Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Estabilidade em análise de agrupamento via modelo AMMI com
reamostragem “bootstrap”

Débora Robert de Godoi

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Esta-
tística e Experimentação Agronômica

Piracicaba
2013

Débora Robert de Godoi
Licenciada em Matemática

**Estabilidade em análise de agrupamento via modelo AMMI com
reamostragem “bootstrap”**

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Dissertação apresentada para obtenção do título de Mestra
em Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba

2013

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Godoi, Débora Robert de
Estabilidade em análise de agrupamento via modelo AMMI com reamostragem
"bootstrap" / Débora Robert de Godoi.- - Piracicaba, 2013.
90 p: il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2013.

1. Interação entre genótipos e ambientes 2. Modelo AMMI 3. Bootstrap 4. Análise de agrupamentos 5. Biplot I. Título

CDD 519.54
G588e

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte -O autor"

DEDICATÓRIA

Aos meus pais,

Cláudia Robert de Godoi e Benedito de Godoi

pelo exemplo de vida, por estarem sempre ao meu lado sempre dando forças quando precisei e por acreditarem mais em mim do que eu mesma.

Ao meu namorado,

Gabriel Tambarussi Avancini

pelo carinho, incentivo, paciência, por toda ajuda, pelo amor sempre constante e por ter ficado ao meu lado em todos os momentos, felizes ou não.

AGRADECIMENTOS

Aos meus pais, Cláudia Robert de Godoi e Benedito Aparecido de Godoi por estarem sempre ao meu lado me apoiando e nunca me deixarem desanimar. Vocês foram fundamentais para que esse sonho se tornasse realidade! Obrigada por tudo, amo vocês!

À minha avó, Therezinha de Camargo Robert, pelo exemplo de vida, pelo apoio, conselhos e incentivo para que eu prosseguisse com os estudos.

Aos meus irmãos, Karina Robert Guedes e Diego Robert Guedes (in memoriam) e aos meus sobrinhos Pyetro Robert Avanzi e Théo Robert Pimenta pelo amor, carinho, alegria e por tudo que passamos juntos. Obrigada pela compreensão da minha ausência em muitos momentos.

Ao Prof. Dr. Carlos Tadeu dos Santos Dias, pela orientação, confiança e pelo incentivo constante dado para a realização deste trabalho.

Ao Prof. Dr. José Baldin Pinheiro, por ter cedido os dados.

À Prof^a Dra. Priscila Neves Faria, um agradecimento em especial, pela ideia do trabalho, paciência e pelos ensinamentos sempre que precisei. Sua colaboração foi fundamental para a realização deste trabalho, muito obrigada por tudo!

Ao Prof. Dr. Antônio Carlos Simões Pião meu orientador de graduação, por sempre ter me incentivado a prosseguir com os estudos e por todo o apoio.

Ao Kuang Hongyu, por toda ajuda no decorrer do trabalho.

Ao Anderson da Silva, pela paciência e pela ajuda com a programação.

Aos amigos do doutorado, Ana Julia Righetto e Everton Batista pelos momentos de distração, risadas e alegria sempre presente, obrigada pelo apoio.

Ao Thiago Gentil, por toda ajuda nesses anos de mestrado e pela paciência sempre presente, muito obrigada!

Aos amigos da UNESP - Rio Claro, por toda nossa história durante os anos inesquecíveis da graduação, e que permanecem meus amigos até hoje. Em especial, Willian Bala, Joane Santos, Ana Paula Inforsato, Tassia Zangiacomo e Tábata Fabri pela alegria constante e por estarem ao meu lado sempre.

Às minhas amigas Débora Marchetti e Letícia Lázaro, que embora distantes, me apoiaram em todos os momentos que precisei. Obrigada pela amizade sempre!

Às amigas que convivi neste último ano, Bruna Merlin e Cecilia Assis,

obrigada pelo apoio e por aguentar os momentos finais de estresse.

À Aurea de Vasconcelos e Clemente de Vasconcelos, pelos ensinamentos, incentivo e ao apoio constante, mesmo distante. Vocês foram a base para que este trabalho acontecesse!

Ao meu primo, Rafael Robert pela ajuda e paciência.

Aos professores do Departamento de Ciências Exatas da ESALQ, pela formação.

Aos funcionários do Departamento de Ciências Exatas da ESALQ, Eduardo Bonilha e Solange de Assis Paes Sabadin pelo auxílio permanente e, em especial, à Luciane Brajão pela ajuda e amizade presente em todos os momentos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo auxílio financeiro na forma de bolsa de estudos.

“Se você pode sonhar, você pode fazer.”

Walt Disney

“Eu tentei 99 vezes e falhei, mas na centésima tentativa eu consegui, nunca desista de seus objetivos mesmo que esses pareçam impossíveis, a próxima tentativa pode ser a vitoriosa.”

Albert Einstein

SUMÁRIO

RESUMO	11
ABSTRACT	13
LISTA DE FIGURAS	15
LISTA DE TABELAS	17
1 INTRODUÇÃO	19
2 REVISÃO BIBLIOGRÁFICA	21
2.1 A interação entre genótipos e ambientes	21
2.1.1 Métodos estatísticos para o estudo da interação $G \times A$	22
2.2 Análise de componentes principais	23
2.3 A representação gráfica em biplot	24
2.4 Modelo AMMI	24
2.5 Medidas de dissimilaridade	25
2.6 Análise de agrupamentos	26
2.7 Dendrograma	27
2.7.1 Métodos hierárquicos	28
2.7.1.1 Métodos aglomerativos	28
2.7.1.2 Métodos divisivos	29
2.7.2 Métodos não - hierárquicos	29
2.8 Reamostragem	30
2.9 Bootstrap	30
3 MATERIAL E MÉTODOS	33
3.1 Análise de variância	33
3.2 Análise de variância conjunta	35
3.3 Autovalores e autovetores	37
3.4 Decomposição em valores singulares	37
3.5 Análise de componentes principais	38
3.6 A representação gráfica biplot	39
3.7 Modelo AMMI	40
3.7.1 Ajuste dos efeitos principais por análise de variância	43
3.7.2 Testes estatísticos para a seleção do modelo AMMI	44
3.8 Determinação do número de grupos no dendrograma	46
3.9 Métodos de agrupamento	47

3.9.1 Procedimento para o método do vizinho mais próximo	47
3.9.2 Procedimento para o método do vizinho mais distante	47
3.9.3 Procedimento para o método da ligação média	48
3.9.4 Procedimento para o método do centroide	48
3.9.5 Procedimento para o método da mediana	48
3.9.6 Procedimento para o método de Ward	48
3.10 Coeficiente de correlação cofenética como critério para avaliar métodos de agrupamento	49
3.11 Estabilidade via método bootstrap	50
4 RESULTADOS E DISCUSSÃO	51
4.1 Adaptabilidade e estabilidade de cultivares de soja por meio da modelagem AMMI	51
4.2 Análise de agrupamentos sobre a matriz de distâncias original	54
4.3 Análise de agrupamentos sobre a matriz de distâncias euclidianas obtida via bootstrap	57
4.4 Correlação cofenética	61
5 CONSIDERAÇÕES FINAIS	67
REFERÊNCIAS	69
APÊNDICES	73
ANEXOS	85

RESUMO

Estabilidade em análise de agrupamento via modelo AMMI com reamostragem “bootstrap”

O objetivo deste trabalho é propor uma nova metodologia de interpretação da estabilidade dos métodos de agrupamento, para dados de vegetação, utilizando a metodologia AMMI e a reamostragem (bootstrap), para ganhar confiabilidade nos agrupamentos formados. Os dados utilizados são provenientes do departamento de genética da Escola Superior de Agricultura “Luiz de Queiroz”, e visam à produtividade de soja. Primeiramente aplica-se a metodologia AMMI e então, é estimada a matriz de distâncias euclidianas - com base nos dados originais e obtidos via reamostragem (bootstrap) - para a aplicação dos métodos de agrupamento (vizinho mais próximo, vizinho mais distante, ligação média, centroide, mediana e Ward). Para a verificação da validade dos agrupamentos formados utiliza-se o coeficiente de correlação cofenética, e pelo teste de Mantel, é apresentada a distribuição empírica dos coeficientes de correlação cofenética. Os agrupamentos obtidos pelos diferentes métodos são, em sua maioria, semelhantes indicando que, em princípio, qualquer um desses métodos seria adequado para a representação. O método que apresenta resultados discrepantes em relação aos outros (tanto para os dados originais, quanto pelos dados obtidos via bootstrap) – na representação gráfica em dendrograma - é método de Ward. Este estudo é promissor na análise da validade de agrupamentos formados em dados de vegetação.

Palavras-chave: Interação entre genótipos e ambientes; Modelo AMMI; Bootstrap; Análise de agrupamentos; Biplot

ABSTRACT

Stability in clustering analysis through the AMMI methodology with bootstrap

The objective of this work is to propose a new interpretation methodology of clustering methods for vegetation data stability, using the AMMI and bootstrap methodology, to gain reliability in the clusters formed. The database used is from the Department of Genetics of Luiz de Queiroz College of Agriculture, aiming soybean yield. Firstly AMMI is applied, then the Euclidian distance matrix is estimated - based on the original data and on the acquired by the bootstrap method - for the application of clustering methods (nearest neighbor, furthest neighbor, average linkage, centroid, median and Ward). In order to assess the validity of clusters formed the cophenetic correlation coefficient is used, and the Mantel test, in order to show the empirical distribution of the cophenetic correlation coefficients. The clusters obtained by different methods are, in most cases, quite similar, indicating that in principle, any of these methods would be suitable for the representation. The method that presents discrepant results (for both the original and bootstrap method obtained data) - on the dendrogram graphical representation, compared to the others - is the Ward's. This study is promising in the analysis of validity of clusters formed in vegetation data.

Keywords: Genotypic and environmental interaction; AMMI Model; Bootstrap; Clustering analysis; Biplot

LISTA DE FIGURAS

Figura 1 - Biplot AMMI ₁ ($PC_1 \times \text{Média}$) para os dados de produtividade de soja .	53
Figura 2 - Biplot AMMI ₂ ($PC_1 \times PC_2$) para os dados de produtividade de soja . .	53
Figura 3 - Dendrograma obtido pelo emprego do método do vizinho mais próximo com base na distância euclidiana da matriz de interações	54
Figura 4 - Dendrograma obtido pelo emprego do método do vizinho mais distante com base na distância euclidiana da matriz de interações	55
Figura 5 - Dendrograma obtido pelo emprego do método do centroide com base na distância euclidiana da matriz de interações	55
Figura 6 - Dendrograma obtido pelo emprego do método da mediana com base na distância euclidiana da matriz de interações	56
Figura 7 - Dendrograma obtido pelo emprego do método de Ward com base na distância euclidiana da matriz de interações	56
Figura 8 - Dendrograma obtido pelo emprego do método da ligação média entre grupos com base na distância euclidiana da matriz de interações	57
Figura 9 - Dendrograma obtido pelo emprego do método do vizinho mais próximo com base na matriz de distâncias euclidianas via bootstrap da matriz de interações	58
Figura 10 - Dendrograma obtido pelo emprego do método do vizinho mais distante com base na matriz de distâncias euclidianas via bootstrap da matriz de interações	58
Figura 11 - Dendrograma obtido pelo emprego do método do centroide com base na matriz de distâncias euclidianas via bootstrap da matriz de interações .	59
Figura 12 - Dendrograma obtido pelo emprego do método da mediana com base na matriz de distâncias euclidianas via bootstrap da matriz de interações .	59
Figura 13 - Dendrograma obtido pelo emprego do método de Ward com base na matriz de distâncias euclidianas via bootstrap da matriz de interações .	60
Figura 14 - Dendrograma obtido pelo emprego do método da ligação média entre grupos com base na matriz de distâncias euclidianas via bootstrap da matriz de interações	60

Figura 15 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do vizinho mais próximo, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do vizinho mais próximo utilizando o teste de Mantel	62
Figura 16 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do vizinho mais distante, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do vizinho mais distante utilizando o teste de Mantel	62
Figura 17 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do centroide, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do centroide utilizando o teste de Mantel	63
Figura 18 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método da mediana, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método da mediana utilizando o teste de Mantel	63
Figura 19 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método de Ward, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método de Ward utilizando o teste de Mantel	64
Figura 20 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método da ligação média, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método da ligação média utilizando o teste de Mantel .	64

LISTA DE TABELAS

Tabela 1 - Análise de variância de um experimento inteiramente aleatorizado . . .	34
Tabela 2 - Análise de variância conjunta para experimentos de um mesmo grupo grupo de genótipos avaliados em a ambientes com r blocos	36
Tabela 3 - Análise de variância para experimentos de um mesmo grupo grupo de genótipos avaliados em a ambientes com r repetições	36
Tabela 4 - Desdobramento da interação $G \times A$ pelo método de Gollob	45
Tabela 5 - Análise de variância conjunta dos dados de produtividade de soja e de- composição das somas de quadrados da interação $G \times A$	51
Tabela 6 - Correlação cofenética entre as matrizes cofenética e de dissimilaridade obtidas conforme o método de agrupamento utilizado	61

1 INTRODUÇÃO

No estudo de variedades, um dos princípios é partir de uma determinada espécie, obter o seu fenótipo que é determinado por dois componentes: a sua composição genética (genótipo) e a contribuição do ambiente em que a variedade se encontra.

Em um experimento, quando é estudada uma mesma variedade em diferentes locais (ambientes), é observada uma diferença na produtividade, e até mesmo no crescimento da mesma. Por exemplo, se estudarmos uma determinada variedade de soja no sudeste, no nordeste e na região sul do Brasil, pelo fato de existir diferenças climáticas, pode ser que a variedade seja extremamente produtiva na região sudeste e pouco produtiva na região nordeste pela diferença existente entre a umidade de solos nesses dois locais. Essa diferença de produtividade ocorre também devido à interação entre genótipos e ambientes (interação genótipos $\times$ ambientes).

Melhoristas de plantas veem essa interação como um fator negativo, mas é possível aproveitar essa interação a favor da pesquisa, existindo a possibilidade de combinar o melhor genótipo com um ambiente que o produza em larga escala. Se um genótipo se adaptasse bem a qualquer ambiente, apenas um ensaio experimental seria suficiente para o seu estudo universal.

Para aproveitar a interação da melhor maneira possível, é necessário fazer um estudo sobre a mesma. Existem vários tipos de estudo, mas em sua maioria são baseados em regressão linear e isso torna o estudo limitado quando se tem falta de linearidade. Os métodos multivariados têm grande destaque nesse tipo de estudo, em especial a metodologia AMMI (additive main effects and multiplicative interaction analysis). Este método combina componentes aditivos para os efeitos principais (genótipos e ambientes) e componentes multiplicativos para a interação genótipos $\times$ ambientes. Por essa análise, identifica-se os genótipos mais produtivos assim como os ambientes que favorecem à produtividade e a análise estatística é de fácil interpretação.

Como a interação genótipos $\times$ ambientes é representada pelos termos multiplicativos do modelo, sendo assim, é de grande importância determinar o número ótimo de termos multiplicativos a serem considerados no modelo AMMI.

