Lista 3 Entrega Multivariada

Davi Wentrick Feijó - 200016806

2023-05-04

Exercício 32 da Lista 4

Suponha que um pesquisador padronizou os dados de um estudo através da transformação de Mahalanobis ($Z = XS^{-1/2}$), em que S é a matriz de variância-covariâncias amostrais. Seria razoável aplicar componentes principais nos dados transformados? Justifique sua resposta.

Vamos usar a seguinte matriz como exemplo! ela vai ser nosso X

4	-2	2	-2
-2	5	1	2
2	1	3	0
-2	2	0	6

Agora vamos calcular a inversa da matriz exemplo e depois tirar sua raiz quadrada para encontrar $S^{-1/2}$

18106456	10863874	-8449679	6035485
10863874	6518325	-5069808	3621291
-8449679	-5069808	3943185	-2816559
6035485	3621291	-2816559	2011829

Agora podemos jogar na formula para encontrar Z

72425825	-21727747	-16899358	-12070971
-21727747	32591623	-5069808	7242582
-16899358	-5069808	11829555	0
-12070971	7242582	0	12070974

Agora podemos tirar a matriz de covariancia de Z

2.010289e+15	-7.847851e+14	-4.195429e+14	-4.055549e+14
-7.847851e+14	5.233261e+14	5.834160e+13	1.873809e+14
-4.195429e+14	5.834160e+13	1.418422e+14	6.187747e+13
-4.055549e+14	1.873809e+14	6.187747e+13	1.102527e+14

Em seguida podemos utilizar a matriz de covariancia de Z para fazer nosso PCA

PC1	PC2	PC3	PC4
-2.134807e+15	-3.073535e+13	-1.397768e+12	0.0312500
1.041086e+15	-1.618627e+14	-8.578442e+12	-0.1250000
5.310786e+14	1.731362e+14	-1.057852e+13	0.0390625
5.626422e+14	1.946187e+13	2.055473e+13	-0.0156250

Aqui temos a variancia explicada com cada PC

	x
PC1	0.9908
PC2	0.0091
PC3	0.0001
PC4	0.0000

Podemos comparar com uma PCA aplicada na matriz de covariancia de X sem passar pela transformação de Mahalanobis

PC1	PC2	PC3	PC4
-17.146338	0.8717701	0.1445018	0
10.337642	-3.0390671	0.0912855	0
-7.175171	-0.7588844	-0.2283287	0
13.983866	2.9261814	-0.0074585	0

A variancia explicada da matriz de covariancia de X

	x
PC1	0.9712
PC2	0.0287
PC3	0.0001
PC4	0.0000

Podemos perceber que a transformacao concentrou quase toda informação na primeira componente .A aplicação da PCA na matriz de Mahalanobis pode ajudar a reduzir a dimensionalidade da matriz e permitir que as informações mais relevantes sejam extraídas.

Exercício 37 da Lista 4 - Johnson e Wichern - Exercício 8.12.

Dados no arquivo Air Pollution (T1-5.DAT). Os dados correspondem a 42 medidas de poluição do ar observadas na área de Los Angeles em um mesmo horário. $X1 : Wind; X2 : SolarRadiation; X3 : CO; X4 : NO; X5 : NO2; X6 : O3; X7 : HC$.

