frontiers in Genetics

ORIGINAL RESEARCH
published: 13 July 2016
doi: 10.3389/fgene.2016.00123

Analysis of Variance Components for Genetic Markers with Unphased Genotypes

Tao Wang *

Division of Biostatistics, Institute for Health and Society, Medical College of Wisconsin, Milwaukee, WI, USA

An ANOVA type general multi-allele (GMA) model was proposed in Wang (2014) on analysis of variance components for quantitative trait loci or genetic markers with phased or unphased genotypes. In this study, by applying the GMA model, we further examine estimation of the genetic variance components for genetic markers with unphased genotypes based on a random sample from a study population. In one locus and two loci cases, we first derive the least square estimates (LSE) of model parameters in fitting the GMA model. Then we construct estimators of the genetic variance components for one marker locus in a Hardy-Weinberg disequilibrium population and two marker loci in an equilibrium population. Meanwhile, we explore the difference between the classical general linear model (GLM) and GMA based approaches in association analysis of genetic markers with quantitative traits. We show that the GMA model can retain the same partition on the genetic variance components as the traditional Fisher's ANOVA model, while the GLM cannot. We clarify that the standard F-statistics based on the partial reductions in sums of squares from GLM for testing the fixed allelic effects could be inadequate for testing the existence of the variance component when allelic interactions are present. We point out that the GMA model can reduce the confounding between the allelic effects and allelic interactions at least for independent alleles. As a result, the GMA model could be more beneficial than GLM for detecting allelic interactions.

Keywords: Fisher's ANOVA model, analysis of variance, general linear model, general multi-allelic model, genetic variance components, orthogonal partition, allelic interactions, least square estimates

OPEN ACCESS

Edited by:
Steven J. Schrodi,
Marshfield Clinic Research Foundation, USA

Reviewed by:
Zhan Ye,
Marshfield Clinic, USA
Charles M. Rowland,
Quest Diagnostics, USA

****Correspondence:***
Tao Wang
taowang@mcw.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 March 2016*
Accepted: *22 June 2016*
Published: *13 July 2016*

Citation:
Wang T (2016) Analysis of Variance Components for Genetic Markers with Unphased Genotypes. Front. Genet. 7:123. doi: 10.3389/fgene.2016.00123

1. INTRODUCTION

Typically, there are two different ways in assessing the statistical association of a categorical variable with a continuous outcome. We can either make a direct comparison of the group means among groups defined by the categorical variable or assess the variation that the categorical variable may contribute to the total variance of the continuous outcome. The former approach is usually conducted via fitting a general linear model (GLM) with or without an adjustment for other covariates. The latter approach is referred as the analysis of variance (ANOVA), which examines a quantitative outcome variable by partitioning its total variance into variance components attributable to different sources of variation. The original ANOVA model was proposed initially by Fisher (1918) and formalized later in Fisher (1925). The traditional ANOVA approach on estimation of variance components was mainly based on Henderson's method I-III by equating the observed sums of squares to their expected values (Henderson, 1953). Currently, for predictors with observed group levels, their variance components can be estimated via ANOVA tables, which are mostly based on the sequential (Type I) sums of squares or partial reductions in (Type III) sums of squares via fitting GLM (Searle, 1971; Searle et al., 1992). For predictors with observed or unobserved group levels, their variance components could also be evaluated via maximum likelihood estimation (MLE) or restricted maximum likelihood (REML) (see Searle et al., 1992).

In genetic association studies, the standard GLM or ANOVA approaches may not be directly applicable to genetic markers with unphased genotypes. In humans, a marker genotype is a combination of one paternal and one maternal allele. But the phase information (i.e., the origin of parental alleles) is often missing from most of the current genotype typing technologies. Appropriate modification on the classical GLM or ANOVA methods is therefore needed in order to overcome this unknown phase problem. In Wang (2011), we introduced several coding schemes for unphased marker genotypes in constructing the dummy-variable based GLM. In Zeng et al. (2005) and Wang (2014), some revised Fisher's ANOVA models were also proposed for ANOVA analysis of quantitative trait loci or genetic markers with unphased genotypes, which were referred as the general biallelic (G2A) or general multi-allelic (GMA) models. The estimation of genetic variance components based on the G2A and GMA models were also briefly explored.

In this study, by applying the GMA model, we further examine estimation of genetic variance components for genetic markers with unphased genotypes from a random sample of individuals in a study population. First, for a single locus GMA model, we derive the least square estimates (LSE) of model parameters in fitting the GMA model based on the genotypic group means and allele frequency estimates. Than we construct estimators of the variance components for one marker in Hardy-Weinberg disequilibrium (HWD). Next, we consider a fully parameterized two-locus GMA model. We derive the LSE of model parameters in fitting the GMA model and develop estimators of the variance components for two genetic markers in an equilibrium population. In both one locus and two loci cases, we also explore the difference between the GLM and GMA in association analysis of genetic markers. We show that the GMA models can provide the same partition on the genetic variance components as the original Fisher's ANOVA models but the GLM cannot. We clarify that the F-statistics based on the partial reductions in sums of squares from the standard ANOVA table for association testing of the fixed allelic effects in a GLM could be inadequate for testing the existence of variance components when higher order fixed allelic interactions are present. In addition, we point out that the GMA model can reduce the confounding between allelic effects and allelic interactions at least when the inheritance of alleles are independent. As the result, the GMA model could be more beneficial than GLM for detecting allelic interactions. Finally, a simulation example is presented to show the difference between GLM and GMA models on partitioning variance components. The performance in model selection between using GLM and GMA models is also examined.

2. MODELS AND RESULTS

Consider a random sample of size N from a study population. Let y_i, $i = 1, \cdots, N$, be their observed phenotypic values for a quantitative trait Y, and g_i, $i = 1, \cdots, N$, be their observed unphased genotypes at certain genetic marker loci. The quantitative trait Y is assumed to be affected by both genetic and environmental effects. Let G denote the true (unobservable) genotypic value, which could be affected by many genetic factors. If we ignore the genetic by environmental interactions, the relationship between the quantitative trait and marker genotypes can be modeled via a GLM

$$y_i = \beta z_i + \mathrm{E}(G|g_i) + \epsilon_i,\ i = 1, \cdots, N, \tag{1}$$

where z_i is a vector of the adjusted environmental covariates with fixed effects β, $\mathrm{E}(G|g_i)$ represents the expected genotypic value of the i-th individual given his/her observed marker genotypes g_i, and ϵ_i is a model residual error contributed by other environmental and genetic factors that cannot be captured by the covariates z_i and marker genotypes g_i. We also assume that ϵ_i, $i = 1, \cdots, N$, are independent and identically distributed (i.i.d) with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{V}(\epsilon_i) = V_\epsilon$. Besides, $\{\epsilon_i, i = 1, \cdots, N\}$ are independent of $\{g_i, i = 1, \cdots, N\}$. Then, the total phenotypic variance $V_Y = V(\mathrm{E}(G|g)) + \mathrm{E}(V(G|g)) = V(\mathrm{E}(G|g)) + V_\epsilon$. In the rest of the paper, we focus on comparing the GLM and GMA modeling on the expected genotypic values $\mathrm{E}(G|g)$ and assessing the genetic variance components contributed by the allelic effects and allelic interactions at the marker loci to the expected genotypic variance $V(\mathrm{E}(G|g))$.

2.1. One-Locus Models

Consider a single marker locus with multiple alleles $A_1, \ldots, A_m$ $(m \geq 2)$. Let $p_j, j = 1, \cdots, m$, be the allele frequencies $(\sum_{j=1}^m p_j = 1)$, and $p_{jk} = P(A_j A_k), j, k = 1, \cdots, m\ (j \leq k)$, be the genotype frequencies $(\sum_{j \leq k} p_{jk} = 1)$, in a study population. For a random sample of size N from the study population, suppose that there are n_{jk} individuals carrying genotypes $A_j A_k$ for $j \leq k, j, k = 1, \cdots, m$ $(N = \sum_{j=1}^m \sum_{k=j}^m n_{jk})$. For notation simplicity, we also let $p_{kj} = p_{jk}$ and $n_{kj} = n_{jk}$ for $k > j$, $j, k = 1, \cdots, m$. Then, $p_j = p_{jj} + \sum_{k \neq j} p_{jk}/2$, for $j = 1, \cdots, m$. Let $n_{j\cdot} = 2n_{jj} + \sum_{k \neq j} n_{jk}$, which is the total number of alleles A_j carried by the sampled individuals $(j = 1, \cdots, m)$. Assume that the observed marker genotypes g_i, $i = 1, \cdots, N$, are i.i.d. and follow a multinomial distribution. Then the MLE of the genotype frequencies are given by $\hat{p}_{jk} = n_{jk}/N$, and the MLE of the allele frequencies are given by $\hat{p}_j = n_{j\cdot}/(2N) = \hat{p}_{jj} + \sum_{k \neq j} \hat{p}_{jk}/2$, for $j = 1, \cdots, m$. Also, let $y_{jki}, i = 1, \cdots, n_{jk}$ be the observed phenotypic values of the individuals carrying genotypes $A_j A_k$. We define the observed genotypic group means $\bar{y}_{jk\cdot} = \bar{y}_{kj\cdot} = \sum_{i=1}^{n_{jk}} y_{jki}/n_{jk}$ for $j, k = 1, \cdots, m$, the allele averaged means $\bar{y}_{j\cdot\cdot} = \bar{y}_{\cdot j\cdot} = (2n_{jj}\bar{y}_{jj\cdot} + \sum_{k \neq j} n_{jk}\bar{y}_{jk\cdot})/n_{j\cdot}$ for $j = 1, \cdots, m$, and the grand mean $\bar{y}_{\cdots} = \sum_{j=1}^m \sum_{k=j}^m n_{jk}\bar{y}_{jk\cdot}/N$. In addition, we define the allele weighted means $\bar{y}^*_{j\cdot} = \bar{y}^*_{\cdot j} = \sum_{k=1}^m \hat{p}_k \bar{y}_{jk\cdot}$ for $j = 1, \cdots, m$, and the weighted overall mean $\bar{y}^*_{\cdot\cdot} = \sum_{j=1}^m \sum_{k=1}^m \hat{p}_j \hat{p}_k \bar{y}_{jk\cdot} = \sum_{j=1}^m \hat{p}_j^2 \bar{y}_{jj\cdot} + \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\hat{p}_j \hat{p}_k \bar{y}_{jk\cdot}$.

