

GABRIELA NUNES DA PIEDADE

**ANÁLISE DE COMPONENTES PRINCIPAIS MODIFICADAS APLICADAS A
DADOS DE PROGRAMAS DE MELHORAMENTO GENÉTICO DE PLANTAS**

Botucatu

2024

GABRIELA NUNES DA PIEDADE

**ANÁLISE DE COMPONENTES PRINCIPAIS MODIFICADAS APLICADAS A
DADOS DE PROGRAMAS DE MELHORAMENTO GENÉTICO DE PLANTAS**

Tese apresentada à Faculdade de
Ciências Agronômicas da Unesp Câmpus
de Botucatu, para obtenção do título de
Doutora em Agronomia/Agricultura.

Orientadora: Dra. Maria Márcia Pereira
Sartori

Coorientadora: Dra. Adelaide de Fátima
Baptista Valente Freitas

Botucatu

2024

P613a Piedade, Gabriela Nunes da
Análise de componentes principais modificadas aplicadas a dados de programas de melhoramento genético de plantas / Gabriela Nunes da Piedade. -- Botucatu, 2024
94 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrônômicas, Botucatu
Orientadora: Maria Márcia Pereira Sartori
Coorientadora: Adelaide de Fátima Baptista Valente Freitas

1. Plantas Melhoramento genético. 2. Análise multivariada. 3. Principal components analysis. I. Título.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ANÁLISE DE COMPONENTES PRINCIPAIS MODIFICADAS APLICADAS A DADOS DE PROGRAMAS DE MELHORAMENTO GENÉTICO DE PLANTAS

AUTORA: GABRIELA NUNES DA PIEDADE

ORIENTADORA: MARIA MÁRCIA PEREIRA SARTORI

COORIENTADORA: ADELAIDE DE FÁTIMA BAPTISTA VALENTE FREITAS

Aprovada como parte das exigências para obtenção do Título de Doutora em Agronomia (Agricultura), pela Comissão Examinadora:

Pesquisadora Dr.^a MARIA MÁRCIA PEREIRA SARTORI (Participação Virtual)
Produção Vegetal / Faculdade de Ciências Agrômicas de Botucatu UNESP

AP 

Prof.^a Dr.^a AMANDA RITHIELI PEREIRA DOS SANTOS (Participação Virtual)
Engenharia Agrícola / Universidade Federal do Pampa

AP 

Prof. Dr. RENATO FERNANDES CANTÃO (Participação Virtual)
Física, Química e Matemática / Universidade Federal de São Carlos

AP 

Prof. Dr. EVANDRO VAGNER TAMBARUSSI (Participação Virtual)
Engenharia Florestal / Universidade Estadual do Centro-Oeste

AP 

Prof. Dr. JOSÉ BRUNO MALAQUIAS (Participação Virtual)
Fitotecnia e Ciências Ambientais / Universidade Federal da Paraíba

AP 

Botucatu, 03 de junho de 2024

Dedico,

*Ao meu padrinho, Aginaldo Nunes (in memoriam)
 e a minha Tia Laura (in memoriam)*

*A todas as mulheres, especialmente às negras, que enfrentam
preconceitos desde o nascimento. Quebrar barreiras, inclusive
educacionais, pode parecer difícil, mas acreditem em sua força!*

AGRADECIMENTOS

A Deus, em quem deposito minha fé e devoção. Como cristã, reconheço sua presença constante em minha vida, sobretudo ao longo destes últimos seis anos, onde nossa intimidade atingiu um nível indescritível, em momentos em que éramos somente Ele e eu. Sua graça divina sustentou-me, e sem ela, eu não estaria aqui hoje!

Aos meus queridos pais, Socorro e Tôxa, e avós Francisquinha, Bispo e Neves (*in memoriam*), por serem meu refúgio e fortaleza, por mesmo na simplicidade, me doarem o mais puro e valioso sentimento: o amor. Por abdicarem dos seus planos para fazerem o possível pelos meus, me apoiando e incentivando. E por acreditarem em mim, até mesmo quando eu não acreditei. Eu amo vocês, eternamente!

A minha família Nunes e Piedade, que tanto nos momentos felizes – aniversários, feriados, almoços de fim de semana – como nos tristes – perdas de avós, bisavós, tia Laura e padrinho Aguinaldo – me acolhiam da forma como a distância permitia, e jamais esquecendo do sentido que é “ser família”.

Aos meus pequenos sobrinhos e afilhados que a cada reencontro, renovavam as minhas esperanças com toda a doçura e divindade que uma criança é capaz de transmitir.

A querida professora Maria Márcia, que além de excelente orientadora, com sua educação, gentileza e respeito, por muitas vezes se mostrou amiga e até mãe, indo além do que sua função exige. Sua personalidade única, nas reuniões do grupo de pesquisa no departamento até em momentos de descontração nos churrascos em sua casa, evidencia a pessoa íntegra e competente, merecedora de todo sucesso e reconhecimento. Como mulher, me inspira a alcançar lugares que sequer cogitei. Com certeza levarei comigo seu modo de liderar e quem sabe, de viver. Obrigada, prof.!

A professora Adelaide, pela valiosa orientação. Sempre acreditei que a educação me levaria longe, mas nunca imaginei que ela me conduziria até Portugal e, ainda mais, à oportunidade de colaborar com uma profissional tão renomada desse país. Professora, sua

generosidade em compartilhar tanto conhecimento de maneira tão clara e organizada (onde tudo se transforma em gráficos) foi fundamental para a conclusão desta tese. Obrigada por fazer parte deste meu sonho!

Aos meus amigos, minha verdadeira família em Botucatu: Gyslane, Amanda, Michelane, Brunna, Isabelly, Politon, Dandára e Mônica. Só a Educação para me juntar a pessoas de lugares e conseqüentemente, culturas tão diferentes, mas com os mesmos propósitos de vida. Foi uma delícia dividir esses anos de vida com vocês. Obrigada pela amizade de cada um!

Ao grupo de pesquisa ModelStat.

A FCA/UNESP, professores e funcionários do Departamento de Produção Vegetal e da Biblioteca.

Ao meu namorado Jhonilson Almeida e sua família.

Aos meus amigos do Maranhão e conterrâneos de Nina Rodrigues.

A minha psicóloga Michele.

Ao meu estimado orientador de iniciação científica, professor Jefferson Siqueira.

Ao grupo de oração Levante Jovem e a Paróquia Sagrado Coração de Jesus.

A cidade de Botucatu.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“Quer saber quanto vale uma saudade? Tenha amor, queira bem e viva ausente”!

BESSA, B. Poesia com rapadura. 1. ed. Fortaleza: Cene, 2017, 152 p.

RESUMO

Os estudos para a obtenção de novas cultivares com características agronômicas desejáveis no melhoramento genético, são constituídos por um grande número de acessos, seguidos por longos períodos de experimentação. Estes conjuntos de dados são tradicionalmente avaliados por meio de análises univariadas, as quais não levam em consideração a dependência que pode existir entre as variáveis. A CDpca (*Clustering and Disjoint Principal Component Analysis*) apresenta-se como uma modificação recente da Análise de Componentes Principais, com a finalidade de facilitar a interpretação destes dados. Assim, o objetivo deste trabalho foi aplicar a CDpca em dados agronômicos da área de melhoramento genético de plantas e aprimorar e/ou adequar essa técnica. A *Clustering and Disjoint Principal Component Analysis* (CDpca) quando comparada a Análise de Componentes Principais (ACP), apresenta algumas vantagens como a facilidade de interpretação de seus resultados. Entretanto, ao comparar os resultados de ambas as técnicas, algumas divergências foram encontradas como: porcentagem de explicação das componentes, variáveis selecionadas para cada componente bem como a escolha de objetos. Outrossim, divergências maiores foram encontradas quando se utiliza três componentes, caso em que a modificação proposta se mostrou como alternativa para tentar solucionar esta questão. Todavia, é interessante ressaltar a necessidade de estudos complementares para melhor entendimento e utilização da CDpca em estudos de melhoramento genético.

Palavras-chave: análise de componentes principais; seleção; análise multivariada.

ABSTRACT

Studies to obtain new cultivars with desirable agronomic characteristics in genetic improvement consist of a large number of accessions, followed by long periods of experimentation. These data sets are traditionally evaluated through univariate analyses, which do not take into account the dependence that may exist between variables. CDpca (Clustering and Disjoint Principal Component Analysis) is a recent modification of Principal Component Analysis, with the purpose of facilitating the interpretation of these data. Thus, the objective of this work was to apply CDpca to agronomic data from the area of plant genetic improvement and to improve and/or adapt this technique. Clustering and Disjoint Principal Component Analysis (CDpca) when compared to Principal Component Analysis (PCA), presents some advantages such as ease of interpretation of its results. However, when comparing the results of both techniques, some divergences were found such as: percentage of explanation of the components, variables selected for each component as well as the choice of objects. Furthermore, greater discrepancies were found when using three components, in which case the proposed modification proved to be an alternative to try to solve this issue. However, it is interesting to highlight the need for complementary studies to better understand and use CDpca in genetic improvement studies.

Keywords: principal component analysis; selection; multivariate analysis.

SUMÁRIO

INTRODUÇÃO GERAL.....	17
CAPÍTULO 1 – ANÁLISE DE COMPONENTES PRINCIPAIS E CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS.....	20
1.1 INTRODUÇÃO.....	20
1.2 ANÁLISE DE COMPONENTES PRINCIPAIS.....	20
1.2.1 Estrutura do Conjunto de Dados.....	21
1.2.2 Forma de Obtenção das Componentes Principais.....	22
1.2.3 Aplicação em Dados Agronômicos.....	24
1.2.4 Limitações da Análise de Componentes Principais.....	25
1.3 <i>Clustering and Disjoint Principal Component Analysis.....</i>	<i>26</i>
1.3.1 Notações.....	27
1.3.2 O modelo.....	28
1.4 APLICAÇÕES DA CDpca.....	30
1.5 CONCLUSÃO.....	31
REFERÊNCIAS.....	32
CAPÍTULO 2 – CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS – UMA NOVA ABRODAGEM PARA OS ESTUDOS DE MELHORAMENTO GENÉTICO VEGETAL.....	34
2.1 INTRODUÇÃO.....	34
2.2 MATERIAL E MÉTODOS.....	35
2.2.1 O Conjunto de Dados.....	35
2.2.2 Pré-processamento dos Dados.....	35
2.2.3 Descrição dos Experimentos Computacionais.....	36
2.2.4 Definição dos Parâmetros.....	37
2.3 RESULTADOS E DISCUSSÃO.....	37
2.3.1 Definição do P e Q.....	37
2.3.2 CDpca.....	43
2.4 CONCLUSÃO.....	46

REFERÊNCIAS	47
CAPÍTULO 3 – ANÁLISE DE COMPONENTES PRINCIPAIS E CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS EM DADOS DE MELHORAMENTO GENÉTICO VEGETAL.....	49
3.1 INTRODUÇÃO.....	49
3.2 MATERIAL E MÉTODOS.....	52
3.2.1 Descrição dos conjuntos de dados.....	52
3.2.2 Análise de Componentes Principais.....	53
3.2.3 <i>Clustering and Disjoint Principal Component Analysis.....</i>	<i>54</i>
3.3 RESULTADOS E DISCUSSÃO.....	55
3.3.1 DADOS 1.....	55
3.3.2 DADOS 2.....	62
3.3.3 PROPOSTA DE MODIFICAÇÃO NA CDpca.....	69
3.4 CONCLUSÃO.....	70
REFERÊNCIAS	72
CONSIDERAÇÕES FINAIS.....	74
REFERÊNCIAS	75
APÊNDICE A.....	77

INTRODUÇÃO GERAL

No melhoramento de plantas, como em outras áreas do conhecimento (genômica, *machine learning*), é usual avaliar as unidades de amostragem (plantas, progênies, populações, acessos, entre outros.), para vários caracteres ou atributos. Tradicionalmente, o tratamento estatístico de cada caráter ou variável, isoladamente, caracteriza o que se conhece por análise univariada.

Esse tipo de análise, embora de grande utilidade, quando associada a um processo de seleção de genótipos, pode resultar em um material não completamente desejável (Godoi, 1985), uma vez que, a seleção com base em um ou poucos caracteres ocasione em materiais com desempenho não tão favorável para caracteres não considerados. Isto pode levar à rejeição precoce do material por parte dos melhoristas, agricultores, consumidores e indústrias.

Pode-se acrescentar que, resultados de análises univariadas para um grande número de caracteres são de difícil manipulação quando se objetiva identificar os genótipos com as melhores combinações de características, o que torna o procedimento praticamente inexequível, especialmente quando o número de genótipos é elevado.

Pimentel Gomes (1990) afirma que, somente havendo independência entre as variáveis, as análises univariadas resolveriam o problema. Assim, para aumentar a probabilidade de êxito num programa de melhoramento, deve-se combinar grande número de caracteres, com análise estatística eficiente em extrair as informações principais contidas numa grande massa de dados.

Desta forma, técnicas de redução de espaço e recursos para seleção de subconjuntos são de grande importância, uma vez que facilitam a visualização e capturam as associações relevantes existentes no conjunto de dados.

As análises multivariadas apresentam-se como instrumento poderoso no estudo de dados, em razão de permitirem combinar as informações múltiplas tomadas nas unidades de amostragem. No melhoramento de plantas, essas técnicas possibilitam prever fenômenos genéticos como mutação, fluxo gênico, reprodução sexuada, deriva genética, baseando-se, simultaneamente em um complexo de variáveis de importância tanto genética quanto evolutiva (Cruz, 1990).

Johnson e Wichern (1992) relacionam os seguintes objetivos aos métodos multivariados de análise: i) redução do volume de dados, representando-os de forma

mais simples possível (interpretação mais fácil), sem sacrifício de informações valiosas; ii) agrupamento e classificação (criação de grupos de objetos ou variáveis similares; ou alternativamente, alocação de objetos a grupos já bem definidos) ; iii) investigação da interdependência entre variáveis (natureza e grau de relacionamento entre variáveis, ou predição dos valores de uma ou mais variáveis com base em observações feitas sobre outras); e iv) construção de hipóteses e testes estatísticos a respeito de parâmetros de populações multivariadas.

As técnicas de maior utilização compreendem um conjunto de métodos estatísticos exploratórios, cujos objetivos centram-se especialmente no estudo da diferenciação entre objetos ou indivíduos, e no estudo do relacionamento entre variáveis (Duarte, 1997).

Embora a aplicação em larga escala seja relativamente recente, as principais técnicas de análise multivariada foram propostas há bastante tempo. Na realidade, todo o desenvolvimento teórico somente alcançou popularidade com o avanço da informática, porque os métodos multivariados, geralmente, exigem operações algébricas com matrizes inexecutáveis de serem implementadas manualmente para grandes volumes de dados, próprios de pesquisas aplicadas (Alencar, 2019). Atualmente, diante dos recursos computacionais disponíveis, esse entrave já não mais existe.

A análise de componentes principais (ACP), que se constitui num procedimento básico do qual derivam vários outros métodos multivariados, foi concebida por Karl Pearson, em 1901 (Souza, 1988). Já na década de 30 a técnica estava completamente consolidada, com os artigos de Harold Hotelling (1933, 1935 e 1936).

A ACP reduz a dimensionalidade dos conjuntos de dados, aumentando a interpretabilidade, mas ao mesmo tempo minimiza a perda de informações. Essa técnica cria variáveis chamadas de componentes principais (CP), não correlacionadas, que são definidas pelo conjunto de dados em questão e onde há a maximização sucessiva da variância, tornando-a uma análise de dados adaptativa (Jolliffe e Cadima, 2016). Por tais características, tem sido amplamente utilizada com diferentes objetivos.

Oliveira, Gravina e Oliveira (2018), utilizaram a ACP como ferramenta para seleção de genótipos de feijão com características agronômicas desejáveis. Esta também já foi utilizada em estudos de divergência fenotípica em linhagens de milho (GReveniotis *et al.*, 2018), para identificar acessos de arroz com desempenho superior

quando submetidos a temperaturas abaixo do ideal durante a fase de desenvolvimento inicial (Moura *et al.*, 2018) e em estudos para identificar e separar diferentes cultivares de soja (Panero *et al.*, 2018).

Entretanto, tal como acontece com outras técnicas estatísticas, muitas adaptações da metodologia básica da ACP para diferentes tipos de dados e estruturas, foram desenvolvidos (Lee *et al.*, 2010; Birnbaum *et al.*, 2013), visando torná-la mais robusta.

Algumas dessas modificações são simplificações das CPs, com o intuito de ajudar na interpretação dos resultados (Jolliffe e Cadima, 2016), como por exemplo, os estudos realizados para decidir os critérios e a quantidade de componentes a serem selecionados. Outras exploram métodos nos quais as funções de dados são vistas como soluções para equações diferenciais, uma abordagem conhecida como análise diferencial (Ramsay e Silverman, 2006).

Nesse contexto, uma nova metodologia foi proposta por Vichi e Saporta (2009), a análise restrita de componentes principais (*Clustering and Disjoint Principal Component Analysis* - CDpca), que visa um agrupamento simultâneo de objetos e um particionamento de variáveis. Ela permite identificar componentes com variância máxima, para cada combinação linear de um subconjunto de variáveis. Simultaneamente, uma partição de objetos também é calculada maximizando a variação entre os agrupamentos. Outra vantagem importante é a facilidade de interpretação dos componentes, uma vez que cada um é caracterizado por um conjunto disjunto de variáveis. Assim como não é necessário usar um método de rotação para melhorar a interpretabilidade dos resultados.

Essa técnica, apesar dos benefícios citados anteriormente, é pouco utilizada, visto que seu complexo sistema ainda não está disponível em softwares executáveis e *scripts* prontos disponíveis em *softwares* programáveis, são escassos e recentes, tornando-se desse modo, fator limitante para sua ampla difusão no meio acadêmico. Quanto aos estudos no melhoramento genético do Brasil, sua aplicação inexistente ou é escassa na literatura.

