Análise Multivariada II
Lista III

Breno Cauã Rodrigues da Silva

1 Análise de Variância Multivariada a 2 Fatores (MANOVA Two Way)

1.1 Introdução

Neste relatório realiza-se a análise dos dados de feijão-vagem utilizando uma análise de variância multivariada a dois fatores (MANOVA Two Way). Os fatores de interesse são:

- **S:** Data de semeadura (níveis 1 a 4);
- **V:** Variedade (níveis 1 a 3).

As variáveis dependentes são:

- `y1`: Precocidade do Rendimento;
- `y2`: Precocidade da Área Foliar Específica (SLA);
- `y3`: Rendimento Total;
- `y4`: SLA Médio.

Cada combinação de níveis de `S` e `V` possui 5 repetições, totalizando 60 observações.

1.2 Análise Exploratória de Dados

Análise exploratória foi feita por medidas de resumo - média e desvio padrão - juntamente de gráficos com o objetivo de evidenciar, caso exista, diferênças entre os fatores ainda na etapa de análise descritiva.

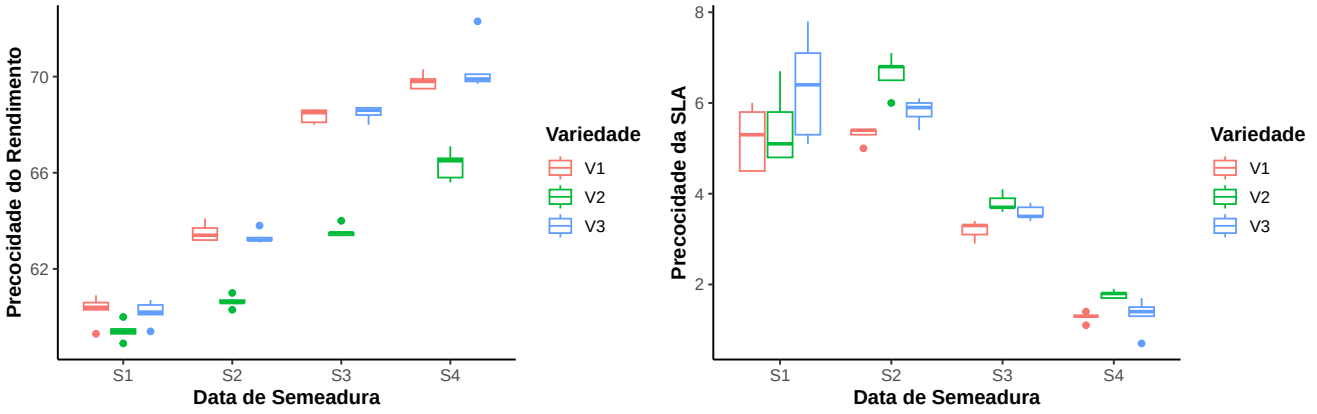
A Tabela 1 apresenta as medidas de resumo para cada combinação entre os fatores e, para cada uma das variáveis y_1, y_2, y_3 e y_4 foi calculada a média e o desvio padrão. Os valores que mais chamam atenção na Tabela 1 são os da variável y_4, apresentando médias distantes uma das outras a depender da iteração dos fatores.

Uma visualização proposta e muito eficaz na comparação de dados como os que esstão sendo analisados é o gráfico de caixa, mais conhecido como *boxplot*. Tal visualização foi desenhada na Figura 1.

Tabela 1: Medidas de Resumo das Variáveis em Análise segmentadas pelos Fatores.

