-se uma queda menos acentuada entre a segunda e a terceira componentes. Os critérios descritos indicam que apenas duas primeiras componentes são suficientes para explicar a maior parte da variação total dos dados, ou seja, podem substituir nas variáveis originais em analises subseqüentes (se for de interesse do pesquisador).

Serão consideradas apenas as duas primeiras componentes são suficientes para explicar a maior parte da variação total dos dados, ou seja, elas podem substituir as variáveis originais em analises subseqüentes (se for de interesse do pesquisador).

Observando que autovalores são representados em ordem decrescente no eixo das abscissas. As quatro primeiras componentes explicam 98,78% da variância total, havendo uma estabilização do gráfico após a quinta componente, sendo considerado neste trabalho duas dimensões. Pode-se observar, também, que as outras componentes apresentam uma baixa explicação, não sendo aconselhável um descarte de variáveis.

Figura 1 – Os valores explicados dos componentes principais de Autovalores e suas Respectivas porcentagens

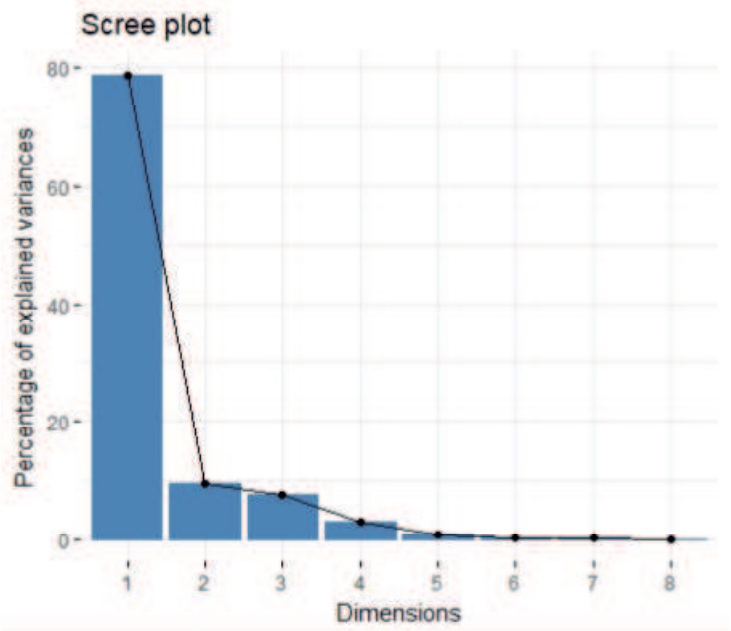
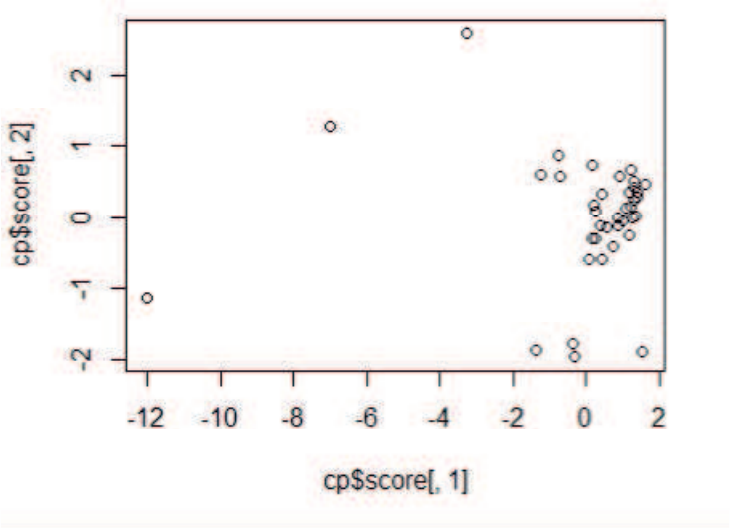