Assim, o objetivo deste trabalho foi: (i) apresentar a CDpca através de uma revisão bibliográfica; (ii) demonstrar o desempenho do algoritmo descrito por Macedo e Freitas (2015) para a execução da CDpca e; (iii) comparar os resultados da CDpca com aqueles obtidos a partir da ACP clássica, quando ambas as metodologias são aplicadas para avaliar experimentos agronômicos.

CAPÍTULO 1

ANÁLISE DE COMPONENTES PRINCIPAIS E CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS

1.1 INTRODUÇÃO

A interpretação de dados complexos em diversas disciplinas científicas é crucial para o avanço do conhecimento e a compreensão dos fenômenos naturais. Nesse contexto, a análise estatística emerge como uma ferramenta essencial e comprometida com a abordagem científica. O seu papel consiste em transformar o cenário inicial de incerteza contido nos dados, em conhecimento consolidado (Freitas, 2022), alinhando-se ao propósito fundamental da ciência de ampliar o entendimento humano sobre o mundo que nos cerca.

No campo da estatística multivariada, a Análise de Componentes Principais (ACP) introduzida por Pearson no início do século XX, tem sido amplamente utilizada para reduzir a complexidade de conjuntos de dados, identificando as variáveis mais influentes e destacando as relações latentes entre observações (Hongyu, *et al.*, 2015). Em contrapartida, a CDpca surge como uma extensão valiosa, incorporando informações para aprimorar a capacidade de discriminação em contextos específicos.

Esta revisão bibliográfica tem como objetivo explorar as nuances e aplicações dessas duas técnicas estatísticas, delineando suas características.

A seguir, serão abordados tópicos relacionados ao entendimento de ambas as técnicas.

1.2 ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (ACP) ou *Principal Component Analysis* (PCA) foi concebida por Karl Pearson (1901) e reformulada por Hotteling (1933). Constitui-se em uma das principais técnicas de análise multivariada, da qual derivam vários outros métodos multivariados (Souza, 1988).

Sua utilização prática é considerada sob o objetivo da redução do espaço paramétrico, simplificando estruturalmente os dados amostrais, de forma que diferenças entre tratamentos, em princípio influenciadas por um conjunto maior de variáveis, possam ser avaliadas em espaços bidimensionais ou tridimensionais de fácil interpretação geométrica (Cruz *et al.*, 2012).

Este método estatístico multivariado permite transformar um conjunto de variáveis iniciais correlacionadas entre si, em um outro conjunto de variáveis não correlacionadas (independentes/ortogonais), chamadas de componentes principais, que resultam de combinações lineares do conjunto inicial. O propósito desta análise consiste em determinar o menor número de componentes principais que consiga explicar a maior variabilidade possível dos dados (Silva, 2005).

1.2.1 Estrutura do Conjunto de Dados

Para realizar a ACP, as médias das unidades amostrais (n) devem estar dispostas em linhas e as variáveis (p) em colunas (Tabela 1).

Tabela 1 – Estrutura dos dados para aplicação da Análise de Componentes Principais (ACP)

Unidade	Variáveis			
Amostral	Y_1	Y_2	...	Y_p
1	y_{11}	y_{12}	...	y_{1p}
2	y_{21}	y_{22}	...	y_{2p}
3	y_{31}	y_{32}	...	y_{3p}
.	.	.		.
.	.	.	...	.
.	.	.		.
n	y_{n1}	y_{n2}	...	y_{np}

Não há consenso entre os estatísticos sobre o tamanho da amostra, mas os melhores resultados ocorrem em análises de amostras grandes (Osborne e Costello, 2004).

As variáveis devem ser numéricas e contínuas e não pode haver valores ausentes nos dados. Além disso, não dependem de qualquer suposição sobre distribuição de probabilidade, ou seja, não requerem uma aceitação normal multivariada (Jolliffe, 2002 e Johnson e Wichern, 2007).

Se as variáveis apresentarem unidades de medidas diferentes, devem ser padronizadas (expressando cada variável em unidades de desvio padrão em relação a uma média igual a zero) e utilizada a matriz de covariância. Caso não seja feita a

padronização, utiliza-se a matriz de correlação. Se os dados não são padronizados a análise será tendenciosa, favorável às variáveis com mais alta variância.

Apesar da ACP encontrar o método matematicamente ótimo (no sentido de minimizar o erro quadrático), ela é sensível a *outliers* nos dados, os quais produzem grandes erros que a ACP tenta evitar. Portanto, é de praxe removê-los.

1.2.2 Forma de Obtenção das Componentes Principais

Seja $X = (X_1, X_2, \dots, X_p)$ um vetor aleatório com vetor de médias $\mu = (\mu_1, \mu_2, \dots, \mu_p)$ e matriz de covariâncias $\Sigma_{p \times p}$. Sejam $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ os autovalores da matriz $\Sigma_{p \times p}$, com os respectivos autovetores normalizados $e_1, e_2, \dots, e_p$, isto é, os autovetores e_i satisfazem as seguintes condições:

- (i) $e_i' e_j = 0$ para todo $i \neq j$
- (ii) $e_i' e_i = 1$ para todo $i = 1, 2, \dots, p$
- (iii) $\Sigma_{p \times p} e_i = \lambda_i e_i$, para todo $i = 1, 2, \dots, p$

Sendo o autovetor e_i denotado por $e_i = (e_{i1}, e_{i2}, \dots, e_{ip})'$. Considere o vetor aleatório $Y = O'X$, onde $O_{p \times p}$ é a matriz ortogonal de dimensão $p \times p$, constituída dos autovetores normalizados da matriz $\Sigma_{p \times p}$, isto é:

$$O_{p \times p} \begin{bmatrix} e_{11} & e_{21} & \dots & e_{p1} \\ e_{12} & e_{22} & \dots & e_{p2} \\ \vdots & \vdots & \dots & \vdots \end{bmatrix} = [e_1 \ e_2 \ \dots \ e_p].$$

O vetor Y é composto de p combinações lineares das variáveis aleatórias do vetor X , com vetor de médias igual a $O' \mu$ e matriz de covariâncias $\Lambda_{p \times p}$, que é uma matriz diagonal, cujos elementos são iguais a $a_{ii} = \lambda_i$, $i = 1, 2, \dots, p$ isto é:

$$\Lambda_{p \times p} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

Portanto, as variáveis aleatórias que constituem o vetor Y são não correlacionadas entre si. Deste modo, surge a ideia de utilizar as combinações lineares em Y , como uma forma alternativa de se representar a estrutura de

covariâncias do vetor X , tentando obter uma redução do espaço de variáveis, passando da dimensão p , para uma dimensão k menor que p . Assim, ao invés de se utilizar o vetor aleatório original na análise de dados, utiliza-se as k combinações lineares principais. Os vetores aleatórios X e Y têm a mesma variância total e generalizada, sendo que o vetor Y tem a vantagem de ser composto por variáveis aleatórias não correlacionadas, facilitando, portanto, a interpretação conjunta dessas.

Assim:

- 1) A j -ésima componente principal da matriz $\Sigma_{p \times p}$, $j = 1, 2, \dots, p$ é definida como:

$$Y_j = e_j' X = e_{j1} X_1 + e_{j2} X_2 + \dots + e_{jp} X_p.$$

- 2) A variância da componente Y_j é:

$$\text{Var} [Y_j] = e_j' \Sigma_{p \times p} e_j = \lambda_j.$$

sendo $\text{cov} [Y_j, Y_k] = 0$, $j \neq k$. Cada autovalor λ_j representa a variância de uma componente principal Y_j . Como os autovalores estão em ordem decrescente, a primeira componente é a de maior variabilidade e a próxima é de menor.

- 3) A proporção de variância total que X é explicada pela j -ésima CP é:

$$\frac{\text{Var} [Y_j]}{\text{Var Total de } X} = \frac{\lambda_j}{\text{Traço} (\Sigma_{p \times p})} = \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}.$$

Para identificar o quanto cada variável explica cada componente principal (CP) e assim definir as variáveis que mais contribuíram na formação de cada CP selecionado, analisa-se seus respectivos autovetores, que representam os coeficientes do CP associados positivamente ou negativamente à cada variável original. A partir destes, é possível obter os escores, que são os valores assumidos de cada CP para os objetos estudados e assim identificar qual tratamento (híbrido por exemplo) melhor se destacou em relação a cada CP.

Os *bipLOTS* têm o objetivo de representar graficamente resultados de ACP, tornando-se possível plotar no mesmo gráfico os autovetores e os escores de forma bidimensional, além de verificar as correlações entre as variáveis (Gabriel, 1971). Variáveis que apresentam vetores com ângulos menores que 90° ($\alpha < 90^\circ$) são

positivamente correlacionados, enquanto aquelas com ângulos maiores que 90° ($\alpha > 90^\circ$) são negativamente correlacionados. Já quando o ângulo é igual a 90° ($\alpha = 90^\circ$), não apresentam correlação (Yan e Fregeau-Reid, 2008).

Pela proporção de explicação da variância total, que o modelo de k componentes principais é responsável, pode-se determinar o número de componentes que se deve reter. Em muitos casos, adota-se modelos que expliquem pelos menos 70% da variação total (Johnson e Wichern, 1998). Outro método utilizado para selecionar os CPs é o critério de Kaiser (1958), o qual determina que o número de componentes deve ser igual ao número de autovalores (variâncias dos CPs), maiores ou iguais à média das variâncias das variáveis analisadas. Quando a ACP é realizada sobre a matriz de correlação, com variáveis padronizadas, esse critério considera que os CPs que apresentam autovalores > 1 ($\lambda_i > 1$), são suficientes para representarem a variabilidade total presente nos dados originais.

Com o uso do *scree plot*, o qual ordena os autovalores do maior para o menor, é possível definir através do ponto da inclinação da curva, o número de CPs que devem ser considerados para análise ($\lambda_i > 1$).

1.2.3 Aplicação em Dados Agronômicos

Em linhas gerais, os métodos de estatística multivariada são utilizados com o propósito de simplificar ou facilitar a interpretação do fenômeno que está sendo estudado através da construção de índices ou variáveis alternativas que sintetizem a informação original dos dados; construir grupos de elementos amostrais que apresentem similaridade entre si; comparar populações; investigar as relações de dependência entre as variáveis respostas associadas aos componentes (Mingoti, 2005).

Por tais características, a ACP tem sido amplamente utilizada com diferentes objetivos: Oliveira, Gravina e Oliveira (2018), utilizaram a ACP como ferramenta para seleção de genótipos de feijão com características agronômicas desejáveis. O estudo seguiu um delineamento experimental em blocos casualizados, com 4 repetições e 30 genótipos. A análise de componentes principais foi conduzida para investigar 8 variáveis: o peso médio da vagem (MPW), número de sementes por vagem (NSP), comprimento da vagem (PL), largura da vagem (PWi), vagens totais (MNP), peso de cem sementes (W100S), rendimento da vagem (PY) e rendimento de grãos (GY). O

PL, NSP, MPW e W100S apresentaram correlação positiva com GY. A análise multivariada sugeriu a seleção das linhas UENF 7-10-1, UENF 7-12-1 e UENF 7-20-1, que apresentou bom rendimento de sementes, bem como de UENF 14-22-3 e UENF 14-23-3, que mostraram aptidão para produção. As linhas localizadas no Grupo II (UENF 14-4-3, UENF 15-6-4, UENF 7-4-1, UENF 7-9-1, UENF 7-7-1, UENF 7-5-1) não apresentaram bom desempenho nas características medidas, podendo ser descartadas.

Com o objetivo de selecionar híbridos de mamona com base na correlação entre as variáveis analisadas, Piedade (2019), utilizou como ferramenta a análise de componentes principais. A avaliação foi realizada utilizando-se 31 híbridos distribuídos em delineamento experimental em blocos casualizados, com 4 repetições, sob as seguintes variáveis: altura média da planta (AP), altura do primeiro cacho (APC), número de nós do caule (NN), número de cachos (NC), número de sementes (NS), diâmetro do caule (DC), número de frutos (NF), massa de 100 grãos (M100) e teor de óleo das sementes (TO). Os três primeiros componentes principais (CPs) explicaram aproximadamente 75,01% de toda a variação contida nas nove variáveis estudadas. O CP1, CP2 e CP3 foram caracterizados como sendo principalmente relacionados à produtividade de grãos de mamona (NS, NC, M100 e NF), ao porte das plantas (DC, APC e AP) e à produção de óleo (TO), respectivamente. Os híbridos 14 e 23 são os mais adequados para produção de grãos em escala comercial, uma vez que apresentaram porte médio, favorecendo a colheita mecanizada. O híbrido comercial 26 apresentou alta produção de óleo, média produção de grãos e médio porte.

A ACP também já foi utilizada em estudos de divergência fenotípica em linhagens de milho (Greveniotis *et al.*, 2018), para identificar acessos de arroz com desempenho superior quando submetidos a temperaturas abaixo do ideal durante a fase de desenvolvimento inicial (Moura *et al.*, 2018) e em estudos para identificar e separar diferentes cultivares de soja (Panero *et al.*, 2018).

1.2.4 Limitações da Análise de Componentes Principais

Tal como acontece com outras técnicas estatísticas, é possível obter em algumas observações um efeito desproporcional sobre os resultados de uma ACP. Assim, apesar de ser uma técnica bastante expressiva, muitas adaptações da

metodologia básica para diferentes tipos de dados e estruturas, foram desenvolvidos (Lee *et al.*, 2010; Birnbaum *et al.*, 2013) visando torná-la mais robusta.

Em estudos de avaliação da divergência genética, por exemplo, no qual procura-se dimensionar a diversidade e agrupar o material, permitindo manejo e uso mais adequados das coleções (Cruz *et al.*, 2012 e Figueiredo Neto *et al.*, 2004), duas técnicas são usualmente utilizadas: a ACP, seguida por uma metodologia de agrupamento. No entanto, esta metodologia de análise pode produzir resultados ambíguos, uma vez que os componentes da ACP podem explicar a variação que mascara a estrutura de agrupamento nos dados (Vichi e Kiers, 2001).

Duas técnicas alternativas têm sido propostas na literatura para classificar simultaneamente objetos e obter uma redução na dimensionalidade de variáveis: k-médias fatoriais (Vichi e Kiers, 2001), a qual minimiza o desvio dentro do agrupamento na redução do espaço, sendo particularmente útil quando o agrupamento de objetos é o principal objetivo da análise. No entanto, nesta redução de dimensionalidade, a explicação da maior variância pode não ser garantida. Por outro lado, em k-médias reduzidas, a maximização do desvio entre o espaço reduzido é útil quando as dimensões que explicam a variância máxima são necessárias. No entanto, em agrupamentos com k-médias reduzidas, a variação mínima pode não ser garantida. Desta forma, ambas as técnicas podem induzir possíveis complicações na interpretação dos componentes.

Nesse contexto, uma nova metodologia foi proposta por Vichi e Saporta (2009), a análise restrita de componentes principais (*Clustering and Disjoint Principal Component Analysis* - CDpca), que visa um agrupamento simultâneo de objetos e um particionamento de variáveis. Esta análise permite identificar componentes com variância máxima, para cada combinação linear de um subconjunto de variáveis. Simultaneamente, uma partição de objetos também é calculada maximizando a variação entre os agrupamentos.

1.3 CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS

A *Clustering and Disjoint Principal Component Analysis* (CDpca) é uma análise de componentes principais restrita proposta por Vichi e Saporta (2009), para agrupamento dos objetos e particionamento das variáveis de forma simultânea.

1.3.1 Notações

$X = (x_{ij})$: Matriz de dados com objetos I em linhas e variáveis J em colunas; se as variáveis forem expressas por diferentes unidades de medidas, são padronizadas para ter média zero e variância unitária;

P, Q : Número desejado de agrupamentos de objetos e de variáveis, respectivamente;

$U = (u_{ip})$: Matriz que define uma alocação dos objetos I em agrupamentos P ; U é uma matriz estocástica binária $I \times P$ e de linha, ou seja, não possui elementos negativos que somam um por linha e, portanto, possui apenas um elemento diferente de zero por linha:

$$\begin{cases} u_{ip} = 1, \text{ se o } i - \text{ésimo objeto pertence ao agrupamento } P \\ u_{ip} = 0, \text{ caso contrário} \end{cases}$$

$V = (v_{jq})$: Matriz que define uma partição das variáveis J em Q agrupamentos; V é uma matriz estocástica binária $J \times Q$ e de linha, ou seja, não possui elementos negativos que somam um por linha e, portanto, possui apenas um elemento diferente de zero por linha:

$$\begin{cases} v_{jq} = 1, \text{ se a } j - \text{ésima variável pertence ao agrupamento } Q \\ v_{jq} = 0, \text{ caso contrário} \end{cases}$$

$\bar{X}$: Matriz do centróide do objeto no espaço original; $\bar{X}$ é uma matriz $P \times J$ definida por:

$$\bar{X} = (U^T U)^{-1} U^T X$$

$A = (a_{jq})$: Matriz das cargas dos componentes; A é matriz $J \times Q$ onde as colunas estão identificando os coeficientes de Q combinações lineares, isto é, os componentes principais da CDPCA, tais que:

$$\begin{cases} \sum_{j=1}^J (a_{jq}a_{jr})^2 = 0, \text{ para qualquer } q \text{ e } r (q \neq r) \\ \sum_{j=1}^J (a_{jq}a_{jr})^2 = 1, \text{ caso contrário} \end{cases}$$

$Y = (y_{iq})$: Matriz de score de componentes onde y_{iq} é o valor do i -ésimo objeto para o q -ésimo componente CDpca; sintetiza a informação comum de um agrupamento de variáveis;

$\bar{Y}$: Matriz de centroide do objeto no espaço reduzido; Y tem dimensão $P \times Q$ sendo definida por $Y = \bar{X} A$.

$E = (e_{ij})$: Matriz de erro de dimensão $I \times J$.

