THE UNIVERSITY OF CHICAGO

OVERCOMING THE "FEAST OR FAMINE" EFFECT: IMPROVED INTERACTION TESTING IN GWAS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
HUANLIN ZHOU

CHICAGO, ILLINOIS
AUGUST 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The six-year journey at the University of Chicago has been a great treasure in my life. I am deeply grateful for all the help and support I received from the university and the Department of Statistics.

First and foremost, I want to express my sincerest gratitude to my advisor, Prof. Mary Sara McPeek. During the past five years, she has been so nice and supportive to me and so patient in teaching me how to conduct professional research. I really appreciate her help and support, her brilliant ideas, inspiration and profound knowledge. I could not have completed my dissertation without her guidance and encouragement.

I also want to thank my thesis committee members, Prof. Dan Nicolae and Prof. Jingshu Wang, for their insightful questions, comments and suggestions that inspired me to have more thoughts on my dissertation.

I am also grateful to my friends and fellow classmates for their help both at school and in daily life, especially Yi Wei, Chih-Hsuan Wu, Zehao Niu and Wanrong Zhu. A special thank you to Yi Wei, who has been my roommate for over two years, for her companionship, care, and for all things that she helped me with.

Last but not least, I want to thank my parents, Chunyang Zhou and Li Ran, for their unwavering love and support. Their concern for me has always been my mental support. I love them dearly and I could never have achieved my current accomplishments without them.

ABSTRACT

In genetic association analysis of complex traits, detection of interaction (either GxG or GxE) can help to elucidate the genetic architecture and biological mechanisms underlying the trait. Detection of interaction in a genome-wide association study (GWAS) can be methodologically challenging for various reasons, including a high burden of multiple comparisons when testing for epistasis between all possible pairs of a set of genome-wide variants, as well as heteroscedasticity effects occurring in the presence of GxG or GxE interaction. In this paper, we address the problem of an even more striking phenomenon that we call the "feast or famine" effect that occurs when testing interaction in a genome-wide context. We show that, even in a simplified setting in which there is no interaction at all (and so no heteroscedasticity) and all SNPs are assumed independent, in a GWAS to detect gene-gene or gene-environment interactions with a fixed genetic variant or environmental factor, the distribution of the genome-wide p-values under the null hypothesis of no interaction is not the i.i.d. uniform one that is commonly assumed. Using standard methods, even if all SNPs are independent, some GWAS's will have systematically underinflated p-values ("feast"), and others will have systematically overinflated p-values ("famine"), which can lead to false detection of interaction, reduced power, inconsistent results across studies, and failure to replicate true signal. This is a surprising result that is specific to detection of interaction in a GWAS, and it may partly explain why such detection has so far proved challenging and difficult to replicate. We show theoretically that the key cause of this phenomenon is which variables are conditioned on in the analysis, and this suggests an approach to correct the problem by changing the way the conditioning is done. Using this insight, we have developed the TINGA (Testing INteraction in GWAS with test statistic Adjustment) method to adjust the interaction test statistics to make their p-values closer to uniform under the null hypothesis. In simulations we show that TINGA controls type 1 error, improves power and reduces the "feast or famine" effect. TINGA allows for covariates and population structure through use

of a linear mixed model and accounts for heteroscedasticity. We apply TINGA to detection of epistasis in a study of flowering time in *Arabidopsis thaliana*.

CHAPTER 1

INTRODUCTION

GWAS usually investigates the associations between genetic variants (or SNPs) and a particular phenotype, in a genome-wide scale. Apart from the ordinary genotype-trait associations, we are also interested in the statistical problem of detecting interactions in a GWAS setting, either gene by environment (GxE) interaction or epistasis, which is the interaction between genes. It is well-known that the effects of a genetic variant can be different for individuals with different environments, such as age [2; 3], sex [4; 5; 6; 7; 8] , lifestyles [9] , cell type [10] and other exposures [11]; the genetic effects can also depend on other variants, either from the same genome [12; 13] or the genome of another species (such as pathogen and host [14], mother and offspring [15]). The two types of interactions are very important in many aspects. For example, detection of such interaction effects can enhance the ability to identify genetic effects that would otherwise be reduced or masked [16]; they are considered as one of the reasons why results of marginal association studies are sometimes hard to replicate [17]; they are believed to account for a large part of missing heritability [18; 19; 20]; and they can lead to a better understanding of genetic architecture of complex traits and diseases [21; 16; 22] and potentially benefit many areas such as public health [23] and agriculture [24; 25]. Extensive prior research has been done to develop appropriate methods for detecting interactions in GWAS, aiming to improve computational efficiency, reduce false positives and increase power [6; 26; 27; 28; 29; 30; 31; 32].

One challenge specific to epistasis detection is that, because of the large number of tests, an exhaustive search for epistatic effects in a GWAS context has a larger computational burden and lower statistical power than an ordinary trait-variant GWAS. To deal with this issue, researchers have developed various methods that corrects for multiple testing while still remaining powerful [33; 34]. Another option is to reduce the number of tests by a two-stage

approach: first select a subset of SNPs that are more likely to be involved in interaction and then test for interaction among them [26; 30; 35; 36; 37].

Moreover, previous work found it hard to replicate interactions in GWAS [38; 39; 40]. This can occur for a variety of reasons. For example, in some cases, an apparent epistatic effect that is detected could be due to an unsequenced causal variant [41; 42; 38]. Another important issue that has been identified is heteroscedasticity [43; 44; 45] that can result under the null model when, for example, interaction is present between one of the two tested variables and some other variable not included in the model or when the null model is misspecified in some other way. If not accounted for, this heteroscedasticity can lead to excess type 1 error for testing interaction [43; 44; 45].

Many scenarios of testing for GxG or GxE in a GWAS context involve fixing one genetic variant or environmental factor and performing an interaction GWAS by testing the fixed variable for interaction with each genetic variant across the genome. The typical approach to inference treats the phenotype as random and the environmental factors and/or genotypes as fixed. Systematically inflated or deflated p-values in such an interaction GWAS have been previously reported, based on both data and simulations [42; 43; 44]. Even under simplified assumptions, in the absence of problems such as heteroscedasticity, it has been noted that type 1 error rates and genomic control inflation factors are highly variable across such interaction GWAS's [43; 44].

