Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Seleção e análise dos modelos PARAFAC e Tucker e gráfico *triplot* com aplicação em interação tripla

Lúcio Borges de Araújo

Tese apresentada para obtenção do título de Doutor em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2009

Lúcio Borges de Araújo
Licenciado em Matemática

Seleção e análise dos modelos PARAFAC e Tucker e gráfico *triplot* com aplicação em interação tripla

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

Tese apresentada para obtenção do título de Doutor em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2009

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

Araújo, Lúcio Borges de
Seleção e análise dos modelos PARAFAC e Tucker e gráfico *triplot* com aplicação em interação tripla / Lúcio Borges de Araújo. - - Piracicaba, 2009.
111p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2009.
Bibliografia.

1. Análise de dados 2. Correlação genética e ambiental 3. Fenótipos 4. Genética estatística I. Título

CDD 519.5

A663s

DEDICATÓRIA

"*Sabemos que Deus age em todas as coisas para o bem daqueles que o amam,*
dos que foram chamados de acordo com o seu propósito."(Romanos 8:28)
"*Que diremos, pois, diante dessas coisas? Se Deus é por nós, quem será contra nós?*
Aquele que não poupou seu próprio Filho, mas o entregou por todos nós,
como não nós dará justamente com ele, e de graça, todas as coisas?"(Romanos 8:31,32)
"*Mas em todas estas coisas somos mais que vencedores,*
por meio daquele que nos amou "(Romanos 8:37)

A minha querida amiga, companheira e principalmente esposa
Mirian Fernandes Carvalho Araújo,
por toda ajuda, companheirismo, amizade, apoio, incentivo e o amor sempre constante.
Te amo muito!
Aos meus pais **Luís Guilherme de Araújo** e **Tânia Maria Borges Araújo**,
por todas oportunidades concedidas.
Aos meus irmãos **Gabriel, Aurélia e Evaldo**,
pela amizade e incentivo.

AGRADECIMENTOS

A Deus, autor e consumador da minha fé, por sua eterna fidelidade e sem o qual nada podemos fazer.

Ao professor Dr. Carlos Tadeu dos Santos Dias pela orientação cultivada pela amizade, apoio e ajuda à elaboração deste trabalho. A sua esposa Elvina e seus filhos Vitor e Laura, pelos momentos prazerosos passados juntos.

Ao Professor Mario Varela, pelas dicas, sugestões e correções na fase final do trabalho

A ESALQ/USP pela estrutura física e humana disponível.

A CNPq pela concessão de bolsas de estudos.

Aos amigos e irmãos em Cristo da Igreja Evangélica na Paulista pela amizade e apoio sempre presente em todos os momentos.

Ao professores do programa de Pós-graduação em Estatística e Experimentação Agronômica Dr. César Gonçalves de Lima, Dra. Clarice Garcia Borges Demétrio, Dr. Décio Barbin, Dr. Edwin Ortega, Dra. Roseli Aparecida Leandro, Dr. Sílvio Zocchi, Dra. Sônia Maria De Stefano Piedade, Dr. Vitor Ozaki, pelos cuidados na formação.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, Solange de Assis Paes Sabadin e Eduardo Bonilha, pelos auxílios permanentes, em especial a Luciane Brajão pela amizade e ajuda sempre que preciso.

Aos amigos do Departamento de Bioestatística da UNESP/BOTUCATU.

Aos ex-professores e atuais colegas de departamento da UFU, em especial ao prof. Ednaldo Carvalho Guimarães e ao prof. Marcelo Tavares.

Aos grandes amigos: Osmar Jesus Macedo, pela alegria de sua companhia e colaboração em tantos momentos, e César Augusto Taconeli, pelos grandes momentos de alegria em Botucatu.

Aos colegas de turma: Ana Alice, Angela, Édila, Vanderly e Wilson e a todos os outros colegas do mestrado e doutorado, em especial, ao Marcelino "Popó" pelo companheirismo de longa data.

A todos que cooperaram direta ou indiretamente na realização deste trabalho, muito obrigado.

SUMÁRIO

RESUMO 8
ABSTRACT 9
LISTA DE FIGURAS 10
LISTA DE TABELAS 11
1 INTRODUÇÃO 12
2 REVISÃO DE LITERATURA 15
2.1 Interação Genótipos × Locais × Anos 15
2.1.1 Graus de interação 15
2.1.2 Avaliação da Interação Genótipos × Ambientes 16
2.1.3 Modelos de ANOVA 19
2.1.4 Fatores Genótipo, Local, Tempo 19
2.2 O que é análise *multiway*? 23
2.2.1 Linhas, Colunas e Tubos; Fatia Fontal, Vertical e Horizontal 23
2.2.2 História dos modelos de análise *multiway* 23
2.2.3 Modelos de componentes com três entradas (PARAFAC) 25
2.2.4 Modelos de Tucker 28
2.2.4.1 Modelos Tucker3 29
2.2.4.1.1 Propriedades do modelo Tucker3 31
2.2.4.2 Modelos de Tucker2 32
2.2.4.3 Modelos de Tucker1 33
2.2.5 Relações entre modelos de componentes de três entradas 34
2.2.5.1 Hierarquia dos Modelos PARAFAC e TUCKER3 35
2.2.5.2 Hierarquia dos Modelos TUCKER3, TUCKER2 e TUCKER1 36
2.3 Graus de liberdade dos modelos *multiway* 37
2.4 Postos de arranjos 38
2.5 Determinação da dimensionalidade de um modelo de Tucker 39
2.5.1 Procedimento $DifFit$ de Timmerman-Kiers 39
2.5.2 Análise residual 40
2.5.3 Critério *st* de Ceulemans-Kiers 41

2.6 Determinação da dimensionalidade de um modelo PARAFAC 42
2.6.1 Procedimentos de dividir ao meio (*Split-half*) 43
2.6.2 Consistência do núcleo 43
2.7 Estabilidade do modelo e poder preditivo por validação 44
2.8 Biplot 46
2.8.1 Decomposição em Valores Singulares 46
2.8.2 *Biplot* padrão 47
2.9 *Joint plot* 49
3 MATERIAL E MÉTODOS 52
3.1 Características dos dados 52
3.2 Análise de dados considerando duas entradas 53
3.2.1 Análise de variância conjunta de duas entradas 53
3.2.2 Análises AMMI 54
3.3 Estimação dos parâmetros dos modelos *multiway* 57
3.3.1 Algoritmo para o modelo PARAFAC 58
3.3.2 Estimativas iniciais para Algoritmo MQA do modelo PARAFAC 59
3.3.3 Algoritmo para o modelo Tucker3 60
3.3.4 Estimativas iniciais para o Algoritmo MQA do modelo Tucker3 62
3.4 Proposta para o *triplot* 62
3.4.1 Produto de elementos por elementos de matrizes 62
3.4.2 Arranjo de três entradas em um gráfico de duas dimensões 63
3.4.3 O produto dos elementos das matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ e suas propriedades 64
3.5 Visualizando o *triplot* 68
3.5.1 Comparação visual dos elementos de uma linha, coluna ou tubo do arranjo . . . 68
3.6 Relações entre linhas, entre colunas e entre tubos 69
3.7 Análise *triplot* de dados de três entradas 70
3.8 Análise de dados considerando três entradas 71
3.8.1 Análise de variância conjunta 71
3.8.2 Generalização da Análises AMMI para o caso de três fatores usando o modelo PARAFAC 73

3.9 Software 76
4 RESULTADOS E DISCUSSÕES 77
4.1 Análise de variância conjunta com dois fatores 77
4.2 Análise AMMI e Biplot para dados de duas entradas 78
4.3 Análise de variância conjunta com três fatores 83
4.4 Modelos de três entradas para a interação tripla 84
4.4.1 Ajuste do Modelo de Tucker3 85
4.4.2 Ajuste do Modelo PARAFAC 92
4.5 Triplot 93
4.6 Comentários Gerais 97
5 CONCLUSÕES 100
REFERÊNCIAS 102
ANEXOS 107

RESUMO

Seleção e análise dos modelos PARAFAC e Tucker e gráfico *triplot* com aplicação em interação tripla

O presente trabalho tem os seguintes objetivos: propor uma sistemática para o estudo e a interpretação da estabilidade e adaptabilidade fenotípica, através de duas técnicas de análise *multiway* (PARAFAC e Tucker3); propor a construção de um gráfico, denominado de Triplot, que possibilita avaliar as relações entre os 3 modos (genótipos, locais e anos); implementar uma rotina computacional para a análise de dados, segundo os modelos *multiway*; implementar uma rotina computacional para a construção do Triplot. 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, pelos municípios de Dourados e Aquidauana, 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. Os resultados indicaram que o gráfico *triplot* e *joint plot*, facilitam o entendimento da interação tripla e traz ao pesquisador informações mais reais sobre a interação tripla, do que a modelagem AMMI de duas entradas; o gráfico *triplot*, ajuda a identificar genótipos, locais e anos estáveis, dentro de um grande grupo de genótipos, locais e anos; de uma maneira geral recomenda-se, utilizar o *triplot* e o *joint plot* juntos, para obter melhores interpretações dos resultados; dentre os genótipos estudados, o genótipo 6 é o que menos contribui para a interação e o os genótipos 12, 9 e 5 são os que mais contribuem para a interação.

Palavras-chaves: Interação genótipos × ambientes × anos; Modelo PARAFAC; Modelo Tucker3; Triplot; Estabilidade; Adaptabilidade

ABSTRACT

Selection and analysis of the PARAFAC and Tucker models and triplot graphic with application in triple interaction

The present work has the following objectives: to propose a systematics for the study and the interpretation of the phenotypic stability and adaptability, through several multiway models (PARAFAC and Tucker3); to propose a graphic, called of Triplot, that it makes possible to evaluate the relations between the 3 ways (genotypes, locations and years); to implement a computational routine for the data analysis, according multiway models; to implement a computational routine for the construction of Triplot. The used data are relative the experiments with 13 genotypes of beans that had been lead in 9 experimental distinct ones constituted by agricultural years of 2000/2001, 2001/2002 and 2005/2006, by Dourados and Aquidauana cities, where the experiments had been installed at the time of waters (Dourados) and also at the time of dries (Dourados and Aquidauana). Each location is constituted of city and time of installation. The results indicated that the graphic triplot and joint plot, facilitate the agreement of triple interaction and bring to the researcher more real information about triple interaction, of what AMMI model of two way; the graphic triplot, helps to identify stabels genotypes, locations and years, inside of a great group of genotypes, location and years; in a general recommend to use triplot and joint plot together, to get better interpretations of the results; the genotype 6 is what less contributes for the triple interaction and genotypes 12, 9 and 5 are the that more contribute for the interaction.

Keywords: Genotypes × locations × years interaction; PARAFAC model; Tucker3 model; Triplot; Stability; Adaptability

LISTA DE FIGURAS

Figura 1 - Particionando um arranjo de três entradas em fatias (arranjos de duas entradas) 24
Figura 2 - Decomposição de um arranjo de três entradas propostas por Harshman (1970) e Carrol e Chang (1970) 25
Figura 3 - O modelo PARAFAC com R components 27
Figura 4 - Representação gráfica do modelo Tucker3 30
Figura 5 - Representação gráfica do modelo Tucker2 33
Figura 6 - Representação gráfica do modelo Tucker1, em que somente o primeiro modo é reduzido 34
Figura 7 - Modelo PARAFAC escrito como um modelo de Tucker3 35
Figura 8 - Representação de dois marcadores de objetos e um marcador de variáveis em um *biplot* 48
Figura 9 - 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 $\underline{\mathbf{Z}}$ 65
Figura 10 - Os marcadores das linhas, colunas, tubos e combinação de uma coluna com um tubo do arranjo $\underline{\mathbf{Z}}$ 66
Figura 11 - *Biplot* para os dados de produção de feijão (ton/ha), com 13 genótipos e 9 ambientes 82
Figura 12 - *Joint plot* projetado dentro da primeira componente do terceiro modo . . 89
Figura 13 - *Joint plot* projetado dentro da segunda componente do terceiro modo . . . 90
Figura 14 - *Scree plot* do número de componentes no modelo PARAFAC e a porcentagem da soma de quadrados explicada pelo modelo 93
Figura 15 - Triplot para os dados de produção de feijão (ton/ha) 95
Figura 16 - Triplot combinando os escores do locais e anos para avaliar a adaptabilidade dos genótipos às combinações de locais e anos. 96

LISTA DE TABELAS

Tabela 1 - Caracterização dos ambientes experimentais 52
Tabela 2 - Esquema da análise de variância para experimentos de um mesmo grupo de g genótipos avaliado em e locais com b blocos 54
Tabela 3 - As matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ para gerar $\underline{\mathbf{Z}}$ 64
Tabela 4 - Elementos do arranjo $\underline{\mathbf{Z}}$ matrizado combinado as colunas tubos 64
Tabela 5 - Esquema da análise de variância para experimentos de um mesmo grupo de genótipos avaliados em l locais e a anos com b blocos 73
Tabela 6 - Análise de variância conjunta para um conjunto de dados com 13 genótipos avaliados em 9 ambientes com 3 blocos 77
Tabela 7 - Médias dos genótipos, ambientes e posição das médias em relação a produtividade 78
Tabela 8 - Valores estimados da interação dupla de 13 genótipos e 9 ambientes (combinação de 3 locais e 3 anos) para a produção em ton/ha 79
Tabela 9 - Teste F_r de Cornelius para determinar o número de termos significativos para a interação $G \times E$ 81
Tabela 10 - Análise de variância conjunta para um conjunto de dados com 13 genótipos avaliados em 3 locais, 3 anos com 3 blocos 83
Tabela 11 - Efeitos da interação tripla para cada combinação de genótipos, locais e anos 84
Tabela 12 - Resultado do procedimento de Timmerman-Kiers para selecionar o modelo de Tucker3 86
Tabela 13 - Escores dos componentes principais para um modelo de Tucker3 (3,2,2) para o arranjo da interação tripla entre genótipos × locais × anos 88
Tabela 14 - Número de componentes utilizado no modelo PARAFAC e a porcentagem da soma de quadrados da interação tripla explicada pelo modelo 92
Tabela 15 - 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. 94

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. Os METs são essenciais porque a presença da interação entre genótipos e ambientes (GE), ou seja, a mudança na performance relativa de genótipos através de diferentes ambientes, complica a avaliação de cultivar. Quando não existe a interação GE, um único cultivar prevaleceria no mundo inteiro e um único experimento bastaria para avaliação de cultivar (Gauch e Zobel, 1996). A interação GE constitui o principal desafio na melhora de cultivares, e a análise de dados provenientes de experimentos multi-ambientais constitui um aspecto importante para o melhoramento genético de plantas. Por isso, melhorias nos métodos usados para análise de dados deve ser de interesse à comunidade de melhoristas.

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). A validez desta prática é baseada normalmente em suposições não declaradas de que os ambientes dos experimentos multi-ambientais pertencem a um único mega-ambiente, que é definido como um grupo de locais no qual um mesmo conjunto de cultivares apresenta-se melhor por vários anos. Normalmente, não se afirma isso, mas a avaliação de cultivar sempre é específica para separar mega-ambientes. Se os ambientes de teste forem suficientemente heterogêneos, os cultivares que são selecionados baseado em rendimento podem não ser o melhor em alguns dos ambientes de teste e, em casos extremos, podem não estar entre os melhores em quaisquer dos ambientes. Assim, a segunda utilidade de análise de dados multi-ambientais, antes de fazer a avaliação de cultivares, 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 de experimentos agrícolas,

duas classes de modelos são comumente utilizadas: modelos lineares e modelos lineares-bilineares. A 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 que, neste caso, cada entrada se refere a genótipos, locais e 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 responder com 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 $\times$ locais $\times$ 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 $\times$ 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.

Na maioria dos estudos, devido a falta de uma ferramenta adequada para estudar a interação entre genótipos, locais e anos, os pesquisadores combinam os fatores locais e anos. Mas de acordo com Varela et. al.(2006), em alguns casos, esta combinação leva a uma perda de informação quando se ajusta um modelo de duas entradas (por exemplo, os modelos AMMI) e quando se faz a estimação dos efeitos da interação. Então, faz-se necessário desenvolver ferramentas que permitam o desdobramento e a interpretação da interação tripla através dos modelos Tucker3 e PARAFAC.

Assim, o presente trabalho tem os seguintes objetivos: propor uma sistemática para o estudo e a interpretação da estabilidade e adaptabilidade fenotípica, através de duas técnicas de análise *multiway*; propor a construção de um gráfico, denominado de Triplot, que possibilita avaliar as relações entre os três modos (genótipos, ambientes e anos); implementar uma rotina computacional para a análise de dados, segundo os modelos *multiway*; implementar uma rotina computacional para a construção do Triplot.

2 REVISÃO DE LITERATURA

2.1 Interação Genótipos × Locais × Anos

Os genótipos analisados em um experimento são avaliados sob uma grande variedade de condições. Eles são expostos a diferentes tipos de solos, níveis de fertilidade, temperaturas e práticas culturais, sendo que estas características são encontradas em uma lavoura e podem ser descritas como um ambiente (DAS, 2005).

Quando os genótipos são comparados em diferentes ambientes, suas performances relativas em outros ambientes podem não ser as mesmas. Um genótipo pode ter alta produção em um ambiente e um segundo genótipo pode ser superior em outros ambientes. Mudança na performance relativa de genótipos através de diferentes ambientes é referida como interação genótipos × ambientes (G × E).

2.1.1 Graus de interação

Todo fator que é parte da planta tem um potencial para causar diferentes performances que é associado com a interação genótipos × ambientes. Variáveis ambientais podem ser classificadas como fatores previsíveis e não previsíveis (ALLARD; BRADSHAW, 1964). Fatores previsíveis são aqueles que ocorrem de maneira sistemática ou ocorre sob controle humano, tais como tipos de solo, data de plantio, espaçamento de linhas, quantidades de nutrientes, profundidade de semeadura etc. Fatores não previsíveis são aqueles que variam de maneira não sistemática, incluindo chuva, temperatura, umidade relativa etc.

Os fatores previsíveis podem ser avaliados individualmente e coletivamente de acordo com suas interações com genótipos (ALLARD; BRADSHAW, 1964). Estudos têm sido feitos para estudar as seguintes interações: genótipos × tipos de solo, genótipos × espaçamentos entre linhas, genótipos × datas de semeadura etc.

Fatores não previsíveis contribuem para a interação entre os genótipos, locais e anos. As interações genótipos × locais (G × L), genótipos × anos (G × A) e genótipos × locais × anos (G × L × A) têm sido avaliadas para diversos genótipos (DAS, 2005).

A performance relativa dos genótipos através dos ambientes, determina a importância de estudar e interpretar uma interação. Não existe interação genótipos × ambientes

quando a performance relativa entre genótipos não muda através de ambientes (CHAVES, 2001).

2.1.2 Avaliação da Interação Genótipos × Ambientes

Fazer uma avaliação da importância da interação genótipos × ambientes requer procedimentos experimentais apropriados. Assim, a compreensão dos passos envolvidos no delineamento, condução, análise e interpretação de tal experimento pode ser útil nesta avaliação.

Os genótipos escolhidos para uma avaliação de uma possível interação é uma importante consideração no delineamento do experimento. Algumas análises da interação genótipos × ambientes não são baseadas em um experimento especificamente delineado para tal proposta, particularmente a avaliação da interação com locais e anos. Ao invés disso, os pesquisadores utilizam dados de genótipos e linhas experimentais que foram avaliados através de locais e anos como parte de um programa com outra finalidade. A principal desvantagem deste procedimento é que os genótipos e experimentos podem não ser uma amostra aleatória de genótipos avaliados. A estimativa da interação genótipos × ambientes obtida com genótipos selecionados pode ser mais alta ou mais baixa do que o valor obtido com indivíduos aleatórios. Mas é preferível usar uma amostra aleatória dos genótipos que estão disponíveis para teste, pois neste caso a quantidade de objetivos do estudo é maior do que considerar os genótipos fixos (ANNICHIARICO, 2002).

Um teste deve ser conduzido em dois ou mais locais e anos para obter estimativas dos efeitos das interações G × L, G × A, G × L × A. Os locais de teste são geralmente aqueles rotineiramente utilizados pelos pesquisadores. Os locais podem ser considerados como efeito fixos quando eles não são aleatoriamente escolhidos de todos possíveis locais na área. Alguns pesquisadores consideram locais como um efeito aleatório, pois os pesquisadores não tem controle sobre as condições climáticas que vão ocorrer nos locais em qualquer ano. Pela mesma razão, anos são considerados como efeitos aleatórios.

Pelo menos duas repetições são necessárias em cada local e em cada ano para obter uma estimativa do erro experimental com o qual testa-se a significância da interação de interesse. Qualquer repetição adicional vai fornecer uma estimativa mais realista do erro

experimental.

A interpretação dos dados inclui considerar as significâncias estatísticas de cada fonte de variação e fazer uma avaliação da importância prática da variação observada entre os valores médios. A interação G×L mede a consistência da performance entre genótipos em diferentes locais. A consistência do desempenho dos genótipos em diferentes anos é indicada pela interação G×A. A interação G×L×A avalia a consistência entre genótipos para cada combinação de ano e locais. Um experimento conduzido em dois locais em dois anos tem 4 combinações entre locais × anos: local 1× ano 1; local 1 × ano 2; local 2 × ano 1; e local 2 × ano 2. A significância da interação G×L×A indica que a performance relativa entre genótipos não foi a mesma em cada combinação de anos e locais. Para todas as interações mencionadas anteriormente, um exame dos valores médios é necessário para determinar se uma interação significativa é devido a mudanças na classificação entre genótipos ou a mudanças nas diferenças entre os genótipos mas, sem ocorrer uma variação na classificação.

A falta de qualquer interação estatisticamente significativa que envolva genótipos, simplifica os testes requeridos pelo programa de melhoramento para desenvolver e selecionar genótipos. Teoricamente, esta falta de significância da interação de genótipos com locais, anos ou com a combinação locais e anos, indica que um teste em local durante um ano poderia ser suficiente para identificar genótipos com potencial genético superior. Genótipo com a melhor performance em determinado local e ano poderia também ser superior a outros locais e anos.

As implicações práticas de uma interação G×E estatisticamente significante depende das causas da interação. Interações G×E não são problemas para os pesquisadores, se estas não são devidas as mudanças na classificação de performance entre os genótipos. Sob estas circunstâncias, um teste em um local determina que certo ano poderia ser usado para identificar os genótipos superiores. O mesmo genótipo poderia ser superior em todos locais e anos, embora as magnitudes de superioridade variassem. Interações G×E significantes, que envolvem mudanças na classificação da performance entre genótipos, são comuns e para determinar a interpretação prática das interações, os pesquisadores deveriam considerar a extensão das mudanças na classificação e seus impactos no melhoramento genético. Julgamentos subjetivos, devem ser feitos e, além disso, outros pesquisadores devem avaliar

o mesmo conjunto de dados para verificar se as formas de agir são as mesmas ou não. As opções disponíveis para os pesquisadores são diferentes para cada tipo de interação (DAS, 2005):

i) G × L: Amplas flutuações nas posições de performance dos genótipos nos locais sugerem que pode ser desejável desenvolver genótipos para diferentes locais por meio de programas e teste de seleção independentes. A perda ao estabelecer programas independentes para diferentes áreas geográficas é substancial, por essa razão, a decisão pode ser difícil. Antes de estabelecer programas de melhoramento independentes, o pesquisador poderia determinar a interação G × L. Se as diferenças entre os locais são devido ao tipo de solo ou a outros fatores que são consistentes de ano para ano, programas de melhoramento independentes podem ser mais apropriados. Diferenças temporais entre locais associadas com condições climáticas não usuais não podem justificar programas independentes.

ii) G × A: Uma ordenação inconsistente entre os genótipos cultivados em diferentes anos é, em algumas situações, mais difícil de tratar que uma interação G × L. Um pesquisador não deve pensar na opção de estabelecer programas de melhoramento independentes para diferentes anos. Uma opção primária disponível é identificar os genótipos que apresentaram uma performance superior através dos anos. Isto envolve os genótipos em vários anos antes da seleção para lançar um genótipo como uma cultivar. Para reduzir o tempo gasto com a melhoria genética, múltiplos locais em um ano, são usados como substituto para os anos. A substituição é efetiva somente quando as divergências nas condições climáticas entre os locais são comparáveis às diferenças entre os anos.

iii) G × L × A: Quando existem modificações nas posições dos genótipos associados com a combinação individual de uma interação local × ano, o pesquisador deve identificar genótipos com performances médias superiores sobre os locais e anos. Por exemplo, uma análise da interação genótipos × ambientes para a produção de tabaco na Carolina do Norte indicou que as interações G × A e G × L não foram significantes (JONES; MATZINGER; COLLINS, 1960), a classificação (posições relativas) entre os genótipos foram similares quando avaliados sobre os locais e as posições relativas dos cultivares

também foram similares quando avaliados sobre os anos. Mas a interação G × L × A foi significativa no experimento. A interação parece estar associada à condição de combinação específica, tais como padrão de chuva e infestação de doença, que provocou uma variação na classificação dos genótipos entre certas combinações de locais × anos. Se um genótipo com uma alta performance média sobre os anos, é escolhido, espera-se que este genótipo tenha uma performance satisfatória no próximo ano, mas este pode não ser o melhor naquela particular época.