Note that in a classical two-way factorial ANOVA model, each individual can receive one and only one level assignment from each of the two factors. At a marker locus, however, an individuals can carry a homozygous genotype "A_jA_j" ($j = 1, \cdots, m$) with two A_j alleles indistinguishable. In general, the allele weighted means $\bar{y}^*_{j\cdot}$ may not be the same as the allele averaged means $\bar{y}_{j\cdot\cdot}$, and the weighted overall mean $\bar{y}^*_{\cdot\cdot}$ may also differ from the grand mean $\bar{y}_{\cdot\cdot\cdot}$.

To assess the allelic effects on the expected genotypic values (or phenotypic values), let us first take a brief review of Fisher's one-locus ANOVA model (see Kempthorne, 1957; Weir and Cockerham, 1977). Since the marker genotypes are unphased, we usually assume that the paternal and maternal gametes share the same set of alleles, have the same allele frequencies, and contribute the same allelic effects at the marker locus. The fact that each allele contributes the same genetic effect regardless of its parental origins implies that the expected genotypic values $G_{jk} = \mathrm{E}(G|g = A_jA_k)$ should satisfy the symmetric property: $G_{jk} = G_{kj}$, for $j \neq k$. So there are totally $m(m+1)/2$ possible distinctive expected genotypic values G_{jk}, $j, k = 1, \ldots, m$. By treating the paternal and maternal gametes as two independent risk factors, the traditional Fisher's one-locus ANOVA model for the expected genotypic values G_{jk} can be written as

$$G_{jk} = \mu + \alpha_j + \alpha_k + \delta_{jk},\ j, k = 1, \ldots, m, \qquad (2)$$

where α_j is referred as the *average* additive effect of the paternal or maternal allele A_j ($j = 1, \ldots, m$), and δ_{jk} the *average* allelic interaction between two parental alleles A_j and A_k ($j, k = 1, \ldots, m$). It is known that not all the parameters in the above model are estimable due to an over-parameterization of the model on the expected genotypic values. Typically, the following constraints can be added on the model parameters (see Kempthorne, 1957).

$$\sum_{j=1}^{m} p_j\alpha_j = 0,\ \ \sum_{j=1}^{m} p_j\delta_{jk} = 0 \ \text{ for } k = 1, \cdots, m. \qquad (3)$$

From the symmetric property of G_{jk}, we also assume that $\delta_{jk} = \delta_{kj}$, for $j \neq k$. With these constraints, if we further incorporate model (2) into GLM (1) and ignore the adjusted covariates, the LSE of parameters are given by $\hat{\mu} = \bar{y}^*_{\cdot\cdot}$, $\hat{\alpha}_j = \bar{y}^*_{j\cdot} - \bar{y}^*_{\cdot\cdot}$ for $j = 1, \cdots, m$, and $\hat{\delta}_{jk} = \bar{y}_{jk\cdot} - \bar{y}^*_{j\cdot} - \bar{y}^*_{k\cdot} + \bar{y}^*_{\cdot\cdot}$ for $j, k = 1, \cdots, m$. It has been known that, under Hardy-Weinberg equilibrium (HWE), the above model (2) can also provide an orthogonal partition of the expected genotypic variance as $V(\mathrm{E}(G|g)) = V_A + V_D$, where $V_A = 2\sum_j p_j\alpha_j^2$ and $V_D = \sum_{j,k} p_jp_k\delta_{jk}^2$ are the so-called additive and dominance variance components. Weir and Cockerham (1977) also explored model (2) on partitioning the expected genotypic variance in HWD. In practice, as pointed out in Wang (2014), the symmetric property of δ_{jk}'s and the irregular constraints (3) could make it difficult to fit model (2) using standard statistical software especially when the adjusted covariates are involved. Besides, the random variables that are responsible for the additive and dominance variance components are not explicitly defined in model (2).

The expected genotypic values can also be modeled via a classical dummy-variable based GLM. As shown in Wang (2011), we can introduce the following indicator variables to describe the inheritance of the two parental alleles for each individual

$$z_{1j} = \begin{cases} 1, & \text{the inherited paternal allele is } A_j \\ 0, & \text{the inherited paternal allele is not } A_j \end{cases},$$

$$z_{2j} = \begin{cases} 1, & \text{the inherited maternal allele is } A_j \\ 0, & \text{the inherited maternal allele is not } A_j \end{cases}$$

for $j = 1, \ldots, m$. Though z_{1j} and z_{2j} cannot be defined on unphased heterozygous genotypes, we can define the following genotype coding variables for unphased genotypes

$$w_j(g) = z_{1j} + z_{2j} = \begin{cases} 2, & \text{if } g = A_jA_j \\ 1, & \text{if } g = A_jA_j^c \\ 0, & \text{if } g = A_j^cA_j^c \end{cases}$$

for $j = 1, \ldots, m$, and

$$v_{jj}(g) = z_{1j}z_{2j} = \begin{cases} 1, & \text{if } g = A_jA_j \\ 0, & \text{otherwise} \end{cases},$$

$$v_{jk}(g) = z_{1j}z_{2k} + z_{1k}z_{2j} = \begin{cases} 1, & \text{if } g = A_jA_k \\ 0, & \text{otherwise} \end{cases}$$

for $j \neq k$, $j, k = 1, \ldots, m$. Here, A_j^c denotes any other allele rather than A_j. By choosing A_m as a reference allele, we can construct the following GLM

$$\mathrm{E}(G|g_i) = \mu_0 + \sum_{j=1}^{m-1} a_jw_j(g_i) + \sum_{j=1}^{m-1}\sum_{k=j}^{m-1} d_{jk}v_{jk}(g_i), \qquad (4)$$

for $i = 1, \ldots, N$, where a_j is usually referred as the *fixed* additive allelic effect of the paternal or maternal allele A_j, and d_{jk} the *fixed* allelic interaction between two parental alleles A_j and A_k, with respect to the reference allele A_m. In terms of the expected genotypic values, we can show that $\mu_0 = G_{mm}$, $a_j = G_{jm} - G_{mm}$ and $d_{jk} = (G_{jk} - G_{km}) - (G_{jm} - G_{mm})$, for $j = 1, \ldots, m-1$ and $k = j, \ldots, m-1$.

Model (4) provides a full re-parameterization of the $m(m+1)/2$ expected genotypic values. Suppose that there are no empty genotype groups for the observed random sample; i.e., $n_{jk} > 0$ for any $j, k = 1, \cdots, m$. When we incorporate model (4) into GLM (1) and ignore the adjusted covariates, the LSE of the expected genotypic values are given by: $\hat{G}_{jk} = \bar{y}_{jk}$, for $j, k = 1, \cdots, m-1$ and $j < k$. Therefore, the LSE of parameters in model (4) can be easily derived in terms of the observed genotypic group means as $\hat{\mu}_0 = \bar{y}_{mm\cdot}$ and $\hat{a}_j = \bar{y}_{jm\cdot} - \bar{y}_{mm\cdot}$ for $j = 1, \cdots, m-1$, and $\hat{d}_{jk} = \bar{y}_{jk\cdot} - \bar{y}_{jm\cdot} - \bar{y}_{km\cdot} + \bar{y}_{mm\cdot}$ for $j, k = 1, \cdots, m-1$ and $j \leq k$. Note that these LSE could be sensitive to phenotypic outliers especially for small genotypic groups. For example, a few individuals may have genotype "A_jA_m" (j $\neq m$) for a rare allele "A_j." If the genotypic group mean $\bar{y}_{jm\cdot}$ is much larger than $\bar{y}_{mm\cdot}$, the LSE $\hat{a}_j$ will be large. By choosing a common allele "A_m" as the reference, we could improve the accuracy of LSE. Through incorporating model (4) into a GLM, we could also perform hypothesis tests on its model parameters via contrasts. In analysis of genetic variance components, however, it has been known that w_j's are often correlated with v_{jk}'s as random variables over the random individuals even when the inheritance of paternal and maternal alleles are independent (Wang and Zeng, 2009). As the result, model (4) leads to a different partition of the expected genotypic variance $V(\mathrm{E}(G|g))$ from the one defined in Fisher's ANOVA model (2).

To further dissect the confounding between the additive and dominance effects in model (4), we can make mean corrections on the indicator variables $\{z_{1i}\}$ and $\{z_{2j}\}$, and introduce the following mean-corrected index variables (see Wang, 2014)

$$x_{1j}(g) = z_{1j}(g) - \mathrm{E}[z_{1j}(g)] = \begin{cases} 1 - p_j, & \text{the paternal allele is } A_j \\ -p_j, & \text{the paternal allele is not } A_j \end{cases}$$

$$x_{2j}(g) = z_{2j}(g) - \mathrm{E}[z_{2j}(g)] = \begin{cases} 1 - p_j, & \text{the maternal allele is } A_j \\ -p_j, & \text{the maternal allele is not } A_j \end{cases}$$