In this paper, we develop a deeper and more detailed understanding of such unexpected behaviour of interaction test results in a GWAS context, which we call the "feast or famine" effect. We frame this problem as resulting from the choice of variables to condition on and show how changing this choice has the potential to resolve the problem. Our framework also explains clearly why the "feast or famine" effect only occurs in interaction GWAS, not in ordinary association GWAS. We implement our ideas in a method we call TINGA (Testing INteraction in GWAS with test statistic Adjustment), in which we adjust the t-statistic for

interaction by re-centering and re-scaling it using the null conditional mean and conditional variance of its numerator, with a more appropriate choice of conditioning variables. In simulations, we demonstrate the ability of TINGA to greatly reduce the "feast or famine" effect while controlling type 1 error and increasing power. We apply TINGA to detect epistasis in a GWAS for flowering time in *Arabidopsis thaliana*.

CHAPTER 2

PROBLEM DESCRIPTION AND INTERPRETATION

We consider the problem of testing for interaction, either $G \times E$ or $G \times G$, in a GWAS context. In a sample of n individuals, let Y be an $n \times 1$ trait vector, and let G be an $n \times m$ matrix of genotypes for a set of m genome-wide variants. Let Z be an $n \times 1$ vector that, in the case of $G \times E$ testing, represents the environmental variable that we wish to test interaction with and in the case of $G \times G$ testing, represents the genotype at a particular variant that we wish to test interaction with (where we assume that Z is removed from the matrix G in that case). In addition, we can allow for an $n \times k$ matrix U of covariates (including intercept). By "testing interaction in a GWAS context", we mean that for each j in $\{1, \ldots, m\}$, we test for interaction between G_j and Z in a linear or linear mixed model for Y, where G_j is the jth column of G.

In this section, we first describe what we call the "feast or famine" effect for testing interaction in a GWAS context. We explain how the "feast or famine" effect can result in some GWAS's having systematically overinflated interaction p-values, reducing power, while others have systematically underinflated p-values, resulting in excess type 1 error. In what follows, we focus our exposition on the t-statistic for testing interaction, but the "feast or famine" effect is very general and applies just as well to, e.g., the likelihood ratio chi-squared test or the F-test for interaction. We show that the "feast or famine" effect does not occur in ordinary GWAS for association between a trait and genetic variants, but only when testing interaction in a GWAS context. After that we describe our TINGA method to correct the interaction test statistic to greatly reduce this effect. We then show the performance of our method via simulation results in the next section.

In the simplest setting in which there are no covariates and no population sub-structure, we let T_j denote the t-statistic for testing interaction between G_j and Z, i.e., for testing

$H_0 : \delta = 0$, in the following linear model:

$$Y = 1_n\alpha + G_j\beta + Z\gamma + (G_j \circ Z)\delta + \epsilon, \tag{2.1}$$

where 1_n is a vector of length n with every entry equal to 1, α, β, γ and δ are unknown scalar parameters, $\epsilon \sim N(0, \sigma_e^2 I_n)$, where σ_e^2 is unknown and I_n is the $n \times n$ identity matrix, and where, for any two vectors a and b, both of length n, we define $a \circ b$ to be the vector of length n with ith element $(a_i - \bar{a})(b_i - \bar{b})$, where, e.g., $\bar{a} = n^{-1}\sum_{i=1}^{n} a_i$. (Note that the test statistics T_j would remain exactly the same if we replaced $G_j \circ Z$ in (2.1) by the element-wise product of the vectors G_j and Z, but choosing to center the variables before multiplying them has various advantages such as reducing potential collinearity and making the coefficients more interpretable.)

2.1 The "feast or famine" effect: what we thought we knew about testing interaction in a GWAS context was wrong

For simplicity, we first focus the exposition on $G \times E$ interaction testing. An essential feature of testing $G \times E$ interaction in a GWAS context is that we obtain a set of m test statistics T_j, $j \in \{1, \ldots, m\}$, where $T_j \equiv T_j(G_j, Z, Y)$, with the same Y and Z used in all the test statistics and only G_j varying. As a thought experiment, imagine the simplest possible null scenario in which Y, Z and the columns of G are mutually independent, with the elements of Y drawn as i.i.d. $N(\mu, \sigma^2)$, the elements of Z drawn as i.i.d. from some distribution F_Z, and the elements of G_j drawn as i.i.d. from some distribution F_{Gj}, for $j = 1, \ldots, m$. What would be the distribution of $(T_1, \ldots, T_m)$ in this case? It is well-known that for any given j, the distribution of T_j in this case is the (central) Student's t distribution on $n - 4$ df, which we denote by $\mathcal{T}_{n-4}$. Thus, it is tempting to assume that $T_1, \ldots, T_m$ must be approximately i.i.d. draws from $\mathcal{T}_{n-4}$, but that is (perhaps surprisingly) incorrect.

In this simple scenario, we show that it is most appropriate to think of $T_1, \ldots, T_m$ as i.i.d. draws from some distribution whose mean is 0 and whose variance is a function of (Y, Z). For some choices of (Y, Z), the variance of the resulting T_j's is larger than 1 (where 1 is the approximate variance of $\mathcal{T}_{n-4}$ for large n), while for other choices of (Y, Z), the variance of the resulting T_j's is smaller than 1. Thus, if we used $\mathcal{T}_{n-4}$ to calculate p-values $p_1, \ldots, p_m$ for $T_1, \ldots, T_m$, respectively, which would be the standard approach, then in one GWAS these p-values might be systematically too big on average, in a second GWAS these p-values might be systematically too small on average, and in a third GWAS, they might be about right (where by "about right" we mean approximately i.i.d. uniform under the null).