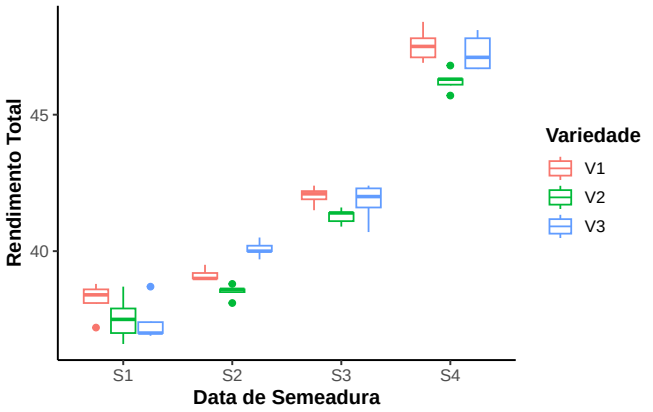
	y_1		y_2		y_3		y_4	
V	Média	DP	Média	DP	Média	DP	Média	DP
S1								
V1	60,30	0,6041523	5,22	0,7049823	38,22	0,6260990	317,0	19,735754
V2	59,42	0,3962323	5,44	0,8142481	37,54	0,8142481	297,2	20,166804
V3	60,18	0,4969909	6,34	1,1545562	37,40	0,7516648	305,6	6,188699
S2								
V1	63,52	0,3834058	5,30	0,1732051	39,14	0,2190890	279,4	8,905055
V2	60,64	0,2509980	6,64	0,4159327	38,52	0,2588436	258,0	6,123724
V3	63,32	0,2774887	5,82	0,2774887	40,08	0,2949576	290,2	7,190271
S3								
V1	68,36	0,2880972	3,20	0,2000000	42,02	0,3420526	280,4	7,266361
V2	63,56	0,2509980	3,80	0,2000000	41,28	0,2774887	247,0	12,000000
V3	68,48	0,2949576	3,58	0,1643168	41,80	0,6892024	287,6	8,532292
S4								
V1	69,80	0,3316625	1,28	0,1095445	47,54	0,5941380	251,6	15,126136
V2	66,32	0,6140033	1,78	0,0836660	46,24	0,3974921	225,8	13,627179
V3	70,36	1,0945319	1,32	0,3768289	47,28	0,6418723	242,4	11,148991

DP: Desvio Padrão.

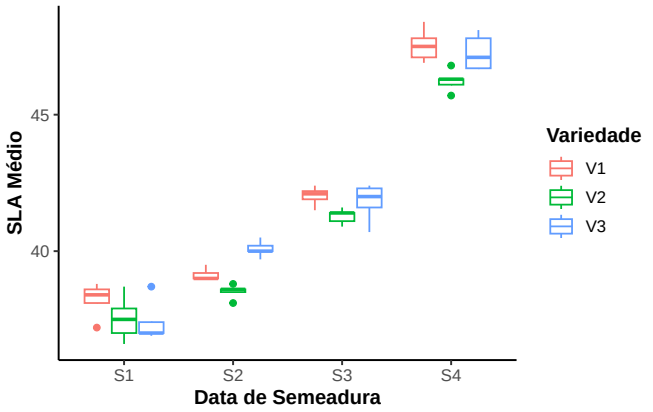


(a) Precocidade do Rendimento (y_1)

(b) Precocidade da SLA (y_2)



(c) Rendimento Total (y_3)



(d) SLA Médio (y_4)

Figura 1: Boxplot das Variáveis Segmentadas pelos Fatores.

1.3 Teste a interação e efeitos principais usando as quatro estatísticas de teste da MANOVA a 2 Fatores

A Tabela 2 apresenta os resultados obtidos para cada teste realizado. São apresentados os graus de liberdade, Estatística, F aproximado e o nível descritivo de cada teste.

1.4 Faça a análise descrevendo a metodologia, procedimentos e conclusão.

Para investigar se há diferenças significativas nas variáveis-resposta multivariadas entre os níveis dos fatores considerados, foi utilizada a Análise Multivariada da Variância (MANOVA). Esta técnica é apropriada quando se deseja avaliar o efeito de variáveis independentes categóricas sobre múltiplas variáveis dependentes quantitativas de forma conjunta, levando em consideração a correlação entre elas.

