A1 - ANÁLISE DE COMPONENTES PRINCIPAIS

Evandro M Saidel Ribeiro

A1.1 Covariância e Correlação
A1.2 Autovalores e Autovetores
A1.3 Componentes Principais
A1.4 Variância explicada
A1.5 PCA: Exemplo no STATA
A1.6 PCA: Exemplo no R

Bibliografia:

• J. **Lattin**, J.D. **Carroll**, P.E. **Green**. Análise de dados Multivariados, Cengage Learning, 2011.

• L.P. **Fávero**, P. **Belfiore**, F.L. **da Silva**, B.L. **Chan**, Análise de Dados: Modelagem Multivariada para tomada de decisões, Elsevier, 2009.

• R.A. **Johnson**, Applied Multivariate Statistical Analysis, Prentice Hall, 1992

A1 Análise de Componentes Principais (PCA)

Tem a finalidade de **substituir** um conjunto de **variáveis correlacionadas** por um conjunto de novas **variáveis não-correlacionadas**, sendo essas, combinações lineares das variáveis iniciais e colocadas em **ordem decrescente de suas variâncias**.

Por exemplo, para p variáveis:

$$CP_1 = e_{11} X_1 + e_{21} X_2 + e_{31} X_3 + \ldots + e_{p1} X_p$$

$$CP_2 = e_{12} X_1 + e_{22} X_2 + e_{32} X_3 + \ldots + e_{p2} X_p$$

$$\vdots$$

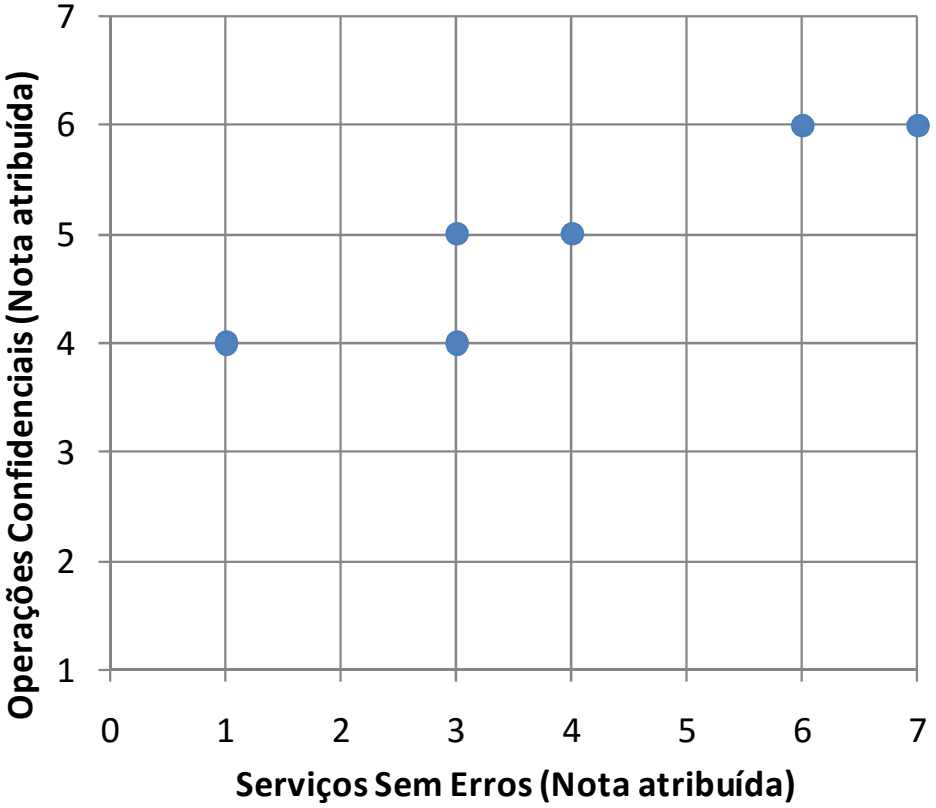
$$CP_p = e_{1p} X_1 + e_{2p} X_2 + e_{3p} X_3 + \ldots + e_{pp} X_p$$

A1 PCA - Exemplo

Um exemplo simples: pesquisa com clientes de banco

	x_1	x_2
Cliente	**SSE**	**OC**
1	7	6
2	1	4
3	3	4
4	3	5
5	6	6
6	4	5

Tabela 1: Resultado da pesquisa aplicada aos clientes de um banco. Foram atribuídas notas de 1 a 7 para: SSE - Serviços Sem Erros; e OC - Operações Confidenciais.



