A Metodologia AMMI: Com Aplicação ao Melhoramento Genético

Carlos Tadeu dos Santos Dias
Kuang Hongyu
Lúcio Borges de Araújo
Maria Joseane Cruz da Silva
Marisol García Peña
Mirian Fernandes Carvalho Araújo
Paulo Canas Rodrigues
Priscila Neves Faria
Sergio Arciniegas Alarcón

A Metodologia AMMI: Com Aplicação ao Melhoramento Genético

Carlos Tadeu dos Santos Dias
Kuang Hongyu
Lúcio Borges de Araújo
Maria Joseane Cruz da Silva
Marisol García Peña
Mirian Fernandes Carvalho Araújo
Paulo Canas Rodrigues
Priscila Neves Faria
Sergio Arciniegas Alarcón

A Metodologia AMMI: Com Aplicação ao Melhoramento Genético

Carlos Tadeu dos Santos Dias
ESALQ/USP

Kuang Hongyu
ESALQ/USP

Lúcio Borges de Araújo
UFU

Maria Joseane Cruz da Silva
ESALQ/USP

Marisol García Peña
ESALQ/USP

Mirian Fernandes Carvalho Araújo
UFU

Paulo Canas Rodrigues
UFBA e ISLA Campus Lisboa, Portugal

Priscila Neves Faria
UFU

Sergio Arciniegas Alarcón
ESALQ/USP

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

A metodologia AMMI: com aplicação ao melhoramento genético ... [recurso eletrônico] / Carlos Tadeu dos Santos Dias ... [et al.]. - - Piracicaba: ESALQ, 2014.
169 p. : il.

ISBN: 978-85-86481-36-9

1. Genética estatística 2. Interação genótipo-ambiente (Modelos matemáticos) 3. Melhoramento genético (Métodos estatísticos) I. Dias, C.T. dos S. II. Hongyu, k. III. Araújo, L. B. de IV. Silva, M. J. C. da V. García Peña, M. VI. Araújo, M. F. C. VII. Canas Rodrigues, P. VIII. Faria, P. N IX. Arciniegas Alarcón, S. X. Título

CDD 575.10212
M593

Agradecimentos

O autor Carlos Tadeu agradece ao CNPq pela bolsa de produtividade e também à FAPESP pelo apoio aos doutores do Estado de São Paulo. A autora Maria Joseane, também agradece ao CNPq pela bolsa de projeto. O autor Paulo C. Rodrigues agradece apoio financeiro parcial à Fundação para a Ciência e Tecnologia, Ministério da Ciência, Tecnologia, e Ensino Superior, Portugal, para realização deste trabalho, através do projeto PTDC/AGR-PRO/2335/2012. Este módulo do minicurso é baseado num trabalho conjunto com H.G. Gauch, J.D. Munkvold, E.L. Heffner, M. Sorrells, da Universidade de Cornell, USA.

Prefácio

Este trabalho é fruto de uma década de pesquisa junto aos alunos de pós-graduação em Estatística e Experimentação Agronômica da ESALQ/USP e pesquisadores do Brasil e do exterior. O objetivo deste livro é introduzir a metodologia AMMI para aqueles que têm e aqueles que não têm formação matemática. Não pretendemos apresentar um livro texto detalhado, mas a intenção é que sirva como uma luz para pesquisadores e estudantes ao nível de graduação e pós-graduação. Em outras palavras, é um livro para estimular a pesquisa e a busca por conhecimento em uma área de métodos estatísticos.

Os leitores precisam ter certo conhecimento prático de estatística elementar, incluindo testes de significância usando a distribuição normal, t, qui-quadrado e F; análise de variância e regressão linear e não-linear. Algum conhecimento de álgebra matricial e geometria são importantes, como em qualquer método multivariado.

Algumas razões pelas quais a metodologia AMMI vem crescendo em usuários e aplicações é o seu forte apelo para explicar o padrão de resposta da interação entre fatores em pesquisas experimentais, o pronto acesso a algum ambiente computacional para realizar a programação e, por conseguinte os cálculos e a aplicação prática no estudo interação entre genótipos e ambientes no melhoramento genético vegetal ou animal, embora a metodologia seja universal, ou seja, pode ser aplicado à qualquer pesquisa nas diferentes áreas do conhecimento humano, que envolva o estudo de fatores e suas interações.

Os capítulos podem ser lidos de forma independente. O primeiro é introdutório, focalizando a interação Genótipo por Ambiente e o enfoque estatístico, apresentando o modelo AMMI, o gráfico *Biplot* e validação cruzada. O capítulo 2 cobre a distribuição empírica dos autovalores da matriz de interação, utilizando reamostragem *bootstrap*. O capítulo 3 discute a divergência genética utilizando reamostragem *bootstrap* e análise de agrupamento. O capítulo 4 trás e aplica um método de correção de autovalores viesados e mostra a eficiência da correção. O capítulo 5 apresenta um teste para confirmar a contribuição de genótipos e ambientes para a interação. O capítulo 6 trata da imputação simples e múltipla para observações ausentes na matriz de interação. O capítulo 7 trás os modelos AMMI bivariados utilizando a análise de procrustes. O capítulo 8 cobre o modelo AMMI no estudo da interação entre QTL e ambiente. O capítulo 9 contém a generalização dos modelos AMMI para três fatores com uso dos modelos PARAFAC e TUCKER, gráficos *Joint plot* e

Triplot. Salientamos que os resultados apresentados aqui foram obtidos pelos autores e portanto, a notação algébrica pode não seguir o mesmo padrão em cada capítulo.

Todo e qualquer comentário sobre o texto desta edição do livro bem como correções, serão bem vindos. Erros que ainda tenham permanecidos são somente de responsabilidade dos autores.

Gostaríamos de agradecer aos avaliadores e supervisores das teses, dissertações e supervisões de pós-doutorado dos autores, bem como a todos que indiretamente contribuíram para essa primeira edição.

Os autores

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Interação Genótipo × Ambiente

A interação ($G \times E$) é definida como o comportamento diferencial de genótipos em função da diversidade ambiental. Neste sentido, na presença da interação, os resultados das avaliações podem mudar de um ambiente para outro, ocasionando mudanças na posição relativa dos genótipos ou mesmo na magnitude das suas diferenças (FALCONER; MACKAY, 1996). Para Chaves (2001), a interação ($G \times E$) deve ser encarada como um fenômeno biológico com suas implicações no melhoramento de plantas e não como um simples efeito estatístico, cumprindo buscar a explicação evolutiva do evento se se quiser tirar proveito de seus efeitos benéficos indesejáveis sobre a avaliação de genótipos e recomendação de cultivares. Diferenças em adaptação de genótipos em populações resultam, evidentemente, de diferenças de constituição gênica para os caracteres importantes nesta adaptação. A reação diferencial às mudanças ambientais pode-se dar desde os mecanismos de regulação gênica até caracteres morfológicos finais.

Segundo Duarte e Vencovsky (1999) a interação ($G \times E$) representa uma das principais dificuldades encontradas pelo melhorista durante sua atividade seletiva. Nas etapas preliminares desse processo (com avaliações normalmente em uma só localidade), a interação ($G \times E$) pode inflacionar as estimativas da variância genética, resultando em superestimativas dos ganhos genéticos esperados com a seleção (ganhos reais inferiores aos previstos). Nas fases finais, em geral, os ensaios são conduzidos em vários ambientes (locais, anos e/ou épocas), o que possibilita o isolamento daquele componente de variabilidade; muito embora, neste momento, a intensidade de seleção seja baixa, o que já minimizaria seus efeitos sobre previsões de ganho genético. Por outro lado, a presença dessa interação, na maioria das vezes, faz com que os melhores genótipos em um determinado local não o sejam em outros. Isso dificulta a recomendação de genótipos (cultivares) para toda a população de ambientes

amostrada pelos testes. Estatisticamente, isso decorre da impossibilidade de interpretar, de forma aditiva, os efeitos principais de genótipos e de ambientes (KANG; MAGARI, 1996).

Cockerham (1963) atribuiu o aparecimento de interações ($G \times E$) como sendo devido a respostas diferenciais do mesmo conjunto gênico em ambientes distintos ou pela expressão de diferentes conjuntos gênicos em diferentes ambientes. Quando um mesmo conjunto de genes se expressa em diferentes ambientes, as diferenças nas respostas podem ser explicadas pela heterogeneidade das variâncias genéticas e experimentais ou por ambas, e , quando diferentes conjuntos de genes se expressam em ambientes distintos, as diferenças nas respostas explicam-se pela inconsistência das correlações genéticas entre os valores de um mesmo caráter em dois ambientes (FALCONER, 1989). Segundo Cruz e Regazzi (1994), a interação ($G \times E$) também pode surgir em função de fatores fisiológicos e bioquímicos próprios de cada genótipo cultivado. Chaves et al. (1989) relatam ainda que a falta de ajuste do modelo estatístico adotado ao conjunto de dados pode ser uma das causas da interação ($G \times E$) significativa.

Várias metodologias têm sido propostas no sentido de entender melhor o efeito da interação ($G \times E$). Algumas dessas propostas são: zoneamento ecológico ou estratificação de ambientes, ou seja, identificar regiões ou subregiões onde o efeito da interação seja não significativo pode levar a identificação de genótipos que se adaptam a ambientes específicos e ainda identificar genótipos com uma ampla adaptação ou estabilidade (RAMALHO et al., 1993). Da importância dessa interação no campo experimental devem-se escolher os métodos estatísticos que melhor expliquem a informação contida nos dados, um daqueles métodos é o modelo de interação multiplicativa, também conhecido como o modelo de efeitos principais aditivos e interação multiplicativa - AMMI, que tem como objetivo selecionar modelos que expliquem o padrão de resposta da interação, deixando fora o ruído presente nos dados (Alarcón, S.A., 2009).

Foram ilustrados por Allard e Bradshaw (1964) e ampliados aqui, alguns tipos de interações em que são considerados dois genótipos em dois ambientes. Nas Figuras 1.1, 1.2 e 1.3 são apresentadas três situações básicas que trazem diferentes consequências para o melhoramento. Em dois ambientes A_1 e A_2 são avaliados dois genótipos G_1 e G_2. Assume-se que o ambiente A_1 é mais favorável para a manifestação do caráter genérico Y, que pode ser qualquer variável métrica. Na Figura 1.1 há ausência de interação, ou seja, a mudança

das condições ambientais afeta igualmente o comportamento dos genótipos e a diferença entre eles permanece constante nos dois ambientes. Observa-se nas Figuras 1.1c, 1.1d e 1.1e que os genótipos podem coincidir e portanto, terem o mesmo comportamento nos dois ambientes ou apresentarem diferenças de um ambiente para outro, mas com ausência da interação. Nas Figuras 1.1 e 1.2 ,exceto Figura 1.1c, a mudança de ambiente afeta desigualmente a manifestação do caráter para os dois genótipos, ou seja, a diferença entre os genótipos varia entre ambientes. Na Figura 1.2, o efeito de cada ambiente não modifica a classificação dos genótipos, sendo o G_1 superior a G_2 em todas condições. Neste caso a interação é denominada simples ou quantitativa. A Figura 1.3 apresenta uma mudança na classificação dos genótipos. Para este tipo, a interação é denominada cruzada ou qualitativa.

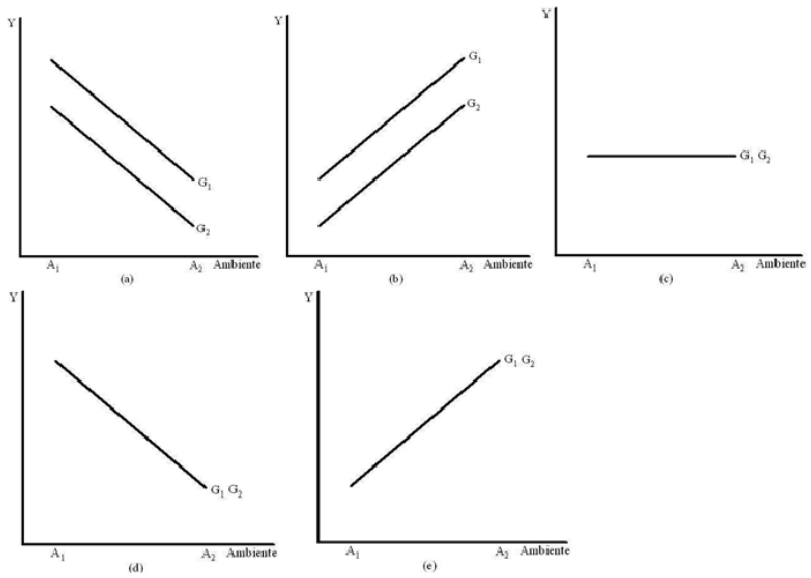

Figura 1.1: Comportamento de dois genótipos (G_1 e G_2) em duas condições ambientais (A_1 e A_2) com ausência de interação

Nas Figuras 1.1 e 1.2, as implicações para o melhoramento é que um mesmo genótipo G_1 é melhor adaptado às duas condições ambientais e uma seleção baseada na média dos ambientes beneficiará sempre o melhor genótipo. Na Figura 1.3, a seleção baseada na média dos ambientes não é capaz de satisfazer o conjunto dos ambientes podendo levar a uma seleção de genótipos mal adaptados a uma situação particular. Nas Figuras 1.1b e 1.2b ocorrem sinergismos e nas Figuras 1.1a e 1.2a antagonismo entre genótipos e ambientes, ao

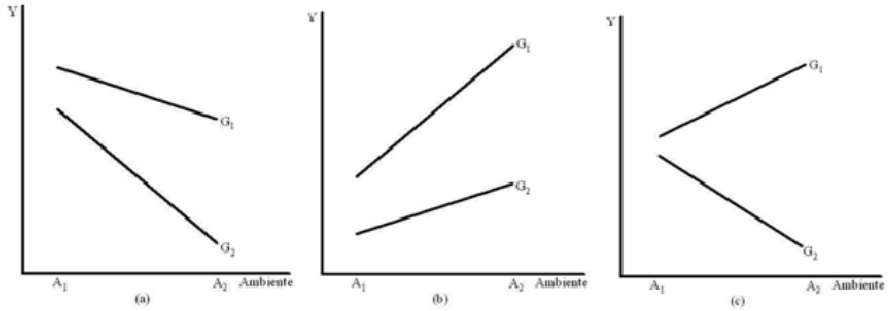

Figura 1.2: Comportamento de dois genótipos (G_1 e G_2) em duas condições ambientais (A_1 e A_2) com interação simples ou quantitativa

passar do ambiente A_1 para o ambiente A_2.

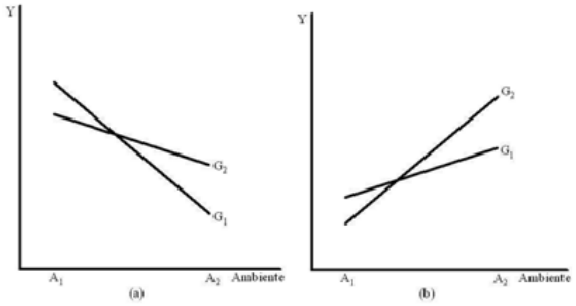

Figura 1.3: Comportamento de dois genótipos (G_1 e G_2) em duas condições ambientais (A_1 e A_2) com interação cruzada ou qualitativa

Para as três figuras anteriores o papel de G_1 e G_2 pode ser permutado com A_1 e A_2, obtendo-se interpretações semelhantes. Quando se consideram vários genótipos avaliados em vários ambientes, a combinação de situações como as das Figuras 1.1, 1.2 e 1.3 formam um emaranhado de situações, difícil de ser interpretado, exigindo métodos adequados de análise da interação genótipos ambientes (Araújo, M.F.C., 2008).

As variáveis ambientais também podem ser classificadas em dois tipos: previsíveis e imprevisíveis (ALLARD e BRADSHAW, 1964). As variáveis previsíveis seriam as características gerais do clima e solos que ocorrem de maneira sistemática ou que estão sob controle do homem. Já as variáveis imprevisíveis correspondem às flutuações climáticas tais como quantidade e distribuição de chuvas, temperatura e outros fatores que não podem ser controlados pelo homem.

A interação genótipos × ambientes, é uma fonte de variação fenotípica, que na maioria dos casos é inseparável da variância ambiental (FALCONER, 1987). Na prática, para verificar a significância da interação de genótipos com ambientes, é necessário repetir o experimento várias vezes, pois se o experimento for realizado somente em um ambiente, poderá ocorrer uma superestimação dos ganhos genéticos (CROSSA, 1990).

Existe uma concordância geral entre melhoristas de plantas de que interação genótipos × ambientes tem um importante significado para a obtenção de variedades superiores. Entretanto, de acordo com Allard (1971) é muito difícil encontrar concordâncias sobre o que se deve conhecer em relação a interação genótipos × ambientes e como utilizá-la.

A natureza da interação genótipos × ambientes também deve ser considerada e não somente a verificação de sua existência (VENCOVSKY e BARRIGA, 1992). Assim a natureza pode ser simples e complexa. A interação de natureza simples indica a presença de genótipos adaptados em um grande número de ambientes, sendo possível fazer uma recomendação generalizada de cultivares. A interação com natureza complexa mostra que existem genótipos adaptados a apenas alguns ambientes, o que traz uma complicação ao pesquisador, quando da recomendação de cultivar.

A existência de interação genótipos × ambientes, produz uma barreira de dificuldades aos melhoristas na identificação de genótipos superiores, tanto no processo de seleção quanto no processo de recomendação de cultivares. Essa interação indica que o comportamento dos genótipos nos experimentos depende principalmente das condições ambientais a que são submetidos. Assim a resposta obtida de um genótipo, em comparação a outro, é variável, sendo que essas variações se apresentam devido a mudança de ambientes (OLIVEIRA; DUARTE; PINHEIRO, 2003; KANG, 1998).

Assim, a interação de genótipos × ambientes deve ser considerada, não como um problema ou um fator indesejável, cujo efeito deve ser minimizado, mas deve ser enfrentada como um fenômeno biológico natural, que deve ser bem conhecido para melhor aproveitá-lo no processo de seleção (CHAVES, 2001). Logo os genótipos que interagem positivamente com os ambientes podem fazer a diferença entre um bom e um ótimo cultivar (DUARTE; VENCOVSKY, 1999).

1.1 A interação $G \times E$ e o enfoque estatístico

A existência de interação genótipos × ambientes tem sido reconhecida há longo tempo de acordo com Freeman (1973), sendo a referência mais antiga feita por Fisher e Mackenzie em 1923, a qual precede a análise de variância conjunta (ANOVA). Desde então, muitos trabalhos tem sido feitos para análises estatísticas da interação genótipos × ambientes, seja por estatísticos, agrônomos, melhoristas e geneticistas.

Os métodos mais simples que utilizam componentes de variância para a interpretação dos resultados quando existem interações, incluem estudos de regressão, métodos baseados em análises modificadas e métodos envolvendo variáveis ambientais externas.

Estudos realizados sobre interação genótipos × ambientes mostraram que a regressão foi utilizada pela primeira vez por Yates e Cochran (1938), analisando grupos de experimentos. Neste caso, o grau de associação entre diferenças varietais pôde ser verificado pelo cálculo da regressão dos rendimentos das variedades isoladas sobre os rendimentos médios de todas as variedades, mostrando assim que a regressão explicava grande parte da interação em uma série de experimentos de cevada. O método de regressão foi também usado por Perkins e Jinks (1968) para estimação de parâmetros em um modelo genético aplicado à biometria.

Outro método usado é aquele no qual os dados são organizados em uma tabela de dupla entrada em que o processo de investigação da interação é feito através da ANOVA. Esta análise envolve vários experimentos, e por isso é possível determinar a magnitude da interação, através da razão do quadrado médio da interação ($QMG \times E$) pelo quadrado médio do resíduo (QMR_{es}). A detecção de significância para a interação não esclarece, contudo, as implicações que estas possam ter sobre o melhoramento, de forma que, estudos de detalhamento deste componente da variação são em geral necessários.

Supondo-se que os genótipos foram avaliados nos diversos ambientes com r repetições em delineamento experimental que permita a estimativa da variação residual em cada ambiente, o modelo mais simples e comum para a análise estatística de um conjunto de dados é dado por (ANOVA para grupos de experimentos):

$$Y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ij} \tag{1.1}$$

sendo que:
Y_{ij} : é a resposta média do i-ésimo genótipo no j-ésimo ambiente;

μ: é uma constante comum às respostas (normalmente a média geral);
g_i: é o efeito do i-ésimo genótipo (i = 1; 2; ...; g);
e_j : é o efeito do j-ésimo ambiente (j = 1; 2;...; e);
$(ge)_{ij}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo ambiente; e
ϵ_{ij} : é o erro experimental médio, assumido independente e $\varepsilon_{ij} \sim N(0; \sigma^2 = r)$.

A forma básica de organização dos dados para análise da interação genótipos × ambientes está representada na Tabela 1, na qual estão representadas as médias dos g tratamentos (genótipos) sobre r repetições, derivadas de um experimento em delineamento apropriado. Quando todos os genótipos são avaliados em todos os ambientes, com o mesmo número de repetições por experimento, diz-se que os dados são balanceados.

Tabela 1.1: Representação dos dados médios de g genótipos avaliados em e ambientes para um caráter genérico Y

	Ambientes				
Genótipo	1	2	...	e	Média $(\overline{Y}_{i.})$
1	Y_{11}	Y_{12}	...	Y_{1e}	$(\overline{Y}_{1.})$
2	Y_{21}	Y_{22}	...	Y_{2e}	$(\overline{Y}_{2.})$
	⋮	⋮	...	⋮	⋮
g	Y_{g1}	Y_{g2}	...	Y_{ge}	$(\overline{Y}_{g.})$
Médias $(\overline{Y}_{.j})$	$(\overline{Y}_{.1})$	$(\overline{Y}_{.2})$	...	$(\overline{Y}_{.e})$	$(\overline{Y}_{..})$

A análise de variância feita sob o modelo (1) com base nos dados da Tabela 1 é bastante simples, sendo calculados os efeitos principais pelas marginais da Tabela 1 e a interação como desvios do modelo, em que $\hat{g}e_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$ são os correspondentes elementos da Tabela 1 para o estudo da interação. O erro experimental $\overline{Y}_{.j}$ é calculado para cada experimento, utilizando-se o resíduo na análise conjunta.

Na solução do modelo (1) visando encontrar os estimadores dos parâmetros, pelo método de mínimos quadrados, admitindo-se as condições marginais:

$$\sum_{i=1}^{g} g_i = \sum_{j=1}^{e} e_j = \sum_{i=1}^{g} (ge)_{ij} = \sum_{j=1}^{e} (ge)_{ij} = \sum_{i=1}^{g} \sum_{j=1}^{e} (ge)_{ij} = 0$$

Assim, a solução é:

$\hat{\mu} = \frac{\sum_{i,j} Y_{ij}}{ge} = \overline{Y}_{..}$
$\hat{g}_i = \frac{\sum_{j} Y_{ij}}{e} - \hat{\mu} = \overline{Y}_{i.} - \overline{Y}_{..}$
$\hat{e}_j = \frac{\sum_{i} Y_{ij}}{g} - \hat{\mu} = \overline{Y}_{.j} - \overline{Y}_{..}$
$\hat{ge}_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$

Esses são estimadores não viesados dos parâmetros μ, g_i, e_j e $(ge)_{ij}$, sujeitos às condições marginais dadas acima. Se essas ou quaisquer outras condições não forem impostas sobre os parâmetros, então os estimadores dos parâmetros individuais são viesados.

A aproximação de mínimos quadrados $\hat{Y}_{ij}$ (ou valor predito) e seu respectivo resíduo, correspondente ao termo geral de interação $(\hat{ge})_{ij}$, são dados por:
$\epsilon_{ij} = Y_{ij} - \hat{Y}_{ij}$
$\hat{Y}_{ij} = \overline{Y}_{i.} + \overline{Y}_{.j} - \overline{Y}_{..}$
Agora, pode-se construir a matriz de interações $GE_{(g\times e)} = [\hat{ge}_{ij}]$:

$$GE_{(g\times e)} = \begin{bmatrix} \hat{ge}_{11} & \hat{ge}_{12} & \dots & \hat{ge}_{1e} \\ \hat{ge}_{21} & \hat{ge}_{22} & \dots & \hat{ge}_{2e} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{ge}_{g1} & \hat{ge}_{g2} & \dots & \hat{ge}_{ge} \end{bmatrix}$$

As somas de quadrados devido aos efeitos principais e interação, referentes aos dados de médias como na Tabela 1, são obtidas por:

$SQ_{Total} = \sum_{i=1}^{g}\sum_{j=1}^{e} Y_{ij}^2 - ge\overline{Y}_{..}^2$, com ge-1 graus de liberdade.
$SQ_G = e\sum_{i=1}^{g} \overline{Y}_{i.}^2 - ge\overline{Y}_{..}^2$, com g-1 graus de liberdade.
$SQ_E = g\sum_{j=1}^{e} \overline{Y}_{.j}^2 - ge\overline{Y}_{..}^2$, com e-1 graus de liberdade.
$SQ_{G\times E} = SQ_{Total} - SQ_G - SQ_E$, com (g-1)(e-1) graus de liberdade.

Os quadrados médios correspondentes são obtidos dividindo-se cada soma de quadrados pelos respectivos graus de liberdade. O quadrado médio do resíduo ou do erro médio ($QMRes$) é calculado pela média ponderada dos

QM's do resíduo (obtido nas ANOVA$'$s individuais de experimentos), assumidos homogêneos, usando seus respectivos graus de liberdade como pesos, ou seja:

$$QM\text{Res} = \frac{\sum_j SQ_{\text{Res}_j}}{\sum_j GL_{\text{Res}_j}}$$

em que $j = 1, 2, \ldots, e$ ambientes ou experimentos.

Uma consideração importante a ser feita para realização do teste F e interpretação dos resultados da análise conjunta da variância, diz respeito à natureza fixa ou aleatória dos efeitos do modelo de análise (CHAVES, 2001). A natureza fixa ou aleatória da interação é determinada pelos efeitos principais. Se genótipos e ambientes são fixos, a interação será fixa. Se pelo menos um dos fatores for aleatório, a interação será aleatória. Na análise de ensaios multiambientais, consideram-se, em geral, os efeitos de genótipos como fixos e os efeitos de ambientes como aleatórios, de tal forma que o efeito da interação genótipos × ambientes é aleatório nesse caso.

O estudo da interação genótipos × ambientes possibilita a identificação de cultivares mais adaptados a determinadas regiões, onde as mesmas poderão expressar o seu potencial genético. Assim, estudos sobre a magnitude de tais interações podem ser úteis na regionalização de cultivares, objetivando indicar áreas onde as mesmas possam expressar o máximo que as condições ambientais particulares permitam, com respeito a respostas de genótipos, e possibilitar a exploração de efeitos específicos de adaptação para determinadas regiões.

No que diz respeito à adaptabilidade e estabilidade de cada genótipo, tais fenômenos não devem ser considerados iguais, apesar de estarem relacionados entre si. A adaptabilidade refere-se à capacidade de os genótipos aproveitarem vantajosamente o estímulo do ambiente e a estabilidade diz respeito à capacidade de os genótipos mostrarem comportamento altamente previsível em razão do estímulo do ambiente.

O conceito de estabilidade foi classificado por Lin, Binns e Lefkovicth (1986) em três tipos. No tipo 1, o genótipo é considerado estável se sua variância entre os ambientes for pequena. Ela pode ser medida pela variância de cada genótipo nos diferentes ambientes e é útil para características que devem ser mantidas, tal como resistência a patógenos e pragas (NUNES, 2000). Na estabilidade tipo 2, o genótipo é considerado estável se sua resposta ao ambiente for paralela à resposta média de todos os materiais avaliados no experimento,

o que ocorre quando o genótipo possui interações mínimas com o ambiente. A estabilidade do tipo 3 é aquela na qual o genótipo seria considerado estável se o quadrado médio dos desvios de regressão for pequeno, classificando-o como de alta confiabilidade de resposta.

1.2 O Modelo de efeitos principais aditivos e interação multiplicativa (Modelo AMMI)

Em geral, um modelo de efeitos principais aditivos e interação multiplicativa (AMMI) pode ser útil para qualquer conjunto de dados provenientes de experimentos com dois fatores de classificação cruzada e é muito apropriado em certas situações descritas por Milliken e Johnson (1989), como por exemplo:

- Quando a interação estiver presente no modelo, mas não existirem diferenças nos tratamentos das linhas, nem nos tratamentos das colunas.
- Quando a interação estiver presente em uma só casela. Esse pode ser o caso no qual a observação seja um dado discrepante, que também pode ocorrer se uma combinação particular de tratamentos dá resultados muito raros quando for aplicada na unidade experimental (tratamentos de controle). A combinação de dois tratamentos de controle pode causar a interação nos dados e um simples modelo aditivo não pode ser ajustado.
- Quando toda a interação estiver em uma só linha (ou coluna). Isto pode ocorrer quando houver vários dados discrepantes na mesma linha (ou coluna).

O modelo AMMI é uma boa alternativa de análise, pois esses modelos ajudam à interpretação e melhor compreensão do fenômeno da interação de fatores, problema que se encontra presente no melhoramento genético de plantas, especificamente no estudo da interação genótipo por ambiente ($G \times E$). Vários autores afirmam que esta metodologia é melhor do que os métodos baseados em regressão. Crossa (1990) argumenta que a análise de regressão linear não é informativa se a linearidade falhar e depende do grupo de genótipos e ambientes incluídos e tende a simplificar modelos de resposta, explicando a variação devida à interação em uma única dimensão, quando na realidade ela pode ser bastante complexa. Esses procedimentos em geral, não informam sobre interações específicas de genótipos com ambientes (se positivas ou negativas),

dificultando explorar vantajosamente os efeitos da interação. É por isso, que Crossa (1990) sugere a aplicação de métodos multivariados como a análise de componentes principais (ACP), a análise de agrupamentos e o procedimento AMMI.

O modelo AMMI combina dois procedimentos estatísticos: análise da variância e a decomposição por valores singulares. Em um único modelo têm-se componentes aditivos para os efeitos principais (linhas ou genótipos e colunas ou ambientes), e componentes multiplicativos para os efeitos da interação. Duarte e Vencovsky (1999) explicam que os efeitos principais, na parte aditiva (média, efeitos genotípicos e ambientais), são ajustados por uma análise de variância comum (univariada) aplicada à matriz de dados, resultando em um resíduo de não aditividade, isto é, na interação ($G \times E$), e essa interação, constituinte da parte multiplicativa do modelo, é, depois, analisada pela decomposição por valores singulares da matriz de resíduos ou interação. Em seguida se apresenta o modelo AMMI de uma forma geral para dois fatores (T e B) de acordo com Milliken e Johnson (1989), nesse caso um fator pode corresponder a genótipos, por exemplo, o fator T, e o outro fator B pode corresponder aos ambientes.

Seja μ_{ij} a resposta esperada quando os níveis de tratamentos T_i e B_j são aplicados em uma dada unidade experimental, em que $i = 1; 2; \ldots; t$ e $j = 1; 2; \ldots; b$. Usando os resultados da decomposição de matrizes, é possível mostrar que qualquer matriz de dimensão $t \times b$ das μ_{ij} sempre pode ser decomposta da seguinte maneira:

$$\mu_{ij} = \mu + \tau_i + \beta_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + \ldots + \lambda_k\alpha_{ki}\gamma_{kj} \tag{1.2}$$

$i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, b$
em que μ representa a média geral, τ_i é o efeito do i-ésimo genótipo, β_j é o efeito do j-ésimo ambiente, $k = \text{posto}(\mathbf{\Omega})$, $\mathbf{\Omega} = (\omega_{ij})$ em que $\omega_{ij} = \mu_{ij} - \overline{\mu}_{i.} - \overline{\mu}_{.j} + \overline{\mu}_{..}$, λ_r (com $r = 1, \ldots, k$) é a raiz quadrada do r-ésimo autovalor das matrizes $\mathbf{\Omega}\mathbf{\Omega}^T$ e $\mathbf{\Omega}^T\mathbf{\Omega}$ de iguais autovalores não nulos, α_{ri} é o i-ésimo elemento (relacionado ao genótipo i) do r-ésimo autovetor de $\mathbf{\Omega}\mathbf{\Omega}^T$ associado a λ_r^2, γ_{rj} é o j-ésimo elemento (relacionado ao ambiente j) do r-ésimo autovetor de $\mathbf{\Omega}^T\mathbf{\Omega}$ associado a λ_r^2 .

Para reduzir o número de possíveis valores para os parâmetros em (2) e sintetizar a apresentação do modelo de efeitos principais aditivos e interação multiplicativa, é assumido sem perda de generalidade que:

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k|$$

$$\sum_i \tau_i = \sum_j \beta_j = 0$$

$$\sum_i \alpha_{ri} = \sum_j \tau_{rj} = 0, \quad parar = 1, 2, \ldots, k$$

$$\sum_i \alpha_{ri}^2 = \sum_j \gamma_{rj}^2 = 1, \quad parar = 1, 2, \ldots, k$$

$$\sum_i \alpha_{ri}\alpha_{r*i} = \sum_i \gamma_{rj}\gamma_{r*i} = 1, parar \neq r^* = 1, 2, \ldots, k \tag{1.3}$$

Nesse modelo, qualquer contraste nas médias μ_{ij}, o qual mede a interação, pode ser escrito como uma combinação linear dos w_{ij}. Também se tem que $k \leq$min(b-1; t-1).

Agora, em uma situação real, dado um conjunto de dados experimentais em uma tabela de dupla entrada com uma observação por casela y_{ij} (essa observação pode ser a média das repetições de cada tratamento em um delineamento balanceado), podem-se considerar modelos da seguinte maneira:

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + \ldots + \lambda_k\alpha_{ki}\gamma_{kj} \tag{1.4}$$

$i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, b$ em que é assumido que os parâmetros satisfazem às restrições dadas em (3) e que os erros ε_{ij} são independentemente distribuídos com média zero e variância comum σ^2. Para obter os resultados dos testes de hipóteses é necessário também assumir normalidade nos erros. Dependendo do número de componentes multiplicativos o modelo (4) é notado por AMMI0, AMMI1 ou AMMIk de forma genérica.

Um conjunto de estimativas de mínimos quadrados dos parâmetros nesse modelo é:

$$\hat{\mu} = \frac{\sum_{i,j} y_{ij}}{t \times b} = \overline{y}_{..}$$

$$\hat{\tau}_i = \frac{\sum_j y_{ij}}{b} - \hat{\mu} = \overline{y}_{i.} - \overline{y}_{..}, \quad i = 1, 2, \ldots, t$$

$$\hat{\beta}_j = \frac{\sum_i y_{ij}}{t} - \hat{\mu} = \overline{y}_{.j} - \overline{y}_{..}, \quad j = 1, 2, \ldots, b$$

$$\hat{\lambda}_r^2 = l_r, \qquad r = 1, 2, \ldots, k$$

em que $l_1 > l_2 > \ldots > l_k > l_{k+1} > \ldots > l_p$ são os autovalores não nulos de $\mathbf{Z}^T\mathbf{Z}$ ou $(\mathbf{Z}\mathbf{Z}^T)$,$\mathbf{Z} = (z_{ij}).z_{ij} = \overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}$
p=min(b-1;t-1);
$\hat{\alpha}_r$: autovetor normalizado de $\mathbf{Z}\mathbf{Z}^T$ correspondente ao autovalor não nulo, lr, $r = 1, 2, \ldots, k$.
γ_r: autovetor normalizado de $\mathbf{Z}^T\mathbf{Z}$ correspondente ao autovalor não nulo, lr, $r = 1, 2, \ldots, k$.
O sinal adequado para $\hat{\lambda}_k$ pode ser obtido assim,
$\hat{\lambda}_k$=sinal$(\hat{\alpha}_r^T\mathbf{Z}\hat{\gamma}_r)$; $r = 1, 2, \ldots, k$.

Além dos erros independentes, se eles tiverem distribuição normal, então os estimadores acima descritos são também os estimadores de máxima verossimilhança. Em experimentos $G \times E$, y_{ij} representa a resposta do i-ésimo genótipo no j-ésimo ambiente, μ é a média geral, τ_i e β_j os efeitos genotípicos e ambientais, λ_r^2 fornece a proporção da variância devido à interação $G \times E$ no r-ésimo componente e α_{ri}, γ_{rj} representam os pesos para o genótipo i e ambiente j naquele componente de interação.

1.2.1 Escolha do número apropriado de termos para descrever a interação

Na literatura existem dois tipos mais utilizados de procedimentos para determinar o número ótimo de termos no modelo AMMI (o k no modelo (4)). Um desses procedimentos consiste em fazer testes de significância dos termos multiplicativos e o outro procedimento consiste em fazer validação cruzada. Na validação cruzada os dados de repetições, para cada combinação de tratamentos são aleatoriamente divididos em dois subconjuntos, um subconjunto de dados para o ajuste do modelo e outro subconjunto para validação. As respostas preditas por um determinado modelo AMMI, são confrontadas com os respectivos dados de validação, calculando-se as diferenças entre esses valores. Obtém-se, em seguida a soma de quadrados dessas diferenças, dividindo-se o resultado pelo número delas. A raiz quadrada desse resultado chama-se diferença preditiva média. Esse método foi estudado com mais detalhes em

Dias (2005). Neste trabalho por causa da estrutura dos dados experimentais, só serão considerados testes estatísticos sobre os componentes multiplicativos, pois não se dispõe das repetições das observações e apenas está disponível a matriz de médias.

Os testes de hipóteses se fazem usando os dados completos, e os critérios adotados para a determinação do número de componentes multiplicativos tem sido objeto de várias pesquisas como as desenvolvidas por Gollob (1968), Mandel (1971), Gauch (1988) e Gauch e Zobel (1988 apud KANG; GAUCH, 1996) entre outros. Mas, levando em conta os resultados e recomendações dos estudos feitos por Milliken e Johnson (1989), Piepho (1995a), Cornelius et al. (1996), Dias e Krzanowski (2003), Dias (2005) e Dias e Krzanowski (2006), os métodos propostos para usar neste curso serão os testes de razão de verossimilhança e o teste FR, os quais serão descritos a seguir:

Testes de razão de verossimilhança

Se um pesquisador deseja usar o modelo dado em (2), então é preciso determinar o número de termos de interação multiplicativa necessário para que o modelo explique adequadamente os dados. Para tomar essa decisão, têm-se problemas muito parecidos aos problemas encontrados na construção dos modelos de regressão. O objetivo é encontrar um modelo parcimonioso (com poucos termos quanto seja possível), e ao mesmo tempo, obter um modelo adequado. Deve-se lembrar que em situações de regressão polinomial, sempre é possível ajustar um modelo de grau (n-1) a n observações, mas, tais modelos não são geralmente bons, pois funcionam bem na predição da resposta média dos valores observados, mas, podem funcionar muito mal na predição de respostas de valores não observados.

Uma situação similar se apresenta para os dados provenientes de uma estrutura de tratamentos em dupla entrada, na qual sempre é possível fazer k = min(b-1; t-1) e ajustar exatamente esses dados com o modelo (2), mas, tal modelo não será provavelmente muito bom por causa do sobreajuste dos dados e pelo fato dele explicar além do padrão de resposta presente nos dados parte do erro de medida. Assim, é desejável um modelo com poucos componentes que ofereça um ajuste ótimo e que explique boa parte do padrão de resposta dos dados. Assumindo por enquanto que se conhece o procedimento necessário para testar as hipóteses, um procedimento razoável pode ser:

1. Testar $H_{01} : \lambda_1 = 0$ vs. $H_{a1} : \lambda_1 \neq 0$ no modelo

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \varepsilon_{ij} \tag{1.5}$$

2. Se não rejeitar H_{01}, conclui-se que os dados são aditivos e deve-se completar uma análise correspondente com o resultado. No entanto, se rejeita-se H_{01}, então se deve testar
$H_{02} : \lambda_2 = 0 \ ; \ \lambda_1 \neq 0$ vs: $H_{a2} : \lambda_2 \neq 0 \ ; \ \lambda_1 \neq 0$
no modelo
$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \varepsilon_{ij}$

3. Se não rejeitar H_{02}, conclui-se que o modelo apropriado é o modelo dado em (5) e deve-se completar a análise de acordo com o resultado encontrado. Se rejeitar H_{02}, então se deve testar
$H_{03} : \lambda_3 = 0 \ ; \ \lambda_2 \neq 0$ vs: $H_{a3} : \lambda_3 \neq 0 \ ; \ \lambda_2 \neq 0; \ \lambda_1 \neq 0$
no modelo
$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \lambda_3 \alpha_{3i} \gamma_{3j} + \varepsilon_{ij}$

4. Continuar dessa forma sucessivamente até não rejeitar a hipótese.

Uma desvantagem do procedimento anterior descrito é parecida à desvantagem encontrada no procedimento de escolha sequencial desenvolvido para problemas de regressão múltipla. Por exemplo, um termo da interação pode explicar apenas uma pouca proporção da variação da interação e por tal razão esse termo pode não ser significativo. Assim, é preferível usar um procedimento parecido ao explicado anteriormente, mas com uma pequena diferença. Podem-se testar hipóteses sucessivas, depois das quais é possível concluir que o valor certo de k é o valor para o qual foi feita a última rejeição. Segundo Milliken e Johnson (1989), em geral o modelo de interação multiplicativa em aplicações sobre dados reais precisa de um máximo de dois componentes e em muitas ocasiões apenas um termo da interação é necessário.

Um teste de razão de verossimilhança para H_{01} versus H_{a1}, pode fazer-se rejeitando H_{01} se

$$U_1 = \frac{l_1}{\sum\limits_{i,j} z_{ij}^2} > C_\alpha$$

em que C_α é o ponto crítico a $\alpha(100\%)$ obtido da tabela com os pontos críticos

da distribuição de U_1 e apresentada em Milliken e Johnson (1989) (p.171) com p = min(t-1; b-1) e n = max(t-1; b-1). l_1, z_{ijj} ,t e b foram definidos anteriormente. Note-se também que U_1 é igual a

$$U_1 = \frac{l_1}{l_1 + l_2 + \ldots + l_p}$$

Agora, uma estatística de razão de verossimilhança para testar H_{02} versus H_{a2} é

$$U_2 = \frac{l_2}{l_2 + l_3 + \ldots + l_p}$$

Rejeita-se H_{02} se U_2 for maior do que o valor crítico da tabela apresentada em Milliken e Johnson (1989) (p.173). Nessa tabela encontram-se os pontos críticos aproximados da distribuição de U_2.

Em geral, as estatísticas de razão de verossimilhança para a hipótese H_{0k} versus H_{ak}, $k = 3, 4, \ldots, p-1$, são dadas por:

$$U_k = \frac{l_k}{l_k + l_k + 1 + \ldots + l_p}$$

Milliken e Johnson (1989) sugerem nestes casos usar os pontos críticos da distribuição de U_1 com p = min(t,b) - k e n = max(t,b) - k, mas, para aqueles experimentos nos quais não existe na tabela o correspondente ponto crítico (grande número de genótipos ou ambientes) Cornelius et al. (1996) apresentam uma transformação da estatística para obter um teste F aproximado. Para testar o k-ésimo termo, o teste F_{teste} aproximado pode ser encontrado como segue:

$$p = min(t-1, b-1); n = max(t-1, b-1),$$

$$Q_k = \frac{[(p-k+1)\,U_k - 1]}{(p-k)}; \qquad c_1^* = \frac{u_{1k} - (n-k+1)}{(n-k+1)(p-k)};$$

$$c_2^* = \frac{(p-k+1)\,(n-k+1)\,u_{2k}^2 - 2u_{1k}^2}{(n-k+1)^2\,[(p-k+1)\,(n-k+1)+2]\,(p-k)^2};$$

$$d^* = c_1^*(1-c_1^*) - c_2^*; \quad a^* = dc_1^*/c_2^*; \quad b^* = d^*(1-c_1^*)/c_2^*;$$

$$F_{teste} = b^* U_k / a^*\,(1-U_k)$$

A estatística F_{teste} tem uma distribuição F aproximada com graus de liberdade ($2a^*$; $2b^*$) e os valores u_{1k} e u_{2k} correspondem à esperança e ao desvio padrão de (l_1/σ^2;). Liu e Cornelius (2001) encontraram através de estudos de

simulação funções polinomiais para esses valores:

$u_{1k} = E_{r,c}(l_1/\sigma^2;) =$ - 0,64679880 + 1,0068336 (r + c) - 7,1495083 $\times 10^{-9} r^2 c^2$
+ 0,082395238 $[(\ln r)^2 + (\ln c)^2]$ + 0,53767438 $(\ln r \ln c)(\ln r + \ln c)$
-0,091580971 $(\ln r \ln c)[(\ln r)^2 + (\ln c)^2]$
+ 0,021644307 $(\ln r \ln c)[(\ln r)^3 + (\ln c)^3]$ + 5,3529799 $\times 10^{-3} (\ln r)^3 (\ln c)^3$
- 0,76227733 $(e^{-r} + e^{-c})$ -0,020829655 $(r/c + c/r)$ + 1,7482806 $\times 10^{-3}$(rc-$|r-c|$)

$u_{2k} = DPr,c(l_1/\sigma^2) =$ -0,015802857(r + c) + 2,3780161 $\times 10^{-9} rc(r^2 + c^2)$
+ 1,7371131 $(\ln r + \ln c)$ - 0,33301620 $[(\ln r)^2 + (\ln c)^2]$
+ 0,11442045$[(\ln r)^3 + (\ln c)^3]$ - 0,035296928 $(\ln r \ln c)[\ln r + \ln c]$
+ 0,033246016 $(r/c + c/r)$ - 5,7298685 $\times 10^{-9}$ $[(r/c)^4 + (c/r)^4]$
+ 1,8747692 $(e^{-r} + e^{-c})$ -1,7476731 $\times 10^{-13} r^2 c^2 (r^2 + c^2)$
+ 6,9946263 $\times 10^{-16} r^3 c^3$ (r + c) + 2,3238523 $\times 10^{-5} |r - c|^2$
em que $r = \max(t, b)$ - k e $c = \min(t, b)$ - k.

Teste F_R

Considere-se o modelo (6) aplicado em um experimento para avaliar genótipos e ambientes:

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \ldots + \lambda_k \alpha_{ki} \gamma_{kj} + \varepsilon_{ij} \qquad (1.6)$$

em que $i = 1, 2, \ldots, b$
τ_i representa os genótipos e β_j representa os ambientes, em que os componentes multiplicativos representam a interação desses fatores $(G \times E)$.
Escrevendo a soma de quadrados da interação ($G\times$ E) tem-se que

$$SQ(G\times E) = \sum_i \sum_j \left(\overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}\right)^2 \text{ com}(t-1)(b-1)\text{graus de liberdade.}$$

Essa soma de quadrados pode ser escrita como:

$$SQ(G \times E) = \sum_{r=1}^{p} l_r$$

em que, como já descrito antes, $l_1 > l_2 > ... > l_k > l_{k+1} > ... > l_p$ são os autovalores não nulos de $\mathbf{Z}^T\mathbf{Z}$ ou($\mathbf{Z}\mathbf{Z}^T$), $\mathbf{Z}$=$\mathbf{z}_{ij}$. $\mathbf{z}_{ij} = \overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}$ e $p = \min(b-1,\, t-1)$.

Então, a idéia é escolher o melhor k, de tal maneira que a soma de quadrados da interação possa ser separada em uma parte determinística (padrão) e outra parte que conterá ruído, assim,

$$SQ(G\times E) = \sum_{r=1}^{p} l_r = \sum_{r=1}^{k} l_r = \sum_{r=k+1}^{p} l_r = SQ(G \times E)_{PADRÃAO} + SQ(G \times E)_{RUÍDO}.$$

A estatística teste de Cornelius reescrita por Dias e Krzanowski (2003) com k termos multiplicativos no modelo é dada por,

$$F_{R,K} = \frac{\left(SQ(G \times E) - \sum_{r=1}^{k} l_r\right)}{f_2 \times QM\,(Erro\,médio)}$$

com $f_2 = (t-1-k)(b-1-k)$. O erro médio é originário das análises individuais de variâncias dos b experimentos. Este é o teste F_R de Cornelius et al. (1996) que sob a hipótese nula de que não mais que k termos determinam a interação, de tal forma que o teste estatístico tem uma distribuição F com f_2 gl e os graus de liberdade do quadrado médio do resíduo. Um resultado significativo para o teste sugere que no mínimo um ou mais termos multiplicativos devem ser adicionados aos k já incluídos.

Apresenta-se na tabela 1.2 a análise da variância a partir de médias segundo o sistema de Cornelius, em que n representa o número de repetições no experimento e IPCA$_i$ é a notação internacional para o i-ésimo componente da interação.

Teste de Gollob

O teste de Gollob (1968) distribui graus de liberdade às Somas de Quadrados SQ$_k$=$n\lambda_k^2$ com $k = 1, 2, \ldots, p$ e n o número de repetições, contando o número de parâmetros no k-ésimo termo multiplicativo. Logo, o teste F é calculado como na análise de variância para modelos lineares. O teste F na análise de variância supõe sob a hipótese nula que, o numerador e o denominador da estatística F são distribuídos independentemente como uma variável qui-quadrado (Cornelius et al. 1996).

O teste F de Gollob não é válido porque os autovalores λ_k^2 são distribuídos como autovalores de uma matriz de Wishart e, portanto não tem distribuição qui-quadrado, além disso, o teste assume que $n\lambda_k^2/\sigma^2$ é distribuído como qui-quadrado e então obviamente não é válido (Dias, 2005).

Tabela 1.2: Esquema da Análise de variância pelo sistema de Cornélius baseado em médias

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*
Genótipos (G)	**t - 1**	SQ(G)
Ambientes (E)	b - 1	SQ(E)
Interação (G×E)	(t – 1)(b – 1)	SQ(G×E)
IPCA1	(t – 1 - 1)(b – 1 - 1)	$\sum_{r=2}^{p} \ell_r$
IPCA2	(t – 1 - 2)(b – 1 - 2)	$\sum_{r=3}^{p} \ell_r$
IPCA3	(t – 1 - 3)(b – 1 - 3)	$\sum_{r=4}^{p} \ell_r$
...	...	...
IPCAk	(t – 1 - k)(b – 1 - k)	$\sum_{r=k}^{p} \ell_r$
...	...	...
IPCAp	-	-
Erro médio	b(t – 1)(n – 1)	
Total	tbn - 1	

$IPCA_k$: (Interaction Principal Component Analysis) modelo com k componentes k=1,2,...,p

No que se refere aos graus de liberdade, o método de Gollob é muito popular, pois o procedimento é fácil de aplicar, uma vez que o número de graus de liberdade para o k-ésimo componente da interação é simplesmente definido como $GL(IPCA_k)$=g+e-1-2k, enquanto muitos outros procedimentos requerem simulações extensivas antes de serem usadas (Dias, 2005).

1.2.2 Avaliação preditiva por validação cruzada.

Em geral, é necessário o uso de procedimentos estatísticos computacionalmente intensivos para fazer predições, daí a importância que tem ultimamente os métodos livres de distribuições teóricas como os baseados em reamostragem jacknife, bootstrap e validação cruzada. O critério preditivo de avaliação prioriza a capacidade de um modelo aproximar suas predições a dados não incluídos na análise (simulando respostas futuras ainda não mensuradas).

Um modelo que seletivamente recupera o padrão e relega ruídos a um resíduo desconsiderado na predição de respostas, pode resultar em melhor precisão do que os próprios dados. Esse é o princípio subjacente à proposta de Gauch (1988) para seleção do modelo AMMI introduzida por ele como *avaliação preditiva*.

Dessa forma, através da validação cruzada, os dados de repetições, para cada

Tabela 1.3: Esquema da Análise de variância pelo sistema de Gollob baseado em médias

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*
Genótipos (G)	t - 1	SQ(G)
Ambientes (E)	b - 1	SQ(E)
Interação (G×E)	(t – 1)(b – 1)	SQ(G×E)
IPCA1	t + b - 1 - (2 × 1)	ℓ_1
IPCA2	t + b - 1 - (2 × 2)	ℓ_2
IPCA3	t + b - 1 - (2 × 3)	ℓ_3
...	...	...
IPCAk	t + b - 1 - (2 × k)	ℓ_k
...	...	...
IPCAp	t + b - 1 - (2 × p)	ℓ_p
Erro médio	b(t – 1)(n – 1)	
Total	tbn - 1	

combinação de genótipos e ambientes, são divididos, por um critério aleatório, em dois subconjuntos: (i) dados para o ajuste do modelo AMMI; e (ii) dados de validação. As respostas preditas do modelo AMMI, são comparadas com os dados de validação, calculando-se as diferenças entre esses valores. Logo, é obtida a soma de quadrados dessas diferenças e o resultado dividido pelo número de respostas preditas. À raiz quadrada desse resultado é chamado de diferença preditiva média (RMSPD), Crosa et al. (1991) sugerem que o procedimento deve ser repetido 10 vezes, obtendo-se uma média dos resultados para cada membro da família de modelos.

Um pequeno valor de RMSPD indica sucesso preditivo do modelo, tal que o melhor modelo é aquele com o menor RMSPD. O modelo selecionado é então usado para para analisar os dados de todas as m repetições, conjuntamente, em uma análise definitiva. (Dias, 2005).