Then, we define the following modified genotype coding variables

$$w_j^*(g) = x_{1j} + x_{2j} = \begin{cases} 2(1-p_j), & \text{if } g = A_jA_j \\ 1 - 2p_j, & \text{if } g = A_jA_j^c \\ -2p_j, & \text{if } g = A_j^cA_j^c \end{cases},$$

$$v_{jj}^*(g) = x_{1j}x_{2j} = \begin{cases} (1-p_j)^2, & \text{if } g = A_jA_j \\ -p_j(1-p_j), & \text{if } g = A_jA_j^c \\ p_j^2, & \text{if } g = A_j^cA_j^c \end{cases}$$

for $j = 1, \cdots, m$, and

$$v_{jk}^*(g) = x_{1j}x_{2k} + x_{1k}x_{2j} = \begin{cases} (1-p_j)(1-p_k) + p_jp_k, \ \text{if } g = A_jA_k \\ -2p_k(1-p_j), \ \text{if } g = A_jA_j \\ -2p_j(1-p_k), \ \text{if } g = A_kA_k \\ -p_k(1-2p_j), \ \text{if } g = A_jA_{jk}^c \\ -p_j(1-2p_k), \ \text{if } g = A_kA_{jk}^c \\ 2p_ip_j, \ \text{if } g = A_{jk}^cA_{jk}^c \end{cases}$$

for $j \neq k, j, k = 1, \ldots, m$. Here, A_{jk}^c denotes an allele which is different from both A_j and A_k. Note that the modified genotype coding variables $w_j^*(g) = w_j(g) - 2p_j$, $v_{jj}^*(g) = v_{jj}(g) - p_jw_j(g) + p_j^2$, and $v_{jk}^*(g) = v_{jk}(g) - p_jw_k(g) - p_kw_j(g) + 2p_jp_k$ are well defined on unphased genotypes, although the mean-corrected index variables x_{1j}, x_{2k} cannot be defined on unphased heterozygous genotypes. By choosing A_m as a reference allele, we can construct the following GMA model (Wang, 2014)

$$\mathrm{E}(G|g_i) = \mu^* + \sum_{j=1}^{m-1} \alpha_j^* w_j^*(g_i) + \sum_{j=1}^{m-1}\sum_{k=j}^{m-1} \delta_{jk}^* v_{jk}^*(g_i), \qquad (5)$$

for $i = 1, \ldots, N$. Still, in model (5), we refer the parameters α_j^* as the *average* additive allelic effects and δ_{jk}^* ($j \leq k$) as the *average* allelic interactions, with respect to the reference allele A_m. Both the GMA model (5) and the GLM (4) can provide a full parameterization of the expected genotypic values G_{jk}. Comparing to model (4), one major advantage of model (5) is that it can retain the same partition on the genetic variance components as the one from Fisher's ANOVA model (2). In fact, under the constraints (3) plus the symmetric property of δ_{jk}, model (2) can be re-written as

$$\mathrm{E}(G|g_i) = \mu + \sum_{j=1}^{m} \alpha_j w_j^*(g_i) + \sum_{j=1}^{m}\sum_{k=j}^{m} \delta_{jk} v_{jk}^*(g_i), \qquad (6)$$

for $i = 1, \ldots, N$. Model (5) is a simplified version of model (6) by further replacing the redundant variables w_m^*, v_{jm}^* and v_{mm}^* by $w_m^* = -\sum_{j=1}^{m-1} w_j^*$, $v_{jm}^* = -v_{jj}^* - \sum_{k=1}^{m-1} v_{jk}^*$ for $j = 1, \cdots, m-1$, and $v_{mm}^* = \sum_{j=1}^{m-1}\sum_{k=j}^{m-1} v_{jk}^*$. We can see that both models (5) and (2) share exactly the same additive and dominance variance components. They become equivalent when we take $\mu^* = \mu$, $\alpha_j^* = \alpha_j - \alpha_m$, and $\delta_{jk}^* = \delta_{jk} - \delta_{jm} - \delta_{km} + \delta_{mm}$, for $j, k = 1, \cdots, m-1$ and $j \leq k$. Note that model (5) does not contain redundant parameters. Therefore, it does not require constraints on its model parameters. Besides, the random variables that are responsible for the additive and dominance variance components are explicitly defined in model (5). In practice, similar to enforcing the constraints (3) on model (2), we can create the variables w_j^* and v_{jk}^* by replacing the allele frequencies p_j's by their MLE $\hat{p}_j$'s. Then, by incorporating model (5) into GLM (1), we can treat these modified genotype coding variables as regular fixed covariates and fit the model using the ordinary LS approach. The hypothesis of $H_0: V_A = 0$ for existence of the additive variance component can also be tested via testing $H_0: \alpha_j^* = 0, j = 1, \cdots, m-1$ for the average additive allelic effects.

The GLM (4) and GMA model (5) can be transformed easily from one to the other. From the relationship between their genotype coding variables, we can show that

$$\begin{cases} \mu^* = \gamma + 2\sum_{j=1}^{m-1} p_ja_j + \sum_{j=1}^{m-1}\sum_{k=1}^{m-1} p_jp_kd_{jk} \\ \alpha_j^* = a_j + \sum_{k=1}^{m-1} p_kd_{jk} \ , \ j = 1, \cdots, m-1 \end{cases}$$

and $\delta_{jk}^* = d_{jk}$ for $j \leq k, j, k = 1, \cdots, m-1$. Here, for notation simplicity, we define $d_{kj} = d_{jk}$, for $j < k$. To fit model (5), instead of solving its normal equations directly, we can derive the LSE of its model parameters from the LSE of model (4) as the following.

$$\begin{cases} \hat{\mu}^* = \sum_{j=1}^{m} \hat{p}_j^2\bar{y}_{jj\cdot} + \sum_{j=1}^{m-1}\sum_{k=j+1}^{m} 2\hat{p}_j\hat{p}_k\bar{y}_{jk\cdot} = \bar{y}_{\cdot\cdot}^* \\ \hat{\alpha}_j^* = \bar{y}_{j\cdot}^* - \bar{y}_{m\cdot}^* \ , \ j = 1, \cdots, m-1 \end{cases}$$

and $\hat{\delta}_{jk}^* = \hat{d}_{jk} = \bar{y}_{jk\cdot} - \bar{y}_{jm\cdot} - \bar{y}_{km\cdot} + \bar{y}_{mm\cdot}$ for $j \leq k, j, k = 1, \cdots, m-1$. Note that the LSE of the average additive allelic effects are calculated from allele weighted means, which could be more robust to phenotypic outliers than the LSE of fixed additive allelic effects. It is also interesting to see that both the fixed additive allelic effects a_j's and the fixed allelic interactions d_{jk}'s may affect the average additive allelic effects α_j^*'s, though the fixed allelic interactions keep the same as the average allelic interactions. In general, the hypothesis test of $H_0 : a_j = 0, j = 1, \cdots, m-1$, for the fixed additive effects in a GLM (4) is not equivalent to the hypothesis test of $H_0 : \alpha_j^* = 0, j = 1, \cdots, m-1$, for the average additive effects in an equivalent GMA model (5). Therefore, the standard F-statistics for testing the fixed additive effects in model (4) could be inadequate for testing the existence of the additive variance component V_A when significant fixed (or average) allelic interactions are present.

In ANOVA, we treat $\{x_{1j}, j = 1, \cdots, m-1\}$ and $\{x_{2k}, k = 1, \cdots, m-1\}$ as random variables over the sampled individuals. Model (5) provides an approximation to the expected genotypic values $\mathrm{E}(G|g)$ by a linear combination of the random variables $x_{1j}, j = 1, \cdots, m-1$ and $x_{2k}, k = 1, \cdots, m-1$, plus their cross-product terms for allelic interactions. To model the genotypic distributions, let $g_i^* = (z_{1j}(g_i)z_{2k}(g_i), j, k = 1, \cdots, m, j \leq k)$ be a vector that describes the unphased genotypic categories for $i = 1, \cdots, N$. As usual, we assume that the marker genotypes $g_1^*, \cdots, g_N^*$ are i.i.d. from a multinomial distribution of $Mult\left(1, \{p_{jk}, j, k = 1, \cdots, m, j \leq k\}\right)$. Under this assumption, we can show that the vectors $g_{1i} = (z_{11}(g_i), \cdots, z_{1m}(g_i)), i = 1, \cdots, N$, of paternal allele indicator variables are i.i.d. from $Mult(1, \{p_1, \cdots, p_m\})$; and the vectors $g_{2i} = (z_{21}(g_i), \cdots, z_{2m}(g_i)), i = 1, \cdots, N$, of maternal allele indicator variables are also i.i.d. from $Mult(1, \{p_1, \cdots, p_m\})$. When the inheritance of paternal and maternal alleles are independent, the study population is under HWE and the mean corrected index variables $\{x_{1j}(g_i), j = 1, \cdots, m\}$ are independent of $\{x_{2j}(g_i), j = 1, \cdots, m\}$ for any randomly sampled individual i. In HWD, however, the mean corrected index variables $\{x_{1j}, j = 1, \cdots, m\}$ could be correlated with $\{x_{2j}, j = 1, \cdots, m\}$. Let us define

$$\begin{cases} D_{jj} = \mathrm{Cov}(z_{1j}, z_{2j}) = \mathrm{E}(x_{1j}x_{2j}) = \mathrm{E}(v_{jj}^*) = p_{jj} - p_j^2 \\ D_{jk} = \mathrm{E}(x_{1j}x_{2k} + x_{1k}x_{2j})/2 = \mathrm{E}(v_{jk}^*)/2 = p_{jk}/2 - p_j p_k. \end{cases}$$

for $j \neq k, j, k = 1, \cdots, m$. Then, $\{D_{jk}, j, k = 1, \cdots, m\}$ measure the departures from HWE, which satisfy $\sum_{j=1}^m D_{jk} = \sum_{k=1}^m D_{jk} = 0$. From the observed genotypes, we have MLE $\hat{D}_{jj} = n_{jj}/N - \hat{p}_j^2$ for $j = 1, \cdots, m$, and $\hat{D}_{jk} = n_{jk}/(2N) - \hat{p}_j\hat{p}_k$ for $j, k = 1, \cdots, m$ and $j \neq k$.

It is more convenient to use GMA model (5) rather than GLM (4) for estimation of genetic variance components. In model (5), let $A = \sum_{j=1}^{m-1} \alpha_j^* w_j^*(g)$ and $D = \sum_{j=1}^{m-1}\sum_{k=j}^{m-1} \delta_{jk}^* v_{jk}^*(g)$ represent the additive and dominance components, respectively. In general, we can partition the expected genotypic variance as $V(\mathrm{E}(G|g)) = V_A + V_D + 2\mathrm{Cov}(A, D)$. In Wang (2014), we have derived formulas for estimating the variance components V_A, V_D and covariance component $\mathrm{Cov}(A, D)$ based on the parameters in model (5). By plugging in LSE of the parameters, we obtain estimators of the variance components V_A, V_D and the covariance component $\mathrm{Cov}(A, D)$ as shown in Appendix A in Supplementary Material. It should be pointed out that these estimators of the variance and covariance components are different from the traditional ANOVA estimators when the marker genotypes are in HWD. Unlike the ANOVA estimators of variance components which could be negative when data are unbalanced, our estimators of the variance components are guaranteed to be non-negative. Similar to the ANOVA estimators, when the model residuals are normally distributed, these estimators become MLE of the variance and covariance components and likely possess the asymptotic normality property (see Searle, 1995). Meanwhile, with variations from both the genotypic group means and the MLE estimates of allele frequencies, these estimators are only asymptotically unbiased, while the original ANOVA estimators are unbiased.

