CRISTIAN TIAGO ERAZO MENDES

AMMI BAYESIANO PARA DADOS ORDINAIS

LAVRAS – MG

2021

CRISTIAN TIAGO ERAZO MENDES

AMMI BAYESIANO PARA DADOS ORDINAIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Márcio Balestre
(in memorian)
Orientador

Dr. Carlos Pereira da Silva
Coorientador
Prof. Júlio Sílvio de Sousa Bueno Filho
Coorientador

LAVRAS – MG

2021

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Mendes, Cristian Tiago Erazo.
AMMI Bayesiano para dados ordinais / Cristian Tiago Erazo Mendes. - 2021.
84 p. : il.

Orientador(a): Márcio Balestre.
Coorientador(a): Carlos Pereira da Silva, Júlio Sílvio de Sousa Bueno Filho.
Tese (doutorado) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. AMMI-Bayesiano. 2. Ensaios Multi Ambientais. 3. Modelos de Limiar. I. Balestre, Márcio. II. da Silva, Carlos Pereira. III. Bueno Filho, Júlio Sílvio de Sousa. IV. Título.

CRISTIAN TIAGO ERAZO MENDES

AMMI BAYESIANO PARA DADOS ORDINAIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 29 de Julho de 2021.

Prof. Dra. Thelma Sáfadi	UFLA
Prof. Dra. Camilla Marques Barroso	UFLA
Prof. Dr. Fábio Mathias Corrêa	Rhodes University
Prof. Dr. Fernando Ribeiro Cassiano	Unimontes

Prof. Dr. Márcio Balestre
Orientador

Dr. Carlos Pereira da Silva
Coorientador

Prof. Júlio Sílvio de Sousa Bueno Filho
Coorientador

LAVRAS – MG
2021

AGRADECIMENTOS

Agradeço primeiramente a Deus, por tudo e todos que Ele colocou no meu caminho ao longo desses 4 anos, sempre me fortalecendo a seguir sempre em frente.

A minha família pela força e confiança, em especial a minha avó Terezinha Soares que esteve ao meu lado mesmo momentos mais dificeis ao longo dessa jornada.

Aos meus colegas do curso de doutorado. Em especial a Vânia, Patrícia, Ernandes, Eleanderson, Rafael pelos momentos de dificuldade que superamos juntos.

Ao colega e amigo Luciano, por todos os ensinamentos, apoio e oportunidades de crescimento que me proporcionou durante o periodo de Doutorado.

Aos meus coorientadores, Carlos Pereira da Silva, e professor Júlio Sílvio de Sousa Bueno Filho, que aceitaram prontamente o desafio de continuar e finalizar o trabalho de TESE, sempre com grandes ensinamentos.

Ao professor Márcio Balestre, com quem tive o prazer de conviver e acompanhar de perto o brilhantismo e energia em suas contribuições que certamente serão sempre lembrados.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES).

À CAPES, pelo apoio e financiamento a pesquisa, que certamente não teria sido concluído se não fosse por esse amparo.

RESUMO

Nesta tese investigamos a implementação do AMMI bayesiano para dados ordinais. Inicialmente fizemos uma revisão sobre aspectos teóricos de inferência bayesiana, da análise de ensaios multi ambientais (MET) e de modelos de limiar. Em seguida apresentamos dois artigos para submissão a revistas científicas. O primeiro é um artigo de revisão sobre o AMMI-bayesiano com exemplificação numérica de sua implementação tecnicamente mais atualizada. O modelo se mostra muito flexível para ajustar dados desbalanceados, não ortogonais e heteroscedásticos, mas depende de respostas contínuas em que se possa supor aproximação normal. O segundo artigo é sobre a análise de dados ordinais na estrutura do modelo AMMI-bayesiano. Um conjunto de dados MET foi artificialmente transformado em respostas ordinais a partir de uma variável observada contínua. Este recurso foi usado para gerar um padrão ouro como referência de análise, que não existe nas aplicações reais. O uso de uma variável latente contínua, usando funções de ligação acumuladas probit permitiu implementar a análise e mostrou-se eficiente em separar genótipos estáveis de instáveis. Com os dados ordinais a interpretação é menos poderosa, porém mais rigorosa e consistente com a análise de dados contínuos.

Palavras-chave: AMMI-Bayesiano, Ensaios Multi Ambientais, Modelos de Limiar, Variáveis Ordinais.

ABSTRACT

In this thesis we study the implementation of Bayesian AMMI for ordinal data. Initially we revisited theoretical aspects of Bayesian analysis, Multi Environment Trials (MET) and threshold models. In the last two sections are presented papers for scientific journals. The first is a review on Bayesian-AMMI literature folowed by a case study of the state of the art implementation. The model has shown flexibility to fit unbalanced, non-orthogonal and heteroscedastic data, but depends on continuous response in which Gaussian assumption is reasonable after scaling. The second deals with Bayesian AMMI to ordinal data. An ordinal data set on MET was artificially generated from continuous responses. This allows for a gold standard on ordinal data analysis, that is not available in actual ordinal data. A latent underlying continuous variable modeled with cumulative probit link allows for a suitable implementation of the analysis. This has shown to be efficient in telling stable from unstable genotypes. Using ordinal models interpretation is less powerful but more rigorous and consistent with continuous data analysis.

Keywords: Bayesian-AMMI, Multi Environment Trials, Ordinal variables, Threshold models.

LISTA DE FIGURAS

Figura 2.1 – Rendimento médio (Rend.) observado para a interação entre genótipos (G) x ambientes (A) 11
Figura 2.2 – Representação de variáveis categóricas na escala subjacente (variável latente) 22

LISTA DE TABELAS

Tabela 2.1 – Tabela de entrada de dados em cada combinação dos níveis de fatores, que pertencem a uma dada categoria de resposta 20

SUMÁRIO

1 INTRODUÇÃO 8

2 REFERÊNCIAL TEÓRICO 10

2.0.1 Interação Genótipo x Ambiente 10

2.1 Inferência Bayesiana 11

2.1.1 Teorema de Bayes 11

2.1.2 Informação *a priori* e *a posteriori* 12

2.1.3 Monte Carlo Via Cadeia de Markov (MCMC) 13

2.1.4 Algoritmo Metropolis-Hasting 15

2.1.5 Algoritmo Gibbs Sampling 15

2.2 Modelo AMMI 16

2.3 Modelo de *Threshold* (limiar) 19

REFERÊNCIAS 23

SEGUNDA PARTE 24

ARTIGO 1 Flexible multi-environment trials analysis via Bayesian-AMMI: A review 26

ARTIGO 2 Amostragem adequada para o AMMI Bayesiano com dados ordinais 48

CONCLUSÃO GERAL 82

1 INTRODUÇÃO

Um das classes de modelos de maior aplicabilidade na análise da interação entre dois fatores qualitativos são os chamados modelos lineares-bilineares. Em especial, quando se modela os efeitos principais de forma aditiva e a interação de forma multiplicativa, temos os chamados AMMI (*Additive Main Effects and Multiplicative Interaction Model*). Neste caso, os parâmetros que modelam a matriz de interações são construídos a partir de sua decomposição em valores singulares.

Estes modelos foram inicialmente aplicados a ensaios não repetidos em que o primeiro parâmetro da interação era associado à estimativa do erro experimental. Mais recentemente foram estendidos para identificar potenciais padrões de interação entre tratamentos e ambientes em experimentos repetidos.

Um tipo especialmente importante de experimento montado em diferentes ambientes são competições de cultivares em fases finais de seleção, com o objetivo de selecionar e recomendar cultivares para diferentes ambientes. Este conjunto de experimentos aleatorizados, repetidos e independentes, mas com o mesmo conjunto de tratamentos em diferentes ambientes, são referidos em genética e melhoramento de plantas como MET (Multi-Environmental Trials). O modelo AMMI vem sendo bastante utilizado para interpretar tais situações, pois permitem identificar grupos de genótipos adequados para ambientes específicos (adaptabilidade), além de identificar os que são relativamente estáveis (estabilidade).

As vantagens oferecidas pela utilização dos modelos lineares-bilineares foram potencializadas pela aplicação da modelagem bayesiana. A análise bayesiana do modelo AMMI foi proposta, aproximadamente, a duas décadas atrás, mas só recentemente adquiriu maior repercussão na literatura com importantes contribuições. Contudo, os trabalhos conduzidos até o momento, utilizando tais modelos, são para dados com respostas contínuas e pressuposição de normalidade. Nesse sentido, seria interessante e útil estender o método a outros contextos.

Em diversos programas de melhoramento, a atribuição de notas ordenadas em escalas subjetivas associadas a caracteres de interesse do melhorista tem um papel muito importante. Como exemplos, podemos destacar o uso de escalas de cinco ou nove pontos para avaliação visual de lesões de doenças em folhas de hortaliças, avaliação de forma comercial ou intensidade de cor de raízes e frutos, estimativa grosseira do grau de cobertura vegetal e parâmetros de propagação em gramíneas forrageiras, avaliação do porte ou outros parâmetros da arquitetura de plantas anuais ou perenes, com objetivos de mecanização ou de aumentar a insolação, avaliação

de qualidades organolépticas dos alimentos processados ou dos frutos *in natura* de espécies alimentares, dentre outras aplicações.

Nestes casos, dois erros de aproximação podem prejudicar a análise, o primeiro seria abusar do uso do teorema central do limite ao se tomar médias de notas em escalas discretas, o que leva a uma estimação imprecisa da distância relativa na população entre estas notas. Para evitar este problema alguns autores recomendam aumentar o número de classes na avaliação, o que nem sempre é conveniente. O segundo problema é a aproximação da proporção populacional em cada classe de nota, que é tão mais precisa quanto menor o número de classes considerado. Para tentar solucionar estes problemas, aprimoraram-se os modelos de análise paramétrica de dados ordinais utilizando variáveis latentes.

Embora o uso de modelos ordinais com variáveis latentes esteja bastante desenvolvido, não vimos aplicações desta técnica na literatura de MET para o estudo de padrões na GEI. Para cobrir esta lacuna, esta tese tem por objetivo estender o modelo AMMI bayesiano para dados ordinais, modelando variáveis latentes tanto para dados simuladas quanto para um ensaio real envolvendo respostas contínuas discretizadas em variáveis ordinais.

No capítulo seguinte apresentamos uma revisão da literatura que embasa a análise bayesiana de dados ordinais e sua possível ligação com o AMMI bayesiano. Nos capítulos subsequentes são apresentados dois artigos, o primeiro consiste em uma revisão da literatura recente sobre o AMMI bayesiano e o segundo trata da análise de MET com dados ordinais utilizando o AMMI bayesiano.

2 REFERÊNCIAL TEÓRICO

2.0.1 Interação Genótipo x Ambiente

A influência do ambiente sobre características fenotípicas tem sido pesquisada continuamente em programas de melhoramento genético com o objetivo de obter as melhores combinações possíveis de genótipos ou cultivares nos ambientes em que se expressa sua mais alta produtividade e também recomendar genótipos com maior adaptabilidade às mais variadas condições ambientais (MALOSETTI; RIBAUT; EEUWIJK, 2013; SILVA; DUARTE, 2006).

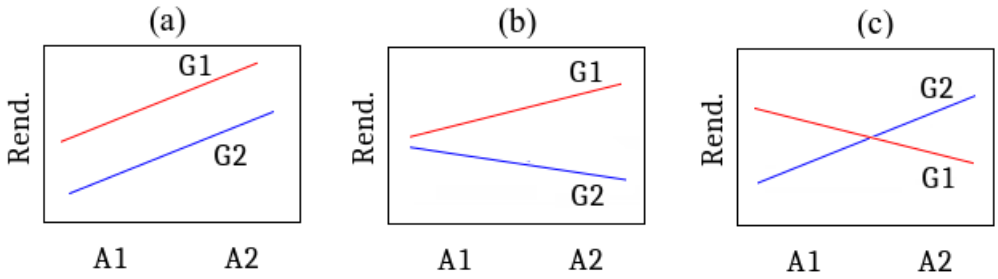
As variações nas respostas dos genótipos aos diferentes ambientes (anos, locais, etc) é a origem da chamada interação entre genótipos x ambientes (*genotype-environment interaction - GEI*). Sua existência dificulta a seleção e a recomendação dos melhores genótipos, mesmo após a realização de experimentos repetidos com a mesma estrutura de tratamentos em vários ambientes, denominados ensaios multiambientais (*Multi Environment Trials - MET*) (YAN; KANG, 2002).

Segundo Crossa (1990), os ensaios multiambientais desempenham um papel fundamental no processo de melhoramento de plantas, sendo possível alcançar uma melhor precisão nas estimativas e predições de rendimentos, a partir da combinação de dados obtidos de experimentos individuais. Além de elucidar padrões de respostas e estabilidade dos genótipos (tratamentos) em diferentes ambientes, pode-se obter informações para auxiliar os melhoristas na seleção de genótipos mais bem adaptados na fase de plantio dos anos seguintes e em novos locais.

Quando a GEI é importante, os efeitos dos genótipos e ambientes não são estatisticamente explicados por modelos aditivos, o que significa que o ranqueamento e desempenho relativo dos genótipos pode ser diferente entre os ambientes.

Na Figura 2.1 são ilustradas 3 situações que podem ocorrer quando dois genótipos (G1 e G2) são avaliados em dois ambientes diferentes (A1 e A2). A figura 2.1a representa o caso onde não há interação, os genótipos se comportam de forma proporcional quando comparados nos diferentes ambientes (paralelismo, que pode no entanto ser afetado pela escala). Na figura 2.1b já observa-se interação onde o desempenho dos genótipos variam, mas sem a mudança na classificação de rendimentos dos genótipos aos diferentes ambientes. Já na figura 2.1c observa-se interação complexa (ou cruzada), em que ocorreu mudança na classificação de rendimentos dos genótipos, sendo esse cenário o mais complicado em termos de seleção e recomendação de genótipos superiores com ampla adaptabilidade aos diferentes ambientes de teste. O mesmo padrão pode ser estendido na presença de vários níveis para cada fator.

Figura 2.1 – Rendimento médio (Rend.) observado para a interação entre genótipos (G) x ambientes (A)



2.1 Inferência Bayesiana

A inferência estatística tem o objetivo de fazer afirmações sobre as características populacionais (não observáveis ou subjacentes) com base em medidas observáveis associadas de uma amostra desta população. Existem vários paradigmas ou estratégias de inferência diferentes, por exemplo a inferência frequentista que considera os parâmetros populacionais constantes fixas desconhecidos e utiliza as informações observadas como a única forma quantificável para o ajuste de distribuições probabilísticas. A generalização neste contexto se dá em imaginar as distribuições de estatísticas em potenciais amostras aleatórias associadas a estes dados. Probabilidades neste paradigma de inferência são interpretadas como frequências relativas em infinitas observações destas amostras (BERNARDO; SMITH, 1994; KOCH, 2007).

Já na inferência bayesiana distribuições de probabilidade não estão associadas apenas a observações serem eventos aleatórios, mas também a declarações ou proposições prévias sobre os próprios parâmetros populacionais que se deseja conhecer. Desta forma é possível quantificar a distribuição conjunta de parâmetros e observações e, mais importante, combinar o conhecimento prévio ao experimental sobre o fenômeno investigado por meio do teorema de Bayes (KOCH, 2007).

2.1.1 Teorema de Bayes

Seja $\mathbf{y}' = (y_1, ..., y_n)$ um vetor de n observações, e $p(\mathbf{y}|\theta)$ a probabilidade condicional de $\mathbf{y}$ dados os k parâmetros do vetor $\theta = (\theta_1, ..., \theta_k)$. Seja $p(\theta)$ a distribuição (caso discreto) ou de densidade (caso contínuo) de probabilidade (*a priori*) de θ e $p(\mathbf{y}, \theta)$ a probabilidade conjunta entre $\mathbf{y}$ e θ, podemos obter as seguintes relações:

$$p(\mathbf{y}|\theta)p(\theta) = p(\mathbf{y}, \theta) = p(\theta|\mathbf{y})p(\mathbf{y}). \tag{2.1}$$

Então, por meio do teorema de Bayes, conhecidas as observações y, a probabilidade condicional de θ é:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{2.2}$$

em que

$$p(\mathbf{y}) = E[p(\mathbf{y}|\theta)] = c^{-1} = \begin{cases} \int p(\mathbf{y}|\theta)p(\theta)d\theta & , \ \theta \ continuo \\ \sum p(\mathbf{y}|\theta)p(\theta) & , \ \theta \ discreto \end{cases} \tag{2.3}$$

Assim, utilizando o teorema de Bayes, a (2.3), pode ser escrita da seguinte forma:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta), \tag{2.4}$$

Ou seja, $p(\theta|\mathbf{y})$ é chamada de distribuição *a posteriori* dos parâmetros dadas as observações, que é proporcional tanto ao conhecimento *a priori* quanto ao conhecimento trazido pelo experimento. É possível calcular a constante normalizadora c, necessária para assegurar que a função $p(\theta|\mathbf{y})$ seja uma distribuição de probabilidades.

A função $p(\mathbf{y}|\theta)$ é a verossimilhança, função de evidência obtida ao se mapear a probabilidade de se encontrar $\mathbf{y}$ para o domínio dos valores paramétricos θ.

2.1.2 Informação *a priori* e *a posteriori*

A inferência condicional usando o teorema de Bayes pode ser considerada como uma atualização da informação prévia promovida pela observação de novos dados. A quantificação do valor relativo destas informações prévias é feita por distribuições de probabilidade. Desta forma torna-se crítica a escolha de quais distribuiçõe usar: tanto para representar a informação prévia quanto para modelar a ocorrência dos dados. Geralmente esse processo tenta conciliar motivações filosóficas e necessidades práticas. A verossimilhança depende da natureza dos dados, de sua forma de medida, de seu delineamento amostral ou experimental e das possíveis distribuições que podem modelar a ocorrência de tais observações, mas também da facilidade algébrica ou computacional em lidar com tais distribuições. As distribuições a priori dependem fundalmentalmente do conhecimento prévio sobre os processos do pesquisador (ou seu grau de crença nos valores paramétricos), do desejo do pesquisador em informar sobre o conhecimento prévio (para muitos fins o pesquisador deseja ignorar o conhecimento prévio e escolher

distribuições a priori que reflitam isso, bem como das propriedades algébricas e, ou numéricas resultantes da combinação das probabilidades e da natureza dos parâmetros da verossimilhança (PRESS, 2002).

A definição da distribuição de probabilidade *a priori* a ser utilizada é, portanto, um passo crucial. Como essa função expressa informações em relação ao grau de conhecimento dos parâmetros, é relativamente difícil caracterizar qual deve ser a distribuição não-informativa, ao passo que se existir alguma informação disponível sobre esses parâmetros (histórica ou de especialistas), é importante utilizar distribuições de probabilidade que reflitam adequadamente este conhecimento (KOCH, 2007).

Duas estratégias muito utilizadas para o estabelecimento de prioris não-informativas, são o Princípio da Razão Insuficiente de Bayes-Laplace que considera a mesma probabilidade para todas as quantidades envolvidas no estudo (priori uniforme) e o método de Jeffreys que minimiza a medida da informação de Fisher (GELMAN et al., 2004). Há também métodos baseados na maximização da distância entre a priori e a posteriori ou maximização da entropia da própria priori (JAYNES, 2003; BERNARDO; SMITH, 1994; PRESS, 2002).

Por outro lado, é muito utilizada a estratégia de se procurar distribuições *a priori* que pertencem a famílias conjugadas, que possuam a característica de preservar a distribuição a posteriori na sua mesma classe. Isto leva a uma facilidade algébrica definida como ”conjugação” e facilita a inferência pois os parâmetros da posteriori são apenas atualizações dos hiperparâmetros da priori na mesma forma algébrica. Sendo assim uma família de distribuições H é conjugada com um modelo observacional $F = p(y|\theta) : \theta \in \Theta$ se a relação é dada da seguinte forma:

$$h(\theta) \in H \Rightarrow h(\theta|y) \propto h(\theta) f(y|\theta) \in H$$

A conjugação traz propriedades bastante interessantes por facilitar o tratamento analítico da distribuição conjunta *a posteriori* e portanto, facilitar a obtenção de seus resumos (GAMERMAN; LOPES, 2006; PAULINO; MURTEIRA; TURKMAN, 2018).

2.1.3 Monte Carlo Via Cadeia de Markov (MCMC)

É frequente que ao se usar inferência bayesiana em problemas complexos não se consiga obter estimadores algébricos e se procure montar amostras da distribuição *a posteriori* conjunta. Isso pode acarretar sobrecarga de cálculos na implementação computacional quando

técnicas mais simples de integração não podem ser utilizadas. O método de Monte Carlo via cadeias de Markov visa gerar valores de amostras da posteriori em cadeias de Markov. Isto é feito simulando observações com o método Monte Carlo nas distribuições convenientes para a implementação da posteriori. Frequentemente este processo é subdividido aproveitando-se propriedades de distribuições condicionais para facilitar a amostragem conjunta.

Suponha que se queira gerar uma amostra de uma distribuição a posteriori $p(\theta|y)$ para $\theta \in \Theta \subset \Re^k$, cuja amostragem direta é difícil de se obter. Pode-se, no entanto, gerar uma cadeia de Markov a partir de um estado inicial θ_0 no domínio conveniente e construir a cadeia em um processo iterativo que segue até se obter a distribuição de equilíbrio para $p(\theta|y)$, considerando eventualmente critérios de parada (BERNARDO; SMITH, 1994).

Sob condições de regularidades adequadas, pode-se verificar por meio de resultados assintóticos, que a saída da amostra de uma cadeia com distribuição de equilíbrio pode ser usada para simular uma amostra aleatória de $p(\theta|y)$ ou para estimar o valor esperado, em relação a $p(\theta|y)$, de uma função $g(\theta)$ de interesse.

Seja $(\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, ...)$ um ponto amostral da cadeia. A avaliação assintótica da cadeia com $t \to \infty$ possui as seguintes propriedades:

i) $\theta \to \theta \sim p(\theta|x)$, em distribuição,

ii) $\frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \to E_{\theta|x}[g(\theta)]$

Como amostras sucessivas da cadeia $\theta^{(t)}$ são correlacionadas, o processo de estabelecimento de resumos da posteriori conjunta fica simplificado se estabelecerem amostras independentes. Isso é feito interrompendo-se a passos arbitrários a série temporal de amostras da cadeia para construir a distribuição conjunta de interesse $p(\theta|\mathbf{y})$. Além disso, funções de interesse dos parâmetros amostrados nas cadeias ganham também distribuições nas quais se pode realizar a inferência (BERNARDO; SMITH, 1994).

Para a construção de cadeias e realização de amostras das distribuições de equilíbrio, existem dois algoritmos bastante utilizados em aplicações da estatística bayesiana. São eles: o algoritmo Metropolis-Hasting e o amstrador de Gibbs, que serão detalhados a seguir.

2.1.4 Algoritmo Metropolis-Hasting

O algoritmo de Metropolis-Hasting é frequentemente utilizado no processo de obtenção de distribuições a posteriori na estatística bayesiana. Ele é mais usado quando não se consegue obter uma amostragem direta da distribuição de destino. A idéia desse algoritmo é gerar valores a partir de distribuições auxiliares (geradoras de propostas) e aceitá-los com uma dada probabilidade. Esse mecanismo de correção é utilizado para que os valores gerados se comportem assintoticamente como observações aleatórias da distribuição alvo.

Cada estado da cadeia de Markov produzida pelo algoritmo Metropolis-Hastings é gerado alternando-se dois passos: 1) a etapa de proposição e 2) a etapa de transição. Estas etapas estão associadas, respectivamente à distribuição geradora de proposições de pontos amostrais e ao cálculo da probabilidade de transição do algoritmo Metropolis-Hastings. A seguir, apresenta-se a descrição dos passos do algoritmo (SORENSEN; GIANOLA, 2002; PRESS, 2002)

Seja $\theta^{(0)}$ o valor inicial da cadeia e suponha que o algoritmo foi executado para obter os valores $\theta^{(0)}, \theta^{(1)}, ..., \theta^{(j-1)}$.

O próximo item da cadeia $\theta^{(j)}$ é produzido pela alternância das duas etapas:

Etapa de proposição: amostre um valor θ da distribuição proposta $q(\theta^{(j-1)}, \theta|y)$ e calcule o valor:

$$\alpha(\theta^{(j-1)}, \theta|\mathbf{y}) = min\left\{1, \frac{h(\theta|y)}{h(\theta^{j-1}|y)}\frac{q(\theta, \theta^{(j-1)}|y)}{q(\theta^{(j-1)}, \theta|y)}\right\} \tag{2.5}$$

em que $h(\cdot)$ é a distribuição de interesse.