Outros autores como Piepho (1994) sugere que o valor médio de RMSPD seja obtido a partir de 1000 randomizações diferentes e não 10 como propôs Crossa et al. (1991). O autor considera uma modificação da partição completamente aleatória dos dados (modelagem e validação) quando o ensaio é em blocos. Neste caso, ele recomenda sortear o bloco inteiro de um ensaio e não fazer componentes para cada combinação de genótipo e ambiente. Assim, a estrutura original de blocos é preservada. Contudo, apesar da coerência lógica desse tipo de proposta, estudos confirmando sua efetividade ainda não estão disponíveis. Gauch e Zobel (1996) sugerem que se faça o conjunto de dados de

validação sempre com uma só observação para cada tratamento. Sendo assim, é mais provável para $m-1$ dados, encontrar um modelo que mais se aproxime do ideal para analisar o conjunto completo dos m dados por tratamento.

Segundo Duarte e Vencovsky (1999), ao avaliar o modelo por validação cruzada, a análise AMMI deve partir das observações individuais propriamente ditas (dados de cada repetição dentro de experimentos). Por outro lado, se o modelo for avaliado por um teste F a análise pode ser feita a partir das médias dos genótipos nos ambientes (experimentos), desde que se disponha dos quadrados médios residuais, obtidos nas análises de variâncias de cada experimento.

Dias e Krzanowski (2003) descrevem dois métodos que otimizam o processo de validação cruzada por validar o ajuste do modelo em cada um dos dados por vez e então combinar essa validação em uma medida simples e geral de ajuste.

Método "leave-one-out"

Dias e Krzanowski (2003) propuseram dois métodos baseados em um procedimento "leave-one-out"completo, que otimiza o processo de validação cruzada. No que segue, assume-se que se deseja predizer os elementos x_{ij} da matriz $\boldsymbol{X}$ por meio do modelo:

$$x_{ij} = \sum_{k=1}^{n} d_k u_{ik} v_{jk} + \varepsilon_{ij}$$

Os métodos são aqueles apresentados em Krzanowski (1987) e Gabriel (2002), no qual prediz-se o valor $\widehat{x}_{ij}^{n}$ de x_{ij} $(i = 1, \ldots, g;\ j = 1, \ldots, e)$ para cada possível escolha de n (o número de componentes), e a medida de discrepância entre o valor atual e predito como

$$PRESS(n) = \sum_{i=1}^{g} \sum_{j=1}^{e} (x_{ij}^{n} - x_{ij})^2$$

Contudo, para evitar viés, os dados x_{ij} não devem ser usados nos cálculos de x_{ij}^{n} para cada i e j. Como conseqüência, apelo a alguma forma de validação cruzada é indicado, e os dois procedimentos diferem na forma com que eles lidam com isso (Dias, 2005). Ambos, entretanto, assumem que a DVS de $\boldsymbol{X}$ pode ser escrita como $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}$.

O procedimento de validação cruzada padrão subdivide $\boldsymbol{X}$ em um certo número de grupos, deleta-se cada grupo por vez a partir dos dados, avalia-se os parâmetros do modelo ajustados a partir dos dados remanescentes, e prediz-se o valor deletado (Wold, 1976, 1978). Krzanowski (1987) argumenta que a predição mais precisa resulta quando cada grupo deletado é tão pequeno quanto possível, que no presente caso é um simples elemento de $\boldsymbol{X}$. Denota-se por $\boldsymbol{X^{(-i)}}$ o resultado de deletar a i-ésima linha de $\boldsymbol{X}$ e centralizar em torno das médias das colunas. Denota-se por $\boldsymbol{X_{(-j)}}$ o resultado de deletar a j-ésima coluna de $\boldsymbol{X}$ e centralizar em torno das médias das colunas, seguindo o esquema dado por Eastment and Krzanowski (1982). Então pode-se escrever

$$\boldsymbol{X_{(-i)}} = \overline{U}\,\overline{D}\,\overline{V}^T \text{ com } \overline{\boldsymbol{U}} = (\overline{u}_{pt}),\ \overline{\boldsymbol{V}} = (\overline{v}_{pt}) \text{ e } \overline{\boldsymbol{D}} = diag(\overline{d}_1, \ldots, \overline{d}_l),$$
$$\boldsymbol{X_{(-j)}} = \widetilde{U}\widetilde{D}\widetilde{V}^T \text{ com } \widetilde{\boldsymbol{U}} = (\widetilde{u}_{pt}),\ \widetilde{\boldsymbol{V}} = (\widetilde{v}_{pt}) \text{ e } \widetilde{\boldsymbol{D}} = diag(\widetilde{d}_1, \ldots, \widetilde{d}_{l-1}).$$

Agora, considere-se o preditor

$$\widehat{x}_{ij}^{n} = \sum_{t=1}^{n} \left(\widetilde{u}_{it}\sqrt{\widetilde{d}_t}\right)\left(\overline{v}_{tj}\sqrt{\overline{d}_t}\right) \tag{1.7}$$

Cada elemento no lado direito da equação (1.7) é obtido da DVS de $\boldsymbol{X}$ centrada na média após omitir a i-ésima linha e a j-ésima coluna. Assim, o valor x_{ij} não é usado no cálculo da predição, e o máximo uso dos dados é feito com os outros elementos de $\boldsymbol{X}$. Os cálculos aqui são exatos, assim não há problema com a convergência como nos procedimentos de maximização que têm sido aplicados ao modelo AMMI, mas que não garantem a convergência (Dias e Krzanowski, 2003).

Gabriel (2002), tomou uma mistura de regressão e aproximação de uma matriz de posto inferior como a base para sua predição. O algoritmo para validação cruzada de aproximações de posto inferior proposto pelo autor é como segue: Para a matrix $\boldsymbol{X}$ (GE), se usa a partição

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \boldsymbol{X}_{1\cdot}^T \\ \boldsymbol{X}_{\cdot 1} & \boldsymbol{X}_{|11} \end{bmatrix}$$

e o ajuste aproximado da sub-matriz $\boldsymbol{X_{|11}}$ de posto n usando a DVS é:

$$\boldsymbol{X_{|11}} = \sum_{k=1}^{n} \boldsymbol{u}_{(k)} d_k \boldsymbol{v}_{(k)}^T = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}},$$

em que $\boldsymbol{U} = [\boldsymbol{u_1}, \ldots, \boldsymbol{u_n}]$, $\boldsymbol{V} = [\boldsymbol{v_1}, \ldots, \boldsymbol{v_n}]$ e $\boldsymbol{D}$= $diag(d_1, \ldots, d_n)$.

Então prediz-se x_{11} por $\widehat{x}_{11} = \boldsymbol{X}_{1.}^T \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{X}_{.1}$ e obtém-se o resíduo de validação cruzada $e_{11} = x_{11} - \widehat{x}_{11}$.

Similarmente, obtém-se o valor ajustado da validação cruzada x_{ij} e os resíduos $e_{ij} = x_{ij} - \widehat{x}_{ij}$ para todos os outros elementos x_{ij}, $i = 1, \ldots, g$; $j = 1, \ldots, n$; $(i, j) \neq (1, 1)$. Cada um irá requerer uma partição diferente de $\boldsymbol{X}$.

Esses resíduos e valores ajustados podem ser sumarizados por

$$PRESS(n) = \frac{1}{ge}\sum_{i=1}^{g}\sum_{j=1}^{e} e_{ij}^2 \quad \text{e} \quad PRECORR(n) = Corr(x_{ij}, \widehat{x}_{ij}|\forall i, j),$$

respectivamente.

Com cada método, a escolha de n pode ser baseada em alguma função apropriada de

$$PRESS(n) = \frac{1}{ge}\sum_{i=1}^{g}\sum_{j=1}^{e}(x_{ij}^n - x_{ij})^2$$

Contudo, as características dessa estatística diferem para os dois métodos. O procedimento de Gabriel produz valores que primeiro decresce e então (usualmente) cresce com n. Por essa razão ele sugere que o valor ótimo de n seja aquele que produz o mínimo da função PRESS. O procedimento de Eastment-Krzanowski produz, geralmente, um conjunto de valores que é monotonicamente não-crescente com n (Dias, 2005). Por isso, sugerem o uso de

$$W_n = \frac{\frac{PRESS(n-1) - PRESS(n)}{D_n}}{\frac{PRESS(n)}{D_r}}$$

em que D_n é o número de graus de liberdade requeridos para ajustar o n-ésimo componente e D_r é o número de graus de liberdade remanescentes após ajustar o n-ésimo componente. Considerações sobre o número de parâmetros a serem estimados juntos com todas as restrições nos autovetores em cada estágio, mostra que $D_n = g + e - 2n$. D_r pode ser obtido por sucessivas subtrações, dando $(g-1)e$ graus de liberdade na matriz centrada na média $\boldsymbol{X}$, isto é, $D_1 = (g-1)e$ e $D_r = D_{r-1} - [g + e - (n-1)2]$, $r = 2, 3, \ldots, (g-1)$, (Wold, 1978). W_n representa o aumento na informação preditiva suprida pelo n-ésimo componente, dividido pela informação preditiva média em cada um dos componentes remanescentes. Assim, importantes componentes devem produzir valores de W_n maiores que a unidade. Baseando-se a escolha de n em W_n pode ser vista como uma natural seleção de um melhor conjunto de variáveis

regressoras ortogonais em análise de regressão múltipla (Dias e Krzanowski, 2003).

Ao nível computacional, a melhor precisão parece ser obtida quando as entradas (x_{ij}) em diferentes colunas de $\boldsymbol{X}$ são comparáveis em tamanho e existe relativamente pouca variação entre os d_i. O procedimento mais estável é, portanto, aquele no qual a média $\overline{x}_j$ e o desvio padrão s_j da coluna j ($j = 1, \ldots, e$) são primeiro calculados dos valores presentes naquela coluna. As entradas existentes x_{ij} de $\boldsymbol{X}$ são então padronizadas para $x'_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$, e estimativas são obtidas pela aplicação de

$$\widehat{x}_{ij} = \boldsymbol{x_{i.}^T V D^{-1} U^T X_{.j}}$$

aos dados padronizados, e então os valores finais são obtidos de

$$\widehat{x}_{ij} = \overline{x}_j + s_j x'_{ij}$$

Voltando ao caso dos dados de genótipo-ambiente, fica claro aqui que $\boldsymbol{X}$ deve ser a matriz de interações previamente denotada por GE. Contudo, desde que se está simplesmente procurando pelo número apropriado de termos multiplicativos no modelo, e qualquer constante aditiva pode ser absorvida no componente ε_{ij} do modelo, pode-se aplicar o procedimento "leave-one-out"diretamente à matriz $\boldsymbol{Y}$ de dados. De fato, isso pode freqüentemente ser preferível dado os valores pequenos tomados por muitos elementos de GE.

Cornelius et al. (1993) compararam resultados de validação cruzada com aqueles obtidos após calcular a estatística PRESS nos modelos multiplicativos em dados MET (MultiEnvironment Trials) completos. A partição dos dados envolveu três repetições para modelagem e uma repetição para validação. Calcularam o RMSPD da estatística PRESS ajustando os valores de PRESS como $[PRESS/ge + 3s^2/4]^{\frac{1}{2}}$, em que g e e denotam o número de genótipos e ambientes no MET e s^2 é a variância residual conjunta dentro de ambientes.

O termo em s^2 é um ajuste para a diferença em variância da validação dos dados nas médias de caselas, para tomar os resultados comparáveis ao RMSPD da divisão $3 - 1$ dos dados. Resultados em um MET com nove genótipos e vinte ambientes mostrou que PRESS é mais sensível a super ajuste do que os dados divididos (Dias, 2005). Neste trabalho será usado a metodologia de Eastment-Krzanowski, por apresentar melhores resultados, conforme (Dias, 2005).

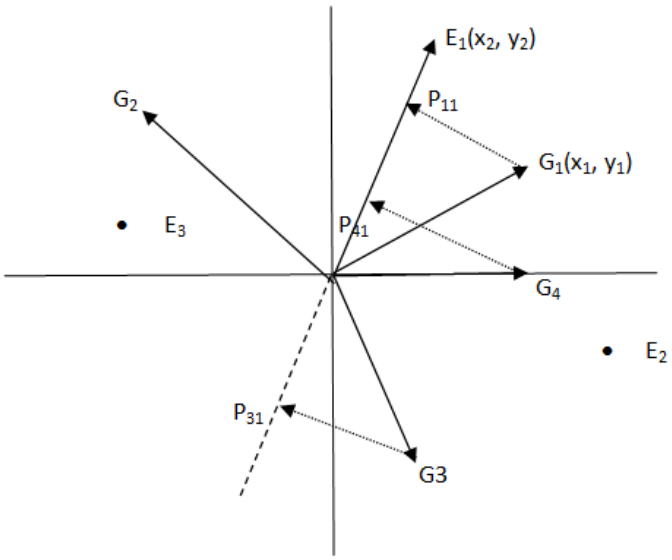

Figura 1.4: A geometria do biplot. Visualizando as projeções dos genótipos G_1, G_2, G_3 e G_4 sobre o ambiente E_1.

1.3 Biplot

1.3.1 Introdução

Muitos estudos observacionais ou experimentais produzem uma tabela de dupla entrada de dados a ser analisada. A origem mais comum de tais dados é de um experimento de dois fatores; se um fator tem g níveis, o segundo tem e níveis, e há n observações repetidas em cada combinação de níveis de fator, então os resultados experimentais serão dispostos em uma tabela tendo $(g \times n)$ linhas e e colunas. O conceito de biplot foi desenvolvido por Gabriel (1971). É uma representação gráfica que apresenta ambas as entradas (por exemplo, cultivares) e os testadores (por exemplo, ambientes) de um conjunto de dados em uma tabela de dupla entrada. Na produção de plantas e em dados genéticos, os testadores também podem ser características, marcadores genéticos, anos etc. O biplot permite a visualização dos dados conforme as seguintes propriedades: a) inter-relação entre as entradas (por exemplo, genótipos); b) inter-relação entre os testadores (por exemplo, ambientes); c) inter-relação entre as entradas e os testadores. Trata-se de uma representação gráfica da informação em uma matriz $n \times e$. O "b" refere-se aos dois tipos de informações contidas em uma matriz de dados. As informações nas linhas pertencem a amostras ou unidades amostrais e aquelas nas colunas pertencem as variáveis. Esta representação gráfica permite a inspeção visual da posição de uma unidade amostral

relativa à outra e a importância relativa de cada uma das variáveis à posição de qualquer unidade. Assim pode-se ver como as unidades amostrais se agrupam e quais variáveis contribuem para sua posição dentro dessa representação.

1.3.2 Fundamentação geométrica do Biplot

A análise Biplot é utilizada com quaisquer tipos de variáveis (contínuas ou discretas), quando a finalidade é aproximar os dados originais e realizar uma análise simultânea das relações entre indivíduos e/ou variáveis. A fundamentação teórica se baseia na aproximação da matriz Y de ordem $(g \times e)$ de posto r por outra matriz de posto q, em que $(q < r)$, por meio de sua DVS.

1.3.3 Construção do Biplot - Gabriel (1971)

A construção de um biplot origina-se dos componentes principais amostrais.

Seja ${}_n\mathbf{X}_p$. Procura-se por uma aproximação ${}_n\mathbf{Y}_p$ de posto 2 da matriz original $\mathbf{X}$, e então obtemos um exato biplot de $\mathbf{Y}$.

A melhor aproximação de posto 2 $\mathbf{Y}$ de $\mathbf{X}$ é obtida pela decomposição singular de $\mathbf{X}$:

$$ {}_n\mathbf{X}_p = {}_n\mathbf{U}_p\mathbf{\Lambda}_p\mathbf{V'}_p $$

com $\mathbf{\Lambda}$=diag$(\lambda_1, \lambda_2, \ldots, \lambda_p)$

$\mathbf{U}$: matriz diagonal de autovetores de ${}_n\mathbf{X}_{pp}\mathbf{X'}_n$;

$\mathbf{V}$: matriz ortogonal de autovetores de ${}_p\mathbf{X'}_{nn}\mathbf{X}_p$.

Assim,

$$ \mathbf{X} = \lambda_\mathbf{1}\mathbf{u_1}\mathbf{v'_1} + \lambda_\mathbf{2}\mathbf{u_2}\mathbf{v'_2} + \ldots + \lambda_\mathbf{p}\mathbf{u_p}\mathbf{v'_p} $$

$$ \mathbf{X} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_\mathbf{i}\mathbf{u_i}\mathbf{v'_i}, \quad \lambda_\mathbf{1} \geq \lambda_\mathbf{2} \geq, \ldots, \geq \lambda_\mathbf{p}. $$

Então,

$$ {}_n\mathbf{Y_p} = \lambda_\mathbf{1}\mathbf{u_1}\mathbf{v'_1} + \lambda_\mathbf{2}\mathbf{u_2}\mathbf{v'_2}, \quad \text{Posto}(\mathbf{Y}) = \mathbf{2} $$

$$ {}_n\mathbf{Y_p} = \left[\begin{array}{cc} \mathbf{u_1} & \mathbf{u_2} \end{array} \right] \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array} \right] \left[\begin{array}{c} \mathbf{v_1'} \\ \mathbf{v_2'} \end{array} \right] $$

$$ {}_n\mathbf{Y_p} = \left[\begin{array}{cc} \mathbf{u_{11}} & \mathbf{u_{21}} \\ \mathbf{u_{12}} & \mathbf{u_{22}} \\ \vdots & \vdots \\ \mathbf{u_{1n}} & \mathbf{u_{2n}} \end{array} \right] \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array} \right] \left[\begin{array}{cccc} v_{11} & v_{12} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \end{array} \right] \tag{1.8} $$

Para obter o biplot, é necessário escrever Y como o produto de duas matrizes $\mathbf{GH'}$, em que $\mathbf{G}$ é uma matriz $(n \times 2)$ e $\mathbf{H}$ é uma matriz $(p \times 2)$. Isso pode ser feito de várias formas, mas (1) sugere três fatorações bem simples:

1.

$$ {}_n\mathbf{Y_p} = \underbrace{\left[\begin{array}{cc} \mathbf{u}_{11}\sqrt{\lambda_1} & \mathbf{u}_{21}\sqrt{\lambda_2} \\ \mathbf{u}_{12}\sqrt{\lambda_1} & \mathbf{u}_{22}\sqrt{\lambda_2} \\ \vdots & \vdots \\ \mathbf{u}_{1n}\sqrt{\lambda_1} & \mathbf{u}_{2n}\sqrt{\lambda_2} \end{array} \right]}_{\mathbf{G}_1} \underbrace{\left[\begin{array}{cccc} v_{11}\sqrt{\lambda_1} & v_{21}\sqrt{\lambda_2} & \ldots & v_{1p}\sqrt{\lambda_1} \\ v_{21}\sqrt{\lambda_2} & v_{22}\sqrt{\lambda_2} & \ldots & v_{2p}\sqrt{\lambda_2} \end{array} \right]}_{\mathbf{H'}_1} \tag{1.9} $$

2.

$$ {}_n\mathbf{Y_p} = \underbrace{\left[\begin{array}{cc} \mathbf{u}_{11}\lambda_1 & \mathbf{u}_{21}\lambda_2 \\ \mathbf{u}_{12}\lambda_1 & \mathbf{u}_{22}\lambda_2 \\ \vdots & \vdots \\ \mathbf{u}_{1n}\lambda_1 & \mathbf{u}_{2n}\lambda_2 \end{array} \right]}_{\mathbf{G}_2} \underbrace{\left[\begin{array}{cccc} v_{11} & v_{21} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \end{array} \right]}_{\mathbf{H'}_2} \tag{1.10} $$

3.

$$ {}_n\mathbf{Y_p} = \underbrace{\left[\begin{array}{cc} \mathbf{u}_{11} & \mathbf{u}_{21} \\ \mathbf{u}_{12} & \mathbf{u}_{22} \\ \vdots & \vdots \\ \mathbf{u}_{1n} & \mathbf{u}_{2n} \end{array} \right]}_{\mathbf{G}_3} \underbrace{\left[\begin{array}{cccc} v_{11}\lambda_1 & v_{21}\lambda_2 & \ldots & v_{1p}\lambda_1 \\ v_{21}\lambda_2 & v_{22}\lambda_2 & \ldots & v_{2p}\lambda_2 \end{array} \right]}_{\mathbf{H'}_3} \tag{1.11} $$

Agora, considere a i-ésima linha de $\mathbf{G}$ ($\mathbf{g_i'}$) e a i-ésima coluna de $\mathbf{H'}$ ($\mathbf{h_i}$).

$$ {}_n\mathbf{Y_p} = \left[\begin{array}{c} \mathbf{g'}_1 \\ \mathbf{g'}_2 \\ \vdots \\ \mathbf{g'}_n \end{array} \right] \left[\begin{array}{cccc} \mathbf{h}_1 & \mathbf{h}_2 & \ldots & \mathbf{h}_p \end{array} \right] $$

No caso 1, temos $\mathbf{g'}_i = \left[\ \mathbf{u}_{1i}\sqrt{\lambda_1}\ \ \mathbf{u}_{2i}\sqrt{\lambda_2}\ \right]$ $(i = 1, 2, \ldots, n)$ e $\mathbf{h'}_j = \left[\ v_{1i}\sqrt{\lambda_1}\ \ u_{2i}\sqrt{\lambda_2}\ \right]$ $1, 2, \ldots, p)$

O biplot consiste em plotar os $(n + p)$ vetores de $\mathbf{g'}_i (i = 1, 2, \ldots, n)$ e $\mathbf{h'}_j (j = 1, 2, \ldots, p)$ em um plano. Os vetores $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ são chamados "efeitos ou marcas de linhas" de $\mathbf{Y}$, enquanto os vetores $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p$ são chamados "efeitos ou marcas de colunas" de $\mathbf{Y}$. Eles são vetores que partem da origem até o ponto de coordenadas dadas pelos elementos de $\mathbf{g}_i$ e $\mathbf{h}_j$.

Cada elemento de y_{ij} de $\mathbf{Y}$ é representado como produto interno $\mathbf{g}_i\mathbf{h}_j$ dos correspondentes vetores "efeito linha" e "efeito coluna". O comprimento da projeção de $\mathbf{g}_i$ sobre $\mathbf{h}_j$ é dado por:

$$\mathbf{L}_{P_{g_i/h_j}} = \frac{|\mathbf{g'}_i\mathbf{h}_j|}{\mathbf{L}_{\mathbf{h}_j}} \Rightarrow |\mathbf{g'}_i\mathbf{h}_j| = \begin{cases} \mathbf{L}_{\mathbf{h}_j}\mathbf{L}_{P_{g_i/h_j}} \\ \text{ou} = L_{\mathbf{g}_j}L_{\mathbf{h}_j}\cos(\theta) \\ \mathbf{L}_{\mathbf{g}_j}\mathbf{L}_{P_{h_i/g_j}} \end{cases}$$

O produto interno dos vetores $(\mathbf{g'}_i\mathbf{h}_j)$ pode ser visualizado como o produto do comprimento de um dos vetores vezes o comprimento da projeção do outro no primeiro. Isto é útil em permitir rápida avaliação visual da estrutura da matriz. Por exemplo, pode-se rapidamente ver quais linhas ou colunas são proporcionais a outras linhas ou colunas (mesma direção dos vetores correspondentes), e quais linhas ou colunas são zeros (ângulo reto entre efeitos linhas e colunas).

A fatoração (1.9), corresponde a uma fatoração geral onde nenhuma ênfase é dada a linha ou coluna de $\mathbf{Y}$. A fatoração (1.10), coloca ênfase nas linhas de $\mathbf{Y}$. A fatoração (1.11), coloca ênfase nas colunas de $\mathbf{Y}$.

Se for de interesse que as relações entre linhas de $\mathbf{Y}$ sejam representadas pelas correspondentes relações dos vetores $\mathbf{g}$, as seguintes condições devem ser satisfeitas para quaisquer duas linhas $\mathbf{y'}_i$ e $\mathbf{y'}_j$ de $\mathbf{Y}$. Mas, $\mathbf{y'}_i\mathbf{y'}_j = \mathbf{g'}_i\mathbf{g'}_j$ equivale a $|\mathbf{y'}_i| = |\mathbf{g'}_j|$, isto é, $\mathbf{L}_{\mathbf{y}_i} = \mathbf{L}_{\mathbf{g}_i}$. Ainda, $|\mathbf{y}_i - \mathbf{y}_j| = |\mathbf{g}_i - \mathbf{g}_j|$ e $\cos(\mathbf{y}_i, \mathbf{y}_j) = \cos(\mathbf{g}_i, \mathbf{g}_j)$.

Essas relações serão satisfeitas se $\mathbf{YY'} = \mathbf{GG'}$ a qual requer $\mathbf{H'H} = \mathbf{I_2}$ a qual é satisfeita por (1.10). Isto é, $\mathbf{YY'} = \mathbf{GH'}(\mathbf{GH'})' = \mathbf{GG'}$. De forma similar, as relações entre colunas de $\mathbf{Y}$ são representadas pela correspondente relação dos vetores $\mathbf{h}$ se $\mathbf{Y'Y} = \mathbf{HH'}$, a qual requer $\mathbf{G'G} = \mathbf{I}_2$ que é satisfeita por (1.11). Isto é, $\mathbf{Y'Y} = (\mathbf{GH'})'\mathbf{GH'} = \mathbf{HG'GH'} = \mathbf{HH'}$

1.3.4 Ilustração

Ex: $\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$

$$\mathbf{G} = \begin{bmatrix} 0,4177 & 0,3624 \\ 0,5172 & 0,4744 \\ 0,4769 & 0,1223 \\ 0,5749 & -0,7923 \end{bmatrix} \begin{bmatrix} \sqrt{100,9824} & 0 \\ 0 & \sqrt{100,9824} \end{bmatrix} = \begin{bmatrix} 4,1977 & 0,5067 \\ 5,1976 & 0,6614 \\ 4,793 & 0,1705 \\ 5,7772 & -1,1047 \end{bmatrix}$$

$$\mathbf{H}' = \begin{bmatrix} \sqrt{100,9824} & 0 \\ 0 & \sqrt{100,9824} \end{bmatrix} = \begin{bmatrix} 0,9969 & 0,0781 \\ -0,0781 & 0,9969 \end{bmatrix} = \begin{bmatrix} 10,0184 & 0,7850 \\ -0,1089 & 1,3900 \end{bmatrix}$$

Programa em SAS para biplot

```
/* AMMI_Exemplo */
proc iml;
reset print;
X={2 3 4,
   1 7 5,
   8 5 9,
   3 4 1};
L=nrow(X);
C=ncol(X);
XTOTC=repeat(X[+,],L,1);
XTOTL=repeat(X[,+],1,C);
XTOTG=repeat(X[+,+],L,C);
MEDIAC=XTOTC/L;
MEDIAL=XTOTL/C;
MEDIAG=XTOTG/(L*C);
GE=X-MEDIAC-MEDIAL+MEDIAG;
call svd (U,LAMBDA,V,GE);
```

```
VL=t(V);
GE1=LAMBDA[1]*U[,1]*VL[1,];
GE2=LAMBDA[2]*U[,2]*VL[2,];
GE3=LAMBDA[3]*U[,3]*VL[3,];
VER=GE1+GE2+GE3;
print "Marcadores para Linhas";
G=U*sqrt(diag(LAMBDA));
print "Marcadores para coluna";
HL=sqrt(diag(LAMBDA))*VL;
H=t(HL);
COODL=G[,1:2];
COODC=H[,1:2];
COODGERA=COODL//(COODC);
cols=ncol(COODGERA);
call symput("cols",compress(char(cols[1,1])));


  MatColNames=("X1":"X&cols");


  create COORDENA from COODGERA [colname=MatColNames];

  append from COODGERA;

quit;

data NOMES;
input NOMES $;
cards;
G1
G2
G3
G4
A1
A2
A3
;
```

```
data TODOS;
merge NOMES COORDENA;
proc plot data=TODOS;
plot X1*X2='*' $ NOMES;
run;
quit;
```

Capítulo 2

Distribuição dos autovalores

Este capítulo visou estudar a distribuição empírica dos autovalores e calcular o intervalo de confiança, com o auxílio da reamostragem *Bootstrap* não-paramétrica no modelo biométrico AMMI (modelos de efeitos principais aditivos e interação multiplicativa). Os ensaios foram realizados em vinte ambientes, o delineamento experimental foi o aleatorizado em blocos completos, com quatro repetições. O modelo AMMI permitiu identificar as melhores combinações entre genótipos e ambientes em relação à variável resposta, os efeitos de ambientes, interação ($G \times E$) e os três dos oitos eixos da análise de componentes principais da interação foram significativos ($p < 0{,}01$), o que levaria à seleção do modelo AMMI3. Os componentes principais explicaram 81,7% da soma de quadrados da interação ($G \times E$). Com a utilização do método *bootstrap* não-paramétrico aplicado à matriz de resíduos, constitui uma eficiente alternativa para encontrar a distribuição empírica dos autovalores e com a aplicação do teste de normalidade de Shapiro-Wilk para cada autovalor, concluímos que o primeiro autovalor até o quinto autovalor apresentaram uma distribuição normal, a partir do sexto autovalor até o oitavo autovalor não apresentaram o mesmo resultado anterior.

Palavras-chave: Distribuição empírica; Autovalores; Intervalo de confiança; *Bootstrap*; Modelos AMMI

2.1 Introdução

Muitos dados coletados em experimentos agronômicos são multivariados, e recebem influência de vários fatores como por exemplo, o ambiente onde foi

realizado o experimento, genótipos, tratamentos agronômicos etc Dias e Krzanowski (2003). A expressão das características das plantas cultivadas está ligada ao controle genético, ao ambiente em que são cultivadas e à interação entre esses dois fatores (Yan e Kang, 2003; Mohammadi e Amri, 2009). A resposta distinta dos genótipos em diferentes condições ambientais é denominada de interação genótipo $\times$ ambiente ($G \times E$), a qual reduz a correlação entre os valores fenotípicos e genotípicos e dificulta a seleção e recomendação de genótipos adaptados e estáveis (Yan e Holland, 2010; Warzecha et al., 2011).

Em virtude disto, diversos métodos estatísticos destinados à avaliação da interação $G \times E$ estão disponíveis no sentido de entender melhor este efeito e a escolha do método mais adequado depende dos dados experimentais, especialmente dos números de ambientes disponíveis, da precisão requerida e do tipo de informação desejada Costa et al. (2010). Sendo assim, o modelo AMMI (modelos de efeitos principais aditivos e interação multiplicativa) nos últimos anos vem se destacando e ganhando grande aplicabilidade (Gollob, 1968; Freire Filho et al., 2003; Warzecha et al., 2011).

Este capítulo foi estudado a metodologia AMMI-*bootstrap* Lavoranti (2003) e a técnica de reamostragem *bootstrap* não-paramétrico. O método *bootstrap* foi proposto por Efron (1979), como um procedimento de reamostragem amplamente utilizado na obtenção de estimativas pontuais e intervalares, bem como na avaliação da acurácia de estimativas e testes. É sabido que as distribuições nulas de estatísticas comumente empregadas a ajustes estatísticos não seguem uma distribuição com uma forma paramétrica conhecida (Karabatsos, 2000; Wang e Chen, 2005). Nesse caso o *bootstrap* é muito útil, pois é uma técnica que não exige diferentes fórmulas para cada problema e pode ser utilizada em casos gerais, não dependendo da distribuição original da estatística do parâmetro estudado.

O *bootstrap* pode ser implementado tanto na estatística não-paramétrica quanto na paramétrica. No caso não-paramétrico, o método *bootstrap* reamostra os dados com reposição, de acordo com uma distribuição empírica estimada, tendo em vista que, em geral, não se conhece a distribuição subjacente aos dados. No caso paramétrico, quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra *bootstrap* é formada realizando-se 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 Lavoranti

(2003).

O objetivo deste trabalho foi propor procedimento empírico de *bootstrap* não-paramétrico aplicada à matriz de resíduos, estimada via metodologia AMMI, para encontrar a distribuição empírica dos autovalores, com a aplicação do teste de normalidade de Shapiro-Wilk para cada autovalor e calcular o intervalo de confiança *Bootstrap*. Com o estudo da distribuição empírica dos autovalores servirá para validar os testes de hipóteses propostos na literatura para identificar o número de $IPCA$ (*Incremental Principal Component Analysis*) para seleção dos modelos AMMI e propor um teste para seleção dos modelos.

2.2 Material e Métodos

Os dados utilizados neste trabalho são os mesmos utilizados por Cornelius e Crossa (1999), Dias e Krzanowski (2003) e Araujo (2005). Foram obtidos pelo CIMMYT (CENTRO INTERNACIONAL DE MEJORAMIENTO DE MAIZ Y TRIGO) em experimentos realizados em 34 países, caracterizando-se, experimentos multiambientais. Foram utilizados 9 genótipos de milho, Cada genótipo foi avaliado em 20 ambientes com 4 blocos, caracterizando assim, um delineamento aleatorizado em blocos.

O modelo AMMI pressupõe componentes aditivos para os efeitos principais de genótipos (g_i) e de ambientes (e_j) e componentes multiplicativos para o efeito da interação $(ge)_{ij}$. Assim, a resposta média do i-ésimo genótipo no j-ésimo ambiente é representada pelo seguinte modelo matemática:

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{n} \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij} + \varepsilon_{ij}; \tag{2.1}$$

em que: y_{ij} é a média do i-ésimo genótipo no j-ésimo ambiente, com i=1, 2, ..., g e j=1, 2, ..., e; μ é uma constante e geralmente é a média geral; g_i e e_j são os efeitos do i-ésimo genótipo e j-ésimo ambiente, respectivamente; λ_k é o k-ésimo valor singular da matriz de interação **GE** com $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$; γ_{ik} e α_{jk} são elementos dos k-ésimos valores singulares correspondentes ao i-ésimo genótipo e j-ésimo ambiente respectivamente; ρ_{ij} é o resíduo da interação $G \times E$; n é o número de eixos ou componentes principais (PC) retidos pelo modelo; r é o número de blocos e ε_{ij} é o erro médio experimental

associado ao i-ésimo genótipo no j-ésimo ambiente, assumidos independentes e $\varepsilon_{ij} \overset{IID}{\sim} N(0, \frac{\sigma^2}{r})$; O qual sugere que os dados observados, possam ser organizados em uma matriz de ordem $g \times e$.

O desenvolvimento da metodologia AMMI-*bootstrap* (Lavoranti, 2003) consiste em se executar o método AMMI seguindo a proposta de Gollob (1968), a qual garante pelo menor rigor dos graus de liberdade, um número maior de fatores multiplicativos e conseqüentemente, capta todo o padrão devido à interação $G \times E$ (Maia et al., 2006). A partir desse modelo, foram determinadas as estimativas das médias dos genótipos (i) nos ambientes (j) $(\hat{Y}_{ij})$ livres de interferência de ruídos, e, a partir desse novo conjunto de dados, foram obtidas as matrizes de resíduos ou dos efeitos $(ge)_{ij}$, $\left|\widehat{\boldsymbol{GE}}_{G\times E} = (\widehat{ge}_{ij})\right|$ representadas por:

$$\widehat{\boldsymbol{GE}}_{G\times E} = \begin{vmatrix} \widehat{ge}_{11} & \widehat{ge}_{12} & \cdots & \widehat{ge}_{1e} \\ \widehat{ge}_{21} & \widehat{ge}_{22} & \cdots & \widehat{ge}_{2e} \\ \vdots & \vdots & \cdots & \vdots \\ \widehat{ge}_{g1} & \widehat{ge}_{g2} & \cdots & \widehat{ge}_{ge} \end{vmatrix},$$

com:

$$\widehat{ge}_{ij} = Y_{ij} - \hat{\overline{Y}}_{i.} - \hat{\overline{Y}}_{.j} + \hat{\overline{Y}}_{..} \ , \tag{2.2}$$

em que $\hat{\overline{Y}}_{i.}$: é a média dos valores estimados do genótipo i; $\hat{\overline{Y}}_{.j}$: é a média dos valores estimados do ambiente j; $\hat{\overline{Y}}_{..}$: é a média geral dos valores estimados.

Das matrizes $\widehat{\boldsymbol{GE}}_{G\times E}$, podemos reamostrar B matrizes *bootstrap* para genótipos $(\boldsymbol{GE}^*_g)$, sendo o sorteio com reposição executado nas linhas; ou também pode reamostrar B matrizes *bootstrap* para ambientes $(\boldsymbol{GE}^*_e)$, sendo o sorteio com reposição executado nas colunas, B é o número de repetição *bootstrap*. Assim, obteve-se (Maia et al., 2006):

$$\boldsymbol{GE}^*_{kl} = \begin{vmatrix} ge^*_{11} & ge^*_{12} & \cdots & ge^*_{1e} \\ ge^*_{21} & ge^*_{22} & \cdots & ge^*_{2e} \\ \vdots & \vdots & \cdots & \vdots \\ ge^*_{g1} & ge^*_{g2} & \cdots & ge^*_{ge} \end{vmatrix},$$

com: $k = g$ ou $k = e$, para genótipos ou ambientes, respectivamente e $l = (1, 2, \cdots, B)$.

A partir da matriz de interação $\widehat{\boldsymbol{GE}}_{G\times E}$, utilizando a técnica de decomposição por valores singulares (DVS) obtendo-se os valores singulares que

é dado por λ_s, com $s = 1, 2, \cdots, p$, $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ e $p = min\{g-1, e-1\}$, ou pela técnica da análise de componentes principais (ACP) obtendo-se um conjunto de p autovalores a partir de $(\widehat{\boldsymbol{GE}}_{G\times E})(\widehat{\boldsymbol{GE}}^{T}_{G\times E})$ ou $(\widehat{\boldsymbol{GE}}^{T}_{G\times E})(\widehat{\boldsymbol{GE}}_{G\times E})$, com $\lambda_1^2 > \lambda_2^2 > \cdots > \lambda_p^2 > 0$.

Da mesma forma podemos encontrar os autovalores das matrizes $\boldsymbol{GE}^*_{kl}$ gerada a partir da $\widehat{\boldsymbol{GE}}_{G\times E}$ e estimar outros conjuntos de autovalores $\lambda_{1i}^2 > \lambda_{2i}^2 > \cdots > \lambda_{pi}^2 > 0$, com $k = g$ ou $k = e$, para genótipos ou ambientes, respectivamente, $l = 1, 2, \cdots, B$ e $i = 1, 2, \cdots, B$. Como as matrizes $\boldsymbol{GE}^*_{kl}$ são as estimativas de $\widehat{\boldsymbol{GE}}_{G\times E}$, temos que os λ_{ki}^2 são estimativas de λ_s^2 com $s = 1, 2, \cdots, p$ Muirhead (1987). Segundo esse autor λ_{11}^2 superestimar λ_1^2, enquanto que λ_{p1}^2 subestimar λ_p^2, e esses λ_{ps}^2 tendem a algum valor central.

Seja uma amostra aleatória baseada em n observações independentes $x_1, x_2, \ldots, x_n$. O erro padrão de uma média $\bar{\mathrm{x}}$ baseada nesta amostra é estimado pela expressão:

$$\widehat{ep}(\bar{x}) = \sqrt{\frac{s^2}{n}}\ , \tag{2.3}$$

em que

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.4}$$

é o estimador não viciado da variância. Note-se que o erro padrão não é uma estimativa de uma quantidade pertinente a uma população, mas uma medida da incerteza da média amostral vista como uma estimativa da média populacional (Altman, 1991). A expressão (2.3) deixa claro que a magnitude desta incerteza diminui conforme o tamanho da amostra n aumenta (Martinez e Louzada-Neto, 2001).

Seja X uma variável aleatória com distribuição F, sendo sua esperança denotada por $\mu_{_F}$ e sua variância denotada por $\sigma^2_{_F}$. Usaremos a notação $X \sim (\mu_{_F}, \sigma^2_{_F})$ e escrevemos $\hat{F}_n \to (x_1, x_2, \ldots, x_n)$ para indicar que $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ é uma amostra aleatória de tamanho n obtida de uma população com função de probabilidade F. A média $\bar{\mathrm{X}}$ é também uma variável aleatória e tem esperança $\mu_{_F}$ e variância $\sigma^2_{_F}/n$, ou seja, $\bar{X} \sim (\mu_{_F}, \sigma^2_{_F}/n)$. Note-se então que X e $\bar{X}$ têm a mesma esperança, entretanto, o desvio padrão de $\bar{X}$ é definido como a raiz quadrada da variância de $\bar{X}$ (Martinez e Louzada-Neto, 2001), ou seja,

$$ep_{_F}(\bar{X}) = \sqrt{\text{var}_{_F}(\bar{X})} = \frac{\sigma_{_F}}{\sqrt{n}} \tag{2.5}$$

Observada uma amostra aleatória de tamanho n oriunda de uma distribuição F, define-se uma função distribuição empírica $\hat{F}_n$ como uma distribuição discreta, que atribui probabilidade n^{-1} a cada valor x_i, $i = 1, \ldots, n$. Uma amostra *bootstrap* $\boldsymbol{x}^* = (x_1{}^*, x_2{}^*, \ldots, x_n{}^*)$ é obtida reamostrando aleatoriamente n vezes, com reposição, as observações $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, onde verifica-se que $\hat{\boldsymbol{F}}_n \to (x_1{}^*, x_2{}^*, \ldots, x_n{}^*)$.

Selecionadas B amostras *bootstrap*, $\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*B}$, de forma independente, estima-se θ em cada uma destas amostras por meio de $\hat{\theta}^{*b} = s(\boldsymbol{x}^{*b})$, $b = 1, 2, \ldots, B$. Uma expressão para o estimador *bootstrap* do erro padrão da estatística $\hat{\theta}$ é dada por

$$\widehat{ep}_{boot} = \sqrt{\frac{\sum_{b=1}^{B} \left(s(\boldsymbol{x}^{*b}) - s(.)\right)^2}{B-1}} \tag{2.6}$$

em que

$$s(.) = \frac{\sum_{b=1}^{B} s(\boldsymbol{x}^{*b})}{B} \tag{2.7}$$

O estimador *bootstrap* ideal de $ep_{_F}(\hat{\theta})$ o limite de $\widehat{ep}_{boot}$ quando B tende para o infinito (EFRON; TIBSHIRANI, 1993), ou seja,

$$\lim_{B\to\infty} \widehat{ep}_{boot} = ep_{\hat{F}}(\hat{\theta}^*) \tag{2.8}$$

O estimador *bootstrap* ideal e sua aproximação (2.8) são chamados estimadores *bootstrap* não paramétricos, já que se baseiam em $\hat{F}$, um estimador não paramétrico de F. Um estimador *bootstrap* paramétrico do erro padrão é baseado em um estimador $\hat{F}$ de F derivado de um modelo paramétrico. Por exemplo, ao invés de estimarmos F pela função distribuição empírica $\hat{F}$, podemos assumir que a população tem distribuição normal (Martinez e Louzada-Neto, 2001).

Sendo $\widehat{ep}_{boot}(\hat{\theta})$, $\hat{\theta} = s(x)$, a estimativa *bootstrap* do erro padrão de $\hat{\theta}$, definida em (2.6), o intervalo de confiança *bootstrap* padrão para θ, com probabilidade de cobertura de aproximadamente $1 - \alpha$, é dado por

$$\hat{\theta} \pm z_{1-\alpha/2}\widehat{ep}_{boot}(\hat{\theta}). \tag{2.9}$$

A maior vantagem deste método reside na sua simplicidade algébrica de encontrar um intervalo de confiança para θ. Entretanto, sendo a equação (2.9) uma seqüência de

$$Z = \frac{\hat{\theta} - \theta}{\widehat{ep}_{boot}(\hat{\theta})} \sim N(0; 1) \tag{2.10}$$

Os números de replicações *bootstrap* necessárias para uma boa estimativa do erro-padrão e do intervalo de confiança é muito importante, foram discutidas por (Efron e Tibshirani, 1993; Kendall e Stuart, 1977; Efron, 1987). Para obter uma boa estimativa do erro-padrão por meio do *bootstrap* são necessárias entre 25 e 200 replicações e que para uma boa estimativa dos limites de confiança seriam necessárias mais de 500 replicações (Efron e Tibshirani, 1993). Além disso, pode-se utilizá-la em diversas situações para estimação de parâmetros, obtenção de intervalos de confiança para os parâmetros analisados, obtenção de distribuição empírica dos estimadores e determinação do tamanho da amostra (Manly, 1997). Utilizamos neste trabalho 1000 replicações para a construção de intervalos de confiança e 100 replicações para o erro-padrão.

Com a técnica de reamostragem *bootstrap* não-paramétrica no modelo AMMI, foram reamostradas 100 matrizes de genótipos $\times$ ambientes, obtendo-se B_1=100 conjuntos de autovalores, $(\lambda^2_{1i} > \lambda^2_{2i} > \cdots > \lambda^2_{pi} > 0)$, em que $i = 1, 2, \cdots, 100$ e $p = min\{(g-1), (e-1)\} = 8$. Para a construção de intervalos de confiança foram reamostradas 1000 matrizes de genótipos $\times$ ambientes, obtendo-se B_2=1000 conjuntos de autovalores, $(\lambda^2_{1i'} > \lambda^2_{2i'} > \cdots > \lambda^2_{pi'} > 0)$, $i' = 1, 2, \cdots, 1000$ e $p = min\{(g-1), (e-1)\} = 8$.

Para cada autovalor foram feitas as análises separadas, como histograma, Q-Q plot, Envelope simulado e com a aplicação do teste de normalidade de Shapiro-Wilk por meio da técnica de reamostragem *bootstrap*.

Portanto, utilizando a metodologia AMMI-*bootstrap* (Lavoranti, 2003) e a técnica de reamostragem *bootstrap* não-paramétrica podemos encontrar a distribuição empírica dos autovalores e construir os intervalos de confiança. Todas as análises, gráficos e a rotina computacional foram implementadas no sistema estatístico software R (2010).

2.3 Resultados e Discussão

Apartir dos resultados utilizando a análise AMMI relativa à produtividade de milho ($kgha^{-1}$), obteve-se os elementos do desdobramento da $\text{SQ}_{G\times E}$ por meio da decomposição por valores singulares (DVS), aplicada à matriz de interações, correspondentes aos quadrados dos oito valores singulares (λ_s^2), com $s = 1, 2, \cdots, 8$, ou equivalentemente, aos oito autovalores de $(\widehat{\boldsymbol{GE}}_{G\times E})(\widehat{\boldsymbol{GE}}_{G\times E}^{T})$ ou $(\widehat{\boldsymbol{GE}}_{G\times E}^{T})(\widehat{\boldsymbol{GE}}_{G\times E})$(Tabela 1).

Avaliando a família dos modelos AMMI (AMMI0, AMMI1, AMMI2, AMMI3, AMMI4, AMMI5, AMMI6, AMMI7 e AMMI8) verificou-se pelo teste F, com os graus de liberdade ajustados pelo método de Gollob (1968) foram significativos para os três dos oitos componentes principais da interação com $p < 0,01$, o que levaria à seleção do modelo AMMI3, e seriam necessários três componentes principais para explicar a interação de forma significativa, pois somados (PC1, PC2 e PC3) explicaram grande porcentagem da variabilidade da interação dos dados (81,7%) e apresentaram o melhor padrão de resposta para os genótipos aos diferentes ambientes (Tabela 2.1). A não-significância para os PC4, PC5, PC6, PC7 e PC8 são desprezíveis e contém apenas ruído (variação aleatória não relacionada com o fenômeno da interação), que pode diminuir a eficiência da interpretação da estabilidade dos genótipos e ambientes na análise gráfica. Assim, a interpretação gráfica, considerando apenas a variação contida nos três primeiros eixos da ACP, é suficiente para avaliar a estabilidade dos genótipos e ambientes (Rocha, 2002; Freire Filho et al., 2003).

Tabela 2.1: Porcentagem da soma de quadrados da interação ($G\times E$) captada por componente principal (PC)

PC	AC(%)[1]	GL[2]	SQ[3]	QM[4]	F_{Gollob}[5]	Valor-p
PC1	56,2	26	140323418	5397054,5	8,95	$<0,01$
PC2	71,3	24	37711634	1571318,1	2,61	$<0,01$
PC3	81,7	22	26058572	1184480,6	1,96	$<0,01$
PC4	90,3	20	21533133	1076656,9	1,79	0,0192
PC5	95,3	18	12371704	687316,9	1,14	0,3093
PC6	98,1	16	6900236	431264,7	0,72	0,7744
PC7	99,1	14	2513071	179505,1	0,30	0,9939
PC8	100,0	12	2292394	191032,8	0,32	0,9858
Total	–	–	249704162	–	–	–

AC(%)[1]: proporção acumulada; GL[2]: graus de liberdade; SQ[3]: soma de quadrados; QM[4]: quadrado médio; F_{Gollob}[5]: teste F de Gollob (1968)

O histograma é uma forma de descrição gráfica de dados quantitativos, agrupados em classes de frequência e a grande vantagem desta ferramenta é permitir facilmente obter informações sobre a distribuição de freqüencias de um determinado grupo de dados. Observando os gráficos de histogramas e Q-Q plot dos valores estimados do primeiro ao quinto autovalores, com $B_1 = 100$ amostras por meio da técnica *bootstrap*, apresentaram uma forma simétrica e estão bem próximos à uma distribuição normal. Também podem ser verificadas pelo Q-Q plot e nota-se que os pontos estão dispostos próximo de linha reta com apenas dois pontos ficaram fora da reta no Q-Q plot do quinto autovalor, mas não influenciaram os resultados e a distribuição dos dados, o que leva a concluir que esses autovalores apresentaram um bom ajuste à distribuição normal (Figura 2.1).

Os gráficos dos Envelopes simulados sob o primeiro ao quinto autovalores com 100 reamostragens utilizando a metodologia *bootstrap*, aplicada à matriz de interação G × E, observando os gráficos do primeiro ao quarto autovalores de que todos os pontos se encontram contidos dentro dos limites de envelope, apenas o quinto autovalor ficou com dois pontos fora dos limites do envelope, mas como dois pontos têm freqüência baixa e não influenciam os resultados e a distribuição dos dados (Figura 2.2). As distribuições dos cinco primeiros autovalores apresentaram uma distribuição normal, pois pelo teste de normalidade de Shapiro-Wilk, reportando o valor da estatística W e o valor p desses cinco autovalores foram: $W = 0,9905$ e valor $p = 0,7074$; $W = 0,981$ e valor $p = 0,1591$; $W = 0,9819$ e valor $p = 0,1865$; $W = 0,9858$ e valor $p = 0,3635$, $W = 0,9804$ e valor $p = 0,1416$, respectivamente.

Os dados do primeiro ao quinto autovalores apresentaram uma média amostral $\hat{\lambda}_1^2 = 35080855$, $\hat{\lambda}_2^2 = 11365486$, $\hat{\lambda}_3^2 = 6234229$, $\hat{\lambda}_4^2 = 3390191$ e $\hat{\lambda}_5^2 = 1693611$, cujos erro padrão são estimados por $\widehat{ep}_{boot}(\hat{\lambda}_1^2) = 796031,3$, $\widehat{ep}_{boot}(\hat{\lambda}_2^2) = 290021$, $\widehat{ep}_{boot}(\hat{\lambda}_3^2) = 156486,3$, $\widehat{ep}_{boot}(\hat{\lambda}_4^2) = 103357,3$ e $\widehat{ep}_{boot}(\hat{\lambda}_5^2)$ $= 70691,93$ com $B_1 = 100$ amostras. Os intervalos de confiança (99%) *bootstrap* com $B_2 = 1000$ reamostragens para a média foram obtidos pelo método padrão para o primeiro ao quinto autovalores: [9856041 ; 53858254], [39997 ; 14816570], [2419947 ; 10913305], [4253955 ; 10349045] e [2640740 ; 6538943], respectivamente.