Under HWE, we would expect an orthogonal partition of the expected genotypic variance as $V(\mathrm{E}(G|g)) = V_A + V_D$. By taking $\hat{D}_{jk} = 0$ for $j, k = 1, \cdots, m$ in Appendix A in Supplementary Material, we obtain the following estimators

$$\widehat{V}_A = 2\sum_{j=1}^{m} \hat{p}_j(\bar{y}_{j\cdot}^* - \bar{y}_{\cdot\cdot}^*)^2,$$

$$\widehat{V}_D = \sum_{j=1}^{m}\sum_{k=1}^{m} \hat{p}_j\hat{p}_k(\bar{y}_{jk\cdot} - \bar{y}_{\cdot\cdot}^*)^2 - 2\sum_{j=1}^{m} \hat{p}_j(\bar{y}_{j\cdot}^* - \bar{y}_{\cdot\cdot}^*)^2$$

If we further incorporate model (5) into GLM (1) and ignore the adjusted covariates, the estimator of residual variance is given by $\widehat{V}_\epsilon = \sum_{j=1}^m \sum_{k=j}^m \sum_{i=1}^{n_{jk}} (y_{jki} - \bar{y}_{jk\cdot})^2/N$. The estimator of total phenotypic variance is given by $\widehat{V}_Y = \sum_{j=1}^m \sum_{k=j}^m \sum_{i=1}^{n_{jk}} (y_{jki} - \bar{y}_{\cdots})^2/N$. We can show that $\widehat{V}_Y = \widehat{V}_A + \widehat{V}_D + \widehat{V}_\epsilon$. Note that when $\hat{D}_{jk} = 0$ for $j, k = 1, \cdots, m$, the allele weighted means and the allele averaged means become the same; i.e., $\bar{y}_{j\cdot}^* = \bar{y}_{j\cdot\cdot}$ for $j = 1, \cdots, m$. Besides, the weighted overall mean $\bar{y}_{\cdot\cdot}^*$ is the same as the grand mean $\hat{y}_{\cdots}$. However, due to possible sampling variations from the sampled individuals, under HWE we could still have some $\hat{D}_{jk} \neq 0, j, k = 1, \cdots, m$, which may lead to a slight deviation from the orthogonal partition of the expected genotypic variance.

When $\hat{D}_{jk} = 0$ for $j, k = 1, \cdots, m$, another nice feature of the GMA model (5) is that the LSE $\{\hat{\alpha}_j^*, j = 1, \cdots, m-1\}$ of its main effects will keep the same in a reduced GMA model when we ignore the allelic interactions. Let us represent the full GMA model (5) in a matrix form with the design matrix $X = (1_N, X_{\alpha^*}, X_{\delta^*})$, where 1_N is a N by 1 vector with all its elements being 1, $X_{\alpha^*} = (w_1^*(g_i), \cdots, w_{m-1}^*(g_i); i = 1, \cdots, N)_{N\times(m-1)}$ and $X_{\delta^*} = (v_{11}^*(g_i), \cdots, v_{1,m-1}^*(g_i), \cdots, v_{m-1,m-1}^*(g_i); i = 1, \cdots, N)_{N\times m(m-1)/2}$, which correspond to the main effects $\alpha^* = (\alpha_1^*, \cdots, \alpha_{m-1}^*)$ and allelic interactions $\delta^* = (\delta_{11}^*, \cdots, \delta_{1,m-1}^*, \cdots, \delta_{m-1,m-1}^*)$, respectively. We can show that $X'X = diag(N, X'_{\alpha^*}X_{\alpha^*}, X'_{\delta^*}X_{\delta^*})$, which is a block diagonal matrix when $\hat{D}_{jk} = 0$, for $j, k = 1, \cdots, m$ (see proof in Appendix B in Supplementary Material). Therefore, the LSE $\hat{\mu}^* = \bar{y}_{\cdots}$ and $\hat{\alpha}^* = (X'_{\alpha^*}X_{\alpha^*})^{-1}X'_{\alpha^*}Y$, which do not depend on X_{δ^*}. In other words, ignoring the average allelic interactions in model (5) does not affect the LSE of the intercept and the average additive effects in this case. The GLM (4) does not have such a property. In a reduced GLM (4) without the fixed allelic interactions, the LSE of its additive effects are $\hat{a}_j = \hat{\alpha}_j^* = \bar{y}_{j\cdot}^* - \bar{y}_{m\cdot}^*$ for $j = 1, \cdots, m-1$, while in a full GLM (4) the LSE become $\hat{a}_j = \bar{y}_{jm\cdot} - \bar{y}_{mm\cdot}$, for $j = 1, \cdots, m-1$.

We can also estimate the genetic variance components with an adjustment for other covariates. First, by incorporating the GMA model (5) into GLM (1), we can fit GLM (1) using the ordinary LS approach. Next, based on the fitted model, we can calculate the fitted additive and dominance components $\hat{A}_i = \sum_{j=1}^{m-1} \hat{\alpha}_j^* w_i(g_i)$ and $\hat{D}_i = \sum_{j=1}^{m-1} \sum_{k=j}^{m-1} \hat{\delta}_{jk}^* v_{jk}(g_i)$, for each individual $i = 1, \cdots, N$. Then, we can estimate V_A and V_D as the sample variances of $\{\hat{A}_i, i = 1, \cdots, N\}$ and $\{\hat{D}_i, i = 1, \cdots, N\}$, respectively. Meanwhile, the covariance component $\mathrm{Cov}(A, D)$ can be estimated as the sample covariance between $\{\hat{A}_i, i = 1, \cdots, N\}$ and $\{\hat{D}_i, i = 1, \cdots, N\}$.

2.2. Two-Locus Models

We consider an extension of the previous one-locus models to two loci. Assume that marker 1 has alleles $A_{11}, \ldots, A_{1m_1}$ ($m_1 \geq 2$) with $p_{11}, \cdots, p_{1m_1}$ being the allele frequencies, and marker 2 has alleles $A_{21}, \ldots, A_{2m_2}$ ($m_2 \geq 2$) with $p_{21}, \cdots, p_{2m_2}$ being the allele frequencies. Let $p_{jkrs} = P(A_{1j}A_{1k}, A_{2r}A_{2s})$ denote the joint genotype frequencies at the two marker loci in a study population. For a random sample of size N from the study population, let n_{jkrs} be the number of individuals who carry unphased genotypes $A_{1j}A_{1k}$ ($j \leq k$) at locus 1, and $A_{2r}A_{2s}$ ($r \leq s$) at locus 2, for $j, k = 1, \cdots, m_1$ and $r, s = 1, \cdots, m_2$ ($N = \sum_{j=1}^{m_1} \sum_{k=j}^{m_1} \sum_{r=1}^{m_2} \sum_{s=r}^{m_2} n_{jkrs}$). Then, the MLE of genotype frequencies $\hat{p}_{jkrs} = n_{jkrs}/N$. Let $y_{jkrs,i}, i = 1, \cdots, n_{jkrs}$, be the observed phenotypic values of individuals carrying the joint genotypes $(A_{1j}A_{1k}, A_{2r}A_{2s})$. We define the observed genotypic group means $\bar{y}_{jkrs\cdot} = \sum_{i=1}^{n_{jkrs}} y_{jkrs,i}/n_{jkrs}$, for $j, k = 1, \cdots, m_1$ and $r, s = 1, \ldots, m_2$. Without distinguishing the origin of parental alleles, we assume that $n_{kjrs} = n_{jkrs}$ for $k > j, j, k = 1, \cdots, m_1$, and $n_{jksr} = n_{jkrs}$ for $r > s, r, s = 1, \cdots, m_2$. Let $n_{jk\cdot\cdot} = \sum_{r=1}^{m_2} \sum_{s=r}^{m_2} n_{jkrs}$ for $j, k = 1, \cdots, m_1$, and $n_{\cdot\cdot rs} = \sum_{j=1}^{m_1} \sum_{k=j}^{m_1} n_{jkrs}$ for $r, s = 1, \cdots, m_2$. We have the MLE of allele frequencies $\hat{p}_{1j} = (2n_{jj\cdot\cdot} + \sum_{k \neq j} n_{jk\cdot\cdot})/(2N)$ for $j = 1, \cdots, m_1$ at locus 1, and $\hat{p}_{2r} = (2n_{\cdot\cdot rr} + \sum_{s \neq r} n_{\cdot\cdot rs})/(2N)$ for $r = 1, \cdots, m_2$ at locus 2. For notation simplicity, we define $\bar{y}_{kjrs\cdot} = \bar{y}_{jkrs\cdot}$ for $k > j$ ($j, k = 1, \cdots, m_1$) and $\bar{y}_{jksr\cdot} = \bar{y}_{jkrs\cdot}$ for $r > s$ ($r, s = 1, \cdots, m_2$). In addition, we define the weighted genotypic group means $\bar{y}^*_{j\cdot rs} = \bar{y}^*_{\cdot jrs} = \sum_{k=1}^{m_1} \hat{p}_{1k}\bar{y}_{jkrs}$, $\bar{y}^*_{jkr\cdot} = \bar{y}^*_{jk\cdot r} = \sum_{s=1}^{m_2} \hat{p}_{2s}\bar{y}_{jkrs}$, $\bar{y}^*_{jk\cdot\cdot} = \bar{y}^*_{kj\cdot\cdot} = \sum_{r,s=1}^{m_2} \hat{p}_{2r}\hat{p}_{2s}\bar{y}_{jkrs}$, $\bar{y}^*_{\cdot\cdot rs} = \bar{y}^*_{\cdot\cdot sr} = \sum_{j,k=1}^{m_1} \hat{p}_{1j}\hat{p}_{1k}\bar{y}_{jkrs}$, $\bar{y}^*_{j\cdots} = \bar{y}^*_{\cdot j\cdot\cdot} = \sum_{k=1}^{m_1} \sum_{r,s=1}^{m_2} \hat{p}_{1k}\hat{p}_{2r}\hat{p}_{2s}\bar{y}_{jkrs}$, and $\bar{y}^*_{\cdot\cdot r\cdot} = \bar{y}^*_{\cdots r} = \sum_{j,k=1}^{m_1} \sum_{s=1}^{m_2} \hat{p}_{1j}\hat{p}_{1k}\hat{p}_{2s}\bar{y}_{jkrs}$, for $j, k = 1, \cdots, m_1$ and $r, s = 1, \cdots, m_2$. The weighted overall mean $\bar{y}^*_{\cdots\cdot} = \sum_{j,k=1}^{m_1} \sum_{r,s=1}^{m_2} \hat{p}_{1j}\hat{p}_{1k}\hat{p}_{2r}\hat{p}_{2s}\bar{y}_{jkrs}$.