2.1.3 Modelos de ANOVA

Considere um experimento fatorial com três fatores: genótipos, locais e anos. Suponha ainda que cada um de "g" genótipos foi avaliado em "l" locais e em "a" anos. Com o objetivo de analisar os dados, análises de variâncias (ANOVAs) preliminares para experimentos individuais podem ser levadas em conta para avaliar a variação entre ambientes pelo erro experimental e, possivelmente, variância genotípica. ANOVA conjunta para um grupo de experimentos ou seus subconjuntos podem ser executadas com objetivos diferentes, como:

i) verificar a ocorrência de efeitos diferentes (isto é, significância das fontes de variação);

ii) estimar e comparar médias para níveis de fatores fixos (em particular, média dos genótipos através de regiões ou dentro de sub-regiões);

iii) estimar os componentes de variância genotípicos e genótipo-ambiental.

A ANOVA também pode representar um passo na análise de adaptação ou na avaliação de medidas de estabilidade do rendimento.

2.1.4 Fatores Genótipo, Local, Tempo

Além dos fatores de genótipo e possível de bloco, a ANOVA conjunta pode incluir o fator local e/ou também um fator de tempo, como ano (colheitas anuais) ou ciclo de colheita, por exemplo, culturas perenes. Alternativamente, poderia haver um fator ambiental, para o qual os níveis são representados através de experimentos individuais. Alguns modelos

de ANOVA também podem incluir um fator de sub-região (seguindo a definição de mega-ambientes) e/ou um fator de grupo de germoplasma que pode coincidir com acúmulo de genes distintos, variedades ou material com padrões de adaptação contrastantes.

Os modelos de ANOVA podem diferir em termos do número e tipo de fatores, como também a relação entre fatores. Em particular, o fator ano pode ser cruzado, ou aninhado dentro do fator local. Um elemento adicional de distinção entre modelos surge da definição de cada fator como aleatório ou fixo. Em geral, aleatoriedade implica que aqueles níveis de fatores são aleatoriamente escolhidos de uma população, para a qual as conclusões para fatores fixos são estendidas, sendo que a extensão de variação é de preocupação primária. Esta definição pode ser aplicada ao fator tempo. O fator de genótipo pode ser definido como aleatório ou fixo, dependendo do objetivo da análise, isto é, quando o objetivo for apoiar decisões relativas a elementos de uma estratégia de melhoramento estimando: componentes de variância, parâmetros genéticos, ganhos genéticos esperados de diferentes estratégias de adaptação ou procedimentos de seleção, etc. Neste caso, os genótipos deveriam ser representativos da base genética, consequentemente o efeito do fator genótipo é aleatório. Reciprocamente, genótipo é um fator fixo quando a ênfase estiver na comparação de material testado para seleção ou recomendação (ANNICHIARICO, 2002).

O fator de local definitivamente é aleatório quando o interesse principal da análise está na estimação de componentes de variância (WRICKE; WEBER, 1986) para locais que são representativos de uma população relevante dentro da região designada. A escolha entre aleatório e fixo pode ser problemática para locais próximos que são investigados por semelhança do efeito da interação de G $\times$ L e da possível identificação de sub-regiões relativamente uniforme por pesquisadores ou por propostas de recomendação, sendo que o efeito aleatório normalmente é o mais apropriado. O local pode ser considerado fixo somente se cada local representar uma área bem definida com manejo da safra relativa, então resultados para um determinado local podem ser estendidos para a área que representa. O fator de ambiente é normalmente aleatório. Finalmente, os grupos de fatores sub-regiões e o germoplasma, se presentes, são considerados fixos.

Três grupos principais de modelos de ANOVA parcialmente hierárquicos são considerados para a análise de conjuntos dos experimentos dispostos em um delineamento

em blocos aleatorizados (ANNICHIARICO, 2002). O primeiro grupo inclui modelos com três fatores: genótipo; local ou ambiente; e bloco dentro de locais ou ambientes.

A resposta observada Y_{ijr} do i-ésimo genótipo no j-ésimo local e r-ésimo bloco é:

$$Y_{ijr} = \mu + g_i + l_j + b_r(l_j) + (gl)_{ij} + \varepsilon_{ijr} \quad (1)$$

em 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 $j = 1, 2, ..., l$;

$(gl)_{ij}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo local;

$b_r(l_j)$: é o efeito do r-ésimo bloco dentro do j-ésimo local, com $r = 1, 2, ..., b$;

ε_{ijr}: é o erro experimental associado ao i-ésimo genótipo, no j-ésimo ambiente e no r-ésimo bloco assumido ser independente e $\varepsilon_{ijr} \sim N(0, \sigma^2)$.

Este modelo é útil para análise de adaptação baseado em experimentos que não estão repetidos no tempo, como freqüentemente é o caso para os genótipos de culturas perenes.

O segundo grupo de modelos de ANOVA inclui quatro fatores: genótipo, local, ano (ou outro fator de tempo) cruzado com o fator local e bloco dentro de locais dentro de anos.

A resposta Y_{ijkr} do genótipo i no local j, ano k e bloco r é:

$$Y_{ijkr} = \mu + g_i + l_j + a_k + b_r(l_j(a_k)) + (gl)_{ij} + (ga)_{ik} + (la)_{jk} + (gla)_{ijk} + \varepsilon_{ijrk} \quad (2)$$

em 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 k-ésimo ano dentro do j-ésimo local, 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 $\varepsilon_{ijrk} \sim N(0, \sigma^2)$.

O terceiro grupo de modelos inclui os mesmos fatores do segundo grupo, mas o fator tempo é aninhado em local. A resposta de Y_{ijkr} é:

$$Y_{ijkr} = \mu + g_i + l_j + a_k(l_j) + b_r(l_j(a_k)) + (gl)_{ij} + (ga)_{ik}(l_j) + \varepsilon_{ijrk} \tag{3}$$

em 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 $j = 1, 2, ..., l$;

$a_k(l_j)$: é o efeito do k-ésimo ano dentro do j-ésimo local, com $k = 1, 2, ..., a$.

$b_r(l_j(a_k))$: é o efeito do r-ésimo bloco dentro do k-ésimo ano dentro do j-ésimo local, com $r = 1, 2, ..., b$;

$(gl)_{ij}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo local;

$(ga)_{ik}(l_j)$: é o efeito da interação do i-ésimo genótipo com o k-ésimo ano dentro do j-ésimo local;

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 $\varepsilon_{ijrk} \sim N(0, \sigma^2)$.

Esse tipo de ANOVA é particularmente útil quando locais diferem ao longo dos anos, embora também possa ser usado como uma alternativa ao procedimento exposto anteriormente, isto é, para testar anos através dos locais.

2.2 O que é análise *multiway*?

Análises *multiway* é a análise de dados que envolve vários fatores. Suponha um experimento no qual foram analisados "g" genótipos em "l" locais, assim os resultados podem ser organizados em uma tabela de duas entradas (ou matriz) de dimensão $g \times l$. Tomando "a" dessas medidas, por exemplo em "a" anos diferentes, os dados podem ser organizados em um arranjo cúbico de dimensões $g \times l \times a$.

2.2.1 Linhas, Colunas e Tubos; Fatia Fontal, Vertical e Horizontal

Para os arranjos de duas entradas é usual fazer uma distinção entre as partes especiais do arranjo, como linhas e colunas. Esta distinção também é feita para arranjos de três entradas e uma divisão adequada é: fatias frontais, horizontais e verticais. Existem três tipos diferentes de fatiar um arranjo de três dimensões $\underline{\mathbf{X}}$ $(I \times J \times K)$, em que "_" indica que $\boldsymbol{X}$ é um arranjo com 3 entradas. O primeiro tipo origina as fatias horizontais $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_I}$, todas de dimensões $(J \times K)$. O segundo origina as fatias verticais $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_J}$, todas tem dimensões $(I \times K)$. E o último tipo origina as fatias frontais $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_K}$, em que todas tem dimensões $(I \times J)$. Esta notação é conveniente, mas nem sempre clara, por exemplo, não é conhecido se $\boldsymbol{X_2}$ é a segunda fatia frontal, vertical ou horizontal. Para evitar esta ambigüidade, pode-se usar, por exemplo, $\boldsymbol{X_{i=2}}$ para a primeira entrada. A Figura 1, ilustra esses casos

2.2.2 História dos modelos de análise *multiway*

A maioria dos trabalhos de análises *multiway* são provenientes da psicometria (registro e medida da atividade intelectual). Os trabalhos pioneiros apareceram na metade

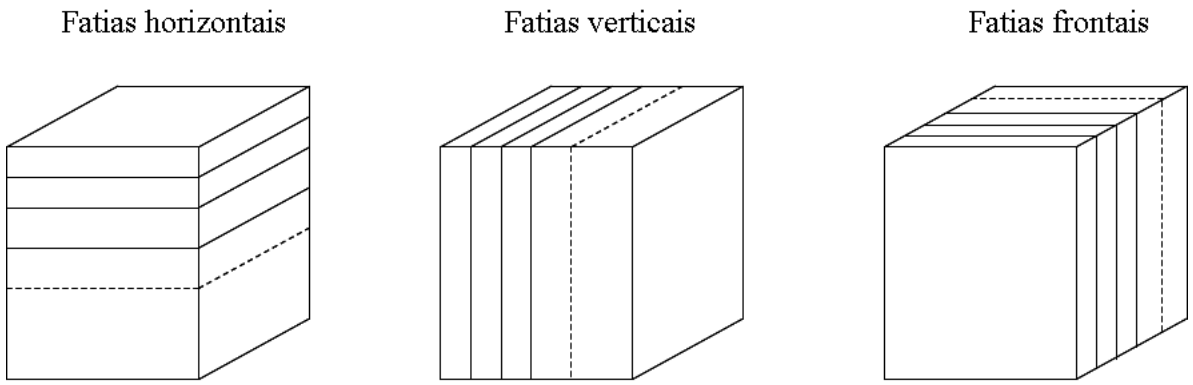

Figura 1 - Particionando um arranjo de três entradas em fatias (arranjos de duas entradas)

século XX e terminaram próximo de 1980, quando os modelos mais importantes e seus algoritmo foram introduzidos.

Algumas das primeiras idéias em análises *multiway* foram publicadas por Cattell (1944, 1952). O princípio da parcimônia de Thurstone (1935) diz que uma estrutura simples pode ser encontrada para descrever uma matriz de dados ou uma matriz de correlações, com a ajuda de fatores. Para a análise simultânea de várias matrizes, Cattell (1944) propôs o uso do princípio do perfil paralelo proporcional. O princípio do perfil paralelo proporcional diz que um conjunto comum de fatores pode ser encontrado e que pode ser ajustado com diferentes dimensões ponderadas para vários dados matriciais no mesmo momento. Isto é o mesmo que encontrar um conjunto comum de fatores para um amontoado de matrizes, ou seja, um arranjo de três entradas.

De acordo com Smilde et. al. (2004) o artigo mais importante de Cattell é o trabalho de 1952, no qual o autor define arranjo *multiway*. Ele definiu objetos, circustâncias/tempo, atributo, escala e observador como as cinco entradas para um arranjo *multiway* idealizado e por razões práticas, reduziu-as a um arranjo de três entradas com pessoas, atributos e circunstâncias.

A decomposição de um arranjo de três entradas foi apresentado primeiramente por Tucker (1963, 1964, 1966). Essa decomposição consiste em encontrar matrizes de cargas $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ e um arranjo núcleo $\underline{\mathbf{G}}$ de três entradas, que foram introduzidos com um exemplo hipotético de 12 indivíduos, 9 tratamentos e 5 observadores. Em outro trabalho independente

foi mostrado uma similaridade entre o arranjo núcleo do modelo de Tucker e a matriz de autovalores na decomposição em valor singular (LEVIN, 1965).

Outros modelos de três entradas foram introduzidos independentemente por Carroll e Chang (1970), que chamaram seu modelo de CANDECOMP (*Canonical Decomposition*) e Harshman (1970) que usou o nome de PARAFAC (*Parallel Factor Analysis*). Entretanto, a idéia de modelagem que está por trás destes dois modelos é a mesma. A proposta básica deste modelo é usar o mesmo fator para descrever a variação em diversas matrizes simultaneamente, embora com diferentes ponderações para cada matriz. Isto é exatamente a idéia definida no perfil paralelo proporcional proposta por Cattell (1944).

O modelo PARAFAC com três entradas consiste em determinar matrizes de cargas $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ com o mesmo número de fatores (Figura 2). Este modelo usualmente produz eixos de coordenadas únicos (não existe liberdade para rotacionar a orientação dos vetores de cargas), enquanto que um modelo de Tucker fornece subespaços únicos.

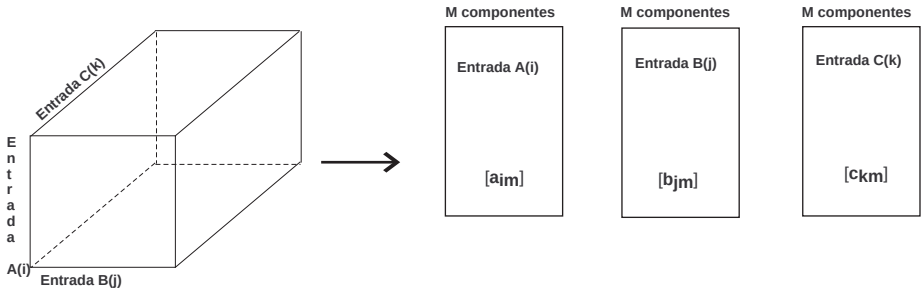

Figura 2 - Decomposição de um arranjo de três entradas propostas por Harshman (1970) e Carrol e Chang (1970)

Outras idéias evoluíram independentemente por Ho et al. (1978; 1980; 1981). Estes autores desenvolveram o método de destruição do rank (*rank annihilation*), que é próximo a idéia de uma decomposição PARAFAC.

2.2.3 Modelos de componentes com três entradas (PARAFAC)

O modelo PARAFAC e o modelo CONDECOMP são aproximadamente os mesmos e serão nomeados aqui como modelos PARAFAC. O modelo PARAFAC é também con-

hecido como decomposição trilinear (SANCHEZ e KOWALSKI, 1990). O modelo PARAFAC é introduzido a seguir usando diferentes notações. 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 (RAO; MITRA, 1971; SCHOTT, 1997).

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 ($\boldsymbol{X = AGB'}$, ver mais detalhes na página 46) truncada em R componentes é:

$$x_{ij} = \sum_{r=1}^{R} a_{ir} g_{rr} b_{jr} + e_{ij}; \qquad i = 1, ..., I \text{ e } ; \; j = 1, ..., J \tag{4}$$

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:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{5}$$

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 3. O modelo representado pela equação (5) 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).

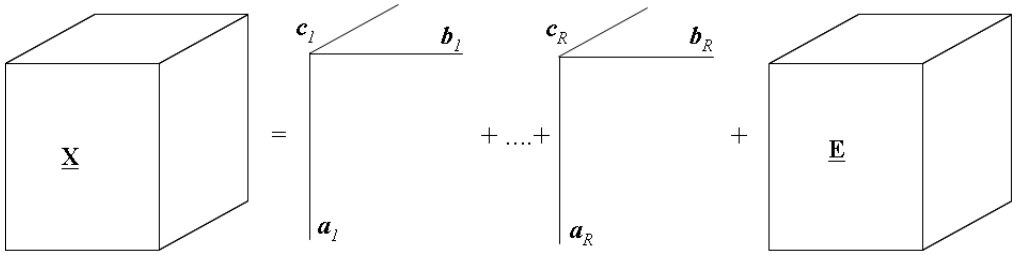

Figura 3 - O modelo PARAFAC com R components

Então, a equação (5) pode ser escrita em termos das matrizes $\boldsymbol{X_k}$, em que $\boldsymbol{X_k}$ é a k-ésima fatia frontal ($I \times J$) do arranjo de três entradas $\underline{\mathbf{X}}$ ($I \times J \times K$):

$$\boldsymbol{X}_k \;=\; \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{B}^{'} + \boldsymbol{E}_k = c_{k1}\boldsymbol{a}_1\boldsymbol{b}_1^{'} + ... + c_{kR}\boldsymbol{a}_R\boldsymbol{b}_R^{'} + \boldsymbol{E}_k \tag{6}$$

em que $\boldsymbol{D}_k$ é uma matriz diagonal com a k-ésima linha de $\boldsymbol{C}$; $\boldsymbol{a}_r$ e $\boldsymbol{b}_r$ são as r-ésimas colunas de $\boldsymbol{A}$ e $\boldsymbol{B}$, respectivamente e $\boldsymbol{E}_k$ é um termo residual. Aqui cada $\boldsymbol{X}_k$ é modelada usando os mesmos componentes, mas com diferentes ponderações, representados por $\boldsymbol{D}_k$.

Assim, todas as fatias $\boldsymbol{X}_k$ são modeladas com perfis paralelos e proporcionais $d_{k1}\boldsymbol{a}_1\boldsymbol{b}_1^{'}$,...,$d_{kR}\boldsymbol{a}_R\boldsymbol{b}_R^{'}$. Isto permite escrever o modelo usando uma notação de componentes simultâneos. Existem três formas completamente equivalentes de escrever o modelo PARAFAC

na notação de componentes simultâneos, devido a simetria da equação 6. Estes são escritos em termos das fatias frontais, horizontais e verticais, respectivamente:

$$\begin{aligned} \boldsymbol{X}_k &= \boldsymbol{A}\boldsymbol{D}_k(\boldsymbol{C})\boldsymbol{B}^{'}, \quad k = 1, ..., K \\ \boldsymbol{X}_i &= \boldsymbol{B}\boldsymbol{D}_i(\boldsymbol{A})\boldsymbol{C}^{'}, \quad i = 1, ..., I \\ \boldsymbol{X}_j &= \boldsymbol{A}\boldsymbol{D}_j(\boldsymbol{B})\boldsymbol{C}^{'}, \quad j = 1, ..., J \end{aligned} \tag{7}$$

sendo que as matrizes $\boldsymbol{D}_k(\boldsymbol{C})$, $\boldsymbol{D}_i(\boldsymbol{A})$ ou $\boldsymbol{D}_j(\boldsymbol{B})$ são interpretadas com operadores que extrai as k-ésimas ou i-ésimas ou j-ésimas colunas das matrizes de cargas apropriadas ($\boldsymbol{C}$, $\boldsymbol{A}$ e $\boldsymbol{B}$ respectivamente).

Quando as matrizes de componentes não têm colunas proporcionais (HARSHMAN[1],1972 apud KROONENBERG, 2008), tem-se a condição necessária para o modelo PARAFAC fornecer estimativas únicas, ou seja, as estimativas de $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ não podem ser modificadas sem mudar o resíduo (não tem liberdade de rotação). Esta propriedade também é chamada de propriedade dos eixos intrínsecos porque com o modelo PARAFAC não somente são gerados subespaços únicos (como na análise de componentes principais), mas também a base de vetores de orientação encontrados são únicos.

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.

2.2.4 Modelos de Tucker

Ledyard Tucker foi um dos pioneiros na análise *multiway*. Ele propôs (TUCKER, 1964; 1966) uma série de modelos, atualmente chamados de análise de com-

[1]HARSHMAN, R.A. Determination and proof of minimum uniqueness conditions for PARAFAC1. **UCLA Working Papers in Phonetics**,Ann Arbor ,v.22, p.30-44, 1972.

ponentes principais de N entradas. Um tratamento extensivo dos modelos de Tucker é dado por Kroonenberg e Leeuw (1980) e Kroonenberg (1983).

2.2.4.1 Modelos Tucker3

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:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{G}\boldsymbol{B}^{'} + \boldsymbol{E} \\ \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{T}_A\boldsymbol{T}_A^{'}\boldsymbol{G}\boldsymbol{T}_B\,\boldsymbol{T}_B^{'}\boldsymbol{B}^{'} + \boldsymbol{E} \\ \boldsymbol{X} &= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{B}}^{'} + \boldsymbol{E} \end{aligned} \tag{8}$$

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 (8) pode ser escrito de outra maneira

$$x_{ij} = \sum_{p=1}^{P}\sum_{q=1}^{Q} \tilde{a}_{ip}\tilde{g}_{pq}\tilde{b}_{jq} + e_{ij} \tag{9}$$

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

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk} \tag{10}$$

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 4.

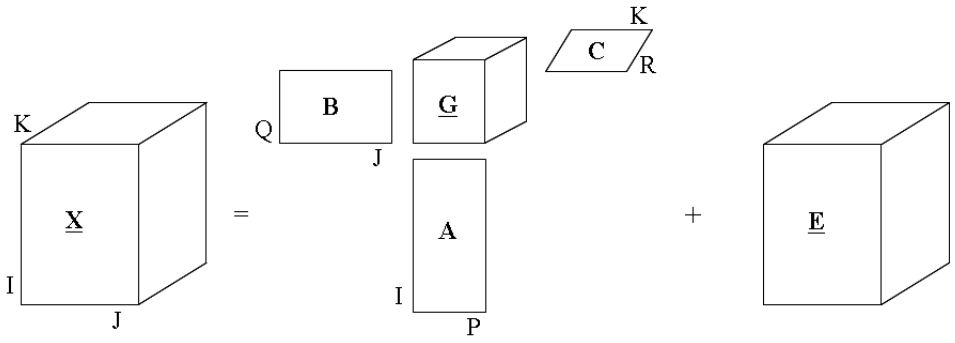

Figura 4 - Representação gráfica do modelo Tucker3

Usando uma matrização adequada para os arranjos $\underline{\mathbf{X}}$ e $\underline{\mathbf{G}}$ com relação ao primeiro modo e a notação do produto de Kronecker, o modelo Tucker3 pode ser escrito como (SMILDE et. al., 2004):

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{G}(\boldsymbol{C} \otimes \boldsymbol{B})' + \boldsymbol{E} \tag{11}$$

sendo que $\boldsymbol{X}$= $[\boldsymbol{X}_1\boldsymbol{X}_2...\boldsymbol{X}_K]$ é uma matrização de $\underline{\mathbf{X}}$ com dimensão $(I \times JK)$ e $\boldsymbol{X}_k$ como definido na equação (6), $\boldsymbol{E}$ é definido similarmente e $\boldsymbol{G}$= $[\boldsymbol{G}_1\boldsymbol{G}_2...\boldsymbol{G}_R]$ é uma matrização do arranjo núcleo $\underline{\mathbf{G}}$ com dimensão $(P \times QR)$, em que $\boldsymbol{G}_r$ é a r-ésima fatia frontal de dimensão $(P \times Q)$ de $\underline{\mathbf{G}}$.

2.2.4.1.1 Propriedades do modelo Tucker3

O modelo Tucker3 tem liberdade de rotação. Isto pode ser visualizado escrevendo, à partir de (11):

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{G}(\boldsymbol{C}' \otimes \boldsymbol{B}') + \boldsymbol{E} \\ &= \boldsymbol{A}\boldsymbol{T}_A\, \boldsymbol{T}_A^{-1}\, \boldsymbol{G}(\boldsymbol{C}' \otimes \boldsymbol{B}') + \boldsymbol{E} \\ &= \tilde{\boldsymbol{A}}\, \tilde{\boldsymbol{G}}\, (\boldsymbol{C}' \otimes \boldsymbol{B}') + \boldsymbol{E} \end{aligned} \tag{12}$$

sendo que $\boldsymbol{T}_A$ é uma matriz ortonormal qualquer, $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{T}_A$ e $\tilde{\boldsymbol{G}} = \boldsymbol{T}_A^{-1}\boldsymbol{G}$. Assim, a transformação de matriz de cargas $\boldsymbol{A}$ pode ser definida similarmente para $\boldsymbol{B}$ e $\boldsymbol{C}$, usando $\boldsymbol{T}_B$ e $\boldsymbol{T}_C$, respectivamente. Conseqüentemente, por existir liberdade de rotação, a ortogonalidade das matrizes de componentes podem ser obtidas sem nenhum custo, definindo as matrizes $\boldsymbol{T}_A$, $\boldsymbol{T}_B$ e $\boldsymbol{T}_C$ apropriadamente. O modelo Tucker3, portanto, não fornece matrizes de componentes únicas por causa de sua liberdade de rotação.

É conveniente fazer as matrizes componentes ortogonais, isto é, $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_{(I\times I)}$, $\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{I}_{(J\times J)}$ e $\boldsymbol{C}\boldsymbol{C}' = \boldsymbol{I}_{(K\times K)}$, assim a soma de quadrados dos elementos de um arranjo núcleo associado com a combinação de certos fatores representa a quantia de variação explicada por aquela combinação de fatores nas diferentes entradas. Se $\underline{\mathbf{X}}$ $(I \times J \times K)$ é um arranjo de três entradas modelado por um modelo de Tucker3 (P, Q, R) e se $\hat{\underline{\mathbf{X}}}$ representa a parte ajustada de $\underline{\mathbf{X}}$, então segue que (Kroonenberg, 1984):

$$\begin{aligned} ||\underline{\mathbf{X}}||^2 &= ||\hat{\underline{\mathbf{X}}}||^2 + ||\underline{\mathbf{E}}||^2 \\ ||\hat{\underline{\mathbf{X}}}||^2 &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}^2 \end{aligned} \tag{13}$$

sendo $\underline{\mathbf{E}}$ um arranjo dos resíduos e $||.||$ é a norma de Frobenius, definida como: $||\underline{\mathbf{X}}|| = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}^2}$, para um arranjo de três entradas $\underline{\mathbf{X}}$ $(I\times J\times K)$. Em outras palavras, a expressão (13) significa que a variação de $\underline{\mathbf{X}}$ é dividida em uma variação não explicada ($\underline{\mathbf{E}}$) e uma variação explicada pelo modelo ($\hat{\underline{\mathbf{X}}}$). Além do mais a soma de quadrados ajustada pode ser dividida em partes relacionadas a cada combinação dos componentes em diferentes direções.