Etapa de transição: amostre um valor $u \sim U(0,1)$. Se $u \leqslant \alpha$ então $\theta^{(j)}$ é igual a θ, caso contrário, mantenha $\theta^{(j)}$ igual a $\theta^{(j-1)}$ e assim, repita as etapas novamente até a convergência da cadeia.

Quando as densidades das distribuições geradoras de propostas são simétricas e $q(\theta, \theta^{(j-1)}|y) = q(\theta^{(j-1)}, \theta|y)$, a probabilidade de aceitação reduz-se a

$$\alpha(\theta^{(j-1)}, \theta|\mathbf{y}) = min\left\{1, \frac{h(\theta|y)}{h(\theta^{j-1}|y)}\right\} \tag{2.6}$$

que corresponde ao caso originalmente considerado por Metropolis et al. (1953).

2.1.5 Algoritmo Gibbs Sampling

O amostrador de Gibbs é um dos algoritmos de mais simples implementação computacional para os métodos MCMC, embora envolva maior complexidade teórica. Este método pode

ser considerado um caso particular do algoritmo Metropolis-Hastings no qual a etapa de proposição é feita escolhendo distribuições *a posteriori* condicionais completas convenientes. Isto leva a probabilidade de aceitação a 100% e potencialmente acelera o processo de amostragem. A seguir apresentam-se os passos para implementação deste algoritmo..

Considere o vetor de parâmetros de um modelo $(\theta_1, \theta_2, ..., \theta_p)$, com densidade *a posteriori* conjunta $h(\theta_1, \theta_2, ..., \theta_p|y)$. Sejam os valores iniciais arbotrários $(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$ no domínio da densidade conjunta $h(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}) > 0$. Simula-se sequencialmente cada elemento do vetor de parâmetros a partir de cada uma das distribuições *a posteriori* condicionais completamente condicionadas, não impostando a ordem da amostragem, ou seja:

$$\theta_1^{(1)} \sim h(\theta_1|\theta_2^{(0)}, ..., \theta_p^{(0)}, y),$$

$$\theta_2^{(1)} \sim h(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_p^{(0)}, y),$$

$$\theta_3^{(1)} \sim h(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, ..., \theta_p^{(0)}, y),$$

$$\vdots$$

$$\theta_3^{(1)} \sim h(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, ..., \theta_p^{(0)}, y),$$

$$\theta_p^{(1)} \sim h(\theta_p|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(1)}, ..., \theta_{p-1}^{(1)}, y),$$

$$\theta_1^{(2)} \sim h(\theta_1|\theta_2^{(1)}, \theta_2^{(1)}, \theta_4^{(1)}, ..., \theta_p^{(1)}, y),$$

$$\vdots$$

Segue-se o processo iterativo até estabilizar a distribuição dos valores de $(\theta_1, \theta_2, ..., \theta_p)$, quando se pode assumir que se toma amostras da distribuição alvo h.

2.2 Modelo AMMI

Dentre os muitos modelos de estudo da interação entre dois fatores, destacam-se no melhoramento de plantas os do tipo aditivo para efeitos principais com interação multiplicativa (AMMI). No contexto de melhoramento de plantas, o objetivo é generalizar conclusões sobre genótipos em vários ambientes, com ênfase na interpretação de possíveis interações. Uma classe de modelos comumente utilizada para descrever a resposta média de genótipos sobre

ambientes é a dos modelos lineares-bilineares, que permitem incorporar covariáveis ambientais e genotípicas diretamente, além da interação na mesma representação.

Gollob (1968) e Mandel (1969), Mandel (1971) foram um dos primeiros trabalhos a desenvolver um modelo com termo bilinear (multiplicativo), a partir de um conjunto de dados dispostos em uma tabela de dupla entrada. A partir de Gauch (1988), Gauch e Zobel (1988) e Zobel, Wright e Gauch (1988) esse modelo passou a ser denominado de modelo de efeitos principais aditivos e interação multiplicativa (AMMI). Generalizações foram realizadas posteriormente por CORNELIUS, CROSSA e SEYEDSADR (1996), descrevendo outras classes de modelos lineares bilineares, por exemplo, modelo de regressão de colunas (CREG), modelo de regressão de linhas (SREG) geralmente chamado de biplot GGE e Modelo Completamente Multiplicativo (COMM).

Por usar menor número de parâmetros para modelar a interação entre os fatores, o AMMI tem sido aplicado em diversas áreas de estudo em busca de obter maior parcimônia, precisão e eficiência. Além disso, análises uni e multivariadas, como a ANOVA do modelo linear fixo, decomposições do tipo PCA e análises de regressão podem ser consideradas como casos particulares da aplicação do AMMI (GOLLOB, 1968; MANDEL, 1969; MANDEL, 1971; GABRIEL, 1978). Por exemplo, se concluirmos ao realizar a análise AMMI, que apenas efeitos principais são relevantes, o modelo pode ser simplificadao para acomodar uma ANOVA. Por outro lado, caso a característica dos dados seja de apresentar apenas a estrutura multiplicativa, o modelo PCA tenderia a ser bem ajustado.

No AMMI, a resposta média de um genótipo i em um ambiente j (y_{ij}) é modelada por:

$$y_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij} \tag{2.7}$$

sendo,

μ é uma constante, geralmente a média geral;

τ_i o efeito do i-ésimo genótipo, efeito linha da matriz GEI, para $i = 1, ..., r$;

δ_j o efeito do j-ésimo ambiente, efeito coluna da matriz GEI, para $j = 1, ..., c$;

λ_k o k-ésimo valor singular obtido a partir da decomposição por valores singulares (DVS) da matriz $GEI'GEI$ ou $GEIGEI'$, com $k = 1, ..., t$, sendo $t \leq min(r-1, c-1)$ (posto da matriz de interação GEI).

α_{ik} o i-ésimo elemento relativo ao genótipo i, do k-ésimo vetor singular referente à matriz $GEIGEI'$;

γ_{jk} o j-ésimo elemento relativo ao ambiente j, do k-ésimo vetor singular referente à matriz $GEI'GEI$;

ε_{ij} o erro experimental admitido ser identicamente e normalmente distribuído, $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Com GE modelado por $\sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk}$, satisfazendo ainda as seguintes restrições:

- $\lambda_1 > \lambda_2 > ... > \lambda_t > 0$
- identificabilidade: $\sum_i \tau_i = \sum_j \delta_j = \sum_i \alpha_{ik} = \sum_{jk} \gamma_j = 0$;
- ortonormalização: $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2$ e $\sum_i \alpha_{ik}\alpha_{ik'} = \sum_j \gamma_{jk}\gamma_{jk'} = 0$ com $k \neq k'$;

Originalmente se ajusta modelos AMMI estimando os efeitos principais (a parte aditiva: média geral, efeitos de genótipos e de ambientes) por um modelo linear fixo aplicado à matriz de médias ($\mathbf{Y}_{(r\times c)}$) e depois se aplica a decomposição de valores singulares (DVS) à matriz de resíduos ($\mathbf{GEI}_{(r\times c)} = [(gei)_{ij}]$, com característica de não-aditividade), que permanecem após o ajuste do efeito principal (PIEPHO, 1995).

A matriz **GEI** é a matriz de interação entre genótipos e ambientes (resíduo dos efeitos principais), que é obtida a partir do desvio relacionado a cada combinação ge_{ij} da **GEI**:

$$\hat{\varepsilon}_{ij} = \hat{ge}_{ij} = Y_{ij} - \hat{Y}_{i.} - \hat{Y}_{.j} + \hat{Y}_{..} \tag{2.8}$$

em que Y_{ij} é a matriz de médias das repetições do genótipo i no ambiente j, $\hat{Y}_{i.}$ é a média de cada a genótipo i, $\hat{Y}_{.j}$ é a média de cada ambiente e $\hat{Y}_{..}$ é a média geral.

A partir dessa matriz de interação é realizada a decomposição em valores singulares (DVS), uma técnica da álgebra linear que consiste em uma fatoração de matrizes na forma:

$$\mathbf{USV}' = \sum_{k=1}^{t} \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\gamma}_k' \tag{2.9}$$

sendo $\mathbf{S}$ uma matriz diagonal contendo os t valores singulares, em ordem decrescente, e as matrizes $\mathbf{U}$ e $\mathbf{V}'$ contendo os vetores singulares $\boldsymbol{\alpha}_k$ e$\boldsymbol{\gamma}_k'$ respectivamente. Uma importante propriedade desse método, é a construção de uma soma de t matrizes de posto unitário e ortogonais entre si, que ao realizar a soma de quadrado de cada elemento, reproduz a soma de quadrados da interação genótipos por ambientes $SQ_{G\times E}$.

O objetivo desse desdobramento é obter uma aproximação da $SQ_{G\times E}$ que explique bem a variabilidade inerente à interação (genótipo x ambiente). Como $\lambda_1^2 > \lambda_2^2 > ... > \lambda_t^2$, é possí-

vel que os primeiros termos possam explicar uma alta proporção da $SQ_{G\times E}$, e assim obter um modelo mais parcimonioso, que possa determinar padrões de resposta importantes e descartar termos de mais alta ordem de mais difícil interpretação como sendo "ruído"ou resíduo do modelo, obtendo assim uma melhor precisão (GAUCH, 2013). De acordo com Duarte e Vencovsky (1999), a interação entre genótipo e ambiente ainda pode ser representada separando o padrão do resíduo adicional, da seguinte maneira:

$$y_{ij} = \mu + \tau_i + \delta_j + \overbrace{\sum_{k=1}^{t^*} \lambda_k \alpha_{ik} \gamma_{jk}}^{padrão} + \overbrace{\sum_{k=t^*+1}^{t} \lambda_k \alpha_{ik} \gamma_{jk}}^{termos\ residuais\ (ruído)} \tag{2.10}$$

Sendo o índice k ($k = 1, 2, ..., t$), referente ao número de termos na interação. Tomando-se os t termos iniciais da DVS para produzir uma aproximação de mínimos quadrados, sendo os demais termos descartados, já que não teriam interpretação prática.

Outra vantagem da utilização dos modelos linear-bilineares, é sobre o uso da análise biplot como ferramenta de estatística descritiva, representando os escores genotípicos e ambientais referente aos termos multiplicativos no modelo. Esses escores plotados, são obtidos a partir da decomposição de valores singulares, da matriz GEI. E auxiliam na visualização de padrões existentes, identificando genótipos e ambientes estáveis, além de adaptações específicas de genótipos entre os ambientes.

O método biplot proposto por Gabriel (1971), consiste na aproximação de matrizes de alta dimensinalidade, em matrizes com baixa dimensão (em geral dois eixos). Cada elemento da matriz é representado pelo produto interno dos vetores linhas $\left(\mathbf{US}^{1/2}\right)$ e colunas $\left(\mathbf{S}^{1/2}\mathbf{V}'\right)$ da matriz de interação $\mathbf{GEI} = \mathbf{USV}' = \left(\mathbf{US}^{1/2}\right)\left(\mathbf{S}^{1/2}\mathbf{V}'\right)$. Sendo que para interpretação no gráfico, os genótipos ou ambientes considerados estáveis são aqueles cujos escores estão proximos à origem, e para estudo de adaptações específicas, é de acordo com os ângulos formados pelos vetores relacionados aos escores genotípicos e ambientais, que se forem menores que 90° indicam adaptações especificas dos genótipos aos ambientes. O modelo AMMI-bayesiano será revisado em detalhes no capítulo seguinte, que incluirá um exemplo numérico de sua implementação.

2.3 Modelo de *Threshold* (limiar)

Características qualitativas são frequentemente empregadas no melhoramento genético animal e de plantas. E são utilizados para representar estados subjetivos das características de

interesse por respostas ordinais, geralmente com múltiplas categorias. Exemplos clássicos são o grau de dificuldade no parto em animais domésticos e resistência a pragas e doenças em animais e plantas (SORENSEN et al., 1995).

Geralmente quando se estuda dados ordinais, a estrutura utilizada é organizada de acordo com a Tabela 2.1. A dimensão $n \times c$, indica que nas linhas são representadas as observações (individuos), como combinações dos níveis dos fatores (com r e w supondo o número de níveis de 2 fatores), e as colunas representam as categorias de resposta $(1,..,C)$, com n_{ij} sendo o número de observações em cada categoria.

Tabela 2.1 – Tabela de entrada de dados em cada combinação dos níveis de fatores, que pertencem a uma dada categoria de resposta

Observações (Indivíduos)	Categorias							
	C_1	C_2	C_3	$\cdots$	C_j	$\cdots$	C_c	TOTAL
1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3j}	$\cdots$	n_{3c}	$n_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	n_{i3}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$n = rw$	n_{n1}	n_{n2}	n_{n3}	$\cdots$	n_{nj}	$\cdots$	n_{nc}	$n_{n.}$

Muitas vezes se utiliza técnicas desenvolvidas para modelos lineares contínuos (aproximações normais) para analisar dados ordinais, mesmo que a estrutura da medida não justifique tal aproximação. Os problemas relatados quando se utiliza de tal estratégia para estudo, dizem respeito à inflação dos os erros de medida quando se observa um número de categorias menor, ou problemas de aproximação da distribuição da classe, quando o tamanho da amostra não é suficientemente grande, ou ainda, quando se tem um excesso de classes e se observa poucos elementos em algumas classes (GIANOLA, 1982; GIANOLA; FOULLEY, 1983). Sabe-se, no entanto, que este tipo de aproximação é muito comumente empregado, devido à relativa robustez e simplicidade dos modelos lineares clássicos (ATKINSON, 1988; MONTESINOS-LÓPEZ et al., 2015).

Na busca por alternativas para análises com dados ordinais, os geneticistas quantitativos introduziram o modelo de *Threshold* no melhoramento genético, para relacionar uma escala contínua subjacente a fenótipos (respostas categorias observadas) (DEMPSTER; LERNER,

1950). Os primeiros conceitos foram utilizados por Wright (1934) em estudos do número de dígitos em cobaias. A partir de Falconer (1965), essa "variável latente", passou a ser conhecida também como "*liability*", na genética de doença. Com o modelo de *Threshold* é possível pontuar ou dimensionar as categorias de resposta, de modo que se uma resposta cair entre os limites que definem a categoria apropriada é possível relacionar a intervalos da distribuição normal (GIANOLA; FOULLEY, 1983).

No modelo de *Thresholds*, as variáveis latentes são representadas pelo vetor $\mathbf{l} = \{l_i\}$ $(i = 1,2,...,n)$, em que cada l_i é obtido a partir da associação a cada i-ésima resposta ordinal observável pertencente ao vetor $\mathbf{Y} = \{y_i\}$, sendo expressa da seguinte forma:

$$l_i = \mathbf{x_i'}\beta + \mathbf{z_i'}\mathbf{a} + e_i \tag{2.11}$$

Em que n é o tamanho da amostra, β representa o parâmetro de efeitos de locação, **a**, representa valores genéticos aditivos e e_i é o resíduo associado a cada observação, $e_i \sim N(0,\sigma^2)$. Os elementos x' e z' são os vetores de incidência em relação aos parâmetros β e **a** respectivamente. Os elementos do vetor **l** são independente e identicamente distribuídos, sendo representado condicionalmente por:

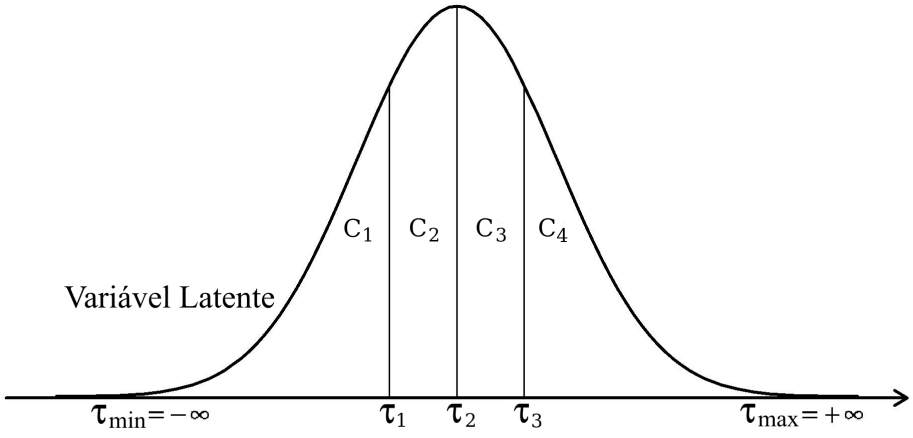
$$(\mathbf{l}|\beta,\mathbf{a},\sigma_e^2) \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma_e^2) \tag{2.12}$$

Como a variável latente é uma variável não-observável, utiliza-se uma pressuposição sobre a variância a fim de se obter a identificabilidade na verossimilhança, sendo $\sigma^2 = 1$. Entretando Bernardo e Smith (1994) comentam que se forem utilizados procedimento adequados para a escolha das prioris para cada parâmetro, a utilização dessa restrição já não é necessária para a realização de inferências sobre as posteriores. Mas como a sua utilização é padrão nesse tipo de análise, ela geralmente é adotada Sorensen e Gianola (2002).

Considerando o $\mathbf{y} = y_i$ ($i = 1,2,...,n$), um vetor de respostas categóricas, em que cada y_i representa uma resposta, pertencente a uma dentre várias categorias independente e mutuamente exclusiva, elas se relacionam em uma escala subjacente as variáveis latentes, de acordo com a definição hipotética de *thresholds*, tal que $\tau_{min} < \tau_1 < \tau_2 < ... < \tau_{c-1} < \tau_{max}$, sendo c a quantidade de categorias consideradas. Além disso, também são definidas que, $\tau_{min} = -\infty$ e $\tau_{max} = \infty$, para que os $c-1$ *thresholds* restantes possam assumir qualquer valor dentro deste intervalo. Um exemplo dessa relação entre as variáveis observáveis e não-observáveis, pode

ser visto na Figura 2.2, em que **l** segue uma distribuição gaussiana, e se l_i for obtida a partir do intervalo de τ_1 e τ_2, a atribuição dessa variável estará na segunda categoria (C_2) de resposta, $y_i = 2$.

Figura 2.2 – Representação de variáveis categóricas na escala subjacente (variável latente)



REFERÊNCIAS

ATKINSON, L. The measurement-statistics controversy: Factor analysis and subinterval data. **Bulletin of the Psychonomic Society**, v. 26, n. 4, p. 361–364, 1988.

BERNARDO, J. M.; SMITH, A. F. **Bayesian theory**. 1st. ed. [S.l.]: Wiley, 1994. (Wiley series in probability and mathematical statistics).

CORNELIUS, P. L.; CROSSA, J.; SEYEDSADR, M. S. **Statistical tests and estimators of multiplicative models for genotype-by-environment interaction In: KANG, M. S.; GAUCH, H. G. (Org.). Genotype-by-environment interaction.** [S.l.]: CRC press, 1996.

CROSSA, J. Statistical analyses of multilocation trials. In: **Advances in agronomy**. [S.l.: s.n.], 1990. v. 44, p. 55–85.

DEMPSTER, E. R.; LERNER, I. M. Heritability of threshold characters. **Genetics**, Genetics, v. 35, n. 2, p. 212–236, 1950.

DUARTE, J. B.; VENCOVSKY, R. Interação genótipos x ambientes: uma introdução à análise "ammi". **Série monografias. Sociedade Brasiliera de Genética**, 1999.

FALCONER, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. **Annals of Human Genetics**, Wiley, v. 29, n. 1, p. 51–76, 1965.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford University Press, v. 58, n. 3, p. 453–467, 1971.

GABRIEL, K. R. Least squares approximation of matrices by additive and multiplicative models. **Journal of the Royal Statistical Society: Series B**, v. 40, n. 2, p. 186–196, 1978.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference**. 2. ed. [S.l.]: Chapman and Hall/CRC, 2006. (Texts in Statistical Science).

GAUCH, H. G. Model selection and validation for yield trials with interaction. **Biometrics**, JSTOR, p. 705–715, 1988.

GAUCH, H. G. A simple protocol for ammi analysis of yield trials. **Crop Science**, v. 53, n. 5, p. 1860–1869, 2013.

GAUCH, H. G.; ZOBEL, R. W. Predictive and postdictive success of statistical analyses of yield trials. **Theoretical and Applied genetics**, Springer, v. 76, n. 1, p. 1–10, 1988.

GELMAN, A. et al. **Bayesian data analysis**. 2nd ed. ed. [S.l.]: Chapman and Hall/CRC, 2004. (Texts in statistical science).

GIANOLA, D. Theory and analysis of threshold characters. **Journal of Animal Science**, Oxford University Press, v. 54, n. 5, p. 1079–1096, 1982.

GIANOLA, D.; FOULLEY, J. L. Sire evaluation for ordered categorical data with a threshold model. **Genetique, selection, evolution**, BioMed Central, v. 15, n. 2, p. 201, 1983.

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

JAYNES, E. T. **Probability theory: The logic of science**. [S.l.]: Cambridge university press, 2003.

KOCH, K.-R. **Introduction to Bayesian Statistics**. 2nd, updated and enl. ed. ed. [S.l.]: Springer, 2007.

MALOSETTI, M.; RIBAUT, J.-M.; EEUWIJK, F. A. van. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. **Frontiers in physiology**, Frontiers, v. 4, p. 44, 2013.

MANDEL, J. The partitioning of interaction in analysis. **Journal of Research of the National Bureau of Standards: Physics and chemistry**, National Bureau of Standards, v. 73, p. 309, 1969.

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

METROPOLIS, N. et al. Equation of state calculations by fast computing machines. **The journal of chemical physics**, AIP, v. 21, n. 6, p. 1087–1092, 1953.

MONTESINOS-LÓPEZ, O. A. et al. Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. **G3: Genes, Genomes, Genetics**, G3: Genes, Genomes, Genetics, v. 5, n. 2, p. 291–300, 2015.

PAULINO, C. D.; MURTEIRA, B.; TURKMAN, M. A. A. **Estatística Bayesiana**. 2ed. ed. [S.l.]: Fundação Calouste Gulbenkian, 2018.

PIEPHO, H. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. **Theoretical and Applied Genetics**, v. 90, n. 3-4, p. 438–443, 1995.

PRESS, S. J. **Subjective and Objective Bayesian Statistics: Principles, Models, and Applications**. 2st. ed. [S.l.]: John Wiley & Sons, 2002. (Wiley Series in Probability and Statistics).

SILVA, W. C. J.; DUARTE, J. B. Métodos estatísticos para estudo de adaptabilidade e estabilidade fenotípica em soja. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 41, p. 23–30, 2006.

SORENSEN, D. et al. Bayesian inference in threshold models using gibbs sampling. **Genetics Selection Evolution**, v. 27, n. 3, p. 229, 1995.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC Methods in Quantitative Genetics**. [S.l.]: Springer-Verlag, 2002. (Statistics for biology and health).

WRIGHT, S. An analysis of variability in number of digits in an inbred strain of guinea pigs. **Genetics**, Genetics, v. 19, n. 6, p. 506–536, 1934.

YAN, W.; KANG, M. S. **GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists**. [S.l.]: CRC press, 2002.

ZOBEL, R. W.; WRIGHT, M. J.; GAUCH, H. G. Statistical analysis of a yield trial. **Agronomy journal**, American Society of Agronomy, v. 80, n. 3, p. 388–393, 1988.

SEGUNDA PARTE

ARTIGO 1

Flexible multi-environment trials analysis via Bayesian-AMMI: A review

Flexible multi-environment trials analysis via Bayesian-AMMI: A review

Abstract

Multi-Environment Trials (MET) are of paramount importance in recommending cultivars. Detecting genotypes that are superior and stable and those that are adapted to specific environments is done finding patterns in genotype x environment interaction (GEI). Additive Main Effect and Multiplicative Interaction (AMMI) models have been widely used for this purpose as it allows visual representation via biplots. Recent advances in AMMI analysis, especially in the Bayesian perspective, with the contribution of inferential regions for the interpretation of biplots are reviewed in this paper and an example is simulated with a controlled scenario with stable and unstable genotypes in different environments is include to illustrate your characteristics and capabilities.

Additional key words: Bayesian analysis, credibility regions, quantitative genetics, selection.

Abbreviations used: AMMI (Additive Main Effect and Multiplicative Interaction); BLUP (Best Linear Unbiased Prediction); BOA (Bayesian Output Analysis)**;** GEI (genotype x environment interaction); HPD (Highest Posterior Density); MET (Multi-Environment Trials); MCMC (Markov Chain Monte Carlo); OLS (Ordinary Least Squares); PC1 (first principal component); PC2 (second principal component); REML (restricted maximum likelihood); SVD (singular value decomposition); VMF (von Mises-Fisher).