Neste estudo, foram considerados dois fatores principais, denominados `S` e `V`, além da interação entre eles (`S:V`). As variáveis-resposta foram avaliadas simultaneamente, e a hipótese nula testada para cada fator/interação foi a de que não há diferença significativa nas médias vetoriais das variáveis-resposta entre os grupos.

O modelo base da MANOVA é expresso por:

$$\mathrm{Y}_{kri} = \mu + S_k + V_r + (S \times V)_{kr} + \varepsilon_{kri},$$

foi conduzida com base em quatro estatísticas multivariadas clássicas:

- **Wilks' Lambda**

Tabela 2: Análise de Variància Multivariada a 2 Fatores.

Efeito	gl	Estatística	F $\approx$	gl_{num}	gl_{den}	Valor p	Signif.
Wilks							
S	3	0,000623	151,940	12	119,35	$<$ 2,2e-16	***
V	3	0,065574	32,682	6	96,00	$<$ 2,2e-16	***
S:V	6	0,135002	5,111	24	158,20	1,076e-10	***
Pillai							
S	3	2,359400	43,280	12	141,00	$<$ 2,2e-16	***
V	3	1,103600	14,157	8	92,00	9,032e-13	***
S:V	6	1,333800	4,002	24	192,00	2,757e-08	***
Hotelling							
S	3	146,107000	531,670	12	131,00	$<$ 2,2e-16	***
V	3	11,670000	64,190	8	88,00	$<$ 2,2e-16	***
S:V	6	3,501000	6,350	24	174,00	5,244e-14	***
Roy							
S	3	140,943000	1656,080	4	47,00	$<$ 2,2e-16	***
V	3	10,251000	121,450	4	28,00	$<$ 2,2e-16	***
S:V	6	2,686000	41,200	6	48,00	4,496e-12	***

Para $\alpha = 0$: '***'; $\alpha = 0.001$: '**'; $\alpha = 0.01$: '*'; $\alpha = 0.05$: '.'.

- **Pillai's Trace**
- **Hotelling–Lawley Trace**
- **Roy's Largest Root**

Esses testes avaliam a razão da variabilidade explicada pelo modelo em relação à variabilidade residual, considerando a estrutura multivariada dos dados. Para cada teste, foram avaliadas as estatísticas correspondentes, graus de liberdade, valores de F aproximados e respectivos valores-p.

A análise foi conduzida no ambiente R, com os resultados organizados em tabela e formatados com o pacote `gt`. A significância estatística foi constatada perante um $\alpha = 0,05$ (5%).

Com base nos resultados obtidos por todos os testes (Wilks, Pillai, Hotelling e Roy), observou-se que:

- **O fator `S` apresentou efeito altamente significativo sobre o conjunto das variáveis-resposta**, com valores-p inferiores a 0.001 em todos os testes. Isso indica que as médias vetoriais diferem entre os níveis de `S`.
- **O fator `V` também apresentou efeito significativo**, com valores-p inferiores a 0.001 nos quatro testes, indicando diferenças entre os níveis de `V` sobre as variáveis-resposta.
- **A interação `S:V` foi igualmente significativa**, sugerindo que o efeito combinado dos fatores altera de forma significativa a distribuição conjunta das variáveis-resposta.

Dessa forma, há evidências estatísticas fortes de que tanto os fatores principais quanto sua interação afetam significativamente o comportamento multivariado das variáveis dependentes analisadas.

2 Análise de Componentes Principais (ACP) em Dados de Heptatlo Feminino

2.1 Primeiros Passos - Leitura, Ajuste e Visualizações Primárias

A Análise de Componentes Principais (PCA) foi utilizada com o objetivo de explorar a estrutura multivariada dos dados do heptatlo feminino. Inicialmente, foi realizada a transformação das variáveis de tempo (`hurdles`, `run200m` e `run800m`), uma vez que, nessas provas, menores valores indicam melhor desempenho. A transformação consistiu em subtrair os tempos originais do maior tempo registrado, mantendo a coerência de interpretação com as demais variáveis (onde valores maiores são melhores).