O histograma do sexto autovalor apresentou uma forma com "dois picos", o lado direito do histograma apresenta a forma de um histograma "despenhadeiro", o lado esquerdo do histograma apresenta um pico e com o valor próxima de zero, isso acontece porque o valor do sexto autovalor foi menor em relação

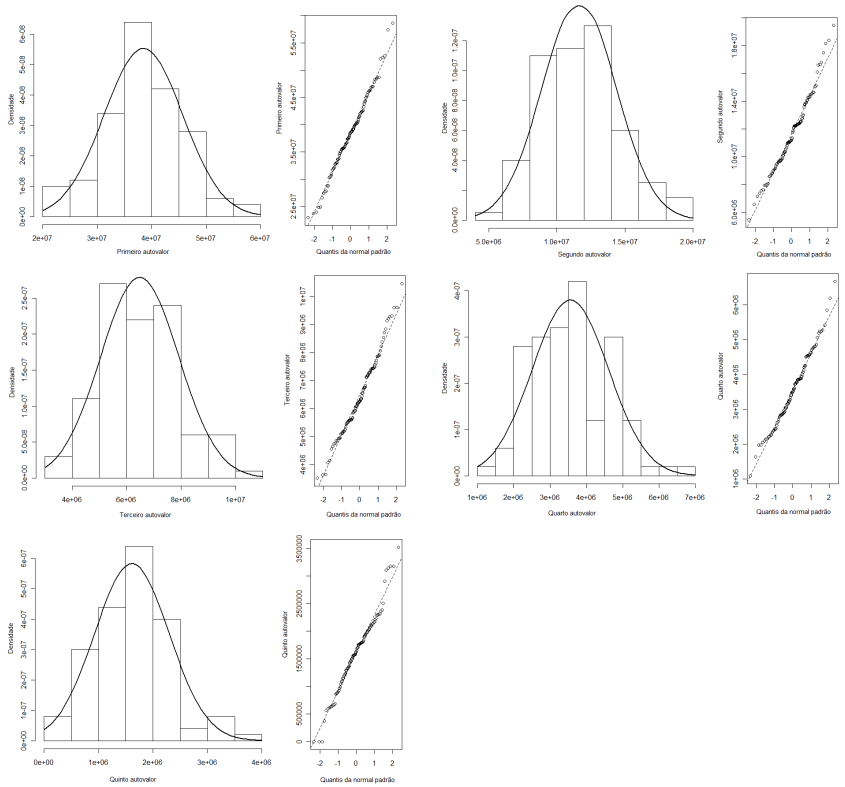

Figura 2.1: Gráficos de Histogramas e Q-Q plot do primeiro ao quinto autovalores com 100 reamostragens aplicada à matriz de interação G × E

aos autovalores anteriores ($\lambda_1^2 > \lambda_2^2 > \cdots > \lambda_8^2 > 0$). O valor de cada autovalor λ_s^2 é igual a variância do componente principal PC_s, em que $s = 1, 2, \cdots, p$; $p = min\{(g-1),(e-1)\} = 8$. O primeiro componente é o que apresenta a maior variância e assim sucessivamente. O histograma do sétimo autovalor apresentou uma forma como "histograma ilha isolada"ou "retângulos isolados", que tem algumas faixas de valores que ficaram isoladas da grande maioria dos dados, gerando uma barra separada. Isso acontece por causa do valor do sétimo autovalor que foi menor do que os outros autovalores anteriores, quase próxima de zero, que foi mostrado no lado esquerdo do histograma e o lado direito tem a forma de um histograma "despenhadeiro"e o histograma do oitavo autovalor apresentou apenas um pico, essas faixas de valores bem próximas a zero, por causa do valor do oitavo autovalor que foi o menor entre todos os autovalores (Figura 2.3).

Observando pelos os gráficos de Q-Q plot, nota-se que para o sexto autovalor uma parte dos pontos estão dispostos próximo à linha reta, mas existem vários pontos que não seguiram a linha reta, para sétimo autovalor apenas poucos

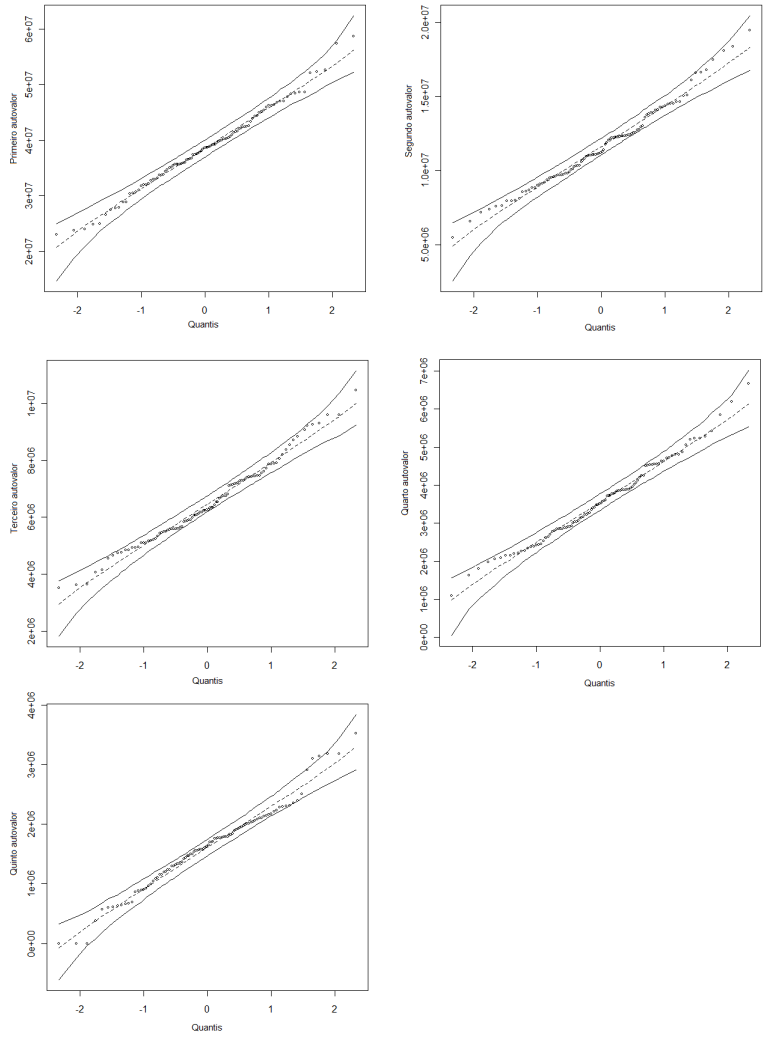

Figura 2.2: Envelope simulado sob o primeiro autovalor com 100 reamostragens aplicada à matriz de interação G × E

pontos estão dispostos próximo à linha reta e o oitavo autovalor quase todos os pontos estão fora da linha reta, logo influenciaram os resultados e a distribuição dos dados, o que leva a concluir que os sexto, sétimo e oitavo autovalores não apresentaram um ajuste à distribuição normal (Figura 2.3).

O gráfico de Envelope simulado do sexto autovalor com 100 reamostragens pelo método *bootstrap* apresentou vários pontos fora dos limites de envelope, com isso foram suficientes para influenciar a distribuição dos dados. Os gráficos do sétimo e oitavo autovalores mostraram que a maior parte dos pontos estão fora dos limites do envelope (Figura 2.4). Pelo teste de normalidade de Shapiro-Wilk, reportando o valor da estatística W e o valor p para esses três autovalores foram: $W = 0,9017$ e valor $p = 0,0042$; $W = 0,5899$ e valor $p = 0,0077$;

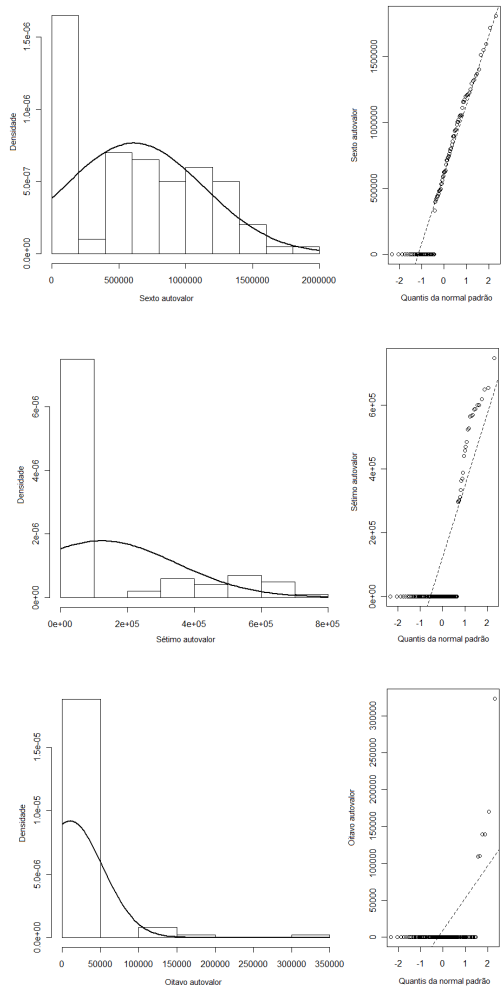

Figura 2.3: Gráfico dos Histogramas e Q-Q plot do sexto ao oitavo autovalores com 100 reamostragens aplicada à matriz de interação G × E

$W = 0,2441$ e valor $p < 0,0022$, respectivamente, portanto o sexto, sétimo e oitavo autovalores não apresentaram uma distribuição normal.

Os dados do sexto ao oitavo autovalores apresentaram uma média amostral $\hat{\lambda}_6^2 = 601648,7$, $\hat{\lambda}_7^2 = 129060,4$ e $\hat{\lambda}_8 = 7350,7$, respectivamente. Cujos erros padrão foram estimado por $\widehat{ep}_{boot}(\hat{\lambda}_6^2) = 49659,09$, $\widehat{ep}_{boot}(\hat{\lambda}_7^2) = 22281,94$ e $\widehat{ep}_{boot}(\hat{\lambda}_8^2) = 3824,227$ com $B_1 = 100$ amostras. Os intervalos de confiança (99%) *bootstrap* para esses três autovalores com $B_2 = 1000$ reamostragens para a média foram obtidos pelo método padrão : [1545877 ; 4197074], [617398 ; 1680491] e [1016415 ; 1260182].

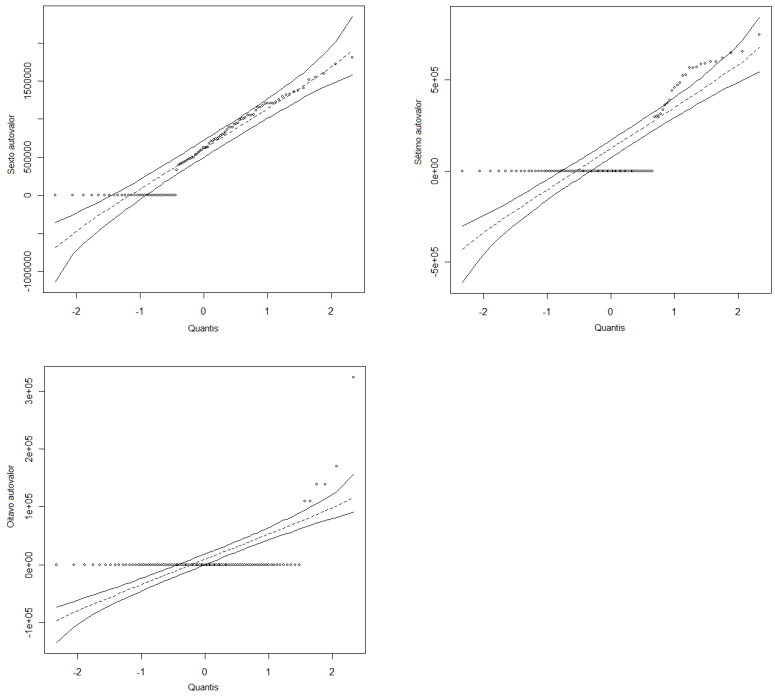

Figura 2.4: Gráfico de Envelope simulado sob o sexto ao oitavo autovalor com 100 reamostragens utilizando a metodologia *bootstrap*, aplicada à matriz de interação genótipos × ambientes

2.4 Conclusão

O método *bootstrap* não-paramétrico aplicado a matriz de resíduo, estimado via metodologia AMMI, constitui uma eficiente alternativa para encontrar a distribuição empírica dos autovalores e calcular o intervalo de confiança. Portanto, com a aplicação do teste de normalidade de Shapiro-Wilk para cada autovalor, concluímos que o primeiro autovalor até o quinto autovalor apresentaram uma distribuição normal, a partir do sexto autovalor até o oitavo autovalor não apresentaram o mesmo resultado anterior.

Capítulo 3

AMMI bootstrap

Modelo de Efeitos Aditivos com Interação Multiplicativa - AMMI

O modelo AMMI tem um modelo específico, e só se este modelo é usado para ajustar dados é que então os resultados serão realmente úteis. Em geral, AMMI é o modelo selecionado quando os dados apresentam efeitos principais e interação significativa (Gauch, 1992). Desta forma, com os dados organizados em uma tabela de dupla entrada, com os genótipos nas linhas e os ambientes nas colunas, a interação G $\times$ E é facilmente estudada.

No entanto, uma consideração importante deve ser feita para realização do teste F e interpretação dos resultados da análise conjunta da variância: a natureza fixa ou aleatória dos efeitos do modelo de análise (Chaves, 2001). A natureza fixa ou aleatória da interação é determinada pelos efeitos principais. Se genótipos e ambientes são fixos, a interação será fixa. Se pelo menos um dos fatores for aleatório, a interação será aleatória. Na análise de ensaios multi-ambientais, consideram-se, em geral, os efeitos de genótipos como fixos e os efeitos de ambientes como aleatórios, de tal forma que o efeito da interação genótipos $\times$ ambientes é aleatória nesse caso. Isto se deve ao fato de normalmente se estar interessado em verificar qual o melhor dos genótipos avaliados em relação à produtividade e se existe uma variabilidade de sua produção em relação ao ambiente onde foi cultivado, assim tem-se um modelo misto com genótipos fixo e ambiente aleatório.

Vantagens do uso do modelo AMMI com reamostragem *bootstrap*

Os escores AMMI têm distribuição de probabilidade desconhecida, fato que dificulta a obtenção de fórmulas via argumentos analíticos. Para medir a precisão desses escores, Lavoranti et al. (2004) apresentou uma metodologia complementar para avaliar a precisão dos escores de genótipos e de ambientes, por meio de técnicas de reamostragem 'bootstrap' aplicadas à análise AMMI juntamente com técnicas de agrupamento, permitindo uma melhoria na qualidade das inferências sobre as adaptabilidades e estabilidades fenotípicas estimadas pelo modelo AMMI e também a obtenção dos grupos formados nos agrupamentos.

Dias (1998) alerta que muitas distâncias genéticas são essencialmente medidas de distâncias geométricas, sem qualquer conteúdo genético, e por isso são denominadas de coeficientes de dissimilaridade ou divergência. Outro ponto considerado importante por certos autores é a distância geográfica como indicadora de divergência genética. Ram e Panwar (1970), trabalhando com arroz, perceberam que os padrões de agrupamento foram relacionados com a distância geográfica.

Assim, para avaliar a diversidade genética de forma simultânea em relação a todas as características de interesse, recomenda-se a utilização de medidas de dissimilaridade (Cruz e Carneiro, 2006). Uma forma prática e eficiente de se obter estas medidas é por meio da análise de agrupamentos (Análise de Cluster - AC), a qual tem por finalidade reunir as variáveis em grupos, de forma que exista máxima homogeneidade dentro do grupo e máxima heterogeneidade entre os grupos (Johnson e Wichern, 1992; Cruz et al., 2004).

Atualmente, diversos coeficientes de similaridade são propostos para a análise das divergências genéticas, principalmente devido à utilização das técnicas multivariadas (Meyer, 2002). Quando um experimento depende de muitas variáveis, não basta conhecer informações estatísticas isoladas para cada variável, mas é necessário também conhecer a totalidade destas informações fornecida pelo conjunto das variáveis. As relações existentes entre as variáveis não são percebidas e assim efeitos antagônicos ou sinérgicos de efeito mútuo entre variáveis complicam a interpretação do fenômeno a partir das variáveis consideradas. A necessidade de entender as relações entre várias variáveis faz com que a análise multivariada seja de grande importância. Com esta análise pode-se reduzir os dados ou simplificar a sua estrutura sem muita perda da

informação contida nos dados obtendo assim uma fácil interpretação dos resultados, principalmente quando as variáveis são correlacionadas.

Desta forma, para determinar quão distante geneticamente uma população ou genótipo é de outra são utilizados os métodos de agrupamentos, com os quais se quantifica ou se estima a heterose, que são analisados pela estatística multivariada permitindo unificar múltiplas informações de um conjunto de caracteres.

Vários métodos podem ser utilizados, dentre eles estão a análise por componentes principais, variáveis canônicas e métodos aglomerativos. Mais especificamente a técnica de componentes principais, apresenta entre seus principais objetivos a possibilidade de eliminar os ruídos presentes nos dados, possibilitando dessa forma uma interpretação segura do padrão de resposta presente nos dados. O ruído sugere que as respostas são imprevisíveis e não interpretáveis sendo parte integrante da variabilidade estranha contida nos dados, enquanto que o padrão responde de forma sistemática, significante e interpretável pois permite quantificar e informar o grau de semelhança ou de diferença entre pares de indivíduos.

Além disso, a aplicação dos métodos multivariados é tida como uma técnica que melhor explora as informações contidas nos dados e, em estudos de divergência genética por métodos aglomerativos, a medida de distância mais amplamente utilizada é a distância euclidiana.

A vantagem dessa estimativa de dissimilaridade é devida ao universo euclidiano e ao fato de minimizar os erros da classificação nos agrupamentos. Essa distância é a mais recomendada quando as unidades de cálculos são escores de componentes principais (Cruz et al., 2004), como é o caso da análise AMMI (additive main effects and multiplicative interaction analysis).

Análise de agrupamento

A análise de agrupamento permite classificar n itens (populações, clones, variedades, indivíduos, etc.) avaliados por um conjunto de m variáveis, cujo objetivo é identificar e separar os itens em grupos, de forma que os mais semelhantes permaneçam no mesmo grupo. Em agrupamentos hierárquicos, tem-se que em todos os casos se desconhece “a priori”o número e a composição dos diferentes grupos ou “clusters”a serem formados.

Inicia-se o processo definindo-se os indivíduos e os objetivos desejados para a aplicação da análise, além dos critérios que irão definir as similaridades entre eles. Assim os dados são dispostos na forma de uma matriz, em que as linhs que representam os indivíduos de interesse e as colunas representam as variáveis. A matriz de distância é gerada a partir de amostras de n itens, totalizando n(n-1)/2 pares de distâncias.

Existem inúmeros métodos de agrupamento, diferenciados nas formas de especificar a medida de proximidade entre indivíduos. Cada critério de análise de agrupamento impõe certo grau de estrutura nos dados. Portanto, recomenda-se diferentes critérios de agrupamento sejam aplicados e comparados entre si (Dias, 1998; Cruz e Carneiro, 2006).

Em se tratando dos métodos de agrupamento, as delimitações podem ser estabelecidas por um exame visual do dendrograma, em que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de indivíduos para determinado grupo (Cruz e Carneiro, 2006), contextualizados neste trabalho pelos genótipos.

É importante ressaltar que o dendrograma ilustra as fusões ou partições efetuadas em cada nível sucessivo do processo de agrupamento, no qual um eixo representa os indivíduos e o outro eixo representa 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 representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

Entretanto, a falta de critérios objetivos para se determinar o ponto de corte no dendrograma (número ótimo de grupos) ainda é um problema em estudos que utilizam a análise de agrupamentos. Um método considerado como "objetivo", dentre os poucos existentes, é o Método de Mojena (1977). Este Método é um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendrograma.

A análise de agrupamento hierárquico consiste no tratamento matemático de cada amostra como um ponto no espaço multidimensional descrito pelas variáveis escolhidas (Moita Neto e Moita, 1970). Também é possível, nesta técnica, tratar cada variável como um ponto no espaço multidimensional descrito pelas amostras, ou seja, podemos ter agrupamento de amostras ou de variáveis de acordo com o interesse em cada situação. Quando uma determinada amostra é tomada como um ponto no espaço das variáveis, é possível calcular a distância deste ponto a todos os outros pontos, constituindo-se as-

sim uma matriz que descreve a proximidade entre todas as amostras estudadas. Existem várias maneiras de calcular a distância entre dois pontos, a mais conhecida e utilizada é a distância euclidiana. Além disso, existem vários algoritmos a serem utilizados na formação dos agrupamentos hierárquicos, contudo todos eles utilizam as informações da matriz de proximidade para criar um dendrograma de similaridade.

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 ou dissimilaridade 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.

Convém ressaltar que a aplicação da análise de agrupamento hierárquico, quando temos variáveis de escalas diferentes, deve ser precedida por um tratamento prévio dos dados. Quando não é feito o pré-tratamento, as variáveis com valores numéricos mais altos serão mais importantes no cálculo que as variáveis com valores numéricos mais baixos. O pré-tratamento mais comumente empregado é a transformação Z, que transforma as medidas de cada variável de tal modo que o conjunto de dados tenha média zero e variância unitária. A finalidade deste procedimento é equalizar a importância estatística de todas as variáveis utilizadas. Tendo por base estas informações, o algoritmo tradicional da estrutura de formação de 'clusters' do tipo hierárquica, de maneira geral, tem os seguintes passos:

1. Considere-se uma base inicial de $Nclusters$ iniciais. Em geral, esses agrupamentos correspondem simplesmente às unidades a serem agrupadas, cada um desses $Nclusters$ contém apenas uma unidade inicialmente. A cada unidade i, está associado um vetor de m características $x_i = [x_{i,1}\ x_{i,2}\ \cdots\ x_{i,m}]$.

2. Calcula-se a distância entre todos os pares formados por elementos dentre esses N clusters iniciais. A distância nesse caso, pode ser qualquer medida de dissimilaridade entre o conjunto de atributos $x_i = [x_{i,1}\ x_{i,2}\ \cdots\ x_{i,m}]$. Para uma discussão sobre as diversas medidas de dissimilaridade, vide Khattree e Naik (2000).

3. Sejam I e J os dois clusters apresentando a menor distância, ou dissimilaridade, entre eles. Agrupa-se então o par I e J em um único novo cluster. O número de *clusters* agora passa a ser $N - 1$.

4. Para os $N - 1$ novos *clusters*, depois da junção descrita no passo 3, calculam-se as distâncias entre todos os pares. Para o par com a menor distância, agrupam-se os elementos em um único novo cluster, de forma que o número de *clusters* existentes passe a ser $N - 2$.

5. Repetem-se os passos 2 a 4 até se obter um único cluster, que deverá conter todos os $Nclusters$ iniciais.

Ao final do processo, o analista terá em mãos uma árvore descrevendo a seqüência de agrupamentos em cada passo do algoritmo. Para um número inicial de N unidades observacionais na base de dados, ao todo ocorrem $N-1$ junções.

Existem na literatura vários métodos para AAH. No agrupamento hierárquico denominado Método do Vizinho Mais Próximo ou Método de Ligação Simples ("Single Linkage Method"), forma-se um grupo inicial de indivíduos mais similares e em seguida são calculadas as distâncias daquele grupo em relação aos demais indivíduos usando-se a menor distância do conjunto. Neste método, as conexões entre indivíduos e grupos são feitas por ligações simples entre pares de indivíduos, ou seja, a distância entre os grupos é definida como sendo aquela entre os indivíduos mais parecidos entre esses grupos (Cruz e Carneiro, 2006). Este método leva a grupos longos e os dendrogramas resultantes deste procedimento são geralmente pouco informativos, devido à informação dos indivíduos intermediários que não são evidentes.

Quando a semelhança entre dois grupos é definida como a máxima distância entre pares de indivíduos tem-se o Método do Vizinho Mais Distante ou Método de Ligação Completa ("Complete Linkage Method"). A similaridade entre dois grupos é definida como aquela apresentada pelos indivíduos de cada grupo que menos se parecem, formando-se todos os pares com um membro de cada grupo. Este método, geralmente leva a grupos compactos e discretos, tendo os seus valores de similaridade relativamente pequenos (Meyer, 2002).

No entanto, se a distância entre os pares de grupos for estimada com base na média aritmética dos indivíduos, tem-se o método hierárquico UPGMA ("Unweighted Pair-Group Method Using Arithmetic Averages") ou Ligação média entre grupo, que evita caracterizar a dissimilaridade por valores extremos entre os genótipos, não considera a estrutura de subdivisão do grupo,

dando pesos iguais a cada indivíduo do grupo calculando a similaridade média de um indivíduo que pretende se juntar ao grupo existente.

Já o método da Mediana ou WPGMA ("Weighted Pair-Group Method Using Arithmetic Averages") difere do UPGMA pelo fato de atribuir aos indivíduos admitidos mais recentemente a um grupo, peso igual aos dos demais indivíduos já pertencentes ao grupo (Romesburg, 1990).

Dentre os métodos de agrupamento hierárquicos, o método de Ward é o que forma grupos de maneira a atingir sempre o menor erro interno entre os vetores que compõe cada grupo e o vetor médio do grupo. Isto equivale a buscar o mínimo desvio padrão entre os dados de cada grupo.

No método de Ward, os grupos de dados são formados em etapas. No princípio, têm-se m grupos; ou seja, um grupo para cada vetor componente da base de dados. Neste estágio inicial o erro interno é nulo para todos os grupos pois cada vetor que compõe cada grupo é o próprio vetor médio do grupo. Igualmente o desvio padrão para cada grupo é nulo. Na etapa subseqüente, cada possibilidade de aglutinação entre os grupos 2 a 2 é verificada, e é escolhido o agrupamento que causa o menor aumento no erro interno do grupo. São m x m verificações. Nota-se que a cada iteração tem-se m - i grupos (i = número de iterações), no entanto, como o número de elementos pertencentes a cada grupo aumenta, é maior o número de cálculos para o erro interno de cada grupo.

Ressalta-se que o agrupamento conduz à perda de informações ao nível de indivíduos, restando apenas informações sobre grupos similares. Por esse motivo, o estudo da divergência pode ser conduzido também simultaneamente, por componentes ou coordenadas principais, isto é particularmente importante quando o número de indivíduos é grande (Cruz e Carneiro, 2006).

Soma de quadrados progressiva (Ward 1963, Orlóci 1978)

O critério de agrupamento de Ward minimiza o aumento na soma de quadrados dentro do grupo formado a cada passo de agrupamento, isto é, $Q_{PQ} = Q_{P+Q} - Q_P - Q_Q$ em que Q_{P+Q} é a soma de quadrados total no grupo P+Q; Q_P e Q_Q são as somas de quadrados dentro dos grupos P e Q. Tem-se que $Q_{P+Q} = Q_{P+Q} = \frac{1}{n_p+n_q} \sum_h \sum_i d_{hi}^2$ para $h = 1, \ldots, n-1$ e $i = h+1, \ldots, n$ objetos, desde

que h e i pertençam ao grupo P ou Q;
$Q_P = \frac{1}{n_p} \sum_h \sum_i d_{hi}^2$ para $h = 1, \ldots, n-1$ e $i = h+1, \ldots, n$ objetos, desde que h e i pertençam ao grupo P e $Q_Q = \frac{1}{n_q} \sum_h \sum_i d_{hi}^2$ para $h = 1, \ldots, n-1$ e $i = h+1, \ldots, n$ objetos, desde que h e i pertençam ao grupo Q.

Reamostragem bootstrap

Existem diversas técnicas de reamostragem que visam estimar parâmetros de uma distribuição de interesse. O método de reamostragem *bootstrap* foi originalmente proposto por Efron (1979) em um influente artigo publicado no Annals of Statistics. É um método de reamostragem que se baseia na construção de distribuições amostrais empíricas de uma estatística de interesse. Uma vantagem em utilizar a técnica de reamostragem *bootstrap* é a generalidade com que pode ser aplicada, pois requer que menos suposições sejam feitas.

Hesterberg et al. (2003) afirmam que a amostra original representa a população da qual foi retirada. Desta forma, tratando a amostra como se ela fosse a "população"realizando sucessivas reamostragens com reposição. A partir destas reamostras, é possível estimar características da população, tais como média, variância, percentis, etc.

A distribuição empírica de uma estatística, gerada pelo método *bootstrap* tem aproximadamente a mesma forma e amplitude que a distribuição amostral da estatística.

Coelho et al. (2008) descreve o método *bootstrap* não paramétrico para obtenção de intervalos de confiança dos coeficientes de um modelo de regressão múltipla. Com esse propósito, o autor considera a seguinte notação:

Seja $U = (y, x_1, x_2, \ldots, x_k)$ em que, $y = (y_1, y_2, \ldots, y_n)$ é o vetor de observações da variável dependente e $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$, $j = 1, \ldots, k$ são os vetores de observações das variáveis independentes, x_j é uma variável contínua para todo j.

Intervalos p-*bootstrap*

1. Amostrar, com reposição de U, uma amostra *bootstrap* $(y_1^*, x_{11}^*, x_{12}^*, \ldots, x_{1k}^*)$, $\ldots$, $(y_n^*, x_{n1}^*, x_{n2}^*, \ldots, x_{nk}^*)$.
2. Da amostra *bootstrap* $(y_1^*, x_{11}^*, x_{12}^*, \ldots, x_{1k}^*), \ldots, (y_n^*, x_{n1}^*, x_{n2}^*, \ldots, x_{nk}^*)$, obter o estimador de mínimos quadrados de β, representado por $\hat{\beta}^*$
3. Repetir os passos I e II um número B grande de vezes.
4. Ordenam-se os valores obtidos, do menor ao maior de forma que $\hat{\beta}^* = (\hat{\beta}_{(1)}^* \leq \hat{\beta}_{(2)}^* \leq \ldots \leq \hat{\beta}_{(B)}^*)$
5. Determinam-se limites de confiança para uma especificada probabilidade α, igual ao nível de significância, sendo $0 < \alpha < 1$, de forma que o intervalo de confiança p-*bootstrap* $100 \times$ (1-α)% é dado por $(\hat{\beta}_{(q_1)}^*; \hat{\beta}_{(q_2)}^*)$ em que $q_1 = [B \times (\alpha/2)]$ e $q_2 = B - q_1$, em que $[x]$ indica o menor número inteiro maior ou igual ao argumento x.

Como exemplo, em V, para ($\alpha = 0.05$ e B=1000) $\Longrightarrow$ ($q_1 = 25$ e $q_2 = 975$). Logo, o intervalo de confiança p-*bootstrap* de 95% é dado por $(\hat{\beta}_{(25)}^*; \hat{\beta}_{(975)}^*)$. Os intervalos de confiança para quaisquer outros parâmetros de interesse são obtidos de maneira similar.

Pode-se então afirmar, com uma probabilidade a de se estar errado, que o intervalo de confiança construído tem alta probabilidade de conter o verdadeiro valor do parâmetro sobre o qual a estimação foi baseada.

Quando uma variável independente não é contínua, deve-se fazer o processo de reamostragem (1, 2 e 3) dentro de cada nível da variável (Wu, 1986; Tibshirani, 1988). Outras alternativas ao intervalo de confiança p-bootstrap são discutidas, por exemplo, em Efron e Tibshirani (1993), Davison e Hinkley (1997) e Carpenter e Bithell (2000).

Para que a aplicação da técnica resulte em valores confiáveis devem ser feitas, a partir da amostra mestre, centenas ou até milhares de reamostras do mesmo tamanho n. A maioria dos autores recomenda a utilização de 1000 reamostras.

Material e métodos

Os experimentos que originaram os dados do presente estudo envolveram linhagens de soja da população com controle de insetos - originadas da seleção para produtividade de grãos, aplicou-se inseticidas durante o ciclo total para o controle de insetos mastigadores e sugadores. Os experimentos foram conduzidos em dois locais do município de Piracicaba - SP (Estação Experimental Anhembi e Fazenda Areão), durante dois anos agrícolas (1999/00, 2000/01) e em dois sistemas de manejo (com controle intensivo de insetos e com controle ecológico de insetos).

Os experimentos desenvolvidos em 1999/2000 incluíram 120 linhagens experimentais F_{10}, (40 de cada população); os experimentos de 2000/01 envolveram 60 linhagens F11, sendo 20 linhagens selecionadas em cada população. As repetições foram estratificadas em conjuntos experimentais, cada um deles com três testemunhas comuns: IAC-100, OCEPAR-4 e IAS-5.

Nos experimentos com controle intensivo de insetos (CII), com base em monitoramento frequente, promoveu-se a aplicação de inseticidas sempre que se detectaram insetos mastigadores e ou dois percevejos/m^2 de pano (método do pano de batida). Nos experimentos com controle ecológico de insetos (CEI), promoveu-se aplicações de inseticidas apenas na ocorrência de grande quantidade de danos nas folhas causadas por insetos mastigadores e ou quando a infestação natural atingiu quatro percevejos/m^2 de pano.

Para a análise da interação considerou-se como ambiente a combinação local e manejo, para cada uma das três populações, sendo as linhagens foram estudadas em dois anos agrícolas (1999/2000 e 2000/2001). Assim, obteve-se quatro ambientes distintos (E1, E2, E3 e E4) a saber: Anhembi-CII (E1), Anhembi-CEI (E2), Areão-CII (E3), Areão-CEI (E4). Esta situação se repetiu para cada ano agrícola. A análise foi realizada com as médias das duas repetições em cada ambiente.

Os experimentos foram conduzidos no delineamento aleatorizado em blocos com as repetições estratificadas em conjuntos experimentais com testemunhas comuns: IAC-100, OCEPAR-4, IAS-5 e Primavera. A cultivar Primavera foi eliminada nos experimentos do ano agrícola 2000/01. Foram considerados fixos os efeitos de populações (linhagens e testemunhas comuns) e o efeito de ambientes foi considerado aleatório.

A metodologia AMMI foi aplicada nas análises fazendo a interação multiplicativa destes fatores com base na análise multivariada por componentes

principais e decomposição por valores singulares. Foi empregado o teste F_{Gollob} proposto por Gollob (1968), feita a associação entre "bootstrap"e AMMI (Lavoranti, 2003) com base na reamostragem da matriz de resíduos (livres de ruídos), obtida dos valores estimados pelo método AMMI proposto por Gollob (1968) e posteriormente aplicou-se a técnica Análise de Agrupamentos.

A reamostragem "bootstrap"foi realizada na versão não-paramétrica e, em concordância com o modelo AMMI, optou-se pela reamostragem na matriz de resíduos, obtida dos valores estimados. Dessa forma, foram aplicados procedimentos estatísticos que possibilitaram as análises gráficas e numéricas das adaptabilidades e estabilidades fenotípicas dos genótipos e dos ambientes envolvidos neste estudo.

Todas as análises estatísticas foram feitas com o software estatístico R (The R Development Core Team, 2008) usando os pacotes: fields (Furrer et al., 2009) e agricolae (De Mendiburu, 2009).

Resultados e Discussão

Antes da obtenção da ANOVA foram analisadas as pressuposições do modelo em cada um dos quatro ambientes. Tendo o modelo atendido a todas as pressuposições, em todos os ambientes, a análise de variância individual (em cada ambiente) foi obtida para avaliação estatística da variabilidade genética entre os tratamentos (linhagens de soja) e da precisão experimental. Uma vez detectada diferenciação entre os tratamentos, realizou-se análise conjunta de variância, e assim, foram feitas as interpretações relativas às significâncias do teste F.

A maioria dos coeficientes de variação experimental no presente estudo esteve abaixo de 20%, indicando boa precisão no controle das causas de variação de ordem sistemática dos ambientes experimentais, para a produtividade de grãos, um caráter quantitativo muito influenciado pelo ambiente. Além disso, a existência de variabilidade genética entre linhagens foi constatada tanto nas análises individuais quanto na análise conjunta.

Na população em estudo, no ano agrícola 2000, foi realizada a análise de variância conjunta (Tabela1) sendo a presença significativa da interação genótipo por ambiente (G $\times$ E), diagnosticada pelo teste F.

Tabela 3.1: Análise de variância conjunta no ano 2000 para dados de produtividade de grãos, em kg/ha, de 44 genótipos de soja avaliados em 4 ambientes com 2 blocos.

Fonte de Variação	GL	SQ	QM	F	Pr (>F)
Ambiente (E)	3	38278785	12759595	242,8261	5,58e-05**
Blocos(E)	4	210185	52546	2,1966	0,070626
Genótipo(G)	43	2392397	55637	2,3259	4,54e-05**
G × E	129	4542867	35216	1,4722	0,006804**
Resíduo	204	4879923	23921		

**: significativo ao nível de probabilidade, p<0,01.
*: significativo ao nível de probabilidade, p<0,05. ns: não significativo.

Verificou-se diferenças significativas ao nível de 1% para genótipos (G), ambientes (E) e interação entre eles (G × E). No entanto, o efeito de blocos dentro de ambientes foi não significativo. Em termos do teste F, a magnitude da fonte de variação de ambientes foi muito superior às demais, sendo ela responsável pela maior parte da variação ocorrida. Pode-se, assim, inferir que os efeitos de locais contribuíram de modo mais acentuado para a variação na produtividade. A significância para genótipos indicou que eles são formados por grupos geneticamente distintos, mostrando uma suficiente disponibilidade de variabilidade para a seleção.

A significância do resultado da interação G × E indicou que as variâncias dos genótipos são diferentes de um ambiente para outro, o que permite inferir que há genótipos ou grupos de genótipos com adaptação específica para determinados ambientes e outros com adaptação geral a todos os ambientes.

Com base nesses resultados, procedeu-se o estudo mais detalhado da interação significativa (Tabela 2).

Tabela 3.2: Porcentagem responsável por cada eixo da interação e porcentagem da soma de quadrados acumulada (PA) por eixo singular, Teste F para os componentes. População PCI, ano 2000.

IPCA	**Porcentagem**	**PA**	**GL**	**SQ**	**QM**	**F**	**P-valor**
IPCA$_1$	49,1	49,1	45	2162507	48055,71	2,01	0,0006
IPCA$_2$	34,8	83,9	43	1534622	35688,88	1,49	0,036
IPCA$_3$	16,1	100	41	710244,2	17323,03	0,72	0,8945
IPCA$_4$	0	100	39	0	0	0	1

F:teste F de Gollob

A análise da interação G × E, por componentes principais mostrou os dois primeiros eixos ($IPCA_1$ e $IPCA_2$) significativos (p<0,05), os quais explicaram 83,9% da porção da $SQ_{G\times E}$. Assim, os escores de genótipos foram obtidos seguindo o modelo $AMMI_2$, sendo que primeiro eixo singular da análise AMMI captura a maior porcentagem de "padrão"e, com acumulação subsequente das dimensões dos eixos, há uma diminuição na porcentagem de "padrão"e um incremento de "ruídos".

Na Figura 1, o gráfico mostra o comportamento dos genótipos mediante a produtividade média de grãos (PG), importante na recomendação de cultivares. Dentre todos os genótipos, verificou-se que aqueles à direita da linha vertical tracejada são os genótipos que ultrapassaram a média geral de PG obtida (991,5312 kg/ha). O genótipo 97-8056 obteve a maior média de produtividade de grãos (1166,500) e, no entanto, a testemunha IAC-100 registrou o pior resultado para esse caráter.

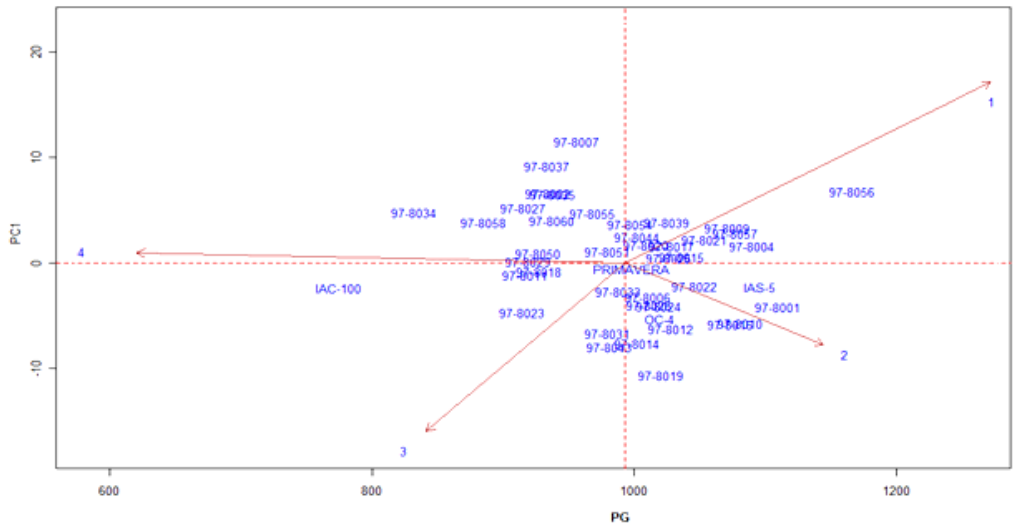

Figura 3.1: Gráfico do comportamento de 44 genótipos de soja em relação ao primeiro componente da interação e ao caráter Produtividade de Grãos (PG).

Pelo gráfico biplot (Figura 2) observou-se que o ambiente que menos contribuiu para a interação foi o ambiente 4, com escore baixo (norma do vetor). Os genótipos considerados estáveis foram 97-8011 (n^o 8), 97-8029 (n^o 24), 97-8050 (n^o 33) e a testemunha IAS-5 (n^o 44). Esses podem ser considerados com alta estabilidade e desta forma, adaptam-se a qualquer um dos quatro ambientes.

Observa-se ainda que o genótipo 97-8056 (n^{oo}37) foi o que apresentou o maior escore de produtividade de grãos (1166,5) e está alta e positivamente relacionado ao ambiente 1, que foi o ambiente que mais contribuiu para a

interação G × E, possuindo o maior escore.

Os genótipos que podem ser recomendados amplamente são os que combinam médias elevadas do caráter PG e estabilidade em relação aos ambientes em estudo, fato este ocorreu com o genótipo 97-8011 (n^o 8) que registrou média 917, o genótipo 97-8029 (n^o 24) com média 919, o 97-8050 (n^o 33) com média 926,75 e a testemunha IAS-5 (n^o 44) com média 1095,625. Todas as médias destes genótipos ficaram acima da média geral de PG. Já o ambiente 1 apresentou a maior média para o caráter produtividade de grãos (1408,602 kg/ha).

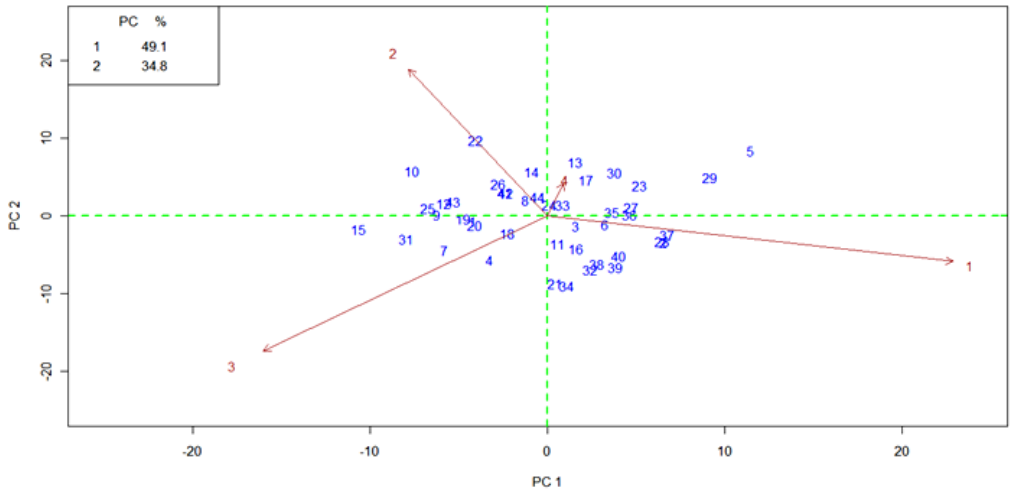

Figura 3.2: Biplot AMMI2 para dados de produtividade de grãos em kg/ha, ano 2000.

Além disso, observou-se que no biplot os aspectos mais relevantes quanto a adaptabilidade dos genótipos ficaram também identificados: os genótipos 97-8012 (n^o 9), 97-8014 (n^o 10), 97-8026 (n^o 22) apresentaram interação positiva no ambiente 2, bem como a testemunha IAS-5 (n^o 43). Já os genótipos com interação positiva no ambiente 3 foram: 97-8010 (n^o 7), 97-8012 (n^o 9), 97-8031 (n^o 25), 97-8019 (n^o 15) e 97-8043 (n^o 31). A interação positiva com o ambiente 4 ocorreu com os genótipos 97-8017 (n^o 13) e 97-8021 (n^o 17) e a interações negativas nesse ambiente ocorreu com os genótipos 97-8006 (n^o 4) e 97-8015 (n^o 11).

A divergência entre os genótipos foi avaliada pelo método hierárquico aglomerativo de Ward, com o emprego da distância Euclidiana. A partir da análise de agrupamentos realizada na matriz que é a média das matrizes de distâncias reamostradas, obteve-se um dendrograma (Figura 3) representativo dos agrupamentos formados entre os genótipos em estudo.

No dendrograma foi registrada a formação de seis grupos distintos, dos

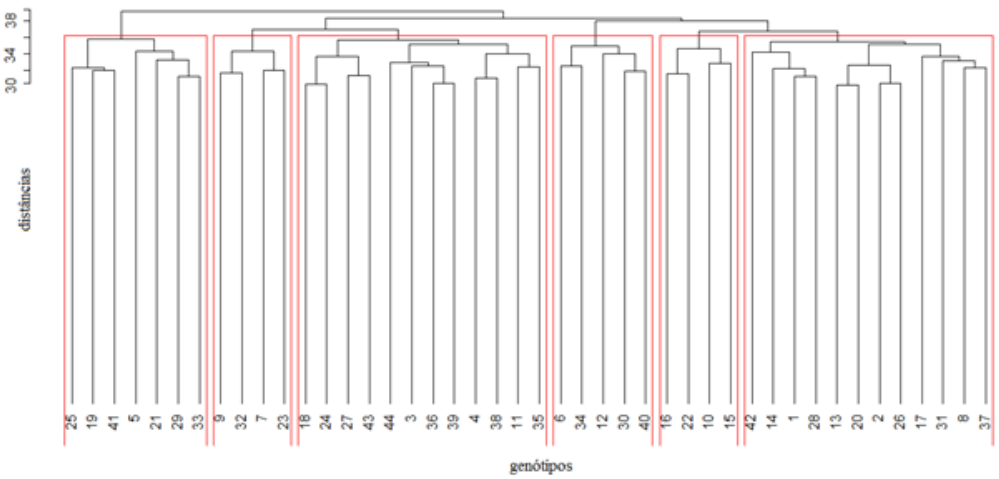

Figura 3.3: Dendrograma das distâncias euclidianas entre os escores "bootstrap"de marcadores de genótipos AMMI2, para dados de produtividade de grãos, em kg/ha.

quais três apresentam testemunhas dentre seus materiais, demonstrando que cada testemunha possui características diversificadas entre si a ponto de serem alocadas em grupos distintos quanto à similaridade. No entanto, os grupos que continham as testemunhas se destacaram dos demais quanto à produtividade de grãos e estabilidade observados nos genótipos neles alocados.

Foi obtido que o grupo da testemunha IAS-5 (n^o 42) apresentou média de produção superior a média geral, sendo o genótipo 97-8017 (n^o 13) o de maior estabilidade, podendo o mesmo ser recomendado para qualquer um dos ambientes.

Já o grupo formado pelas testemunhas Primavera (n^o 44) e OC-4 (n^o 43) apresentou média produtiva similar à média geral, sendo o grupo que apresentou o maior número de genótipos estáveis, 97-8004 (n^o 3) e 97-8015 (n^o 11). Por outro lado, quando se analisa a redução relativa da produção quando submetido ao ataque de insetos percebe-se que estes materiais apresentaram um dos menores valores médios de redução, justificando assim sua proximidade com a testemunha tolerante (IAC-100).

Finalizando as análises, por meio do uso do *bootstrap* foi possível obter intervalos aproximados d e100(1-α)% de confiança para o parâmetro de interesse (traço da matriz de covariâncias da matriz G $\times$ E). A análise do intervalo interpercentílico baseado nos percentis bootstrap, referente à estimativa do parâmetro, mostrou que tal estimativa (igual a 2.203.686) está contida no intervalo definido pelo percentil 10 (igual a 1.881.002), o limite inferior, e o

percentil 90 (igual a 2.738.849), o limite superior. Este resultado confirma que a reamostragem realizada foi precisa uma vez que a estatística de interesse, definida como sendo o traço da matriz de variâncias e covariâncias da matriz de interação, estava contida no intervalo obtido.

Capítulo 4

Correção de autovalores

4.1 Introdução

Em muitos casos, o pesquisador está interessado em avaliar o desempenho de vários genótipos (tratamentos) em diversos ambientes (locais e/ou ano). Quando um conjunto de experimentos é planejado para vários locais é necessário considerar o delineamento individual em cada local e a combinação total dos delineamentos com os locais (interação genótipo × ambiente).

Várias metodologias estatísticas têm sido propostas para a interpretação da interação genótipo × ambiente e os pesquisadores ainda continuam na busca de uma ferramenta estatística que permita extrair grande parte da informação possível desta fonte de variação. Entre tantas ferramentas disponíveis na literatura os métodos de regressão linear simples (Eberhart & Russel, 1966) e regressão linear múltipla (Silva & Barreto, 1985) têm sido as metodologias mais utilizadas, mas estas técnicas possuem limitações e têm sido alvo de várias críticas, como no caso em que a linearidade falha.

Na busca de novas ferramentas para o estudo da interação genótipo × ambiente, o modelo AMMI (Mandel, 1961, 1969, 1971; Gollob, 1968) vem se destacando e ganhando grande aplicabilidade nos últimos anos (Duarte & Vencovsky, 1999).

Essa técnica multivariada é baseada no uso dos autovalores e autovetores provenientes da matriz de interação genótipo × ambiente. Araújo & Dias (2005), utilizando técnicas de simulação multivariada verificaram o problema de superestimação e subestimação de autovalores estimados da maneira convencional. Muirhead (1987) apresenta três métodos para corrigir autovalores estimados a partir das matrizes de convariâncias amostral. Este autor alerta que nem sempre essas correções mantêm a ordem decrescente de valores e dessa forma, é sugerido que se use regresão isotônica para ordenar esses dados.

4.2 Correção dos autovalores

Para corrigir o viés dos autovalores provenientes da matriz $(\boldsymbol{GE})(\boldsymbol{GE})^{\boldsymbol{t}}$ ou $(\boldsymbol{GE})^{\boldsymbol{t}}(\boldsymbol{GE})$, onde $(\boldsymbol{GE})(\boldsymbol{GE})^{\boldsymbol{t}} \sim W_g(e,\boldsymbol{\Sigma})$, usa-se quaisquer dos três métodos apresentados por Muirhead (1987):

i) A correção apresentada pelo primeiro método resulta um novo autovalor, dado por:

$$\phi_k^{(1)} = \frac{e\lambda_k^2}{\alpha_k^{(1)}} = \frac{e\lambda_k^2}{e-g+1+2\lambda_k^2 \sum_{k\neq k^*}^{p} \frac{1}{\lambda_k^2-\lambda_{k^*}^2}} \tag{4.1}$$

ii) No segundo método, um autovalor é corrigido por:

$$\phi_k^{(2)} = \frac{e\lambda_k^2}{\alpha_k^{(2)}} = \frac{e\lambda_k^2}{e-g-1+2\lambda_k^2 \sum_{k\neq k^*}^{p} \frac{1}{\lambda_k^2-\lambda_{k^*}^2}} \tag{4.2}$$

iii) E o terceiro método produz um autovalor, dado por:

$$\begin{aligned}
\phi_k^{(3)} &= \frac{e\lambda_k^2}{e+g+1-2k} - \frac{ec\lambda_k^2 \ln(e\lambda_k^2)}{b+\sum_{k=1}^{p}[\ln(e\lambda_k^2)]^2} \\
&= \frac{e\lambda_k^2(b+\sum_{k=1}^{p}[\ln(e\lambda_k^2)]^2 - (e+g+1-2k)c\ln(e\lambda_k^2))}{(e+g+1-2k)(b+\sum_{k=1}^{p}[\ln(e\lambda_k^2)]^2)} \\
&= \frac{\alpha_k^{(3)}}{d_k}
\end{aligned} \tag{4.3}$$

com $b = \dfrac{5,8(g-2)^2}{(e+g-1)^2}$, $c = \dfrac{6(g-2)}{5(e+g-1)^2}$,

$\alpha_k^{(3)} = e\lambda_k^2(b+\sum_{k=1}^{p}[\ln(e\lambda_k^2)]^2 - (e+g+1-2k)c\ln(e\lambda_k^2))$

e $d_k = (e + g + 1 - 2k)(b + \sum_{k=1}^{p}[\ln(e\lambda_k^2)]^2$.

onde:

$\phi_k^{(l)}$: é a correção do k-ésimo autovalor da matiz $(\boldsymbol{GE})(\boldsymbol{GE})^{\boldsymbol{t}}$ (ou $(\boldsymbol{GE})^{\boldsymbol{t}}(\boldsymbol{GE})$), pelo método $l = 1, 2, 3$;

g : é número de genótipos do experimento;

e : é número de ambientes do experimento;

λ_k^2 : é o k-ésimo autovalor da matriz $(\boldsymbol{GE})(\boldsymbol{GE})^{\boldsymbol{t}}$ (ou $(\boldsymbol{GE})^{\boldsymbol{t}}(\boldsymbol{GE})$), com $k = 1, 2, ..., p$ e onde $p = min\{g - 1, e - 1\}$;

Os autovalores obtidos pelas expressões (4.1), (4.2) e (4.3), nem sempre se apresentam na ordem: $\phi_1 \geq \phi_2 \geq ... \geq \phi_p$. Para colocá-los em ordem decrescente é necessário modificar os autovalores obtidos pelos métodos de correção, já que os autovalores provenientes da DVS apresetam uma ordem decrescente. Essa modificação pode ser realizada por regressão isotônica como definido por Robertson et al. (1988).

Lin & Perlman (1985) apresentam o procedimento de Stein, um algoritmo para modificar autovalores obtidos pelos métodos de correção (4.1), (4.2) e (4.3), ordenando-os de forma decrescente.

