河南大學

Undergraduate Thesis

Significance and Stability Analysis of Gene-Environment Interaction

Authored by: Meng'en Qin,
Statistics,
School of Mathematics and Statistics,
Henan University.

Instructed by: Xiaohui Yang,
Professor of Statistics,
School of Mathematics and Statistics,
Henan University.

April 26, 2025

Abstract

Genotype-by-Environment (GxE) interactions influence the performance of genotypes across diverse environments, reducing the predictability of phenotypes in target environments. In-depth analysis of GxE interactions facilitates the identification of how genetic advantages or defects are expressed or suppressed under specific environmental conditions, thereby enabling genetic selection and enhancing breeding practices. This paper introduces two key models for GxE interaction research. Specifically, the study includes significance analysis based on the mixed effect model to determine whether genes or GxE interactions significantly affect phenotypic traits; stability analysis, which further investigates the interactive relationships between genes and environments, as well as the relative superiority or inferiority of genotypes across environments. Additionally, this paper presents RGxEStat, a lightweight interactive tool, which is developed by the author and integrates the construction, solution, and visualization of the aforementioned models. Designed to eliminate the need for breeders and agronomists to learn complex SAS or R programming, RGxEStat provides a user-friendly interface for streamlined breeding data analysis, significantly accelerating research cycles.

Keywords: G×E interaction; mixed effect model; stability analysis model

Content

Chapter 1 Introduction

1.1 Background and Motivations

Gene-environment interactions refer to the modification of genetic factors by environmental factors, which can affect the performance of genotype in different environments and reduce the predictability of phenotypes in the target environment[1]. Understanding the complex behavior characteristics of organisms requires not only genetic information, but also the environment in which they live. The in-depth study of the interaction between genes and environments can make people have a good grasp of the relationship between individual living environment and phenotypic traits, and between genes and phenotypic traits. This analysis will help researchers identify how genetic advantages or defects are realized or suppressed in a specific environment, and then they can conduct gene selection, animal and plant breeding, pharmacogenomics, conservation biology research and so on.

When a phenotype is of great economic significance, researchers usually analyze the stability between environmental factors and the genes corresponding to the phenotype. These phenotypes usually include reproductive fitness, length, height, weight, yield and disease resistance of organisms. For botanists, selecting genotypically superior crop seeds in the target environment is an important goal of plant breeding plan. The target environment is the production environment used by growers and farmers. In order to identify excellent genotypes in multiple environments, plant breeding needs to design and carry out cross-location and cross-year completely random regional trials, especially in the final stage of variety development. Gene-environment interaction is considered to exist when the performance difference of a genotype in different environments leads to changes in size or rank. Genotypes show great performance differences in different environments due to the influence of environment on gene expression. Although crop varieties with high yield and stable performance are difficult to identify, they are of great value. Therefore, when the interaction between genes and environment is significant, it is very valuable to use statistical models to analyze the stability of genes and environment to select varieties with high yield and stable performance in agriculture.

Biostatisticians and breeders often use the analysis of variance (ANOVA) to determine the existence and magnitude of genes and gene-environment interactions. When analyzing the interactions between genes and the environment of a group of superior varieties, genotypes are

usually considered as fixed effects and the environment is random. However, in order to estimate the breeding value or yield by using the best linear unbiased prediction (BLUP), genotypes are considered as random effects and the environment is fixed[2,33]. Some statisticians also tend to consider genotypes as random effects, as long as the goal is to choose varieties with the best traits[3]. ANOVA measures the variance components caused by different fixed and random factors (such as gene, location, year and replicate) and their interactions. However, ANOVA has limitations in exploring the response of genotypes to the environment, including assuming the homogeneity of variance among different environments[21]. If the gene action and the interaction between gene and environment are significant, then in turn, additional stability statistics can be calculated for gene stability analysis.

So far, researchers in related fields have put forward several statistical methods for stability analysis. The most widely used method is the single-factor stability model based on regression and variance estimation. Regression statistics as the stability measurement was proposed by Yates and Cochran[4] and later improved by Eberhart and Russell[5]. According to the regression model, stability is expressed as the sum of squares of the trait means, the slope of the regression line and the regression deviation. The high trait mean of the genotype is a prerequisite for stability. The slope of the regression line represents the response of the genotype to the environmental index, which comes from the average performance of all genotypes in each environment. If there is no significant statistical difference between the slope and 1, then the genotype is applicable for all environments. A slope greater than 1 describes the genotype with higher sensitivity to environmental changes (lower than average stability) and higher specific adaptability to high-yield environments. A slope less than 1 provides greater resistance to environmental changes (higher than average stability), thus increasing the specificity of adaptability to low-yield environments. Perkins and Jinks put forward another similar regression coefficient, and they used the interaction effect between genes and environment to regress environmental effects[6].

There are many variance statistics to measure gene stability. Wricke put forward the stability ecovalence (W_i^2) in 1962 [7]. The ecovalence of genotype indicates its contribution to the sum of squares of gene-environment interactions in all environments. Because W_i^2 is expressed as the sum of squares, it is impossible to test the significance. Shukla proposed a stability variance (σ_i^2), which is an unbiased estimate of the variance of gene-environment interactions plus the sum of error terms related to genotype [8]. σ_i^2 is a linear combination of Wricke's stability ecovalence (W_i^2). Shukla's stability parameter measures the contribution of a

genotype to the interaction between genes and environment and error term. Therefore, genotypes with low σ_i^2 are regarded to be stable. According to Kang et al., the rank correlation coefficient between W_i^2 and σ_i^2 is 1, which means they both are equivalent in ranking the stability of genotypes. Kang developed the yield stability (YS_i) in 1993[9]. This index is a nonparametric statistic, which needs to be calculated by using the mean of traits and Shukla's stability variance. YS_i gives equal weight to the mean of traits and σ_i^2, and the gene is considered stable if its YS_i is greater than the mean YS_i[10,11]. YS_i is computed based on the procedure outlined by Mekbib[10] in this thesis. The stability concept P_i proposed by Lin and Binns in 1988 is derived from the average sums of squares of years nested in locations[12]. High stability is represented by low P_i, i.e., the low temporal variation of gene trait values. Francis and Kannenberg proposed the coefficient of variation (CV_i) in 1978[13], a genotype with low coefficient of variation are relatively stable.

For multi-factor stability models, the additive main effects multiplicative interaction model (AMMI) and the genotypic main effect plus genotypic-by-environment interaction model (GGE) are gaining popularity in analyzing multi-environment experiment data because of their graphical display[14]. Proponents of AMMI and GGE methods disagree on the best way to analyze multi-environment breeding data[15,16], even though both methods present similar results. Yan et al. proposed GGE biplots based on the singular value decomposition of environment-centered or within-environment standardized bidirectional (gene and environment) data matrix[18]. GGE biplot is constructed by the first two principal components (PC1 and PC2), which accounts for the maximum variability in data. The GGE biplot graphically shows two-way data matrix and visualizes the interrelationships and interactions between environments and genes [17]. In GGE biplots, genotype effect and gene-environment interaction effect are the two sources of variation associated with gene evaluation and original environment identification. AMMI model combines ANOVA with principal component analysis (PCA), where the former is used to characterize the main effects of genotype and environment, and the latter is to characterize the interaction (interaction principle component, IPCs)[19,20]. The AMMI biplot separates different genotypes according to their IPC scores. Therefore, it is easy to qualitatively evaluate the differences in gene stability and environment adaptability. The closer the IPC is to zero, the more stable the gene is in the test environment. Although the researchers have put forward various statistical methods to evaluate the stability, which reflects different aspects of gene-environment interaction effect, none of them can comprehensively explain the performance of genotypes in different environments. Stability statistics (variation) are best used

in conjunction with trait performance (e.g., average yield)

Regarding the program implementation of the above methods, Piepho released a SAS mixed model program in 1999 to calculate the single factor stability statistics[27]. Hussein et al. also provided a SAS program (SASGxESTAB) in 2000 to calculate the single factor stability statistics and their correlation[26]. However, due to the age and version updates, the above SAS program is not available on the listed servers. Recently, Dia et al. proposed brand new SAS codes to analyze the interaction between genes and environment[2]. However, after the author's practice, it is found that some of its functions can't be realized due to code problems, and because of the lack of interactive interface, it also requires users to have high SAS programming skills, which is very unfavorable to breeders and farmers. Driven by this, this paper presents an easy-to-operate interactive software developed by the author in R language, which integrates significance and stability analysis models to help breeders and agronomists better analyze the gene-environment interactions for variety selection.