Figura 2 – Os scores considerando duas componentes principais em duas dimensões

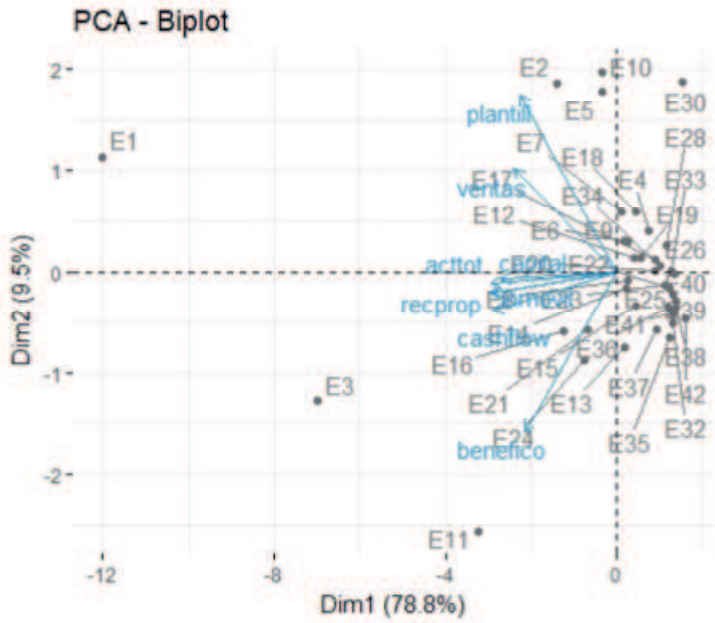


Na Figura 2 verifica-se que os estados estão distribuídos de acordo com sua representatividade em relação à grau de correlação entre as variáveis. Os estados que estão mais afastados da origem são os que melhor contribuem para representação dos nomes de cada componente.

Na Figura 3 observa-se a dispersão das variáveis e das empresas simultaneamente,

pode-se concluir que as variáveis actol, recprop, capital e Bimobil explica melhor a primeira dimensão, enquanto que as variáveis:, ventas y benefico contribuem melhor para denominar a segunda componente principal.

Figura 3 – Gráfico da dispersão das variáveis

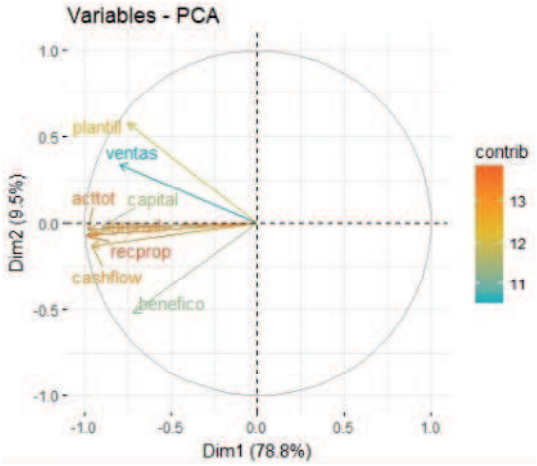


A primeira componente principal CIfra de vendas (ventas) explica 78,8% e a segunda Capital social (capital) 9,5 % da variação total. As variáveis primeira componente principal, 6,3000 e a segunda 0.7575 respectivamente; a variável ganho em Rentabilidade neto (inmovol) praticamente não contribui em nada, pois o seu valor é muito baixo, 0,001317, ou seja menos de 1%. Assim, a primeira componente pode ser interpretada como um índice de desempenho global das empresas. Como os ganhos são positivos, quando maior for o ganho bruto e o patrimônio da empresa, maior será a variaçâo em relação as outras variáveis.O componente principal de melhor desempenho global das empresas observadas. São as empresas E1, E21 e E3 respectivamente tiveram os melhores índices de desempenho, respectivamente, enquanto a empresa que ficou com o pior índice foi a E40.

Na Figura 4 tem-se o mapa fatorial mais importante, sobrepondo-se às projeções dos pontos objetos (empresas) a pontos variáveis. Observa-se que todas as variáveis com qualidade de representação maior ou igual a 0,5 foram projetadas no mapa. As variáveis de maior correlação estão mais próximo a circunferência. Mediante esta figura verifica-se que as variáveis, que melhor representam a primeira componente em relação a segunda componente, são aquelas que estão bem próximas ao círculo unitário.