Introduction

Additive Main Effect and Multiplicative Interaction (AMMI) models are widely used to analyse cross classified data. Main effects are estimated using information from the margins and we search for patterns in the remaining interaction, that is of paramount importance (Cornelius & Seyedsadr, 1997). One of the most important applications of it is on Multi-Environment Trials (MET) used to evaluate genotype x environment interactions (GEI).

AMMI combines linear and nonlinear techniques in a single analysis. Model fitting is usually done in two steps. First we adjust main effects (mean squares or more sophisticated methods). Remaining GEI matriz, residuals of the first step, goes into Singular Value Decomposition (SVD) to detect patterns (usually linear combinations).

This approach allows for more parsimonious models reducing the number of bilinear terms kept from the initial model, through statistical criteria to describe GEI (Piepho, 1995; Dias & Krzanowski, 2003; Hadasch, et al. 2017). Interaction patterns can be directly depicted in biplot graphs (Gabriel, 1971; Kempton, 1984; Greenacre, 2010).

Despite its advantages over previous (linear) methods, usual AMMI shares the drawbacks of fixed effect models specifications. In this case, inference on distribution of statistics for nonlinear parameters as variance components and singular values are very cumbersome and usually approximated. Yang et al. (2009) alert for the abuse of interpretation of biplots as most published applications bring no uncertainty measures for plotted scores.

The first proposed methods, both fiducial and pure frequentist to draw confidence regions on biplots are problematic. The first uses strong assumptions on parametric models for scores distributions (Dennis & Gower, 1994). The second relies on computer intensive sampling that could fail to restore patterns in line and columns of the interaction matrix (Maya, 2006, Lavoranti, 2007; Yang et al., 2009; Yan et al., 2010). However, more recent studies (Hu & Yang, 2013a,b) have proposed new methods of bootstrapping in an attempt to remove deficiencies in the classical methods prior to Yang et al. (2009).In the Bayesian perspective, the characteristic of this approach can partially help on this problems as MCMC (Markov Chain Monte Carlo) sampling from posterior distributions allows for greater flexibility on complex parametric models, thus preserving desired patterns in GEI.

First proposal of Bayesian AMMI (Viele & Srinivasan, 2000), that in a groundbreaking paper solves the main problem of correct specification of bilinear parameters distribution assuring orthonormality to singular values. Liu (2001) PhD thesis expanded on many aspects of conjugate models for priors for direct implementation of Gibbs Sampling, reducing

computational time and making for reliable and stable posterior samples. Crossa et al. (2011) and Perez-Elizalde et al. (2012) has shown how to make inferential biplots using bivariate credibility regions and how to use previous experiments to elicit prior parameters.

Recent advances in Bayesian-AMMI application to unbalanced data, heterogeneous variance and model selection can be found in (Oliveira et al., 2015; Silva et al., 2015; Jarquin et al., 2016; Silva et al., 2019; Romão et al., 2019). The technique is not yet widely used and a clear presentation of its capabilities is needed to make its use routine to MET analysis. Note that despite the publication of a specific R package (*bayesammi*, from Yaseen et al., 2018) allowing for point estimates of biplot using posterior averages, credibility regions are still difficult to draw.

In this paper we review Bayesian-AMMI, it's strengths, pitfalls and implications, aiming to amplify the application of the method. We illustrate its flexibility with a simulated example, with controlled scenario (stable and unstable genotypes known), that could help breeders on cultivar recommendation.

Development

Bayesian-AMMI model specification

In what follows, we rewrite Viele & Srinivasan (2000) and Liu (2001) model specification. The conditional expected value of the realizations can be written as:

$$E\left[\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{g}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right] = \boldsymbol{\mu}_{\mathbf{y}} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \sum_{k=1}^{t} \lambda_k diag(\mathbf{Z}\boldsymbol{\alpha}_k)\mathbf{X}_2\boldsymbol{\gamma}_k \qquad (1)$$

where $\mathbf{y}$ represents $n = rl$ phenotypic responses, related to r genotypes and l repetitions, being $l = bc$ are b (blocks) effects nested in c (environments). Vector $\boldsymbol{\beta}$ represents hierarchical effects of blocks within environments and $\mathbf{g}$ the genotype effect. The parameters that describe the multiplicative components of the model allows the single value decomposition of the matrix $GEI_{r\times c}$, where λ_k is the k-th singular value and, $\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k$ are to be k-th singular vectors related to genotypes and environments respectively, with $k = 1,...,t$ and $t = \min(r,c)$ being the rank of matrix $GEI_{r\times c}$. Matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Z}$ are design matrices (known 0 or 1 constants). Singular value decomposition is strictly orthonormal and there is an eigenvalues are ordered $\left(\lambda_1 \geq \ldots \geq \lambda_t \geq 0\right)$.

Following the usual experimental assumptions for a continuous variable in randomized experiments, conditional distribution for realizations follows a Gauss-Markov Normal model:

$$\mathbf{y} \mid \boldsymbol{\mu}_{\mathbf{y}}, \sigma_e^2 \sim N_n(\boldsymbol{\mu}_{\mathbf{y}}, \mathbf{I}_n\sigma_e^2) \qquad (2)$$

where $\mathbf{I}_n$ is an identity matrix of order *n*

Prior density:

In order to complete the Bayesian specification of the model, the prior distributions will be the same as those presented in Oliveira et al. (2015):

$\boldsymbol{\beta} \sim$ constant;

$\boldsymbol{g} \mid \boldsymbol{\mu}_g, \sigma_g^2 \sim N\ (\mathbf{0}, \mathbf{I}\sigma_g^2)$; in which $\sigma_g^2 \sim \frac{1}{\sigma_g^2}$;

$\lambda_k \sim$ constant;

$\alpha_k \sim$ spherical uniform on the correct subspace;

$\boldsymbol{\gamma}_k \sim$ spherical uniform on the correct subspace;

$\sigma_e^2 \sim \frac{1}{\sigma_e^2}$.

These *prior* specifications are non-informative for environmental effects and experimental variance, and minimally informative for bilinear parameters. For effects of genotypes, a hierarchical prior distribution was used, being assigned a multivariate normal distribution for $\boldsymbol{g}$ with uncertainty assigned in relation to the variance, the same prior specification that would yield restricted maximum likelihood/ Best Linear Unbiased Prediction (REML/BLUP) type of random effects in mixed models fiducial inference. For the components variance, genotypic and residual, non-informative priors were assigned, proportional to Jeffreys' priors. Note that Viele & Srinivasan (2000), Liu (2001), Crossa et al. (2011) and Perez-Elizalde et al. (2012) a common constant (general average) is estimated. The authors also used prior distributions with large values for scale parameters and degrees of freedom equal to one for scaled inverse chi-square distributions, so that the whole process is presented as non-informative. In our case there is no need for it. It is debatable whether breeders would like to use available information *a priori* (on averages and relatedness of genotypes and also on environments). However, our presentation allows for a clean comparison with classic or frequentist AMMI analysis.

Posterior complete conditional distributions:

The likelihood function for the model described in (2) is:

$$L\left(\boldsymbol{\mu}_{\mathbf{y}} \mid \mathbf{y}\right) = \frac{1}{\left(2\pi\right)^{\frac{n}{2}} \left|I\sigma_e^2\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^{\top}\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\right\}$$

And the joint distribution is given by:

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto p\left(\mathbf{y}\,|\,\boldsymbol{\mu}_{\mathbf{y}}, \sigma_e^2\right) p(\mathbf{g}\,|\,\boldsymbol{\mu}_g, \sigma_g^2)\, p(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_\beta, \sigma_\beta^2)\, p(\sigma_g^2\,|\,v_g, S_g^2)\, p(\sigma_e^2\,|\,v_e, S_e^2) \times$$
$$\times \prod_{k=1}^{t} p(\lambda_k\,|\,\boldsymbol{\mu}_{\lambda_k}, \sigma_{\lambda_k}^2)\, p(\boldsymbol{\alpha}_k)\, p(\boldsymbol{\gamma}_k)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\beta}, \sigma_g^2, \sigma_e^2)$.

The simplified expression of the joint posterior distribution, after the substitutions of the appropriate expressions of the prior distributions for hyperparameters, is given by:

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto (\sigma_e^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^{\top}\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\right\}(\sigma_g^2)^{-\frac{n_g}{2}} exp\left\{-\frac{1}{2\sigma_g^2}\mathbf{g}^{\top}\mathbf{g}\right\} \times \frac{1}{\sigma_g^2}\frac{1}{\sigma_e^2}$$

After some algebraic manipulations, the posterior complete conditional distributions for each of the model parameters can be written as:

$$\boldsymbol{\beta}\,|\cdots \sim N\left[(\mathbf{X}_1^{\top}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\top}(\mathbf{y}-\mathbf{Z}\mathbf{g}-\boldsymbol{\Theta}), (\mathbf{X}_1^{\top}\mathbf{X}_1)^{-1}\sigma_e^2\right], \text{ where } \boldsymbol{\Theta} = \sum_{k=1}^{t}\lambda_k diag(Z\alpha_k)X_2\gamma_k. \tag{2}$$

$$\mathbf{g}\,|\cdots \sim N\left[\left(\mathbf{Z}^{\top}\mathbf{Z}+\mathbf{I}\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}\mathbf{Z}^{\top}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\boldsymbol{\Theta}), \left(\mathbf{Z}^{\top}\mathbf{Z}+\mathbf{I}\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}\sigma_e^2\right] \tag{3}$$

$$\sigma_g^2\,|\ldots \sim \text{Scaled}-\chi^{-2}\left[n_g, \mathbf{g}^{\top}\mathbf{g}\right]. \tag{4}$$

$$\sigma_e^2\,|\ldots \sim \text{Scaled}-\chi^{-2}\left[n, \left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^{\top}\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\right]. \tag{5}$$

$$\lambda_k\,|\cdots \sim N^{+}\left[\left(\phi_k^{\top}\phi_k\right)^{-1}\phi_k^{\top}\boldsymbol{\Delta}_{k'}, \left(\phi_k^{\top}\phi\right)^{-1}\sigma_e^2\right] \tag{6}$$

where $\boldsymbol{\Delta}_{k'} = \mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\mathbf{g}-\sum_{k'\neq k}^{t}\lambda_{k'}diag(\mathbf{Z}\boldsymbol{\alpha}_{\mathbf{k}'})\mathbf{X}_2\boldsymbol{\gamma}_{\mathbf{k}'}$ and $\phi_k = diag(\mathbf{Z}\boldsymbol{\alpha}_k)\mathbf{X}_2\boldsymbol{\gamma}_k$.

$p(\boldsymbol{\alpha}_k\,|\ldots) \propto exp\left\{\frac{\lambda_k}{\sigma_e^2}\boldsymbol{\alpha}_k^{\top}\boldsymbol{\mu}_{\alpha_k}\right\}$, where $\boldsymbol{\mu}_{\alpha_k} = \boldsymbol{\Lambda}_k^{\top}\left(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\mathbf{g}\right)$ and $\boldsymbol{\Lambda}_k = diag\left(\mathbf{X}_2\boldsymbol{\gamma}_k\right)\mathbf{Z}$.

$p(\boldsymbol{\gamma}_k\,|\ldots) \propto exp\left\{\frac{\lambda_k}{\sigma_e^2}\boldsymbol{\gamma}_k^{\top}\boldsymbol{\mu}_{\gamma_k}\right\}$, where $\boldsymbol{\mu}_{\gamma_k} = \boldsymbol{\Omega}_k^{\top}\left(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\mathbf{g}\right)$ and $\boldsymbol{\Omega}_k = diag\left(\mathbf{Z}\boldsymbol{\alpha}_k\right)\mathbf{X}_2$.

One of the main difficulties with respect to the application of the Bayesian method is to correctly sample the singular vectors since the supports of the *a posteriori* conditional distributions are not trivial. Vielle & Srinivasan (2000) showed how to get around this difficulty by sampling the vectors in the corrected subspace from orthogonal linear transformations.

Sampling of singular vectors is performed using auxiliary variables $\boldsymbol{\alpha}_k^* = \mathbf{H}_k^{\top}\boldsymbol{\alpha}_k$ and $\boldsymbol{\gamma}_k^* = \mathbf{R}_k^{\top}\boldsymbol{\gamma}_k$ defined in the corrected subspace, being $\mathbf{H}_k$ and $\mathbf{R}_k$ linear transformation

matrices. Those variables follow a von Mises-Fisher (VMF) full conditional distribution as demonstrated in Liu (2001), given by:

$$\boldsymbol{\alpha}_k^* \mid \ldots \sim VMF\left(r-m, \frac{c_k \lambda_k}{\sigma_e^2}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}_k} \right),$$

where $c_k \lambda_k / \sigma_e^2$ is the concentration parameter, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}_k} = c_k^{-1} \mathbf{H}_k^\top \boldsymbol{\mu}_{\alpha_k}$ the directional average with $c_k = \sqrt{(\mathbf{H}_k^\top \boldsymbol{\mu}_{\alpha k})^\top \mathbf{H}_k^\top \boldsymbol{\mu}_{\alpha_k}}$ = $\sqrt{\boldsymbol{\mu}_{a_k}^\top \mathbf{H}_k \mathbf{H}_k^\top \boldsymbol{\mu}_{\alpha_k}}$, being r-m, the corrected subspace dimension $(m = t-1)$ and

$$\boldsymbol{\gamma}_k^* \mid \text{outros} \sim VMF\left(c-m, \frac{d_k \lambda_k}{\sigma_e^2}, \tilde{\boldsymbol{\mu}}_{\gamma_k} \right).$$

where $d_k \lambda_k / \sigma_e^2$ is the concentration parameter, $\tilde{\boldsymbol{\mu}}_{\gamma_k} = d_k^{-1} \mathbf{R}_k^\top \boldsymbol{\mu}_{\gamma_k}$ the directional average with $d_k = \sqrt{\boldsymbol{\mu}_{a_k}^\top \mathbf{R}_k \mathbf{R}_k^\top \boldsymbol{\mu}_{\alpha_k}}$.

MCMC sampling

To draw samples from the joint posterior distribution for all parameters we used the Gibbs sampler in MCMC (Markov chain Monte Carlo). Algorithm is described in Oliveira et al. (2015) and Silva et al. (2015). First 4,000 were discarded (burn-in) and samples were drawn at a jump interval of 10 observations (thinning). Resulting effective chain has 14,000 realizations. Raftery & Lewis (1992) diagnostics and Heidelberger & Welch (1983) convergence criterion were used to check sample quality of the resulting chains. Those methods are implemented in the BOA (Bayesian Output Analysis) package. For all data analysis we used R (TEAM, R CORE, 2019). The script is available at the UFV (Universidade Federal de Viçosa) and Crop Science Journal repositories.

Biplot analysis

To infer about the adaptability (specific and broad) of genotypes to environmental diversity, based on the results of the biplot in the AMMI model, we used the internal product between genotypic and environmental scores. For those, we built credibility regions considering only the first two principal components (PC1 and PC2).

It is possible to form subgroups of scores whose regions overlap and that belong to the same quadrants in the Cartesian plane represented in Figure 1. Those subgroups are evidence of positive and negative interactions between genotypes and environments. In most cases, however, interactions are more complex and this pattern is not so clear (Denis & Gower,

1994). Ellipses that contain the origin indicate stability of genotypes and environments that share such scores estimates, and are evidence of not relevant interaction (Junior et al., 2018).

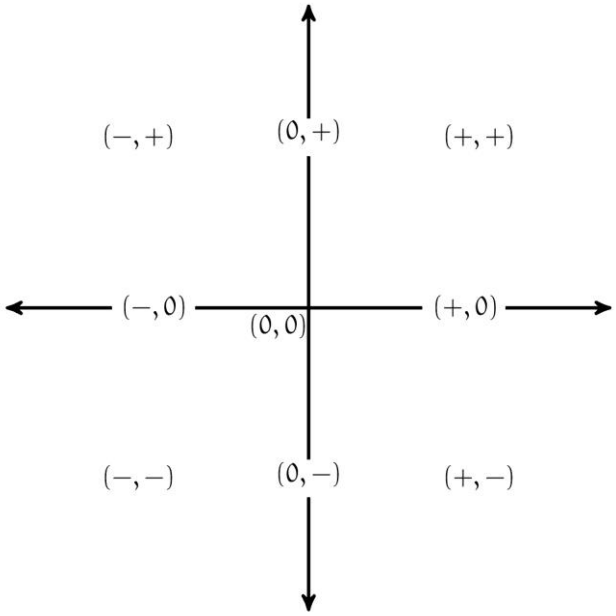


Figure 1. Cartesian plane representation of potential outcomes for inner product of genotypes and environments scores.

a) Example

Simulated data

In this study we simulated a dataset with 12 genotypes (G1-G12) in 20 environments (E1-E20), each in complete randomized block designs with 3 replications (blocks). The main effects of genotypes (G) and environments (E) were generated from Gaussian distributions, $N(0,5)$ and $N(0,1)$, respectively. For the effect of interaction, were simulated 2 response patterns. Of the 12 genotypes, 8 were evaluated in environments generated from Gaussian distributions $N(0,1)$ (G1 to G8, stable genotypes), and the other group formed by genotypes generated from Gaussian distributions $N(5,11)$ (G9 to G12, unstable genotypes), thus defining the stability and instability of this dataset so that a clearer comparison can be made (observed in Figure S1 [suppl], representing the effects on the GE interaction). Figure S2 [suppl] shows a summary of the average yield $(\bar{y})$ generated for genotypes in the different environments.

Evaluating posterior chains

Figure 2 shows the traces of the chains of the residual σ_e^2 and genotypic σ_g^2 variance components and their respective *a posteriori* marginal densities, visually reinforcing the convergence of the chains, corroborating the results obtained by the convergence tests. For the other parameters, were also carried out and verified convergence tests in the chain.

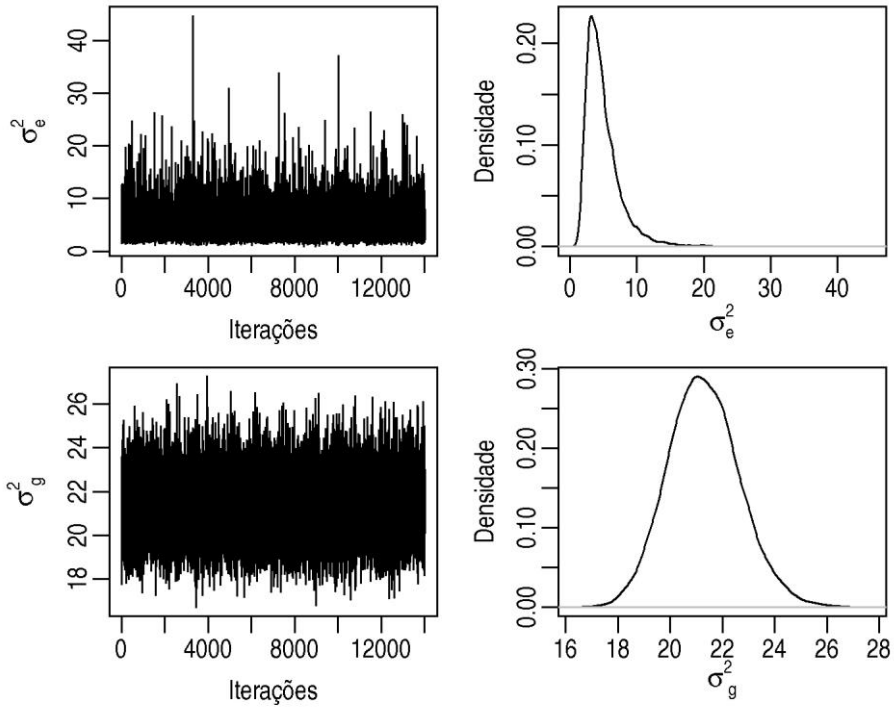


Figure 2. Graph of the traces and densities of the variance components.

Results

Caterpillar plot in Figure 3 shows Highest Posterior Density (HPD) regions with 95% credibility for genotype effects. Overlaps indicate similar effects. It is possible to highlight the G1 and G7 genotypes as those substantially above population mean for the simulated trait. Those are of greatest interest in terms of selection and recommendation, but we still need to analyze the stability and adaptability, i.e., GEI.

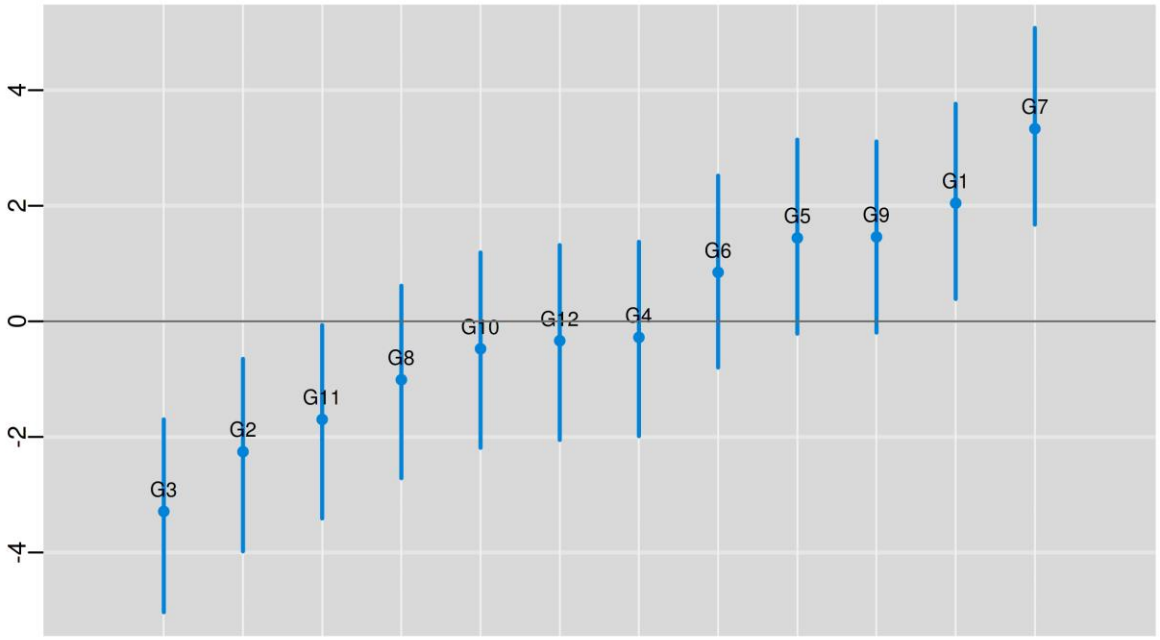


Figure 3. 95% credibility regions for highest posterior density (HPD) of genotype effects.

Point and interval posterior summaries for singular values and variance components are presented in Table 1. Ordinary Least Squares (OLS) estimates referring to the application of frequentist AMMI analysis are also presented for comparisons purposes (for those we used R package agricolae (Mendiburu, 2020). Note that despite noninformative prior distributions chosed for singular values, there is a substantial shrinkage of eigenvalue estimates in Bayesian predictions as compared to OLS estimates, especially from λ_4 (being almost four times larger than λ_5). This fact was also reported by several authors using Bayesian-AMMI (Liu, 2001; Oliveira et al., 2015; Silva et al., 2015; Crossa et al., 2011). 75% of GEI variance were explained by Bayesian-AMMI with the first two axes as opposed to 68.5% by OLS. Another representation of the variability decay and shrinkage in estimates is depicted in Figure S3 (represents the densities of the accumulated proportions of the explanation of the variability of the data, λ_k^2).

Table 1. Least squares estimates and summaries of marginal posterior distribution for singular values and variance components.

Parameters	OLS estimates	Bayesian estimates			
		Mean	SD	LL	UL

λ_1	67.539	66.019	2.68	60.657	71.121
λ_2	50.391	48.461	2.745	43.229	54.029
λ_3	39.116	36.769	2.728	31.465	42.109
λ_4	30.521	27.595	2.798	22.263	33.128
λ_5	16.684	7.706	4.336	0.009	14.853
λ_6	14.317	3.218	2.681	<0.001	8.534
λ_7	12.079	1.417	1.509	<0.001	4.572
λ_8	8.368	0.671	0.879	<0.001	2.575
λ_9	8.091	0.325	0.515	<0.001	1.352
λ_{10}	4.817	0.159	0.297	<0.001	0.719
λ_{11}	3.521	0.079	0.176	<0.001	0.374
σ_g^2	-	4.269	2.7283	1.365	10.099
σ_e^2	-	21.25	1.366	18.618	23.974

SD: Standard deviation, LL: Lower limit, UL: Upper limit, λ_k : k-th singular value, σ_g^2 : genotypic variance, σ_e^2 :residual variance.

