Introdução

Considere o seguinte cenário:

- Um analista financeiro está interessado em determinar a "saúde financeira" das firmas de uma determinada indústria. Foi feita uma pesquisa que permitiu identificar 120 variáveis financeiras que poderiam ser usadas para levar a cabo tal propósito. Obviamente, seria intratável interpretar 120 indicadores financeiros para ter acesso à "saúde financeira" de uma firma. O trabalho do analista será simplificado se estas 120 variáveis poderem ser reduzidas a um nº inferior de novas variáveis.

Tal como no exemplo anterior, em muitos estudos o nº de variáveis consideradas é demasiado grande para ser tratável, tornando-se, muitas vezes, absolutamente necessário reduzir a dimensão da análise para que a situação se torne compreensível, isto é, torna-se necessário usar uma **técnica de redução de dados**.

A **ANÁLISE FACTORIAL** (ou Análise de Factores Comuns) e a **ANÁLISE DE COMPONENTES PRINCIPAIS** são técnicas estatísticas cujo objectivo é representar ou descrever um número de variáveis iniciais a partir de um menor número de variáveis hipotéticas (os factores \ componentes principais). Isto é, permite identificar novas variáveis (os factores \ componentes principais), em menor número que o conjunto inicial, mas sem perda significativa da informação contida neste conjunto.

O propósito geral destas técnicas é encontrar uma maneira de condensar (sumariar) a informação contida num conjunto de variáveis originais, num conjunto menor de variáveis perdendo o mínimo possível de informação. Tratam-se portanto de técnicas de redução de dados que investigam os inter-relacionamentos (correlações) entre as variáveis e os descrevem, se possível, em termos de um menor número de variáveis chamadas **factores \ componentes principais**.

A Análise de Componentes Principais é considerada por muitos autores um dos muitos tipos de Análise Factorial. É de salientar, no entanto, que apesar das várias tentativas para esclarecer o assunto, ainda existe muita confusão no que diz respeito à distinção entre Análise Factorial e Análise de Componentes Principais. Uma das razões que poderá contribuir para tal, é o facto de que, em muitos packages estatísticos (como por exemplo o SPSS), a Análise de Componentes Principais pode ser levado a cabo como um procedimento de Análise Factorial.

De facto, embora ambos os tipos de análise permitam uma redução de dados, a Análise Factorial está mais preocupada em explicar a estrutura de covariâncias entre as variáveis. Contrariamente, o objectivo da Análise de Componentes Principais, não é explicar as correlações entre as variáveis mas apenas encontrar combinações lineares das variáveis iniciais que expliquem o máximo possível da variação existente nos dados e os permitam descrever e reduzir.

A **ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)** constitui um método estatístico multivariado que permite transformar um conjunto de variáveis iniciais correlacionadas entre si, num outro conjunto de variáveis não correlacionadas (independentes / ortogonais), as chamadas **componentes principais**, que resultam de combinações lineares do conjunto inicial.

O propósito desta análise é determinar as componentes principais de forma a explicar o mais possível da variação total dos dados com o menor número possível de componentes.

As **componentes principais** são calculadas por ordem decrescente de importância, isto é, a primeira explica o máximo possível da variância dos dados originais, a segunda explica o máximo possível da variância ainda não explicada, e assim por diante. A última componente principal será a que menor contribuição dá para a explicação da variância total dos dados originais. Porque cada combinação linear explica o máximo possível da variância não explicada e terá de ser ortogonal a qualquer outra combinação já definida, o conjunto de todas as combinações encontradas constitui uma solução única.

Análise de Componentes Principais

A ACP é uma técnica de análise exploratória multivariada que transforma um conjunto de variáveis correlacionadas num conjunto menor de variáveis independentes, combinações lineares das variáveis originais, designadas por componentes principais.

Descrita desta forma, a ACP é geralmente encarada como um método de redução dos dados mas, para alem deste objectivo, uma das principais vantagens da ACP é permitir resumir a informação de várias variáveis correlacionadas (e portanto de alguma forma redundantes) em uma ou mais combinações lineares independentes (as componentes principais) que representem a maior parte da informação presente nas variáveis originais.

Adicionalmente, as componentes principais podem ser utilizadas em análises posteriores, nomeadamente em técnicas estatísticas (por exemplo, regressão linear múltipla) que exigem que as variáveis em estudo sejam independentes.

Exemplo 1: O exemplo seguinte é um exemplo simples em que se considera apenas uma componente principal.

Suponha que conhecíamos o peso e a altura de 10 indivíduos e que, com estes dois indicadores descrevíamos a estatura física de cada um deles. Poder-se-ía, no entanto, descrever esta mesma estatura física utilizando apenas uma variável que estivesse relacionada com os indicadores iniciais, por exemplo, de uma forma linear:

$$\text{ESTATURA} = \alpha\ \text{Altura} + \beta\ \text{Peso}$$

em que α e β indicariam a contribuição, respectivamente, da altura e do peso para a estatura física de cada indivíduo.

Passamos então a descrever a estatura física de um indivíduo com um valor apenas, resultante da combinação linear da sua altura e peso, perdendo alguma da informação inicial, mas ganhando em termos de simplificação e de uma compreensão mais imediata do aspecto físico de cada indivíduo.

Esta simplificação é muito útil quando num determinado estudo existem dezenas de indicadores a considerar.

(Reis, E. (1993). Análise factorial das componentes principais: um método de reduzir sem perder informação, Temas em Métodos Quantitativos para Gestão nº2, Giesta – ISCTE)

Exemplo 2: Suponhamos que estamos interessados em medir o nível de "performance" em matemática dos alunos de uma certa escola. Para isso precisaríamos apenas de registar as notas em matemática desses alunos, isto é, necessitaríamos apenas de considerar uma característica de cada aluno.

