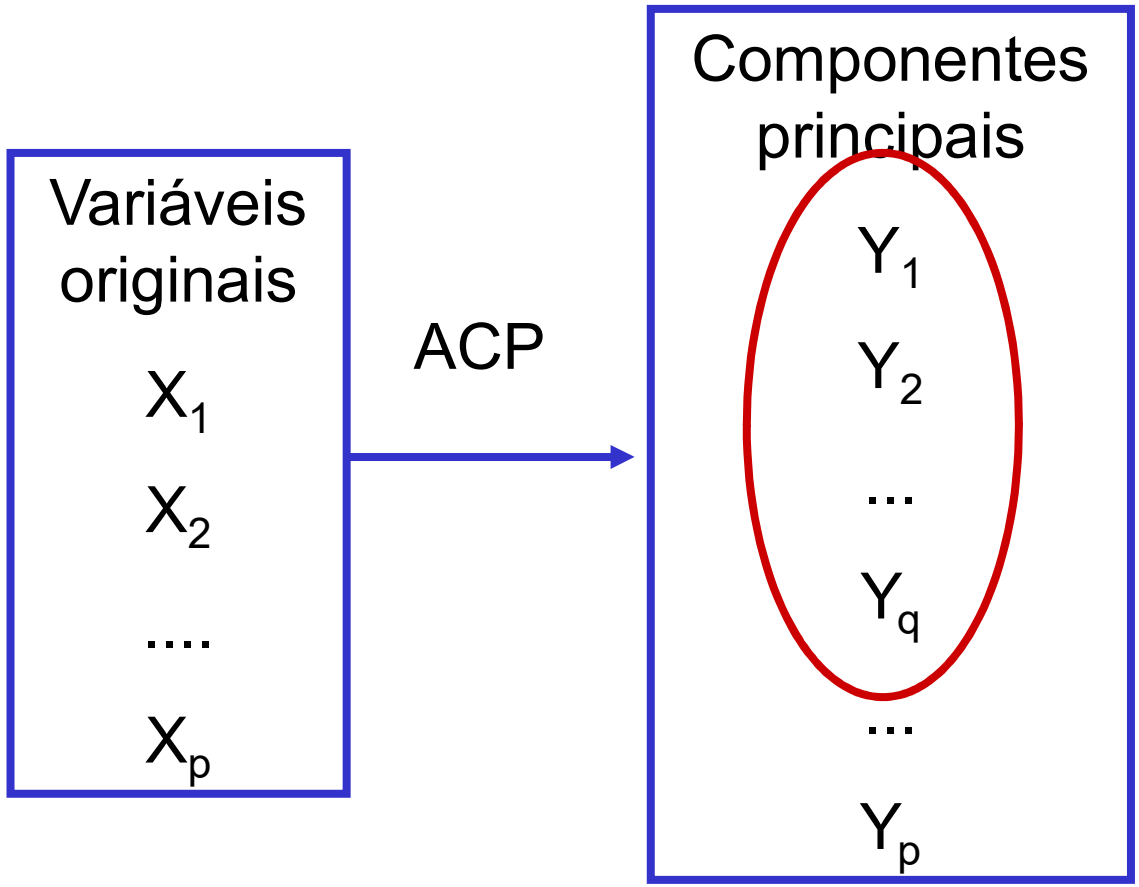
Análise de Componentes Principais

principal components analysis

Situação hipotética



q primeiras componentes resumam, por exemplo, 80% do comportamento geral das p variáveis originais

Principais Objetivos

- Redução da dimensionalidade dos dados
- Obtenção de combinações interpretáveis
- Descrição e entendimento da estrutura de correlação

Componentes Principais

Algebricamente: são combinações lineares das variáveis originais

Geometricamente: são as coordenadas dos pontos amostrais em um sistema de eixos obtido pela rotação do sistema de eixos original, na direção de variabilidade máxima

Componentes

$X_1, X_2, ..., X_p$: variáveis originais

$Y_1, Y_2, ..., Y_p$: componentes principais

$$Y_1 = \ell_{11} X_1 + \ell_{12} X_2 + ... + \ell_{1p} X_p = \underset{\sim}{\ell_1}^T \underset{\sim}{X}$$

$$Y_2 = \ell_{21} X_1 + \ell_{22} X_2 + ... + \ell_{2p} X_p = \underset{\sim}{\ell_2}^T \underset{\sim}{X}$$