1.2 Thesis Outline

This thesis first introduces how to perform significance analysis based on the mixed effect model, including determining the significance of genes, gene-environment interactions, etc. Subsequently, when these effects are significant, single-gene and multi-gene stability models are established to analyze the stability and expression performance of genes in different environments. Finally, this paper presents RGxEStat developed by the author in R language, a portable interactive software that integrates the above significance analysis and stability analysis models. The thesis is organized as follows:

Chapter 1 mainly focuses on the research background and related works of gene-environment interaction analysis;

Chapter 2 introduces some basic theories and knowledge needed for later significance and stability analysis, including fixed effect, random effect, principal component analysis and singular value decomposition;

In chapter 3, the mixed effect model is constructed to analyze the significance of gene effect, gene-environment interaction effect and so on, and meanwhile determine the variation degree caused by those effects;

Chapter 4 develops single-gene and multi-gene stability models for breeding gene selection, and analyzes the stability ability of different genes in various environments and the positive and negative effects of gene-environment interaction on phenotypic response values;

Chapter 5 presents RGxEStat developed by the author, and uses two public breeding

datasets to show its operation process and analysis results;

Chapter 6 is the summary and prospect of this thesis.

Chapter 2 Fundamentals

This section mainly introduces some basic theories and methods used in the models in the third and fourth chapters, including fixed effects and random effects, principal component analysis and singular value decomposition.

2.1 Fixed Effects and Random Effects

In the mixed effect model, fixed effect and random effect are two key elements that deserve careful discrimination, which are used to describe different parameter types in the regression model. First, consider the following general linear regression equation:

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_i, \tag{2.1}$$

where i represents different individuals, t denotes different periods, α and β are the regular intercept and slope. Consider a specific case where y_{it} denotes the output value of the i-th firm in the t-th year and x_{it} is the scale of the i-th firm in the t-th year, and ε_i is the random error term. Equation (2.1) shows that the output value of each firm is determined by a constant (α), which will be adjusted according to the scale of the firm (βx_{it}). It seems reasonable to use the general linear regression equation to model this example with a fixed effect of firm scale. However, the fact that the firm scale (x_{it}) changes within each firm is ignored, which will lead to the independence of the model residuals, and violates the basic assumption of regression.

In order to ensure the independence of the residuals and considering the fact that each firm has multiple observations, it is necessary to separate and estimate the individual characteristic quantity of each firm from the residual (ε_i). Thus, the following regression equation (2.2) is obtained.

$$y_{it} = \alpha + \beta x_{it} + \gamma_i + \varepsilon_{it}. \tag{2.2}$$

This equation takes into account the random effects of individual firms (individual characteristics γ_i), allowing each firm to have a separate intercept term $(\alpha + \gamma_i)$. Therefore, this model is called random intercept and fixed slope regression model. If analyzed more deeply, the individual random effects can be split into random intercept component and a random slope part:

$$y_t = \alpha + \beta x_{it} + \alpha_i + \beta_i x_{it} + \varepsilon_{it} = (\alpha + \alpha_i) + (\beta + \beta_i) x_{it} + \varepsilon_{it}. \tag{2.3}$$

In this equation, each firm has a separate intercept $(\alpha + \alpha_i)$ and slope $(\beta + \beta_i)$, hence it is called random intercept and random slope model. At this point, each individual is estimated to have a

unique trajectory. The intercept of these trajectories consists of the group horizontal intercept plus the deviation of individual random effect in the intercept term, and the slope of these trajectories consists of the group horizontal slope plus the deviation of individual random effect in the slope term. Note that the individual random effect may be positive or negative, because it indicates the way a given individual deviates from the group.

In summary, fixed effects are the group level intercept and slope ($\alpha + \beta x_{it}$). These effects are traditional main effects and interactions, and they are fixed factors or treatments that do not change randomly. Random effects are individual characteristic part ($\alpha_i + \beta_i x_{it}$), which explains the random variability across samples and are randomly selected factors or treatments in a certain range in experiments or research. In Chapter 3, when analyzing the variability of gene-environment interaction, different factors affecting phenotypic response values are regarded as fixed effects or random effects depending on the research purposes.

2.2 Principle Component Analysis and Singular Value Decomposition

Principal component analysis (PCA) is a widely used data dimensionality reduction algorithm. When the data set to be analyzed contains the observations of more than three interrelated quantitative variables, it will be very difficult to visualize the data because each variable can be treated as a different dimension. PCA can extract essential information from high-dimensional data tables and linearly combine the information into a set of new variables called principal components, which can be used to replace the original data, so as to reduce the dimension of the data. The core idea of PCA is to reconstruct the k −dimensional feature data on the basis of the original n −dimension feature data, and the k −dimension is a brand-new orthogonal feature data and the above-mentioned principal components. The workflow of PCA is to sequentially find a set of mutually orthogonal coordinate axes from the original dimension data space: the first new coordinate axis is chosen to be the direction with the largest variance in the original data; the second new coordinate axis is the direction that maximize the variance in the plane orthogonal to the first coordinate axis; the third coordinate axis is the direction with the largest variance in the plane orthogonal to the first and second axes; by analogy, we can get n such coordinate axes. By attaining new coordinate dimensions in this way, most of the variance of the original data is contained in the first k dimensions, and the variance in the following dimensions is almost zero. Therefore, it is reasonable to keep the first k features (principal components) that contain most of the variability of the original data, and ignore the features (redundancy) that contain almost zero variance. In fact, by calculating the covariance matrix of the original data matrix and then attaining the eigenvalues and eigenvectors of

covariance matrix, the matrix composed of eigenvectors corresponding to the largest k eigenvalues (i.e., the largest variance) is the reduced-dimension mapping matrix. Since we can calculate the eigenvalues and eigenvectors of the covariance matrix by eigenvalue decomposition and singular value decomposition, there are two implementations of PCA algorithm. These two implementations will be described in detail below.

Eigenvalue decomposition. For the covariance matrix of original data $A \in R^{n\times n}$, the k largest eigenvalues of the matrix, $\lambda_1, \lambda_2, \cdots, \lambda_k$, can be obtained by solving the following characteristic equation:

$$|A - \lambda E| = 0. \tag{2.4}$$

The eigenvectors $v_1, v_2, \cdots, v_k \in R^{n\times 1}$ can be obtained by taking the obtained $\lambda_1, \lambda_2, \cdots, \lambda_k$ into the following linear equations respectively:

$$(A - \lambda E)x = 0. \tag{2.5}$$

The eigenvectors $v_1, v_2, \cdots, v_k$ are formed into a matrix $[v_1, v_2, \cdots, v_k] \in R^{n\times k}$, i.e., the mapping matrix. The PCA is realized by multiplying the original data by the mapping matrix and each column in the reduced data matrix is a principal component.

Singular value decomposition. Eigenvalue decomposition can only be performed for $n \times n$ square matrices. In practices, the eigenvalue decomposition method usually fails since most of the matrices encountered are not square matrices. In this case, SVD can be applied to arbitrary matrix decomposition, including non-square matrices. For any matrix $A \in R^{m\times n}$, there is always a singular value decomposition:

$$A = U\Sigma V^T, \tag{2.6}$$

where $U \in R^{m\times m}$, the column vectors inside it are orthogonal and are called left singular vectors; $\Sigma \in R^{m\times n}$, the elements in Σ, except for the diagonal elements, are 0 and the diagonal elements are called singular values (also eigenvalues); $V^T \in R^{n\times n}$, and the orthogonal column vectors in it are called right singular vectors. To decompose the matrix A using SVD, the eigenvalues and eigenvectors of AA^T are first obtained and U is composed of unitized eigenvectors; then compute the eigenvalues and eigenvectors of A^TA and V is formed by these unitized eigenvectors; finally take the square root of the eigenvalues of AA^T or A^TA to form Σ. The eigenvalues in Σ are sorted from large to small, and the largest k eigenvalues are selected, and the corresponding k eigenvectors in the right singular matrix are used as column vectors to form the dimensionality reduction mapping matrix. The original data is multiplied by this mapping matrix to realize the dimensionality reduction. Using SVD for

PCA, the original data can be downscaled in two directions (row and column). The eigenvectors in the left singular matrix can be used to compress the original data in the row direction, and the right singular matrix is to reduce the dimensionality along column direction.

So far, this thesis has introduced the principles of PCA and the use of singular value decomposition to realize PCA. In Chapter 4, we will use principal component analysis and singular value decomposition to analyze the gene-environment interactions.

