Exercício Escolar 1: Análise de Componentes Principais

Estatística Multivariada II
Autora: Maria Nilza de Sousa Ramos
Professor: José Clelto Barros Gomes

1. Introdução

A Análise de Componentes Principais (ACP) é uma técnica estatística utilizada para reduzir a dimensionalidade de um conjunto de dados, transformando um grande número de variáveis correlacionadas em um menor número de variáveis não correlacionadas, chamadas de componentes principais. A ACP visa explicar a maior parte da variabilidade dos dados originais com o menor número possível de componentes principais.

2. Objetivo

O objetivo deste relatório é aplicar a ACP em três conjuntos de dados para identificar as componentes principais e analisar a variância explicada por cada uma delas.

3. Metodologia

Em todos os bancos de dados fornecidos possuíam mais observações do que variáveis, portanto foi possível utilizar a ACP. Por sua vez, esta foi aplicada utilizando a linguagem de programação Python e o notebook do código pode ser encontrado na plataforma Kaggle.

4. Desenvolvimento

4.1. Conjunto de Dados 1

O primeiro conjunto de dados possui 11 variáveis relacionadas à meteorologia, englobando registros de temperatura, umidade, evaporação e ventos. Após padronização e aplicação da ACP, foram obtidas as seguintes porcentagens de explicação das variáveis:

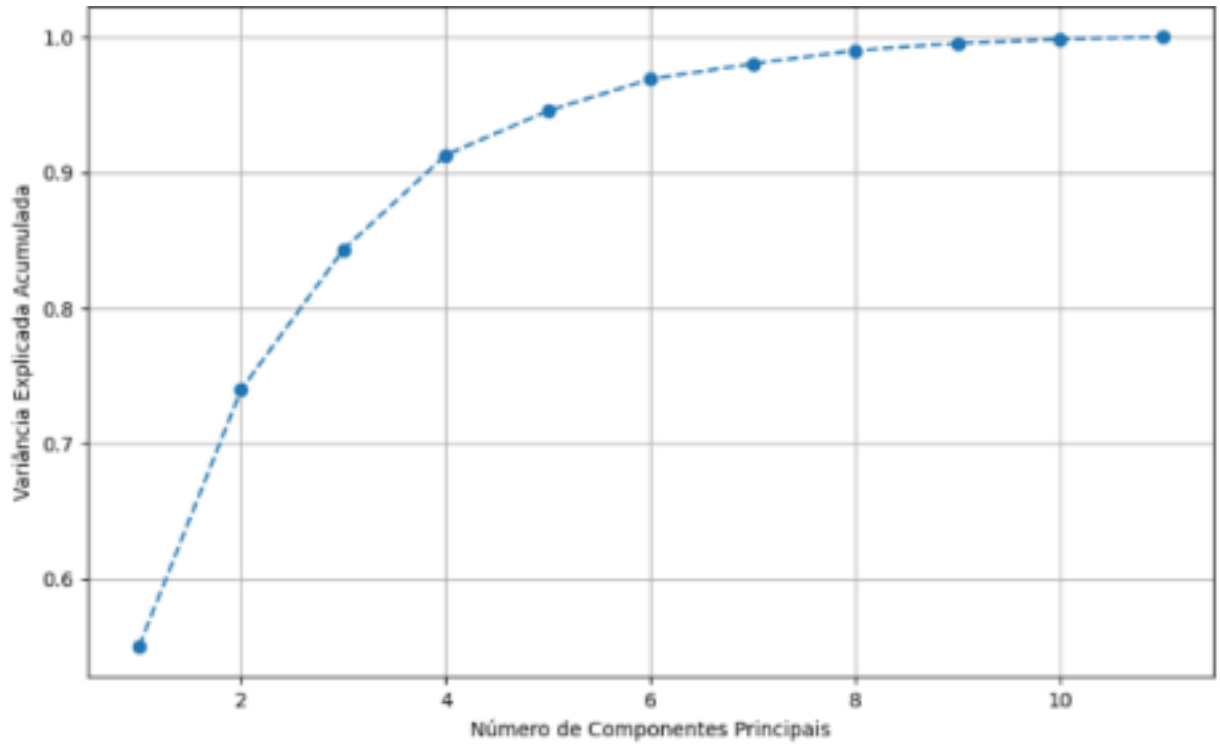
Tabela 1: Porcentagem das Variâncias Explicadas para cada Componente Principal

