

Atendendo solicitação do(a) autor(a), o texto completo desta tese/dissertação será disponibilizado somente a partir de 03/06/2026

At the author's request, the full text of this thesis/dissertation will not be available online until Jun. 03, 2026

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



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 coorientação. 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

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 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 agrônômicos.

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.