Outro método multivariado que possui grande destaque é a análise de agrupamentos (cluster analysis). Esta análise tem por objetivo reunir, por algum critério de classificação as unidades amostrais em grupos, de tal modo que exista homogeneidade dentro dos grupos e heterogeneidade entre os grupos. Esta análise é feita em dois

passos: primeiramente determina-se a medida de dissimilaridade que será utilizada, posteriormente é tomada a decisão de qual método de agrupamento é o ideal para a análise dos dados. Escolher qual é o método de agrupamento ideal não é uma tarefa simples, já que existem muitos métodos disponíveis e cada um dos métodos podem levar o pesquisador a diferentes resultados.

Com o avanço dos recursos computacionais, a análise multivariada teve altos ganhos, economizou tempo e conseguiu otimizar os resultados. Um exemplo disso é a reamostragem “bootstrap”. Esse recurso computacional vem sendo bastante utilizado para avaliar a estabilidade dos agrupamentos obtidos a partir das matrizes de dissimilaridade. Sendo assim, esse procedimento pode ser utilizado para encontrar um ponto de equilíbrio que facilite a estimativa dos grupos a serem considerados (ALBUQUERQUE, 2005).

O objetivo deste trabalho é propor uma sistemática para o estudo da estabilidade dos métodos em análise de agrupamento, utilizando a reamostragem “bootstrap” juntamente com a metodologia AMMI auxiliando para a tomada de decisões sobre os agrupamentos a serem formados.

2 REVISÃO BIBLIOGRÁFICA

2.1 A interação entre genótipos e ambientes

A experimentação agronômica frequentemente necessita da análise conjunta de experimentos, que tem como objetivo generalizar os resultados obtidos nos experimentos para toda uma área ou região. Quando isto é feito, surge a interação entre genótipos e ambientes ($G \times A$), que reflete às diferentes respostas dos genótipos às diferentes condições ambientais, sendo um dos aspectos mais importantes na seleção e na recomendação dos genótipos com alta produtividade. O potencial genético de um cultivar não depende apenas do genótipo, mas também das condições climáticas, do tipo de solo e das mudanças que podem ocorrer no ambiente (HONGYU, 2012).

O fenótipo é resultante da ação conjunta do genótipo, do ambiente e da interação $G \times A$. O ambiente reflete as diferentes sensibilidades dos genótipos às variações do ambiente, resultando assim, na mudança de seus desempenhos (PIOVESAN, 2007).

Quando um experimento é realizado com o objetivo de testar um mesmo grupo de genótipos em diversos ambientes, o pesquisador pode obter respostas fenotípicas distintas dos diferentes genótipos para cada um dos ambientes em teste. Essa diferença é devido à interação $G \times A$ que pode ser provocada por fatores fisiológicos, adaptativos, entre outros fatores (CRUZ; CARNEIRO, 2003).

Em melhoramento genético, o principal objetivo é a recomendação dos genótipos que garantam uma elevada produtividade nos mais diversos ambientes. A seleção é feita avaliando-se o desempenho dos genótipos nos mais variados ambientes (local, ano, época de semeadura).

Mas a interação não deve ser vista como uma dificuldade, mas sim como uma oportunidade de melhorar o conhecimento da pesquisa e caso seja uma interação positiva, existirá uma oportunidade de ganhos elevados (PIOVESAN, 2007). Adaptações específicas de genótipos a ambientes podem fazer a diferença entre uma boa e uma excelente variedade (GAUCH; ZOBEL, 1996). Se não existisse tal interação, apenas um ensaio experimental (ainda com repetições) seria o suficiente, já que uma variedade adaptar-se-ia facilmente na maior parte dos ambientes (DUARTE; VENCOSKY, 1999).

A interação $G \times A$ pode ser simples ou complexa. A interação simples possibilita que o cultivar seja recomendado de forma generalizada enquanto a complexa torna a recomendação restrita a ambientes específicos (RAMALHO et al., 1993).

Em melhoramento genético, a influência que o ambiente exerce sobre cada genótipo tem merecido atenção especial por parte dos melhoristas de plantas, já que as interações entre genótipos e ambientes podem interferir de forma positiva ou negativa nos resultados finais (GONZALEZ, 1988).

Segundo Chaves (2001), a interação $G \times A$ resulta da resposta diferencial dos genótipos à variação ambiental e não deve ser vista como um simples efeito estatístico, mas como um fenômeno biológico em suas aplicações no melhoramento de plantas. Entretanto, pode-se dizer que unindo os conhecimentos em estatística com o conhecimento do fenômeno biológico é possível tirar informações valiosas para entender a complexidade da interação.

O estudo da interação $G \times A$ ganhou grande importância, pois existe a possibilidade do genótipo mais produtivo em um determinado ambiente não o ser em outro, o que influencia o ganho de seleção e dificulta a recomendação de genótipos para toda a população de ambientes amostrada pelos testes (DUARTE; VENCOSKY, 1999).

A interação $G \times A$ pode ser reduzida utilizando-se genótipos que possuam ampla adaptabilidade e sejam estáveis, ou estratificando-se a região em questão em sub-regiões que tenham características ambientais semelhantes de tal modo que a interação seja não significativa (RAMALHO et al., 1993).

2.1.1 Métodos estatísticos para o estudo da interação $G \times A$

Para que seja possível explorar os efeitos positivos da interação, é necessário utilizar métodos estatísticos que sejam capazes de captar esse tipo de informação.

Nos últimos trinta anos, várias metodologias foram propostas com o objetivo de diminuir os desafios impostos pela interação $G \times A$. Atualmente os pesquisadores ainda buscam instrumentos estatísticos que sejam cada vez mais efetivos nesta tarefa. Entre os métodos tradicionais, tem-se o zoneamento agrônomo que consiste em estratificar a população de ambientes em sub-regiões, dentro das quais a interação seja não significativa permitindo-se assim, recomendações regionalizadas (DUARTE; ZIMMERMANN, 1991).

Outro método bastante utilizado é a regressão linear simples e múltipla (BRASIL, 1990), que consiste em identificar genótipos estáveis, isto é, identificar os genótipos que possuam uma razoável produtividade nos diferentes ambientes e contribuam menos para a interação, podendo assim, serem recomendados para diversos ambientes.

Segundo Crossa et al. (1990), os métodos de regressão linear embora sejam

bastante utilizados, possuem muitas limitações quanto sua utilização e acrescenta que se a linearidade falha, a análise de regressão não é informativa. A análise de regressão é altamente dependente do grupo de genótipos e ambientes incluídos e tende a simplificar modelos de resposta, explicando a variação devida à interação em uma única dimensão por meio do coeficiente de regressão, quando na realidade ela pode ser bastante complexa (DUARTE; VENCOVSKY, 1999). Procedimentos univariados como a análise de variância têm dificuldades em detectar a interação entre os fatores (ZOBEL et al., 1988).

Segundo Crossa et al. (1990), a utilização dos métodos estatísticos multivariados é útil para explorar as informações contidas nos dados e oferece a possibilidade de avaliar um conjunto de características levando em consideração as correlações existentes. Dentre os métodos multivariados, pode-se citar a análise de componentes principais, análise de agrupamentos, análise discriminante e também, a análise AMMI (additive main effects and multiplicative interaction analysis) que vem ganhando grande aplicabilidade nos últimos anos.

2.2 Análise de componentes principais

A técnica de análise de componentes principais (ACP) foi descrita inicialmente por Karl Pearson (1901) (MANLY, 2008). Este é o método mais simples quanto à sua utilização, mas é considerado como o principal método da análise multivariada pela sua fundamentação.

A finalidade da ACP é a análise dos dados visando sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais com menor perda possível da informação da variância contida nos dados.

Esta análise é bastante útil quando os vetores de características possuem muitas dimensões, dificultando-se assim a representação gráfica. Para reduzir a massa de dados, são feitas combinações lineares das variáveis originais que representam a seleção de novos sistemas de coordenadas obtidos pela rotação do sistema original. Esses novos eixos ortogonais (novas variáveis) são chamados de componentes principais e os valores das novas variáveis são chamados de escores dos componentes principais (HONGYU, 2012).

2.3 A representação gráfica em biplot

Os métodos gráficos mostram de maneira clara e rápida ao pesquisador o comportamento dos dados em questão. Com a utilização dos recursos computacionais, houve um maior desenvolvimento sobre os recursos gráficos e uma maior utilização dos mesmos nos trabalhos e pesquisas científicas.

Biplot é uma técnica multivariada que foi proposta por Gabriel (1971), com o objetivo de representar graficamente uma matriz de dados, de tal modo que esta representação permita observar em um plano as relações entre as linhas e as colunas desta matriz. Esta representação gráfica é baseada na fatoração da matriz de dados por meio da decomposição em valores singulares (SOUZA, 2010).

Esta representação é de grande valia para diagnosticar o melhor modelo estatístico para um conjunto de dados. Kempton (1984) utilizou biplot em dados que visavam a produtividade, e a interpretação foi feita nos modelos de análise de variância, regressão linear, ACP e AMMI. Esse autor cita que o biplot, ao seu ver, é uma excelente ferramenta para o estudo da interação $G \times A$ (ARAÚJO, 2008).

No método AMMI, a utilização do biplot vem da decomposição em valores singulares aplicada à matriz de interações que será representada, geralmente, por meio de sua aproximação de posto dois ou três. Ainda neste método, o gráfico biplot mostra a produtividade média dos genótipos e dos ambientes e também estuda a adaptabilidade e estabilidade dos mesmos. Para a construção do gráfico, o posto da matriz obtida pela decomposição em valores singulares deverá ser dois ou três (em aplicações, o mais comum são esses dois casos) resultando assim, num biplot em duas ou três dimensões, respectivamente.

Os gráficos em biplot que são feitos utilizando a análise AMMI em artigos para o estudo da interação $\mathbf{GA}$, normalmente apresentam o primeiro eixo de interação (IPCA_1). Mesmo assim, os gráficos são construídos em duas dimensões, utilizando-se o eixo das abcissas para representar os efeitos principais (médias de genótipos e ambientes) para o referido eixo de interação.

2.4 Modelo AMMI

A análise AMMI é um método uni-multivariado proposto inicialmente por Mandel (1971) que combina análise de variância para os efeitos principais (genótipos e ambientes) com a ACP para os efeitos provocados pela interação $G \times A$. Esta análise permite

um maior detalhamento sobre a soma de quadrados interação $G \times A$ ($SQ_{G \times A}$)(ZOBEL et al., 1988).

A análise AMMI ganhou grande aplicabilidade nos últimos anos pelo fato de explicar a interação $G \times A$ de maneira eficiente, quando comparada aos métodos univariados. Segundo Zobel et al. (1988), os procedimentos univariados como, por exemplo, a análise de variância possuem limitações em encontrar interação entre os fatores. Existem casos em que a $SQ_{G \times A}$ atinge cerca de 50% da SQ_{Total} e ainda assim, a interação não atinge significância estatística. O método de regressão linear também nem sempre se mostra eficiente. Há análises em que o método AMMI captou cerca de 71% da $SQ_{G \times A}$ enquanto o método de regressão linear captou apenas em torno de 8%. Sendo assim, o método AMMI mostra-se mais eficiente quando o pesquisador está interessado em estudar essa interação para recomendação de genótipos com ambientes que tornem a produção de uma determinada cultura promissora.

Utilizando a representação gráfica biplot, tem-se que o método AMMI permite uma análise mais detalhada da interação $G \times A$, com a vantagem de selecionar genótipos mais produtivos, estimar mais precisamente as respostas genotípicas e possibilitar uma fácil interpretação dos resultados (FARIA, 2012).

Os escores AMMI têm distribuição de probabilidade desconhecida e isso dificulta a obtenção de fórmulas via argumentos analíticos. Para medir a precisão desses escores, Lavoranti (2003) apresentou uma metodologia complementar para avaliar a precisão dos escores de genótipos e de ambientes, por meio de técnicas de reamostragem “bootstrap”, permitindo uma melhoria na qualidade das inferências sobre as adaptabilidades e estabilidades fenotípicas estimadas pelo modelo AMMI (FARIA, 2012).

2.5 Medidas de dissimilaridade

Um critério que quantifique a distância entre dois objetos é chamado “medida de dissimilaridade”. Quanto menor for o valor dessa medida, significa que mais similares são os objetos em comparação. Muitas medidas de dissimilaridade têm sido propostas e utilizadas em análise de agrupamentos, mas cada uma delas produz um determinado tipo de agrupamento, sendo assim, deve-se utilizar a mesma medida para toda a análise.

As distâncias mais utilizadas em estudos genéticos são: a distância euclidiana, o quadrado da distância euclidiana, a distância euclidiana média, a distância ponderada e a distância generalizada de Mahalanobis (1936). Esta última, leva em consideração

a magnitude das variâncias e covariâncias residuais que existem entre os objetos em estudo (SILVA, 2012). Outras medidas de dissimilaridade podem ser encontradas na literatura.

Alguns autores consideram a distância euclidiana uma medida de dissimilaridade insatisfatória para análise estatística. Isso se deve ao fato de cada coordenada ter o mesmo peso para o cálculo da distância. A vantagem da utilização da distância euclidiana deve-se ao universo euclidiano e ao fato de minimizar os erros da classificação nos agrupamentos. Essa medida é recomendada quando as unidades de cálculos são escores de componentes principais como é o caso da metodologia AMMI (FARIA, 2012).

Neste trabalho fez-se o uso dessa distância por ser uma das mais utilizadas em trabalhos científicos, pela praticidade e por ser facilmente encontrada em programas computacionais (FARIA, 2012).

2.6 Análise de agrupamentos

O objetivo da análise de agrupamentos é obter grupos homogêneos por um esquema que dê a possibilidade de reunir os indivíduos em um determinado número de grupos, de modo que exista homogeneidade dentro dos grupos e heterogeneidade entre os grupos (JOHNSON; WICHERN, 1992). Em outras palavras, a análise de agrupamentos consiste em classificar n itens (variedades, por exemplo) avaliados por um conjunto de m variáveis, com o objetivo de separar os itens em grupos, de tal modo que os itens mais semelhantes permaneçam em um mesmo grupo (FARIA, 2012).

A análise de agrupamentos é uma técnica multivariada que vem sendo bastante utilizada, pois frequentemente o pesquisador necessita classificar variedades. Por exemplo, surge a necessidade de agrupar genótipos que têm comportamento semelhante quanto à produtividade, ou seja, dado um conjunto de n genótipos, tem-se por objetivo agrupá-los em subgrupos de tal modo que cada subgrupo seja internamente homogêneo e que os grupos formados sejam heterogêneos entre si.

O procedimento da análise se divide em duas etapas: primeiramente relaciona-se com a estimação de uma medida de dissimilaridade (ou similaridade) entre as unidades amostrais e depois adota-se uma técnica de agrupamento para formação dos grupos (CRUZ; REGAZZI, 1994).