Componente Principal	1	2	3	4	5	6	7	8	9	10	11
Variância Explicada (%)	55.03	18.96	10.33	6.97	3.26	2.34	1.13	0.98	0.53	0.29	0.18
Variância Explicada Acumulada (%)	55.03	73.99	84.32	91.29	94.55	96.88	98.01	99.00	99.53	99.82	100

A análise revelou que os quatro primeiros componentes principais explicam aproximadamente 91% da variância total dos dados, a partir da quinta componente em diante a variância explicada é inferior a 5%.

Podemos aferir serem baixas visualmente também no screeplot da Figura 1 a seguir, onde da quinta componente em diante há uma angulação bem baixa com relação ao eixo horizontal.

Figura 1: Screeplot com a Variância Explicada Acumulada para cada Componente Principal



Isso indica que a maior parte da variabilidade nos dados pode ser explicada por apenas

quatro componentes, permitindo uma significativa redução na dimensionalidade dos dados com mínima perda de informação.

Com a redução de dimensionalidade de 11 para 4 temos as seguintes componentes principais, onde cada uma combina linearmente com as 11 variáveis por observação:

CP1: -0.3294982 -0.35265814 -0.39072946 -0.38043982 -0.2313135 -0.36070001 0.08902412 0.26058246 0.31153742 0.02271676 -0.33563171

CP2: 0.07805701 -0.19822345 -0.05633223 -0.05218239 -0.53510112 -0.2405401 -0.03042597 -0.49019644 -0.36869958 -0.46107206 0.12157464

CP3: 0.096284 0.10708815 0.11487751 0.13650179 0.00263238 0.11534703 0.78124554 0.09970789 0.20212851 -0.48968358 -0.18217539

CP4: -0.27723066 -0.22547128 -0.14110691 -0.0119847 -0.06393415 0.1351658 0.56009945 -0.16517059 -0.21456211 0.49868976 0.44065022

4.2. Conjunto de Dados 2

O segundo conjunto de dados são os recordes femininos na modalidade de atletismo referente a 54 países. Os autovalores ordenados de forma decrescente da matriz de correlação obtidos foram:

Tabela 2: Autovalores ordenado de forma decrescente para a matriz de correlação do Conjunto de Dados 2

	1	2	3	4	5	6	7
Autovalores	5,81	0,63	0,28	0,12	0,091	0,055	0,014

Os autovetores associados a suas respectivas componentes principais (PC) por sua vez são:

Tabela 3: Autovetores associados aos autovalores ordenados para a matriz de correlação do Conjunto de Dados 2

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
100 m/s	-0,3778	-0,4072	-0,1406	0,5871	-0,1671	0,5397	0,0889
200 m/s	-0,3832	-0,4136	-0,1008	0,1941	0,0935	-0,7449	-0,2657

400 m/s -0,3680 -0,4594 0,2370 -0,6454 0,3273 0,2401 0,1266 **800 m/min** -0,3948 0,1612 0,1475 -0,2952 -0,8191 -0,0165

-0,1952 **1500 m/min** -0,3893 0,3091 -0,4220 -0,0667 0,0261 -0,1890 -0,7308 **3000 m/min** -0,3761 0,4232 -0,4061 -0,0802

0,3517 0,2405 -0,5715 **Maratona min** -0,3552 0,3892 0,7411 0,3211 0,2470 -0,0483 0,0821

Interpretação das duas primeiras Componentes Principais (CP1 e CP2):

PC1 está fortemente correlacionado com todas as variáveis de corrida (100m, 200m, 400m, 800m, 1500m, 3000m e Maratona), sugerindo que ele representa um fator geral de desempenho em corrida. PC2 pode capturar captura variações entre corridas de 400m em diante, indicando um diferencial se tratando de distâncias.

Como podemos ver na tabela 4, as duas primeiras componentes principais explicam aproximadamente 92% da variância total dos dados, indicando que a maior parte da informação pode ser capturada por esses componentes.

