JAQUICELE APARECIDA DA COSTA

AUTOENCODER, ANÁLISE VIA COMPONENTES PRINCIPAIS E INDEPENDENTES APLICADOS NO RECONHECIMENTO DE PADRÕES DE POPULAÇÕES

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientadora: Camila Ferreira Azevedo

Coorientadores: Ana Carolina C. Nascimento
Isabela de Castro Sant'Anna
Moysés Nascimento

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
C837a
2022

Costa, Jaquicele Aparecida da, 1990-
Autoencoder, análise via componentes principais e independentes aplicados no reconhecimento de padrões de populações / Jaquicele Aparecida da Costa. – Viçosa, MG, 2022.
1 tese eletrônica (62 f.): il. (algumas color.).

Inclui apêndice.
Orientador: Camila Ferreira Azevedo.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.537
Modo de acesso: World Wide Web.

1. Sistemas de reconhecimento de padrões. 2. Inteligência computacional. 3. Redes neurais (Computação). 4. Análise dimensional. 5. *Oryza sativa* - Populações. 6. Marcadores genéticos - Métodos estatísticos. I. Azevedo, Camila Ferreira, 1988-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 006.37

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

JAQUICELE APARECIDA DA COSTA

AUTOENCODER, ANÁLISE VIA COMPONENTES PRINCIPAIS E INDEPENDENTES APLICADOS NO RECONHECIMENTO DE PADRÕES DE POPULAÇÕES

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 10 de junho de 2022.

Assentimento:

Jaquicele Aparecida da Costa.
Jaquicele Aparecida da Costa
Autora

Camila Ferreira Azevedo
Camila Ferreira Azevedo
Orientadora

AGRADECIMENTOS

A Deus, por me conduzir com o Teu amor, me fortalecer a cada dia e por me ajudar a enxergar sinais da eternidade em tudo que foi desenvolvido na pesquisa. À Virgem Maria, minha mãezinha, por interceder por tudo que faltava.

Aos meus pais, Maria Helena e José Raimundo, por não medirem esforços para que eu realizasse todos os meus sonhos, e meu amado irmão, Josimar, por trazer paz nos momentos mais difíceis com o seu jeito de ser. Agradeço também às minhas amigas/irmãs, Daniela e Simone, que fazem parte da família. Amo vocês!

Aos meus amigos da RCC Viçosa, agradeço pelos testemunhos e momentos de partilhas que fortaleceram minha fé. Obrigada por me ensinarem a edificar tudo, de modo especial, o presente trabalho, na verdadeira ROCHA. As orações foram como abraços diários de Deus!

Aos amigos do PPESTBIO: Alex Temoteo, Ana Carolina, Gabriela França, Gabriely Lazzarini, Leísa Pires, Lucas da Silveira e Roberta Amorim, pelas conversas, brincadeiras, almoços e jantares. A presença de vocês foi cuidado de Deus comigo ao longo do doutorado.

Agradeço imensamente a Doutora e orientadora Camila Ferreira Azevedo pela disponibilidade, por compartilhar todo conhecimento e, principalmente, por mostrar que vale a pena se dedicar a pesquisa. Mais uma vez, muito obrigada por tudo!

Ao Laboratório de Inteligência Computacional e Aprendizado Estatística (LICAE) e ao Grupo de Estudo em Estatística Aplicada e Biometria (GESTBIO), agradeço pela parceria, confraternizações e por todas as amizades que foram construídas no decorrer dos trabalhos.

À Universidade Federal de Viçosa, pela oportunidade de realizar a pós-graduação.

Aos Doutores e coorientadores Ana Carolina Campana Nascimento, Isabela de Castro Sant'Anna e Moysés Nascimento, pela disponibilidade, confiança, incentivo e pelos saberes transmitidos ao longo de todo o mestrado e doutorado.

Aos membros da banca examinadora, Doutor Cosme Damião e Cruz, Doutora Gabi Nunes Sila, Doutor Ivan de Paiva Barbosa e Doutor Moysés Nascimento, pelas valiosas contribuições neste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pelos saberes transmitidos.

Aos amigos e funcionários do Departamento de Estatística e do Programa de Pós-Graduação em Estatística Aplicada, de forma especial, Nayara e Junior, respectivamente, por se preocuparem e estarem sempre prontos para ouvir e acolher os estudantes. Vocês foram essenciais na minha trajetória!

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)-Código de Financiamento 001, pela concessão da bolsa de estudos.

Enfim, agradeço imensamente aos familiares, amigos e tantas outras pessoas que, de alguma forma, me ajudaram. Obrigada!

“Cabe ao homem formular projetos em seu coração, mas do Senhor vem a resposta língua. Todos os caminhos parecem puros ao homem, mas o Senhor é quem pesa os corações. Confia teus negócios ao Senhor e teus planos terão bom êxito.”

(Provérbios 16, 1-3)

RESUMO

COSTA, Jaquicele Aparecida da, D.Sc., Universidade Federal de Viçosa, junho de 2022. **Autoencoder, Análise via Componentes Principais e Independentes Aplicados no Reconhecimento de Padrões de Populações**. Orientadora: Camila Ferreira Azevedo. Coorientadores: Ana Carolina Campana Nascimento, Isabela de Castro Sant'Anna e Moysés Nascimento.

Nos últimos tempos, diante do grande volume de informações, é essencial o desenvolvimento de metodologias que visam reduzir o tempo e esforço computacional da análise de dados com alta dimensionalidade. Nos estudos que buscam associações ou o reconhecimento de padrões, há um grande número de variáveis que apresentam informações sobrepostas ou correlacionadas, o que impossibilita a identificação de grupos divergentes, além de exigir um grande esforço computacional. A genética utiliza milhares de marcadores moleculares do tipo SNPs (*Single nucleotide polymorphisms*) para estimar os valores genéticos genômicos dos indivíduos, classificar genótipos dentro de determinados grupos e reconhecer padrões na população para direcionar os estudos de diversidade genética. Os principais métodos usados para redução de dimensionalidade são baseados em Análise via Componentes Principais (PCA), a versão esparsa da Análise via componentes Principais (SPCA) e Análise via Componentes Independentes (ICA). Outra técnica em destaque é a metodologia que combina os métodos PCA e ICA que é denominada Análise via Componentes Principais Independentes (IPCA), mas ainda pouco utilizada em banco de dados genômicos. Mais recentemente, têm se destacado os métodos fundamentados em inteligência artificial, como as redes neurais, sendo o Autoencoder um tipo de rede neural que também busca reduzir o espaço dimensional e reconstruir os dados com perda mínima de informação. Assim, o primeiro capítulo desta tese é uma revisão bibliográfica sobre os métodos estatísticos e baseados em inteligência computacional, destacando as vantagens e desvantagens ao utilizar cada uma das metodologias, além de apresentar as técnicas para agrupar e determinar o número ótimo de grupos nos estudos que visam reconhecer padrões. O segundo capítulo propõe a aplicação da PCA, SPCA e IPCA no reconhecimento de padrões de subpopulações do arroz asiático (*Oryza Sativa*) utilizando 36.901 marcadores moleculares e 413 genótipos, a fim de buscar uma técnica que seja eficiente e possa reduzir o tempo computacional na discriminação dos mesmos. As técnicas, PCA,

SPCA e IPCA, apresentaram resultados similares, tais como a matriz confusão, porcentagem de acerto e correlação cofenética. O método Autoencoder foi menos eficiente, mas foi capaz de formar grupos mais compactos, menor variância dentro dos grupos, e mais dissimilares entre eles, maior variância entre os grupos, quando comparado com os métodos estatísticos tradicionais. Diante disso, foi proposto utilizar os componentes obtidos via PCA, SPCA e IPCA, como variáveis de entrada no Autoencoder. A proposta provocou melhorias no Autoencoder, sendo que o PCA-AUT (componentes principais como variáveis de entrada no Autoencoder) foi mais eficiente que os métodos estatísticos e o próprio Autoencoder, além de reduzir ainda mais o espaço dimensional para discriminar os genótipos de arroz. Além disso, a técnica conseguiu capturar parte da variabilidade mensurada antes de aplicar qualquer método de redução dimensional.

Palavras-chave: Inteligência computacional. Redes Neurais. Redução de dimensionalidade. *Oryza sativa*. Marcadores Moleculares.

ABSTRACT

COSTA, Jaquicele Aparecida da, D.Sc., Universidade Federal de Viçosa, June, 2022. **Autoencoder, Independent and Principal Component Analysis Applied in the Recognition of Population Patterns**. Adviser: Camila Ferreira Azevedo. Co-advisers: Ana Carolina Campana Nascimento, Isabela de Castro Sant'Anna and Moysés Nascimento.

In recent times, given the large volume of information, it is essential to develop methodologies that aim to reduce the time and computational effort of analyzing high-dimensional data. Many variables present overlapping or correlated information in studies that aim to look for associations or recognize patterns, making it impossible to identify dissimilar groups and demanding a great computational effort. Genetics uses thousands of molecular markers such as SNPs (Single nucleotide polymorphisms) to estimate individuals' genomic-genetic values, classify genotypes within specific groups, and recognize patterns in the population to guide genetic diversity studies. The main methods used for dimensionality reduction are based on Principal Component Analysis (PCA), the sparse version of Principal Component Analysis (SPCA), and Independent Component Analysis (ICA). Another technique that stands out is the methodology that combines the PCA and ICA methods, called Independent Principal Component Analysis (IPCA), which is still rarely used in genomic databases. More recently, methods based on artificial intelligence, such as neural networks, have been highlighted. The Autoencoder is a type of neural network that also seeks to reduce dimensional space and reconstruct data with minimal loss of information. Thus, the first chapter is a literature review on statistical methods and methods based on computational intelligence, highlighting the advantages and disadvantages of using each of the methodologies and presenting the techniques to group and determine the optimal number of groups in studies that aim to recognize patterns. The second chapter proposes the application of PCA, SPCA, and IPCA in recognition of patterns of subpopulations of Asian rice, Oryza sativa, using 36,901 molecular markers and 413 genotypes to search for a technique that is efficient and can reduce the computational time in discrimination against them. PCA, SPCA, and IPCA presented similar results, such as the confusion matrix, percentage of correct answers, and cophenetic correlation. The Autoencoder method was less efficient, but it could form more compact groups, with minor variance within groups, and more dissimilar between them,

with greater variance between groups, compared to traditional statistical methods. Therefore, it was proposed to use the components obtained via PCA, SPCA, and IPCA as input variables in the Autoencoder. The proposal led to improvements in the Autoencoder. The PCA-AUT (main components as input variables in the Autoencoder) was more efficient than the statistical methods and the Autoencoder itself, in addition to further reducing the dimensional space to discriminate rice genotypes. Furthermore, the technique captured part of the measured variability before applying any dimensional reduction method.

Keywords: Computational Intelligence. Neural Networks. Dimensionality Reduction. Oryza sativa. Molecular Markers.

SUMÁRIO

INTRODUÇÃO GERAL

A seleção genômica utiliza milhares de marcadores SNPs (*Single Nucleotide Polymorphisms*) amplamente distribuídos ao longo do genoma (Meuwissen et al., 2001) para predizer os valores genéticos genômicos (*Genomic Estimated Breeding Values* - GEBVs) dos indivíduos, classificar ou reconhecer padrões, o que possibilita reduzir o intervalo de gerações no processo de seleção, e, consequentemente, minimizar os custos de um programa de melhoramento, e direcionar os estudos de diversidade genética ao identificar grupos de genótipos dissimilares. No entanto, as análises apresentam alguns desafios, como alta dimensionalidade, ou seja, o número de observações é inferior ao número de varáveis explicativas (marcadores), o que exige, em muitos casos, um elevado tempo computacional, e há a presença de multicolinearidade, marcadores altamente correlacionados devido ao desequilíbrio de ligação entre os marcadores, inviabilizando a aplicação do tradicional método de estimação de parâmetros, a saber, o método dos quadrados mínimos ordinários, nos casos de predição (Resende et al., 2014), e nos casos de classificação ou reconhecimento de padrões, as informações sobrepostas não permitem diferenciar os genótipos.

Dentre as várias metodologias utilizadas para solucionar os problemas citados como a predição de valores genômicos, Resende et al. (2014) destacam os métodos de regressão explícita, como baseados em *shrinkage*: RR-BLUP, LASSO, Rede Elástica (*Elastic Net*- EN); métodos bayesianos: como Bayes A, Bayes B, Fast BayesB, Bayes Cπ e BLASSO; regressão implícita, como Regressão Kernel, RKHS(*Reproducing Kernel Hilbert Spaces)* e Redes Neurais; e métodos estatísticos de redução de dimensionalidade baseados em Análise de componentes principais (*Principal Component Analysis* - PCA), Análise de componentes independentes (*Indpendent Component Analysis* - ICA) e Análise de componentes principais independentes (*Independent Principal Component Analysis* – IPCA), ainda pouco explorada em bancos de dados genômicos e que pode ser usado para a predição, mas também pode ser utilizado em estudos de reconhecimento de padrões e de diversidade genética. Os métodos estatísticos de redução de dimensionalidade se destacam pela facilidade de aplicação e entendimento da teoria e já foram aplicados em vários estudos com o objetivo de predição (de los Campos et al., 2013; Azevedo et al., 2013, 2014; Costa et al., 2020, 2021).

Ainda no contexto de redução de dimensionalidade, mas sob a abordagem de inteligência computacional, tem se destacado o Autoencoder, uma rede neural artificial que utiliza um número reduzido de neurônios artificiais em relação ao número de observações que são apresentadas como entradas e algoritmos de aprendizagem para treinamento da rede a fim de reduzir a dimensão dos dados e reconstruí-los de forma a conduzir a uma perda mínima de informação (Fusi et al, 2016). A vantagem das metodologias baseadas em redes neurais é a capacidade de capturar estruturas complexas por meio das funções de ativação que são empregadas, além de não exigir quaisquer pressuposições. Em contrapartida, as abordagens estatísticas tradicionais, assumem que a relação entre os marcadores é linear, o que pode dificultar o reconhecimento de padrões em populações que não atendam tal exigência.

O primeiro capítulo desta tese é uma revisão bibliográfica sobre os métodos estatísticos e baseados em inteligência computacional, destacando as vantagens e desvantagens ao utilizar cada uma das metodologias, além de apresentar as técnicas para agrupar e determinar o número ótimo de grupos nos estudos que visam reconhecer padrões. O segundo capítulo propõe a aplicação da PCA, SPCA e IPCA no reconhecimento de padrões de seis subpopulações do arroz asiático, *Oryza sativa*, utilizando 36.901 marcadores moleculares e 413 genótipos, a fim de buscar uma técnica que seja eficiente e possa reduzir o tempo computacional na discriminação dos mesmos.

O arroz asiático, *Oryza sativa*, destaca-se por ser consumido em grande escala pela população mundial e por ser uma cultura cujo genoma foi o primeiro a ser totalmente sequenciado e ser uma planta autógama, o que leva a conjectura de que os grupos de genótipos mais dissimilares geneticamente advêm de diferentes ambientes (Garris et al., 2005). Historicamente, já existe uma diferenciação entre os dois principais grupos: Indica e Japônica, mas algumas pesquisas já possibilitaram a identificação de cinco grupos diferentes (Garris et al., 2005; Ammiraju et al., 2006; Zhao et al., 2010 e 11), sendo as subespécies identificadas: *Indica*, AUS, *Temperate Japônica*, *Tropical Japônica*, *Aromatic*. Assim, diante do aumento da população mundial, torna-se essencial desenvolver técnicas capazes de extrair genótipos de grupos contrastantes que possam ser genitores na busca pela potencialização da cultura de arroz (Pandey et al., 2009).