Lista-se os produtos $e\lambda_k^2$ (numerador das expressões de correção 4.1 e 4.2) ou d_k (denominador da expressão de correção 4.3) em uma coluna e em outra coluna lista-se os valores do denominador de (4.1), $\alpha_k^{(1)}$, ou os valores do denominador de (4.2), $\alpha_k^{(2)}$, ou ainda os valores do numerador da expressão (4.3), $\alpha_k^{(3)}$:

$$\begin{array}{cc} e\lambda_1^2 \ ou \ d_1 & \alpha_1^{(1)} \ ou \ \alpha_1^{(2)} \ ou \ \alpha_1^{(3)} \\ e\lambda_2^2 \ ou \ d_2 & \alpha_2^{(1)} \ ou \ \alpha_2^{(2)} \ ou \ \alpha_2^{(3)} \\ \vdots \quad \vdots & \vdots \quad \vdots \quad \vdots \\ e\lambda_p^2 \ ou \ d_p & \alpha_p^{(1)} \ ou \ \alpha_p^{(2)} \ ou \ \alpha_p^{(3)} \end{array}$$

Passo 1 Fazendo todos α_k's positivos:

a) Inicia-se pelo final da lista e procura-se para cima até que se encontre o primeiro par ($e\lambda_k^2$, α_k) com α_k negativo.

b) Soma-se este par com o par imediatamente acima dele, substituindo-os pelo par $(e\lambda_k^2 + e\lambda_{k-1}^2,\ \alpha_k + \alpha_{k-1})$, para que na lista um par seja menor do que o próximo.

c) Repete-se (a) e (b), para a nova lista até que todos α_k sejam positivos.

Passo 2 Reordenando as razões $\dfrac{e\lambda_k^2}{\alpha_k}$ de forma que estejam em ordem decrescente:

Lista-se as razões $\dfrac{e\lambda_k^2}{\alpha_k}$ à direita de cada par $(e\lambda_k^2,\ \alpha_k)$ obtido no passo **1**. Um par $(e\lambda_k^2,\ \alpha_k)$, exceto o par no final da lista, é chamado de par violado se a razão $\dfrac{e\lambda_k^2}{\alpha_k}$ não foi maior do que a razão $\dfrac{e\lambda_{k+1}^2}{\alpha_{k+1}}$.

a) Inicia-se pelo final da lista encontrada no passo **1** e procede-se para cima até o primeiro par violado ser encontrado.

b) Soma-se este par violado com o par imediatamente acima dele, substitui-se esses dois pares e suas razões pelo par $(e\lambda_k^2 + e\lambda_{k+1}^2,\ \alpha_k + \alpha_{k+1})$ e sua razão $\dfrac{e\lambda_k^2 + e\lambda_{k+1}^2}{\alpha_k + \alpha_{k+1}}$, formando uma nova lista de razões.

c) Reinicia-se no par imediatamente após o par trocado em **(b)** e procede-se para cima até o próximo par violado ser encontrado; então, repete-se **(b)**.

d) Repete-se **(c)** até todas razões $\dfrac{e\lambda_k^2}{\alpha_k}$ estarem em ordem decrescente.

Passo 3 Cada razão no final da lista é obtida por bloco acumulado de um ou mais pares consecutivos $(e\lambda_k^2,\ \alpha_k)$ na lista original.

4.3 Eficiência da correção dos autovalores

As correções sugeridas anteriormente devem ser avaliadas por algum procedimento estatístico. A estatística $PRESS$ ("Prediction Sum of Squares"), proposta por Allen (1971), consiste em um critério para escolha das variáveis regressoras a serem incluídas no modelo. Especificamente, $PRESS$ é a soma de quadrado das diferenças entre um valor observado e um valor predito pelo

modelo selecionado. Para os modelos AMMI, Cornelius et al. (1993) definiram a estatística $PRESS$ como:

$$PRESS = \sum_{i=1}^{g}\sum_{j=1}^{e}(y_{ij} - \hat{y}_{ij})^2, \tag{4.4}$$

em que:

y_{ij} : é a média de r repetição do i-ésimo genótipo localizado no j-ésimo ambiente.

$\hat{y}_{ij}$: é a média predita pelo modelo selecionado para o i-ésimo genótipo no j-ésimo ambiente.

Utilizando a estatística $PRESS$, Cornelius et al. (1993) ajustaram um valor para a medida RMSPD que faz uma comparação aproximada com outros RMSPDs provenientes de modelos ajustados por validação cruzada. O ajuste é feito por:

$$RMSPD_{(PRESS)} = \sqrt{\frac{PRESS}{ge} + \frac{(r-1)}{r}\frac{QM_{EM}}{r}} \tag{4.5}$$

em que r é o número de repetição ou o número de blocos em cada ambiente do experimento.

Na maioria dos estudos, existe um grande interesse na comparação do Erro Médio Quadrático do modelo ($QM_{EM}(modelo)$) selecionado, com o Erro Médio Quadrático (QM_{EM}) do experimento. Nachit et al. (1992) utilizaram essa comparação para encontrar uma aproximação do número de repetições que falta para o modelo AMMI completo apresentar uma performance igual ao modelo AMMI selecionado, ou seja, indica o número de repetições que se ganha ao analisar os dados com o modelo selecionado. Essa medida é obtida por:

$$R_{AMMI} = \frac{\dfrac{QM_{EM}}{r}}{QM_{EM}(modelo)}. \tag{4.6}$$

Então, para fazer uma estimativa do $QM_{EM}(modelo)$, Piepho (1994) sugere a seguinte expressão:

$$\hat{QM}_{EM}(modelo) = (RMSPD_{(PRESS)})^2 - \frac{QM_{EM}}{r} \tag{4.7}$$

Este foi o procedimento utilizado no presente estudo para avaliar o ganho em termos de número de repetições ao se fazer a correção dos autovalores nos modelos AMMI.

4.4 Exemplo

Considere os dados de um experimentos realizados em 34 países. Foram utilizados 20 genótipos de trigo sendo que um genótipo é do tipo "durum" e os outros 19 são do tipo "bread". Cada genótipo foi avaliado em 34 ambientes com 4 blocos.

Na Tabela 4.1, apresenta-se à análise de variância conjunta bem como o desdobramento da interação genótipo × ambiente efetuada com os dados observados. Verifica-se, ao nível de 1% de significância, que o efeito de genótipos, o efeito de ambientes e o efeito da interação genótipo × ambiente são significativos e suas somas de quadrados (SQ) correspondem a $1,49\%$, $72,7\%$ e $9,97\%$, respectivamente, da soma de quadrados total.

Tabela 4.1: Análise de variância conjunta do Experimento com 20 genótipos avaliado em 34 ambientes com 4 blocos e decomposição das somas de quadrados da interação genótipo × ambiente

Fonte de Variação	GL	SQ	QM	F	valor p
Blocos/ambiente	102	13.961.185,00	136.874,36	0,29	1,00
Ambiente (E)	33	4.333.925.428,00	131.331.073,58	217,34	$< 0,01$
Genótipo (G)	19	89.066.441,00	4.687.707,42	4,95	$< 0,01$
Interação (G×E)	627	594.108.485,00	947.541,44	1,97	$< 0,01$
IPCA 1	51	179.059.116,00	3.510.963,06	7,31	$< 0,01$
IPCA 2	49	91.403.060,00	1.865.368,57	3,88	$< 0,01$
IPCA 3	47	78.977.904,00	1.680.380,94	3,50	$< 0,01$
IPCA 4	45	53.975.744,00	1.199.460,98	2,50	$< 0,01$
IPCA 5	43	34.225.218,00	795.935,30	1,66	$< 0,01$
IPCA 6	41	27.621.529,60	673.695,84	1,40	0,04
IPCA 7	39	24.062.812,00	616.995,18	1,28	0,11
IPCA 8	37	23.218.126,00	627.516,92	1,31	0,10
IPCA 9	35	15.530.032,40	443.715,21	0,92	0,59
IPCA 10	33	14.201.897,60	430.360,53	0,90	0,64
IPCA 11	31	12.867.389,60	415.077,08	0,86	0,68
IPCA 12	29	10.048.801,60	346.510,40	0,72	0,86
IPCA 13	27	7.612.139,20	281.931,08	0,59	0,95
IPCA 14	25	5.691.899,60	227.675,98	0,47	0,99
IPCA 15	23	4.627.279,60	201.186,07	0,42	0,99
IPCA 16	21	4.282.728,00	203.939,43	0,42	0,99
IPCA 17	19	2.440.969,88	128.472,10	0,27	0,99
IPCA 18	17	2.308.250,76	135.779,46	0,28	0,99
IPCA 19	15	1.953.586,04	130.239,07	0,27	0,99
Resíduo	1.938	930.547.529,00	480.158,68	-	-
Total	2.719	5.961.609.068,00	-	-	-
Média (kg/ha)	3.990,80				
CV (%)	14,90				

Na mesma tabela, é feito um ajuste da interação por decomposição em

valores singulares (DVS), aplicada à matriz de interação genótipo × ambiente.

Assim na Tabela 4.1 é apresentada a análise de cada componente pelo teste F, com os graus de liberdade ajustados pelo método de Gollob (1968). Nota-se, ao nível de 1% de significância que os cinco primeiros componentes são significativos para o modelo, sendo que o primeiro componente retém $30,13\%$ da $SQ_{G\times E}$, o segundo contém $15,38\%$, $13,29\%$ é retido pelo terceiro componente, $9,08\%$ pelo quarto e $5,76\%$ pelo quinto componente. Esses cinco componentes juntos represetam $73,64\%$ da $SQ_{G\times E}$, que é considerada como resposta padrão presente na $SQ_{G\times E}$ com 235 graus de liberdade (37% dos graus de liberdade da interação).

É possível que estejam viesados os autovalores encontrados, pois foram obtidos de maneira usual e de acordo com o que foi verificado por Araújo & Dias (2005), autovalores obtidos de maneira usual podem apresentar viés. Nos modelos AMMI, a retenção de ruído por esses componentes indica que a estimativa do modelo AMMI (como média) não é perfeita (Gauch Jr, 1992).

Utilizando as sugestões de Muirhead (1987), obtêm-se os valores dos autovalores corrigidos $\phi^{(1)}$, $\phi^{(2)}$ e $\phi^{(3)}$, obtidas pelas eq. (4.1), (4.2) e (4.3), respectivamente, apresentados na Tabela 4.2. Percebe-se que os novos autovalores não satisfazem algumas condições que deveriam ser respeitadas. A ordem decrescente dos valores não é verificada para nenhum dos métodos de correção. Os valores de $\phi^{(1)}$ e $\phi^{(2)}$ apresentam outro problema que é fato de assumirem valores negativos, sendo que estes não podem assumir valores negativos pois são provenientes da soma de quadrados da interação genótipo × ambiente. Para superar esses problemas utiliza-se da regressão isotônica, mais propriamente o algoritimo de Stein apresentado por Lin & Perlman (1985). Os valores $\phi^{(1)*}$, $\phi^{(2)*}$ e $\phi^{(3)*}$ também são apresentados na Tabela 4.2 e referem-se aos autovalores corrigidos por regressão isotônica de $\phi^{(1)}$, $\phi^{(2)}$ e $\phi^{(3)}$, respectivamente.

A regressão isotônica mostrou-se muito eficaz para realizar os ajuste necessários nos autovalores corrigidos. Verifica-se ainda que $\sum_{k=1}^{19}\phi_k^{(1)*}$ equivale a 76% da $SQ_{G\times E}$ e o restante da $SQ_{G\times E}$ (24%) pode representar ruído detectado pela correção que estava presente na interação. A soma total dos autovalores corrigidos pelo método 2 e ajustados ($\sum_{k=1}^{19}\phi_k^{(2)*}$) é equivalente a 80% e 20% representa um suposto o ruído excluído da $SQ_{G\times E}$ pela correção; e $\sum_{k=1}^{19}\phi_k^{(3)*}$ representa 28% da $SQ_{G\times E}$, sendo que a correção considerou 72% da interação genótipos × ambientes como sendo supostamente ruídos, ou heterogeneidade devido a genótipos e/ou ambientes (Snee, 1982).

Os métodos de correção mostraram-se bastantes distintos quando se com-

Tabela 4.2: Correção dos autovalores da matriz $(\boldsymbol{GE})(\boldsymbol{GE})^{\boldsymbol{t}}$ e os autovalores ajustados pela regressão isotônica

λ^2	$\phi^{(1)}$	$\phi^{(1)*}$	$\phi^{(2)}$	$\phi^{(2)*}$	$\phi^{(3)}$	$\phi^{(3)*}$
44.764.779,00	26.268.136,00	26.268.136,00	27.207.272,00	27.207.272,00	6.286.366,50	6.286.366,50
22.850.765,00	11.426.689,00	12.962.548,00	11.772.993,00	13.443.869,00	4.147.974,00	4.168.439,50
19.744.476,00	15.350.388,00	12.962.548,00	16.086.043,00	13.443.869,00	4.189.740,40	4.168.439,50
13.493.936,00	9.286.514,50	9.286.514,50	9.678.314,90	9.678.314,90	3.383.543,80	3.383.543,80
8.556.304,50	4.787.919,20	4.957.793,30	4.950.884,20	5.179.251,30	2.512.930,80	2.512.930,80
6.905.382,40	3.759.430,20	4.957.793,30	3.883.808,10	5.179.251,30	2.305.817,10	2.330.966,40
6.015.703,00	2.391.437,20	4.957.793,30	2.448.698,20	5.179.251,30	2.260.432,40	2.330.966,40
5.804.531,50	-9.118.598,00	4.957.793,30	-8.347.242,00	5.179.251,30	2.432.846,10	2.330.966,40
3.882.508,10	2.142.554,10	3.602.683,10	2.214.438,40	3.831.408,40	1.847.233,70	1.876.806,40
3.550.474,40	3.884.208,80	3.602.683,10	4.151.360,30	3.831.408,40	1.883.214,10	1.876.806,40
3.216.847,40	13.991.748,00	3.602.683,10	18.802.428,00	3.831.408,40	1.903.167,50	1.876.806,40
2.512.200,40	4.050.462,30	3.602.683,10	4.474.868,40	3.831.408,40	1.667.989,70	1.667.989,70
1.903.034,80	2.732.130,30	2.732.130,30	2.984.144,90	2.984.144,90	1.420.000,30	1.420.000,30
1.422.974,90	1.710.600,20	2.419.297,40	1.840.767,30	2.739.715,80	1.195.173,10	1.195.173,10
1.156.819,90	1.125.535,60	2.419.297,40	1.193.863,70	2.739.715,80	1.093.842,40	1.116.427,80
1.070.682,00	-3.052.528,00	2.419.297,40	-2.614.122,00	2.739.715,80	1.140.977,10	1.116.427,80
610.242,47	402.691,41	2.349.044,50	418.953,91	2.739.715,80	744.352,27	772.206,06
577.062,69	-1.219.441,00	2.349.044,50	-1.084.618,00	2.739.715,80	802.991,83	772.206,06
488.396,51	-1.485.228,00	2.349.044,50	-1.259.859,00	2.739.715,80	784.362,66	772.206,06
148.527.121,0[a]		112.758.807,0[a]		119.238.403,0[a]		41.975.675,0[a]
		0,76[b]		0,80[b]		0,28[b]

a: Total da coluna; *b*: $\frac{\sum_{k=1}^{19} \phi_k^{()*}}{\sum_{k=1}^{19} \lambda_k^2}$

para as taxas de correção. O método 3 mostrou-se bastante rigoroso e apresentou as maiores taxas de correções, ou seja, antes de aplicar o teste F para os componentes. A correção descartou grande parte da $SQ_{G\times E}$ (28%), considerando essa parte que foi descartada como ruído. O método 2 foi o que se mostrou menos rigoroso e apresentou as menores taxas de correções, sendo que este método considerou 80% da $SQ_{G\times E}$. Já o método 1 mostrou-se intermediário aos métodos 2 e 3, mas apresentou resultados próximos do método 2.

Na Tabela 4.3 apresenta-se a análise de cada componente pelo teste F, com os graus de liberdade ajustados pelo método de Gollob (1968), para os autovalores corrigidos pelo método 1 (eq. (4.1)). Considerando um nível de significância de 1%, tem-se que os quatros primeiros componentes são significativos, sendo que o primeiro, o segundo, o terceiro e o quarto componente retém, respectivamente, $23,30\%$, $11,50\%$, $11,50\%$ e $8,24\%$ da $SQ_{G\times E}$. Os quatros componentes representam $54,52\%$ da $SQ_{G\times E}$, com 192 graus de liberdade, correspondente a 30,62 % dos graus de liberdade da interação, enquanto os outros $45,48\%$ da $SQ_{G\times E}$ são supostamente ruídos presente nos dados, descartados conjuntamente pelo teste F de Gollob e pelo método de correção 1.

A Tabela 4.4 apresenta a análise de cada valor singular pelo teste F, com os graus de liberdade ajustados pelo método de Gollob (1968), para os autovalores corrigidos pelo método 2 (eq. (4.2)). Considerando um nível de significância de 1%, tem-se que os quatro primeiros componentes são significativos, sendo

Tabela 4.3: Análise do Teste F para os valores singulares corrigidos pelo método 1

	GL	SQ ($\phi_k^{(1)*}$)	QM	F	valor p
1	51	26.268.136,00	515.061,49	4,29	$< 0,01$
2	49	12.962.548,00	264.541,80	2,20	$< 0,01$
3	47	12.962.548,00	275.798,89	2,30	$< 0,01$
4	45	9.286.514,50	206.366,99	1,72	$< 0,01$
5	43	4.957.793,30	115.297,52	0,96	0,54
6	41	4.957.793,30	120.921,79	1,01	0,46
7	39	4.957.793,30	127.122,91	1,06	0,37
8	37	4.957.793,30	133.994,41	1,12	0,29
9	35	3.602.683,10	102.933,80	0,86	0,70
10	33	3.602.683,10	109.172,22	0,91	0,61
11	31	3.602.683,10	116.215,58	0,97	0,51
12	29	3.602.683,10	124.230,45	1,03	0,41
13	27	2.732.130,30	101.190,01	0,84	0,69
14	25	2.419.297,40	96.771,90	0,81	0,73
15	23	2.419.297,40	105.186,84	0,88	0,63
16	21	2.419.297,40	115.204,64	0,96	0,51
17	19	2.349.044,50	123.633,92	1,03	0,42
18	17	2.349.044,50	138.179,09	1,15	0,29
19	15	2.349.044,50	156.602,97	1,30	0,19

que o primeiro, o segundo, o terceiro e o quarto componente retêm, respectivamente, $24,13\%$, $11,92\%$, $11,92\%$ e 8.58% da $SQ_{G\times E}$. A união dos quatros componentes representam $56,56\%$ da $SQ_{G\times E}$, com 192 graus de liberdade, que representam 30,62% do graus de liberdade da interação. Os outros $43,44\%$ da $SQ_{G\times E}$ são supostamente ruídos presente nos dados, descartados pelo teste F de Gollob e pelo método de correção 2.

Tabela 4.4: Análise do Teste F para os valores singulares corrigidos pelo método 2

	GL	SQ ($\phi_k^{(2)*}$)	QM	F	valor p
1	51	27.207.272,00	533.475,92	4,44	$< 0,01$
2	49	13.443.869,00	274.364,67	2,29	$< 0,01$
3	47	13.443.869,00	286.039,77	2,38	$< 0,01$
4	45	9.678.314,90	215.073,66	1,79	$< 0,01$
5	43	5.179.251,30	120.447,70	1,00	0,46
6	41	5.179.251,30	126.323,20	1,05	0,38
7	39	5.179.251,30	132.801,32	1,11	0,30
8	37	5.179.251,30	139.979,76	1,17	0,22
9	35	3.831.408,40	109.468,81	0,91	0,61
10	33	3.831.408,40	116.103,28	0,97	0,52
11	31	3.831.408,40	123.593,82	1,03	0,42
12	29	3.831.408,40	132.117,53	1,10	0,32
13	27	2.984.144,90	110.523,89	0,92	0,58
14	25	2.739.715,80	109.588,63	0,91	0,58
15	23	2.739.715,80	119.118,08	0,99	0,47
16	21	2.739.715,80	130.462,66	1,09	0,35
17	19	2.739.715,80	144.195,57	1,20	0,24
18	17	2.739.715,80	161.159,75	1,34	0,15
19	15	2.739.715,80	182.647,72	1,52	0,08

Na Tabela 4.5 apresenta-se a análise de cada valor singular pelo teste F, com os graus de liberdade ajustados pelo método de Gollob (1968), para os autovalores corrigidos pelo método 3 (eq. (4.3)). Considerando um nível de

significância de 1%, tem-se que todos valores singulares são não significativos, assim, o método 3 e o teste F de Gollob supõe que toda interação genótipo × ambiente é composta por ruídos. Logo, o modelo meramente aditivo seria o melhor para analisar os dados.

Tabela 4.5: Análise do Teste F para os valores singulares corrigidos pelo método 3

	GL	SQ ($\phi_k^{(3)*}$)	QM	F	valor p
1	51	6.286.366,50	123.262,09	1,03	0,42
2	49	4.168.439,50	85.070,19	0,71	0,93
3	47	168.439,50	3.583,82	0,03	1,00
4	45	383.543,80	8.523,20	0,07	1,00
5	43	512.930,80	11.928,62	0,10	1,00
6	41	330.966,40	8.072,35	0,07	1,00
7	39	330.966,40	8.486,32	0,07	1,00
8	37	330.966,40	8.945,04	0,07	1,00
9	35	876.806,40	25.051,61	0,21	1,00
10	33	876.806,40	26.569,89	0,22	1,00
11	31	876.806,40	28.284,08	0,24	1,00
12	29	667.989,70	23.034,13	0,19	1,00
13	27	420.000,30	15.555,57	0,13	1,00
14	25	195.173,10	7.806,92	0,07	1,00
15	23	116.427,80	5.062,08	0,04	1,00
16	21	116.427,80	5.544,18	0,05	1,00
17	19	72.206,06	3.800,32	0,03	1,00
18	17	72.206,06	4.247,42	0,04	1,00
19	15	72.206,06	4.813,74	0,04	1,00

Nota-se que, quando aplicou-se o método de Gollob (1968) com os métodos de correção dos autovalores, houve uma redução no número de componentes retidos no modelo.

A medida $RMSPD_{PRESS}$ reflete os desvios das predições do modelo AMMI para cada combinação de genótipos e ambientes, assim, a preferência deve recair sobre o modelo cuja diferença preditiva seja menor. Pelos valores da $RMSPD_{PRESS}$ obtidos para cada método de correção, verifica-se que a menor medida $RMSPD_{PRESS}$ foi obtida para o modelo ajustado para os autovalores corrigidos pelo método 2, mas ficou muito próximo da medida obtida quando se utilizou o método 1 para corrigir os mesmos autovalores. Ao utilizar o método 3 e ajustar o modelo AMMI, verificou-se o maior valor para a medida $RMSPD_{PRESS}$.

Tabela 4.6: $RMSPD_{PRESS}$ e R_{AMMI} para o melhor modelo AMMI selecionado após a correção dos autovalores pelos métodos 1, 2 e 3

	$RMSPD_{PRESS}$	R_{AMMI}
método 1	409,16	2,53
método 2	408,06	2,58
método 3	555,38	0,64

De acordo com interpretação dada por Nachit et al. (1992), a medida

R_{AMMI} indica uma aproximação para o número de repetições que falta para o modelo AMMI completo apresentar uma performance igual ao modelo AMMI selecionado, ou seja, a medida R_{AMMI} indica o número de repetições que se ganha ao analisar os dados com o modelo selecionado. Na Tabela 4.6 também encontra-se o ganho em termos do número de repetições ao se fazer a correção dos autovalores dos modelos AMMI. Verificou-se, assim, que os métodos 1 e 2 se mostraram-se melhores, fornecendo um benefício de repetições próximo de três. Já o método de correção 3 é o que apresentou o menor ganho em número de repetições, aproximadamente uma repetição.

Capítulo 5

Contribuição para a interação

5.1 Introdução

Na realização de um experimento, quando o mesmo grupo de genótipos é avaliado em diferentes ambientes, geralmente ocorre, de o ambiente alterar diferentemente o mesmo caráter em diferentes genótipos, ou seja, pode ocorrer uma interação entre fatores. Assim, a interação entre genótipos e ambientes (genótipos $\times$ ambientes) é resultante da resposta diferencial de cada genótipo à variação ambiental.

A presença da interação é essencial para a eficiência do processo podendo afetar positiva ou negativamente a sua finalização. O problema resulta da deficiência dos genótipos manterem um padrão consistente de desempenho nos ambientes. Logo, a interação não é apenas um problema, mas também uma oportunidade a ser aproveitada.

Os melhoristas de plantas compreendem que a interação genótipos $\times$ ambientes é de suma importância para a obtenção de variedades superiores. Entretanto, há a necessidade de entender a contribuição de cada genótipo e ambiente para a interação, uma vez que as avaliações do comportamento diferencial dos genótipos em função da diversidade ambiental, pode ocasionar mudanças na posição relativa dos genótipos como por exemplo, a produtividade, ou até mesmo na magnitude das suas diferenças.

A proposta de um teste F como foco de análise neste capítulo, deve-se a importância por parte dos melhoristas em conhecer a contribuição de cada genótipo e ambiente para a interação, contribuindo para a seleção de bons genótipos e ambientes nos seus estudos.

5.2 Contribuição para a interação (Teste MLPT)

Seja a matriz de interação $\mathbf{GE}_{(g\times e)} = (\widehat{ge}_{ij})$ dada da seguinte forma:

$$\mathbf{GE}_{(g\times e)} = \begin{pmatrix} \widehat{ge}_{11} & \cdots & \widehat{ge}_{1e} \\ \widehat{ge}_{21} & \cdots & \widehat{ge}_{2e} \\ \widehat{ge}_{31} & \cdots & \widehat{ge}_{3e} \\ \cdots & \cdots & \cdots \\ \widehat{ge}_{g1} & \cdots & \widehat{ge}_{ge} \end{pmatrix}$$

Como o objetivo é verificar a contribuição de cada genótipo e ambiente para a interação G×E, as hipóteses testadas são:

i) H_{01}: o i-ésimo genótipo não contribui para a interação G×E, para todo $i = 1, \ldots, g$
H_{a1}: o i-ésimo genótipo contribui para a interação G×E, para algum $i = 1, \ldots, g$

A soma de quadrados devido ao efeito de genótipo da matriz de interação **GE** é obtida por:

$$SQ_{G_i(G\times E)} = \sum_{j=1}^{e} (\widehat{ge}_{ij})^2 \ (i = 1, 2, \ldots, g), \text{ com } \frac{(g-1)(e-1)}{g} \text{ graus de liberdade,}$$

e o quadrado médio correspondente é obtido dividindo-se cada soma de quadrados pelos respectivos graus de liberdade, ou seja:

$$QM_{G_i(G\times E)} = \frac{SQ_{G_i(G\times E)}}{\frac{(g-1)(e-1)}{g}} = \frac{g(SQ_{G_i(G\times E)})}{(g-1)(e-1)}, \ i = 1, \ldots, g.$$

Observa-se que a $SQ_{G_1(G\times E)} + \cdots + SQ_{G_i(G\times E)} = SQ_{G\times E}$, com $(g-1)(e-1)$ graus de liberdade.

Tem-se que a soma de quadrados devido ao efeito de genótipo dividido pelos respectivos graus de liberdade, segue distribuição qui-quadrado não-central com parâmetro de não-centralidade θ_1, ou seja:

$$\frac{SQ_{G_i(G\times E)}}{\frac{(g-1)(e-1)}{g}} \sim \chi^2(\theta_1)$$

com $[(g-1)(e-1)]/g$ graus de liberdade em que $\theta_1 = \dfrac{\sum_{j=1}^{e}(\widehat{ge}_{ij})^2}{\sigma^2}$

Assim, verifica-se que:

$$F(G_i(G\times E)) = \frac{QM_{G_i(G\times E)}}{QM_{Res}} \sim F_{([(g-1)(e-1)]/g,(ge-1)(r-1))} \text{ para todo } i \text{ sob } H_{01}.$$

Pois sob a hipótese $H_0 : \theta_1 = 0$ versus $H_a : \theta_1 > 0$, que é equivalente a hipótese H_{01} versus H_{a1}, a estatística $F(G_i(G \times E))$ tem distribuição F central com $[(g-1)(e-1)]/g$ e $(ge-1)(r-1)$ graus de liberdade.

Ao nível α de significância, rejeita-se H_0 ou H_{01} quando,

$$F(G_i(G\times E)) = \frac{QM_{G_i(G\times E)}}{QM_{Res}} \geq F_{(\alpha,[(g-1)(e-1)]/g,(ge-1)(r-1))} \text{ para } i = 1,\ldots,g.$$

Através do teste F, aplicado a ANOVA, levando em consideração a decomposição dos $(g-1)(e-1)$ graus de liberdade da interação G×E com $[(g-1)(e-1)]/g$ graus de liberdade para os genótipos, resulta na Tabela 5.1:

Tabela 5.1: Esquema da ANOVA com teste F para obtenção de genótipos que contribuem significativamente para a interação G×E

Fonte de Variação	GL	SQ	QM	F
B d. E	$(r-1)$	$SQ_{Bd.E}$	$QM_{Bd.E}$	
G	$(g-1)$	SQ_G	QM_G	
E	$(e-1)$	SQ_E	QM_E	
G×E	$(g-1)(e-1)$	$SQ_{G\times E}$	$QM_{G\times E}$	
$G_1(G\times E)$	$\dfrac{(g-1)(e-1)}{g}$	$SQ_{G_1(G\times E)}$	$QM_{G_1(G\times E)}$	$\dfrac{QM_{G_1(G\times E)}}{QM_{Res}}$
...	...	...	...	...
$G_g(G\times E)$	$\dfrac{(g-1)(e-1)}{g}$	$SQ_{G_g(G\times E)}$	$QM_{G_g(G\times E)}$	$\dfrac{QM_{G_g(G\times E)}}{QM_{Res}}$
Resíduo	$(ge-1)(r-1)$	SQ_{Res}	QM_{Res}	
Total	$ger-1$	SQ_{Total}		

$G_i(G\times E)$: é o efeito do i-ésimo genótipo dentro da interação, com $i = 1,\ldots,g$.

ii) H_{02}: o j-ésimo ambiente não contribui para a interação G×E, para todo $j = 1,\ldots,e$

H_{a2}: o j-ésimo ambiente contribui para a interação G×E, para algum $j = 1,\ldots,e$

A soma de quadrados devido ao efeito de ambiente da matriz de interação **GE** é obtida por:

$$SQ_{E_j(G\times E)} = \sum_{i=1}^{g}(\widehat{ge}_{ij})^2 \ (j = 1,2,\ldots,e), \text{ com } \frac{(g-1)(e-1)}{e} \text{ graus de liberdade,}$$

e o quadrado médio correspondente é obtido dividindo-se cada soma de quadrados pelos respectivos graus de liberdade, ou seja:

$$QM_{E_j(G\times E)} = \frac{SQ_{E_j(G\times E)}}{\frac{(g-1)(e-1)}{e}} = \frac{e(SQ_{E_j(G\times E)})}{(g-1)(e-1)},\ j = 1, \ldots, e.$$

Observa-se que a $SQ_{E_1(G\times E)} + \cdots + SQ_{E_j(G\times E)} = SQ_{G\times E}$, com $(g-1)(e-1)$ graus de liberdade.

Tem-se que a soma de quadrados devido ao efeito de ambiente dividido pelos respectivos graus de liberdade, segue distribuição qui-quadrado não-central com parâmetro de não-centralidade θ_2, ou seja:

$$\frac{SQ_{E_j(G\times E)}}{\frac{(g-1)(e-1)}{e}} \sim \chi^2(\theta_2)$$

com $[(g-1)(e-1)]/e$ graus de liberdade em que $\theta_2 = \dfrac{\sum_{i=1}^{g}(\widehat{ge}_{ij})^2}{\sigma^2}$

Assim verifica-se que:

$$F(E_j(G\times E)) = \frac{QM_{E_j(G\times E)}}{QM_{Res}} \sim F_{([(g-1)(e-1)]/e,(ge-1)(r-1))} \text{ para todo } j \text{ sob } H_{02}.$$

Pois sob a hipótese $H_0 : \theta_2 = 0$ versus $H_a : \theta_2 > 0$, que é equivalente a hipótese H_{02} versus H_{a2}, a estatística $F(E_j(G \times E))$ tem distribuição F central com $[(g-1)(e-1)]/e$ e $(ge-1)(r-1)$ graus de liberdade.

Ao nível α de significância, rejeita-se H_0 ou H_{02} quando,

$$F(E_j(G\times E)) = \frac{QM_{E_j(G\times E)}}{QM_{Res}} \geq F_{(\alpha,[(g-1)(e-1)]/e,(ge-1)(r-1))} \text{ para } j = 1, \ldots, e.$$

Através do teste F, aplicado a ANOVA, levando em consideração a decomposição dos $(g-1)(e-1)$ graus de liberdade da interação G×E com $[(g-1)(e-1)]/e$ graus de liberdade para os ambientes, resulta no seguinte esquema:

Tabela 5.2: Esquema da ANOVA com teste F para obtenção de ambientes que contribuem significativamente para a interação G×E

Fonte de variação	GL	SQ	QM	F
B d. E	$e(r-1)$	$SQ_{B\ d.\ E}$	$QM_{B\ d.\ E}$	
G	$(g-1)$	SQ_G	QM_G	
E	$(e-1)$	SQ_E	QM_E	
G×E	$(g-1)(e-1)$	$SQ_{G\times E}$	$QM_{G\times E}$	
$E_1(G\times E)$	$\frac{(g-1)(e-1)}{e}$	$SQ_{E_1(G\times E)}$	$QM_{E_1(G\times E)}$	$\frac{QM_{E_1(G\times E)}}{QM_{Res}}$
...	...	...	...	...
$E_e(G\times E)$	$\frac{(g-1)(e-1)}{e}$	$SQ_{E_e(G\times E)}$	$QM_{E_e(G\times E)}$	$\frac{QM_{E_e(G\times E)}}{QM_{Res}}$
Resíduo	$e(g-1)(r-1)$	SQ_{Res}	QM_{Res}	
Total	$ger-1$	SQ_{Total}		

$E_j(G\times E)$: é o efeito do j-ésimo ambiente para a interação, com $j=1,\ldots,e$.

A proposta dos graus de liberdade igualitários para genótipos e ambientes, vai de encontro à proposta de Gauch Jr (1992) para obtenção dos graus de liberdade dos componentes multiplicativos de um modelo $AMMI$.

5.3 Exemplo

Considere os dados obtidos pelo CIMMYT (Centro Internacional de Mejoramiento de Maiz y Trigo) em experimentos realizados em vários países. Foram utilizados genótipos de milho e trigo sendo que em todos os experimentos utilizou-se o delineamento aleatorizado em blocos. Cada conjunto tem a seguinte descrição:

> Conjunto 1: 20 genótipos de trigo, sendo que um genótipo é do tipo trigo "durum" e os outros 19 são do tipo trigo "bread". Cada genótipo foi avaliado em 34 ambientes com 4 blocos;
>
> Conjunto 2: 9 genótipos de milho avaliados em 20 ambientes com 4 blocos.

A Tabela 5.3 corresponde a ANOVA efetuada com o Conjunto 1.

Através do teste F, aplicado a ANOVA, levando em consideração a decomposição dos $(g-1)(e-1)=(20-1)(34-1)=627$ graus de liberdade da interação G×E com $[(g-1)(e-1)]/g=[(20-1)(34-1)]/20=31,35$ graus de liberdade para os genótipos, resulta na Tabela 5.4:

Nota-se que somente os genótipos 3, 4, 6 e 15 são não significativos ao nível de 5% de significância, ou seja, tais genótipos não contribuem significativamente para a interação genótipos × ambientes. Os demais genótipos são significativos e contribuem mais para a interação. Sendo assim, tais genótipos

Tabela 5.3: ANOVA do Conjunto 1 com 20 genótipos de trigo avaliados em 34 ambientes com 4 blocos

Fonte de Variação	GL	SQ	QM	F	valor-p
B d. E	102	257862519	2528064	7,14	¡0,0001
G	19	89066441	4687707	13,23	¡0,0001
E	33	4333925428	131331074	370,67	¡0,0001
G×E	627	594108485	947541	2,67	¡0,0001
Resíduo	1938	686646195	354307		
Total	2719	5961609068			

Tabela 5.4: Teste F, aplicado ao Conjunto 1, para obtenção de genótipos que contribuem significativamente para a interação G×E

Fonte de Variação	GL	SQ	QM	F	valor-p
$G_1(G \times E)$	31,35	38419779	1225511,30	3,4589006	¡0,0001
$G_2(G \times E)$	31,35	18000775	574187,39	1,6205947	0,0164
$G_3(G \times E)$	31,35	15051875	480123,62	1,3551077	0,0910
$G_4(G \times E)$	31,35	13438384	428656,59	1,2098465	0,1970
$G_5(G \times E)$	31,35	26944156	859462,70	2,4257598	¡0,0001
$G_6(G \times E)$	31,35	15771832	503088,74	1,4199248	0,0619
$G_7(G \times E)$	31,35	34115953	1088228,20	3,0714306	¡0,0001
$G_8(G \times E)$	31,35	26152950	834224,88	2,3545282	¡0,0001
$G_9(G \times E)$	31,35	18638713	594536,31	1,6780278	0,0109
$G_{10}(G \times E)$	31,35	27052040	862903,98	2,4354725	¡0,0001
$G_{11}(G \times E)$	31,35	18457831	588766,55	1,6617431	0,0122
$G_{12}(G \times E)$	31,35	38927562	1241708,50	3,5046158	¡0,0001
$G_{13}(G \times E)$	31,35	20045716	639416,78	1,8046990	0,0042
$G_{14}(G \times E)$	31,35	38666042	1233366,60	3,4810713	¡0,0001
$G_{15}(G \times E)$	31,35	13098543	417816,37	1,1792509	0,2278
$G_{16}(G \times E)$	31,35	34294875	1093935,40	3,0875388	¡0,0001
$G_{17}(G \times E)$	31,35	41353606	1319094,30	3,7230305	¡0,0001
$G_{18}(G \times E)$	31,35	20927058	667529,76	1,8840455	0,0022
$G_{19}(G \times E)$	31,35	37625900	1200188,20	3,3874282	¡0,0001
$G_{20}(G \times E)$	31,35	97124895	3098082,80	8,7440729	¡0,0001

podem ser descartados pelos melhoristas, uma vez que apresentam respostas heterogêneas quanto aos ambientes.

Quanto aos ambientes, observa-se que os ambientes 2, 3, 8, 11, 13, 17, 21, 28, 30 e 31 são os que não contribuem para a interação genótipos × ambientes, sendo esses ambientes que podem ser escolhidos pelos melhoristas por serem ambientes de resposta homogênea aos genótipos para essa variável. Os demais ambientes contribuem significativamente, ao nível de 5% de significância, para a interação.

A Tabela 5.5 corresponde a ANOVA efetuada com o Conjunto 2.

Através do teste F, aplicado a ANOVA, levando em consideração a decomposição dos $(g-1)(e-1) = (9-1)(20-1) = 152$ graus de liberdade da

Tabela 5.5: ANOVA do Conjunto 2 com 9 genótipos de milho avaliados em 20 ambientes com 4 blocos

Fonte de Variação	GL	SQ	QM	F	valor-p
B d. E	60	118813053,8	1980217,6	3,28	¡0,0001
G	8	79828574,7	9978571,8	16,55	¡0,0001
E	19	989593771,8	52083882,7	86,40	¡0,0001
G×E	152	249704161,7	1642790,5	2,73	¡0,0001
Resíduo	480	289366499,0	602847,0		
Total	719	1727306061,0			

interação G×E com $[(g-1)(e-1)]/g = [(9-1)(20-1)]/9 \approx 16,89$ graus de liberdade para os genótipos, resulta na Tabela 5.6:

Tabela 5.6: Teste F, aplicado ao Conjunto 2, para obtenção de genótipos que contribuem significativamente para a interação G×E

Fonte de Variação	GL	SQ	QM	F	valor-p
$G_1(G \times E)$	16,89	35147999	2081131,50	3,4521727	¡0,0001
$G_2(G \times E)$	16,89	17268730	1022490,60	1,6961034	0,0405
$G_3(G \times E)$	16,89	10872797	643784,06	1,0679064	0,3829
$G_4(G \times E)$	16,89	42526664	2518026,20	4,1768918	¡0,0001
$G_5(G \times E)$	16,89	26242317	1553821,40	2,5774728	0,0006
$G_6(G \times E)$	16,89	15653251	926837,25	1,5374339	0,0779
$G_7(G \times E)$	16,89	22544078	1334846,70	2,2142384	0,0037
$G_8(G \times E)$	16,89	53853715	3188706,80	5,2894142	¡0,0001
$G_9(G \times E)$	16,89	25594611	1515470,40	2,5138562	0,0008

Tem-se para o Conjunto 2 que somente os genótipos 3 e 6 são não significativos ao nível de 5% de significância, sendo assim, os demais genótipos são os que contribuem mais para a interação genótipos × ambientes.

Quanto aos ambientes, observa-se que os ambientes 2, 4, 5, 6, 7, 9, 10, 14, 16, 17 e 20 são os que não contribuem para a interação genótipos × ambientes, enquanto que os demais ambientes contribuem significativamente, ao nível de 5% de significância, para a interação.

Capítulo 6

Introdução aos métodos de imputação

6.1 Introdução

Os métodos de imputação, também conhecidos como métodos de substituição, foram criados com a finalidade de resolver os problemas ocorridos em experimentos, cuja unidade observacional não fornece resposta em alguma ou algumas das variáveis devido a alguns fatores que surgem durante a realização da pesquisa. Estas unidades passaram a ser conhecidas como dados ausentes ou faltantes, cuja existência pode interferir nos resultados da pesquisa, produzindo respostas não confiáveis.

Os fatores causadores deste problema variam conforme a área de conhecimento. Em análise de séries temporais a falta de alguma informação, pode ocorrer devido à ausência de valores de precipitação em séries históricas. Em uma instituição financeira, unidades ausentes podem ser causadas devido a recusa de clientes para fornecer informação sobre a variável quantidade de meses de conta corrente, pois o seu preenchimento não é obrigatório. Em pesquisas clínicas, este problema ocorre, quando nem todos os exames exigidos são realizados pelos pacientes. Em experimentos agrícolas, a falta de resposta podem ocorrer porque os animais morrem ou porque as plantas estão danificadas (Krzanowski, 1988).

Este problema também tem se tornado presente em estudos de melhoramento genético, em que a falta de genótipos em alguns ambientes geram matrizes incompletas e, conseqüentemente, dificultam o uso de técnicas multivariadas, pois exigem para sua aplicação uma matriz de dados completa.

De forma geral, os métodos de imputação se baseiam na estimação de unidades ausentes, utilizando algum método estatístico conforme o mecanismo que gerou a falta (ausência totalmente aleatória-*Missing complete et Random-MCA*, ausência aleatória-*Missing at Random-MAR* e ausência não aleatória-

Missing Not at Random-MNAR). Esta estatística nos fornecerá um ou mais valores, os quais serão candidatos a substituírem as unidades que não forneceram resposta, produzindo um vetor ou uma matriz de dados completa.

Os métodos de imputação foram introduzidos por Rubin (1976). A princípio, a idéia fundamentou-se em estimar os valores ausentes uma única vez para cada valor ausente, o qual ficou conhecido como imputação simples. Porém, foi verificado que tais técnicas produziam estimativas desviadas dos verdadeiros valores (valores que poderiam ter ocorrido se as unidades tivessem fornecido respostas).

Visando obter estimativas mais próximas do valor real, Rubin (1987) desenvolveu métodos que forneceram estimativas com menos viés comparado com os métodos simples. Estes métodos ficaram conhecidos como imputação múltipla, pois os valores ausentes eram estimados pela junção de várias estimativas geradas a partir das unidades observadas. Com o avanço da tecnologia, tais métodos foram ganhando espaço nas pesquisas. Os métodos de simulação bayesiana foram criados e, posteriormente, utilizados para a resolução de unidades ausentes. Em seguida, novos métodos foram surgindo e outros são continuamente estudados nos dias atuais.

Métodos que não exigem suposição sobre a distribuição ou estrutura dos dados, também foram desenvolvidos como, por exemplo, o método de imputação múltipla livre de distribuição, o qual utiliza a técnica de decomposição por valor singular. Detalhes podem ser visto em Bergamo et al. (2008), Arciniegas-Alarcón e Dias (2009).

Este capítulo apresentará, de forma introdutória, os conceitos de alguns métodos de imputação, focalizando no método de imputação múltipla com enfoque bayesiano, com aplicação em problemas de dados faltantes em estudos de melhoramento Genético. Visto que, a idéia base é direcionar os leitores ao uso das ferramentas fornecidas pelo software SAS com o surgimento deste problema em suas pesquisas.

6.2 Padrões de dados ausentes

Os padrões de dados ausentes são importantes para identificar a forma com que as unidades ausentes estão distribuídas em um vetor ou matriz de dados, descrevendo a localização dos mesmos (Enders, 2010). Esta localização pode ser expressa de várias maneiras.

- **Padrão univariado** (*Univariate Pattern*): apresenta uma falta de dados

isoladamente em uma variável, o que é comum em estudos experimentais.

- **Padrão de não-resposta** (*Unit Nonresponse Pattern*): freqüentemente ocorre em pesquisas realizadas por meio de questionários como o censo, pesquisas domiciliares, em que alguns ítens são respondidos pelos indivíduos e outros são recusados, assim item sem resposta é considerado como unidade ausente.

- **Padrão monótono** (*Monotone Pattern*): geralmente ocorre em pesquisas clínicas, em que os indivíduos participantes da pesquisa em algum momento não podem continuar no estudo devido à alguns fatores, por exemplo, reação de alguma droga em análise. Este tipo de padrão de dados em falta é característico de experimentos longitudinais, sendo as variáveis medidas ao longo do tempo.

- **Padrão geral** (*General Pattern*): padrão conhecido como arbitrário que consiste numa dispersão de unidades ausentes por toda a matriz de dados. Aparentemente é aleatório, porém pode existir uma relação entre a falta de valores de uma variável e a tendência da falta de dados referente à outra variável medida.

Considerando uma matriz de dados com 4 variáveis em estudos, os padrões supracitados tem as características mostradas na figura 6.1. As áreas sombreadas representam a localização dos valores em falta no conjunto de dados.

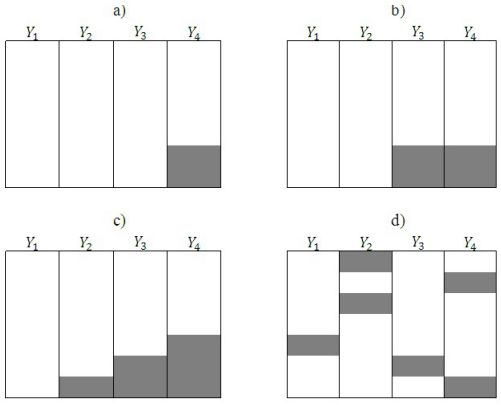

Figura 6.1: Alguns padrões de comportamento de dados ausentes: (a) Padrão univariado, (b) Padrão de não resposta, (c) Padrão monótono e (d) Padrão geral

6.3 Distribuição dos dados ausentes-Teoria de Rubin

A teoria de Rubin consiste em dividir a matriz de dados em dois subconjuntos, um contendo apenas os valores observados $\mathbf{Y}_{obs}$ e outro contendo os valores que não foram observados ($\mathbf{Y}_{aus}$). Desta forma podemos representar a matriz de dados da seguinte maneira,

$$\mathbf{Y}_{com} = (\mathbf{Y}_{obs}, \mathbf{Y}_{aus})$$

assim, uma matriz de dados retangular $(n \times p)$ em que as linhas são representadas pelos amostra aleatória de indivíduos $(i = 1, 2, ..., n)$, de alguma distribuição de probabilidade multivariada p-dimensional, e as colunas representam as p variáveis $(j = 1, 2, ..., p)$, em que, os valores das variáveis para o i-ésimo indivíduo podem ser agrupadas em um vetor $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{ip})^T$.

Considerando uma variável da matriz de dados multivariada p-dimensional observa-se que,

$$\mathbf{Y}_{i1} = \{Y_{11}, Y_{21}, ..., Y_{n1}\} = \{Y_{11}, Y_{21}, Y_{r1}, Y_{r+1}..., Y_{n1}\}$$

sendo $\mathbf{Y}_{obs} = \{Y_{11}, Y_{21}, Y_{r1}\}$ que corresponde aos valores que foram observados e $\mathbf{Y}_{aus} = \{Y_{r1+1}..., Y_{n1}\}$ refere-se aos valores ausentes da variável. Observa-se ainda que, o conjunto de dados contém r valores observados e $m = n - r$ valores ausentes. Sendo assim, o autor cria uma variável indicadora R, com o objetivo de estudar o comportamento dos dados ausentes, a qual fornece uma distribuição de probabilidade da falta completa, ou seja, uma distribuição de probabilidade indicando se R_i assume o valor 0 se o indivíduo não apresentar resultado sobre a variável em estudo, caso contrário assumirá o valor 1.

$$\mathbf{R} = \begin{cases} 1, & \text{se } Y_{ij} \text{ é observado;} \\ 0, & \text{se } Y_{ij} \text{ é não observado.} \end{cases}$$

Tal distribuição, depende da forma com que os dados ausentes se distribuem ao longo da matriz de dados, sendo de suma importância quando se pretende verificar a causa da falta dos dados ausentes.

6.4 Mecanismos que levam a falta de dados

Os mecanismos de dados ausentes consistem em verificar as relações existentes entre os valores perdidos e a probabilidade de ausência, informando o que a gerou (Enders, 2006). Estes mecanismos são classificados como:

- **Ausência totalmente aleatória**: a ausência ocorre de forma totalmente aleatória se a probabilidade da falta de dados sobre a variável $\mathbf{Y}$ não está relacionada com alguma outra variável medida e, não tem relação com os valores de $\mathbf{Y}$. Este mecanismo é representado por MCAR (*Missing completely at Random*) e a distribuição da falta completa contém um parâmetro ϕ que é importante para expressar a probabilidade de que $\mathbf{R}$ assuma um valor 0 ou 1. Neste mecanismo, a falta completa não está relacionada com os dados, assim esta distribuição pode ser expressa por $p(\mathbf{R} \mid \boldsymbol{\phi})$.

- **Ausência de forma aleatória**: a ausência de dados ocorre de forma aleatória se a probabilidade de uma variável ausente depende das informações disponíveis na matriz de dados que contém as variáveis medidas, porém, em muitas situações experimentais esta ausência não é completamente aleatória. Este mecanismo é conhecido na literatura por *Missing at Random* representado por MAR. Sua distribuição indica que a probabilidade da falta completa ($\mathbf{R}$) depende da proporção de dados observados por meio de algum parâmetro ϕ, que relaciona $\mathbf{Y}$ e $\mathbf{R}$. A distribuição pode ser expressa como $p(\mathbf{R} \mid \mathbf{Y_{obs}}, \boldsymbol{\phi})$

- **Ausência não aleatória**: em algumas situações experimentais, a causa da ausência depende de algumas informações que não foram observadas ou da variável em si. Neste caso, a falta é considerada não aleatória, sendo conhecida como *Missing Not at Random* e representado por MNAR. A distribuição de dados em falta, indica a probabilidade da falta completa assumir um valor de 0 ou 1 dependendo de $\mathbf{Y}_{obs}$ e $\mathbf{Y}_{mis}$. Esta distribuição pode ser expressa por $p(\mathbf{R} \mid \mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \boldsymbol{\phi})$.

Os dados MCAR e MAR são chamados de dados ausentes ignoráveis, já os dados MNAR são conhecidos como não ignoráveis (Nunes, 2007). Na prática, os dados MNAR são mais complicados de serem analisados comparado com os demais. De forma simples e compreensível, Schafer e Graham (2002) apresenta graficamente, no caso univariado, os mecanismos de ausência de dados (Figura 6.2). Na figura 6.2 temos $\mathbf{X}$, que representa a variável que sofreu falta de dados, $\mathbf{Y}$ é a variável que contém valores em falta, $\mathbf{Z}$ representa uma determinada variável que não foi observada e que pode ter influenciado na ausência de dados e, $\mathbf{R}$ representa a falta completa (*missingness*), a qual expressa os valores em falta por meio de uma variável indicadora que assume o valor 0 quando não foi observado e 1 quando observado.

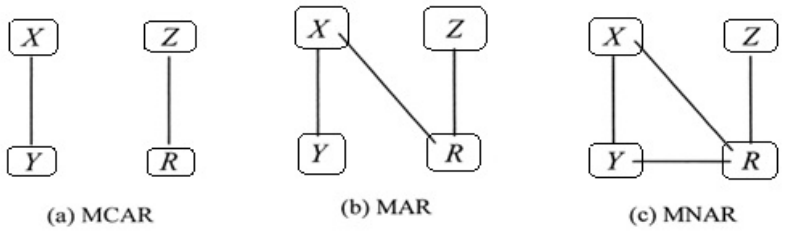

Figura 6.2: Representação gráfica: (a) ausência completamente aleatória, (b) ausência de forma aleatória, (c) ausência de forma não aleatória

Ainda se observa que, quando os dados em falta seguem um mecanismo MCAR, a falta completa (**R**) não tem relação com a variável em falta (**Y**) e com a outra variável que foi medida (**X**), ou seja, a causa do valor ausente está relacionado com outras variáveis que não foram medidas ou por causas aleatórias que não podem ser controladas, porém quando ocorre um mecanismo MCAR, a falta completa apresenta uma relação tanto com os fatores aleatórios ou variáveis que não foram medidas como também existe uma relação com a própria variável que contém a falta. Em relação ao mecanismo MNAR, verifica-se que a falta completa (**R**) está relacionada com todos os outros componentes (**X**, **Y** e **Z**).