Tabela 4: Porcentagem das Variâncias Explicadas para cada Componente Principal do Conjunto de Dados 2

CP1 CP2 CP3 CP4 CP5 CP6 CP7

Variância Explicada (%) 82,97 8,98 3,99 1,78 1,30 0,78 0,20
82,97 91,95 95,94 97,71 99,02 99,80 1

Variância Explicada Acumulada (%)

Ao levar em conta a primeira componente e ranquear temos respectivamente os seguintes países na liderança: USA, GER, RUS, CHN, FRA, GBR, CZE, POL, ROM e AUS.

Foi feita ainda uma outra análise após converter todos os dados para m/s, os autovalores para a matriz de covariância em ordem decrescente obtidos foram:

Tabela 5: Autovalores ordenado de forma decrescente para a matriz de correlação do Conjunto de Dados 2 transformada

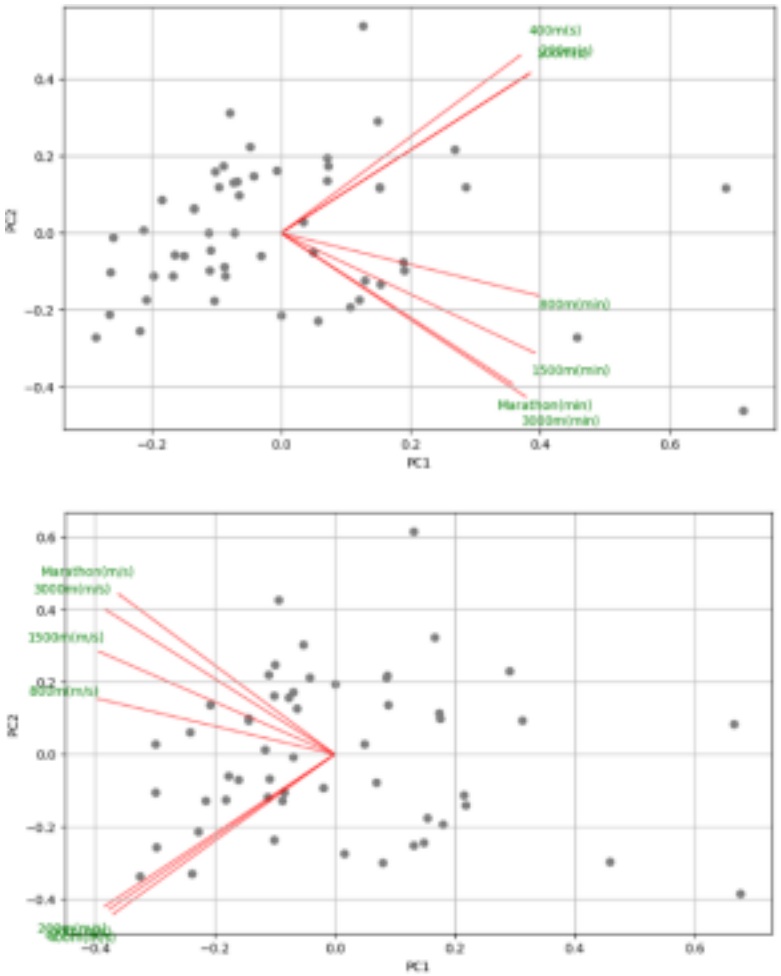
1 2 3 4 5 6 7

Autovalores 0.73214696 0.08607185 0.03338003 0.01497734 0.00206554 0.00616758 0.00885102

Apesar de alterações nos autovalores e autovetores, a análise das duas primeiras componentes principais, com variância explicativa acumulada de 92,6%, não se alterou. Pode ser observado nas figuras 2 e 3 que apesar de uma troca de sinais, a CP1 permaneceu associada a um fator geral de desempenho na corrida e a CP2 se diferenciou para distâncias inferiores a 800m.

Figura 2: Biplot das Componentes Principais 1 e 2 para o Conjunto de Dados 2.

Figura 3: Biplot das Componentes Principais 1 e 2 para o Conjunto de Dados 2 transformado.