CAPÍTULO 1

REVISÃO DE LITERATURA

1. Reconhecimento de padrões

Nos últimos tempos, o grande volume de informações disponíveis contribuiu para a busca de metodologias alternativas às técnicas estatísticas, visando possibilitar a visualização dos mesmos em uma dimensão inferior, construir modelos estatísticos mais parcimoniosos, facilitando a interpretação dos parâmetros; e redução de tempo e esforço computacional na execução das análises (Liu et al., 2014; James et al., 2013). Dentre estas metodologias alternativas que englobam inteligência artificial e aprendizado de máquinas, destacam-se algumas aplicações, como a identificação de anomalias com o desenvolvimento da computação pervasiva (Erfani et al., 2016; Ye et al., 2012), análise de fatores na econometria (Bai e Wang, 2016; Mullainathan e Spiess, 2017), análise de grupos de dados de espectroscopia NIR (Crase et al., 2021) e estudos de reconhecimento de padrões (Silva et al., 2010).

Os estudos de reconhecimento de padrões são essenciais nas tomadas de decisão, tendo aplicações em diversas áreas, por exemplo, na medicina para diagnóstico precoce do câncer (Cheng e Zhan, 2017; Harding et al., 2017), análises de neuroimagem para a identificação de doenças (Klöppel et al., 2012), avaliação de crédito (Nazari e Alidadi, 2013) e nos estudos que visam avaliar a variabilidade genética presente entre e dentro das espécies (Ellegren e Galtier, 2016; Ogwu e Osawaru, 2016; Petersen et al., 2013). A variabilidade genética pode ser mensurada utilizando informações diretamente do DNA, por meio de marcadores moleculares (Cruz et al., 2011). No entanto, as análises que utilizam marcadores do tipo SNPs, abundantemente presentes no genoma, enfrentam os desafios de multicolinearidade e de alta dimensionalidade, o que dificulta identificar diferentes grupos de genótipos a partir de um denso conjunto de informações genotípicas.

Nesse contexto, os métodos de redução de dimensionalidade são amplamente utilizados, principalmente, pela facilidade de aplicação e entendimento da teoria. Estes métodos utilizam variáveis latentes (componentes) que consistem em combinações lineares das variáveis originais, auxiliando a resumir as informações contidas nas variáveis originais e a reduzir o espaço dimensional do problema e, assim, facilitar o entendimento das mesmas, bem como o reconhecimento de padrões.

Os principais métodos de redução são: Análise via Componentes Principais (*Principal Component Analysis* - PCA), a versão esparsa da PCA, intitulada SPCA (*Supervised Principal Component Analysis*), e a Análise via Componentes Principais Independentes (*Independent Principal Component Analysis* – IPCA). Todas estas técnicas assumem que a relação entre os componentes e as variáveis originais é linear e se diferenciam no processo de construção dos componentes.

A Análise de Componentes Principais (PCA - Kendall e Hotelling, 1957) é uma das técnicas mais utilizadas, pois permite reduzir o espaço dimensional por meio de variáveis latentes (componentes) não correlacionadas (não existe uma relação linear entre elas), permitindo que as mesmas possam ser avaliadas individualmente. Além disso, os componentes principais são construídos de forma a maximizar a variância dos mesmos, sendo que os primeiros componentes explicam grande parte da variabilidade total presente nos dados. No entanto, Lee e Batzoglou (2003) e Huang e Zheng (2006) apontam que uma das limitações da PCA é descrever componentes menos representativos nos casos em que os dados não apresentam distribuição gaussiana, pois, nestes casos, a correlação igual a zero não implica em independência, que se caracteriza por não existir qualquer relação, linear ou não, entre os componentes.

Alternativamente, foi proposta a Análise de Componentes Independentes (*Independent Component Analysis* - ICA - Jutten e Hérault, 1991), o qual constrói componentes que são independentes entre si. No entanto, este método demanda elevado tempo computacional para dados de alta dimensionalidade (Costa et al., 2020; 2021), o que inviabiliza, em muitos casos, a aplicação da ICA. Além disso, Yao et al. (2012) apontam que os componentes independentes enfrentam problemas de convergência na região ótima, e, assim, podem ser obtidos componentes diferentes ao se reanalisar o mesmo conjunto de dados. Dessa forma, foi proposta a Análise de Componentes Principais Independentes (IPCA -Yao et al., 2012), método que combina ou que tenta contornar simultaneamente as limitações da ICA e da PCA. A IPCA já foi aplicada em áreas como hidrologia (Boluwade et al., 2016), detecção e quantificação de metabolitos polares (Jiang et al., 2014) e compressão, detecção e reconhecimento de faces (Alorf, 2016).

Outra técnica de redução dimensional, mas no contexto de inteligência computacional, é a rede neural Autoencoder. O autoencoder, ao contrário dos métodos estatísticos, não necessita de quaisquer pressuposições e consegue

capturar relações lineares ou não-lineares entre os dados. Para isso, utiliza um número de neurônios artificiais na camada oculta inferior em relação ao número de observações que são apresentadas como entradas, funções de ativação para capturar as relações entre os dados de entrada e algoritmos de aprendizagem não supervisionados para treinamento da rede a fim de reduzir a dimensão dos dados e reconstruí-los de forma a conduzir a uma perda mínima de informação (Fusi et al, 2016).

Alguns exemplos das aplicações dessas metodologias no reconhecimento de padrões são: a aplicação da PCA no reconhecimento de padrões de cultivares de arroz utilizando características morfológicas (Sinha e Mishra, 2013), da aplicação do SPCA na identificação de séries temporais similares para detectar falhas (Fontes e Pereira, 2016) e na aplicação da IPCA na compressão de imagens para reconhecimento de faces em que a IPCA se mostrou mais eficiente do que a PCA (Alorf, 2016). A rede Autoencoder já foi aplicada no processo de fabricação visando identificar os padrões de defeitos (Yu et al, 2019).

2. Métodos de redução de dimensionalidade

A seguir são descritos cada um dos principais métodos de redução de dimensionalidade apresentados na literatura. Para isso, em um banco de dados, considere X (dimensão $I \times J$ sendo I o número de observações e J o número de variáveis).

2.1. Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (*Principal Component Analysis* - PCA), introduzida por Kendall (1957) e Hotelling (1957), faz uso da Decomposição em Valores Singulares (*Singular value Decomposition*- SVD) ou da Decomposição Espectral (*Spectral Decomposition* - SD) no processo de construção dos componentes principais, sendo que as duas abordagens conduzem ao mesmo resultado, conforme apresentado por Costa (2018).

Utilizando a SVD em uma matriz centrada de dados X $(I \times J)$ faz-se a seguinte decomposição:

$$Z = XP, \tag{2}$$

em que as colunas de Z são os componentes principais, X é a matriz de incidência de marcadores e P uma matriz ortogonal que corresponde aos n_{PCR} autovetores da matriz de covariância de X denota por Σ. A matriz Σ tem em sua diagonal principal as variâncias das variáveis de $X_1, X_2, \dots, X_J$, ou seja, $\sigma_1^2, \sigma_2^2, \dots, \sigma_J^2$, respectivamente, e fora da diagonal as covariâncias entre as variáveis.

A porcentagem de variação explicada pelo k-ésimo componente é dada por $\frac{\lambda_k}{\sum_{j=1}^{J} \sigma_j^2}$ em que $\sum_{j=1}^{J} \sigma_j^2 = traço(\Sigma)$, ou seja, a variância total presente nas variáveis X, e λ_k é o k-ésimo autovalor. Já a porcentagem de variação total explicada pelos n_{PCR} primeiros componentes principais é dada por meio de $\frac{\sum_{k=1}^{K} \lambda_k}{\sum_{j=1}^{J} \sigma_j^2}$. Dessa forma, ao indicar uma porcentagem de explicação desejada, é possível reduzir o número de variáveis latentes utilizadas em relação ao número de variáveis explicativas, o que permite reduzir o espaço dimensional a ser analisado. A escolha de n_{PCR} componentes, por exemplo, para explicar entre 70% e 80% da variabilidade total dos dados, como sugere Ferreira (2012). Outro ponto relevante acerca da PCR é que os componentes principais são ortogonais, isto é, $Cor(Z_j, Z_k) = 0$ $(j \neq k)$, o que permite avaliar as informações extraídas pelos componentes de forma isolada. No entanto, a obtenção de componentes ortogonais é mais relevante quando os dados assumem uma distribuição gaussiana, que é o único caso em que a correlação nula implica em independência estatística.

Neste sentido, Yao et al. (2012) descrevem que a PCA pode não ser eficiente nos casos em que os dados correspondem a informações biológicas, por exemplo, em expressão gênica, que vem mostrando apresentar distribuição “super-gaussiana”. O termo super-gaussiana se refere a distribuições de probabilidade leptocúrticas e o termo sub-gaussiana se refere a distribuições de probabilidade platicúrtica. Vale ressaltar que a sub ou super-gaussianidade podem ser avaliadas por meio da curtose, dada por $k = \frac{E((X-\mu)^4)}{\sigma^4} - 3$, sendo X uma variável aleatória com $E(X) = \mu$ e $V(X) = \sigma^2$. As variáveis com distribuições gaussianas possuem $k = 0$, as variáveis com distribuição super-gaussianas possuem $k > 0$ e as variáveis com distribuição sub-gaussianas apresentam $k < 0$ (Chissom, 1970; Yao et al., 2012). Outro exemplo de não-gaussianidade na genética seriam os dados referentes a marcadores moleculares SNPs (*Single nucleotide polymorphisms*) que são por natureza discretos.

Outra desvantagem da PCA é que os componentes principais são combinações lineares das variáveis explicativas, sendo que os pesos associados a estas variáveis, normalmente, são diferentes de zero. Nestes casos, a interpretação e a identificação de quais variáveis estão sendo mais relevantes na construção dos componentes se tornam mais complexas (Zou et al., 2006). E, para solucionar este problema, com o intuito de identificar e manter as variáveis explicativas que mais contribuem para a variabilidade total dos dados, foi proposta a versão esparsa do PCA (SPCA- *Sparse Principal Component Analysis*) que será apresentada a seguir.

2.2. Análise via Componentes Principais Esparsa (SPCA)

A SPCA foi proposta por Shen e Huang (2008) com o objetivo de obter componentes principais que tenham um número reduzido de cargas não nulas atribuídas a cada variável original. Para isso, pode-se utilizar a SVD penalizada (*Penalized Singular value Decomposition* - PSVD) que utiliza a penalização LASSO (*Least Absolute Shrinkage and Selection Operator* - Tibshirani, 1996), e implicitamente, selecionam as variáveis que contribuem para os componentes dando peso diferente de zero a elas. Podemos reescrever a formulação da SVD para que seja possível um entendimento melhor da PSVD. A SVD de uma matriz X visa encontrar as quantidades $\boldsymbol{u_i}$, d_i e $\boldsymbol{v_i}$ tal que:

$$\boldsymbol{X} = \sum_{i=1}^{m} d_i \boldsymbol{u_i} \boldsymbol{v_i}$$

em que o i-ésimo componente principal é definido por $Z_i = u_i d_{ii}$ ou, de forma equivalente, $\boldsymbol{Z_i} = \boldsymbol{X v_i}$. Assim, podemos reescrever a expressão anterior como sendo uma regressão entre a variável dependente $\boldsymbol{Z_i}$ e a variável explicativa X e v_i sendo o coeficiente de regressão. Dessa forma, aplica-se a regressão LASSO nestes termos.

Para impor uma penalização LASSO na PCA, Zou et. al (2006) propõe resolver o seguinte problema de otimização:

$$\hat{\beta}_{lasso} = \min \left\{ \left\| Z_i - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

em que $\hat{v}_i = \frac{\hat{\beta}}{\|\beta\|}$ e λ é uma constante não negativa que pode ser obtida por meio de um procedimento de validação cruzada.

2.3. Análise via Componentes Independentes (ICA)

A Análise via Componentes Independentes (*Independent Component Analysis* - ICA) foi proposta por Jutten e Hérault (1991) e Comon (1994). A ICA decompõe a matriz centrada de dados X como:

$$X = AS, \tag{2}$$

em que S é uma matriz de componentes independentes e A é denominada matriz de misturas, sendo esta, geralmente, desconhecida, não quadrada e não ortonormal.

Diante disso, para determinar os componentes independentes, é necessário encontrar uma matriz K para que as novas variáveis encontradas XK sejam não correlacionadas, ou seja, a covariância de XK seja igual a identidade (variâncias iguais a um, covariâncias iguais a zero e, consequentemente, colunas ortogonais). Posteriormente, deve-se encontrar uma matriz R que faça com que a matriz resultante XKR tenha colunas independentes. Então, define-se a matriz A como sendo uma função do produto de uma matriz K, denominada matriz de ortogonalização, e a matriz R que garante a independência entre os componentes.

Para o procedimento de ortogonalização dos dados é aplicada a decomposição ortogonal na matriz de covariância amostral de X, denotada por Σ, obtendo $\Sigma = P\Lambda^{\frac{-1}{2}}P'$ em que P é composta pelos autovetores em suas colunas e Λ é uma matriz diagonal de autovalores da matriz de covariância amostral de X. A matriz K é então definida como $P_r\Lambda_r^{\frac{-1}{2}}$, sendo r o número de componentes independentes desejado, ou seja, P_r é a matriz com as r primeiras colunas da matriz P e Λ_r é uma matriz com as r primeiras linhas e colunas da matriz Λ (autovalores associados a esses primeiros autovetores). Assim, a matriz de dados ortogonais será obtida por meio de XK.

A independência entre as colunas de S é atingida com base na maximização da não-gaussianidade, uma vez que a matriz de mistura não pode ser estimada para variáveis gaussianas, como pode ser visto em Hyvarinen et al. (2001). Este processo pode ser realizado, principalmente, por meio da medida de curtose, porém esta é muito sensível a *outliers*, ou por meio da negentropia, utilizada no algoritmo *FastICA* proposto Hyvärinen (1998), que é mais robusto quando comparada à curtose.

A negentropia é definida como:

$$J(R) = H\left(R_{gaussiana}\right) - H(R), \quad (3)$$

em que $H(R) = -\int_R f_R(r)\, lnf_R(r)dr$ é a entropia de uma variável aleatória R com função densidade de probabilidade $f_R(.)$ e $H\left(R_{gaussiana}\right)$ é a entropia de uma variável aleatória R com distribuição gaussiana.

Geralmente, o cálculo da entropia é complexo e de difícil/impossível obtenção analítica, sendo necessário fazer algumas aproximações para a expressão anterior, como a seguir:

$$J(R) \propto \left\{E[G(R)] - E\left[G\left(R_{gaussiana}\right)\right]\right\}^2,$$

em que G é uma função não quadrática e a escolha da função G influencia na aproximação da negentropia (Hyvärinen, 1999). As funções G mais empregadas com esse intuito são $G_1(r) = \frac{1}{a} logcosh(ar)\ e\ G_2(r) = -exp\left(\frac{-r^2}{2}\right)$, em que a é uma constante $(1 \leq a \leq 2)$.