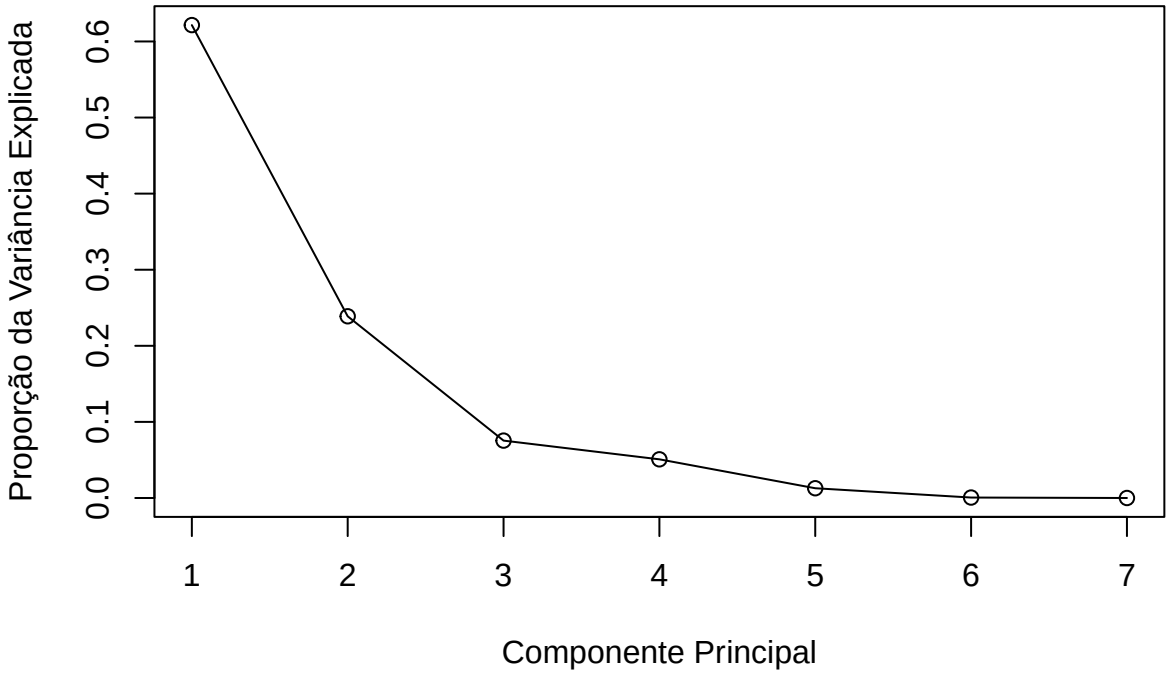
wind	solar_radiation	CO	NO	NO2	O3	HC
8	98	7	2	12	8	2
7	107	4	3	9	5	3
7	103	4	3	5	6	3
10	88	5	2	8	15	4
6	91	4	2	8	10	3
8	90	5	2	12	12	4
9	84	7	4	12	15	5
5	72	6	4	21	14	4
7	82	5	1	11	11	3
8	64	5	2	13	9	4
6	71	5	4	10	3	3
6	91	4	2	12	7	3
7	72	7	4	18	10	3
10	70	4	2	11	7	3
10	72	4	1	8	10	3
9	77	4	1	9	10	3
8	76	4	1	7	7	3
8	71	5	3	16	4	4
9	67	4	2	13	2	3
9	69	3	3	9	5	3
10	62	5	3	14	4	4
9	88	4	2	7	6	3
8	80	4	2	13	11	4
5	30	3	3	5	2	3
6	83	5	1	10	23	4
8	84	3	2	7	6	3
6	78	4	2	11	11	3
8	79	2	1	7	10	3
6	62	4	3	9	8	3
10	37	3	1	7	2	3
8	71	4	1	10	7	3
7	52	4	1	12	8	4
5	48	6	5	8	4	3
6	75	4	1	10	24	3
10	35	4	1	6	9	2
8	85	4	1	9	10	2
5	86	3	1	6	12	2
5	86	7	2	13	18	2
7	79	7	4	9	25	3
7	79	5	2	8	6	2
6	68	6	2	11	14	3
8	40	4	3	6	5	2

As seguintes questões são feitas no livro:

(a) Resumir os dados em em menos de 7 dimensões (se possível) através de análise de componentes principais utilizando a matrix de covariâncias S e apresentar suas conclusões.

Vale notar que vamos optar por normalizar os dados já que estao em escalas diferentes que podem acabar se distoando somente pela forma de medida (como no exemplo acima a radiação solar)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
wind	-2.2924613	-1.0737278	-1.0805437	-0.0312221	-0.2467956	0.0381333	0
solar_radiation	3.5530204	-0.9880229	-0.2487641	-0.7244098	0.0761370	0.0019894	0
CO	-0.9614146	-0.1074319	0.7858461	-0.2884078	-0.3613112	-0.0925438	0
NO	-1.5230543	0.1033952	0.9543866	-0.3377302	0.1934729	0.0943268	0
NO2	0.7628518	2.7374241	-0.4473302	-0.0037972	-0.1013611	0.0116899	0
O3	1.6955688	-0.6191695	0.3419890	1.1685692	-0.0916879	0.0204639	0
HC	-1.2345109	-0.0524672	-0.3055837	0.2169980	0.5315459	-0.0740594	0