In Figure 5 we show the marginal posterior distributions for singular values from Bayesian-AMMI. For the initial values $(\lambda_1 \text{ a } \lambda_4)$ the distributions are approximately symmetric. From λ_5 on, densities are skewed to the right as modes, medians and averages increasingly approaching zero as i increases in λ_i, which agrees with the theory of principal component analysis. An interpretation in terms of directional data is that hyperspheres from just the initial eigenvalues explain most of the variability in the data.

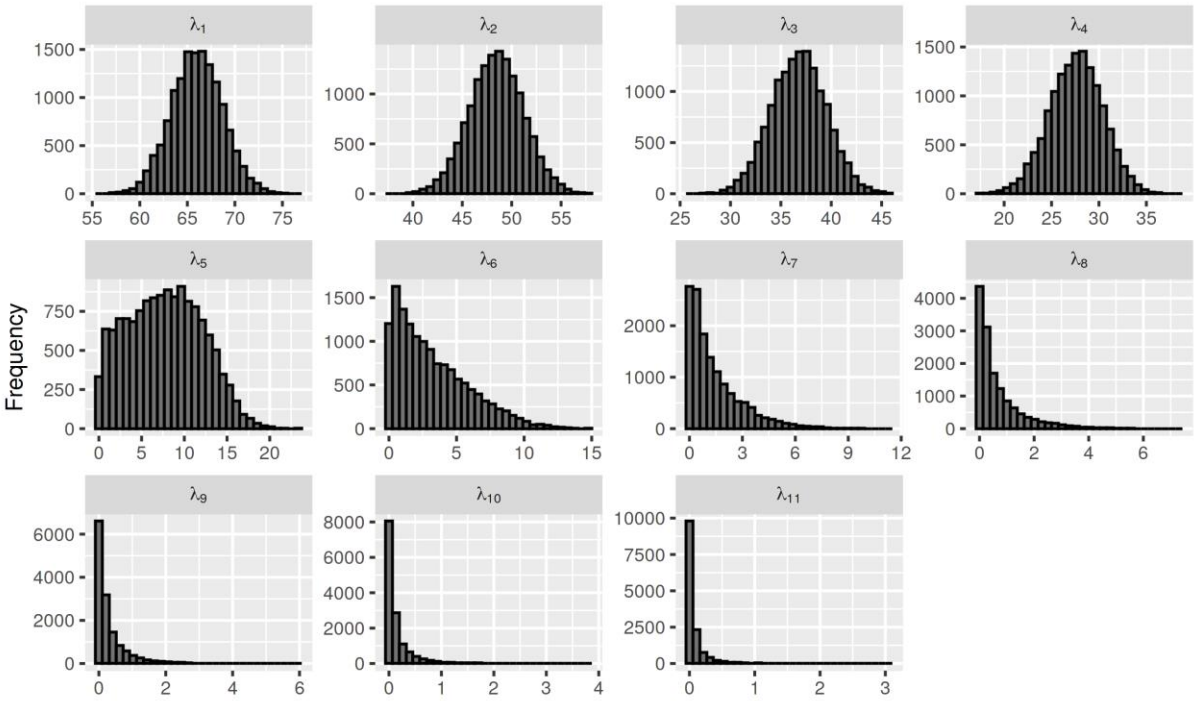


Figure 5. Marginal posterior distribution approximation for singular values from the interaction of 12 genotypes in 20 environments.

A cornerstone of AMMI interpretations is the biplot analysis of GEI. In Figure 6 (a) are depicted genotypic and environmental scores that represent GEI and their 95% credibility regions (that are not present in figure 6-b). Interpretation is easy on overlappings of regions within quadrants drawn with the two principal components (PC1 e PC2). The farthest the region is from the origin, the greater the contribution of the genotype and environment to the interaction. Genotypes or environments whose credibility region contains the origin $(0,0)$ are stable and were not plotted to simplify the interpretations.

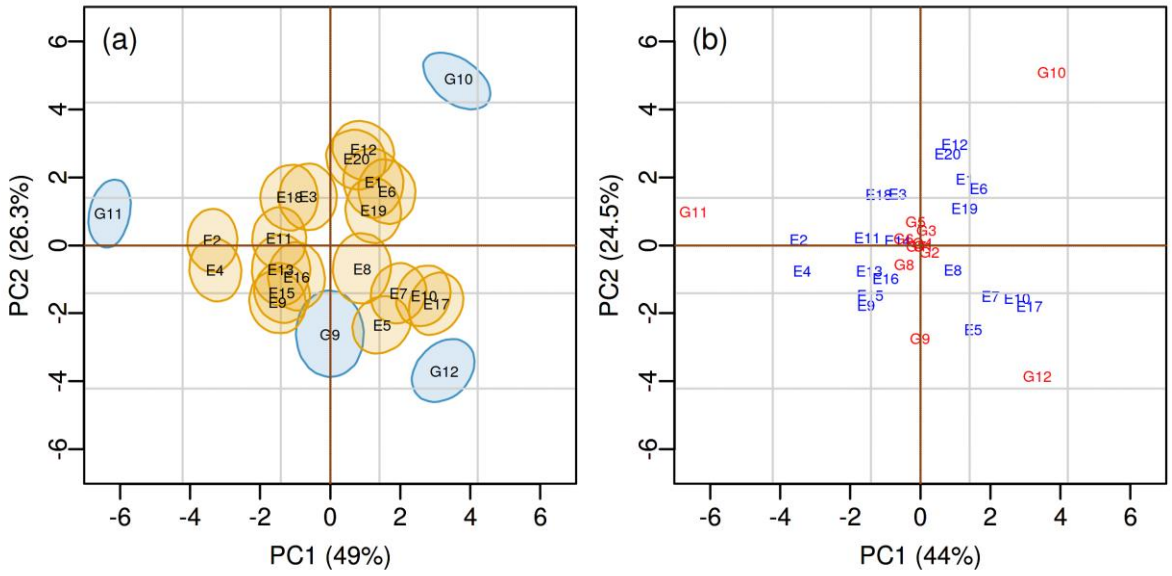


Figure 6. 95% credibility regions for genotypic (G) and environmental (E) scores. AMMI-2 biplot: (a) Bayesian and (b) frequentist.

Using Bayesian-AMMI we identify the subgroup {G1, G2, G3, G4, G5, G6, G7, G8} of stable genotypes, as well as stable environment E14. Credibility regions for scores that do not contain the origin can have the foccus in remaining interesting GEI. Genotypes G9, G10, G11 and G12 are very different in their GEI patterns, being adapted to contrasting environments.

For environments we could identify the following subgroups: {E14}, {E1, E6, E12, E19}, {E3, E20}, {E18}, {E2, E4, E11, E13, E16}, {E9, E15}, {E5, E7, E10, E17}, {E8}. Suggestions for the formation of homogeneous subgroups are represented in more detail in Table S1.

Frequentist AMMI from Figure 6 (b) is almost the same, showing similar patterns in a simulated / controlled scenario (no missing data, many times difficult to observe with data from real experiments), beyond the disadvantage of not having inferential regions for easy interpretation of scores, as emphasized by Yang et al. (2009).

Combining information and inference presented in Figures 3 and 6 could help the breeder in critical decisions for many stages of the breeding program. Best cultivars G1 and G7 are also considerably stable in biplot and should be recommended. Genotypes G9, G10, G11 and G12 are not stable but could be well adapted to some environments. On the other hand, G10, G11 and G12 in general should not be recommended.

General Remarks

Bayesian-AMMI is still not widely used to multi-environment trials and cultivar recommendations. Methodological tools as described by Crossa et al. (2011) and Perez-

Elizalde et al. (2012) could be more explicit on how to use resulting biplots. Junior et al. (2018) and Oliveira et al. (2017) presented a practical and useful way to interpret those graphs. Complex overlapping of regions in more than a single quadrant and large imprecision in evaluating regions has also been a problem (Perez-Elizalde et al., 2012). Enhancing the power of those methods is of course, based on better designed experiments and their precisions, but also could use some prior information on historical data and genotypic similarities.

Josse et al. (2014), used normal priors for singular vectors and rescaled them for orthonormality. Silva et al. (2015), on the other hand, used Jeffrey's priors for eigenvalues (variance components) yielding shrinkage similar to those verified by Cornelius & Seyedsadr (1997) and Cornelius & Crossa (1999). This allows for more parsimonious models. Heterogeneous variances could as well be modelled in a relatively straightforward fashion as described by Silva et al. (2019). All those contributions could enhance the precision of biplots, increasing their power.

Methods presented here could be extended to other bilinear models by just deleting main effects in linear terms and relaxing corresponding assumptions.

Conclusion

Bayesian-AMMI deals with a plethora of parametric methods to make posterior inference on biplots, being an advantage in the use of this approach, in addition to the use of information obtained in previous experiments, leading to a gain in the precision of the parameter estimates.

We used non-informative priors to make things comparable, but practical breeding programs could make it better by using well elicited proper priors for main effects and their covariances.

Computational costs used to be a disadvantage on using Bayesian methods, however, computers and algorithms have been developed that makes all this analysis feasible in personal computers in minutes.

Acknowledgements

This study was funded by CAPES and CNPQ with scholarships for Master and PhD students.

References

Cornelius PL, Seyedsadr MS, 1997. Estimation of general linear-bilinear models for two-way tables. J. Statist. Comput. Simulation 58 (4): 287-322. DOI: 10.1080/00949659708811837.

Cornelius PL, Crossa J, 1999. Prediction assessment of shrinkage estimators of multiplicative models for multienvironment cultivar trials. Crop Sci 39 (4): 998-1009. DOI: 10.2135/cropsci1999.0011183X003900040007x.

Crossa J, Perez-Elizalde S, Jarquin D, Cotes JM, Viele K, Liu G, Cornelius PL, 2011. Bayesian estimation of the additive main effects and multiplicative interaction model. Crop Sci 51 (4): 1458-1469. DOI: 10.2135/cropsci2010.06.0343.

da Silva CP, de Oliveira LA, Nuvunga JJ, Pamplona AKA, Balestre M, 2015. A Bayesian shrinkage approach for AMMI models. Plos one 10(7): e0131414. DOI: 10.1371/journal.pone.0131414.

da Silva CP, de Oliveira LA, Nuvunga JJ, Pamplona AKA, Balestre M, 2019. Heterogeneity of variances in the Bayesian AMMI model for multienvironment trial studies. Crop Sci, 59(6): 2455-2472. doi: 10.2135/cropsci2018.10.0641.

de Oliveira LA, da Silva CP, Nuvunga JJ, da Silva AQ, Balestre M, 2015. Credible intervals for scores in the AMMI with random effects for genotype. Crop Sci 55(2): 465-476. DOI: 10.2135/cropsci2014.05.0369.

de Oliveira LA, da Silva CP, Teodoro PE, Torres FE, Corrêa AM, Bhering, LL, 2018. Performance of Cowpea Genotypes in the Brazilian Midwest Using the Bayesian Additive Main Effects and Multiplicative Interaction Model. Agron J, 110(1): 147-154. DOI: 10.2134/agronj2017.03.0183.

Denis JB, Gower JC, 1996. Asymptotic confidence regions for biadditive models: Interpreting genotype × environment interactions**.** Appl Statistics 45(4): 479-493. DOI: 10.2307/2986069.

Dias CTS, Krzanowski WJ, 2003. Model selection and cross validation in additive main effects and multiplicative interaction models. Crop Science 43 (3): 865-873. DOI: 10.2135/cropsci2003.8650.

Gabriel KR, 1971. The biplot graphic display of matrices with application to principal componente analysis. Biometrika 58 (3): 453-467. DOI: 10.1093/biomet/58.3.453.

Greenacre M, 2010. Correspondence analysis of raw data. Ecology 91(4): 958-963. DOI: 10.1890/09-0239.1

Heidelberger P, Welch PD, 1983. Simulation run length control in the presence of an initial transient. Ops Res. 31(6): 1109-1144. DOI: 10.1287/opre.31.6.1109.

Hadasch S, Forkman J, Piepho HP, 2017. Cross-Validation in AMMI and GGE Models: A Comparison of Methods. Crop Sci. 57(1): 264-274. DOI: 10.2135/cropsci2016.07.0613.

Hu Z, Yang RC, 2013a. A new distribution-free approach to constructing the confidence region for multiple parameters. PloS one 8 (12): e81179. DOI: 10.1371/journal.pone.0081179.

Hu Z, Yang RC, 2013b. Improved statistical inference for graphical description and interpretation of genotype × environment interaction. Crop Sci. 53(6): 2400-2410. DOI: 10.2135/cropsci2013.04.0218.

Jarquín D, Pérez-Elizalde S, Burgueño J, Crossa J, 2016. A hierarchical Bayesian estimation model for multienvironment plant breeding trials in successive years. Crop Sci. 56(5): 2260-2276. DOI: 10.2135/cropsci2015.08.0475.

Josse J, van Eeuwijk F, Piepho HP, Denis JB, 2014. Another look at Bayesian analysis of AMMI models for genotype-environment data. J Agric Biol Environ Stat 19 (2): 240-257. DOI: 10.1007/s13253-014-0168-z.

Júnior LAYB, de Silva CP, de Oliveira LA, Nuvunga JJ, Pires LPM, Von Pinho RG, Balestre M, 2018. AMMI bayesian models to study stability and adaptability in maize. Agron J. 110 (5): 1765-1776, 2018. doi: 10.2134/agronj2017.11.0668.

Kempton RA, 1984. The use of biplots in interpreting variety by environment interactions. J. Agric. Sci. 103(1): 123-135. DOI: 10.1017/S0021859600043392.

Lavoranti OJ, Dias CDS, Kraznowski WJ, 2007. Phenotypic stability via ammi model with bootstrap re-sampling**.** Pesq Flor Bras 54: 45-52.

Liu, G, 2001. Bayesian Computation for General Linear–Bilinear Models. master thesis, University of Kentucky, Lexington.

Maia MCC, Vello NA, Lavoranti OJ, Dias CDS, Vencovsky R, Rocha MDM, Nunes GDS, 2006. AMMI Bootstrap no estudo da interação genótipos por ambientes em soja. Rev. Mat Est 24(3): 7-24.

Mendiburu F, 2020. Agricolae: Statistical Procedures for Agricultural Research. R package version. 1.2.

Perez-Elizalde S, Jarquin D, Crossa J, 2012. A general Bayesian estimation method of linear-bilinear models applied to plant breeding trials with genotype × environment interaction. J Agric Biol Environ Stat, 17(1): 15-37. DOI: 10.1007/s13253-011-0063-9.

Piepho HP, 1995. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. Theor Appl Genet 90:(3-4): 438-443. DOI: 10.1007/BF00221987.

Team, R Core, 2019. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Raftery AE, Lewis S, 1992. How many iterations in the Gibbs sampler? In: J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, **Bayesian Statistics**. Oxford University Press, Oxford, UK. 763–773.

Romão RF, Nuvunga JJ, da Silva CP, da Oliveira LA, Mendes CTE, Balestre M, 2019. Predictive ability of ammi and factorial analytical models in the study of unbalanced multi-environment data. Gen Mol Res, 18(3). DOI: 10.4238/gmr18176.

Viele K, Srinivasan C, 2000. Parsimonious estimation of multiplicative interaction in analysis of variance using Kullback-Leibler Information. J Stat Plan Infer 84(1-2): 201-219. DOI: 10.1016/S0378-3758(99)00151-2.

Yan W, Glover, KD, Kang MS, Yang RC, Crossa J, Burgueño J, 2010. Comment on “Biplot Analysis of Genotype × Environment Interaction: Proceed with Caution,” by R.-C. Yang, J. Crossa, P.L. Cornelius, and J. Burgueño in Crop Science 2009 49:1564–1576. Crop Sci 50(4): 1121–1123. DOI: 10.2135/cropsci2010.01.0001le.

Yang RC, Crossa J, Cornelius PL, Burgueño J, 2009. Biplot analysis of genotype × environment interaction: Proceed with caution. Crop Sci 49(5): 1564-1576. DOI: 10.2135/cropsci2008.11.0665.

Yaseen M, Crossa J, Perez-Elizalde S, Jarquin D, Cotes JM, Viele K, Liu G, Cornelius PL, 2018. bayesammi: Bayesian Estimation of the Additive Main Effects and Multiplicative Interaction Model. R package version. 0.1.0.

Supplementary material

Table S1. Genotype to environment adaptability suggestions based on biplot analysis (Bayesian-AMMI).

Grupos	Genótipos	Ambientes
$(0,0)$	G1, G2, G3, G4, G5, G6, G7, G8	E14
$(+,+)$	G10	E1, E6, E12, E19
$(0,+)$	-	E3, E20
$(-,+)$	-	E18
$(-,0)$	G11	E2, E4, E11, E13, E16
$(-,-)$	-	E9, E15
$(0,-)$	G9	-
$(+,-)$	G12	E5, E7, E10, E17
$(0,+)$	-	E8

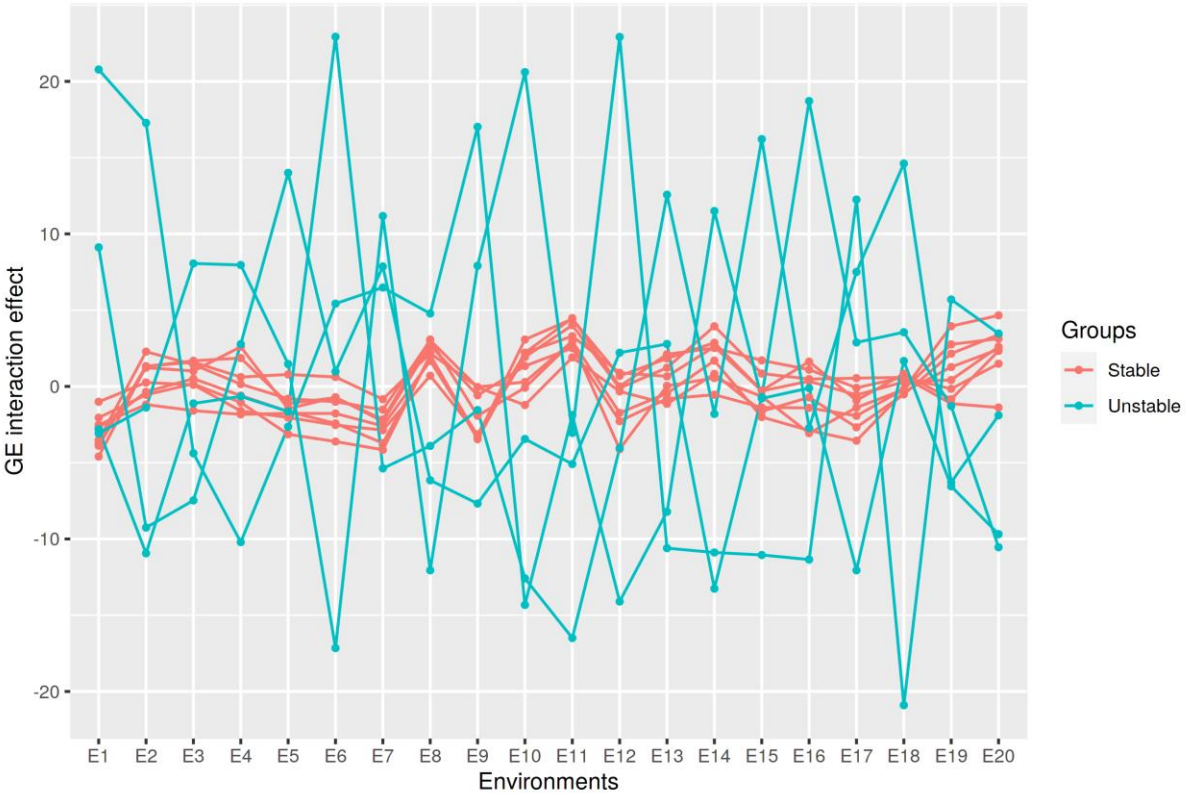


Figure S1. Graphical representation of the GEI interaction effect used in the simulation, separating the genotypes into two subgroups (stable and unstable) in the environments.

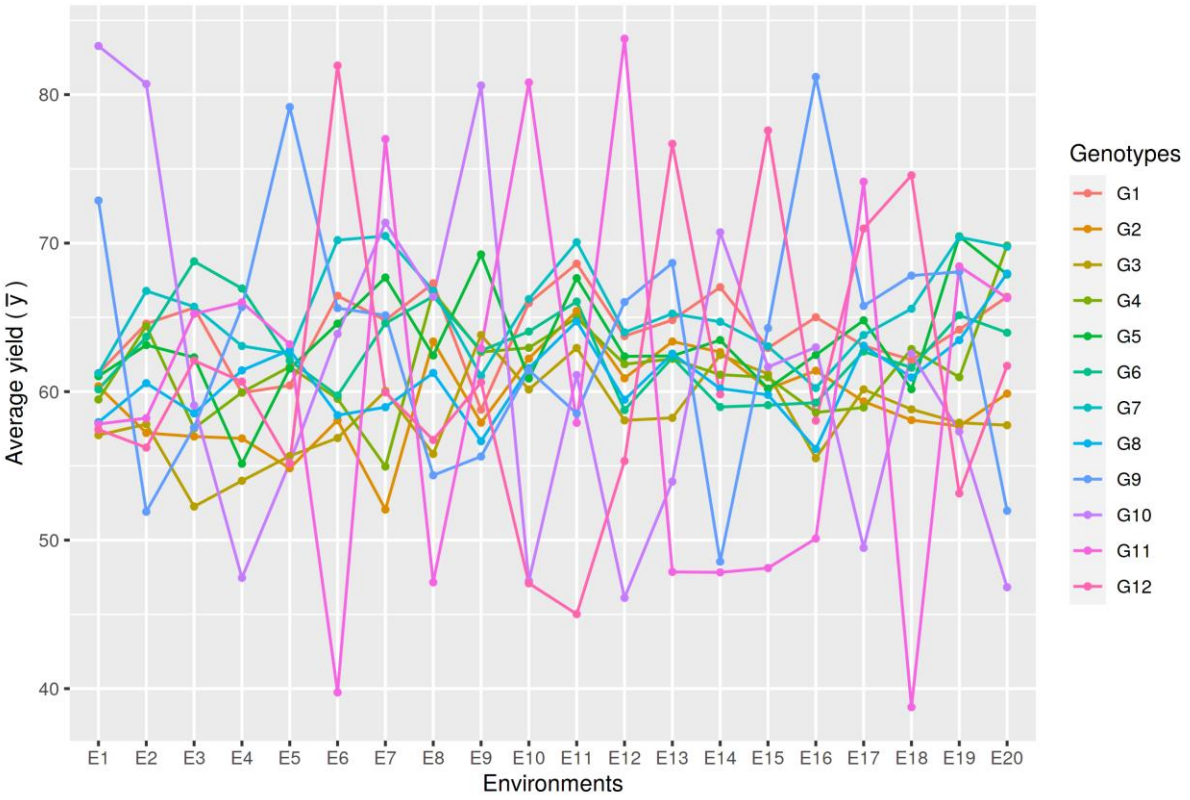


Figure S2. Graphical representation of the relationship of each genotype in the environments by means of the average yield response $(\bar{y})$.