Após a convergência do algoritmo *FastICA*, obtém-se uma matriz R, que torna as colunas da matriz XK independentes e, consequentemente as colunas de S, uma vez que os componentes independentes podem ser obtidos via:

$$S = XKR. \quad (4)$$

Para determinar a porcentagem de explicação de cada componente independente, Hyvärinen (1999) relata que é necessário assumir que os componentes independentes tenham variância igual a 1 e média igual a 0. Dessa forma, a porcentagem da variabilidade presente nas variáveis explicativas X que é explicada pelos componentes independentes pode ser mensurada por meio de $\frac{n\sum_{j=1}^{m} a_{jr}^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2}$, em que a_{jk} é o elemento da i-ésima linha e j-ésima coluna da matriz de misturas A ($j = 1,2, \dots, m\ e\ r = 1, \dots, min(n,m) - 1$), x_{ij} é o elemento da i-ésima linha e j-ésima coluna da matriz centrada na média das variáveis explicativas X ($j = 1,2, \dots, m$) e n é o número de observações (Bingham e Hyvärinen, 2000; Helwig e Hong, 2013).

Diferentemente dos componentes principais, cada componente independente explica uma pequena parte da variância total dos dados. Além disso, também não é possível determinar a ordem com que os componentes independentes são extraídos. A ICA contempla uma das deficiências abordadas pela PCA, ou seja, os casos em que os dados não apresentam distribuição gaussiana, visto que, somente em casos em que se é verificada a normalidade dos dados, a correlação igual a zero entre as

variáveis implica em independência das mesmas. No entanto, a ICA apresenta algumas desvantagens, como na convergência do algoritmo *FastICA*, pois conduz a diferentes resultados quando as análises são reavaliadas (Yao et al., 2012), além de demandar um alto esforço computacional (Costa et al., 2020).

2.4. Análise via Componentes Principais Independentes (IPCA)

Diante das vantagens e desvantagens apresentadas pela ICA e PCA, Yao et al. (2012) propuseram a Análise de Componentes Principais Independentes (*Independent Principal Component Analysis*- IPCA), com o objetivo de obter componentes independentes e com a mesma propriedade de reprodutibilidade da PCA, ou seja, a ordenação dos componentes e que os primeiros componentes explicam grande parte da variabilidade presente nos dados.

Para isso, são definidos os componentes principais utilizando a decomposição espectral:

$$Z = XP, \tag{5}$$

em que P é a matriz com dimensão $I \times n_{PCR}$ que corresponde aos n_{PCR} autovetores da matriz de covariância de X.

Para alcançar a independência entre os componentes é necessário aplicar o algoritmo *FastICA* nos vetores de carregamento da matriz P. Mas antes, é necessário que a matriz de covariância de P seja igual a identidade. Observe que P ao conter autovetores, já possui as colunas não correlacionadas e para que a variância seja igual à 1, basta fazer $\check{P} = \sqrt{n-1}\,P'$. Assim, tem-se que

$$S^* = \check{P}R^* \tag{6}$$

em que $\check{R}$ é a matriz que garante a independência dos componentes.

Portanto, os componentes principais independentes W são definidos como:

$$W = XS^* \tag{7}$$

Em seguida, Yao et al. (2012) sugerem ordenar os componentes principais independentes baseando-se no valor da curtose dos vetores de carregamento independentes, visto que o interesse é a extração de componentes mais independentes possíveis, ou seja, com alto valor de curtose.

Também é possível criar a versão esparsa da IPCA, denominada de *soft-Threshouding* (sIPCA) que possibilita a identificação das variáveis que mais

contribuem para determinar os componentes principais. Para tanto, a penalização dos vetores de carregamento é dada por:

$$\hat{s}_{jk} = sign\,(s_{jk})(|s_{jk}| - \gamma)^{+} \quad (8)$$

em que γ é o limiar a ser aplicado em cada carga k do componente s_j $(k = 1 \ldots p, j = 1 \ldots m)$.

2.5. Redes Neurais Artificiais

Nos últimos tempos, as Redes Neurais Artificiais (RNAs) vem ganhando destaque maior nas ciências agrárias, podendo citar os trabalhos de Nascimento et al. (2013), Rosado et al. (2020), Bhering et al (2015) e Oda et al (2022). As RNAs consistem em uma técnica da Inteligência Computacional que simula o cérebro humano e, exige na sua aplicação, o treinamento e o ajuste dos pesos sinápticos e seus limiares (Cruz e Nascimento, 2018). A primeira arquitetura de RNA foi proposta por McCulloch e Pitts (1943) e restringia-se a resolver problemas de classificação do tipo binário. Rosenblatt (1957) a fim de aumentar a eficiência do método, apresentou a topologia de um modelo em que os neurônios são dispostos em uma única camada, conforme a Figura 1.

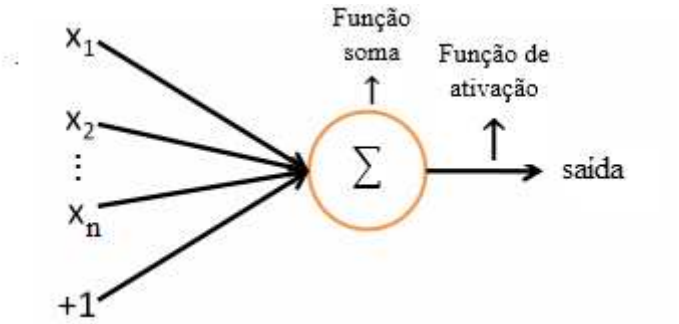

Figura 1: Representação de *perceptron* com uma camada. Fonte: Adaptado de Ng (2011).

Na Figura 1 são apresentados um conjunto de dados de entrada (variáveis de entrada - $x_1, x_2, \ldots x_n$) e saídas (respostas desejadas - y). Os sinais de entrada são multiplicados por pesos correspondentes $(w_0, w_1, w_2, \ldots w_n)$, sendo que $w_0 + \sum_{i=1}^{n} w_i\, x_i$ é definida como a soma ou a porta do limiar. O y_r a ser testado é dado por $y_r = g(\sum_{i=0}^{n} w_i\, x_i)$ em que $x_0 = 1$ e $g()$ e a função de ativação, escolhida a depender do problema utilizado. Os pesos e o limiar são ajustados sucessivamente para que o erro

(ε), diferença entre a resposta desejada (y) e a saída de rede (y_r) ($\varepsilon = y - y_r$), seja o menor possível. Em relação a teoria desenvolvida por Mcculloch e Pitts (1943), Cruz e Nascimento (2018) ressaltam que Rosenblatt (1957) agregou uma regra de aprendizagem, mas, ainda assim, a metodologia resolvia os problemas linearmente separáveis, mas não contemplavam os casos cujo interesse era o ajuste de modelos ou predição de valores cuja variável resposta assumia uma distribuição contínua.

O problema foi resolvido com a implementação da rede Adaline (Widrow e Hoff, 1960), que utiliza o algoritmo de treinamento da Regra Delta de Aprendizagem para ajustar os pesos (Bishop, 2006). Na sua abordagem mais simples, apresentada por Cruz e Nascimento (2018) para o modelo Adaline, o algoritmo consiste em minimizar o erro quadrático dado por:

$$E = \frac{1}{2}\sum_{j=0}^{n}\left(y_j - y_{rj}\right)^2 = \frac{1}{2}\sum_{j=0}^{n}\left(y_j - \vec{w}\vec{x}_j\right)^2,$$

o que implica em $\frac{\partial E}{\partial \vec{w}} = -\sum \varepsilon_j \, x_{ij}$, com $\varepsilon_j = y_i - \vec{w}\vec{x}_i$ para uma observação particular. A Regra Delta de Aprendizagem utiliza os príncipios da teoria do gradiente descente (GD) e determina que o ajuste do peso é dado por:

$$w^{i(\tau+1)} = w^{i(\tau)} + \eta \sum \varepsilon_j \, x_{ij} \ ,$$

tal que o termo de incremento no peso associado a i-ésima entrada é expresso por $\nabla w^{(i)} = \frac{\partial E}{\partial \vec{w}} = -\eta \sum \varepsilon_j \, x_{ij}$, em que η é a taxa de aprendizado e τ é a etapa de interação ou época.

A possibilidade de generalizar a aplicação das redes neurais para resolver problemas que não eram linearmente separáveis veio com a rede intitulada rede Perceptron Multicamadas (*Multilayer Perceptron* -MLP), apresentada na Figura 2, e se deve ao desenvolvimento do algoritmo de treinamento *backpropagation*, apresentado por Rumelhart et al. (1986). A redes MPL inclui camadas ocultas para capturarem melhor a estrutura dos dados, seja ela linear ou não, e transmitir toda informação da camada de entrada para a camada de saída através dos pesos.

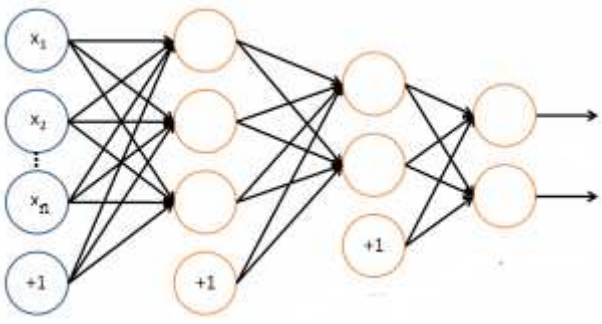

Figura 2: Esquema de um *Perceptron Multilayer*. Fonte: Adaptado de Ng (2011).

O número de camadas ocultas e a quantidade de neurônios depende do conhecimento subjetivo do pesquisador, da experimentação e monitoramento das redes, visto que a utilização de uma grande quantidade de camadas e neurônios pode ocasionar o *overfitting* (a rede memorizar os dados de treinamento), e, uma quantidade insuficiente de neurônios e camadas ocultas pode não ser capaz de capturar a estrutura não linear dos dados (Bishop, 2016). A função de ativação tem a capacidade de capturar a não linearidade das informações apresentadas na camada de entrada (Cruz e Nascimento, 2018), e exige-se que sejam parcialmente ou totalmente diferenciáveis de acordo com o algoritmo *backpropagation* utilizado.

O algoritmo de treinamento *backpropagation* consiste em definir uma função de custo e a utilização de algum método de otimização que permita a atualização dos pesos após cada iteração (Rumelhart et al., 1995). Na fase *forward,* o processo consiste em apresentar as unidades de entrada, definir valores iniciais para os pesos e propagar a informação por meio das camadas ocultas até a saída de rede. Já a fase *backward* atualiza os pesos da saída de rede em direção às entradas utilizando a técnica do gradiente descendente, sendo que no processo das camadas intermediárias ou ocultas o potencial de ativação, dado pela função de ativação aplicada na matriz de pesos multiplicado pelas entradas da camada imediatamente anterior, é utilizado como saída de rede em busca do ajuste de pesos que apontam para o negativo do gradiente, ou seja, em direção ao mínimo da função de custo.

Em relação ao processo de ajuste dos pesos, podem ser utilizadas diferentes técnicas baseadas em gradiente com a finalidade de alcançar a convergência do algoritmo, otimizar o tempo computacional e reduzir o esforço computacional. Segundo Sharma e Venugopalan (2014), as técnicas ou passos utilizados para o treinamento da rede, podem ser agrupados em três tipos (métodos que se enquadram na categoria): Gradiente Descendente (*Gradient Descent backpropagation algorithm, Gradient Descent with momentum, Resilence backpropagation* e *Globally Resilient Backpropagation*), Gradiente Conjugado *(Scaled Conjugate Gradient, Conjugate Gradient backpropagation with Fletcher-Reeves Updates* e *Conjugate Gradient backpropagation with Polak-Riebre Updates*) e Quase-Newton (*Levenberg–Marquardt Backpropagation* e *Bayesian regularization*)

Além disso, os algoritmos de treinamento podem ter várias nomenclaturas para diferenciar as diferentes formas de ajuste dos pesos: por local em que o reajuste dos pesos ocorre após a exposição de cada observação; ou por lote em que o reajuste dos pesos ocorre após a apresentação de todas as observações da entrada e da saída de rede (Cruz e Nascimento, 2018; Silva et al., 2010). No primeiro caso, tem-se, por exemplo, o Gradiente Descendente estocástico (*Stochastic Gradient Descent* - SGD), amplamente utilizado por ter um desempenho computacional mais rápido quando comparado ao GD, que utiliza todas as observações em cada iteração, mas apresenta a desvantagem de ter a possibilidade de não alcançar o mínimo local devido à uma fase estacionária do método (Ruder et al., 2016; Song et al. 2013).

2.5.1. Autoencoder

Dentre os diferentes tipos de RNAs, o Autoencoder destaca-se por ser uma rede neural do tipo *feed-forward* de aprendizado não supervisionado (Lewis, 2016), que consiste em apresentar à rede apenas os vetores de entrada, para agrupar, buscar associações ou sumarizar os dados para facilitar a visualização dos mesmos em dimensões inferiores (Bishop, 2006). Já o aprendizado supervisionado consiste em um conjunto de vetores de entrada e um valor alvo, com o objetivo de classificar ou realizar predição.

A rede neural Autoencoder já foi aplicada em vários contextos como na identificação dos genes mais relevantes que contribuem para o câncer de mama (Dannae et al., 2017), reconstrução de imagens (Wang et al., 2012 ; Yang et al., 2016), como método de redução de dimensionalidade na área da computação (Wang et al., 2016), identificação de manuscritos (Almotiri et al., 2017), na biometria para eliminar os ruídos na extração das características que permitem identificar individualmente o gado (Kumar et al., 2018) e no melhoramento de plantas para identificação de SNPs mais relevantes (Cudic et al, 2018).

O Autoencoder é um tipo de RNA caracterizada por possuir o mesmo número de neurônios na camada de entrada e na camada de saída e, por apresentar como característica ter na camada intermediária um número de neurônios inferior à camada imediatamente anterior. Este último ponto possui um aspecto semelhante aos métodos de redução de dimensionalidade, que consiste em projetar os dados em uma dimensão inferior utilizando variáveis latentes, os componentes, para explicar grande parte da variabilidade total dos dados. Ladjal et al. (2019) destacam que os métodos

de redução de dimensionalidade fazem uma transformação linear e os primeiros componentes já explicam grande parte dos dados e em ordem crescente, enquanto o Autoencoder, através da função de ativação, possibilita uma transformação não linear dos dados.

Conforme Lewis (2016), o Autoencoder consiste em dois processos: o *encoder* que é a transformação que ocorre da camada de entrada para a camada oculta, sendo que neste processo há a redução (ou aumento) da dimensionalidade ao utilizar um número menor (ou maior) de neurônios em relação a quantidade de neurônios da camada de entrada, e a transformação da camada oculta para a camada de saída que é denominada *decoder*. Este processo é descrito na Figura 3, considerando o Autoencoder padrão com apenas uma camada oculta.