A liberdade de rotação do modelo Tucker3 também pode ser usada para rotacionar o arranjo núcleo para uma estrutura simples assim como é comum na análise de duas entradas. Impor as restrições $\boldsymbol{AA}^{'} = \boldsymbol{I}_{(I\times I)}$, $\boldsymbol{BB}^{'} = \boldsymbol{I}_{(J\times J)}$ e $\boldsymbol{CC}^{'} = \boldsymbol{I}_{(K\times K)}$ não é suficiente para obter uma solução única. Para obter estimativas únicas dos parâmetros, impor ortogonalidade das matrizes de cargas não seria suficiente, mas $\boldsymbol{A}$ poderia também conter os autovetores $\boldsymbol{X}(\boldsymbol{CC}^{'} \otimes \boldsymbol{BB}^{'})\boldsymbol{X}^{'}$ correspondendo aos autovalores decrescentes daquela mesma matriz. Restrições similares poderiam ser colocadas em $\boldsymbol{B}$ e $\boldsymbol{C}$.

2.2.4.2 Modelos de Tucker2

Nos modelos Tucker3 (P, Q, R) de um arranjo $\underline{\mathbf{X}}$ $(I \times J \times K)$ todos três modos são reduzidos, isto é, usualmente $P < I$, $Q < J$ e $R < K$. Também existem os modelos em que somente dois dos três modos são reduzidos, que são chamados de modelos de Tucker2. Isto origina a três modelos especiais, dependendo de qual modo é reduzido. Suponha que um modelo de Tucker3 é ajustado para $\underline{\mathbf{X}}$ $(I \times J \times K)$, mas $\boldsymbol{C}$ é escolhida ser a matriz identidade $\boldsymbol{I}$de dimensão $K \times K$. Assim, não há o interesse na redução no terceiro modo pois a base não é modificada. Então o modelo de Tucker2 pode ser escrito como:

$$\boldsymbol{X}_{(K\times IJ)} = \boldsymbol{I}\boldsymbol{G}_{(K\times PQ)}(\boldsymbol{B}\otimes \boldsymbol{A})^{'} + \boldsymbol{E} = \boldsymbol{G}_{(K\times PQ)}(\boldsymbol{B}\otimes \boldsymbol{A})^{'} + \boldsymbol{E} \tag{14}$$

em que $\boldsymbol{G}_{(K\times PQ)}$ é uma matriz $(K \times PQ)$ e esta é uma versão matrizada apropriada do arranjo núcleo "estendido" $\underline{\mathbf{G}}$ $(P \times Q \times K)$; $\boldsymbol{X}_{K\times IJ}$ é uma versão matrizada apropriada de $\underline{\mathbf{X}}$. Na notação de somatório o modelo de Tucker2 é:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}b_{jq}g_{pqk} + e_{ijk} \tag{15}$$

em que x_{ijk}, a_{ip}, b_{jq} são elementos de $\underline{\mathbf{X}}$ $(I \times J \times K)$, $\boldsymbol{A}$ $(I \times P)$ e $\boldsymbol{B}$ $(J \times Q)$, respectivamente. Além disso, g_{pqk} é um elemento do arranjo núcleo estendido $\underline{\mathbf{G}}$ $(P \times Q \times K)$. Comparando as equações (15) com (10) verifica-se que um símbolo do somatório está ausente. Isto é uma conseqüência da não redução de um dos modos. Uma representação do modelo de Tucker2 é mostrada na Figura 5. Ao comparar esta com a Figura 4, vê-se que a matriz $\boldsymbol{C}$ agora está ausente, ou melhor, $\boldsymbol{C}$= $\boldsymbol{I}$.

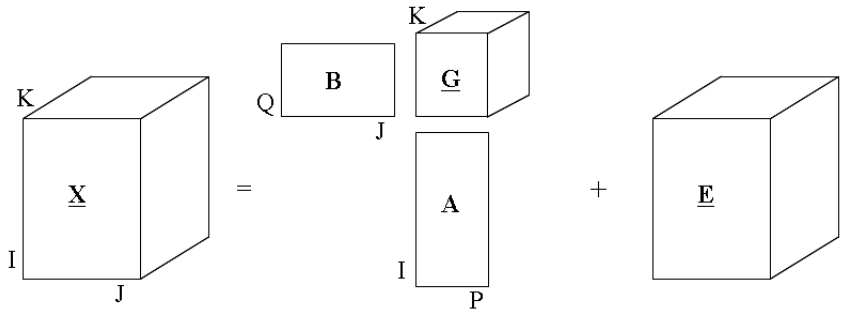

Figura 5 - Representação gráfica do modelo Tucker2

A equação (14) mostra que o modelo de Tucker2 também tem liberdade de rotação pois $\boldsymbol{G}$ pode ser pós-multiplicada por $\boldsymbol{U}\otimes \boldsymbol{V}$ e $(\boldsymbol{B}\otimes\boldsymbol{A})'$ pré-multiplicada por $(\boldsymbol{U}\otimes \boldsymbol{V})^{-1}$ resultando em $(\boldsymbol{B}(\boldsymbol{U}')^{-1}\otimes \boldsymbol{A}(\boldsymbol{V}')^{-1})'$ sem mudanças no ajuste. Assim, as matrizes componentes $\boldsymbol{A}$ e $\boldsymbol{B}$ podem ser ortogonais sem mudança no ajuste.

2.2.4.3 Modelos de Tucker1

Como já foi dito, no caso do modelo de Tucker3, três modos são reduzidos e no modelo de Tucker2 têm dois modos reduzidos. Seguindo nessa linha de raciocínio, é definido como o modelo de Tucker1 os modelos que têm somente um modo reduzido. Existem três diferentes modelos de Tucker1 para um dado arranjo $\underline{\mathbf{X}}$ $(I \times J \times K)$, dependendo de qual modo será reduzido. No modelo de Tucker2 (Figura 5), o terceiro modo não foi reduzido, pela substituição de $\boldsymbol{C}$ por $\boldsymbol{I}$. Se o modo $\boldsymbol{B}$ também for substituído por $\boldsymbol{I}$, então o segundo e o terceiro modos não são reduzidos. O modelo de Tucker1 resultante é:

$$\boldsymbol{X}_{(\boldsymbol{K}\times\boldsymbol{IJ})} = \boldsymbol{I}\boldsymbol{G}_{(\boldsymbol{K}\times\boldsymbol{JP})}(\boldsymbol{I}\otimes\boldsymbol{A})' + \boldsymbol{E} \tag{16}$$

sendo que $\boldsymbol{G}_{K\times JP}$ tem dimensão $K \times JP$ e somente o terceiro modo é reduzido. O modelo (16) é modificado se $\underline{\mathbf{X}}$ for matrizada, para:

$$\boldsymbol{X}_{(\boldsymbol{I}\times\boldsymbol{JK})} = \boldsymbol{A}\boldsymbol{G}_{(\boldsymbol{P}\times\boldsymbol{JK})} + \boldsymbol{E} \tag{17}$$

em que $\boldsymbol{X}_{(I\times JK)}$ é a matrização de $\underline{\mathbf{X}}$ de dimensão $(I \times JK)$; $\boldsymbol{G}_{P\times JK}$ é uma matrização de $\underline{\mathbf{G}}$ de dimensão $(P\times JK)$. Se $\boldsymbol{X}_{I\times JK}$ for substituída por $\boldsymbol{X}$, $\boldsymbol{A}$ por $\boldsymbol{T}$ e $\boldsymbol{G}_{P\times JK}$ por $\boldsymbol{P}'$, então

é obtido uma solução pela análise de componentes principais (ACP) usual. Isto é exatamente o que o modelo de Tucker1 representa: uma ACP na matrização apropriada de $\underline{\mathbf{X}}$. Assim, algoritmos para encontrar $\boldsymbol{A}$ e $\boldsymbol{G}_{P\times JK}$ estão disponíveis, como por exemplo a decomposição por valores singulares.

A equação (17) mostra que o modelo de Tucker1 também tem liberdade de rotação. Assim, $\boldsymbol{A}$ pode ser escolhida ortogonal sem perda na proporção da variabilidade total explicada pelo modelo. Uma representação gráfica do modelo de Tucker1 é dado na Figura 6. Como no modelo de Tucker2, existem diferentes modelos de Tucker1 para certo conjunto de dados dependendo de qual modo será reduzido.

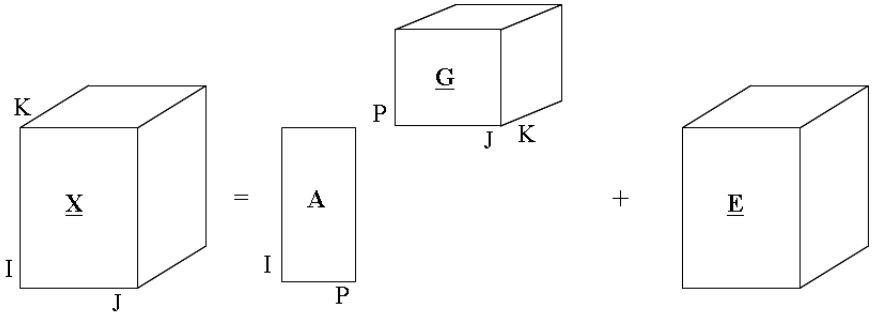

Figura 6 - Representação gráfica do modelo Tucker1, em que somente o primeiro modo é reduzido

2.2.5 Relações entre modelos de componentes de três entradas

Os modelos de três entradas e suas propriedades foram apresentados anteriormente e, para um principiante, pode ser muito difícil decidir qual modelo será utilizado. Assim, a fim de decidir qual modelo escolher para utilizar em determinada situação, é importante ter um bom entendimento das diferenças e similaridades entre os modelos.

Uma importante questão é como os modelos PARAFAC e Tucker3 são relacionados. O modelo PARAFAC fornece eixos únicos (solução única) enquanto que no modelo Tucker3 esta característica não é observada devido a liberdade de rotação. Um Tucker3 pode ser transformado (rotacionado) e simplificado para ser parecido com um modelo PARAFAC, e isto às vezes pode ser feito com pouca ou sem perda no ajuste. Existe também uma hierarquia, isto é, dentro da família dos modelos Tucker (Tucker3, Tucker2 e Tucker1, mais

detalhes na seção 2.2.5.1). Além disso, as propriedades estatísticas dos dados, como ruídos e erros sistemáticos, apresentam-se também como uma importante regra na escolha do modelo.

Nem sempre há uma resposta clara ou uma definição adequada do modelo de três entradas para determinada aplicação. Às vezes, um ou mais modelos têm que ser ajustados, a fim de encontrar o melhor modelo. Em muitos casos, os modelos apresentam-se igualmente bons e a seleção pode ser baseada em critérios práticos como precisão de algoritmo ou velocidade. Assim, conhecendo as opções de modelos de três entradas junto com suas propriedades e relações, têm-se uma base que serve para fazer escolhas objetivas dos modelos adequados para determinada situação.

2.2.5.1 Hierarquia dos Modelos PARAFAC e TUCKER3

O modelo PARAFAC geral de R-componentes de um arranjo $\underline{\mathbf{X}}$ $(I \times J \times K)$ é dado pela equação (5). Introduzindo o termo g_{pqr} no somatório com $g_{pqr} = 1$ caso $p = q = r$ e $g_{pqr} = 0$, caso contrário. Assim, a equação (5) pode ser reescrita como a equação (10) que é o modelo de Tucker3 de $\underline{\mathbf{X}}$. Logo, o modelo PARAFAC pode ser entendido como um modelo de Tucker3 restrito (Figura 7). Por esta razão, existe uma relação hierárquica entre os modelos PARAFAC e Tucker3. Um modelo PARAFAC com R-componentes sempre tem um ajuste pior ou igual que o modelo de Tucker3 (R, R, R). No entanto, isto não significa necessariamente que o modelo de Tucker3 (R, R, R) é o preferido. Na Figura 7 a linha tracejada no arranjo $\underline{\mathbf{G}}$, refere-se a diagonal de um cubo, sendo chamada de superdiagonal.

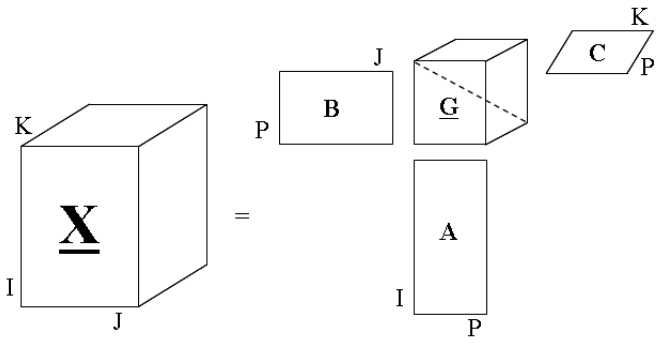

Figura 7 - Modelo PARAFAC escrito como um modelo de Tucker3

É menos intuitivo, mas apesar disso é instrutivo, ver que o modelo de Tucker3 também pode ser representado como um modelo PARAFAC restrito, embora use muito mais componentes.

2.2.5.2 Hierarquia dos Modelos TUCKER3, TUCKER2 e TUCKER1

Considere um arranjo de três entradas $\underline{\mathbf{X}}$ $(I \times J \times K)$ e diferentes modelos para este arranjo. O modelo de Tucker3 para este arranjo é dado por (10). O modelo de Tucker2 (15) pode substituir o número de componentes em um modo pela dimensão deste modo. Por exemplo, substituindo o número de componentes no terceiro modo por K, o modelo de Tucker3 indica:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{K} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk}. \tag{18}$$

em que a_{ip}, b_{jq}, c_{kr} e g_{pqr} são como definidos na equação (10). Devido a liberdade de rotação dos modelos de Tucker3, a matriz de cargas $\boldsymbol{C}$ $(K \times K)$ é rotacionada para uma matriz identidade $\boldsymbol{I}$ $(K \times K)$. Então $c_{kr} = 1$ para $k = r$ e zero para qualquer outro valor. Reescrevendo a equação (18) tem-se:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}b_{jq}c_{kk}g_{pqk} + e_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}b_{jq}f_{pqk} + e_{ijk} \tag{19}$$

em que $f_{pqk} = c_{kk}g_{pqk}$. A equação 19 pode ser escrita, em termos matriciais, como:

$$\boldsymbol{X}_{(K\times IJ)} = \boldsymbol{C}\boldsymbol{G}_{(K\times PQ)}(\boldsymbol{B}\otimes \boldsymbol{A})^{'} + \boldsymbol{E} = \boldsymbol{F}_{(K\times PQ)}(\boldsymbol{B}\otimes \boldsymbol{A})^{'} + \boldsymbol{E} \tag{20}$$

sendo $\boldsymbol{F}_{(K\times PQ)} = \boldsymbol{C}\boldsymbol{G}_{(K\times PQ)} = \boldsymbol{I}\boldsymbol{G}_{(K\times PQ)}$. No arranjo de três entradas $\underline{\mathbf{F}}$ de dimensão $(P \times Q \times K)$, com o terceiro modo não reduzido é chamado de arranjo núcleo estendido no modelo de Tucker2. Existem ainda outros dois tipos de modelos de Tucker2: aqueles nos quais o primeiro ou segundo modo não é reduzido.

Para explicar a hierarquia dos modelos de Tucker3 e Tucker2, suponha um modelo de Tucker3 $\boldsymbol{X}=\boldsymbol{A}\boldsymbol{G}(\boldsymbol{C} \otimes \boldsymbol{B})^{'} + \boldsymbol{E}$ e um modelo de Tucker2 em que o terceiro modo não é reduzido $\boldsymbol{X}=\boldsymbol{F}(\boldsymbol{C} \otimes \boldsymbol{B})^{'} + \boldsymbol{E}$. Os parâmetros $\boldsymbol{B}$ e $\boldsymbol{C}$ são diferentes nos dois modelos, mas é importante notar que o modelo de Tucker2 é mais flexível que o modelo de Tucker3 porque no modelo de Tucker3 a parte correspondente a $\boldsymbol{F}(K \times PQ)$ é especificamente

parametrizado como $\boldsymbol{AG}$. Logo, o modelo de Tucker2 vai se ajustar melhor que os modelos de Tucker3 e, assim, o modelo de Tucker3 pode ser visto como um modelo de Tucker2 restrito.

O Modelo de Tucker2 pode até ser construído menos restrito, por exemplo, por não comprimir o segundo modo. Isto pode ser feito pela formulação do modelo de Tucker1

$$x_{ijk} = \sum_{p=1}^{P} a_{ip} h_{pjk} + e_{ijk}. \tag{21}$$

em que x_{ijk}, a_{ip} e e_{ijk} são elementos de $\underline{\mathbf{X}}$, $\boldsymbol{A}$ $(I \times P)$ e $\underline{\mathbf{E}}$ $(I \times J \times K)$ respectivamente. Os valores h_{pjk} são os elementos do arranjo $\underline{\mathbf{H}}$ $(P \times J \times K)$. O segundo e o terceiro modo de $\underline{\mathbf{X}}$ não foram reduzidos, porque as cargas não são estimadas nestes modos. Este modelo é menos restrito que o modelo Tucker2 e vai além disso, este modelo, ajuste-se melhor ao dados.

2.3 Graus de liberdade dos modelos *multiway*

Os primeiros a terem a idéia em atribuir graus de liberdade aos ajustes dos modelos de três entradas segundo Kroonenberg (2008) foram Weesie e Van Houwelingen (1983)[2]. Eles discutiram os graus de liberdade para o modelo Tucker3, mas o princípio pode ser estendido aos modelos de Tucker2 e PARAFAC. Este princípio consiste em:

$$\begin{aligned} df &= \text{n.º de observações - n.º de médias removidas - n.º de SQ empatadas} \\ &- \text{n.º de dados perdidos - n.º de parâmetros livres} \end{aligned} \tag{22}$$

com o número de parâmetros livres, f_p,

$$\begin{aligned} f_p &= \text{n.º de parâmetros independentes na matriz de componentes observada} \\ &+ \text{n.º de elementos no arranjo núcleo} \\ &- \text{n.º de indeterminação rotacional para matrizes de componente (Tucker) ou} \\ &- \text{n.º de restrições de comprimento unitário (PARAFAC)} \end{aligned} \tag{23}$$

[2]WESSIE, J.; VAN HOUWELINGEN, H. **GEPCAM user's manual:** Generalized principal components analysis with missing values. Utrecht: University of Utrecht, Institute of Mathematical Statistics, 1983. 47p. Technical report.

O número de observações é calculado como um critério direto: simplesmente conta-se o número de observações. Para arranjos de três entradas é obviamente IJK. Porém, ao calcular graus de liberdade, deve-se levar em conta a regra do produto-máximo. Esta regra diz que "em casos em que o tamanho de uma das entradas seja maior que o produto das outras duas, devem ser feitos ajustes especiais (para o cálculo do número de parâmetros livres), porque em tais casos segundo (KIERS; HARSHMAN, 1997), o modo maior pode ser reduzido consideravelmente sem perda de informação para os componentes das outras duas entradas e para o arranjo núcleo. Especificamente, quando $I > JK$, os dados podem ser reduzidos a um conjunto de dados de dimensão $JK \times J \times K$. Então, em tais casos, no cálculo do número de parâmetros livres f_p, I deve ser substituído por JK"(CEULEMANS; KIERS, 2006). Uma conseqüência disto, por exemplo, é que para ajustar perfeitamente um arranjo de dados $I \times J \times K$ com $I > JK$, precisa-se de JK, J, K componentes para as três entradas ajustar perfeitamente aos dados. Assim, no cálculo dos graus de liberdade, isto deve ser levado em consideração.

Para os modelos *multiway*, o número de parâmetros livres é o seguinte:

$$\begin{aligned} \text{PARAFAC}f_p &= I \times S + J \times S + K \times S + S - 3S \\ \text{Tucker2}f_p &= I \times P + J \times Q + P \times Q \times K - P^2 - Q^2 \\ \text{Tucker3}f_p &= I \times P + J \times Q + K \times R + P \times Q \times R - P^2 - Q^2 - R^2. \end{aligned} \tag{24}$$

2.4 Postos de arranjos

No caso de arranjos de duas entradas (matrizes) existe uma relação direta entre a dimensão da matriz núcleo e o seu posto. Entretanto, o caso de *multiway* é consideravelmente mais complexo porque um arranjo núcleo de mesma ordem pode ter diferentes postos, por exemplo, um arranjo de dimensão $2 \times 2 \times 2$ pode ter qualquer posto 2 ou posto 3, pode gerar alguma dúvida sobre os verdadeiros graus de liberdade neste caso. A pergunta sobre qual é o real posto de arranjos *multiway* é muito complicado, mas tem-se observado grandes progressos para determinar o posto de arranjos de três entradas, veja (TEN BERGE, 2004).

2.5 Determinação da dimensionalidade de um modelo de Tucker

Um aspecto importante da seleção de dimensionalidade para modelos de Tucker é que nem todas as combinações de dimensões são manejáveis. A razão para isto é que em cada passo do algoritmo, a matriz para a qual os autovalores são calculados tem o posto igual ao produto do número de componentes dos outros modos. Se aquele posto for menor que o modo atual, o algoritmo falhará pois terá uma solução redundante, ou seja, existe outra solução que explica a mesma variabilidade, mas com um número menor de componentes. Outra forma de olhar para isto é notar que o posto do arranjo núcleo é o posto do arranjo de três entradas ortogonal, e este pode ser ortogonal somente quando o produto dos números de componentes forem maiores ou iguais ao número de componentes no modo restante (WANSBEEK; VERHEES, 1989).

2.5.1 Procedimento $DifFit$ de Timmerman-Kiers

Dado um modelo de Tucker3, Timmerman e Kiers (2000) sugeriu um procedimento para selecionar a dimensionalidade semelhante ao scree plot na análise de componentes principais. Inicialmente (1º filtro) selecionam-se todos os possíveis modelos de Tucker3 que satisfazem a seguinte condição proposta por Kruskal (1989) :

$$\begin{aligned} P &\leq QR \\ Q &\leq PR \\ R &\leq PQ \end{aligned} \tag{25}$$

O próximo passo (2º filtro), consiste em selecionar, dentro de uma determinada classe de modelos de Tucker3, com o mesmo número total de componentes $S = P + Q + R$, o modelo que tem a maior proporção da soma de quadrados ajustada, SQ_S, ou equivalentemente, a menor soma de quadrados residual ou deviance. Para comparar classes com S diferente, calcula-se $dif_S = SQ_S - SQ_{S-1}$. Somente serão considerados os dif_S que são consecutivamente (em ordem decrescente) maiores. Estes autores ainda definiram um valor de relevância (*salience value*): $b_S = dif_S/dif_{S*}$, em que dif_{S*} é o maior valor depois de dif_S. Assim, seleciona-se o modelo que têm o valor de b_S mais alto, e este procedimento é chamado de critério $DifFit$.

Neste trabalho de Timmerman e Kiers, ainda foi proposto um ponto de corte para dif_S, sendo que modelos com valores abaixo deste limite não devem ser levados em conta. O dif_S deve ser maior que a proporção média da variabilidade explicada, tomada sobre todos possíveis valores de S (esta proporção é dada por $\|\underline{\mathbf{Z}}\|/S_{min}$, em que $S_{mim} = min(I; JK) + min(J; IK) + min(K; IJ) - 3$). Isto é equivalente ao critério de Kaiser da análise de componentes principais (JOHNSON; WICHERN, 1998; BARROSO; ARTES, 2003), em que os autovalores devem ser maiores que o valor médio da variância explicada por cada componente.

Para visualizar o procedimento proposto por Timmerman e Kiers, é necessário construir uma versão do *scree plot*, o *scree plot multiway*, no qual as somas de quadrados residuais para cada dimensionalidade são colocadas em gráficos contra a soma S do número de componentes em cada uma das entradas. O procedimento de $DifFit$ é essencialmente uma maneira para definir o número ótimo S no *scree plot multiway*, e assim a dimensionalidade S ótima do modelo é obtida quando a variação da explicação entre dimensionalidades consecutivas passa ser pequena. Timmerman e Kiers (2000) apresentaram este procedimento para o modelo de Tucker3, mas trabalha da mesma forma para outros modelos de Tucker.

Na proposta inicial de Timmerman e Kiers (2000), a soma de quadrados residual para cada modelo era calculado ajustando o modelo de Tucker3 para cada dimensionalidade fixa através do algoritmo ALS. Mas Kiers e Der Kinderen (2003) mostraram que é suficiente aproximar os modelos por um único cálculo, usando o método original de mínimos quadrados de Tucker, e calcular a soma de quadrado residual para os modelos em consideração usando o arranjo central, o que promoverá uma grande economia no tempo gasto com os ajustes dos modelos.

2.5.2 Análise residual

Uma alternativa para o procedimento de Timmerman-Kiers é a avaliação da soma de quadrados residual junto com os graus de liberdade. Novamente um gráfico pode ser construído com a soma de quadrados residual d de cada modelo, mas agora colocado em um gráfico contra os graus de liberdade df.