$$\mathbf{X}' = \begin{bmatrix} 7 & 1 & 3 & 3 & 6 & 4 \\ 6 & 4 & 4 & 5 & 6 & 5 \end{bmatrix}$$

Cálculo de covariâncias e correlações:

A1.1 Covariância e Correlação

Determinação da covariância ($\mathbf{S}$) e da correlação ($\mathbf{R}$) entre as variáveis x_1 e x_2 (SSE e OC, respectivamente)

$$\mathbf{X} = \begin{bmatrix} 7 & 6 \\ 1 & 4 \\ 3 & 4 \\ 3 & 5 \\ 6 & 6 \\ 4 & 5 \end{bmatrix} \Rightarrow \mathbf{X}_{desv} = \begin{bmatrix} 7 & 6 \\ 1 & 3 \\ 3 & 4 \\ 3 & 5 \\ 6 & 6 \\ 4 & 5 \end{bmatrix} - \begin{bmatrix} 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \\ 4 & 5 \end{bmatrix} \Rightarrow \mathbf{X}_{desv} = \begin{bmatrix} 3 & 1 \\ -3 & -2 \\ -1 & -1 \\ -1 & 0 \\ 2 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$\mathbf{S} = \frac{1}{5}\mathbf{X}'_{desv}\,\mathbf{X}_{desv}$$

$$\mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$$

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}919 \\ 0{,}919 & 1 \end{bmatrix}$$

A1.1 Covariância e Correlação - STATA

Cálculos no STATA: Statistics ▶ Summaries, tables and tests ▶ Summary and descriptive statistics ▶ Correlation and Covariances

```
. correlate
(obs=6)
```

	SSE	OC
SSE	1.0000	
OC	0.9186	1.0000

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}919 \\ 0{,}919 & 1 \end{bmatrix}$$

```
. correlate, covariance
(obs=6)
```

	SSE	OC
SSE	4.8	
OC	1.8	.8

$$\mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$$

Analisando a variância total (Exemplo clientes banco)

Variância total: soma das variâncias, ou seja,

Variância Total = Tr ($\mathbf{S}$) = $S_{11}+S_{22}+S_{33}+....+S_{pp}$

Para o nosso exemplo: $\mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$

Variância Total = Tr ($\mathbf{S}$) = $S_{11}+S_{22}$ = 4,8 + 0,8 = 5,6

Se usarmos **variáveis padronizadas**, a variância total é o traço da matriz de correlação, que é igual ao número de variáveis, ou seja, p.

Para o nosso exemplo: $\mathbf{R} = \begin{bmatrix} 1 & 0{,}919 \\ 0{,}919 & 1 \end{bmatrix}$