This can easily be observed in simulations. Fig 2.1 shows four histograms, each of which depicts the p-values $p_1, \ldots, p_m$ for a $G \times E$ GWAS obtained as described above, where n is 1,000, m is 5,000, F_Z is taken to be Bernoulli(.2), and F_{Gj} is taken to be Bernoulli(f_j) for $j = 1, \ldots, m$, where $f_1, \ldots, f_m$ are drawn as i.i.d. Unif(.1, .9), to mimic the genotypes of a haploid organism or an inbred line. In Panel A of Fig 2.1, the p-values are seen to be systematically overinflated, while in Panel B of Fig 2.1, the p-values are seen to be systematically underinflated. The information in Table 2.1 supports this conclusion, where we can see that for Panel A, the s.d. of the interaction t-statistics is < 1 and the genomic control inflation factor is < 1, while for Panel B the opposite holds. We repeated this experiment 400 times, and in each replicate, we tested whether the 5,000 p-values were i.i.d. Uniform(0,1) distributed under the null hypothesis (which is equivalent to testing whether the 5,000 interaction t-statistics are i.i.d. $\mathcal{T}_{n-4}$ distributed) using the two-sided equal local levels (ELL) test as implemented in qqconf [1] (See S1 R script to calculate p-values for the two-sided equal local levels test for i.i.d. uniformity for an R script to perform the test). In 190 out of 400, i.e., 47.5%, of the replicates, the two-sided ELL test for uniformity was rejected at level .05, clearly showing that the t-statistics for interaction in a GWAS are not i.i.d. $\mathcal{T}_{n-4}$ distributed under the null hypothesis.

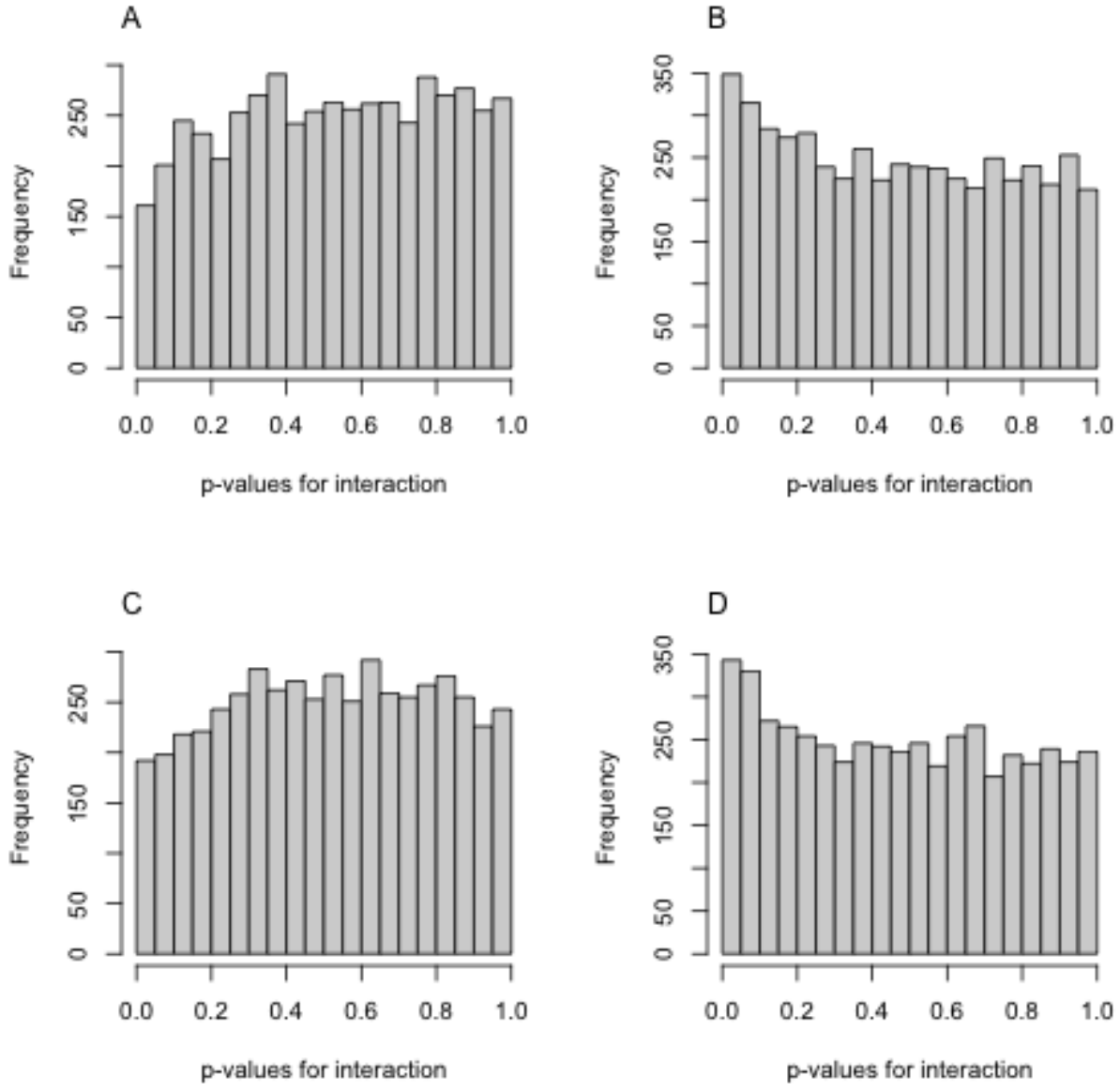
This effect seems to be very general and also occurs when, e.g., F_Z and F_{Gj} are taken to be Gaussian or Binomial, as we show later. Furthermore, if instead of a t-test for interaction, we apply a likelihood ratio chi-squared test or F-test for interaction to the same simulated data sets, we get essentially indistinguishable histograms to those in Fig 2.1 (which is perhaps not surprising since they are asymptotically equivalent tests), and the same 190 replicates out of 400 are rejected by the ELL test for uniformity of the p-values, showing that the likelihood ratio chi-squared test and F-test for interaction are also subject to the "feast or famine" effect.