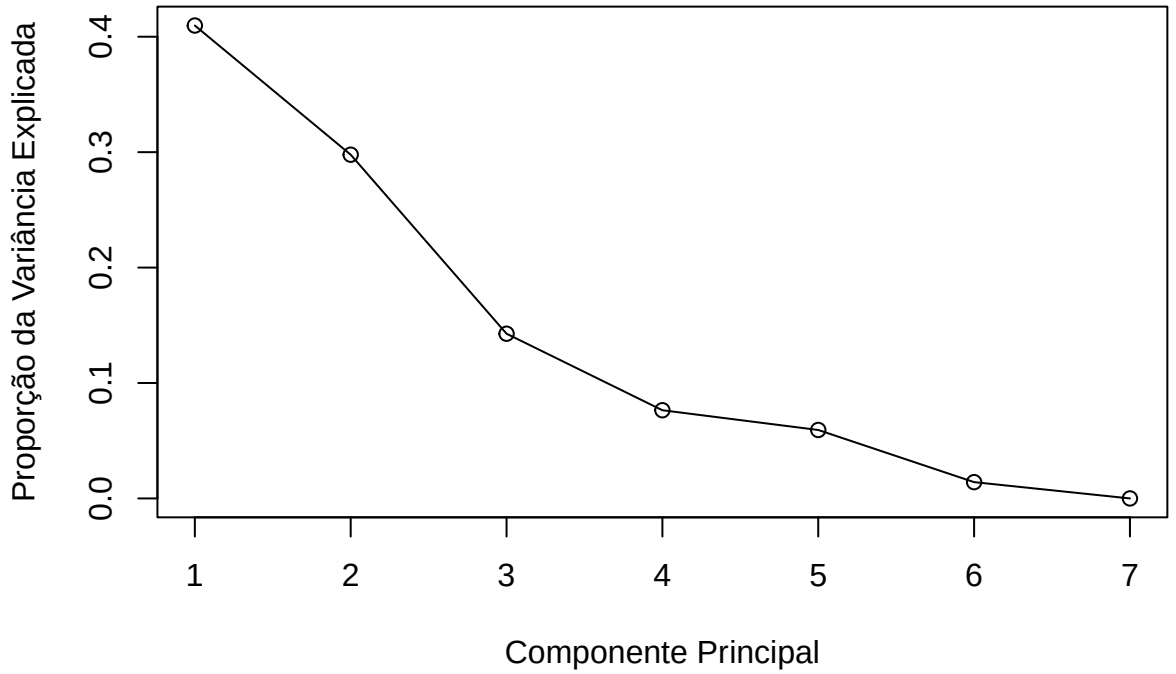



```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7
## 0.6215 0.2388 0.0754 0.0508 0.0128 0.0006 0.0000
```

```
## As 3 primeiras componentes explicam: 0.9357
```

(b) Resumir os dados em em menos de 7 dimensões (se possível) através de análise de componentes principais utilizando a matrix de correlações R e apresentar suas conclusões.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
wind	-3.3246202	0.7167773	0.5441323	0.4925443	-0.3436494	-0.0276421	0
solar_radiation	-0.6355540	-2.0939084	0.1406882	-1.2141415	-0.1809864	-0.0648603	0
CO	1.7592694	-0.1784703	0.5763529	0.5403962	-0.3407019	-0.5341095	0
NO	0.9515666	1.2199173	1.4710781	-0.3637897	0.6941522	0.2305068	0
NO2	1.2736086	0.8062578	-0.6395591	-0.0537019	-1.0168599	0.3661664	0
O3	0.2412495	-1.8675923	-0.5060371	0.9643400	0.5759132	0.2541099	0
HC	-0.2655199	1.3970185	-1.5866553	-0.3656474	0.6121322	-0.2241711	0




```
##    PC1    PC2    PC3    PC4   <NA>   <NA>   <NA>
## 0.4097 0.2978 0.1427 0.0764 0.0593 0.0141 0.0000

## As 3 primeiras componentes explicam: 0.8502
```

(c) A escolha da matriz para análise faz alguma diferença? Explique.

A matriz de correlação é calculada a partir da matriz de dados original, dividindo cada valor pelo desvio padrão da variável correspondente. Isso permite que todas as variáveis tenham a mesma escala, o que é importante quando se deseja avaliar a correlação entre elas. Quando a matriz de correlação é usada no PCA, as componentes principais resultantes são ortogonais e não estão correlacionadas entre si.

Por outro lado, a matriz de covariância é calculada a partir da matriz de dados original, sem ajustes para diferentes escalas. Isso significa que as variáveis com variações maiores terão mais peso na análise do que as variáveis com variações menores. Quando a matriz de covariância é usada no PCA, as componentes principais resultantes também são ortogonais, mas podem estar correlacionadas entre si.

Em resumo, a principal diferença entre o PCA com matriz de correlação e o PCA com matriz de covariância está na maneira como as variáveis são escalonadas antes da análise. Se as variáveis tiverem diferentes escalas, é importante usar a matriz de correlação para garantir que todas as variáveis tenham o mesmo peso na análise. Se as variáveis estiverem na mesma escala ou se a escala não for um problema, a matriz de covariância pode ser usada.

(D) Os dados podem ser resumidos em 3 ou menos dimensões?

Como podemos percerber tanto com a matriz de covariancia quanto a matriz de corelções explicam bem a variancia dos dados em 3 componentes. Contudo vale ressaltar que a matriz de correlçoes explica menos com 3 PCs que a matriz de covariancia (85% vs 93%).