6.5 Abordagens para o tratamento de dados ausentes

A ocorrência de dados ausentes em uma matriz de dados, pode influenciar nos resultados da pesquisa, no poder dos testes estatísticos, causando diminuição, e também viés nas estimativas. Este problema também dificulta as análises estatísticas, pois para a utilização da maioria das técnicas estatísticas multivariadas é necessário uma matriz de dados completa. Como alternativa, existe diversas maneiras de lhe dar com este problema. Um deles são os métodos tradicionais, que envolvem a retirada dos indivíduos que não apresentaram resposta para uma das variáveis. Outro são os métodos de imputação simples, que consistem em substituir os valores ausentes por apenas uma estimada, obtida a partir dos dados disponíveis na matriz. E também existe, os métodos de imputação múltipla que são similares ao método simples, porém estimam vários valores por meio de algum modelo estatístico para cada observação em falta e posteriormente, são agrupadas em apenas um valor. Alguns métodos, são descritos a seguir para ambas classes e podem ser encontrados com detalhes nas obras de Rubin (1987), Enders (2010) e Medina e Galván

(2007).

6.5.1 Métodos tradicionais

Em uma análise de dados, quando o pesquisador se depara com indivíduos que não forneceram informações a uma das variáveis, estes são excluídos, deixando apenas os indivíduos com informações completas. Este procedimento é característico do método de eliminação, conhecido como análise de dados completos ou *listwise deletion*. Existe também, o método de análise de dados disponíveis ou *pairwise deletion*. Ambos assumem o mecanismo MCAR para descrever o comportamento das unidades ausentes.

Estes métodos são simples de implementação e também, padrão em alguns programas estatísticos. Porém, não são classificados como a melhor solução para o caso de dados ausentes, pois com o aumento das unidades em falta e com a violação das suposições sobre o mecanismo exigido, apresentam diminuição no poder do teste, perda de informação e também resultados falseados, assim como estimativas dos parâmetros tendenciosos. Desta forma, seu uso não é recomendável. Uma breve descrição sobre tais métodos, é apresentado a seguir.

- *Listwise deletion (LD)*: consiste em eliminar os indivíduos que contém um ou mais valores ausentes e prossegui-se as análises com os dados disponíveis. Segundo Medina e Galván (2007) a partir do momento que se exclui indivíduos, considera-se que a subamostra de indivíduos excluídos tem a mesma característica dos indivíduos completos e que, a falta de resposta foi gerada de forma aleatória, o que na realidade não ocorre. Este problema gera conseqüências, como estimativas dos parâmetros viesadas.

- *Pairwise deletion (PD)*: este método consiste em atenuar a perda de dados, estimando os parâmetros de interesse separadamente, para cada variável em estudo. Sendo assim, quando se tem uma base de dados com mais de uma variável, elimina-se os valores ausentes para cada uma delas e posteriormente, estima-se os parâmetros. Este método apresenta desvantagens quando se pretende aplicar algum método multivariado, pois para cada variável em estudo, ao se excluir os valores ausentes gera-se tamanhos amostrais diferentes, o que não é considerável pelos testes multivariados. Problemas também podem ser causados quando se pretende calcular o coeficiente de correlação, pois os valores obtidos podem se encontrar fora do intervalo [0,1]. Em algumas situações, este método se limita a comparações de resultados (Medina e Galván, 2007).

6.5.2 Imputação simples

Com a finalidade de atenuar os problemas existentes nos métodos tradicionais, Rubin (1976) propôs um método que, em vez de excluir os valores ausentes, usaria as informações adquiridas pelos indivíduos presentes com a tentativa de recuperar as informações perdidas por meio da estimação. Estas técnicas são conhecidas como métodos de imputação.

A figura Figura 6.3, apresenta um banco de dados com p variáveis e n indivíduos, sendo que alguma informação não foi coletada em uma ou mais variáveis.

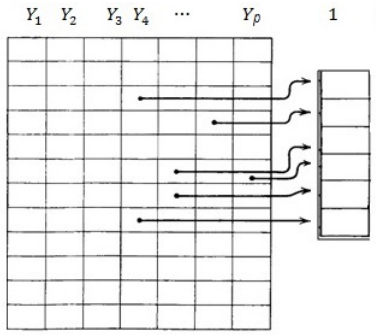

Figura 6.3: Conjunto de dados com uma única imputação para cada valor ausente

Observa-se que para cada valor ausente sua estimação se deu apenas uma única vez. Este processo é conhecido como imputação simples ou única.

De forma geral, a imputação simples é definida como sendo uma técnica que substitui valores em falta por valores estimados, a partir das unidades completas em um conjunto de dados específico conforme o mecanismo que o gerou. Tal método tem suas vantagens: gera um conjunto de dados completo, suas estimativas são mais coerentes que os métodos tradicionais por utilizarem mais informações ao se estimar os valores ausentes, não diminui a tamanho amostral real e suaviza os erros padrão. Mesmo diante destas vantagens, o método de imputação simples apresenta desvantagens: fornece estimativas desviadas do valor real e os erros padrão subestimados. Dentre as técnicas de imputação simples se destacam:

- **Imputação média**: consiste em substituir os valores ausentes que corresponde a uma determinada variável em estudo, pelo valor médio dessa variável. Outro método similar é calcular o valor mediano ou a moda desta variável para ser o doador das unidades em falta. A medida que a porcentagem de falta aumenta em cada variável, o método vai perdendo

a precisão nas estimativas e, conseqüentemente, falsas informações sobre elas, serão obtidas.

- **Imputação do vizinho mais próximo**: é um método que utiliza o valor mais próximo de uma variável auxiliar, correspondente ao valor ausente. Este método pode proporcionar valores eqüidistastes. Neste caso, se faz um sorteio aleatório dos mesmos, para ser incluído na respectiva casela vazia (Rosas et al., 2009).

- **Imputação *hot deck***: é um método que tem como foco, identificar o indivíduo com valor mais parecido com o individuo com valor ausente, baseando-se com uma determinada variável auxiliar. Desta forma, todos os indivíduos são divididos em grupos com características semelhantes e, posteriormente, o valor a ser imputado na variável referente ao valor ausente é retirado aleatoriamente de um conjunto que contém características semelhantes (He, 2006).

- **Imputação pela regressão**: é um método também conhecido como imputação da média condicional, o qual consiste em estimar os valores ausentes por meio de valores previstos de um modelo de regressão linear com os dados disponíveis das variáveis completas, ou seja, com os indivíduos que forneceram respostas em todas as variáveis. No caso bivariado, o processo torna-se mais simples. Porém no caso em que envolve mais de duas variáveis, o processo se torna complexo, pois para cada variável que contém valores ausentes pode existir diferentes combinações de modelo de regressão. Considerando, por exemplo, um conjunto de dados com três variáveis em estudo (Y_1,Y_2 e Y_3) e que, em cada uma existe valores em falta. Desconsiderando os indivíduos com informações em todas as variáveis, podemos ter os seguinte padrões de falta de dados: 1) Falta em apenas Y_1 ($\hat{y}_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3$); 2) Falta em Y_2 ($\hat{y}_2 = \beta_0 + \beta_1 y_1 + \beta_2 y_3$); 3) Falta em Y_3 ($\hat{y}_3 = \beta_0 + \beta_1 y_1 + \beta_2 y_2$); 4) Falta em Y_1 e Y_2 ($\hat{y}_1 = \beta_0 + \beta_1 y_3$ e $\hat{y}_2 = \beta_0 + \beta_1 y_3$); 5) Falta em Y_1,Y_3 ($\hat{y}_1 = \beta_0 + \beta_1 y_2$ e $\hat{y}_3 = \beta_0 + \beta_1 y_2$) e; 6) Falta em Y_2,Y_3 ($\hat{y}_2 = \beta_0 + \beta_1 y_1$ e $\hat{y}_3 = \beta_0 + \beta_1 y_1$) Enders (2010). Apesar de ser um método que atenua o viés das estimativas comparado com os métodos supracitados, ainda não é o método mais recomendado, pois apresenta medida de correlação superestimado mesmo quando os valores em falta seguem um comportamento MCAR. Com tentativa de eliminar este problema, foi proposto o método de imputação pela regressão estocástica.

- **Imputação pela regressão estocástica**: este método consiste em estimar os valores ausentes de forma similar ao método de imputação pela regressão, diferindo apenas no fato de que, após a estimação da equação da reta de regressão é adicionado aos valores previstos, um termo residual distribuído normalmente com média zero e variância igual a variância residual do modelo regressão estimado. Por exemplo, se temos duas variáveis em questão Y_1 e Y_2, sendo a equação estimada $\hat{y}_1 = \beta_0 + \beta_1 y_2$, então adiciona-se para cada valor previsto um termo residual, ou seja, $\hat{y}_1 = \beta_0 + \beta_1 y_2 + z_i$, em que $z_i \sim N(0, \sigma^2{}_\varepsilon)$. Este método elimina o viés causado pelo método de regressão e, recupera a variabilidade perdida pelos valores em falta.

- **Imputação pela máxima verossimilhança**: consiste em estimar os parâmetros de uma determinada distribuição utilizando as informações das variáveis presentes no conjunto de dados, sem a necessidade de excluir os indivíduos que não forneceram informações para mais de uma delas, ou seja, o método utiliza todos os indivíduos que fornece pelo menos uma informação para o conjunto de variáveis em estudo. Desta forma, o cálculo da log-verossimilhança, para cada indivíduo, depende apenas das variáveis e dos parâmetros que contém informações, fornecendo diferentes valores para cada caso de falta associado ao individuo do banco de dados. Tal método, tem como foco identificar o conjunto de parâmetros que maximizam a função de log-verossimilhança. Este método de imputação, produz estimativas imparciais dos parâmetros quando os dados em falta apresentam um mecanismo MCAR ou MAR, porém se torna desapropriado quando os dados seguem um mecanismo MNAR. Por ser um método presente na maioria dos programas estatísticos, seu uso tem sido amplamente presente nas pesquisas. Detalhes sobre seu processo de imputação pode ser encontrado em Enders (2010).

Segundo Enders (2010), dentre os métodos simples citados, o método de imputação por máxima verossimilhança produz estimativa dos parâmetros com menos viés.

6.5.3 Imputação múltiplas

Mesmo com o surgimento dos métodos simples, o aumento do viés continuou sendo um problema para se obter estimativas mais precisas dos dados em falta. Desta forma, foram desenvolvidas novas técnicas com a finalidade

de obter estimativas que refletisse a incerteza sobre as previsões dos dados em falta. Estas técnicas são conhecidas como métodos de imputação múltipla, as quais geram um conjunto de estimativas razoáveis que representam a incerteza sobre o valor a ser imputado (Rubin, 1987).

A imputação múltipla consiste em gerar m estimativas para cada valor em falta, por meio de algum método de imputação adequado conforme o mecanismo que gerou a falta. Considerando um conjunto de dados multivariado com p variáveis, observa-se que, para cada valor em falta, m imputações são realizadas, construindo uma matriz auxiliar composta por m vetores de valores estimados para os m valores em falta (figura 6.4).

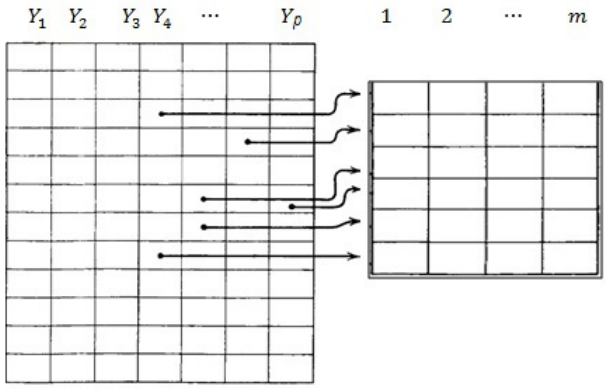

Figura 6.4: Conjunto de dados com m imputações para cada valor ausente

Após a obtenção destes m vetores, os valores estimados são substituídos na matriz de dados original, fornecendo m conjuntos de dados completos ou imputados (Y_{obs},Y_{aus}). Esta metodologia foi, primeiramente, proposta como uma ferramenta para resolução de valores em falta, em amostras grande, nas pesquisas públicas. Em seguida, adaptado para outros contextos estatísticos (Reiter et al., 2007).

A imputação múltipla é realizada por meio de três fases:

1. **Fase de imputação**: Os dados ausentes são preenchidos em m ($m > 1$) tempos gerando m conjuntos de dados completos por meio de técnicas adequadas de imputação. De acordo com a literatura m fica entre 3 e 10 imputações. Este passo é realizado por meio de algoritmo iterativo contendo dois passos (Passo-I e Passo-P).

2. **Fase de análise**: Os m conjuntos de dados completos são analisados usando técnicas estatísticas padrão.

3. **Fase de agrupamento**: Os resultados dos m conjuntos de dados completos são combinados para produzir inferência dos resultados a serem imputados.

Diversas técnicas de imputação múltiplas estão implementadas em programas estatísticos como o sistema computacional SAS (Berglund, 2010; Yuan, 2010) por meio dos processos MI (Multiple Imputation) e MIANALYZE (Multiple Imputation Analyze), e o software R por meio do pacote MICE (Buuren; Oudshoorn, 2000).

6.5.4 Imputação múltipla com enfoque bayesiano

A inferência bayesiana visa diminuir o desconhecimento sobre o valor do parâmetro ($\boldsymbol{\theta}$) de interesse, incorporando a opinião subjacente do pesquisador antes da amostra ser coletada. Esta opinião pode assumir diferentes graus, os quais são representados por meio de modelos probabilísticos. Diferente da inferência freqüentista, tal metodologia considera θ como um escalar ou um vetor aleatório desconhecido que é quantificado por meio de uma distribuição de probabilidade conhecida como distribuição a priori (Rossi, 2011).

Existem diversas aplicações dos métodos bayesianos conforme a área em estudo. Esta abordagem tem sido utilizada como alternativa para lidar com problemas de dados em falta, fundamentada na probabilidade condicional dos valores ausentes dado os valores completos, a qual é considerada uma técnica de imputação múltipla em que as fases do processo bayesiano são incorporadas em dois passos (Passo-P e Passo-I).

No decorrer desta seção, o procedimento da imputação múltipla por meio da inferência bayesiana defini as distribuições a posteriores do vetor de médias e matriz de covariâncias de dados, provindos de uma distribuição normal multivariada.

Conceitos básicos de Inferência bayesiana

A inferência bayesiana expressa as informações sobre os parâmetros desconhecidos após observar os dados, fundamentando-se no teorema de Bayes representado por

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}} \tag{6.1}$$

em que $p(\boldsymbol{\theta})$ é a distribuição a priori de $\boldsymbol{\theta}$, $p(\boldsymbol{y}|\boldsymbol{\theta})$ a informação proveniente dos dados (função de verossimilhança) e $p(\boldsymbol{\theta}|\boldsymbol{y})$ a probabilidade condicional do parâmetro dado os valores observados (distribuição a posteriori).

Na distribuição a priori, o pesquisador inclui sua opinião em relação ao parâmetro de interesse utilizando seu conhecimento prévio. Tal informação fornece a cada parâmetro, valores igualmente prováveis, sendo conhecida como priori não informativa. Na distribuição a posteriori são realizadas atualizações dessas informações sobre o parâmetro após observar os dados. Isto acontece devido a junção da distribuição a priori com a função de verossimilhança. Na equação 6.1 temos que,

$$\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int p(\boldsymbol{y},\boldsymbol{\theta})\,d\boldsymbol{\theta} = p(\boldsymbol{y}) \tag{6.2}$$

o que nos permite reescrever a equação 6.1 da seguinte forma,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y},\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{6.3}$$

Observa-se que $1/p(\boldsymbol{y})$ não depende de $\boldsymbol{\theta}$, é considerado como uma constante normalizadora de $p(\boldsymbol{\theta}|\boldsymbol{y})$. Desta forma, este termo pode ser ignorado na expressão 6.3, pois a sua ausência não altera a forma da distribuição por ser uma constante. Assim, a distribuição a posteriori pode ser simplificada, ou seja,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \longrightarrow \textit{posteriori} \propto \textit{verossimilhança} \times \textit{priori} \tag{6.4}$$

Em casos como a seleção de modelos a constante $1/p(\boldsymbol{y})$ excluída se torna fundamental, detalhes sobre o processo bayesiano pode ser encontrado em Ehlers (2003). O processo bayesiano consiste de três fases.

1. Definir uma distribuição a priori para o parâmetro de interesse;
2. Utilizar uma função de verossimilhança para resumir evidências dos dados em relação aos valores dos diferentes parâmetros e;
3. Combinar as informações a partir da distribuição a priori juntamente com a verossimilhança gerando uma distribuição a posteriori.

Quando o interesse é sobre um determinado conjunto de $\boldsymbol{\theta}$, obtém-se a distribuição marginal de $\boldsymbol{\theta}_i$ (MARTINS FILHO et al. 2008), ou seja,

$$p(\boldsymbol{\theta}_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}_i} p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

Os cálculos das distribuições marginais, em geral, são de difícil solução. Devido a isto, existem algoritmos iterativos apropriados conhecidos como MCMC (*Markov Chain-Monte Carlo*), mas que exigem distribuições condicionais completas. Um deles é conhecido como algoritmo Metropolis-Hastings, o qual é utilizado em casos que as distribuições condicionais completas tem formas conhecidas. Quando esta condição não é satisfeita, utiliza-se um caso especial de Metropolis-Hastings conhecido como Amostrador de Gibbs (Reis et al., 2009).

Distribuição *a priori* do vetor de médias e da matriz de covariâncias

Nesta fase do processo bayesino, o pesquisador determina uma distribuição a priori para os parâmetros de interesse. Neste caso, considera-se uma matriz de dados $\mathbf{Y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \cdots, \boldsymbol{y}_k]$ com distribuição normal multivariada, em que, cada coluna representa uma variável medida sobre uma amostra de tamanho n retirada da população. Uma distribuição *a priori* para o vetor de médias pode ser não informativa, conhecida como priori de Jeffrey ou uma distribuição *a priori* de famílias conjugadas (Gelman et al., 1995). Por padrão esta priori é um plano multidimensional na superfície que atribui peso igual a cada combinação de valores médios. Entretanto, para a matriz de covariâncias recomenda-se uma priori conjugada que pertence a uma família de distribuições, sendo a mais adequada é uma priori baseada na distribuição Wishart Invertida (Enders, 2010). Detalhes podem ser obtidos em Gelman et al. (1995).

De acordo com a forma da distribuição descrita em Schafer (1997), em que $\mathbf{Y}^T = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]$ é uma matriz contendo n vetores independentes $\boldsymbol{y}_i$, com distribuição normal multivariada e vetor de médias ($\boldsymbol{\mu} = \mathbf{0}$) e matriz de covariâncias $\boldsymbol{\Sigma}$. Se $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$ é uma matriz de somas de quadrados e produtos cruzados, em que, $\mathbf{A^{-1}}$ tem distribuição Wishart. Então, $\mathbf{X} = \mathbf{A}$ tem uma distribuição Wishart invertida com função de densidade de probabilidade denotada por,

$$f(\mathbf{X}) = \frac{|\boldsymbol{\Lambda}|^{v/\mathbf{2}}|\mathbf{X}|^{-(v+p+\mathbf{1})/\mathbf{2}}}{\mathbf{2}^{vp/\mathbf{2}}\pi^{p(p-\mathbf{1})/\mathbf{4}}\prod_{i=\mathbf{1}}^{p}\boldsymbol{\Gamma}(\frac{v+\mathbf{1}-\mathbf{i}}{\mathbf{2}})}\exp\{-(\mathbf{1/2})\mathbf{tr}(\boldsymbol{\Lambda}\mathbf{X}^{\mathbf{-1}})\}$$

sendo $\mathbf{X}$ e $\boldsymbol{\Lambda}$ matrizes de dimensão, $p \times p$, positivas definidas, $\Gamma_p(.)$ uma função gama multivariada e v o grau de liberdade correspondente a distribuição Wishart invertida com $v \geq p$ e $\boldsymbol{\Lambda} > 0$.

Fazendo $v = 0$ e $\boldsymbol{\Lambda} = \mathbf{0}$ na função de densidade da distribuição Wishart

invertida, teremos uma distribuição *a priori* multivariada de Jeffrey dada por,

$$p(\Sigma) \propto |\Sigma|^{(k+1)/2}$$

em que o determinante $|\Sigma|$ é um valor escalar, que quantifica a variação total da matriz de covariâncias populacional conhecida como variância generalizada.

Função de verossimilhança

Após a escolha da priori a ser utilizada, defini-se a função de verossimilhança dos dados observados. A função de verossimilhança é definida com base na distribuição normal multivariada. Por definição, a função de máxima verossimilhança tem por finalidade, encontrar os valores dos parâmetros que maximizam a probabilidade dos dados amostrado dado o modelo estatístico assumido, a qual é expressa por

$$L(\boldsymbol{\theta}, \boldsymbol{y}_i) = \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\theta}) \tag{6.5}$$

sendo $i = 1, 2, \cdots, n$. A função de máxima verossimilhança de uma distribuição normal é dada por

$$L(\boldsymbol{\theta}, \boldsymbol{y}_i) = (\frac{1}{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}}) \exp\{-\frac{1}{2}\sum_{i=1}^{p}(\boldsymbol{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\} \tag{6.6}$$

Aplicando a função logarítmica na equação 6.6, a função pode ser reescrita da seguinte forma

$$L(\boldsymbol{\theta}, \boldsymbol{y}_i) = -\frac{n}{2}\log(2\pi)^p - \frac{n}{2}\log|\ \boldsymbol{\Sigma}\ | - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) \tag{6.7}$$

Com a presença de observações ausentes em um conjunto de dados, a função log-verossimilhança dos dados observados passa a ser maximizada considerando os diferentes padrões de dados em ausentes (Lin, 2010). Desta forma, o log da equação 6.5 pode ser reescrita da seguinte maneira

$$\ell(L(\boldsymbol{\theta}, \boldsymbol{y}_{obs})) = \sum_{g=1}^{G}\log L_g(\boldsymbol{\theta}, \boldsymbol{y}_{obs})$$

em que $\ell(L(\boldsymbol{\theta}, \boldsymbol{y}_{obs}))$ é a função log-verossimilhança do g-ésimo padrão de dados em falta. Assim, a função log-verossimilhança para uma distribuição normal

multivariada com diferente padrões de dados ausentes é expressa como

$$\ell(L(\boldsymbol{\theta}, \boldsymbol{y}_{obs})) = -\frac{n_g}{2} \log \mid \boldsymbol{\Sigma}_g \mid - \frac{1}{2} \sum_{ig}^{n_g} (\boldsymbol{y}_{ig} - \boldsymbol{\mu}_g)^T {\boldsymbol{\Sigma}_g}^{-1} (\boldsymbol{y}_{ig} - \boldsymbol{\mu}_g) \qquad (6.8)$$

em que n_g é o número de observações no g-ésimo padrão de dados ausentes, $\boldsymbol{\mu}_g$ é o vetor de médias correspondente, $\boldsymbol{\Sigma}_g$ é a matriz de covariância e $\boldsymbol{y}_{ig}$ é o vetor de valores observados correspondente para as variáveis observadas no caso i.

Distribuição *a posteriori* do vetor de médias e matriz de covariâncias

Nesta fase, o objetivo é encontrar a distribuição a posteriori do vetor de médias e da matriz de covariâncias utilizando as informações a priori com a informações dos dados coletado. Considerando, para o vetor de médias, uma distribuição *a priori* não informativa, conseqüentemente a distribuição *a posteriori* será uma distribuição normal multivariada, ou seja,

$$p(\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\Lambda}) \sim NM(\hat{\boldsymbol{\mu}}, n^{-1}\boldsymbol{\Sigma})$$

Para a matriz de covariâncias, a distribuição *a posteriori* baseada na junção da distribuição *a priori* (distribuição Wishart Invertida) e a função de máxima verossimilhança da distribuição normal multivariada terá a seguinte forma

$$p(\boldsymbol{\Sigma}|\hat{\boldsymbol{\mu}}) \sim WI(n-1, \hat{\boldsymbol{\Lambda}})$$

em que $p(\boldsymbol{\Sigma}|\hat{\boldsymbol{\mu}})$ é a distribuição *a posteriori* da matriz de covariâncias que segue uma distribuição Wishart Invertida, com $v = n - 1$ graus de liberdades e $\hat{\boldsymbol{\Lambda}}$ uma matriz de somas de quadrados e produtos cruzados.

Fases da Imputação múltipla com enfoque bayesiano

O procedimento realizado pelo método de imputação múltipla com enfoque bayesiano, utiliza um algoritmo iterativo que repete os passos I e P até obter m conjuntos de dados imputados. O passo I tem a função de estimar os valores faltantes por meio de um método estatístico adequado e, o passo P utiliza os dados imputados para gerar novas estimativas dos parâmetros de suas respectivas posteriores (Enders, 2010). Estas estimativas são geradas com o método e simulação Monte Carlos. A seuir, será apresentado uma descrição simples e intuitiva do processo de imputação múltipla com enfoque bayesiano, proporcionando uma visão geral do método.

- **Passo I**

Considerando um modelo estatístico de uma regressão linear, em que uma variável Y_i pode ser expressa em função de variáveis explicativas $(X_1, X_2, X_3, ..., X_p)$ temos que:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2} + ... + \boldsymbol{\beta}_p X_{ip} + \varepsilon_i \tag{6.9}$$

em que $i = \{1, 2, ..., n\}$, Y_i é a variável que contém a falta e pretendemos estimar, X_{ip} são as variáveis que não contém valores ausentes para as pontuações faltantes em Y_i e, ε_i é o erro aleatório que se distribui normalmente, com vetor de médias ($\boldsymbol{\mu} = \mathbf{0}$) e variância residual ($\boldsymbol{I}\sigma^2$). A parti da equação 6.9 tem-se a seguinte equação estimada,

$$\hat{Y}_i = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 X_{i1} + \hat{\boldsymbol{\beta}}_2 X_{i2} + ... + \hat{\boldsymbol{\beta}}_p X_{ip} \tag{6.10}$$

a qual prever as observações faltantes em Y_i e para recuperar a variabilidade residual perdida, devido aos valores estimados, adiciona-se um termo aleatório z_i em cada valor estimado, ou seja,

$$\hat{Y}_i = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 X_{i1} + \hat{\boldsymbol{\beta}}_2 X_{i2} + ... + \hat{\boldsymbol{\beta}}_p X_{ip} + z_i \tag{6.11}$$

sendo z_i, o erro aleatório gerado pela distribuição normal com vetor de médias zero e variância residual dos dados disponíveis. Com os dados disponíveis, por meio da equação 6.11, os valores estimados são gerados. Quando uma matriz de dados apresenta mais de uma variável com valores ausentes, conseqüentemente apresenta diferentes padrões de valores ausentes, gerando diferentes equações de regressões para cada padrão (o que foi visto na imputação simples pela regressão). Tendo os valores estimados, imputa-se os valores ausentes, formando uma matriz de dados completa no ciclo t.

- **Passo P**

Com o banco de dados completo, gerado pelo Passo I no ciclo t, estima-se novos vetores de médias ($\boldsymbol{\mu}_t$) e matriz de covariâncias ($\boldsymbol{\Sigma}_t$). Nesta fase do processo, usa-se a distribuição *a posteriori* da matriz de covariâncias e com o método de simulação Monte carlos, gera-se uma nova matriz de covariâncias ($\boldsymbol{\Sigma}_t{}^*$). Sabe-se que, a distribuição *a posteriori* da matriz de covariâncias é dada por

$$p(\boldsymbol{\Sigma}|\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}) \sim WI(n-1, \hat{\boldsymbol{\Lambda}}_{\boldsymbol{t}})$$

em que $\hat{\mathbf{\Lambda}}_t = (n-1)\hat{\mathbf{\Sigma}}_t$, em cada tempo t teremos estimativas de matriz de covariâncias diferentes. Com ${\Sigma_t}^*$ e $\boldsymbol{\mu}_t$ utiliza-se a distribuição *a posteriori* do vetor de médias dado por

$$p(\boldsymbol{\mu}|\, \boldsymbol{Y}, \boldsymbol{\Sigma}) \sim NM(\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}, n^{-1}{\boldsymbol{\Sigma}_{\boldsymbol{t}}}^*)$$

novamente com o método de simulação Monte Carlos, gera-se um novo vetor de médias (${\boldsymbol{\mu}_t}^*$). Com estes parâmetros simulados, volta-se ao passo I para gerar um novo conjunto de dados imputados. O processo continua até obter m conjuntos de dados completos.

Fase de análise e agrupamento

A parti dos vários conjuntos de dados completos, gerados pela fase de imputação, calcula-se estimativas de parâmetros e erros padrão. De modo que, nesta fase do processo de imputação, o objetivo é combinar todas as análises dos m conjuntos de dados completos, em um único conjunto de resultados.

Rubin (1987) descreveu fórmulas simples para agrupar as estimativas dos parâmetros e erros padrão dos m conjuntos de dados completos com base na média aritmética das estimativas da fase de análise.

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i \tag{6.12}$$

em que $\hat{Q}_i$ é uma estimativa do i-ésimo parâmetro considerado correspondente aos seu conjunto de dados imputado ($m = 1, 2, 3, 4, 5$), o qual é considerado como uma variável aleatória. O autor observou que a combinação dos erros padrão é mais complexa, pois envolve duas fontes de variação de amostragem:

- **A variância dentro das imputações**: Definida como média aritmética das m variâncias amostrais descrita como:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \hat{U}_i \tag{6.13}$$

sendo $\hat{U}_i$ a variância do m-ésimo conjunto de dados imputados ($m = 1, 2, 3, 4, 5$).

- **A variação entre imputações**: Que quantifica a variabilidade de uma estimativa em todas as m imputações, sendo simplesmente, a variância do parâmetro estimado em todas as m imputações.

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{Q}_i - \bar{Q}) \tag{6.14}$$

Após o cálculo das estimativas combinadas ($\bar{Q}$, $\bar{U}$ e B), o próximo passo é a obtenção da variância combinada descrita por:

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{6.15}$$

sendo $(1+\frac{1}{m})$ a correção de números infinitos de imputações, a raiz da variância fornece o valor do desvio padrão combinado (S.E). Em seguida pode-se realizar testes de hipóteses e construir intervalos de confiança para a média ($\bar{Q}$) por meio de uma aproximação t-Student, ou seja, $(\bar{Q} - Q)/S.E \sim t$ com $v_m = (m-1)[1 + \bar{U}/(1+m)^{-1}B]^2$ graus de liberdades, sendo Q a média real da variável em estudo (Medina e Galván, 2007). O autor mostra uma medida para determinar o incremento relativo da variância, devido a presença das unidades ausentes, $r = ((1 + m^{-1})B)/\bar{U}$, uma taxa de unidades ausentes que se aproxima de $\lambda = r/(1+r)$ ou $\lambda = ((r+2)/(v_m+3))/(1+r)$ e a eficiência relativa $ER = (1 + \frac{\lambda}{m})^{-1}$. Uma descrição detalhada da construção dos passos de imputação pode ser obtida em Enders (2006).

6.5.5 Exemplo-Imputação múltipla com enfoque bayesiano

O conjunto de dados refere-se à uma das três procedências estudadas por Lavoranti (2003), provenientes de estudos experimentais conduzidos nas regiões Sudeste do Brasil. Nesta aplicação foi utilizada a procedência Bellthorpe St. Forest-QLD/14.431, sendo a variável em análise a altura em metros aos cinco anos de idade. Foram consideradas, apenas, as médias de alturas para cada genótipo de *Eucalyptus grandis* nos quatros primeiros ambientes (Barra Ribeiro-RS, Telêmico Borba-PR, Boa Esperança do Sul-SP, Guanhães-MG). Desse conjunto de dados, foi realizado uma retirada aleatória de 10%, gerando um padrão arbitrário de observações em falta, as quais são apresentadas a seguir:

```
data eucal;
input amb1 amb2 amb3 amb4;
cards;
17.37 23.84 19.18 19.61
15.23   .   15.43 17.46
14.49 23.37 18.22 19.22
16.51 23.17  .    19.78
15.90 23.28 17.32   .
```

```
18.69 24.15 17.55   .
17.84   .   17.55 19.63
18.21   .   17.26 19.65
17.56 23.79 15.64 19.17
17.87 23.65 18.09 20.93
18.93 23.26 17.67 20.51
18.83 23.78 17.80 20.31
18.09   .   17.30 19.99
17.31 22.97 17.79 18.62
18.30 23.38 16.90 19.18
17.06 23.62 17.11 18.29
18.13   .   17.37 18.83
17.18 22.98 18.14 19.85
19.65 24.21 18.20 20.90
17.79 24.44 16.62 19.62
14.37 19.77 14.20 14.18
16.77 20.45 16.49 18.03
17.16 22.98 16.01 18.28
  .     .   15.36 15.10
15.12 22.62 16.39 20.18
17.20 22.84 17.13 19.13
;
```

Para ilustrar o processo de imputação com enfoque bayesiano foi utilizado o PROC MI do software SAS. Para realizar a primeira etapa da imputação múltipla (fase de imputação) no SAS, utiliza-se a sentença a seguir.

```
proc mi data=eucal MU0=17.1973 22.8404 17.1285 19.1346 seed=68619
out=outmi;
mcmc chain=multiple displayinit initial=em(itprint);
var amb1
amb2 amb3 amb4 amb5 amb6 amb7;
run;
```

Em que, a opção "MU0"indica os parâmetros das médias em cada ambiente antes das observações em falta, serem geradas. A opção "OUT"guarda os resultados dos m conjuntos de observações imputadas. Por padrão o procedimento gera cinco conjuntos de dados em falta, mas pode-se usar a opção "NIMPUTE=número de imputação desejada"para alterar este número. A opção

"CHAIN"indica que o procedimento utiliza cadeias múltiplas e completa com um burn-in de 200 iterações (por padrão) antes de cada imputação.

A opção "ITPRINT"em "INITIAL=EM"permite que o processo mostre uma tabela com o histórico de iteração, realizado pelo algoritmo EM. Estas estimativas serão utilizadas como parâmetros iniciais no algoritmo de imputação. Com a opção "DISPLAYINIT"o processo mostra, a as estimativas iniciais do vetor médias e matriz de covariâncias a serem utilizadas no processo de imputação. Com o "PROC PRINT DATA=outmi", o procedimento permite visualizar os cinco conjuntos de dados imputados.

Destes conjuntos de dados imputados, realiza-se a segunda etapa da imputação múltipla com o PROC UNIVARIATE, pelo qual calcula-se as estimativas das médias e os desvios padrão em cada conjunto de dados imputado para cada ambiente.

```
proc univariate data=outmi noprint;
var amb1 amb2 amb3 amb4;
output out=resultado mean= Mamb1 Mamb2 Mamb3 Mamb4
stderr= Samb1 Samb2 Samb3 Samb4;
by _Imputation_;
run;

proc print data=resultado;
run;
```

Na sentença "VAR", indica-se as variáveis sobre as quais pretende-se obter as estimativas dos parâmetros. Neste exemplo, as variáveis são os ambiente. Estas informações são guardadas em um arquivo nomeado de "RESULTADO".

Em seguida realiza-se a terceira etapa da imputação múltipla, a fase de agrupamento. Para obter as estimativas agrupadas pelas equações 8.13, 8.14 e 8.15, utiliza-se o PROC MIANALYZE. Novamente, os valores das médias originais de cada ambiente antes de gerar a falta, são incluídos. A opção "edf"indica os graus de liberdade a ser considerado nos testes de hipóteses. Na sentença "MODELEFFECTS"e "STDERR"indica-se as estimativas que se pretende agrupar.

```
proc mianalyze data=resultado edf=25 MU0= 17.1973 22.8404 17.1285
19.1346;
modeleffects Mamb1 Mamb2 Mamb3 Mamb4;
stderr Samb1 Samb2 Samb3 Samb4;
```

```
run;
```

Os resultados do PROC MIANALYZE é mostrado no output do SAS, como segue.

```
The SAS System          17:32 Sunday, March 3, 2013  33

                    The MIANALYZE Procedure

                       Model Information

                Data Set                  WORK.RESULTADO
                Number of Imputations     5


                      Variance Information

                -----------------Variance-----------------
Parameter         Between         Within          Total       DF

Mamb1            0.001999       0.091012       0.093411   22.534
Mamb2            0.004169       0.070073       0.075076   21.158
Mamb3            0.000426       0.046195       0.046706   22.945
Mamb4            0.000684       0.094143       0.094965   23.004

                      Variance Information

                      Relative       Fraction
                      Increase        Missing       Relative
Parameter     in Variance    Information     Efficiency

Mamb1            0.026359       0.026003       0.994826
Mamb2            0.071398       0.068705       0.986445
Mamb3            0.011060       0.010998       0.997805
Mamb4            0.008724       0.008686       0.998266


                      Parameter Estimates

Parameter        Estimate      Std Error    95% Confidence Limits        DF

Mamb1           17.120187       0.305632     16.48721     17.75316   22.534
Mamb2           23.084816       0.274001     22.51526     23.65437   21.158
Mamb3           17.093622       0.216115     16.64649     17.54075   22.945
Mamb4           19.016028       0.308163     18.37855     19.65351   23.004

                      Parameter Estimates

                                                          t for H0:
Parameter         Minimum         Maximum         Theta0   Parameter=Theta0   Pr > |t|

Mamb1           17.051056       17.152775       17.197300              -0.25     0.8031
Mamb2           23.007301       23.169572       22.840400               0.89     0.3824
Mamb3           17.074620       17.127998       17.128500              -0.16     0.8732
Mamb4           18.980686       19.046240       19.134600              -0.38     0.7039
```

No output observa-se, os resultados das estimativas agrupadas das médias ("Estimate"), das variâncias entre as imputações ("Between"), das variâncias dentro das imputações ("Within") em cada ambiente. Apresenta-se também, resultados dos teses de hipóteses e seus respectivos intervalos de confiança, como também os valores da eficiência relativa em cada ambiente. Detalhes sobre o PROC MI e MIANALYZE podem ser encontrados em SAS (2004).

Capítulo 7

Modelos AMMI: Metodologia alternativa para experimentos multiambientes bivariados

Geralmente a análise de dados de dupla entrada é feita através da análise de variância - ANOVA, mais há outros estudos nos quais a interação é de grande importância, por exemplo, no melhoramento genético, em que o objetivo é selecionar genótipos com ótimos desempenhos em diferentes ambientes. A pouca eficiência na análise da interação dos genótipos com os ambientes ($G \times E$) da ANOVA representa um problema aos melhoristas, que devem tirar proveito dessa interação para os seus estudos. Os modelos aditivos com interação multiplicativa - AMMI, trazem vantagens na seleção de genótipos quando comparados com métodos convencionais, pois proporcionam uma melhor análise da interação ($G \times E$), além de permitir combinar componentes aditivos e multiplicativos em um mesmo modelo; estes modelos são eficientes na análise quando se tem apenas uma variável resposta, mas quando há mais de uma, ainda não existe um procedimento geral para realizar a análise. O presente capítulo propõe uma metodologia de análise quando se têm modelos AMMI bivariados, realizando análises individuais das variáveis respostas seguidas de uma análise de procrustes, que permite fazer comparações dos resultados obtidos nas análises individuais e finalmente uma confirmação destes resultados através da análise multivariada de variância - MANOVA. Os resultados obtidos permitem concluir que as análises AMMI e procrustes proporcionam uma boa alternativa de análise para os modelos AMMI bivariados.

Este módulo do minicurso está baseado em García-Peña e Dias (2009).

7.1 Introdução

Em alguns estudos, é comum encontrar a análise de dois fatores cada um com diferentes níveis, eles proporcionam uma tabela de dados de dupla entrada, geralmente a análise destes dados é feita através da ANOVA, cumprindo algumas pressuposições como a que o número de repetições em cada combinação do fator deve ser maior do que 1, homogeneidade de variâncias, normalidade dos resíduos, erros independentes e identicamente distribuídos e efeitos aditivos no modelo, mas há outros estudos nos quais é importante a interação ainda que as pressuposições não sejam satisfeitas. Por exemplo, no melhoramento genético vegetal, o objetivo é selecionar genótipos com ótimos desempenhos em diferentes ambientes. A baixa eficiência na análise da interação dos genótipos com os ambientes ($G \times E$) pela ANOVA pode representar um problema aos melhoristas, que devem tirar proveito dessa interação para os seus estudos.

Segundo Lavoranti (2003) as posições críticas dos estatísticos que atuam em programas de melhoramento genético vegetal, referem-se à falta de uma análise criteriosa da estrutura da interação ($G \times E$) como um dos principais problemas para a recomendação de cultivares. Tradicionalmente, a análise dessa estrutura é superficial, não detalhando os efeitos da complexidade da interação.

Os modelos aditivos com interação multiplicativa (AMMI) são uma boa opção para a análise da interação ($G \times E$), pois permitem um detalhamento maior da soma de quadrados da interação e conseqüentemente, traz vantagens na seleção de genótipos, quando comparados a outros métodos tradicionais de análise como a ANOVA. A utilização dessa teoria parece ser uma alternativa eficiente para os programas de melhoramento, já que permite combinar em um único modelo estatístico, componentes aditivos para os efeitos principais, como genótipos e ambientes, e componentes multiplicativos para os efeitos da interação.

Quando um experimento depende de muitas variáveis, não basta conhecer informações estatísticas isoladas para cada variável, também é necessário conhecer a totalidade destas informações fornecida pelo conjunto das variáveis. As relações existentes entre as variáveis não são percebidas e assim efeitos antagônicos ou sinérgicos de efeito mútuo entre variáveis complicam a interpretação do fenômeno a partir das variáveis consideradas. A necessidade de entender as relações entre várias variáveis faz que a análise multivariada seja de grande importância. Com esta análise podem-se reduzir a dimensionalidade

dos dados ou simplificar a estrutura sem muita perda da informação contida nos dados, obtendo assim uma fácil interpretação dos resultados.

Os modelos AMMI têm sido muito eficientes na análise de dados quando se tem apenas uma variável resposta, mas quando há mais de uma ainda não existe um procedimento claro de análise, por isso, no presente trabalho propõe-se uma metodologia de análise para este caso, usando algumas técnicas de análise multivariada. Serão usados testes paramétricos para a seleção do modelo levando em conta também a validação cruzada dada a sua importância por ser uma ferramenta de seleção do modelo que não depende de alguma distribuição de probabilidade.

A análise dos modelos AMMI com mais de uma variável resposta será feita através de análises individuais e em seguida comparadas com a análise de procrustes. Esta técnica permite fazer a comparação dos resultados obtidos para cada uma das variáveis nas mesmas condições. A comparação de dois conjuntos de coeficientes de componentes principais poderia indicar se existem fontes comuns de variação ou não; no entanto, uma simples comparação dos coeficientes pode ser enganadora. Nesse sentido, a análise de procrustes evita que esta situação, em particular, aconteça.

Com as ferramentas apresentadas, pretende-se oferecer ao pesquisador novas possibilidades na análise dos modelos AMMI, no que diz respeito à análise de variáveis respostas.

7.2 Material e métodos

7.2.1 Características dos dados

Os dados utilizados são do Instituto Agronômico de Campinas, e referem-se a dados de feijão comum (Phaseolus vulgaris), compostos por 19 genótipos avaliados em 18 ambientes e com 3 repetições, o delineamento experimental foi o aleatorizado em blocos. A parcela experimental foi constituída por quatro linhas de 4 metros de comprimento, espaçadas de 0,50 metros entre si, com 10 a 12 plantas viáveis por metro linear e a área útil da parcela correspondendo às duas linhas centrais. Os dados são de produtividade de grãos e tecnológicos, relativos à qualidade alimentícia do feijão; as variáveis respostas foram produtividade (kg/ha) e tempo de cozimento (min.).

7.2.2 Modelos AMMI

O modelo AMMI usa dois métodos na sua análise: análise de variância e a decomposição singular; no modelo se unem os termos aditivos dos efeitos principais e os termos multiplicativos para os efeitos da interação. Na primeira fase a análise de variância é aplicada à matriz de médias ($Y_{(g\times e)}$) composta pelos efeitos principais na parte aditiva (média geral, efeitos genotípicos e ambientais), resultando em um resíduo de não aditividade, isto é, na interação (G×E), dada por $(\widehat{ge})_{ij}$, essa interação constitui a parte multiplicativa do modelo, na segunda fase a interação é analisada pela decomposição por valores singulares (DVS) da matriz de interações ($GE_{(g\times e)} = [(ge_{ij})]$), ou por análise de componentes principais (PCA) como prefere referir um grande número de autores (Duarte e Vencovsky, 1999).

O modelo AMMI para dois fatores (G e E) é apresentado como:

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{p} \lambda_k \gamma_{ik} \alpha_{jk} + \varepsilon_{ij}; \tag{7.1}$$

com $(ge)_{ij}$ interação genótipo-ambiente modelada por $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij}$, $i = 1, 2, \ldots, g$, $j = 1, 2, \ldots, e$, em que Y_{ij} é a resposta média do i-ésimo genótipo no j-ésimo ambiente; μ a média geral; g_i efeito do i-ésimo genótipo; e_j efeito do j-ésimo ambiente; λ_k raiz quadrada do k-ésimo autovalor das matrizes $\boldsymbol{(GE)(GE)^T}$ e $\boldsymbol{(GE)^T(GE)}$ de iguais autovalores não nulos (λ_k^2 é o k-ésimo autovalor); $[GE_{ge} = (ge_{ij})]$ matriz de interações obtida como resíduo do ajuste aos efeitos principais por ANOVA, aplicada à matriz de médias, $(k = 1, 2, \ldots, p)$, em que p é o número de raízes características não nulas $p = (1, 2, \ldots, \min\{g-1, e-1\})$; γ_{ik} i-ésimo elemento (relacionado ao genótipo i) do k-ésimo autovetor de $\boldsymbol{(GE)(GE)^T}$ associado a λ_k^2; α_{jk} j-ésimo elemento (relacionado ao ambiente j) do k-ésimo autovetor de $\boldsymbol{(GE)^T(GE)}$ associado a λ_k^2; ρ_{ij} ruídos presentes nos dados; ε_{ij} erro experimental médio, $\varepsilon_{ij} \sim N(0, \frac{\sigma^2}{m})$, em que m é o número de repetições; n número de componentes retidos no modelo ($n < p$).

Sob as restrições $\sum_{i=1}^{g} g_i = \sum_{j=1}^{e} e_j = \sum_{i=1}^{g} (ge)_{ij} = \sum_{j=1}^{e} (ge)_{ij} = 0$, as estimativas da média geral (μ) e os efeitos principais (g_i e e_j) são obtidos no contexto simples de uma ANOVA de dupla entrada aplicada à matriz de médias ($Y_{(g\times e)}$). Os resíduos deste ajuste para os efeitos principais, equivalem ao termo das interações $GE_{g\times e} = [(ge)_{ij}]$ e os termos multiplicativos da interação são esti-

mados através da decomposição singular - DVS.

7.2.3 Métodos de seleção do número de componentes de interação

Existem diferentes métodos para selecionar o número de componentes retidos no modelo, mas todos concordam em que o número deve ser o menor possível, para assim explicar a estrutura da interação. Duarte e Vencovsky (1999) apresentam o método de Gollob como um dos mais usados, o qual distribui graus de liberdade às somas de quadrados $SQ_k = m\lambda_k^2$, contando o número de parâmetros no k-ésimo termo multiplicativo. Para o k-ésimo componente de interação IPCA_k (interaction principal component analysis), $\text{GL}(\text{IPCA}_k)=g+e-1-2k$, logo o teste F é calculado como na análise da variância para modelos lineares.

Piepho (1995a) mostra o teste F_R proposto por Cornelius, mais robusto que o proposto por Gollob em 1968, a estatística é definida como

$$F_R = \frac{SQ_{G\times E} - \sum_{k=1}^{n} \lambda_k^2}{f_2 \; QM_{\text{erro médio}}},$$

em que $f_2 = (g-1-n)(e-1-n)$ com n o número de termos multiplicativos incluídos no modelo. A estatística F_R, sob a hipótese nula de que não haja mais do que n termos determinando a interação, tem uma distribuição F aproximada com f_2 e $GL_{\text{erro médio}}$ graus de liberdade. Um resultado significativo pelo teste sugere que pelo menos um termo multiplicativo ainda deve ser adicionado aos n já ajustados. Os graus de liberdade do numerador de F_R são iguais aos graus de liberdade para toda a interação menos os graus de liberdade atribuídos pelo método de Gollob aos n primeiros termos. Assim, pelo sistema de Gollob e Cornelius, a análise da variância conjunta completa tem a estrutura como mostrada na Tabela 7.1. Note-se que não aparece o efeito bloco pois a análise da variância é calculada a partir das médias.

Dias e Krzanowski (2003), apresentam o método de Eastment e Krzanowski (1982) baseado no procedimento "leave-one-out" completo que otimiza o processo de validação cruzada por validar o ajuste do modelo em cada um dos dados por vez e então combinar essa validação em uma medida simples e geral de ajuste.

Assume-se que se deseja predizer os elementos x_{ij} da matriz $\mathbf{X}$ por meio do modelo:

Tabela 7.1: Análise da variância conjunta completa calculada a partir das médias usando os sistemas de Gollob e Cornelius

Fontes de variação	GL[1] Gollob	SQ[2] Gollob	GL Cornelius	SQ Cornelius
Genótipo (G)	$g-1$	SQ_G	-	-
Ambiente (E)	$e-1$	SQ_E	-	-
Interação (GE)	$(g-1)(e-1)$	SQ_{GE}	-	-
$IPCA_1$[3]	$g+e-1-(2\times 1)$	λ_1^2	$(g-1-1)(e-1-1)$	$\sum_{k=2}^{p}\lambda_k^2$
$IPCA_2$	$g+e-1-(2\times 2)$	λ_2^2	$(g-1-2)(e-1-2)$	$\sum_{k=3}^{p}\lambda_k^2$
$IPCA_3$	$g+e-1-(2\times 3)$	λ_3^2	$(g-1-3)(e-1-3)$	$\sum_{k=4}^{p}\lambda_k^2$
...	...	...	...	...
$IPCA_p$	$g+e-1-(2\times p)$	λ_p^2	-	-
Erro médio	$ge(m-1)$	$SQ_{\text{erro médio}}$	-	-
Total	$gem-1$	SQ_{TOTAL}	-	-

[1] GL: Graus de liberdade
[2] SQ: Soma de quadrados
[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k=1,2,\ldots,p$.

$x_{ij}=\sum_{k=1}^{n} d_k u_{ik} v_{jk}+\varepsilon_{ij}$, em que d_i é a raiz quadrada dos autovalores da matriz $\boldsymbol{XX^T}$, a i-ésima coluna $v_i=(v_{i1},\ldots,v_{ip})$ da matriz $\boldsymbol{V_{p\times p}}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\boldsymbol{XX^T}$ e a j-ésima coluna $u_j=(u_{1j},\ldots,u_{nj})^T$ da matriz $\boldsymbol{U_{n\times p}}$ e o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\boldsymbol{XX^T}$, ε_{ij} é o ruído.

O método prediz o valor $\widehat{x}_{ij}^n$ de x_{ij} $(i=1,\ldots,g;\ j=1,\ldots,e)$ para cada possível escolha de n (o número de componentes), a medida de discrepância entre o valor atual e predito é

$$PRESS(n)=\sum_{i=1}^{g}\sum_{j=1}^{e}(x_{ij}^n-x_{ij})^2 \tag{7.2}$$

contudo, para evitar o viés, os dados x_{ij} não devem ser usados nos cálculos de x_{ij}^n para cada i e j. O método assume que a DVS de $\mathbf{X}$ pode ser escrita como $\mathbf{X}=\mathbf{UDV^T}$, e utiliza a seguinte estatística para determinar o número de componentes no modelo:

$$W_n=\frac{\frac{PRESS(n-1)-PRESS(n)}{D_n}}{\frac{PRESS(n)}{D_r}} \tag{7.3}$$

em que D_n é o número de graus de liberdade requeridos para ajustar o n-ésimo componente e D_r é o número de graus de liberdade remanescentes após ajustar o n-ésimo componente.