Nos gráficos de análise residual, a avaliação do modelo pode ser facilitada dese-

nhando linhas diretas entre os espaços dos modelos com uma constante $k = d/df$. Se a origem está incluída no gráfico, estas linhas podem ser vistas como um critério para retirar a origem com um declive k. Se k = 1, então cada parâmetro adicionado ou cada grau de liberdade perdido conduz ao ganho de uma unidade na soma de quadrados residual. Se dois modelos estão ligados por uma linha que tem um declive íngreme (ou seja, k grande), então pela troca de alguns graus de liberdade pode-se adquirir uma redução grande na soma de quadrados residual, ou em outras palavras, um modelo ajustando-se consideravelmente melhor. Por outro lado, se o declive é muito baixo (ou seja, k é pequeno), precisa-se sacrificar muitos graus de liberdade para ter uma redução pequena nos resíduos ou um pequeno aumento no ajuste. O princípio de parcimônia pode ser aplicado aqui, dando preferência para modelos com bom ajuste e poucos parâmetros.

Embora esta análise residual tenha sido projetada para ajudar escolher entre modelos de Tucker que diferem em dimensionalidade, é igualmente possível incluir modelos de outras classes. Em outras palavras, poderia ser vantajoso adicionar os modelos de Tucker2 e modelos PARAFAC apropriados nesta análise residual. Para modelos de PARAFAC, não existe uma maneira correta para determinar o número de graus de liberdade, mas dada uma decisão sobre o número de graus de liberdade, estes modelos podem ser incluídos no gráfico da análise residual (KROONENBERG, 2008).

2.5.3 Critério *st* de Ceulemans-Kiers

Um método de seleção de modelos em gráficos de análise residual foi proposto por Ceulemans e Kiers (2006). O critério de seleção é o critério *st* que, baseado na soma de quadrados residual e graus de liberdade em vez da soma de quadrados do ajuste e número de parâmetros livres, é definido como:

$$st_i = \frac{\frac{d_{i-1} - d_i}{df_{i-1} - df_i}}{\frac{d_i - d_{i+1}}{df_i - df_{i+1}}}, \qquad \text{com } i = 1, \ldots, n \tag{26}$$

em que d_i é o resíduo, df_i é o número de graus de liberdade do modelo i e n é o número máximo de modelos que se pode ajustar. O procedimento para seleção do modelo consiste nos seguintes passos:

1. Determine os valores df e d de todas as soluções das quais deseja escolher;
2. Para cada valor de n observa-se df, retendo somente a melhor solução em termos do ajuste;
3. Ordene as n soluções pelos valores de df e denote-os por s_i $(i = 1, \ldots, n)$;
4. Exclua todas as soluções s_i para qual uma solução s_j $(j < i)$ existe tal que $d_j < d_i$;
5. Considere consecutivamente todos os grupos de soluções iguais adjacentes. Exclua a solução mediana, se seu valor ficar situado acima ou na linha que conecta aos seus vizinhos no gráfico da análise residual;
6. Repita o passo 5 até que nenhuma solução possa ser excluída;
7. determine os valores de st das soluções obtidas, de acordo com a equação (26).
8. selecione a solução com o menor valor de st;

Em resumo, passos $1 - 6$ servem para determinar quais dos modelos devem fazer parte da avaliação do st e os passos 7 e 8 servem para achar o menor ângulo $\phi_{i-1,i+1}$ entre as linhas que conectam o i-ésimo modelo com seu anterior $(i - 1)$ com o seu modelo subseqüente $(i + 1)$.

Um comentário importante feito por Ceulemans e Kiers (2006) é que o critério de seleção do modelo proposto por eles não deve ser seguido rigorosamente, e critérios que levam em consideração um bom conhecimento do assunto também podem fazer um papel de seleção, de forma que um modelo pouco inferior ao indicado pelo método às vezes pode ser escolhido devido a suas qualidades de interpretação. Este trabalho ainda discute as vantagens do critério st em cima do método de $DifFit$ (TIMMERMAN; KIERS, 2000; KIERS; DER KINDEREN, 2003) para tipos diferentes de modelos. A principal comparação entre os métodos é a equivalência entre os dois métodos no caso de um modelo de Tucker.

2.6 Determinação da dimensionalidade de um modelo PARAFAC

Em princípio, selecionar a dimensionalidade de um modelo de PARAFAC é muito mais simples que selecionar a dimensionalidade de um modelo de Tucker, já que todos

os modos têm o mesmo número de componentes. Em algumas áreas da ciência, por exemplo a psicologia, é comum que o número máximo de componentes nos modelos PARAFAC seja dois ou três, porém para alguns modelos físicos com aplicações químicas podem ser ajustados com grandes quantidades de componentes. Por exemplo, Março et al. (2005) mostram que são necessários seis componentes no modelo PARAFAC para estudar substâncias químicas em flores de hibisco. Murphy et al. (2006) analisando o conteúdo químico na troca de água em um lastro de navios, indicaram um modelo PARAFAC com nove componentes.

2.6.1 Procedimentos de dividir ao meio (*Split-half*)

Uma maneira para determinar a dimensionalidade de um modelo PARAFAC é avaliar a estabilidade da solução, dividindo o conjunto de dados ao meio e executar uma análise separada em ambas as partes. Se houver uma solução ótima, esta deve aparecer em ambas as análises.

Uma advertência é que deve haver objetos suficientes no modo que está sendo dividido. Em outras palavras, ambas as partes devem ser grandes o bastante para minimizar a influência individual de dados específicos. Quanto é esse "grande o bastante" é difícil de dizer em geral, porque depende muito do nível de ruído e da qualidade da estrutura dos dados. Harshman e DeSarbo (1984) discutiram e demonstraram o procedimento de dividir ao meio em grandes detalhes e com exemplos ilustrativos. Eles também sugeriram fazer duas (ou mais) divisões ortogonais. Para análises de dados dividido ao meio, o requisito é que os dados sejam divididos aleatoriamente em quatro partes mais ou menos iguais, por exemplo A, B, C, e D, que são combinados em quatro novos conjuntos de dados: (A+B e C+D), e (A+C e B+D). Kiers e Van Mechelen (2001) sugerem uma aproximação semelhante para o modelo de Tucker3.

2.6.2 Consistência do núcleo

Uma aproximação interessante para selecionar o número de componentes do modelo PARAFAC foi proposta por Bro (1998). O autor propôs o princípio de consistência do núcleo, para avaliar quantos componentes no modelo PARAFAC são necessários para descrever os dados, e em Bro e Kiers (2003) este princípio é nomeado de CONCORDIA ou

diagnóstico de consistência do núcleo. O método consiste em avaliar quão distante está a ordem do núcleo derivada do ajuste de um modelo PARAFAC de um modelo ideal, ou seja, de um arranjo núcleo superdiagonal $\underline{\mathbf{G}}$.

Um arranjo núcleo do modelo PARAFAC, que é um arranjo núcleo calculado a partir dos componentes de um modelo PARAFAC, é uma superdiagonal se para um modelo de três entradas somente g_{111}, g_{222}, $g_{333}\ldots$ tem valores consideráveis e todos os outros elementos de arranjo são próximos a zero. Na aproximação de Bro (1998), o arranjo núcleo é padronizado de tal forma que os elementos da superdiagonal são iguais a 1 (arranjo superidentidade). Para um modelo de três entradas PARAFAC, a discrepância do arranjo ideal é:

$$CONCORDIA \;=\; 1 - \frac{\sum_{p=1}^{S}\sum_{q=1}^{S}\sum_{r=1}^{S}(g_{pqr} - i_{pqr})^2}{\sum_{p=1}^{S}\sum_{q=1}^{S}\sum_{r=1}^{S}(i_{pqr})^2}. \tag{27}$$

em que i_{pqr} é um elemento de um arranjo superidentidade, de modo que $i_{pqr} = 1$ se $p = q = r$ e $i_{pqr} = 0$, caso contrário.

O diagnóstico de consistência de núcleo tornará um valor muito baixo (negativo) quando os componentes dentro de uma entrada forem altamente correlacionados. Se isto não é desejável, os i_{pqr} podem ser substituídos por g_{pqr} no denominador; a medida é chamada consistência do núcleo normalizada. Segundo os autores é duvidoso se isto faria qualquer diferença com relação a decisão sobre o número de componentes a reter no modelo, mas com relação a consistência do núcleo, para valores próximos de 1, significa que o modelo é informativo. O grau de superdiagonalidade também pode ser usado para o mesmo propósito. Esta medida é igual à soma de quadrados dos elementos da superdiagonal dividida pela soma total de quadrados dos elementos de núcleo. Todas as três medidas serão iguais a 1 no caso de uma arranjo superdiagonal.

2.7 Estabilidade do modelo e poder preditivo por validação

Uma das principais preocupações na construção de modelo é responder a pergunta "se o modelo encontrado será adequado em novas amostras". A idéia básica é dividir o conjunto em duas partes e ajustar um modelo em uma parte dos dados e então comparar com os dados não envolvidos na estimação. Esse é um procedimento de validação. Quando

tais diferenças forem pequenas, é dito que as estimativas de parâmetro têm poder de boa predição. É claro que, para tal aproximação nada é feito para as insuficiências da amostra original, mas pelo menos dá uma certa informação em quão bom o modelo se mostrará em outras amostras da mesma população (KROONENBERG, 2008).

Uma outra maneira de avaliar o poder preditivo de um modelo é através de um procedimento de reamostragem jackknife. Riu e Bro (2003) retiraram uma fatia completa de cada vez, tendo como objetivo usar jackknife para estimar os erros padrões e procurar pontos discrepantes, e não avaliar o poder preditivo do modelo. Uma versão mais elegante desta proposta foi formulada por Louwerse; Smilde e Kiers (1999), baseado em Eastment e Krzanowski (1982), no qual também removem fatias inteiras e desenvolveram uma maneira sofisticada para combinar os resultados da retirada da fatia, de forma que o poder preditivo possa ser avaliado. Outra proposta é remover cada observação, ou vários dados ao mesmo tempo, sendo que estas observações são consideradas como perdidas. Eles usaram um algoritmo EM para estimar os dados indicados como perdidos, e compararam os valores estimados com os valores observados.

O poder preditivo de um modelo geralmente é avaliado pela soma de quadrados dos erros preditivos ($PRESS$), sendo calculada comparando os valores de todos os dados originais com os seus valores estimados, com base nos modelos sem os dados em questão. Assim, para o caso de três entradas, x_{ijk} é comparado com $\widehat{x}_{ijk}^{PQR}$ e o poder preditivo, $PRESS_{PQR}$, é

$$PRESS_{PQR} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\widehat{x}_{ijk}^{PQR} - x_{ijk})^2. \tag{28}$$

Além disso, Louwerse; Smilde e Kiers (1999) desenvolveram a estatística $W_{PQR,(P-1)QR}$, para medir a diferença relativa do $PRESS_{PQR}$ entre dois modelos que diferem em um único componente (aqui apresentada para a primeira entrada). A estatística W é definida como

$$W_{(P-1)QR,PQR} = \frac{\dfrac{PRESS_{(P-1)QR} - PRESS_{PQR}}{df_{(P-1)QR} - df_{PQR}}}{\dfrac{PRESS_{PQR}}{df_{PQR}}} \tag{29}$$

em que *df* são os graus de liberdade do modelo. Para evitar valores negativos, que podem acontecer quando a redução dos graus de liberdade não é acompanhada por uma diminuição suficiente no $PRESS$, um limite inferior de zero pode ser introduzido.

Uma estratégia para minimizar o número de modelos que têm que ser avaliados, ou seja, para encontrar modelos com baixos valores suficientemente de $PRESS$, também foi sugerido por Louwerse; Smilde e Kiers (1999). A essência desta estratégia é: dado um modelo com um número específico de componentes, o próximo conjunto de modelos a ser avaliado são aqueles com um componente a mais em cada entrada. O melhor destes modelos é tomado como uma base para o próximo passo. Este processo será interrompido se o $PRESS$ aumentar ou não diminuir suficientemente (por exemplo, menor que 10^{-6}).

2.8 Biplot

Biplot é uma representação gráfica em que as linhas e as colunas são apresentadas em um gráfico com duas ou três-dimensões, sendo que a construção do *biplot* é baseada na decomposição em valores singulares da matriz de dados. A grande importância deste gráfico é a possibilidade da inspeção visual da posição de uma unidade amostral relativa a outra e a importância relativa de cada uma das variáveis à posição de qualquer unidade. Por consequência, pode-se ver como as unidades amostrais se agrupam e quais variáveis contribuem para sua posição dentro dessa representação.

2.8.1 Decomposição em Valores Singulares

Suponha uma matriz $\boldsymbol{X}$ com informação de I objetos em J variáveis. A Decomposição em Valor Singular (DVS) da matrix $\boldsymbol{X}$ é definida como:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{'} \tag{30}$$

que também pode ser escrita em notação de somatórios como

$$x_{ij} = \sum_{s=1}^{P} \lambda_s u_{is} v_{js} \tag{31}$$

em que $P = min(I, J)$, ou seja, são necessários P termos para reproduzir exatamente a matriz original $\boldsymbol{X}$. Os escalares λ_s são os valores singulares organizados em ordem decrescente, $\mathbf{u_s}$ é um vetor de objetos e $\mathbf{v_s}$ é um vetor de variável. Em ambos os casos os vetores são ortonormais. $\boldsymbol{U}$ e $\boldsymbol{V}$ são matrizes que contêm os vetores $\mathbf{u_s}$ e $\mathbf{v_s}$ em suas colunas, respectivamente.

Para encontrar uma aproximação de menor dimensão para $\boldsymbol{X}$, tem-se que minimizar a distância entre a matriz original e a matriz de aproximação $\widehat{\mathbf{X}}$. Esta distância entre os elementos das duas matrizes, $\boldsymbol{X}= (x_{ij})$ e $\widehat{\mathbf{X}}= (\widehat{x}_{ij})$, é definido como:

$$d(\boldsymbol{X},\widehat{\mathbf{X}}) = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \widehat{x}_{ij})^2}, \tag{32}$$

e Eckart e Young (1936) mostraram que a melhor aproximação Q-dimensional de mínimos quadrados da matriz $\boldsymbol{X}$ é obtida pela decomposição em valor singular de $\boldsymbol{X}$, somando somente os Q ($Q < P$) primeiros termos da equação (31).

Os primeiros Q vetores $\mathbf{u_s}$ e $\mathbf{v_s}$, com Q usualmente dois ou três, são usados como coordenadas para o gráfico de representação dos dados. Eles podem ser combinados com os valores singulares λ_s em diferentes formas, das quais as duas mais comuns são:

$$\widehat{x}_{ij} = \sum_{s=1}^{Q} u_{is}(\lambda_s v_{js}) = \sum_{s=1}^{Q} y_{is}z_{js}, \tag{33}$$

$$\widehat{x}_{ij} = \sum_{s=1}^{Q}(u_{is}\lambda_s^{1/2})(\lambda_s^{1/2}v_{js}) = \sum_{s=1}^{Q} y^*_{is}z^*_{js}, \tag{34}$$

em que y e z são as coordenadas dos objetos e das variáveis, sendo que somente as coordenadas principais das variáveis são escalonadas e y^* e z^* são as coordenadas dos objetos e das variáveis, em que coordenadas principais são simetricamente escalonadas.

2.8.2 *Biplot* padrão

Um *biplot* padrão é a apresentação da tabela de interação entre objetos e variáveis $\boldsymbol{X}$, decomposta no produto $\boldsymbol{Y}\boldsymbol{Z}'$, em que $\boldsymbol{Y}$ é de dimensão $I \times Q$ e $\boldsymbol{Z}$ é de dimensão $J \times Q$. Usando uma decomposição de duas-dimensões, cada elemento $\widehat{x}_{ij}$ de $\widehat{\mathbf{X}}$ pode ser escrito como:

$$\widehat{x}_{ij} = y_{i1}z_{j1} + y_{i2}z_{j2} \tag{35}$$

que é o produto interno dos vetores linhas (y_{i1}, y_{i2}) e (z_{j1}, z_{j2}). Um *biplot* é obtido representando cada linha como um ponto Y_i com coordenadas (y_{i1}, y_{i2}) e cada coluna como um ponto Z_j com coordenadas (z_{j1}, z_{j2}) em um gráfico de duas dimensões. Estes pontos são geralmente chamados de marcadores de linhas e marcadores de colunas, respectivamente. Se escrever Y_i''

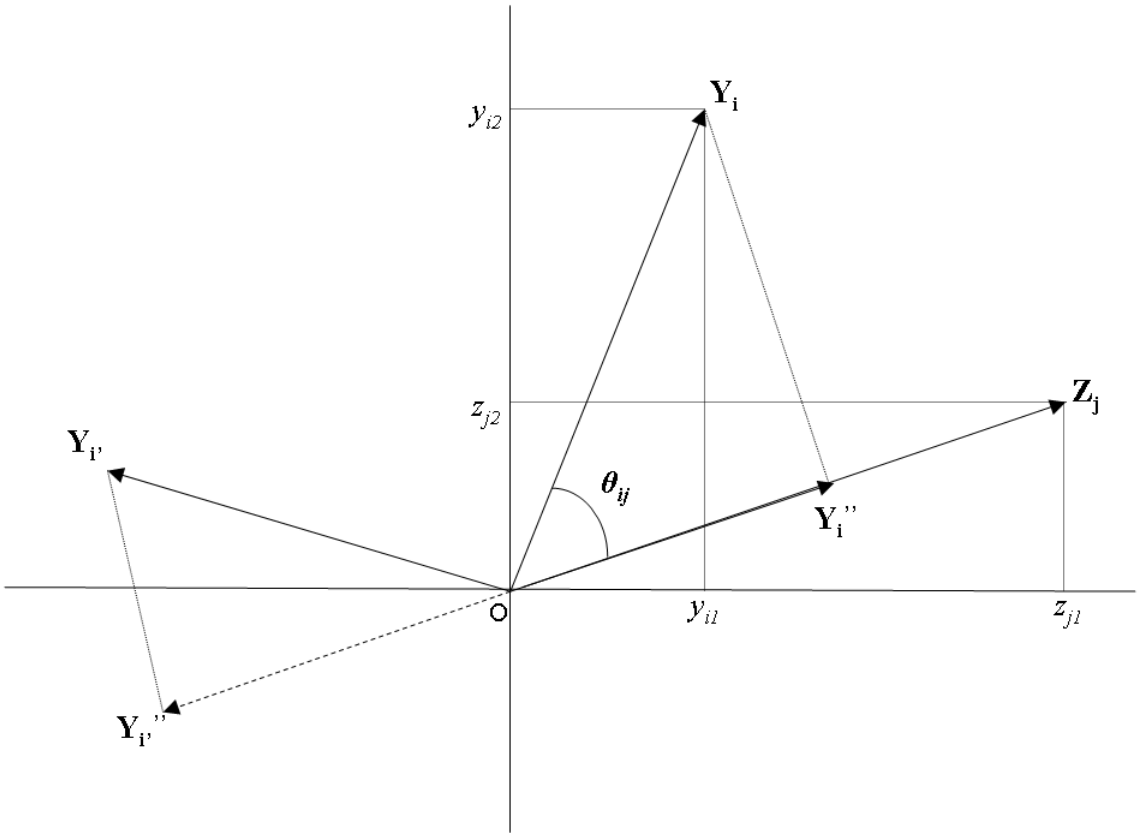

Figura 8 - Representação de dois marcadores de objetos e um marcador de variáveis em um *biplot*

para a projeção ortogonal de Y_i no segmento $0Z_i$, θ_{ij} para o ângulo entre os segmentos $0Y_i$ e $0Z_j$, e escrever $|0Z_j|^2$ para o comprimento do vetor $0Z$ (representação geométrica apresentada na Figura 8), tem-se que :

$$\widehat{x}_{ij} = |0Z_j||0Y_i| \cos(\theta_{ij}) = |0Z_j||0Y_i^{''}| \tag{36}$$

A equação (36) mostra que $\widehat{x}_{ij}$ é proporcional ao comprimento de $0Y_i^{''}$. Assim, a relação ou a interação entre dois objetos com a mesma variável pode ser avaliada comparando os comprimentos das projeções dentro daquela variável. Além disso, a relação ou interação entre um vetor objeto $0Y_i$ e a variável $0Z_j$ é positivo se eles tem um ângulo agudo, e negativo se eles tem um ângulo obtuso. Quando a projeção do marcador Y_i dentro do vetor variável $0Z_j$ coincide com a origem, $\widehat{x}_{ij}$ é igual a zero e o objeto tem aproximadamente o valor médio, para esta variável, se os dados x_{ij} foram centrados. Um valor positivo para $\widehat{x}_{ij}$ indica que o i-ésimo objeto tem um alto escore na variável j, com relação ao escore médio desta, e valores

negativos indicam que o i-ésimo objeto tem um escore relativamente baixo na variável i.

2.9 *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:

$$\boldsymbol{A}^*_r = \left(\frac{I}{J}\right)^{1/4} \boldsymbol{A}\boldsymbol{U}_r\boldsymbol{\Lambda}^{1/2}_r \tag{37}$$

$$\boldsymbol{B}^*_r = \left(\frac{J}{I}\right)^{1/4} \boldsymbol{B}\boldsymbol{V}_r\boldsymbol{\Lambda}^{1/2}_r. \tag{38}$$

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.

3 MATERIAL E MÉTODOS

3.1 Características dos dados

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 1. Têm-se ainda que em cada experimento foi utilizado um delineamento em blocos ao acaso, com 3 blocos em cada experimento.

Tabela 1 - 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.

3.2 Análise de dados considerando duas entradas

Em grande parte dos trabalhos de pesquisa em que se tem três fatores, os pesquisadores consideram a combinação de dois destes, como sendo um fator. No melhoramento genético, em que são considerados como fatores genótipos, locais e anos, os melhoristas têm combinados os fatores locais e anos, sendo que a combinação de cada nível dos locais com cada nível do fator ano gera um nível de um único fator denominado de ambiente. No entanto com esse procedimento perde-se muitas informações de associação entre os níveis de fatores que são combinados, além do real efeito da interação tripla.

3.2.1 Análise de variância conjunta de duas entradas

Com o objetivo de verificar se existe a interação entre genótipos e ambiente, realiza-se uma análise de variância conjunta que envolve o estudo de todos os genótipos em todos os ambientes. De acordo com o descrito por Annichiarico (2002), pode-se assumir que o efeito de genótipos seja fixo e o efeito dos ambientes como aleatório, obtendo o efeito da interação genótipos $\times$ ambientes aleatório. Os dados serão representados pelo seguinte modelo matemático:

$$Y_{ijr} = \mu + g_i + e_j + (ge)_{ij} + b_r(e_j) + \varepsilon_{ijr} \tag{39}$$

sendo que:

Y_{ijr}: é o valor observado do i-ésimo genótipo no j-ésimo ambiente e no r-ésimo bloco, com $i = 1, 2, ..., g$, $j = 1, 2, ..., e$ e $r = 1, 2, ..., b$;

μ: é uma constante, geralmente a média;

g_i: é o efeito do i-ésimo genótipo;

e_j: é o efeito do j-ésimo ambiente;

$(ge)_{ij}$: é o efeito da interação do i-ésimo genótipo com o j-ésimo ambiente;

$b_r(e_j)$: é o efeito do r-ésimo bloco dentro j-ésimo ambiente;

ε_{ijr}: é erro experimental associado ao i-ésimo genótipo, no j-ésimo ambiente e no r-ésimo bloco assumido ser independente e $\varepsilon_{ijr} \sim N(0, \sigma^2)$.

Na Tabela 2 apresenta-se o esquema da análise de variância para o modelo (39) com os graus de liberdade e esperanças dos quadrados médios (KUTNER et. al., 2005).

Tabela 2 - Esquema da análise de variância para experimentos de um mesmo grupo de g genótipos avaliado em e locais com b blocos

Fontes de Variação	Graus de liberdade	E [QM]
Blocos d. ambientes (B d. E)	$e(b-1)$	$\sigma^2 + g\sigma^2_{Bd.E}$
Genótipos (G)	$(g-1)$	$\sigma^2 + be\phi_g + b\sigma^2_{GE}$
Ambientes (E)	$(e-1)$	$\sigma^2 + bg\sigma^2_E$
Interação $(G \times E)$	$(g-1)(e-1)$	$\sigma^2 + b\sigma^2_{GE}$
Resíduo(Res)	$e(g-1)(b-1)$	σ^2
Total	$(geb-1)$	

E[QM]: Esperanças dos Quadrados Médios; $\phi_g = \frac{\sum_{i=1}^{g} g_i^2}{g-1}$

3.2.2 Análises AMMI

Sendo a interação significativa, o próximo passo é fazer a decomposição da $SQ_{G\times E}$, para descartar um resíduo adicional presente nessa soma de quadrados. Essa decomposição é feita utilizando o fator analítico proposto por Gollob (1968) e Mandel (1969, 1971) e tem a seguinte expressão:

$$(ge)_{ij} = \sum_{k=1}^{p} \lambda_k \alpha_{ik} \gamma_{jk}, \tag{40}$$

em que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, e α_{ik}, γ_{jk} satisfazem o contraste de orto-normalização $\sum_i \alpha_{ik}\alpha_{ik}^{'} = \sum_j \gamma_{jk}\gamma_{jk}^{'} = 0$ para $k \neq k^{'}$ e $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$.