Table 2.1: **Summary statistics for the examples in Fig 2.1**

Panel	T_j mean	T_j s.d.	genomic control λ	ELL p-value
A	.015	.93	.88	2.2e-10
B	-.002	1.09	1.19	3.6e-12
C	.013	.94	.92	9.5e-9
D	-.010	1.09	1.16	3.5e-12

For each panel of Fig 2.1, T_j mean is the mean and T_j s.d. is the s.d. of the interaction t-statistics whose p-values are displayed in the panel. The genomic control λ is based on the squares of the interaction t-statistics in each panel. The ELL p-value is the p-value for testing the null hypothesis that the interaction p-values are uniformly distributed under the null hypothesis, as described in [1].

Figure 2.1: **Histograms of p-values for t-tests for interaction in a GWAS when the null hypothesis is true**



Each histogram is based on a replicate of (Y, Z) and 5,000 genotypes, $G_1, \ldots, G_{5000}$. In each histogram, interaction is tested between Z and G_j in the linear model in (2.1) for $j = 1, \ldots 5{,}000$, as described in the text, and the 5,000 p-values are computed using the $\mathcal{T}_{n-4}$ distribution and are displayed in the histogram. Panels A and B represent two different replicates of a null simulation as described in the text. In Panel C, the same (Y, Z) replicate is used as in Panel A, and a new set of 5,000 genotypes is simulated and used in the interaction tests. Similarly, in Panel D, the same (Y, Z) replicate is used as in Panel B, and a new set of 5,000 genotypes is simulated and used in the interaction tests.

2.2 A deeper understanding

We want to emphasize that we are not simply saying that $p_1, \ldots, p_m$ are positively correlated. A further key point is that for a particular $G \times E$ GWAS, i.e., for a particular choice of (Y, Z), it is, in principle, predictable based on (Y, Z) whether the p-values of $p_1, \ldots, p_m$ will be systematically too large, systematically too small or about right. For example, in Fig 2.1, when we keep (Y, Z) the same as in Panel A and simulate a completely new and independent set of genotypes G for testing interaction, as in Panel C, we again see over-inflation of the p-values. Similarly, when we keep (Y, Z) the same as in Panel B and simulate a completely new and independent set of genotypes G for testing interaction, as in Panel D, we again see under-inflation of the p-values. This is further supported by the information in Table 2.1.

Thus, use of standard methods would be expected to result in loss of power ("famine") in some GWAS's (e.g., the (Y, Z) used in Panels A and C) and excessive type 1 ("feast") error in other GWAS's (e.g., the (Y, Z) used in Panels B and D).

To understand why this happens, it is helpful to think about which variables we are conditioning on. The ordinary t-statistic for interaction was developed in a non-GWAS context in which it made sense to condition on G_j and Z and treat Y as random, and in that case, the null conditional distribution of T_j can be proven to be the same $\mathcal{T}_{n-4}$ distribution for any values of G_j, Z in the simple setting described above. As a direct consequence of this, it is also true that the unconditional distribution of T_j is $\mathcal{T}_{n-4}$. In other words, if we randomly choose a $G \times E$ GWAS (i.e., randomly choose (Y, Z)), and then randomly choose a null SNP j from that GWAS, then T_j has distribution $\mathcal{T}_{n-4}$. However, in any particular $G \times E$ GWAS, Z and Y are fixed, and only G_j is varying, so it is more appropriate to consider the null conditional distribution of the t-statistic for interaction where we condition on Z and Y and treat G_j as random (see, e.g. [43]). We show that even in the simple case described above, conditional on (Y, Z), the distribution of T_j depends on (Y, Z) and is not $\mathcal{T}_{n-4}$. In fact, in the slightly more general null hypothesis scenario when G_j has some

marginal effect on Y but no interaction with Z, we show that not only the null conditional variance of T_j but even its null conditional mean depends on (Z, Y).

These same ideas apply to testing $G \times G$ interaction in a GWAS context if we think of setting Z to be the genotype of one particular variant, we exclude Z from the columns of G, and we consider a GWAS in which we test for interaction between Z and G_j for $j = 1, \ldots, m$ in model (2.1) using a t-test for interaction. The upshot is that for some $G \times E$ or $G \times G$ GWAS's, i.e., for some realizations of (Y, Z), use of a $\mathcal{T}_{n-4}$ distribution to assess significance of interaction will systematically overstate the evidence for interaction ("feast"), while for other $G \times E$ or $G \times G$ GWAS's, it will systematically understate the evidence for interaction ("famine"). Whether there is feast or famine will depend on the luck of what value of (Y, Z) is observed. This statistical phenomenon could be an important explanation of the difficulty in detecting and replicating epistasis and gene-environment interaction that has long been observed.

With this conditioning explanation in mind, one way of thinking of the "feast or famine" effect is that if we average across many interaction GWASs, then the t-statistic for interaction has correct type 1 error, but its false positives are excessively concentrated in some GWASs, and its false negatives are excessively concentrated in some other GWASs. The good news is that our conditioning explanation implies that by doing conditional calculations, such as we describe below, we should in principle be able to alleviate or entirely eliminate this effect.

2.3 Why doesn't ordinary (non-interaction) GWAS have the "feast or famine" phenomenon?

We have argued that when testing interaction in a GWAS context, we are actually conditioning on Y and Z and letting G_j be random, and that the t-statistic for interaction does not have a t-distribution under the null hypothesis when we condition on (Y, Z). By a similar argument, we could point out that in an ordinary (non-interaction) GWAS, we are

conditioning on Y and letting G_j be random, rather than the reverse. Does this also cause a problem for the t-statistic for association? The answer is no. The problem we describe does not occur for ordinary (non-interaction) GWAS, but is specific to interaction GWAS, as we now explain.

First, consider the t-statistic for association in an ordinary GWAS. We consider a slightly more general scenario than before in which there may be additional covariates U in the model (where U includes an intercept). Suppose the model we use for testing association is

$$Y = U\alpha + G_j\beta + \epsilon \tag{2.2}$$

where Y is $n \times 1$, U is $n \times k$, and G_j is $n \times 1$, all as defined before, α is an (unknown) $k \times 1$ vector, β is the unknown scalar parameter of interest, and $\epsilon \sim N(0, \sigma_e^2 I_n)$, where σ_e^2 is unknown.