...

$$Y_p = \ell_{p1} X_1 + \ell_{p2} X_2 + ... + \ell_{pp} X_p = \underset{\sim}{\ell_p}^T \underset{\sim}{X}$$

Componentes Principais

$$Y_i = \ell_{i1} X_1 + \ell_{i2} X_2 + \ldots + \ell_{ip} X_p$$

$$Y_i = \underset{\sim}{\ell}_i^T \underset{\sim}{X}$$

$$\underset{\sim}{l}_i = \begin{pmatrix} l_{i1} \\ l_{i2} \\ \vdots \\ l_{ip} \end{pmatrix} \qquad \underset{\sim}{\mathbf{X}} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

Primeira CP

$$Y_1 = \ell_{11} X_1 + \ell_{12} X_2 + \ldots + \ell_{1p} X_p = \underset{\sim}{\ell}_1^T \underset{\sim}{X}$$

Encontrar $\underset{\sim}{\ell}_1 = (\ell_{11}, \ell_{12}, \ldots, \ell_{1p})^T$ tal que:

$\text{Var}(Y_1) = \lambda_1$ seja máxima

Sujeita à restrição:

$$l_{11}^2 + l_{12}^2 + \cdots + l_{1p}^2 = l_1^T l_1 = 1$$

Segunda CP

$$Y_2 = \ell_{21} X_1 + \ell_{22} X_2 + \ldots + \ell_{2p} X_p = \underset{\sim}{\ell}_2^T \underset{\sim}{\mathbf{X}}$$

Encontrar $\underset{\sim}{\ell}_2 = (\ell_{21}, \ell_{22}, \ldots, \ell_{2p})^T$ tal que:

$\mathrm{Var}(Y_2) = \lambda_2$ seja máxima

Sujeita às restrições:

$$l_{21}^2 + l_{22}^2 + \cdots + l_{2p}^2 = l_2^T l_2 = 1$$

$\mathrm{Cov}(Y_1, Y_2) = 0$

i- ésima CP

$$Y_i = \ell_{i1} X_1 + \ell_{i2} X_2 + \ldots + \ell_{ip} X_p = \underset{\sim}{\ell}_i^T \underset{\sim}{\mathbf{X}}$$

Encontrar $\underset{\sim}{\ell}_i = (\ell_{i1}, \ell_{i2}, \ldots, \ell_{ip})^T$ tal que:

$\mathrm{Var}(Y_i) = \lambda_i$ seja máxima

Sujeita às restrições:

$$l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = l_i^T l_i = 1$$

$\mathrm{Cov}(Y_i, Y_k) = 0$, para $k < i$

Solução

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ são os autovalores de $\underset{\sim}{\boldsymbol{\Sigma}}$

$\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_p$ são os respectivos autovetores.

Caso $\underset{\sim}{\boldsymbol{\Sigma}}$ seja desconhecida, substituí-la por $\underset{\sim}{\mathbf{S}}$

Solução

$$\mathbf{y} = \Gamma^T \mathbf{x}$$

Γ: matriz cujas colunas
são os autovetores de Σ

Características das componentes

Variável	Variância
X_1	σ_1^2
X_2	σ_2^2
...	...
X_p	σ_p^2
Total	$\sigma_T^2 = \sum_{j=1}^{p} \sigma_j^2$

Componentes	Variância
Y_1	λ_1
Y_2	λ_2
...	...
Y_p	λ_p
Total	$\lambda_T = \sum_{j=1}^{p} \lambda_j$

$$\sigma_T^2 = \lambda_T$$

Características das componentes

Componentes	Variância	% de explicação
Y_1	λ_1	$100\ \lambda_1 / \sigma_T^2$
Y_2	λ_2	$100\ \lambda_2 / \sigma_T^2$
...	...	...
Y_p	λ_p	$100\ \lambda_p / \sigma_T^2$
Total	$\sigma_T^2 = \lambda_T$	

As componentes são não correlacionadas

Resultado

$$\mathrm{Corr}(Y_i, X_k) = \frac{\alpha_{ik}\sqrt{\lambda_i}}{\sigma_k}$$

Variáveis Padronizadas

As componentes principais também podem ser obtidas a partir das variáveis padronizadas, ou seja, a partir da matriz de correlação.

$$\mathbf{Y_i = \varepsilon_i^T (V^{1/2})^{-1} (X - \mu)}$$

$$\mathbf{V^{1/2} = diag(\sigma_1, \sigma_2, \ldots, \sigma_p)}$$

Resultado

Os resultados podem ser diferentes quando se faz a análise utilizando a matriz de covariância e a matriz de correlação.