1.3.2 O modelo

O modelo CDpca resulta da aplicação da ACP na matriz de dados transformados, onde cada objeto é substituído por seu centroide. Por sua vez, os centroides são obtidos aplicando o algoritmo K -médias na matriz de dados originais (Freitas, Macedo e Vichi, 2020). Assim, a matriz de dados é ajustada pelo modelo da seguinte forma (Vichi e Saporta, 2009):

$$\begin{aligned} X &= U\bar{X} + E_1 \text{ (K- médias aplicada a X),} \\ X &= U\bar{Y}A^T + E_1 + E_2 \text{ (ACP aplicada a } U\bar{X}), \\ X &= U\bar{Y}A^T + E \text{ (modelo da CDPCA),} \end{aligned} \quad (1)$$

onde $E = E_1 + E_2$ com $E_1 + E_2$ as matrizes de erro ($I \times J$) decorrentes de k -médias e ACP, respectivamente.

A partir do modelo CDpca (1), é possível ver que $E = X - U\bar{Y}A^T$. Portanto, o problema do CDpca visa minimizar a norma da matriz do erro E , resultando no seguinte problema de otimização:

$$\min_{U, \bar{Y}, A} \|X - U\bar{Y}A^T\|_2^2, \quad (2)$$

sujeito às condições acima para as matrizes U (ou seja, U é formada por dados binários e matriz de linha estocástica), $\bar{Y}$ ($\bar{Y}$ é uma matriz do objeto centroide no espaço reduzido) e A (A é uma matriz ortonormal em colunas, onde cada linha contribui para uma única coluna).

Pode-se provar que o problema (2) é equivalente a maximizar o termo não associado ao erro do modelo, equivalente a:

$$\max_{U, \bar{Y}, A} \|U \bar{Y} A^T\|_2^2 \quad (3)$$

sujeito a U ser uma matriz estocástica binária e linha e A uma matriz ortonormal coluna a coluna, ou seja, $A^T A = I$, onde cada linha contribui para uma única componente. Devido à característica binária das matrizes de atribuição U e V , o problema de otimização (3) para obter a decomposição CDpca é bastante difícil de resolver.

Sendo, $\bar{Y} = \bar{X} A$ e A com colunas ortonormal $A^T A = I$, quando $\|U \bar{Y} A^T\|_2^2 = \|U \bar{X} A\|_2^2$. Assim, o problema (3) é equivalente a:

$$\max_{U, \bar{X}, A} \|U \bar{X} A\|_2^2 \quad (4)$$

Para resolver o problema (4), um algoritmo de mínimos quadrados alternados (ALS), desenhado com quatro etapas, foi proposto em Vichi e Saporta (2009).

Macedo e Freitas (2015), mostraram que cada iteração do algoritmo ALS pode ser resumidamente descrito por duas etapas básicas: a atribuição de objetos via k-médias e a redução do espaço de atributos via aplicação da ACP aos centroides resultantes.

Vale ressaltar que, ao final de cada iteração do ALS, um novo carregamento da matriz A é construída e a função objetivo (finita) em (3) é atualizada. Se a função objetivo aumenta com este carregamento, então a matriz anterior de carregamentos A é alterado por este novo, caso contrário, a matriz anterior A é preservada para a próxima iteração do algoritmo. Essa estratégia garante a convergência da sequência de valores da função objetivo a um ponto estacionário, que se espera ser, pelo menos, um máximo local do problema (3). Este procedimento pode ser considerado como uma heurística e assim, para aumentar a possibilidade de atingir o máximo global, foi sugerido por Vichi e Saporta (2009) executar o algoritmo várias vezes para diferentes

matrizes de atribuição inicial U e V , escolhidos aleatoriamente no início de cada execução.

1.4 APLICAÇÕES DA CDpca

A primeira publicação sobre a CDpca consta em Vichi e Saporta (2009), onde os autores apresentam a técnica como uma nova metodologia denominada *Clustering and Disjoint Principal Component Analysis* (CDpca), porque, segundo os autores, precisa-se encontrar componentes associados a classes disjuntas de variáveis.

Nesse trabalho, o desempenho da CDpca foi avaliado, aplicando-a a dois conjuntos de dados: um primeiro, que descreve o cenário macroeconômico de curto prazo (setembro de 1999) de vinte países da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), considerando seis principais indicadores econômicos; e o segundo conjunto de dados que apresenta as características socioeconômicas de 103 países de acordo com 17 variáveis.

Os autores concluem em síntese, que, com a CDpca obtém-se uma dupla vantagem: primeiro, identifica-se uma classificação de variáveis e uma classificação de objetos; segundo, obtém uma redução de dimensionalidade da matriz de dados através de um conjunto reduzido de centróides para objetos e um conjunto reduzido de componentes (combinações lineares) para variáveis.

Outra vantagem importante destacada pelos autores é a fácil interpretabilidade dos componentes, uma vez que cada um é caracterizado por um conjunto disjundo de variáveis. Portanto, com a CDpca não é possível ter uma variável observada caracterizando dois componentes diferentes como acontece frequentemente com a ACP. Ou seja, para a CDpca, não é necessário utilizar um método de rotação para melhorar a interpretabilidade dos resultados.

Macedo e Freitas (2015), apresentaram um procedimento heurístico iterativo para realizar a CDpca em dados bidirecionais, o qual corresponde a uma versão em duas etapas dos mínimos quadrados alternados (ALS) original, envolvendo uma etapa para a alocação de objetos via *k-means* e outra etapa para a redução do espaço de atributos via aplicação de PCA nos centróides resultantes.

Outro procedimento utiliza uma estrutura algorítmica de aproximação baseada em uma abordagem de programação semidefinida (SDP) e foi chamado de SDP em duas etapas (Macedo 2015), porque dois problemas de SDP são considerados para

clusterizar objetos e atributos. Este processo também pode ser subdividido em duas fases, envolvendo uma primeira fase em que os clusters de objetos e atributos são inicialmente estimados usando projeções e modelos SDP, e uma segunda fase em que um procedimento de arredondamento baseado em k-médias aplicado no espaço reduzido de centróides é executado para refinar soluções. Ambos os algoritmos, ALS e SDP, foram implementados e testados em R (Development Core 2019) em vários conjuntos de dados.

Freitas, Macedo e Vichi (2021), utilizaram conjuntos de dados de expressão gênica simulados e reais onde o número de variáveis era maior que o número de objetos, a fim de comparar empiricamente o desempenho dos dois procedimentos iterativos citados anteriormente, ALS e SDP. No geral, ambos os procedimentos produziram resultados satisfatórios em termos de precisão da solução – a abordagem SDP em duas etapas parece ser uma ferramenta importante em termos de obtenção de resultados mais rápidos e com grande capacidade de recuperar o número real de clusters de objetos, enquanto o algoritmo ALS supera ao fornecer mais resultados precisos no espaço reduzido dos componentes, identificando clusters mais claramente homogêneos, compactos e bem separados.

1.5 CONCLUSÃO

A Análise de Componentes Principais (ACP) destaca-se como uma ferramenta valiosa, explorada em diversos contextos científicos. No entanto, é crucial reconhecer que novas modificações, como a *Clustering and Disjoint Principal Component Analysis* (CDpca), têm emergido, evidenciando vantagens significativas em sua aplicação.

REFERÊNCIAS

- BIRNBAUM A.; JOHNSTONE I. M.; NADLER B.; PAUL, D. Minimax bounds for sparse PCA with noisy high-dimensional data. **Ann. Statist.** v. 41, n. 3, p. 1055–1084, 2013. (doi:10.1214/12-AOS1014).
- CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos Biométricos Aplicados ao Melhoramento Genético**. Viçosa: UFV, 2012, 514 p.
- FIGUEREDO NETO, A.; ALMEIDA, F. de A. C.; GOUVEIA, J. P. G. de; NÓBREGA, M. B. M.; CARNEIRO, R. M.; PEDROZA, J. P. Divergência genética em acessos de mamona (*Ricinus communis* L.) baseada nas características de semente. **Revista de Biologia e Ciências da Terra**, v. 4, p. 1-10, 2004.
- FREITAS A.; MACEDO E.; VICH M. An empirical comparison of two approaches for CDPCA in high-dimensional data. **Statistical Methods & Applications**, v. 30, n. 3, p. 1007-1031, 2021.
- FREITAS, A. R. de. **Estatística experimental na agropecuária**. Brasília: Embrapa, 2022, 457 p.
- GABRIEL, K. R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. **Biometrika**, v. 58, p. 453-467, 1971.
- HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. O. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. **Engineering and Science**, v. 5, n. 1, 2016.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417– 441, 1933. <https://doi.org/10.1037/h0071325>
- JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 4. ed. New York: Prentice-Hall, 1998.
- JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 5. ed. New York: Prentice-Hall, 2007.
- JOLLIFFE I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Phil. Trans. R. Soc. A** v. 374, n. 13, 2016.
- JOLLIFFE, I. **Principal Component Analysis**. 2. ed. New York: Springer, 2002.
- KAISER, H. F. The varimax criterion for analytic rotation in fator analysis. **Psychometrika**, v. 23, n. 3, p. 187-200, 1958.
- LEE, D.; LEE, W.; LEE, Y.; PAWITAN, Y. Super-sparse principal component analyses for highthroughput genomic data. **BMC Bioinformatics**. v. 11, n. 296, 2010.
- MACEDO, E. Abordagem SDP em duas etapas para agrupamento e redução de dimensionalidade. **Computação Stat. Optim. Inf.**, v. 3, n. 3, p. 294–311, 2015.
- MACEDO, E.; FREITAS, A. O algoritmo de mínimos quadrados alternados para CDPCA. **Springer**, v. 499, p. 173–191, 2015.
- MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: UFMG, 2005, 295 p.

MOURA, D. S.; BRITO, G. G.; MORAES, Í. L.; FAGUNDES, P. R. R.; CASTRO, A. P.; DEUNER, S. Cold Tolerance in Rice Plants: Phenotyping Procedures for Physiological Breeding. **Journal of Agricultural Science**, v. 10, n. 1, p. 313–324, 2018.

OLIVEIRA, T. R. A. D.; GRAVINA, G. D. A.; OLIVEIRA, G. H. F. D.; ARAÚJO, K. C.; ARAÚJO, L. C. D.; DAHER, R. F.; VIVAS, M.; GRAVINA, L. M.; CRUZ, D. P. D. The GT biplot analysis of green bean traits. **Ciência Rural**, v. 48, n. 6, 2018.

OSBORNE, J. W.; COSTELLO, A. B. Sample Size and Subject to Item Ratio in Principal Components Analysis. **Practical Assessment, Research & Evaluation**, v. 9, 2004.

PANERO, J. S.; SILVA, H. E. B. da; PANERO, P. S.; SMIDERLE, O. J.; PANERO, F. S.; FARIA, F. S. E. D. V.; RODRIGUEZ, A. F. R. Separation of Cultivars of Soybeans by Chemometric Methods Using Near Infrared Spectroscopy. **Journal of Agricultural Science**, v. 10, n. 4, p. 351–360, 2018.

PIEDEDE, G. N. Principal Component Analysis for Identification of Superior Castor Bean Hybrids. **Journal of Agricultural Science**, v. 11, n. 9, 2019.

R Development Core Team. R: uma linguagem e ambiente para computação estatística. 2019 <http://www.R-project.org/>

RAMSAY J. O.; SILVERMAN B. W. Functional data analysis. **Springer Series in Statistics**. New York: Springer, 2006.

SILVA, N. da. Aplicativo computacional para utilização de componentes principais em experimentação agrônômica, 2005.

SOUZA, J. de. Análise em componentes principais e suas aplicações, métodos estatísticos nas ciências sociais. Brasília: **Thesaurus**, v. 2, 1988, 68 p.

VICHI, M.; KIERS, H. A. L. Factorial k-means analysis for two-way data. **Computational Statistics and Data Analysis**, v. 37, p. 49–64, 2001.

VICHI, M.; SAPORTA, G. Clustering and disjoint principal component analysis. **Computational Statistics and Data Analysis**. v. 53, p. 3194–3208, 2009.

YAN, W.; FRÉGEAU-REID, J. Breeding line selection based on multiple traits. **Crop Science**. v. 48, p. 417–423, 2008.

CAPÍTULO 2

CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS – UMA NOVA ABORDAGEM PARA OS ESTUDOS DE MELHORAMENTO GENÉTICO VEGETAL

2.1 INTRODUÇÃO

A estatística está em constante desenvolvimento devido ao avanço contínuo da ciência, que tanto cria quanto aprimora técnicas já existentes. Muito desse progresso deve-se às ferramentas computacionais que facilitam sua execução, utilizando um menor tempo para realizar a análise, e permitindo melhor compreensão dos resultados obtidos.

A Análise de Componentes Principais (ACP) mais popularmente conhecida como *Principal Component Analysis* (PCA), por exemplo, que se constitui em um procedimento básico do qual derivam-se vários outros métodos multivariados, foi concebida por Karl Pearson, em 1901 (Souza, 1988), mas consolidada apenas na década de 30 com os artigos de Harold Hotelling (1933, 1935 e 1936). Apesar de ser uma técnica bastante expressiva, muitas adaptações da metodologia básica para diferentes tipos de dados e estruturas foram desenvolvidas (Lee *et al.*, 2010; Birnbaum *et al.*, 2013).

Algumas dessas modificações são simplificações das componentes principais, com o intuito de ajudar na interpretação dos resultados (Jolliffe e Cadima, 2016), como por exemplo, os estudos realizados para decidir os critérios e a quantidade de componentes a serem selecionadas. Outras, exploram métodos nos quais as funções de dados são vistas como soluções para equações diferenciais, uma abordagem conhecida como análise diferencial (Ramsay e Silverman, 2006).

Nesse contexto, uma nova metodologia foi proposta por Vichi e Saporta (2009), a análise restrita de componentes principais - *Clustering and Disjoint Principal Component Analysis* (CDpca), que visa um agrupamento simultâneo de objetos e um particionamento de variáveis. Esta análise, além de permitir uma redução na dimensionalidade da matriz de dados, identifica componentes com variância máxima, para cada combinação linear de um subconjunto de variáveis. Simultaneamente, uma partição de objetos também é calculada maximizando a variação entre os

agrupamentos. Outra vantagem importante ressaltada pelos autores é a facilidade de interpretação dos componentes, uma vez que cada um é caracterizado por um conjunto disjunto de variáveis, assim como também não é necessário usar um método de rotação para melhorar a interpretabilidade dos resultados.

No contexto do *big data*, no qual existem dados em grande volume, estas análises exploratórias atualizadas mostram-se como alternativas eficientes. Já na área de estudos de melhoramento vegetal, tais vantagens apresentadas podem trazer grandes benefícios para avaliação dos dados gerados nos experimentos. Entretanto, pouco se sabe sobre essa técnica. Desta forma, objetivou-se neste capítulo compreender a CDpca: quais seus pressupostos, como organizar o banco de dados, definir cada parâmetro para análise e interpretar os seus resultados.

2.2 MATERIAL E MÉTODOS

2.2.1 O conjunto de Dados

O conjunto de dados utilizado no presente estudo, pertence a área de melhoramento vegetal e caracteriza-se por apresentar mais objetos (amostras) que atributos (variáveis). Neste, foram avaliados 31 híbridos de mamona (*Ricinus communis* L.), sendo 28 desenvolvidos pelo Programa de Melhoramento da Faculdade de Ciências Agrônômicas da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), e três híbridos comerciais: 26, 27 e 28.

A avaliação foi realizada em relação aos seguintes parâmetros: altura da planta (AP), altura do primeiro cacho (APC) e diâmetro do caule (DC) em centímetros; número de nós (NN); número de cachos (NC); número de sementes (NS) e número de frutos (NF); massa de 100 sementes (M100) em gramas e teor de óleo de sementes (TO) em percentual. Assim, foi possível obter uma matriz X com 31 amostras (I) e 9 variáveis (J), tendo a matriz X dimensão I:J (31:9).

2.2.2 Pré-processamento dos Dados

Como a *Clustering and Disjoint Principal Component Analysis* (CDpca), é uma técnica que combina a análise de agrupamento com a análise de componentes principais, a normalização dos dados é recomendada. Isso ocorre porque a CDpca

usa a matriz de covariância dos dados para calcular os componentes principais, sendo sensível às diferenças de escala entre as variáveis. Se as variáveis não estiverem normalizadas, aquelas com maiores magnitudes podem dominar a análise, resultando em uma representação distorcida dos dados (à semelhança do método ACP padrão).

Ressalta-se que a normalização dos dados é uma etapa pré-processamento e deve ser aplicada antes de realizar a CDpca ou qualquer outro algoritmo de análise de dados, para garantir resultados mais precisos e confiáveis. Neste estudo, a normalização dos dados foi feita com a padronização *lz-score1*.

Como pressupostos para se fazer a análise com a CDpca, os dados precisam ser numéricos e não é possível ter dados faltantes no conjunto que será avaliado. Além disso, recomenda-se a avaliação de números discrepantes (*outliers*).

2.2.3 Descrição dos Experimentos Computacionais

Os experimentos numéricos foram realizados em um computador com processador Intel Core i5-1135G7 CPU@2.40 GHz, Windows 11 Pro (64 bits) e 16GB RAM, utilizando a função CDpca disponível no pacote biplotbootGUI (LIBRERO, *et al.* 2019) R versão 4.3.2 (R Core Team, 2023).

A função apresenta os seguintes argumentos, conforme descrito na Tabela 1.

Tabela 1 – Argumentos da função CDpca

Argumentos	Descrição
data	Matriz numérica (I x J)
class	Um vetor numérico contendo a classificação real dos objetos nos dados, ou <i>NULL</i> se a classe dos objetos for desconhecida
P	Um valor inteiro indicando o número de grupos de objetos
Q	Um valor inteiro indicando o número de grupos de variáveis
SDPinitial	Um valor lógico indicando se as matrizes de atribuição inicial U e V são geradas aleatoriamente (padrão) ou se uma estrutura algorítmica baseada em uma abordagem de programação semidefinida é preferida (<i>TRUE</i>)

<code>tol</code>	Um valor positivo (baixo) que indica o termo máximo para a diferença entre dois valores consecutivos da função objetivo. Um valor de tolerância de 10^{-5} é indicado por padrão
<code>maxit</code>	O número máximo de iterações de uma execução do algoritmo ALS
<code>r</code>	Número de execuções do algoritmo ALS para a solução final
<code>CDpcaplot</code>	Um valor lógico indicando se um gráfico adicional é criado (mostrando os dados projetados nos dois primeiros componentes principais da CDpca)

2.2.4 Definição dos Parâmetros

Os argumentos de entrada P , Q , tol , $maxit$ e r da função devem ser definidos previamente em qualquer aplicação da metodologia CDpca.