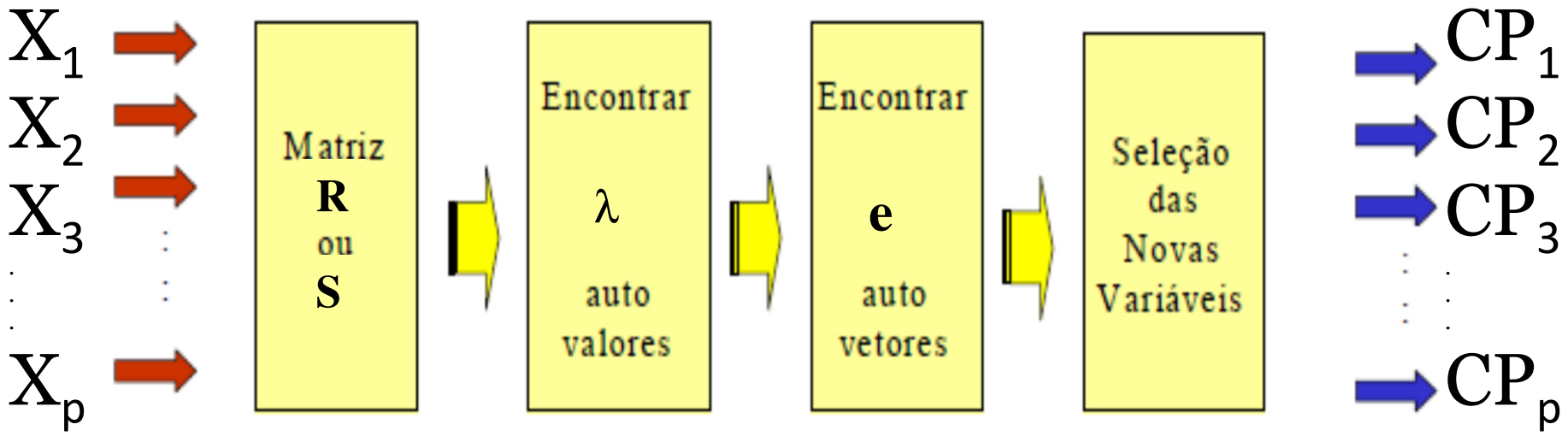
Variância Total = 2

A1.2 Autovalores e Autovetores

Queremos encontrar uma combinação linear (componentes principais) das variáveis originais, de forma que estas combinações não estejam correlacionadas, mas tenham alta variância.

$$CP_1 = e_{11} X_1 + e_{21} X_2$$

$$CP_2 = e_{12} X_1 + e_{22} X_2$$



p variáveis

p componentes principais

A1.2 Autovalores e Autovetores

Matriz Autovetor Autovalor

$$\mathbf{S}\mathbf{e}_i = \lambda_i \mathbf{e}_i$$

Os autovalores de $\mathbf{S}$, (λ_1, λ_2, . . ., λ_p) correspondem a variância associada a cada componente principal (CP_1, CP_2, CP_3... CP_p).

Os coeficientes e_{ij} correspondem aos elementos dos auto-vetores normalizados e ortogonais $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$,..., $\mathbf{e}_p$ dos respectivos autovalores da matriz de covariância (ou de correlação).

A1.2 Autovalores e Autovetores - Exemplo

A matriz de covariância obtida foi $\mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$

Autovalores são calculados por: $\det \begin{vmatrix} s_{11} - \lambda & s_{12} \\ s_{21} & s_{22} - \lambda \end{vmatrix} = 0$

Como $s_{12} = s_{21}$, os autovalores são dados por:

$$\lambda = \frac{(s_{11} + s_{22})}{2} \pm \frac{\sqrt{(s_{11} + s_{22})^2 - 4(s_{11}s_{22} - s_{12}^2)}}{2} \quad \begin{cases} \lambda_1 = 5{,}491 \\ \lambda_2 = 0{,}109 \end{cases}$$

Variância total= $\lambda_1 + \lambda_2 = 5{,}491 + 0{,}109 = 5{,}6 = \text{Tr}(\mathbf{S})$

E os autovetores de $\mathbf{S}$? $\mathbf{S}\mathbf{e}_1 = \lambda_1 \mathbf{e}_1$ $\mathbf{S}\mathbf{e}_2 = \lambda_2 \mathbf{e}_2$

A1.2 Autovalores e Autovetores - Exemplo

Considerando a matriz de covariância, os autovetores são obtidos supondo vetores $\mathbf{v}_i$ que satisfaçam

$$\mathbf{S}\mathbf{v}_i = \lambda_i \mathbf{v}_i \quad \text{ou seja:} \quad \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix} \begin{bmatrix} v_{1i} \\ v_{2i} \end{bmatrix} = \lambda_i \begin{bmatrix} v_{1i} \\ v_{2i} \end{bmatrix}$$

No caso mais geral supomos $v_{ii} = 1$

Para este exemplo esta suposição resulta em:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0{,}3837 \end{bmatrix} \qquad \mathbf{v}_2 = \begin{bmatrix} -0{,}3837 \\ 1 \end{bmatrix}$$

Estes vetores não estão normalizados!

A1.2 Autovalores e Autovetores - Exemplo

Os vetores normalizados são obtidos dividindo $\mathbf{v}_i$ pelo módulo de $\mathbf{v}_i$. No nosso exemplo os módulos são:

$$\left|\mathbf{v}_1\right| = 1{,}0711 \qquad \left|\mathbf{v}_2\right| = 1{,}0711$$

Os autovetores normalizados são:

$$\mathbf{e}_1 = \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \begin{bmatrix} 0{,}9336 \\ 0{,}3583 \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \begin{bmatrix} -0{,}3583 \\ 0{,}9336 \end{bmatrix}$$

As componentes dos auto-vetores correspondem aos coeficientes dos componentes principais.