Antes de aplicar a decomposição, é necessário organizar os dados em uma tabela de dupla entrada (ou matriz de ordem $g \times e$) com as médias dos b blocos (ou repetições) para cada combinação de genótipos e ambientes:

$$\mathbf{Y}_{g\times e} = \begin{pmatrix} Y_{11} & Y_{12} & ... & Y_{1e} \\ Y_{21} & Y_{22} & ... & Y_{2e} \\ ... & ... & ... & ... \\ Y_{g1} & Y_{g2} & ... & Y_{ge} \end{pmatrix}. \tag{41}$$

O modelo AMMI pressupõe componentes aditivos para os efeitos principais de genótipos e ambientes e componentes multiplicativos para o efeito de interação. Então, a resposta média sobre b blocos do i-ésimo genótipo no j-ésimo ambiente é representada por:

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{q} \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij} + \varepsilon_{ij} \tag{42}$$

sendo que:

Y_{ij}: é a resposta média do i-ésimo genótipo no j-ésimo ambiente, com $i = 1, 2, ..., g$ e $j = 1, 2, ..., e$;

μ: é uma constante, geralmente a média;

g_i: é o efeito do i-ésimo genótipo;

e_j: é o efeito do j-ésimo ambiente;

λ_k: é a raiz quadrada do k-ésimo autovalor da matriz $\boldsymbol{(GE)(GE)^t}$ (ou $\boldsymbol{(GE)^t(GE)}$), com $k = 1, 2, ..., q$ e onde $q < p$ determina uma aproximação de mínimos quadrados para a matriz $\boldsymbol{GE}$ pelos q primeiros termos da DVS e $p = min\{g - 1, e - 1\}$;

α_{ik}: é o i-ésimo elemento do vetor coluna $\boldsymbol{\alpha_k}$ associado a λ_k;

γ_{jk}: é o j-ésimo elemento do vetor linha $\boldsymbol{\gamma_k}$ associado a λ_k;

ρ_{ij}: é o resíduo adicional;

ε_{ij}: é erro experimental associado ao i-ésimo genótipo no j-ésimo ambiente, assumido ser independente e $\varepsilon_{ij} \sim N(0, \frac{\sigma^2}{b})$;

A matriz $\boldsymbol{GE}$ é a interação entre genótipos $\times$ ambientes (matriz de resíduos), em que cada elemento $(ge)_{ij}$ da matriz $\boldsymbol{GE}$ é encontrado pela seguinte relação:

$$(ge)_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..} \tag{43}$$

em que:

Y_{ij}: é a média das repetições do genótipo i no ambiente j, com $i = 1, 2, ..., g$ e $j = 1, 2, ..., e$;

$\overline{Y_{i.}}$: é a média do genótipo i;

$\overline{Y_{.j}}$: é a média do ambiente j;

$\overline{Y_{..}}$: é a média geral do experimento.

Existem várias técnicas estatísticas para selecionar o número de componentes adequado na decomposição da $SQ_{G\times E}$. Entre estes métodos destaca-se o teste F_r proposto por Cornelius (1993) que é considerado um método robusto. A estatística F_r, sob a hipótese nula de que não haja mais do que n termos significativos para a interação, tem uma distribuição F aproximada com $f_2 = (g - 1 - n)(e - 1 - n)$ graus de liberdade e os graus de liberdade do resíduo. Sob essa hipótese, o numerador da expressão (44) é aproximadamente uma variável qui-quadrado com f_2 graus de liberdade(PIEPHO, 1995):

$$F_r = \frac{SQ_{G\times E} - \sum_{k=1}^{n} \lambda_k^2}{f_2 QM_{Res}} \tag{44}$$

em que:

λ_k : é a raiz quadrada do k-ésimo autovalor da matriz $\boldsymbol{(GE)(GE)^t}$;

QM_{Res} : é o Quadrado médio do resíduo;

Assim, um resultado significativo pelo teste sugere que pelo menos um termo multiplicativo ainda deve ser adicionado aos n já ajustados. Logo, F_r pode ser visto como um teste para a significância dos $n+1$ primeiros termos da interação.

3.3 Estimação dos parâmetros dos modelos *multiway*

A seguir é apresentado como se ajusta os modelos descritos anteriormente. Os algoritmos são baseados em Mínimos Quadrados Alternados (MQA). O princípio do MQA é antigo (YATES, 1933) e consiste simplesmente em dividir os parâmetros em vários conjuntos. Cada conjunto de parâmetros é estimado em um sentido de Mínimos Quadrados (MQ) condicionalmente aos parâmetros restantes, ou seja, fixado os demais conjuntos de parâmetros. A estimação dos parâmetros é repetida iterativamente até não observar mudanças significativas nos valores dos parâmetros ou no ajuste do modelo aos dados.

O benefício do algoritmo MQA é a simplicidade envolvida nos passos quando comparados com algoritmos que trabalham no problema inteiro, garantindo a convergência do algoritmo MQA. A desvantagem é que em alguns casos a taxa de convergência pode ser baixa (BRO, 1998), ou seja, os algoritmos procedem com passos muito pequenos provocando pouca melhora por passo, tornando assim a convergência um processo demorado.

Para um arranjo $\underline{\mathbf{X}}$ considere o modelo

$$\underline{\mathbf{X}} \;=\; f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, ...) + \boldsymbol{E}. \tag{45}$$

Para estimar os parâmetros $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, etc. um algoritmo pode ser formulado como:

1. Inicialize os parâmetros $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$;
2. $\boldsymbol{A}$ é a solução para $\min_{\boldsymbol{A}} \|\underline{\mathbf{X}} - f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, ...)\|^2$;
3. $\boldsymbol{B}$ é a solução para $\min_{\boldsymbol{B}} \|\underline{\mathbf{X}} - f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, ...)\|^2$;
4. Estime os conjuntos de parâmetros restantes de forma similar;
5. Retorne para o passo 2 até convergir.

O primeiro passo envolve estimativas iniciais para os parâmetros, sendo que estas estimativas iniciais dependem do modelo que se deseja estudar, mas Kroonenberg (2008)

recomenda escolher várias estimativas iniciais baseadas em alguma lógica matemática e comparar com as estimativas finais. No último passo avalia-se a convergência do algoritmo e, o processo iterativo termina quando não ocorrem mudanças nos parâmetros ou no ajuste do modelo. Nota-se que no algoritmo apresentado acima, os parâmetros foram divididos em conjuntos em que são organizados em matrizes, mas isto é apenas um exemplo e a idéia é dividir os parâmetros em conjuntos tão pequenos quanto possível, a fim de evitar o aparecimento de um mínimo local ou uma baixa taxa de convergência.

É importante destacar ainda que, uma das principais características dos modelos de três entradas é que para fazer o ajuste do modelo é necessário definir o número de componentes antes de executar o algoritmo enquanto que nos modelos de duas entradas, primeiro aplica-se o algoritmo, por exemplo a decomposição em valor singular e depois escolhe o modelo que melhor se ajusta aos dados.

3.3.1 Algoritmo para o modelo PARAFAC

Os algoritmos para ajustar os modelos PARAFAC são usualmente baseados em MQA. Este é vantajoso pois o algoritmo é simples de implementar, simples de incorporar restrições e tem convergência garantida, mas em alguns casos a velocidade com que o algoritmo converge é baixa.

O modelo PARAFAC de três entradas é definido como:

$$\boldsymbol{X}_{(I\times JK)} = \boldsymbol{A}(\boldsymbol{C}\odot\boldsymbol{B})^{'} + \boldsymbol{E} \tag{46}$$

em que $\odot$ é o produto de Khatri-Rao (RAO; MITRA, 1971; SCHOTT, 1997) e a função perda de mínimos quadrados correspondente é:

$$\min_{\boldsymbol{A},\,\boldsymbol{B},\,\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{A}(\boldsymbol{C}\odot\boldsymbol{B})^{'}\|^2. \tag{47}$$

Para ajustar o modelo usando o algoritmo MQA é necessário buscar uma atualização para $\boldsymbol{A}$ dado $\boldsymbol{B}$ e $\boldsymbol{C}$; uma atualização para $\boldsymbol{B}$ dado $\boldsymbol{A}$ e $\boldsymbol{C}$ e uma atualização para $\boldsymbol{C}$ dado $\boldsymbol{A}$ e $\boldsymbol{B}$. Para estimar $\boldsymbol{A}$ dado $\boldsymbol{B}$ e $\boldsymbol{C}$, pode-se formular o seguinte problema

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z}^{'}\|^2 \tag{48}$$

em que $\boldsymbol{Z}=\boldsymbol{C}\odot\boldsymbol{B}$. Então, para dados $\boldsymbol{B}$ e $\boldsymbol{C}$, torna-se um problema de duas entradas para encontrar o mínimo quadrado ótimo $\boldsymbol{A}$ no modelo $\boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z}^{'}+\boldsymbol{E}$, que tem como solução, quando $\boldsymbol{Z}$ é de posto (coluna) completo:

$$\boldsymbol{A}=\boldsymbol{X}\boldsymbol{Z}(\boldsymbol{Z}^{'}\boldsymbol{Z})^{-1}. \tag{49}$$

Por causa da simetria do problema, $\boldsymbol{B}$ e $\boldsymbol{C}$ podem ser atualizadas de maneira similar, fazendo a matrização adequada do arranjo $\underline{\mathbf{X}}$. Assim um algoritmo MQA para ajustar um modelo PARAFAC pode ser descrito como, para certo arranjo $\underline{\mathbf{X}}$ de dimensão $I\times J\times K$ e procurando o ajuste de dimensão R:

1. Inicialize $\boldsymbol{B}$ e $\boldsymbol{C}$;

2. $\boldsymbol{Z}=(\boldsymbol{C}\odot\boldsymbol{B})$;
   $\boldsymbol{A}=\boldsymbol{X}_{I\times JK}\boldsymbol{Z}(\boldsymbol{Z}^{'}\boldsymbol{Z})^{-1}$;

3. $\boldsymbol{Z}=(\boldsymbol{C}\odot\boldsymbol{A})$;
   $\boldsymbol{B}=\boldsymbol{X}_{J\times IK}\boldsymbol{Z}(\boldsymbol{Z}^{'}\boldsymbol{Z})^{-1}$;

4. $\boldsymbol{Z}=(\boldsymbol{A}\odot\boldsymbol{B})$;
   $\boldsymbol{C}=\boldsymbol{X}_{K\times IJ}\boldsymbol{Z}(\boldsymbol{Z}^{'}\boldsymbol{Z})^{-1}$;

5. Retorne ao passo 2 até que as mudanças no ajuste do modelo sejam relativamente pequenas.

3.3.2 Estimativas iniciais para Algoritmo MQA do modelo PARAFAC

Assim, como em qualquer processo iterativo, uma boa estimativa inicial dos parâmetros pode acelerar o algoritmo e diminuir o risco de convergir para um mínimo local ou mesmo para estimativas diferentes da verdadeira. Uma boa estimativa inicial é caracterizada como aquela que conduz diretamente a um mínimo global. Harshman e Lundy (1984) defendem que o algoritmo deve ser iniciado de vários pontos iniciais aleatórios e se a mesma solução é encontrada várias vezes, existe a probabilidade de se ter encontrado um mínimo local, ou seja, valores iniciais aleatórios nem sempre fornecem uma boa estimativa. Smilde et. al. (2004) chamam de início racional (*rational start*) uma estimativa inicial ótima, sendo

que o início racional é baseado em uma solução aproximada baseada nos métodos GRAM (*Generalized Rank Annihilation Method*) ou Decomposição Trilinear Direta (DTD). A vantagem de usar um início racional é que provavelmente a solução verdadeira está próxima da encontrada. Um problema deste método é avaliar se o algoritmo convergiu para o mínimo global ou não. Um procedimento similar ao início racional é o início semi-racional (*semi-rational start*) que é baseado em vetores singulares. Ambos métodos geralmente são bons, mas as vezes eles conduzem a um mínimo local.

Assim, um procedimento recomendado por Smilde et. al. (2004) é ajustar o modelo para o início racional, para o início semi-racional, bem como, para vários pontos iniciais aleatórios. Se a mesma solução é obtida para todos reajustes, este provavelmente não seja mínimo local.

3.3.3 Algoritmo para o modelo Tucker3

Suponha o seguinte modelo de Tucker3:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{G}(\boldsymbol{C} \otimes \boldsymbol{B})^{'} + \boldsymbol{E}. \tag{50}$$

Como cada elemento g_{pqr} do arranjo $\underline{\mathbf{G}}$ de dimensão $P \times Q \times R$ especifica a importância da combinação de cada componente, faz-se necessário estimar cada elemento g_{pqr}. O arranjo núcleo $\underline{\mathbf{G}}$ pode ser determinado condicionalmente em $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ fazendo uma regressão simples dos dados sobre $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ obtendo:

$$\boldsymbol{G}_{(P\times QR)} \;=\; (\boldsymbol{A}^{'}\boldsymbol{A})^{-1}\boldsymbol{A}^{'}\;\boldsymbol{X}_{(I\times JK)}\;[(\boldsymbol{C}^{'}\boldsymbol{C})^{-1}\boldsymbol{C}^{'} \otimes (\boldsymbol{B}^{'}\boldsymbol{B})^{-1}\boldsymbol{B}^{'}]. \tag{51}$$

Para bases ortogonais o arranjo núcleo pode ser encontrado como:

$$\boldsymbol{G}_{(P\times QR)} \;=\; \boldsymbol{A}^{'}\;\boldsymbol{X}_{(I\times JK)}\;(\boldsymbol{C}\otimes\boldsymbol{B}). \tag{52}$$

Originalmente, o algoritmo, proposto para ajustar o modelo de Tucker3 não foi um algoritmo de mínimos quadrados. Posteriormente, um algoritmo baseado no princípio de mínimos quadrados alternados foi introduzido. O algoritmo mais importante é chamado de TUCKALS3 para ajustar o modelo em sentido de mínimos quadrados com vetores de cargas ortogonais (KROONENBERG; de LEEUW, 1980).

Por definição de $\boldsymbol{G}$ em (52) segue que o modelo de Tucker3 do arranjo $\underline{\mathbf{X}}$ com matrizes de cargas ortogonais pode ser escrito como:

$$\boldsymbol{AG}(\boldsymbol{C}^{'}\otimes\boldsymbol{B}^{'}) \;=\; \boldsymbol{AA}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})(\boldsymbol{C}^{'}\otimes\boldsymbol{B}^{'}). \tag{53}$$

Para as matrizes $\boldsymbol{B}$ e $\boldsymbol{C}$ fixas, segue que o valor ótimo de $\boldsymbol{A}$ é encontrado como:

$$\min_{\boldsymbol{A}}\|\boldsymbol{X} - \boldsymbol{AA}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})(\boldsymbol{C}^{'}\otimes\boldsymbol{B}^{'})\|^2$$

que é equivalente a:

$$\min_{\boldsymbol{A}}\|\{\boldsymbol{X}-\boldsymbol{AA}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})(\boldsymbol{C}^{'}\otimes\boldsymbol{B}^{'})\}(\boldsymbol{C}\otimes\boldsymbol{B})\|^2 \tag{54}$$

pois $(\boldsymbol{C}\otimes\boldsymbol{B})$ é ortonormal. Esta expressão pode ser reescrita como:

$$\begin{aligned} & \min_{\boldsymbol{A}}\|\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})-\boldsymbol{AA}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})(\boldsymbol{C}^{'}\otimes\boldsymbol{B}^{'})(\boldsymbol{C}\otimes\boldsymbol{B})\|^2 \\ = \;& \min_{\boldsymbol{A}}\|\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})-\boldsymbol{AA}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})\|^2 \\ = \;& \min_{\boldsymbol{A}}\|(\boldsymbol{I}-\boldsymbol{AA}^{'})\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})\|^2. \end{aligned} \tag{55}$$

A matriz $(\boldsymbol{I}-\boldsymbol{AA}^{'})$ projeta no complemento ortogonal do espaço coluna de $\boldsymbol{A}$, ou seja, $(\boldsymbol{I}-\boldsymbol{AA}^{'})$ produz os resíduos após a projeção no espaço coluna de $\boldsymbol{A}$. Assim, as componentes tem que ser encontradas de tal forma que após projeção de $\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})$, a variação residual é mínima. Isto é o que a análise de componentes principais faz e a solução é tomada como os P primeiros vetores singulares de $\boldsymbol{X}(\boldsymbol{B}\otimes\boldsymbol{A})$.

Para $\boldsymbol{A}$ e $\boldsymbol{C}$ fixas, procedimentos similares podem ser encontrados para $\boldsymbol{B}$ por tomar os Q primeiros vetores singulares da matrix $\boldsymbol{X}_{(J\times IK)}(\boldsymbol{C}\otimes\boldsymbol{A})$. Assim, um algoritmo MQA para ajustar um modelo de Tucker3 com matrizes de cargas ortogonais para um arranjo $\underline{\mathbf{X}}$ de dimensão $I\times J\times K$, sendo que se deseja ajustar um modelo com dimensões $(P\times Q\times R)$, pode ser implementado como:

1. Inicialize $\boldsymbol{B}$ e $\boldsymbol{C}$;
2. $\boldsymbol{A}$ igual a P vetores singulares de $\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})$;
3. $\boldsymbol{B}$ igual a Q vetores singulares de $\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{A})$;

4. $\boldsymbol{C}$ igual a R vetores singulares de $\boldsymbol{X}(\boldsymbol{B}\otimes\boldsymbol{A})$;

5. Retorne ao passo 2 até que as mudanças no ajuste do modelo seja relativamente pequeno;

6. Faça $\boldsymbol{G}=\boldsymbol{A}^{'}\boldsymbol{X}(\boldsymbol{C}\otimes\boldsymbol{B})$.

3.3.4 Estimativas iniciais para o Algoritmo MQA do modelo Tucker3

Para inicializar o algoritmo é necessário fornecer valores iniciais para $\boldsymbol{B}$ e $\boldsymbol{C}$. Usualmente estes valores são tomados como os vetores singulares de uma decomposição em valores singulares de um arranjo matrizado adequadamente. Para grandes arranjos, este procedimento pode consumir tempo e outros procedimentos baseados no cálculo da decomposição em valores singulares de matrizes menores que a dos dados podem ser utilizados (ANDERSSON; BRO, 1998).

3.4 Proposta para o *triplot*

3.4.1 Produto de elementos por elementos de matrizes

O produto de três matrizes para gerar um arranjo de três entradas é possível se o número de colunas da primeira matriz é igual ao número de colunas da segunda matriz que é igual ao número de colunas da terceira matriz. O arranjo resultante terá um número de linhas igual ao número de linhas da primeira matriz, o número de colunas igual ao número de linhas da segunda matriz e o número de tubos será igual ao número de linhas da terceira matriz. Pode ser conveniente chamar a primeira matriz de matriz linha, a segunda de matriz coluna e a terceira de matriz tubo.

Seja as seguintes matrizes linha ($\boldsymbol{A}$), coluna ($\boldsymbol{B}$) e tubo ($\boldsymbol{C}$):

$$\boldsymbol{A}=\begin{bmatrix} 3 & 5 \\ -2 & 1 \end{bmatrix};\quad \boldsymbol{B}=\begin{bmatrix} 2 & 3 \\ -4 & -3 \end{bmatrix};\quad \boldsymbol{C}=\begin{bmatrix} 5 & 2 \\ 3 & -1 \end{bmatrix}.$$

Assim o produto de elementos por elementos das matrizes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ é um

arranjo de três entradas $\underline{\mathbf{Z}}$ de dimensão $(2 \times 2 \times 2)$:

$$\underline{\mathbf{Z}} = \boldsymbol{A} \odot \boldsymbol{B} \odot \boldsymbol{C} = \begin{bmatrix} 3 & 5 \\ -2 & 1 \end{bmatrix} \odot \begin{bmatrix} 2 & 3 \\ -4 & -3 \end{bmatrix} \odot \begin{bmatrix} 5 & 2 \\ 3 & -1 \end{bmatrix}$$

$$Z(:,:,1) = \begin{bmatrix} 60 & -90 \\ -14 & -34 \end{bmatrix}$$

$$Z(:,:,2) = \begin{bmatrix} -3 & -21 \\ -15 & 27 \end{bmatrix}.$$

Assim cada um dos $2 \times 2 \times 2 = 8$ elementos de $\underline{\mathbf{Z}}$ são calculados como:

$$Z_{111} = a_{11}b_{11}c_{11} + a_{12}b_{12}c_{12} = 3 \times 2 \times 5 + 5 \times 3 \times 2 = 60;$$

$$Z_{112} = a_{11}b_{11}c_{21} + a_{12}b_{12}c_{22} = 3 \times 2 \times 3 + 5 \times 3 \times (-1) = -3;$$

$$Z_{121} = a_{11}b_{21}c_{11} + a_{12}b_{22}c_{12} = 3 \times (-4) \times 5 + 5 \times (-3) \times 2 = -90;$$

$$Z_{122} = a_{11}b_{21}c_{21} + a_{12}b_{22}c_{22} = 3 \times (-4) \times 3 + 5 \times (-3) \times (-1) = -21;$$

$$Z_{211} = a_{21}b_{11}c_{11} + a_{22}b_{12}c_{12} = (-2) \times 2 \times 5 + 1 \times 3 \times 2 = -14;$$

$$Z_{212} = a_{21}b_{11}c_{21} + a_{22}b_{12}c_{22} = (-2) \times 2 \times 3 + 1 \times 3 \times (-1) = -15;$$

$$Z_{221} = a_{21}b_{21}c_{11} + a_{22}b_{22}c_{12} = (-2) \times (-4) \times 5 + 1 \times (-3) \times 2 = -34;$$

$$Z_{222} = a_{21}b_{21}c_{21} + a_{22}b_{22}c_{22} = (-2) \times (-4) \times 3 + 1 \times (-3) \times (-1) = 27;$$

desde que $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ tenha posto igual a 2, estas matrizes podem ser representadas por um gráfico de duas dimensões, semelhante ao biplot proposto por Gabriel (1971). Neste caso, como este gráfico de duas dimensões representa o efeito de três fatores, é conveniente chama-lo de *triplot*.

3.4.2 Arranjo de três entradas em um gráfico de duas dimensões

Sejam as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ na forma da Tabela 3 , em que as colunas são nomeadas de x e ye 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 4).

Um gráfico de duas dimensões pode ser obtido se os valores x e y da Tabela 3 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). 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 3 - 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 4 - 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

3.4.3 O produto dos elementos das matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ e suas propriedades

Como indicado anteriormente, o elemento Z_{111} do arranjo $\underline{\mathbf{Z}}$ é calculado como:

$$Z_{111} = a_{11}b_{11}c_{11} + a_{12}b_{12}c_{12} = 3 \times 2 \times 5 + 5 \times 3 \times 2 = 60.$$

Na Figura 10, 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.