Como a CDpca se baseia no K-médias para agrupamento dos objetos (P) e na ACP para escolha do agrupamento das variáveis (Q), realizou-se neste estudo ambas as técnicas isoladamente para definição destes parâmetros. A ACP foi realizada nos dados utilizando o pacote *prcomp* e a análise de agrupamento, o pacote *vegan* do R.

Vichi e Saporta (2009) destacam que, com base em experiências numéricas em conjuntos de dados com $I > J$, é recomendado executar o algoritmo ALS (r) pelo menos 30 vezes, para diferentes matrizes de atribuição inicial U e V escolhidos aleatoriamente no início de cada execução. O número máximo de iterações ($maxit$) deve variar de 500 a 1.000, de modo a aumentar a chance de encontrar a solução ótima global e, conseqüentemente, reduzir a sensibilidade do algoritmo ALS nas matrizes iniciais U e V . Estes autores também definiram o valor de tolerância (tol) como 10^{-5} para CDpca. Desta forma, definiu-se neste estudo o $maxit = 500$, $r = 1.000$ e $tol = 10^{-5}$.

Para Freitas, Macedo e Vichi (2021), essa escolha particular de argumentos de entrada deve levar à estabilidade dos resultados numéricos.

2.3 RESULTADOS E DISCUSSÃO

2.3.1 Definição do P e Q

Realizou-se a ACP para definir o parâmetro Q. Pela proporção de variância total que os componentes principais (CPs) explicam, pode-se determinar o número de componentes que se deve reter. Para Rencher (2002), este percentual corresponde a no mínimo, 70%. Neste caso, os três primeiros CPs explicaram aproximadamente 75,01% de toda a variação contida nas nove variáveis estudadas, sendo que destes, 40,92% foram explicadas pela CP1, 22,11% pela CP2 e os demais 11,98% pela CP3 (Tabela 2). Conforme Cruz *et al.*, (2012) esta análise possibilita uma considerável simplificação nos cálculos estatísticos e interpretação dos resultados em relação aos outros métodos alternativos.

Diferente da ACP, a CDpca não permite realizar a análise com $Q = 1$ ou $Q = 9$, ou seja, uma só componente não pode explicar a estrutura dos dados e nem são gerados o número de variáveis igual ao número de componentes. Assim como o número de grupos $P = 1$ ou 2 indicou erro (subscrito fora dos limites ou argumentos inconformáveis).

Tabela 2 – Componentes principais, autovalores, proporção da variância explicada e proporção acumulada pelos componentes da Análise de Componentes Principais

Componentes Principais	Autovalores (λ)	Proporção (%)	Proporção Acumulada (%)
CP1	1,9190	0,4092	0,4092
CP2	1,4107	0,2211	0,6303
CP3	1,0383	0,1198	0,7501
CP4	0,9692	0,1044	0,8545
CP5	0,7334	0,0607	0,9142
CP6	0,5756	0,0368	0,9510
CP7	0,5533	0,0341	0,9851
CP8	0,3331	0,0123	0,9974
CP9	0,1520	0,0026	1,0000

Para definir o parâmetro P, realizou-se a análise de agrupamentos. Com base na matriz de distância generalizada de Mahalanobis (D^2) foi realizado o agrupamento das progênes pelos métodos Ligação Simples, Ligação Completa, UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) e Método de Ward. Determinou-se o coeficiente de correlação cofenético (CCC) para verificar e selecionar o método mais representativo (Sokal e Rohlf, 1962). O coeficiente de correlação cofenético (CCC) do método UPGMA baseado na distância generalizada de

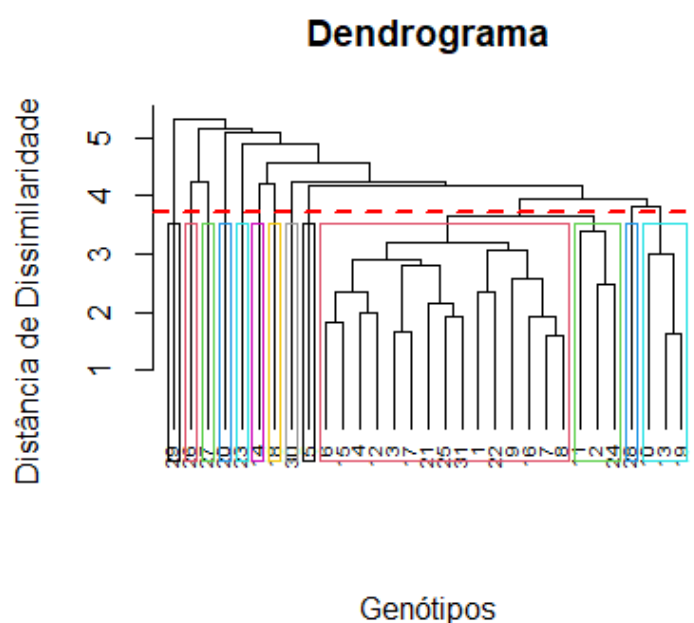
Mahalanobis foi de 0,82. Este valor indica uma boa representatividade entre a matriz de dissimilaridade e o dendrograma (Rohlf, 1970; Cruz *et al.*, 2012). Quanto menor for o valor de CCC, maior é a distorção verificada entre as distâncias de dissimilaridade genética apresentada pelo dendrograma e matriz de dissimilaridade genética original obtida ao utilizar o método UPGMA (Sokal e Rohlf, 1962) (Tabela 3).

Tabela 3 – Correlação cofenética para vários métodos na distância de Mahalanobis

Métodos	Correlação cofenética
Ligação Simples	0,7878
Ligação Completa	0,6914
UPGMA	0,8231
Método De Ward	0,7878

Pelo dendrograma, observa-se que as progênes, de acordo com suas dissimilaridades, se dividiram em 13 grupos (Figura 1).

Figura 1 - Dendrograma obtido pelo método UPGMA com base na distância generalizada de Mahalanobis (D^2) para agrupamento de 31 genótipos de *Ricinus communis* L.



A partir desses resultados, pode-se definir então $P = 13$ e $Q = 3$. Entretanto, para verificar melhor estas análises, simulou-se P variando de 2 a 20 e $Q = 2$ e 3. A partir de então, foram registrados os valores de: tempo de execução (em segundos), *iter* (número total de iterações usadas no melhor *loop* para calcular a melhor solução), *bcd* (desvio entre cluster), *bcdT* (desvio entre clusters sobre a variabilidade total), F (valor da função objetivo para maximizar), E (norma de erro para o modelo CDpca obtido) e %var (porcentagem de variabilidade total explicada) (Tabelas 4 e 5).

Observa-se que em ambos os casos, os parâmetros *bcd*, *bcdT* e F aumentaram conforme o aumento no número de grupos, assim como o E diminuiu. Já o Tempo e %var oscilaram. Para $Q = 2$ o número total de iterações usadas no melhor *loop* para calcular a melhor solução variou de 4 a 8 e para $Q = 3$ variou de 3 a 10.

Levando em consideração o critério para a escolha da ACP, a porcentagem acima de 70% de explicação dos dados só foi alcançada quando $Q = 3$, o que sugere a escolha de 3 componentes.

Já em relação a escolha de P , observou-se que a maior diferença na %var é encontrada ao mudar o valor de 5 para 6 (70,85 para 72,08). A partir de então, as mudanças são mínimas. A maior %var foi equivalente a 72,63 quando $P = 15, 17$ e 19.

Tabela 4 – Tempo de execução (em segundos), iter (número total de iterações usadas no melhor *loop* para calcular a melhor solução), bcd (desvio entre cluster), bcdT (desvio entre clusters sobre a variabilidade total), F (valor da função objetivo para maximizar), E (norma de erro para o modelo CDpca obtido) e %var (porcentagem de variabilidade total explicada), para número de componentes (Q) = 2 e diferentes escolhas para o número de grupos (P) pela CDpca

Q	P	Tempo	Iter	bcd	bcdT	F	E	%var
2	2	18,40	4	64,98	28,18	78,63	0,46	44,81
	3	30,18	6	66,55	38,53	107,49	0,42	59,82
	4	47,46	4	75,62	43,86	122,37	0,40	59,93
	5	57,43	5	84,61	48,01	133,94	0,39	58,63
	6	39,82	8	85,94	51,08	142,53	0,38	61,42
	7	44,30	8	90,38	53,39	148,95	0,37	61,04
	8	73,56	8	92,10	54,85	153,02	0,36	61,54
	9	47,17	7	94,13	55,81	155,71	0,36	61,27
	10	52,98	7	95,43	56,54	157,75	0,36	61,22
	11	44,86	4	95,98	57,17	159,51	0,35	61,56
	12	45,03	8	96,71	57,58	160,64	0,35	61,51
	13	54,41	7	97,52	58,11	162,13	0,35	61,58
	14	44,94	6	97,90	58,35	162,81	0,35	61,59
	15	32,36	7	98,26	58,54	163,33	0,35	61,56
	16	35,75	4	98,67	58,79	164,04	0,35	61,58
	17	38,56	6	98,91	58,93	164,42	0,35	61,56
	18	38,12	4	99,24	59,13	164,96	0,34	61,57
	19	33,32	4	99,33	59,19	165,14	0,34	61,57
	20	32,19	4	99,51	59,29	165,41	0,34	61,56

Tabela 5 – Tempo de execução (em segundos), iter (número total de iterações usadas no melhor *loop* para calcular a melhor solução), bcd (desvio entre cluster), bcdT (desvio entre clusters sobre a variabilidade total), F (valor da função objetivo para maximizar), E (norma de erro para o modelo CDpca obtido) e %var (porcentagem de variabilidade total explicada), para número de componentes (Q) = 3 e diferentes escolhas para o número de grupos (P) pela CDpca

Q	P	Tempo	Iter	bcd	bcdT	F	E	%var
3	2	21,99	3	64,98	28,18	78,63	0,46	53,01
	3	42,88	5	65,23	39,79	111,04	0,42	70,63
	4	90,71	10	67,63	46,38	129,40	0,39	70,87
	5	54,31	5	75,37	51,68	144,20	0,37	70,85
	6	64,74	5	78,93	55,06	153,63	0,36	72,08
	7	62,82	6	83,55	58,38	162,89	0,35	72,21
	8	49,24	10	87,47	61,41	171,33	0,34	72,54
	9	95,67	7	89,89	62,96	175,67	0,33	72,38
	10	69,65	5	91,42	64,23	179,25	0,32	72,62
	11	85,54	5	93,13	65,45	182,59	0,32	72,61
	12	123,38	6	94,33	66,18	184,65	0,31	72,50
	13	87,92	4	94,99	66,61	185,83	0,31	72,46
	14	50,13	5	95,70	67,18	187,44	0,31	72,54
	15	53,94	5	96,31	67,69	188,86	0,31	72,63
	16	63,09	7	97,20	68,09	189,96	0,30	72,38
	17	56,40	7	97,20	68,33	190,65	0,30	72,63
	18	51,28	4	97,87	68,77	191,86	0,30	72,60
	19	53,64	5	98,34	69,12	192,85	0,30	72,63
	20	60,36	4	98,75	69,32	193,41	0,30	72,56

Após estas simulações, para analisar a classificação das variáveis e dos objetos pela CDpca, definiu-se então $P = 15$ e $Q = 3$. Aqui serão apresentados seus principais resultados.

2.3.2 CDpca

A três componentes apresentaram juntas uma porcentagem de explicação de 72,42% (Tabela 6).

Tabela 6 – Porcentagem de explicação de cada componente da CDpca

Variabilidade		
[,1]	[,2]	[,3]
38,46	17,43	16,53

As variáveis número de cachos (NC), número de frutos (NF), número de sementes (NS), teor de óleo (TO) e massa de 100 grãos (M100), foram mais expressivas na primeira componente. As variáveis número de nós (NN) e altura do primeiro cacho (APC), na segunda componente. E as variáveis altura da planta (AP) e diâmetro do caule (DC) na terceira componente (Tabelas 7 e 8).

Tabela 7 – Matriz V – Partição de variáveis da CDpca para os seguintes parâmetros: número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo de sementes (TO) e massa de 100 sementes (M100)

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	1	0	0
NF	1	0	0
NS	1	0	0
AP	0	0	1
DC	0	0	1
NN	0	1	0
APC	0	1	0
TO	1	0	0
M100	1	0	0

Tabela 8 – Matriz A – Matriz de *loading* de componentes da CPDCA para os seguintes parâmetros: número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo de sementes (TO) e massa de 100 sementes (M100)

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	0,4711	0	0
NF	0,4816	0	0
NS	0,5125	0	0
AP	0	0	0,6479
DC	0	0	0,7617
NN	0	0,6667	0
APC	0	0,7453	0
TO	0,1999	0	0
M100	0,4993	0	0

Os escores são os valores dos novos indivíduos após a CDpca. É comum utilizá-los para análises estatísticas ou ordenação dos elementos amostrais (Tabela 9).

Tabela 9 – Matriz Y – Matriz de escores de componentes da CDpca para os seguintes parâmetros: número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo de sementes (TO) e massa de 100 sementes (M100)

Objetos	Componentes		
	[,1]	[,2]	[,3]
1	-1,40	0,81	1,46
2	1,24	0,24	-0,21
3	0,88	-0,37	-0,53
4	-0,40	-0,63	-0,06
5	0,89	-1,67	-0,76
6	-2,85	-0,47	-0,17
7	-1,37	-0,63	-0,62
8	-0,52	0,34	-0,33
9	1,10	1,52	-0,72
10	1,44	0,32	0,07
11	0,70	1,41	0,84
12	-1,10	-2,04	-1,22
13	-0,11	1,16	0,79
14	4,39	0,71	-1,91

15	-2,74	-1,49	-0,19
16	-0,49	-0,97	-0,58
17	0,79	-0,49	-0,38
18	1,02	3,54	-0,82
19	-1,00	1,15	0,57
20	-0,66	0,40	-0,38
21	-0,80	-0,16	-1,67
22	-2,27	1,35	0,74
23	4,01	0,34	-0,40
24	2,44	-0,52	1,14
25	0,91	0,23	-0,23
26	2,01	-1,22	1,23
27	1,42	-0,30	2,62
28	-2,45	-1,30	-0,17
29	-3,25	1,92	4,08
30	-1,13	-1,88	-1,34
31	-0,72	-1,28	-0,86

Os métodos de agrupamento têm por finalidade separar um grupo original de observações em vários subgrupos, de forma a obter homogeneidade dentro e heterogeneidade entre os subgrupos. Na CDpca, esta análise não precisa ser realizada separadamente (Tabela 10).

Tabela 10 – Classificação dos 31 objetos (híbridos) nos 15 grupos (P) definidos, na CDpca

Grupos	Objetos
1	11, 13
2	3, 5, 17
3	4, 7, 16, 31
4	1, 19, 22
5	21
6	27
7	2, 10, 25
8	29
9	18
10	24, 26
11	8, 20
12	12, 30
13	9
14	14, 23
15	6, 15, 28

Observou-se em análises posteriores do mesmo *script*, que existem pequenas oscilações nos resultados de alocação dos indivíduos nos grupos, indicando instabilidade no agrupamento.

2.4 CONCLUSÃO

É evidente que a interpretação dos resultados da *Clustering and Disjoint Principal Component Analysis* (CDpca) não se mostra uma tarefa difícil. Além disto, por apresentar o diferencial de realizar também o agrupamento dos indivíduos estudados, sem precisar da complementação de outra técnica, pode ser bastante explorada, emergindo como uma alternativa promissora para aplicação em estudos de melhoramento genético.

Neste trabalho, seguiu-se uma abordagem que pode ser mais aprofundada, já que a implementação da CDpca demanda análises prévias para definir seus parâmetros, os quais ainda não estão totalmente esclarecidos, bem como a estabilidade nos resultados.

REFERÊNCIAS

- BIRNBAUM A.; JOHNSTONE I. M.; NADLER B.; PAUL, D. Minimax bounds for sparse PCA with noisy high-dimensional data. **Ann. Stat.** v. 41, p. 1055–1084, 2013. (doi:10.1214/12-AOS1014).
- CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos Biométricos Aplicados ao Melhoramento Genético**. Viçosa: UFV, v. 11, n. 2, 514 p. 2012.
- FREITAS A.; MACEDO E.; VICHÍ M. An empirical comparison of two approaches for CDPCA in high-dimensional data. **Statistical Methods & Applications**, v. 30, n. 3, p. 1007-1031, 2021.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417– 441, 1933. <https://doi.org/10.1037/h0071325>
- HOTELLING, H. The most predictable criterion. **Journal of Educational Psychology**, v. 26, n. 2, p. 139–142. 1935
- HOTELLING, H. Relations between two sets of variates, **Biometrika**, v. 28, Issue 3-4, p. 321–377, 1936.
- JOLLIFFE I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Phil. Trans. R. Soc. A** v. 374, n. 13, 2016.
- LEE, D.; LEE, W.; LEE, Y.; PAWITAN, Y. Super-sparse principal component analyses for highthroughput genomic data. **BMC Bioinformatics**. v. 11, n. 296, 2010.
- LIBRERO, A. B. N.; VILLARDÓN, M. P. G.; FREITAS, A. biplotbootGUI: Bootstrap on Classical Biplots and Clustering Disjoint Biplot, **R package**, v. 1.2, 2019. <http://www.R-project.org/package=biplotbootGUI>
- R Development Core Team (2023) R: a language and environment for statistical computing. <http://www.Rproject.org/>
- RAMSAY J. O.; SILVERMAN B. W. Functional data analysis. **Springer Series in Statistics**. New York: Springer, 2006.
- RENCER, A. C. **Methods of Multivariate Analysis**. 2. ed. Inc. Publication. p. 727, 2002.
- ROHLF, F. J. Adaptative hierarquical clustering schemes. **Systematic Zoology**, v. 19, n. 1, p. 58-82, 1970.
- SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, v. 11, p. 33-40, 1962.