Define $P_U = I - U(U^T U)^{-1} U^T$, an $n \times n$ symmetric matrix. We note that the t-statistic for testing $H_0 : \beta = 0$ in the model in (2.2) can be written as

$$S_j = \frac{(G_j^T P_U Y)\sqrt{n-k-1}}{\sqrt{(Y^T P_U Y)(G_j^T P_U G_j) - (G_j^T P_u Y)^2}} \tag{2.3}$$

(see the proof in section 2.4).

From this formula, it is clear that the t-statistic is symmetric in G_j and Y. That is to say, if we switch Y and G_j in 2.2 and regress G_j on Y and U, we will get the same form of t-statistic S_j. In other words, it is equivalent to testing for the marginal effect of Y on G_j in the model

$$G_j \sim N(U\alpha + Y\beta, \sigma^2 I) \tag{2.4}$$

The symmetry between G_j and Y in the ordinary (non-interaction) t-statistic for association means that in large samples, the distribution of the t-statistic under the null hypothesis of

no association would be approximately the same regardless of whether we conditioned on G_j and let Y be random or conditioned on Y and let G_j be random. The only difference would be that G_j would typically be a Binomial or Bernoulli random variable (genotype) and Y might commonly be a conditionally normal random variable (phenotype). In very small sample sizes, the difference between the underlying distributions of G_j and Y would change the conditional distribution of the t-statistic for association depending on which one you conditioned on, but in typical GWAS sample sizes, the central limit theorem will take effect, and the conditional distribution of the t-statistic for association will be approximately the same in both cases.

This difference between ordinary (non-interaction) GWAS and interaction GWAS can be seen in simulations. We performed $r = 5{,}000$ replicates of a null simulation similar to that in section 2.1, except that instead of F_Z being Bernoulli(.2), we made F_Z Bernoulli(p_{Zk}) in replicate k, where $p_{Z1}, \ldots, p_{Zk}$ are i.i.d. Unif(.1, .9). As before, frequencies of G_j are iid draws from Unif(0.1, 0.9), $Y \sim N(\mu, \sigma^2)$ and is independent of Z and G_j's. In replicate k, we tested interaction between Z and G_j ($H_0 : \delta = 0$ in Model (2.1)) for $j = 1, \ldots m$, obtaining interaction t-statistics $T_1^{(k)}, \ldots, T_m^{(k)}$. We also tested marginal association between G_j and Y in a model with no other covariates except intercept, obtaining ordinary association t-statistics $S_1^{(k)}, \ldots, S_m^{(k)}$ as in (2.3). We obtain the interaction p-values for $T_j^{(k)}$ using the $\mathcal{T}_{n-4}$ distribution and the ordinary association p-values for $S_j^{(k)}$ using the $\mathcal{T}_{n-2}$ distribution. In this simulation, when we apply the two-sided ELL test for uniformity at level .05 to the interaction p-values from each replicate, we reject 29.3% of the 5,000 replicates as being significantly non-uniform. In contrast, when we apply the same ELL test to the ordinary association p-values from each replicate, we reject just 4.8% of the 5,000 replicates, which is not significantly different from the nominal 5% rate. This verifies that the ordinary GWAS p-values are showing the expected behavior, while the "feast or famine" effect is only showing up in the interaction p-values. This can be seen also in Fig 2.2 Panel A which depicts

a histogram of the genomic control inflation factors for each replicate for the interaction GWAS's in red and for the ordinary (non-interaction) GWAS's in purple. The narrower purple histogram reflects the expected sampling variability of the GCIF based on 5,000 i.i.d. test statistics. In contrast, the wider red histogram reflects the additional spread due to the "feast or famine effect", i.e., the fact that conditional on (Y, Z) the p-values may be systematically over- or under-inflated compared to uniform. Fig 2.2 Panel B is similar but for a simulation in which F_Z is Binomial$(2, p_{Zk})$ in replicate k instead of Bernoulli(p_{Zk}) and F_{Gj} is Binomial$(2, p_{Gj})$ instead of Bernoulli(p_{Gj}). In Figure 2.3, a similar pair of histograms can be seen for the case when both Z and G are normally distributed.

For the case when Y follows a linear mixed model, i.e., the model is as in (2.2) except that

$$\epsilon \sim N(0, \Sigma),\ \Sigma = \sigma_g^2 K + \sigma_e^2 I_n$$

where K is a GRM, it is also true that the Wald test statistic for association (i.e., the Wald test for $H_0 : \beta = 0$) is symmetric between G_j and Y when Σ is known. Thus, in this case also, ordinary GWAS association testing is essentially not affected by whether we condition on G_j and let Y be random or condition on Y and let G_j be random.