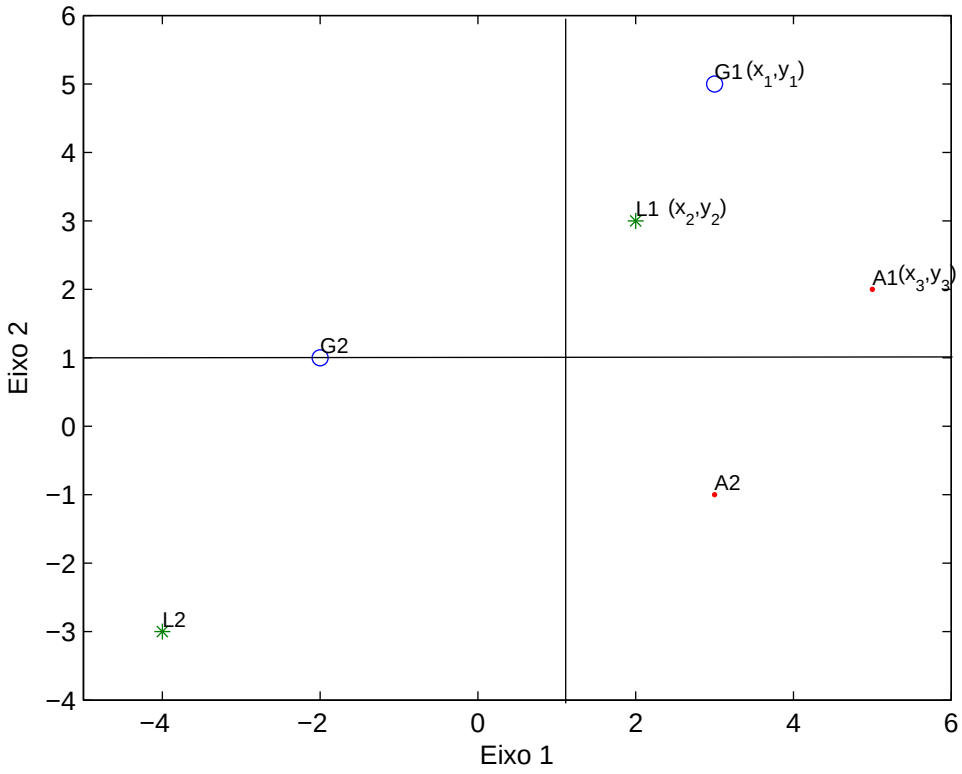

Figura 9 - Um *triplot* que apresenta as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$. Os elementos de $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ são multiplicados segundo o produto de Hadamard para produzir o arranjo $\underline{\mathbf{Z}}$

Da Figura 10, têm-se as seguintes relações:

$$
\begin{aligned}
x_1 &= \overline{OG_1}cos(\alpha_2+\alpha_3) & y_1 &= \overline{OG_1}sen(\alpha_2+\alpha_3)\\
&= \overline{OG_1}cos(\beta_1+\theta_1) & &= \overline{OG_1}sen(\beta_1+\theta_1)\\
x_2 &= \overline{OL_1}cos(\alpha_1+\alpha_2+\alpha_3) & y_2 &= \overline{OL_1}sen(\alpha_1+\alpha_2+\alpha_3)\\
x_3 &= \overline{OA_1}cos(\alpha_3) & y_3 &= \overline{OA_1}sen(\alpha_3)\\
x_2x_3 = u_1 &= \overline{OL_1A_1}cos(\theta_1) & y_2y_3 = v_1 &= \overline{OL_1A_1}sen(\theta_1)
\end{aligned}
$$

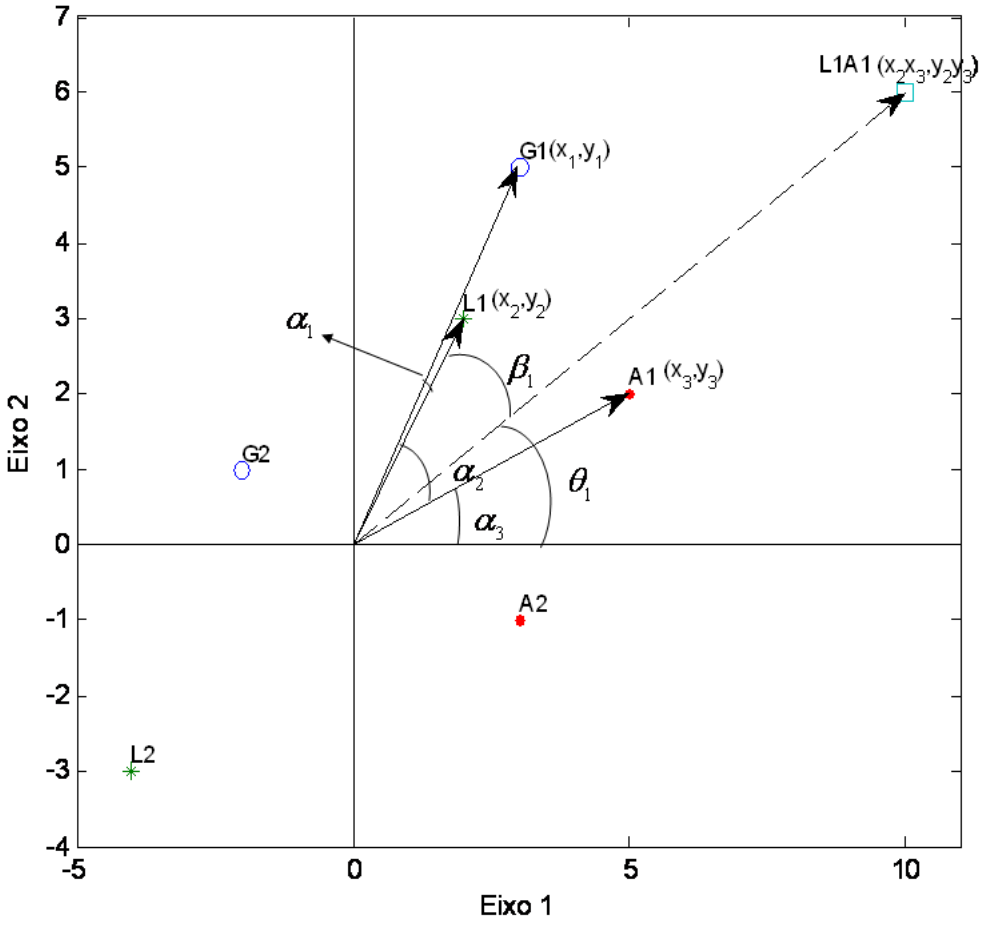

Figura 10 - Os marcadores das linhas, colunas, tubos e combinação de uma coluna com um tubo do arranjo $\underline{\mathbf{Z}}$

Assim,

$$\begin{aligned} z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_1u_1 + y_1v_1 \\ z_{111} &= \overline{OG_1}cos(\beta_1+\theta_1)\overline{OL_1A_1}cos(\theta_1) + \overline{OG_1}sen(\beta_1+\theta_1)\overline{OL_1A_1}sen(\theta_1) \\ z_{111} &= \overline{OG_1}\ \overline{OL_1A_1}cos(\beta_1+\theta_1-\theta_1) \\ z_{111} &= \overline{OG_1}\ \overline{OL_1A_1}cos(\beta_1). \end{aligned} \tag{56}$$

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 10).

Para vizualizar z_{111} diretamente do *triplot*, a equação 56 pode ser escrita como:

$$z_{111} = \overline{OG_1}cos(\beta_1)\overline{OL_1A_1} = \overline{OP_{G_1}}\,\overline{OL_1A_1} \quad (57)$$

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 (56) pode ser escrita como:

$$\begin{aligned} z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_2u_2 + y_2v_2 \\ z_{111} &= \overline{OL_1}cos(\beta_2+\theta_2)\overline{OG_1A_1}cos(\theta_2) + \overline{OL_1}sen(\beta_2+\theta_2)\overline{OG_1A_1}sen(\theta_2) \\ z_{111} &= \overline{OL_1}\,\overline{OG_1A_1}cos(\beta_2+\theta_2-\theta_2) \\ z_{111} &= \overline{OL_1}\,\overline{OG_1A_1}cos(\beta_2) = \overline{OP_{L_1}}\,\overline{OG_1A_1} \end{aligned} \quad (58)$$

ou

$$\begin{aligned} z_{111} &= x_1x_2x_3 + y_1y_2y_3 = x_3u_3 + y_3v_3 \\ z_{111} &= \overline{OA_1}cos(\beta_3+\theta_3)\overline{OG_1L_1}cos(\theta_3) + \overline{OA_1}sen(\beta_3+\theta_3)\overline{OG_1L_1}sen(\theta_3) \\ z_{111} &= \overline{OA_1}\,\overline{OG_1L_1}cos(\beta_3+\theta_3-\theta_3) \\ z_{111} &= \overline{OA_1}\,\overline{OG_1L_1}cos(\beta_3) = \overline{OP_{A_1}}\,\overline{OG_1L_1} \end{aligned} \quad (59)$$

em que $\overline{OP_{L1}}$ 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 (56), (58) e (59) podem ser generalizadas como:

$$z_{ijk} = \overline{OG_i}cos(\beta_{G_i;j*k})\overline{OL_jA_k} \quad (60)$$

$$z_{ijk} = \overline{OL_j}cos(\beta_{L_j;i*k})\overline{OG_iA_k} \quad (61)$$

$$z_{ijk} = \overline{OA_k}cos(\beta_{A_k;i*j})\overline{OG_iL_j} \quad (62)$$

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 $\beta_{G_i;j*k} = \beta_{L_j;i*k} = \beta_{A_k;i*j} = 90^o$;

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.

3.5 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 (60), (61) e (62), 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 $\boldsymbol{C}$.

3.5.1 Comparação visual dos elementos de uma linha, coluna ou tubo do arranjo

A equação (60) pode ser reescrita como

$$z_{ijk} \;=\; \overline{OG_i}cos(\beta_{G_i;j*k})\overline{OL_jA_k} = \overline{OP_{G_{i;j*k}}}\;\overline{OL_jA_k} \tag{63}$$

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:

$$\frac{z_{ijk}}{\overline{OL_jA_k}} \;=\; \overline{OP_{G_{i;j*k}}}\,. \tag{64}$$

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 (61) pode ser escrita como:

$$z_{ijk} = \overline{OL_j} cos(\beta_{L_j;i*k}) \overline{OG_iA_k} = \overline{OP_{L_{j;i*k}}}\ \overline{OG_iA_k} \quad (65)$$

em que $\overline{OP_{L_{j;i*k}}}$ é a projeção $\overline{OL_j}$ no vetor $\overline{OG_iA_k}$. Assim, para certos elementos G_i e A_k, têm-se que $\overline{OG_iA_k}$ é comum para todos os elementos L_j. Dessa forma, a equação (65), pode ser reescrita como:

$$\frac{z_{ijk}}{\overline{OG_iA_k}} = \overline{OP_{L_{j;i*k}}} \quad (66)$$

e a magnitude relativa dos elementos da j-ésima coluna $\underline{\mathbf{Z}}$ para a combinação da linha i com o tubo k, $\frac{z_{ijk}}{\overline{OG_iA_k}}$, pode ser visualizada pela comparação de suas projeções ($\overline{OP_{L_{j;i*k}}}$) sobre o vetor $\overline{OG_iA_k}$.

Novamente, como foi feito para as equações (60) e (61), também pode ser feito para a equação (62):

$$z_{ijk} = \overline{OA_k} cos(\beta_{A_k;i*j}) \overline{OG_iL_j} = \overline{OP_{A_{k;i*j}}}\ \overline{OG_iL_j} \quad (67)$$

em que $\overline{OP_{A_{k;i*j}}}$ é a projeção de $\overline{OA_k}$ no vetor $\overline{OG_iL_j}$. Como para cada elemento G_i e L_j, $\overline{OG_iL_j}$ é comum para todas as linhas A_k, a equação (67) pode ser reescrita como:

$$\frac{z_{ijk}}{\overline{OG_iL_j}} = \overline{OP_{A_{k;i*j}}}\ . \quad (68)$$

Assim, a magnitude relativa dos elementos no k-ésimo tubo de $\underline{\mathbf{Z}}$ para a combinação da linha i com a coluna j, $\frac{z_{ijk}}{\overline{OG_iL_j}}$, pode ser visualizada pela comparação das projeções do vetor $\overline{OA_k}$ sobre o vetor $\overline{OG_iL_j}$.

3.6 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.

3.7 Análise *triplot* de dados de três entradas

Até o momento assumiu-se que as três matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ tinham duas colunas e a combinação delas através de um produto tensorial gerava um arranjo de três entradas $\underline{\mathbf{Z}}$. Quando $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ são apresentados em um gráfico bidimensional, que foi chamado de *triplot*, todos os elementos do arranjo $\underline{\mathbf{Z}}$ podem ser visualizados e comparados através de linhas, colunas e tubos.

Todos os dados de pesquisa experimental de três entradas podem ser visualizados em um *triplot*, como por exemplo, dados de produção de vários genótipos avaliados em vários locais e repetidos por vários anos. Para apresentar um conjunto de dados de três entradas de genótipos $\times$ locais $\times$ anos em um *triplot*, uma condição inicial é achar suas três matrizes de componentes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$. O processo de decompor um arranjo de três entradas em três matrizes de componente é chamada decomposição PARAFAC (página 25) e alguns autores o chamam de decomposição por valores singulares generalizada (SMILDE et. al., 2004), que é essencial para o processo inverso de multiplicação de matrizes.

Assumindo-se que os dados de "g" genótipos testados em "l" locais e "a" anos estão em um arranjo de três entradas $\underline{\mathbf{Z}}$ de dimensão $g \times l \times a$. Quando $\underline{\mathbf{Z}}$ é submetido a decomposição PARAFAC , pode-se obter P componentes principais resultantes, em que P é o posto do arranjo $\underline{\mathbf{Z}}$. Cada componente principal é feito de três partes: uma combinação da importância dos genótipos (a_{ir}), uma combinação da importância dos locais (b_{jr}) e uma

combinação da importância dos anos (c_{kr}). Então, cada elemento de $\underline{\mathbf{Z}}$ é recuperado por

$$z_{ijk} = \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp} \tag{69}$$

em que:

a_{ip}: é o elemento que está na i-ésima linha e na p-ésima coluna da matriz de componentes $\boldsymbol{A}$ do modelo PARAFAC;

b_{jp}: é o elemento que está na j-ésima linha e na p-ésima coluna da matriz de componentes $\boldsymbol{B}$ do modelo PARAFAC;

c_{kp}: é o elemento que está na k-ésima linha e na p-ésima coluna da matriz de componentes $\boldsymbol{C}$ do modelo PARAFAC;

Para representar os dados de genótipo × locais × anos em um *triplot* q-dimensional, deve-se selecionar matrizes $\boldsymbol{A}$ de dimensão $(g \times q)$, $\boldsymbol{B}$ de dimensão $(l \times q)$ e $\boldsymbol{C}$ de dimensão $(a \times q)$. Submetendo um conjunto de dados de três entradas a um modelo PARAFAC, as matrizes resultantes, $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ podem, teoricamente, ser apresentadas em um *triplot* q-dimensional, porém somente os *triplots* bidimensionais são fáceis de visualizar e interpretar. Um *triplot* bidimensional pode ser construído usando os dois primeiros componentes principais, sendo que tal *triplot* é significante quando os primeiros dois componentes principais têm uma aproximação suficiente para a matriz $\underline{\mathbf{Z}}$.

3.8 Análise de dados considerando três entradas

3.8.1 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. Como discutido na seção 2.1.4, Annichiarico (2002) sugere assumir o efeito dos genótipos como fixo, o efeito de locais como aleatório e o efeito de anos também como 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:

$$Y_{ijkr} = \mu + g_i + l_j + a_k + b_r(l_j(a_k)) + (gl)_{ij} + (ga)_{ik} + (la)_{jk} + (gla)_{ijk} + \varepsilon_{ijrk} \qquad (70)$$

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 5 apresenta-se o esquema da análise de variância para o modelo (70), com os graus de liberdade (GL) e esperanças dos quadrados médios ($E[QM]$), de acordo com Kutner et. al. (2005).

Tabela 5 - 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; $\phi_g = \frac{\sum_{i=1}^{g} g_i^2}{g-1}$

3.8.2 Generalização da Análises 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 proposto por Harshaman(1970) e Caroll e Chang (1970), e tem a seguinte expressão:

$$(gla)_{ijk} \;=\; \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp}, \tag{71}$$

P: é o número de componentes em que o arranjo de três entradas da interação tripla pode ser decomposta;

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

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

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

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 (72) o arranjo cúbico é apresentado na forma matrizada):

$$\underline{\mathbf{Y}}_{\,g\times l\times a} = \left(\begin{array}{cccc|cccc|c|cccc} Y_{111} & Y_{121} & ... & Y_{1l1} & Y_{112} & Y_{122} & ... & Y_{1l2} & ... & Y_{11a} & Y_{12a} & ... & Y_{1la} \\ Y_{211} & Y_{221} & ... & Y_{2l1} & Y_{212} & Y_{222} & ... & Y_{2l2} & ... & Y_{21a} & Y_{22a} & ... & Y_{2la} \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ Y_{g11} & Y_{g21} & ... & Y_{gl1} & Y_{g12} & Y_{g22} & ... & Y_{gl2} & ... & Y_{g1a} & Y_{g2a} & ... & Y_{gla} \end{array}\right). \qquad (72)$$

O modelo AMMI para a interação pressupõe componentes aditivos para os efeitos principais de genótipos, locais e anos e componentes multiplicativos para os efeitos de interações duplas e triplas. Então, a resposta média sobre r repetições ou blocos do i-ésimo genótipo no j-ésimo ambiente é representada por:

$$Y_{ijk} = \mu + g_i + l_j + a_k + (gl)_{ij} + (ga)_{ik} + (la)_{jk} + \sum_{p=1}^{Q} a_{ip}b_{jp}c_{kp} + \rho_{ijk} + \varepsilon_{ijk} \qquad (73)$$

sendo que:

Y_{ijk} : é a resposta média do i-ésimo genótipo no j-ésimo local no k-ésimo ano, com $i = 1, 2, ..., g$, $j = 1, 2, ..., l$ e $k = 1, 2, ..., a$;

μ: é uma constante, geralmente a média;

g_i: é o efeito do i-ésimo genótipo;

l_j: é o efeito do j-ésimo local;

a_k: é o efeito do k-ésimo ano;

$(gl)_{ij}$: 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;

a_{ip}: é o i-ésimo elemento da p-ésima coluna da matriz de componentes $\boldsymbol{A}$;

b_{jp}: é o j-ésimo elemento da p-ésima coluna da matriz de componentes $\boldsymbol{B}$;

c_{kp}: é o k-ésimo elemento da p-ésima coluna da matriz de componentes $\boldsymbol{C}$;

ρ_{ijk}: é o resíduo adicional da interação tripla que não é explicado pelos Q primeiros componentes do modelo PARAFAC;

ε_{ijk} : é erro experimental associado ao i-ésimo genótipo no j-ésimo local no k-ésimo ano, assumido ser $\varepsilon_{ijk} \sim N(0, \frac{\sigma^2}{b})$ e todos os ε_{ijk} independentes.

O arranjo cúbico $\underline{\mathbf{Z}}$ é um arranjo com as interações entre genótipos $\times$ locais $\times$ anos (arranjo de resíduos) obtida do modelo (70), ou seja, cada elemento $(gla)_{ijk}$ do arranjo de três entradas $\underline{\mathbf{Z}}$ é estimado pela seguinte relação:

$$\widehat{(gla)}_{ijk} = Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i.k} - \overline{Y}_{.ij} + \overline{Y}_{i..} + \overline{Y}_{.j.} + \overline{Y}_{..k} - \overline{Y}_{...} \tag{74}$$

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.

3.9 Software

Usou-se o Software Matlab (2007) juntamente com *Toolbox N-way* (ANDERSSON; BRO, 2000) para ajustar os modelos PARAFAC e Tucker3, em que os algoritmos para o modelo de Tucker3 e modelo PARAFAC estão implementados. Neste mesmo software foi implementado uma rotina computacional para gerar o gráfico *triplot*, seguindo a proposta aqui apresentada e mostrada em anexo.

4 RESULTADOS E DISCUSSÕES

4.1 Análise de variância conjunta com dois fatores

A análise de variância conjunta dos dados considerando 13 genótipos, 9 ambientes (combinação de 3 locais e 3 anos) é apresentada na Tabela 6. Por esta tabela, percebe-se que os efeitos de genótipos, ambientes e interação genótipos × ambientes são significativos. Neste caso, um dos principais resultados de interesse é a soma de quadrados da interação ($SQ_{G\times E} = 58,64$), que representa 28% da variabilidade total dos dados.

Tabela 6 - Análise de variância conjunta para um conjunto de dados com 13 genótipos avaliados em 9 ambientes com 3 blocos

Fonte de variação	GL	SQ	QM	F	valor-p
B d. E	18	0,70	0,04	1,00	0,4715
G	12	31,60	2,63	4,31	<,0001
E	8	109,95	13,74	343,51	<,0001
G × E	96	58,64	0,61	15,5	<,0001
Resíduo	216	8,51	0,04		
Total	350	209,40			

As estimativas das médias dos genótipos e dos ambientes, com relação a produtividade em toneladas por hectare, são apresentados na Tabela 7. Por esta tabela, observa-se que o genótipo 2 apresentou a maior produtividade, seguidos dos genótipos 8 e 1. Já os genótipos 11, 10 e 12 apresentaram as menores produtividades médias. Com relação aos ambientes, têm-se que os ambientes 9, 6 e 3 (que corresponde ao município de Aquidauana na época das secas) apresentaram as maiores médias e as menores produtividades médias foram averiguadas para os ambientes 8, 2 (município de Dourados na época das secas) e 7 (município de dourados na época das águas).

Tabela 7 - Médias dos genótipos, ambientes e posição das médias em relação a produtividade

Genótipo	Médias	Posição		Ambiente	Média	Posição
G_1	2,11	3		E_1	1,62	6
G_2	2,26	1		E_2	1,56	8
G_3	1,95	4		E_3	2,17	3
G_4	1,65	7		E_4	1,83	4
G_5	1,91	5		E_5	1,83	5
G_6	1,61	9		E_6	2,34	2
G_7	1,64	8		E_7	0,49	9
G_8	2,25	2		E_8	1,56	7
G_9	1,60	10		E_9	2,53	1
G_{10}	1,34	12				
G_{11}	1,59	11				
G_{12}	1,29	13				
G_{13}	1,82	6				

4.2 Análise AMMI e Biplot para dados de duas entradas

Considerando a expressão (43), e ao aplicá-la à matriz de médias (expressão (72)), obtém-se a matriz de interações $\boldsymbol{GE}$, apresentada na Tabela 8.

A próxima etapa da análise corresponde ao ajuste da interação pela decomposição em valores singulares ($\boldsymbol{GE}$=$\boldsymbol{USV}'$), aplicada a matriz $\boldsymbol{GE}$ (Tabela 8). Esta matriz terá posto $p = min(12, 8) = 8$, consequentemente a $SQ_{G\times E}$ pode ser decomposta em até 8 componentes. A decomposição da matriz $\boldsymbol{GE}$ é apresentada a seguir:

Tabela 8 - Valores estimados da interação dupla de 13 genótipos e 9 ambientes (combinação de 3 locais e 3 anos) para a produção em ton/ha

Ambiente/ Genótipo	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
G_1	-0,413	0,473	-0,532	-0,030	0,272	-0,094	-0,312	0,332	0,304
G_2	0,659	-0,210	-0,205	0,169	0,028	-0,251	-0,035	-0,003	-0,152
G_3	0,290	0,071	-0,042	0,126	0,527	-0,068	-0,119	-0,471	-0,314
G_4	-0,058	-0,250	-0,278	0,137	0,593	-0,325	-0,076	0,580	-0,324
G_5	0,202	-0,798	0,936	-0,050	0,045	0,615	-0,311	-0,216	-0,422
G_6	-0,005	-0,172	-0,006	0,024	0,512	0,160	0,037	-0,321	-0,230
G_7	0,134	-0,273	0,349	0,222	-0,537	-0,377	0,846	-0,132	-0,232
G_8	0,771	-0,027	0,260	0,764	0,195	-0,231	-0,558	-0,611	-0,563
G_9	0,792	-0,071	-0,662	0,303	-0,298	0,144	0,248	0,186	-0,641
G_{10}	-0,730	-0,303	-0,083	-0,629	-0,402	0,243	0,187	0,849	0,868
G_{11}	-0,234	0,683	-0,129	-0,295	-0,067	0,409	-0,100	-0,275	0,008
G_{12}	-0,559	0,543	0,039	-0,309	-0,333	-0,367	0,225	-0,234	0,995
G_{13}	-0,848	0,334	0,353	-0,431	-0,535	0,143	-0,032	0,315	0,703

$$\mathbf{U} = \begin{pmatrix} -0,171 & -0,443 & -0,095 & 0,227 & -0,030 & 0,133 & 0,189 & -0,310 \\ 0,193 & -0,100 & 0,193 & -0,076 & -0,069 & 0,276 & -0,456 & 0,485 \\ 0,211 & -0,126 & -0,253 & 0,032 & -0,105 & -0,293 & -0,214 & 0,205 \\ 0,082 & -0,231 & 0,262 & 0,424 & -0,430 & -0,107 & 0,356 & 0,329 \\ 0,197 & 0,709 & -0,148 & 0,331 & 0,124 & 0,063 & -0,042 & 0,071 \\ 0,121 & 0,013 & -0,152 & 0,172 & -0,083 & -0,509 & -0,210 & -0,468 \\ 0,063 & 0,290 & 0,356 & -0,560 & -0,242 & -0,311 & 0,331 & -0,063 \\ 0,437 & -0,015 & -0,306 & -0,145 & -0,143 & 0,555 & 0,154 & -0,306 \\ 0,260 & -0,214 & 0,486 & -0,136 & 0,561 & 0,055 & -0,018 & -0,178 \\ -0,473 & 0,170 & 0,362 & 0,274 & 0,032 & 0,133 & -0,317 & -0,217 \\ -0,109 & -0,138 & -0,340 & -0,062 & 0,557 & -0,260 & 0,172 & 0,317 \\ -0,376 & -0,097 & -0,239 & -0,435 & -0,253 & 0,037 & -0,342 & 0,001 \\ -0,435 & 0,183 & -0,126 & -0,045 & 0,083 & 0,228 & 0,397 & 0,133 \end{pmatrix};$$

$$\mathbf{V} = \begin{pmatrix} 0,583 & -0,056 & 0,190 & -0,197 & 0,194 & 0,295 & -0,459 & 0,368 \\ -0,206 & -0,514 & -0,417 & -0,326 & 0,290 & -0,009 & 0,363 & 0,294 \\ 0,017 & 0,731 & -0,318 & -0,071 & -0,263 & 0,099 & 0,273 & 0,307 \\ 0,366 & -0,121 & 0,023 & -0,166 & -0,214 & 0,280 & 0,342 & -0,686 \\ 0,234 & -0,288 & -0,287 & 0,543 & -0,409 & -0,419 & -0,164 & 0,071 \\ -0,034 & 0,282 & -0,124 & 0,343 & 0,733 & -0,155 & -0,086 & -0,327 \\ -0,103 & 0,098 & 0,452 & -0,449 & -0,076 & -0,674 & -0,026 & -0,050 \\ -0,272 & -0,105 & 0,609 & 0,441 & -0,026 & 0,278 & 0,326 & 0,233 \\ -0,586 & -0,029 & -0,128 & -0,117 & -0,230 & 0,305 & -0,570 & -0,209 \end{pmatrix};$$

$$\mathbf{S} = diag \begin{pmatrix} 3,079 & 1,768 & 1,588 & 1,484 & 1,058 & 0,821 & 0,599 & 0,246 \end{pmatrix}.$$

O desdobramento da $SQ_{G\times E}$, correspondente aos quadrados dos valores singulares, que estão na diagonal da matriz $\boldsymbol{S}$ é apresentado na Tabela 9. Nesta tabela também é apresentado o teste F_r de Cornelius para determinar o número de componentes adequado para explicar a $SQ_{G\times E}$.