$$CP_1 = e_{11}\, X_1 + e_{21}\, X_2 \qquad CP_1 = 0{,}9336\, X_1 + 0{,}3583\, X_2$$

$$CP_2 = e_{12}\, X_1 + e_{22}\, X_2 \qquad CP_2 = -0{,}3583\, X_1 + 0{,}9336\, X_2$$

A1.3 Componentes Principais

O próximo passo é calcular o valor de cada componente principal para os dados originais:

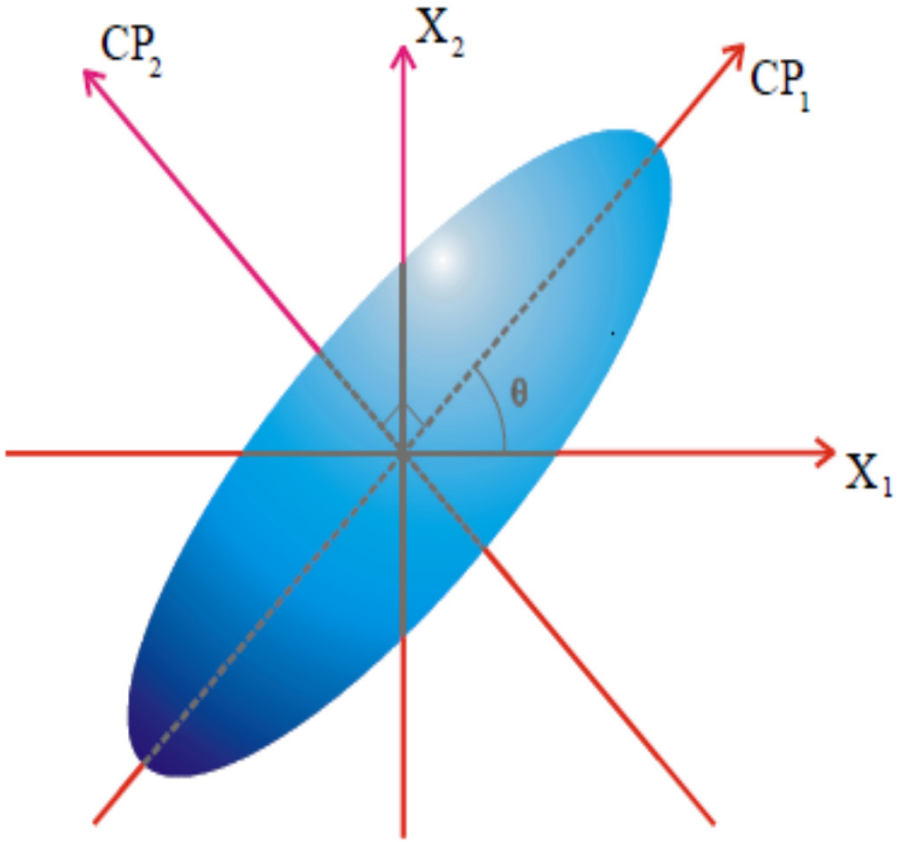
	Variáveis Originais		Variáveis Geradas pela PCA	
Cliente	**SSE**	**OC**	**CP1**	**CP2**
1	7	6	8,685	3,094
2	1	4	2,367	3,376
3	3	4	4,234	2,660
4	3	5	4,592	3,593
5	6	6	7,751	3,452
6	4	5	5,526	3,235

$CP_1 = 0{,}9336\ SSE + 0{,}3583\ OC$

$CP_2 = -0{,}3583\ SSE + 0{,}9336\ OC$

A1.3 Componentes Principais

Para duas variáveis, x_1 e x_2, espera-se que os dados estejam correlacionados, os dados num diagrama de dispersão são representados pela elipse na figura abaixo.



CP_1 tem maior variância, corresponde ao maior eixo da elipse.

CP_2 tem menor variância e é perpendicular ao eixo maior.

As variâncias são proporcionais aos autovalores.

A1.3 Componentes Principais – Número

O número de componentes principais é igual ao número de variáveis. Na análise fatorial o interesse será ter menos componentes principais do que variáveis originais.