Se, em vez disso, quisermos medir a "performance" global dos alunos, necessitamos de seleccionar várias características tais como: Inglês, História, Educação Física, Educação Visual, Geografia, Português, etc.

Estas características, embora estejam relacionadas umas com as outras, podem não conter a mesma quantidade de informação, e de facto algumas características podem ser completamente redundantes.

Se eventualmente estivermos interessados em "explicar" as diferenças entre alunos, temos de seleccionar as características que discriminam verdadeiramente um aluno de outro e rejeitar as que não têm poder discriminatório, o que não é uma tarefa fácil. Alternativamente, poderíamos aplicar a ACP para determinar combinações lineares das características seleccionadas **- as componentes principais**.

Poderia acontecer que grande parte da variação de aluno para aluno residisse apenas em 3 componentes principais. Poderíamos então direccionar o nosso estudo para estas 3 quantidades; as outras componentes principais variam tão pouco de um aluno para outro, que o estudo delas diria pouco acerca da variação individual.

Exemplo 3: Este exemplo é uma aplicação de um estudo desenvolvido por P. Doyle e J. Saunders (1985) a uma empresa industrial - a Boliet - cuja actividade principal era o processamento de pasta a partir de resina de pinheiro, que era posteriormente vendida como matéria prima a fabricantes de papel e resinas sintéticas. A especialização da Boliet visava fundamentalmente o processamento de produtos derivados de resina que lhes abriria um importante mercado especializado consumidor de colas industriais.

A pesquisa feita envolveu a recolha de informação quer ao nível dos consumidores, quer das empresas concorrentes, e permitiu identificar 6 variáveis específicas que afectavam a escolha do produto por parte do consumidor e 4 variáveis afectas às empresas:

1- suavidade	1- distância ao fornecedor
2- viscosidade	2- serviço de apoio
3- estabilidade da cor	3- reputação
4- cor inicial	4- cobertura geográfica
5- aderência	
6- preço	

Foi sobre este conjunto de 10 variáveis que se aplicou a ACP, tendo-se extraído **4 factores** explicando 78% da variância inicial e interpretados do seguinte modo:

FACTOR I: Descreve o poder do fornecedor;
FACTOR II: Campo de aplicação do produto;
FACTORES III e **IV**: Características técnicas que definam o uso do produto.

(Reis, E. (1993). Análise factorial das componentes principais: um método de reduzir sem perder informação, Temas em Métodos Quantitativos para Gestão nº2, Giesta – ISCTE)

Exemplo 4: Como é que os consumidores avaliam os bancos? Foi pedido aos consumidores inquiridos que classificassem a importância de 15 atributos bancários. Foi usada uma escala de 1 a 5 pontos, onde 1 significa não importante e 5 significa muito importante. Os dados foram analisados através da análise de components principais.

A solução resultou em 4 factores, que foram designados por serviços tradicionais, conveniência, visibilidade e competência.

Os serviços tradicionais incluem taxas interessantes em empréstimos, reputação na comunidade, preços baixos nos serviços bancários, atendimento personalizado, extractos mensais de leitura fácil e facilidade na obtenção de empréstimos.

Conveniência inclui localização dos balcões, localização de caixas multibanco, rapidez do serviço e horário conveniente do banco.

O factor visibilidade inclui recomendações dos amigos e familiares, estrutura física atraente, comunidade envolvente e facilidade na obtenção de empréstimos.

Competência consiste na competência dos empregados e as capacidades demonstradas nos serviços auxiliares do banco.

Conclui-se que os consumidores avaliam os bancos usando os 4 factores básicos acima referidos e os bancos devem ser excelentes nestes factores para projectarem uma boa imagem.

Apostando nestes factores o banco JPMorgan Chase & Co. tornou-se o segundo maior banco Americano com lucro de 2,26 biliões de dolares no 1º trimestre de 2005, valor esse 17% superior aos 1,93 biliões de dolares obtidos no mesmo período em 2004.

(Malhotra, N.K. (2006). Marketing Research: An Applied Orientation, 5º Edition, Person Prentice Hall, New Jersey)

Exemplo 5: O índice PSI20 é um exemplo de uma combinação linear das cotações na bolsa de valores das 20 empresas com maior volume de negócios. As vantagens de tal índice são óbvias: é claramente mais fácil para o analista de mercados avaliar a evolução do mercado através de um índice do que com 20 variáveis que registam a cotação de 20 empresas.

(Maroco, J. (2003). Análise Estatística – Com utilização do SPSS, Edições Sílabo, Lisboa)

PRELIMINARES

Para estudar as relações entre duas variáveis aleatórias X e Y pode-se analisar a covariância e o coeficiente de correlação linear.

A **covariância entre X e Y** representa-se por **Cov(X,Y)** ou $\sigma_{X,Y}$, e define-se por:

$$Cov(X,Y)=E[(X - \mu_X)(Y - \mu_Y)]$$

A covariância descreve a relação linear ou ligação entre duas variáveis e a sua mútua dependência, fornecendo-nos uma indicação do modo como X e Y variam uma relativamente à outra.

A covariância está expressa nas unidades de X e nas de Y, simultaneamente, o que por vezes introduz algumas dificuldades. Para ultrapassar esta situação, pode calcular-se o coeficiente de correlação linear entre X e Y.