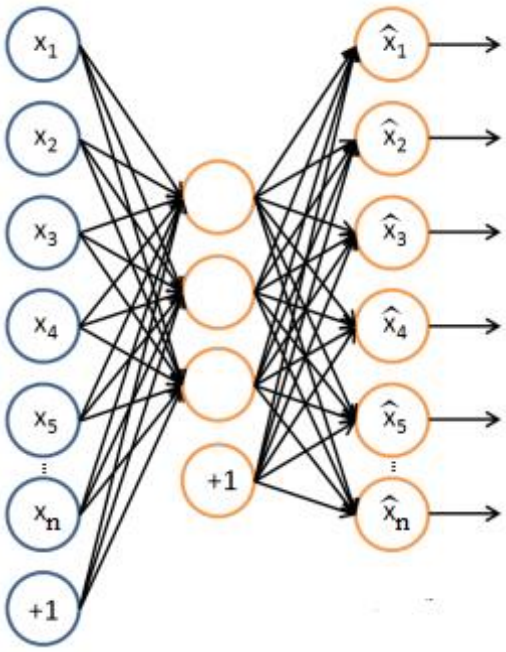

Figura 3: Estrutura do Autoencoder padrão. Fonte: Adaptado de Ng (2011).

A função *encoder, denotada* por f, *e decoder,* uma função linear g, é dada por:

$$h(x) = \lambda\big(f(Wx + b_h)\big) + \mathrm{b}_g,$$

em que x é o vetor de entrada dos atributos, f é a função de ativação, b_h é o vetor de viés da camada oculta, W é a matriz de pesos estimados, b_g é o vetor de viés após o reajuste dos pesos utilizando o algoritmo do gradiente descendente *backpropagation* para treinar a rede buscando minimizar o erro $(g(\hat{x}) - x)^2$, dado pela diferença quadrática entre a saída de rede após os dados serem reconstruídos $(g(\hat{x}))$ e os dados

de entrada (x). E, geralmente, utiliza-se $\lambda = W^T$ para reduzir a quantidade de parâmetros a serem ajustados (Lewis, 2016).

3. Métodos de agrupamento

Por estas metodologias descritas anteriormente é possível reduzir o espaço dimensional e/ou reconhecer padrões do banco de dados estudado, mas não permitem o agrupamento. Assim, para determinar o número de grupos e quais grupos os dados estão estruturados deve-se utilizar técnicas como os métodos de agrupamento. A aplicação desses métodos exige a utilização de uma medida de dissimilaridade, como por exemplo, a distância euclidiana ou a distância generalizada de *Mahalanobis*. Neste capítulo será descrito apenas a distância euclidiana e o método de Agrupamento Hierárquico Aglomerativo. Vale ressaltar que os dados moleculares utilizados no próximo capítulo são codificados em 0, 1 e 2.

3.1. Matriz de distância Euclidiana

Dada uma matriz X com dimensão $n \times m$ em que n o número de observações e m o número de variáveis, tem-se que a matriz de distância euclidiana D é uma matriz de dissimilaridade com dimensão $n \times n$. A matriz D é composta por elementos d_{ij} $(i, j = 1, .. m)$ que se refere a distância entre a i-ésima e j-ésima observação:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

em que $d_{ij} = \sqrt{\sum_{k=1}^{m}(X_{ik} - X_{jk})^2}$, com X_{ik} e X_{jk} correspondendo, respectivamente, ao valor da variável k para os indivíduos i e j, com $i, j = 1, \ldots, n$.

3.2. Análise de Agrupamento Hierárquico Aglomerativo

Os métodos de análise de agrupamento têm por objetivo formar grupos de unidades experimentais homogêneos ou similares dentro do grupo e heterogêneos ou dissimilares entre os grupos, utilizando alguma medida de dissimilaridade (ou similaridade) e um critério de classificação (Cruz et al., 2011). Dentre as várias técnicas de agrupamentos existentes, destaca-se os métodos hierárquicos

aglomerativos descritos detalhadamente por Cruz et al. (2011) e apresentados resumidamente a seguir.

a) Método do vizinho mais próximo ou Método da ligação simples

Dado dois indivíduos, i e j, pertencentes à um mesmo grupo, a distância entre este grupo e outro indivíduo k é dada por: $d_{(ij)k} = min\{d_{ik}, d_{jk}\}$.

A distância entre dois grupos formados pelos indivíduos (i e j) e (k e l), é dada por:

$$d_{(ij)(kl)} = min\{d_{ik}; d_{il}; d_{jk};, d_{jk};, d_{jl}\}$$

A partir disso, o passo inicial do método é determinar os dois indivíduos mais similares através da matriz de distância e, em seguida, calcular a distância entre o grupo formado por esses dois indivíduos e os outros restantes, conforme definido acima. Então, é construída uma nova matriz de dissimilaridade e verificado qual indivíduo é mais similar com o grupo formado no passo anterior. O processo é repetido até que o último grupo seja formado por todos os indivíduos. Tal processo pode ser visualizado graficamente por meio de um dendrograma, que contempla, também, os pontos de junção dos grupos.

b) Método do vizinho mais distante ou Método da ligação completa

Em contraposição ao Método do vizinho mais próximo, tem-se o Método do vizinho mais distante, em que a distância entre um grupo, formado pelos progenitores i e j, e o indivíduo k é dada por:

$$d_{(ij)k} = max\{d_{ik}, d_{jk}\}.$$

Já a distância entre dois grupos, formados por (i e j) e (k e l) é dada por:

$$d_{(ij)(kl)} = max\{d_{ik}; d_{il}; d_{jk};, d_{jk};, d_{jl}\}.$$

De modo similar ao método anterior, mas considerando as distâncias acima, também pode ser visualizado graficamente por meio de um dendograma.

c) Método de ligação média não ponderada (*Unweighted pair-group method using arithmetic averages* - UPGMA)

A distância entre um indivíduo e um grupo ou entre dois grupos é definida, respectivamente, por uma média aritmética não ponderada da seguinte forma:

$$d_{(ij)(k)} = \frac{d_{ik}+d_{jk}}{2}.$$

A distância entre dois grupos formados por indivíduos (i e j) e (k, l e m) é dada por:

$$d_{(ij)(klm)} = \frac{d_{ik} + d_{il} + d_{im} + d_{jk} + d_{jl} + d_{jm}}{6}.$$

Os passos são similares aos métodos do vizinho mais próximo e mais distante, considerando o cálculo de distâncias apresentado anteriormente. De forma análoga, também pode ser visualizado graficamente por meio de um dendrograma.

4. Escolha do número de grupos

Segundo Cruz et al. (2011), a escolha do número de grupos formados pelos métodos de agrupamento é subjetiva, podendo ser feita *a priori* pelo pesquisador, por meio de uma avaliação das ramificações do dendograma ou utilizando algum critério estatístico. Dentre os critérios estatísticos, podem ser citados três critérios que auxiliarão na determinação do número de grupos e na partição (Kassambara, 2017). São eles:

i) *Within-cluster sums of squares* (WCSS): Considere n observações de p variáveis. Então, a soma de quadrados dentro de K grupos é dada por:

$$WCSS_k = \sum_{k=1}^{K} \sum_{i \in A_k} \sum_{j=1}^{p} \left(x_{ij} - \mu_{kj}\right)^2$$

em que K é o número total de grupos, $i = 1,2,..,n$, A_k é o conjunto de objetos no k-ésimo grupo e $\mu_{kj} = \frac{\sum_{i \in A_k} x_{ij}}{n_k}$ sendo n_k o número de elementos do grupo A_k (ou seja, a média da variável j sob o k-ésimo grupo).O melhor agrupamento é o valor K que conduz ao valor mínimo de $WCSS$ (Lisboa et al., 2013).

ii) *Average Silhouette Width (ASW)*: A largura da silhueta da i-ésima observação é dada por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

em que:

$a(i)$ é a distância euclidiana média entre a i-ésima observação e todas as observações i' que pertencem ao grupo A_k, ou seja, $a(i) = \frac{\sum_{i\,e\,i' \in A_k} d_{ii'}}{n_k - 1}$ e $d_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$;

a distância euclidiana média entre a i-ésima observação de A_k e todos os objetos do grupo A_r $(A_r \neq A_k)$ é dada por $d(i, A_r) = \frac{\sum_{i' \in A_r} d_{ii'}}{n_r}$;

$b(i)$ é o mínimo da distância euclidiana média entre a i-ésima observação de A_k e todos os grupos A_r $(k, r = 1,2, \dots, K$ com $k \neq r)$, ou seja, $b(i) = \min_{\substack{r=1,\dots,K \\ k \neq r}} d(i, A_r)$.

A largura da silhueta média de cada grupo *k* considerando todos os seus objetos é dada por $ASW_k = \frac{1}{n_k} \sum_{i=1}^{n_k} s(i)$ com $i \in A_k$ e a largura da silhueta média total é dada por $ASW = \frac{1}{K} ASW_k$. O arranjo do agrupamento e o número de grupos ideal é o valor de K que apresenta maior valor de largura de silhueta média (Rousseeuw, 1987).

iii) Estatística GAP: Considere D_k como sendo a soma das distâncias dos pares entre todas as observações do k-ésimo grupo, $D_k = \sum_{i,i' \in A_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$. Então, a variação total dentro dos grupos é dada por:

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

Para calcular a estatística GAP, é necessário seguir os passos abaixo:

1. Agrupar os dados observados, variando o número de agrupamentos de $K = 1,2 \dots, K_{max}$, e calcule o valor de W_K para cada;
2. Gerar N conjuntos de referência com uma distribuição uniforme aleatória;
3. Agrupar cada um desses conjuntos de referência com um número variável de grupos $K = 1,2 \dots, K_{max}$ e calcule o valor de W_{Kn} para cada $(n = 1,2, \dots, N)$;
4. Calcular a estatística GAP é definida por Tibshirani et al. (2001) e dada por:

$$GAP(K) = \frac{1}{N} \sum_{n=1}^{N} \log(W_{Kn}) - \log(W_K)$$

Por essa metodologia as partições e o número de grupos são determinados pelo valor de K que maximiza a expressão $GAP(K)$ (Tibshirani et al., 2001).

5. Correlação cofenética

A aplicação das técnicas de agrupamento provoca distorções na matriz de distâncias, sendo necessário uma medida capaz de quantificar a relação entre a matriz de distância D obtida antes do agrupamento e a matriz C, conhecida por matriz cofenética, obtida após a construção do dendrograma (Saraçli et al., 2013). Assim, Sokal e Rohlf (1962) propuseram o coeficiente de correlação cofenética como sendo uma das medidas mais utilizadas para avaliar a qualidade de ajuste de um certo agrupamento, e esta é dada por:

$$r = \frac{\sum_{i=1}^{n} c_i d_i - \frac{\sum_{i=1}^{n} c_i \sum_{i=1}^{n} d_i}{n}}{\sqrt{\sum_{i=1}^{n} c_i^2 - \frac{(\sum_{i=1}^{n} c_i)^2}{n}} \sqrt{\sum_{i=1}^{n} d_i^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}}}$$

em que d_{12}, d_{13}, ..., $d_{n-1,n}$ correspondem aos elementos da matriz D, matriz de distância entre os elementos antes da construção do dendrograma ou da determinação do número de grupos; e c_{12} , c_{13} , ..., $c_{n-1,n}$ são elementos da matriz cofenética C, sendo o elemento c_{rs} o valor mínimo da distância, ou nível de fusão, dos grupos compostos pelos indivíduos r e s após a construção do dendrograma ou da determinação do número de grupos, indicando a similaridade conforme a técnica de agrupamento utilizada. Os valores altos de correlação cofenética indicam que o agrupamento promoveu poucas distorções entre as distâncias de pares avaliados (Cruz et al., 2011).

6. Variância entre e dentro de grupos

Segundo Cruz et al. (2011), uma das alternativas para avaliar a variância entre e dentro dos grupos é por meio da análise de variância (ANOVA) que preliminarmente requer a verificação de algumas pressuposições, sendo estas: a relação entre as variáveis dependente e independente deve ser linear, e ambas devem ser ortogonais com os erros, sendo estes normalmente distribuídos, homocedásticos e independentes. Assim, a análise univariada de k grupos compostos por N_i indivíduos cada $(i = 1, \dots, k)$ é feita através do ajuste do modelo de regressão entre a variável dependente (X) e a variável independente (P), o qual é dado por:

$$x_{ij} = \mu + P_i + \epsilon_{ij}, \text{ para } i = 1, \dots, k$$

em que μ é a média geral; P_i é o efeito do i-ésimo grupo; ϵ_{ik} representa o erro experimental.

A Tabela 1 apresenta as fontes de variação do estudo que são a variabilidade total dos dados e o que foi explicado pela regressão e o que é devido ao resíduo.

Tabela 1: Descrição das fontes de variação com seus respectivos graus de liberdade e quadrados médios para a obtenção da variância entre e dentro de grupos.

Fonte	Graus de liberdade	Quadrado Médio
Regressão (entre os grupos)	$k-1$	$\text{Variância entre} = \frac{SQRegressão}{k-1}$
Resíduos (dentro de grupos)	$\sum_{i=1}^{k} N_i - k$	$\text{Variância dentro} = \frac{SQResíduos}{\sum_{i=1}^{k} N_i - k}$

em que $SQRegressão = \sum_{i=1}^{n}(\hat{x}_{ij} - \bar{x}_{ij})^2$ e $SQResíduos = \sum_{i=1}^{I}(x_{ij} - \hat{x}_{ij})^2$ denotam a soma de quadrado de regressão e de resíduos, respectivamente.

7. Referências

Almotiri, J; Elleithy, K; Elleithy, A. Comparison of autoencoder and Principal Component Analysis followed by neural network for e-learning using handwritten recognition. **In: 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). IEEE**, p. 1-5, 2017.

Alorf, A. A. Performance evaluation of the PCA versus improved PCA (IPCA) in image compression, and in face detection and recognition. **In: 2016 Future Technologies Conference (FTC). IEEE**, 2016.

Ammiraju, J. S. S.; Luo, M.; Goicoechea, J. L.; Wang, W.; Kudrna, D.; Mueller, C.; Talag, J.; Kim, H.; Sisneros, N. B.; Blackmon, B.; Fang, E.; Tomkins, J. B.; Brar, D.; Mackill, D.; Maccouch, S.; Kurata, N.; Lambert, G.; Galbraith, D.W.; Arumuganathan, K.; Rao, K.; Walling, J. G.; Gill, N. Y.U.Y.; Sanmiguel, P.; Soderlund, C.; Jackson, S.; Wing, R. A. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, v.16, 2006.

Azevedo, C. F.; Resende, M. D. V. D.; Silva, F. F.; Lopes, P. S.; Guimarães, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, p. 619-626, 2013.

Azevedo, C. F.; Silva, F. F.; de Resende, M. D. V.; et al. Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal of Animal Breeding and Genetics,** v. 131, 2014.
Bai, J.; Wang, P. Econometric analysis of large factor models. **Annual Review of Economics**, v. 8, 2016.

Bhering, L. L.; Cruz, C. D.; Peixoto, L. D. A.; Rosado, A. M.; Laviola, B. G.; Nascimento, M.. Application of neural networks to predict volume in eucalyptus. **Crop Breeding and Applied Biotechnology**, v. 15, 2015.