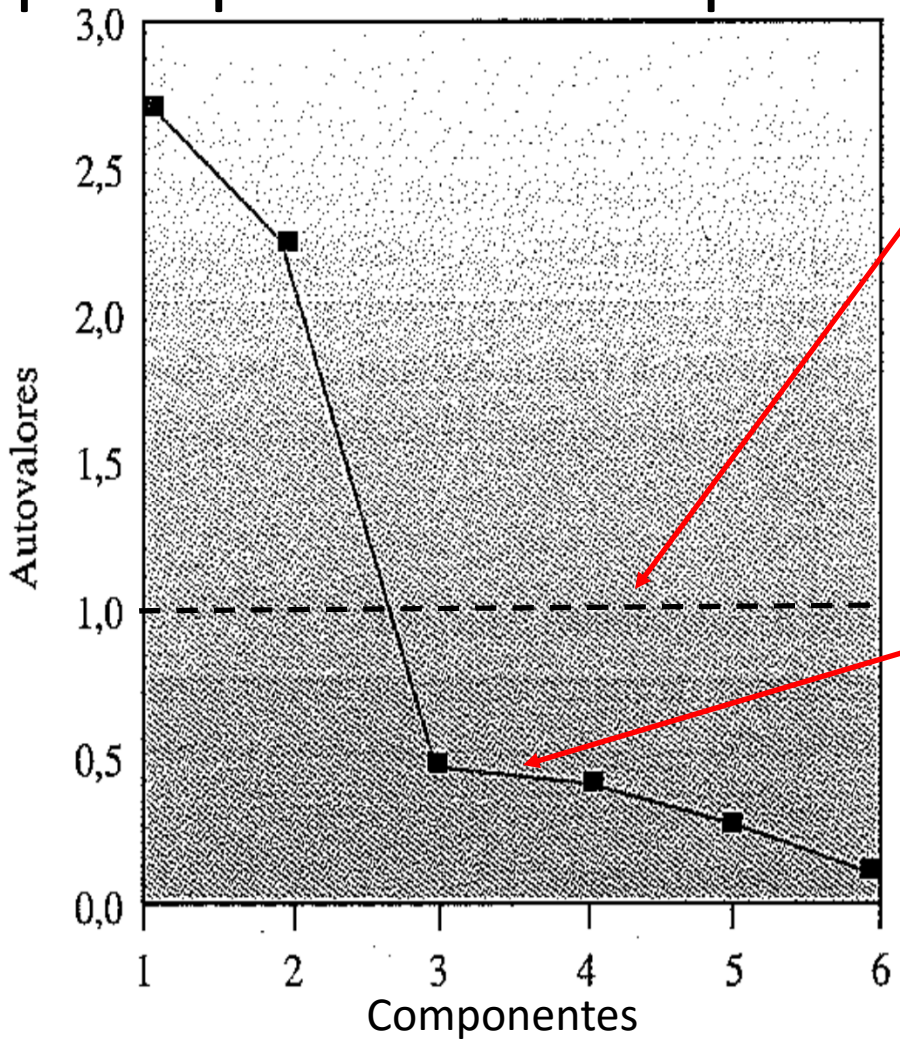
(1)Podemos estipular o número de componentes (k) com base na variância acumulada, por exemplo:

Variância acumulada = $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\times 100\% \geq 70\%$ sendo $k < p$

(2) Uma maneira visual é através do gráfico de declive (***scree plot***) . Gráfico dos autovalores ordenados do maior para o menor. (scree: declive do fundo de um precipício)

A1.3 Componentes Principais – Scree plot

Num ex. com 6 autovalores teremos 6 componentes principais. Critérios para considerar componentes:



(a) Critério de Kaiser. Autovalores maiores que um. Por este critério: 2 componentes.

(b) Autovalores até o declive (cotovelo). Por este critério: 3 componentes

A1.4 Variância explicada

A contribuição de cada componente principal $\mathbf{CP}_i$ é medida em termos da proporção da variância total explicada, que é dada por:

$$\frac{\text{Var}(\mathbf{CP}_i)}{\sum_{i=1}^{p}\text{Var}(\mathbf{CP}_i)}.100 = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}.100 = \frac{\lambda_i}{\text{Tr}(\mathbf{S})}.100$$

Se usarmos variáveis padronizadas: $\Sigma\lambda_i = \text{Tr}(\mathbf{R}) = p$

A1.4 Variância explicada

No exemplo dos clientes do banco:

Porcentagem da variação total dos dados explicada por $\mathbf{CP}_1$: $\frac{\lambda_1}{\lambda_1+\lambda_2}.100 = \frac{5{,}491}{5{,}6}.100 = 98{,}05\,\%$

Porcentagem da variação total dos dados explicada por $\mathbf{CP}_2$: $\frac{\lambda_2}{\lambda_1+\lambda_2}.100 = \frac{0{,}109}{5{,}6}.100 = 1{,}95\,\%$