Considere que por exemplo, deseja-se analisar dados de n indivíduos que serão estudados por meio de p variáveis. A estimação das medidas de dissimilaridade consiste na conversão da matriz de observações $n \times p$ em uma matriz quadrada e simétrica $n \times n$

de dissimilaridades individuais, isto é, medidas de distância que são calculadas entre pares de indivíduos. A intersecção da i -ésima linha com a j -ésima coluna é dada pela distância entre o i -ésimo e o j -ésimo indivíduo; quanto menor for esse valor, maior é a similaridade, ou seja, os indivíduos são próximos em relação ao conjunto de variáveis (ABREU et al., 2002).

Segundo Barroso & Artes (2003), as distâncias são as medidas de dissimilaridade mais utilizadas no estudo de bancos de dados com variáveis quantitativas. Para identificar quais são esses subgrupos formados, utiliza-se a análise de agrupamentos (cluster analysis). Existem dois métodos para a análise: os métodos hierárquicos e os métodos não-hierárquicos.

2.7 Dendrograma

Dendrograma é uma representação gráfica bidimensional em forma de árvore utilizada para ilustrar a análise de agrupamentos feita sobre um conjunto de dados. Essa representação ilustra todo o procedimento de agrupamentos por meio de uma estrutura de árvore. Se for feito um corte em um determinado nível do gráfico, este corte representará o número de grupos existentes nesse nível e dos indivíduos que os formam (ALBUQUERQUE, 2005).

A interpretação de um dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas, portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O gráfico denominado por dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade de todo o conjunto de amostras utilizado no estudo. Um caso particular é verificado quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre estas variáveis do conjunto de dados estudado.

Vale salientar que o dendrograma ilustra as partições feitas em cada nível sucessivo do processo de agrupamento. Neste, um eixo representa os indivíduos e o outro eixo representa a variabilidade ou as distâncias obtidas após a utilização de uma metodologia de agrupamento. Os ramos da árvore fornecem a ordem das $n - 1$ ligações, em que o primeiro nível apresenta a primeira ligação, o segundo nível apresenta a segunda ligação, e assim sucessivamente. Por outro lado, a falta de critérios objetivos para determinar o número ótimo de grupos é uma dificuldade encontrada em estudos que utilizam análise de

agrupamentos.

2.7.1 Métodos hierárquicos

Nestes métodos os indivíduos são classificados em grupos em diferentes etapas de modo ordenado (hierárquico), produzindo ao final da análise um gráfico em forma de árvore (dendrograma). Estes agrupamentos podem ser utilizados tanto para agrupar indivíduos quanto para agrupar variáveis. Quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre elas. Os métodos hierárquicos podem ser subdivididos em métodos aglomerativos e métodos divisivos (SARTORIO, 2008).

2.7.1.1 Métodos aglomerativos

São os mais comuns de se encontrar em trabalhos científicos. Estes métodos são baseados em uma medida de dissimilaridade escolhida a priori, reduzindo a um único grupo ao final. Em geral, o algoritmo para n objetos é:

1. Inicia-se com n grupos, cada um com um único elemento e com uma matriz de distâncias simétrica de ordem n ;
2. Busca-se na matriz de distâncias o par de grupos com a menor distância (mais similar);
3. Fundir dois grupos x e y e nomeá-lo, por exemplo, como xy . Calcular novamente e rearranjar as distâncias na matriz de distâncias utilizando a medida de parença e o método aglomerativo escolhido;
4. Eliminar as linhas e as colunas correspondentes aos grupos x e y ;
5. Acrescentar uma nova linha e uma nova coluna com as distâncias entre o grupo xy e os demais grupos;
6. Repetir os passos 2 e 3 $n - 1$ vezes até que todos os objetos estejam agrupados em um único grupo.

2.7.1.2 Métodos divisivos

Estes métodos partem de um único grupo com os n elementos e por divisões sucessivas são divididos em $2, 3, \dots$, grupos, de tal modo que os indivíduos em um subgrupo estarão distantes dos indivíduos de outro subgrupo. Esses indivíduos são divididos novamente em subgrupos e o processo se repete até que cada indivíduo forme um grupo (SARTORIO, 2008). Por este método não ser tão abrangente em pesquisas científicas, não abordaremos o mesmo neste trabalho.

2.7.2 Métodos não - hierárquicos

Nestes métodos, o número de subgrupos k que serão formados é fixado inicialmente, isto é, é feita uma classificação dos n indivíduos em k subgrupos. É necessário fixar antecipadamente critérios que produzam medidas sobre a qualidade da partição produzida. Com o objetivo de tornar a classificação coerente, é feita a transferência de indivíduos de uma classe para outra de tal modo que os subgrupos sejam internamente homogêneos e externamente heterogêneos. Este método é utilizado quando o pesquisador já sabe quantos grupos o seu experimento deve formar. De maneira geral, os métodos não - hierárquicos seguem os seguintes passos:

1. Selecionar k centroides de grupos, em que k é o número de grupos inicial e arbitrário;
2. Designar cada observação ao grupo mais próximo;
3. Realocar cada observação a um dos k grupos de acordo com uma regra de parada determinada anteriormente;
4. Se não existirem mais realocações de pontos ou se a realocação satisfizer o critério na regra de parada, deve-se parar o processo. Caso contrário, deve-se retornar ao passo 2.

Existem vários métodos de agrupamentos, esses que se diferenciam na maneira como especificar a medida de proximidade entre os indivíduos. Recomenda-se que sejam aplicados diferentes métodos e que sejam feitas as comparações entre os mesmos (DIAS, 1998).

2.8 Reamostragem

Reamostragem é uma abordagem paramétrica ou não-paramétrica que tem conquistado apoio por ser uma alternativa aos métodos clássicos utilizados de inferência paramétrica. Com sua utilização não é necessário assumir que o intervalo de confiança para um determinado parâmetro segue uma distribuição normal, por exemplo. Pode-se gerar intervalos de confiança para parâmetros, o que é difícil com a utilização de técnicas de inferência paramétrica.

De maneira geral, a reamostragem desconsidera a distribuição amostral assumida de uma estatística e calcula uma distribuição empírica, isto é, a real distribuição da estatística ao longo de milhares de amostras, essas que podem ser obtidas a partir da reamostragem dos dados.

2.9 Bootstrap

Bootstrap é um método de reamostragem computacionalmente intensivo que foi proposto por Efron (1979). Este método foi desenvolvido para fornecer medidas de intervalos de confiança, testes de hipóteses e situações inferenciais mais complexas (EFRON; TIBSHIRANI, 1993). O método bootstrap também pode ser utilizado para estimar o viés e a variância de estimadores.

Este método se baseia na construção de distribuições amostrais empíricas de uma estatística de interesse. Uma vantagem de utilizá-lo é a generalidade de sua aplicação, pois não é necessário se fazer suposições (FARIA, 2012).

O objetivo geral do método é fazer com que o pesquisador trate sua amostra como se ela fosse a população que deu origem aos dados e usar amostragem com reposição da amostra original para gerar pseudo-amostras de mesma dimensão. Por meio dessas pseudo-amostras, é possível estimar características da população tais como média, variância, percentis, entre outros (BARROS, 2005).

Os resultados obtidos por bootstrap são de grande valia, por exemplo, a distribuição empírica de uma estatística gerada pelo método possui aproximadamente a mesma forma e amplitude que a distribuição estatística amostral (FARIA, 2012).

O bootstrap pode ser utilizado tanto na estatística paramétrica quando na não-paramétrica, depende do conhecimento do problema. No primeiro caso, o método reamostra dados com reposição, de acordo com uma distribuição empírica estimada, o que é feito pelo fato de, em geral, não se conhecer a distribuição subjacente aos dados. No

segundo caso, quando se tem informações suficientes para afirmar algo sobre a distribuição dos dados, a amostra bootstrap é formada de maneira a realizar a amostragem diretamente nessa distribuição com parâmetros desconhecidos, substituídos por estimativas paramétricas. A distribuição da estatística de interesse aplicada aos valores da amostra bootstrap, condicional aos dados observados, é definida como a distribuição bootstrap dessa estatística (HONGYU, 2012).

Para que a aplicação do método resulte em estimativas confiáveis, devem ser feitas, a partir da amostra original centenas ou até milhares de reamostras de mesmo tamanho. Muitos autores aconselham a utilização de 1000 reamostras (FARIA, 2012).

3 MATERIAL E MÉTODOS

3.1 Análise de variância

A análise de variância é utilizada com o objetivo de comparar dois ou mais tratamentos. Considerando-se um delineamento inteiramente aleatorizado, cada observação Y_{ij} pode ser decomposta conforme o modelo a seguir:

$$Y_{ij} = \mu + t_i + \varepsilon_{ij} \quad i = 1, \dots, I \text{ e } j = 1, \dots, J ;$$

em que:

Y_{ij} é a observação do i -ésimo tratamento na j -ésima repetição;

μ é uma constante (costuma ser a média geral);

t_i é o efeito do i -ésimo tratamento;

ε_{ij} é o erro associado ao i -ésimo tratamento na j -ésima repetição, considerando-se $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$.

Como foi dito, o interesse é verificar se existe diferença entre os tratamentos em teste, o que equivale a testar as seguintes hipóteses (HONGYU, 2012):

$$\begin{cases} H_0 : t_1 = t_2 = \dots = t_I = 0 \\ H_1 : t_i \neq 0, \text{ para pelo menos algum } i \end{cases}$$

Observe que se H_0 for verdadeira, todos os tratamentos possuem a mesma média de tratamentos. A análise de variância é baseada na decomposição da variação total da variável resposta que, em partes, pode ser atribuída aos tratamentos e ao erro experimental. Essas variações são medidas por meio das somas de quadrados definidas a seguir:

$$SQTrat = \frac{1}{J} \sum_{i=1}^I Y_{i.}^2 - C \quad \quad SQTotal = \sum_{i=1}^I \sum_{j=1}^J Y_{ij}^2 - C$$

em que:

C é a correção para a média dada por $C = \frac{\left(\sum_{i=1}^I \sum_{j=1}^J Y_{ij} \right)^2}{IJ}$;
 $Y_{i.}$ é o total do i -ésimo tratamento dada por $Y_{i.} = \sum_{j=1}^J Y_{ij}$.

A soma de quadrados dos resíduos é obtida da seguinte maneira:

$$SQRes = SQTotal - SQTrat$$

Tem-se que a $SQTrat$ também é chama de “variação entre”, que é a variação entre os diferentes tratamentos e a $SQRes$ é chamada de “variação dentro” que é função das diferenças existentes entre as repetições de um mesmo tratamento.

Utilizando as somas de quadrados apresentadas, pode-se organizar uma tabela (Tabela 1). Para testar a hipótese H_0 , é utilizado o teste F de Snedecor. É importante lembrar que este teste é válido somente se os pressupostos assumidos para os erros do modelo forem satisfeitos (HONGYU, 2012).

Tabela 1 - Análise de variância de um experimento inteiramente aleatorizado

Fonte de Variação	Graus de Liberdade	Somas de Quadrados	Quadrado Médio	$F_{\text{calculado}}$
Tratamentos	$I - 1$	$SQTrat$	$QMTrat$	$\frac{QMTrat}{QMRes}$
Resíduo	$I(J - 1)$	$SQRes$	$QMRes$	
Total	$IJ - 1$	$SQTotal$		

em que:

$$QMTrat = \frac{SQTrat}{I - 1} \quad \text{e} \quad QMRes = \frac{SQRes}{I(J - 1)}$$

Pode-se mostrar que o quociente $\frac{QMTrat}{QMRes}$ segue uma distribuição F com $(I - 1)$ e $I(J - 1)$ graus de liberdade, supondo que y_{ij} são variáveis aleatórias independentes, todos os tratamentos possuem variâncias iguais a σ^2 e Y_{ij} segue distribuição normal com média μ_i e variância σ^2 . Por esses motivos, os pressupostos da análise de variância devem ser testados em qualquer análise.

Se $F_{\text{calculado}} > F_{\text{tabelado}}$ ao nível α de significância, rejeita-se a hipótese H_0 e conclui-se que pelo menos um contraste das médias de tratamentos difere ao nível α

de significância. Por outro lado, se $F_{calculado} < F_{tabelado}$, não rejeita-se a hipótese H_0 e conclui-se que não há evidências de diferença significativa entre os tratamentos ao nível α de significância, baseado na amostra obtida.

3.2 Análise de variância conjunta

Na pesquisa agrônômica existe a dificuldade de conduzir um experimento em um determinado local, com o objetivo de tirar conclusões generalizadas dos seus resultados para toda uma região, considerando-se que as condições climáticas variam bastante de um local para outro. Experimentos que envolvem vegetação, por exemplo, o solo e as condições climáticas têm grande influência sobre a produção (FERREIRA, 2000).

Para resolver este problema tem-se a análise conjunta de experimentos, que permite que o pesquisador generalize resultados de um experimento para toda uma região. O modelo para a análise estatística de um conjunto de observações provenientes de grupos de experimentos em blocos é (HONGYU, 2012).

$$Y_{ijk} = \mu + g_i + a_j + (ga)_{ij} + b_{k(j)} + \varepsilon_{ijk}$$

em que:

Y_{ijk} é o valor observado do i -ésimo genótipo no j -ésimo ambiente e no k -ésimo bloco, com

$i = 1, 2, \dots, g$; $j = 1, 2, \dots, a$ e $k = 1, 2, \dots, r$;

μ é uma constante (costuma ser a média geral);

g_i é o efeito do genótipo i ;

a_j é o efeito do ambiente j ;

$(ga)_{ij}$ é o efeito da interação específica do genótipo i com o ambiente j ;

$b_{k(j)}$ é o efeito do k -ésimo bloco dentro do j -ésimo ambiente;

ε_{ijk} é o erro experimental associado ao i -ésimo genótipo, no j -ésimo ambiente e no k -ésimo bloco, independente e com distribuição normal com média 0 e variância σ^2 .

Segue na Tabela 2 o quadro da análise de variância conjunta:

Tabela 2 - Análise de variância conjunta para experimentos de um mesmo grupo de genótipos avaliados em a ambientes com r blocos

Fonte de Variação	Graus de Liberdade	Quadrado Médio
Genótipos (G)	$g - 1$	QM_G
Ambientes (A)	$a - 1$	QM_A
Blocos dentro de ambientes (BdA)	$a(r - 1)$	$QM_{B.d.A.}$
Interação ($G \times A$)	$(g - 1)(a - 1)$	$QM_{G \times A}$
Resíduo (Res)	$a(g - 1)(r - 1)$	QM_{Res}
Total	$gar - 1$	

O modelo mais comum para a análise estatística de um conjunto de dados com essa estrutura é dado por (DUARTE; VENCOSKY, 1999):

$$Y_{ij} = \mu + g_i + a_j + (ga)_{ij} + \varepsilon_{ij}$$

em que:

Y_{ij} é o valor observado do i -ésimo genótipo no j -ésimo ambiente, com $i = 1, 2, \dots, g$ e $j = 1, 2, \dots, a$;

μ é uma constante (costuma ser a média geral);

g_i é o efeito do genótipo i ;

a_j é o efeito do ambiente j ;

$(ga)_{ij}$ é o efeito da interação específica do genótipo i com o ambiente j ;

ε_{ij} é o erro experimental médio, assumido independente e $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$.