Bingham, E.; Hyvärinen, A. A fast fixed-point algorithm for independent component analysis of complex valued signals. **Int. J. Neural Syst**, v. 10, 2000.

Bishop, Christopher M. Pattern recognition and machine learning. **Springer**, 2006.

Boluwade, A.; Madramootoo, C. A. Independent principal component analysis for simulation of soil water content and bulk density in a Canadian Watershed. **International Soil and Water Conservation Research**, v. 4, 2016.

Cheng, T.; Zhan, X. (2017). Pattern recognition for predictive, preventive, and personalized medicine in cancer. **EPMA Journal**, v. 8, 2017.

Chissom, B. S. Interpretation of the kurtosis statistic. **The American Statistician**, v. 24, 1970.

Comon, P. Independent component analysis, a new concept?. **Signal Process**, v. 36, 1994.

Costa, J. A. D.; Azevedo, C. F.; Nascimento, M.; Resende, M. D. V. D.; Nascimento, A. C. C. Genomic prediction with the additive-dominant model by dimensionality reduction methods. **Pesquisa Agropecuária Brasileira**, v. 55, 2020.

Crase, S.; Hall, B.; Thennadil, S. N. (2021). Cluster Analysis for IR and NIR Spectroscopy: Current Practices to Future Perspectives. **Cmc-Computers Materials & Continua**, v. 69, 2021.

Cruz, C D.; Ferreira, M. F.; Pessoni, L. A. Biometria Aplicada ao estudo da diversidade genética. Suprema. Visconde do Rio Branco, Minas Gerais. p. 13-17, 2011.

Cruz, C. D.; Nascimento, M. Inteligência computacional aplicada ao melhoramento genético. **Editora UFV**. Viçosa, Minas Gerais. 414 pp. Il, 2018.

Cudic, M.; Baweja, H.; Parhar, T.; Nuske, S. Prediction of Sorghum bicolor Genotype from In-situ Images Using Autoencoder-identified SNPs. **In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE**, p. 23-31, 2018

da Costa, J. A.; Azevedo, C. F.; Nascimento, M.; Silva, F. F.; de Resende, M. D. V.; Nascimento, A. C. C. Determination of optimal number of independent components in yield traits in rice. **Scientia Agricola**, v.79, p. e20200397, 2021

de Los Campos, G.; Hickey, J. M.; Pong-Wong, R.; Daetwyler, H. D.; Calus, M. P. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, p. 327-345, 2013.

Erfani, S. M.; Rajasegarar, S.; Karunasekera, S.; Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. **Pattern Recognition**, v. 58, 2016.
Fontes, C. H.; Pereira, O. Pattern recognition in multivariate time series–A case study applied to fault detection in a gas turbine. **Engineering Applications of Artificial Intelligence**, v. 49, 2016.

Fusi, S.; Miller, E.K.; Rigotti, M. Why neurons mix: high dimensionality for higher cognition. **Current opinion in neurobiology**, v. 37, 2016.

Garris, A. J.; Tai, T. H.; Coburn, J., Kresovich, S., McCouch, S. Genetic structure and diversity in Oryza sativa L. **Genetics**, v. 169, 2005.

Harding, S. M.; Benci, J. L.; Irianto, J.; Discher, D. E.; Minn, A. J.; Greenberg, R. A. Mitotic progression following DNA damage enables pattern recognition within micronuclei. *Nature*, *v.* 548, 2017.

Helwig, N.E.; Hong, S. A critique of tensor probabilistic independent component analysis: implications and recommendations for multi-subject fMRI data analysis. J. **Neurosci. Methods**. v. 213, 2013.

Hotelling, H. The relations of the newer multivariate statistical methods to factor analysis. **British Journal of Mathematical and Statistical Psychology**, v. 10, 1957.

Huang, D.; Zheng, C. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. **Bioinformatics**, v. 22, 2006.

Hyvarinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley & Sons, Hoboken, 2001.

Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit. **Adv. Neural Inf. Process. Syst,** v. 10, 1998.

Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. **IEEE Trans. Neural Netw**, v. 10, 1999.

Liu, Xiufeng; Iftikhar, Nadeem; Xie, Xike. Survey of real-time processing systems for big data. In: **Proceedings of the 18th International Database Engineering & Applications Symposium**, 2014

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning. **Springer**, New York, 2013.

Jiang, M.; Jiao, Y.; Wang, Y.; Xu, L.; Wang, M.; Zhao, B.; Jia, L.; Pan, H.; Zhu, Y.; Gao, X. Quantitative profiling of polar metabolites in herbal medicine injections for multivariate statistical evaluation based on independence principal component analysis. **PloS one**, v. 9, 2014.

Jutten, C.; Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. **Sigmal Processing**, v. 24, 1991.

Kassambara, A.; Mundt, F. Package 'factoextra'. **Extract and visualize the results of multivariate data analyses**, v. 76, 2017.

Kendall, M. G. **A Course in Multivariate Analysis**. London: Griffin, 1957.

Klöppel, S.; Abdulkadir, A.; Jack Jr, C. R.; Koutsouleris, N.; Mourão-Miranda, J.; Vemuri, P. Diagnostic neuroimaging across diseases. **Neuroimage**, v. 61, 2012.

Kumar, S.; Pandey, A.; Satwik, K. S. R.; Kumar, S.; Singh, S. K.; Singh, A. K.; Mohan, A. Deep learning framework for recognition of cattle using muzzle point image pattern. **Measurement**, v. 116, 2018.

Ladjal, S.; Newson, A.; Pham, C. H. A PCA-like autoencoder. **arXiv preprint arXiv:1904.01277**, 2019.

Lee s.; Batzoglou, S. Application of independent component analysis to microarrays. **Genome Biology**, v. 4, 2003.

Lewis, N. D. Deep learning made easy with R. A gentle introduction for data science, 2016.

Lisboa, P. J.; Etchells, T. A.; Jarman, I. H.; Chambers, S. J. Finding reproducible cluster partitions for the k-means algorithm. **BMC bioinformatics**, v. 14, 2013.

Mcculloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v.5, 1943.

Meuwissen, T.H.E.; Hayes, B. J.; Goddard, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v.157, 2001.

Mullainathan, S.; Spiess, J. Machine learning: an applied econometric approach. **Journal of Economic Perspectives**, v. 31, 2017.

Nascimento, M.; Peternelli, L.A..; Cruz, C. D.; Nascimento, A. C. C.; Ferreira, R. D. P.; Bhering, L. L.;Salgado, C. C. . Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breeding and Applied Biotechnology, v. 13,2013.

Nazari, M.; Alidadi, M. (2013). Measuring credit risk of bank customers using artificial neural network. **Journal of Management Research,** v. 5, 2013.

Ng, Andrew. Sparse autoencoder. **CS294A Lecture notes**, v. 72, 2011.

Oda, M. C., Sediyama, T., Cruz, C. D., Nascimento, M., & Matsuo, É. Adaptability and yield stability of soybean genotypes by mean Eberhart and Russell methods, artificial neural networks and centroid. **Agronomy Science and Biotechnology**, v. 8, 2022.

Pandey, P., Anurag, P. J., Tiwari, D. K., Yadav, S. K., Kumar, B. Genetic variability, diversity and association of quantitative traits with grain yield in rice (Oryza sativa L.). **Journal of bioscience**, v. 17, 2009.

Petersen, J. L.; Mickelson, J. R.; Cothran, E. G.; Andersson, L. S.; Axelsson, J.; Bailey, E.; McCue, M. E. Genetic diversity in the modern horse illustrated from genome-wide SNP data. **PloS one**, v. 8, 2013.

Rani, M. P.; Arumugam, G. An efficient gait recognition system for human identification using modified ICA. **International journal of computer science and information technology**, v. 2, 2010.

Rawat, P.; Shankhdhar, D.; Shankhdhar, S. C.. Plant growth promoting potential and biocontrol efficiency of phosphate solubilizing bacteria in rice (Oryza sativa L**.). Int J Curr Microbiol Appl Sci**, v. 9, 2020

Resende, M. D. V.; Silva, F. F; Azevedo, C. F. Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), **Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. Visconde do Rio Branco: Suprema, v. 1, p. 881, 2014.

Rosado, R. D. S.; Cruz, C. D.; Barili, L. D.; de Souza Carneiro, J. E.; Carneiro, P. C. S.; Carneiro, V. Q.; da Silva, J. T.; Nascimento, M. Artificial Neural Networks in the Prediction of Genetic Merit to Flowering Traits in Bean Cultivars. **Agriculture**, v. 10, , 2020.

Rosenblatt, F. **The perceptron, a perceiving and recognizing automaton Project Para**. Cornell Aeronautical Laboratory, 1957.

Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, 1987.

Ruder, S. An overview of gradient descent optimization algorithms. **arXiv preprint arXiv:1609.04747**, 2016.

Rumelhart, D. E.; Durbin, R.; Golden, R.; Chauvin, Y. Backpropagation: The basic theory. **Backpropagation: Theory, architectures and applications**, p. 1-34, 1995.

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. Nature, v. 323, 1986.

Rui, W.; Feng, Y.; Jiang, M.; Liang Wang, X.; Shi, Z. Z.. Pattern recognition of Glycyrrhiza uralensis Metabonomics on rats with MixOmics package of R software. **Procedia Engineering**, v. 24, 2011.

Sharma, B.; Venugopalan, K. Comparison of neural network training functions for hematoma classification in brain CT images. **IOSR Journal of Computer Engineering**, v. 16, 2014.

Saraçli, S.; Doğan, N.; Doğan, İ. Comparison of hierarchical cluster analysis methods by cophenetic correlation. **Journal of inequalities and Applications**, v. 2013, 2013.

Shen, Haipeng; Huang, Jianhua Z. Sparse principal component analysis via regularized low rank matrix approximation. **Journal of multivariate analysis**, v. 99, 2008.

Silva, I. N. DA; Spatti, D. H.; Flauzino, R. A. **Redes neurais Artificiais Para Engenharia e Ciências Aplicadas - Fundamentos Teóricos e Aspectos Práticos**. 2 ed. São Paulo: Artliber Editora Ltda, 2010.

Sinha, A. K.; Mishra, P. K. Morphology based multivariate analysis of phenotypic diversity of landraces of rice (Oryza sativa L.) of Bankura district of West Bengal. **Journal of Crop and Weed**, v. 9, 2013.

Sokal, R. R.; Rohlf, F. James. The comparison of dendrograms by objective methods. **Taxon**, v. 11, 1962.

Song, S.; Chaudhuri, K.; Sarwate, A. D. Stochastic gradient descent with differentially private updates. In: **2013 IEEE Global Conference on Signal and Information Processing**. IEEE, 2013.

Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. **Stat. Soc. Series B Stat. Methodol**, v. 58, 1996.

Tibshirani, Robert; Walther, Guenther; Hastie, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, 2001.

Wang, H.; Misztal, I.; Aguilar, I.; Legarra, A.; Muir, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, v. 94, p. 73-83, 2012.

Wang, Y.; Yao, H.; Zhao, S. Auto-encoder based dimensionality reduction. **Neurocomputing**, v. 184, 2016.

Widrow, B.; Hoff, M. E. Adaptive switching circuits, IRE WESCON Convention Record, v. 4, IRE, New York, pp. 96–104, 1960.

Xie, R.; Wen, J.; Quitadamo, A.; Cheng, J.; Shi, X. A deep auto-encoder model for gene expression prediction. **BMC genomics**, v. 18, p. 39-49, 2017.

Yang, Y.; Wu, Q. M J.; Wang, Y. Autoencoder with invertible functions for dimension reduction and image reconstruction. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 48, 2016.

Yao, F.; Coquery, J.; Lê Cao, K.A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. **BMC bioinformatics**, v. 13, 2012.

Ye, J.; Dobson, S.; Mckeever, S. Situation identification techniques in pervasive computing: A review. **Pervasive and mobile computing**, v. 8, 2012.

Yu, Jianbo; Zheng, Xiaoyun; Wang, Shijin. A deep autoencoder feature learning method for process pattern recognition. **Journal of Process Control**, v. 79, 2019.

Zhao, K..; Wright, M.; Kimball, J., Eizenga, G.; McClung, A.; Kovach, M., Tyagi, W.; Ali, M. L.; Tung, C. W.; Reynolds, A.; Bustamante, C. D.;cCouch, S. R. Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. **PloS one**, v.5, 2010.

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, J. G., Islam, A.R., Reynolds, A., Mezey, J., Mcclung, A. M., Bustamante, C. D., Mcclung, A. M. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature communications**, v. 2, 2011.

Zou, Hui; Hastie, Trevor; Tibshirani, Robert. Sparse principal component analysis**. Journal of computational and graphical statistics**, v. 15, 2006.

CAPÍTULO 2

AUTOENCODER, ANÁLISE VIA COMPONENTES PRINCIPAIS E INDEPENDENTES APLICADOS NO RECONHECIMENTO DE PADRÕES DE POPULAÇÕES

Resumo

O objetivo deste trabalho é comparar metodologias estatísticas de redução de dimensionalidade: Análise via componentes principais (PCA), Análise via componentes principais supervisionada (SPCA) e Análise via componentes principais independentes (IPCA); e de inteligência artificial: Autoencoder; para o reconhecimento de padrões utilizando para isso dados genômicos. Os estudos de reconhecimento de padrões são essenciais nos programas de melhoramento e permitem avaliar a variabilidade entre e dentro das subpopulações. Atualmente, temos disponível um elevado número de marcadores SNPs *(Single Nucleotide Polymorphisms)* por espécie. O uso desses marcadores, em métodos de agrupamento, visando construir subgrupos distintos e homogêneos, ocasiona elevado tempo computacional e um elevado consumo de memória do computador. Desta forma, aplicaremos os métodos de redução de dimensionalidade e a rede Autoencoder antes do método de agrupamento UPGMA (*Unweighted pair-group method using arithmetic averages*) com o intuito de reduzir o esforço computacional sem que haja perda relevante de informações. Os dados utilizados neste trabalho é um banco de dados público de arroz Asiático *Oryza sativa*, um dos mais consumidos no mundo, composto por 413 genótipos, os quais foram genotipados para 44.100 marcadores SNPs**.** Os métodos PCA, SPCA e IPCA identificaram o mesmo número de subpopulações da análise preliminar e apresentaram os maiores valores de correlação cofenética e porcentagem de acerto. O Autoencoder apresentou menor variância dentro e maior variância entre os grupos avaliados. A partir disso, foram propostas algumas versões do Autoencoder: PCA-AUT, SPCA-AUT e IPCA-AUT, que, respectivamente, utilizam os componentes obtidos via PCA, SPCA e IPCA como variáveis de entrada. Dentre todas as metodologias avaliadas, a SPCA-AUT apresentou menor tempo computacional e maior porcentagem de acerto.

Palavras-chave: Inteligência computacional, Redes Neurais, Autoencoder Modificado, métodos de redução dimensional, *Oryza sativa*.