Chapter 3 Significance Analysis of Gene-Environment Interaction

This chapter mainly analyzes the significance of the gene-environment interactions. First, a mixed effect model based on genes, environment and gene-environment interaction is constructed, and then the significance of the model is analyzed to determine the significant effects that affect the phenotypic response. Moreover, the model can estimate the variance of the response value caused by gene and gene-environment interaction effect.

3.1 Mixed Effect Model

Generally, G varieties planted in E environments with R random replications generate $G \times E \times R$ observations. The "env" is usually a combination of location and year. Each combination of genotype and environment is called a "treatment", so there is a total of $G \times E$ treatments.

Table 3.1 Degree of freedom and effect category of variation source terms

Source of variation	Degree of freedom	Fixed or random effect				
		Case 1	Case 2	Case 3	Case 4	Case 5
CLT	**G-1**	random	fixed	fixed	random	fixed
LC	**L-1**	random	fixed	random	fixed	fixed
YR	**Y-1**	random	fixed	random	random	random
RP(LC∗YR)	**(R-1)LY**	random	random	random	random	random
CLT∗YR	**(G-1)(Y-1)**	random	fixed	random	random	random
CLT∗LC	**(G-1)(L-1)**	random	fixed	random	random	fixed
LC∗YR	**(L-1)(Y-1)**	random	fixed	random	random	random
CLT∗YR∗LC	**(G-1)(L-1)(Y-1)**	random	fixed	random	random	random

The mixed effect model constructed in this thesis should include four factors: genotype (CLT), location (LC), year (YR) and replication (RP) nested in location and year. According to different analysis purposes, genotype, location and year can be defined as random or fixed effects, as shown in Table 3.1, where G, L, Y and R are the numbers of varieties (genotypes), locations, years and replications respectively. Genotypes can be considered as random when the aim is to estimate variance components, genetic parameters, genetic gains obtained from breeding selection or different breeding strategies, etc. On the contrary, when the purpose is to compare experimental materials for selection or recommendation, genotypes are fixed factors. Similarly, when researchers are primarily interested in estimating variance components of sites

that represent relevant population in the target area, location is considered random. When it is interesting to make explicit comparisons of a level, the locations are fixed, and each location represents a clearly-defined area relative to crop management. Years and replicate trials are usually regarded as random factors.

Furthermore, as shown in Table 3.2, five different mixed effect models can be deduced in different cases. In Table 3.2, $*$ shows the interaction effect of different terms, i.e., the combination of grouping variables; 1 is the global intercept term of the model, indicating the benchmark value of Trait when all other effects are zero.

Table 3.2 Mixed effect models in different cases

Case	Mixed effect model
Case 1	$Trait \sim 1+1\|YR+1\|LC+1\|CLT+1\|(YR*LC)+1\|(YR*CLT)+1\|(LC*CLT)+1\|(YR*LC*CLT)+1\|(YR*LC*RP)$ (3.1)
Case 2	$Trait \sim YR*LC*CLT+1\|(YR*LC*RP)$ (3.2)
Case 3	$Trait \sim 1+1\|YR+1\|LC+CLT+1\|(YR*LC)+1\|(YR*CLT)+1\|(LC*CLT)+1\|(YR*LC*CLT)+1\|(YR*LC*RP)$ (3.3)
Case 4	$Trait \sim 1+1\|YR+LC+1\|CLT+1\|(YR*LC)+1\|(YR*CLT)+1\|(LC*CLT)+1\|(YR*LC*CLT)+1\|(YR*LC*RP)$ (3.4)
Case 5	$Trait \sim 1+1\|YR+LC+CLT+LC*CLT+(YR*LC)+1\|(YR*CLT)+1\|(LC*CLT)+1\|(YR*LC*CLT)+1\|(YR*LC*RP)$ (3.5)

This work uses the lmer function of lme4 package[22] to fit the mixed effect models, calculate the estimates and standard deviations of the fixed effect terms and the estimates of the variation components of the random effect terms. In order to obtain the response predictions of a certain trait by using the mixed effect model, it is necessary to compute the estimates of the random terms. The best linear unbiased prediction (BLUP) of random effect is the predicted value of the random effect at the individual or subgroup level, and its formula is as follows:

$$\hat{u} = E(u|y) = GZ^T V^{-1}\left(y - X\hat{\beta}\right), \tag{3.6}$$

$$V = ZGZ^T + R, \tag{3.7}$$

where G is the covariance matrix of random effects, Z the design matrix of random effects, V the total covariance matrix, R the residual covariance matrix, X the design matrix of fixed effects, and $\hat{\beta}$ the estimate of fixed effects. In the experiment, the ranef function[22] of lme4 package is used to estimate BLUPs of random effects. The estimates of random effects tend to be "shrunk" towards the population mean relative to the fixed effect estimates. Then the best

linear unbiased predictions corresponding to each genotype is the summation of all fixed effect and random effect estimates.

3.2 Significance Analysis

After constructing the mixed effect model (Table 3.2), it is necessary to analyze the significant effect items affecting the phenotype response, and estimate the variation magnitude of each part in the model, which is beneficial to simplifying the model and also a prerequisite for later stability analysis.

For the fixed effect terms in the model, the lmer function will automatically use Satterthwaite approximation method to calculate the degree of freedom and p value when fitting. Satterthwaite method is suitable for equilibrium designs or simple models, but may not be accurate enough in complex models, such as those with random replications and high-order interaction effects. Here, the mixed function in afex package[29] is employed to compute the degree of freedom and p value of F test by Kenward-Roger approximation.

On the other hand, it is often more complicated to calculate the significance of random effect terms, and we need construct auxiliary mixed effect models. When the gene (CLT), location (LC) and year (YR) are all random effects (case 1), in order to validate the significance of YR, it is necessary to construct an auxiliary mixed effect model as follows:

$$Trait \sim 1 + 1|LC + 1|CLT + 1|(YR * LC) + 1|(YR * CLT) + 1|(LC * CLT) + 1|(YR * LC * CLT) + 1|(YR * LC * RP). \quad (3.8)$$

Compared with formula (3.1), formula (3.8) only lacks one effect term whose significance needs to be analyzed. Now all that is required to verify the significance of YR is to analyze the differences between models (3.8) and (3.1). It is the same operations to verify the significance of other random terms. The significance of the difference between two mixed effect models is estimated by likelihood ratio test to obtain p value. Here, likelihood ratio is the probability of known data under a given model. The logic of likelihood ratio test is to compare the two models against each other. According to Wilk's theorem, the negative double of the log likelihood ratio of the two models approximates the chi-square distribution with k degrees of freedom, where k is the number of random effects in the test. In this paper, the anova function in stats package is utilized to test the log-likelihood ratio between the models without and with the factor to be tested, which gives the corresponding chi-square value, degree of freedom and p value. Using this method, the significance of random effects can be attained.

Until now, this paper has developed the mixed effect model of phenotype responses, and tested the significance of fixed effect and random effect terms in the model, so as to judge

whether genes and the interaction between genes and environment have a statistically significant impact on gene traits. Only if the interaction between genes and environment is significant, the later model for gene stability analysis is meaningful. At the same time, according to the model, we can also estimate the trait variance caused by various significant components such as genes and gene-environment interactions, and the larger the variance is, the greater influence it has on the gene performance.

Chapter 4 Stability Analysis of Gene-Environment Interaction

After determining that gene-environment interactions statistically exist by significance analysis, this thesis continues to analyzes the adaptability of genotypes (variety) in different environments and the mutual assistance between genotypes and environments. This chapter will first introduce the single-gene stability model based on grouped regression, and then the multi-gene stability model, which includes additive main effect and multiplicative interaction and gene main effect plus gene-environment interaction. It is worth noting that single-gene model can only analyze a single genotype (variety) at a time, and quantitatively investigate its stability in different environments. Multi-gene model can simultaneously analyze the adaptability and interaction patterns of all genotypes (varieties) in different environments. It processes the data of all genotypes and environments, and study the gene-environment interaction by multivariate methods (such as principal component analysis and singular value decomposition).