O quadro da ANOVA para este modelo segue na Tabela 3, assumindo-se um delineamento experimento inteiramente aleatorizado:

Tabela 3 - Análise de variância para experimentos de um mesmo grupo de genótipos avaliados em a ambientes com r repetições

Fonte de Variação	Graus de Liberdade	Quadrado Médio
Genótipos (G)	$g - 1$	QM_G
Ambientes (A)	$a - 1$	QM_A
Interação ($G \times A$)	$(g - 1)(a - 1)$	$QM_{G \times A}$
Erro médio/ r	$ga(r - 1)$	QM_{EM}
Total	$gar - 1$	

A análise feita utilizando a média exige cálculo a parte da estimativa da

variância residual média (QM_{EM}), assumida comum a toda série de experimentos. Isto é feito obtendo-se a média ponderada dos quadrados médios residuais (QM_{Res}) obtida de todas as ANOVA's individuais de experimentos, em que os graus de liberdade são os pesos, isto é (HONGYU, 2012):

$$QM_{EM} = \frac{\sum_{j=1}^a SQ_{Res_j}}{\sum_{j=1}^a GL_{Res_j}}$$

3.3 Autovalores e autovetores

Definição 1: Seja V um espaço vetorial sobre $\mathbb{R}$ e seja $T : V \rightarrow V$ um operador linear. Um vetor $\underline{\mathbf{v}} \in V$, $\underline{\mathbf{v}} \neq \underline{\mathbf{0}}$, é um autovetor de T se existe um escalar $\lambda \in \mathbb{R}$ tal que $T(\underline{\mathbf{v}}) = \lambda \underline{\mathbf{v}}$. Neste caso, λ é um autovalor de T associado ao autovetor $\underline{\mathbf{v}}$.

Definição 2: Seja V um espaço vetorial de dimensão finita, T um operador linear em V e B uma base qualquer de V . O polinômio $P_t(\lambda) = \det([T]_B - \lambda \mathbf{I})$ é chamado o polinômio característico de T .

Observação: Pode-se mostrar que λ é um autovalor de T se, e somente se, $P_T(\lambda) = 0$.

Definição 3: A raiz quadrada de um autovalor λ é chamada de valor singular. Nesse texto será denotado por $\sqrt{\lambda}$.

3.4 Decomposição em valores singulares

Teorema: (*Teorema da decomposição em valores singulares*) Seja $\mathbf{A}_{m \times k}$ uma matriz de números reais. Então existe uma matriz ortogonal $\mathbf{U}_{m \times m}$ com suas colunas formadas por vetores singulares à esquerda de $\mathbf{A}\mathbf{A}^T$; $\mathbf{V}_{k \times k}$ é uma matriz ortogonal com suas colunas formadas por vetores singulares à direita de $\mathbf{A}^T\mathbf{A}$ tal que (JOHNSON; WICHERN, 1992):

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

em que a matriz $\mathbf{\Sigma}$ é uma matriz diagonal formada pelos valores singulares de $\mathbf{A}\mathbf{A}^T$ ou $\mathbf{A}^T\mathbf{A}$, isto é, uma matriz diagonal com valores reais e não negativos.

A decomposição em valores singulares também pode ser expressa como uma

matriz de expansão que depende do posto r de $\mathbf{A}$. Especificamente, existem r constantes positivas $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_r}$; r vetores ortonormais (de comprimento unitário e ortogonal) $m \times 1$ $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ e r vetores ortonormais $k \times 1$ $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$ tais que:

$$\mathbf{A} = \sum_{i=1}^r \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$$

em que $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$, $\mathbf{V}_r = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$ e $\mathbf{\Sigma}_r$ é uma matriz diagonal $r \times r$, com $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_r}$ em sua diagonal.

3.5 Análise de componentes principais

Matematicamente, a técnica consiste basicamente em transformar um conjunto original de variáveis $X_1, X_2, \dots, X_p$ e encontrar combinações lineares dessas e produzir outro conjunto de mesma dimensão $Z_1, Z_2, \dots, Z_p$ que sejam não correlacionados na ordem de sua importância ($Var[Z_1] \geq Var[Z_2] \geq \dots \geq Var[Z_p]$) e que descreva a variação dos dados. Os índices de Z representam os componentes principais.

Quando as variáveis são não correlacionadas, não é aconselhável utilizar a ACP, pois a análise é altamente satisfatória no caso em que as variáveis originais são altamente correlacionadas, positiva ou negativamente (MANLY, 2008).

Como foi dito anteriormente, o procedimento para a ACP inicia-se com dados de p variáveis para n indivíduos. O primeiro componente principal é a combinação linear das variáveis $X_1, X_2, \dots, X_p$, isto é,

$$Z_1 = c_{11}X_1 + c_{12}X_2 + \dots + c_{1p}X_p$$

com a seguinte restrição:

$$c_{11}^2 + c_{12}^2 + \dots + c_{1p}^2 = 1$$

Deste modo, $Var[Z_1]$ é a maior possível sobre a restrição que foi imposta. O segundo componente principal é escolhido de tal modo que $Var[Z_2]$ seja a maior possível:

$$Z_2 = c_{21}X_1 + c_{22}X_2 + \dots + c_{2p}X_p$$

com a restrição:

$$c_{21}^2 + c_{22}^2 + \dots + c_{2p}^2 = 1$$

e deve-se ter também que Z_1 e Z_2 sejam não correlacionadas. As combinações lineares são obtidas assim sucessivamente até o p -ésimo componente principal.

Uma ACP envolve encontrar os autovalores de uma matriz de covariâncias amostral. Uma matriz de covariâncias tem a seguinte forma:

$$\mathbf{\Sigma} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

em que os elementos da diagonal principal a_{ii} correspondem à variância da variável X_i e o restante dos elementos a_{ij} da matriz tratam da covariância entre as variáveis X_i e X_j . Tem-se também que as variâncias dos componentes principais correspondem aos autovalores da matriz $\mathbf{\Sigma}$. Vale salientar que em uma matriz de variâncias e covariâncias não é possível obter autovalores negativos, ou seja, os autovalores se comportam de tal modo que:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0,$$

ou seja, o i -ésimo λ corresponde ao i -ésimo componente principal.

Na parte final da ACP deve-se evitar componentes que expliquem uma pequena parcela de todo o conjunto. Por exemplo, podemos ter 15 componentes principais, porém os três primeiros explicam 95% da variância total, neste âmbito, utiliza-se apenas os três primeiros componentes para análise, já que estes explicam grande parte da variância total das variáveis originais.

3.6 A representação gráfica biplot

Seja a seguinte aproximação DVS para a matriz de interações $\mathbf{GA}$ de posto conhecido p (DUARTE; VENCOSKY, 1999):

$$\widehat{\mathbf{GA}} = \sum_{i=1}^n \sqrt{\lambda_i} \hat{\boldsymbol{\gamma}}_i \hat{\boldsymbol{\alpha}}_i^T = \tilde{\mathbf{U}} \tilde{\mathbf{\Sigma}} \tilde{\mathbf{V}}^T$$

em que:

$\tilde{\mathbf{U}}$ é a matriz cujas colunas possuem os n primeiros vetores $\boldsymbol{\gamma}$;

$\tilde{\Sigma}$ é a matriz diagonal com os primeiros valores singulares;

$\tilde{\mathbf{V}}^T$ é a matriz cujas linhas possuem os n primeiros vetores $\boldsymbol{\alpha}^T$;

ou, de forma equivalente pode-se escrever (DUARTE; VENCovsky, 1999):

$$\widehat{\mathbf{GA}}_{g \times a} = \overbrace{\left(\tilde{\mathbf{U}} \tilde{\Sigma}^{\frac{1}{2}} \right)}^{\tilde{\mathbf{G}}} \underbrace{\left(\tilde{\Sigma}^{\frac{1}{2}} \tilde{\mathbf{V}}^T \right)}_{\tilde{\mathbf{H}}}$$

em que:

$\tilde{\mathbf{G}}$ possui g vetores denominados marcadores de linhas, cada um com n componentes (coordenadas dos genótipos para cada eixo selecionado);

$\tilde{\mathbf{H}}$ possui a marcadores de colunas, cada um com n componentes.

Como os eixos singulares são ortogonais, pode-se plotar os g genótipos e os a ambientes em um eixo cartesiano com n eixos (geralmente esse n é um, dois ou três). Multiplicando-se uma linha i de $\tilde{\mathbf{G}}$ por uma coluna j de $\tilde{\mathbf{H}}$ tem-se a estimativa AMMI_n para a interação do genótipo i com o ambiente j .

Com isso, pode-se verificar interações positivas entre genótipos e ambientes (isso ocorre quando as coordenadas/escores possuem o mesmo sinal), assim como é possível a verificação de interações negativas (quando as coordenadas/escores são negativas), o que facilita na recomendação de cultivares, já que existe a possibilidade de fazer a melhor/pior combinação possível entre genótipos e ambientes.

É importante enfatizar que o termo “biplot” não é utilizado em qualquer modelo estatístico em particular. É um tipo de gráfico que possui dois marcadores que, neste âmbito foram adotados como sendo um marcador para genótipos e o outro para ambientes. Embora a maioria dos biplots encontrados em artigos sejam de duas dimensões, isso não denomina o nome biplot, mas sim os dois tipos de marcadores que, na prática, pode ser representado em uma, duas ou três dimensões (GAUCH JÚNIOR, 1992).

3.7 Modelo AMMI

A resposta média sobre r repetições ou blocos de um genótipo i em um ambiente j é dada por (DUARTE; VENCovsky, 1999):

$$Y_{ij} = \mu + g_i + a_j + \sum_{k=1}^n \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk} + \rho_{ij} + \varepsilon_{ij} \quad (1)$$

Note que $(ga)_{ij}$ que estava presente no modelo de análise de variância conjunta foi modelado por:

$$\sum_{k=1}^n \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk} + \rho_{ij}$$

e sob as seguintes restrições:

$$\sum_i g_i = \sum_j a_j = \sum_i (ga)_{ij} = \sum_j (ga)_{ij} = 0$$

Os demais termos do modelo resultam da decomposição em valores singulares da matriz de interações $\mathbf{GA}_{(g \times a)} = [(\widehat{ga})_{ij}]$, com $(\widehat{ga})_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$, em que:

Y_{ij} é o valor observado do i -ésimo genótipo no j -ésimo ambiente, com $i = 1, 2, \dots, g$ e $j = 1, 2, \dots, a$;

$\bar{Y}_{i.}$ é a média do i -ésimo genótipo dada por $\bar{Y}_{i.} = \frac{1}{a} \sum_{j=1}^a Y_{ij}$;

$\bar{Y}_{.j}$ é a média do j -ésimo ambiente dada por $\bar{Y}_{.j} = \frac{1}{g} \sum_{i=1}^g Y_{ij}$;

$\bar{Y}_{..}$ é a média geral dada por $\bar{Y}_{..} = \frac{\sum_{i,j} Y_{ij}}{ga}$.

A ilustração das partes aditiva e multiplicativa segue abaixo:

$$Y_{ij} = \underbrace{\mu + g_i + a_j}_{\text{parte aditiva}} + \underbrace{\sum_{k=1}^p \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk} + \varepsilon_{ij}}_{\text{parte multiplicativa}}$$

A parte multiplicativa pode ser desdobrada como segue abaixo:

$$Y_{ij} = \mu + g_i + a_j + \underbrace{\sum_{k=1}^n \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk}}_{\text{padrão}} + \underbrace{\sum_{k=n+1}^p \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk} + \varepsilon_{ij}}_{\text{ruído}}$$

em que: $n < p$ e $\sum_{k=n+1}^p \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk} = \rho_{ij}$.

Da equação (1) apresentada anteriormente, tem-se que:

Y_{ij} é a resposta média do genótipo i no ambiente j ;

μ é uma constante (costuma ser a média geral);

g_i é o efeito fixo do genótipo i ;

a_j é o efeito fixo do ambiente j ;

$\sqrt{\lambda_k}$ é o k -ésimo valor singular das matrizes $(\mathbf{GA})(\mathbf{GA})'$ e $(\mathbf{GA})'(\mathbf{GA}) \Rightarrow \lambda_k$ é o k -ésimo autovalor;

γ_{ik} é o i -ésimo elemento associado ao genótipo i do k -ésimo autovetor de $(\mathbf{GA})(\mathbf{GA})'$ associado a λ_k ;

α_{jk} é o j -ésimo elemento associado ao ambiente j do k -ésimo autovetor de $(\mathbf{GA})'(\mathbf{GA})$ associado a λ_k ;

O índice k ($k = 1, \dots, p$ em que $p = \min(g - 1, a - 1)$) é o posto da matriz $\mathbf{GA}$, tomando até n no somatório ($n < p$), determinando uma aproximação de mínimos quadrados para

$\mathbf{GA}$ pelos n primeiros termos da decomposição de valores singulares; n é o número de eixos (componentes principais) selecionados, que descrevem o padrão da interação $G \times A$;

ρ_{ij} é o ruído associado ao termo $(ga)_{ij}$ da interação do genótipo i com o ambiente j ;

ε_{ij} é o erro experimental médio.

Por meio da metodologia AMMI, busca-se recuperar a parcela fortemente determinada dos genótipos e dos ambientes da $SQ_{G \times A}$. Pelas propriedades citadas na DVS, tem-se que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$, isto é, as primeiras parcelas são as que possuem as maiores proporções da variação na matriz $\mathbf{GA}$.

Por este motivo, o método AMMI é visto como um modelo que separa a parte padrão (lei geral da formação que comanda os primeiros eixos da ACP ou DVS) do

ruído (parcela não fortemente determinada por genótipos e ambientes que será descartada com o objetivo de melhorar a capacidade preditiva do modelo) na análise da $SQ_{G \times A}$. Assim, a interação do genótipo i com o ambiente j é descrita por:

$$\sum_{k=1}^n \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk}$$

descartando-se o ruído $\rho_{ij} = \sum_{k=n+1}^p \sqrt{\lambda_k} \gamma_{ik} \alpha_{jk}$, buscando-se melhorar a capacidade preditiva do modelo.

Resumidamente, as vantagens do método AMMI são: análise detalhada da interação $G \times A$; seleção das melhores combinações entre genótipos e ambientes visando ganhos na produtividade; estimativas precisas das respostas genotípicas; fácil interpretação gráfica utilizando o biplot.