SOUZA, J. de. Análise em componentes principais e suas aplicações, métodos estatísticos nas ciências sociais. Brasília: **Thesaurus**, v. 2, 1988, 68 p.

VICHI, M.; SAPORTA, G. Clustering and disjoint principal component analysis. **Computational Statistics and Data Analysis**. v. 53, p. 3194–3208, 2009.

CAPÍTULO 3

ANÁLISE DE COMPONENTES PRINCIPAIS E CLUSTERING AND DISJOINT PRINCIPAL COMPONENT ANALYSIS EM DADOS DE MELHORAMENTO GENÉTICO VEGETAL

3.1 INTRODUÇÃO

O melhoramento genético vegetal é uma das áreas da Agronomia que se destaca por exigir amplo conhecimento sobre os efeitos bióticos, abióticos e suas interações. Além disso, desempenha um papel fundamental na garantia de uma alimentação de qualidade e em quantidade suficiente para a população.

Em decorrência de mudanças climáticas frequentes e abruptas têm-se encurtado a vida útil de materiais vegetais superiores lançados, o que gera pressão constante para o desenvolvimento contínuo de novos materiais que atendam às necessidades mundiais, concomitantemente a busca pela mitigação aos efeitos ambientais.

Como forma de superar tais desafios, os programas de melhoramento têm trabalhado com ampla variabilidade de material genético, em busca de genótipos com características favoráveis que possam ser inseridos tanto em materiais altamente já melhorados como a soja (*Glycine max* (L.) Merrill) quanto em culturas pouco exploradas, mas que apresentem alto potencial produtivo e econômico como a mamona (*Ricinus communis* L.).

A avaliação da produção vegetal normalmente implica em estudos onde são avaliados os efeitos dos tratamentos a partir da mensuração de inúmeras variáveis em diversos indivíduos. Nos programas de melhoramento genético, por exemplo, os trabalhos desenvolvidos para a obtenção de novas cultivares com características agronômicas desejáveis, como a superação dos níveis atuais de produtividade e ampla adaptação, são primeiramente constituídos por muitos acessos (variabilidade), seguidos por longos períodos de estudo e experimentação em vários locais. Consequentemente, grandes conjuntos de dados são gerados, atualmente denominados como *big data*, tornaram-se mais comuns e, muitas vezes, de difícil análise e interpretação.

De modo geral, essas variáveis são dependentes (altura x diâmetro, por exemplo) porque possuem componentes temporais, e são descritas tradicionalmente por meio de análises estatísticas univariadas, que, embora de grande utilidade, podem comprometer as interpretações e as conclusões destes estudos, justamente por não explorar a dependência entre as variáveis analisadas (Fikdalski *et al.*, 2007), resultando em seleção de material não completamente desejável (Godoi, 1985).

Pimentel Gomes (1990) afirma que, somente havendo independência entre as variáveis (situação quase impossível na prática), as análises univariadas resolveriam o problema. Assim, para aumentar a probabilidade de êxito, num programa de melhoramento, deve-se combinar grande número de caracteres, com análise estatística eficiente em extrair as informações principais contidas numa grande massa de dados e gerar bancos de dados característicos para futuro aprendizado de máquinas.

Desta forma, as técnicas de redução de espaço e recursos para seleção de subconjuntos são de grande importância, uma vez que facilitam a visualização e capturam as associações relevantes existentes no conjunto de dados componentes (Mingoti, 2005).

As técnicas de análise multivariada apresentam-se como instrumentos poderosos na análise de dados, em razão de permitirem combinar as informações múltiplas tomadas nas unidades de amostragem. Segundo Cruz (1990), no melhoramento de plantas, essas técnicas possibilitam prever fenômenos genéticos, baseando-se, simultaneamente, em um complexo de variáveis de importância tanto genética quanto evolutiva.

As técnicas de maior utilização compreendem um conjunto de métodos estatísticos exploratórios, cujos objetivos centram-se especialmente, no estudo da diferenciação entre objetos ou indivíduos e no estudo do relacionamento entre variáveis.

Dentre estas, destaca-se a Análise de Componentes Principais (ACP) mais conhecida como PCA do inglês *Principal Component Analysis*, que se constitui num procedimento básico do qual derivam vários outros métodos multivariados, que reduz a dimensionalidade dos conjuntos de dados, aumentando a interpretabilidade, mas ao mesmo tempo minimizando a perda de informações. Essa técnica cria variáveis chamadas de componentes principais (CP), não correlacionadas, que são definidas

pelo conjunto de dados em questão e onde há a maximização sucessiva da variância, tornando-a uma análise de dados adaptativa (Jolliffe e Cadima, 2016).

Entretanto, tal como acontece com outras técnicas estatísticas, é possível obter em algumas observações um efeito desproporcional sobre os resultados de uma ACP. Dentro do espaço de cada modo, a interpretação desses eixos é feita de acordo com os valores de seus componentes (cargas). Portanto, a interpretação às vezes pode ser difícil, levando a uma caracterização imprecisa desses novos eixos. Assim, é desejável que cada componente principal ou eixo possua poucas entidades que contribuam para a variabilidade do componente de maneira relevante (Martin-Barreiro *et al.*, 2021).

Visando a redução de objetos e variáveis, é comum nos estudos do melhoramento aplicar duas técnicas sequencialmente: primeiro executando uma ACP e subsequentemente aplicando um algoritmo de agrupamento nas pontuações dos componentes dos objetos, usando assim apenas os primeiros componentes. No entanto, DeSarbo *et al.* (1990), De Sorte e Carroll (1994) e Vichi e Kiers (2001) advertem contra essa abordagem, denominada “*análise em tandem*”, pois a ACP pode identificar dimensões que não necessariamente contribuem muito para a percepção da estrutura de agrupamento nos dados e que, ao contrário, podem obscurecer ou mascarar as informações taxonômicas (Vichi e Saporta, 2009).

Assim, apesar de ser uma técnica bastante expressiva, muitas adaptações da metodologia básica para diferentes tipos de dados e estruturas, foram desenvolvidos (Lee *et al.*, 2010; Birnbaum *et al.*, 2013) visando torná-la mais robusta.

Nesse contexto, uma nova metodologia foi proposta por Vichi e Saporta (2009), a análise restrita de componentes principais (*Clustering and Disjoint Principal Component Analysis* - CDpca), a qual, em síntese, apresenta dupla vantagem: identificar uma classificação de variáveis e uma classificação de objetos; possibilita obter uma redução na dimensionalidade da matriz de dados através de um conjunto reduzido de centroides para objetos e um conjunto reduzido de componentes (combinações lineares) para variáveis. Outra vantagem importante é a facilidade de interpretação dos componentes, uma vez que cada um é caracterizado por um conjunto disjunto de variáveis.

Portanto, com a CDpca, não é possível ter uma variável observada explicando mais de um componente, como acontece com frequência para a ACP. Assim como também não é necessário usar um método de rotação para melhorar a

interpretabilidade dos resultados e nem de agrupamento já que esta técnica também faz esse procedimento.

Apesar dos grandes benefícios citados anteriormente, a CDpca ainda é pouco utilizada, visto que seu complexo sistema ainda não está disponível em softwares executáveis e *scripts* prontos disponíveis em *softwares* programáveis, são recentes e escassos, tornando-se desse modo, fator limitante para sua ampla difusão no meio acadêmico.

Quanto aos estudos de melhoramento genético do Brasil, a aplicação da CDpca inexistente. Assim, o objetivo deste estudo foi demonstrar o desempenho do algoritmo descrito por Macedo e Freitas (2015) para a execução da CDpca e comparar os resultados com aqueles obtidos a partir da ACP clássica, quando ambas as metodologias são aplicadas para avaliar diversas variáveis em dados reais de dois diferentes experimentos da área de melhoramento genético de plantas, de modo a divulgar e ampliar sua utilização.

3.2 MATERIAL E MÉTODOS

O desempenho e a eficiência da ACP e da CDpca, foram avaliados sendo aplicadas ao mesmo grupo de dados, a fim de comprovar as vantagens e desvantagens citadas na literatura e dessa forma possibilitar a utilização de uma nova técnica nos estudos agronômicos, especificamente na área do melhoramento vegetal.

3.2.1 Descrição dos conjuntos de dados

Os conjuntos de dados utilizados no presente estudo são caracterizados por serem todos da área do melhoramento vegetal e apresentarem mais objetos (amostras) do que atributos (variáveis):

Dados 1: Foram avaliados 31 híbridos de mamona (*Ricinus communis* L.), sendo 28 desenvolvidos pelo Programa de Melhoramento da Faculdade de Ciências Agronômicas da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), e 3 híbridos comerciais foram usados como testemunhas: híbridos 26, 27 e 28. A avaliação foi realizada em relação aos seguintes parâmetros: altura da planta (AP), altura do primeiro cacho (APC) e diâmetro do caule (DC) em centímetros; número de nós (NN); número de cachos (NC); número de sementes (NS) e número de frutos (NF);

massa de 100 sementes (M100) em gramas e teor de óleo de sementes (TO) em percentual. Assim, foi possível obter uma matriz X com 31 amostras (n) e 9 variáveis (m), tendo a dimensão n:m (31:9).

Dados 2: Foram avaliadas 23 progênies de mamona (*Ricinus communis* L.), advindas de três tipos de polinização do cultivar FCA-PB – autofecundação, polinização cruzada e polinização livre – em relação aos seguintes atributos: produtividade do grão (PROD) e produtividade do óleo (PRODO) em kg ha⁻¹; altura (ALT) em centímetros; teor de óleo (TO) em porcentagem e peso de 100 grãos (PCG) em gramas. Assim, foi possível obter uma matriz X com 69 amostras (n) e 5 variáveis (m), tendo a dimensão n:m (69:5).

Para os dados 2, têm-se o conhecimento prévio sobre os grupos: três tipos de polinização (P = 3) autofecundação, polinização cruzada e polinização livre advindas do cultivar de mamona FCA-PB. Sendo assim, testes simulatórios para identificar o P não foram necessários.

3.2.2 Análise de Componentes Principais

Os dados foram submetidos a análise de componentes principais (ACP), com a intenção de reduzir o grande número de variáveis para um conjunto menor e mais significativo (representado pelos componentes), mantendo o máximo de informação existente. Como as variáveis originais não possuem a mesma unidade, foi utilizada a matriz de correlação como forma de padronização prévia dos dados (centrar em relação à média e reduzir à unidade de desvio padrão). Isso porque a ACP, com dados em escalas diferenciadas, faz com que as variáveis de maior variância entrem com maior peso nas primeiras componentes principais (CPs) em relação as de menor variância (Wilks, 2006).

Para identificar o quanto cada variável explica cada componente principal (CP) e assim definir as variáveis que mais contribuíram na formação de cada CP selecionado, analisou-se seus respectivos autovetores, que representam os coeficientes do CP associados positivamente ou negativamente à cada variável original. A partir destes, é possível obter os escores, que são os valores assumidos de cada CP para os objetos estudados, neste caso, os híbridos, e assim identificar qual híbrido melhor se destacou em relação a cada CP.

O processamento dos dados foi realizado com o auxílio do software R 4.3.2.

3.2.3 Clustering and Disjoint Principal Component Analysis

Os experimentos numéricos foram realizados em um computador com processador Intel Core i5-1135G7 CPU@2.40 GHz, Windows 11 Pro (64 bits) e 16GB RAM, utilizando a função CDpca disponível no pacote biplotbootGUI (Librero, *et al.* 2019) R versão 4.3.2.

A função apresenta os seguintes argumentos, conforme descrito na Tabela 1.

Tabela 1 – Argumentos da função CDpca

Argumentos	Descrição
data	Matriz numérica (I x J)
class	Um vetor numérico contendo a classificação real dos objetos nos dados, ou <i>NULL</i> se a classe dos objetos for desconhecida
P	Um valor inteiro indicando o número de grupos de objetos
Q	Um valor inteiro indicando o número de grupos de variáveis
SDPinitial	Um valor lógico indicando se as matrizes de atribuição inicial U e V são geradas aleatoriamente (padrão) ou se uma estrutura algorítmica baseada em uma abordagem de programação semidefinida é preferida (<i>TRUE</i>)
tol	Um valor positivo (baixo) que indica o termo máximo para a diferença entre dois valores consecutivos da função objetivo. Um valor de tolerância de 10^{-5} é indicado por padrão
maxit	O número máximo de iterações de uma execução do algoritmo ALS
r	Número de execuções do algoritmo ALS para a solução final
CDpcaplot	Um valor lógico indicando se um gráfico adicional é criado (mostrando os dados projetados nos dois primeiros componentes principais da CDpca)

Definiu-se neste estudo o $maxit = 500$, $r = 1.000$ e $tol = 10^{-5}$ para ambos os grupos de dados. Entretanto, para o primeiro conjunto o $P = 3$ e $Q = 15$, já para o segundo $P = 2$ e $Q = 3$.

Como pressupostos para se fazer a análise com a CDpca, os dados precisam ser numéricos e não é possível ter dados faltantes no conjunto que será avaliado. Além disso, recomenda-se a avaliação de números discrepantes (*outliers*).

Ressalta-se que a normalização dos dados também é uma etapa pré-processamento e deve ser aplicada antes de realizar a CDpca para garantir resultados mais precisos e confiáveis. Neste estudo, a normalização dos dados foi feita com a padronização |z-score|.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Dados 1

Pela proporção de variância total que os componentes principais (CPs) explicam, pode-se determinar o número de componentes que se deve reter. Para Rencher (2002), este percentual corresponde a no mínimo, 70%. Neste caso, os três primeiros CPs explicaram 75,01% de toda a variação contida nas nove variáveis estudadas, sendo que destes, 40,92% foram explicadas pela CP1, 22,11% pela CP2 e os demais 11,98% pela CP3 (Tabela 2).

Tabela 2 – Componentes principais, autovalores, proporção da variância explicada e proporção acumulada pelos componentes da Análise de Componentes Principais

Componentes Principais	Autovalores (λ)	Proporção (%)	Proporção Acumulada (%)
CP1	1,9190	0,4092	0,4092
CP2	1,4107	0,2211	0,6303
CP3	1,0383	0,1198	0,7501
CP4	0,9692	0,1044	0,8545
CP5	0,7334	0,0607	0,9142
CP6	0,5756	0,0368	0,9510
CP7	0,5533	0,0341	0,9851
CP8	0,3331	0,0123	0,9974
CP9	0,1520	0,0026	1,0000

Na CDpca, as três componentes são responsáveis por 72,42% da variabilidade total dos dados, menos do que a porcentagem da ACP (Tabela 3).

Com o intuito de identificar as variáveis que mais contribuíram na formação de cada componente, analisou-se seus respectivos autovetores.

Tabela 3 – Porcentagem de explicação de cada componente da CDpca

Variabilidade		
[,1]	[,2]	[,3]
38,46	17,43	16,53

Assim, é possível observar que, o CP1 apresentou grandes associações positivas com as variáveis NS, NC, M100 e NF. Já no CP2, observa-se que as variáveis DC, APC, e AP apresentaram maiores autovetores positivos. Entretanto, na CP3 a APC, TO e NN foram as variáveis que se associaram mais fortemente. É possível notar então que a variável APC destaca-se nas CP2 e CP3, devendo neste caso ser decisão do experimentador definir em qual será considerada (Tabela 4).

Tabela 4 – Autovetores do número de cachos (NC), número de frutos (NF), número de sementes (NS), número de nós (NN), altura do primeiro cacho (APC), altura da planta (AP), diâmetro do caule (DC), teor de óleo (TO) e massa de 100 grãos (M100), nos três componentes principais (CP1, CP2 e CP3) da Análise de Componentes Principais

Variáveis	CP1	CP2	CP3
NC	0,4723	-0,0542	0,1424
NF	0,4382	0,0071	-0,2522
NS	0,4728	0,0710	-0,1464
NN	-0,2872	0,3518	-0,4393
APC	-0,0190	0,5246	-0,5046
AP	0,1723	0,4896	0,3853
DC	-0,0375	0,5513	0,3162
TO	0,1719	-0,1718	-0,4457
M100	0,4668	0,1411	-0,0026

Assim, o CP1, CP2 e CP3 podem ser caracterizados como sendo principalmente relacionado à produtividade de grãos de mamona, ao porte das plantas e à produção de óleo, respectivamente, contribuindo desta forma com o processo de escolha dos melhores híbridos, quando baseada no objetivo dos produtores ou dos programas de melhoramento.

Na CDpca, a partição das variáveis é dada pela matriz V (Tabela 5), uma matriz binária, que facilita a visualização e interpretação, diminuindo desta forma erros na análise, e pela matriz de *loading* (Tabela 6), que também apresenta as mesmas

características citadas anteriormente, uma vez que as variáveis que não são importantes na CP em questão, têm seu valor zerado.

Na CP1, destacaram-se as variáveis NC, NF, NS, TO e M100. Na CP2, as variáveis NN e APC e na CP3 as variáveis AP e DC. Neste caso, não houve dualidade da variável APC em mais de uma componente (Tabelas 5 e 6) como ocorreu nos resultados da ACP (Tabela 4). Isto evita deixar a tomada de decisão por parte do pesquisador, tornando-a menos subjetiva e sim, mais assertiva.