4.1 Single-Gene Stability Analysis

In the single-gene stability model, the average trait observation (ENVTrait) needs to be calculated by YR-LC-RP groups. Let's assume that there are $(E_1, E_2, \cdots, E_k)$ levels of ENV and $(R_1, R_2, \cdots, R_M)$ levels of RP, and then construct the following group linear regression model:

$$Trait_j = \beta_{0j} + \beta_{1j} \cdot ENVTrait + \sum_{i=2}^{K} \beta_{E_{ij}} \cdot I(ENV = E_i) + \sum_{i=2}^{M} \beta_{R_{ij}} \cdot I(RP = R_i) + \epsilon_j, \quad (4.1)$$

where j represents the j-th genotype (variety), $I(\cdot)$ is an indicative function and ϵ_j an error term, and we take the first level E_1, R_1 as a reference by default.

In this model, we particularly focus on the regression slope (β_{1j}) and regression deviation (s_d^2). β_{1j} is the slope of the j-th variety to the continuous variable ENVTrait, which indicates the response of genotypes to ENVTrait, and ENVTrait comes from the average performance of all genotypes in each environment. Therefore, if there is no statistically significant difference between β_{1j} with 1, the variety has strong adaptability to all environments; β_{1j} greater than 1 describes genotypes that are more sensitive to environmental changes (lower than average stability) and have higher mutual assistance with high-yield environments; β_{1j} less than 1 provides greater resistance to environmental changes (higher than the average stability), so cultivars have better adaptability to low-yield environments of drought and water shortage. Similarly, we can deduce the same conclusions by analyzing the relationship between s_d^2 and 0. Here, we need to test the regression slope ($H0$: $\beta_{1j} = 1$) by T-test and the regression deviation

($H0: s_d^2 = 0$) by F-test. We perform T-test and F-test by using lm function, dplyr[31] and tidyr[32] packages in R.

Environment index is the average performance of all genotypes in each environment. In order to compute the variance statistics of gene stability, the following Gaussian generalized linear model is needed additionally:

$$Trait \sim LC + YR + YR * LC + LC * YR * RP + CLT + CLT * LC + CLT * YR + CLT * LC * YR. \quad (4.2)$$

Subsequently, we need to use the variations and residuals of the model to calculate Shukla's σ_i^2, ssquares and Wricke's stability ecovalence W_i^2 and Kang's yield stability YS_i by the Stability.par function of agricolae package[30].

Although the single-gene stability model can numerically analyze the stability of each genotype, it cannot solve the problems such as which genes have higher yield in which environment, so it is necessary to develop the multi-gene stability model.

4.2 Multi-Gene Stability Analysis

This section introduces the multi-gene stability analysis models, and analyzes the adaptability and capacity of multiple genotypes in different environments, so as to facilitate the cross-sectional comparison of multiple genotypes. The models include additive main effect and multiplicative interaction (AMMI) and the gene main effect plus gene-environment interaction (GGE).

4.2.1 Additive Main Effect and Multiplicative Interaction

For AMMI, it is a combination of ANOVA and PCA. PCA calculates the gene score and environment score, and adds the product of the two to the overall mean as the trait prediction of the genotype in the environment. ANOVA calculates the gene deviation (the difference from the overall mean) and the environment deviation, and adds the sum of them to the overall mean as the trait estimation of the genotype. ANOVA model leaves a non-additive residual, i.e., the interactions between gene and environment. ANOVA is an additive model and PCA is a multiplicative model, so AMMI is called the additive main effect and multiplicative interaction model.

The additive principal effect and multiplicative interaction model firstly estimates the additive main effect of two-way data table (gene and environment) by ANOVA, and then applies PCA to the residual term (interaction) of additive ANOVA model to approximate $N \leq \min(I - 1, J - 1)$ interaction principal components (IPC), obtaining the estimates of the multiplicative term of AMMI model. I is the number of genotypes (rows) and J number of environments

(columns) considered in the two-way data table. For simplicity, supposing a completely random trial design, the AMMI model can be written as:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{n,j} + \rho_{i,j} + \epsilon_{i,j,k}, \quad (4.3)$$

where $y_{i,k,k}$ is the trait response of the i-th genotype (variety) in the j-th environment of the k-th replication, μ is the overall mean, α_i the gene-induced deviation from the overall mean, β_j the environment-induced deviation from the overall mean, λ_n the singular value of the n-th axis of IPC, $\gamma_{i,n}$ and $\delta_{n,j}$ the IPC scores of the i-th genotype and the j-th environment of n-th axis, $\rho_{i,j}$ the residual of all multiplication terms not included in the model, $\epsilon_{ij,k}$ the experiment error and N the number of maintaining principle components.

Expressed in the form of matrix:

$$\boldsymbol{Y} = \mathbf{1}_i \mathbf{1}_j^T \mu + \alpha_i \mathbf{1}_j^T + \mathbf{1}_i \beta_j^T + \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + \varepsilon, \quad (4.4)$$

where $\boldsymbol{Y} \in R^{I \times J}$ is the trait mean across repeated trials or blocks, and each column of $\boldsymbol{Y}$ represents the vector of genotypic means as obtained from the phenotypic analysis of a corresponding trial by an appropriate mixed model that accounts for experimental design features or spatial trends. $\mathbf{1}_i \mathbf{1}_j^T \mu$ is a $(I \times J)$ matrix with the grand mean μ in all positions, $\alpha_i \mathbf{1}_j^T$ is the $(I \times J)$ main effect matrix of genes, and $\mathbf{1}_i \beta_j^T$ is the $(I \times J)$ main effect matrix of environment. The gene-environment interactions $\boldsymbol{Y}^* = \boldsymbol{Y} - \mathbf{1}_i \mathbf{1}_j^T \mu - \alpha_i \mathbf{1}_j^T - \mathbf{1}_i \beta_j^T$ are approximated by the product of $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{U}$ is a $(I \times N)$ matrix whose columns contains left singular vectors; $\boldsymbol{D}$ is a $(N \times N)$ diagonal matrix containing the singular value of $\mathbf{Y}^*$; $\boldsymbol{V}$ is a $(J \times N)$ matrix whose columns contain the right singular vectors. The residual term $\varepsilon \in R^{I \times J}$ in the equation contains both the misfitting term and the error term of the model.

It is quite crucial to choose the number of multiplication terms *N* (i.e. the number of principal components retained) in AMMI, because it will affect the subsequent results[23]. Forkman and Piepho put forward a new method on how to select the number of principal components in AMMI based on parametric bootstrap resampling[24]. This thesis draws on this method, and uses agricolae package[30] to realize AMMI model construction and biplot analysis in RGxEStat.

4.2.2 Gene Main Effect plus Gene-Environment Interaction

Unlike AMMI model, GGE model considers gene main effects and gene-environment interaction effects, which are also referred as locus regression model. In AMMI, main effects

and interaction effects are estimated separately, and bilinear parameters specifically describe gene-environment interactions. In GGE, the main effect of genotype is not much different from gene-environment interaction effect. Therefore, the multiplicative part of the model describes the effects of gene, gene-environment interaction at the same time. In addition, the stability and adaptability of genotype can be directly explained by examining the GGE biplots. GGE biplots mainly analyze the first two principal components, which are obtained based on the singular value decomposition of standardized two-way (gene and environment) data matrix in the environment center or within environment. The first principal component describes the performance (or fitness), and the second principal component describes the effects of gene-environment interaction (or stability).

The GGE model can be formulated as:

$$\boldsymbol{Y} = \boldsymbol{X_1\beta} + \sum_{k=1}^{t} \lambda_k \mathrm{diag}(\boldsymbol{Z\alpha_k})\boldsymbol{X_2\gamma_k} + \boldsymbol{\varepsilon}, \tag{4.5}$$

where the vector $\boldsymbol{Y}$ contains $\left(v \times (r \times j)\right)$ phenotypic responses, v are the number of genotypes, r the number of repeated trials or blocks, and j the number of environments (combination of location and year). $\boldsymbol{\beta} \in R^{(r\times j)\times 1}$ represents the effect vector of replications within locations, which is obtained from the appropriate mixed effect model. λ_k, $\boldsymbol{\alpha_k}$, $\boldsymbol{\gamma_k}$ respectively denotes the singular value and the genotype and environment singular vectors related to the k-th principal component, where $k = 1, \cdots, t$ and $t = min(v-1, j)$ is the rank of the gene and gene-environment interaction matrix $GGE_{(r\times j)\times c}$. In addition, $\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{Z}$ are the design matrices; $\boldsymbol{\varepsilon} \in R^{[v\times(r\times j)]\times 1}$ is the experiment error and $\boldsymbol{\varepsilon}|\sigma_e^2 \sim N_{v\times(r\times j)}\left(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_{v\times(r\times j)}\right)$; σ_e^2 is the residual variance, $\boldsymbol{0} \in R^{[v\times(r\times j)]\times 1}$ the zero vector，$\boldsymbol{I}_{v\times(r\times j)}$ the unit matrix.