3.7.1 Ajuste dos efeitos principais por análise de variância

A primeira etapa para se realizar a análise AMMI, é realizar uma análise de variância comum aplicada aos dados (o modelo já foi apresentado anteriormente). As soluções dos mínimos quadrados ordinários para o sistema de equações normais sob as restrições $\sum_i g_i = \sum_j a_j = 0$ são dados por (DUARTE; VENCOSKY, 1999):

$$\hat{\mu} = \bar{Y}_{..} \quad \hat{g}_i = \bar{Y}_{i.} - \bar{Y}_{..} \quad \hat{a}_j = \bar{Y}_{.j} - \bar{Y}_{..}, \quad \text{em que:}$$

$$\bar{Y}_{..} = \frac{\sum_{i,j} Y_{ij}}{ga}$$

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} \quad \Leftrightarrow \quad (\widehat{ga})_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

A partir deste resultado, é possível construir a matriz de interações mostrada a seguir ($\mathbf{GA}_{(g \times a)}$), objeto de estudo na segunda etapa da análise:

$$\begin{bmatrix} (\widehat{ga})_{11} & (\widehat{ga})_{12} & \cdots & (\widehat{ga})_{1a} \\ (\widehat{ga})_{21} & (\widehat{ga})_{22} & \cdots & (\widehat{ga})_{2a} \\ \vdots & \vdots & \ddots & \vdots \\ (\widehat{ga})_{g1} & (\widehat{ga})_{g2} & \cdots & (\widehat{ga})_{ga} \end{bmatrix}$$

3.7.2 Testes estatísticos para a seleção do modelo AMMI

Métodos que verifiquem o número de eixos a serem retidos na modelagem AMMI de tal modo que expliquem o padrão relacionado à interação é foco de pesquisas há anos. Todos esses métodos são semelhantes à ACP, isto é, levam em consideração a proporção da $SQ_{G \times A}$ acumulada até o n -ésimo eixo.

Um dos procedimentos mais utilizados em pesquisas é determinar o número de graus de liberdade associados à cada parcela desdobrada da $(SQ_{G \times A})$. Após isso ser feito, é necessário um teste F avaliando a significância dos eixos, com o objetivo de verificar quais serão retidos. Com esse procedimento, será obtida uma tabela semelhante à de análise de variância tradicional, a diferença é que na parte da interação, haverá um desdobramento pelo número de eixos. O ponto de parada que determina a seleção do modelo $(AMMI_0, AMMI_1, \dots, AMMI_n)$ é baseado na significância do teste F para os eixos da interação (HONGYU, 2012).

Existem vários métodos para atribuir graus de liberdade para um modelo AMMI (GAUCH & ZOBEL, 1996), os mais conhecidos são o de Gollob (1968) e o de Mandel (1971). Vários autores alertam que não existe uma boa concordância entre os métodos, a escolha entre um desses acaba recaindo sobre considerações de natureza teórica e prática.

O método de Gollob é bastante utilizado pelo fato de ser extremamente fácil de ser aplicado por envolver apenas as mais simples operações aritméticas, isto é:

$$GL_{IPCA_n} = g + a - 1 - 2n$$

em que $IPCA_n$ indica o n -ésimo componente da interação.

Por outro lado, Mandel parte de uma definição geral de graus de liberdade associados a cada parcela λ_k que envolve esperança matemática e variância populacional é:

$$GL_{IPCA_n} = \frac{E(\lambda_k)}{\sigma^2}$$

Dessa expressão foram feitas simulações que resultaram em três tabelas que estão relacionadas ao número de graus de liberdade, essas que exigem regras especiais para consulta, razão pela qual reduziu sua aplicação em pesquisas. Comparando-se os métodos de Gollob

e Mandel, tem-se que o sistema de Mandel na maior parte das vezes resulta em uma diminuição não linear dos graus de liberdade para sucessivos termos da interação e também, os graus de liberdade podem ser fracionários (ARAÚJO, 2008).

A metodologia de Mandel perdurou como sendo a melhor por muitos anos. Atualmente essa metodologia tem recebido várias críticas consistentes que a tornou menos indicada que a metodologia de Gollob. As críticas são sobre as pressuposições que são feitas por Mandel em suas simulações. Por esse motivo, a metodologia de Gollob tem sido a mais utilizada.

Segundo Gauch Júnior (1992), o sistema de Gollob faz o uso da definição clássica para graus de liberdade, isto é, o número de parâmetros estimados menos o número de restrições. Do modelo AMMI, cada eixo de interação possui um valor singular σ_k , g elementos para o vetor γ_k e a elementos para o vetor α_k , o que totaliza $g + a + 1$ parâmetros estimados (DUARTE; VENCOSKY, 1999).

Após retirar-se os efeitos principais, os autovetores ficam sujeitos às seguintes restrições: $\sum_i \gamma_{ik} = 0$ e $\sum_j \alpha_{jk} = 0$. Ortonormalizando os vetores, tem-se também que: $\sum_i \gamma_{ik}^2 = 1$ e $\sum_j \alpha_{jk}^2 = 1$. Após esses procedimentos, necessita-se que cada eixo $IPCA_n$ seja ortogonal a todos os outros $n - 1$ eixos anteriores. Com isso, existem $2(n - 1)$ restrições adicionais $\sum_i \gamma_{in} \gamma_{ik} = 0$ e $\sum_j \alpha_{jn} \alpha_{jk} = 0$, $k = 1, 2, \dots, n - 1$. Deste modo, os graus de liberdade associados a cada eixo são como segue:

$$GL_{IPCA_k} = g + a + 1 - [2 + 2 + 2(k - 1)] = g + a - 1 - 2k$$

Fazendo-se uma tabela na estrutura de análise de variância utilizando a metodologia de Gollob, tem-se:

Tabela 4 - Desdobramento da interação $G \times A$ pelo método de Gollob

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Genótipos (G)	$g - 1$	SQ_G
Ambientes (A)	$a - 1$	SQ_A
Interação $G \times A$	$(g - 1)(a - 1)$	$SQ_{G \times A}$
IPCA ₁	$g + a - 1 - (2 \times 1)$	λ_1
IPCA ₂	$g + a - 1 - (2 \times 2)$	λ_2
$\vdots$	$\vdots$	$\vdots$
IPCA _{n}	$g + a - 1 - (2 \times p)$	λ_n
Erro médio	$ga(r - 1)$	SQ_{Erro}
Total	$gar - 1$	SQ_{Total}
IPCA _{k} é o k -ésimo componente da interação.		

Piepho (1995) investigou a robustez de testes utilizados para a seleção do modelo AMMI, quanto às suposições de homocedasticidade e normalidade dos erros. Segundo Piepho (1995), o teste F aplicado no método de Gollob é liberal no fato de selecionar mais termos multiplicativos do que o verdadeiro modelo contempla. O teste considerado mais robusto é o F_R proposto por Cornelius et al. (1992). Piepho recomenda que se for utilizado outros testes para análise, deve-se verificar a validade das suposições.

O teste F_R tem como hipótese nula o fato de que não exista mais do que n termos determinando a interação. Essa estatística F_R tem uma distribuição F aproximada com $f_2 = (g - n - 1)(a - n - 1)$ graus de liberdade. O teste estatístico F_R com n termos multiplicativos é:

$$F_R = \frac{SQ_{G \times A} - \sum_{i=1}^n \lambda_i^2}{f_2 \cdot QM_{\text{Erro médio}}},$$

Se o teste F_R for significativo, tem-se que pelo menos um termo multiplicativo deve ser adicionado aos n que já foram ajustados. De um modo geral, tem-se que o teste F_R pode ser considerado um teste de significância dos $n + 1$ primeiros termos da interação. Quando nenhum termo multiplicativo for ajustado ($n = 0$) o teste é equivalente ao teste F para a interação $G \times A$ na análise de variância (ARAÚJO, 2008). Os graus de liberdade do numerador de F_R são os mesmos que os graus de liberdade para toda a interação menos os graus de liberdade atribuídos por Gollob para os n primeiros termos (DUARTE; VENCovsky, 1999).

3.8 Determinação do número de grupos no dendrograma

Para determinar o ponto de corte no dendrograma e definir o número de grupos a ser considerado, Mojena (1977) sugeriu um procedimento baseado nas distâncias presentes no dendrograma. De maneira geral, o procedimento tem por objetivo selecionar o número de grupos no passo j que, primeiramente, satisfizer a inequação (SILVA, 2012):

$$\alpha_j > \theta_k$$

em que: α_j é o valor da distância do nível de fusão correspondente ao passo j ($j = 1, 2, \dots, g - 1$), θ_k é o referencial de corte dado por:

$$\theta_k = \bar{\alpha} + k\hat{\sigma}_{\alpha}$$

em que: $\bar{\alpha}$ e $\hat{\sigma}_\alpha$ são, respectivamente, a média e o desvio padrão dos valores de α ; k é uma constante cujo valor a ser adotado é de $k = 1,25$ como regra de parada para a definição do número de grupos a ser considerado (MILLIGAN; COOPER, 1985). Deste modo, tem-se que:

$$\bar{\alpha} = \frac{1}{g-1} \sum_{j=1}^{g-1} \alpha_j \quad \text{e} \quad \hat{\sigma}_\alpha = \sqrt{\frac{\sum_{j=1}^{g-1} \alpha_j^2 - \frac{1}{g-1} \left(\sum_{j=1}^{g-1} \alpha_j \right)^2}{g-2}}$$

3.9 Métodos de agrupamento

3.9.1 Procedimento para o método do vizinho mais próximo

1. De posse da matriz de dados, calcula-se a matriz de distâncias euclidianas ($\mathbf{D}$), em que d_{ij} são os elementos da matriz;
2. Localiza-se o valor mínimo d_{ij} presente na matriz $\mathbf{D}$. Os elementos correspondentes a este valor, serão reunidos em um grupo, sobrando $n - 1$ agrupamentos a serem formados;
3. Baseando-se na matriz $\mathbf{D}$, determina-se a distância entre o novo agrupamento e os demais elementos presentes na matriz, por meio da relação: $d_{(i,j),I} = \min(d_{i1}, d_{i2}), I = 1, \dots, n - 2$ e constrói-se uma nova matriz de distâncias $\mathbf{D}'$;
4. Localiza-se em $\mathbf{D}'$ o menor valor d_{ij} e, em seguida, agrupa-se os elementos correspondentes a essa distância, formando-se um novo agrupamento;

O processo é repetido até que todos os objetos sejam unidos a um único agrupamento.

3.9.2 Procedimento para o método do vizinho mais distante

1. De posse da matriz de dados, calcula-se a matriz de distâncias euclidianas ($\mathbf{D}$), em que d_{ij} são os elementos da matriz;
2. Localiza-se o valor mínimo d_{ik} presente na matriz $\mathbf{D}$. Os elementos correspondentes a este valor, serão reunidos em um grupo, sobrando $n - 1$ agrupamentos a serem formados;

3. Baseando-se na matriz $\mathbf{D}$, determinou-se a distância entre o novo agrupamento e os demais elementos presentes na matriz, por meio da relação: $d_{(i,j),I} = \max(d_{i1}, d_{i2})$, $I = 1, \dots, n - 2$ e constrói-se uma nova matriz de distâncias $\mathbf{D}^*$;
4. Localiza-se em $\mathbf{D}^*$ o maior valor d_{ij} e, em seguida, agrupa-se os elementos correspondentes a essa distância, formando-se um novo agrupamento;

O processo é repetido até que todos os objetos sejam unidos a um único agrupamento.

3.9.3 Procedimento para o método da ligação média

Este método é baseado na média das distâncias entre todos os indivíduos. Considere G_1 e G_2 dois grupos de indivíduos e $d[G_1, G_2]$ a distância entre eles, sendo assim, esse método pode ser representado por:

$$d[G_1, G_2] = \sum_{i \in G_1} \sum_{j \in G_2} \frac{d_{ij}}{g_1 g_2}$$

3.9.4 Procedimento para o método do centroide

Este método é baseado na distância entre os centroides dos grupos G_1 e G_2 , em outras palavras “centroide” significa que são definidos como a média das coordenadas de todos os objetos de um mesmo grupo (JOHNSON; WICHERN, 1992).

3.9.5 Procedimento para o método da mediana

Trata-se de um caso particular do método do centroide, isto é, quando determina-se a distância entre agrupamentos utilizando-se o centro de massa, não é considerado o número de elementos que existem em cada um dos agrupamentos, deste modo, o vetor médio que representa o novo agrupamento, pode ficar situado entre os elementos do agrupamento com maior número de elementos. O método da mediana surgiu para ponderar a medida de distância pelo número de elementos de cada agrupamento. Segundo Barroso & Artes (2003), este método é uma modificação do método do centroide para a independência da distância do tamanho dos grupos.

3.9.6 Procedimento para o método de Ward

É baseado no fato de que, a alocação de um elemento a um grupo é feita maximizando a homogeneidade dentro dos grupos ou minimizando o total das somas de

quadrados dentro dos grupos (soma de quadrados dos erros), calculada por:

$$\sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})'(\mathbf{y}_i - \bar{\mathbf{y}})$$

em que $\mathbf{y}_i$ é o vetor multivariado de medidas associado ao i -ésimo objeto e $\bar{\mathbf{y}}$ é a média de todos os itens (SARTORIO, 2008).

3.10 Coeficiente de correlação cofenética como critério para avaliar métodos de agrupamento

Após terem sido aplicados os métodos de agrupamentos hierárquicos, deve-se obter o coeficiente de correlação cofenética entre a matriz de distâncias originais e a matriz cofenética. Este coeficiente quantifica a semelhança entre a matriz de distâncias e a matriz cofenética; quanto maior for o valor da correlação, menor será a distorção provocada pelo agrupamento (SILVA, 2012). Segundo Manly (2008), o valor da correlação entre as duas matrizes pode ser testado pela aplicação do teste de aleatorização de Mantel (1967). A correlação cofenética é calculada da seguinte maneira (BUSSAB et al., 1990):

$$r_{cof} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c})(d_{ij} - \bar{d})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (d_{ij} - \bar{d})^2}},$$

em que:

c_{ij} é o valor de dissimilaridade entre os indivíduos i e j , obtidos a partir da matriz cofenética;

d_{ij} é o valor de dissimilaridade entre os indivíduos i e j , obtidos a partir da matriz dissimilaridade;

n é a dimensão da matriz;

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n c_{ij} \quad \text{e} \quad \bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij}$$

3.11 Estabilidade via método bootstrap

Para verificar a estabilidade dos métodos em análise de agrupamentos, utilizou-se a análise via reamostragem “bootstrap” e comparou-se os resultados obtidos com a análise AMMI, verificando se havia coerência a formação dos grupos. Para a análise de agrupamentos via bootstrap, foram considerados os seguintes passos (ALBUQUERQUE, 2006):

1. Foi considerada a matriz de dados original, **GA** (que é a matriz de interação entre genótipos e ambientes utilizada na metodologia AMMI);

$$\mathbf{GA} = \begin{bmatrix} (\widehat{ga})_{11} & (\widehat{ga})_{12} & \cdots & (\widehat{ga})_{1a} \\ (\widehat{ga})_{21} & (\widehat{ga})_{22} & \cdots & (\widehat{ga})_{2a} \\ \vdots & \vdots & \ddots & \vdots \\ (\widehat{ga})_{g1} & (\widehat{ga})_{g2} & \cdots & (\widehat{ga})_{ga} \end{bmatrix}$$

2. Sobre a matriz original do item 1. encontrou-se a matriz de distâncias euclidianas entre os g genótipos para posterior aplicação dos algoritmos de agrupamento;
3. Após o cálculo da matriz de distâncias, aplicou-se os métodos de agrupamento;
4. Aplicou-se reamostragem “bootstrap” (10.000) sobre a matriz **GA** e repetiu-se os passos 2. e 3.;
5. Aplicação do coeficiente de correlação cofenética para analisar o grau de associação entre os métodos para os resultados obtidos antes da reamostragem e após a reamostragem “bootstrap”;
6. Mostrou-se a distribuição empírica dos coeficientes de correlação cofenética obtidos no item 5.