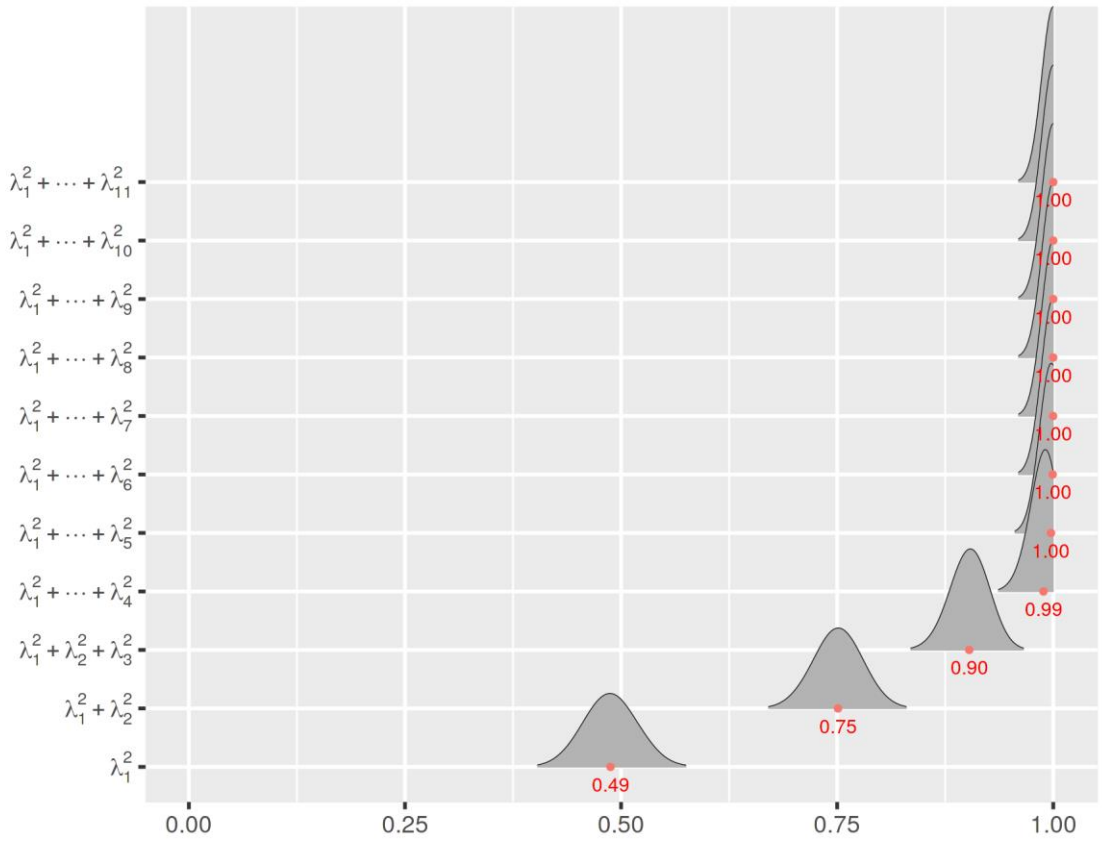


Figure S3. Aggregate proportion of GEI variance explained by adding a new dimension

ARTIGO 2

Amostragem adequada para o AMMI Bayesiano com dados ordinais

O artigo será traduzido e submetido à uma revista na área de estatística com enfase em análise de dados multiambientais (versão preliminar)

Amostragem adequada para o AMMI Bayesiano com dados ordinais

RESUMO

Neste trabalho apresentamos o AMMI bayesiano aplicado a dados ordinais em experimentos realizados em vários ambientes (MET). Foram utilizadas uma simulação de dados ordinais e dados reais de MET com uma variável contínua, discretizada após padronização para fins de referência de análise. A implementação da variável latente com função de ligação acumulada probit e a amostragem da posteriori condicional conjunta de parâmetros threshold e desta variável (seguindo Cowles, 1996) permitiram aplicar o AMMI e reconhecer padrões na interação consistentes com o que se esperava na simulação e na variável real de referência. O método proposto se mostrou relativamente menos poderoso, mas mais rigoroso do que análises tradicionais supondo que dados ordinais são contínuos. A implementação aqui apresentada é mais complexa do que se poderia esperar para contagens ou proporções, e pretendemos em trabalhos futuros estender a análise AMMI em tais configurações de modelos não normais.

Palavras-chave: AMMI-bayesiano, Dados Ordinais, Modelos de Limiar.

INTRODUÇÃO

O estudo da interação genótipo x ambiente (GEI) é parte importante de programas de melhoramento genético de plantas, dado que procura-se não apenas altos rendimentos, mas também estabilidade e adaptabilidade fenotípica de genótipos em relação ao ambiente de destino. A presença de GEIs significativas acarreta complicações na seleção e recomendação dos genótipos em diversos ambientes, podendo reduzir a correlação entre os valores fenotípicos futuros e valores genotípicos estimados e até mesmo provocar mudanças do seu ordenamento e dos genótipos selecionados para cada ambiente. Neste caso, é preciso investigar as causas e implicações destas interações (KANG; GORMAN,1989; ROMAGOSA; FOX, 1993)

Diversas técnicas estatísticas foram desenvolvidas e empregadas ao longo dos anos, para o estudo de dados de ensaios multiambientes (MET), como a análise da variância conjunta de experimentos, a análise de regressão em indicadores ambientais e técnicas não paramétricas que, no entanto, são limitadas na identificação de padrões de interação (FINLAY; WILKINSON, 1963; LIN; BINNS, 1988; CROSSA, 1990). Dentre estas técnicas destaca-se o modelo de efeitos principais aditivos e interação

multiplicativa (AMMI), que combina técnicas uni e multivariadas para estimação dos parâmetros e permite avaliar estabilidade e adaptabilidade em uma unica abordagem (ZOBEL et al., 1988; CROSSA, 1990**)**. Além disso, o padrão de respostas de genótipos a diferentes ambientes é organizado graficamente em biplots e pode ser interpretado com base nas propriedades de produtos internos entre escores genotípicos e ambientais (GABRIEL, 1971; KEMPTON, 1984).

Apesar destas vantagens, a análise AMMI clássica apresenta as reconhecidas limitações de modelos de efeitos fixos, como as dificuldades em acomodar dados desbalanceados ou com heterogenidade de variâncias e dificuldades de quantificação e de interpretação da incerteza nos escores do biplot. Isto motivou a busca por métodos mais flexíveis, inicialmente com versões de modelos lineares-bilineares mistos e com a adoção de estruturas fator-analíticas na matriz de variância-covariância genética para ambientes (PIEPHO, 1997, 1998; SMITH et al, 2001; PIEPHO; MOHRING, 2006). Nestes modelos a heterogeneidade de variância dos erros entre os ambientes, e a correlação espacial dentro de cada ambiente podem ser modeladas sem grandes dificuldades. Além disso, conjuntos de dados incompletos podem ser analisados de forma mais direta.

Outra opção interessante é a utilização da inferência bayesiana com suas diversas possibilidades de adicionar informações *a priori* às analises. Embora haja evidência de pouca utilização prática, a literatura sobre o uso da inferência bayesiana no melhoramento de plantas já tem mais de uma década, sendo que alguns trabalhos apontam vantagens expressivas de sua aplicação (EDWARD; JENNINK, 2006; COTES et al., 2006; ORELLANA; EDWARDS; CARRIQUIRY, 2014; NUVUNGA et al., 2019). Especificamente para os dados de MET, podemos citar (CROSSA et al., 2011; JOSSE 2014) que abordaram análises bayesianas para modelos lineares bilineares (AMMI-Bayesiano) como um procedimento paramétrico flexível para incorporar inferência ao biplot.

A primeira versão bayesiana do modelo AMMI foi apresentada por Viele e Srnivasan (2000) que mostraram como lidar com as restrições paramétricas e amostrar corretamente os vetores singulares utilizando metodos de Monte Carlo via cadeias de Marcov (MCMC). Contribuições importantes ao método foram dadas por Liu (2001), que derivou um conjunto completo de densidades condicionais a posteriori permitindo a amostragem do tipo Gibbs Sampling de todos os parâmetros, tornando o algoritmo mais rápido e estável. Crossa et al. (2011) e Perez-Elizalde et al. (2012) deram continuidade

aos trabalhos pioneiros de Viele e Srinivasan (2000) e Liu (2001), mostrando como elicitar distribuições *a priori* de experimentos anteriores e, especialmente, incorporar elementos de inferências aos biplot. Outros trabalhos surgiram trazendo novas contribuições visando o aperfeiçoamento do método **(**JOSSE et al, 2014; de OLIVEIRA et al., 2015; da SILVA et al., 2015; JARQUIN et al., 2016; da SILVA et al., 2019; JÚNIOR et al., 2018**).**

Exemplos encontrados na literatura sobre o AMMI Bayesiano pressupõe respostas contínuas em delineamentos aleatorizados, com boa justificativa para a aproximação normal dos dados experimentais. Por outro lado, uma parte importante da atividade de pesquisa no melhoramento vegetal se baseia em atribuição de notas subjetivas em escalas ordinais, como por exemplo, o grau de resistência ou de suceptibilidade das cultivares a pragas e doenças, notas visuais para a arquitetura da planta, escores subjetivos associados à propagação ou à produtividade da parcela em gramíneas, notas para o formato de tubérculos e raízes, notas para o formato e, ou, a aparência de frutos que serão consumidos in natura, dentre outras. Todos estes caracteres podem ser analisados com aproximações normais das escalas originais e, ou, com transformações de dados, mas acreditamos que da mesma forma que em outros aspectos do melhoramento, isto pode levar a perdas de informação (CORRÊA et al, 2016) . Abordagens utizando método linear generalizado para os modelos lineares-bilineares foram desenvolvidas (VAN EEUWIJK, 1995; GABRIEL, 1998). Nossa hipótese de trabalho é que no caso do AMMI bayesiano, a incorporação de distribuições adequadas aos dados de natureza ordinal pode ser vantajosa.

Uma forma direta para a análise bayesiana de dados ordinais é a utilização de modelos de variáveis latentes (comumente referidos na literatura de doenças de plantas como *liability*, do Inglês suscetibilidade) divididas por *thresholds* (limiares) em que as respostas ordinais saltam degrais de uma escala discreta (GIANOLA; FOULLEY, 1983). O modelo indica que o valor observado pertence a uma categoria determinada pela ocorrência do valor da variável latente entre os seus limites (SÖRENSEN et al., 1995). Trabalhos pioneiros na utilização desse modelo foram abordados no melhoramento animal por Dempster e Lerner (1950) e na suscetibilidade a doenças genéticas humanas por Falconer (1965, 1967). Algoritmos para a implementação desse método podem ser encontrados em diversos trabalhos (ALBERT; CHIB, 1993; COWLES, 1996; NANDRAM; CHEN, 1996)**.**

Este trabalho tem por objetivo desenvolver a análise AMMI bayesiana para dados ordinais e verificar o seu poder no reconhecimento de padrões no estudo da GEI, comparando esta análise com a aproximação normal. Exemplificaremos o uso e verificaremos o poder do método com um conjunto de dados simulados e outro exemplo de uma variável contínua discretizada proveniente de experimentos reais.

MATERIAL E MÉTODOS

Conjunto de dados simulados

O conjunto de dados ordinais foi simulado de acordo com os seguintes passos:

i) Inicialmente foram simulados dados contínuos, com efeitos principais de genótipos (G) e ambientes (E) geradas de distribuições gaussianas, $N(0,5)$ e $N(0,1)$, respectivamente. Foram simulados efeitos de 15 genótipos (G1-G15) avaliados em 11 ambientes (E1-E11), utilizando delineamento em blocos completos casualizados com três repetições.

ii) Para o efeito de interação, foram simulados 2 padrões de resposta de acordo com a figura 1. Genótipos foram avaliados em ambientes originados de distribuições Gaussianas $N(0,1)$ com/sem a restrição de possuírem sinais negativos (G1-G10, genótipos estáveis), e outro grupo formado por genótipos avaliados em ambientes, gerados a partir distribuições Gaussianas $N(5,11)$ (G11-G15, genótipos instáveis).

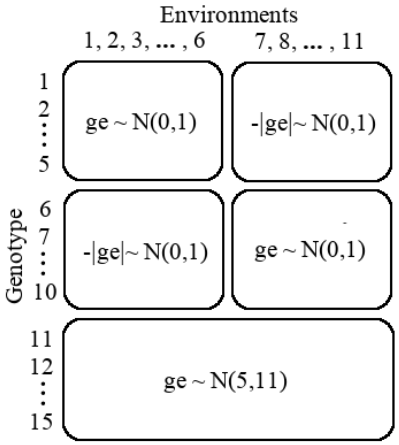


Figure 1: Esquema de simulação dos padrões de conjunto de dados da interação GE

iii) O passo seguinte foi a categorização da resposta, feita com atribuição de limiares em pontos da distribuição Normal Padrão. Para esse estudo, foram considerados 4 categorias. Os dados do experimento simulado foram padronizados e os quartis (0.25, 0.5, 0.75) da normal padrão foram tomados como limiares.

Conjunto de dados reais

O conjunto de dados multiambientais, foi obtido no site do Indian Agricultural Statistics Research Institute (IASRI, 2021). Foi realizado um estudo com 12 variedades de mostarda (G1-G12), em 6 localidades na Índia (A1-A6), em um delineamento em blocos casualizados completos com três repetições. O rendimento de sementes foi medido em t/ha. Para a produção de dados ordinais, efetuou-se a padronização dos dados e a atribuição de quatro categorias com o mesmo procedimento descrito para os dados simulados.

Modelo AMMI-Bayesiano para dados ordinais

Seja o vetor $\mathbf{y} = \{y_i\}\ (i = 1, 2, ..., n)$ com $n = rw$ respostas categóricas ordinais e mutuamente exclusivas, sendo r indicando os níveis do efeito de genótipos repetidos e $w = bd$ vezes, sendo b o número de blocos e d o número de ambientes, respectivamente. O vetor $\mathbf{l} = \{l_i\}\ (i = 1, 2, \ldots, n)$ composto por n valores da variável latente (*liability*) representando em uma escala subjacente o vetor $\mathbf{y}$ com base em $c-1$ *thresholds* $(\tau_{\min} < \tau_1 < \tau_2 < \ldots < \tau_{c-1} < \tau_{máx})$ em que $\tau_{\min} = -\infty$ e $\tau_{máx} = \infty$, é equivalente à:

$$y_i = \begin{cases} 1 & \text{se } -\infty < l_i < \tau_1 \\ 2 & \text{se } \tau_1 < l_i < \tau_2 \\ & \vdots \\ c & \text{se } \tau_{c-1} < l_i < \infty \end{cases}$$

O modelo AMMI para $\mathbf{l}$ com notação vetorial é dado por:

$$\mathbf{l} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \sum_{k=1}^{s} \lambda_k diag(\mathbf{Z}\boldsymbol{\alpha}_\mathbf{k})\mathbf{X}_2\boldsymbol{\gamma}_\mathbf{k} + \varepsilon \qquad (1)$$

sendo $\boldsymbol{\beta}_{w\times 1}$ e $\boldsymbol{\delta}_{r\times 1}$ os vetores contendo os efeitos de blocos hierarquizados dentro de ambientes e o vetor de efeitos principais de genótipos, respectivamente. Os parâmetros λ_k, $\boldsymbol{\alpha}_k$ e $\boldsymbol{\gamma}_k$ representam, respectivamente, valores singulares e vetores singulares, genotípicos e ambientais, associados ao k-ésimo componente principal da decomposição por valor singular (DVS) da matriz de interação $GEI_{r\times w}$, com $k = 1, \ldots, s$ sendo $s = \min(r, w)$.

Os termos bilineares do modelo (1) estão sujeitos à restrição de ordem em relação a $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s \geq 0)$, e também a restrição de ortonormalidade dos vetores singulares, ou seja, $\boldsymbol{\alpha}_k^{\top}\boldsymbol{\alpha}_{k'} = \boldsymbol{\gamma}_k^{\top}\boldsymbol{\gamma}_{k'} = 0$, para $k \neq k'$ e $\boldsymbol{\alpha}_k^{\top}\boldsymbol{\alpha}_{k'} = \boldsymbol{\gamma}_k^{\top}\boldsymbol{\gamma}_{k'} = 1$ para $k = k'$. As matrizes de delineamento $\mathbf{X}_{1(n\times l)}$ e $\mathbf{Z}_{(n\times r)}$ estão associadas aos vetores $\boldsymbol{\beta}$ e $\boldsymbol{\delta}$. O vetor de erros aleatórios $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0},\sigma_e^2\mathbf{I})$, em que N_n é normal multivariada com vetor de médias igual a zero e matriz de variância e covariância igual a $\sigma_e^2\mathbf{I}$, em que $\sigma_e^2 = 1$, valor adotado aqui por ser padrão em análises de dados ordinais. Ressaltando que se forem adotadas prioris adequados a todos os parâmetros, essa restrição não seria necessariamente obrigatória para convergências a posteriori (BERNARDO; SMITH, 1994; SORENSEN; GIANOLA, 2002).

Distribuições *a priori*

Para os parâmetros lineares e bilineares do modelo AMMI, as prioris são as mesmas utilizadas em Oliveira (2016), já para o Threshold foi atribuido uma priori não informativa:

$\boldsymbol{\tau} \sim$ constante;

$\boldsymbol{\beta} \sim$ constante;

$\boldsymbol{\delta} \mid \boldsymbol{\mu}_{\delta}, \sigma_{\delta}^2 \sim N\ (\mathbf{0}, \mathbf{I}\sigma_{\delta}^2)$ sendo que $\sigma_{\delta}^2 \sim 1/\sigma_{\delta}^2$ (priori de Jeffreys);

$\lambda_k \mid \mu_{\lambda_k}, \sigma_{\lambda_k}^2 \sim N^{+}\left(\mathbf{0}, 10^8\right)$, sendo $\lambda_k \sim$ constante;

$\alpha_k \sim$ uniforme esférica no subespaço corrigido;

$\gamma_k \sim$ uniforme esférica no subespaço corrigido;

Para os parâmetros representando os thresholds, efeito principal de ambiente e valores singulares foram definidos prioris não informativas, proporcionais a uma constante. Para os vetores singulares atribuíram-se prioris esféricas uniformes, com objetivo de declarar um conhecimento vago sobre esses parâmetros. Para efeito de genótipos considerou-se uma priori hierárquica em dois níveis, sendo que a incerteza sobre a componente de variância σ_{δ}^2 é declarada por uma priori de Jeffreys, com essa escolha admite-se uma distribuição comum para genótipos semelhante ao que ocorre em modelos mistos em que efeitos de genótipos são considerados aleatórios.

Distribuição conjunta *a posteriori*

Dada a observação $y_i = j$ com $(j = 1,2,...,c)$, pertencente a uma determinada categoria condicionada aos thresholds, $\boldsymbol{\tau}' = (\tau_{\min} < \tau_1 < \tau_2 < \ldots < \tau_{c-1} < \tau_{\text{máx}})$, e aos parâmetros lineares e bilineares, a função de verossimilhança é expressa hierarquicamente pelo seguinte produto:

$$p(\mathbf{y} \mid \mathbf{l}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}) = p(\mathbf{y} \mid \mathbf{l}, \boldsymbol{\tau})\, p(\mathbf{l} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}) \qquad (1)$$

em que a primeira componente $p(\mathbf{y} \mid \mathbf{l}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{c} I(\tau_{j-1} < l_i < \tau_j) I(y_i = j) \right]$ (sendo $I(\cdot)$ uma função indicadora $(1,0)$, que assume 1, caso o valor da observação pertença a uma determinada categoria, e atribui zero para as demais). Na segunda componente, a variável latente, dado os parâmetros do modelo, é condicionalmente independente e identicamente distribuída. O vetor $\mathbf{l}$ segue uma distribuição normal multivariada dada por:

$$\mathbf{l} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\beta} \sim N_n(\boldsymbol{\mu}_y, \mathbf{I}_n \sigma_e^2) \qquad (2)$$

com $\boldsymbol{\mu}_y = E[\mathbf{l}] = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \sum_{k=1}^{s} \lambda_k diag(\mathbf{Z}\boldsymbol{\alpha}_k)\mathbf{X}_2\boldsymbol{\gamma}_k$, sendo a esperança do modelo (1) e a variância residual com restrição, $\sigma_e^2 = 1$.

Sendo assim, a distribuição conjunta a posteriori para os parâmetros a partir da atualização das informações a priori por meio da função de verossimilhança é dada por:

$$\begin{aligned} p(\mathbf{l}, \boldsymbol{\tau}, \boldsymbol{\eta} \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \mathbf{l}, \boldsymbol{\tau})\, p(\mathbf{l} \mid \boldsymbol{\eta})\, p(\boldsymbol{\tau}, \boldsymbol{\eta}) \\ &\propto p(\mathbf{y} \mid \mathbf{l}, \boldsymbol{\tau}) \left[\prod_{i=1}^{n} p(l_i \mid \boldsymbol{\eta}) \right] p(\boldsymbol{\delta} \mid \boldsymbol{\mu}_\delta, \sigma_\delta^2)\, p(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \sigma_\beta^2)\, p(\sigma_\delta^2 \mid \nu_\delta, S_\delta^2) \times \\ &\quad \times \left[\prod_{k=1}^{s} p(\lambda_k \mid \boldsymbol{\mu}_{\lambda_k}, \sigma_{\lambda_k}^2)\, p(\boldsymbol{\alpha}_k)\, p(\boldsymbol{\gamma}_k) \right] \end{aligned}$$

em que $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_\delta^2)$.

Distribuições *a posteriori* condicionais completas

As distribuições *a posteriori* para os parâmetros $\mathbf{l}$ e $\boldsymbol{\tau}$ do modelo, foram obtidas de acordo com a proposta de Cowles (1996) em relação ao trabalho de Albert e Chib

(1993), em que a convergência da cadeia ocorre de forma mais objetiva quando realizada conjuntamente. Esse procedimento requer um número menor de iterações para a obtenção das estimativas dos parâmetros, sendo defina a distribuição a posteriori condicional conjunta para $\mathbf{l}$ e $\boldsymbol{\tau}$ a partir da equação: $p(\mathbf{l},\boldsymbol{\tau}|\cdots)=\prod_{i=1}^{n}\sum_{j=1}^{c}\phi(\boldsymbol{\mu}_{y_i},1)$ (mais detalhes em Apêndice S1**).**

As distribuições a posteriori condicionais completas para os parâmetros lineares e bilineares do modelo são fornecidos a seguir, para mais detalhes sobre os desenvolvimentos algébricos, consultar de Oliveira et al. (2015).

Distribuições condicionais completas *a posteriori* para os parâmetros lineares:

$$\boldsymbol{\beta}|\cdots\sim N\left[(\mathbf{X}_1^{\top}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\top}(\mathbf{l}-\mathbf{Z}\boldsymbol{\delta}-\boldsymbol{\Theta}),(\mathbf{X}_1^{\top}\mathbf{X}_1)^{-1}\right],\ \text{onde}\ \ \boldsymbol{\Theta}=\sum_{k=1}^{t}\lambda_k diag(Z\alpha_k)X_2\gamma_k. \quad (3)$$

$$\boldsymbol{\delta}|\cdots\sim N\left[\left(\mathbf{Z}^{\top}\mathbf{Z}+\mathbf{I}\frac{1}{\sigma_{\boldsymbol{\delta}}^2}\right)^{-1}\mathbf{Z}^{\top}\left(\mathbf{l}-\mathbf{X}_1\boldsymbol{\beta}-\boldsymbol{\Theta}\right),\left(\mathbf{Z}^{\top}\mathbf{Z}+\mathbf{I}\frac{1}{\sigma_{\boldsymbol{\delta}}^2}\right)^{-1}\right]. \quad (4)$$

$$\sigma_{\boldsymbol{\delta}}^2|\ldots\sim \text{Escalada}-\chi^{-2}\left[n_{\boldsymbol{\delta}},\boldsymbol{\delta}^{\top}\boldsymbol{\delta}\right]. \quad (5)$$

Distribuições condicionais completas *a posteriori* para os parâmetros bilineares:

$$\lambda_k|\cdots\sim N^{+}\left[\left(\phi_k^{\top}\phi_k\right)^{-1}\phi_k^{\top}\boldsymbol{\Delta}_{k'},\left(\phi_k^{\top}\phi\right)^{-1}\right] \quad (6)$$

em que $\boldsymbol{\Delta}_{k'}=\mathbf{l}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\delta}-\sum_{k'\neq k}^{t}\lambda_{k'}diag(\mathbf{Z}\boldsymbol{\alpha}_{\mathbf{k}'})\mathbf{X}_2\boldsymbol{\gamma}_{\mathbf{k}'}$, $\phi_k=diag(\mathbf{Z}\boldsymbol{\alpha}_k)\mathbf{X}_2\boldsymbol{\gamma}_k$ e $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_t\geq 0$**.**

$p(\boldsymbol{\alpha}_k|\ldots)\propto exp\left\{\lambda_k\boldsymbol{\alpha}_k^{\top}\boldsymbol{\mu}_{\alpha_k}\right\}$, em que $\boldsymbol{\mu}_{\alpha_k}=\boldsymbol{\Lambda}_k^{\top}\left(\mathbf{l}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\delta}\right)$ e $\boldsymbol{\Lambda}_k=diag\left(\mathbf{X}_2\boldsymbol{\gamma}_k\right)\mathbf{Z}$

$p(\boldsymbol{\gamma}_k|\ldots)\propto exp\left\{\lambda_k\boldsymbol{\gamma}_k^{\top}\boldsymbol{\mu}_{\gamma_k}\right\}$, em que $\boldsymbol{\mu}_{\gamma_k}=\boldsymbol{\Omega}_k^{\top}\left(\mathbf{l}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\delta}\right)$ e $\boldsymbol{\Omega}_k=diag\left(\mathbf{Z}\boldsymbol{\alpha}_k\right)\mathbf{X}_2$.