O **coeficiente de correlação linear entre X e Y** representa-se por $\rho_{X,Y}$, e define-se por:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X.\sigma_Y}$$

Verifica-se que $-1 \leq \rho_{X,Y} \leq 1$.

MODELO 1:

Componentes principais obtidas a partir da matriz de covariâncias ($\sum$)

Seja $X^T = [X_1 \quad X_2 \quad \ldots \quad X_p]$ o vector das variáveis aleatórias observadas, com média $\mu = [\mu_1 \quad \mu_2 \quad \ldots \quad \mu_p]^T$ e matriz de covariâncias $\sum$.

Queremos encontrar as componentes principais Y_1, Y_2,..., Y_p:

$$Y_j = a_j^T X = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p \Leftrightarrow Y = P^T X$$

(onde $P = [a_1 \quad a_2 \quad \ldots \quad a_p]$)

De forma a que :

- Y_1, Y_2,..., Y_p sejam não correlacionadas entre si;
- $Var(Y_1) \geq Var(Y_2) \geq \ldots \geq Var(Y_p)$.

Solução única:

- a_i é o vector próprio normalizado associado a λ_i;
- λ_i é o i-ésimo maior valor próprio da matriz de covariâncias $\sum$.

Propriedades:

Sejam

$$D = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_p \end{bmatrix} \quad \text{e} \quad C = \left[\sqrt{\lambda_1}a_1 \quad \dots \quad \sqrt{\lambda_p}a_p\right] = PD^{\frac{1}{2}}$$

Temos

- $Var(Y_j) = \lambda_j$; $E(Y_j) = a_j^T E(X)$ e $Var(Y_1) \geq Var(Y_2) \geq \dots$
- $Cov(Y_i, Y_j) = 0$, para $(i \neq j)$

- $Cov(X_i, Y_j) = \lambda_i a_{ij}$ logo $\rho_{X_i,Y_j} = \dfrac{\sqrt{\lambda_i}a_{ij}}{\sigma_{X_i}}$

(loading da variável X_i na componente Y_j)

- $\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(X_i)$, daqui sai que $\dfrac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$ é a proporção da variância total explicada pela componente Y_j
- $\Sigma = CC^T$
- $Var(X_i) = \sum_{j=1}^{p} \lambda_j a_{ij}^2 = \underline{(\lambda_1 a_{i1}^2 + \dots + \lambda_k a_{ik}^2)} + \lambda_{k+1} a_{ik+1}^2 + \dots + \lambda_p a_{ip}^2 = \underline{h_i} + \lambda_{k+1} a_{ik+1}^2 + \dots + \lambda_p a_{ip}^2$

(h_i - comunalidade = porção da $Var(X_i)$ explicada pelas primeiras *k* componentes)

MODELO 2:

Componentes principais obtidas a partir da matriz de correlações (ρ)

Sejam $X_1^{'} = \frac{X_1 - \mu_1}{\sigma_{X_1}}$, $X_2^{'} = \frac{X_2 - \mu_2}{\sigma_{X_2}}$,..., $X_p^{'} = \frac{X_p - \mu_p}{\sigma_{X_p}}$ as variáveis aleatórias observadas estandardizadas, com matriz de correlações ρ.

Queremos encontrar as componentes principais Y_1, Y_2,..., Y_p:

$$Y_j = a_j^T X^{'} = a_{1j} \frac{X_1 - \mu_1}{\sigma_1} + a_{2j} \frac{X_2 - \mu_2}{\sigma_2} + ... + a_{pj} \frac{X_p - \mu_p}{\sigma_p} \Leftrightarrow Y = P^T X^{'}$$

(onde X' é o vector das variáveis estandardizadas)

De forma a que :

- Y_1, Y_2,..., Y_p sejam não correlacionadas entre si;
- $Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p)$.

Solução única:

- a_i é o vector próprio normalizado associado a λ_i;
- λ_i é o i-ésimo maior valor próprio da matriz de correlações ρ.

Propriedades:

Sejam

$$D = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_p \end{bmatrix} \quad \text{e} \quad C = \begin{bmatrix} \sqrt{\lambda_1}a_1 & \dots & \sqrt{\lambda_p}a_p \end{bmatrix} = PD^{\frac{1}{2}}$$

Temos

- $Var(Y_j) = \lambda_j$; $E(Y_j) = 0$ e $Var(Y_1) \geq Var(Y_2) \geq \dots$
- $Cov(Y_i, Y_j) = 0$, para $(i \neq j)$

- $Cov(X_i', Y_j) = \lambda_i a_{ij}$ logo $\rho_{X_i', Y_j} = \sqrt{\lambda_i}a_{ij}$

  (loading da variável X_i' na componente Y_j)

- $\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(X_i') = p$, daqui sai que $\frac{\lambda_j}{p}$ é a proporção da variância total das variáveis estandardizadas explicada pela componente Y_j

- $\rho = CC^T$

- $1 = Var(X_i') = \sum_{j=1}^{p} \lambda_j a_{ij}^2 = \underline{\lambda_1 a_{i1}^2 + \dots + \lambda_k a_{ik}^2} + \lambda_{k+1} a_{ik+1}^2 + \dots + \lambda_p a_{ip}^2 = \underline{h_i} + \lambda_{k+1} a_{ik+1}^2 + \dots + \lambda_p a_{ip}^2$

  (h_i - comunalidade = porção da $Var(X_i')$ explicada pelas primeiras *k* componentes)

NOTA: Na prática, em geral, não são conhecidas as matrizes Σ e ρ, por isso temos que usar estimativas.