4 RESULTADOS E DISCUSSÃO

4.1 Adaptabilidade e estabilidade de cultivares de soja por meio da modelagem AMMI

Os dados utilizados neste trabalho foram provenientes do departamento da genética da Escola Superior de Agricultura “Luiz de Queiroz” - ESALQ pelo professor José Baldin Pinheiro. O experimento foi realizado visando a produtividade de soja provenientes de 5 ambientes (Goiânia, Itumbiara, Santa Helena, Rio Verde e Jataí) com 24 genótipos, 4 repetições e no delineamento aleatorizado em blocos. A safra foi dos anos de 2003 e 2004.

Os dados referentes à produtividade foram inicialmente submetidos à análise de variância individual por local e, posteriormente, à análise de variância conjunta em que foram considerados todos os locais para se avaliar a variabilidade genética entre os tratamentos e a presença de interação $G \times A$. Após verificar a presença da interação, foi realizada a análise da adaptabilidade e estabilidade fenotípica via modelagem AMMI.

Os resultados da análise de variância conjunta com relação à produtividade de soja (Tabela 4) mostraram que os efeitos dos genótipos, ambientes e da interação $G \times A$ foram significativos ao nível de 1%, isto é, os genótipos apresentaram comportamento diferente em cada um dos ambientes, sendo assim, foi feito um estudo mais aprofundado sobre a interação $G \times A$ para identificar a magnitude da mesma (KLEFENS, 2009).

Tabela 5 - Análise de variância conjunta dos dados de produtividade de soja e decomposição das somas de quadrados da interação $G \times A$

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor de p
Ambientes	4	6.374.608	1.593.652	0,000004178***
Repetições	15	1.097.361	73.157	0,0021062**
Genótipos	23	2.652.777	115.338	0,00000002042***
Genótipos $\times$ Ambientes	92	4.693.037	51.011	0,0003429***
IPCA ₁	26	2.902.487,3	111.634,13	0
IPCA ₂	24	752.867,3	31.369,47	0,401
IPCA ₃	22	597.574,8	27.162,49	0,5819
IPCA ₄	20	440.108	22.005,4	0,795
Resíduo	345	10.332.354	29.949	-

Teste F de Gollob (1968).

*** : significativo ao nível de probabilidade, $p < 0,001$.

** : significativo ao nível de probabilidade, $p < 0,01$.

Observando na análise conjunta de variância, temos que a $SQ_{G \times A} = 4.693.037$ corresponde exatamente à soma dos autovalores $\left(\sum_{k=1}^n \lambda_k\right)$, que pode estar inflacionada por conta da presença de ruídos na variável resposta; por este motivo é necessário um ajuste da interação por meio da decomposição em valores singulares aplicada à matriz de interações **GA**. Essa matriz tem posto 5 e, portanto, pode ser decomposta em 5 componentes principais. Avaliando os modelos AMMI ($AMMI_0$, $AMMI_1$, $AMMI_2$, $AMMI_3$ e $AMMI_4$) verificou-se pelo teste F que o único modelo significativo foi o $AMMI_1$ (Tabela 4).

Os genótipos mais produtivos (em ordem decrescente) foram G23, G2, G21, G14, G18 e G24. Percebe-se também que os ambientes que tiveram maior produtividade foram A4, seguido de A3 (Figura 1). Por outro lado, os genótipos que tiveram menor produtividade foram G9, G5, G19 e G4 (em ordem crescente).

Os genótipos mais estáveis foram G19, G6, G13, G22, G24, G14, G23 estes, que podem ser cultivados em quaisquer ambientes que foram estudados neste experimento. Os genótipos G23, G22, G24 e G14 foram estáveis e tiveram a produtividade acima da média, o que indica que esses genótipos relacionam produtividade com estabilidade (Figura 1). É importante observar que alguns genótipos que apresentaram produtividade acima da média não foram necessariamente os mais estáveis, como por exemplo, os genótipos G2 e G21.

O ambiente A4 é o mais produtivo, porém nem todos os genótipos se adaptam bem a ele, já que não se trata de um ambiente estável. Os genótipos G9, G8, G16 e G2 têm uma adaptação específica com este ambiente (Figura 2).

Os genótipos G19, G14 e G23 têm adaptabilidade específica aos ambientes A3 e A2, assim como os genótipos G18, G17, G12, G21 e G1 têm adaptabilidade específica aos ambientes A1 e A5 (Figura 2).

Nota-se que os ambientes A1 e A5 tiveram comportamentos semelhantes, assim como A2 e A3. Isto é importante de se observar, já que existem ocasiões em que é necessário substituir um dos locais por outro mais representativo da região.

Outro fato importante de se observar é que, embora o ambiente A4 seja o mais produtivo, ele se difere completamente dos outros ambientes estudados, não podendo ser substituído por nenhum outro.

Figura 1 - Biplot AMMI₁ (PC₁ × Média) para os dados de produtividade de soja

Figura 2 - Biplot AMMI₂ (PC₁ × PC₂) para os dados de produtividade de soja

4.2 Análise de agrupamentos sobre a matriz de distâncias original

Utilizando a matriz de interações **GA**, aplicou-se o algoritmo para encontrar a matriz de distâncias euclidianas e logo após, foram aplicados os métodos de agrupamento apresentados anteriormente (vizinho mais próximo, vizinho mais distante, centroide, mediana, Ward e ligação média entre grupos) e, por fim, foram obtidos os dendrogramas apresentados nas figuras de 3 a 8.

Embora as estruturas obtidas pelos agrupamentos tenham sido similares, pode-se observar que houve divergência entre alguns dos agrupamentos formados. Em outras palavras, uma estrutura de agrupamento obtida em um dos métodos pode ser completamente diferente da estrutura obtida por outro método. Essa ideia confere com a afirmação feita por Johnson & Wichern (1992), que dificilmente os dendrogramas obtidos por diferentes métodos de agrupamentos tenham semelhança (ALBUQUERQUE, 2005).

Figura 3 - Dendrograma obtido pelo emprego do método do vizinho mais próximo com base na distância euclidiana da matriz de interações

Figura 4 - Dendrograma obtido pelo emprego do método do vizinho mais distante com base na distância euclidiana da matriz de interações

Figura 5 - Dendrograma obtido pelo emprego do método do centróide com base na distância euclidiana da matriz de interações

Figura 6 - Dendrograma obtido pelo emprego do método da mediana com base na distância euclidiana da matriz de interações

Figura 7 - Dendrograma obtido pelo emprego do método de Ward com base na distância euclidiana da matriz de interações

Figura 8 - Dendrograma obtido pelo emprego do método da ligação média entre grupos com base na distância euclidiana da matriz de interações

Analisando cada um dos métodos de agrupamentos, percebe-se que em praticamente todos, os genótipos G2 e G19 ou formaram grupos individuais ou formaram juntamente um grupo. Comparando esse resultado com a metodologia AMMI aplicada anteriormente, faz total sentido principalmente para o genótipo G19 que, como foi dito, é um genótipo estável mas que não é semelhante aos outros, ou seja, com relação aos genótipos aos quais este está sendo comparado, não é um genótipo simples de ser substituído.

4.3 Análise de agrupamentos sobre a matriz de distâncias euclidianas obtida via bootstrap

Como foi apresentado anteriormente, encontrou-se a matriz de interações **GA** a partir dos dados originais, sobre a mesma, aplicou-se a metodologia AMMI e, em seguida, foi calculada a matriz de distâncias euclidianas para aplicação dos métodos de agrupamentos. Feito isso, aplicou-se bootstrap na matriz de interações **GA**, calculou-se a média das matrizes obtidas e o processo foi repetido.

Os grupos presentes aos dendrogramas obtidos a partir de 10.000 reamostragens foram obtidos utilizando o método de Mojena (1977) para a definição do ponto de corte e seguem nas figuras de 9 a 14.

Figura 9 - Dendrograma obtido pelo emprego do método do vizinho mais próximo com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

Figura 10 - Dendrograma obtido pelo emprego do método do vizinho mais distante com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

Figura 11 - Dendrograma obtido pelo emprego do método do centroide com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

Figura 12 - Dendrograma obtido pelo emprego do método da mediana com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

Figura 13 - Dendrograma obtido pelo emprego do método de Ward com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

Figura 14 - Dendrograma obtido pelo emprego do método da ligação média entre grupos com base na matriz de distâncias euclidianas via bootstrap da matriz de interações

4.4 Correlação cofenética

Os valores das correlações cofenéticas (Tabela 5) foram, em sua maioria elevados (maiores que 0,7) tanto para os dados originais quanto para os dados após a reamostragem “bootstrap”. Em outras palavras, isso significa que a representação das matrizes de dissimilaridade está compatível e que, independente do método de agrupamento utilizado, a representação é válida. O único método que não obteve correlação cofenética alta foi o método de Ward, sendo o menos desejável para utilização nesta análise. O método de Ward tende a formar grupos homogêneos levando em consideração o número total de indivíduos para formação dos grupos o que, na nossa análise não é favorável principalmente pelo genótipo G19.

Tabela 6 - Correlação cofenética entre as matrizes cofenética e de dissimilaridade obtidas conforme o método de agrupamento utilizado

Métodos de agrupamento	Matriz	
	Original	Bootstrap
Vizinho mais próximo	0,69	0,77
Vizinho mais distante	0,72	0,80
Centroide	0,74	0,79
Mediana	0,72	0,77
Ward	0,59	0,58
Ligação média	0,80	0,82

Para todos os métodos estudados foi esboçada a distribuição empírica dos coeficientes de correlação cofenética via bootstrap com o intuito de verificar se a distribuição possui uma forma simétrica. Foi realizado também o teste de Mantel para verificar se a correlação que existe entre a matriz de distâncias obtida via bootstrap e a matriz cofenética é significativa.

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método do vizinho mais próximo baseado na distância euclidiana foi de 0,77; indicando uma boa representação das distâncias no dendrograma. Esse valor foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 15 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do vizinho mais próximo, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do vizinho mais próximo utilizando o teste de Mantel

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método do vizinho mais distante baseado na distância euclidiana foi de 0,80; indicando uma boa representação das distâncias no dendrograma. Esse valor foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 16 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do vizinho mais distante, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do vizinho mais distante utilizando o teste de Mantel

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método do centroide baseado na distância euclidiana foi de 0,79; indicando uma boa representação das distâncias no dendrograma. Esse valor foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 17 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método do centroide, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método do centroide utilizando o teste de Mantel

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método da mediana baseado na distância euclidiana foi de 0,77; indicando uma boa representação das distâncias no dendrograma. Esse valor foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 18 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método da mediana, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método da mediana utilizando o teste de Mantel

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método de Ward baseado na distância euclidiana foi de 0,58 que embora seja um valor baixo em relação aos outros, foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 19 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método de Ward, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método de Ward utilizando o teste de Mantel

Como foi apresentado na Tabela 6, o coeficiente de correlação cofenética (r) para o método da ligação média baseado na distância euclidiana foi de 0,82; indicando uma boa representação das distâncias no dendrograma. Esse valor foi significativo pelo teste de Mantel ao nível de 1% de significância, baseado em 1.000 reamostragens.

Figura 20 - (a) Correlação cofenética para as matrizes de dissimilaridade obtidas com o método da ligação média, baseado na distância euclidiana (b) Distribuição empírica do coeficiente de correlação cofenética obtida via bootstrap para o método da ligação média utilizando o teste de Mantel

Os métodos apresentados têm o mesmo objetivo, de agrupar indivíduos semelhantes dado um critério de parecença, e como foi possível observar, podem apresentar resultados bastante parecidos. Sobre qual critério utilizar, o pesquisador deve conhecer as propriedades, qualidades e deficiências de cada um desses e escolher àquele que melhor se adequar com a pesquisa em questão.

O principal problema encontrado na aplicação da análise de agrupamentos, é o fato de não existir um critério objetivo para a classificação dos grupos, um critério que determine o ponto de corte. Existem diversos métodos para determinar o ponto de corte, cada pesquisador deve utilizar aquele que melhor represente a realidade dos dados. Outro fator que dificulta a análise de agrupamentos, é se o pesquisador tiver um número alto de dados, pois a visualização do dendrograma fica comprometida. Esse problema pode ser resolvido fazendo-se o uso de um método não-hierárquico.

Para dados de genótipos, conciliar a análise de agrupamentos com a análise AMMI foi de grande valia. Quando se faz primeiramente a análise AMMI, já é possível verificar quais genótipos têm propriedades semelhantes, então já é de se esperar que genótipos semelhantes sejam de um mesmo grupo na análise de agrupamentos. Sendo assim, sempre que os dados forem desse tipo, é interessante unir as duas análises.

Com os softwares se desenvolvendo cada vez mais e pela rapidez com que o mesmos computam os resultados, a reamostragem “bootstrap” tem sido bastante utilizada em pesquisas com a finalidade de se estudar estabilidade, distribuições empíricas, intervalos de confiança, dentre outros assuntos nos quais a mesma é extramamente valiosa quando é desejado alta qualidade na informação sobre estabilidade dos dados em estudo.

5 CONSIDERAÇÕES FINAIS

O principal objetivo deste trabalho foi identificar a estabilidade de agrupamentos formados em dados de genótipos de soja e, em paralelo, foi feita a análise AMMI dos mesmos dados para comparação futura dos resultados.

Concluiu-se que independente do método de agrupamento utilizado, em sua maioria os resultados foram semelhantes, e comparando os mesmos com a metodologia AMMI, é possível conseguir excelentes resultados.

Essa metodologia é extremamente eficiente para identificação de genótipos divergentes, difíceis de se substituir em fins de melhoramento genético. Em paralelo, a metodologia é eficiente também para identificação de genótipos similares para reduções futuras que sejam necessárias em um programa de melhoramento e que não provoque redução do potencial desejado.

REFERÊNCIAS

- ABREU, V.M.N.; SILVA, M.A.; CRUZ, C.D.; FIGUEIREDO, E.A.P.; ABREU, P.G.; Desempenho e Predição de Híbridos e Análise de Agrupamento de Características de Matrizes de Frangos de Corte. **Revista Brasileira de Zootecnia**, Viçosa, v.31, n.2, p.617-626, 2002.
- ALBUQUERQUE, M.A. **Estabilidade em análise de agrupamento (cluster analysis)**. 2005. 64p. Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernambuco, Recife, 2005.
- ALBUQUERQUE, M.A.; FERREIRA, R.L.C.; SILVA, J.A.A.; SANTOS, E.S.; STOSIC, B.; SOUZA, A.L. Estabilidade em análise de agrupamento: Estudo de caso em ciência florestal. **Revista Árvore**, Viçosa, v.30, n.2, p. 257-265, 2006.
- ARAÚJO, M.F.C. **Teste estatístico para contribuição de genótipos e ambientes na matriz de interação GE**. 2008. 113p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2008.
- BARROS, E.A.C., **Aplicações de Simulação Monte Carlo e Bootstrap**. 2005. 52p. Monografia (Estatística) - Universidade Estadual de Maringá, Maringá, 2005.
- BARROSO, L.P.; ARTES, R., **Análise Multivariada**. Lavras: UFLA, 2003. 157p.
- BRASIL, E.M. **Comparação de métodos no estudo da interação genótipo $\times$ ambiente em milho (*Zea mays* L.)**. 1990. 214p. Dissertação (Mestrado em Agronomia) - Universidade Federal de Goiás, Goiânia, 1990.
- BUSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.
- CHAVES, L.J. Interação de cultivares com ambientes. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S., VALADARES, M.C. **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. p.673-713.
- CORNELIUS, P.L.; SEYEDSADR, M.; CROSSA, J. Using the shifted multiplicative mode to search for “separability” in crop cultivar trials. **Theoretical and Applied Genetics**, New York, v.84, p.161-172, 1992.
- CROSSA, J.; GAUCH, H.G.; ZOBEL, R.W. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. **Crop Science**, Madison, v.30, n.3, p.493-500, 1990.
- CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 2003. 585p.
- CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 1994. 390p.