Pela Tabela 9, observa-se que o primeiro eixo singular da interação captura $48,5\%$, o segundo $16,0\%$, o terceiro $12,9\%$, o quarto $11,3\%$, o quinto $5,7\%$, o sexto $3,5\%$, o sétimo $1,8\%$ e o oitavo $0,3\%$. Consequentemente, o modelo AMMI com dois componentes explica $64,5\%$ da soma de quadrados da interação entre genótipos e ambientes como resposta padrão e $35,5\%$ desta soma de quadrados é considerada como sendo ruídos presente nos dados. Esta tabela também apresenta os resultados do teste dos resíduo AMMI, correspondentes a cada um dos possíveis modelos AMMI, sendo assim, é possível avaliar todos modelos AMMI através do método de seleção proposto por Cornelius (1993). Portanto, nota-se que o modelo com 6 componentes é o melhor modelo como descritor da resposta padrão, sendo que este modelo consegue explicar $97,9\%$ da $SQ_{G\times E}$.

A última etapa da análise AMMI consiste na representação gráfica dos genótipos e ambientes em um gráfico denominado de *biplot*. Para isso faz-se necessário a determinação das coordenadas para os eixos singulares da interação, ou seja, as matrizes $\boldsymbol{G}$ e

Tabela 9 - Teste F_r de Cornelius para determinar o número de termos significativos para a interação $G \times E$

Eixo	valores singulares(λ_k)	Autovalor (λ_k^2)	Proporção explicada	Proporção acumulada	GL	QM	F	valor-p
1	3,079	9,480	0,485	0,485	77	0,131	9,952	0,000
2	1,768	3,127	0,160	0,645	60	0,116	8,805	0,000
3	1,588	2,523	0,129	0,774	45	0,098	7,471	0,000
4	1,484	2,202	0,113	0,887	32	0,069	5,267	0,000
5	1,058	1,120	0,057	0,944	21	0,052	3,966	0,000
6	0,821	0,675	0,035	0,979	12	0,035	2,661	0,002
7	0,599	0,359	0,018	0,997	5	0,012	0,921	0,468
8	0,246	0,060	0,003	1,000	0	-	-	-

$\boldsymbol{H}'$ tal que $\boldsymbol{GE}=\boldsymbol{GH}'$, em que $\boldsymbol{G}=\boldsymbol{US}^{\frac{1}{2}}$ e $\boldsymbol{H}'=\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{V}'$. Assim, para construir um *biplot* de duas dimensões é suficiente tomar as duas primeiras colunas de $\boldsymbol{G}$ e as duas primeiras linhas de $\boldsymbol{H}'$ para ter os marcadores de genótipos e ambientes, respectivamente, que correspondem a representação gráfica de um modelo AMMI com dois componentes.

Para o conjunto de dados em questão (que é composto por 13 genótipos e 9 ambientes), um *biplot* de duas dimensão representará $64,5\%$ da $SQ_{G\times E}$ de acordo com a Tabela 9, pois neste caso utiliza-se apenas dois primeiros componentes para obter as coordenadas dos genótipos e ambientes. O gráfico *biplot* é apresentado na Figura 11.

A Figura 11 ilustra o *biplot* resultante do conjunto de dados utilizado e a partir dele são feitas as devidas interpretações procurando identificar genótipos e ambientes que menos contribuem para a interação entre genótipos $\times$ ambientes. Logo, por este gráfico, nota-se que os genótipos que menos contribuíram para a interação (pontos próximos da origem, que indica quais são os genótipos estáveis) foram os genótipos G_2, G_3, G_4, G_6, G_7 e G_{11}, mas entretanto para fins de recomendação de cultivares deseja-se uma alta performance na produtividade, que pode ser avaliada pelas médias (DUARTE; VENKOVSKY, 1999). Assim, entre estes genótipos destacam-se os genótipos G_2 e G_3, que tiveram a primeira e a quarta maior

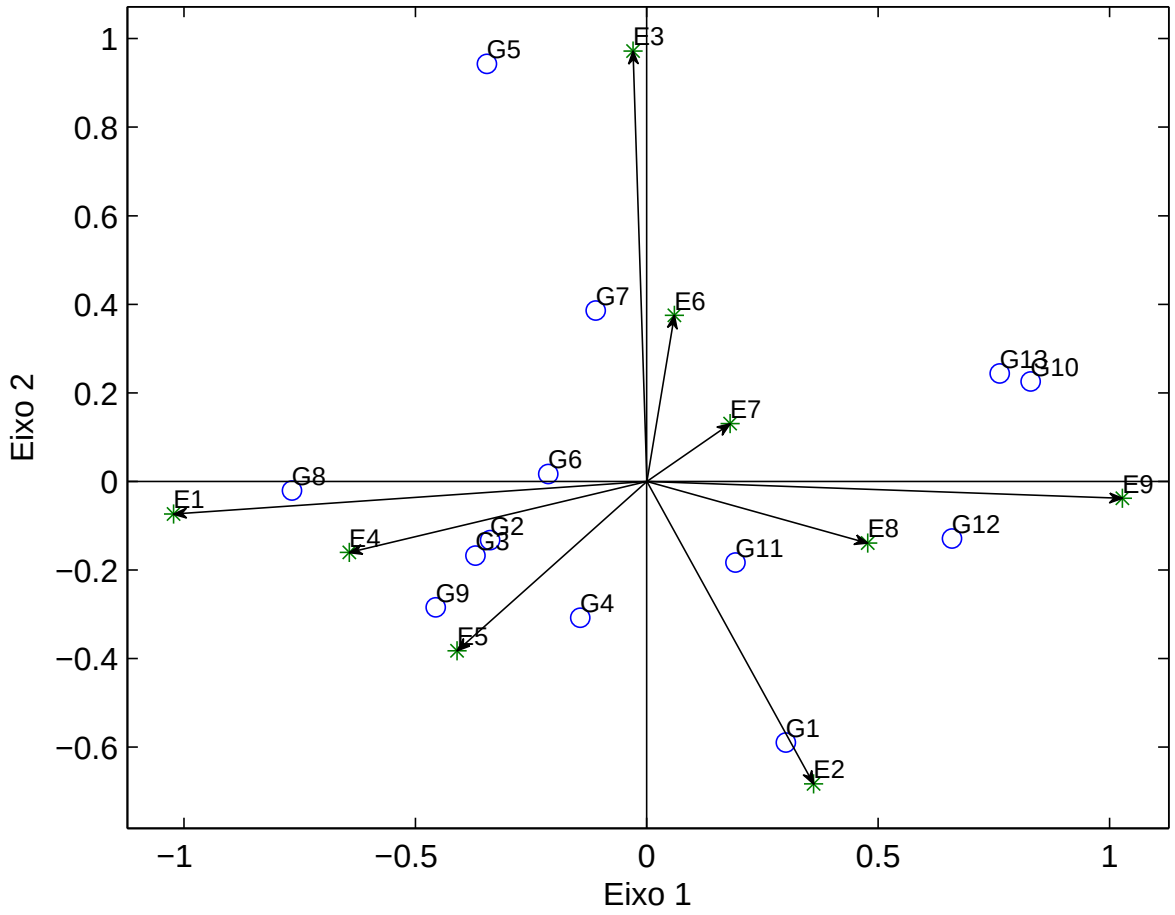

Figura 11 - *Biplot* para os dados de produção de feijão (ton/ha), com 13 genótipos e 9 ambientes

média, respectivamente, em termos da produtividade, enquanto que os genótipos G_4, G_6, G_7 e G_{11}, que também são estáveis, tiveram algumas das piores médias. Os demais genótipos tiveram adaptações específicas a determinados ambientes (coordenadas dos genótipos estão próximas a coordenadas dos ambientes), ou seja, G_1 adaptou-se especificamente ao ambiente E_2, G_5 adaptou-se bem ao ambiente E_3, G_8 aos ambientes E_1 e E_4, G_9 aos ambientes E_4 e E_5, G_{10} ao ambiente E_9, G_{12} aos ambientes E_8 e E_9 e G_{13} ao ambiente E_9.

Por outro lado, o ranqueamento dos genótipos num ambiente estável será de maior confiança para o melhorista. Assim, entre os ambientes destacam-se E_6 (combinação

do local Aquidauana na época das secas com o ano de 2001/2002), E_7 (combinação do local Dourados na época das águas com o ano de 2005/2006) e E_8 (combinação do local Dourados na época das secas com o ano de 2005/2006) como ambientes estáveis (por estarem próximos da origem), enquanto que os demais locais não são estáveis e tem uma grande contribuição para a interação $G \times E$.

4.3 Análise de variância conjunta com três fatores

Pela Tabela 10, que corresponde à parte da 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 10 - 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			

O resultado de maior interesse nessa tabela é a soma de quadrados da interação genótipos × locais × anos, $SQ_{G\times L\times A} = 18,269$, que pode estar inflacionada devido a presença de ruído na variável resposta.

4.4 Modelos de três entradas para a interação tripla

A grande vantagem deste modelo, com relação aos outros modelos multiplicativos de duas entradas, é a possibilidade do estudo simultâneo de diversos fatores. Isso significa, por exemplo, que é possível fazer uma decomposição da interação tripla entre genótipos × locais × anos, tornando as conclusões mais precisas e reais do que aquelas obtidas com modelos multiplicativos para duas entradas.

Desta forma, é possível construir um arranjo cúbico de dimensão $(13 \times 3 \times 3)$ com os efeitos das interações triplas entre genótipos × locais × anos, de modo que nas linhas estão os genótipos, nas colunas os locais e nos tubos os anos. Estes efeitos das interações triplas estão apresentados na Tabela 11.

Tabela 11 - 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

Ao comparar a Tabela 11 com a Tabela 8, percebe-se uma das principais diferenças entre os modelos com três fatores e os modelos com dois fatores. Os resulta-

dos da análise baseados em um modelo com dois fatores, obtido da combinação de outros dois fatores (locais e anos), parecem que não são apropriados, pois os efeitos das interações mostraram-se diferentes. As principais diferenças foram encontradas no genótipo 10, no local 3 e no ano 3, em que a estimação da interação dupla foi de $0,868$ e a interação tripla foi $-0,109$, sendo que esta diferença não está refletida no *biplot* (Figura 11). Resultados similares foram observados para outras combinações, por exemplo: genótipo 8, no local 3 e no ano 3; genótipo 10, no local 1 e no ano 1. De uma maneira geral, ao comparar estas duas tabelas com os efeitos de interação, pode-se observar que há vários valores estimados que são altos para a interação dupla e que não aparecem (estão próximos de zero) para a interação tripla; e por outro lado, existem vários valores estimados que são baixos na interação dupla (próximos de zero) e na interação tripla têm valores altos. Assim, a idéia de fazer a combinação de dois fatores traz prejuízos para a análise da interação entre os fatores em questão (genótipo, local e ano) superestimando ou subestimando a interação. Portanto, faz-se necessário utilizar uma metodologia adequada para interpretar a interação tripla, ou seja, o uso da metodologia *multiway* (por exemplo, modelos Tucker3 e PARAFAC).

4.4.1 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 12. 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 12). 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

Tabela 12 - 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

dif e b_S: são explicados na página 39

($\|\underline{\mathbf{Z}}\|/S_{min} = (SQ_{G\times L\times A}/b)/(min(I;JK)+min(J;IK)+min(K;IJ)-3) = (18,269/3)/12 = 0,5074$), mas como todos os dif_S são maiores 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 13. 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.

O arranjo núcleo $\underline{\mathbf{G}}$ (na tabela 13) 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 13, 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 12), 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 13).

O primeiro *joint plot* (Figura 12) corresponde ao *biplot* da matriz $\boldsymbol{\Delta}_1 = \boldsymbol{A}\boldsymbol{G}_1\boldsymbol{B}'$,

Tabela 13 - Escores dos componentes principais para um modelo de Tucker3 (3,2,2) para o arranjo da interação tripla entre genótipos × locais × anos

Genótipos ($\boldsymbol{A}$)				Locais ($\boldsymbol{B}$)			Anos ($\boldsymbol{C}$)		
	p_1	p_2	p_3		q_1	q_2		r_1	r_2
G_1	0,2633	0,0854	-0,2875	L_1	-0,1498	0,8026	A_1	-0,7903	0,2051
G_2	-0,1587	0,0489	-0,4110	L_2	0,7700	-0,2716	A_2	0,2175	-0,7870
G_3	-0,0006	-0,2277	-0,1433	L_3	-0,6202	-0,5311	A_3	0,5728	0,5819
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	Arranjo núcleo ($\boldsymbol{G}$)					
G_9	-0,1383	0,6940	-0,0839		r_1		r_2		
G_{10}	-0,0376	0,2526	0,0580		q_1	q_2		q_1	q_2
G_{11}	0,3361	0,3268	0,2070	p_1	-1,6769	0,4374		-0,1383	-0,3829
G_{12}	0,5517	-0,3125	-0,0566	p_2	-0,2772	-0,4625		0,9434	0,3448
G_{13}	0,2543	0,0463	0,3161	p_3	0,1206	0,8824		0,3372	0,3579

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 $49,88\%$ 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 12 as seguintes relações, de acordo com o que foi explicada na página 50:

- 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);

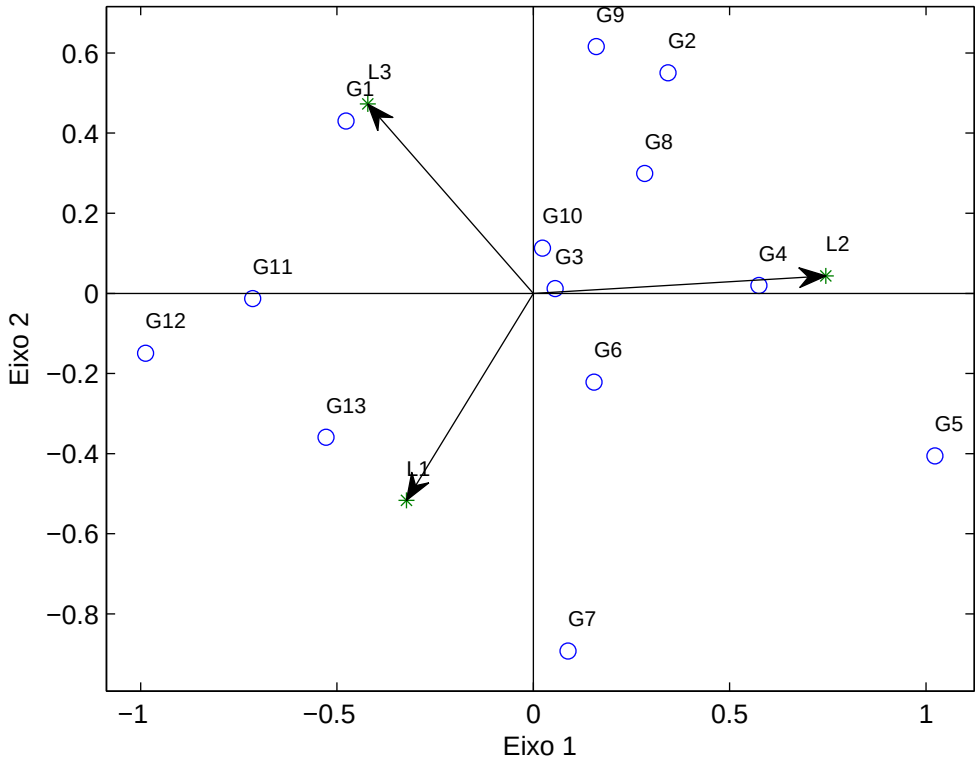

Figura 12 - *Joint plot* projetado dentro da primeira componente do terceiro modo

- 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 $L1$;

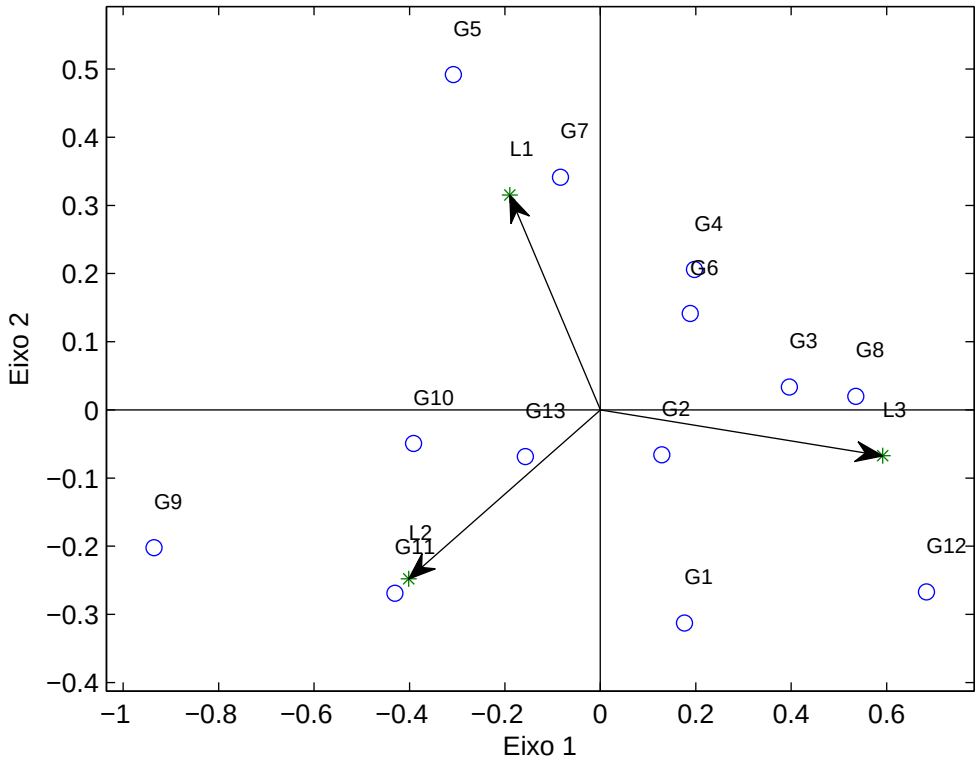

Figura 13 - *Joint plot* projetado dentro da segunda componente do terceiro modo

- 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 13) 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 13 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.

4.4.2 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 14). Estes ajustes foram obtidos utilizando como estimativa inicial os valores obtidos pela Decomposição Trilinear Direta (página 59)e assumindo um critério de convergência de 10^{-6}. 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 14). As estimativas do modelo com dois componentes são apresentadas na Tabela 15, 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 14 - 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

Ao observar os valores da Tabela 15 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

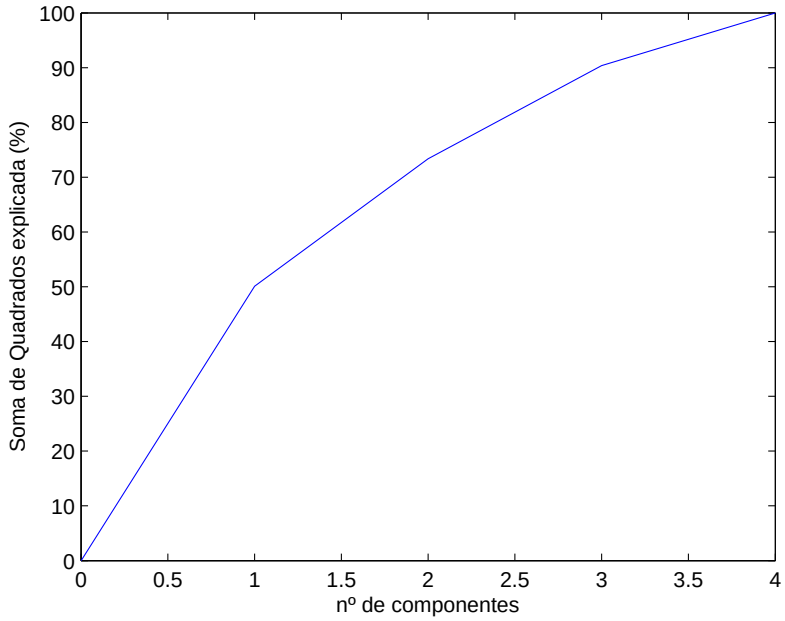

Figura 14 - *Scree plot* do número de componentes no modelo PARAFAC e a porcentagem da soma de quadrados explicada pelo modelo

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.

4.5 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

Tabela 15 - 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						

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 15, 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 $\times$ locais $\times$ 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.

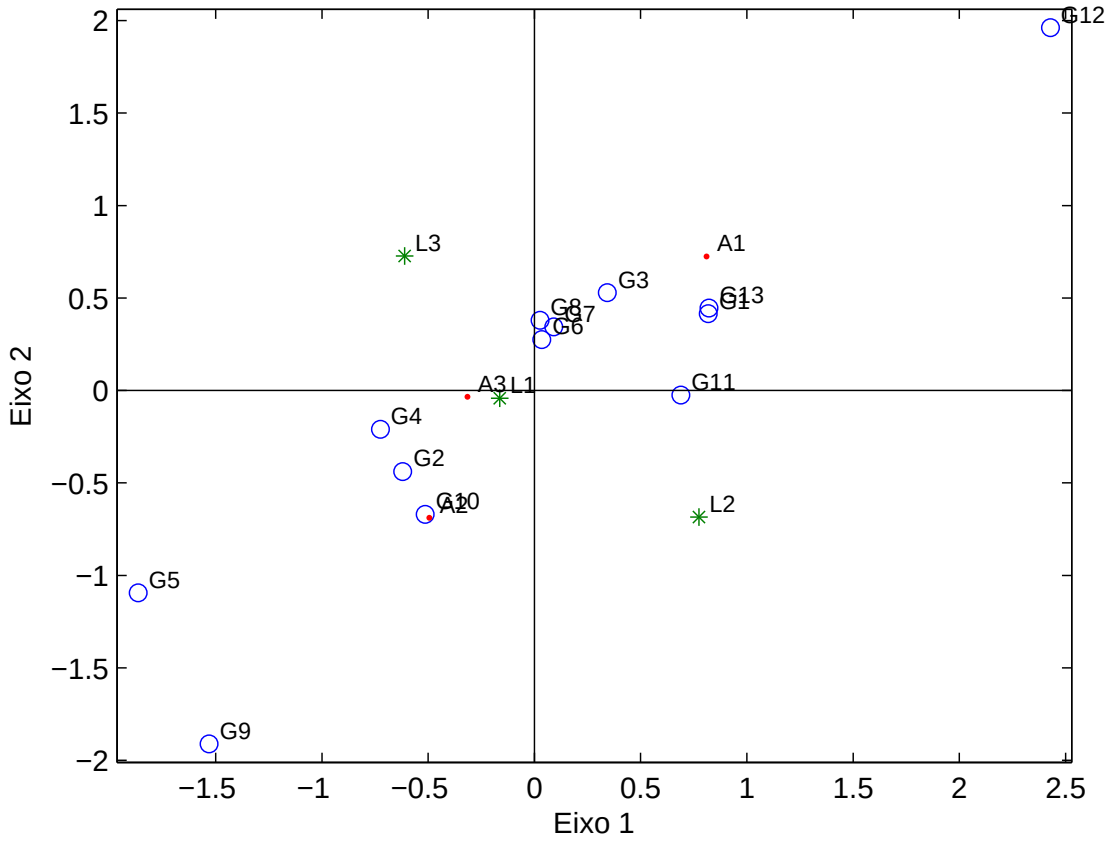

Figura 15 - Triplot para os dados de produção de feijão (ton/ha)

Baseando-se na Figura 15, 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 16.

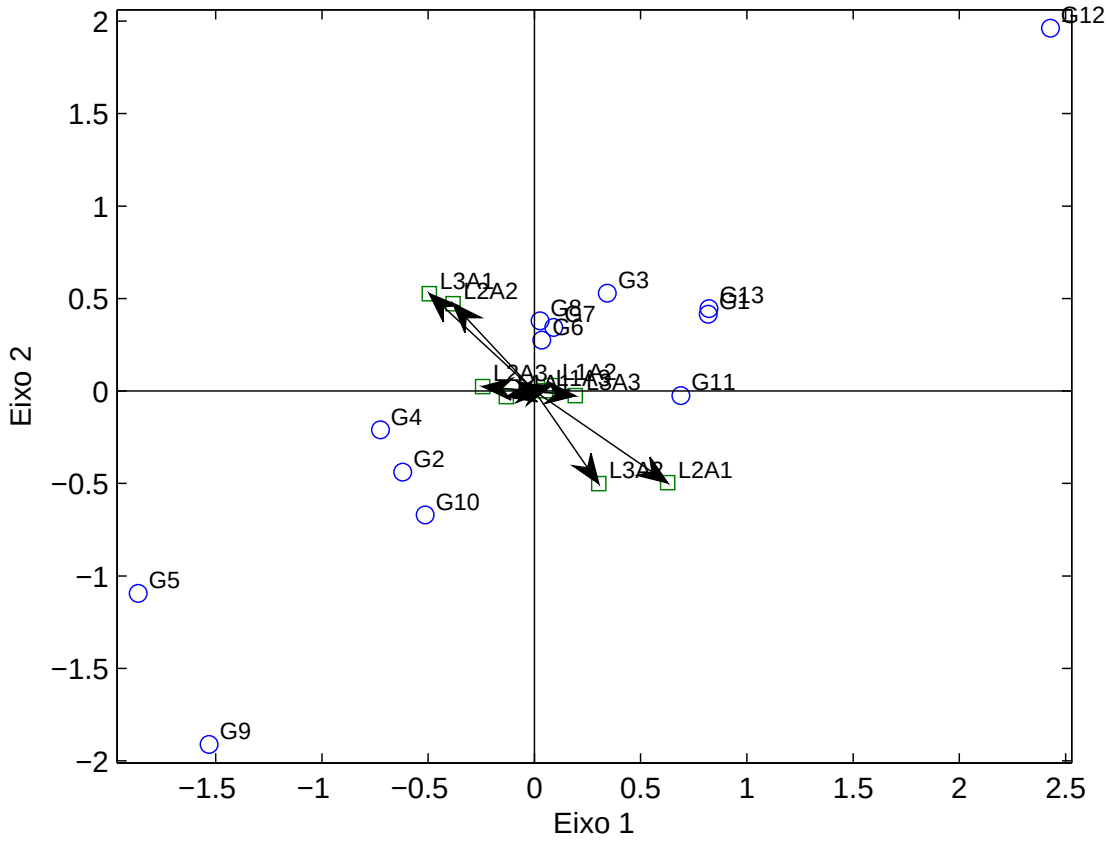

Figura 16 - Triplot combinando os escores do locais e anos para avaliar a adaptabilidade dos genótipos às combinações de locais e anos.