Estimativa de Σ $S = (s_{jk})$ matriz $(p \times p)$ onde $s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \overline{x_j}\right)\left(x_{ik} - \overline{x_k}\right)$

Estimativa de ρ $R = DSD$ onde $D = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$

Exemplo (exercício 1):

A tabela seguinte apresenta os preços (em unidades monetárias - u.m.) de 5 produtos alimentares em 23 cidades.

Tabela I

Cidades	Pão X_1	Hambúrguer X_2	Leite X_3	Laranjas X_4	Tomates X_5
1	24,50	94,50	73,90	80,10	41,60
2	26,50	91,00	67,50	74,60	53,30
3	29,70	100,80	61,40	104,00	59,60
4	22,80	86,60	65,30	118,40	51,20
…	…	…	…	…	…

As médias e as variâncias amostrais das 5 variáveis são as seguintes:

Tabela II

	Média	Desvio padrão	Variância	% de variância total
Pão - X_1	25.2913	2.507	6.284	1.688
Hambúrguer - X_2	91.8565	7.555	57.077	15.334
Leite - X_3	62.2957	6.95	48.306	12.978
Laranjas - X_4	102.9913	14.239	202.756	54.472
Tomates - X_5	48.7652	7.603	57.801	15.528
Total			372.224	100

Suponhamos que estamos interessados em formar uma medida do “Consumer Price Index” (CPI), isto é, estamos interessados em formar uma soma ponderada dos preços dos vários produtos alimentares, que nos dê uma indicação de quão caros ou baratos são os produtos alimentares, em geral, numa dada cidade. A análise de componentes principais é uma técnica apropriada para desenvolver tal tarefa.

Os valores próprios da matriz amostral de covariâncias (S) são:

λ_1=218.999 λ_2=91.723 λ_3=37.663 λ_4=20.811 λ_5=3.029

Os vectores próprios normalizados associados aos dois primeiros valores próprios são respectivamente:

$$a_1=\begin{bmatrix}0.028\\0.2\\0.042\\0.939\\0.276\end{bmatrix} \qquad a_2=\begin{bmatrix}0.165\\0.632\\0.442\\-0.314\\0.528\end{bmatrix}$$

Os valores próprios da matriz amostral de correlações (R) são:

λ_1=2.42247 λ_2=1.10467 λ_3=0.73848 λ_4=0.49361 λ_5=0.24077

Os vectores próprios normalizados associados aos dois primeiros valores próprios são respectivamente:

$$a_1=\begin{bmatrix}0.496\\0.576\\0.34\\0.225\\0.506\end{bmatrix} \qquad a_2=\begin{bmatrix}-.309\\-.044\\-.43\\0.797\\0.287\end{bmatrix}$$

No que se segue, vamos assumir que apenas a primeira componente principal é usada como medida do CPI.

Tabela III

Cidades	Coluna 1	Coluna 2	Coluna 3	Coluna 4
1	109,3560	-1,51881	-,2272	-,14598
2	106,5064	-1,71137	,2817	,18099
3	137,6432	,39267	2,2480	1,44431
4	145,9721	,95549	-,3412	-,21921
…	…	…	…	…

Coluna 1 - scores da 1ª componente principal obtida a partir dos dados da tabela I

Coluna 2 - são os scores da coluna 1 estandardizados, que são obtidos subtraindo os scores da coluna 1 pela média da 1ª componente principal e dividindo pelo seu desvio padrão.

Coluna 3 - scores da 1ª componente principal obtida a partir dos dados estandardizados, isto é, obtida a partir da matriz amostral de correlações (R).

Coluna 4 - são os scores da coluna 3 estandardizados, que são obtidos subtraindo os scores da coluna 3 pela média da 1ª componente principal e dividindo pelo seu desvio padrão.

Como exemplo dos modelos apresentados, vamos responder às alíneas a) e b) do exercício 1.

a) Considerando os dados na sua forma original:

(i) Determine as expressões para as duas primeiras componentes principais.

$$Y_1 = 0{,}028X_1 + 0{,}2X_2 + 0{,}042X_3 + 0{,}939X_4 + 0{,}276X_5$$
$$Y_2 = 0{,}165X_1 + 0{,}632X_2 + 0{,}442X_3 - 0{,}314X_4 + 0{,}528X_5$$

(ii) Determine a percentagem de variância total explicada pela 1ª componente principal, pela 2ª componente principal e pelas 3 últimas componentes principais.

Sabemos que $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = \sum_{i=1}^{5} Var(X_i) = 372{,}224$

de facto $\sum_{i=1}^{5} \lambda_i = 218{,}999 + 91{,}723 + 37{,}663 + 20{,}811 + 3{,}029 = 372{,}225$

% de variância total explicada por Y_1 = $\frac{\lambda_1}{\sum_{i=1}^{5} \lambda_i} \times 100\% = \frac{218{,}999}{372{,}224} \times 100\% = 58{,}84\%$

% de variância total explicada por Y_2 = $\frac{\lambda_2}{\sum_{i=1}^{5}\lambda_i} \times 100\% = \frac{91{,}723}{372{,}224} \times 100\% = 24{,}64\%$

% de variância total explicada por Y_3, Y_4 e Y_5 =

$$= \frac{\lambda_3 + \lambda_4 + \lambda_5}{\sum_{i=1}^{5}\lambda_i} \times 100\% = \frac{37{,}663 + 20{,}811 + 3{,}029}{372{,}224} \times 100\% = 16{,}52\%$$

ou

$$= 100\% - 24{,}64\% - 58{,}84\% = 16{,}52\%$$

(iii) Determine os scores das duas primeiras componentes principais para a cidade 1.

$$Y_1^1 = 0{,}028 \times 24{,}5 + 0{,}2 \times 94{,}5 + 0{,}042 \times 73{,}9 + 0{,}939 \times 80{,}1 + 0{,}276 \times 41{,}6 = 109{,}3853$$

$$Y_2^1 = 0{,}165 \times 24{,}5 + 0{,}632 \times 94{,}5 + 0{,}442 \times 73{,}9 - 0{,}314 \times 80{,}1 + 0{,}528 \times 41{,}6 = 93{,}2437$$

(iv) Diga quais as variáveis que mais influenciam a 1ª componente principal.