Para amostrar os vetores singulares $\boldsymbol{\alpha}_k$ utiliza-se transformações lineares ortogonais devido a dificuldade nos suportes das condicionais a posteriori que não são triviais de serem obtidas (VIELLE; SRINIVASAN, 2000). Assim, utiliza-se da transformação linear $\boldsymbol{\alpha}_k=\mathbf{H}_k\boldsymbol{\alpha}_k^{*}$, em que $\boldsymbol{\alpha}_k^{*}$ é amostrado em uma esfera unitária no subespaço corrigido $r-m$ dimensional, sendo que o vetor singular é obtido no correto subespaço r dimensional, utilizando-se da matriz $\mathbf{H}_k$ (ortornormal e ortogonal aos demais vetores $\boldsymbol{\alpha}_{k'}$ $(k'\neq k)$) (LIU, 2001). Para amostrar $\boldsymbol{\alpha}_k^{*}$ segue:

$$\boldsymbol{\alpha}_k^* \mid ... \sim VFM\left(r-m, \frac{c_k\lambda_k}{\sigma_e^2}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}_k} \right),$$

em que $c_k\lambda_k/\sigma_e^2$ é o parâmetro de concentração $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\alpha}_k} = c_k^{-1}\mathbf{H}_k^{\top}\boldsymbol{\mu}_{\alpha_k}$ com $c_k = \sqrt{(\mathbf{H}'_k\boldsymbol{\mu}_{\alpha k})'\,\mathbf{H}'_k\boldsymbol{\mu}_{\alpha_k}} = \sqrt{\boldsymbol{\mu}'_{a_k}\mathbf{H}_k\mathbf{H}'_k\boldsymbol{\mu}_{\alpha_k}}$ e $m = s-1$.

Analogamente, para amostrar $\boldsymbol{\gamma}_k$, obtêm-se a distribuição de $\boldsymbol{\gamma}_k^* = \mathbf{R}_k^{\top}\boldsymbol{\gamma}_k$ no subespaço corrigido $c-m$ como:

$$\boldsymbol{\gamma}_k^* \mid ... \sim VMF\left(c-m, \frac{d_k\lambda_k}{\sigma_e^2}, \tilde{\boldsymbol{\mu}}_{\gamma_k} \right).$$

sendo $d_k\lambda_k/\sigma_e^2$ o parâmetro de concentração e $\tilde{\boldsymbol{\mu}}_{\gamma_k} = d_k^{-1}\mathbf{R}_k^{\top}\boldsymbol{\mu}_{\gamma_k}$, em que $\mathbf{R}_k$ é ortonormal e ortogonal aos demais vetores $\boldsymbol{\gamma}_{k'}$ com $d_k = \sqrt{\boldsymbol{\mu}_{a_k}^{\top}\mathbf{R}_k\mathbf{R}_k^{\top}\boldsymbol{\mu}_{\alpha_k}}$, de forma que $\boldsymbol{\gamma}_k = \mathbf{R}_k\boldsymbol{\gamma}_k^*$.

Algoritmo: Gibbs Sampling com passo de Metropolis-Hasthings

A estimação dos parâmetros do modelo AMMI Bayesiano aqui proposto é realizada tomando uma amostra considerada suficientemente grande da distribuição a posteriori. Utilizou-se o Amostrador de Gibbs de forma direta para amostrar os parâmetros lineares e bilineares de acordo com o trabalho de Oliveira et al. (2015) e o passo de aceitação/rejeição do algoritmo Methropolis-Hastings para amostras da distribuição conjunta da variavel latente e dos thresholds, condicionados aos demais parâmetros do modelo, (conforme Cowles, 1996).

As cadeias de Markov (MCMC) são obtidas de acordo com os seguintes passos:

Primeira Etapa: Amostragem dos Parâmetros Lineares e Bilineares

A1 - atribuir valores iniciais aos parâmetros $\boldsymbol{\Phi}^0 = [\boldsymbol{\beta}^0, \boldsymbol{\delta}^0, (\sigma_\delta^2)^0, \lambda^0, \boldsymbol{\alpha}^0, \boldsymbol{\gamma}^0, \boldsymbol{\tau}^0, \mathbf{l}^0]$.

A2- A partir desses valores iniciais a t-ésima iteração pode ser obtida da seguinte forma:

a) Gerar $\boldsymbol{\beta}^t \mid \boldsymbol{\delta}^{t-1}, (\sigma_{\boldsymbol{\delta}}^2)^{t-1}, \lambda^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\tau}^{t-1}, \mathbf{l}^{t-1}$ a partir da condicional a posteriori (3);

b)Gerar $\boldsymbol{\delta}^t \mid \boldsymbol{\beta}^t, (\sigma_{\boldsymbol{\delta}}^2)^{t-1}, \lambda^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\tau}^{t-1}, \mathbf{l}^{t-1}$ a partir da condicional a posteriori (4);

c) Gerar $(\sigma_{\boldsymbol{\delta}}^2)^t \mid \boldsymbol{\beta}^t, \boldsymbol{\delta}^t, \lambda^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\tau}^{t-1}, \mathbf{l}^{t-1}$ a partir da condicional a posteriori (5);

d) Gerar a t-ésima observação dos parâmetros bilineares por meio da sequência d1), d2) e d3) abaixo, para $k = 1, 2, ..., s$:

d1) Gerar $\lambda_k^t \,|\, \boldsymbol{\beta}^t, \boldsymbol{\delta}^t, (\sigma_g^2)^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\gamma}^{t-1}, \boldsymbol{\tau}^{t-1}, \mathbf{l}^{t-1}$ a partir da condicional a posteriori (6);

d2) Gerar $(\boldsymbol{\alpha}_k)^t \,|\, \boldsymbol{\beta}^t, \boldsymbol{\delta}^t, (\sigma_g^2)^t, \lambda_k^t, \boldsymbol{\gamma}_k^{t-1}, \boldsymbol{\tau}^{t-1}, \mathbf{l}^{t-1}$:

d2-i) Gerar $\left(\boldsymbol{\alpha}_k^*\right)^l$ da $VFM\left(r-s, c_k\lambda_k^t, \boldsymbol{\mu}_{\alpha_k}\right)$;

d2-ii) obter $\left(\boldsymbol{\alpha}_k\right)^t = \mathbf{H}_k\left(\boldsymbol{\alpha}_k^*\right)^t$.

d3) Gerar $(\boldsymbol{\gamma}_k)^l \,|\, \boldsymbol{\beta}^l, \boldsymbol{\delta}^l, (\sigma_{\boldsymbol{\delta}}^2)^l, \lambda_k^l, \boldsymbol{\alpha}_k^l, \boldsymbol{\tau}^{l-1}, \mathbf{l}^{l-1}$:

d3-i) Gerar $\left(\boldsymbol{\gamma}_k^*\right)^t$ da $VFM\left(c-s, d_k\lambda_k^t, \boldsymbol{\mu}_{\gamma_k}\right)$;

d3-ii) obter $(\boldsymbol{\gamma}_k)^t = \mathbf{D}_k\left(\boldsymbol{\gamma}_k^*\right)^t$.

Segunda Etapa: Amostragem da variavel latente e thresholds

A ideia está na escolha da densidade alvo e da densidade apropriada que seja fácil de amostrar e com boa probabilidade de aceitação. Amostra-se a distribuição conjunta a partir do produto de uma distribuição independente e outra condicional: $p(\mathbf{l}, \boldsymbol{\tau} \,|\, \cdots) = p(\mathbf{l} \,|\, \cdots)\, p(\mathbf{l} \,|\, \boldsymbol{\tau}, \cdots)$. A amostra $p(\mathbf{l}, \boldsymbol{\tau} \,|\, \cdots)$ é aceita com um passo Metropolis-Hastings. Após manipulações algébricas (mais detalhes em Cowles (1996)). O algoritmo é expresso da seguinte forma:

f) Gerar o t-ésimo candidato $\boldsymbol{\tau}_{\text{new}}$ para $\boldsymbol{\tau}^t$.

f1) Para $(j = 2, \ldots, c-1)$, gerar candidato a partir $\tau_{(j)\,\text{new}} \sim NT(\tau_j^{(t-1)}, 1)$, truncada em $(\tau_{(j-1)\,\text{new}}^{(t)}, \tau_{(j+1)}^{(t-1)})$ com $\tau_{\min} = -\infty$, $\tau_1 = \tau_{\text{Q}(1/c)}$ e $\tau_{\text{c=máx}} = \infty$, obtida pela condicional a posteriori (4), (sendo $\text{Q}_{(1/c)}$, quantil da normal padrão).

f2) Obter a probabilidade de aceitação $\alpha = \min(1, \text{R})$ para o vetor $\boldsymbol{\tau}$, em que R é obtido de acordo com a expressão, considerando $\sigma_\tau = 1$:

$$R = \prod_{i=1}^{n} \frac{\Phi\left(\tau_{(\text{y}_i)\,\text{new}} - \boldsymbol{\mu}_{\text{y}_i}^{(t-1)}\right) - \Phi\left(\tau_{(\text{y}_i-1)\,\text{new}} - \boldsymbol{\mu}_{\text{y}_i}^{(t-1)}\right)}{\Phi\left(\tau_{(\text{y}_i)}^{(t-1)} - \boldsymbol{\mu}_{\text{y}_i}^{(t-1)}\right) - \Phi\left(\tau_{(\text{y}_i-1)}^{(t-1)} - \boldsymbol{\mu}_{\text{y}_i}^{(t-1)}\right)}$$
$$\times \prod_{j=2}^{c-1} \frac{\Phi\left(\left(\tau_{(j+1)}^{t-1} - \tau_{(j)}^{t-1}\right)/\sigma_\tau\right) - \Phi\left(\left(\tau_{(j-1)\,\text{new}} - \tau_{(j)}^{t-1}\right)/\sigma_\tau\right)}{\Phi\left(\tau_{(j+1)\,\text{new}} - \tau_{(j)\,\text{new}}/\sigma_\tau\right) - \Phi\left(\tau_{(j-1)}^{(t-1)} - \tau_{(j)\,\text{new}}/\sigma_\tau\right)}$$

f3) A partir da probabilidade α definida, se aceito, $\boldsymbol{\tau}^t = \boldsymbol{\tau}_{\text{new}}$, caso contrário, $\boldsymbol{\tau}^t = \boldsymbol{\tau}^{t-1}$.

g) Gerar $\mathbf{l}^t \,|\, \boldsymbol{\beta}^t, \boldsymbol{\delta}^t, (\sigma_{\boldsymbol{\delta}}^2)^t, \lambda^t, \boldsymbol{\alpha}^t, \boldsymbol{\gamma}^t, \boldsymbol{\tau}^t$ a partir da condicional a posteriori (2)

RESULTADOS E DISCUSSÃO

Para as amostras de todas as cadeias simuladas verificaram-se boas propriedades, segundo os critérios de monitoramento utilizados. O fator de Independência (I) foi sempre menor que um (I<1) (RAFTERY; LEWIS, 1995) e todas as cadeias passaram no teste de estacionaridade (HEIDELBERGER; WELCH, 1983).

Dados simulados

Na Figura 2 são apresentados os traços e as respectivas densidades para os thresholds τ_2 e τ_3 que em conformidade com os resultados dos critérios de monitoramento utilizados, sugerem que as amostras são provenientes de distribuições *a posteriori* marginais estacionárias (para os dados reais os gráficos de traços das cadeias dos threshold são apresentados na Figura S1 (em Apêndice) o que também indica estacionariedade das distribuições). Padrões semelhantes foram observados para as cadeias de todos os parâmetros amostrados.

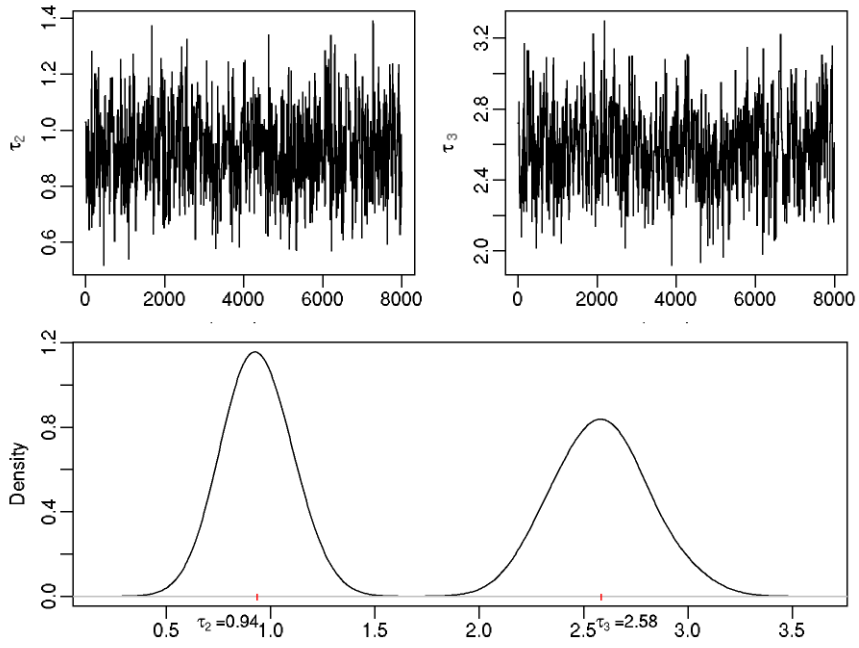


Figura 2: Gráfico de traços e densidades dos thresholds τ_2 e τ_3

No algoritmo empregado, apenas são amostrados k-2 thresholds, para k categorias, isso porque o τ_1 é fixado. Existem outras opções de algoritmos para a amostragem dos thresholds na literatura. Um desses algoritmos foi apresentado por

Gianola e Foulley (1983) em uma análise aplicada ao melhoramento genético animal com ênfase na predição do mérito genético representados em uma escala subjacente, sendo utilizado informações a priori sobre a população do qual a amostra foi selecionada. Albert e Chib (1993), por sua vez, desenvolveram uma metodologia para obter os parâmetros do modelo threshold utilizando regressão probit para modelar os dados latentes para duas ou mais categorias de respostas ordinais. Sorensen et al. (1995) também desenvolveu um algoritmo de Gibbs Sampling para inferências em modelos de threshold em um contexto de genética quantitativa, derivando as distribuições condicionais a posteriori para parâmetros do modelo de threshold.

Aqui utilizamos o algoritmo de Cowles (1996), que propôs um passo de Metropolis Hasting no algoritmo de Albert e Chib (1993), em que a geração dos dados latentes e estimativas dos parâmetros de thesholds são realizados conjuntamente, acelerando o processo de amostragem. Um algoritmo mais rápido foi proposto por Nandram e Chen (1996), mas a variável latente resultante gera interpretações mais complexas sobre os parâmetros genéticos por estimar uma componente extra para a variância da liability (CORRÊA et al., 2016). Adotamos o algoritmo de Cowles (1996), no entanto, pois embora não seja o mais rápido permite uma conjugação relativamente direta e boa adaptação a algoritmos pré-existentes de análise AMMI bayesiana (da SILVA et al. (2015); de OLIVEIRA et al. (2015)).

Na Figura 3 são apresentadas as médias a posteriori e as respectivas regiões de máxima densidade a posteriori (HPD) (a 95% de credibilidade) para efeitos principais de genótipos (G1-G15), em ordem crescente de magnitude para os dados simulados categorizados. Regiões HPD que se sobrepõem indicam efeitos semelhantes. Destaca-se neste grupo os genótipos G10, G2 e G15, em que os intervalos não incluem o valor zero (a uma probabilidade 95%) indicando que os mesmos possuem efeito médio superior a média geral. Entretando, para seleção e recomendação, é preciso estudar os efeitos da GEI.

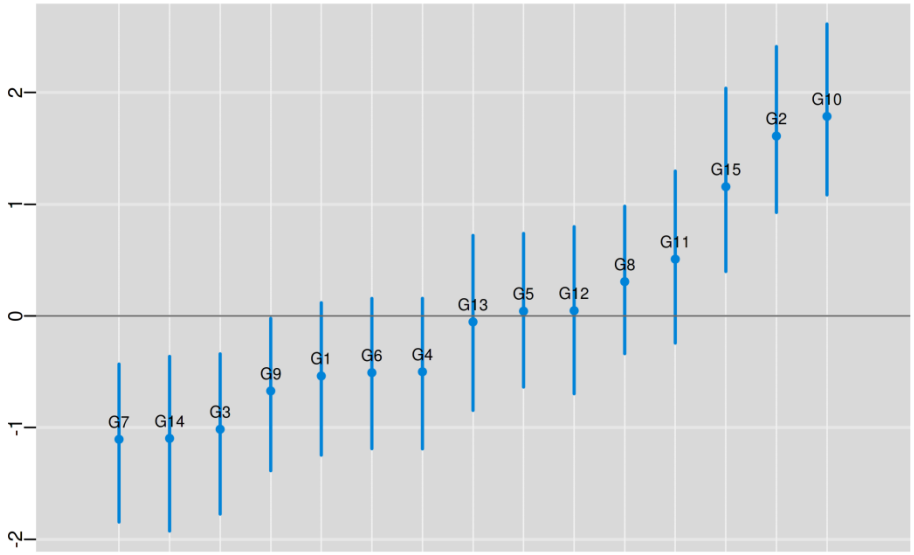


Figura 3: Regiões de máxima densidade a posteriori (HPD) a 95% de credibilidade para efeitos principais de genótipos para os dados simulados categorizados.

Os histogramas das distribuições a posteriori de valores singulares, para os dados simulados categorizados, exibidos na Figura 4, como pode ser visto, têm densidades consistentes com as restrições impostas no modelo $(\lambda_k > \lambda_{k+1})$. Visto que para os primeiros valores singulares $(\lambda_1 \text{ à } \lambda_4)$, a forma das distribuições são aproximadamentes simétricas (se assemelhando ao formato de sinos). Para λ_k's subsequentes, as distribuições tendem a aproximar-se cada vez mais do valor zero, o que sinaliza que eles contribuem cada vez menos para explicar a variabilidade dos dados. Efeitos de encolhimento mais acentuado são observados a partir do λ_6 indicando que esse componente principal e os demais subsequentes não são importantes para explicar o efeito da GEI. Os dois primeiros componentes principais explicaram 71,55% da soma de quadrado total da GEI recuperada pelo modelo. Essas informações são apresentadas na Tabela S1 que mostra resumos a posteriori para valores singulares e para o componente da variância genotípica.

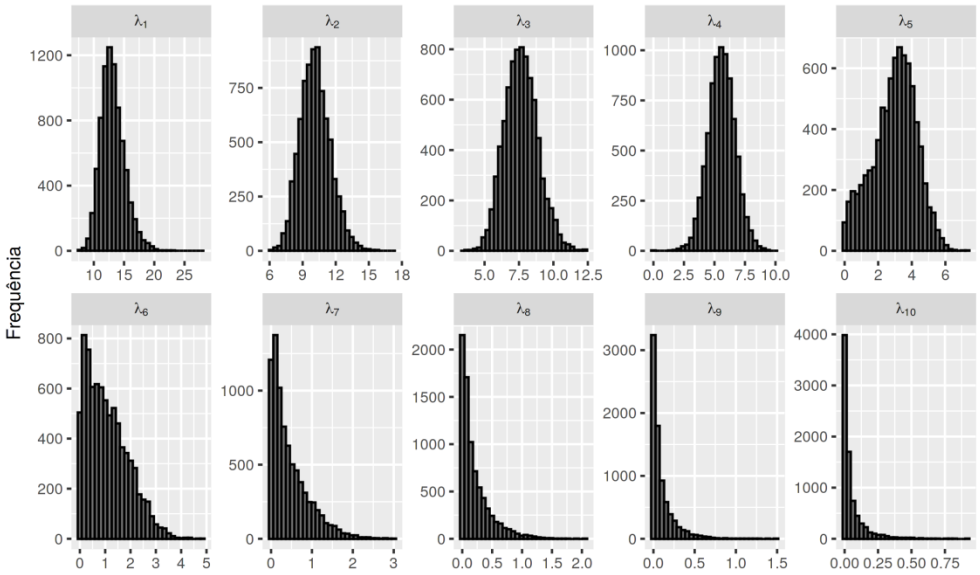


Figura 4: Histogramas das distribuições marginais a posteriori dos valores singulares para os dados simulados categorizados.

Ao invés de considerar todos os componentes principais que, neste exemplo, seriam 11 (grau de liberdade da matriz da GEI), optamos por não amostrar o 11° componente, como ocorre na análise padrão do modelo AMMI, em que se perde um grau de liberdade na matriz da interação pelas restrições marginais incorporadas para se ter identificabilidade na estimação. Viele e Srinivasan (2000), Crossa et al. (2011) e Perez-Elizalde et al. (2012) assumem essas mesmas restrições ao estimarem uma média geral no modelo. A nossa motivação é diferente, pois hierarquizamos blocos dentro de ambientes e não consideramos uma média geral em nosso modelo de forma que o mesmo não tem as restrições de que os efeitos devam ter soma igual a zero nas linhas e colunas. Entretanto, a estimação do vetor singular ambiental correspondente ao último componentes principal tem apenas um grau de liberdade para ser amostrado, ou seja, teríamos apenas uma mudança de sinal na cadeia e isso pode tornar o algoritmo mais lento e instável, o que procuramos evitar (VIELE; SRINIVASAN, 2000; CROSSA et al., 2011; de OLIVEIRA et al., 2015). Os últimos componentes geralmente não têm contribuição importante para explicar os dados, o que é uma das principais propriedades da análise de componentes principais.

Na Figura 5, regiões bivariadas (a 95% de credibilidade) foram incorporadas para os escores genotípicos e ambientais que descrevem a GEI no biplot AMMI-2 ((a) dados simulados, (b) dados categorizados e (c) dados transformados utilizando função logarítmica). As interpretações são realizadas de acordo com posições e sobreposições dessas regiões no plano determinado pelos eixos PC1 e PC2 (JUNIOR et al., 2018).

Regiões de credibilidade que englobaram a origem (0,0) sinalizam que os respectivos genótipos e ambientes não tem contribuição importante para explicar a GEI (definidos como estáveis) e não foram plotados para simplificar as interpretações. Nestes gráficos, observam-se sobreposições complexas (regiões se estendendo por mais de um quadrante, subgrupos sobrepondo parte de outros subgrupos, ou ainda, mais de uma configuração aceitável de formação de subgrupos). Como pode ser constatada a complexidade é mais acentuada para os biplots exibidos nas Figuras (a) e (b).