2.4 Proof of equation 2.3

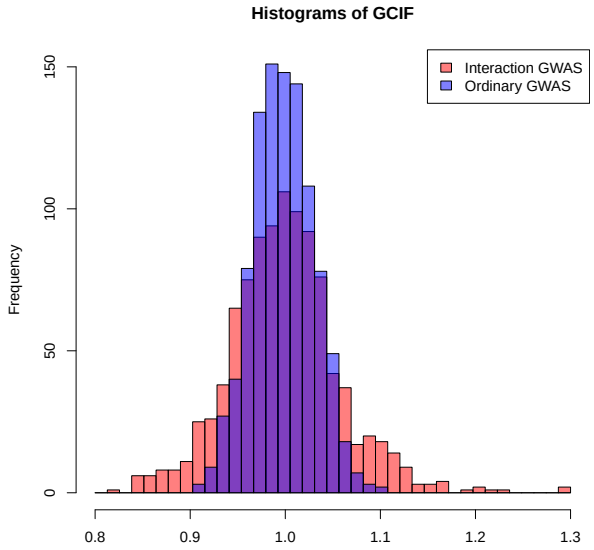
For the linear model

$$Y \sim N(U\alpha + G_j\beta, \sigma^2 I), \tag{2.5}$$

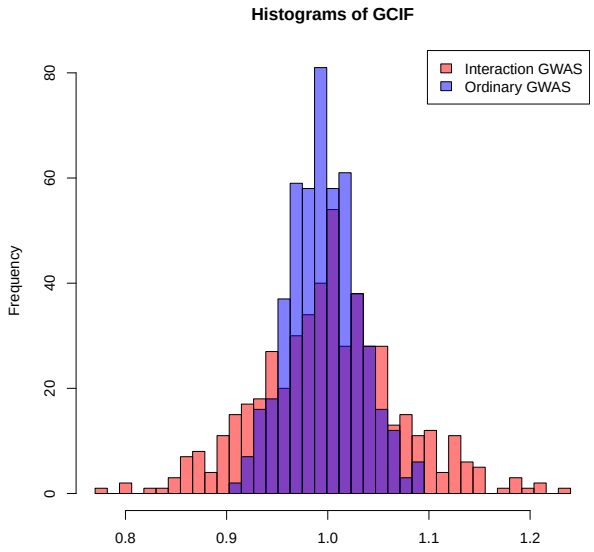
let $\hat{\beta}$ be the estimated value for β. Let $P_U = I - U(U^T U)^{-1} U^T$, note that P_U is a projection matrix so it is symmetric and idempotent. $P_U Y$ is the residual of Y after regressing out U; similarly, $P_U G_j$ is the residual of G_j after regressing out U. Then from the properties of partial regression, we know that $\hat{\beta}$ is the same as the estimated coefficient for G_j in the

Figure 2.2: **Marginal vs. interaction GCIF**

(a)

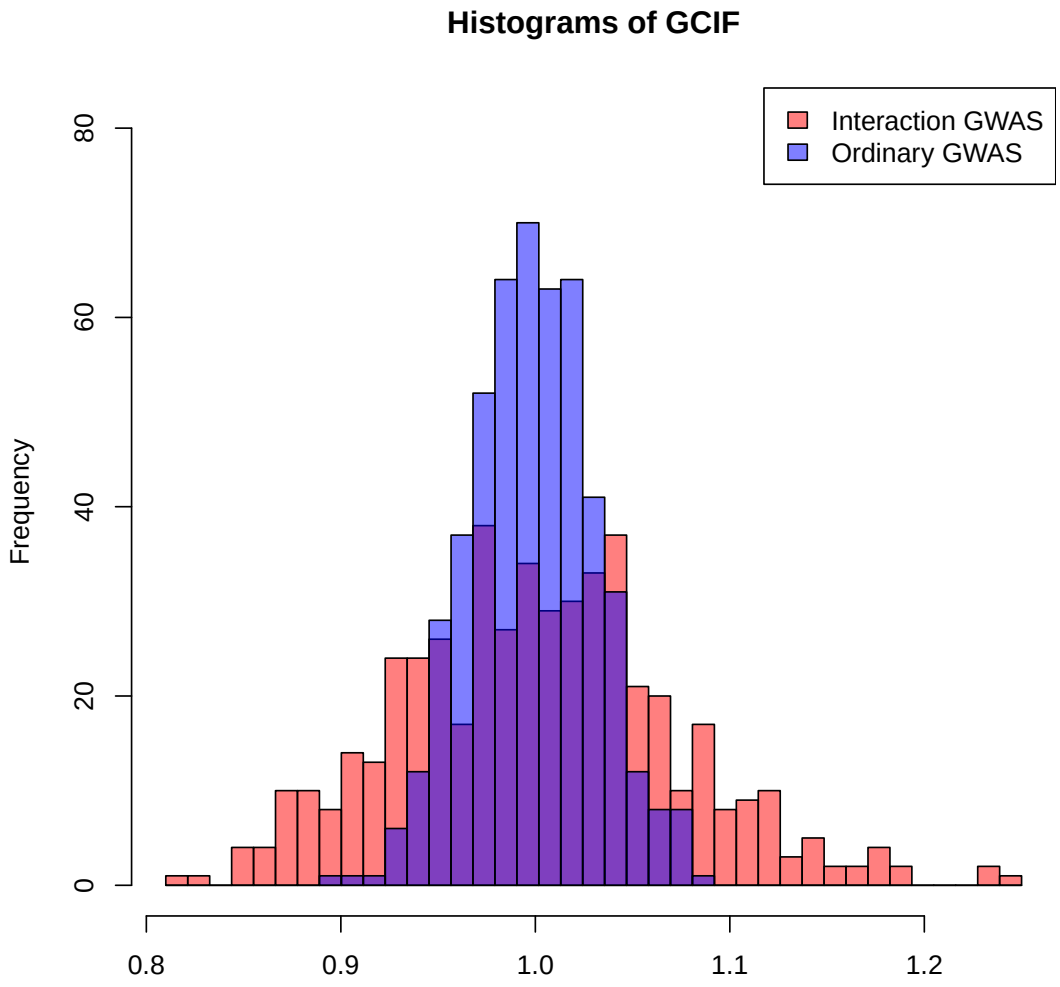


(b)



Y is simulated without GRM and under completely null. Purple: genomic control inflation factors of the 5000 marginal association tests between $G_1, ..., G_{5000}$ and Y; Red: GCIF of the 5000 interaction tests between Z and $G_1, ..., G_{5000}$. (a). Both Z and G_j's are Bernoulli distributed; (b). Both Z and G_j's are Binomial distributed with 2 trials.