Loadings na componente principal Y_1:

$$\rho_{Y_1,X_1} = \frac{\sqrt{\lambda_1}a_{11}}{\sigma_{X_1}} = \frac{\sqrt{218{,}999}\times 0{,}028}{2{,}507} = 0{,}17$$

$$\rho_{Y_1,X_2} = \frac{\sqrt{\lambda_1}a_{21}}{\sigma_{X_2}} = \frac{\sqrt{218{,}999}\times 0{,}2}{7{,}555} = 0{,}39$$

$$\rho_{Y_1,X_3} = \frac{\sqrt{\lambda_1}a_{31}}{\sigma_{X_3}} = \frac{\sqrt{218{,}999}\times 0{,}042}{6{,}95} = 0{,}089$$

$$\rho_{Y_1,X_4} = \frac{\sqrt{\lambda_1}a_{41}}{\sigma_{X_4}} = \frac{\sqrt{218{,}999}\times 0{,}939}{14{,}239} = 0{,}98$$

$$\rho_{Y_1,X_5} = \frac{\sqrt{\lambda_1}a_{51}}{\sigma_{X_5}} = \frac{\sqrt{218{,}999}\times 0{,}276}{7{,}603} = 0{,}54$$

Temos $Y_1 = 0{,}028X_1 + 0{,}2X_2 + 0{,}042X_3 + 0{,}939X_4 + 0{,}276X_5$.

Os pesos das variáveis X_i na componente principal Y_1, indicam que a 1ª componente principal é muito mais influenciada por X_4 (preço das laranjas) do que pelas outras variáveis.

De facto, pela análise dos loadings, conclui-se que é a variável X_4 que apresenta um grau de associação linear mais forte com a 1ª componente principal, sendo portanto esta a variável que mais influencia na formação dos scores de Y_1.

(v) Os scores da 1ª componente principal para cada cidade em estudo, estão registados na coluna 1 da tabela III. Por vezes os scores das componentes principais são estandardizados. A coluna 2 da tabela II apresenta os scores estandardizados, que são obtidos subtraindo os scores da coluna 1 pela média da 1ª componente principal e dividindo pelo seu desvio padrão. Tendo em conta que assumimos que apenas a 1ª componente principal é usada como medida do CPI, diga quais são as cidades mais caras e quais as mais baratas.

Temos que

média de Y_1 = $0{,}028\overline{x_1} + 0{,}2\overline{x_2} + 0{,}042\overline{x_3} + 0{,}939\overline{x_4} + 0{,}276\overline{x_5} =$

$= 0{,}028 \times 25{,}2913 + 0{,}2 \times 91{,}8565 + 0{,}042 \times 62{,}2957 + 0{,}939 \times 102{,}9913 + 0{,}276 \times 48{,}7652 = 131{,}86$

desvio padrão de Y_1 = $\sqrt{\lambda_1} = 74{,}8$

Tendo em conta que assumimos que apenas a 1ª componente principal é usada como medida do CPI, analisando a coluna 2 da tabela III, concluímos que as cidades mais caras são 10, 4 e 18 (por ordem decrescente do CPI) e as cidades mais baratas são 2, 13 e 1 (por ordem crescente de CPI).

b) <u>Considerando os dados estandardizados:</u>

(i) Determine as expressões para as duas primeiras componentes principais.

$$Y_1 = 0{,}496X_1' + 0{,}576X_2' + 0{,}34X_3' + 0{,}225X_4' + 0{,}506X_5'$$

$$Y_2 = -0{,}309X_1' - 0{,}044X_2' - 0{,}43X_3' + 0{,}793X_4' + 0{,}287X_5'$$

onde $X_1' = \dfrac{X_1 - 25{,}2913}{2{,}507}$, $X_2' = \dfrac{X_2 - 91{,}8565}{7{,}555}$, $X_3' = \dfrac{X_3 - 62{,}2957}{6{,}95}$,

$X_4' = \dfrac{X_4 - 102{,}9913}{14{,}239}$ e $X_5' = \dfrac{X_5 - 48{,}7652}{7{,}603}$

(ii) Determine a percentagem de variância total explicada pela 1ª componente principal, pela 2ª componente principal e pelas 3 últimas componentes principais.

Temos que $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = \sum_{i=1}^{5} Var(X_i^{'}) = 5$.