Em seguida, os dados foram padronizados e submetidos à PCA. A matriz de correlação revelou padrões interessantes entre as variáveis. Foi possível observar correlações positivas entre algumas variáveis atléticas como `longjump`, `shot` e `javelin`, indicando um possível grupo de atletas com perfil mais técnico de força e explosão. Variáveis como `run800m` e `hurdles` apresentaram correlações distintas, sugerindo que podem representar outros aspectos do desempenho físico, como resistência e velocidade. Vejamos a Figura 2.

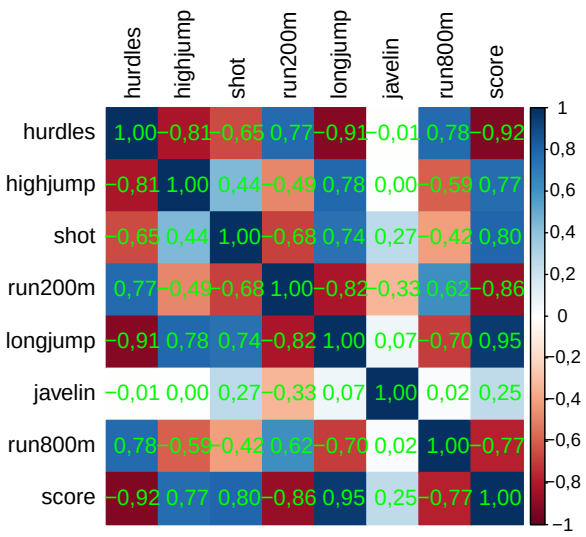


Figura 2: Mapa de Calor da Matriz de Correlação Linear de Perason.

Após isso, foi aplicada, de fato, a análise de componentes principais.

2.2 Gráficos de Apoio

Após a aplicação, a Figura 3 mostra o primeiro gráfico a ser analisado. Gráfico de Autovalores ou, popularmente *Scree Plot*, nos fornece um método para determinar o número de componentes principais.

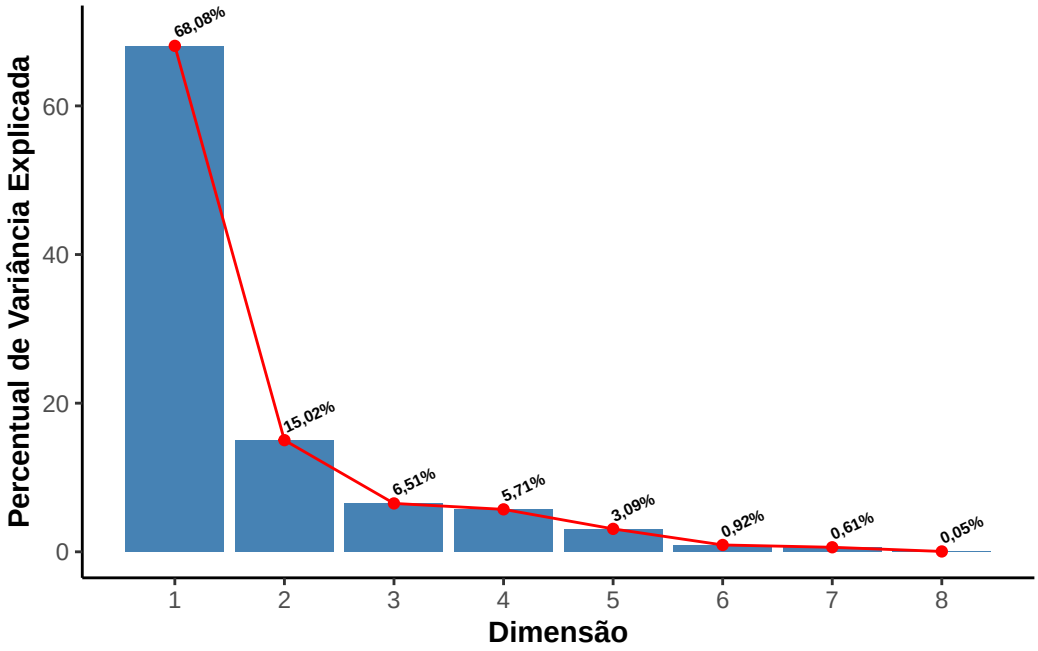


Figura 3: Gráfico de Autovalores.