Abstract

The objective of this study is to compare statistical methodologies for dimensionality reduction: Principal Component Analysis (PCA), Supervised Principal Component Analysis (SPCA), and Independent Principal Component Analysis (IPCA); and artificial intelligence: Autoencoder; for pattern recognition using genomic data. Pattern recognition studies are essential in breeding programs and allow the assessment of variability between and within subpopulations. We currently have many SNPs (Single Nucleotide Polymorphisms) markers per species. Using these markers in clustering methods to build distinct and homogeneous subgroups causes high computational time and high consumption of computer memory. In this way, we will apply the dimensionality reduction methods and the Autoencoder network before the UPGMA clustering method (Unweighted pair-group method using arithmetic averages) to reduce the computational effort without significant loss of information. The data used in this work is a public database of Asian rice Oryza sativa, one of the most consumed in the world, composed of 413 genotypes, which were genotyped for 44,100 SNP markers. The PCA, SPCA, and IPCA identified the same number of subpopulations as the preliminary analysis and presented the highest values of cophenetic correlation and percentage of correct answers. The Autoencoder showed the lowest variance within and the highest variance between the evaluated groups. Based on this, some versions of the Autoencoder were proposed: PCA-AUT, SPCA-AUT, and IPCA-AUT, which, respectively, use the components obtained via PCA, SPCA, and IPCA as input variables. The SPCA-AUT presented the lowest computational time and the highest percentage of correct answers among all the methodologies evaluated.

Keywords: Computational Intelligence, Neural Networks, Modified Autoencoder, dimensional reduction methods, *Oryza sativa.*

1 Introdução

O arroz (*Oryza sativa* L.) é consumido por mais da metade da população mundial, mas, nos últimos tempos, houve uma queda na razão entre produção e tamanho populacional. Diante do aumento da população mundial, torna-se de extrema urgência a identificação de cultivares altamente produtivos (Rawat et al., 2020; Ashfaq et al., 2012; Muralidhara et al., 2019; (Kubo e Purevdorj, 2004). As duas principais variedades do arroz *Oryza sativa* são Indica, cultivado em regiões tropicais, responsável por mais de 50% da produção mundial e de alto valor nutritivo (Datta et al. 2006); e o Japônica, geralmente, cultivado em regiões temperadas, tem ganhado destaque por ser responsável por 20% da produção a nível mundial e se desenvolver em áreas que ainda possibilitam a expansão da produção de arroz (Lee et al., 2003; Cordero-Lara e Karla, 2020). Dentro dessas variedades, a utilização de marcadores moleculares permitiu identificar no grupo Indic*a*, duas subespécies, Indica e o AUS, e dentre as variedades do grupo Japônica tem-se o Temperate japônica, Tropical Japônica e Aromatic (Garris et al., 2005; Zhao et al., 2010; Huang et al., 2012).

Nesse sentido é de suma importância o desenvolvimento de pesquisas que visam avaliar a variabilidade genética entre os genótipos do arroz (Thomson et al., 2007; Rabbani et al., 2008; Vanniarajan et al., 2012; Akinwale et al., 2011; Ranjith et al., 2018). O estudo da variabilidade genética é essencial nos programas de melhoramento com o intuito de reconhecer padrões de similaridade entre os indivíduos ou orientar as pesquisas relacionadas à diversidade genética (Silva Júnior et al., 2020), pois permite avaliar a perda da variabilidade entre e dentro das espécies, e, consequentemente, nortear as buscas por estratégias que buscam conservar ou potencializar o material genético das espécies (Cruz et al., 2011). Dentre as várias metodologias utilizadas para avaliar a diversidade genética entre os progenitores, destacam-se os métodos preditivos em que são utilizadas medidas de dissimilaridade para detectar diferenças morfológicas, fisiológicas ou moleculares (Cruz et al., 2014, 2012). Vários estudos, como o de Xia et al. (2019) e Semagn et al. (2012), utilizaram marcadores moleculares para o estudo da diversidade genética. Uma vez que eles extraem informações diretamente do DNA, espera-se um aumento da eficiência na diferenciação dos genótipos.

No entanto, análises que utilizam marcadores moleculares trazem milhares de informações, o que pode dificultar o reconhecimento de padrões dos dados e

demandar um grande esforço computacional. Além disso, segundo Silva Júnior et al. (2020), abordagens baseadas em métodos de reconhecimento de padrões podem ser utilizadas para avaliar o comportamento dos genótipos, mas até o momento são pouco exploradas. Estas metodologias podem ser baseadas, por exemplo, em métodos estatísticos de redução de dimensionalidade como a Análise de Componentes Principais (PCA), Análise Componentes Principais Supervisionado (SPCA), a Análise de Componentes Principais Independentes (IPCA) e baseadas em inteligência artificial como a rede o Autoencoder, descrito por Lewis (2016) como uma rede neural do tipo *feed-forward* de aprendizado não supervisionado.

Além disso, vários autores desenvolveram trabalhos utilizando componentes principais como variáveis de entrada na análise de discriminante (Lee et al., 2016). Assim, diante da capacidade do Autoencoder em discriminar os grupos que são formados, foi proposto e avaliado a junção das metodologias de redução de dimensionalidade e a rede Autoencoder criando as metodologias PCA-AUT, SPCA-AUT e IPCA-AUT. Nestes procedimentos, os componentes construídos pelos métodos de redução de dimensionalidade foram utilizados para extrair as informações do conjunto de dados avaliado e, em seguida, utilizados como variáveis de entrada da rede Autoencoder.

Diante disso, o presente trabalho propõe a aplicação de metodologias de redução: PCA, SPCA, IPCA, metodologia que ainda não foi aplicada nos estudos relacionados à diversidade genética, o Autoencoder e as versões propostas: PCA-AUT, SPCA-AUT e IPCA-AUT, em dados genômicos do arroz asiático, Oryza *sativa*, visando o reconhecimento de padrões de subpopulações. Posteriormente, foram aplicados conjuntamente os métodos UPGMA (*Unweighted pair-group method using arithmetic averages*) e ASW *(Average Silhouette Width)* para agrupar e determinar o número de grupos verificados nos dados, respectivamente, uma vez que, os métodos de redução buscam associações entre os genótipos avaliados, mas não são se enquadram no contexto de técnicas de agrupamento e não são capazes de quantificar o número de grupos formados.

2 Materiais e Métodos

2.1 Dados reais

O banco de dados utilizado nesse estudo faz parte do Projeto OryzaSNP e do Projeto OMAP (Ammiraju et al., 2006; Zhao et al., 2011). Os dados correspondem a 413 genótipos de arroz Asiático, *Oryza sativa*, oriundos de 82 países, genotipados para 44.000 marcadores SNPs. O SNPs que apresentaram MAF (Minor Allele Frequency – Frequência do Menor Alelo) menor do que 5% e taxa de atendimento menor que 70% (call-rate) foram eliminados das análises obtendo-se ao final do processo 36.901 marcadores. Assim, a matriz de incidência dos marcadores moleculares, denotada por X, apresenta dimensão 413 (número de genótipos) × (36.901 marcadores), sendo codificada por 0, 1, ou 2; em que X_{ij} corresponde ao número de alelos do j-ésimo marcador para o i-ésimo genótipo.

2.2 Definição de grupos iniciais

Geralmente, a *priori*, o número de grupos é desconhecido, ou os dados utilizados podem não possibilitar a captura das subpopulações presentes nos dados. Dessa forma, para avaliar os métodos de redução e o Autoencoder de modo adequado, foi realizada uma análise preliminar utilizando o método de agrupamento UPGMA na matriz de incidência dos marcadores e utilizando o método ASW (*Average Silhouette Width* - Rousseeuw, 1987) para determinar o número de grupos. Para uma observação i pertencente a um grupo A_k ($k = 1, \dots, 10$ – índice dos grupos), define-se:

1. A distância euclidiana média entre a i-ésima observação e todos as observações que pertencem ao grupo A_k, é dada por *a(i)*= $\frac{\sum_{i\, e\, i' \in A_k} d_{ii'}}{C_{n_k,2}}$, em que $C_{n_k,2} = \frac{n_k!}{(n_k-2)!2!}$, $n_k = |A_k|$ (número de observações no grupo A_k) e $d_{ii'} = \sum_{j=1}^{p}\left(x_{ij} - x_{ii'}\right)^2$ a distância euclidiana entre as observações i e i' com x_{ij} (i= $1, \dots, n$ e $j = 1, \dots, m$) .
2. O mínimo da distância euclidiana média entre a i-ésima observação e todas as observações do grupo A_r sendo $A_r \neq A_k$,é dada por: b*(i)*= $\min_{A_r \neq A_k} \frac{\sum_{i\, e\, i' \in A_r} d_{ii'}}{C_{n_r,2}}$ com $C_{n_r,2} = \frac{n_r!}{(n_r-2)!2!}$ e , $n_r = |A_r|$ (número de observações no grupo A_r).

Assim, calcula-se a silhueta da i-ésima observação dada por: $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, e, consequentemente, a largura da silhueta média considerando todas

as observações: $ASW = \frac{1}{N}\sum_{i=1}^{N} S(i)$. O arranjo do agrupamento e o número de grupos ideal é o valor de k que maximiza a expressão ASW. A partir da determinação deste número, as variâncias entre e dentro de grupos foram calculadas utilizando o procedimento de ANOVA (*Analysis of variance*).

2.3 Métodos estatísticos para reconhecimento de padrões

Foram aplicados os métodos de redução de dimensionalidade, PCA, SPCA e]IPCA, assim como a rede neural Autoencoder, na matriz de incidência dos marcadores. Essas metodologias, descritas a seguir, buscam construir de componentes (ou neurônios), que são combinações lineares (métodos de redução dimensional) ou não lineares (Autoencoder) das variáveis, visando a redução de dimensionalidade. Espera-se que o número de componentes (ou neurônios) seja inferior ao número de marcadores utilizados no processo de agrupamento. Antes da aplicação dos métodos, as variáveis foram padronizadas. Além disso, posteriormente, essas combinações lineares também serão utilizadas como variáveis de entrada na rede Autoencoder.

Na PCA, proposta por Kendall (1957) e Hotelling (1957), a decomposição em valores singulares é aplicada na matriz de incidência dos marcadores $X\ (n, m)$, sendo $X = UDV'$, em que $D\ (min\ (n, m) \times min\ (n, m))$ é uma matriz diagonal com elementos da forma $d_j = \sqrt{\lambda_j}$, sendo λ_j o j-ésimo autovalor não-nulo de $X'X$ ou de XX', V $(m \times min\ (n, m))$ é uma matriz com os autovetores de $X'X$ e U é uma matriz $\big(n \times min(n, m)\big)$ com os autovetores de XX' associados a λ. O j-ésimo componente principal é definido como $Z_j = u_j d_j$, u_j é o j-ésimo vetor de carregamento e a porcentagem de variação total explicada pelo j-ésimo componente é dada por $\frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j}$. Os primeiros componentes explicam grande parte da variação total, o que permite reduzir o espaço dimensional a ser analisado, além dos componentes principais serem ortogonais.

A versão esparsa do PCA (sPCA) foi proposta por Shen e Huang (2008) com o objetivo de obter componentes que tenham um número reduzido de cargas não-nulas visando facilitar a interpretação. A esparsidade faz com que alguns elementos da matriz V sejam iguais a zero, selecionando variáveis que tenham maior contribuição para a porcentagem de explicação dos componentes por meio da penalização do

LASSO (*least absolute shrinkage and selection operator* - Tibshirani, 1996). No entanto, a PCA e a sPCA são métodos de redução mais eficientes nos casos em que os dados apresentam distribuição gaussiana, visto que é o único caso em que a correlação igual a zero implica em independência, o que contribui para que os componentes sejam avaliados de forma independente. Para resolver esse problema, foi proposto a ICA- *Independent Component Analysis* (Jutten e Hérault, 1991; Comon, 1994), que utiliza a decomposição ortogonal para obter componentes ortogonais e o algoritmo *FastICA*, desenvolvido por Hyvärinen (1998), para alcançar a independência dos mesmos.

A IPCA, proposta por Yao et al. (2012), contempla as vantagens da PCA, os primeiros componentes explicarem grande parte da porcentagem de explicação da variabilidade total dos dados, e da ICA que apresenta a vantagem de obter componentes independentes, no entanto, demanda alto esforço computacional em suas análises (Costa, 2020). O passo inicial da IPCA consiste em extrair a matriz V, que contém os vetores de carregamento dos componentes principais (obtidos na PCA), e aplicar os procedimentos da ICA em V. Primeiramente, a matriz V é decomposta em $V = P\Lambda^{\frac{-1}{2}}P'$, em que P contém os autovetores da matriz de covariância e Λ é uma matriz diagonal com seus elementos sendo os autovalores. A partir disso, determina-se o número de componentes, construindo $K = P_r\Lambda_r^{\frac{-1}{2}}$, em que r são os primeiros autovetores e autovalores contidos em P e Λ, respectivamente. A maximização de independência entre as colunas dos componentes principais independentes, denotados por W, consiste em obter uma matriz R que torna as colunas VK independentes. A matriz R é obtida por meio do algoritmo *FastICA*. Assim, os componentes independentes principais são dados por $W = XS$, sendo $S = VKR$.

O Autoencoder (AUT), segundo Lewis (2016), é um tipo de rede neural que utiliza dois processos *encoder* e *decoder*. O processo *encoder* reduz a dimensionalidade do problema usando a função dada por: $h(x) = (Wx + b_h)$, em que x é o vetor de entrada dos marcadores, b_h é o vetor de viés da camada oculta e W é a matriz de pesos; já o processo *decoder* reconstrói os dados originais e utiliza a função dada por: $g(\hat{x}) = \lambda f(Wx + b_h) + \mathrm{b}_g$, em que f é a função de ativação (geralmente se utiliza a sigmoide), b_g é o vetor de viés após o reajuste dos pesos e geralmente, $\lambda = W^T$ para reduzir a quantidade de parâmetros a serem ajustados. O

processo de reajuste dos pesos é feito utilizando o algoritmo do gradiente descendente *backpropagation* para minimizar o erro.

O PCA-Autoencoder (PCA-AUT), SPCA-Autoencoder (SPCA-AUT) e IPCA-Autoencoder (IPCA-AUT) propõe que as variáveis latentes (componentes), obtidos nas análises da PCA, SPCA e IPCA, respectivamente, sejam utilizados como variáveis de entrada no tradicional método Autoencoder.

Após a aplicação dos métodos de redução de dimensionalidade, do Autoencoder e da combinação entre eles, foi aplicado o método de agrupamento UPGMA na matriz que contém os vetores de componentes da PCA, SPCA e IPCA, e na camada intermediária, no caso do Autoencoder. Além disso, mais uma vez, foi utilizado o método ASW para determinar o número de grupos.

2.4 Medidas para comparação dos métodos de reconhecimento de padrões

As abordagens foram avaliadas sob um processo de validação direcionada considerando 5-*fold*, ou seja, em cada *fold* foi considerado 20% dos genótipos de cada uma das subpopulações já conhecidas para comporem a população de validação e 80% para a população de treinamento, visto que o número disponível de genótipos de cada uma das subpopulações foi diferente. As porcentagens de acerto, nas populações de treinamento e validação, foram utilizadas para avaliar a capacidade dos métodos em agrupar corretamente os elementos em cada um dos grupos, considerando a formação dos grupos na análise preliminar, isto é, antes da aplicação dos métodos de redução de dimensionalidade. Após a determinação do número de componentes associado a maior porcentagem de acerto na população de validação, foram construídas matrizes de confusão, calculando as variâncias entre e dentro de grupos e calculando a correlação cofenética entre D0 e D1, matrizes de distância euclidiana da análise preliminar e após a redução de dimensionalidade, respectivamente (Sokal e Rohlf, 1962).