Cada componente principal sintetiza a máxima proporção de variância contida nos dados

Componentes principais - Correlações

$$\lambda_1 = 5{,}491 \quad \lambda_2 = 0{,}109 \quad \mathbf{e}_1 = \begin{bmatrix} 0{,}9336 \\ 0{,}3583 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} -0{,}3583 \\ 0{,}9336 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 4{,}8 & 1{,}8 \\ 1{,}8 & 0{,}8 \end{bmatrix}$$

Correlações entre as componentes principais e a variáveis originais podem ser calculadas por:

$$\text{corr}(CP_i, x_j) = \frac{e_{ij}\sqrt{\lambda_i}}{\sqrt{s_{jj}}}$$

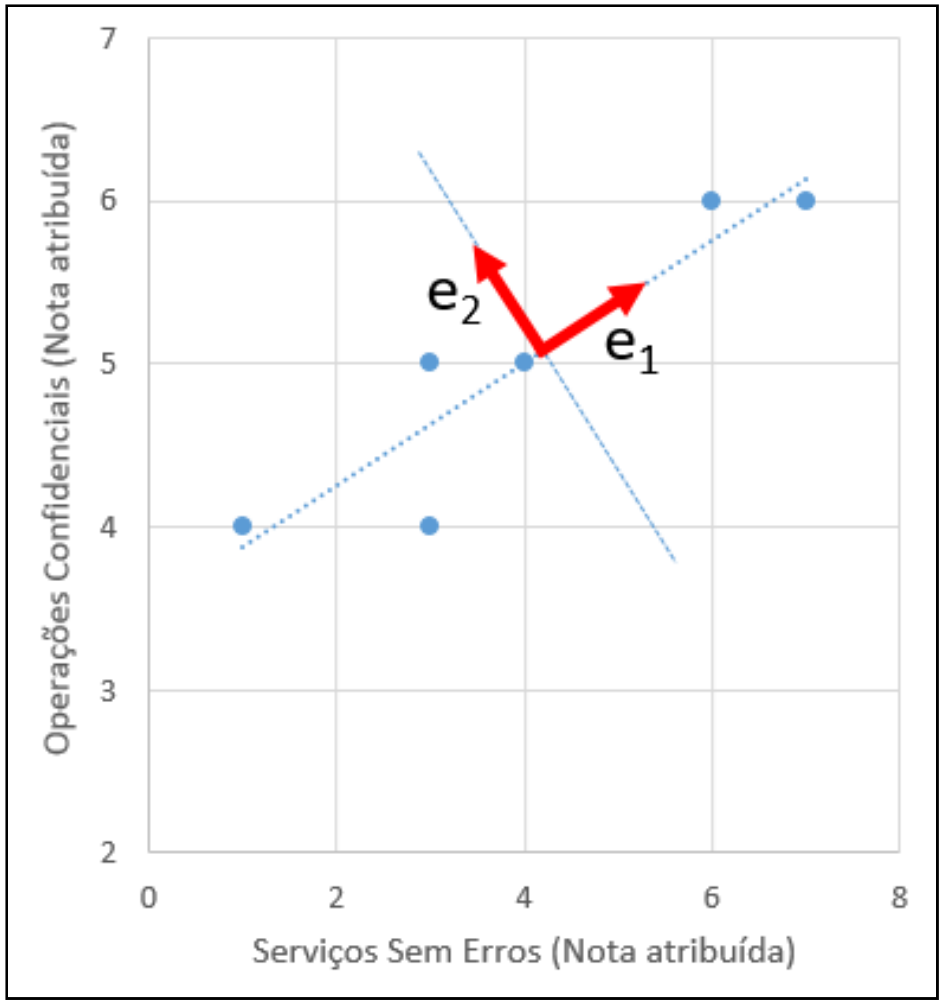
Correlações entre CP e X

	SSE	OC
CP1	0,9985	0,9386
CP2	-0,0541	0,3451

A primeira componente possui maior correlação com as variáveis, ela tem maior importância.

Componentes Principais - ilustração

Na figura abaixo os autovetores foram inseridos próximos ao ponto médio dos valores das variáveis



As notas para "Operações Confidenciais" e "Serviços Sem Erros" estão correlacionadas. Podemos encontrar um componente principal que explica 98,05% da variação total dos dados.