Ainda assim, houve diferença no ranking tendo na liderança USA, RUS, CHN, GER, GBR, FRA,CZE, POL, ROM e AUS. Já nas três últimas colocações antes e após a mudança permaneceram PNG, COK e SAM.

4.3. Conjunto de Dados 3

O segundo conjunto de dados são os recordes masculinos na modalidade de atletismo referente a 54 países em 8 categorias. Os autovalores ordenados de forma decrescente da matriz de covariância obtidos para esses dados foram:

Tabela 6: Autovalores ordenados para a matriz de covariância do Conjunto de Dados 3.

	1	2	3	4	5	6	7	8
Autovalores	6.7033	0.6384	0.2275	0.2058	0.0976	0.0707	0.0469	0.0097

Os autovetores associados a suas respectivas componentes principais (PC) por sua vez são:

Tabela 7: Componentes Principais do Conjunto de Dados 3.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8
100 m (s)	-0.3324	-0.5294	-0.3439	0.3807	0.2997	-0.3620	0.3476	-0,0657
200 m (s)	-0.3461	-0.4704	0.0038	0.2170	-0.5414	0.3486	-0.4399	0.0608
400 m (s)	-0.3391	-0.3453	0.0671	-0.8513	0.1330	0.0771	0.1136	-0.0035
800 m (min)	-0.3530	0.0895	0.7827	0.1343	-0.2273	-0.3413	0.2589	-0.0393
1500 m (min)	-0.3660	0.1537	0.2443	0.2330	0.6516	0.5298	-0.1470	-0.0397
5000 m (min)	-0.3698	0.2948	-0.1829	-0.0546	0.0718	-0.3591	-0.3283	0.7057
10000 m (min)	-0.3659	0.3336	-0.2440	-0.0871	-0.0613	-0.2731	-0.3511	-0.6972
Maratona (min)	-0.3543	0.3866	-0.3346	0.0181	-0.3379	0.3752	0.5942	0.0693

Interpretação das duas primeiras Componentes Principais (PC1 e PC2):

A análise fica extremamente parecida com os dados do ranking nacional feminino, onde a CP1 está correlacionada com todas as variáveis, entretanto na CP2 há captura de um diferencial entre corridas inferiores e a partir de 800m. Este diferencial pode ser visto na figura a seguir:

Figura 4: Biplot das Componentes Principais 1 e 2 para o Conjunto de Dados 3.

Assim como no caso anterior, CP1 e CP2 somam aproximadamente 92% da variância explicada total dos dados, indicando que a informação pode ser capturada pelos dois primeiros componentes principais.

Tabela 8: Porcentagem das Variâncias Explicadas para cada Componente Principal do Conjunto de Dados 3.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8
Variância Explicada (%)	83,79	7,98	2,84	2,57	1,22	0,884	0,587	0,121
Variância Explicada Acumulada (%)	83,79	91,77	94,62	97,19	98,41	99,29	99,88	100

Ao levar em conta a primeira componente principal temos um ranking diferente do feminino com os seguintes países na liderança: Estados Unidos da América, Grã-Bretanha, Quênia, França, Austrália, Itália, Brasil, Alemanha, Portugal e Canadá.