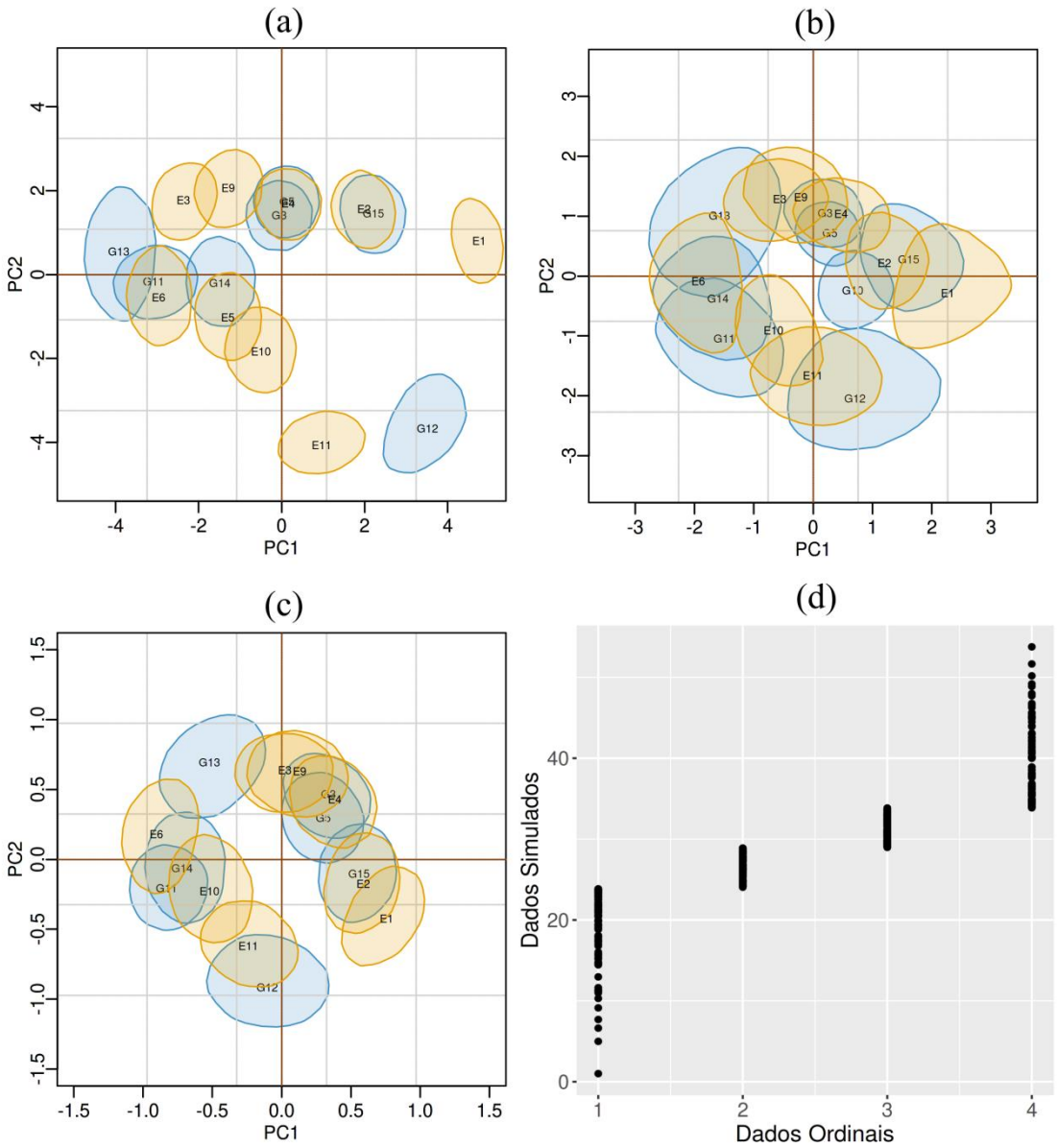


Figura 5: Representação gráfica de escores genotípicos e ambientais e regiões bivariadas a 95% de credibilidade para dados (a) simulados (b) categorizados (c)

utilizando transformação logarítmica nos dados ordinais. (d) Agrupamento em categorias a partir dos dados simulados.

Vale ressaltar que a configuração exibida na Figura 5c foi incluída para fins de comparação, já que a aproximação normal com ou sem a transformação de dados é comumente utilizada. Como observado na Figura 5d, os dados foram categorizados de forma que a quantidade de observações nas categorias foram semelhantes e também não houve uma variação mais acentuada dentro de cada categoria. Isso favorece uma aproximação da distribuição dos dados transformados a distribuição normal, e assim apresentando certas semelhanças no padrão biplot com dados reais. Argumentamos, entretanto, que a imprecisão se deve mais a distinção das amplitudes das regiões de credibilidade, bem como amplitudes relativamente menores quando comparado com o biplot para dados categóricos (Figura 5b).

Visualizando com mais cuidado é possivel perceber, por exemplo, que as posições para escores ambientais são melhores preservadas no biplot categórico e isso é mais nítido para os ambientes E1, E2 e E15 e o mesmo é percebido em relação aos genótipos G12, G3 e G5. Uma diferença clara é observada em relação ao genótipo G10, que na análise de dados categóricos saiu da origem e isso poderia ser indício de um aumento do erro tipo 2; contudo isso parece ter sido um fato mais isolado e a região de credibilidade se manteve bem próximo a origem e com amplitude relativamente menor. Fora essas constatações, as posições relativas entre genótipos e ambientes foram, de forma geral, preservadas tanto no caso dos dados transformados, como para os dados categóticos, tomando o biplot para dados contínuos (Figura 5a) como o padrão. A menor amplitude das regiões para o biplot da Figura 5c pode ser questionado, já que o erro na categorização deveria ser propagado (Figura 5d).

Resultados para conjuntos de dados não tão bem comportados, diferentemente do nosso exemplo, poderiam levar a resultados mais discrepantes. Uma vantagem para o método threshold aqui apresentado é que essa análise seria mais robusta para conjuntos de dados em que as distribuições de frequência são assimétricas ou com poucas categorias. Nesses casos muitos autores recomendam que as variáveis ordinais devam ser modeladas diretamente como ordinais, evitando quaisquer transformações (RHEMTULLA et al., 2012; Li, 2016; ROBITZSCH, 2020). Além disso, segundo Corrêa et al. (2016), a análise de threshold é robusta em relação a assimetria ou superdisperção dos dados, mais ainda, se a variável latente segue pressuposição de normalidade. Com o auxilio dos thresholds, os modelos podem ser utilizados em

diversas situações experimentais, mesmo se a variavel observavel não seja normal, obtendo estimativas precisas.

Quanto à categorização de uma variável contínua, o primeiro problema causado por esse procedimento seria a perda de informações, sendo essa perda menor quando se consideram vários grupos e muito mais grave com apenas duas categorias. Outro grande problema da dicotomização é a não utilização das informações dentro da categoria em que todas as quantidades acima ou abaixo do ponto de corte são tratadas como iguais; o que pode distorcer as relações e comprometer resultados e interpretações. Naggara et al. (2011) enfatiza essa ideia e ressaltam que o principal obstáculo imposto à categorização de variáveis contínuas pode não apenas perder a informação, mas sobretudo cometer erro, gerando resultados tendenciosos. Em nosso trabalho, verificamos que a análise biplot preservou, de forma geral, os padrões da análise com dados contínuos, ou seja, não verificamos a ocorrência sistemática de erros do tipo 1 ou tipo 2, como observado em Vargha et al. (1996), o que justificaria de certa forma nossa abordagem. Contudo, ressaltamos que esse exemplo foi apenas ilustrativo e não estamos incentivando pesquisados a modelarem dados contínuos de MET categoricamente, mas sim apresentando um método com justificativa teórica para tratar dados ordinais. Mais detalhes e implicações em análise de variáveis contínuas categorizadas podem ser encontrados, por exemplo, em Ragland, (1992), Taylor e Yu, (2002), Selvin (2004) e Royston et al. (2006).

Para exemplificação das interpretações práticas no biplots considera-se a análise a Figura 5(b) (biplot para dados categorizados), que de forma geral se estende aos demais biplots. Não obstante a complexidade das sobreposições das regiões de credibilidade, subgrupos de genótipos e ambientes semelhantes com relação aos efeitos da GEI (aqui referidos por subgrupos homogêneos) podem ser sugeridos. Esses subgrupos homogêneos baseiam-se em propriedades do produto interno, ou seja, na observação de ângulos formados entre os vetores determinados pelos escores (sobreposições entre as elipses e posições nos quadrantes). Destacam-se assim os subgrupos homogêneos de ambientes: {E1, E2}, {E3, E4, E9}, {E6}, {E10, E11} e subgrupos de genótipos, {G3, G15}, {G10, G12}, {G11, G14} e {G13}. Além disso, genótipos e ambientes não representados no biplot compõem subgrupos homogêneos separáveis de genótipos e ambientes estáveis.

A possibilidade de utilizar um método paramétrico flexível para fazer inferência no biplot como apresentado na Figura 5 foi, justamente, o que motivou o resgate dos

trabalhos pioneiros de Viele e Srnivasan (2000) e de Liu (2001). Argumentações da necessidade de se considerar a incorporação de incerteza aos escores que descrevem a interação, e as soluções do método frequentista para esse problema, bem como as críticas a esses métodos, podem ser consultados em Denis Gower (1994), Yang et al. (2009), Yan (2014) e de Oliveira et al. (2015).

Existem abordagens conduzidas com modelos lineares generalizados para o AMMI (VAN EEUWIJK, 1995; GABRIEL, 1998; ACORSI et al., 2017). Mas, para essas propostas ainda não foram apresentadas regiões de confiança tal qual apresentadas para modelos frequentistas usuais e, como já enfatizado, esses métodos tem sido controversos na literatura. Modelos lineares mistos são flexíveis e eficientes, como já ressaltado, e abordagens considerando limiares e utilizando uma variável latente foram utilizada por Foulley e Gianola (1996) e Montesinos-López et al. (2016), motivando a nossa abordagem. Trabalhos recentes utilizando essa modelagem em dados multiambientes foram desenvolvidos, apresentando como finalidade de estudo, obter modelos para representar interação genotipo x ambiente além da interação por caracteristícas, na predição genômica utilizando abordagem bayesiana (MONTESINOS-LÓPEZ et al., 2015). Entretando para o estudo de padrões da GEI não foi encontrado até o momento nenhum trabalho relacionado com resposta ordinal.

Além disso, considerar um modelo linear, ainda com efeito aleatório para GEI não seria o mais adequado, e, nesse ponto, concordamos com Cornelius, Crossa e Seyedsadr (1996) de que essa suposição implica que a interação de qualquer célula não informa sobre a resposta observada em qualquer outra (não contém informações transferíveis). Exceto a sua contribuição para a estimação da variância desconhecida como argumentam esses autores. Então nossa ideia foi utilizar o modelo de threshold e propor um procedimento que explore tanto as propriedades dos modelos de reconhecimento de padrões, quanto as flexibilidades que a modelagem bayesiana oferece, tal como a possibilidade de incorporar incerteza diretamente aos escores no biplot.

Na figura 6 são representadas as médias a posteriori e regiões HPD das normas dos vetores genotípicos (a) e ambientais (b), determinados pelos escores em relação aos dois primeiros componentes principais (PC1 e PC2) para dados simulados categorizados. Essa é uma medida proposta para avaliar a contribuição dos respectivos genótipos e ambientes para o efeito da GEI. Pontos mais afastados da origem são próprios de genótipos ou ambientes que possuem maior contribuição para GEI,

refletindo as informações apresentadas na Figura 5. Genótipos com médias mais próximas ao valor zero (estáveis) apresentaram menor variabilidade e o inverso também é observado para os genótipos que mais contribuem com a interação. Esse comportamento, de forma geral, também é observado para ambientes com exceção ao ambiente E8.

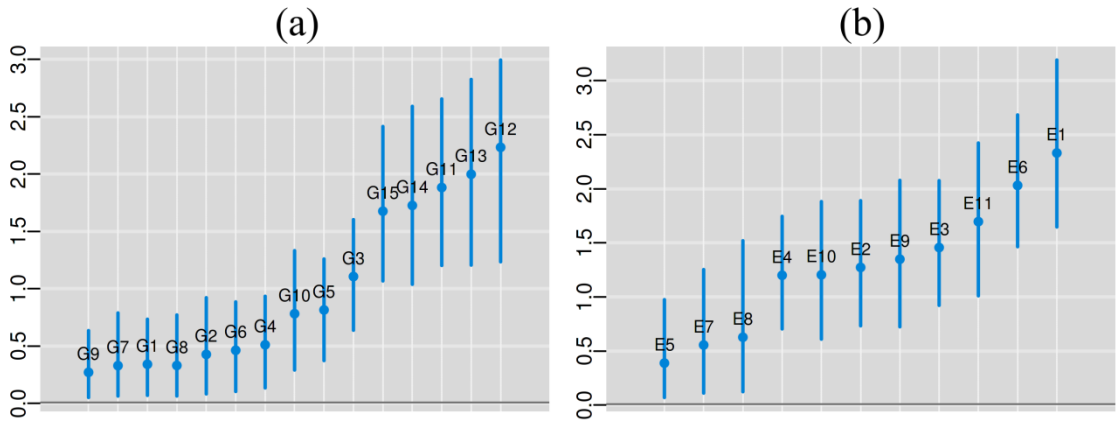


Figura 6: Médias a posteriori e regiões HPD das normas dos vetores genotípicos (a) e ambientais (b) determinados pelos escores em relação aos dois primeiros componentes principais (PC1 e PC2) para dados simulados categorizados.

Conectando as informações sobre os efeitos principais de genótipos (Figura 3) e análise da GEI (Figuras 5b e 6a), constata-se que um dos genótipos com melhores rendimentos marginais, G2 é o que menos contribui com a GEI e tem recomendação ampla para o conjunto de ambientes alvo. O genótipo G10 também não mostrou contribuição tão contundente com a interação (Figura 6a), mas a região de credibilidade para os escores genotípicos não inclui a origem do biplot (Figura 5b). Por sua vez, G15 apresenta contribuição expressiva para o efeito da GEI (Figura 6) e possui recomendação específica para os ambientes E2 e E1.

Esse mesmo comportamento, em geral, também foi observado em relação aos ambientes. Como representado na figura 6, ambientes mais instáveis tendem a ter maior variabilidade para as normas vetoriais (a exceção fica para E8). Dos ambientes o subgrupo {E5, E7, E8} são os que menos contribuem para a GEI, sendo interpretados como estáveis pela análise biplot (Figura 5b).

Dados reais categorizados

Densidades dos efeitos principais dos genótipos e suas respectivas HPD (a 95% de credibilidade) são apresentadas na Figura 7 ((a) dados continuos e (b) dados categorizados). Essas densidades estão em concordância com as distribuições a

posteriori propostas aos efeitos principais de genótipos, sendo aproximadamente simétricas e com curvas se assemelhando ao comportamento gaussiano. Destaca-se, que a mudança de ranqueamento após a transformação foi mínima, sendo mantido o padrão dos genótipos superiores formado pelo sugrupo (G10, G8, G11), tanto para os dados reais quanto para os dados categorizados, embora se observem amplitudes maiores para os intervalos do modelo AMMI ordinal. É preciso ressaltar que na análise AMMI da Figura 7a os dados originais são contínuos distribuídos normalmente e a análise com os dados categorizados tende a perder informações, sendo esse apenas um exemplo ilustrativo. Contudo, os resultados mostram que o padrão do ranqueamento, de forma geral, foi mantido.

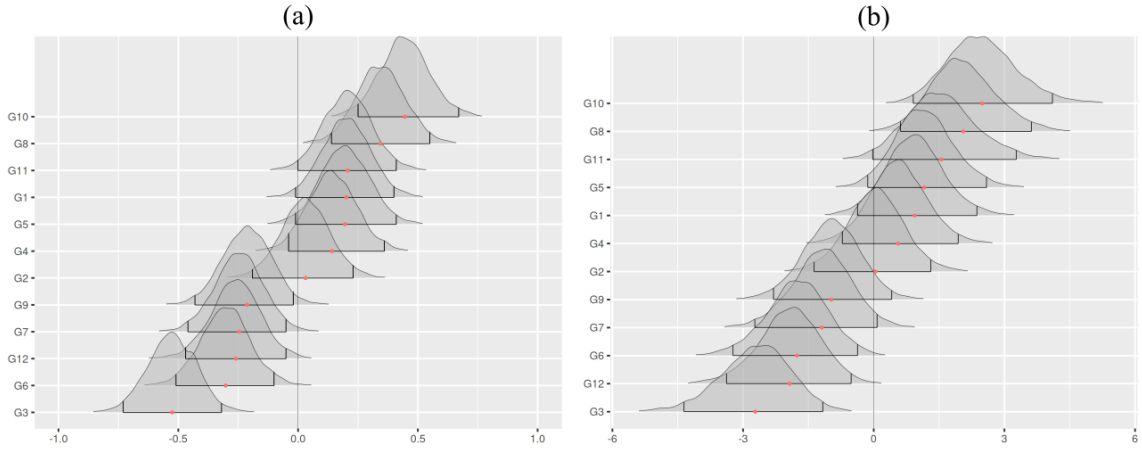


Figura 7: Densidades para efeitos principais de genótipos com HPD a 95% de credibilidade para os (a) dados reais (b) dados categorizados.

Na Tabela 2 são apresentados resumos pontuais e intervalares das distribuições *a posteriori* para valores singulares e componentes da variância dos dados categorizados e dos dados reais. Os dois primeiros componentes principais explicaram 94.95% da soma de quadrado total da GEI para dados ordinais, e 88.40 % para os dados reais e, portanto, capturaram informações substanciais em ambas as análises.

Tabela 2: Resumos pontuais (média e desvio padrão) e intervalos HPD a 95% de credibilidade para as distribuições a posteriori dos valores singulares e componentes de variância para dados reais e categorizados.

Dados Reais				
Parâmetros	Média	Desvio Padrão	HPD a 95% de credibilidade	
			Limite Inferior	Limite Superior
λ_1	2.419	0.122	2.183	2.655
λ_2	1.515	0.124	1.277	1.754
λ_3	0.841	0.13	0.574	1.087

λ_4	0.493	0.146	0.201	0.777
λ_5	0.256	0.131	<0.001	0.471
σ_g^2	0.098	0.06	0.0334	0.223
σ_e^2	0.044	0.006	0.0338	0.056
Dados Categorizados (Ordinais)				
Parâmetros	Média	Desvio Padrão	HPD a 95% de credibilidade	
			Limite Inferior	Limite Superior
λ_1	12.044	2.276	8.432	17.113
λ_2	6.779	1.448	4.054	9.689
λ_3	2.531	1.469	0.003	5.131
λ_4	1.059	0.901	<0.001	2.846
λ_5	0.486	0.517	<0.001	1.557
σ_g^2	3.251	2.441	0.833	8.377

Na figura 8 são apresentados histogramas das distribuições marginais *a posteriori* dos valores singulares para os dados reais ordinais. Percebe-se, como já ressaltado para dados simulados, que as distribuições se deslocam para cada vez mais para a direita e são mais simétricas para os dois primeiros valores singulares. Aqui não utilizamos nenhum critério para seleção do número de eixos a serem mantidos para explicar a GEI. Na análise bayesiana de modelos multiplicativos, isso tem sido feito a partir de critérios de informação como AIC, AICM e BIC, além do fator de Bayes como apresentado em Liu (2001), Jarquim et al. (2014) e da Silva et al., (2015). Mas, observando as distribuições dos valores singulares (Figura 8) e os resumos a posteriori (Tabela 2) acreditamos que o modelo com dois componentes (AMMI2) seria o mais adequado. Cabe ressaltar, entretanto, que embora modelos mais complexos possam ser selecionados, a análise biplot usual (AMMI ou GGE) tem sido aquela considerando os dois primeiros eixos e isso tem sido justificado pela interpretação biológica ou ainda por questão de parcimônia (YAN et al., 2000; YAN et al., 2007; GAUCH; PIEPHO; ANNICCHIARICO, 2008). Se o objetivo fosse predizer a resposta ou rendimento nominal de um genótipo em um determinado ambiente, a etapa de seleção de modelos seria essencial.

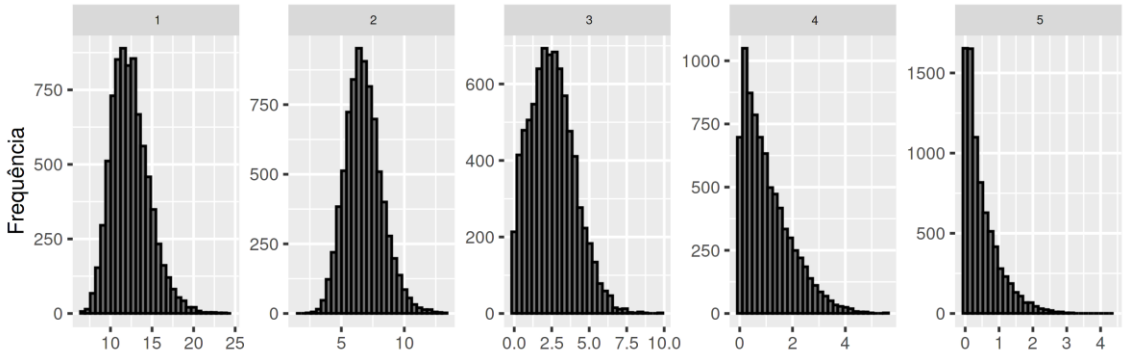


Figura 8: Histogramas das distribuições marginais a posteriori dos valores singulares para dados reais categorizados.

O biplot AMMI-2 (para (a) dados reais e (b) dados categorizados, (c) categóricos transformados utilizando função logaritmica) compostos pelos escores genotípicos e ambientais são apresentados na Figura 9. Regiões bivariadas, a 95% de credibilidade, foram incorporadas e permitem considerar as incertezas sobre esses escores. Observa-se que o padrão na representação biplot foi, relativamente, mantido após a categorização, como aconteceu com os dados simulados, apesar da perda de informação, ocasionando em um aumento da variabilidade (amplitude das regiões são maiores, Figura 9b). As regiões de credibilidade que incluiram a origem também não foram plotados, indicando que os respectivos genótipos (ou ambientes) não possuem efeito importante para a GEI.

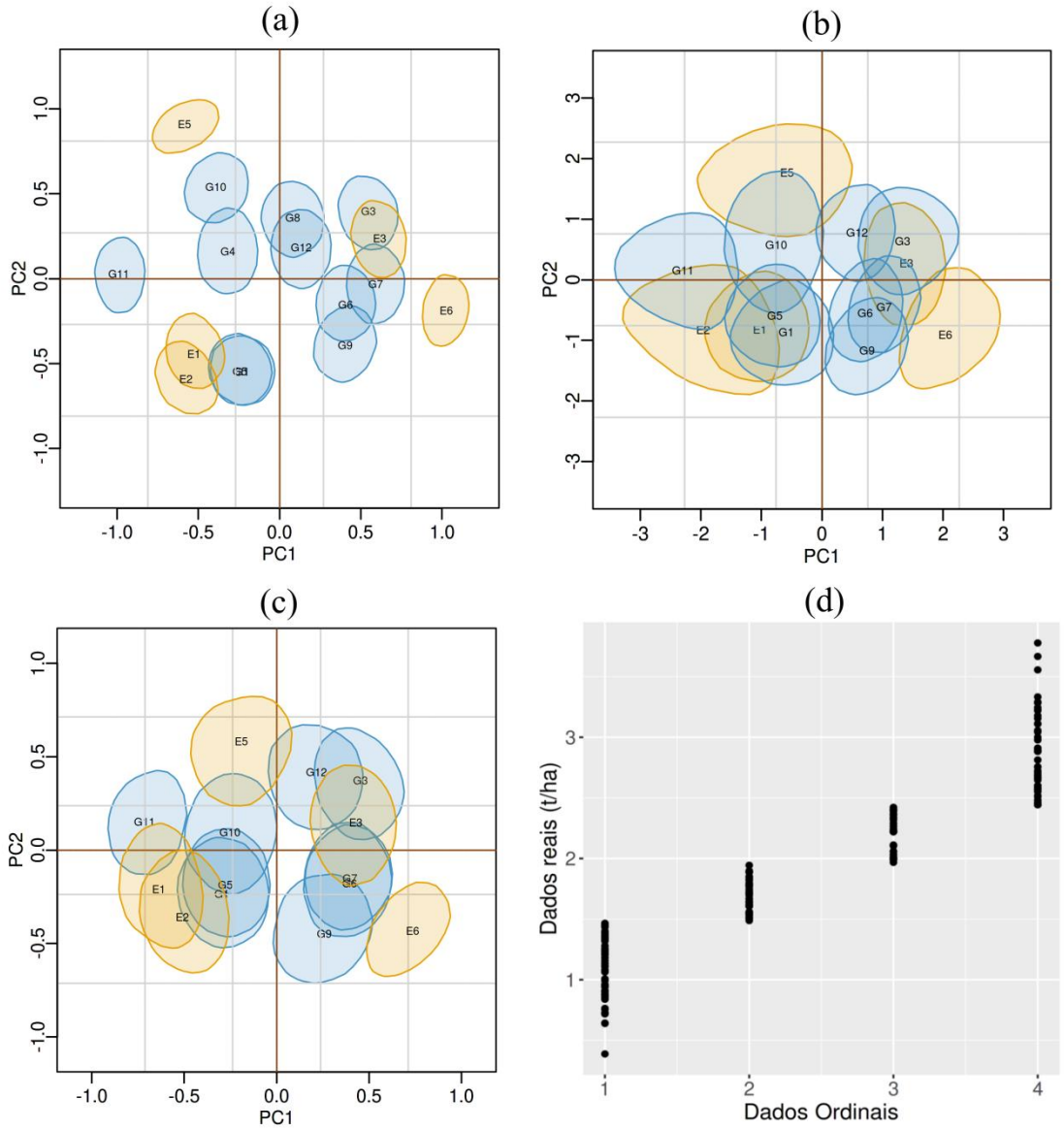


Figura 9: Representação gráfica de escores genotípicos e ambientais e regiões bivariadas a 95% de credibilidade para dados (a) reais (b) categorizados (c) utilizando transformação logarítmica nos dados ordinais. (d) Agrupamento em categorias a partir dos dados reais.