A validação cruzada subdivide $\mathbf{X}$ em certo número de grupos, deleta-se cada grupo por vez a partir dos dados, avalia-se os parâmetros do modelo ajustados a partir dos dados remanescentes, e prediz-se o valor deletado Wold (1978). Krzanowski (1987) argumenta que a predição mais precisa resulta

quando cada grupo deletado é tão pequeno quanto possível, que no presente caso é um simples elemento de $\mathbf{X}$. Denota-se por $\mathbf{X}^{(-\mathbf{i})}$ o resultado de deletar a i-ésima linha de $\mathbf{X}$ e centralizar em torno das médias das colunas. Denota-se $\mathbf{X}_{(-\mathbf{j})}$ o resultado de deletar a j-ésima coluna da $\mathbf{X}$ e centralizar em torno das médias das colunas, seguindo o esquema, pode-se escrever

$$\mathbf{X}^{(-\mathbf{i})} = \overline{\mathbf{U}}\,\overline{\mathbf{D}}\,\overline{\mathbf{V}}^{\mathbf{T}} \text{ com } \overline{\mathbf{U}} = (\overline{u}_{pt}),\ \overline{\mathbf{V}} = (\overline{v}_{pt}) \text{ e } \overline{\mathbf{D}} = diag(\overline{d}_1, \ldots, \overline{d}_l),$$
$$\mathbf{X}_{(-\mathbf{j})} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^{\mathbf{T}} \text{ com } \widetilde{\mathbf{U}} = (\widetilde{u}_{pt}),\ \widetilde{\mathbf{V}} = (\widetilde{v}_{pt}) \text{ e } \widetilde{\mathbf{D}} = diag(\widetilde{d}_1, \ldots, \widetilde{d}_{l-1}).$$

Agora, considere-se o preditor

$$\widehat{x}_{ij}^{n} = \sum_{t=1}^{n} \left(\widetilde{u}_{it}\sqrt{\widetilde{d}_t}\right)\left(\overline{v}_{tj}\sqrt{\overline{d}_t}\right) \tag{7.4}$$

cada elemento no lado direito da expressão (7.4) é obtido da DVS de $\mathbf{X}$ centrada na média após omitir a i-ésima linha e a j-ésima coluna. Assim, o valor x_{ij} não é usado no cálculo da predição, e o máximo uso dos dados é feito com os outros elementos de $\mathbf{X}$.

7.2.4 Análise de procrustes

É muito comum que sejam obtidas configurações (conjuntos de dados) no desenvolvimento de uma pesquisa e a comparação entre os resultados por meios gráficos é necessária. A possibilidade de comparação entre as configurações ou conjuntos surge porque os pontos de todas as configurações referem-se às mesmas n entidades, mas é apenas a comparação entre os pontos correspondentes de duas configurações que é de interesse. A comparação de dois conjuntos de r componentes é equivalente à comparação de dois subespaços r-dimensionais correspondentes deste espaço comum. É possível que dois conjuntos de vetores sejam diferentes um do outro, mas estejam definindo o mesmo subespaço. É necessária uma técnica analítica que proporcione uma medida numérica de quanto duas ou mais representações gráficas diferem. Têm-se duas opções: uma é assumir que os n pontos de cada um dos conjuntos de configurações referem-se às mesmas n entidades e tentar definir a quantidade numérica do desvio da coincidência dos conjuntos de pontos. A outra, é assumir que os p eixos do conjunto de configurações são os mesmos e tentar definir a quantidade numérica do desvio da coincidência dos conjuntos de subespaços definidos em relação a estes eixos (Krzanowski, 2000).

Sejam $\mathbf{X}$ e $\mathbf{Y}$ duas matrizes com dimensão $(n \times p)$ e $(n \times q)$ que representam as coordenadas dos n pontos em cada uma das configurações, supondo que $p > q$, a segunda configuração está em um subespaço do espaço p-dimensional e pode ser tratado como se estivesse neste último espaço adicionando $(p - q)$ colunas de zeros ao lado direito de $\mathbf{Y}$, convertendo-a em uma matriz $(n \times p)$. Sem perda de generalidade, pode-se assumir que $p = q$, e que esta condição é atingida adicionando um número apropriado de colunas de zeros na menor matriz dentre as duas.

O objetivo é comparar as duas configurações, assumindo que as linhas das duas matrizes referem-se à mesma entidade. Com a pressuposição de $p = q$, podem-se desenhar as duas configurações com os mesmos p eixos ortogonais. A estatística de procrustes é composta por três passos, o primeiro é a translação que é um deslocamento de todos os pontos através de uma distância constante no mesmo sentido, ou seja, uma translação fixa de toda a configuração; o segundo é a rotação que consiste em um deslocamento fixo de todos os pontos através de um ângulo constante, mantendo a distância de cada ponto ao centróide, significa, uma rotação fixa de toda a configuração; e finalmente dilatação que é o estiramento ou encolhimento de todos os pontos através de uma constante em uma linha reta do ponto ao centróide da configuração, isto é, dilatação uniforme de toda a configuração. Levando em conta isto, a estatística de procrustes pode ser calculada como

$$M^2 = c^2 \text{ traço}(\mathbf{Y}\mathbf{Y}^\mathbf{T}) - 2c \text{ traço}(\mathbf{X}\mathbf{Q}^\mathbf{T}\mathbf{Y}^\mathbf{T}) + \text{ traço}(\mathbf{X}\mathbf{X}^\mathbf{T}) \qquad (7.5)$$

em que $c = \frac{\text{traço}(\mathbf{X}\mathbf{Q}^\mathbf{T}\mathbf{Y}^\mathbf{T})}{\text{traço}(\mathbf{Y}\mathbf{Y}^\mathbf{T})}$ é o parâmetro de dilatação e $\mathbf{Q} = \mathbf{V}\mathbf{U}^\mathbf{T}$ é a matriz de rotação, com é a decomposição por valores singulares da matriz $\mathbf{X}^\mathbf{T}\mathbf{Y}$. O procedimento adequado é manter uma configuração fixa e igualar a outra a esta. Segundo Krzanowski (2000) deve-se fixar a configuração cujas coordenadas estão dadas por $\mathbf{X}$ e igualar a configuração com coordenadas $\mathbf{Y}$ a esta. A correspondência será obtida através da realização dos passos descritos anteriormente, na seqüência, de forma tal a tornar o valor final de M^2 para a correspondência das configurações tão pequena quanto possível. Se o procedimento é feito sobre a mesma matriz, quer dizer, $\mathbf{X} = \mathbf{Y}$, a estatística $M^2 = 0$, o que indica que as duas configurações se ajustam exatamente, então quanto menor o valor da estatística maior similaridade entre as configurações.

7.2.5 Análise multivariada da variância - MANOVA

Geralmente quando se tem mais de uma variável medida por parcela nos delineamentos de experimentos é feita a Análise de Variância Multivariada - MANOVA. A Análise de Variância - ANOVA tem uma generalização quando as variáveis são vetores levando a análise de uma matriz de somas de quadrados e produtos cruzados. Considerando o caso de dupla entrada, é suposto que se tem mge observações independentes geradas pelo modelo

$$\mathbf{Y_{iju}} = \mu + \mathbf{g_i} + \mathbf{e_j} + (\mathbf{ge})_{\mathbf{ij}} + \varepsilon_{\mathbf{iju}}, \text{ com } i = 1, 2, \ldots, g,\ j = 1, 2, \ldots, \text{ e } u = 1, 2, \ldots, m$$

em que $\mathbf{g_i}$ é o efeito da i-ésima linha, $\mathbf{e_j}$ é o efeito da j-ésima coluna, $(\mathbf{ge})_{\mathbf{ij}}$ o efeito da interação entre a i-ésima linha e a j-ésima coluna, $\varepsilon_{\mathbf{iju}}$ é o termo do erro que é assumido independente $N_p(\mathbf{0}, \mathbf{\Sigma})$ para todo i, j, u e m é o número de repetições. É necessário que o número de observações em cada casela (i, j) deva ser o mesmo, tal que a soma de quadrados total e matriz de produtos pode ser decomposta. O interesse está em testar a hipótese nula da igualdade de $\mathbf{g_i}$, igualdade de $\mathbf{e_j}$ e a igualdade de $(\mathbf{ge})_{\mathbf{ij}}$, Mardia et al. (2003).

Os métodos descritos foram implementados no módulo IML (Interactive matrix programming) do pacote SAS (Statistical Analysis System), o programa utilizado está disponível em García (2009).

7.3 Resultados e discussão

Para cada uma das variáveis respostas foi feita a análise de variância dando como resultado o ajuste dos efeitos principais por ANOVA (primeira etapa da análise AMMI). A correlação entre as variáveis respostas é $-0,5052$ (valor-$p < 0,0001$), foi obtida com os dados originais, o que indica uma correlação moderada inversa entre elas, enquanto a produtividade é alta o tempo de cozimento é baixo. Para la variável produtividade os genótipos, assim como os ambientes, são estatisticamente significativos com valores $F = 4,59$ e $F = 173,67$ (valores-$p < 0,0001$), da mesma maneira, os efeitos genotípicos e ambientais para a variável tempo de cozimento apresentam os valores $F = 2,73$ e $F = 28,86$ (valores-$p < 0,0001$), também é de grande interesse a soma de quadrados $G \times E = 24242129,90$ e $G \times E = 7181,03$, objeto da decomposição DVS, na segunda etapa da análise, para a variável resposta produtividade esta interação representou 9% da soma de quadrados total e para o tempo de cozimento 36%.

As estimativas das médias de genótipos e ambientes, assim como a média geral, ajustadas pelo modelo sem interação, são mostradas na Tabela 7.2. Nela pode-se observar que os genótipos com maior produtividade média são G11, G10, G15 e G13 e com menor os G8, G7, G1 e G9, enquanto ao tempo de cozimento os genótipos com maior tempo de cozimento em média são G15, G1, G10 e G8 e com menor G7, G5, G19 e G13, já para os ambientes as maiores produtividades em média foram obtidas nos ambientes A5, A13, A12 e A1 e com menor os A17, A18, A16 e A6, para o tempo de cozimento em média os maiores tempos pertencem aos ambientes A17, A18, A9 e A7 e os menores os A12, A1, A11 e A8. É interessante notar que existem diferenças entre as duas variáveis respostas, por exemplo, o genótipo G13 que para a produtividade é um dos que possui um valor alto, para o tempo de cozimento é um dos menores, o mesmo acontece com os ambientes, aqueles com maior produtividade apresentam menor tempo de cozimento, este é o caso de A1 e A12, isto devido à correlação entre as duas variáveis repostas.

Tabela 7.2: Médias de produtividade e tempo de cozimento para Genótipos e Ambientes

	Média ($\overline{Y_{i.}}$)			Média ($\overline{Y_{.j}}$)	
Genótipo	Produtividade	Tempo C.[1]	Ambientes	Produtividade	Tempo C.
G1	2148,8802	32,4707	A1	3272,7193	21,4323
G2	2396,8422	31,0046	A2	2317,0614	30,1028
G3	2289,8026	28,1580	A3	1881,3596	31,4130
G4	2337,5195	28,3856	A4	2018,4211	31,3944
G5	2277,3450	27,1902	A5	4076,2676	25,1932
G6	2324,1702	29,5344	A6	1681,3597	29,6974
G7	2051,3359	26,2743	A7	1965,4386	35,2854
G8	2001,3272	31,2035	A8	2017,7874	23,0312
G9	2171,3037	29,1393	A9	2301,7983	35,5588
G10	2486,6924	32,2815	A10	2407,9437	28,4053
G11	2557,7411	29,7409	A11	2879,1491	22,2446
G12	2282,2393	29,3757	A12	3345,8333	20,1737
G13	2445,2254	27,6054	A13	3460,6065	29,3437
G14	2298,2628	28,8691	A14	2419,9561	23,2409
G15	2474,3609	33,1915	A15	2202,4035	33,3089
G16	2270,0485	28,1839	A16	1244,0877	31,0360
G17	2270,0319	29,4756	A17	828,2456	40,5781
G18	2371,4182	29,9026	A18	1165,2807	38,6161
G19	2335,9345	27,5165			
			Média Geral ($\overline{Y_{..}}$)	2304,7622	29,4475

[1] Tempo de Cozimento

Na segunda etapa da análise AMMI a interação $G \times E$ é o objeto da decomposição DVS. Segundo a regra de Gollob e o teste F para a variável produtividade, 5 dos 17 eixos de interação são significativos ($F = 2,2206$ e valor-$p = 0,0005$), o que levaria à seleção do modelo $AMMI5$, enquanto pelo teste F_R de Cornelius et al. (1996), o modelo selecionado é o $AMMI4$, já que somente a partir de $IPCA_4$ o resíduo AMMI torna-se não significativo ($F = 1,2074$ com valor-$p = 0,0508$). Para a variável tempo de cozimento o

método de Gollob e o teste F indicam que 7 dos 17 eixos de interação são significativos, o que indica a seleção do modelo $AMMI7$, entretanto, o teste F_R de Cornelius, seleciona o modelo $AMMI6$, pois a partir de $IPCA_6$ o resíduo AMMI torna-se não significativo ($F = 1,0817$ com valor-$p = 0,2690$). O método de Eastment e Krzanowski, através da estatística W, não selecionou componente para reter no modelo, isto significa que o modelo seria aditivo para a produtividade e o tempo de cozimento; mas diante da maior simplicidade representativa do modelo e das propriedades do teste F_R, sugere-se, o modelo $AMMI4$ como o melhor descritor do padrão de resposta diferencial dos genótipos aos ambientes para a variável produtividade e o modelo $AMMI6$ para o tempo de cozimento.

A última etapa da análise AMMI consiste na representação gráfica dos genótipos e ambientes no chamado *biplot*, Gabriel (2002). Para isso, faz-se necessária a determinação de suas coordenadas para os eixos singulares de interação. A partir das mesmas matrizes $\boldsymbol{U, S, V}$, resultantes da decomposição por valor singular - DVS da matriz GE, obtém-se novamente os resultados de interesse. De acordo com Gauch (1988), o primeiro eixo singular da análise AMMI captura a maior porcentagem de "padrão" e com acumulação subseqüente das dimensões dos eixos, há uma diminuição na porcentagem de "padrão" e um incremento de "ruídos" Com isso, apesar da porção pequena de $SQ_{G\times E}$ explicada pelos dois primeiros eixos (40,1%) para produtividade e (40,7%) para tempo de cozimento, espera-se capturar a maior parte do "padrão" devido a interação $G \times E$. Dessa forma, os escores de genótipos e de ambientes são plotados só até o segundo eixo, para cada uma das variáveis respostas. Vários autores utilizaram o mesmo tipo de representação apesar de terem verificado uma menor proporção da $SQ_{G\times E}$ explicada, a saber: 60% (Eyherabide et al., 1997); 57,6% (Flores et al., 1996); 54,6% (Crossa et al., 1990); 44,6% (Pereira e Costa, 1998); 28,6% (Arias, 1996); 27,1% (Crossa et al., 1991).

Para a variável produtividade as Figuras 7.1 e 7.2 ilustram as duas representações (os biplots $AMMI1$ e $AMMI2$, respectivamente) resultantes dos conjuntos de coordenadas, para o biplot $AMMI1$ (médias vs. $IPCA_1$) e para biplot $AMMI2$ ($IPCA_1$ vs $IPCA_2$). A partir deles são feitas, então, as devidas interpretações, procurando identificar genótipos e ambientes que menos contribuíram para a interação $G \times E$; combinações de genótipos e ambientes desejáveis em termos de adaptabilidade; relações entre os eixos de interação e características genotípicas e ambientais conhecidas.

Pela Figura 7.1, pode-se concluir que os genótipos mais estáveis e que menos contribuíram para a interação $G \times E$, captada pelo primeiro eixo ($IPCA_1$), foram G4, G7 G12, G13, G17 e G19; enquanto entre os ambientes nesse sentido são A3, A4, A7, A9, A12, A14, A15 e A18. Além de ser estáveis os genótipos indicam ser amplamente adaptados aos ambientes no teste. Os genótipos estáveis devem apresentar também um bom desempenho, neste caso uma alta produtividade, o que é avaliado através de suas médias. Assim, entre os genótipos estáveis destaca-se o G13 com uma produtividade média elevada e entre os ambientes o A12.

Examinando a Figura 7.2, os genótipos e ambientes mais estáveis são G9, G12, G13 e G14; A4, A9, A15 e A18. Estes genótipos tiveram uma má classificação em produtividade média (G9 foi décimo nono, G12 foi décimo segundo e G14 foi décimo), só o G13 parece ter uma boa produtividade, ele foi o quarto. Para este conjunto, as produtividades médias elevadas parecem estar associadas a adaptações específicas. Observa-se que o primeiro eixo singular deve ser determinado por características contrastantes entre os ambientes A6 - A8 e o par A11 - A13, o segundo eixo parece resultante principalmente das diferenças entre A2, A8 e A11. Já para os genótipos, o primeiro eixo parece estar relacionado a aspectos determinantes da divergência entre os genótipos G11-G15 e G16, enquanto o segundo à divergência entre G1 e G16. Também observa-se a adaptabilidade dos genótipos nos ambientes, por exemplo, G16 com o A8, os G11 e G15 com os A13 e A11, os G1 e G6 com o A2. Compete ao melhorista identificar tais características para assim discernir melhor os mecanismos determinantes da interação.

Para a variável tempo de cozimento se apresentam os biplot $AMMI1$ (médias vs. IPCA1) e o biplot AMMI2 ($IPCA_1$ vs $IPCA_2$). Da Figura 7.3, pode-se dizer que os genótipos mais estáveis e que menos contribuíram para a interação $G \times E$, captada pelo primeiro eixo ($IPCA_1$), estão situados na faixa horizontal em torno de zero em relação ao eixo $IPCA_1$, estes foram G2, G5, G9, G12, G13 e G14; enquanto entre os ambientes nesse sentido estão A1, A2, A3, A8, A9, A11, A12, A13 e A14. Além de ser estáveis os genótipos indicam ser amplamente adaptados aos ambientes no teste. É importante que os genótipos estáveis apresentem também um bom desempenho, neste caso um baixo tempo de cozimento, isto é avaliado através de suas médias. Assim, entre os genótipos estáveis destaca-se o G5 com um tempo de cozimento médio baixo e nos ambientes o A12, observa-se que o G13 ainda sendo um pouco maior em média que o G5 também proporciona um bom tempo de cozimento.

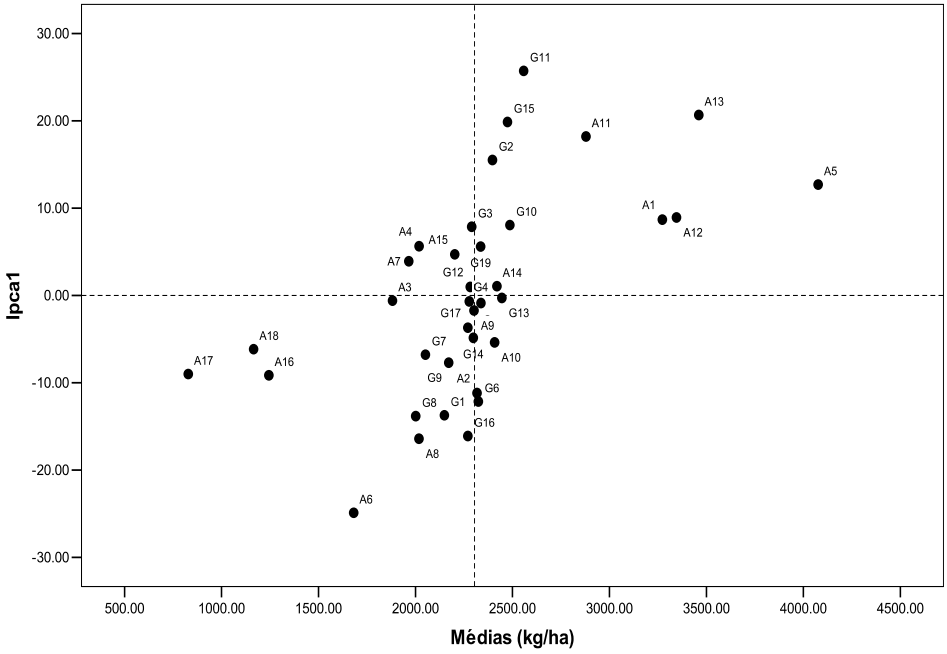

Figura 7.1: Biplot AMMI1 para dados de produtividade de grãos (kg/ha), em feijoeiro, com dezenove genótipos (G) e dezoito ambientes (A). A figura captura 23,6% de variabilidade

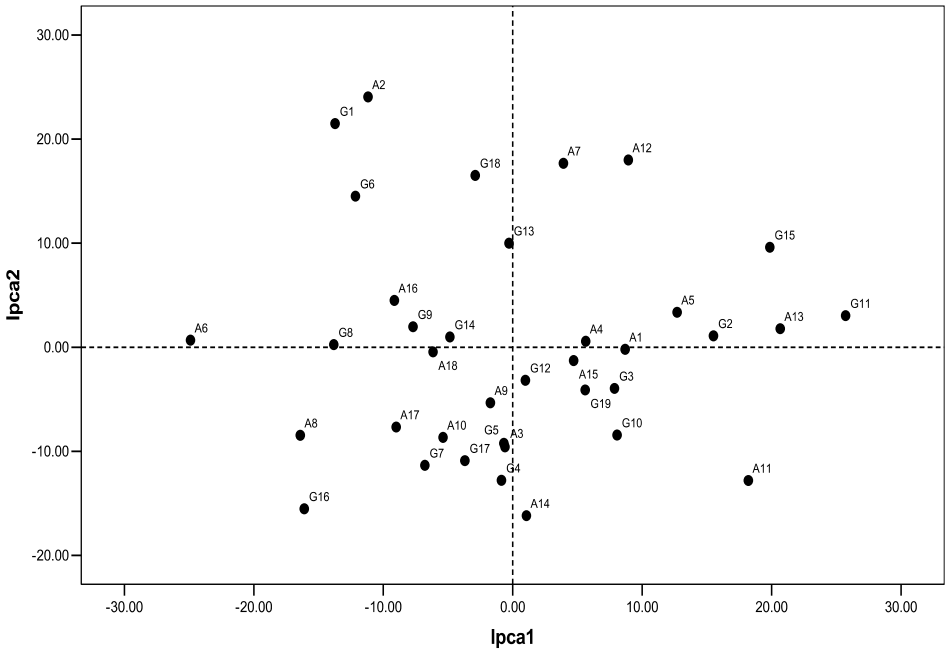

Figura 7.2: Biplot AMMI2 para dados de produtividade de grãos (kg/ha), em feijoeiro, com dezenove genótipos (G) e dezoito ambientes (A). A figura captura 40,1% de variabilidade

Observando a Figura 7.4, os genótipos e ambientes mais estáveis são G2, G5, G10 e G14; A1, A2, A7, A8, A11 e A14. Estes genótipos tiveram a

seguinte classificação em tempo de cozimento médio, G2 foi décimo quinto, G5 foi décimo oitavo, G10 foi terceiro e G14 foi décimo segundo, G5 parece ter um bom tempo de cozimento. Observa-se que o primeiro eixo singular pode estar determinado por características contrastantes entre os ambientes A6 - A9 - A10 e A18, o segundo eixo parece resultante principalmente das diferenças entre A4 e A6 - A8. Já para os genótipos, o primeiro eixo parece estar relacionado a aspectos determinantes da divergência entre os genótipos G11-G13-G18 e G8-G15, enquanto o segundo à divergência entre G3 - G15 e G12 - G17. Também observa-se a adaptabilidade dos genótipos nos ambientes, por exemplo, G8 e G15 com o A16, o G18 com A19, o G13 com o A10. Compete ao melhorista identificar tais características para assim discernir melhor os mecanismos determinantes da interação. Com os resultados encontrados nas duas análises, pode-se recomendar o genótipo G13 e o ambiente A12 como estáveis, com uma alta produtividade e um baixo tempo de cozimento.

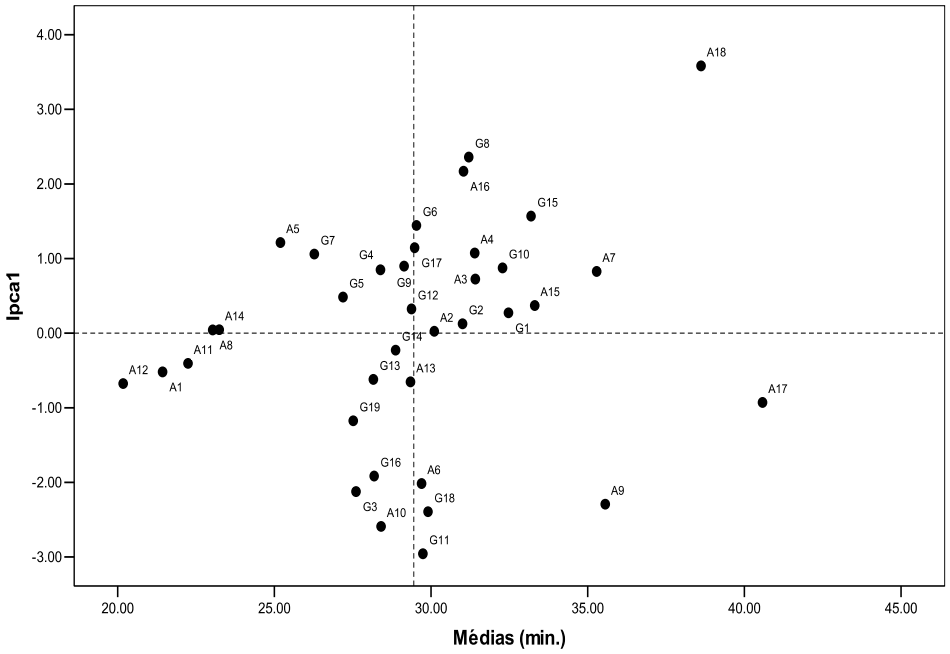

Figura 7.3: Biplot AMMI1 para dados de tempo de cozimento (min.), em feijoeiro, com dezenove genótipos (G) e dezoito ambientes (A). A figura captura 21,9% de variabilidade

A estatística de procrustes M^2 definida na equação (7.5) quantifica a diferença de duas configurações de pontos, neste caso os marcadores, quanto menor o valor da estatística, as configurações serão mais similares. Foi usada esta estatística para comparar os resultados após realizar as análises AMMI individuais para cada uma das variáveis respostas. A estatística M^2 foi calcu-

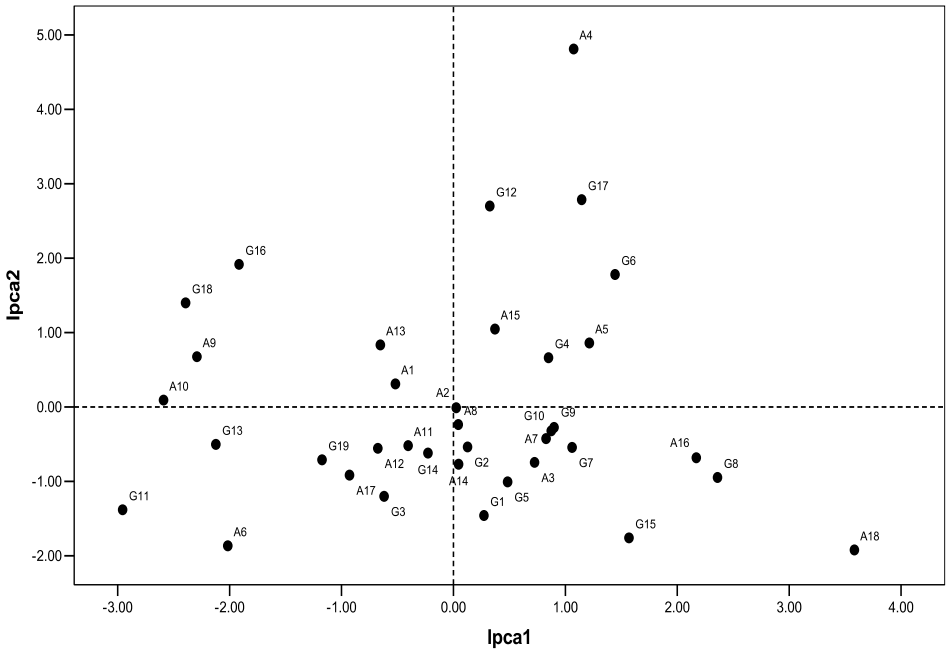

Figura 7.4: Biplot AMMI1 para dados de tempo de cozimento (min.), em feijoeiro, com dezenove genótipos (G) e dezoito ambientes (A). A figura captura 40,7% de variabilidade

lada fazendo a rotação da matriz de marcadores da variável produtividade em relação à matriz de marcadores da variável tempo de cozimento, as matrizes inicialmente foram padronizadas, gerando assim o fator de dilatação $c = 1$. Neste trabalho procura-se avaliar a similaridade entre as análises individuais e fazer uma recomendação geral de genótipos e ambientes.

Depois de realizar as duas análises individuais, decidiu-se escolher um modelo que explique conjuntamente segundo as duas variáveis respostas. Para isto compararam-se as matrizes de marcadores levando em conta cada um dos modelos escolhidos, por exemplo, para a variável produtividade o método de Cornelius selecionou um $AMMI4$ e para o tempo de cozimento um $AMMI6$, então foram aplicados os dois modelos nas duas variáveis e obtidas as respectivas matrizes de marcadores para genótipos e ambientes. Com estas matrizes de marcadores encontraram-se todas as combinações possíveis entre eles. Foram a seguir comparadas através da análise de procrustes, quanto menor o valor da estatística de procrustes indica um melhor ajuste considerando duas variáveis respostas.

Na Tabela 7.3, observa-se os valores da estatística M^2 para os marcadores de genótipos segundo os modelos escolhidos pelo método de Cornelius, ou seja, $AMMI4$ para a variável produtividade e $AMMI6$ para o tempo de cozimento

e as combinações entre eles, o menor valor da estatística foi encontrado quando as duas variáveis respostas foram modeladas com o modelo $AMMI4$, indicando que as matrizes de marcadores de genótipos são mais similares do que com a combinação dos outros modelos.

Tabela 7.3: Análise de procrustes (M^2) para os marcadores de genótipos

	Modelo para Tempo de Cozimento	
Modelo para Produtividade	AMMI4	AMMI6
AMMI4	70,8369	92,0335
AMMI6	95,0152	96,3995

Na Tabela 7.4, encontram-se os valores da estatística M^2 para os marcadores de ambientes segundo os modelos escolhidos pelo método de Cornelius, do mesmo jeito que na Tabela 7.3, neste caso o menor valor da estatística foi encontrado quando as duas variáveis respostas foram modeladas com o modelo $AMMI4$, indicando mais uma vez que as matrizes de marcadores de ambientes são mais similares do que com a combinação dos outros modelos.

Tabela 7.4: Análise de procrustes (M^2) para os marcadores de ambientes

	Modelo para Tempo de Cozimento	
Modelo para Produtividade	AMMI4	AMMI6
AMMI4	62,7695	75,7061
AMMI6	91,6764	90,3906

Portanto, segundo a estatística de procrustes o modelo que gera maior similaridade entre as matrizes de marcadores tanto para genótipos quanto para ambientes é o modelo $AMMI4$, então, poderia-se recomendar o genótipo G13 e o ambiente A12 para futuros estudos, pois apresentam um bom desempenho nas duas variáveis respostas.

Para o conjunto original de dados de 19 genótipos, 18 ambientes, 3 repetições e duas variáveis respostas, foi feita a análise de variância multivariada - MANOVA, foram calculados os testes Lambda de Wilks, o traço de Pillai, o traço de Hotelling-Lawley e a maior raiz de Roy para cada um dos efeitos (genótipos, ambientes e interação). Os resultados para os quatro testes multivariados, nos quais a hipótese nula é $H_{0Gen} = Gen_1 = Gen_2 = \cdots = Gen_g$, em que $g = 19$, ou seja, que não existe diferença entre os níveis do fator genótipo, por meio dos testes pode-se rejeitar a hipótese nula ($F = 17,07$, $F = 16,13$, $F = 18,03$ e $F = 28,51$ respectivamente com valores-$p < 0,0001$), isto é, o

genótipo tem um efeito significativo considerando as duas variáveis respostas produtividade e tempo de cozimento, simultaneamente. Igualmente para testar a significância dos ambientes, a hipótese nula é $H_{0Amb} = Amb_1 = Amb_2 = \cdots = Amb_e$, em que $e = 18$, isto é, que não há diferença entre os níveis do fator ambiente, a hipótese nula é rejeitada ($F = 229,21$, $F = 167,25$, $F = 309,70$ e $F = 534,77$ com valores-$p < 0,0001$), ou seja, o efeito que tem o ambiente sobre as variáveis respostas produtividade e tempo de cozimento é significativo. Finalmente, testa-se a hipótese nula $H_{0AmbGen} = GE_{11} = GE_{12} = \ldots = GE_{ge}$ de não efeito de interação, é rejeitada ($F = 4,85$, $F = 3,82$, $F = 6,06$ e $F = 10,41$ com valores-p$< 0,0001$), indicando que existe um efeito significativo da interação entre os genótipos e os ambientes sobre as duas variáveis respostas.

Dado que a análise MANOVA não proporciona uma boa interpretação sobre a interação, isto é, só permite saber se ela é ou não significativa, decidiu-se fazer uma análise de comparações post hoc para identificar quais das interações $(ge)_{ij}$ são significativamente diferentes das outras, nas variáveis respostas, produtividade e tempo de cozimento, foram realizadas comparações múltiplas de Scheffe. Ao realizar as comparações encontrou-se que as interações para a variável produtividade não diferem significativamente e na variável tempo de cozimento somente diferem as interações G10A7 e a G12A1.

Também foi testado o efeito dos genótipos nos níveis dos ambientes, sendo significativos para os ambientes A1, A2, A3, A5, A6, A7, A8, A11, A12, A13 e A14, na variável resposta produtividade, indicando certa responsabilidade pela significância da interação e para a variável tempo de cozimento em todos os ambientes. Entretanto, o efeito dos ambientes nos níveis dos genótipos é significativo em todos os genótipos tanto para a variável produtividade como para o tempo de cozimento, coincidindo com o resultado das comparações post hoc realizadas.

Da análise AMMI seguida de procrustes, encontrou-se que o G13 e o A12, poderiam ser uma boa opção para futuros estudos, pois proporcionam estabilidade e um bom desempenho nas duas variáveis respostas, agora com os resultados obtidos da análise MANOVA, o A12 resultou ser um dos possíveis responsáveis pela interação, dado que o efeito dos ambientes nos níveis dos genótipos foi significativo para todos os níveis, poderia-se dizer que o G13 em combinação com o A12, proporcionam uma boa escolha, confirmando os resultados da análise AMMI.

7.4 Conclusões

A análise de procrustes encontrou um modelo que apresenta alta similaridade entre as matrizes de marcadores para genótipos e ambientes nas duas variáveis respostas, sendo um modelo com poucos componentes, o que indica que este modelo esta captando o padrão dos dados para as duas variáveis e deixando fora o ruído.

Se encontrou que a análise AMMI seguida de procrustes pode ser uma alternativa para a análise de experimentos com duas variáveis respostas. Os resultados encontrados através da metodologia proposta foram confirmados usando a análise multivariada da variância e comparações múltiplas.

7.5 Sugestões

Apesar de que a análise MANOVA confirmou os resultados da análise AMMI, recomenda-se realizar uma análise mais exaustiva da interação na MANOVA, tentando encontrar as interações responsáveis da significância da interação geral, por exemplo, usar a análise canônica de variáveis na MANOVA chamada CVA para o efeito da interação, como é apresentado em Lejeune e Caliński (2000).

Dado que as análises AMMI e procrustes apresentaram bom resultados conjuntamente, recomenda-se explorar a combinação das duas análises quando se têm mais de duas variáveis respostas, em que a análise procrustes já não compara matrizes se não subespaços.

Capítulo 8

Modelo AMMI no estudo da interação entre QTL e ambiente

A interação entre genótipo e ambiente (IGA) e entre quantitative trait loci (QTL) e ambiente (IQA) são fenômenos comuns em ensaios multi-localização e representam um grande desafio para melhoradores de plantas que pretendem desenvolver genótipos mais adaptados a diferentes condições atmosféricas.

O modelo de efeitos principais aditivos e interação multiplicativa (AMMI) é uma ferramenta largamente utilizada na análise de ensaios multi-localização (e.g. rendimento) quando os dados são apresentados na forma de matriz de dupla entrada, com genótipos nas linhas e ambientes (combinação local/ano) nas colunas.

Neste módulo do minicurso o modelo AMMI será usado na detecção de QTL e no estudo da interação entre QTL e ambiente. Nomeadamente, descreveremos a *AQ analysis* (i.e. obtenção dos QTL scans tendo por base os valores preditos pelo ajustamento do modelo AMMI aos dados fenotípicos) que permite obter picos mais elevados (em termos de LOD scores) para os QTL (Gauch et al., 2011). Além disso a ordenação dos ambientes por parâmetros AMMI que sumarizem a IGA revela padrões consistentes e tendências sistemáticas que muitas vezes têm interpretação ecológica ou biológica (Gauch et al., 2011). Os métodos propostos serão ilustrados com um conjunto de dados sobre scores de germinação pré-colheita em trigo (*Triticum aestivum* L.) em 14 ambientes.

8.1 Introdução

Muitos fenótipos (características) avaliados em experiências agrícolas são quantitativos (e.g. rendimento por hectare, número de sementes produzidas por planta, etc.). A variação nessas características quantitativas é usualmente

devida ao efeito de várias localizações genéticas e a factores ambientais. O conhecimento sobre o número, localização, efeito e identidade de tais localizações genéticas (denominadas de *quantitative trait loci*, QTL) poderão conduzir a novas descobertas biológicas. Esta informação sobre QTL pode ser usada para ajudar na seleção e melhoramento de colheitas agrícolas (Broman e Sen, 2009). Assim, QTL podem ser definidos como localizações genéticas que contribuem para a variação de uma característica quantitativa. O mapeamento de QTL é uma técnica para tentar identificar QTL numa população experimental resultante do cruzamento de dois progenitores.

Um dos maiores desafios em estatística genética é encontrar melhores genótipos ao longo de uma ampla variação de condições agroecológicas e também ao longo dos anos. Este é também um desafio para agricultores, melhoradores de plantas e geneticistas, se bem que os agricultores e os melhoradores de plantas têm interesses diferentes: os agricultores pretendem genótipos melhores para o seu clima e tipo de solo e os melhoradores de plantas pretendem desenvolver genótipos que tenham uma boa performance em localizações diferentes e heterogêneas. Para alcançar este objetivo, são conduzidos ensaios multi-localização (*multi-environment trials*, MET) em que uma série de genótipos é avaliada em diferentes condições ambientais e em diferentes anos. Os dados provenientes destes MET é usualmente sumarizada numa tabela de dupla entrada com genótipos nas linhas e ambientes (combinações de localização e ano) nas colunas. Na maior parte destas tabelas de dupla entrada é possível encontrar diferenças na estabilidade dos fenótipos (e.g. rendimento) ao longo dos ambientes, i.e. os efeitos genotipicos e ambientais não são simplesmente aditivos e a interação entre genótipo e ambiente (IGA) está presente nos dados. IGA é definido como a alteração do ranking genético dos genótipos para diferentes ambientes, e.g., um genótipo com boa performance em condições humidas pode apresentar uma má performance em condições de seca. A IGA pode ser expressa como *crossovers*, quando dois genótipos diferentes apresentam alteração no ranking de performance quando avaliados em ambientes diferentes; ou respostas inconsistentes de alguns genótipos ao longo dos ambientes sem mudança em termos de ranking. O estudo e entendimento destas interações representam um grande desafio, com o objetivo de melhorar características complexas (e.g. rendimento) para diferentes gradientes ambientais.

Com o desenvolvimento de marcadores moleculares e técnicas de mapeamento, os pesquisadores podem ir mais além e analisar todo o genoma para detectar localizações específicas para os genes que influenciam a característica

quantitativa. Estas localizações são os QTL e, quando estes QTL apresentam uma expressão diferente ao longo dos ambientes, estamos perante interação entre QTL e ambiente (IQA), que representa a base da IGA. Um bom entendimento destas interações permite aos pesquisadores selecionar melhores genótipos para diferentes gradientes ambientais e, consequentemente, melhorar colheitas e as suas produções em países desenvolvidos e em desenvolvimento, com base no seu clima e características dos solos.

O modelo de efeitos principais aditivos e interação multiplicativa (AMMI) é uma ferramenta largamente utilizada na análise de ensaios multi-localização que permite particionar a interação em $N = min(I-1, J-1$ termos, em que I é o número de genótipos e J o número de ambientes. O modelo AMMI combina a análise de variância (ANOVA) e a decomposição em valores singulares (DVS), sendo que a ANOVA é aplicada primeiramente e permite extrair os efeitos principais aditivos, e a DVS é aplicada aos resíduos da ANOVA (i.e. matriz da IGA) de forma a decompor a interação em N partes (Gauch, 1988, 1992).

Neste módulo do minicurso o modelo AMMI será usado na detecção de QTL e no estudo da interação entre QTL e ambiente. Primeiramente será aplicado um modelo AMMI parcimônio aos dados fenotípicos, e seguidamente os valores preditos com esse modelo AMMI serão usados para obter os QTL scans, o que permite incluir informação de outros ambientes no scan de um determinado ambiente (Jiang e Zeng, 1995). Esta técnica é denominada de *AQ analysis* (Gauch et al., 2011) e permite obter picos mais elevados (em termos de LOD scores) para os QTL quando efetuado um QTL scan. Além disso a ordenação dos ambientes por parâmetros AMMI que sumarizem a IGA revela padrões consistentes e tendências sistemáticas, que muitas vezes têm interpretação ecológica ou biológica (Gauch et al., 2011). Os métodos propostos serão ilustrados com um conjunto de dados sobre scores de germinação pré-colheita em trigo (*Triticum aestivum* L.) em 14 ambientes.

8.2 Materiais e métodos

8.2.1 Dados genotípicos e fenotípicos

O conjunto de dados usado na ilustração desta estratégia para mais facilmente detectar QTL e estudar IQA, é proveniente de um ensaio multi-localização em que a característica fenotípica de interesse são os scores de germinação pré-colheita (*preharvest sprouting*, PHS) em trigo (*Triticum aestivum* L.), cujos

métodos experimentais estão descrito em Munkvold et al. (2009). Nos scores da PHS, o zero corresponde a nenhuma evidência de germinação na espiga enquanto o nove corresponde a a uma germinação extrema na espiga. Esta população foi derivada de um cruzamento entre a variedade resistente a PHS Cayuga e a variedade suscetível a PHS Caledonia. Foram considerados 197 genótipos avaliados em 14 ambientes, e considerados 205 marcadores mapeados em 42 grupos de ligação (*linkage groups*).

8.2.2 Análise estatística

O ajustamento do modelo AMMI aos dados pode ser efetuado através do software MATMODEL (Gauch, 2007). Considere-se uma tabela de dupla entrada de observações fenotípicas (e.g. rendimento ou scores de germinação pré-colheita) com I genótipos nas linhas e J ambientes (combinações de localização e ano) nas colunas, com replicações.

O modelo AMMI (Gauch, 1988, 1992) combina a ANOVA e a DVS, sendo a ANOVA aplicada primeiramente para obter os efeitos principais aditivos e a DVS aplicada à matriz de GEI (i.e. a matriz que contém os resíduos da ANOVA). Considerando o número máximo de componentes principais $N = min(I-1, J-1)$, o modelo AMMI pode ser escrito como

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sum_{n=1}^{N} \lambda_n \gamma_{in} \delta_{jn} + \theta_{ij}, \tag{8.1}$$

em que Y_{ij} representa os dados fenotípicos (e.g. rendimento) para o genótipo i no ambiente j, μ a média global, α_i a diferença entre a média global e o genótipo i, β_j a diferença entre a média global e o ambiente j, λ_n é o valor singular para a componente principal n, γ_{in} e δ_{jn} são os vectores singulares esquerdos e direitos, e θ_{ij} o resíduo para o genótipo i no ambiente j.

Os *QTL scans* podem ser obtidos usando o software QTL Cartographer 2.5 (Wang et al., 2007) através do método *composite interval mapping*. Os QTL significativos, com um nível de significância de 0.05, são determinados com um teste de permutações com 1000 permutações (Churchill e Doerge, 1994).

8.3 Resultados e discussão

Os dados fenotípicos obtidos de ensaios multi-localização podem variar consoante dois aspetos importantes: (i) nivel de ruido; e (ii) complexidade da IGA e, consequentemente, da IQA.

O painel da esquerda da Figura 8.1 apresenta os QTL scans para os 11 ambientes em estudo, obtidos com base nos dados originais e ordenados pelo nome do local e ano, o que em termos biológicos representa uma ordem aleatória (Gauch et al., 2011).

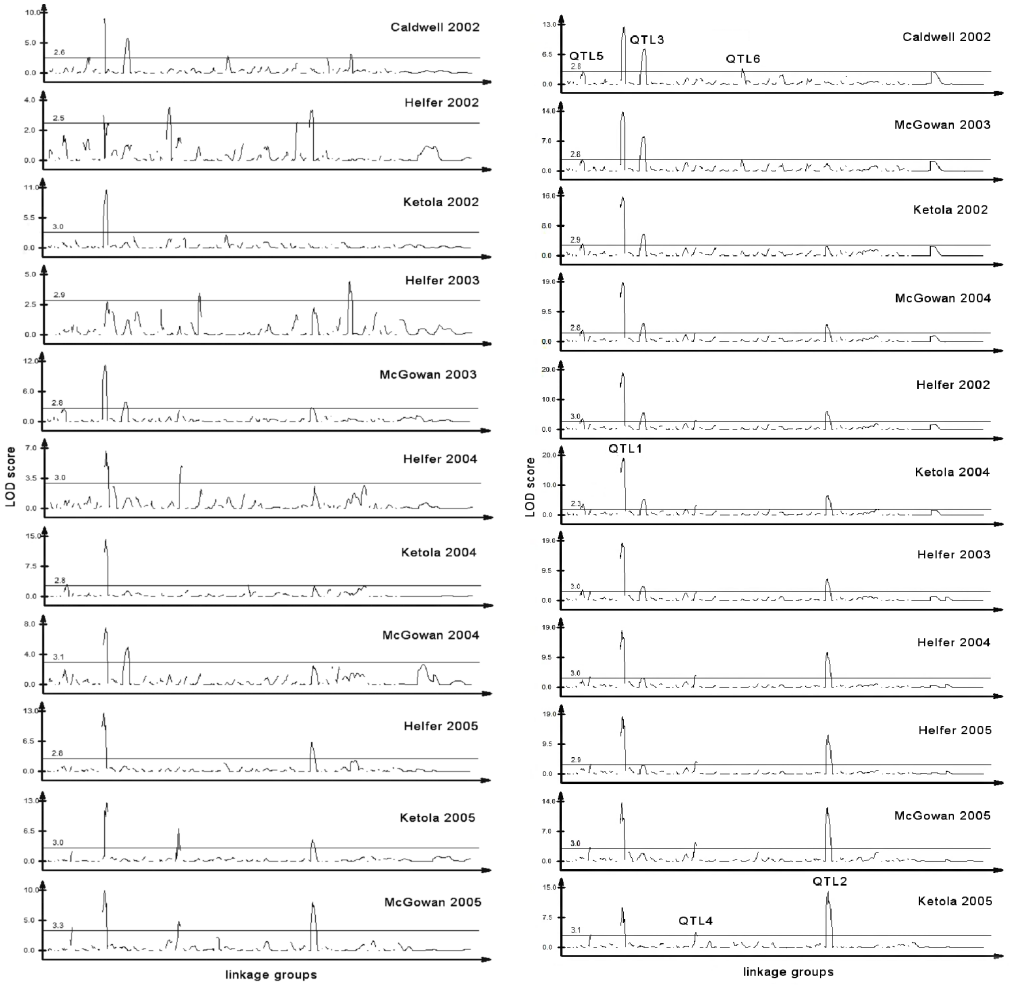

Figura 8.1: QTL scans para os 11 ambientes da população Cayuga x Caledonia: (i) com base nos dados originais e ordenados pelo nome do local e ano (esquerda); e (ii) com base nos valores ajustados pelo modelo AMMI1 e ordenados pelos scores ambientais da IPC1 (Gauch et al., 2011).

A Tabela 8.1 apresenta a ANOVA para o modelo AMMI3, isto é, o modelo AMMI com três componentes principais para a interação (*interaction principal*

components, IPC). Note-se que genótipos, ambientes e IGA representam 35.0%, 31.6%, e 33.4% da soma dos quadrados (SQ) dos tratamentos. O ruído na IGA pode ser estimado pelo produto entre os graus de liberdade (gl) da interação e o quadrado médio (QM) do erro, nomeadamente 2195, que por diferença do total de 2661 implica um sinal da IGA de 466, ou 17.5% Gauch (1992).

Tabela 8.1: ANOVA para o modelo AMMI3 (Gauch et al., 2011). A média geral é 4.097.

Fonte de variação	gl	SQ	QM	Probabilidade
Total	4306	10370.35	2.408	
Tratamentos	2166	7974.12	3.682	0.0000000
Genótipos	196	2789.94	14.234	0.0000000
Ambientes	10	2523.39	252.339	0.0000000
IGA	1960	2660.79	1.358	0.0011753
IPC1	205	577.31	2.816	0.0000000
IPC2	203	366.51	1.805	0.0000017
IPC3	201	321.77	1.601	0.0003115
Resíduo	1351	1395.20	1.033	0.8994796
Erro	2140	2396.23	1.120	

gl: graus de liberdade; SQ: soma dos quadrados; QM: quadrados médios.

A Figura 8.2 apresenta os QTL scans para os efeitos principais (médias dos 11 ambientes) e scores das IPC1, IPC2 e IPC3, obtidas através do ajustamento do modelo AMMI3 (Gauch et al., 2011). Este tipo de gráficos foi introduzido por Romagosa et al. (1996). O QTL1 apresenta um efeito principal e uma interação na IPC2 e, similarmente, o QTL2 apresenta um efeito principal e uma interação na IPC1. Os QTL3, QTL4 e QTL5 apenas apresentam efeitos principais e o QTL6 apenas apresenta uma interação na IPC1.

Com base na ANOVA apresentada na Tabela 8.1 e noutros procedimentos usuais para uma melhor escolha de um modelo AMMI mais parcimônio, optou-se pelo AMMI1, ou seja, um modelo AMMI com apenas uma componente principal. Assim, com base no modelo AMMI1 obtiveram-se os valores preditos para cada combinação genótipo ambiente. Estes valores preditos são então usados na obtenção dos QTL scans, como descrito anteriormente (isto é, a *AQ analysis*). Esses QTL scans, ordenados pelos scores da IPC1, são apresentados no painel da direita da Figura 8.1.

Ao comparar os QTL scans (painéis da esquerda e da direita) na Figura 8.1 é possível verificar diferenças significativas. Por um lado, os LOD scores obtidos através da *AQ analysis* são mais elevados e, consequentemente os QTL são detectados mais facilmente. Por outro lado, ao ordenar os QTL scans obtidos através da *AQ analysis* usando os scores da IPC1, é visível uma clara tendência

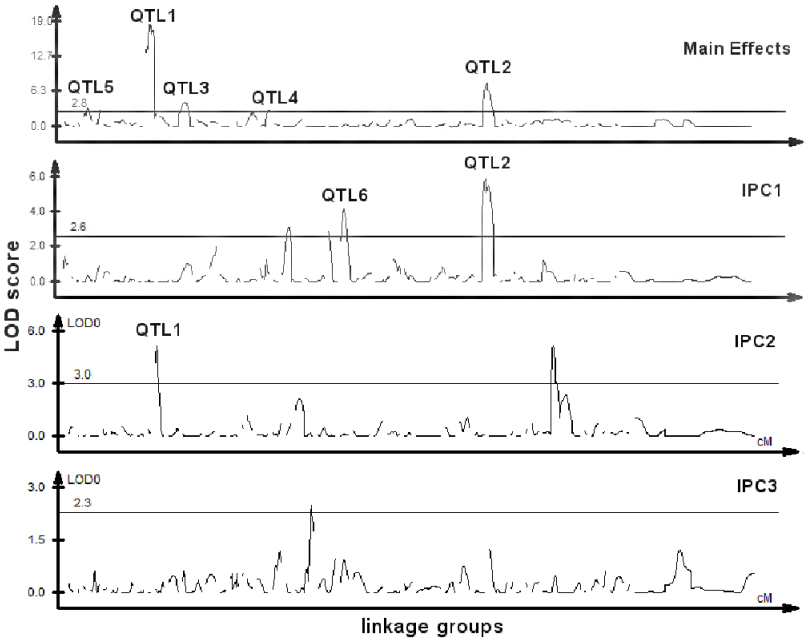

Figura 8.2: QTL scans para os efeitos principais, IPC1, IPC2 e IPC3 (Gauch et al., 2011).

e similaridade entre ambientes vizinhos.

A Figura 8.3 apresenta um resumo geral dos resultados da *AQ analysis*. A abscissa apresenta os scores da IPC1 com a ordenação dos 11 ambientes da esquerda para a direita, que é a mesma que a ordenação dos QTL scans na Figura 8.1 (painel da direita) de cima para baixo. A ordenada apresenta os LOD scores. Os QTL2 e QTL4 aumento para a direita, enquanto os QTL3, QTL5 e QTL6 aumentam para a esquerda. Uma regressão linear ajusta-se bem a esses QTL, mas não ao QTL1, que apresenta uma resposta quadrática, com um pico no meio. Pode-se assim verificar que todos os seis QTL apresentam interação com o ambiente.