The vector $\boldsymbol{Y}$ obeys a multivariate normal distribution conditioned on the model parameters,

$$\boldsymbol{Y}|\boldsymbol{\beta}, \lambda, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_e^2 \sim \boldsymbol{N}(\boldsymbol{\mu_y}, \boldsymbol{I_{v\times(r\times j)}}\sigma_e^2), \tag{4.6}$$

where $\boldsymbol{\mu_y} = \boldsymbol{X_1\beta} + \sum_{k=1}^{t} \lambda_k \mathrm{diag}(\boldsymbol{Z\alpha_k})\boldsymbol{X_2\gamma_k}$.

Compared with the AMMI biplot, GGE biplots are highly attractive. Based on the analysis of the inner product property between singular vectors (genotypes and environments) in GGE model, more different kinds of biplots can be potted to identify genotypes with high yield and wide adaptability (stability), and the winning varieties are also suggested in each environment.

This work employs GGEBiplotGUI package[25] to construct GGE model, and integrates it

into RGxEStat software, so that users can perform interactive GGE model analysis and draw various biplots without running any code.

Chapter 5 Experiment Results and Analysis

This chapter will use RGxEStat to perform the significance and stability analysis on two public available breeding datasets, and show the experiment results of the models. The interface of RGxEStat is shown in Figure 5.1, including significance analysis, single-gene stability model analysis and multi-gene stability model analysis. Code and datasets are available at https://github.com/mason-ching/RGxEStat.

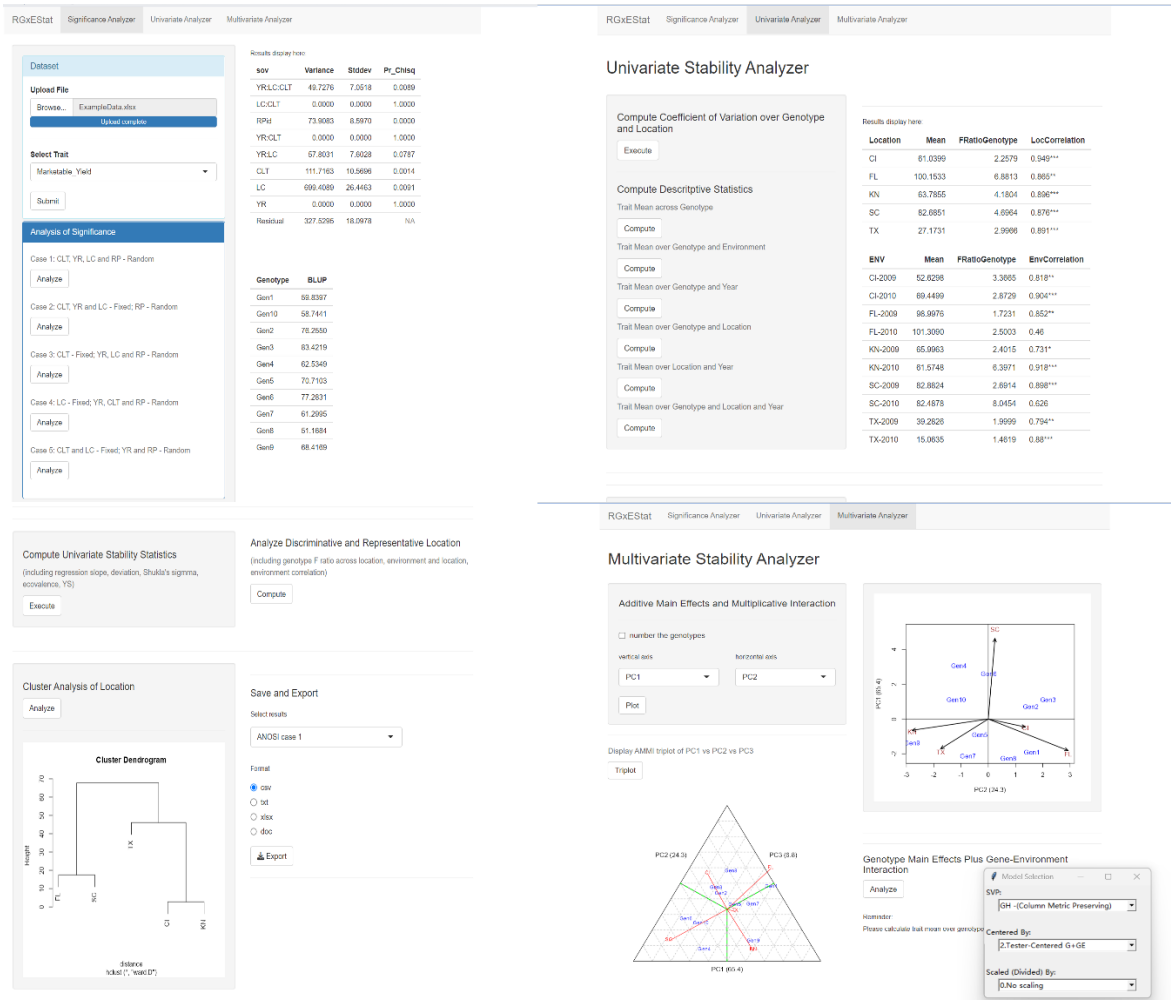


Figure 5.1 Interface of RGxEStat

5.1 Data Sources

The experiment data selected in this thesis consists of two parts: watermelon breeding data in the southern part of the United States from 2009 to 2010; oat field random trial data from agridat package. The watermelon breeding data in the southern United States were obtained from random and replicated block trials across 2 years, 10 varieties, 5 locations, 4 replications，with the marketable yield as the analyzed trait. The oat field trial data includes 6 environments,

24 varieties and the corresponding marketable yield.

Table 5.1 shows the template of multi-year multi-location random trial data (breeding data) of gene and environment, in which YR, LC, RP, CLT and MY represent year, location, replication, variety (genotype) and marketable yield, respectively.

Table 5.1 Demo of watermelon yield trial data in the southern United States in 2009-2010

YR	LC	RP	CLT	MY
2009	KN	1	EarlyCanda	56.236
2009	KN	1	CalhounGray	74.167
2009	KN	1	GeorgiaRattlesnake	55.873
2010	TN	2	EarlyCanda	32.601
2010	TN	2	CalhounGray	74.167
2010	TN	2	GeorgiaRattlesnake	64.794

5.2 Significance Analysis Results

This section begins with modeling watermelon breeding data from the southern region of the United States by constructing the mixed effect model of case 1. Subsequently, we perform a significant analysis of all the fixed effect and random effect terms in the model by F test and likelihood ratio test respectively, and determine the significant effects influencing the watermelon yield. The analysis results are shown in Table 5.2.

Table 5.2 Significance analysis of watermelon breeding data

Source of variance	Variance	Standard deviation	p value of Chi-square test
YR ∗ LC ∗ CLT	49.74	7.05	0.008
YR ∗ LC ∗ RP	73.91	8.60	0.000
LC ∗ CLT	0.00	0.00	1.000
YR ∗ CLT	0.00	0.00	1.000
YR ∗ LC	57.81	7.60	0.078
CLT	111.68	10.57	0.001
LC	699.36	26.45	0.009
YR	0.00	0.00	1.000
residual	327.53	18.10	-

In this example, the chi-square p values of the random effects location-genotype (LC ∗ CLT), year-genotype (YR ∗ CLT), year-location (YR ∗ LC) and year (YR) are greater than significant level $\alpha = 0.05$, so it can be concluded that these terms have no effect on watermelon yield. Meanwhile, it also indicates that the mixed effect model is over-fitted and the model is more

complicated than the analyzed data. It is often the case that random effect variance estimated as zero appear when those effects have too few or small number of levels. The alternate option is to use MCMCglmm package[28] to implement Markov chain Monte Carlo (MCMC) simulation to obtain the significance probability of random effects. Subsequently, the oat field data are analyzed for significance probability by developing the mixed effect model of case 2. The results are displayed in Table 5.3.

Table 5.3 Significance analysis of oat field data

Source of variance	Mean of squared error	F value	p value of F test
LC ∗ CLT	0.411	2.989	0.018
CLT	0.554	3.547	0.013
LC	0.334	2.856	0.037
residual	0.156	-	-

Since year is not recorded in the oat field random trial data, there is no effect term related to the year. From Table 5.3, it can be seen that the F-test p values of location-genotype (LC ∗ CLT), location (LC) and genotype (CLT) are all greater than significant level $\alpha = 0.05$, so it can be assumed that these effects have an impact on oat yield.