Como observado, os genótipos G4 e G8 foram para origem nos biplots para dados transformados e categórico, o que representaria a ocorrência de erro tipo 2, se considerássemos que análise para dados contínuos se constituísse o padrão verdadeiro. Maiores discrepâncias sobre interpretações relativas à adaptabilidade e estabilidade não são observadas em relação a dados transformados e categorizados, a não ser pelas amplitudes das regiões de credibilidade. Como já argumentado, o problema na categorização de uma variável contínua está na perda da informação e a Figura 9d exemplifica essa situação. Embora se observe variação dos dados, em cada categoria,

eles são considerados proporcionalmente com mesma quantidade. Mesmo assim, a análise, salvo algumas discrepâncias, foi robusta para identificar padrões.

Para explorar melhor as possibilidades da proposta deste trabalho, apresentamos na Figura 10 as distribuições dos valores preditos (B+G+GEI) dos genótipos (G10, G8 e G11) que são os mais produtivos em média (Figura 7), limitados ainda em categorias pelas estimativas médias à posteriori dos Thresholds (-0,64; 1,97 e 4,22). Destaca-se que o genótipo G10 obteve maior densidade na categoria 4 (categoria mais alta), nos ambientes E2, E3 e E5. O genótipo G8 foi o melhor classificado nos ambientes E3 e E5. Por fim, G11 teve densidades mais amplas nos ambientes E2 e E5, ou seja, seriam mais adaptados a estes ambientes.

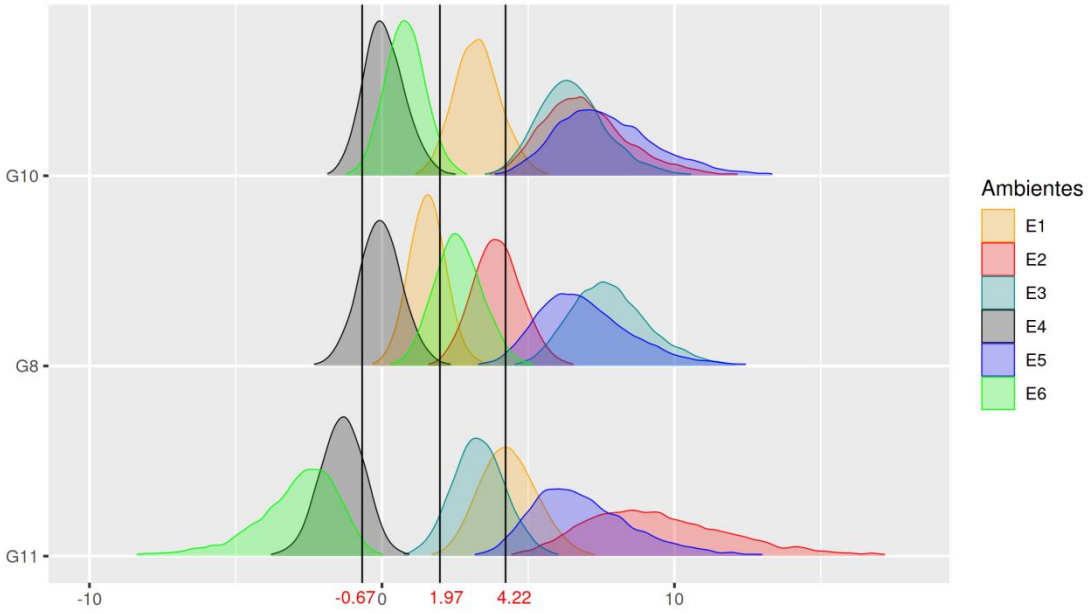


Figura 10: Rendimento nominal dos genótipos superiores {G10, G8 e G11} relacionados a todos os ambientes{E1 à E8} juntamente com as estimativas médias à posteriori dos thresholds (-0,67; 1,97 e 4,22), para dados reais categorizados.

CONCLUSÃO

Neste trabalho apresentamos um método que permite o reconhecimento de padrões aplicado em escalas ordinais, e exemplificamos sua funcionalidade com dados simulados e dados categorizados (a título de comparação). No cenário controlado, o método foi eficiente para identificar os genótipos que foram simulados como estáveis ou instáveis. Para dados reais, efeitos principais de genótipos e biplots também apresentaram padrões semelhantes. Esses resultados mostram que o método se mostrou relativamente menos poderoso, mas, mais rigoroso do que análises tradicionais supondo que dados ordinais são contínuos.

A implementação do modelo ordinal com funções de ligação acumuladas probit é mais complexa do que se poderia esperar para contagens ou proporções, e pretendemos em trabalhos futuros estender a análise AMMI em tais configurações de modelos não normais.

REFERÊNCIAS

ACORSI, et al.. Applying the generalized additive main effects and multiplicative interaction model to analysis of maize genotypes resistant to grey leaf spot. **The Journal of Agricultural Science**, v.155 , n.6 ,p. 939-953, 2017.

ALBERT, J. H.; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, v. 88,n. 422, p. 669-679, 1993.

BERNARDO, J. M.; SMITH, A. F. Bayesian theory. **John Wiley & Sons**, 2009.

CORNELIUS, P. L.; CROSSA, J.; SEYEDSADR, M. S. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: KANG, M. S.; GAUCH, H. G. (Org.). **Genotype-by-environment interaction**. Boca Raton: CRC, p. 199-234, 1996.

CORRÊA, Fábio Mathias et al. Bayesian algorithms for analysis of categorical ordinal data. **Revista Brasileira de Biometria**, v. 34, n. 4, p. 597-620, 2016.

COTES, J. M. et al. A Bayesian approach for assessing the stability of genotypes. **Crop Science**, v. 46, n. 6, p. 2654-2665, 2006.

COWLES, M. K. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. **Statistics and Computing**, v. 6, n. 2, p. 101-111, 1996.

CROSSA, J. Statistical analyses of multilocation trials. In: **Advances in agronomy**. Academic Press, 1990. p. 55-85.

CROSSA, J. et al. Bayesian estimation of the additive main effects and multiplicative interaction model. **Crop Science**, v. 51, n. 4, p. 1458-1469, 2011.

da SILVA, C. P. et al. A Bayesian Shrinkage approach for AMMI Models. **PLoS ONE**, v. 10, n. 7, p. e0131414, 2015.

da SILVA, C. P. et al. Heterogeneity of variances in the Bayesian AMMI model for multienvironment trial studies. **Crop Science**, v. 59, n. 6, p. 2455-2472, 2019.

de OLIVEIRA, L. A. et al. Credible intervals for scores in the AMMI with random effects for genotype. **Crop Science**, v. 55, n. 2, p. 465-476, 2015.

DEMPSTER, E. R.; LERNER, I. M. Heritability of threshold characters. **Genetics**, v. 35, n. 2, p. 212, 1950.

DENIS, J. B.; GOWER, J. C. Asymptotic covariances for parameters of biadditive models. **Utilitas Mathematica**, v. 46, p. 193-205, 1994.

EDWARDS, J. W.; JANNINK, J. L. Bayesian modeling of heterogeneous error and genotype environment interaction variances. **Crop Science**, v. 46, n.2, p. 820-833, 2006.

FALCONER, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. **Annals of human genetics**, v. 29, n. 1, p. 51-76, 1965.

FALCONER, D. S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. **Annals of human genetics**, v. 31, n. 1, p. 1-20, 1967.

FIENBERG S.E.; LIEVESLEY D.; ROLPH J. Statistics for Social Science and Public Policy, Springer, New York, v.1, 1999.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant-breeding programme. **Australian Journal of Agricultural Research**, v. 14, p., n. 6, 742-754, 1963.

FOULLEY, J. L.; GIANOLA, D. Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model. **Genetics Selection Evolution**, v. 28, n. 3, p. 249-273, 1996.

GABRIEL, K. R.. Generalised bilinear regression. **Biometrika**,v.85, n.3 , p.689-700, 1998.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal components analysis. **Biometrika**, v. 58, n. 3, p. 453-467, 1971.

GAUCH, H. G.; PIEPHO, H. P.; ANNICCHIARICO, P. Statistical analysis of yield trials by AMMI and GGE: further considerations. **Crop Science**, v.48, n3, p.866-889, 2008.

GIANOLA, Daniel. Theory and analysis of threshold characters. **Journal of animal Science**, v. 54, n. 5, p. 1079-1096, 1982.

GIANOLA, D.; FOULLEY, J. L. Sire evaluation for ordered categorical data with a threshold model. **Genetique Selection Evolution**, v. 15, n. 2, p. 201-224, 1983.

HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. **Operations Research**, Landing, v. 31, n. 6, p. 1109-1144, 1983.

IASRI. Analysis of Data from Designed Experiments, 2021. Disponível em: < https://drs.icar.gov.in/Analysis%20of%20data/combined_anlysis_alpha.html/> . Acesso em: 20 de maio de 2021.

JARQUÍN, D. et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. **Theoretical and applied genetics**, v. 127, n. 3, p. 595-607, 2014.

JARQUÍN, D. et al. A hierarchical Bayesian estimation model for multi-environment plant breeding trials in successive years. **Crop Science**, v. 56, n.5, p. 2260-2276, 2016.

JOSSE, J. et al. Another look at Bayesian analysis of AMMI models for genotype-environment data. **Journal of Agricultural, Biological and Environmental Statistics,** v. 19, n. 2, p. 240-257, 2014.

JÚNIOR, L. A. Y. B. et al. AMMI Bayesian models to study stability and adaptability in maize. **Agronomy Journal**, v. 110, n. 5, p. 1765-1776, 2018.

KANG, M. S.; GORMAN, D. P. Genotype× environment interaction in maize. **Agronomy Journal**, v. 81, n. 4, p. 662-664, 1989.

KEMPTON, R. A. The use of biplots in interpreting variety by environment interactions. **Journal of Agricultural Science**, v. 103, n. 1, p. 123-135, 1984.

LI, C.-H. The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. **Psychological methods**, v. 21, n. 3, p. 369, 2016.

LIN, C.-S.; BINNS, M. R. A superiority measure of cultivar performance for cultivar× location data. **Canadian journal of plant science**, v. 68, n. 1, p. 193-198, 1988.

LIU, G. **Bayesian computations for general linear-bilinear models**. 2001. 169 p. Thesis (Doctor of Philosophy) - University of Kentucky, Lexington, 2001.

MONTESINOS-LÓPEZ, O. A. et al. Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. **G3: Genes, Genomes, Genetics**, v. 5, n. 2, p. 291-300, 2015.

MONTESINOS-LÓPEZ, O. A. et al. A genomic Bayesian multi-trait and multi-environment model. **G3: Genes, Genomes, Genetics**, v. 6, n. 9, p. 2725-2744, 2016.

NAGGARA, O. et al. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. **American Journal of Neuroradiology**, v. 32, n. 3, p. 437-440, 2011.

NANDRAM, B.; CHEN, M.-H. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. **Journal of Statistical Computation and Simulation**, v. 54, n. 1-3, p. 129-144, 1996.

NUVUNGA, J. J. et al. Bayesian factor analytic model: An approach in multiple environment trials. **PloS one**, v. 14, n. 8, p. 1-26, 2019.

ORELLANA, M. A.; EDWARDS, J.; CARRIQUIRY, A. L. Heterogeneous variances in multi-environment yield trials for corn hybrids. **Crop Science**, v. 54, n. 3, p. 1048, 2014.

PEREZ-ELIZALDE, S.; JARQUIN, D.; CROSSA, J. A general Bayesian estimation method of linear–bilinear models applied to plant breeding trials with genotype× environment interaction. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 17, n. 1, p. 15-37, 2012.

PIEPHO, H. P. Analyzing genotype-environment data by mixed models with multiplicative terms. **Biometrics**, v. 53, n. 2, p. 761-766, 1997.

PIEPHO, H. P. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. **Theoretical Applied Genetics**, v. 97, p. 195-201, 1998.

PIEPHO, H. P.; MÖHRING, J. Selection in Cultivar Trials—Is It Ignorable?. **Crop Science**, v. 46, n. 1, p. 192-201, 2006.

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian Statistics**. Oxford: Oxford University, 1992. p. 763-773.

RAGLAND, David R. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. **Epidemiology**, p. 434-440, 1992.

RHEMTULLA, M.; BROSSEAU-LIARD, P. É.; SAVALEI, V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. **Psychological methods**, v. 17, n. 3, p. 354, 2012.

ROBITZSCH, A. Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In: **Frontiers in Education**. Frontiers, v.5 p. 177, 2020.

ROMAGOSA, I.; FOX, P. N. Genotype× environment interaction and adaptation. In:Hayward MD, Bosemark NO, Romagosa I (eds) **Plant breeding: principles and prospects**, Chapman and Hall, p.373-390, 1993.

ROYSTON, P.; ALTMAN, D. G.; SAUERBREI, W. Dichotomizing continuous predictors in multiple regression: a bad idea. **Statistics in medicine**, v. 25, n. 1, p. 127-141, 2006.

SELVIN S. Statistical Power and Sample Size Calculations: Statistical Analysis of Epidemiological Data. **Oxford University Press**, New York, p.75–92, 2004.

SMITH, A.; CULLIS, B.; THOMPSON, R. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend, **Biometrics**, v. 57, p. 1138-1147, 2001

SORENSEN, D. A. et al. Bayesian inference in threshold models using Gibbs sampling. **Genetics Selection Evolution**, v. 27, n. 3, p. 229-249, 1995.

SORENSEN, D.; GIANOLA, D. Likelihood, Bayesian and MCMC methods in quantitative genetics, Springer, New York, 2002.

TAYLOR, J. M. G.; YU, M., Menggang. Bias and efficiency loss due to categorizing an explanatory variable. **Journal of Multivariate Analysis**, v. 83, n. 1, p. 248-263, 2002.

VAN EEUWIJK, Fred A. Multiplicative interaction in generalized linear models. **Biometrics**, p. 1017-1032, 1995.

VARGHA, A. et al. Dichotomization, partial correlation, and conditional independence. **Journal of Educational and Behavioral statistics**, v. 21, n. 3, p. 264-282, 1996.

VIELE, K.; SRINIVASAN, C. Parsimonious estimation of multiplicative interaction in analysis of variance using Kullback-Leibler information. **Journal of Statistical Planning and Inference**, v. 84, n. 1-2, p. 201-219, 2000.

YAN, W et al. Cultivar evaluation and mega-environment investigation based on the GGE biplot. **Crop Science**, v. 40, n. 3, p. 597-605, 2000.

YAN, W. et al. GGE biplot vs. AMMI analysis of genotype-by-environment data. **Crop science**, v. 47, n. 2, p. 643-653, 2007.

YAN, W. Crop variety trials: Data management and analysis. John Wiley & Sons, New York, 361 p., 2014.

YANG, R. C. et al. J. Biplot analysis of genotype environment interaction: proceed with caution. **Crop Science**. v. 49, n. 5, p. 1564-1576, 2009.

ZOBEL, R. W.; WRIGHT, M. J.; GAUCH JR, H. G. Statistical analysis of a yield trial. **Agronomy journal**, v. 80, n. 3, p. 388-393, 1988.

APÊNDICE

Apêndice S1: Distribuição a posteriori conjunta para $\mathbf{l}$ e $\boldsymbol{\tau}$ condicionada aos demais parâmetros:

$$p(\mathbf{l},\boldsymbol{\tau}\,|\,\mathbf{y},\boldsymbol{\eta}) = p(\mathbf{l}\,|\,\boldsymbol{\tau},\mathbf{y},\boldsymbol{\eta})\,p(\boldsymbol{\tau}\,|\,\mathbf{y},\boldsymbol{\eta}) \tag{2}$$

a distribuição condicional para $\boldsymbol{\tau}$ (segunda expressão) é proporcional ao produto da probabilidade de cada observação pertencer a uma categoria e da informação a priori sobre $\boldsymbol{\tau}$ dada por:

$$p(\boldsymbol{\tau}\,|\,\mathbf{y},\boldsymbol{\eta}) \propto p(\mathbf{y}\,|\,\boldsymbol{\eta},\boldsymbol{\tau})\,p(\boldsymbol{\tau}) \tag{3}$$

Se a variável categórica ordinal $\mathrm{y}_i = j \;\; (j = 1,2,\ldots,c)$, a sua probabilidade de pertencer a uma categoria especifica é $p(\mathrm{y}_i = j) = \pi_{i,j}$ (probit), sendo:

$$\begin{aligned}\pi_{i,j} &= \int_{\tau_{j-1}}^{\tau_j} \phi(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i})\, dl_i \\ &= p(\tau_{j-1} < l_i < \tau_j) \\ &= \Phi(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}) - \Phi(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i})\end{aligned}$$

em que $\phi(\cdot)$ representa a distribuição normal padrão e $\Phi(\cdot)$ a distribuição normal padrão acumulada. Se as probabilidades $\pi_{i,j}$ forem independentes entre si, a distribuição amostral para os dados observados é dada por uma distribuição multinomial, e possuindo a seguinte forma (FIENBERG; LIEVESLEY; ROLPH ,1999):

$$\begin{aligned}p(\mathbf{y}\,|\,\boldsymbol{\eta},\boldsymbol{\tau}) &= \prod_{i=1}^{n}\sum_{j=1}^{c}\pi_{i,j} \\ &= \prod_{i=1}^{n}\sum_{j=1}^{c}\Phi(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}) - \Phi(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i})\end{aligned}$$

Retornando a expressão (3), assumindo uma priori não informativa para $\boldsymbol{\tau}$, a distribuição condicional torna-se:

$$\begin{aligned}p(\boldsymbol{\tau}\,|\,\mathbf{y},\boldsymbol{\eta}) &\propto p(\mathbf{y}\,|\,\boldsymbol{\eta},\boldsymbol{\tau})\,p(\boldsymbol{\tau}) \\ p(\boldsymbol{\tau}\,|\,\mathbf{y},\boldsymbol{\eta}) &\propto \prod_{i=1}^{n}\sum_{j=1}^{c}\left[\Phi(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}) - \Phi(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i})\right]\end{aligned} \tag{4}$$

Para a variável latente $\mathbf{l}$ (primeira expressão em (2)), a distribuição condicional *a posteriori* dada o restante dos parâmetros segue:

$$p(\mathbf{l}\,|\,\boldsymbol{\tau},\mathbf{y},\boldsymbol{\eta}) \propto p(\mathbf{y}\,|\,\mathbf{l},\boldsymbol{\tau})\, p(\mathbf{l}\,|\,\boldsymbol{\tau},\boldsymbol{\eta})\, p(\boldsymbol{\tau})\, p(\boldsymbol{\eta})$$

Se $\mathbf{l}$ segue uma distribuição gaussiana, e seus elementos são condicionalmente independentes, a distribuição para cada l_i, dado os dados observados $\mathbf{y}$ e ao restante dos parâmetros $\boldsymbol{\eta} = \left(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_\delta^2\right)$, segue uma NT (normal truncada) com média igual a $\boldsymbol{\mu}_{\mathrm{y}_i}$ e variância $\sigma_e^2 = 1$, limitada pelos thresholds τ_{j-1} e τ_j:

$$p(l_i\,|\cdots) \propto p(\mathrm{y}_i = \mathrm{j}\,|\,l_i, \boldsymbol{\tau})\, p(l_i\,|\,\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\tau})$$

$$p(l_i\,|\cdots) \propto \left[\sum_{j=1}^{c} I(\tau_{j-1} < \mathrm{l}_i < \tau_j) I(\mathrm{y}_i = j)\,|\,\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\tau}\right]$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{c} \frac{\phi(\boldsymbol{\mu}_{\mathrm{y}_i}, 1)}{\Phi\left(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}\right) - \Phi\left(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i}\right)} \tag{5}$$

Como $p(\mathrm{y}_i = \mathrm{j}\,|\,l_i, \boldsymbol{\tau})$ é conhecida, essa expressão é absorvida pela constante de normalização, restando apenas a condicional da variável latente (GIANOLA,1982; SORENSEN; GIANOLA, 2002).

Assim, voltando a expressão (2), a distribuição *a posteriori* conjunta para $\mathbf{l}$ e $\boldsymbol{\tau}$ segue:

$$p(\mathbf{l}, \boldsymbol{\tau}\,|\,\mathbf{y}, \boldsymbol{\eta}) = p(\mathbf{l}\,|\,\boldsymbol{\tau}, \mathbf{y}, \boldsymbol{\eta})\, p(\boldsymbol{\tau}\,|\,\mathbf{y}, \boldsymbol{\eta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{c} \frac{\phi(\boldsymbol{\mu}_{\mathrm{y}_i}, 1)}{\Phi\left(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}\right) - \Phi\left(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i}\right)} \times \prod_{i=1}^{n} \sum_{j=1}^{c} \left[\Phi\left(\tau_j - \boldsymbol{\mu}_{\mathrm{y}_i}\right) - \Phi\left(\tau_{j-1} - \boldsymbol{\mu}_{\mathrm{y}_i}\right)\right]$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{c} \phi(\boldsymbol{\mu}_{\mathrm{y}_i}, 1) \tag{6}$$

Figura S1: Gráfico de traços e densidades dos thresholds τ_2 e τ_3 para dados categorizados.

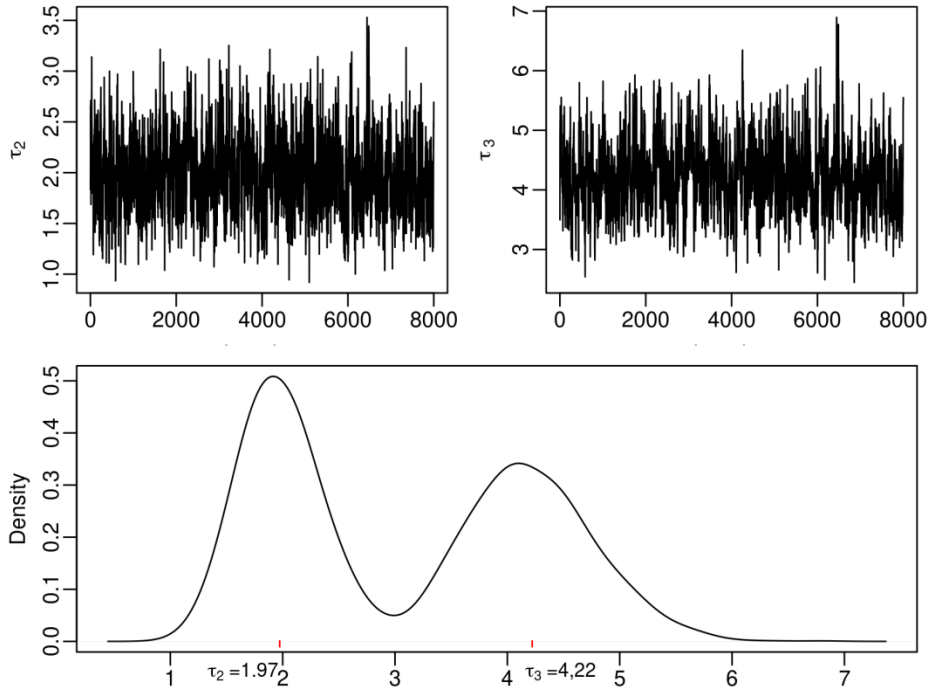


Tabela S1: Resumos pontuais (média e desvio padrão) e intervalos HPD a 95% de credibilidade para as distribuições a posteriori dos valores singulares e componentes de variância para dados simulados categorizados.

Parâmetros	Média	Desvio Padrão	HPD a 95% de credibilidade	
			Limite Inferior	Limite Superior
λ_1	12.649	1.758	9.568	16.311
λ_2	9.861	1.209	7.393	12.126
λ_3	7.439	1.029	5.491	9.571
λ_4	5.571	0.945	3.735	7.428
λ_5	3.396	1.02	1.169	5.249
λ_6	1.321	0.86	<0.001	2.862
λ_7	0.534	0.481	<0.001	1.499
λ_8	0.24	0.269	<0.001	0.811
λ_9	0.112	0.155	<0.001	0.424
λ_{10}	0.057	0.095	<0.001	0.243
σ_g^2	0.917	0.48	0.325	1.944

CONCLUSÃO GERAL

- No primeiro artigo fica claro que o modelo AMMI bayesiano é muito flexível e poderia acomodar desbalanceamentos, ausências de ortogonalidade e heterogeneidade de variÂncias, mas depende de respostas contínuas em que se possa supor aproximação normal.
- No segundo artigo construimos uma variável latente contínua usando funções de ligação acumuladas probit para respostas ordinais e verificamos que o AMMI bayesiano permite interpretação menos poderosa, mas mais rigorosa e consistente com a análise de dados contínuos que em geral não existirão nas aplicações.