O exemplo indica que estas variáveis podem formar um único fator: "Confiança", por exemplo. Esta será a tarefa da **Análise Fatorial**

A1.5 PCA: Exemplo no STATA (covariância)

Statistics ▶ Multivariate analysis ▶ Factor and principal componentes analysis

▶Principal componentes analysis (PCA)

```
. pca SSE OC, covariance

Principal components/covariance                 Number of obs    =          6
                                                Number of comp.  =          2
                                                Trace            =        5.6
    Rotation: (unrotated = principal)           Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.49072	5.38145	0.9805	0.9805
Comp2	.109275	.	0.0195	1.0000

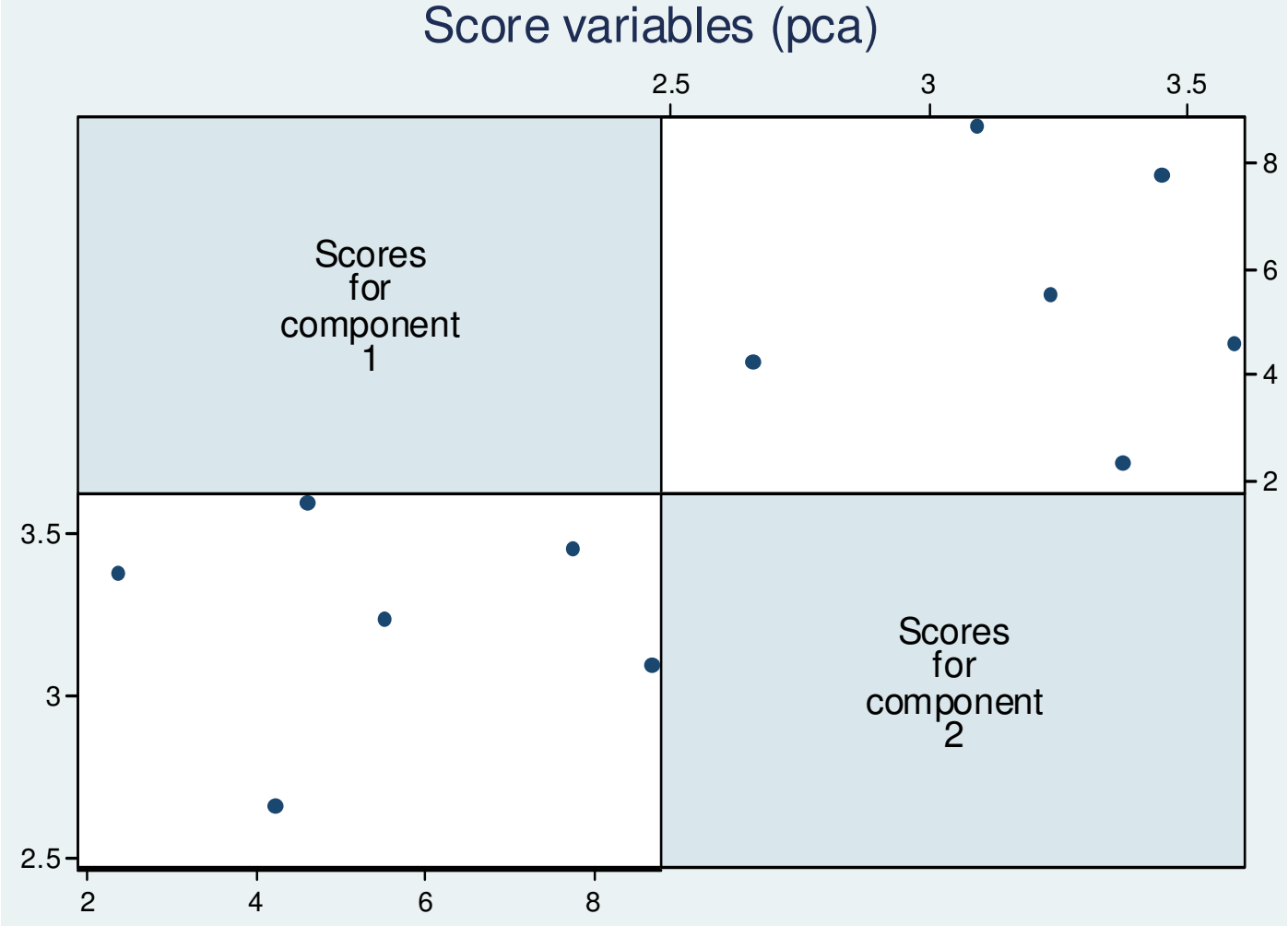
Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
SSE	0.9336	-0.3583	0
OC	0.3583	0.9336	0

A1.5 PCA: Exemplo no STATA (covariância)

SCORES Statistics ▶ Multivariate analysis ▶

Factor and principal componentes analysis ▶ Post estimation ▶ Score variables Plot



Variáveis Geradas pela PCA

CP1	CP2
8,685	3,094
2,367	3,376
4,234	2,660
4,592	3,593
7,751	3,452
5,526	3,235

A1.5 PCA: Exemplo no STATA (correlação)

Statistics ▶ Multivariate analysis ▶ Factor and principal componentes analysis

▶Principal componentes analysis (PCA)