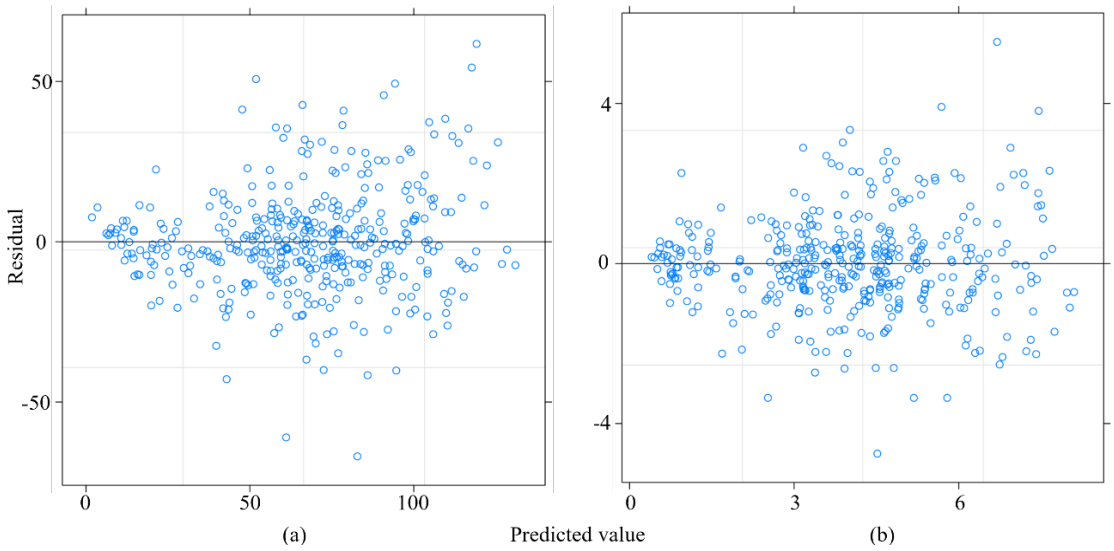


Figure 5.2 Scatter plots of yield predictions versus model residuals for (a) watermelon breeding data and (b) oat field random trial data.

Finally, the RGxEStat can automatically calculate the best linear unbiased predictions of fixed effects and random effects in the model, and get the yield estimates. The predictions of yield and the residual error of the model calculated by RGxEStat are shown in Figure 5.2. After significant analysis, we can judge that gene effects, gene-environment effects will affect the yield of watermelon and oat, so it is necessary to carry out subsequent stability analysis to select

varieties with excellent high yield and strong stability.

5.3 Results of Single-Gene Stability Analysis

In this section, we will show RGxEStat to model single-gene stability of watermelon breeding data. RGxEStat can fast obtain single-gene stability statistics, including single-gene regression slope (β_{1j}), regression deviation (s_d^2), Shukla's σ_i^2, ssquares, Wricke's stability ecovalence (W_i^2) and Kang's yield stability statistic (YS_i). The results of single-gene stability analysis of watermelon breeding data are shown in Table 5.4.

In Table 5.4, ns indicates not significant, and + means that the genotype can be judged to be stable according to Kang's yield stability. In the β_{1j} column, ** indicates that the t-test p values of the regression slope corresponding to this genotype is less than the significance level $\alpha = 0.01$ (*** and * are the significance levels of 0.001 and 0.05), so $H0: \beta_{1j} = 1$ is rejected, stating that the regression slope is significantly different from 1. Similarly, in s_d^2 column, ** represents F-test p value of the regression deviation corresponding to the genotype is less than the significant level $\alpha = 0.01$, and the rejection of $H0: s_d^2 = 0$ indicates that the regression deviation is significantly different from 0. From the statistical analysis in the table, it can be concluded that the watermelon varieties of CalhounGray, FiestaF1, GeorgiaRattlesnake, Legacy and StarbriteF1 have strong adaptability to the environment and are stable genotypes.

Table 5.4 西瓜育种数据单基因稳定性分析表

CLT	β_{1j}	s_d^2	σ_i^2	ssquares	W_i^2	YS_i
CalhounGray	1.301	124.670	61.347 ns	15.761 ns	279.747	10+
CrimsonSweet	1.341	1450.035**	439.125 ns	567.988 ns	1488.636	4
EarlyCanada	0.249**	686.251**	253.230 ns	285.370 ns	893.772	2
FiestaF1	1.639	657.873	300.285 ns	385.997 ns	1044.349	11+
GeorgiaRattlesnake	0.945	220.063	52.213 ns	44.863 ns	250.518	8+
Legacy	1.056	428.070	287.394 ns	262.488 ns	1003.097	7+
Mickylee	0.618	705.481**	188.105 ns	195.126 ns	685.374	3
Quetzali	0.965	96.532	82.809 ns	86.103 ns	348.425	1
StarbriteF1	1.388	221.137	157.241 ns	78.371 ns	586.607	12+
SugarBaby	0.498**	332.181**	264.187 ns	308.430 ns	928.836	-1

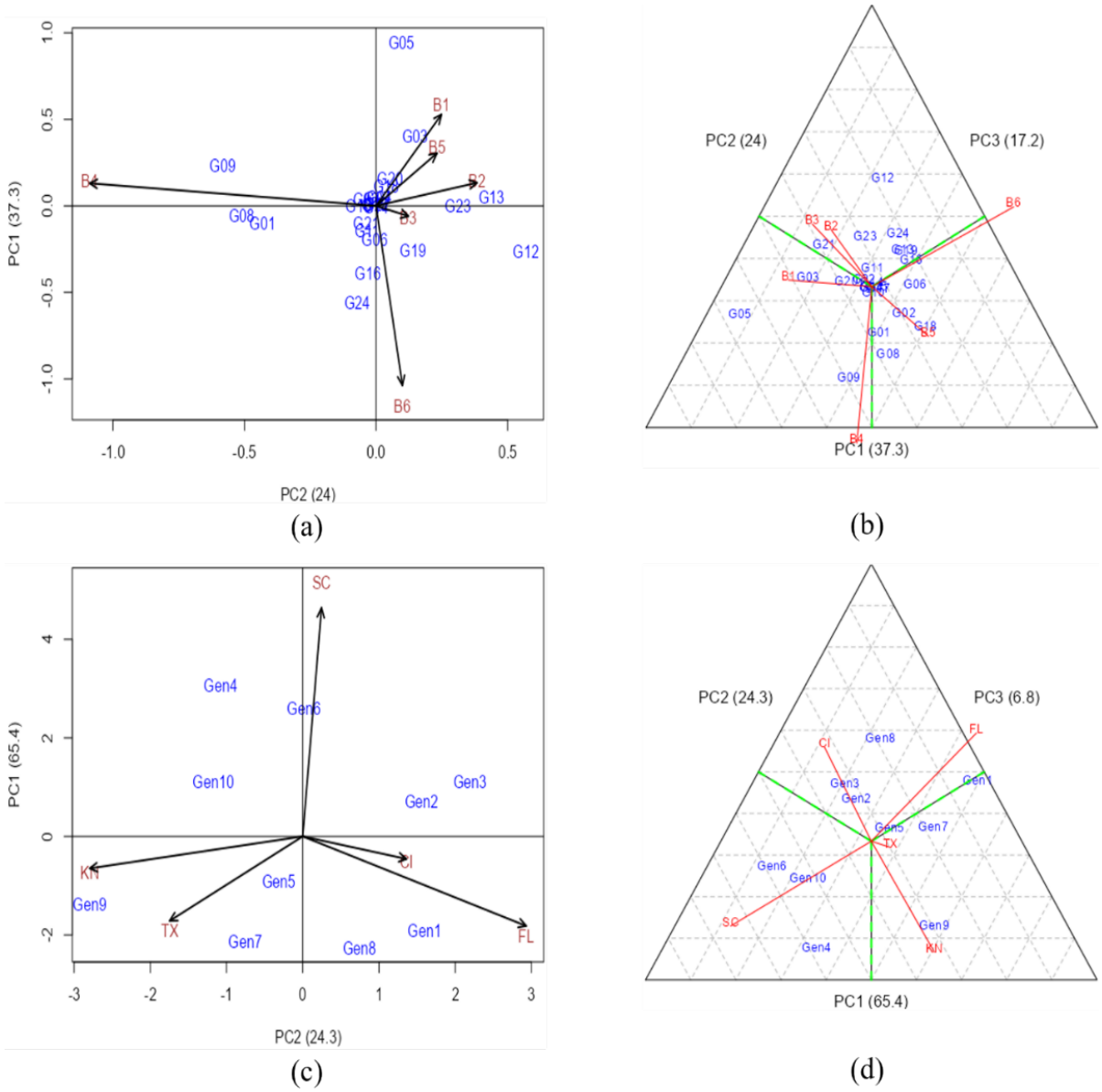
5.4 Results of Multi-Gene Stability Analysis

In this section, we mainly present the principal component biplots of stability analysis model, and utilize them to compare the superiority between multiple genotypes and multiple

environments and determine the mutual assistance between genotypes and environments.