DIAS, L.A.S. **Análises multidimensionais**. In: ALFENAS, A.C. (Ed.). **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microorganismos. Viçosa: UFV, 1998. p.405-475.

DUARTE, J.B.; VENCOSKY, R. **Interação genótipo × ambiente**: uma introdução à análise “AMMI”. Ribeirão Preto: Sociedade Brasileira de Genética, 1999. 60p.

DUARTE, J.B.; ZIMMERMANN, M.J.O. Selection of locations for common bean (*Phaseolus vulgaris* L.) germoplasm evaluation. **Revista Brasileira de Genética**, Ribeirão Preto, v.14, n.3, p.765-770, 1991.

EFRON, B. Bootstrap methods: another look at jackknife. **Annals of Statistics**, Hayward, v.7, n.1, p.1-26, 1979.

EFRON, B.; TIBSHIRANI, R. **An introduction to the bootstrap**. New York: Chapman & Hall, 1993. v.57, (Monographs on Statistics and Applied Probability).

FARIA, P.N. **Utilização de técnicas multivariadas na análise da divergência genética via modelo AMMI com reamostragem “bootstrap”**. 2012. 130p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2012.

FERREIRA, P.V. **Estatística experimental aplicada à agronomia**. Maceió: ADUFAL, 2000. 422p.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**. Great Britain, v.58, n.03, p.453-467, 1971.

GAUCH JÚNIOR, H.G. **Statistical analysis of regional yield trials**: AMMI analysis of factorial designs. Amsterdam: Elsevier, 1992. 278p.

GAUCH, H.G.; ZOBEL, R.W. AMMI analysis of yield trials. In: KANG, M.S.; GAUCH, H.G. (Ed.). **Genotype by environment interaction**. New York: CRC Press, 1996. p.416-428.

GONZALEZ, L.M.A.R. **Estudo da interação de genótipos × ambientes para a cultura de algodão (*Gossypium* sp) na região oriental do Paraguai**. 1988. 91p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1988.

GOLLOB, H.F. A statistical model which combines features of factor analytic and analysis of variance techniques. **Psychometrika**, Greensboro, v.33, n.1, p.73-115, 1968.

HILL, R.R.; ROSENBERGER, J.L. Methods for combining data from germplasm evaluation trials. **Crop Science**, Madison, v.25, p.467-470, 1985.

HONGYU, K. **Distribuição empírica dos autovalores associados à matriz de interação dos modelos AMMI pelo método bootstrap não-paramétrico**. 2012. 103p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2012.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. New Jersey, USA: Englewood Cliffs, 1992. 642p.

KEMPTON, R.A. The use of biplots in interpreting variety by environment interactions. **Journal of Agricultural Science**, Cambridge, v.103, p.123-135, 1984.

KLEFENS, P.C.O. **O biplot na análise fatorial multivariada**. 2009. 103p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2009.

LAVORANTI, O.J. **Estabilidade e adaptabilidade fenotípica através da reamostragem "bootstrap" no modelo AMMI**. 2003. 184p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

MAHALANOBIS, P.C. On the generalized distance in statistics. **Proceedings of The National Institute of Sciences of India**, India, v.12, p.49-55, 1936.

MANDEL, J. A new analysis of variance model for non-additive data. **Technometrics**, Alexandria, v.13, n.1. 1971. p.1-18.

MANLY, B.F.J. **Métodos Estatísticos Multivariados: Uma Introdução**, 3ª edição, Bookman, Porto Alegre, 2008. 229p.

MANTEL, N. The detection of disease clustering and generalizes regression approach. **Cancer Research**, Chestnut, v.27, p.209-220, 1967.

MILLIGAN, G.W.; COOPER, M.C. An examination of procedures for determining the number of cluster in a data set. **Psychometrika**, Williamsburg, v.50, p.159-179, 1985.

MOJENA, R. Hierarchical grouping methods and stopping rules: an evaluation. **The Computer Journal**, Kingston, v.20, n.4, p.359-363, 1977.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, Philadelphia, v.6, n.2, p.559-572, 1901.

PIEPHO, H.P. Robustness of statistical test for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trial. **Theoretical Applied of Genetics**, Berlin, v.90, p.438-443, 1995.

PIOVESAN, P. **Validação cruzada com correção de autovalores e regressão isotônica nos modelos AMMI**. 2007. 78p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

RAMALHO, M.A.P.; SANTOS, J.B.; ZIMMERMANN, M.J.O. **Genética quantitativa em plantas autógamas: aplicações ao melhoramento do feijoeiro**. Goiânia: UFG, 1993. 271p.

SARTORIO, S.D. **Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R**. 2008. 131p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

SILVA, A.R. **Métodos de agrupamento: Avaliação e aplicação ao estudo de divergência genética em acessos de alho**. 2012. 83p. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2012.

SOUZA, E.C. **Os métodos Biplot e escalonamento multidimensional nos delineamentos experimentais**. 2010. 135p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2010.

ZOBEL, R.W.; WRIGHT, M.J.; GAUCH, H.G. Statistical analysis of a yield trial. **Agronomy Journal**, Madison, v.80, p.388-393, 1988.

APÊNDICES

APÊNDICE A - Programação feita no software R

```
rm(list = ls(all = TRUE)) ## Limpando a memória ##

setwd("C:/Users/User/Desktop/Dissertação Débora") ## Definindo o diretório ##

## Entrando com o conjunto de dados ##

dados <- read.csv2("conjuntodedados.csv", head = TRUE); dados

library(agricolae) ## Pacote para realizar a análise AMMI ##

## Nomeando as colunas da matriz de dados ##

Ambientes <- dados[, 1]; Ambientes
Genotipos <- dados[, 3]; Genotipos
Blocos <- dados[, 2]; Blocos
Producao <- dados[, 4]; Producao

## Análise AMMI ##

## Comando para realizar a análise AMMI ##

Modelo <- AMMI(Ambientes, Genotipos, Blocos, Producao, MSE, number = FALSE,
               graph = "biplot")

## Exibe a interação de cada genótipo com um respectivo ambiente, isto é, ##
## a combinação de todos os genótipos com todos os ambientes ##

Modelo$genXenv

## Agora, vamos montar nossa matriz de interação GA ##

ListaGA <- Modelo$genXenv; ListaGA

A1 <- ListaGA[1:24]; A1
A2 <- ListaGA[25:48]; A2
A3 <- ListaGA[49:72]; A3
A4 <- ListaGA[73:96]; A4
A5 <- ListaGA[97:120]; A5

M <- cbind(A1, A2, A3, A4, A5); M
dimnames(M)[1] <- list(c(1, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 20,
                        21, 22, 23, 24, 3, 4, 5, 6, 7, 8, 9)); M

MatrizGA <- rbind(M[1, ], M[12, ], M[18, ], M[19, ], M[20, ], M[21, ], M[22, ],
                 M[23, ], M[24, ], M[2, ], M[3, ], M[4, ], M[5, ], M[6, ],
                 M[7, ], M[8, ], M[9, ], M[10, ], M[11, ], M[13, ], M[14, ],
                 M[15, ], M[16, ], M[17, ])

## Nomeando as linhas da matriz GA##

dimnames(MatrizGA)[1] <- list(paste("G", 1:24, sep = "")); MatrizGA
```

```

## Construção do gráfico do PC1 pela produção de grãos ##

Modelo$biplot
Biplot <- Modelo$biplot[, 1:4]; Biplot

PC1 <- Modelo$biplot$PC1; PC1
PC2 <- Modelo$biplot$PC2; PC2
Prod <- Modelo$biplot$Producao; Prod

length(PC1)
length(Prod)

## Fazendo o gráfico ##

par(cex = 0.8)

plot(Prod, PC1, cex = 0, text(Prod, PC1, cex = 0.8, labels = row.names(Biplot),
    col = "blue"), main = "Biplot AMMI", frame = TRUE)

MEANS <- mean(Prod)
abline(h = 0, v = MEANS, lty = 2, col = "red")

Biplot
amb <- subset(Biplot, type == "ENV"); amb

Produ <- amb$Producao; Produ
PC1amb <- amb$PC1; PC1amb
s <- seq(1, length(Produ)); s

arrows(MEANS, 0, 0.9*(Produ[s] - MEANS) + MEANS, 0.9*PC1amb[s], col = "brown", code = 2,
    lwd = 1.8, length = 0.1)

## Encontrando a matriz de distâncias euclidianas referente à matriz de interação GA ##

distancias <- dist(MatrizGA, method = "euclidean"); distancias # Matriz de distâncias

round(distancias, 2) ## Arredondando para duas casas decimais ##

require(cluster) # Pacote para realizar a AA

## Obtendo a separação dos grupos no dendrograma segundo o valor k fornecido ##

## Definindo k para o Determinação do ponto de corte ##

k = 1.25 ## MILLIGAN, COOPER; 1985 ##

## Iniciando os métodos de agrupamento ##

## Método da ligação média entre grupos ##

UPGMA <- hclust(distancias, method = "average")

```

```

## Determinação do ponto de corte ##

PCUPGMA <- mean(UPGMA$height) + k*sd(UPGMA$height); PCUPGMA

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(UPGMA, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método da ligação média entre grupos)")
abline(h = PCUPGMA, v = NULL, col = 4, lty = 2)
G <- rect.hclust(UPGMA, h = PCUPGMA, which = c(1:3), border = 1:6); G

## Calculando o coeficiente de correlação cofenética ##

cofUPGMA <- cophenetic(UPGMA)
cor(distancias, cofUPGMA)

## Método de Ward ##

Ward <- hclust(distancias, method = "ward")

## Determinação do ponto de corte ##

PCWard <- mean(Ward$height)+k*sd(Ward$height); PCWard

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(Ward, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método de Ward)")
abline(h = PCWard, v = NULL, col = 4, lty = 2)
H <- rect.hclust(Ward, h = PCWard, which = c(1:3), border = 1:6)

## Calculando o coeficiente de correlação cofenética ##

cofWard <- cophenetic(Ward)
cor(distancias, cofWard)

## Método do centroide ##

cen <- hclust(distancias, method = "centroid")

## Determinação do Determinação do ponto de corte ##

PCcen <- mean(cen$height) + k*sd(cen$height); PCcen

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(cen, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método do centroide)")
abline(h = PCcen, v = NULL, col = 4, lty = 2)

```

```

I <- rect.hclust(cen, h = PCcen, which = c(1:3), border = 1:6)

## Calculando o coeficiente de correlação cofenética ##

cofcen <- cophenetic(cen)
cor(distancias, cofcen)

## Método da mediana ##

med <- hclust(distancias, method = "median")

## Determinação do ponto de corte##

PCmed <- mean(med$height) + k*sd(med$height); PCmed

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(med, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
      main="Dendrograma", sub="(Método da mediana)")
abline(h = PCmed, v = NULL, col = 4, lty = 2)
J <- rect.hclust(med, h = PCmed, which = c(1:3), border = 1:6)

## Calculando o coeficiente de correlação cofenética ##

cofmed <- cophenetic(med)
cor(distancias, cofmed)

## Método do vizinho mais distante ##

distante <- hclust(distancias, method = "complete")

## Determinação do ponto de corte##

PCdistante <- mean(distante$height) + k*sd(distante$height); PCdistante

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(distante, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
      main="Dendrograma", sub="(Método do vizinho mais distante)")
abline(h = PCdistante, v = NULL, col = 4, lty = 2)
K <- rect.hclust(distante, h = PCdistante, which = c(1:4), border = 1:6)

## Calculando o coeficiente de correlação cofenética ##

cofdistante <- cophenetic(distante)
cor(distancias, cofdistante)

## Método vizinho mais próximo ##

prox <- hclust(distancias, method = "single")

```

```

## Determinação do ponto de corte ##

PCprox <- mean(prox$height) + k*sd(prox$height); PCprox

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(prox, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
      main="Dendrograma", sub="(Método do vizinho mais próximo)")
abline(h = PCprox, v = NULL, col = 4, lty = 2)
L <- rect.hclust(prox, h = PCprox, which = c(1:3), border = 1:6)

## Calculando o coeficiente de correlação cofenética ##

cofprox <- cophenetic(prox)
cor(distancias, cofprox)

## Reamostragem bootstrap da matriz de interação GA ##

nr <- 10000 ## Número de reamostragens ##

boot <- lapply(1:nr, function(i) MatrizGA[, sample(ncol(MatrizGA), ncol(MatrizGA),
                                                    replace = T)])

msoma <- matrix(0, ncol = 5, nrow = 24); msoma

for(i in 1:nr){
  msoma <- msoma + boot[[i]]
}
msoma
MatrizGA2 <- msoma/nr; MatrizGA2

## Análise de agrupamentos para a nova matriz de interação obtida ##
## via método bootstrap ##

distanciasboot <- dist(MatrizGA2, method = "euclidean"); distanciasboot

round(distanciasboot, 2) ## Arredondando para duas casas decimais ##

## Refazendo todos os métodos de agrupamentos para a nova matriz de distâncias ##

## Método da ligação média entre grupos ##

UPGMAb <- hclust(distanciasboot, method = "average")

## Determinação do ponto de corte ##

PCUPGMAb <- mean(UPGMAb$height) + k*sd(UPGMAb$height); PCUPGMAb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

```



```

plot(UPGMab, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método da ligação média entre grupos)")
abline(h = PCUPGMab, v = NULL, col = 4, lty = 2)
Gb <- rect.hclust(UPGMab, h = PCUPGMab, which = c(1:3), border = 1:6); Gb

## Coeficiente de correlação cofenética ##

cofUPGMab <- cophenetic(UPGMab)
vcofUPGMab <- c(cofUPGMab)
vcofUPGMab

vdistanciasb <- c(distanciasboot)
vdistanciasb

cor(vcofUPGMab, vdistanciasb)

plot(vdistanciasb, vcofUPGMab, xlab = "Matriz de distâncias",
     ylab = "Matriz cofenética (ligação média)", main="Correlação cofenética")
text(7, 0.5, "r = 0,82**")

## Fazendo o teste de Mantel ##

require(ape) ## Pacote que realiza o teste de Mantel ##
mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(UPGMab)),
            nperm = 1000, graph = T, xlab = "Estatística - z", ylab = "Densidade",
            main = "Teste de Mantel (Ligação média entre grupos)")

## Método de Ward ##

Wardb <- hclust(distanciasboot, method = "ward")

## Determinação do ponto de corte ##

PCWardb <- mean(Wardb$height) + k*sd(Wardb$height); PCWardb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(Wardb, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método de Ward)")
abline(h = PCWardb, v = NULL, col = 4, lty = 2)
Hb <- rect.hclust(Wardb, h = PCWardb, which = c(1:3), border = 1:6); Hb

## Coeficiente de correlação cofenética ##

cofWardb <- cophenetic(Wardb)
vcofWardb <- c(cofWardb); vcofWardb

cor(vcofWardb, vdistanciasb)

plot(vdistanciasb, vcofWardb, xlab = "Matriz de distâncias",
     ylab = "Matriz cofenética (Ward)", main="Correlação cofenética")