Tabela 5 – Matriz V – Partição das variáveis número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo (TO) e massa de 100 grãos (M100), nos três componentes principais (CP1, CP2 e CP3) da CDpca, dos híbridos de mamona

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	1	0	0
NF	1	0	0
NS	1	0	0
AP	0	0	1
DC	0	0	1
NN	0	1	0
APC	0	1	0
TO	1	0	0
M100	1	0	0

Tabela 6 – Matriz A – Matriz de *loading* de componentes nas variáveis número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo (TO) e massa de 100 grãos (M100), nos três componentes principais (CP1, CP2 e CP3) da CDpca, dos híbridos de mamona

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	0,4809483	0	0
NF	0,4818184	0	0
NS	0,5035033	0	0
AP	0	0	0,6062765
DC	0	0	0,7952539
NN	0	0,7418628	0
APC	0	0,6705517	0

TO	0,2063418	0	0
M100	0,4903542	0	0

A seleção dos híbridos a partir dos escores, fica facilitada através da caracterização dos CPs realizada anteriormente: para produtividade de grãos é necessário que eles possuam altos valores positivos na CP1; já em relação ao porte da planta e a produtividade de óleo, os melhores híbridos devem apresentar altos escores negativos nas CP2 e CP3. Assim, os híbridos 14 e 23 foram os indicados na CP1. Na CP2, os híbridos 30 e 12, e na CP3, além dos híbridos comerciais 26 e 27, o híbrido 24 apresentou desempenho semelhante para este componente (Tabela 7).

Tabela 7 – Escores dos híbridos de *Ricinus communis* L. em relação aos três componentes (CP1, CP2 e CP3) da Análise de Componentes Principais

HÍBRIDOS	COMPONENTES PRINCIPAIS		
	CP1	CP2	CP3
1	-1.5628	1.5698	-0.0999
2	0.9724	0.1679	-0.2366
3	0.9953	-0.5492	0.2444
4	-0.5322	-0.4541	-0.2635
5	0.8725	-1.5546	-0.1869
6	-2.8328	-0.9213	0.1218
7	-1.4180	-1.2245	-0.1246
8	-0.3910	-0.2140	0.0805
9	1.4344	0.7384	1.1222
10	1.3982	0.5639	0.1826
11	0.5995	2.1935	0.7488
12	-1.1383	-2.2651	-0.1408
13	0.1121	1.2808	0.2751
14	4.7972	-0.4887	0.7811
15	-2.9734	-1.4110	-0.1589
16	-0.6820	-1.3314	-0.5598
17	0.7555	-0.4813	0.0675
18	1.8385	2.1191	2.3254
19	-0.7975	1.2632	0.8229
20	-0.4344	-0.6040	0.4695
21	-0.2501	-1.0996	1.5946
22	-2.2975	1.5422	0.6103
23	3.9006	0.0303	-0.3778
24	1.7581	0.5268	-1.7058
25	1.0211	0.3183	0.4261
26	1.9706	-0.3249	-2.6905
27	1.2770	1.4260	-2.6675
28	-2.3590	-0.6839	1.0581

29	-4.2702	3.9196	-1.3968
30	-1.2132	-2.5917	-0.5812
31	-0.5506	-1.4505	0.2397

Observa-se que, desta forma, não se identificou nenhum híbrido dentre os estudados que apresentasse desempenho superior concomitantemente aos três componentes, ou seja, híbridos com alta produtividade de grãos e de óleo e porte baixo.

Na CDpca a caracterização das componentes não ficou tão evidente como na ACP. Entretanto, os híbridos 14 e 23 também se destacaram na CP1 como na ACP. Na CP2, foi o híbrido 18 e na CP3 os híbridos 29 e 27 (Tabela 8).

Tabela 8 – Matriz Y – Matriz de escores dos híbridos de *Ricinus communis* L. em relação aos três componentes (CP1, CP2 e CP3) da CDpca

Objetos	Componentes		
	[,1]	[,2]	[,3]
1	-1.3977996	0.8107646	1.45776532
2	1.2426029	0.2405323	-0.20589100
3	0.8842681	-0.3679099	-0.52761833
4	-0.4055223	-0.6303476	-0.05626284
5	0.8900815	-1.6754277	-0.76236362
6	-2.8487358	-0.4736755	-0.17312737
7	-1.3677051	-0.6327433	-0.61909509
8	-0.5222711	0.3351584	-0.33227788
9	1.1054514	1.5225768	-0.71629567
10	1.4405500	0.3264145	0.06890242
11	0.7021334	1.4080672	0.83613240
12	-1.1033747	-2.0364842	-1.21890171
13	-0.1120998	1.1559304	0.78728599
14	4.3908718	0.7145414	-1.90662508
15	-2.7406802	-1.4924837	-0.18950618
16	-0.4938207	-0.9683666	-0.58211616
17	0.7885710	-0.4943575	-0.37980529
18	1.0254669	3.5427053	-0.81970027
19	-1.0000766	1.1487836	0.57045446
20	-0.6620734	0.3995599	-0.37969130
21	-0.8011881	-0.1595327	-1.66727491
22	-2.2667456	1.3492154	0.74570363
23	4.0146445	0.3367555	-0.40097245
24	2.4419533	-0.5237835	1.14228050
25	0.9119635	0.2269971	-0.23249508
26	2.0129430	-1.2221006	1.23086483
27	1.4232313	-0.2995156	2.61796552
28	-2.4501658	-1.2968432	-0.17631584
29	-3.2586726	1.9186892	4.08296066

30	-1.1274348	-1.8853618	-1.33866853
31	-0.7163665	-1.2777582	-0.85531114

A análise de agrupamento é importante pois possibilita classificar indivíduos em grupos de forma a maximizar sua homogeneidade dentro dos grupos e discriminar a heterogeneidade entre os grupos (Tabela 9). Na CDpca esta análise é feita automaticamente, o que se demonstra como vantagem.

De acordo com análise prévia (Capítulo 2), o número de grupos (P) foi definido como 15. Desta forma, o agrupamento de objetos seguiu o disposto na Tabela 9.

Observa-se que os híbridos 14 e 23, ambos destacados na CP1 da ACP bem como da CDpca, se mantiveram no mesmo grupo, o que sugere estabilidade e superioridade em relação às suas características, independente da técnica estatística utilizada. O mesmo aconteceu para os híbridos 12 e 30, que se destacaram na CP2 quando a ACP foi realizada, já que ambos ficaram alocados no mesmo grupo. Já entre os híbridos que tiveram destaque na CP3, o 24 e 26 ficaram juntos, porém o híbrido 27 ficou alocado em outro grupo (Figuras 1 e 2).

Ressalta-se que neste caso os híbridos 26 e 27 são comerciais, sendo esta divergência esperada. Quanto aos demais agrupamentos relacionados aos híbridos destacados em cada CP da CDpca, o híbrido 18, que demonstrou superioridade na CP2, se manteve em um agrupamento sozinho. Já na CP3, mesmo os híbridos 27 e 29 tendo se destacado em conjunto, na análise de agrupamento se mantiveram isolados também (Figuras 1 e 2).

Estes resultados mostram uma certa conformidade entre si, sendo indícios de validação da análise de agrupamento da CDpca. Porém, como já relatado ocorrer certas alterações a cada rodagem do *script* para se fazer esta análise, estudos posteriores ainda são necessários.

Tabela 9 – Classificação dos híbridos de *Ricinus communis* L. pela CDpca

Grupos	Objetos
1	11, 13
2	3, 5, 17
3	4, 7, 16, 31
4	1, 19, 22
5	21
6	27
7	2, 10, 25
8	29
9	18

10	24, 26
11	8, 20
12	12, 30
13	9
14	14, 23
15	6, 15, 28

Figura 1- *Biplot* dos componentes principais 1 e 2 pela Análise de Componentes Principais

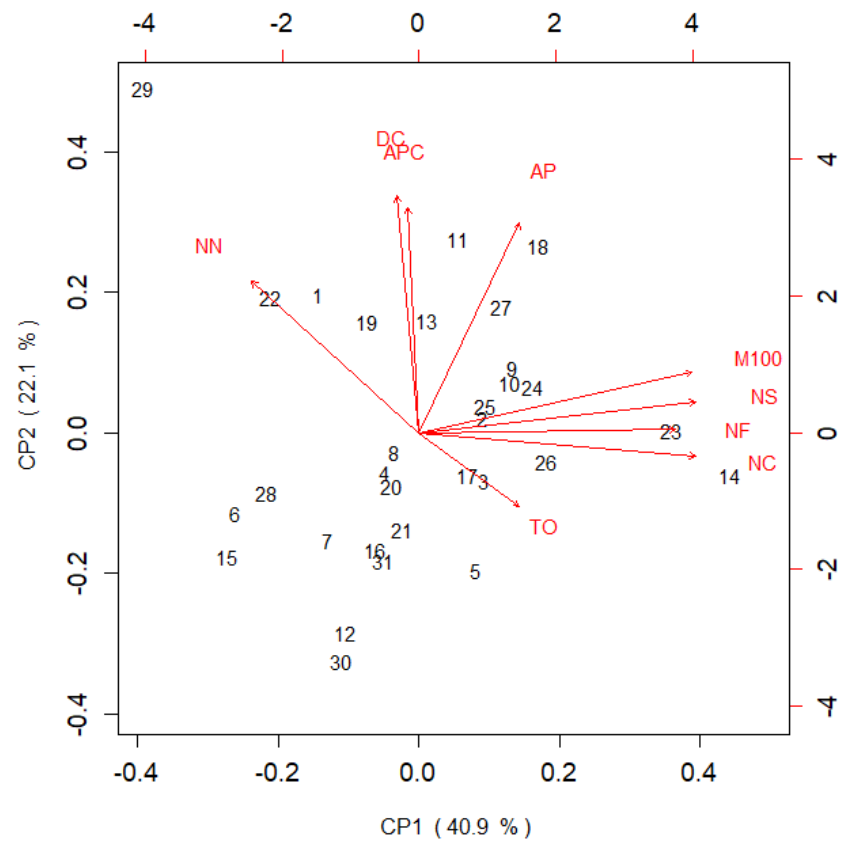
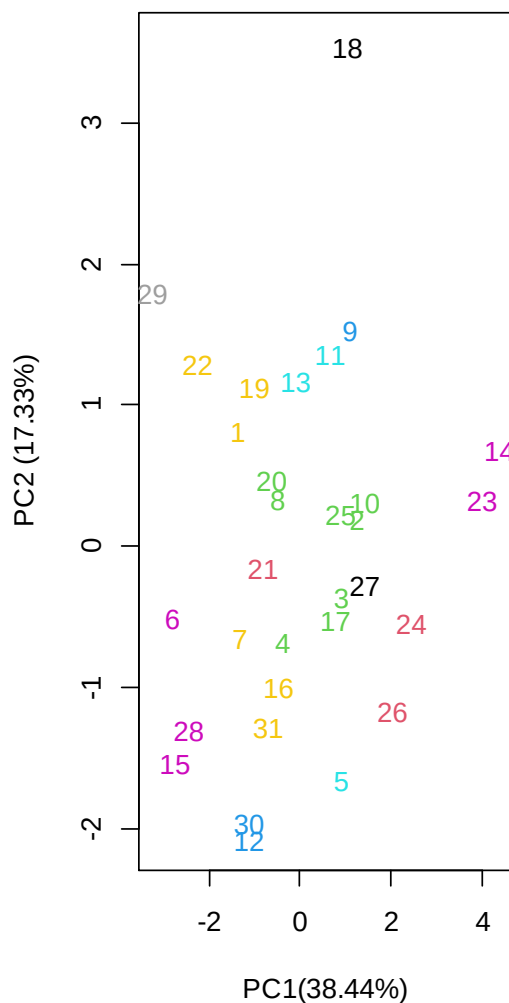


Figura 2 - CDpca plot dos componentes principais 1 e 2 pela CDpca

3.3.2 Dados 2

Pela ACP, as duas primeiras CPs explicaram aproximadamente 76,27% de toda a variação contida nas cinco variáveis estudadas, sendo que destas, 43,64% foram explicadas pela CP1 e 32,63% pela CP2 (Tabela 10).

Tabela 10 – Componentes principais, autovalores, proporção da variância explicada e proporção acumulada pelos componentes da Análise de Componentes principais.

Componentes Principais	Autovalores (λ)	Proporção (%)	Proporção Acumulada (%)
CP1	1,4772	43,64	43,64
CP2	1,2773	32,63	76,27
CP3	0,8231	13,55	89,82
CP4	0,7119	10,14	99,96
CP5	0,0452	00,41	100

Na CDpca, as duas componentes foram responsáveis por 76,97% da variabilidade total dos dados (Tabela 11), uma porcentagem um pouco maior que a anterior (Tabela 10).

Tabela 11 – Porcentagem de explicação de cada componente pela CDpca

Variabilidade	
[,1]	[,2]
40,22	36,75

Com o intuito de identificar as variáveis que mais contribuíram na formação de cada componente selecionado, analisou-se seus respectivos autovetores. Assim, é possível observar que, o CP1 apresentou grandes associações positivas com as variáveis PO e PROD. Já no CP2, observa-se que as variáveis TO (positiva), PROD (negativa), PCG e AL (positiva) apresentaram maiores autovetores.

É possível notar então que a variável PROD destaca-se nas CP1 e CP2, devendo neste caso ser decisão do experimentador definir em qual componente esta variável será considerada (Tabela 12). Verifica-se que este cenário é comum nas análises de componentes principais.

Tabela 12 – Autovetores da produtividade (PROD), altura (AL), teor de óleo (TO), produtividade de óleo (PO), e produtividade corrigida (PCG), nos dois componentes (CP1 e CP2) da Análise de Componentes Principais

Variáveis	CP1	CP2
PROD	0,543	-0,465
AL	0,312	0,453
TO	0,323	0,478
PO	0,591	-0,376
PCG	0,392	0,457

Na CDpca, destacaram-se as variáveis PROD e PO na CP1, e TO, PCG e AL na CP2. Neste caso, não há dualidade das variáveis em mais de uma componente (Tabelas 13 e 14).

Tabela 13 – Matriz V – Partição das variáveis: produtividade (PROD), altura (AL), teor de óleo (TO), produtividade de óleo (PO), e produtividade corrigida (PCG), nos dois componentes (CP1 e CP2) da CDpca

Variáveis	Componentes	
	CP1	CP2
PROD	1	0
AL	0	1
TO	0	1
PO	1	0
PCG	0	1

Tabela 14 – Matriz A – Matriz de *loading* nos dois componentes (CP1 e CP2) das variáveis: produtividade (PROD), altura (AL), teor de óleo (TO), produtividade de óleo (PO), e produtividade corrigida (PCG), da CDpca

Variáveis	Componentes	
	CP1	CP2
PROD	0,6902	0
AL	0	0,5112
TO	0	0,6304
PO	0,7237	0
PCG	0	0,5841

A seleção dos híbridos a partir dos escores, mostra que, pela ACP, na CP1 os híbridos 35 e 62 apresentaram maiores escores e na CP2 os híbridos 27 (negativa) e 52 (positiva) (Tabela 15).

Tabela 15 – Escores dos híbridos de *Ricinus communis* L. em relação as duas componentes (CP1 e CP2) da Análise de Componentes Principais

HÍBRIDOS	Componentes	
	CP1	CP2
1	0,7942	-1,2693
2	0,6666	-1,2496
3	1,5038	0,9671

4	-0,4366	0,3258
5	2,4450	-0,5440
6	-0,0624	-1,4790
7	-0,2309	-0,2677
8	-1,3794	-0,1682
9	0,3668	-1,3726
10	-1,5462	-0,1601
11	-1,2389	-0,9624
12	1,0403	-0,5421
13	-0,7478	0,4468
14	-2,4725	-0,7179
15	1,4771	2,4151
16	-1,0545	-1,2348
17	-0,8983	-0,0559
18	-0,1609	-0,8271
19	-1,6035	-0,0020
20	-0,7800	-1,1698
21	-3,5446	0,6727
22	-0,2671	-0,3236
23	0,4511	-0,3606
24	0,0873	1,3854
25	0,5343	-0,5956
26	-0,2051	0,9135
27	2,6858	-3,0643
28	0,8855	0,7205
29	0,1833	1,9303
30	0,0915	0,2136
31	-2,2049	1,6759
32	-0,5832	2,1281
33	-1,4636	0,1235
34	0,9488	-0,0726
35	4,3882	1,4243
36	0,5942	-1,1388
37	0,0511	-2,7912
38	1,5398	-0,7629
39	1,0795	0,6620
40	-1,8234	1,2393
41	-1,8634	0,9997
42	-1,4125	0,5767
43	0,0191	1,3388
44	1,0492	0,8296
45	0,7763	-2,1307
46	-0,4051	0,1322
47	0,6145	0,8247
48	1,9464	0,9068
49	-0,1134	0,9765
50	0,1089	-0,1936
51	-1,0654	-1,6704
52	1,8991	2,6362
53	-0,7451	0,6519

54	-1,7332	-1,1378
55	-0,0158	1,7181
56	-1,4323	1,3578
57	-1,7452	-0,9463
58	0,3246	2,4368
59	-1,2416	0,8798
60	0,5189	-0,5572
61	-0,4236	1,9572
62	4,5490	-0,3613
63	2,9468	1,6601
64	-0,2479	-2,2206
65	-1,7475	-0,8436
66	0,0357	-1,4423
67	0,7568	-1,6167
68	-1,8837	0,8760
69	1,1199	-1,0727

Já a seleção dos híbridos a partir dos escores, pela CDpca, mostra que, na CP1 os híbridos 35 e 62 apresentaram maiores escores e na CP2 os híbridos 27 e 62 (Tabela 16). Comparando ambos resultados, observa-se discrepância apenas na seleção do híbrido 52 pela ACP com o 62 pela CDpca.

Tabela 16 – Escores dos híbridos de *Ricinus communis* L. em relação as duas componentes principais (CP1 e CP2) pela CDpca.