Let $G_{jkrs} = \mathrm{E}(G(g)|g = A_{1j}A_{1k}, A_{2r}A_{2s})$ be the expected genotypic value of individuals with the joint genotypes $(A_{1j}A_{1k}, A_{2r}A_{2s})$, for $j, k = 1, \ldots, m_1$ and $r, s = 1, \ldots, m_2$. Without distinguishing the origin of parental alleles, we assume that $\{G_{jkrs}\}$ satisfy the symmetric properties: $G_{jkrs} = G_{kjrs} = G_{jksr} = G_{kjsr}$, for $j < k$ and $r < s$. In general, there are totally $m_1m_2(m_1+1)(m_2+1)/4$ possible distinctive expected genotypic values. By treating the paternal and maternal gametes as two independent risk factors, the Fisher's two-locus ANOVA model for the expected genotypic values G_{jkrs} can be written as (see Kempthorne, 1957)

$$\begin{aligned} G_{jkrs} = {} & \mu + \alpha_{1j} + \alpha_{1k} + \delta_{1jk} + \alpha_{2r} + \alpha_{2s} + \delta_{2rs} + (\alpha\alpha)_{jr} \\ & + (\alpha\alpha)_{js} + (\alpha\alpha)_{kr} + (\alpha\alpha)_{ks} + (\alpha\delta)_{j,rs} + (\alpha\delta)_{k,rs} \\ & + (\delta\alpha)_{jk,r} + (\delta\alpha)_{jk,s} + (\delta\delta)_{jk,rs} \end{aligned} \tag{7}$$

for $j, k = 1, \ldots, m_1$ and $r, s = 1, \ldots, m_2$. The parameters α_{1j} and δ_{1jk} are referred as the average additive and dominance effects at locus 1, α_{2r} and δ_{2rs} the average additive and dominance effects at locus 2, $(\alpha\alpha)_{jr}$ the additive by additive interactions, $(\alpha\delta)_{j,rs}$ the additive by dominance interactions, $(\delta\alpha)_{jk,r}$ the dominance by additive interactions, and $(\delta\delta)_{jk,rs}$ the dominance by dominance interactions. Still, due to an over-parameterization of the model for the expected genotypic values, not all the model parameters are estimable. From the symmetric property of the expected genotypic values, we usually assume that $\delta_{1jk} = \delta_{1kj}$, $\delta_{2rs} = \delta_{2sr}$, $(\alpha\delta)_{j,rs} = (\alpha\delta)_{j,sr}$, $(\delta\alpha)_{jk,r} = (\delta\alpha)_{kj,r}$, $(\delta\delta)_{jk,rs} = (\delta\delta)_{kj,rs} = (\delta\delta)_{jk,sr}$. In addition, the following constraints need to be added on the model parameters (Gallais, 1974; Weir and Cockerham, 1977).

$$\begin{aligned} & \sum_{j=1}^{m_1} p_{1j}\alpha_{1j} = 0, \ \sum_{r=1}^{m_2} p_{2r}\alpha_{2r} = 0, \ \sum_{j=1}^{m_1} p_{1j}\delta_{1jk} = 0, \ \sum_{r=1}^{m_2} p_{2r}\delta_{2rs} = 0, \\ & \sum_{j=1}^{m_1} p_{1j}(\alpha\alpha)_{jr} = 0, \ \sum_{r=1}^{m_2} p_{2r}(\alpha\alpha)_{jr} = 0, \ \sum_{j=1}^{m_1} p_{1j}(\alpha\delta)_{j,rs} = 0, \\ & \sum_{r=1}^{m_2} p_{2r}(\alpha\delta)_{j,rs} = 0, \sum_{j=1}^{m_1} p_{1j}(\delta\alpha)_{jk,r} = 0, \ \sum_{r=1}^{m_2} p_{2r}(\delta\alpha)_{jk,r} = 0, \\ & \sum_{j=1}^{m_1} p_{1j}(\delta\delta)_{jk,rs} = 0, \sum_{r=1}^{m_2} p_{2r}(\delta\delta)_{jk,rs} = 0 \end{aligned} \tag{8}$$

In other words, a weighted sum of the average genetic effects of a genetic component is zero over any index. It has been known that model (7) under constraints (8) can provide an orthogonal partition on the genetic variance components when the inheritance of four paternal and maternal alleles at the two loci are independent (Kempthorne, 1957; Weir and Cockerham, 1977). Still, as pointed out in Wang (2014), it is difficult to estimate the parameters in model (7) under the complicated constraints (8) when other adjusted covariates are involved. Besides, the random variables that constitute the random sources of the genetic variance components are not explicitly defined in model (7).

To model the expected genotypic values through a GLM, we can re-list all the sampled individuals by an index variable $k = 1, \cdots, N$. Similar to the one-locus case, we introduce the indicator variables $z_{1j}^{(1)}$ and $z_{2k}^{(1)}$ for inheritance of the paternal and maternal alleles at locus 1, and $z_{1r}^{(2)}$ and $z_{2s}^{(2)}$ for inheritance of the paternal and maternal alleles at locus 2. Then, for phase-unknown genotypes, we define the genotype coding variables w_{1j}, v_{1jk} for $j, k = 1, \cdots, m_1$ ($j \leq k$) at locus 1, and w_{2r}, v_{2rs} for $r, s = 1, \cdots, m_2$ ($r \leq s$) at locus 2, in the same way as we did in the one-locus case. By including the within locus additive and dominance effects as well as the locus-by-locus interactions (i.e., epistases), a fully parameterized two-locus GLM model can be written as (Wang, 2011),

$$
\begin{aligned}
\mathrm{E}(G|g_i) = \mu_0 &+ \sum_{j=1}^{m_1-1} a_{1j} w_{1j} + \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1} d_{1jk} v_{1jk} + \sum_{r=1}^{m_2-1} a_{2r} w_{2r} \\
&+ \sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} d_{2rs} v_{2rs} + \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} (aa)_{jr} w_{1j} w_{2r} \\
&+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (ad)_{j,rs} w_{1j} v_{2rs} \\
&+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1} (da)_{jk,r} v_{1jk} w_{2r} \\
&+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (dd)_{jk,rs} v_{1jk} v_{2rs} \qquad (9)
\end{aligned}
$$

for $i = 1, \ldots, N$. A nice feature of the above GLM is that we can easily establish the relationship between its model parameters and the expected genotypic values by starting from the lowest order parameter $\mu_0 = G_{m_1 m_1 m_2 m_2}$. Suppose that there are no empty genotypes; i.e., $n_{jkrs} > 0$ for any $j, k = 1, \cdots, m_1$ and $r, s = 1, \cdots, m_2$. When we incorporate the above model (9) into a GLM (1) and ignore the adjusted covariates, the LSE of parameters in model (9) can be derived as shown in Appendix C in Supplementary Material. Similar to the one-locus GLM, the LSE of these fixed allelic effects are highly dependent upon the genotypic group means, which could be sensitive to phenotypic outliers in small genotypic groups. More importantly, model (9) cannot provide the same partition of the expected genotypic variance $V(\mathrm{E}(G|g))$ as the one defined in the original Fisher's ANOVA model (7).

To construct a two-locus GMA model for analysis of genetic variance components, we can introduce the mean-corrected index variables $x_{1j}^{(1)}, x_{2k}^{(1)}$ for $j, k = 1, \cdots, m_1$ at locus 1, and $x_{1r}^{(2)}, x_{2s}^{(2)}$ for $r, s = 1, \cdots, m_2$ at locus 2. Then we define the modified genotype coding variables w_{1j}^*, v_{1jk}^* for $j, k = 1, \cdots, m_1$ at locus 1, and w_{2r}^*, v_{2rs}^* for $r, s = 1, \cdots, m_2$ at locus 2 in the same way as we did in the one-locus case. By including the within locus additive and dominance effects as well as the locus-by-locus interactions (i.e., epistases), we can build a fully parameterized two-locus GMA model as the following.

$$
\begin{aligned}
\mathrm{E}(G|g_i) = \mu^* &+ \sum_{j=1}^{m_1-1} \alpha_{1j}^* w_{1j}^* + \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1} \delta_{1jk}^* v_{1jk}^* + \sum_{r=1}^{m_2-1} \alpha_{2r}^* w_{2r}^* \\
&+ \sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} \delta_{2rs}^* v_{2rs}^* + \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} (\alpha\alpha)_{jr}^* w_{1j}^* w_{2r}^* \\
&+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (\alpha\delta)_{j,rs}^* w_{1j}^* v_{2rs}^* \\
&+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1} (\delta\alpha)_{jk,r}^* v_{1jk}^* w_{2r}^* \\
&+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (\delta\delta)_{jk,rs}^* v_{1jk}^* v_{2rs}^* \qquad (10)
\end{aligned}
$$

for $i = 1, \ldots, N$. Similar to the one-locus case, the original Fisher's ANOVA model (7) under constraints (8) and the symmetric properties of the dominance effects can be re-written as

$$
\begin{aligned}
\mathrm{E}(G|g_i) = \mu &+ \sum_{j=1}^{m_1} \alpha_{1j} w_{1j}^* + \sum_{j=1}^{m_1}\sum_{k=j}^{m_1} \delta_{1jk} v_{1jk}^* + \sum_{r=1}^{m_2} \alpha_{2r} w_{2r}^* \\
&+ \sum_{r=1}^{m_2}\sum_{s=r}^{m_2} \delta_{2rs} v_{2rs}^* + \sum_{j=1}^{m_1}\sum_{r=1}^{m_2} (\alpha\alpha)_{jr} w_{1j}^* w_{2r}^* \\
&+ \sum_{j=1}^{m_1}\sum_{r=1}^{m_2}\sum_{s=r}^{m_2} (\alpha\delta)_{j,rs} w_{1j}^* v_{2rs}^* \\
&+ \sum_{j=1}^{m_1}\sum_{k=j}^{m_1}\sum_{r=1}^{m_2} (\delta\alpha)_{jk,r} v_{1jk}^* w_{2r}^* \\
&+ \sum_{j=1}^{m_1}\sum_{k=j}^{m_1}\sum_{r=1}^{m_2}\sum_{s=r}^{m_2} (\delta\delta)_{jk,rs} v_{1jk}^* v_{2rs}^*. \qquad (11)
\end{aligned}
$$