Figure 2.3: **GCIF** Z Normal, G_j Normal



Y is simulated without GRM and under completely null. Purple: genomic control inflation factors of the 5000 marginal association tests between $G_1, ..., G_{5000}$ and Y; Red: GCIF of the 5000 interaction tests between Z and $G_1, ..., G_{5000}$. Both Z and G_j's are Normal distributed with mean $\sim$ Unif$(-10, 10)$ and variance 1

following linear regression

$$P_U Y \sim N(P_u G_j \theta, \omega^2 I), \tag{2.6}$$

which regress the residual of Y on the residual of G_j. Then we can get

$$\hat{\beta} = \hat{\theta} = \frac{G_j^T P_U Y}{G_j^T P_U G_j} \tag{2.7}$$

Then go back to our original model 2.5, the variance of $\hat{\beta}$ can be obtained by

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{(G_j^T P_U G_j)^2} G_j^T P_U G_j = \frac{\sigma^2}{G_j^T P_U G_j} \tag{2.8}$$

Then the t-statistic for testing $H_0 : \beta = 0$ in model 2.5 will be

$$T = \frac{\hat{\beta}}{\sqrt{\hat{\text{Var}}(\hat{\beta})}} = \frac{G_j^T P_U Y}{\sqrt{\hat{\sigma}^2 G_j^T P_U G_j}}, \tag{2.9}$$

where $\hat{\sigma}^2$ is the estimated variance from model 2.5.

Let $M = (U, G_j)$, $P_M = I - M(M^T M)^{-1} M^T$. Then P_M is the projection matrix that regresses out all predictors in model 2.5. Then

$$\hat{\sigma}^2 = \frac{Y^T P_M Y}{n - k - 1} \tag{2.10}$$

We claim that

$$P_M = P_U - \frac{P_U G_j G_j^T P_U}{G_j^T P_U G_j} := P \tag{2.11}$$

Proof

Multiply the RHS of equation 2.11 to Y, we have

$$PY = P_U Y - \frac{P_U G_j G_j^T P_U Y}{G_j^T P_U G_j} = P_U Y - \hat{\beta} P_U G_j \tag{2.12}$$

Since P_M is the projection matrix that regress M out, we have

$$P_M Y = Y - U\hat{\alpha} - G_j\hat{\beta} \tag{2.13}$$

where $\hat{\alpha},\ \hat{\beta}$ are the least square estimators of $\alpha,\ \beta$ in model 2.5. Multiply both sides of above equation by P_U, we get

$$P_U P_M Y = P_U Y - P_U U\hat{\alpha} - P_U G_j\hat{\beta} = P_U Y - P_U G_j\hat{\beta} = PY \tag{2.14}$$

since $P_U U = 0$.

Notice that

$$\begin{gathered} P_U P_M = (I - U(U^T U)^{-1}U^T)P_M = P_M - U(U^T U)^{-1}U^T P_M \\ (U(U^T U)^{-1}U^T P_M)^T = P_M U(U^T U)^{-1}U^T = 0 \end{gathered} \tag{2.15}$$

The second equation holds because M contains U as its columns and $P_M U = 0$. Then we have

$$P_M Y = PY \tag{2.16}$$

Since it holds for all $Y \in \mathbb{R}^n$, we have $P_M = P$.

Plug $P_M = P$ in the formula for $\hat{\sigma}^2$, we get

$$\begin{gathered} \hat{\sigma}^2 = \frac{1}{n-k-1}\left(Y^T P_U Y - \frac{(G_j^T P_U Y)^2}{G_j P_U G_j}\right) \\ T = \frac{G_j^T P_U Y}{\sqrt{\hat{\sigma}^2 G_j^T P_U G_j}} = \frac{G_j^T P_U Y\sqrt{n-k-1}}{\sqrt{(Y^T P_U Y)(G_j P_U G_j) - (G_j^T P_U Y)^2}} \end{gathered} \tag{2.17}$$

□

If Y has some variance structure,

$$Y \sim N(U\alpha + G_j\beta, \sigma_T^2\Sigma), \tag{2.18}$$

we could just multiply everything by $\hat{\Sigma}^{-\frac{1}{2}}$ and then apply the same procedure as the constant and independent variance case.

CHAPTER 3

METHODOLOGY

3.1 TINGA method for correcting t-statistics for interaction in a GWAS

To address the "feast or famine" effect in interaction GWAS, we propose to correct the interaction t-statistics for a given GWAS by subtracting off their null conditional mean and dividing by their conditional s.d. given the (Y, Z) observed for that GWAS. We call this approach TINGA for "Testing INteraction in GWAS with test statistic Adjustment".

In the most general case, we consider testing for interaction in the model

$$Y = U\alpha + G_j\beta + Z\gamma + (G_j \circ Z)\delta + \epsilon, \tag{3.1}$$

where Y, U, G_j, β, Z, γ, $(G_j \circ Z)$ and δ are as defined before, α is a $k \times 1$ vector of unknown coefficients, and $\epsilon \sim N(0, \Sigma\sigma_T^2)$, where either $\Sigma = I_n$ in the case of a linear model, or else $\Sigma = h^2K + (1-h^2)I_n$ where h^2 is an unknown heritability parameter, in the case of a linear mixed model. Then the t-statistic for interaction can be written as

$$T = \frac{\sqrt{n-k-3}\left(G_j \circ Z\right)^T P_M Y}{\sqrt{\left(G_j \circ Z\right)^T P_M \left(G_j \circ Z\right) \cdot Y^T P_M Y - (\left(G_j \circ Z\right)^T P_M Y)^2}} \tag{3.2}$$

where the "M" in P_M stands for "marginal", and P_M is a symmetric matrix that removes the marginal effects of G_j, Z, and U, where in the simplest case U represents just the intercept, but it may contain additional covariates as needed. We let M be the $n \times (k+2)$ matrix M whose columns are G_j, Z, and the columns of U. Then in the case of a linear model, we have $P_M = I_n - M(M^TM)^{-1}M^T$, and in the case of a linear mixed model, we have $P_M = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}M\left(M^T\hat{\Sigma}^{-1}M\right)^{-1}M^T\hat{\Sigma}^{-1}$, where $\hat{\Sigma}$ is Σ with the estimated value of h^2

plugged in. The proof is similar to the Proof of equation 2.3.

In the LMM context, the test based on T_j is commonly called the "Wald test". In fact, the ordinary t-test for interaction is also a Wald test, so this term is not a useful way of distinguishing the LMM-based test from the ordinary one. We refer to the test based on T_j as the "t-test" in both cases, and, when needed, we specify whether it is performed in an LMM or a linear model.