```
. pca SSE OC

Principal components/correlation                  Number of obs    =          6
                                                  Number of comp.  =          2
                                                  Trace            =          2
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Compl	1.91856	1.83712	0.9593	0.9593
Comp2	.0814413	.	0.0407	1.0000

Principal components (eigenvectors)

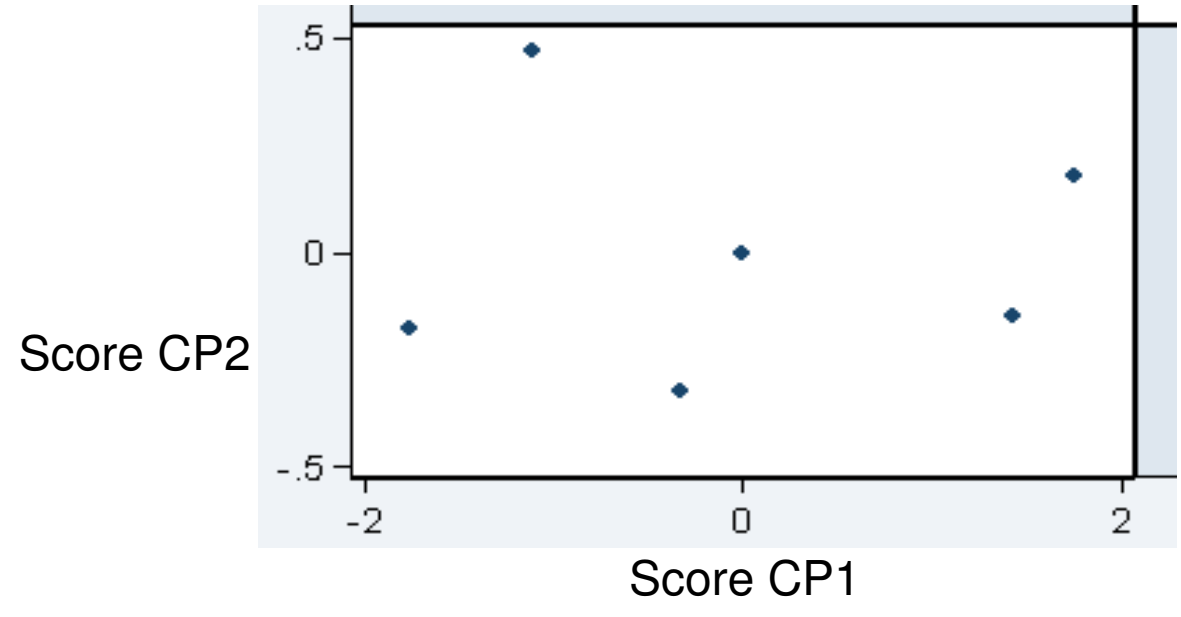
Variable	Compl	Comp2	Unexplained
SSE	0.7071	0.7071	0
OC	0.7071	-0.7071	0

A1.5 PCA: Exemplo no STATA (correlação)

SCORES Statistics ▶ Multivariate analysis ▶

Factor and principal componentes analysis ▶ Post estimation ▶ Score variables Plot

	Variáveis Originais		Variáveis Geradas pela PCA	
Cliente	**SSE**	**OC**	**CP1**	**CP2**
1	7	6	1,8	0,2
2	1	4	-1,8	-0,2
3	3	4	-1,1	0,5
4	3	5	-0,3	-0,3
5	6	6	1,4	-0,1
6	4	5	0,0	0,0
Média	4	5		
Desv.Pad	2,191	0,894		



A1.6 PCA: Exemplo no R

Arquivos: **ExemploBanco.csv** e **ExemploBanco.R**