Esses resultados poderão vir a contribuir para o estudo de novas pesquisas na area, na area acadêmica e difundir novos métodos de cultivos e novas e idéias na análise de componentes principais.

Figura 4 – Gráfico da distribuição das variáveis



3 Conclusão

Diante do estudo realizado, podo-se concluir que:

O uso da técnica multivariada de análise de componentes principais pode auxiliar bastante o pesquisador na construção de novas de decisões a serem tomadas, baseando-se em informação de mais de uma característica. Na decisão pela melhor solução, recomenda-se que o pesquisador avalie a qualidade dos agrupamentos obtidos, compare as variâncias internas dos blocos e a variância total.

Foi observado que a primeira componente principal explicou (78.8% da variação explicada, enquanto que a segunda componente principal $9,5\%$ da variação.

As variáveis que mais contribuíram para a primeira componente principal foram CIfra de vendas (ventas) e Capital social (capital). Enquanto, a segunda componente principal teve contribuição das variáveis ganho em Rentabilidade neto (inmovol). Assim, a primeira componente pode ser interpretada como um índice de desempenho global das empresas. Como os ganhos são positivos, quando maior for o ganho bruto e o patrimônio da empresa, maior será a variaçâo em relação as outras variáveis.

As empresas E1, E21 e E3 respectivamente tiveram os melhores índices de desempenho, respectivamente, enquanto a empresa E40 ficou com o pior índice.

Considerando as 40 empresas maiores da espanha e variáveis métricas consideradas observou-ser que o uso da técnica de componente principais é bastante eficaz, uma vez houve uma redução de dimensão de maneira expressiva para duas componentes principais, explicando mais de 88.21% da variância total.

Referências

CATTELL, R. B. The scree test for the number of factors. *Multivariate behavioral research*, Taylor & Francis, v. 1, n. 2, p. 245–276, 1966. Citado 2 vezes nas páginas 32 e 34.

FERREIRA, D. F. *Estatística multivariada*. [S.l.]: Editora UFLA, 2008. Citado na página 13.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, Warwick & York, v. 24, n. 6, p. 417, 1933. Citado na página 14.

JOHNSON, R.; WICHERN, D. *Multivariate statistics, a practical approach*. [S.l.]: Chapman & Hall Boca Raton, 1988. Citado 3 vezes nas páginas 12, 13 e 15.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, Springer, v. 23, n. 3, p. 187–200, 1958. Citado na página 32.

KENDALL. *anse multivariada*. Segunda edi. [S.l.], 1980. Citado na página 15.

KENDALL, M. G. *Multivariate Analysis*. London: Griffin, 1975. Citado 2 vezes nas páginas 12 e 14.

MARDIA, K. Linear-circular correlation coefficients and rhythmometry. *Biometrika*, JSTOR, p. 403–405, 1976. Citado na página 14.

MARDIA K. V.; KENT, J. T. B. J. M. *Multivariate Analysis*. London: Academic, 1979. Citado na página 15.

MORRISON, D. et al. Surface compositions of the satellites of saturn from infrared photometry. *The Astrophysical Journal*, v. 207, p. L213–L216, 1976. Citado 2 vezes nas páginas 14 e 15.

MORRISON, D. F. *Multivariate statistical methods*. 3. ed. New York: McGraw-Hill, 1990. Citado na página 12.

PEARSON, K. Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, v. 6, n. 2, p. 566, 1901. Citado na página 14.

PEÑA, D. *Análisis de datos multivariantes*. [S.l.]: McGraw-Hill Madrid, 2002. v. 24. Citado 5 vezes nas páginas 18, 20, 22, 26 e 27.

REGAZZI, A. Anse multivariada, notas de aula inf 766, departamento de informca da universidade federal de via. v.2, p. notas de aula INF 766, 2000. Citado 4 vezes nas páginas 13, 16, 26 e 30.

URIEL, E.; MANZANO, J. A. *Análisis multivariante aplicado*. [S.l.]: Paraninfo, 2002. v. 76. Citado na página 16.