A correlação, em geral, é a melhor opção quando as variâncias são muito heterogêneas.

Como obter as variáveis originais a partir das CPs ?

$$\mathbf{y} = \Gamma^T \mathbf{x}$$

$$\Gamma \mathbf{y} = \Gamma \Gamma^T \mathbf{x}$$

$$\mathbf{x} = \Gamma \mathbf{y}$$

% de Explicação

$$X_j = \alpha_{1j} Y_1 + \alpha_{2j} Y_2 + \cdots + \alpha_{pj} Y_p$$

$$Var(X_j) = \alpha_{1j}^2 Var(Y_1) + \alpha_{2j}^2 Var(Y_2) + \cdots + \alpha_{pj}^2 Var(Y_p)$$

$$\sigma_j^2 = Var(X_j) = \sum_{i=1}^{p} \alpha_{ij}^2 \lambda_i$$

A porcentagem da variância da j-ésima variável explicada pela i-ésima componente principal é:

$$\frac{\alpha_{ij}^2 \lambda_i}{\sigma_j^2}$$

Exemplo1: Deinter

Taxa de Delitos por 100.000 habitantes				
	Homicídio			Roubo e furto
Deinter	doloso	Furto	Roubo	de veículos
SJRP	10,85	1500,80	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,90	275,89
GSP	42,55	797,16	520,73	602,63
Média	25,25	1310,96	427,56	332,64
DP	14,36	239,48	330,76	275,01

Matriz de covariâncias

	HD	F	R	RFV
HD	206			
F	-1.526	57.353		
R	4.19	-20.612	109.401	
RFV	3.156	-41.428	80.242	75.628

Explicação (covariância)

CP	Autovalores	% de Explicação	% Acumulada
1	188.433	77,7	77,7
2	51.813	21,3	99,0
3	2.327	1,0	100,0
4	15	0,0	100,0

Coeficientes (covariância)

	Y_1	Y2	Y3	Y4
X_1	0.029	0.006	0.117	-0.993
X2	-0.310	0.866	-0.389	-0.050
X3	0.716	0.484	0.496	0.082
X4	0.624	-0.125	-0.768	-0.073

Correlações (covariância)

	Y_1	Y2	Y3	Y4
X_1	0.877	0.095	0.393	-0.268
X2	-0.562	0.823	-0.078	-0.001
X3	0.940	0.333	0.072	0.001
X4	0.985	-0.103	-0.135	-0.001

Matriz de correlações

	HD	F	R	RFV
HD	1			
F	-0,444	1		
R	0,882	-0,26	1	
RFV	0,8	-0,629	0,882	1

Explicação (correlação)

CP	Autovalores	% de Explicação	% Acumulada
1	3.01	75.2	75.2
2	0.80	20.0	95.2
3	0.19	4.6	99.8
4	0.01	0.2	100,0

Coeficientes (correlação)

	Y_1	Y2	Y3	Y4
Z_1	0.533	0.213	0.769	0.283
Z_2	-0.361	0.870	-0.108	0.317
Z_3	0.526	0.440	-0.233	-0.690
Z_4	0.557	-0.056	-0.586	0.586

Correlações (correlação)

	Y_1	Y2	Y3	Y4
Z_1	0.924	0.191	0.331	0.025
Z_2	-0.626	0.778	-0.047	0.028
Z_3	0.912	0.394	-0.100	-0.061
Z_4	0.966	-0.050	-0.253	0.051

% Explicada da Variância de X

	Y1	Y2
X1	76.9	0.9
	85.4	3.6
X2	31.6	67.8
	39.2	60.5
X3	88.3	11.1
	83.2	15.5
X4	97.0	1.1
	93.2	0.2

Quantas Componentes usar?