8.3.1 Predição de QTL scans

A elevada relação entre os scores da IPC1 e os QTL, evidente nas Figuras 8.1 (painel da direita) e 8.3, levanta a questão sobre a possibilidade de prever QTL scans tendo por base apenas os IPC1 scores. Para analisar essa possibilidade foram utilizados dados de mais três ambientes, observados durante o ano de 2006, resultando num total de 14 ambientes em análise. Quando se adiciona (ou remove) dados, é sempre possível que os parâmetros do modelo AMMI sofram alterações radicais, apesar disso ser menos provável quando os ambientes adicionados (ou removidos) são similares aos restantes. Neste caso, os

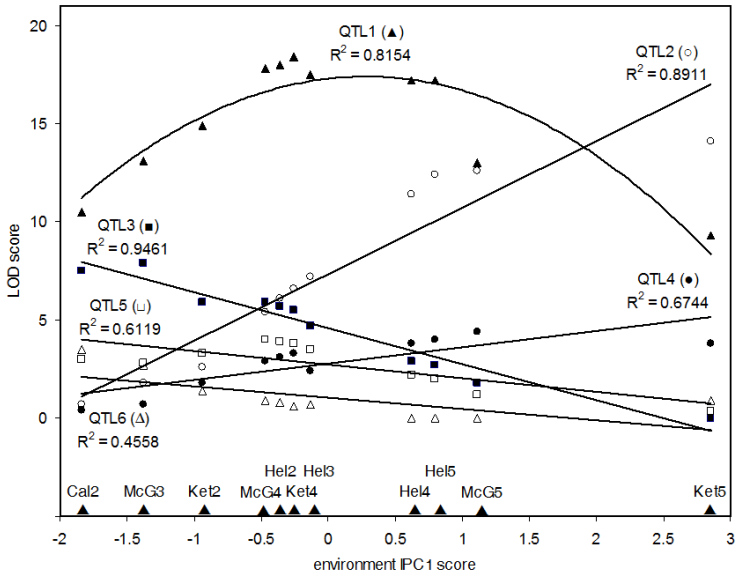

Figura 8.3: LOD scores dos seis QTL detectados, ao longo dos scores da IPC1 (Gauch et al., 2011).

parâmetros do modelo AMMI para 11 e 14 ambientes são bastante similares, tal como confirmado pela correlação de 0.9943 entre os IPC1 scores para os 11 ambientes em comum, ao considerar o conjunto de dados formado por 11 e 14 ambientes separadamente. Assim, a previsão para o QTL scan de cada um dos três novos ambientes pode ser obtida simplesmente considerando o scan do ambiente antigo com o IPC1 score mais próximo do novo. Na Figura 8.4 são apresentados os QTL scans para cada um dos novos ambientes, juntamente com o QTL scan dos ambientes antigos com o IPC1 score mais próximo do do novo ambiente. Nos três casos, os QTL scans preditos estão praticamente sobrepostos aos correspondentes QTL scans observados. De referir que, desta forma, cada QTL scan é predito com apenas um valor, o IPC1 score obtido apenas com base nos dados fenotípicos (Gauch et al., 2011).

8.4 Conclusões

A utilização de modelos AMMI permite melhorar a detecção de QTL, assim como alcançar um melhor entendimento da interação entre QTL e ambiente. Este objetivo é alcançado através da *AQ analysis*, isto é, quando os QTL scans são realizados nos valores preditos do modelo AMMI mais parcimônio, em vez de os obter com base nos dados originais. O entendimento de como os QTL interagem com o ambiente permite aos melhoradores de plantas obter

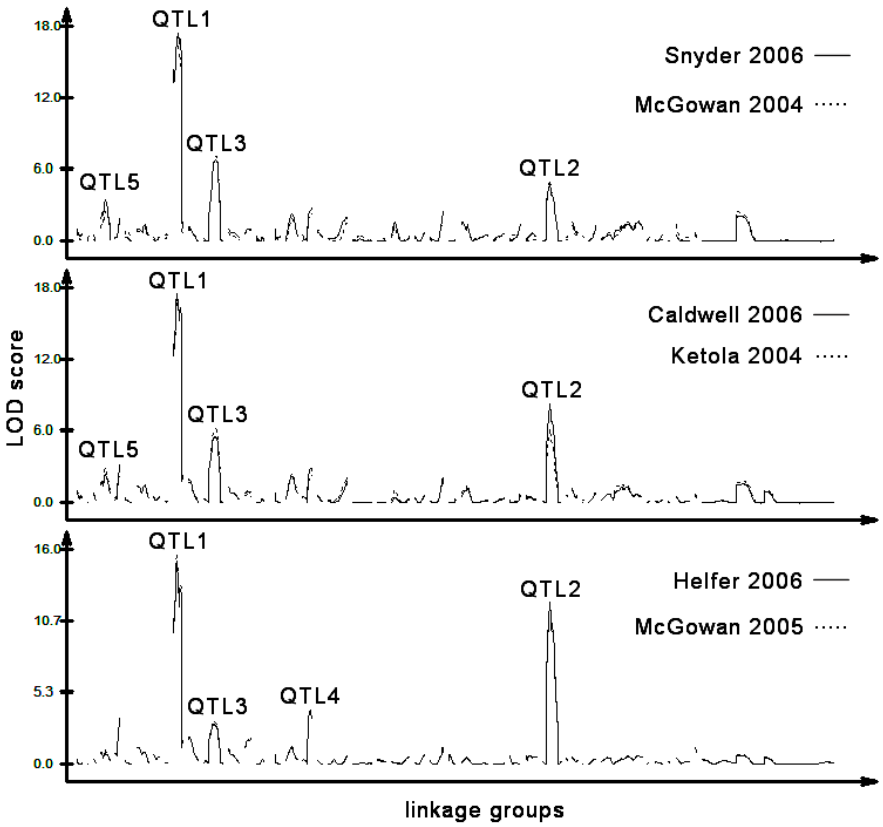

Figura 8.4: QTL scans para os três novos ambientes de 2006, em comparação com os ambientes antigos com IPC1 score mais próximo do IPC1 score do novo ambiente.

informação para melhorar a colheita em diferentes ambientes (Figura 8.3). Esta análise também permite aos melhoradores de plantas selecionar localizações que otimizem a detecção de um determinado QTL, tal como a localização Hel para o QTL1 na Figura 8.3. Essa localização maximiza consistentemente o LOD score para o QTL1 ao longo dos anos.

Apesar da ilustração apresentada ser direcionada para melhoramento vegetal, as estratégias apresentadas para detectar e entender IQA são baseadas em princípios estatísticos de igual aplicabilidade em populações microbiais e vegetais quando estudadas em vários ambientes, e podem ser adaptadas a estudos genéticos em animais e humanos (Gauch et al., 2011).

Capítulo 9

Generalização dos modelos AMMI

9.1 Introdução

Os experimentos multi-ambientais (MET) são conduzidos através de vários anos para os principais produtos agrícolas no mundo, constituindo um passo caro mas essencial para a liberação de um novo genótipo de um produto agrícola e, conseqüentemente, a recomendação de cultivar.

O objetivo primário de um MET é identificar cultivares superiores. A prática mais comum usada para este fim é comparar o rendimento de um genótipo em vários ambientes de teste (normalmente combinações de locais e anos). O segundo objetivo de análise de dados multi-ambientais, deveria ser investigar as relações entre os ambientes de teste e a possibilidade de diferenciação do mega-ambiente (YAN; HUNT, 2002).

Para a descrição da resposta média de genótipos em ambientes e para o estudo e interpretação da interação genótipos $\times$ ambientes (GE) em METs agrícolas, duas classes de modelos são comumente utilizadas: modelos lineares e modelos lineares-bilineares. Em princípio, as abordagens para a análise da interação GE incluem a apresentação dos dados em tabela de duas entradas (matriz), sendo que cada casela desta tabela contém a resposta média de cada genótipo em cada ambiente.

Considere agora o caso em que os METs são avaliados através de vários anos (ou seja, genótipos $\times$ locais $\times$ anos) (GLA), em que os dados podem ser organizados em arranjo de três entradas onde, neste caso, as entradas se referem a genótipos, locais ou anos.

Em alguns casos o investigador pode estar interessado em saber se existe uma estrutura comum encoberta pelos locais com relação aos anos e como os vários genótipos respondem através da estrutura formada por ambientes e anos. Alguns genótipos podem ter altas respostas em alguns locais, mas não em

outros e, alguns locais podem estar mais associados com alguns genótipos do que a outros por alguns anos. Um procedimento para ganhar uma compreensão clara em arranjo GLA de três-entradas é determinar uma estrutura dimensional menor, expressado em componentes principais, para a interação genótipos × locais × anos e, então, estudar as relações entre estes componentes. Esta aproximação é mais útil que combinar dois dos três fatores de maneira que os dados formem um arranjo de duas entradas. Outro procedimento menos útil é excluir um fator diretamente (por exemplo anos) e analisar um arranjo de duas entradas dos genótipos × locais em cada ano e, neste caso, o problema está em encontrar uma interpretação global para os anos.

Para os dados organizados em arranjo de três-entradas existem alguns modelos para analisá-los, como por exemplo, os modelos propostos por Tucker: Tucker1, Tucker2 e Tucker3 e o modelo proposto por Harshman que é denominado de modelo PARAFAC, que fornecem uma decomposição trilinear dos dados organizados no arranjo.

9.2 Modelos PARAFAC

O modelo PARAFAC é também conhecido como decomposição trilinear (SANCHEZ e KOWALSKI, 1990). As notações mais utilizadas são aquelas com somatórios e componentes simultâneos e as menos utilizadas são aquelas que usam produtos de Kronecker, produto tensorial (produto de Hadamard) e produtos de Khatri-Rao.

O modelo PARAFAC é introduzido através da generalização da decomposição em valor singular. Um modelo de duas entradas para a matriz $\boldsymbol{X}$ $(I \times J)$, com elementos x_{ij}, baseado na sua decomposição em valor singular $(\boldsymbol{X} = \boldsymbol{AGB}')$ truncada em R componentes é:

$$x_{ij} = \sum_{r=1}^{R} a_{ir} g_{rr} b_{jr} + e_{ij}; \qquad i = 1, ..., I \text{ e } ; \; j = 1, ..., J \tag{9.1}$$

em que:

a_{ir}: é o elemento que está na i-ésima linha e na r-ésima coluna da matriz de autovetores $\boldsymbol{A}$;

g_{rr}: é o elemento que está na r-ésima linha e na r-ésima coluna da matriz de autovalores $\boldsymbol{G}$;

b_{jr}: é o elemento que está na j-ésima linha e na r-ésima coluna da matriz de autovetores $\boldsymbol{B}$;

e_{ij}: é o elemento que está na i-ésima linha e na j-ésima coluna matriz residual, que contém a variação não explicada pelo modelo com R componentes.

Suponha um arranjo de três entradas $\underline{\mathbf{X}}$ de dimensões $(I \times J \times K)$, com elementos x_{ijk}, a expressão generalizada para um modelo PARAFAC é:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{9.2}$$

em que:

R: é o número de componentes usados no modelo PARAFAC;

a_{ir}: é o elemento que está na i-ésima linha e na r-ésima coluna da matriz de componentes $\boldsymbol{A}$;

b_{jr}: é o elemento que está na j-ésima linha e na r-ésima coluna da matriz de componentes $\boldsymbol{B}$;

c_{kr}: é o elemento que está na k-ésima linha e na r-ésima coluna da matriz de componentes $\boldsymbol{C}$;

e_{ijk}: é o elemento que está na i-ésima linha, na j-ésima coluna e no k-ésimo tubo do arranjo residual, que contém a variação não explicada pelo modelo com R componentes.

Uma descrição gráfica deste modelo é apresentada na Figura 9.1. O modelo representado pela equação (9.2) é um modelo trilinear: fixando dois parâmetros (por exemplo, a e b), x_{ijk} é expresso como um função linear dos parâmetros remanescentes (por exemplo, c).

Os parâmetros em $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ podem ser estimados com diferentes algoritmos. Os fatores são estimados simultaneamente, ao contrário da análise de componentes principais, em que os componentes podem ser estimados um de cada vez. Isto ocorre porque os componentes no modelo PARAFAC são não ortogonais e, portanto, dependem um do outro. Estimando os componentes do modelo PARAFAC seqüencialmente, como na análise de componentes principais (ACP), fornece resultados diferentes quando comparados com a estimativa de componentes simultâneos, e a aproximação seqüencial não fornece uma solução de mínimos quadrados.

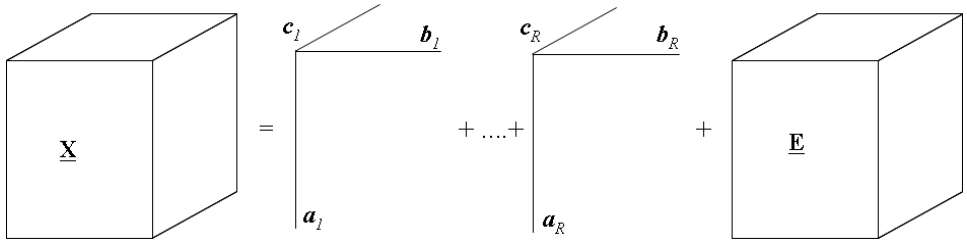

Figura 9.1: O modelo PARAFAC com R components

9.3 Modelos TUCKER

Uma possível generalização do modelo de componentes principais para dados de duas entradas é usar uma matriz núcleo não diagonal $\tilde{\boldsymbol{G}}$. Para isto considere a decomposição em valores singulares de uma matriz $\boldsymbol{X}$ $(I \times J)$ e as matrizes $\boldsymbol{T}_A$ e $\boldsymbol{T}_B$ quaisquer ortonormais:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{G}\boldsymbol{B}^{'} + \boldsymbol{E} \\ \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{T}_A\boldsymbol{T}_A^{'}\boldsymbol{G}\boldsymbol{T}_B\,\boldsymbol{T}_B^{'}\boldsymbol{B}^{'} + \boldsymbol{E} \\ \boldsymbol{X} &= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{B}}^{'} + \boldsymbol{E} \end{aligned} \tag{9.3}$$

sendo $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{T}_A$, $\tilde{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{T}_B$, $\tilde{\boldsymbol{G}} = \boldsymbol{T}_A^{'}\boldsymbol{G}\boldsymbol{T}_B$ e que o modelo (9.3) pode ser escrito de outra maneira

$$x_{ij} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\tilde{a}_{ip}\tilde{g}_{pq}\tilde{b}_{jq} + e_{ij} \tag{9.4}$$

em que "$\sim$"em cima de $\boldsymbol{G}$, $\boldsymbol{A}$ e $\boldsymbol{B}$ é usado para indicar a diferença entre as matrizes núcleo convencionais, e $\tilde{a}_{ip}$, $\tilde{g}_{pq}$ e $\tilde{b}_{jq}$ são elementos das matrizes $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{G}}$ e $\tilde{\boldsymbol{B}}$, respectivamente. Diferente da decomposição em valores singulares, o modelo (9.3) não tem a exigência de que $\tilde{\boldsymbol{A}}$ e $\tilde{\boldsymbol{B}}$ tenha o mesmo número de componentes, permitindo que p e q assuma valores até P e Q, respectivamente, e $\tilde{\boldsymbol{G}}$ seja de dimensão $(P \times Q)$, fazendo com que o número de componentes seja diferentes nos dois modos. A matriz núcleo $\tilde{\boldsymbol{G}}$ não diagonal significa explicitamente que no modelo existe interações entre os fatores. Esta é uma propriedade importante dos modelos de Tucker em geral. Na ACP tradicional, vetores de cargas interagem aos pares. Por exemplo, o segundo vetor de escores interage com o segundo vetor de cargas pela magnitude definida pelo segundo valor singular. No modelo (9.4) todos vetores podem interagir. Por exemplo, o primeiro vetor de escores interage com o terceiro vetor de carga, com uma magnitude definida pelo elemento $\tilde{g}_{13}$.

O modelo (9.4) pode ser generalizado para um arranjo de três entradas $\underline{\mathbf{X}}$, com elementos x_{ijk}

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk} \tag{9.5}$$

sendo que e_{ijk} é um elemento do arranjo $\underline{\mathbf{E}}(I \times J \times K)$; a_{ip}, b_{jq} e c_{kr} são elementos típicos das matrizes de cargas $\boldsymbol{A}(I \times P)$, $\boldsymbol{B}(J \times Q)$ e $\boldsymbol{C}(K \times R)$; e g_{pqr} é um elemento típico do arranjo núcleo $\underline{\mathbf{G}}(P \times Q \times R)$. Este é o modelo Tucker3 de $\underline{\mathbf{X}}$ (P, Q, R), em que a notação (P, Q, R) é usada para indicar que o modelo tem P, Q, R fatores em três entradas diferentes. A representação gráfica do modelo Tucker3 é dado na Figura 9.2.

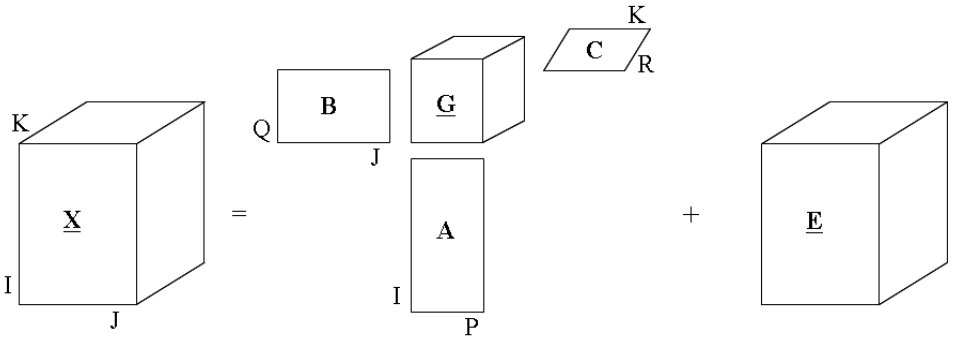

Figura 9.2: Representação gráfica do modelo Tucker3

9.4 Modelos AMMI para interação tripla

9.4.1 Análise de variância conjunta

Com o objetivo de verificar se existe a interação entre genótipos, locais e anos, realiza-se uma análise de variância conjunta que envolve o estudo de todos os genótipos em todos os locais e todos os anos, sendo que em cada local tem-se um delineamento aleatorizado em blocos. Com o efeito dos genótipos fixo, o efeito de locais aleatório e o efeito de anos também aleatório, obtendo os efeitos das interações duplas (genótipos × locais, genótipos × anos e locais × anos) e triplas (genótipos × locais × anos) como aleatórias. Os dados serão representados pelo seguinte modelo matemático:

$$Y_{ijkr} = \mu + g_i + l_j + a_k + b_r(l_j(a_k)) + (gl)_{ij} + (ga)_{ik} + (la)_{jk} + (gla)_{ijk} + \varepsilon_{ijrk} \tag{9.6}$$

sendo que:

μ: é uma constante comum a todos os efeitos, normalmente a média geral;

g_i: é o efeito do i-ésimo genótipo, com $i = 1, 2, ..., g$;

l_j: é o efeito do j-ésimo local, com $j = 1, 2, ..., l$;

a_k: é o efeito do k-ésimo ano, com $k = 1, 2, ..., a$;

$b_r(l_j(a_k))$: é o efeito do r-ésimo bloco dentro do j-ésimo local dentro do k-ésimo ano, com $r = 1, 2, ..., b$;

$(gl)_{ij}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo local;

$(ga)_{ik}$: é o efeito da interação do i-ésimo genótipo com o k-ésimo ano;

$(la)_{jk}$: é o efeito da interação do j-ésimo local com o k-ésimo ano;

$(gla)_{ijk}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo local com o k-ésimo ano;

ε_{ijrk}: é o erro experimental associado ao i-ésimo genótipo, no j-ésimo ambiente, no k-ésimo ano e no r-ésimo bloco assumido ser independente e $\varepsilon_{ijrk} \sim N(0, \sigma^2)$.

Na Tabela 9.1 apresenta-se o esquema da análise de variância para o modelo (9.6), com os graus de liberdade (GL) e esperanças dos quadrados médios ($E[QM]$).

Tabela 9.1: Esquema da análise de variância para experimentos de um mesmo grupo de genótipos avaliados em l locais e a anos com b blocos

Fontes de Variação	Graus de liberdade	E[QM]
B d. L d. A	$la(b-1)$	$\sigma^2 + g\sigma^2_{bloco}$
Genótipos (G)	$(g-1)$	$\sigma^2 + bl\phi_g + ba\sigma^2_{GL} + bl\sigma^2_{GA} + b\sigma^2_{GLA}$
Locais (L)	$(l-1)$	$\sigma^2 + bga\sigma^2_L + bg\sigma^2_{LA}$
Anos (A)	$(a-1)$	$\sigma^2 + bgl\sigma^2_A + bg\sigma^2_{LA}$
Interação ($G \times L$)	$(g-1)(l-1)$	$\sigma^2 + ba\sigma^2_{GL} + b\sigma^2_{GLA}$
Interação ($G \times A$)	$(g-1)(a-1)$	$\sigma^2 + bl\sigma^2_{GA} + b\sigma^2_{GLA}$
Interação ($L \times A$)	$(l-1)(a-1)$	$\sigma^2 + bg\sigma^2_{LA}$
Interação ($G \times L \times A$)	$(g-1)(l-1)(a-1)$	$\sigma^2 + b\sigma^2_{GLA}$
Resíduo	$la(g-1)(b-1)$	σ^2
Total	$(glab-1)$	

E[QM]: Esperanças dos Quadrados Médios; B d. L d. A: Blocos dentro de locais dentro de anos; $\phi_g = \frac{\sum_{i=1}^{g} g_i^2}{g-1}$

9.4.2 Generalização da Análise AMMI para o caso de três fatores usando o modelo PARAFAC

Sendo a interação tripla significativa, o próximo passo é fazer a decomposição da $SQ_{G\times L\times A}$, para descartar um resíduo adicional presente nessa soma de quadrados. Essa decomposição é feita utilizando o modelo PARAFAC.

Antes de aplicar a decomposição, é necessário organizar os dados em um arranjo cúbico de dimensões $g \times l \times a$ com as médias dos r blocos para cada combinação de genótipos, locais e anos (na equação (9.7) o arranjo cúbico é apresentado na forma matrizada):

$$\underline{\mathbf{Y}}_{\,g\times l\times a} = \left(\begin{array}{ccc:ccc:c:ccc} Y_{111} & \ldots & Y_{1l1} & Y_{112} & \ldots & Y_{1l2} & \ldots & Y_{11a} & \ldots & Y_{1la} \\ Y_{211} & \ldots & Y_{2l1} & Y_{212} & \ldots & Y_{2l2} & \ldots & Y_{21a} & \ldots & Y_{2la} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ Y_{g11} & \ldots & Y_{gl1} & Y_{g12} & \ldots & Y_{gl2} & \ldots & Y_{g1a} & \ldots & Y_{gla} \end{array}\right) \tag{9.7}$$

O arranjo cúbico $\underline{\mathbf{Z}}$ é um arranjo com as interações entre genótipos $\times$ locais $\times$ anos (arranjo de resíduos) obtida do modelo (9.6), ou seja, cada elemento $(gla)_{ijk}$ do arranjo de três entradas $\underline{\mathbf{Z}}$ é estimado pela seguinte relação:

$$\widehat{(gla)}_{ijk} = Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i.k} - \overline{Y}_{.ij} + \overline{Y}_{i..} + \overline{Y}_{.j.} + \overline{Y}_{..k} - \overline{Y}_{...} \tag{9.8}$$

em que:

$\widehat{(gla)}_{ijk}$: é o efeito da interação tripla estimada para o genótipo i no local j e no ano k;

Y_{ijk} : é a média das b repetições do genótipo i no local j e no ano k;

$\overline{Y}_{ij.}$: é a média dos elementos da i-ésima linha com a j-ésima coluna do arranjo de interação, obtida de ba observações;

$\overline{Y}_{i.k}$: é a média dos elementos da i-ésima linha com o k-ésimo tubo do arranjo de interação, obtida de bl observações;

$\overline{Y}_{.jk}$: é a média dos elementos da j-ésima coluna com o k-ésimo tubo do arranjo de interação, obtida de bg observações;

$\overline{Y}_{i..}$: é a média dos elementos da i-ésima fatia horizontal do arranjo de interação, obtida de bla observações;

$\overline{Y}_{.j.}$: é a média dos elementos da j-ésima fatia vertical do arranjo de interação, obtida de bga observações;

$\overline{Y}_{..k}$: é a média dos elementos da k-ésima fatia frontal do arranjo de interação, obtida de bgl observações;

$\overline{Y}_{...}$: é a média geral do experimento, obtida de $bgla$ observações.

9.5 Gráficos para interação tripla

9.5.1 Joint Plot

Um *joint biplot* na análise *multiway* é semelhante a um *biplot* padrão e todos os princípios de interpretação do *biplot* padrão podem ser utilizados. A diferença nesta construção é que o *joint plot* é construído como um *biplot* para dois fatores dada a matriz de componente do modelo Tucker3 referente ao terceiro fator (modo) ou fator de referência (modo de referência). Cada *joint plot* é construído usando diferentes fatias do arranjo núcleo. O fatiamento é feito para cada componente do modo de referência. Cada fatia contém o poder de ligação ou os pesos para os componentes dos modos apresentados no gráfico. Os coeficientes no componente associado ao modo de referência pondera inteiramente o *joint plot* por seus valores, de forma que os *joint plot* são pequenos para os pequenos valores no componente e grande para aqueles com grandes coeficientes.

O ponto inicial para construir um *joint plot* após ajustar um modelo de Tucker3 é obter uma matriz $\boldsymbol{\Delta}_r = \boldsymbol{A}\boldsymbol{G}_r\boldsymbol{B}^{'} = \boldsymbol{A}^*_r\boldsymbol{B}^{*'}_r$ de dimensão $I \times J$, com $r = 1, 2, \ldots, R$ ou uma matriz $\boldsymbol{\Delta_k} = \boldsymbol{A}\boldsymbol{H}_k\boldsymbol{B}^{'} = \boldsymbol{A}^*_k\boldsymbol{B}^{*'}_k$ de dimensão $I \times J$,com $k = 1, 2, \ldots, K$,após ajustar um modelo de Tucker2. Para cada fatia do núcleo, $\boldsymbol{G}_r$(ou $\boldsymbol{H}_k$), é necessário construir um *joint plot* para a matriz de componentes $\boldsymbol{A}^*$ $(J \times P)$ e $\boldsymbol{B}^*$ $(J \times Q)$.

O procedimento para a construção de um *joint plot* é o seguinte (KROONENBERG, 1994). A fatia do arranjo núcleo $\boldsymbol{G}_r$ $(P \times Q)$ é decomposta via decomposição em valor singular em

$$\boldsymbol{G}_r = \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}^{'}_r$$

e os vetores singulares $\boldsymbol{U}_r$ e $\boldsymbol{V}^{'}$ são combinados com as matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$, respectivamente, e a matriz diagonal $\boldsymbol{\Delta}_r$ com os valores singulares é dividido

entre as duas matrizes de forma que:

$$\boldsymbol{A}_r^* = \left(\frac{I}{J}\right)^{1/4} \boldsymbol{A}\boldsymbol{U}_r\boldsymbol{\Lambda}_r^{1/2} \tag{9.9}$$

$$\boldsymbol{B}_r^* = \left(\frac{J}{I}\right)^{1/4} \boldsymbol{B}\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{1/2}. \tag{9.10}$$

As colunas das matrizes de componentes ajustadas estão se referindo aos eixos do *joint plot*. Quando a matriz $\boldsymbol{G}_r$ (ou $\boldsymbol{H}_r$) não é quadrada, o seu posto é $M = min(P, Q)$, e somente M *joint biplot* podem ser apresentados. O procedimento completo rotaciona cada matriz de componentes para uma matriz ortonormal, seguido por um alongamento (ou encolhimento) dos componentes rotacionados. O tamanho do alongamento ou do encolhimento dos eixos é regulado pela raiz quadrada de $\lambda_{mm}^{(r)}$ e pela raiz quarta de $(\frac{J}{I})$. Note que se existe uma grande diferença na variabilidade explicada pelos eixos, isto é, entre $(\lambda_{mm}^r)^2$ e $(\lambda_{m'm'}^r)^2$, pode ocorrer uma dispersão visual considerável no gráfico, pois os coeficientes dos componentes são multiplicados por $(\lambda_{mm}^r)^{1/2}$.

Como $\boldsymbol{A}_r^*\boldsymbol{B}_r^{*'}=\boldsymbol{\Delta}_r$, cada elemento δ_{ij}^r é igual ao produto interno de $\boldsymbol{a}_i^*\boldsymbol{b}_j^{*'}$, e isto proporciona um alongamento na ligação entre a i-ésima linha da matriz de componentes $\boldsymbol{A}$ e a j-ésima linha da matriz de componentes $\boldsymbol{B}$, controlado pela r-ésima fatia do arranjo núcleo. Exibindo simultaneamente os dois modos em um gráfico, podem ser obtidas conclusões visuais sobre as relações entre eles. O espaçamento e a ordem das projeções dos objetos em uma variável correspondem ao tamanho do produto interno entre eles e, assim, a importância relativa daquela variável para os objetos.

Uma das vantagens do *joint plot* é que a interpretação das relações de variáveis e objetos podem ser feitas diretamente, sem envolver os eixos das componentes ou seus rótulos. Outra característica do *joint plot* é que por meio da fatia do arranjo central $\boldsymbol{G}_r$ ($\boldsymbol{H}_k$), os eixos das coordenadas *joint plot* são escalonados de acordo com a importância relativa, de forma que visualmente uma impressão correta da dispersão dos componentes é criada. Porém, no escalonamento simétrico dos componentes (como descrito anteriormente), as distâncias entre os objetos não são aproximações da distância Euclidiana, nem os ângulos entre as variáveis representam correlações. O *joint plot* para o modelo de Tucker3 é utilizado para investigar o significado dos objetos com respeito às variáveis explicitamente, dado um componente do terceiro modo. Para o modelo de Tucker2, o *joint plot* provê a informação sobre as relações entre objetos e variáveis dadas a um nível do terceiro modo (KROONENBERG,

2008).

Quanto a interpretação de um *joint plot* (VARELA, et al., 2006), suponha um gráfico que é projetado sobre o r-ésimo componente principal da terceira entrada tal que, no *joint plot* aparecem todos os níveis das duas primeiras entradas. Em seguida, selecione, a partir de matrizes $\boldsymbol{C}$ (matriz das componentes principais da terceira entrada), os níveis deste fator com maior peso no r-ésimo componente (positivos ou negativos), pois são estes valores que determinam os níveis da terceira entrada. Suponha que a matriz $\boldsymbol{C}$ tem um valor positivo e elevado associado ao k-ésimo nível da terceiro entrada, então proximidades entre os níveis da primeiro e da segunda entrada (por exemplo, i-ésimo nível do primeiro fator e o j-ésimo nível do segundo fator) indicam que a interação tripla entre i-ésimo nível da primeira entrada, j-ésimo nível da segunda entrada e k-ésimo nível da terceira entrada é positiva. Em contrapartida, se o i-ésimo nível do primeiro fator está muito distante do j-ésimo nível do segundo fator, indica que a interação tripla associada com i-ésimo nível da primeira entrada, j-ésimo nível da segunda entrada e k-ésimo nível da terceira entrada é negativa.

Suponha que a matriz $\boldsymbol{C}$ tem um alto valor negativo associado ao k-ésimo nível do terceiro fator, então proximidades entre os níveis do primeiro fator e do segundo fator (por exemplo, i-ésimo nível do primeiro fator e o j-ésimo nível do segundo fator) no *joint plot* indicam que a interação tripla entre i-ésimo nível da primeira entrada, j-ésimo nível da segunda entrada e k-ésimo nível da terceira entrada é negativa. Em contrapartida, se o i-ésimo nível do primeiro fator está muito distante do j-ésimo nível do segundo fator, indica que a interação tripla associada com i-ésimo nível da primeira entrada, j-ésimo nível da segunda entrada e k-ésimo nível da terceira entrada é positiva.

Em geral, os níveis de uma entrada localizado no centro do *joint plot* são considerados um conjunto que tem um desempenho médio em todos os outros modos.

9.5.2 Triplot

Arranjo de três entradas em um gráfico de duas dimensões

Sejam as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ na forma da Tabela 9.2 , em que as colunas são nomeadas de x e y e G_1, G_2, L_1, L_2,A_1, A_2 representam, por exemplo, genótipos, locais e anos. De forma semelhante, também é possível apresentar os três fatores como um arranjo $\underline{\mathbf{Z}}$ (Tabela 9.3).

Um gráfico de duas dimensões pode ser obtido se os valores x e y da Tabela 9.2 são representados em um plano cartesiano, em que cada linha de $\boldsymbol{A}$ é representado por um ponto. De modo semelhante, cada linha de $\boldsymbol{B}$ e $\boldsymbol{C}$- também é representado por um ponto (Figura 9.3). Este gráfico pode ser chamado de *triplot* pois apresenta as linhas das matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$. O *triplot* não apresenta somente as linhas das matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$, mas também apresenta os seus produtos elementos por elementos, que é o arranjo $\underline{\mathbf{Z}}$.

Tabela 9.2: As matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ para gerar $\underline{\mathbf{Z}}$

	x	y		x	y		x	y
Matriz linha ($\boldsymbol{A}$)			Matriz coluna ($\boldsymbol{B}$)			Matriz tubo ($\boldsymbol{C}$)		
G_1	3	5	L_1	2	3	A_1	5	2
G_2	-2	1	L_2	-4	-3	A_2	3	-1

Tabela 9.3: Elementos do arranjo $\underline{\mathbf{Z}}$ matrizado combinado as colunas tubos

	A_1		A_2	
	L_1	L_2	L_1	L_2
G_1	60	-90	-3	-21
G_2	-14	-34	-15	27

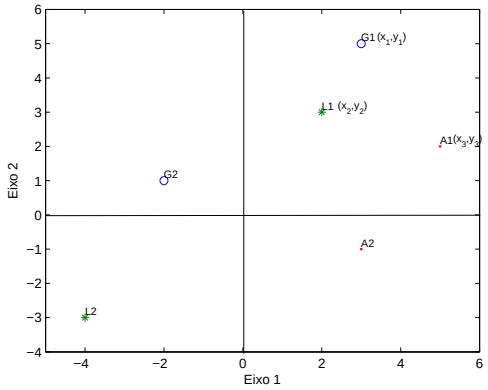

Figura 9.3: Um *triplot* que apresenta as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$. Os elementos de $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ são multiplicados segundo o produto de Hadamard para produzir o arranjo $\underline{\mathbf{Z}}$

O produto dos elementos das matrizes A, B e C e suas propriedades

Na Figura 9.4, a_{11}, b_{11} e c_{11} são denotados por x_1, x_2 e x_3, respectivamente e a_{12}, b_{12} e c_{12} são denotados por y_1, y_2 e y_3, respectivamente. Além disso

$$Z_{111} = x_1x_2x_3 + y_1y_2y_3.$$

A distância da origem O ao marcador de G_1 é chamada de vetor de G_1 e representado por $\overline{OG_1}$; as distâncias entre O e o marcador de L_1 e O e o marcador de A_1 são chamados de vetores de L_1 e A_1, representadas por $\overline{OL_1}$ e $\overline{OA_1}$, respectivamente.

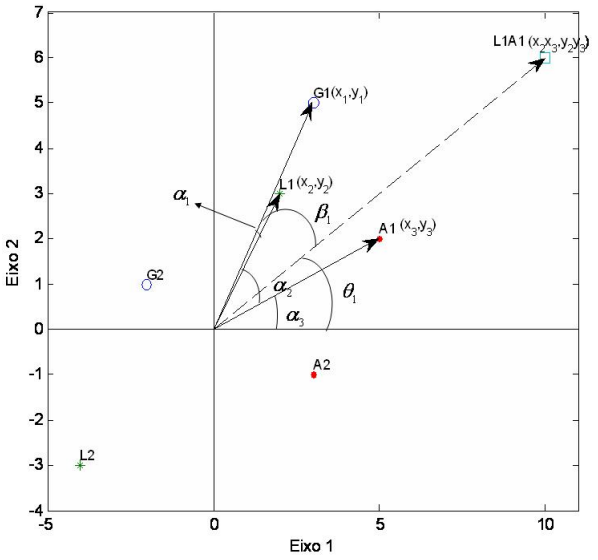

Figura 9.4: Os marcadores das linhas, colunas, tubos e combinação de uma coluna com um tubo do arranjo $\underline{\mathbf{Z}}$

Da Figura 9.4, têm-se as seguintes relações:

$$\begin{array}{ll}
x_1 = \overline{OG_1}cos(\alpha_1+\alpha_2+\alpha_3) & y_1 = \overline{OG_1}sen(\alpha_1+\alpha_2+\alpha_3) \\
\quad = \overline{OG_1}cos(\beta_1+\theta_1) & \quad = \overline{OG_1}sen(\beta_1+\theta_1) \\
x_2 = \overline{OL_1}cos(\alpha_2+\alpha_3) & y_2 = \overline{OL_1}sen(\alpha_2+\alpha_3) \\
x_3 = \overline{OA_1}cos(\alpha_3) & y_3 = \overline{OA_1}sen(\alpha_3) \\
x_2x_3 = u_1 = \overline{OL_1A_1}cos(\theta_1) & y_2y_3 = v_1 = \overline{OL_1A_1}sen(\theta_1)
\end{array}$$

Assim,

$$\begin{aligned}
z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_1u_1 + y_1v_1 \\
z_{111} &= \overline{OG_1}cos(\beta_1+\theta_1)\overline{OL_1A_1}cos(\theta_1) + \overline{OG_1}sen(\beta_1+\theta_1)\overline{OL_1A_1}sen(\theta_1) \\
z_{111} &= \overline{OG_1}\,\overline{OL_1A_1}cos(\beta_1+\theta_1-\theta_1) \\
z_{111} &= \overline{OG_1}\,\overline{OL_1A_1}cos(\beta_1). \qquad (9.11)
\end{aligned}$$

Então, o primeiro elemento da primeira linha, da primeira coluna, do primeiro tubo de $\underline{\mathbf{Z}}$ é produto dos vetores da primeira linha de $\boldsymbol{G}(\overline{OG_1})$, o vetor $\overline{OL_1A_1}$ e o cosseno de (β_1) que é o ângulo entre estes vetores (Figura 9.4).

Para vizualizar z_{111} diretamente do *triplot*, a equação 9.11 pode ser escrita como:

$$z_{111} = \overline{OG_1}cos(\beta_1)\overline{OL_1A_1} = \overline{OP_{G_1}}\ \overline{OL_1A_1} \tag{9.12}$$

em que $\overline{OP_{G_1}} = \overline{OG_1}cos(\beta_1)$ é a projeção do vetor $\overline{OG_1}$ no vetor $\overline{OL_1A_1}$.

Alternativamente a equação (9.11) pode ser escrita como:

$$\begin{aligned} z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_2u_2 + y_2v_2 \\ z_{111} &= \overline{OL_1}cos(\beta_2+\theta_2)\overline{OG_1A_1}cos(\theta_2) + \overline{OL_1}sen(\beta_2+\theta_2)\overline{OG_1A_1}sen(\theta_2) \\ z_{111} &= \overline{OL_1}\ \overline{OG_1A_1}cos(\beta_2+\theta_2-\theta_2) \\ z_{111} &= \overline{OL_1}\ \overline{OG_1A_1}cos(\beta_2) = \overline{OP_{L_1}}\ \overline{OG_1A_1} \end{aligned} \tag{9.13}$$

ou

$$\begin{aligned} z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_3u_3 + y_3v_3 \\ z_{111} &= \overline{OA_1}cos(\beta_3+\theta_3)\overline{OG_1L_1}cos(\theta_3) + \overline{OA_1}sen(\beta_3+\theta_3)\overline{OG_1L_1}sen(\theta_3) \\ z_{111} &= \overline{OA_1}\ \overline{OG_1L_1}cos(\beta_3+\theta_3-\theta_3) \\ z_{111} &= \overline{OA_1}\ \overline{OG_1L_1}cos(\beta_3) = \overline{OP_{A_1}}\ \overline{OG_1L_1} \end{aligned} \tag{9.14}$$

em que $u_2 = x_1x_3$, $u_3 = x_1x_2$, $v_2 = y_1y_3$, $v_3 = y_1y_2$, β_2 é o ângulo entre vetores $\overline{OL_1}$ e $\overline{OG_1A_1}$, β_3 é o ângulo entre vetores $\overline{OA_1}$ e $\overline{OG_1L_1}$, $\overline{OP_{L1}}$ é a projeção do vetor $\overline{OL_1}$ no vetor $\overline{OG_1A_1}$ e $\overline{OP_{A1}}$ é a projeção do vetor $\overline{OA_1}$ no vetor $\overline{OG_1L_1}$.

As equações (9.11), (9.13) e (9.14) podem ser generalizadas como:

$$z_{ijk} = \overline{OG_i}cos(\beta_{G_i;j*k})\overline{OL_jA_k} \tag{9.15}$$

$$z_{ijk} = \overline{OL_j}cos(\beta_{L_j;i*k})\overline{OG_iA_k} \tag{9.16}$$

$$z_{ijk} = \overline{OA_k}cos(\beta_{A_k;i*j})\overline{OG_iL_j} \tag{9.17}$$

em que z_{ijk} é o elemento de $\underline{\mathbf{Z}}$ da linha i, coluna j e tubo k, com $i = 1,\dots I$, $j = 1,\dots J$, e $k = 1,\dots K$; $\overline{OG_i}$, $\overline{OL_j}$, $\overline{OA_k}$, $\overline{OL_jA_k}$, $\overline{OG_iA_k}$ e $\overline{OG_iL_j}$ são os vetores de G_i, L_j, A_k, L_jA_k, G_iA_k e G_iL_j, respectivamente, que é a distância entre a origem do *triplot* e os marcadores G_i, L_j, A_k, L_jA_k (produto de Hadamard entre L_j e A_k), G_iA_k (produto de Hadamard entre G_i e A_k) e G_iL_j (produto

de Hadamard entre G_i e L_j); e $\beta_{G_i;j*k}$, $\beta_{L_j;i*k}$ e $\beta_{A_k;i*j}$ são os ângulos entre os vetores $\overline{OG_i}$ e $\overline{OL_jA_k}$, $\overline{OL_j}$ e $\overline{OG_iA_k}$ e $\overline{OA_k}$ e $\overline{OG_iL_j}$, respectivamente.

Observe que $\overline{OG_i}$, $\overline{OL_j}$, $\overline{OA_k}$, $\overline{OL_jA_k}$, $\overline{OG_iA_k}$ e $\overline{OG_iL_j}$, nunca serão negativos, por serem o comprimentos de vetores, mas o $cos(\beta_{G_i;j*k})$, $cos(\beta_{L_j;i*k})$ e $cos(\beta_{A_k;i*j})$, podem ser positivos ou negativos, dependendo de $\beta_{G_i;j*k}$, $\beta_{L_j;i*k}$ e $\beta_{A_k;i*j}$. Conseqüentemente, o sinal de z_{ijk} é determinado somente por $\beta_{G_i;j*k}$, $\beta_{L_j;i*k}$ e $\beta_{A_k;i*j}$, ou seja:

i) z_{ijk} será zero se $\beta_{G_i;j*k} = \beta_{L_j;i*k} = \beta_{A_k;i*j} = 90^0$;

ii) z_{ijk} será positivo se os ângulos $\beta_{G_i;j*k}$, $\beta_{L_j;i*k}$ e $\beta_{A_k;i*j}$ forem obtuso;

iii) z_{ijk} será negativo se os ângulos $\beta_{G_i;j*k}$, $\beta_{L_j;i*k}$ e $\beta_{A_k;i*j}$ forem agudo.

9.5.3 Visualizando o *triplot*

O *triplot* não apresenta somente as linhas de $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$, mas também o arranjo $\underline{\mathbf{Z}}$. Além disso, com base nas equações (9.15), (9.16) e (9.17), o *triplot* permite visualizar as relações entre as linhas, entre as colunas e entre os tubos do arranjo $\underline{\mathbf{Z}}$. Uma aplicação adicional é a possibilidade de identificação visual de quais linhas de certa matriz de componentes têm os maiores valores, para certa combinação das linhas das outras matrizes (produto de Hadamard entre as linhas das outras matrizes). E por último, uma outra aplicação é que este gráfico pode ser utilizado para agrupar as linhas de cada matriz de componentes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$.

Comparação visual dos elementos de uma linha, coluna ou tubo do arranjo

A equação (9.15) pode ser reescrita como

$$z_{ijk} \;=\; \overline{OG_i}cos(\beta_{G_i;j*k})\overline{OL_jA_k} = \overline{OP_{G_{i;j*k}}}\;\overline{OL_jA_k} \tag{9.18}$$

em que $\overline{OP_{G_{i;j*k}}}$ é a projeção do vetor $\overline{OG_i}$ dentro do vetor $\overline{OL_jA_k}$. Como para certos valores dados de L_j e A_k, tem-se que $\overline{OL_jA_k}$ é comum para todas as linhas G_i e, portanto:

$$\frac{z_{ijk}}{\overline{OL_jA_k}} \;=\; \overline{OP_{G_{i;j*k}}}\,. \tag{9.19}$$

Em outras palavras, a magnitude relativa dos elementos que estão na i-ésima linha de $\underline{\mathbf{Z}}$, para uma combinação da coluna j com o tubo k, $\frac{z_{ijk}}{\overline{OL_jA_k}}$, pode ser comparada através da projeção ($\overline{OP_{G_{i;j*k}}}$) sobre $\overline{OL_jA_k}$.

De forma análoga, a equação (9.16) pode ser escrita como:

$$z_{ijk} \;=\; \overline{OL_j} cos(\beta_{L_j;i*k}) \overline{OG_i A_k} = \overline{OP_{L_{j;i*k}}}\; \overline{OG_i A_k} \tag{9.20}$$

em que $\overline{OP_{L_{j;i*k}}}$ é a projeção $\overline{OL_j}$ no vetor $\overline{OG_i A_k}$. Assim, para certos elementos G_i e A_k, têm-se que $\overline{OG_i A_k}$ é comum para todos os elementos L_j. Dessa forma, a equação (9.20), pode ser reescrita como:

$$\frac{z_{ijk}}{\overline{OG_i A_k}} \;=\; \overline{OP_{L_{j;i*k}}} \tag{9.21}$$

e a magnitude relativa dos elementos da j-ésima coluna $\underline{\mathbf{Z}}$ para a combinação da linha i com o tubo k, $\frac{z_{ijk}}{\overline{OG_i A_k}}$, pode ser visualizada pela comparação de suas projeções ($\overline{OP_{L_{j;i*k}}}$) sobre o vetor $\overline{OG_i A_k}$.

Novamente, como foi feito para as equações (9.15) e (9.16), também pode ser feito para a equação (9.17):

$$z_{ijk} \;=\; \overline{OA_k} cos(\beta_{A_k;i*j}) \overline{OG_i L_j} = \overline{OP_{A_{k;i*j}}}\; \overline{OG_i L_j} \tag{9.22}$$

em que $\overline{OP_{A_{k;i*j}}}$ é a projeção de $\overline{OA_k}$ no vetor $\overline{OG_i L_j}$. Como para cada elemento G_i e L_j, $\overline{OG_i L_j}$ é comum para todas as linhas A_k, a equação (9.22) pode ser reescrita como:

$$\frac{z_{ijk}}{\overline{OG_i L_j}} \;=\; \overline{OP_{A_{k;i*j}}}\;. \tag{9.23}$$

Assim, a magnitude relativa dos elementos no k-ésimo tubo de $\underline{\mathbf{Z}}$ para a combinação da linha i com a coluna j, $\frac{z_{ijk}}{\overline{OG_i L_j}}$, pode ser visualizada pela comparação das projeções do vetor $\overline{OA_k}$ sobre o vetor $\overline{OG_i L_j}$.

9.5.4 Relações entre linhas, entre colunas e entre tubos

Relações entre as linhas podem ser visualizadas pelos ângulos entre os seus vetores. Pode-se estabelecer como regra que os ângulos entre os vetores das linhas do arranjo de $\underline{\mathbf{Z}}$ aproxima-se da correlação entre as linhas do arranjo.

Note que o cosseno do ângulo entre os vetores de duas linhas é determinado somente pelos valores na matriz $\boldsymbol{A}$ e não tem nada a ver com valores de $\boldsymbol{B}$ e $\boldsymbol{C}$, enquanto que o cálculo do coeficiente de correlação é baseado no arranjo $\underline{\mathbf{Z}}$, que é dependente de $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$. Conseqüentemente, os ângulos entre as linhas de $\boldsymbol{A}$ no *triplot* deve ser relativamente relacionado com o coeficiente de

correlação entre as linhas de $\underline{\mathbf{Z}}$, mas nenhuma correspondência perfeita deve ser esperada.

O mesmo raciocínio pode ser considerado para as colunas e tubos de $\underline{\mathbf{Z}}$, ou seja, o cosseno do ângulo entre as linhas de $\boldsymbol{B}$ é aproximadamente a correlação entre as colunas de $\underline{\mathbf{Z}}$ e o cosseno do ângulo entre as colunas de $\boldsymbol{C}$ é próximo do coeficiente de correlação dos tubos de $\underline{\mathbf{Z}}$.

Esta propriedade é muito útil para visualizar via *triplot*, as interrelações entre as linhas, entre as colunas e entre os tubos de um conjunto de dados organizados em um arranjo de três entradas.

9.6 Exemplos

Os dados a serem utilizados são relativos a experimentos com 13 genótipos de feijão que foram conduzidos em 9 experimentos distintos constituídos pelos anos agrícolas de 2000/2001, 2001/2002 e 2005/2006, nos municípios de Dourados e Aquidauana no estado de Mato Grosso do Sul, sendo que os experimentos foram instalados na época das águas (Dourados) e também na época da seca (Dourados e Aquidauana). Cada local é constituído de município e uma época de instalação, conforme representados na Tabela 9.4. Têm-se ainda que em cada experimento foi utilizado um delineamento em blocos ao acaso, com 3 blocos em cada experimento.

Tabela 9.4: Caracterização dos ambientes experimentais

Município	Época	Local[1]	Ano agrícola
Dourados	"das águas"	L1	2000/2001 (A1)
Dourados	"das secas"	L2	2000/2001 (A1)
Aquidauana	"das secas"	L3	2000/2001 (A1)
Dourados	"das águas"	L1	2001/2002 (A2)
Dourados	"das secas"	L2	2001/2002 (A2)
Aquidauana	"das secas"	L3	2001/2002 (A2)
Dourados	"das águas"	L1	2005/2006 (A3)
Dourados	"das secas"	L2	2005/2006 (A3)
Aquidauana	"das secas"	L3	2005/2006 (A3)

[1]O fator local consiste na combinação de municípios com épocas

Para cada um dos genótipos, em cada um dos ambientes, foram avaliadas as seguintes variáveis respostas:

1. Número médio de vagens por planta (VAG): medido na colheita durante o processo de arranquio;

2. Número médio de sementes por vagem (SEM): obtido na colheita durante o processo de trilha;
3. Massa de 100 sementes (MCS): medida após a colheita e expresso em gramas;
4. Produtividade de grãos (PROD): medida após a colheita e expressa em ton/ha;

sendo que neste trabalho será considerado somente a variável produtividade de grãos.

9.6.1 Análise de variância conjunta com três fatores

Pela Tabela 9.5, que corresponde à análise de variância conjunta efetuada com os dados observados, verifica-se que o efeito de blocos dentro de locais dentro de anos e os efeitos principais de genótipos, locais e anos são não significativos, enquanto que os efeitos das interações duplas (genótipos × locais, genótipos × anos e locais × anos) e interação tripla (genótipos × locais × anos) são significativos.

Tabela 9.5: Análise de variância conjunta para um conjunto de dados com 13 genótipos avaliados em 3 locais, 3 anos com 3 blocos

Fontes de variação	GL	SQ	QM	F	valor-p
B d. L d. A	18	0,70	0,04	1,00	0,4608
G	12	31,60	2,63	2,02	0,0622
L	2	65,21	32,60	4,11	0,1071
A	2	13,02	6,51	0,82	0,5030
G × L	24	21,19	0,88	2,32	0,0065
G × A	24	19,17	0,80	2,10	0,0143
L × A	4	31,71	7,91	203,28	<0,0001
G × L × A	48	18,27	0,38	9,74	<0,0001
Resíduo	216	8,51	0,04		
Total	350	209,40			

9.6.2 Modelos de três entradas para a interação tripla

construindo um arranjo cúbico de dimensão (13 × 3 × 3) com os efeitos das interações triplas entre genótipos × locais × anos, de modo que nas linhas

Tabela 9.6: Efeitos da interação tripla para cada combinação de genótipos, locais e anos

	A1			A2			A3		
	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3
G_1	-0,004	0,271	-0,267	0,172	-0,137	-0,036	-0,168	-0,135	0,303
G_2	0,313	-0,230	-0,083	-0,077	0,108	-0,030	-0,236	0,122	0,114
G_3	0,085	-0,078	-0,007	-0,168	0,290	-0,122	0,083	-0,212	0,129
G_4	0,137	-0,363	0,226	0,001	0,150	-0,151	-0,137	0,212	-0,075
G_5	0,142	-0,588	0,447	-0,200	0,165	0,036	0,059	0,424	-0,482
G_6	0,037	-0,118	0,080	-0,227	0,274	-0,047	0,190	-0,156	-0,034
G_7	-0,337	-0,029	0,366	0,052	0,007	-0,060	0,285	0,022	-0,306
G_8	0,111	-0,214	0,103	0,196	0,100	-0,296	-0,306	0,114	0,193
G_9	0,325	-0,030	-0,295	-0,194	-0,287	0,481	-0,131	0,316	-0,185
G_{10}	0,033	0,021	-0,054	0,024	-0,188	0,163	-0,057	0,167	-0,109
G_{11}	-0,131	0,463	-0,332	-0,101	-0,196	0,297	0,232	-0,267	0,035
G_{12}	-0,352	0,543	-0,191	0,241	0,012	-0,253	0,111	-0,555	0,444
G_{13}	-0,357	0,350	0,007	0,281	-0,298	0,018	0,077	-0,051	-0,025

estão os genótipos, nas colunas os locais e nos tubos os anos. Estes efeitos das interações triplas estão apresentados na Tabela 9.6.