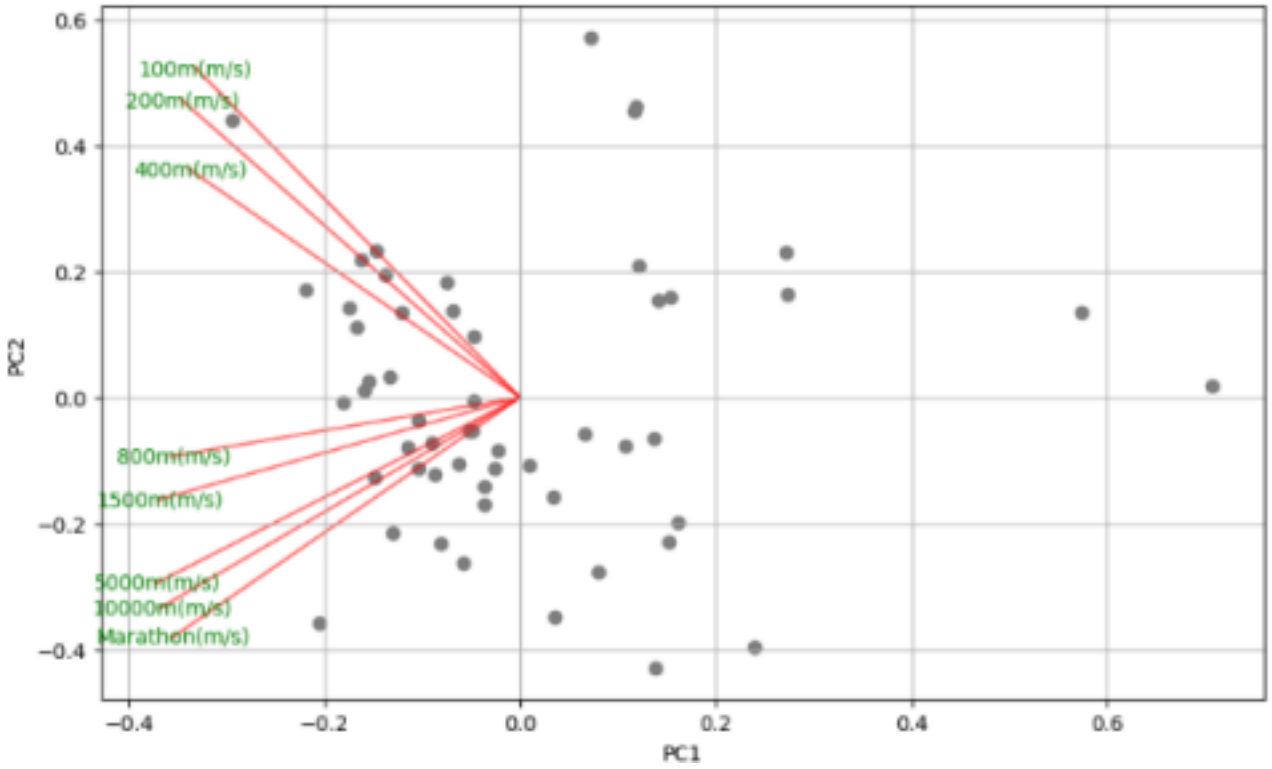
Assim como no conjunto de dados 2, foi feita ainda uma outra análise após converter todos os dados para m/s. Os oito autovalores para a matriz de covariância em ordem decrescente obtidos foram 0.49404995, 0.0462238, 0.01391228, 0.0133208, 0.00112071, 0.00322038, 0.00752255 e 0.00574921. Apesar das diferenças numéricas causadas pela transformação dos dados, e diferir a metodologia da CPA original, que foi com base na correlação, e a CPA da transformada, baseada na matriz de variância-covariância. Tanto o ranking de países quanto a análise das duas primeiras componentes mantiveram-se parecidas. A variância explicativa acumulada também se aproximou de 92%.

Tabela 9: Autovalores ordenados para a matriz de covariância do Conjunto de Dados 3.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8
Variância Explicada (%)	82,82	8,46	2,99	2,68	1,35	0,89	0,67	0,15
Variância Explicada Acumulada (%)	82,82	91,28	94,27	96,95	98,29	99,18	99,85	100

O ranking possuiu alteração na décima colocação após a transformação, deixando de ser do Canadá e passando a ser da Bélgica. As três últimas colocações em ambas as análises baseadas na CP1 foram de Singapore, Samoa e Cook Islands.

Figura 5: Biplot das Componentes Principais 1 e 2 para o Conjunto de Dados 3 transformado.



Podemos conferir na figura 5 o diferencial na componente 2 entre corridas inferiores e a partir de 800m e a correlação na componente 1 entre todas as diferentes categorias. Exatamente como antes da transformação, mas espelhada.

5. Conclusão

A matriz de correlação é a melhor maneira para obter as componentes principais, pois é equivalente às obtidas a partir da matriz de covariâncias das variáveis originais padronizadas. Determinar a proporção da variância de cada componente principal com relação a todas é um ótimo modo de determinar até onde se pode reduzir a dimensionalidade. Nos últimos dois bancos de dados, transformar os valores de forma a ficarem todas as variáveis de velocidade no SI (m/s) não causou mudança na análise final.