- Critério de Kaiser
- Até acumular certa porcentagem da variância total explicada
- Até acumular certa porcentagem da varância de cada variável
- Critério scree-test

Exemplo 2: Melões

- NFT: total de melões por hectare
- PT: peso médio dos melões (kg)
- PROD: Produção (kg/ha)
- NFP: nº. médio de melões por planta
- IF: índice de formato
- BRIX: teor de açúcar (graus Brix)

Fonte: Profs. Fábio Gurgel e Daniel Ferreira-UFLA

Matriz de Correlação

	NFT	PT	PROD	NFP	IF	BRIX
NFT	1.00					
PT	-0.19	1.00				
PROD	0.48	0.42	1.00			
NFP	0.99	-0.20	0.47	1.00		
IF	-0.10	0.06	0.03	-0.10	1.00	
BRIX	0.16	0.33	0.32	0.15	-0.22	1.00

Autovalores e % Explicação

CP	Autovalor	%	% Acumulada	Variação %
1	2.438	40.6	40.6	
2	1.572	26.2	66.8	14.4
3	1.101	18.4	85.2	7.8
4	0.583	9.7	94.9	8.7
5	0.299	5.0	99.9	4.7
6	0.007	0.1	100.0	4.9

Coeficientes e Correlações

Coeficientes			
Variável	**CP1**	**CP2**	**CP3**
NFT	0.598	0.249	0.055
PT	0.024	-0.713	0.162
PROD	0.459	-0.360	0.267
NFP	0.595	0.262	0.057
IF	-0.114	0.006	0.872
BRIX	0.254	-0.480	-0.370

Correlações			
Variável	**CP1**	**CP2**	**CP3**
NFT	0.934	0.312	0.058
PT	0.037	-0.894	0.170
PROD	0.717	-0.451	0.280
NFP	0.929	0.329	0.060
IF	-0.178	0.008	0.915
BRIX	0.397	-0.602	-0.388

% Explicação Individual Acumulada

Variável	CP1	CP2	CP3
NFT	87.2	96.9	97.3
PT	0.2	80.1	83.0
PROD	51.4	71.8	79.6
NFP	86.3	97.1	97.4
IF	3.2	3.2	86.9
BRIX	15.7	51.9	67.0

Interpretação das CPs

CP1: indicador da produção das plantas

CP2: indicador das características de sabor dos frutos

CP3: indicador das características físicas dos frutos

Exemplo: crimes (variáveis padronizadas)

Unidades amostrais: São José do Rio Preto, Ribeirão Preto, Bauru, Campinas, Sorocaba, São Paulo, São josé dos Campos, Santos, Grande São Paulo (n=9)

Variáveis: Homicídio doloso, Furto, Roubo e Roubo e Furto de Veículos (p=4)