2.5 Recursos Computacionais

As análises foram executadas no *software* R (R *Development Core Team*, 2020) disponível em (http://cran.r-project.org). Os pacotes e as funções do *software* R correspondentes na aplicação das metodologias estudadas são apresentados na Tabela 1.

Tabela 1. Métodos avaliados e respectivas ferramentas de implementação no *software* R.

Metodologia	Pacote	Função	Referência
PCA	*pls*	*Pcr*	(Mevik et al., 2016)
IPCA	*mixOmics*	*Ipca*	(Le Cao et al., 2018)
sPCA	*mixOmics*	*Ipca*	(Le Cao et al., 2018)
Autoencoder	*keras*	*Keras*	(Allaire e Cholle, 2018)
ASW	*factoextra*	*fviz_nbclust*	(Kassambara e Mundt, 2017)

3 Resultados e discussão

Análise preliminar (D0)

Na etapa inicial das análises, sem aplicar os métodos de redução de dimensionalidade e a rede Autoencoder, e utilizando a matriz de incidência dos marcadores, foi utilizado o método UPGMA para agrupar os genótipos e o método ASW para determinar o número de grupos que seria utilizado como base na etapa posterior. Essas análises preliminares identificaram três grupos, sendo que o grupo 1 aloca um maior número de genótipos representando cerca de 62.71% (259 genótipos) do total (413); o grupo 2 aloca cerca de 23.24% (96 genótipos) e o grupo 3 aloca cerca de 14.04% (58 genótipos).

Comparação entre os métodos de reconhecimento de padrões e a Análise preliminar

Os genótipos foram subdivididos em população de validação e população de treinamento usando o processo 5-*fold*. Em seguida, foram aplicados os métodos de redução dimensionalidade (PCA, SPCA e IPCA) e o Autoencoder, além do UPGMA em seus respectivos componentes e valores das camadas intermediárias. Nas Figuras 1A) e 1B) são expostos o número de componentes em função da porcentagem de acerto para a população de treinamento e validação referente a cada metodologia. A maior porcentagem média de acerto na validação apresentada por cada método foi

de 80.15% para a PCA, SPCA e IPCA e 68.65% para o Autoencoder. Assim, o menor número de componentes/neurônios associados a esta porcentagem de acerto são: 2 (PCA), 2 (SPCA), 2 (IPCA) e 4 (Autoencoder). Em relação à porcentagem de explicação da variabilidade total presente nos dados genômicos considerando esse número de componentes foram, respectivamente, de aproximadamente 72.29% para os métodos PCA, SPCA e IPCA. Destaca-se que os métodos atingiram com vários números de componentes a porcentagem de acerto máxima. No entanto, utilizamos aquele o menor número de componentes, visto que o aumento do número de componentes contribui também para o aumento do esforço e tempo computacional.

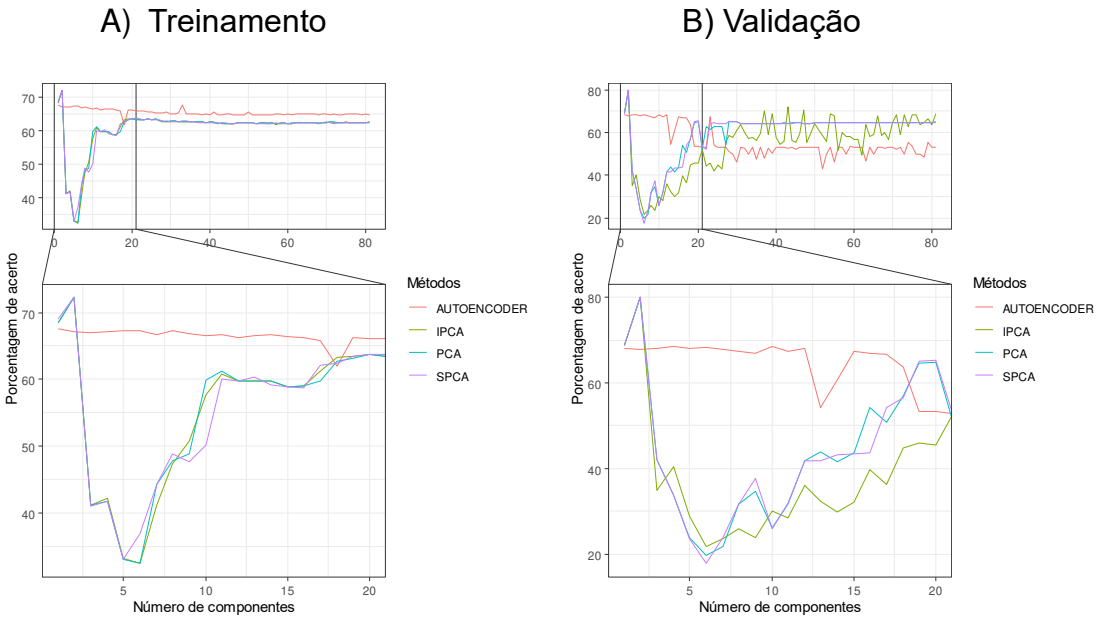

Figura 1: Porcentagem de acerto versus número de componentes considerando o agrupamento pelo método UPGMA (*unweighted pair-group method using arithmetic averages*). A) População de treinamento; B) População de Validação. PCA: Análise via componentes principais, SPCA: Análise via componentes principais supervisionada, IPCA: Análise via componentes principais independentes.

Na Tabela 2 é apresentada a tabela confusão correspondente a cada método utilizando o número de componentes determinado na etapa anterior e todo o conjunto de dados genômicos. É possível observar que os métodos de redução de dimensionalidade, PCA, SPCA e IPCA, foram idênticos na determinação do número de grupos. Diante dos resultados, é possível verificar que a esparsidade na matriz de

carregamento para estes dados não foi capaz de conduzir a uma maior porcentagem de acerto quando comparado a PCA. Este fato pode sugerir que todas as variáveis genômicas contribuem para a distinção destes genótipos, ou seja, os carregamentos devem ser diferentes de zero. Além disso, a independência entre os componentes também não aumentou a porcentagem de acerto em relação a PCA e SPCA, o que sugere que não há uma relação não linear entre as variáveis explicativas (Azevedo et al., 2013).

O Autoencoder por sua vez conduziu a formação de apenas dois grupos (Tabela 2), apresentando assim uma porcentagem de acerto muito inferior (68.65%) a apresentada pelos métodos de redução de dimensionalidade (80.15%). Como consequência deste agrupamento, a alocação dos genótipos nos grupos ficou confusa e sem um padrão de discriminação. Observa-se ainda na Figura 1, em que são apresentadas as porcentagens de acerto no treinamento e na validação em função do número de componentes, ainda no caso do Autoencoder, o agrupamento é ineficiente no treinamento e na validação, o que segundo James et al. (2013) e Silva et al. (2016), caracteriza a ocorrência do *underfitting.* Considerando quatro neurônios a porcentagem de acerto no treinamento foi de 67.09% e na validação de 68.65%. As demais metodologias apresentaram desempenho superior no treinamento e na validação ao considerar dois componentes para a PCA, SPCA e IPCA.

Tabela 2: Matriz de confusão considerando os métodos de redução de dimensionalidade: PCA, SPCA e IPCA; e o Autoencoder, número de grupos determinado pelo pacote ASW- *Average Silhouette Width* e as respectivas porcentagens de acerto.

Métodos	Ngrupos	Grupos de genótipos (real)	Grupos preditos			PAcerto
			G1	G2	G3	
PCA	3	G1	208	50	1	80,15
		G2	2	77	17	
		G3	12	0	46	
SPCA	3	G1	208	50	1	80,15
		G2	2	77	17	
		G3	12	0	46	
IPCA	3	G1	208	51	0	80,15
		G2	2	77	17	
		G3	12	0	46	
AUT	2	G1	206	52	-	68,65
		G2	19	77	-	
		G3	12	46	-	

PCA: Análise via componentes principais, SPCA: Análise via componentes principais supervisionada, IPCA: Análise via componentes principais independentes; AUT: Autoencoder; PAcerto: Porcentagem de acerto; Ngrupos: Número de grupos.

Em relação à variância média entre (83.98) e dentro (0.59) de grupos apresentada pela análise preliminar (D0), é de interesse que as variâncias obtidas pelos métodos de redução e o Autoencoder sejam próximos dele (Tabela 3). A comparação pode ser feita com ressalva, visto que não temos evidências de que as variâncias após a aplicação dos métodos de redução de dimensionalidade são comparáveis. Neste sentido, o Autoencoder, mesmo identificando apenas dois grupos, foi o método que apresentou menor variância dentre e maior variância entre os grupos,

seguido pela SPCA e IPCA, que apresentaram resultados próximos, e pela PCA. Os métodos SPCA e IPCA, que identificaram o mesmo número de grupos da população D0, apresentaram menor variância dentro dos grupos em relação a análise preliminar. Conforme reportado por Raj et al. (2015), o redimensionamento dos dados contribui para a redução da variabilidade dentro dos grupos avaliados e, consequentemente, os métodos de agrupamento aplicados posteriormente têm a discriminação dos grupos mais facilitada.

Tabela 3: Metodologias estatística de reconhecimento de padrão e seus respectivos: Ngrupos (número de grupos determinados pelo ASW - *Average Silhouette Width*, variância dentro os grupos (desvio padrão), variância entre os grupos (desvio padrão), correlação cofenética entre a matriz de distância original e matriz após a aplicação dos métodos de reconhecimento e p-valor associado ao Teste de Mantel para cada um dos valores de correlação obtidos.

Métodos	Número de grupos	Variância dentro	Variância entre	Cor.	p-valor
D0	3	0.59 (0.27)	83.98 (54.58)	-	-
PCA	3	0.07 (0.02)	38.11 (1.08)	0.97	1e-04
SPCA	3	0.07 (0.02)	38.10 (1.08)	0.97	1e-04
IPCA	3	0.08 (0.02)	37.75 (0.98)	0.96	1e-04
AUT	2	0.01 (0.00)	72.43 (16.70)	0.81	1e-04

PCA: Análise via componentes principais, SPCA: Análise via componentes principais supervisionada, IPCA: Análise via componentes principais independentes.

Além disso, os métodos PCA, SPCA e IPCA apresentaram os maiores valores de correlação entre a matriz de distância original e matriz cofenética formada a partir do agrupamento nos vetores de componentes, indicando que essas metodologias foram mais eficientes em recuperar a matriz de distância original. Utilizando o Teste de Mantel, todos os métodos apresentaram p-valor=1e-04<1% de probabilidade, ou seja, baseado em 9999 reamostragens, os valores de correlação cofenética foram significativos.

Diante da capacidade do Autoencoder em discriminar os grupos que são formados, ou seja, maior variância entre os grupos e menor variância entre, além da PCA, SPCA e IPCA capturar três grupos das subpopulações, foi proposto e avaliado a junção destes métodos criando as metodologias PCA-AUT, SPCA-AUT e IPCA-AUT. Nestes procedimentos, os componentes construídos pelos métodos de redução de dimensionalidade foram utilizados para extrair as informações do conjunto de dados avaliado e, em seguida, utilizados como variáveis de entrada da rede Autoencoder e posteriormente, o método UPGMA visando discriminar os grupos.

De forma semelhante, utilizou-se a validação para determinar o número de neurônios que seriam utilizados na camada oculta e o número de componentes que seriam utilizados como variáveis de entrada. Foi necessário apenas um neurônio na camada oculta, reduzindo ainda mais a dimensão do estudo quando comparado à análise anterior, em que foram necessários dois componentes para atingir uma porcentagem de acerto de 80.15%; e como variáveis de entrada foram necessários 74 componentes na PCA-AUT, 50 componentes na SPCA-AUT e 1 componente na IPCA-AUT.

Observa-se que os métodos PCA-AUT, SPCA-AUT e IPCA-AUT apresentaram maiores porcentagens de acerto (Tabela 4) quando comparados ao Autoencoder padrão (Tabela 2). Além disso, verificou-se que os métodos PCA-AUT e SPCA-AUT foram ligeiramente mais eficientes que os métodos PCA, SPCA e IPCA para agrupar os genótipos de arroz. A redução da dimensionalidade do problema beneficiou a estrutura da rede do Autoencoder bem como o reconhecimento de padrões.

Tabela 4: Matriz de confusão considerando os métodos de redução de dimensionalidade e o Autoencoder, número de grupos determinado pelo pacote ASW-*Average Silhouette Width* e as respectivas porcentagens de acerto.

	Ngrupos	Grupos de genótipos (real)	G1	G2	PAcerto
AUT-PCA	2	G1	244	15	82,08
		G2	1	95	
		G3	5	53	
AUT-SPCA	2	G1	245	14	82,57
		G2	0	96	
		G3	3	55	
AUT-IPCA	2	G1	207	52	68,76
		G2	19	77	
		G3	12	46	

PCA: Análise via componentes principais, SPCA: Análise via componentes principais supervisionada, IPCA: Análise via componentes principais independentes; AUT: Autoencoder; PAcerto: Porcentagem de acerto.

Observou-se ainda que a Autoencoder não foi eficiente no caso em que as variáveis de entrada foram independentes, ou seja, obtidas via análise de componentes principais independentes (IPCA-AUT). Este cenário justifica-se pela capacidade do Autoencoder capturar alguma estrutura, linear ou não, nos dados, e na ausência de qualquer estrutura, é extremamente difícil impor a redução de dimensionalidade e recuperar parte da informação presente nos dados através da reconstrução. Além disso, na Tabela 5, mostra-se que os métodos PCA-AUT e SPCA-AUT conseguiram capturar parte da variabilidade apresentada na análise (D0).

Tabela 5: Metodologias estatística de reconhecimento de padrão e seus respectivos: Ngrupos (número de grupos determinados pelo ASW- *Average Silhouette Width*, variância dentro os grupos (desvio padrão), variância entre os grupos (desvio padrão), correlação cofenética entre a matriz de distância original e matriz após a aplicação dos métodos de reconhecimento e p-valor associado ao Teste de Mantel para cada um dos valores de correlação obtidos.

Métodos	Número de grupos	Variância dentro	Variância entre	Cor.	p-valor
D0	3	0.59 (0.27)	83.98 (54.58)	-	-
AUT- PCA	2	2,28e-06 (3,22)	81,60 (1,34)	1	1e-04
AUT- SPCA	2	1,46e-03 (3,27)	81,60 (1,34)	1	1e-04
AUT- IPCA	2	0.08 (0.02)	75,46 (2,39)	0.95	1e-04

PCA: Análise via componentes principais, SPCA: Análise via componentes principais supervisionada, IPCA: Análise via componentes principais independentes.

Ademais, na Tabela 6, é possível verificar que a análise inicial D0 conseguiu alocar em três grupos as seis subpopulações: Indica (87 genótipos), AUS (57 genótipos), Temperate Japônica -TEJ (96 genótipos), Tropical Japônica - TRJ (97 genótipos), Aromatic (14 genótipos) e Admixed - ADMIX (62 genótipos); que foram identificadas no Projeto OryzaSNP e do Projeto OMAP (Ammiraju et al., 2006; Zhao et al., 2011).