5.4.1 Results on Additive Main Effect and Multiplicative Interaction

RGxEStat utilizes agricolae package to establish and analyze the additive main effect and multiplication interaction model. Taking watermelon breeding data and oat field data as examples, the biplots of AMMI model drawn by RGxEStat is shown in Figure 5.3, where the number of principal components retained is set to 4.



(a) (b) (c) (d)

Figure 5.3: (a) PC1 v.s. PC2 biplot of oat field data, (b) PC1 v.s. PC2 v.s. PC3 triplot of oat field data, (c) PC1 v.s. PC2 biplot of watermelon breeding data, (d) PC1 v.s. PC2 v.s. PC3 triplot of watermelon breeding data.

In order to make biplots more concise, Gen1-10 is used to replace watermelon varieties

EarlyCanada, CalhounGray, StarbriteF1, CrimsonSweet, GeorgiaRattlesn, FiestaF1, Mickylee, SugarBaby, Legacy and Quetzali respectively in this section and thereafter. According to the model formula (4.4), the interaction between genes and environments can be represented by multiplying the gene score and environment score in each principal component and summing them along the principal components. Figure 5.3 provides a very effective depiction of the gene-environment interactions. In Figure 5.3, genotypes and environments near the axis origin will not interact with each other (the product of their scores will be close to zero); genotypes and environments far away from the axis origin show great interaction, and therefore have high yield instability. Some people argue that the Euclidean distance from the origin should be regarded as a measure of instability. When genotypes and environments are close to each other, the interaction is positive. If two objects are close, their scores (coordinates) will have the same sign, so their products will be positive. When genotypes and environments are far away from each other, the interaction is negative.

For example, for oat varieties, G03 variety in B1 environment, G13 in B2 environment, G05 in B1, B5, B2 environment have good yields (compared with the average level), while watermelon varieties Gen7 and Gen9 have particularly excellent yields in KN and TX environment.

5.4.2 Results on Gene Main Effect plus Gene-Environment Interaction

The GGE model decomposes the total effects of genes and gene-environment interactions by SVD and PCA, and then makes various biplots according to PC1 and PC2. In this section, only the watermelon breeding data is taken as an example to show the analysis process of GGE model by RGxEStat. RGxEStat can plot various GGE biplots with one click to identify varieties with high yield and high stability, and can also suggest which varieties are suitable for planting in a specific environment. Figure 5.4 illustrates a wide variety of GGE biplots, including discrimitiveness and representativeness, variety yield and stability, environment ranking, which won where/what, variety ranking and relationship between environments. The PC1 v.s. PC2 biplot is not shown here, because it is similar to Figure 4.3 (c).

Figures 5.4 (a) and (c) are used to judge the discrimination and representativeness of different environments, i.e., which environment can better distinguish high-yield and high-stability varieties (the length of line segment in Figure 5.4 (a)) and which environment has strong representativeness for the target ecological zone. The arrowed line in (a) is the mean environment axis. The direction indicated by the arrowhead on it is the evaluation of the discrimination and representativeness of the test locations. The angle between the line segment

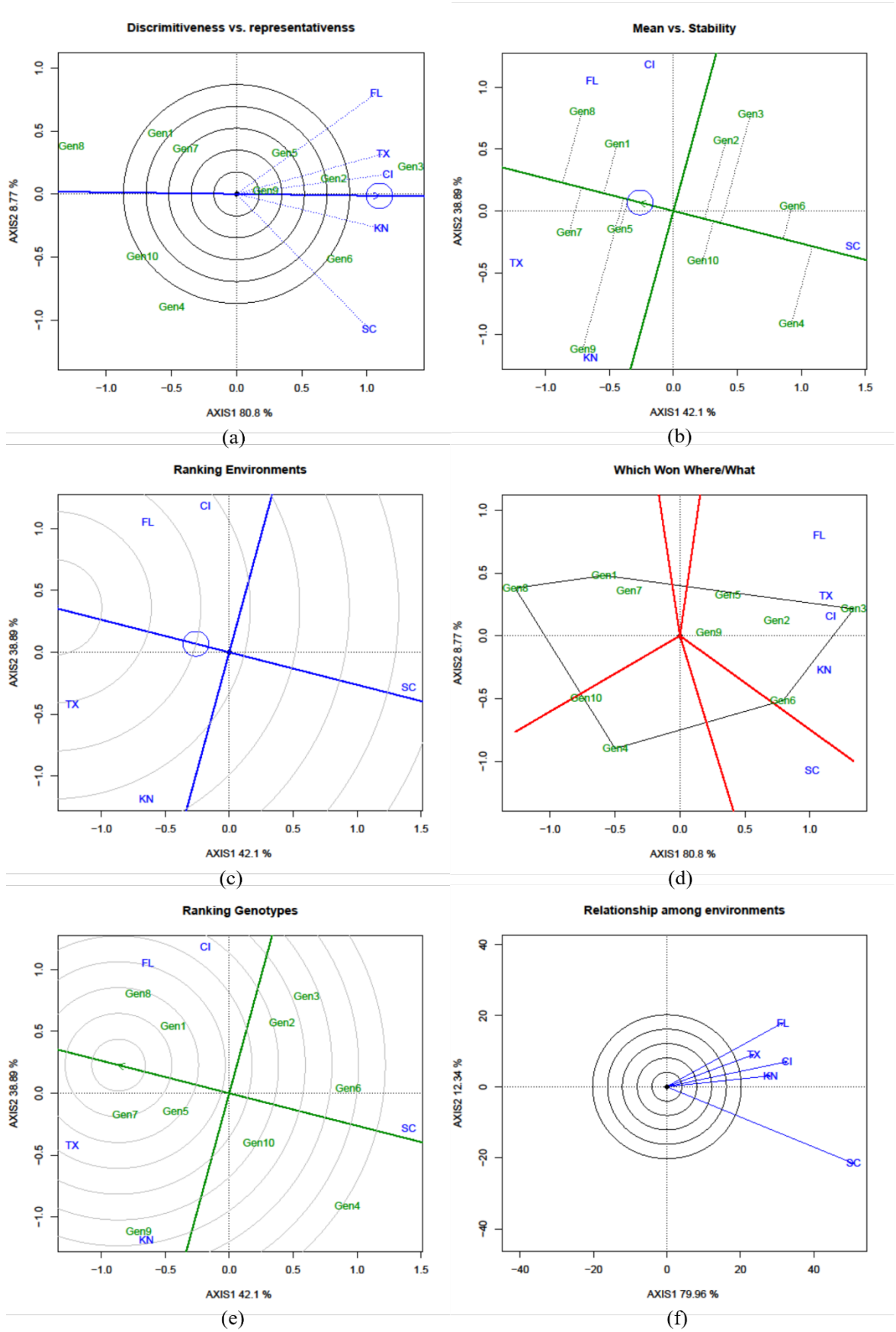


Figure 5.4 GGE biplots: (a) discrimitiveness v.s. representativeness, (b) yield mean v.s. stability, (c) ranking environments, (d) which won where/what, (e) ranking genotypes, (f) relationship among environments.

of the test site and the mean environmental axis is a measure of its representativeness of the target environment. The smaller the angle, the stronger the representativeness. If the angle between a test site and the mean environment axis is obtuse, it is not suitable as a trial environment. A test site without discrimination is useless. Discriminatory but unrepresentative test sites can eliminate unstable varieties. Only those with both discriminative and representative test sites can choose the best varieties with high and stable yield, such as the environment CL and TX in the biplot. Figure 5.4 (c) draws concentric circles according to the mean environment axis, where the smaller the circle is, the stronger the distinctiveness and representativeness of the environment.