Considerando dimensão 2,

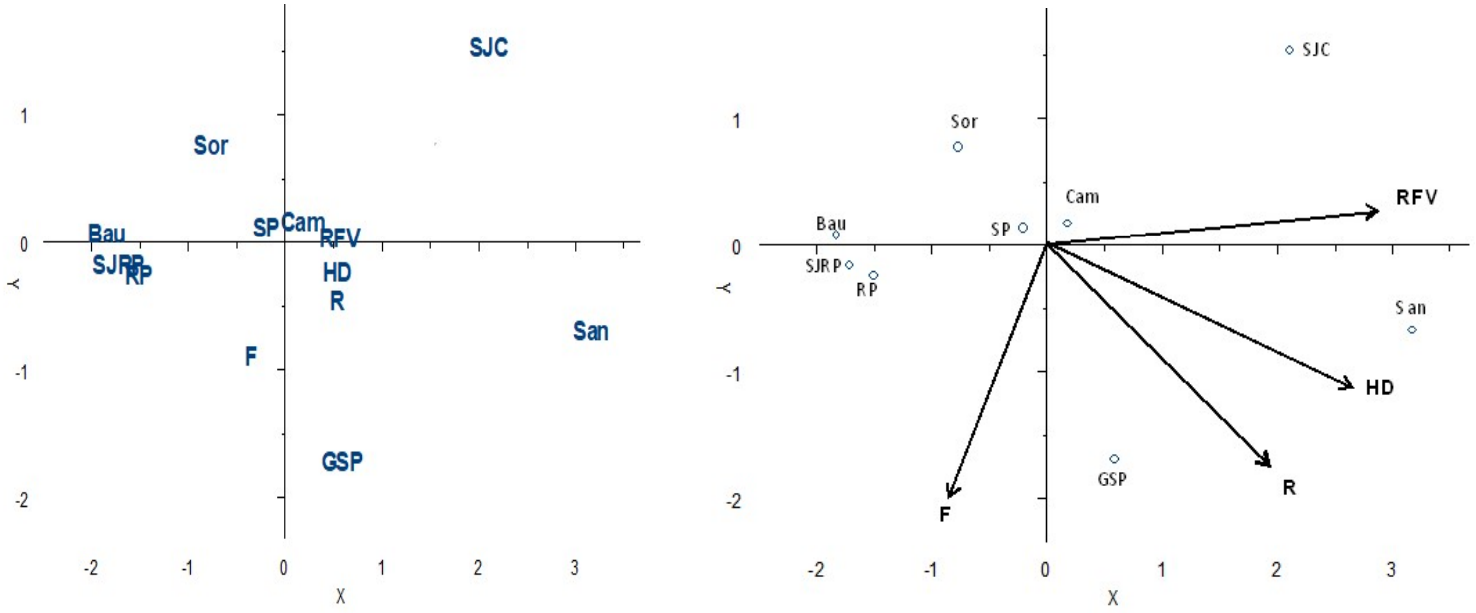
$$X_{(9x4)} = U_{(9x2)}\ D_{(2x2)}\ V'_{(2x4)}$$

$$X_{(9x4)} = A_{(9x2)}\ B_{(2x4)}$$

Exemplo: crimes (variáveis padronizadas)

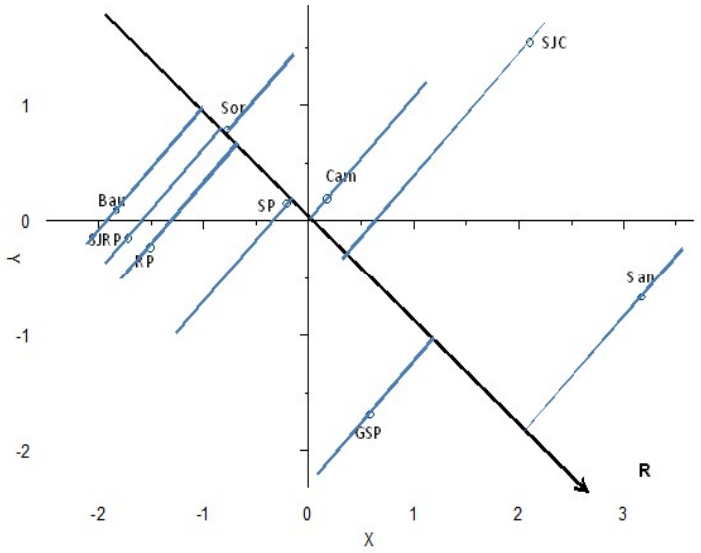
Item	Rótulo	Abscissa	Ordenada
São José do Rio Preto	SJRP	-1,72	-0,15
Ribeirão Preto	RP	-1,51	-0,23
Bauru	Bau	-1,83	0,09
Campinas	Cam	0,17	0,18
Sorocaba	Sor	-0,77	0,78
São Paulo	SP	-0,21	0,14
São José dos Campos	SJC	2,11	1,54
Santos	San	3,17	-0,67
Grande São Paulo	GSP	0,59	-1,68
Homicídio Doloso	HD	0,53	-0,21
Furto	F	-0,36	-0,87
Roubo	R	0,53	-0,44
Roubo e Furto de Veículos	RFV	0,56	0,06

Exemplo: crimes (variáveis padronizadas)



Qualidade da representação: (3,01 + 0,80) / 4 = 0,952

Exemplo: crimes (variáveis padronizadas)



Ordem no gráfico: San > GSP > SJC > Cam > SP > RP > Sor > SJRP > Bau

Ordem observada: **SP** > San > GSP > **Cam** > SJC > **Sor** > RP > SJRP > Bau