No grupo 1 estão todos os genótipos das subpopulações TEJ (96), Aromatic (14) e TRJ (97) e 83.87% (52) dos genótipos da subpopulação Admix, no grupo 2 estão todos os genótipos da subpopulação IND (87) e no grupo 3 e estão todos os genótipos da subpopulação AUS (57) e apenas um genótipo da subpopulação Admix.

Tabela 6: Descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os três grupos que foram formados pelo método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipos.

Subpopulações	Grupo 1	Grupo 2	Grupo 3
TEJ (96)	96	0	0
IND (87)	0	87	0
AUS (57)	0	0	57
AROMATIC (14)	14	0	0
TRJ (97)	97	0	0
ADMIX (62)	52	9	1
Total	259	96	58

TEJ: *Temperate Japonica*, IND: Indica e TRJ: *Tropical Japonica*.

Conforme o estudo de Zhao et al. (2011), que analisou a matriz de parentesco das subpopulações, exceto da AROMATIC, se verificou que o compartilhamento de alelos é maior entre as subpopulações TEJ e TRJ. Este fato explica o resultado encontrado neste estudo, o qual os dados não permitiram que o UPGMA não foi capaz de separar estas duas subpopulações e as alocou no mesmo grupo 1. O menor compartilhamento encontrado por estes autores foi entre as subpopulações Japônica (TEJ e TRJ) e Indica, e neste estudo estas subpopulações foram agrupadas em grupos distintos, grupos 1 e 2, respectivamente. Estes autores também encontraram que a subpopulação AUS tem um moderado compartilhamento com a Indica, no entanto, neste estudo o UPGMA alocou AUS e Indica em grupos distintos, grupo 3 e 2, respectivamente. A dificuldade de agrupar a subpopulação ADMIX pelo UPGMA também é justificável pelo fato desta subpopulação ser originada da mistura de inúmeras subpopulações.

No APÊNDICE I foi apresentado a descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os grupos que foram formados pelo método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipo após a aplicação dos métodos PCA, SPCA, IPCA, AUT, PCA-AUT, SPCA-AUT e IPCA-AUT. Em relação a alocação dos genótipos e subpopulações em cada um dos grupos, os mesmos

resultados foram encontrados pelo PCA e SPCA (APÊNDICE I -Tabela 7) e que diferiram apenas por um genótipo dos resultados encontrados pelo IPCA (APÊNDICE I - Tabela 8). No entanto, esta diferença entre agrupamentos não impactou a porcentagem de acerto destes métodos, que apresentaram igual porcentagem, pois nas três metodologias a alocação deste genótipo foi feita incorretamente. Ainda, as modificações propostas no Autoencoder (PCA-AUT, SPCA- AUT e IPCA-AUT) revelaram que as técnicas foram eficientes para reconhecer o padrão das duas principais subpopulações ao alocar o Indica e grande parte da variedade AUS no mesmo grupo e o Temperate japônica, Aromatic e Tropical japônica em um grupo distinto (APÊNDICE I -Tabela 10).

Como o objetivo desse estudo foi reduzir o tempo computacional da utilização de dados genômicos no reconhecimento de padrão, o mesmo foi alcançado. Tendo em vista os resultados apresentados na Tabela 6 o tempo computacional de todos os métodos de redução de dimensionalidade e a rede Autoencoder foram inferiores a análise preliminar D0. A PCA apresentou um tempo ligeiramente superior em relação aos demais métodos. Além disso, as metodologias que uniam os métodos de redução e o Autoenconder foram mais eficientes e reduziram o tempo computacional quando comparadas ao Autoencoder padrão.

Tabela 6. Número de componentes (Nc) e tempo computacional (TC) em segundos, considerando a Análise via Componentes Principais (PCA), Análise via Componentes Principais Esparsa (SPCA)Análise via Componentes Principais Independentes (IPCA) e autoencoder aplicados na matriz de marcadores moleculares no reconhecimento de padrões de genótipos de arroz.

Métodos	Nc	TC
D0	-	1398,22
PCA	2	0,49
SPCA	2	0,48
IPCA	2	0,47
AUT	4	0,63
PCA-AUT	1	0,46
SPCA-AUT	1	0,47
IPCA-AUT	1	0,42

4 Conclusões

Considerando a capacidade do método em agrupar os genótipos do arroz *Oryza Sativa* com menor variância dentro e maior variância entre bem como menor esforço computacional, a metodologia mais eficiente foi a técnica proposta SPCA-AUT, que utiliza componentes principais esparsos como variáveis de entrada no Autoencoder. A SPCA-AUT pode ser aplicada com sucesso em estudos de reconhecimento padrão e de diversidade genética nos programas de melhoramento.

5 Referências

Akinwale, M. G.; Gregorio, G.; Nwilene, F.; Akinyele, B. O.; Ogunbayo, S. A.; Odiyi, A. C. Heritability and correlation coefficient analysis for yield and its components in rice (Oryza sativa L.). **African Journal of plant science**, v. 5, 2011.

Allaire, J. J. et al. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza**. Genome Research**, v.16, 2006.

Ammiraju, J. S. S.; Luo, M.; Goicoechea, J. L.; Wang, W.; Kudrna, D.; Mueller, C.; Talag, J.; Kim, H.; Sisneros, N. B.; Blackmon, B.; Fang, E.; Tomkins, J. B.; Brar, D.; Mackill, D.; Maccouch, S.; Kurata, N.; Lambert, G.; Galbraith, D.W.; Arumuganathan, K.; Rao, K.; Walling, J. G.; Gill, N. Y.U.Y.; Sanmiguel, P.; Soderlund, C.; Jackson, S.; Wing, R. A. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, v.16, 2006.

Ashfaq, M.; Khan, A. S.; Ullah Khan, S. H.; Ahmad, R.. Association of Various Morphological Traits with Yield and Genetic Divergence in Rice (Oryza sativa). **International Journal of Agriculture & Biology**, v.14, 2012.

Azevedo, C. F.; Resende, M. D. V. D.; Silva, F. F.; Lopes, P. S.; Guimarães, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, 2013.

Comon, P. Independent component analysis, a new concept?. **Signal Process,** v.36, 1994.

Cordero-Lara, Karla I. Temperate japonica rice (Oryza sativa L.) breeding: History, present and future challenges. **Chilean journal of agricultural research**, 2020.

Costa, J. A. D. et al. Genomic prediction with the additive-dominant model by dimensionality reduction methods. **Pesquisa Agropecuária Brasileira**, v. 55, 2020.

da Costa, J. A., Azevedo, C. F., Nascimento, M., e Silva, F. F., de Resende, M. D. V., & Nascimento, A. C. C. A comparison of regression methods based on dimensional reduction for genomic prediction, v. 20, 2021.

Cruz, C D.; Ferreira, M. F.; Pessoni, L. A. Biometria Aplicada ao estudo da diversidade genética. **Suprema**. Visconde do Rio Branco, Minas Gerais. p. 13-17, 2011.

Cruz, C. D.; Regazzi, A. J.; Carneiro, P. C. S. Modelos Biométricos aplicados ao melhoramento genético. **Editora UFV** . Viçosa, Minas Gerais , v.1, 2012.

Cruz. C. D; Carneiro, P.C.S.; Regazzi, A. J. Modelos Biométricos aplicados ao melhoramento genético. **Editora UFV**. Viçosa, Minas Gerais, 2018

Datta, K.; Datta, S. K. Indica rice (Oryza sativa, BR29 and IR64). **Agrobacterium Protocols**, 2006

Garris, A. J.; Tai, T. H.; Coburn, J.; Kresovich, S.; McCouch, S. Genetic structure and diversity in Oryza sativa L. **Genetics,** v. 169, 2005.

Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit. **Adv. Neural Inf. Process. Syst**, v.10, 1998.

Hotelling, H. The relations of the newer multivariate statistical methods to factor analysis. **British Journal of Mathematical and Statistical Psychology**, v. 10, 1957.

Huang, X.; Kurata, N.; Wang, Z. X.; Wang, A.; Zhao, Q.; Zhao, Y.; Han, B. A map of rice genome variation reveals the origin of cultivated rice. **Nature,** v. 490, 2012.

Juliana, P.; Poland, J.; Huerta-Espino, J.; Shrestha, S.; Crossa, J.; Crespo-Herrera, L.; Singh, R. P. Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. **Nature genetics**, v. 51, 2019.

Jutten, C.; Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. **Sigmal Processing**, v. 24, 1991.

Kassambara, A; Mundt, F. Package 'factoextra'. **Extract and visualize the results of multivariate data analyses**, v. 76, 2017.

Kendall, M. G. A Course in Multivariate Analysis. London: Griffin, 1957.

Kubo, M.; Purevdorj, M. The future of rice production and consumption. **Journal of Food Distribution Research,** v. 35, 2004.

Le Cao, K. A.; Rohart, F.; Gonzalez, I.; Le Cao, M. K. A. Package 'mixOmics'. 2018.

Lee, K. S.; Choi, W. Y.; Ko, J. C.; Kim, T. S.; Gregorio, G. B. Salinity tolerance of japonica and indica rice (Oryza sativa L.) at the seedling stage. **Planta**, v. *216*, 2003.

Lee, L. C.; Liong, C. Y.; Osman, K.;Jemain, A. A.. Comparison of several variants of principal component analysis (PCA) on forensic analysis of paper based on IR spectrum. **In AIP Conference Proceedings** AIP Publishing LLC, v. 1750, 2016.

Lewis, N. D. Deep learning made easy with R. **A gentle introduction for data science**, 2016.

Mevik, B. H.; Wehrens, R.; Liland, K. H. Pls: Partial Least Squares and Principal Component Regression. **R package version 2.6-0**. https://CRAN.R-project.org/package=pls, 2016.

Rabbani, M. A.; Pervaiz, Z. H.; Masood, M. S. Genetic diversity analysis of traditional and improved cultivars of Pakistani rice (Oryza sativa L.) using RAPD markers. **Electronic journal of biotechnology**, v. 11, 2008.

Raj, M. P.; Swaminarayan, P. R.; Saini, J. R.; Parmar, D. K. Applications of pattern recognition algorithms in agriculture: a review. **International Journal of Advanced Networking and Applications**, v. 6, 2015.

Ranjith, P.; Sahu, S.; Dash, S. K.; Bastia, D. N.; Pradhan, B. D. Genetic diversity studies in Rice (Oryza sativa L.). **Journal of Pharmacognosy and Phytochemistry**, v. 7, 2018.

Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, 1987.

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. Nature, v. 323, 1986.

Semagn, K.; Magorokosho, C.; Vivek, B. S.; Makumbi, D.; Beyene, Y.; Mugo, S.; Prasanna Bm; Warburton, M. L. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. **BMC genomics**, v. 13, 2012.

Silva, I.N.; Spatti, D.H.; Flauzino, R.A. Redes Neurais Artificiais para Engenharia e Ciências Aplicadas. Fundamentos Teóricos e Aspectos Práticos. Editora Artliber, São Paulo – SP, 2016.

Silva Júnior, A. C. D.; Silva, M. J. D.; Cruz, C. D., Nascimento; M., Azevedo, C. F.; Soares, P. C. Patterns recognition methods to study genotypic similarity in flood-irrigated rice. **Bragantia**, v. 79, 2020.

Shen, Haipeng; Huang, Jianhua Z. Sparse principal component analysis via regularized low rank matrix approximation. **Journal of multivariate analysis**, v. 99, 2008.

Sokal, Robert R.; Rohlf, F. James. The comparison of dendrograms by objective methods. **Taxon**, v. 11, 1962.

Tibshirani, R. Regression shrinkage and selection via the lasso. **J R Stat Soc Series B Stat Methodol**. v. 58, 1996.

Thomson, M. J.; Septiningsih, E. M.; Suwardjo, F.; Santoso, T. J.; Silitonga, T. S.; Mccouch, S. R. Genetic diversity analysis of traditional and improved Indonesian rice (Oryza sativa L.) germplasm using microsatellite markers. **Theoretical and Applied Genetics**, v. 114, 2007.

Vanniarajan, C.; Vinod, K. K.; Pereira, Andy. Molecular evaluation of genetic diversity and association studies in rice (Oryza sativa L.). **Journal of genetics**, v. 91, 2012.

Xia, W.; Luo, T.; Zhang, W.; Mason, A. S.; Huang, D.; Huang, X.; Tang, W.; Dou, Y.; Zhang, C.; Xiao, Y. Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in Elaeis guineensis. **Frontiers in plant science**, v.10, 2019.

Yao, F.; Coquery, J.; Lê Cao, K.A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. **BMC bioinformatics**, v. 13, 2012.

Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., McCouch, S. R. (2010). Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. **PloS one**, v.5, e10780, 2010

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, J. G., Islam, A.R., Reynolds, A., Mezey, J., Mcclung, A. M., Bustamante, C. D., Mcclung, A. M. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature communications**, v. 2, 2011.

Zou, H.; Hastie, T.; Tibshirani, R. Sparse principal component analysis. **Journal of computational and graphical statistics**, v. 15, 2006.

APÊNDICE I

Tabela 7: Descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os três grupos que foram formados após a aplicação da PCA e SPCA e do método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipos.

Subpopulações	G1	G2	G3
TEJ (96)	77	19	0
IND (87)	0	70	17
AUS (57)	11	0	46
AROMATIC (14)	11	3	0
TRJ (97)	78	19	0
ADMIX (62)	45	16	1
Total	222	127	64

TEJ: *Temperate Japonica*, IND: Indica e TRJ: *Tropical Japonica*.

Tabela 8: Descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os três grupos que foram formados após a aplicação da IPCA e do método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipos.

Subpopulações	G1	G2	G3
TEJ (96)	77	19	0
IND (87)	0	70	17
AUS (57)	11	0	46
AROMATIC (14)	11	3	0
TRJ (97)	78	19	0
ADMIX (62)	45	17	0
Total	222	128	63

TEJ: *Temperate Japonica*, IND: Indica, AUS: TRJ: *Tropical Japonica*.

Tabela 9: Descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os dois grupos que foram formados após a aplicação da AUT e do método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipos.

Subpopulações	G1	G2
TEJ (96)	77	19
IND (87)	17	70
AUS (57)	11	46
AROMATIC (14)	11	3
TRJ (97)	78	19
ADMIX (62)	43	19
Total	237	176

TEJ: *Temperate Japonica*, IND: Indica, AUS: TRJ: *Tropical Japonica*.

Tabela 10: Descrição das seis principais subpopulações do arroz Asiático, *Oryza sativa* (número de genótipos) e os dois grupos que foram formados após a aplicação da PCA-AUT, SPCA-AUT e IPCA-AUT e do método de agrupamento UPGMA (*unweighted pair-group method using arithmetic averages*) com os respectivos números de genótipos.

Métodos	Grupos	TEJ	IND	AUS	ARO	TRJ	ADM
PCA-AUT	G1	96	0	3	9	95	45
	G2	0	87	54	5	2	17
SPCA-AUT	G1	96	1	5	7	96	45
	G2	0	86	52	7	1	17
IPCA-AUT	G1	77	17	11	11	78	44
	G2	19	70	46	3	19	18

TEJ: *Temperate Japonica*, IND: Indica e TRJ: *Tropical Japonica*.