HÍBRIDOS	CP1	CP2
1	-0,0912	1,5038
2	-0,2185	1,3421
3	1,8037	0,2340
4	-0,1783	-0,5531
5	1,6579	1,8867
6	-0,8870	1,1916
7	-0,2540	0,0559
8	-1,1564	-0,7073
9	-0,4501	1,3422
10	-1,0778	-0,7117
11	-1,5983	-0,0421
12	0,5266	1,0683
13	-0,3287	-0,7355
14	-2,4248	-1,0301
15	2,6632	-0,9993
16	-1,5526	0,2506
17	-0,7578	-0,5885

18	-0,6101	0,4844
19	-1,2857	-0,9612
20	-1,3452	0,4476
21	-2,4971	-2,5854
22	-0,4752	0,0833
23	0,0535	0,6705
24	0,8870	-0,8994
25	0,0351	0,8138
26	0,4109	-0,9149
27	0,3563	4,1408
28	1,2031	-0,0263
29	1,3115	-1,3984
30	0,2461	-0,0025
31	-0,7585	-2,6346
32	0,7850	-2,1044
33	-1,0461	-1,0113
34	0,6531	0,5651
35	4,4010	1,3303
36	-0,1804	1,2071
37	-1,6473	2,3822
38	0,7530	1,5178
39	1,2656	0,1738
40	-0,7069	-2,0829
41	-0,9289	-1,8936
42	-0,7942	-1,3505
43	-0,7782	0,9110
44	1,3345	-0,1523
45	-0,6058	2,1510
46	-0,2593	-0,2642
47	0,9400	-0,2144
48	2,1561	0,4559
49	0,4929	-0,9137
50	-0,0041	0,1535
51	-1,8424	0,6910
52	3,1290	-1,0152
53	-0,2485	-0,8925
54	-2,0297	-0,1029
55	1,0161	-1,3618
56	-0,4076	-1,9384
57	-1,9425	-0,2218
58	1,6859	-1,7077
59	-0,4693	-1,3620
60	0,1028	0,7993
61	0,7586	-1,7458
62	3,3405	2,8755
63	3,3026	0,3824
64	-1,5337	1,6443
65	-1,9448	-0,2838
66	-0,7791	1,1527
67	-0,4137	1,8144

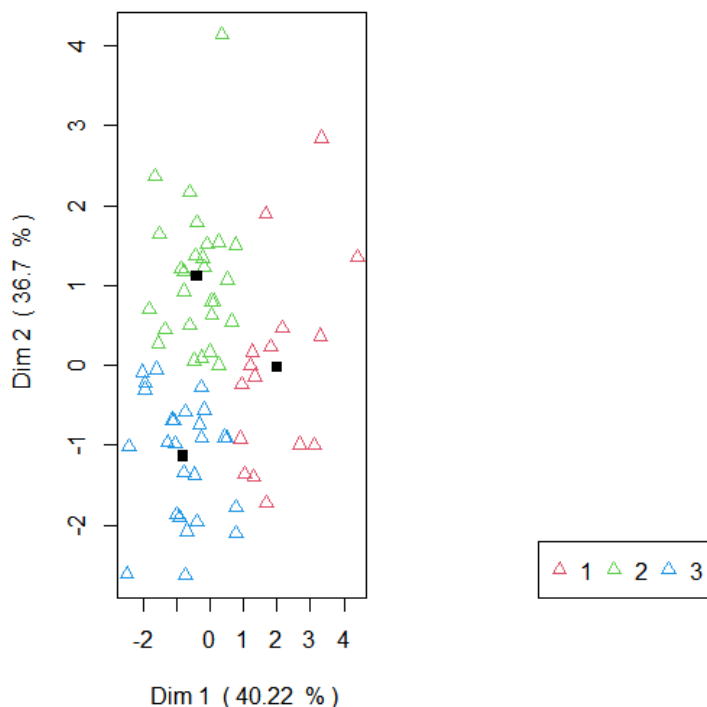
68	-1,0027	-1,8625
69	0,2446	1,5490

A análise de agrupamento é importante pois possibilita classificar indivíduos em grupos de forma a maximizar sua homogeneidade dentro dos grupos e discriminar a heterogeneidade entre os grupos (Tabela 17, Figura 3). Na CDpca esta análise é feita automaticamente, o que se apresenta como vantagem.

Tabela 17 – Classificação dos 69 genótipos de *Ricinus communis* L. nos três grupos pela CDpca

Grupos	Objetos
1	4, 8, 10, 11, 13, 14, 17, 19, 21, 26, 31, 32, 33, 40, 41, 42, 46, 49, 53, 54, 56, 57, 59, 61, 65, 68
2	3, 5, 15, 24, 28, 29, 35, 39, 44, 47, 48, 52, 55, 58, 62, 63,
3	1, 2, 6, 7, 9, 12, 16, 18, 20, 22, 23, 25, 27, 30, 34, 36, 37, 38, 43, 45, 50, 51, 60, 64, 66, 67, 69

Figura 3 - Classificação da CDpca para os 69 genótipos de *Ricinus communis* L. nos três grupos



Destacando os híbridos selecionados anteriormente, observa-se que apenas o híbrido 27 foi alocado em um grupo diferente dos demais (35, 52 e 62). Nota-se também que a classificação original dos híbridos de acordo com os três métodos de polinização, não foi mantida.

3.3.3 Proposta De Modificação Na CDpca

Na tentativa de estabilizar os resultados da CDpca a cada rodagem e manter os resultados da CP3 similares aos da PCA, foi realizada uma modificação no script original da função, onde fixou-se a matriz V.

Testes simulatórios foram realizados utilizando o conjunto Dados 1.

As três componentes foram responsáveis por 69,77% da variabilidade total dos dados (Tabela 18).

Tabela 18 – Porcentagem de explicação de cada componente da CDpca modificada

Variabilidade		
[,1]	[,2]	[,3]
36,32	16,89	16,56

Na CP1, destacaram-se as variáveis NC, NF, NS e M100. Na CP2, as variáveis NN e APC e na CP3 as variáveis AP, DC e TO (Tabelas 19 e 20). Este resultado se assemelha ao da PCA.

Tabela 19 – Matriz V – Partição das variáveis número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo (TO) e massa de 100 grãos (M100), nos três componentes principais (CP1, CP2 e CP3) da CDpca modificada

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	1	0	0
NF	1	0	0
NS	1	0	0
AP	0	0	1
DC	0	0	1
NN	0	1	0

APC	0	1	0
TO	0	0	1
M100	1	0	0

Tabela 20 – Matriz A – Matriz de *loading* de componentes nas variáveis número de cachos (NC), número de frutos (NF), número de sementes (NS), altura da planta (AP), diâmetro do caule (DC), número de nós (NN), altura do primeiro cacho (APC), teor de óleo (TO) e massa de 100 grãos (M100), nos três componentes principais (CP1, CP2 e CP3) da CDpca modificada

Variáveis	Componentes		
	[,1]	[,2]	[,3]
NC	0,4862	0	0
NF	0,4752	0	0
NS	0,5267	0	0
AP	0	0	0,7037
DC	0	0	0,6759
NN	0	0,7240	0
APC	0	0,6898	0
TO	0	0	-0,2191
M100	0,5101	0	0

E o agrupamento dos objetos seguiu o disposto na Tabela 21.

Tabela 21 – Classificação dos híbridos *Ricinus communis* L. pela CDpca modificada

Grupos	Objetos
1	2, 3, 10, 17, 25
2	23
3	4, 7, 16, 31
4	29
5	5
6	18
7	22
8	8, 20, 21
9	24, 26, 27
10	9
11	11
12	12, 30
13	1, 13, 19
14	14
15	6, 15, 28

3.4 CONCLUSÃO

A *Clustering and Disjoint Principal Component Analysis* (CDpca) quando comparada a Análise de Componentes Principais (ACP), apresenta algumas vantagens como a facilidade de interpretação de seus resultados. Entretanto, ao comparar os resultados de ambas as técnicas, algumas divergências foram encontradas como: porcentagem de explicação das componentes, variáveis selecionadas em cada componente bem como a escolha de objetos. Além disto, divergências maiores foram encontradas quando se trabalha com 3 componentes, a qual a modificação proposta se mostrou como alternativa para tentar resolver esta questão. Todavia, é interessante ressaltar a necessidade de estudos complementares para melhor entendimento e utilização da CDpca em estudos de melhoramento genético.

REFERÊNCIAS

BIRNBAUM A.; JOHNSTONE I. M.; NADLER B.; PAUL, D. Minimax bounds for sparse PCA with noisy high-dimensional data. **Ann. Statist.** v. 41, n. 3, p. 1055–1084, 2013. (doi:10.1214/12-AOS1014).

CRUZ, C. D. Aplicação de algumas técnicas multivariadas no melhoramento de plantas. 188 p. Tese (Doutorado) – ESALQ/USP, Piracicaba, 1990.

DESARBO, W. S.; JEDIDI, K.; COOL, K.; SCHENDEL, D. Simultaneous Multidimensional Unfolding and Cluster Analysis: An Investigation of Strategic Groups, **Marketing Letters**, v. 2, p. 129-146, 1990.

DE SORTE, G.; CARROL, J.D. K-means clustering in a low-dimenNew Approached in Classification and Data Analysis. **Springer**, p. 212-219, 1994.

FIDALSKI, J.; TORMENA, C. A.; SCAPIM, C. A. Espacialização vertical e horizontal dos indicadores de qualidade para um latossolo vermelho cultivado com citros. **Revista Brasileira de Ciência do Solo**, v. 31, n. 1, p. 9-19, 2007.

GODOI, C. R. de M. Análise estatística multidimensional. In: Simpósio de estatística aplicada a experimentação agronômica, v. 1. Piracicaba: USP/ESALQ, 1985. 187 p.

JOLLIFFE I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Phil. Trans. R. Soc. A** v. 374, n. 13, 2016.

LEE, D.; LEE, W.; LEE, Y.; PAWITAN, Y. Super-sparse principal component analyses for highthroughput genomic data. **BMC Bioinformatics**. v. 11, n. 296, 2010.

MACEDO, E.; FREITAS, A. The alternating least-squares algorithm for CDPCA. Optimization in the natural sciences, **Communications In Computer And Information Science**, v. 499, Springer, p. 173–191, 2015.

BARREIRO, C. M.; FIGUEROA, J. R.; CABEZAS, X.; LEIVA, V.; VILLARDON, M. G. Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data. **Sensors**, 2021, v. 21, 4094 p.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: UFMG, 295 p. 2005.

LIBRERO, A. B. N.; VILLARDÓN, M. P. G.; FREITAS, A. biplotbootGUI: Bootstrap on Classical Biplots and Clustering Disjoint Biplot, **R package**, v. 1.2, 2019. <http://www.R-project.org/package=biplotbootGUI>

PIMENTEL GOMES, F. **Curso de estatística experimental**. 13. ed. Piracicaba: Nobel, 1990. 468 p.

R Development Core Team (2023) R: a language and environment for statistical computing. <http://www.Rproject.org/>

RENCHE, A. C. **Methods of Multivariate Analysis**. 2. ed. Inc. Publication. p. 727. 2002.

VICHI, M.; KIERS, H. A. L. Factorial k-means analysis for two-way data.

Computational Statistics and Data Analysis, v. 37, p. 49–64, 2001.

WILKS, D. S. **Statistical Methods in the Atmospheric Sciences**. 2. ed., Academic Press, London, 2006.

VICHI, M.; SAPORTA, G. Clustering and disjoint principal component analysis.

Computational Statistics and Data Analysis. v. 53, p. 3194–3208, 2009.

CONSIDERAÇÕES FINAIS

A Análise de Componentes Principais (ACP) destaca-se como uma ferramenta valiosa, explorada em diversos contextos científicos. No entanto, é crucial reconhecer que novas modificações, como a *Clustering and Disjoint Principal Component Analysis* (CDpca) têm emergido, evidenciando vantagens significativas em sua aplicação.

A CDpca, ao se diferenciar da ACP convencional, apresenta uma abordagem inovadora e promissora, revelando potencialidades ainda não totalmente exploradas. Explorar e disseminar a compreensão sobre a CDpca torna-se imperativo para o conhecimento geral, promovendo uma visão abrangente das possibilidades oferecidas por essa abordagem. Sua aplicação não apenas enriquece a gama de técnicas disponíveis, mas também proporciona uma perspectiva mais refinada e adaptável às nuances presentes em diversas disciplinas científicas.

Além disso, a necessidade de tornar a CDpca acessível e de fácil manipulação não pode ser subestimada. Ao simplificar os processos envolvidos, pode-se ampliar o alcance e a adoção dessa técnica, permitindo que pesquisadores, cientistas e profissionais de diferentes áreas, inclusive, do melhoramento genético, possam integrar eficazmente a CDpca em suas análises e investigações.

REFERÊNCIAS

ALENCAR, F. R. G. de. Recursos genéticos de *Vigna Unguiculata*: variabilidade genética e potencial de uso no melhoramento genético. 2019.

BIRNBAUM A.; JOHNSTONE I. M.; NADLER B.; PAUL, D. Minimax bounds for sparse PCA with noisy high-dimensional data. **Ann. Statist.** v. 41, n. 3, p. 1055–1084, 2013. (doi:10.1214/12-AOS1014).

CRUZ, C. D. Aplicação de algumas técnicas multivariadas no melhoramento de plantas. 188 p. Tese (Doutorado) – ESALQ/USP, Piracicaba, 1990.

DUARTE, J. B. Princípios e utilização de técnicas multivariadas no melhoramento de plantas. Monografia. Departamento de Genética. ESALQ-USP. Piracicaba, 1997.

GODOI, C. R. de M. Análise estatística multidimensional. In: Simpósio de estatística aplicada a experimentação agrônômica, v. 1. Piracicaba: USP/ESALQ, 1985. 187 p.

GREVENIOTIS, V.; BOULOUMPASI, E.; TSAKIRIS, I.; SIOKI, E.; IPSILANDIS, C. Evaluation of Elite Open-Pollinated Maize Lines in Two Contrasting Environments. **Journal of Agricultural Science**, v. 10, n. 1, p. 85–101, 2018. <https://doi.org/10.5539/jas.v10n1p85>

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417– 441, 1933. <https://doi.org/10.1037/h0071325>

HOTELLING, H. Relations between two sets of variates, **Biometrika**, v. 28, p. 321–377, 1936. <https://doi.org/10.1093/biomet/28.3-4.321>.

HOTELLING, H. The most predictable criterion. **Journal of Educational Psychology**, v. 26, n. 2, p. 139–142, 1935. <https://doi.org/10.1037/h0058165>

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 3. ed. New Jersey: Prentice-Hall Intern., 1992. 642 p.

JOLLIFFE I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Phil. Trans. R. Soc. A**. v. 374, n. 13, 2016.

LEE, D.; LEE, W.; LEE, Y.; PAWITAN, Y. Super-sparse principal component analyses for highthroughput genomic data. **BMC Bioinformatics**. v. 11, n. 296, 2010.

MOURA, D. S.; BRITO, G. G.; MORAES, Í. L.; FAGUNDES, P. R. R.; CASTRO, A. P.; DEUNER, S. Cold Tolerance in Rice Plants: Phenotyping Procedures for Physiological Breeding. **Journal of Agricultural Science**, v. 10, n. 1, p. 313–324, 2018.

OLIVEIRA, T. R. A. D.; GRAVINA, G. D. A.; OLIVEIRA, G. H. F. D.; ARAÚJO, K. C.; ARAÚJO, L. C. D.; DAHER, R. F.; VIVAS, M.; GRAVINA, L. M.; CRUZ, D. P. D. The GT biplot analysis of green bean traits. **Ciência Rural**, v. 48, n. 6, 2018.

PANERO, J. S.; SILVA, H. E. B. da; PANERO, P. S.; SMIDERLE, O. J.; PANERO, F. S.; FARIA, F. S. E. D. V.; RODRIGUEZ, A. F. R. Separation of Cultivars of Soybeans by Chemometric Methods Using Near Infrared Spectroscopy. **Journal of Agricultural Science**, v. 10, n. 4, p. 351–360, 2018.

PIMENTEL GOMES, F. **Curso de estatística experimental**. 13. ed. Piracicaba: Nobel, 1990. 468 p.

RAMSAY J. O.; SILVERMAN B. W. Functional data analysis. **Springer Series in Statistics**. New York: Springer, 2006.

SOUZA, J. de. Análise em componentes principais e suas aplicações, métodos estatísticos nas ciências sociais. Brasília: **Thesaurus**, v. 2, 1988, 68 p.

VICHI, M.; SAPORTA, G. Clustering and disjoint principal component analysis. **Computational Statistics and Data Analysis**. v. 53, p. 3194–3208, 2009.

APÊNDICE A – Script da CDpca Modificada em R

```
#####

#      Function RandMat      #
#####

RandMat <- function(dim1, dim2) {
# Generates a random binary and row stochastic matrix.
#
# Args:
# dim1: number of rows.
# dim2: number of columns.
#
# Returns:
# The random matrix (dim1xdim2) with only one nonzero element per row.
#
  U0 <- matrix(0, dim1, dim2)
  U0[1:dim2, 1:dim2] <- diag(dim2)
  for (j in (dim2+1):dim1){
    p <- sample(1:dim2, 1)
    U0[j, p] <- 1
  }
  U0
} # end RandMat function

#####

# SDP initialization #
```

```
#####
SDPinitialization <- function(dim1,dim2,D,data){
# Generates a binary and row stochastic matrix
# using SDP-based approach
#
# Args:
# dim1: number of rows.
# dim2: number of columns.
#
# Returns:
# The random matrix (dim1xdim2) with only one nonzero element per row.
#
# approximate solution for clustering
em <- cbind(rep(1,dim1))
matlem <- diag(dim1)-(1/dim1)*em%*%t(em)
if (dim1 == dim(data)[1]){
  data <- D
}else{
  data <- t(D)
}
So <- matlem%*%data%*%t(data)%*%matlem
matF <- matrix(svd(So)$v[, 1:(dim2-1)], ncol = dim2-1)
matFFT <- matF%*%t(matF)
# approximate solution to the SDP-based model for clustering
Z.bar <- matFFT + 1/dim1 *em%*%t(em)

# refine solutions
cent <- Z.bar%*%data
centunique <- unique(cent)
```

```

t <- dim(centunique)[1]

# initial kmeans step
indcentroids <- sort(sample(1:t, dim2, replace=FALSE), decreasing=FALSE)
centroids <- centunique[indcentroids, ]
solkmeans <- kmeans(data, centroids, iter.max = 100000)
M <- matrix(0, dim1, dim2)
for (i in 1:dim1){
  M[i, solkmeans$cluster[i]] <- 1
}
M
}

#####
#          Function CDpca          #
#####

CDpca <- function(data, class, Vfix, P, Q, SDPinitial, tol=10^(-5), maxit, r, cdpcaplot)
{
# CDpca performs a clustering and disjoint principal components analysis
# on the given numeric data matrix and returns a list of results
#
# Args:
# data: data frame (numeric).
# class: vector (numeric) or 0, if classes of objects are unknown.
# fixAtt: vector (numeric) or 0, for a selection of attributes.
# Vfix: 0 if V is not prefixed previously and 1 if V is based on the maximum
coefficient by original variable until Q principal components
# SDPinitial: 1 or 0, for random initialization of U and V
# nnloads: 1 or 0, for nonnegative loadings
# cdpcaplot: integer: 1 or 0, if no plot is displayed.