Também, foi desenhada a Figura 4, que apresenta o *biplot* - visualização que sobrepõe o gráfico de indivíduos e o gráfico de cargas fatoriais. Usado para avaliar a estrutura dos dados e as cargas fatoriais dos primeiros dois componentes em um gráfico.

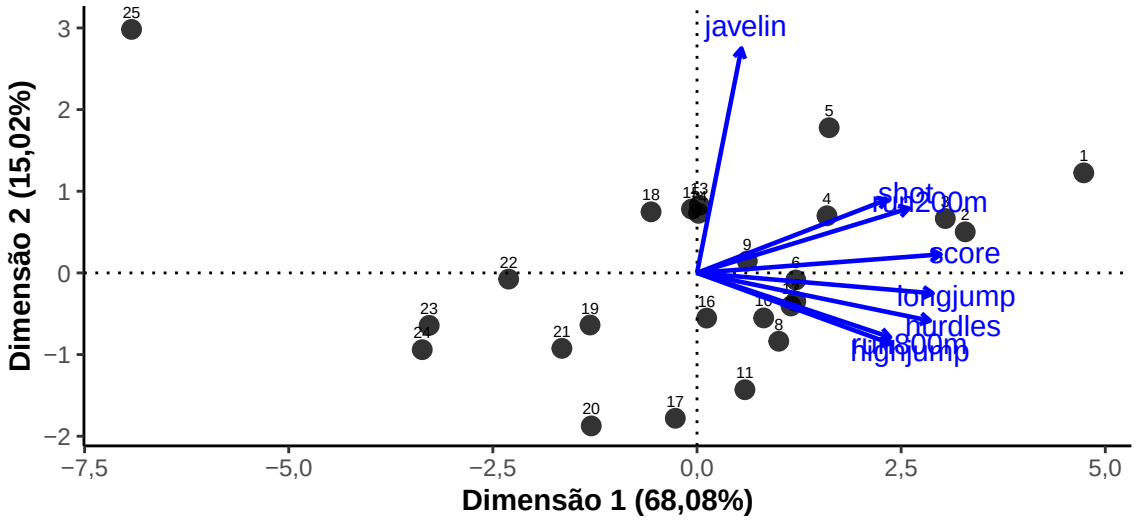


Figura 4: Biplot dos Componentes Principais 1 e 2.

Por fim, foi construída a Figura 5. Este gráfico mostra as variáveis que são mais correlacionadas com o CP_1 e CP_2 são as mais importantes para explicar a variabilidade dos dados. As variáveis que não são correlacionadas com nenhum CP ou correlacionadas com as últimas dimensões (últimos CP's) são variáveis com baixa contribuições e candidatas a serem removidas para simplificar a análise.