Model (10) is actually a simplified version of model (11) by further removing the redundant parameters. The two models (10) and (11) are equivalent when we take

$$
\begin{aligned}
\mu^* &= \mu, \;\; \alpha_{1j}^* = \alpha_{1j} - \alpha_{1m_1}, \;\; \alpha_{2k}^* = \alpha_{2k} - \alpha_{2m_2} \\
\delta_{1jk}^* &= \delta_{1jk} - \delta_{1jm_1} - \delta_{1km_1} + \delta_{m_1 m_1} \\
\delta_{2rs}^* &= \delta_{2rs} - \delta_{2rm_2} - \delta_{2sm_2} + \delta_{m_2 m_2} \\
(\alpha\alpha)_{jr}^* &= (\alpha\alpha)_{jr} - (\alpha\alpha)_{jm_2} - (\alpha\alpha)_{m_1 r} + (\alpha\alpha)_{m_1 m_2} \\
(\alpha\delta)_{j,rs}^* &= [(\alpha\delta)_{j,rs} - (\alpha\delta)_{j,rm_2} - (\alpha\delta)_{j,sm_2} + (\alpha\delta)_{j,m_2 m_2}] \\
&- [(\alpha\delta)_{m_1,rs} - (\alpha\delta)_{m_1,rm_2} - (\alpha\delta)_{m_1,sm_2} + (\alpha\delta)_{m_1,m_2 m_2}] \\
(\delta\alpha)_{jk,r}^* &= [(\delta\alpha)_{jk,r} - (\delta\alpha)_{jm_1,r} - (\delta\alpha)_{km_1,r} + (\delta\alpha)_{m_1 m_1,r}] \\
&- [(\delta\alpha)_{jk,m_2} - (\delta\alpha)_{jm_1,m_2} - (\delta\alpha)_{km_1,m_2} + (\delta\alpha)_{m_1 m_1,m_2}] \\
(\delta\delta)_{jk,rs}^* &= [(\delta\delta)_{jk,rs} - (\delta\delta)_{jk,rm_2} - (\delta\delta)_{jk,sm_2} + (\delta\delta)_{jk,m_2 m_2}] \\
&- [(\delta\delta)_{jm_1,rs} - (\delta\delta)_{jm_1,rm_2} - (\delta\delta)_{jm_1,sm_2} + (\delta\delta)_{jm_1,m_2 m_2}] \\
&- [(\delta\delta)_{km_1,rs} - (\delta\delta)_{km_1,rm_2} - (\delta\delta)_{km_1,sm_2} + (\delta\delta)_{km_1,m_2 m_2}] \\
&+ [(\delta\delta)_{m_1 m_1,rs} - (\delta\delta)_{m_1 m_1,rm_2} - (\delta\delta)_{m_1 m_1,sm_2} \\
&+ (\delta\delta)_{m_1 m_1,m_2 m_2}]
\end{aligned}
$$

for $j, k = 1, \cdots, m_1 - 1$ ($j \leq k$) and $r, s = 1, \cdots, m_2 - 1$ ($r \leq s$). Therefore, both model (10) and (11) share exactly the same genetic components as the ones defined in the original Fisher's two-locus ANOVA model (7). In model (10), let $A_1 = \sum_{j=1}^{m_1-1} \alpha_{1j}^* w_{1j}^*$, $D_1 = \sum_{j=1}^{m_1-1} \sum_{k=j}^{m_1-1} \delta_{1jk}^* v_{1jk}^*$, $A_2 = \sum_{r=1}^{m_2-1} \alpha_{2r}^* w_{2r}^*$, $D_2 = \sum_{r=1}^{m_2-1} \sum_{s=r}^{m_2-1} \delta_{2rs}^* v_{2rs}^*$, $A_1 A_2 = \sum_{j=1}^{m_1-1} \sum_{r=1}^{m_2-1} (\alpha\alpha)_{jr}^* w_{1j}^* w_{2r}^*$,

$A_1D_2 = \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1}(\alpha\delta)^*_{j,rs}w^*_{1j}v^*_{2rs}$, $D_1A_2 = \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}(\delta\alpha)^*_{jk,r}v^*_{1jk}w^*_{2r}$ and $D_1D_2 = \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1}(\delta\delta)^*_{jk,rs}v^*_{1jk}v^*_{2rs}$, which represent the additive and dominance at locus 1, additive and dominance at locus 2, additive by additive, additive by dominance, dominance by additive and dominance by dominance genetic components, respectively. Each genetic component consists of a set of average allelic effects (additive or interaction) contributed by alleles from the same locus or the same set of loci, and at each locus at most two alleles could be involved. We refer the number of alleles involved in each average allelic effect of a genetic component as the order of the genetic component, and the number of non-redundant average allelic effects involved in a genetic component as the degrees of freedom (df) of the genetic component. Note that genetic components of the same order may have different df. For example, the additive by additive component A_1A_2 has its order being 2 and $df = (m_1 - 1)(m_2 - 1)$, while the dominance component D_1 at locus 1 has its order being 2 and $df = (m_1 - 1)^2$.

Both the GLM (9) and GMA model (10) provide a re-parameterization of the expected genotypic values. From the relationship between their genotype coding variables, these two models can be transformed equivalently from one to the other. We can represent the parameters in GMA model (10) in terms of the parameters in its equivalent GLM (9) as shown in Appendix D in Supplementary Material. It is interesting to see that an average allelic effect can be affected by not only its corresponding fixed allelic effect but also other higher-order fixed allelic effects. As a result, the hypothesis test of an average allelic effect in the GMA model (10) is not equivalent to the hypothesis test of the corresponding fixed allelic effect in an equivalent GLM (9) when higher order fixed allelic effects are present. From Appendices C, D in Supplementary Material, we can also derive the LSE of parameters for the GMA model (10) as the following

$$\begin{aligned}
\widehat{\mu}^* &= \bar{y}^*_{\cdots\cdot}, \ \ \widehat{\alpha}^*_{1j} = \bar{y}^*_{j\cdots} - \bar{y}^*_{m_1\cdots}, \ \ \widehat{\alpha}^*_{2r} = \bar{y}^*_{\cdot\cdot r\cdot} - \bar{y}^*_{\cdot\cdot m_2\cdot} \\
\widehat{\delta}^*_{1jk} &= \bar{y}^*_{jk\cdot\cdot} - (\bar{y}^*_{jm_1\cdot\cdot} + \bar{y}^*_{m_1k\cdot\cdot}) + \bar{y}^*_{m_1m_1\cdot\cdot} \\
\widehat{\delta}^*_{2rs} &= \bar{y}^*_{\cdot\cdot rs} - (\bar{y}^*_{\cdot\cdot rm_2} + \bar{y}^*_{\cdot\cdot m_2s}) + \bar{y}^*_{\cdot\cdot m_2m_2} \\
\widehat{(\alpha\alpha)}^*_{jr} &= \bar{y}^*_{j\cdot r\cdot} - (\bar{y}^*_{j\cdot m_2\cdot} + \bar{y}^*_{m_1\cdot r\cdot}) + \bar{y}^*_{m_1\cdot m_2\cdot} \\
\widehat{(\alpha\delta)}^*_{j,rs} &= \bar{y}^*_{j\cdot rs} - (\bar{y}^*_{m_1\cdot rs} + \bar{y}^*_{j\cdot rm_2} + \bar{y}^*_{j\cdot m_2s}) \\
&\quad +(\bar{y}^*_{j\cdot m_2m_2} + \bar{y}^*_{m_1\cdot rm_2} + \bar{y}^*_{m_1\cdot m_2s}) - \bar{y}^*_{m_1\cdot m_2m_2} \\
\widehat{(\delta\alpha)}^*_{jk,r} &= \bar{y}^*_{jkr\cdot} - (\bar{y}^*_{jkm_2\cdot} + \bar{y}^*_{jm_1r\cdot} + \bar{y}^*_{km_1r\cdot}) \\
&\quad +(\bar{y}^*_{jm_1m_2\cdot} + \bar{y}^*_{km_1m_2\cdot} + \bar{y}_{m_1m_1r\cdot}) - \bar{y}^*_{m_1m_1m_2\cdot}
\end{aligned}$$

and $\widehat{(\delta\delta)}^*_{jk,rs} = \widehat{(dd)}_{jk,rs}$, for $j, k = 1, \ldots, m_1 - 1, j \leq k$; and $r, s = 1, \ldots, m_2 - 1, r \leq s$. Similar to the one-locus GMA model, the LSE of these average allelic effects depend on the weighted genotypic group means. Except the LSE of the highest order effects for dominance by dominance interactions, the LSE of other lower order average allelic effects could be more robust to phenotypic outliers in small genotypic groups than the LSE of their corresponding fixed allelic effects.