```

```

# Q: integer, number of clusters of variables.
# P: integer, number of clusters of objects.
# tol: real number, small positive tolerance.
# maxit: integer, maximum of iterations.
# r: number of runs of the cdpca algorithm for the final solution.
#
# Returns:
# iter: iterations used in the best loop for computing the best solution
# loop: best loop number
# timebestloop: computation time on the best loop
# timeallloops: computation time for all loops
# Y: the component score matrix
# Ybar: the object centroids matrix in the reduced space
# A: the component loading matrix
# U: the partition of objects
# V: the partition of variables
# F: function to maximize
# bcdev: between cluster deviance
# bcdevTotal: between cluster deviance over the total variability
# tableclass: cdpca classification
# pseudocm: pseudo confusion matrix of the real and cdpca classifications
# Enorm: error norm for the obtained cdpca model
#CDPCA is applied on normalized data (mean zero and unit variance)
data.sd <- data.frame(scale(data, center = TRUE, scale = TRUE))
I <- dim(data)[1] # number of objects I
J <- dim(data)[2] # number of variables J
Xs <- as.matrix(data.sd*sqrt(I/(I-1))) # matrix of normalized data (with variance
divided by I)
# matriz X (I x J) (obs x vars), numeric matrix argument
tfbest <- matrix(0, r, 1) # computational time at each loop

```

```
# Run CDPCA ALS algorithm r times
```

```
for (loop in 1:r) {
```

```
  t1 <- proc.time()
```

```
  # Initialization
```

```
  iter <- 0
```

```
# SDP initialization or not?
```

```
if (SDPinitial){
```

```
  U <- SDPinitialization(I, P, Xs, data)
```

```
  V <- SDPinitialization(J, Q, Xs, data)
```

```
}else{
```

```
  U <- RandMat(I, P)
```

```
  V <- RandMat(J, Q)
```

```
}
```

```
#####
```

```
#MODIFICATION TO INCLUDE THE RESTRICTION TO FIX THE MATRIX V FROM  
STANDARD PCA ----- Gabriela
```

```
if (Vfix==1)
```

```
  { V=matrix(0,J,Q)
```

```
    for (j in 1:Q)
```

```
  {
```

```
    indi=which.max(abs(prcomp(Xs,scale=T)$rotation[1:dim(Xs)[2],j]))
```

```
    V[indi,j]=1
```

```
  }
```

```
SumMarginal=apply(V,1,sum)
```

```
With0=which(SumMarginal!=1)
```

```
if (length(With0)!=0)
```

```

    {
        for (k in 1:length(With0))
            {

indj=which.max(abs(prcomp(Xs,scale=T)$rotation[With0[k],1:Q]))

                V[With0[k],]=0
                V[With0[k],indj]=1
            }
        }
    }

#####
#####

#Set of the variables

X.bar <- diag(1/colSums(U))%*%t(U)%*%Xs      # (PxJ) object centroid matrix. It
identifies the P centroids in the variable space

X.group <- U%*%X.bar                          # (IxJ) matrix. Each object in
the data matrix is replaced by its centroid.

vJ <- matrix(1:J, ncol = 1)

zJ <- matrix(0, J, 1)

zQ <- matrix(0, 1, Q)

# matrix A

A <- matrix(0, J, Q)

for (j in 1: Q) {
    Jxx <- which(V[, j] == 1)
    len.Jxx <- length(Jxx)
    if (sum(V[, j])>1) {
        if (I >= len.Jxx) {
            # CASE 1: I >= J (more observations than variables)

```

```

S.group <- t(X.group[, Jxx])%*%X.group[, Jxx]
A[Jxx, j] <- powerMethod(S.group)$vector

}else{
  # CASE 2: I < J ( more variables than observations)
  SS.group <- X.group[, Jxx]%*%t(X.group[, Jxx])
  PMinial <- powerMethod(SS.group)

  A[Jxx, j] <- t(X.group[, Jxx])%*%PMinial$vector/sqrt(PMinial$value)
} # end if
}else{
  A[Jxx, j] <- 1
} # end if
} # end for

A0 <- A

Y.bar <- X.bar%*%A # (PxQ) object centroid matrix. It
identifies the P centroids in the reduced space of the Q principal components.

F0 <- sum(diag(t(U%*%Y.bar)%*%U%*%Y.bar)) # Objective function to
maximize.

Fmax <- F0
conv <- 2*tol
while (conv > tol) {
  iter <- iter+1

  Y <- Xs%*%A # (IxQ) component score matrix. It
identifies the I objects in the reduced space of the Q principal components.

  U <- matrix(0, I, P)

  # update U
  for (i in 1:I) {
    dist <- rep(0, P)
    for (p in 1:P) {

```

```

    dist[p] <- sum((Y[i, ]-Y.bar[p, ])^2)
  } # end for
  min.dist <- which.min(dist)
  U[i, min.dist] <- 1
} # end for
su <- colSums(U)
while (sum(su == 0) > 0) {
  ind.max <- which.max(su) # p2
  su.max <- max(su) # m2
  ind.min <- which.min(su) # p1
  su.min <- min(su) # m1
  ind.nzU <- which(U[, ind.max] == 1)
  ind.sel <- ind.nzU[1:floor(su.max)/2]
  U[ind.sel, ind.min] <- 1
  U[ind.sel, ind.max] <- 0
  su <- colSums(U)
} # end while

```

Given U and A compute X.bar

```

X.bar <- diag(1/colSums(U))%*%t(U)%*%Xs
Y.bar <- X.bar%*%A
X.group <- U%*%X.bar

```

Update V and A

#####

#MODIFICATION TO INCLUDE THE RESTRICTION TO FIX THE MATRIX V FROM
STANDARD PCA ----- Gabriela

```

if (Vfix==1)

```

```

{

```

```

    #Update V

```

```

V0=V
V=matrix(0,J,Q)
for (j in 1:Q)
{
    indi=which.max(abs(prcomp(Xs,
scale=T)$rotation[1:(dim(X.group)[2]),j])) # X.group
    V[indi,j]=1
}
SumMarginal=apply(V,1,sum)
With0=which(SumMarginal!=1)
if (length(With0)!=0)
{
    for (k in 1:length(With0))
    {
        indj=which.max(abs(prcomp(Xs,scale=T)$rotation[With0[k],1:Q])) #X.group
        V[With0[k],]=0
        V[With0[k],indj]=1
    }
}

# Update A
A <- matrix(0, J, Q)
for (j in 1: Q) {
    Jxx <- which(V[, j] == 1)
    len.Jxx <- length(Jxx)
    if (sum(V[, j])>1) {
        if (I >= len.Jxx) {
            # CASE 1: I >= J (more observations than variables)
            S.group <- t(X.group[, Jxx])%*%X.group[, Jxx]

```

```

A[Jxx, j] <- powerMethod(S.group)$vector

}else{
  # CASE 2: I < J ( more variables than observations)
  SS.group <- X.group[, Jxx]%*%t(X.group[, Jxx])
  PMinial <- powerMethod(SS.group)

  A[Jxx, j] <- t(X.group[, Jxx])%*%PMinial$vector/sqrt(PMinial$value)
} # end if
}else{
  A[Jxx, j] <- 1
} # end if
} # end for

#####

Y.bar <- X.bar%*%A
F <- sum(diag(t(U%*%Y.bar)%*%U%*%Y.bar))
if (F > Fmax) {
  Fmax <- F
  A0 <- A
}else{
  A <- A0
  V <- V0
} # end if

}      #introduzido para a Gabriela
#####

else   #introduzido para a Gabriela
{      #introduzido para a Gabriela
  for (j in 1:J) {
    posmax <- which(V[j, ] == 1)

```

```

for (g in 1:Q) {
  V[, ] <- diag(Q)[g, ]
  #####
  if (g!=posmax) { #### if 1
    for (gg in c(g,posmax)) {
xx <- V[, gg]
Jxx <- which(xx == 1)
len.Jxx <- length(Jxx)
A[, gg] <- zJ
if (sum(xx) > 1) {
  if (I >= len.Jxx) {
    # CASE 1: I >= J (more observations than variables)
    S.group <- t(X.group[, Jxx])%*%X.group[, Jxx]
    A[Jxx, gg] <- matrix(powerMethod(S.group)$vector, nrow = len.Jxx)

  }else{
    # CASE 2: I < J (more variables than observations)
    SS.group <- X.group[, Jxx]%*%t(X.group[, Jxx])
    PMgeneral <- powerMethod(SS.group)

    A[Jxx, gg] <- matrix((t(X.group[,
Jxx])%*%PMgeneral$vector/sqrt(PMgeneral$value))[, 1], nrow = len.Jxx)
  } # end if
}else{
  if (sum(xx)==1) {
    A[Jxx, gg] <- 1
  } # end if
} # end if
  } # end for
} # end #### if 1
#####

```

```

Y.bar <- X.bar%*%A
F <- sum(diag(t(U%*%Y.bar)%*%U%*%Y.bar))
if (F > Fmax) {
  Fmax <- F
  posmax <- g
  A0 <- A
}else{
  A <- A0
} # end if
} # end for
V[j, ]=diag(Q)[posmax, ]
} # end for

}      #introduzido para a Gabriela

```

```

Y <- Xs%*%A
Y.bar <- X.bar%*%A
F <- sum(diag(t(U%*%Y.bar)%*%U%*%Y.bar))
conv <- F-F0
if (conv > tol) {
  F0 <- F
  A0 <- A
}else{
  break
} # end if
if (iter == maxit) {
  print("Maximum iterations reached.")
  break
} # end if

```

```

} # end while

# Computation time for each loop
t2 <- proc.time()
tfinal <- t2-t1
tfbest[loop] <- tfinal[3]

#Results to be observed in each run of CDpca
# BetClusDevTotal <- F/(I*J)*100 # between cluster deviance
(two different formula for the same thing)
BetClusDev <- (F/sum(diag(t(Y)%*%Y)))*100 # between cluster
deviance
tabperloop <- data.frame(cbind(loop, iter, tfinal[3], BetClusDev, F, conv))
rownames(tabperloop) <- c(" ")
colnames(tabperloop) <- c("Loop", "Iter", "Loop time", "Between cluster
deviance(%)", "F", "Convergence")
print(tabperloop) # Loop display
if (loop == 1) {
  Vcdpca <- V
  Ucdpca <- U
  X.barcdpca <- diag(1/colSums(Ucdpca))%*%t(Ucdpca)%*%Xs
  Acdpca <- A
  Ycdpca <- Xs%*%Acdpca
  Y.barcdpca <- X.barcdpca%*%Acdpca
  Fcdpca <- F
  loopcdpca <- 1
  itercdpca <- iter
  convcdpca <- conv
} # end if
if (F > Fcdpca) {
  Vcdpca <- V

```

```

Ucdpca <- U
X.barcdpca <- diag(1/colSums(Ucdpca))%*%t(Ucdpca)%*%Xs
Acdpca <- A
Ycdpca <- Xs%*%Acdpca
Y.barcdpca <- X.barcdpca%*%Acdpca
Fcdpca <- F
loopcdpca <- loop
itercdpca <- iter
convcdpca <- conv
} # end if
} # end for loop

# Computation time for all loops
tftotal=sum(tfbest)

# maximum between cluster deviance
BetClusDevTotal <- Fcdpca/(I*J)*100
BetClusDev <-
(sum(diag(t(Ucdpca%*%Y.barcdpca)%*%(Ucdpca%*%Y.barcdpca)))/sum(diag(t(Ycd
pca)%*%Ycdpca)))*100

# error in cdpca model
Enormcdpca <- 1/I*norm(Xs-Ucdpca%*%Y.barcdpca%*%t(Acdpca), "F")

# Table Real vs CDPCA classification
classcdpca <- Ucdpca%*%as.matrix(1:ncol(Ucdpca))
if (!is.null(class)){
class <- data.frame(class)
maxclass <- max(class)
}

```

```
# Variability of Ycdpca and in decreasing order
```

```
varYcdpca <- var(Ycdpca)
```

```
d <- round(diag(varYcdpca)*100/sum(diag(var(Xs))), 2)
```

```
#####alterado para a Gabriela
```

```
dorder <- d[order(d, decreasing = TRUE)]
```

```
tablevar <- data.frame(1:Q, d)
```

```
colnames(tablevar) <- c("Dim", "Var (%)")
```

```
# Presentation of the matrices associated to the CDPCA component sorted by their
variances.
```

```
Yorder <- t(t(rbind(d, Ycdpca))[order(d, decreasing = TRUE), ])[-1, ]
```

```
Ybarorder <- t(t(rbind(d, Y.barcdpca))[order(d, decreasing=TRUE), ])[-1, ]
```

```
Aorder <- t(t(rbind(d, Acdpca))[order(d, decreasing = TRUE), ])[-1, ]
```

```
Vorder <- t(t(rbind(d, Vcdpca))[order(d, decreasing = TRUE), ])[-1, ]
```

```
#
```

```
# We can check the model using these column sort matrices and observe that
```

```
# Ucdpca%*%Y.barcdpca%*%t(Acdpca) = Ucdpca%*%Ybarorder%*%t(Aorder)
```

```
if ( is.null(class) ) {
```

```
  realclasstrue <- 2
```

```
  pseudocm <- NULL
```

```
  tabrealcdpca <- data.frame(classcdpca)
```

```
  colnames(tabrealcdpca) <- c("CDPCA Class")
```

```
}else{
```

```
  realclasstrue <- 3
```

```
  tabrealcdpca <- data.frame(class, classcdpca)
```

```
  colnames(tabrealcdpca) <- c("Real Class", "CDPCA Class")
```

```
  # pseudo-confusion matrix
```

```

pseudocm <- table(tabrealcdpca)
} # end if

# PLOTS: CDPCA classification
if (cdpcaplot) {

  displaygraphics <- par(no.readonly = TRUE)
  par(mfrow = c(1, realclasstrue))
  if (realclasstrue == 3) {
    # plot1
    matplot(Yorder[, 1], Yorder[, 2], xlab = "Dim 1", ylab = "Dim 2", type = "n", main =
"Real classification")
    for (i in 1:maxclass) { points(Yorder[, 1][class == i], Yorder[, 2][class == i], pch =
1, col = i+1) }
    # plot2
    matplot(Yorder[, 1], Yorder[, 2], xlab = paste(" Dim 1  (", (dorder[1]), " % )"), ylab
= paste(" Dim 2  (", (dorder[2]), " % )"), type = "n", main = "CDPCA classification")
    for (i in 1:P) { points(Yorder[, 1][classcdpca == i], Yorder[, 2][classcdpca == i], pch
= 2, col = i+1) }
    points(Ybarorder[, 1], Ybarorder[, 2], pch = 15) # introduce the centroids into the
plot
    # plot3 (legend)
    matplot(Y.barcdpca, type = "n", axes = FALSE, xlab = "", ylab = "")
    legend("topleft", pch = 1, col=c(2:(maxclass+1)), legend = c(1:maxclass), ncol =
maxclass)
    legend("bottomleft", pch = 2, col=c(2:(P+1)), legend = 1:P, ncol = P)
  }else{
    # plot2
    matplot(Yorder[, 1], Yorder[, 2], xlab = paste(" Dim 1  (", (dorder[1]), " % )"), ylab
= paste(" Dim 2  (", (dorder[2]), " % )"), type = "n", main = "CDPCA classification")
    for (i in 1:P) { points(Yorder[, 1][classcdpca == i], Yorder[, 2][classcdpca == i], pch
= 2, col = i+1)}
  }
}

```

```

    points(Ybarorder[, 1],Ybarorder[, 2], pch = 15) # introduce the centroids into the
plot
    # plot3 (legend)
    matplot(Y.barcdpca, type = "n", axes = FALSE, xlab = "", ylab = "")
    legend("bottomleft", pch = 2, col=c(2:(P+1)), legend = 1:P, ncol = P)
  }    # end if
} #end if cdpcaplot

```

```

# OUTPUT

```

```

list(timeallloops = tftotal,
      timebestloop = tfbest[loopcdpca],
      loop = loopcdpca,
      iter = itercdpca,
      bcdevTotal = BetClusDevTotal ,
      bcdev = BetClusDev ,
      V = Vorder,
      U = Ucdpca,
      A = Aorder,
      F = Fcdpca,
      Enorm = Enormcdpca,
      tableclass = tabrealcdpca,
      pseudocm = pseudocm,
      Y = Yorder,
      Ybar = Ybarorder,
      dorder = dorder,
      Xscale = Xs
    )
} # end CDpca function

#####

```

```

#      Function CDpcaSummary      #
#####

CDpcaSummary <- function(obj) {
  cat("Summary of Cluster and Disjoint Principal Component Analysis:\n")
  print(rbind("Loop with best result:" = obj$loop, "Number of iterations:" = obj$iter,
    "F:" = round(obj$F, 5), "Error:" = round(obj$Enorm, 5),
    "Between cluster deviance (%):" = round(obj$bcdev, 2)))
  print(rbind("Explained variance by CDpca components" = obj$dorder))
  cat("Pseudo Confusion Matrix:\n")
  print(obj$pseudocm)
  print(obj$RES)
  invisible(obj)
} # end CDpcaSummary function

```