% de variância total explicada por Y_1 = $\frac{\lambda_1}{5} \times 100\% = \frac{2,42247}{5} \times 100\% = 48,45\%$

% de variância total explicada por Y_2 = $\frac{\lambda_2}{5} \times 100\% = \frac{1,10467}{5} \times 100\% = 22,09\%$

% de variância total explicada por $Y_3, Y_4 \; e \; Y_5$ =

$$= \frac{\lambda_3 + \lambda_4 + \lambda_5}{5} \times 100\% = \frac{0,73848 + 0,49361 + 0,24077}{5} \times 100\% = 29,46\%$$

ou

$$= 100\% - 22,09\% - 48,45\% = 29,46\%$$

(iii) Determine os scores das duas primeiras componentes principais para a cidade 1.

$$Y_1^1 = 0{,}496 \times \left(\frac{24{,}5 - 25{,}2913}{2{,}507}\right) + 0{,}576 \times \left(\frac{94{,}5 - 91{,}8565}{7{,}555}\right) + 0{,}34 \times \left(\frac{73{,}9 - 62{,}2957}{6{,}95}\right) +$$
$$+ 0{,}225 \times \left(\frac{80{,}1 - 102{,}9913}{14{,}239}\right) + 0{,}506 \times \left(\frac{41{,}6 - 48{,}7652}{7{,}603}\right) = -0{,}23$$

$$Y_2^1 = -0{,}309 \times \left(\frac{24{,}5 - 25{,}2913}{2{,}507}\right) - 0{,}044 \times \left(\frac{94{,}5 - 91{,}8565}{7{,}555}\right) - 0{,}43 \times \left(\frac{73{,}9 - 62{,}2957}{6{,}95}\right) +$$
$$+ 0{,}797 \times \left(\frac{80{,}1 - 102{,}9913}{14{,}239}\right) + 0{,}287 \times \left(\frac{41{,}6 - 48{,}7652}{7{,}603}\right) = -2{,}19$$

(iv) Diga quais as variáveis que mais influenciam a 1ª componente principal.

Loadings na componente principal Y_1:

$$\rho_{Y_1,X_1'} = \sqrt{\lambda_1}\, a_{11} = \sqrt{2{,}42247} \times 0{,}496 = 0{,}77$$
$$\rho_{Y_1,X_2'} = \sqrt{\lambda_1}\, a_{21} = \sqrt{2{,}42247} \times 0{,}566 = 0{,}896$$
$$\rho_{Y_1,X_3'} = \sqrt{\lambda_1}\, a_{31} = \sqrt{2{,}42247} \times 0{,}34 = 0{,}53$$
$$\rho_{Y_1,X_4'} = \sqrt{\lambda_1}\, a_{41} = \sqrt{2{,}42247} \times 0{,}225 = 0{,}35$$
$$\rho_{Y_1,X_5'} = \sqrt{\lambda_1}\, a_{51} = \sqrt{2{,}42247} \times 0{,}506 = 0{,}79$$

Temos $Y_1 = 0{,}496X_1' + 0{,}576X_2' + 0{,}34X_3' + 0{,}225X_4' + 0{,}506X_5'$.

Podemos ver, pelos pesos da 1ª componente principal, que nenhuma das variáveis domina a formação dos scores da componente, mas as que mais influenciam os scores são as variáveis X_1, X_2 e X_5.

De facto, pela análise dos loadings, conclui-se que as variáveis X_1, X_2 e X_5 são as variáveis mais fortes associadas a Y_1, logo são estas as mais influentes na formação dos scores.

(v) Os scores da 1ª componente principal para cada cidade em estudo, estão registados na coluna 1 da tabela III. Por vezes os scores das componentes principais são estandardizados. A coluna 2 da tabela II apresenta os scores estandardizados, que são obtidos subtraindo os scores da coluna 1 pela média da 1ª componente principal e dividindo pelo seu desvio padrão. Tendo em conta que assumimos que apenas a 1ª componente principal é usada como medida do CPI, diga quais são as cidades mais caras e quais as mais baratas.

Pela análise da coluna 4 da tabela III sai que as cidades mais caras são 10, 16 e 3 (por ordem decrescente do CPI) e as cidades mais baratas são 22, 20 e 11 (por ordem crescente de CPI).

EFEITOS DAS UNIDADES DE MEDIDA

As componentes principais obtidas a partir da matriz de covariâncias Σ têm a desvantagem de não serem invariantes perante alterações nas escalas de medida das variáveis iniciais.

Quando as escalas de medida das variáveis são consideravelmente diferentes, as suas variâncias vão, também, ter valores numéricos consideravelmente diferentes, e as variáveis com maior variância vão "dominar" as primeiras componentes principais (já que as primeiras componentes principais são obtidas de forma a explicarem o máximo possível da variância total dos dados).

Por isso nesta situação deve-se estandardizar as variáveis iniciais o que corresponde a derivar as componentes principais a partir da matriz de correlações (ρ), a não ser que haja razão para crer que a variância de uma variável é um indicador da sua importância.

Como exemplo, vamos responder à alínea c) do exercício 1.

c) O que lhe parece mais adequado para o problema em questão: aplicar a análise de componentes principais aos dados originais ou aos dados estandardizados? Justifique.

Da análise feita na alínea a) deste exercício, podemos concluir que a 1ª componente principal, apesar de ser uma soma ponderada de todos os preços, é muito mais afectada pelo preço das laranjas. A razão principal do preço das laranjas dominar a formação dos scores da 1ª componente principal, é a existência de uma grande variação no preço das laranjas entre as várias cidades. De facto, a variância do preço das laranjas, X_4, é muito maior comparada com a dos preços dos outros produtos alimentares (a variável X_4 é responsável por 54,47% da variância total).

Em geral o peso de uma variável numa componente principal é afectado pela variância relativa dessa variável. Se não quisermos que esta variância relativa afecte os pesos, então os dados devem ser estandardizados.

Na alínea b), ao estandardizarmos os dados para aplicar a ACP, verificámos que nenhuma das variáveis dominava a formação dos scores da 1ª componente principal, apesar de se evidenciar uma maior influência das variáveis X_1, X_2 e X_5.