When the two markers are unlinked and in HWE, the inheritance of four paternal and maternal alleles at the two loci are independent. Therefore, the four sets of indicator variables $\{z^{(1)}_{1j}, j = 1, \cdots, m_1\}$, $\{z^{(1)}_{2k}, k = 1, \cdots, m_1\}$, $\{z^{(2)}_{1r}, r = 1, \cdots, m_2\}$, and $\{z^{(2)}_{2s}, s = 1, \cdots, m_2\}$ are independent with each other, although the variables within each set could still be correlated. In this case, all the genetic components are independent of each other, and the GMA model (10) can provide an orthogonal partition of the expected genotypic variance. In Wang (2014), we have derived formulas for computing the genetic variance components based on the model parameters in a general multi-locus GMA model. For the two-locus GMA model (10), by plugging in the LSE of its model parameters, we obtain estimators of the genetic variance components as shown in Appendix E in Supplementary Material. It should be pointed out that Weir and Cockerham (1977) also derived estimates of the genetic variance components based on the model parameters in Fisher's two-locus ANOVA model (7). But they did not construct the estimators of genetic variance components in terms of the weighted genotypic group means. Note that the orthogonal partition on the genetic variance components implies that the genetic components constitute independent random sources contributing to the expected genotypic variance. Therefore, we could estimate and test for each genetic component separately. Bonferroni criterion can also be applied to correct for the association testing of multiple genetic components.

When the inheritance of paternal and maternal alleles at the two loci are dependent, non-zero covariances among different genetic components may present. The dependency among inheritance of the paternal and maternal alleles at the two markers can be measured by $D_{jkrs} = p_{jkrs}/(2-1_{\{j=k\}})(2-1_{\{r=s\}}) - p_jp_kp_rp_s$, for $j, k = 1, \cdots, m_1$ and $r, s = 1, \cdots, m_2$. Let $\widehat{D}_{jkrs} = \hat{p}_{jkrs}/(2 - 1_{\{j=k\}})(2 - 1_{\{r=s\}}) - \hat{p}_j\hat{p}_k\hat{p}_r\hat{p}_s$. Similar to the one-locus case, when $\widehat{D}_{jkrs} = 0$ for $j, k = 1, \cdots, m_1$ and $r, s = 1, \cdots, m_2$, we can show that the LSE of the average additive, dominance, additive by additive, additive by dominance, dominance by additive and dominance by dominance effects in the full model (10) can keep consistent when some components are excluded from the model.

In general, we can always incorporate the two-locus GMA model (10) into a regression model (1). By treating the modified genotype coding variables w^*_{1j}, w^*_{2r}, v^*_{1jk}, and v^*_{2rs} as regular fixed covariates, we can fit the regression model using the ordinary LS approach. Based on the fitted model, we can calculate the fitted genetic components for each individual $i = 1, \cdots, N$. Then the genetic variance components and their possible covariances can be estimated as the sample variances and covariances from the fitted values of these genetic components. Similar to the one-locus case, these estimators of the variance and covariance components are different from the traditional ANOVA estimators when the two marker are linked or their genotypes are in HWD. As the estimators of variance components coming from the sample variances, they are guaranteed to be non-negative. When the model residuals are normally distributed, these estimators of the variance and covariance components become MLE and likely possess the asymptotic normality property. Besides, these variance and covariance estimators are likely asymptotically unbiased.

2.3. A Simulation Example

Consider two biallelic marker loci with allele frequencies $p_1 = 0.4$ and $q_1 = 0.6$ for alleles "1" and "0" at locus 1, and $p_2 = 0.2$ and $q_2 = 0.8$ for alleles "1" and "0" at locus 2. Assume that the two marker loci are in linkage and gametic equilibria. The expected genotypic values at the two marker loci are simulated based on a two-locus GLM model (9) with $\mu_0 = 10$ and $a_{11} = a_{21} = d_{111} = d_{211} = (aa)_{11} = 1$ and $(ad)_{1,11} = (da)_{11,1} = (dd)_{11,11} = 0$. From Appendix D in Supplementary Material, we can show that this GLM is equivalent to a GMA model (10) with $\mu^* = 11.72$ and $\alpha^*_{11} = 1.8$, $\alpha^*_{21} = 2$, $\delta^*_{111} = 1$, $\delta^*_{211} = 1$, $(\alpha\alpha)^*_{11} = 1$, $(\alpha\delta)^*_{1,11} = 0$, $(\delta\alpha)^*_{11,1} = 0$, and $(\delta\delta)^*_{11,11} = 0$. Based on the true expected genotypic values and the genotypic distribution, we also know that the total expected genotypic variance, which is $V_G = 3.09$, has an orthogonal partition which consists of eight variance components contributed by $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$. Besides, these eight components $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$ contribute 50.74, 1.87, 41.53, 0.83, 5.03, 0.00, 0.00, 0.00% of the total expected genotypic variance V_G. For each random sample of size n, we first generate genotypes of individuals independently based on the genotypic distribution. Then, based on the genotypes, we create dummy variables $w_1 = w_{11}, v_1 = v_{111}, w_2 = w_{21}, v_2 = v_{211}, ww = w_{11} * w_{21}, wv = w_{11} * v_{211}, vw = v_{111} * w_{21}, vv = v_{111} * v_{211}$ and mean-corrected index variables $w^*_1 = w_{11}, v^*_1 = v_{111}, w^*_2 = w_{21}, v^*_2 = v_{211}, ww^* = w_{11} * w_{21}, wv^* = w_{11} * v_{211}, vw^* = v_{111} * w_{21}, vv^* = v_{111} * v_{211}$. The phenotypic values is a sum of the genotypic values and the residuals with the residuals being independent of the genotypes. To generate phenotypic values, we assume that the residuals $\epsilon_i, i = 1, \cdots, n$, are i.i.d. and normally distributed with the residual variance being $V_\epsilon = 17.51$, which account for 85% of the total phenotypic variance (i.e., the broad sense heritability is about 15%).

First, we show that the GLM and GMA models provide different partitions of the expected genotypic variance. To minimize the sampling variation, we simulate a random sample with $n = 100{,}000$. We fit both GLM and GMA models to the random sample using SAS "proc glm" (SAS Institute, INC.). For the fitted GLM, we have LSE $\hat{\mu}_0 = 9.99$, $\hat{a}_{11} = 1.01$, $\hat{a}_{21} = 1.05$, $\hat{d}_{111} = 0.97$, $\hat{d}_{211} = 0.95$, $\widehat{(aa)}_{11} = 0.97$, $\widehat{(ad)}_{1,11} = 0.02$, $\widehat{(da)}_{11,1} = -0.03$ and $\widehat{(dd)}_{11,11} = 0.07$, which are close to the true values. From the fitted GLM, we calculate estimates of the variance components contributed by $w_1, v_1, w_2, v_2, ww, wv, vw, vv$ as 0.4894, 0.1274, 0.3533, 0.0345, 0.4153, 0.00, 0.00, 0.00, respectively. The summation of these variance components is 1.4199, which is about 46% of the total expected genotypic variance estimate $V_G = 3.0896$. In other words, in the GLM based partition of the expected genotypic variance, about 54% is contributed by the covariance components due to the collinearity among the variables $w_1, v_1, w_2, v_2, ww, wv, vw, vv$. For the fitted GMA, we have LSE $\hat{\mu}^* = 11.72$, $\hat{\alpha}^*_{11} = 1.78$, $\hat{\alpha}^*_{21} = 2.02$, $\hat{\delta}^*_{111} = 0.96$, $\hat{\delta}^*_{211} = 0.98$, $\widehat{(\alpha\alpha)}^*_{11} = 0.97$, $\widehat{(\alpha\delta)}^*_{1,11} = 0.05$, $\widehat{(\delta\alpha)}^*_{11,1} = -0.02$ and $\widehat{(\delta\delta)}^*_{11,11} = 0.07$, which are also close to the true values. From the fitted GMA model, we compute estimates of the variance components contributed by $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$ as 1.5321, 0.0534, 1.3079, 0.0244, 0.1462, 0.00, 0.00, 0.00, respectively. The summation of these variance components is 3.064, which is about 99% of the total expected genotypic variance estimate $V_G = 3.0896$. This indicates that the GMA model leads to a partition of the expected genotypic variance, which is almost orthogonal with only about 1% coming from the covariance components. Note also that these estimates of the variance components contributed by $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$ account for approximately 49.59, 1.73, 42.33, 0.79, 4.73, 0.00, 0.00, 0.00% of the expected genotypic variance estimate $V_G = 3.0896$, which are close to the true proportions 50.74, 1.87, 41.53, 0.83, 5.03, 0.00, 0.00, 0.00% of the variables $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$ contributed to the expected genotypic variance.

We also look at the difference between the partition of the expected genotypic variance and the Type III sums of squares for this random sample. Both the GLM and GMA models have the same regression sum of squares $SSR = 307788.89$ and mean square error $MSE = 17.52$. In the fitted GLM, the Type III sums of squares for the eight variables $w_1, v_1, w_2, v_2, ww, wv, vw, vv$ are 13415.25, 3475.46, 8472.60, 840.54, 4149.25, 0.24, 1.04, and 0.67, respectively. The summation of these Type III sums of squares is 30355.05, which is less than 10% of the total SSR. In the fitted GMA, the Type III sums of squares for the eight variables $w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*$ are 153184.54, 5340.61, 130751.17, 2438.75, 14620.38, 2.92, 0.44, and 0.67, respectively. The summation of these Type III sums of squares is 306339.50, which is about 99.5% of the total SSR. This indicates that these Type III sums of squares also provide an orthogonal partition of the SSR. Or, in other words, the hypothesis tests of the eight genetic components are approximately orthogonal.

Next, we examine the performance in model selection between using the GLM and GMA models. We consider varied sample sizes of $n = 500$, 1000, 2000, and 5000. Under each simulation scenario, we run 1000 simulation. For each simulation sample, we first apply one commonly used method—the forward stepwise selection for model selection on $\{w_1, v_1, w_2, v_2, ww, wv, vw, vv\}$ and $\{w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*\}$, separately. We run "proc glmselect" in SAS with the criterion p=0.05 for both entry and stay in the model. For the GMA model selection on index variables $\{w^*_1, v^*_1, w^*_2, v^*_2, ww^*, wv^*, vw^*, vv^*\}$, as these variables are independent in the underlying true model, we can rank them in the order of $w^*_1, w^*_2, ww^*, v^*_1, v^*_2, vw^*, wv^*$, and vv^* according to their variance contributions from the largest to the smallest. Intuitively, we expect that the selected models would include the higher ranked variables more often than the lower ranked ones. We mainly focus on the top five variables $w^*_1, w^*_2, ww^*, v^*_1, v^*_2$ and classify the selected models into 10 categories: I = all the five variables $w^*_1, w^*_2, ww^*, v^*_1, v^*_2$ are selected; II = $w^*_1, w^*_2, ww^*, v^*_1$ are selected but not v^*_2; III = $w^*_1, w^*_2, ww^*, v^*_2$ are selected but

TABLE 1 | Counts of selected GLM and GMA model types for “Stepwise Selection”.