Figures 5.4 (b) and (e) reflect the productivity and stability of varieties. In Figure 5.4 (e), concentric circles are made according to the mean environment axis and if the circle where varieties are located is smaller, the better yield-environment interaction of varieties is, the better the variety is. However, the figure 5.4 (b) of mean and stability also requires the mean environment axis (straight line with arrow) and the mean environment value. There is also a straight line perpendicular to the mean environment axis through the center. The direction of the mean environment axis refers to the trend of the approximate average yield of varieties in all environments. Make a vertical line (green dotted line) between varieties and the mean environment axis. The longer the vertical line between varieties and the mean environment axis, the more unstable the varieties are. Thus, Gen1, Gen5 and Gen7 have stability and high yield.

Figure 5.4 (d) mainly illustrates the variety with the highest yield in each environment according to the interaction between the variety and the environment. The figure connects the farthest points in all directions with straight lines to form a polygon, and divides the biplot into several sectors by making the center perpendicular to each side. The variety is distributed in the sectors. For the environments in a specific sector, the variety at the top corner of the sector has the best yield. For example, Gen3 variety has the highest yield in FL, TX, CL and KN environments.

Figure 5.4 (f) depicts the relationship between the environments. It draws a line segment from the coordinate center to each environment, mainly to evaluate the distinctiveness and similarity of the environments. The cosine value of the angle ($0^{\circ} - 180^{\circ}$) between the two environment line segments is their correlation coefficient. The angle between the two line segments is less than 90 degrees, which indicates a positive correlation and the two environments are similar in ranking the variety quality. The angle between the two line segments greater than 90 degrees indicates negative correlation and the ranking of variety

quality in the two environments is mostly opposite; The angle equal to 90 degrees means that the two environments are not related. The smaller angle close to zero degrees means that the two environments may be duplicated trial sites, and the evaluation of varieties will not be affected if one of them is removed. The length of the line segment represents the ability of the environment to discriminate between varieties, and the longer the line segment, the stronger the discrimination.

Chapter 6 Summary and Outlook

This thesis introduces in detail the statistical methods for gene-environment interaction analysis, including significance analysis based on mixed effect model, single-gene stability model and multi-gene stability model. Then it presents RGxEStat, a portable interactive software integrating the above analysis methods and models. It is hoped that this software can avoid the need for breeders and agronomists to learn complex SAS or R programming and provide them with a simple and easy-to-operate tool for gene-environment interaction analysis, and thus shorten their research time.

Currently, this field mainly relies on statistical models and methods to analyze multi-environment breeding data and gene-environment interactions. Then, with the wide application of artificial intelligence and deep learning methods in various cross-cutting fields, in the future, the author hopes to develop some breeding tools based on deep learning to analyze and model complex high-dimensional nonlinear breeding data and nonlinear components in the interaction. Besides, it is also important to process multi-trait, multi-gene and multi-environment data simultaneously in agricultural production, because breeders always want to select varieties with more excellent traits at the same time. These excellent traits may be controlled by the same gene or have nothing to do with each other, so it is very challenging to analyze the gene-environment interaction data for multiple phenotypic traits. At present, there is still a lack of academic research in this field, and this issue is also a direction worth exploring in the future.

References

[1] Kumar, R., Dia, M., & Wehner, T. C. (2013). Implications of mating behavior in watermelon breeding. *HortScience*, *48*(8), 960-964.

[2] Dia, M., Wehner, T. C., & Arellano, C. (2016). Analysis of genotype×environment interaction (G×E) using SAS programming. *Agronomy Journal*, *108*(5), 1838-1852.

[3] Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*, *143*(6), 449-462.

[4] Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, *28*(4), 556-580.

[5] Eberthart, S. A., & Russel, W. A. (1966). Stability parameters for comparing varieties. *Crop Science*, *6*(1), 36-40.

[6] Perkins, J. M., & Jinks, J. L. (1968). Environmental and genotype-environmental components of variability. *Heredity*, *23*(3), 339-356.

[7] Wricke, G. (1962). Evaluation method for recording ecological differences in field trials. *Z Pflanzenzücht*, *47*(1), 92-96.

[8] Shukla, G. K. (1972). Genotype stability analysis and its application to potato regional trails. *Crop Science*, *11*, 184-190.

[9] Kang, M. S. (1993). Simultaneous selection for yield and stability in crop performance trials: Consequences for growers. *Agronomy Journal*, *85*(3), 754-757.

[10] Mekbib, F. (2003). Yield stability in common bean (Phaseolus vulgaris L.) genotypes. *Euphytica*, *130*, 147-153.

[11] Fan, X. M., Kang, M. S., Chen, H., Zhang, Y., Tan, J., & Xu, C. (2007). Yield stability of maize hybrids evaluated in multi-environment trials in Yunnan, China. *Agronomy journal*, *99*(1), 220-228.

[12] Lin, C., & Binns, M. (1988). A method for analyzing cultivar x location x year experiments: A new stability parameter. *Theoretical Aplied Genetics*, *76* (3), 425–430.

[13] Francis, T. R., & Kannenberg, L. W. (1978). Yield stability studies in short-season maize. I. A descriptive method for grouping genotypes. *Canadian Journal of plant science*, *58*(4), 1029-1034.

[14] Casanoves, F., Baldessari, J., & Balzarini, M. (2005). Evaluation of multienvironment trials of peanut cultivars. *Crop Science*, *45*(1), 18-26.

[15] Gauch Jr, H. G. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop*

science, *46*(4), 1488-1500.

[16] Yan, W., Kang, M. S., Ma, B., Woods, S., & Cornelius, P. L. (2007). GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop science*, *47*(2), 643-653.

[17] Yan, W., & Kang, M. S. (2002). GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC press.

[18] Yan, W., Hunt, L. A., Sheng, Q., & Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop science*, *40*(3), 597-605.

[19] Gauch, H. J. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs* (pp. xi+-278).

[20] Crossa, J., Gauch Jr, H. G., & Zobel, R. W. (1990). Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. *Crop science*, *30*(3), 493-500.

[21] Zobel, R. W., Wright, M. J., & Gauch Jr, H. G. (1988). Statistical analysis of a yield trial. *Agronomy journal*, *80*(3), 388-393.

[22] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*, 1-48.

[23] Gauch Jr, H. G. (2013). A simple protocol for AMMI analysis of yield trials. *Crop science*, *53*(5), 1860-1869.

[24] Forkman, J., & Piepho, H. P. (2014). Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics*, *70*(3), 639-647.

[25] Frutos, E., Galindo, M. P., & Leiva, V. (2014). An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment*, *28*, 1629-1641.

[26] Hussein, M. A., Bjornstad, A. S., & Aastveit, A. H. (2000). SASG× ESTAB: A SAS program for computing genotype×environment stability statistics. *Agronomy journal*, *92*(3), 454-459.

[27] Piepho, H. P. (1999). Stability analysis using the SAS system. *Agronomy Journal*, *91*(1), 154-160.

[28] Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software*, *33*, 1-22.

[29] Henrik Singmann, Ben Bolker, Jake Westfall and Frederik Aust (2016). afex: Analysis of Factorial Experiments. R package version 0.16-1. https://CRAN.R-project.org/package=afex.

[30] de Mendiburu F (2023). agricolae: Statistical Procedures for Agricultural Research. R package version 1.3-7, https://CRAN.R-project.org/package=agricolae.

[31] Wickham H, François R, Henry L, Müller K, Vaughan D (2025). dplyr: A Grammar of Data Manipulation. R package version 1.1.4.9000, https://github.com/tidyverse/dplyr.

[32] Wickham, H. (2016) tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions. R package Version 0.4.1. http://CRAN.R-project.org/package=tidyr.

[33] Dia, M., Wehner, T. C., & Arellano, C. (2017). RGxE: An R program for genotype x environment interaction analysis. *American Journal of Plant Sciences*, *8*(7), 1672-1698.

Acknowledgements

In the process of graduation thesis research, the most important person to thank is Prof. Yang Xiaohui. Under her guidance, I identified the selected topic, and kept pushing it forward until the completion of the study. As I was exposed to a new direction, I was not very familiar with it at first. When I encountered difficulties, Miss Xiaohui was always able to communicate with me at the first time, formulate a research plan and answer my various weird questions. At the same time, I would like to thank the graduate students supervised by Miss Xiaohui for their advice on code operation, thesis writing and format, so that I can finish this graduation thesis smoothly. The time spent with the teacher and her graduate students is really enjoyable and happy.