```

```

text(7, 1, "r = 0,58")

## Fazendo o teste de Mantel ##

mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(Wardb)),
             nperm = 1000, graph = T, xlab = "Estatística - z", ylab = "Densidade",
             main = "Teste de Mantel (Ward)")

## Método do centroide ##

cenb <- hclust(distanciasboot, method = "centroid")

## Determinação do ponto de corte ##

PCcenb <- mean(cenb$height) + k*sd(cenb$height); PCcenb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(cenb, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método do centroide)")
abline(h = PCcenb, v = NULL, col = 4, lty = 2)
Ib <- rect.hclust(cenb, h = PCcenb, which = c(1:3), border = 1:6); Ib

## Coeficiente de correlação cofenética ##

cofcenb <- cophenetic(cenb)
vcofcenb <- c(cofcenb); vcofcenb

cor(vcofcenb, vdistanciasb)

plot(vdistanciasb, vcofcenb, xlab = "Matriz de distâncias",
     ylab = "Matriz cofenética (centroide)", main = "Correlação cofenética")
text(7, 0.5, "r = 0,79")

## Fazendo o teste de Mantel ##

mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(cenb)),
             nperm = 1000, graph = T, xlab = "Estatística - z", ylab = "Densidade",
             main = "Teste de Mantel (centroide)")

## Método da mediana ##

medb <- hclust(distanciasboot, method = "median")

## Determinação do ponto de corte ##

PCmedb <- mean(medb$height) + k*sd(medb$height); PCmedb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

```

```

plot(medb, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método da mediana)")
abline(h = PCmedb, v = NULL, col = 4, lty = 2)
Jb <- rect.hclust(medb, h = PCmedb, which = c(1:3), border = 1:6); Jb

## Coeficiente de correlação cofenética ##

cofmedb <- cophenetic(medb)
vcofmedb <- c(cofmedb); vcofmedb

cor(vcofmedb, vdistanciasb)

plot(vdistanciasb, vcofmedb, xlab = "Matriz de distâncias",
     ylab = "Matriz cofenética (mediana)", main = "Correlação cofenética")
text(7, 0.5, "r = 0,77")

## Fazendo o teste de Mantel ##

mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(medb)),
            nperm = 1000, graph = T, xlab = "Estatística - z",
            ylab = "Densidade", main = "Teste de Mantel (mediana)")

## Método do vizinho mais distante ##

distanteb <- hclust(distanciasboot, method = "complete")

## Determinação do ponto de corte ##

PCdistanteb <- mean(distanteb$height) + k*sd(distanteb$height); PCdistanteb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(distanteb, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
     main="Dendrograma", sub="(Método do vizinho mais distante)")
abline(h = PCdistanteb, v = NULL, col = 4, lty = 2)
Kb <- rect.hclust(distanteb, h = PCdistanteb, which = c(1:4), border = 1:6); Kb

## Coeficiente de correlação cofenética ##

cofdistanteb <- cophenetic(distanteb)
vcofdistanteb <- c(cofdistanteb); vcofdistanteb

cor(vcofdistanteb, vdistanciasb)

plot(vdistanciasb, vcofdistanteb, xlab = "Matriz de distâncias",
     ylab = "Matriz cofenética (vizinho mais distante)", main = "Correlação cofenética")
text(7, 0.8, "r = 0,80")

## Fazendo o teste de Mantel ##

mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(distanteb)),

```

```

nperm = 1000, graph = T, xlab = "Estatística - z",
ylab = "Densidade", main = "Teste de Mantel (vizinho mais distante)")

## Método vizinho mais próximo ##

proxb <- hclust(distanciasboot, method = "single")

## Determinação do ponto de corte ##

PCproxb <- mean(proxb$height) + k*sd(proxb$height); PCproxb

## Esboçando o gráfico com a determinação do ponto de corte e exibindo ##
## os agrupamentos ##

plot(proxb, hang = -1, cex = 0.8, ylab = "Distâncias", xlab = "Genótipos",
      main="Dendrograma", sub="(Método do vizinho mais próximo)")
abline(h = PCproxb, v = NULL, col = 4, lty = 2)
Lb <- rect.hclust(proxb, h = PCproxb, which = c(1:4), border = 1:6); Lb

## Coeficiente de correlação cofenética ##

cofproxb <- cophenetic(proxb)
vcofproxb <- c(cofproxb); vcofproxb

cor(vcofproxb, vdistanciasb)

plot(vdistanciasb, vcofproxb, xlab = "Matriz de distâncias",
      ylab = "Matriz cofenética (vizinho mais próximo)", main = "Correlação cofenética")
text(7, 0.5, "r = 0,77")

## Fazendo o teste de Mantel ##

mantel.test(as.matrix(distanciasboot), as.matrix(cophenetic(proxb)),
            nperm = 1000, graph = T, xlab = "Estatística - z",
            ylab = "Densidade", main = "Teste de Mantel (vizinho mais próximo)")

```


ANEXOS

ANEXO A - Distâncias

Uma medida d_{ij} é definida como uma distância entre os pontos i e j se:

1. $d_{ij} \geq 0 \forall i, j$;
2. $d_{ii} = 0$;
3. $d_{ij} = d_{ji}$;
4. $d_{ij} \leq d_{im} + d_{mj}$.

A distância euclidiana do elemento i ao elemento j é calculada da seguinte maneira:

$$d_{ij} = \sqrt{\sum_{k=1}^p (y_{ik} - y_{jk})^2}$$

em que:

d_{ij} é a distância do elemento i ao j , $i, j = 1, 2, \dots, n$,
 y_{ik} e y_{jk} são os valores observados da variável k , $k = 1, 2, \dots, p$ para os indivíduos i e j .

A distância de Mahalanobis generalizada entre um grupo de valores com média $\underline{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^T$ e matriz de covariâncias Σ para um vetor multivariado $\underline{x} = (x_1, x_2, \dots, x_p)^T$ é definida da seguinte maneira:

$$D^2 = \sqrt{(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})}$$

e representa a distância generalizada de $\underline{x}$ a $\underline{\mu}$ (SARTORIO, 2008).

A distância de Mahalanobis pode ser definida como uma medida de dissimilaridade entre dois vetores aleatórios $\underline{x}$ e $\underline{y}$ que possuam mesma distribuição com matriz de covariâncias Σ :

$$d(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})^T \Sigma^{-1} (\underline{x} - \underline{y})}$$

Se $\Sigma = \mathbf{I}$, em que $\mathbf{I}$ é a matriz identidade, a distância de Mahalanobis se reduz à distância Euclidiana. A diferença entre a distância Euclidiana e a distância de Mahalanobis é que a segunda leva em consideração a correlação do conjunto de dados.

ANEXO B - Exemplo de autovalores e autovetores

Considere a matriz quadrada $\mathbf{A}_{2 \times 2}$ a seguir:

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$$

Para calcular os autovalores dessa matriz deve-se fazer $|\mathbf{A} - \lambda \mathbf{I}| = 0$, em que $\mathbf{I}$ representa a matriz identidade de ordem dois. Logo,

$$\det \left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = 0 \Leftrightarrow \det \left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = 0$$

$$\det \begin{pmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{pmatrix} = 0 \Leftrightarrow (4 - \lambda)(3 - \lambda) - 2 = 0$$

$$\lambda^2 - 7\lambda + 12 - 2 = 0 \Rightarrow \lambda^2 - 7\lambda + 10 = 0 \Leftrightarrow (\lambda - 5)(\lambda - 2) = 0 \Leftrightarrow \lambda_1 = 2 \text{ ou } \lambda_2 = 5$$

Portanto, os autovalores associados à matriz $\mathbf{A}$ são $\lambda_1 = 2$ e $\lambda_2 = 5$. Agora, basta encontrar os autovetores associados a cada um desses autovalores fazendo $(\mathbf{A} - \lambda \mathbf{I})\underline{\mathbf{x}} = \underline{\mathbf{0}}$. Fazendo primeiramente para $\lambda_1 = 2$:

$$\begin{aligned} \left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \\ &\Leftrightarrow \begin{cases} 2x_1 + 2x_2 = 0 \\ x_1 + x_2 = 0 \end{cases} \Leftrightarrow x_1 = -x_2 \end{aligned}$$

Sendo assim, $\underline{\mathbf{v}}_1 = (-x_2, x_2)$ é o autovetor associado a λ_1 . Basta fazer o mesmo para λ_2 :

$$\begin{aligned} \left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - 5 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \\ &\Leftrightarrow \begin{cases} -x_1 + 2x_2 = 0 \\ x_1 - 2x_2 = 0 \end{cases} \quad x_1 = 2x_2 \end{aligned}$$

Logo, $\underline{\mathbf{v}}_2 = (2x_2, x_2)$ é o autovetor associado a λ_2 .

ANEXO C - Exemplo de decomposição em valores singulares

Considere a matriz quadrada $\mathbf{A}_{2 \times 2}$ a seguir:

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$$

Para encontrar sua decomposição em valores singulares primeiramente deve-se encontrar as matrizes $\mathbf{A}\mathbf{A}^T$ e $\mathbf{A}^T\mathbf{A}$ e encontrar seus respectivos autovalores.

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 20 & 10 \\ 10 & 10 \end{bmatrix}$$

Encontrando os autovalores da matriz $\mathbf{A}\mathbf{A}^T$:

$$\det \left(\begin{bmatrix} 20 & 10 \\ 10 & 10 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = 0 \Leftrightarrow \det \left(\begin{bmatrix} 20-\lambda & 10 \\ 10 & 10-\lambda \end{bmatrix} \right) = 0 \Leftrightarrow \lambda^2 - 30\lambda + 100 = 0 \Rightarrow$$

$$\Rightarrow \lambda_1 = 26,18 \text{ e } \lambda_2 = 3,82$$

Encontrando os autovetores associados a λ_1 e a λ_2 . Para $\lambda_1 = 26,18$ tem-se:

$$\left(\begin{bmatrix} 20 & 10 \\ 10 & 10 \end{bmatrix} - 26,18 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} -6,18 & 10 \\ 10 & -16,18 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow$$

$$\Rightarrow x_1 = 1,618 x_2$$

Logo, $\mathbf{u}_1 = (1,618 x_2; x_2)$ é o autovetor associado a λ_1 , podendo ser representado, por exemplo, por $\mathbf{u}_1 = (1,618; 1)$. Agora, é necessário ortonormalizar este vetor (deixá-lo com comprimento unitário), para isto basta fazer:

$$\frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}$$

em que $\|\mathbf{u}_1\|$ é a norma do vetor $\mathbf{u}_1$ e é calculada da seguinte forma: $\|\mathbf{u}_1\| = \sqrt{x^2 + y^2}$, para x e y coordenadas do vetor $\mathbf{u}_1$.

$$\|\mathbf{u}_1\| = \sqrt{1,618^2 + 1^2} = \sqrt{3,618}$$

$$\frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} = \frac{(1,618; 1)}{1,902} \Rightarrow \mathbf{u}_1 = \begin{bmatrix} 0,850 \\ 0,525 \end{bmatrix}$$

Para o autovalor λ_2 deve-se repetir o raciocínio: obter seu autovetor e ortonormalizá-lo. Fazendo isso, deve-se obter:

$$\tilde{\mathbf{u}}_2 = \begin{bmatrix} -0,525 \\ 0,850 \end{bmatrix}$$

Esses autovalores e autovetores foram obtidos para a matriz $\mathbf{A}\mathbf{A}^T$. Deve-se fazer o mesmo procedimento para a matriz $\mathbf{A}^T\mathbf{A}$.

Os autovalores da matriz $\mathbf{A}^T\mathbf{A}$ são $\lambda_1 = 26,18$ e $\lambda_2 = 3,82$ (são exatamente os mesmos que os autovalores da matriz $\mathbf{A}\mathbf{A}^T$). Os autovetores ortonormalizados são:

$$\tilde{\mathbf{v}}_1 = \begin{bmatrix} 0,767 \\ 0,640 \end{bmatrix} \text{ e } \tilde{\mathbf{v}}_2 = \begin{bmatrix} -0,640 \\ 0,767 \end{bmatrix}$$

A decomposição em valores singulares (DVS) para a matriz inicial $\mathbf{A}$ é dada pelas parcelas (são duas parcelas pelo posto da matriz $\mathbf{A}$ ser igual a dois):

$$\sqrt{\lambda_1}\tilde{\mathbf{u}}_1\tilde{\mathbf{v}}_1^T = \sqrt{26,18} \cdot \begin{bmatrix} 0,850 \\ 0,525 \end{bmatrix} \cdot \begin{bmatrix} 0,767 & 0,640 \end{bmatrix} = \begin{bmatrix} 3,341 & 2,788 \\ 2,065 & 1,723 \end{bmatrix} = \mathbf{A}_1$$

$$\sqrt{\lambda_2}\tilde{\mathbf{u}}_2\tilde{\mathbf{v}}_2^T = \sqrt{3,82} \cdot \begin{bmatrix} -0,525 \\ 0,850 \end{bmatrix} \cdot \begin{bmatrix} -0,640 & 0,767 \end{bmatrix} = \begin{bmatrix} 0,658 & -0,788 \\ -1,065 & 1,276 \end{bmatrix} = \mathbf{A}_2$$

Fazendo-se $\mathbf{A}_1 + \mathbf{A}_2$ pode-se observar que é obtida a matriz $\mathbf{A}$. Outra maneira de obter a DVS de $\mathbf{A}$ é fazendo $\mathbf{U}\mathbf{S}\mathbf{V}^T$, ou seja:

$$\begin{array}{cccccc} \mathbf{u}_1 & \mathbf{u}_2 & \sqrt{\lambda_1} & \sqrt{\lambda_2} & \mathbf{v}_1 & \mathbf{v}_2 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \underbrace{\begin{bmatrix} 0,850 & -0,525 \\ 0,525 & 0,850 \end{bmatrix}}_{\mathbf{U}} \cdot \underbrace{\begin{bmatrix} \sqrt{26,18} & 0 \\ 0 & \sqrt{3,82} \end{bmatrix}}_{\mathbf{S}} \cdot \underbrace{\begin{bmatrix} 0,767 & 0,640 \\ -0,640 & 0,767 \end{bmatrix}}_{\mathbf{V}^T} \end{array}$$