Não existe nenhuma razão para crer que alguns produtos alimentares sejam mais importantes na dieta diária do que outros. Consequentemente, ao formar o índice CPI o preço das laranjas não deve receber um peso maior devido à sua variação. Por isso deve-se estandardizar os dados antes de aplicar a ACP.

Vamos responder à alínea d) do exercício 1.

d) Mediante a resposta à alínea anterior diga qual lhe parece ser efectivamente a cidade mais cara e a mais barata.

Uma vez que chegámos à conclusão que devemos usar dados estandardizados a resposta é dada com base na alínea b) (v).
Deste modo a cidade mais cara é a cidade 10 e a mais barata é a cidade 22.

QUANTAS COMPONENTES PRINCIPAIS SE DEVEM RETER?

Quando aplicamos a análise de componentes principais com o objectivo de reduzir o nº de variáveis em estudo, esperamos que as primeiras componentes expliquem uma proporção significativa da variância total dos dados, isto é, esperamos que os dados possam ser representados por um pequeno nº de componentes principais sem que haja uma perda significativa de informação. Põe-se então uma questão: O que se entende por "perda significativa de informação"? Isto é, quantas componentes principais se devem reter?

Consideremos os seguintes exemplos:

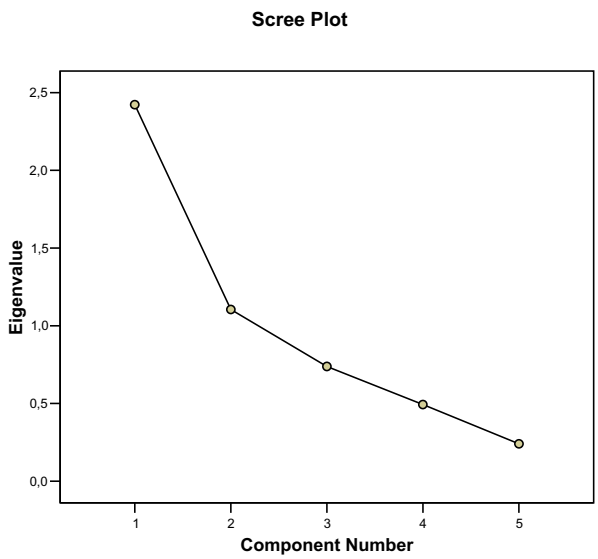
- Um grupo de cientistas tinha à sua disposição 100 variáveis para tomar uma decisão muito importante relativa a uma nave espacial. Verificaram que 5 componentes principais explicavam 99% da variação total das 100 variáveis. No entanto, dada a importância e o risco que envolvia a tomada de tal decisão, os cientistas consideraram 1% de variação não explicada (i.e. de perda de informação) como sendo uma percentagem substancial, e por isso optaram por usar as 100 variáveis que tinham à disposição para tomar a decisão.
- Suponha, agora, que as 100 variáveis representavam preços de vários produtos alimentares. Neste caso, poderá acontecer que 1% de variação não explicada seja considerada não substancial e então as 5 componentes principais poderiam ser usadas no estudo em vez das 100 variáveis.

Os exemplos anteriores ilustram que o número de componentes principais a reter numa análise depende da quantidade de informação que estamos dispostos a perder (i.e., da quantidade de variância não explicada que podemos admitir).

Existem, no entanto, várias regras práticas para determinar quantas componentes excluir da análise:

1. Reter as componentes suficientes para explicar 80 a 90 % da variância total.
2. Excluir as componentes cujos valores próprios são inferiores à média. No caso da análise ser feita a partir da matriz de correlações devemos excluir as componentes cujos valores próprios são inferior a 1 (**critério de Kaiser**).

3. Representar graficamente a percentagem de variância explicada por cada componente principal. Quando esta percentagem se reduz e a curva passa a ser quase paralela ao eixo das abcissas, são de excluir as componentes correspondentes (**Scree-test**).



Como exemplo, vamos responder à alínea e) do exercício 1.

e) Utilizando o critério de Kaiser diga quantas componentes principais deveriam ter sido retidas e usadas para medir o CPI.

Consideremos as componentes principais obtidas a partir dos dados estandardizados.

Temos $\lambda_1 > 1,\ \lambda_2 > 1$ e $\lambda_3 < 1,\ \lambda_4 < 1,\ \lambda_5 < 1$.

Logo devem ser retidas as duas primeiras componentes principais.

ALGUNS ASPECTOS IMPORTANTES DA INTERPRETAÇÃO DO SPSS

Quando se apresentam os resultados de análise de componentes principais é vulgar apresentar em vez dos vectores próprios a_j os seus transformados:

$$a_j^* = \lambda_j^{1/2} a_j = \sqrt{\lambda_i}\, a_j$$

Note que, enquanto para os vectores próprios a_j tínhamos $a_j^T a_j = 1$, agora temos $a_j^{*T} a_j^* = \sum_{i=1}^{p} a_{ij}^{*2} = \lambda_j$, isto é, a soma dos quadrados dos elementos de a_j^* é igual a λ_j.

O output do SPSS fornece os transformados a_j^* em vez dos vectores próprios a_j.