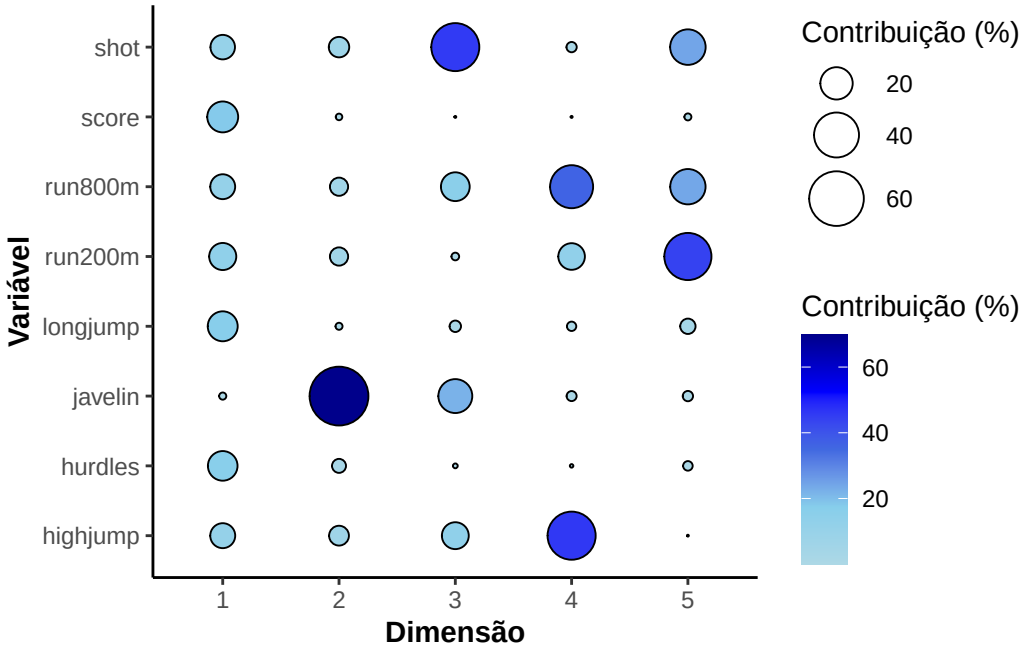


Figura 5: Gráfico de Contribuições das Variáveis nos Componentes Pricipais.

2.3 Conclusão da Análise de Componentes Principais

A Análise de Componentes Principais (ACP) aplicada aos dados do heptatlo feminino teve como objetivo explorar a estrutura de correlações entre variáveis de desempenho atlético e verificar como os escores das componentes se relacionam com a pontuação geral (`score`). A transformação das variáveis de tempo (`hurdles`, `run200m`, `run800m`) foi uma etapa essencial para alinhar a direção da interpretação — maiores valores sempre indicando melhor desempenho. Após a padronização, a ACP foi executada considerando as correlações entre as variáveis.

A matriz de correlação revelou dois grupos principais de variáveis:

- Um primeiro grupo composto por variáveis de potência/explosão, como `shot`, `longjump` e `highjump`, que apresentaram **forte correlação entre si**;
- Um segundo grupo mais voltado à resistência e velocidade, representado por `run800m`, `hurdles` e `run200m`, o qual mostrou **correlação negativa com o grupo anterior**, refletindo a natureza multidimensional do desempenho atlético.

O **scree plot** indicou que as **duas primeiras componentes principais explicam juntas cerca de 83% da variância total**, sendo a **Dimensão 1** (aproximadamente 68%) fortemente associada à variabilidade geral do desempenho técnico. A **Dimensão 2**, responsável por cerca de 15% da variância, representa aspectos secundários, mas ainda relevantes, ligados a atributos específicos.

O **biplot**, que tem como um de seus objetivos apresentar as **cargas fatoriais**, evidenciou nitidamente o agrupamento de algumas variáveis. Pode-se notar que todas as variáveis exercem influência na Dimensão 1, porém, com bem menos influência da variável `javelin`. Em contrpartida, as variáveis `highjump`, `run800m`, `hurdles` e `longjump` são mais influentes.

O **biplot** também mostra que grande parte das observações estou centradas. Talvez, com mais rigidez, possamos dizer que a amostra 1 possa ser um *outlier*, pois está mais distante da maioria. Entretanto, sem fazer muito esforço pode-se idetificar que a observação 25 pode ser problemática, já que a sua distância perante as demais observações é bastante evidente.

O **gráfico das contribuições das variáveis** também trouxe um panorama claro: variáveis como `score`, `longjump` e `hurdles` foram as que **mais contribuíram para a formação da Dimensão 1**, enquanto `javelin` é a que contribuí significativamente para a Dimensão 2, reforçando a dualidade técnica da estrutura dos dados.

3 Para Dados de Progênies de Eucalyptus sp