Model	GLM				GMA			
types	*n* = 500	*n* = 1000	*n* = 2000	*n* = 5000	*n* = 500	*n* = 1000	*n* = 2000	*n* = 5000
I	0	15	196	605	13	71	285	737
II	3	132	401	342	118	294	433	239
III	0	27	70	33	43	98	99	22
IV	89	266	176	9	370	361	160	2
V	288	213	71	11	456	176	23	0
VI	247	140	47	0	0	0	0	0
VII	82	77	9	0	0	0	0	0
VIII	265	129	30	0	0	0	0	0
IX	26	1	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0
Total	1000	1000	1000	1000	1000	1000	1000	1000

TABLE 2 | Counts of selected GLM and GMA model types for “Adaptive LASSO”.

Model	GLM				GMA			
types	*n* = 500	*n* = 1000	*n* = 2000	*n* = 5000	*n* = 500	*n* = 1000	*n* = 2000	*n* = 5000
I	1	18	91	363	4	17	58	367
II	17	87	246	372	33	89	226	346
III	8	18	41	19	17	20	38	13
IV	87	180	226	153	194	336	371	197
V	271	227	125	25	602	432	237	55
VI	184	212	150	46	0	0	0	0
VII	39	34	13	0	0	0	0	0
VIII	227	145	71	10	0	0	0	0
IX	61	17	6	1	8	0	0	0
X	105	62	31	11	142	106	70	22
Total	1000	1000	1000	1000	1000	1000	1000	1000

not v_1^*; IV= w_1^*, w_2^*, ww^* are selected but not v_1^*, v_2^*; V= w_1^*, w_2^* are selected but not ww^*, v_1^*, v_2^*; VI = w_1^*, ww^* are selected but not w_2^*; VII = w_2^*, ww^* are selected but not w_1^*; VIII = ww^* is selected but w_1^*, w_2^* are missed; IX=w_1^* is selected but miss w_2^*, ww^*, or w_2^* is selected but miss w_1^*, ww^*; X=w_1^*, w_2^*, ww^* are missed. For the GLM model selection on the dummy variables $\{w_1, v_1, w_2, v_2, ww, wv, vw, vv\}$, we also define the similar 10 categories I-VIII based on the variables $w_1, v_1, w_2, v_2, ww, wv, vw, vv$. Under each simulation scenario, the counts of selected GLM and GMA model types for stepwise selection are listed in **Table 1**.

For the stepwise selection, the GMA model shows a clear advantage over the GLM on choosing the Type I (the true model), Type I+II, Type I+II+III, Type I+II+III+IV, or Type I-V models. Meanwhile, the GLM tends to miss one of the main fixed effects (Type VI, VII, or VIII) when the sample size ≤2000, even though a Type VI, VII, or VIII GLM could imply an equivalent GMA model of Type IV. Overall, as the sample size increases, the model selection improves using either GLM or GMA models. But the selected GLM have a bigger variety, while the selected GMA models are more predictable.

It has been known that the conventional stepwise model selection has problems such as the stepwise F-statistics may no longer follow the F-distributions and the aggravated collinearity among the selected variables (Harrell, 2001). Several regularized regression methods have been developed to deal with the model selection problem especially for high-dimensional data. Here we also apply one of the regularized regression methods - the adaptive LASSO, using “proc glmselect” in SAS. We adopt the Bayesian information criterion (BIC) for model selections. The standardization of the variables is also made prior to the model selection. Note that the standardization of the variables $\{w_1, v_1, w_2, v_2, ww, wv, vw, vv\}$ or $\{w_1^*, v_1^*, w_2^*, v_2^*, ww^*, wv^*, vw^*, vv^*\}$ can only affect the scales of the regression coefficients but does not change the correlation structure among each set of variables, while making mean-corrections on the indicator variables $\{z_{1j}^{(1)}, z_{2k}^{(1)}\}$ and $\{z_{1r}^{(2)}, z_{2s}^{(2)}\}$ can lead to different correlation structures between the dummy variables $\{w_1, v_1, w_2, v_2, ww, wv, vw, vv\}$ and the index variables $\{w_1^*, v_1^*, w_2^*, v_2^*, ww^*, wv^*, vw^*, vv^*\}$. The counts of selected GLM and GMA model types for adaptive LASSO are listed in **Table 2**.

For the adaptive LASSO, it appears that the GMA model performs slightly better on choosing Type I (the true model), Type I+II, Type I+II+III, Type I+II+III+IV, or Type I-V models when sample size=500. However, this advantage diminishes for larger sample sizes. Still, the selected GLM appear to have a bigger variety, while the selected GMA models are more predictable. On the other hand, it seems that the selected GMA models could miss all the three top variables w_1^*, w_2^*, ww^* more likely than the GLM models for w_1, w_2, ww. A brief comparison between **Tables 1**, **2** also reveals that the stepwise selection has a better performance than the adaptive LASSO under our simulation setting. But this may not be a fair comparison as the selected variables in our simulation setting is very limited. Besides, except BIC, many other criteria are available for model selections in "proc glmselect." Further exploration on those criteria is needed.

3. DISCUSSION

In this study, we applied the GMA models on analysis of variance components for genetic markers with unphased genotypes. We pointed out that the traditional Fisher's ANOVA model does not explicitly specify the random variables that contribute to the various genetic variance components. The constraints on their model parameters also complicate the model fitting. Meanwhile, the classical dummy-variable based GLM does not provide the same partition on the genetic variance components as the original Fisher's ANOVA model. The hypothesis tests of the fixed allelic effects in a GLM based on the partial (Type III) reduction in sums of squares could also be inadequate for testing the existence of variance components when allelic interactions are present. Alternatively, the GMA model can retain the same partition on the genetic variance components as the traditional Fisher's ANOVA model. Similar to the classical GLM, the GMA model does not require constraints on its model parameters and can be fitted using the ordinary least square approach. As the result, the GMA model allows us to estimate and test for the genetic variance components more conveniently than the classical GLM.

The classical GLM is appealing in genetic association studies due to its simplicity in interpretation of the model parameters, which often represent certain comparisons of the expected genotypic values in different genotype groups. However, the GLM-based approach faces challenges in dealing with allelic interactions because the fixed lower order allelic effects are often confounded with the higher order allelic interactions in a GLM even when the paternal and maternal alleles are independently inherited. In order to test for a particular fixed allelic interaction based on the classical GLM, we need to include all its lower order effects in the model to make this allelic interaction interpretable. Besides, ignoring certain higher order interactions in the model could also affect the definition of this particular allelic interaction due to their potential confounding, as pointed out in Zeng et al. (2005). On the other hand, analysis of genetic variance components using ANOVA type models such as GMA provides an alternative way on assessing the allelic effects and interactions. A nice feature of the GMA model is that the genetic components are independent of each other when multiple genetic markers are unlinked and in equilibrium populations. Therefore, at least in equilibrium populations, each genetic component can be treated as a random source of variation and tested individually regardless of its lower or higher order genetic components. A statistically significant higher-order genetic variance component implies a variation contributed by varied allele types from a set of loci, which could be more perceivable than a significant fixed allelic interaction for claiming allelic interactions.

As shown in Wang (2014), the extension of GMA to multiple loci is straightforward. The GMA model could also be applied to other phenotypic outcomes. For example, under the generalized linear mixed model framework, we can consider the genetic components contributing to a g-transformed (g - a link function) expected outcomes. For survival outcomes, we can examine the genetic components contributing to the hazard functions as well. However, it should be pointed out that the estimation of genetic variance components for genetic markers relies on the genotypic distribution of the markers in a study population. When a sample does not come from the simple random sampling, an adjustment for the sampling strategy is needed in order to make appropriate statistical inference in the general population. Currently, the genome-wide association studies (GWAS) often adopt the case-control design. When the case and control sampling rates are known, we could assess the variance components of genetic markers based on their odds ratio estimates. A thorough exploration on applying the GMA type models to a case-control based GWAS could be a research topic in the future.

AUTHOR CONTRIBUTIONS

TW planned the study, conducted the derivation, and wrote the manuscript.

ACKNOWLEDGMENTS

The author would like to thank his colleague Dr. Kwang Woo Ahn in Division of Biostatistics at Medical College of Wisconsin for his advice on implementation of the adaptive LASSO method.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00123

REFERENCES

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh; London: Oliver & Boyd.

Gallais, A. (1974). Covariances between arbitrary relatives with linkage and epistasis in the case of linkage disequilibrium. *Biometrics* 30, 429–446.

Harrell, F. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* New York, NY: Springer-Verlag.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* 9, 226–252.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics.* New York, NY: Wiley.

Searle, S. R. (1971). *Linear Models.* New York, NY: John Wiley & Sons, Inc.

Searle, S. R. (1995). An overview of variance component estimation. *Metrika* 42, 215–230.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* New York, NY: John Wiley & Sons, Inc.

Wang, T. (2011). On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. *BMC Genet.* 12:82. doi: 10.1186/1471-2156-12-82

Wang, T. (2014). A revised fisher model on analysis of quantitative trait loci with multiple alleles. *Front. Genet.* 5:328. doi: 10.3389/fgene.2014.00328

Wang, T., and Zeng, Z. B. (2009). Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium. *BMC Genet.* 10:52. doi: 10.1186/1471-2156-10-52

Weir, B. S., and Cockerham, C. C. (1977). "Two-locus theory in quantitative genetics," in *Proceedings of the International Conference on Quantitative Genetics*, eds E. Pollak, O. Kempthorne, and T. B. Bailey (Ames: Iowa State University Press), 247–269.

Zeng, Z.-B., Wang, T., and Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* 169, 1711–1725. doi: 10.1534/genetics.104.035857

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer ZY and handling editor declared their shared affiliation, and the handling editor states that the process nevertheless met the standards of a fair and objective review.