Portanto, faz-se necessário utilizar uma metodologia adequada para interpretar a interação tripla, ou seja, o uso da metodologia *multiway* (modelos Tucker3 e PARAFAC).

Ajuste do Modelo de Tucker3

Para selecionar o melhor modelo de Tucker3, foi utilizado o procedimento de Timmerman-Kiers. Os resultado deste procedimento estão apresentados na Tabela 9.7. As estimativas de $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ foram obtidas utilizando a solução proposta por Tucker (1966) como soluções iniciais, sendo que esta solução inicial é o "*default*"do *Toolbox N-way*.

Inicialmente é possível ajustar 117 modelos Tucker3, mas após aplicar o primeiro filtro, que consiste em verificar quais dos modelos satisfazem as condições de Kruskal (1989) (para o conjunto de dados deste trabalho as condições são: $P \leq QR = 9$, $Q \leq PR = 27$ e $R \leq PQ = 27$), restou a possibilidade de escolher entre 28 modelos (Tabela 9.7). Após o segundo filtro, ou seja, dentro de uma mesma classe de modelos com $S = P + Q + R$ componentes, deve-se selecionar aqueles que tem a maior soma de quadrados, assim restou 12 modelos. O próximo passo foi calcular a quantidade $dif_S = SQ_S - SQ_{S-1}$, sendo que para a solução trivial ($P = 1$; $Q = 1$; e $R = 1$) Schepers; Ceulemans e Van Mechelen (2008) sugerem que o $dif_S = 0$ e também deve-se calcular o ponto de corte para os dif_S ($\|\underline{\mathbf{Z}}\|/S_{min} = (SQ_{G\times L\times A}/b)/(min(I; JK) + min(J; IK) + min(K; IJ) - 3) = (18,269/3)/12 = 0,5074$), mas como todos os dif_S são mai-

Tabela 9.7: Resultado do procedimento de Timmerman-Kiers para selecionar o modelo de Tucker3

Após Primeiro filtro					Após Segundo filtro						
P	Q	R	S	SQ	P	Q	R		SQ	dif	b_s
1	1	1	3	49,84	1	1	1	3	49,84	0,00	0,00
1	2	2	5	51,01	2	2	1	5	67,72	17,88	1,05
2	1	2	5	66,75	2	2	2	6	73,38	5,66	
2	2	1	5	67,72	3	2	2	7	90,38	17,00	1,77
2	2	2	6	73,38	4	2	2	8	100,00	9,62	
1	3	3	7	52,04	4	3	2	9	100,00	0,00	
3	1	3	7	67,10	5	3	2	10	100,00	0,00	
3	3	1	7	67,72	6	3	2	11	100,00	0,00	
2	2	3	7	73,38	6	3	3	12	100,00	0,00	
2	3	2	7	73,38	7	3	3	13	100,00	0,00	
3	2	2	7	90,38	8	3	3	14	100,00	0,00	
2	3	3	8	73,38	9	3	3	15	100,00	0,00	
3	2	3	8	90,38							
3	3	2	8	90,38							
4	2	2	8	100,00							
3	3	3	9	90,38							
4	2	3	9	100,00							
4	3	2	9	100,00							
4	3	3	10	100,00							
5	2	3	10	100,00							
5	3	2	10	100,00							
5	3	3	11	100,00							
6	2	3	11	100,00							
6	3	2	11	100,00							
6	3	3	12	100,00							
7	3	3	13	100,00							
8	3	3	14	100,00							
9	3	3	15	100,00							

SQ: Soma de quadrados explicada pelo modelo com P, Q e R componentes

ores que o ponto de corte, nenhuma solução deve ser desconsiderada. Ainda, deve-se considerar somente os dif_S que estão em ordem decrescente, assim a solução (2,2,2) foi desconsiderada. Por último deve-se calcular a relação $b_S = \frac{dif_S}{dif_S^*}$ e então, segundo Timmerman e Kiers (2000), deve-se escolher o modelo que resultou no maior b_S. Logo, para este conjunto de dados o procedimento de Timmerman e Kiers sugere que deve-se selecionar o modelo de Tucker3 (3,2,2).

As matrizes $\boldsymbol{A}$ com três componentes, $\boldsymbol{B}$ com duas componentes, $\boldsymbol{C}$ com duas componentes e o arranjo núcleo $\underline{\mathbf{G}}$ são apresentados na Tabela 9.8. Estas componentes explicam $90,38\%$ da soma de quadrados da interação tripla entre genótipos $\times$ locais $\times$ anos, sendo que as três componentes, p_1, p_2 e p_3, da matriz $\boldsymbol{A}$ (referente aos genótipos) explicam $52,05\%$, $21,34\%$ e $17,00\%$, respectivamente. As duas componentes, q_1 e q_2, da matriz $\boldsymbol{B}$ (que refere-se

aos locais) explicam $64,48\%$ e $25,90\%$, respectivamente e na matriz $\boldsymbol{C}$ (matriz referente aos anos), as duas componentes r_1 e r_2 explicam $67,12\%$ e $23,26\%$, respectivamente.

Tabela 9.8: Escores dos componentes principais para um modelo de Tucker3 (3,2,2) para o arranjo da interação tripla entre genótipos × locais × anos

Genótipos ($\boldsymbol{A}$)			
	p_1	p_2	p_3
G_1	0,2633	0,0854	-0,2875
G_2	-0,1587	0,0489	-0,4110
G_3	-0,0006	-0,2277	-0,1433
G_4	-0,2806	-0,1755	-0,1231
G_5	-0,5519	-0,0694	0,2732
G_6	-0,0696	-0,1914	0,0626
G_7	-0,0646	-0,2284	0,5738
G_8	-0,1035	-0,2491	-0,3853
G_9	-0,1383	0,6940	-0,0839
G_{10}	-0,0376	0,2526	0,0580
G_{11}	0,3361	0,3268	0,2070
G_{12}	0,5517	-0,3125	-0,0566
G_{13}	0,2543	0,0463	0,3161

Locais ($\boldsymbol{B}$)		
	q_1	q_2
L_1	-0,1498	0,8026
L_2	0,7700	-0,2716
L_3	-0,6202	-0,5311

Anos ($\boldsymbol{C}$)		
	r_1	r_2
A_1	-0,7903	0,2051
A_2	0,2175	-0,7870
A_3	0,5728	0,5819

Arranjo núcleo ($\boldsymbol{G}$)				
	r_1		r_2	
	q_1	q_2	q_1	q_2
p_1	-1,6769	0,4374	-0,1383	-0,3829
p_2	-0,2772	-0,4625	0,9434	0,3448
p_3	0,1206	0,8824	0,3372	0,3579

O arranjo núcleo $\underline{\mathbf{G}}$ (na tabela 9.8) apresenta as relações entre as componentes e entre esses valores a relação mais importante é entre as primeiras componentes de cada fator, $g_{111} = -1,6769$. Esta quantia indica que a combinação da primeira componente dos genótipos com a primeira componente dos locais com a primeira componente dos anos explicam juntas $(-1,6769)^2/6,0896 \times 100\% = 46,17\%$ da $SQ_{G\times L\times A}$ e a relação menos importante é a relação entre a terceira componente dos genótipos com a primeira componente dos locais com a primeira componente dos anos que explicam juntas $(-0,1206)^2/6,0896 \times 100\% = 0,23\%$ da $SQ_{G\times L\times A}$.

Ainda pela Tabela 9.8, percebe-se que a primeira componente da matriz $\boldsymbol{C}$ (Anos), é caracterizada por um contraste entre o ano 1 (-0,7903) e o ano 3 (0,5728) e a segunda componente é caracterizada por um contraste entre o ano 2 (-0,7870) e o ano 3 (0,5819). Assim, ao construir um *joint plot*, que projeta os genótipos e locais dentro da primeira componente dos anos, as conclusões serão restritas somente ao ano 1 e ao ano 3 (Figura 9.5), mas quando projetar os genótipos e locais dentro da segunda componente dos anos, as conclusões serão válidas para o ano 2 e ano 3 (Figura 9.6).

O primeiro *joint plot* (Figura 9.5) corresponde ao *biplot* da matriz $\boldsymbol{\Delta}_1 = \boldsymbol{A}\boldsymbol{G}_1\boldsymbol{B}^{'}$, em que $\boldsymbol{G}_1$ é a primeira fatia frontal do arranjo núcleo $\underline{\mathbf{G}}$, obtido ao ajustar modelo de Tucker3 (3,2,2). Este *joint plot* é projetado dentro da primeira componente do fator ano (r_1) e esta componente explica $67,12\%$ da $SQ_{G\times L\times A}$, sendo que a primeira componente deste gráfico corresponde a

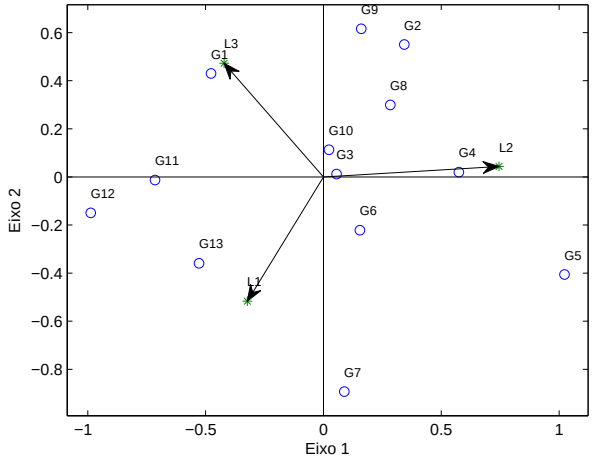

Figura 9.5: *Joint plot* projetado dentro da primeira componente do terceiro modo

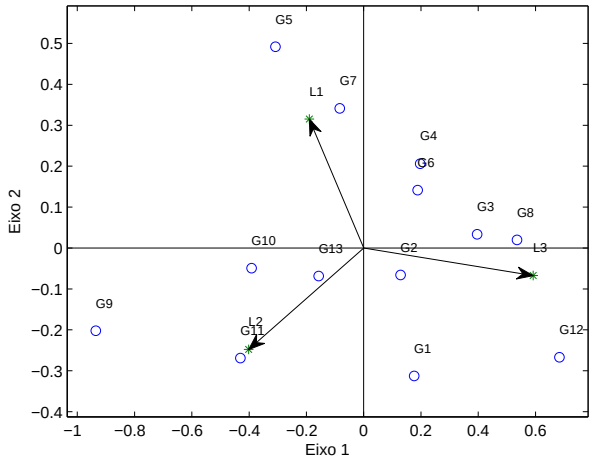

Figura 9.6: *Joint plot* projetado dentro da segunda componente do terceiro modo

49, 88% e a segunda componente explica 17, 23% da soma de quadrados da interação genótipos × locais × anos. O gráfico representa a interação entre genótipos × ambientes no ano 1 (2000/2001) e no ano 3 (2005/2006). Assim, em relação ao ano 1 (c_{11} é negativo), observa-se pela Figura 9.5 as seguintes relações:

- O genótipo 1 teve uma interação negativa com $L3$ (Aquidauana na época das secas), positiva com $L2$ (Dourados na época da secas) e não interage com o local $L1$ (Dourados na época das águas);
- Para os genótipos 11, 12 e 13, observa-se uma interação positiva com $L2$

e negativa com $L1$ e $L3$;

- Os genótipos 4 e 5 teve uma interação positiva com $L1$ e $L3$ e negativa com $L2$;
- Para os genótipos 2, 8 e 9, nota-se que a interação é negativa com $L2$ e $L3$ e positiva com $L1$;
- O genótipo 7 teve interação negativa com $L1$, positiva com $L3$ e não interage com o local $L1$;

ainda neste gráfico , com relação ao ano 3 (c_{31} é positivo) observa-se as relações

- O genótipo 1 teve uma interação positiva com $L3$, negativa com $L2$ e não interage com o local $L1$;
- Para os genótipos 11, 12 e 13, observa-se uma interação negativa com $L2$ e positiva com $L1$ e $L3$;
- Os genótipos 4 e 5 teve uma interação negativa com $L1$ e $L3$ e positiva com $L2$;
- Para os genótipos 2, 8 e 9, nota-se que a interação é positiva com $L2$ e $L3$ e negativa com $L1$;
- O genótipo 7 teve um interação positiva com $L1$, negativa com $L3$ e não interage com o local $L2$;

e com relação aos genótipos 3, 6 e 10, que estão no centro deste gráfico, pode-se dizer que todos tem uma baixa interação com todos os locais no ano 1 e no ano 3 e, conseqüentemente, são genótipos estáveis.

De maneira semelhante, o segundo *joint plot* (Figura 9.6) corresponde ao *biplot* da matriz $\boldsymbol{\Delta}_2 = \boldsymbol{A}\boldsymbol{G}_2\boldsymbol{B}'$, em que $\boldsymbol{G}_2$ é a segunda fatia frontal do arranjo núcleo $\underline{\mathbf{G}}$. Para este *joint plot*, que é projetado no segundo componente do fator ano (r_2), a $SQ_{G\times L\times A}$ explica $23,26\%$, sendo que o primeiro eixo deste gráfico corresponde a $21,49\%$ e o segundo eixo explica $1,94\%$ da soma de quadrados da interação genótipos $\times$ locais $\times$ anos. O gráfico representa a interação entre genótipos $\times$ locais no ano 2 (2001/2002) e no ano 3 (2005/2006). Assim, em relação ao ano 2 (c_{22} é negativo), observa-se pela Figura 9.6 as seguintes relações:

- O genótipo 1 teve uma interação negativa com $L2$ e $L3$, positiva com $L1$;

- O genótipo 4 teve interação negativa com $L1$ e $L3$, positiva com $L2$;
- Para os genótipos 9, 10 e 11, observa-se uma interação positiva com $L1$ e $L3$ e negativa com $L2$;
- Os genótipos 5 e 7 teve uma interação positiva com $L2$ e $L3$ e negativa com $L1$;
- Para os genótipos 3, 8 e 12, nota-se que a interação é positiva com $L1$ e $L2$ e negativa com $L3$;

ainda neste gráfico, com relação ao ano 3 (c_{32} é positivo) observa-se as relações:

- O genótipo 1 teve uma interação positiva com $L2$ e $L3$, negativa com $L1$;
- O genótipo 4 teve interação positiva com $L1$ e $L3$, negativa com $L2$;
- Para os genótipos 9, 10 e 11, observa-se uma interação negativa com $L1$ e $L3$ e positiva com $L2$;
- Os genótipos 5 e 7 teve uma interação negativa com $L2$ e $L3$ e positiva com $L1$;
- Para os genótipos 3, 8 e 12, nota-se que a interação é negativa com $L1$ e $L2$ e positiva com $L3$;

e com relação aos genótipos 2, 6 e 13, que estão no centro deste gráfico, pode-se dizer que estes genótipos tem uma baixa interação com todos os locais no ano 2 e no ano 3 e, conseqüentemente, são genótipos estáveis.

Ajuste do Modelo PARAFAC

Inicialmente, ajustou-se o modelo PARAFAC com um, dois, três e quatro componentes e observou a porcentagem da soma de quadrados da interação tripla explicada pelo modelo (Tabela 9.9). Estes ajustes foram obtidos utilizando como estimativa inicial os valores obtidos pela Decomposição Trilinear Direta. Assim, decidiu-se utilizar o modelo com dois componentes, pois este explica 73,37% da soma de quadrados da interação genótipos × locais × anos e o acréscimo que se têm, na soma de quadrados explicada pelo modelo, ao aumentar uma componente, é de 17,00% (Figura 9.7). As estimativas do modelo com dois componentes são apresentadas na Tabela 9.10, sendo que as estimativas iniciais para as matrizes de componentes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ foram obtidas pelo método Decomposição Trilinear Direta, que é o padrão do *Toolbox N-way*.

Tabela 9.9: Número de componentes utilizado no modelo PARAFAC e a porcentagem da soma de quadrados da interação tripla explicada pelo modelo

Número de componentes	1	2	3	4
Soma de quadrados explicada(%)	50,11	73,37	90,37	100,00
Acréscimo na soma de quadrados explicada(%)		23,26	17,00	9,63

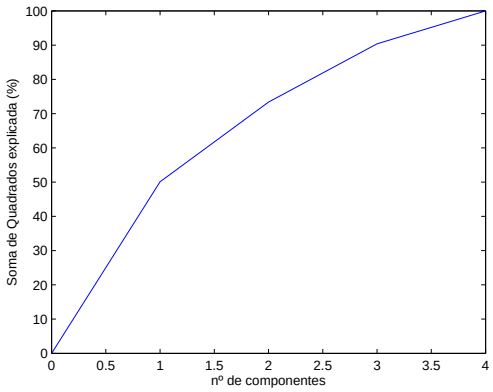

Figura 9.7: *Scree plot* do número de componentes no modelo PARAFAC e a porcentagem da soma de quadrados explicada pelo modelo

Tabela 9.10: Primeiro e segundo escores dos componentes principais para genótipos ($\boldsymbol{a}_1$ e $\boldsymbol{a}_2$), locais ($\boldsymbol{b}_1$ e $\boldsymbol{b}_2$) e anos ($\boldsymbol{c}_1$ e $\boldsymbol{c}_2$) para os dados do exemplo.

	Matriz $\boldsymbol{A}$			Matriz $\boldsymbol{B}$			Matriz $\boldsymbol{C}$	
	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$		$\boldsymbol{b}_1$	$\boldsymbol{b}_2$		$\boldsymbol{c}_1$	$\boldsymbol{c}_2$
G_1	0,8186	0,4155	L_1	-0,1632	-0,0423	A_1	0,8099	0,7239
G_2	-0,6195	-0,4390	L_2	0,7745	-0,6850	A_2	-0,4947	-0,6890
G_3	0,3431	0,5289	L_3	-0,6112	0,7273	A_3	-0,3152	-0,0350
G_4	-0,7248	-0,2098						
G_5	-1,8643	-1,0944						
G_6	0,0351	0,2747						
G_7	0,0913	0,3441						
G_8	0,0263	0,3799						
G_9	-1,5310	-1,9111						
G_{10}	-0,5145	-0,6704						
G_{11}	0,6893	-0,0257						
G_{12}	2,4284	1,9615						
G_{13}	0,8221	0,4458						

Ao observar os valores da Tabela 9.10 com relação a matriz de componentes principais $\boldsymbol{A}$, nota-se que dentro da primeira componente os maiores valores (tanto positivo quanto negativo) estão relacionados aos genótipos 5 (negativo), 9 (negativo) e 12 (positivo), sendo portanto, que a primeira componente ($\boldsymbol{a}_1$)

está relacionada diretamente com o genótipo 12 e esta componente está relacionada inversamente com os genótipos 5 e 9. Mas esta componente ainda tem relações com outros genótipos. Dentro deste componente também há escores próximos de zeros (genótipos 6 e 8), que indica que esta componente não tem relação com estes genótipos. Já para a segunda componente ($\boldsymbol{a}_2$) também observa-se uma relação positiva com o genótipo 12 e relação negativa com os genótipos 5 e 9, e o genótipo que tem uma relação muito baixa com esta componente é o genótipo 11.

Para a matriz de componentes da segunda entrada, $\boldsymbol{B}$, que está relacionada aos locais, nota-se que a primeira componente ($\boldsymbol{b}_1$) está relacionada positivamente com o local 2 (município de Dourados na época da seca), negativamente com o local 3 (município de Aquidauana na época da seca) e não apresentou relação com o local 1 (município de Dourados na época das águas). Na componente dois ($\boldsymbol{b}_2$) apresentou relação positiva com o local 3, relação negativa com o local 2 e também não apresentou relação com o local 1.

Com relação a matriz de componentes $\boldsymbol{C}$ da terceira entrada (que apresenta informações sobre os anos), percebe-se que a componente 1 ($\boldsymbol{c}_1$) depende positivamente do ano 1 (ano agrícola 2000/2001) e negativamente do ano 2 (ano agrícola 2001/2002) e ano 3 (ano agrícola 2005/2006). Na segunda componente ($\boldsymbol{c}_2$) é dominado positivamente também pelo ano 1, negativamente pelo ano 2, já o ano 3 não interfere na segunda componente.

9.6.3 Triplot

A interpretação do *triplot*, quanto à interação entre genótipos × locais × ano, pode ser feita observando a magnitude e o sinal dos escores de genótipos, locais e anos dos eixos, como é feito para o já conhecido *biplot*. Assim, escores baixos, próximos de zero, são característicos dos genótipos, locais e anos que contribuíram pouco ou quase nada para a interação, caracterizando-se como estáveis (DUARTE; VENCOVSKY, 1999). Portanto, no *triplot*, serão considerados como estáveis os genótipos, os locais e os anos que estiverem próximos da origem, ou seja, com escores próximos de zero.

Assim, observando a Figura 9.8, nota-se que os genótipos 3, 6, 7 e 8, são os que estão mais próximos da origem, portanto são os genótipos estáveis, mas os genótipos 5, 9 e 12, são os que estão mais distantes da origem, portanto são estes genótipos que mais contribuem para a interação entre genótipos × locais × anos. Ainda, em relação aos genótipos pode-se construir alguns grupos de

genótipos que apresentam características semelhantes: grupos 1: genótipos 5 e 9, grupo 2: genótipos 2, 4 e 10 , grupo 3: genótipos 3, 6, 7 e 8, grupo 4: genótipos 1, 11 e 13 e grupo 5: genótipo 12.

Com relação aos locais, observa-se que o local 1 (município de Dourados na época das águas) é um local estável, pois está perto da origem e os locais 2 (município de Dourados na época das secas) e 3 (município de Aquidauana na época das secas) são os que contribuem para a interação, sendo que os anos não apresentam uma formação de grupos, indicando que cada um destes locais tem características próprias. Agora, ao observar os níveis do fator ano, percebe-se que os anos 1 (2000/2001) e 2 (2001/2002) contribuíram para a interação tripla e o ano 3 (2005/2006) foi um ano estável. Dentre os níveis deste fator também não houve formação de grupos, de modo que cada ano teve suas próprias características.

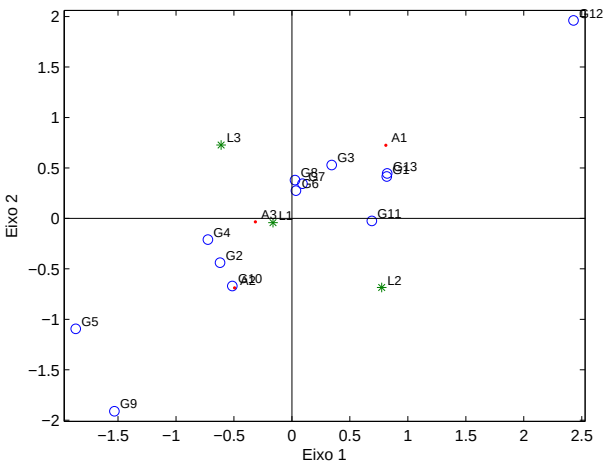

Figura 9.8: Triplot para os dados de produção de feijão (ton/ha)

Baseando-se na Figura 9.8, não é possível fazer uma avaliação da adaptabilidade dos genótipos. Assim, fez-se a combinação dos escores dos locais e anos, para avaliar a adaptação dos genótipos e os resultados são apresentados na Figura 9.9.

Pela Figura 9.9, nota-se que as combinações dos locais L3 (município de Dourados na época das secas) e L2 (município de Aquidauana na época das secas) com os anos A1 (2000/2001) e A2 (2001/2002), foram os efeitos que mais contribuíram para a interação. Com relação a adaptabilidade, percebe-se que os genótipos 11 e 12 apresentaram adaptação específica a combinação do local 2 e ano 1, pois os ângulos formado pelos vetores dos genótipos 11 e 12

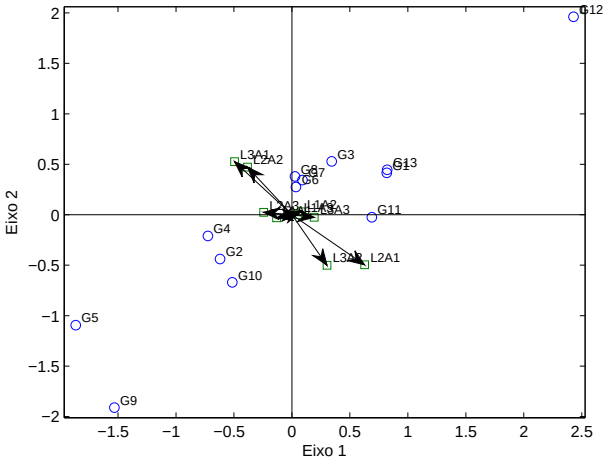

Figura 9.9: Triplot combinando os escores do locais e anos para avaliar a adaptabilidade dos genótipos às combinações de locais e anos.

com vetor do local 2 combinado com o ano 1 são agudos, ou seja, ao fazer a projeção dos genótipos no vetor da combinação L2A1, nota-se que eles estão na mesma direção e portanto têm interação tripla positiva (fato que é observado na Tabela 9.6). De forma análoga, a projeção do genótipo 5 na combinação L2A1 está em direção oposta, portanto a interação tripla é negativa.

9.6.4 Comentários Gerais

Ambas as metodologias utilizadas nos modelos AMMI de três entradas mostraram-se fáceis de ser aplicadas, principalmente na questão de esforço computacional, e os resultados mostraram que os modelos AMMI de três entradas são fáceis de serem interpretados. Mas ambas metodologias de três entradas apresentam vantagens e desvantagens.

Com relação ao modelo PARAFAC utilizado para modelar a interação tripla e consequentemente, utilizado para construir o *triplot*, pode-se citar como vantagem o fato de ser construído somente um gráfico, o que facilita organizar os resultados para os dados. Outra vantagem é que em um único gráfico é possível observar qual nível de cada um dos fatores contribuem e qual nível não contribuem para a interação. Mas, por outro lado, esta técnica apresenta algumas desvantagens, como é o caso de uma situação com um número muito elevado de genótipos, de ambientes e de anos, neste caso o *triplot* ficaria muito carregado o que dificultaria as conclusões, embora não impeça de tirar tais conclusões. Também pode-se usar pesos nos componentes de cada fator, para aumentar os

vetores no gráfico *triplot*, como sugerido por Galindo Villardón (1986) para o *biplot*. Outra solução seria acrescentar um terceiro eixo, resultando num *triplot* tridimensional, que poderia facilitar a visualização. Ainda como desvantagem pode-se citar que o modelo PARAFAC recuperou uma porcentagem menor da soma de quadrados da interação tripla, mas essa desvantagem pode ser questionada, pois não se sabe qual a verdadeira proporção de resposta padrão e qual a proporção de ruído dentro $SQ_{G\times L\times A}$.

Com relação ao outro modelo AMMI de três entradas que utilizou o modelo de Tucker3 para encontrar as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ e depois construir o *joint plot*, pode-se relatar a seguinte vantagem que é o fato do modelo de Tucker3 ter recuperado uma alta quantidade da soma de quadrados da interação genótipos $\times$ locais $\times$ anos, mas como foi citado para o modelo de PARAFAC, esta vantagem também pode ser questionada, pois não se sabe exatamente qual a verdadeira proporção de resposta padrão e a proporção de resposta que é ruído. Outra vantagem desta metodologia é que os *joint plot* ficam menos carregados, pois um dos fatores não é colocado no gráfico. Com relação as desvantagens, cita-se o fato de que o número de *joint plot* a ser construído é igual ao número de componentes que têm o fator que receberá a projeção e, portanto, o fator que receberá a projeção será aquela que tem o menor número de componentes, logo a medida que aumentar o número de *joint plot* ficará mais difícil agrupar as conclusões para o conjunto de dados. Outra desvantagem deste método é que não fica claro, no *joint plot*, a contribuição do fator que está recebendo a projeção do *joint plot* para a interação tripla, ou seja, visualmente não é possível saber se este fator têm uma contribuição alta ou uma contribuição baixa para a interação. Para solucionar este problema é necessário fazer a projeção sobre outro fator e, consequentemente, aumentará a dificuldade de organizar os resultados e tirar conclusões gerais sobre o conjunto de dados.

Referências Bibliográficas

ALLAN, F.E.; WISHART, J. A method of estimating the yield of a missing plot in field experimental work. In: DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley. 1985. chap. 5, p. 93-162.

ALTMAN, D.G. *Practical statistics for medical research*. London: Chapman; Hall, 1991. 611p.

ARAÚJO, L.B. *Seleção e análise dos modelos PARAFAC e Tucker e gráfico triplot com aplicação em interação tripla*. 2009. 111p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2009.

ARAÚJO, L.B. *Métodos de correção de autovalores e regressão isotônica nos modelos AMMI*. 2005. 75p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

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.

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. dos S. Imputação de dados em experimentos com interação genótipo por ambiente: Uma apliação a dados de algodão. *Revista Brasileira de Biometria*, São Paulo, v.27, n.1, p.125-138, 2009.

ARIAS, E. R. A. *Adaptabilidade e estabilidade das cultivares de milho avaliadas no Estado do Mato Grosso do Sul e avanço genético obtido no período de 1986/87 a 1993/94*. Tese de Doutorado, Universidade Federal de Lavras, Estatística e Experimentação Agropecuária, Lavras - MG. 1996.

BARTLETT, M.S. Some examples of statistical methods of research in agri-

culture and applied biology. In: DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley. 1985. chap. 5, p. 93-162.

BERGAMO, G.C. *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação*. 89p. 2007. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

BERGAMO, G.C.; DIAS, C.T. dos S.; KRZANOWSKI, W.J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, Piracicaba, v.65, n.4, p. 422-427, 2008.

BIRKES, D.; DODGE, Y.; SEELY, J. Spanning tests for estimable contrasts in classification models. *The Annals of Statistics*, Corvallis, v.4, n.1, p.86-107, 1976.

BROMAN, K.W.; SEN, S. *A Guide to QTL Mapping with R/qtl*. New York: Springer-Verlag, 2009.

CALIŃSKI, T.; CZAJKA, S.; DENIS, J.B.; KACZMAREK. Z. EM and ALS algorithms applied to estimation of missing data in series of variety trials. *Biuletyn Oceny Odmian*, Poznan, v.24-25, p.7-31, 1992.

CARPENTER, J.; BITHELL, J.Bootstrap Con?dence Intervals: When, which, what? A Practical Guide for Medical Statistician, *Statistics in Medicine*, London, v. 19, p.1141-1164, 2000.

CHAVES, J.L. Interação de cultivares com ambientes. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S.; VALADARES, M.C. *Recursos genéticos e melhoramento de plantas*. Rondonópolis: Fundação MT, 2001. p.673-713

CHAVES, L.J.; VENCOVSKY, R.; GERALDI, I.O. Modelo não linear aplicado ao estudo da interação de genótipos × ambientes em milho. *Pesquisa agropecuária Brasileira*, v.24, n.2, p. 259-269, 1989.

CHURCHILL, G.A.; DOERGE, R.W.: Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, v. 138, p. 963–971, 1994.

COCHRAN, W.G.; COX, G. *Experimental designs*. New York: John Wiley, 1957. 611p.

COCKERHAM, C.C. Estimation of genetics variance. In: HANSON, W.D.; ROBINSON, H.F. Ed. *Statistical genetics and plant breeding*. Madison: Na-

tional Academy of Sciences, 1963. chap. 2, p.53-94.

COELHO-BARROS, EMÍLIO AUGUSTO et al. Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. *Revista Colombiana de Estadística*, Bogotá, v.31, n.1, p.111-129, Jun 2008.

CORNELIUS, P. L.; CROSSA J.; SEYEDSADR M. S. *Tests and estimators of multiplicative models for variety trials*, 1993. In: DIAS, C. T. S.; KRZANOWSKI, W. J. *Model selection and cross-validation in additive main effect and multiplicative interaction (AMMI) models*. Crop Science, v.43, p.865-873, 2003.

CORNELIUS, P.L.; CROSSA J.; SEYEDSADR M.S. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: KANG, M.S.; GAUCH, H.G. *Genotype-by.environment interaction*. Boca Raton: CRC Press, 1996. chap. 8, p.199-234.

CORNELIUS, P. L.; CROSSA, J. Prediction assessment of shrinkage estimators of multiplicative models for multi-environment trials. *Crop Science*, Madison, v.39, p.998-1009, 1999.

CORNISH, E.A. The estimation of missing values in quasi-factorial designs. In: DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley, 1985. chap. 5, p. 93-162.

_____. The analysis of quasi factorial designs with incomplete data: lattice squares, 1941. In: DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley, 1985. chap. 5, p. 93-162.

COSTA, E. F. N.; SOUZA, J.C.; LIMA, J. L.; CARDOSO, G. A. Interação entre genótipos e ambientes em diferentes tipos de híbridos de milho. *Pesquisa Agropecuária Brasileira*, Brasília, v.45, n.12, p.1433-1440, 2010.

CROSSA, J.; GAUCH, H. G.; ZOBEL, R. W. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials, *Crop Science* 30(3), 493-500, 1990.

CROSSA, J.; FOX, P. N.; PFEIFER, W. H.; RAJARAM, S.; GAUCH, H. G. AMMI adjustment for statistical analysis of an international wheat yield trial, *Theoretical Applied of Genetics* 81, 27-37, 1991.

CROSSA, J. Statistical analyses of multilocation trials. In: DUARTE, J.B.; VENCOVSKY, R. *Interação genótipo × ambiente: uma introdução à análise*

"AMMI". Riberão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografías).

CRUZ, C.D.; REGAZZI, A.J. Modelos biométricos aplicados ao melhoramento genético. Viçosa: UFV, 1994. 390p.

CRUZ, C.D.; CARNEIRO, P.C.S. *Modelos biométricos aplicados ao melhoramento genético*. Viçosa: UFV, 2006. 585p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. *Modelos biométricos aplicados ao melhoramento genético*. 3ed. Viçosa: UFV, 2004. v.1 480p.

DAVISON, A.C.; HINKLEY, D.V. *Bootstrap Methods and their Application* Cambridge:. Cambridge University Press,. 1997. With 1 IBM-PC floppy disk (3.5 inch; HD). v.1: Cambridge Series in Statistical and Probabilistic Mathematics.

De MENDIBURU, F. *agricolae*: Statistical Procedures for Agricultural Research. R package version 1.0-7. Vienna, Austria: The R Foundation for Statistical Computing, 2009.

DENIS, J.B.; BARIL C.P. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. *Biuletyn Oceny Odmian*, Poznan, v.24-25, p.33-45, 1992.

DIAS, C.T.S. *Métodos para a escolha de componentes em modelo de efeito principal aditivo e interação multiplicativa*. 73 p. 2005. Tese (livre-docência no Departamento de Ciências Exatas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

DIAS, C.T.S.; KRZANOWSKI, W.J. Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Science*, Madison, v.43, p.865-873, 2003.

_____. Choosing components in the additive main efect and multiplicative interaction (AMMI) models. *Scientia Agricola*, Piracicaba, v.63, n.2, p.169-175, 2006.

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.

DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley,

1985. 499p.

DODGE, Y.; ZOPPE, A. Adjusting the EM algorithm for design of experiments with missing data. In: *International Conference on Information Technology Interfaces*, 26., 2004. Cavtat. Proceedings. Cavtat: s. ed, 2004. p.9-12.

DUARTE, J.B.; VENCOVSKY, R. *Interação genótipo × ambiente: uma introdução à análise "AMMI"*. Riberão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografías).

EASTMENT, H. T.; KRZANOWSKI, W. J. Cross-validatory choice of the number of components from a principal component analysis, *Technometrics* 24, 73-77, 1982.

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

_____. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, New York, n.82, p.171-185, 1987.

EFRON, B.; TIBSHIRANI, R. *An introduction to the bootstrap*. New York: Chapman; Hall, 1993. 436p.

EHLERS, R. S. Introdução a inferência Bayesiana. `http://www.leg.ufpr.br/~paulojus/CE227/ce227/`, 2003.

ENDERS, C.K. *Applied Missing Data Analysis*. Guilford Press, Inc.72 Spring Street, New York, 2010.

ENDERS, C.K. A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research. *Psychosomatic Medicine*, p.68:427-436, 2006.

EYHERABIDE, G. H.; ALVAREZ, M. P.; PRESELLO, D.; COLAZO, J. C.; DAMILANO, A.; FERNÁNDEZ, A. Estabilidad del rendimiento de cultivares de maíz en la área de la EEA Pergamino en el trienio 1994/95 - 1996/97, *Revista Tecnología Agropecuaria* INTA Pergamino may/ago, 51-54, 1997.

FALCONER, D.S. *Introduction to quantitative genetics*. Harlow: Longman, 1989, 438p.

FALCONER, D.S; MACKAY, T.F.C. *Introduction to quantitative genetics*. Harlow: Longman, 1996, 446p.

FARIAS, F.J.C. *Índice de seleção em cultivares de algodoeiro herbáceo*. 121p. 2005. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

FLORES, F.; MORENO, M. T.; A., M.; CUBERO, J. I. Genotype-environment interaction in faba bean: comparison of AMMI and principal coordinate models, *Field Crops Research* 47,117-127, 1996.

FREIRE FILHO, F. R.; RIBEIRO, V. Q.; ROCHA, M. M.; LOPES, Â. C. A. Adaptabilidade e estabilidade da produtividade de grãos de genótipos de caupi enramador de tegumento mulato. *Pesquisa Agropecuária Brasileira*, Brasília, v.38, n.5, p.591-598, 2003.

FURRER, R; NYCHKA, D.; SAIN, S. *fields*: Tools for Spatial Data. R package version 6.01, URL http://CRAN.R-project.org/package=fields. 2009.

GABRIEL, K. R. Le biplot-outil d'exploration de données multidimensionelles, *Journal de la Societe Francaise de Statistique* 143, 5-55, 2002.

GALINDO VILLARDÓN, M.P. Una alternativa de representación simultanea: HJ-Biplot. *Questiio*, Barcelona, v.10, p.13-23, 1986.

GARCÍA, P. M. *Análise dos modelos AMMI bivariados*, Dissertação Mestrado, Universidade de São Paulo, 2009.

GARCIA-PEÑA, M.; DIAS, C. T. S. Analysis of bivariate additive models with multiplicative interaction (AMMI). *Rev. Bras. Biom.*, São Paulo, v.27, n.4, p.586-602, 2009.

GAUCH, H.G.; RODRIGUES, P.C.; MUNKVOLD, J.D.; HEFFNER, E.L; SORRELS, M.: New Strategies for Detecting and Understanding QTL by Environment Interactions. *Crop Science*, v. 51, p. 96–113, 2011.

GAUCH, H.G.: Model Selection and Validation for Yield Trials with Interaction. *Biometrics*, v. 44, p. 705–715, 1988.

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

GAUCH, H.G. *MATMODEL version 3.0: Open source software for AMMI and related analyses*. Available at http://www.css.cornell.edu/staff /gauch. Crop and Soil Sci., Cornell Univ., Ithaca, NY., 2007.

GAUCH, H.G.; ZOBEL, R.W. Predictive and postdictive sucess of statistical analysis of yield trials. In: KANG, M.S.; GAUCH, H.G. *Genotype-by-environment interaction*, Boca Raton: CRC Press, 1996. chap. 8. p. 199-234.

_____. Imputing missing yield trial data. *Theoretical and Applied Genetics*, New York, v.79, p.753-761, 1990.

GELMAN, A., RUBIN, D. B., CARLIN, J., STERN, H. *Bayesian data analysis*. Chapman & Hall,London, 1995.

GODFREY, A.J.R; WOOD, G.R.; GANESALINGAM, S.; NICHOLS, M.A.; QIAO, C.G. Two-stage clustering in genotype-by-environment analyses with missing data. *Journal of Agricultural Science*, Cambridge, v.139, p.67-77, 2002.

GOLLOB, H.F. A statistical model which combines feature of factor analitic and analysis of variance techniques. *Psychometrika*, New York, v.33, p.73-115, 1968.

HARTLEY, H.O. A plan for programming analysis of variance for general purpose computers. *Biometrics*, Washington, v.12, n.2, p.110-122, 1956.

HE, Y. Missing Data Imputation for Tree-based Models. 2006. 81 p. Doctor of Philosophy in Statistics - University of carfornia, Los Angeles, 2006.

HEALY, M; WESTMACOTT, M. Missing values in experiments analyzed on automatic computers. In: LITTLE, R. J; RUBIN D.B. *Statistical analysis with missing data*. 2nd ed. New York: John Wiley, 2002. chap. 2, p.24-40.

HESTERBERG, T.; MOORE, D.S.; MONAGHAN, S.; CLIPSON, A.; EPSTEIN, R. Bootstrap methods and permutation tests. In: *The practice of business statistics: using data for decisions* (5th ed., pp. 14-1-14-70). New York: W.H. Freeman, 2003. cap. 18.

JIANG, C.; ZENG, Z.-B.: Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, v. 140, p. 1111–1127, 1995.

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

KANG, M.S.; MAGARI, R. New developments in selecting for phenotypic stability in crop breeding. In: DUARTE, J.B.; VENCOVSKY, R. *Interação genótipo × ambiente: uma introdução à análise "AMMI"*. Riberão Preto:

Sociedade Brasileira de Genética, 1999. 60p. (Série Monografías).

KARABATSOS, G. A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, v.1, p.152-176, 2000.

KENDALL, M. G.; STUART, M. A. *The advanced theory of statistics*. 4th ed. London: Charles Griffin, 1977. v.1.

KROONENBERG, P.M., Three-mode principal component analysis: illustrated with an example from attachment theory. In: LAW, E.G.; SNYDER, C.W.; HATTIE, J.A.; McDONALD, R.P. *Research Methods for Multimode Data Analysis*. New York: Praeger, 1984. p.64-103.

KROONENBERG, P.M. *Applied Multiway Data Analysis*. New Jersey: Wiley-Interscience, 2008. 579p.

KRUSKAL, J.B. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In: COPPI, R.; BOLASCO, S. *Multiway Data Analysis*. Amsterdam: Elsevier, 1989. p.8-18.

KRZANOWSKI, W. J. Cross-validation in principal component analysis, *Biometrics* 43, 575-584, 1987.

KRZANOWSKI, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. *Biometrical Letters*, [s.l.], v. 25, n. 1-2, p. 31-39, 1988.

KRZANOWSKI, W. J. *PPrinciples of multivariate analysis: A user's perspective*, Oxford University Press, 2000.

LAVORANTI, O.J. *Estabilidade e adaptabilidade fenotípica através da reamostragem "bootstrap"no modelo AMMI*. 2003. 166p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

LAVORANTI, O. J.; DIAS, C. T. S.; KRZANOWSKI, W. J. *Análise da divergência genética via modelo AMMI com reamostragem bootstrap*. In: 49ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, Uberlândia, v. 1. p. 16, 2004.

LEJEUNE, M.; CALIŃSKI, T. Canonical analysis applied to multivariate analysis of variance, *Journal of Multivariate Analysis* 72, 100-119, 2000.

LI, C.C. Analysis of unbalanced data. A pre-program introduction. In:

DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley. 1985. chap. 5, p. 93-162.

LIN, T. H. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Qual Quant*, p. 44:277-287, 2010.

LIU, G.; CORNELIUS, P. L. Simulations and derived approximations for the means and standard deviations of the characteristic roots of a wishart matrix. *Communications in Statistics - Simulation and Computation*, London, v.30, n.4, p.963-989. 2001.

LITTLE, R. J; RUBIN D.B. *Statistical analysis with missing data*. 2nd ed. New York: John Wiley, 2002. 381p.

MAIA, M. C. M.; VELLO, N. A.; ROCHA, M. de M.; PINHEIRO, J. B.; SILVA JÚNIOR, N. F. da. Adaptabilidade e estabilidade de linhagens experimentais de soja selecionadas para caracteres agronômicos através de método uni-multivariado. *Bragantia*, Campinas, v.65, p.215-226, 2006.

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

MANLY, B.F.J. *Randomization, bootstrap and Monte Carlo methods in biology*. 2nd ed. New York: Chapman; Hall, 1997. 399p

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. *Multivariate analysis*. Amsterdam: Academic Press. 2003. 521p.

MARTINEZ, E.Z.; LOUZADA-NETO, F. Estimação intervalar via bootstrap. *Revista de Matemática e Estatística*, São Paulo, v.19, p.217-251, 2001.

MEDINA, F.; GALVÁN, M. *Estúdios Estadísticos y prospectivos*. Imputación de datos: Teoría y Prática. División de Estadística y Proyecciones Económicas. Naciones Unidas- CEPAL, Santiago de Chile, p. 10-34, Julio de 2007.

MEYER, A.S. *Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes*. 2002. 106p (Dissertação de Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2002.

MILLIKEN G.A.; JOHNSON D.E. *Analysis of messy data*. New York: Chap-

man e Hall, 1989. v.2, 199p.

MOHAMMADI, R.; AMRI, A. Analysis of genotype x environment interactions for grain yield in durum wheat. *Crop Science*, Madison, v.49, n.4, p.1177-1186, 2009.

MOITA NETO, J. M.; MOITA, G.C. Uma Introdução à Análise Exploratória de Dados Multivariados. *Química Nova*, São Paulo, SP: v. 21, n. 4, p. 467-469, RAM, J. ; PANWAR, D.V.S. Intraspecific divergence in rice. *The Indian Journal of Genetics Plant Breeding*, New Delhi, v. 30, n. 1, p. 1-10, 1970.

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

MUIRHEAD, R.J. Developments in eigenvalue estimation. In: GUPTA, A.K. (Ed.) *Advances in multivariate statistical analysis*. Dordrecht: Reidel, 1987. p.277-288.

MUNKVOLD, J.D.; TANAKA, J.; BENSCHER, D.; SORRELS, M.E.: Mapping quantitative trait loci for preharvest sprouting resistance in white wheat. *Theoretical and Applied Genetics*, v. 119, p. 1223–1235, 2009.

NUNES, L. N. *Métodos de imputação de dados aplicados na área da saúde*. 2007. 120 p. Tese (Doutorado em Epidemiologia) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.

NUNES, G.H.S. *Interação genótipos × ambientes em eucalipto: implicações sobre a seleção e formas de atenuar seu efeito*. 2000. 160p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2000.

PEREIRA, A. d. S.; COSTA, D. M. d. Análise de estabilidade de produção de genótipos de batata no Rio Grande do Sul, *Pesquisa Agropecuária Brasileira* 33(4), 405-409, 1998.

PIEPHO, H. P. *Best linear unbiased prediction (BLUP) for regional yiels trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis*, 1994. In: In: DUARTE, J.B.; VENCOVSKY, R. *Interação genótipo × ambiente: uma introdução à análise "AMMI"*. Ribeirão Preto: Sociedade Brasileira de Genética, 1999. p.60 (Série Monografias, 9).

PIEPHO, H.P. Robustness of statistical test for multiplicative terms in the

additive main efects and multiplicative interaction model for cultivar trial. *Theoretical and Applied Genetics*, New York, v.90, p.438-443, 1995a.

_____. Methods for estimating missing genotype-location combinations in multilocation trials - an empirical comparison. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, Stuttgart, v.26, n.4, p.335-349, 1995b.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, 2010. ISBN 3-900051-07-0.

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.

REITER, J. P., RAGHUNATHAN, T. E. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, p. 102:1462-1471, 2007.

REIS, R.L., MUNIZ, J.A., SILVA, F. F., SÁFADI, T., AQUINO, L.H. Inferência Bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taca de fecundação cruzada. *Ciência Rural*, Santa Maria, v.39, n.6, p.1752-1759, ago, 2008.

REIS, R.L., MUNIZ, J.A., SILVA, F. F., SÁFADI, T., AQUINO, L.H. Abordagem bayesiana da sensitividade de modelos para o coeficiente de endogamia. *Ciência Rural*, Santa Maria, v.38, n.5, p.1258-1265, set, 2009.

ROCHA, M. M. *Seleção de linhagens experimentais de soja para adaptabilidade e estabilidade fenotípica*. 2002. 174 f. Tese (Doutorado em Genética e Melhora mento de Plantas) - Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2002.

ROMAGOSA, I.; ULLRICH, S.E.; HAN, F.; HAYES, P.M.: Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theoretical and Applied Genetics*, v. 93, p. 30–37, 1996.

ROMESBURG, H.C. *Cluster analysis for researchers*. Robert E. Krieger Publishing, 1990. Malabar, Florida. 334p.

MUÑOZ ROSAS, J. F.; ALVAREZ VERDEJO, E. Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus. *Revista*

de métodos cuantitativos para la enonomía y la empresa (7). p. 3-30. Junio de 2009.

ROSSI, R. M. *Introdução aos métodos Bayesianos na análise de dados zootécnicos com o uso do WinBUGS e R*. 191 p.:il., tabs., Maringá: Eduem, 2011.

RUBIN, D.B. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Applied Statistics, London*, v.21, n.2, p.136-141, 1972.

_____. Inference and missing data. *Biometrika*, Oxford, v.63, n. 3, p.581-592, 1976.

_____. *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, 258 p., New York, 1987.

_____. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, Alexandria, v.91, n.434, p.473-489, 1996.

SANCHEZ, E.; KOWALSKI, B.R. Tensorial resolution: a direct trilinear decomposition. *Journal of Chemometrics*. Chichester, v.4, p.29-45, 1990.

SAS INSTITUTE SAS/IML 9.1 User.s guide. Carey: *SAS Institute Inc.*, 2004. 1040p.

SCHAFER, J.L.; GRAHAM, J.W. Missing Data: Our View of the State of the Art. *Psychological Methods*, Vol.7, No. 2, p. 147-177, 2002.

SCHEPERS, J.; CEULEMANS, E.; VAN MECHELEN, I. Selecting among multi-mode partitioning models of different complexities: a comparison of four model selection criteria. *Journal of Classification*, New York, v.25, p.67-85, 2008.

TIBSHIRANI, R. Correction to Discussion of: Jackknife, Bootstrap and other Resampling Methods in Regression Analysis, *Annals of Statistics*, Madison, v.16, n.1, p.479, 1988.

TIMMERMAN, M. E.; KIERS, H.A.L. Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, London, v.53, n.1, p.1-16, 2000.

TUCKER, L. Some mathematical notes on three-mode factor analysis. *Psy-*

chometrika. New York, v.31, p.279-311, 1966.

VAN EEUWIJK, F.A.; KROONENBERG, P.M. Multiplicative models for interaction in three-way ANOVA, with applications to plant breeding. *Biometrics*, Oxford, v.54, n.4, p.1315-1333, 1998.

WANG, S.; BASTEN, C.J.; ZENG, Z.-B. *Windows QTL Cartographer 2.5.*. Department of Statistics, North Carolina State University, Raleigh, NC, 2007.

WANG, W.C.; CHEN, C.T. Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement*, v.65, p.376-404, 2005.

WARZECHA, T.; ADAMSKI, T.; KACZMAREK, Z.; SURMA, M.; CHEŁKOWSKI, J.; WIŚNIEWSKA, H.; KRYSTKOWIAK, K.; KUCZYŃSKA, A. Genotype-by-environment interaction of barley DH lines infected with Fusarium culmorum (W.G.Sm.)Sacc. *Field Crops Research*, Amsterdam, v.120, p.21-30, 2011.

WOLD, S. *Pattern recognition by means of disjoint principal component models*, 1976. In: DIAS, C. T. S.; KRZANOWSKI, W. J. *Model selection and cross-validation in additive main effect and multiplicative interaction (AMMI) models*. Crop Science, v.43, p.865-873, 2003.

WOLD, S. Cross-validatory estimation of the number of components in factor and principal component models, *Crop Science* 20, 397-405, 1978.

WU, C. F.J. Jackknife, Bootstrap and other Resampling Methods in Regression Analysis, *Annals of Statistics*, Madison, v.14, n.4, 1261-1350, 1986.

YAN, W.; HUNT L.A. Biplot analysis of multi-environment trail data. In: KANG, M.S *Quantitative Genetics, Genomics and Plant Breeding*. New York: CAB Publishing, 2002. p.289-303.

YAN, W.; HOLLAND, J.B. A Heritability-adjusted GGE biplot for test environment evaluation. *Euphytica*, Wageningen, v.171, n.3, p.355-369, 2010.

YAN, W.; KANG, M.S. *GGE Biplot analysis:* a graphical tool for breeders, geneticists, and agronomists. Flórida:Boca Raton, 2003.

YATES, F. The analysis of replicated experiments when the .eld results are incomplete. In: DODGE Y. *Analysis of experiments with missing data*. New

York: John Wiley, 1985. chap. 5, p. 93-162.