Pela Figura 16, 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 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 11). De forma análoga,

a projeção do genótipo 5 na combinação L2A1 está em direção oposta, portanto a interação tripla é negativa. Logo, o gráfico *triplot*, assim como *joint plot*, representam a verdadeira estruturam da interação tripla em um gráfico de duas dimensões.

4.6 Comentários Gerais

De um modo geral, ao fazer a estimação da interação duas entrada (Tabela 8) ao invés de fazer a estimação da interação de três entradas (Tabela 11), comete-se um erro na obtenção deste efeitos, por exemplo, para o Genótipo 8 no ambiente 8 (combinação do local 2 como o ano 3) a interação dupla é $-0,611$ e a interação tripla este efeito é $0,114$; observa-se que além de diferentes os efeitos têm sinais opostos. Isto acontece pois a expressão (43), que é utilizada para obter interação dupla é diferente da expressão (74) usada para calcular a interação de tripla (que é a verdadeira interação entre os três fatores).

Assim como Varela et al. (2006) encontrou problemas de superestimação e subestimação da interação para o modelo AMMI de duas entradas, quando se faz a combinação de dois fatores (locais e anos), neste trabalho também foi encontrado os mesmos problemas na estimação dos efeitos da interação. Portanto, a utilização desta técnica oferece desvantagens, ou seja, esta metodologia pode fornecer resultados e conduzir a conclusões que não são observáveis no conjunto de dados devido a superestimação e subestimação dos efeitos de interação.

Diante deste problema, faz-se necessário aplicar uma metodologia adequada e conseqüentemente, a extensão dos modelos AMMI de duas entradas para o modelo AMMI de três entradas (tanto utilizando os modelos PARAFAC quanto aos modelos de Tucker3), oferece uma aproximação natural para avaliar a resposta de genótipos em diferentes locais e diferentes anos.

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.

Portanto, de uma forma geral, percebe-se que as vantagens de um modelo supre as desvantagens do outro modelo e vice-versa, logo recomenda-se , sempre que possível, utilizar as duas abordagens para concluir sobre os dados.

Ainda, ao observar os resultados, percebe-se que o genótipo 6, foi o mais estável (o genótipo que menos contribuiu para a interação) e os genótipos 12, 9 e 5 são os que mais contribuíram para a interação.

5 CONCLUSÕES

Os resultados parciais obtidos nessa pesquisa permitem extrair as seguintes conclusões:

a) O uso dos modelos AMMI de duas entradas não é adequada para estudar a interação de três fatores, pelo fato de ocorrer superestimação e subestimação dos efeitos da interação;

b) Os gráficos *triplot* e *joint plot* facilitam o entendimento da interação tripla e traz ao pesquisador informações mais reais, sobre a interação tripla, do que a modelagem AMMI de duas entradas;

c) Estas metodologias também facilitam o entendimento da interação tripla no sentido de separar a resposta padrão do ruído, de modo que nos primeiros eixos há uma maior quantidade de resposta padrão e nos últimos eixos há uma maior quantidade de ruído;

d) O gráfico *triplot* ajuda a identificar genótipos, locais e anos estáveis, dentro de um grande grupo de genótipos, locais e anos, utilizando apenas um gráfico;

e) O gráfico *triplot* pode apresentar uma interpretação difícil, se o número de níveis dos fatores forem grandes, sendo que esta dificuldade pode ser resolvida pelo uso do *joint plot*;

f) De uma maneira geral, recomenda-se utilizar o *triplot* e o *joint plot* juntos para obter melhores interpretações dos resultados;

g) O *Toolbox N-way*, é uma ferramenta importante e de fácil utilização para o ajuste de modelos *multiway*. Conseqüentemente, é fácil fazer uma decomposição da interação entre genótipos $\times$ locais $\times$ anos;

h) Dentre os genótipos estudados, o genótipo 6 foi o que menos contribui para a interação e os genótipos 12, 9 e 5 são os que mais contribuem para a interação.

PROPOSTAS FUTURAS DE PESQUISAS

Como continuidade deste estudo, seria interessante a realização de pesquisas para:

- Comparar qual dos modelos de três entradas (PARAFAC ou Tucker3) é melhor para separar a resposta padrão da resposta ruído;
- Propor um *triplot* baseado no modelo de Tucker3 e comparar com o *triplot* baseado no modelo PARAFAC;
- Propor um teste de hipótese para verificar quais genótipos, locais e anos, contribuem para a interação tripla;
- Propor uma metodologia para o estudo de dados desbalanceados, como por exemplo, quando o número de locais difere de um ano para outro;
- Propor métodos para imputação de dados em um arranjo de três entradas, baseados nos modelos PARAFAC e Tucker3.

REFERÊNCIAS

ALLARD, R.W.; BRADSHAW, A.D. Implications of genotype-environmental interactions in applied plant breeding. **Crop Science**, Madison, v.4, p.503-507, 1964.

ANDERSSON, C.A.; BRO, R. Improving the speed of multi-way algorithms: Part I. Tucker3 **Chemometrics and Intelligent laboratory Systems**, Amsterdam, v.42, p.93-103, 1998.

ANDERSSON, C.A.; BRO, R. The N-way Toolbox for MATLAB **Chemometrics and Intelligent laboratory Systems**, Amsterdam, v.d2, p.1-4, 2000.

ANNICHIARICO, P. **Genotype x Environment Interactions - Challenges and Opportunities for Plant Breeding and Cultivar Recommendations**. Rome: FAO Corporate Document Repository, 2002. 126p.

BARROSO,L.P.; ARTES,R. **Análise Multivariada**. Lavras: SEAGRO, 2003. 156p.

BRO, R. Multiway analysis in the food industry. Models, algorithms and applications. 1998. 290p. Tese(Ph.D thesis) - University of Amsterdam, Amsterdam, 1998.

BRO, R.; KIERS, H.A.L. A new efficient method for determining the number of components in PARAFAC models. **Journal of Chemometrics**, Chichester, v.17, p.273-286, 2003

CARROLL, J.D.; CHANG, J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'EckartYoing´ decomposition. **Psychometrika**, New York, v.35, p.283-319, 1970.

CATTELL, R.B. 'Parallel proportional profiles' and other principles for determining the choice of factors by rotation. **Psychometrika**, New York, v.9, p.267-283, 1944.

CATTELL, R.B. The Three basic factor-analytic research designs - their interrelations and derivatives. **Psychological Bulletin**, Lancaster, v.49, p.499-521, 1952.

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 - plantas**. Rondonópolis: Fundação MT, 2001. p.673-713.

CEULEMANS, E.; KIERS, H.A.L. Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method.**British Journal of Mathematical and Statistical Psychology**, London, v.59, p.133-150, 2006

CORNELIUS, P. L. Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. **Crop Science**, Madison, v. 33, p. 1186-1193, 1993.

DAS, L. D. V. **Genetics and Plant Breeding** . New Delhi: New Age International Publishers, 2005. 396p.

DUARTE, J.B.; VENCOVSKY, R. **Interação genótipo x ambiente:** uma introdução à análise "AMMI". Ribeirão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografias, 9).

EASTMENT, H.T.; KRZANOWSKI, W.J. Cross-validatory choice of the number of components from a principal component analysis. **Technometrics**, Alexandria, v.24, p.73-77, 1982.

ECKART, C.; YOUNG, G. The approximation of one matrix by another of lower rank. **Psychometrika**, New York, v.1, p.211-218, 1936.

GABRIEL, K.R. The biplot graphic display of matrices with applications to principal components analysis. **Biometrika**, Cambridge, v.58, p.453–467, 1971

GALINDO VILLARDÓN, M.P. Una alternativa de representación simultanea: HJ-Biplot. **Qüestiio**, Barcelona, v.10, p.13–23, 1986.

GAUCH, H.G.; ZOBEL, R.W. AMMI analysis of yield trials. In: KANG, M.S.; GAUCH, H.G. **Genotype-by-Environment Interaction**. Boca Raton: CRC Press, 1996. p. 1–40.

GOLLOB, H.F. A statistical model which combines features of factor analitic and analysis of variance techniques. **Psychometrika**, New York, v.33, p.73-115, 1968.

HARSHMAN, R.A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory"multi-modal factor analysis. **UCLA Working Papers in Phonetics**,Ann Arbor ,v.16, p.1-84, 1970.

HARSHMAN, R.A. Determination and proof of minimum uniqueness conditions for PARAFAC1. **UCLA Working Papers in Phonetics**,Ann Arbor ,v.22, p.30-44, 1972.

HARSHMAN, R.A.; DeSARBO, W.S. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In: LAW, H.G.; SNYDER JR. C.W.; HATTIE, J.A.; MCDONALD, R. P. **Research methods for multimode data analysis**. New York: Praeger, 1984. p. 602-642.

HARSHMAN, R.A.;LUNDY, M.E. The PARAFAC model for three-way factor analysis and multidimensional scaling. In: LAW, E.G.; SNYDER, C.W.; HATTIE, J.A.; McDONALD, R.P. **Research Methods for Multimode Data Analysis**. New York: Praeger, 1984. p.122-215.

HO C.N.; CHRISTIAN G.D.; DAVIDSON E.R. Application of the method of rank annihilation to quantitative analyses of multicomponent fluorescence data from the video fluorometer. **Analytical Chemistry**, Washington, v.50, p.1108–1113, 1978.

HO C.N.; CHRISTIAN G.D.; DAVIDSON E.R. Application of the method of rank annihilation to fluorescence multicomponent fluorescence data from the video fluorometer. **Analytical Chemistry**, Washington, v.52, p.1071–1079, 1980.

HO C.N.; CHRISTIAN G.D.; DAVIDSON E.R. Simultaneous multicomponet rank annihilation and applications to multicomponent fluorescence data acquired by the video fluorometer. **Analytical Chemistry**, Washington, v.53, p.92–98, 1981.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical anlysis**. Madison: Prentice Hall, 1998. 816p.

JONES, G.L.; MATZINGER, D.F; COLLINS, W.K. A comparison of Flue-cured tabacco varieties over locations and years with implications on optimum plot allocation. **Agronomy Journal**. Madison, v.52, p.195-199, 1960.

KIERS, H.A.L.; DER KINDEREN, A. A fast method for choosing the numbers of components in Tucker3 analysis. **British Journal of Mathematical and Statistical Psychology**, London, v.56, p.119-125, 2003.

KIERS, H.A.L.; HARSHMAN, R.A. Relating two proposed methods for speedup of algorithms for fitting two- and three-way principal component and related multilinear models. **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v.36, n.1, p.31-40, 1997.

KIERS, H.A.L.; VAN MECHELEN, I. Three-way component analysis: Principles and illustrative application. **Psychological Methods**, Washington, v.6, p.84-110, 2001.

KROONENBERG, P.M. **Three-Mode Principal Component Analysis:** Theory and applications. Leiden: DSWO Press, 1983. 398p.

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. The TUCKALS line: A suite programs for tree-way data analysis. **Computational Statistics and Data Analysis**, Amsterdam,v.18,p.73-96, 1994.

KROONENBERG, P.M. **Applied Multiway Data Analysis**. New Jersey: Wiley-Interscience, 2008. 579p.

KROONENBERG, P.M.; LEEUW, J. Principal component analysis of three-mode data by means of alternating least squares algorithms. **Psychometrika**, New York, v. 45,p.69-97, 1980.

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.

KUTNER, M.H. et. al. **Applied linear statistical models**. New York: McGraw-Hill, 2005. 1396p.

LEVIN, J. Three-mode factor analysis. **Psychological Bulletin**, Lancaster, v.64, p.442-452, 1965.

LOUWERSE, D.J.; SMILDE, A.K.; KIERS, H.A.L. Cross-validation of multiway component models. **Journal of Ckemometrics**, Chichester, v.13, p.491-510, 1999.

MANDEL, J. The partitioning of interactions in analysis of variance. **Journal of Research of the National Bureau of Standards , Series B**, Washington, v.73, p.309-328, 1969.

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

MARÇO, P.H. et al. Exploratory analysis of simultaneous degradation of anthocyanins in the calyces of flowers of the *Hibiscus sabdariffa* species by PARAFAC model. **Analytical Sciences**, Tokyo, v. 21, p. 1523-1527, 2005.

MATLAB. **The Language of technical Computing R2007a**. São Paulo. 2007.

MURPHY, K. R. et al. Optimized parameters for fluorescence-based verification of ballast water exchange by ships. **Environmental Science** & **Technology**, Easton, v.40, p.2357-2362, 2006.

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

RAO, C.R.; MITRA, S.K. **Generalized Inverse of Matrices and its Applications**. New York: John Wiley & Sons, 1971. 240p.

RIU, J.; BRO, R. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v.65, p.35-69, 2003.

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.

SCHOTT J.R. **Matrix Analysis for Statistics**. New York: John Wiley & Sons, 1997. 426p.

SMILDE, A.; BRO, R.; GELADI, P. **Multi-way Analysis With Applications in the Chemical Sciences**. Chichester: John Wiley & Sons, 2004. 369p.

SANCHEZ, E.; KOWALSKI, B.R. Tensorial resolution: a direct trilinear decomposition. **Journal of Chemometrics**. Chichester, v.4, p.29-45, 1990.

THURSTONE, L.L. **The vector of mind**. Chicago: University of Chicago, 1935. 32p.

TEN BERGE, J.M.F. Simplicity and typical rank of three-way arrays, with applications to Tucker-3 analysis with simple cores. **Journal of Chemometrics**, Chichester, v.18, p.17-21, 2004.

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. Implications of factor analysis of three-way matrices for measurement of change. In: HARRIS, C.W. **Problems in measuring change**. Madison: University of Wisconsin Press, 1963. p.122-137.

TUCKER, L. The extension of factor to three-dimensional matrices. In: FREDERIKSEN, N.; GULLIKSEN, H. **Contributions to mathematical psychology**. New York: Holt, Rinehart & Winston, 1964. p.110-127.

TUCKER, L. Some mathematical notes on three-mode factor analysis. **Psychometrika**. New York, v.31, p.279-311, 1966.

VARELA, M. et. al. Analysis of a three-way interaction including multi-attributes. **Australian Journal of Agricultural Research**. Sydney, v.57, p.1185–1193, 2006.

WANSBEEK, T.; VERHEES, J. Models for multidimensional matrices in econometrics and psychometrics. In: COPPI, R.; BOLASCO, S. **Multiway data analysis**, Amsterdam: Elsevier, 1989. p.543-552.

WESSIE, J.; VAN HOUWELINGEN, H. **GEPCAM user's manual:** Generalized principal components analysis with missing values. Utrecht: University of Utrecht, Institute of Mathematical Statistics, 1983. 47p. Technical report.

WRICKE, G.; WEBER, W.E. **Quantitative genetics and selection in plant breeding**. Berlin: Gruyter, 1986. 405p.

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.

YATES, F. The analysis of replicated experiments when to field results are incomplete. **The Empire Journal of Experimental Agriculture**. Oxford, v.1, p.129-142, 1933.

ANEXOS

ANEXO A - Programa em Matlab para construção do *triplot*

```
newData1 = importdata('C:\Lucio\Tese\dados\medias.xls');
vars = fieldnames(newData1);
for i = 1:length(vars)
    assignin('base', vars{i}, newData1.(vars{i}));
end
r=3;
I=max(data(:,3));
J=max(data(:,2));
K=max(data(:,1));
X=reshape(data(:,4),I,J,K);
% dados organizados em um arranjo
X_I_JK=reshape(permute(X,[1,2,3]),I,J*K);
% media da i-ésima fatia horizontal
Xi_barra=mean(X_I_JK',1);
 % arranjo de media da k-ésima fatia horizontal
Xi_barra=reshape((ones(J*K,1)*Xi_barra)',[I,J,K]);
X_J_IK=reshape(permute(X,[2,1,3]),J,I*K);
% media da j-ésima fatia vertical
Xj_barra=mean(X_J_IK',1);
 % arranjo de media da j-ésima fatia vertical
Xj_barra=reshape(kron(kron(Xj_barra',ones(I,1)),ones(1,K)),[I,J,K]);
X_K_IJ=reshape(permute(X,[3,1,2]),K,I*J);
% media da k-ésima fatia frontal
Xk_barra=mean(X_K_IJ',1);
% media geral
X_barra=mean(Xk_barra);
% arranjo de media da k-ésima fatia frontal
Xk_barra=reshape((ones(I*J,1)*Xk_barra),[I,J,K]);
% arranjo de media geral
X_barra=X_barra(1*ones(I,J,K));
Xjk_barra=reshape(kron(reshape(mean(X,1),1,K*J),ones(I,1)),[I,J,K]);
Xik_barra=reshape(kron(mean(X,2),ones(1,J)),[I,J,K]);
Xij_barra=reshape(kron(ones(1,1,K),mean(X,3)),[I,J,K]);
% estimação do efeito de interação tripla
Int=X - Xij_barra - Xik_barra - Xjk_barra + Xi_barra + Xj_barra + Xk_barra - X_barra;
SQInt=r*(sum(sum(sum(Int.^2))));
% Ajuste do modelo PARAFAC
[Factors_P,it,err] = parafac(Int,2);
[A_P B_P C_P]=fac2let(Factors_P);
xgen_P=A_P(:,1);
ygen_P=A_P(:,2);
xamb_P=B_P(:,1);
yamb_P=B_P(:,2);
xyear_P=C_P(:,1);
yyear_P=C_P(:,2);
xmin=min([min(xgen_P),min(xamb_P),min(xyear_P)]);
xmax=max([max(xgen_P),max(xamb_P),max(xyear_P)]);
ymin=min([min(ygen_P),min(yamb_P),min(yyear_P)]);
ymax=max([max(ygen_P),max(yamb_P),max(yyear_P)]);
labels_g=strcat({'G'},num2str((1:I)','%d'));
labels_a=strcat({'L'},num2str((1:J)','%d'));
```

```
labels_y=strcat({'A'},num2str((1:K)','%d'));
figure;
hold;
box;
plot(xgen_P,ygen_P,'o',xamb_P,yamb_P,'*',xyear_P,yyear_P,'.')
text(xgen_P+0.05,ygen_P,labels_g,'fontsize',8,'verticalalignment','bottom');
text(xamb_P+0.05,yamb_P,labels_a,'fontsize',8,'verticalalignment','bottom');
text(xyear_P+0.05,yyear_P,labels_y,'fontsize',8,'verticalalignment','bottom');
line((xmin-1):(xmax+2),((xmin-1):(xmax+2))-((xmin-1):(xmax+2)),'color','k');
line(((ymin-1):(ymax+2))-((ymin-1):(ymax+2)),(ymin-1):(ymax+2),'color','k');
axis([(xmin-0.1) (xmax+0.1) (ymin-0.1) (ymax + 0.1)]);
xlabel('Eixo 1');
ylabel('Eixo 2');

%triplot combinando os fatores locais e anos
xly=kron(xamb_P,xyear_P);
yly=kron(yamb_P,yyear_P);
labels_ly=['L1A1'; 'L1A2'; 'L1A3'; ...
    'L2A1'; 'L2A2'; 'L2A3'; ...
    'L3A1'; 'L3A2'; 'L3A3'];
xmin2=min([min(xgen_P),min(xly)]);
xmax2=max([max(xgen_P),max(xly)]);
ymin2=min([min(ygen_P),min(yly)]);
ymax2=max([max(ygen_P),max(yly)]);
plot(xgen_P,ygen_P,'o',xly,yly,'S')
text(xgen_P+0.05,ygen_P,labels_g,'fontsize',8,'verticalalignment','bottom');
text(xly+0.05,yly,labels_ly,'fontsize',8,'verticalalignment','bottom');
line((xmin2-1):(xmax2+2),((xmin2-1):(xmax2+2))-((xmin2-1):(xmax2+2)),'color','k');
line(((ymin2-1):(ymax2+2))-((ymin2-1):(ymax2+2)),(ymin2-1):(ymax2+2),'color','k');
axis([(xmin2-0.1) (xmax2+0.1) (ymin2-0.1) (ymax2 + 0.1)]);
xlabel('Eixo 1');
ylabel('Eixo 2');
[orx,ory] = dsxy2figxy(gca,0,0);
[lx,ly] = dsxy2figxy(gca,xly,yly);
for i = 1:(J*K)
    annotation('arrow',[orx lx(i)],[ory ly(i)]); % flecha l1
end
```

ANEXO B - Programa em Matlab para construção do *joint plot*

```
newData1 = importdata('C:\Lucio\Tese\dados\medias.xls');
vars = fieldnames(newData1);
for i = 1:length(vars)
    assignin('base', vars{i}, newData1.(vars{i}));
end
r=3;
I=max(data(:,3));
J=max(data(:,2));
K=max(data(:,1));
X=reshape(data(:,4),I,J,K);
% dados organizados em um arranjo
X_I_JK=reshape(permute(X,[1,2,3]),I,J*K);
% media da i-ésima fatia horizontal
Xi_barra=mean(X_I_JK',1);
 % arranjo de media da k-ésima fatia horizontal
Xi_barra=reshape((ones(J*K,1)*Xi_barra)',[I,J,K]);
X_J_IK=reshape(permute(X,[2,1,3]),J,I*K);
% media da j-ésima fatia vertical
Xj_barra=mean(X_J_IK',1);
 % arranjo de media da j-ésima fatia vertical
Xj_barra=reshape(kron(kron(Xj_barra',ones(I,1)),ones(1,K)),[I,J,K]);
X_K_IJ=reshape(permute(X,[3,1,2]),K,I*J);
% media da k-ésima fatia frontal
Xk_barra=mean(X_K_IJ',1);
% media geral
X_barra=mean(Xk_barra);
% arranjo de media da k-ésima fatia frontal
Xk_barra=reshape((ones(I*J,1)*Xk_barra),[I,J,K]);
% arranjo de media geral
X_barra=X_barra(1*ones(I,J,K));
Xjk_barra=reshape(kron(reshape(mean(X,1),1,K*J),ones(I,1)),[I,J,K]);
Xik_barra=reshape(kron(mean(X,2),ones(1,J)),[I,J,K]);
Xij_barra=reshape(kron(ones(1,1,K),mean(X,3)),[I,J,K]);
% estimação do efeito de interação tripla
Int=X - Xij_barra - Xik_barra - Xjk_barra + Xi_barra + Xj_barra + Xk_barra - X_barra;
% Ajuste do modelo Tucker3
P=input('Número de componentes para a 1º entrada: ');
Q=input('Número de componentes para a 2º entrada: ');
R=input('Número de componentes para a 3º entrada: ');
W = [P Q R];
[Factors_T,G,SSE]=tucker(Int,W);
[A_T B_T C_T]=fac2let(Factors_T);
RR=input('Projetar o Joint plot dentro de qual componente da 3º entrada: ');
[U_jp,S_jp,V_jp]=svds(G(:,:,RR));
A_jp=((I/J)^(1/4))*A_T*U_jp*(S_jp.^(1/2));
B_jp=((J/I)^(1/4))*B_T*V_jp*(S_jp.^(1/2));
labels_g=strcat({'G'},num2str((1:I)','%d'));
labels_l=strcat({'L'},num2str((1:J)','%d'));
labels_a=strcat({'A'},num2str((1:K)','%d'));
figure;
```

```
hold
plot(A_jp(:,1),A_jp(:,2),'o',B_jp(:,1),B_jp(:,2),'*')
text(A_jp(:,1),(A_jp(:,2)+0.05),labels_g,'fontsize',8,'verticalalignment','bottom');
text(B_jp(:,1),(B_jp(:,2)+0.05),labels_l,'fontsize',8,'verticalalignment','bottom');
xmin=min(min([min(A_jp(:,1)),min(B_jp(:,1))]));
xmax=max(max([max(A_jp(:,1)),max(B_jp(:,1))]));
ymin=min(min([min(A_jp(:,2)),min(B_jp(:,2))]));
ymax=max(max([max(A_jp(:,2)),max(B_jp(:,2))]));
line((xmin-10):(xmax+10),((xmin-10):(xmax+10))-((xmin-10):(xmax+10)),'color','k');
line(((ymin-10):(ymax+10))-((ymin-10):(ymax+10)),(ymin-10):(ymax+10),'color','k');
axis([ (xmin - 0.1) (xmax + 0.1) (ymin - 0.1) (ymax + 0.1) ]);
xlabel('Eixo 1');
ylabel('Eixo 2');
box;
[orx,ory] = dsxy2figxy(gca,0,0);
[lx,ly] = dsxy2figxy(gca,B_jp(:,1),B_jp(:,2));
for i = 1:J
    annotation('arrow',[orx lx(i)],[ory ly(i)]); % flecha l1
end
```