A matriz que contém os transformados a_j^* é designada, no output do SPSS, por "*Component Matrix*".

$$C = \left[a_1^* \quad a_2^* \quad \ldots \quad a_p^* \right] = \begin{bmatrix} \sqrt{\lambda_1} a_{11} & \sqrt{\lambda_2} a_{12} & \ldots & \sqrt{\lambda_p} a_{1p} \\ \sqrt{\lambda_1} a_{21} & \sqrt{\lambda_2} a_{22} & \ldots & \sqrt{\lambda_p} a_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \sqrt{\lambda_1} a_{p1} & \sqrt{\lambda_{1p}} a_{p2} & \ldots & \sqrt{\lambda_p} a_{pp} \end{bmatrix}$$

Note que, no caso das componentes principais serem obtidas a partir dos dados estandardizados, temos que:

$\rho_{X_i', Y_j} = \sqrt{\lambda_i} a_{ij} = a_{ij}^*$ (loading da variável X_i' na componente Y_j)

Então a matriz C (Component Matrix) dos transformados a_j^*, é uma matriz de loadings, e portanto pode ser usada para interpretar as componentes principais:

$$\mathrm{C} = \begin{bmatrix} \rho_{X_1', Y_1} & \rho_{X_1', Y_2} & \cdots & \rho_{X_1', Y_p} \\ \rho_{X_2', Y_1} & \rho_{X_2', Y_2} & \cdots & \rho_{X_2', Y_p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{X_p', Y_1} & \rho_{X_p', Y_2} & \cdots & \rho_{X_p', Y_p} \end{bmatrix}$$

NOTAS:

- A soma dos quadrados da coluna j de C é igual a λ_j (que dividido por *p* dá a proporção da variância total explicada pela j-ésima componente).
- É fácil de ver que a soma dos quadrados dos elementos da linha i de C é igual a 1 ($= Var(X_i')$).

Na prática, podemos estar interessados apenas nas primeiras k componentes principais. Neste caso só nos interessam as primeiras k colunas. A matriz C (Component Matriz) terá então apenas k colunas. O SPSS permite-nos reter o número de componentes que quisermos.

A soma de quadrados dos elementos da coluna j de C continuará a ser igual a λ_j (não alterámos as colunas), mas a soma dos quadrados dos elementos da linha i já não é igual a 1, mas sim a

$$\sum_{j=1}^{k}(\lambda_j^{1/2}a_{ij})^2 = \sum_{j=1}^{k}\lambda_j a_{ij}^{\ 2} = comunalidade = h_i =$$

= proporção de variância da variável X_i' explicada pelas k componentes principais retidas na análise

É claro que quando consideradas todas as componentes principais as comunalidades vêm todas iguais a 1, indicando que a proporção de variância de cada variável explicada por todas as componentes principais é igual a 1.

QUANDO É QUE A ANÁLISE DE COMPONENTES PRINCIPAIS É UMA TÉCNICA APROPRIADA?

Há casos em que poderá não ser possível explicar uma proporção significativa de variância apenas com algumas componentes principais. Em tais casos poderemos ser obrigados a usar todas as componentes principais (tantas como o nº de variáveis originais) para explicar uma quantidade significativa de variação. Isto acontece, geralmente, quando as variáveis não estão correlacionadas entre si.

Se as variáveis não estão correlacionadas entre si, então cada componente principal explicará a mesma quantidade de variância. Nestes casos não é possível atingir o objectivo de redução de dados. Por outro lado, se as variáveis estão perfeitamente correlacionadas entre si então a 1ª componente principal explicará toda a variância dos dados. Isto é, quanto maior for a correlação entre as variáveis maior redução de dados conseguiremos atingir e vice-versa.

Esta discussão sugere que a análise de componentes principais é mais apropriada se as variáveis estiverem inter-relacionadas, pois só assim é possível reduzir o nº de variáveis a um nº menor de componentes principais sem perda significativa de informação. Se não conseguirmos atingir tal objectivo então a ACP poderá não ser apropriada.

Existem testes estatísticos para determinar se as variáveis estão significativamente correlacionadas entre elas, como por exemplo o teste de **esfericidade de Bartlett** e o **KMO**.

O teste de **esfericidade de Bartlett**, que pode ser usado para dados estandardizados, testa a hipótese da matriz das correlações ser a matriz identidade (isto é, as variáveis serem não correlacionadas). A estatística de teste para o teste de esfericidade de Bartlett tem distribuição de Qui-Quadrado. Um valor elevado da estatística de teste favorecerá a rejeição da hipótese nula (teste unilateral à direita). Se a hipótese nula não poder ser rejeitada, então deve-se reconsiderar a utilização da ACP.

No entanto, este teste é sensível ao tamanho das amostras no sentido de que para amostras grandes até pequenas correlações poderão ser estatisticamente significantes, pelo que se torna preferível usar o KMO.

O **Kaiser-Meyer-Olkin** (**KMO**), é uma estatística que varia entre zero e um e compara as correlações simples com as correlações parciais observadas entre as variáveis.

Kaiser adjectiva os valores do KMO como se apresentam:

KMO	Análise Componentes Principais
1-0,9	Muito Boa
0,8-0,9	Boa
0,7-0,8	Média
0,6-0,7	Razoável
0,5-0,6	Má
<0,5	Inaceitável

Como exemplo, vamos responder à alínea f) do exercício 1.

f) Utilizando o seguinte output do SPSS, verifique se a ACP é uma técnica apropriada neste caso.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,662
Bartlett's Test of Sphericity	Approx. Chi-Square	28,251
	df	10
	Sig.	,002

- O KMO de 0,662 mostra que há uma correlação razoável entre as variáveis.
- O teste de esfericidade de Bartlett tem associado um p-value de 0,002 o que leva à rejeição da matriz das correlações na população ser a identidade, para um nível de significância superior a 0,002, evidenciando portanto que existe correlação entre algumas variáveis. Deste modo a ACP é uma técnica apropriada.