For both the linear and LMM cases, we define the numerator of the t-statistic to be

$$N_j \equiv N_j(G_j, Z, Y) = \left(G_j \circ Z\right)^T P_M Y. \tag{3.3}$$

Then the regular interaction t-statistic in (3.2) can be rewritten as

$$T_j = \frac{N_j - E_0(N_j|G_j, Z)}{\sqrt{\widehat{\mathrm{Var}}(N_j|G_j, Z)}} = \frac{N_j}{\sqrt{\widehat{\mathrm{Var}}(N_j|G_j, Z)}}, \text{ as } E_0(N_j|G_j, Z) = 0, \tag{3.4}$$

where both $E_0(N_j|G_j, Z)$ and $\mathrm{Var}(N_j|G_j, Z)$ are calculated based on Model (3.1), $E_0(N_j|G_j, Z)$ has the additional assumption $\delta = 0$, which is the null hypothesis, and $\widehat{\mathrm{Var}}$ denotes estimated variance.

For testing interaction in a GWAS context, we propose to replace T_j by a "corrected" statistic

$$\tilde{T}_j = \frac{N_j - \hat{E}_0(N_j|Z, Y)}{\sqrt{\widehat{\mathrm{Var}}(N_j|Z, Y)}}, \tag{3.5}$$

where the difference from Eq (3.4) is that we condition on (Z, Y) instead of on (G_j, Z). The remaining challenge of the methods development is to obtain appropriate estimators $\hat{E}_0(N_j|Z, Y)$ and $\widehat{\mathrm{Var}}(N_j|Z, Y)$.

3.2 Key idea and framework

The key idea of our proposed TINGA method is to get the correct conditional variance and null mean of the t-statistic numerator for the GWAS study design. This calculation basically contains following steps:

1. We approximate N_j by $\tilde{N}_j \equiv \tilde{N}_j(G_j, Z, Y)$, where $\tilde{N}_j$ is quadratic in G_j.
2. We calculate $E_0(\tilde{N}_j|Z,Y)$ and approximate $\mathrm{Var}(\tilde{N}_j|Z,Y)$ as functions of $E_0(G_j|Z,Y)$ and $\mathrm{Var}(G_j|Z,Y)$.
3. We calculate $E_0(G_j|Z,Y)$ and $\mathrm{Var}(G_j|Z,Y)$ theoretically based on a suitable model.
4. We obtain estimates $\hat{E}_0(G_j|Z,Y)$ and $\widehat{\mathrm{Var}}(G_j|Z,Y)$ for the quantities in step 3.
5. We plug the estimates from step 4 into the expressions for $E_0(\tilde{N}_j|Z,Y)$ and $\mathrm{Var}(\tilde{N}_j|Z,Y)$ from step 2 to obtain $\hat{E}_0(N_j|Z,Y)$ and $\widehat{\mathrm{Var}}(N_j|Z,Y)$, respectively, and calculate $\tilde{T}$ in (3.5).

3.2.1 Step 1: approximate N_j by a quadratic function of G_j

Case with no covariates Firstly we start with a simpler model for 3.1 which does not contain any covariates except the intercept:

$$Y \sim N(\alpha + \beta_1 G_j + \gamma Z + \delta(G_j \circ Z),\ \sigma_T^2 \Sigma)$$

Where $\Sigma = \frac{\sigma_g^2}{\sigma_T^2} K + \frac{\sigma_e^2}{\sigma_T^2} I$ is assumed known, $\left(G_j \circ Z\right)_k = (G_{kj} - \bar{G_j})(Z_k - \bar{Z})$. When testing for interaction between Z and each of the G_j's, we test $H_0 : \delta = 0$ *vs.* $H_1 : \delta \neq 0$.

Note that the Wald statistic will be

$$T = \frac{\sqrt{n-4}\left(G_j \circ Z\right)^T P_M Y}{\sqrt{\left(G_j \circ Z\right)^T P_M \left(G_j \circ Z\right) \cdot Y^T P_M Y - (\left(G_j \circ Z\right)^T P_M Y)^2}}$$

Where $P_M = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}M\left(M^T\hat{\Sigma}^{-1}M\right)^{-1}M^T\hat{\Sigma}^{-1},\ M = (1,\ G_j,\ Z)$. P_M is a symmetric matrix that removes the effect of M in the above linear mixed model. We are mainly interested in the numerator, which we have defined to be $N_j = \left(G_j \circ Z\right)^T P_M Y$.

For approximating N_j, it is useful to note that the matrix P_M can be calculated explicitly using an iterative method. To describe the interative method, we first consider the ordinary linear regression model $Y \sim N(M\beta,\ \sigma^2 I)$ where $\Sigma = I$ and M contains intercept term 1, variable G_j and variable Z. Then we have

$$\begin{gathered} H = P_0 = I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T \text{ is the projection matrix that projects out } \mathbf{1}, \\ P_1 = P_0 - \frac{P_0 G_j G_j^T P_0}{G_j^T P_0 G_j} \text{ is the matrix that further regresses out } G_j \\ \text{so that } (1,\ G_j) \text{are regressed out, and} \\ P_2 = P_1 - \frac{P_1 Z Z^T P_1}{Z^T P_1 Z} \text{ is the matrix that regresses out } (1,\ \ G_j,\ \ Z) \end{gathered} \tag{3.6}$$

(by a similar argument as in the Proof of equation 2.3).

To generalize above formula to the LMM case, we let $C^T C = \hat{\Sigma}^{-1}$, and define $CM = \widetilde{M} = \left(\widetilde{1},\ \widetilde{G_j},\ \widetilde{Z}\right)$. Then $CY = \widetilde{Y} \sim \left(\widetilde{M}\beta + \gamma\widetilde{\left(G_j \circ Z\right)},\ \sigma_T^2 I\right)$, where $\widetilde{w}$ denotes Cw for any vector w.

Then we can write

$$P_M = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}M\left(M^T\hat{\Sigma}^{-1}M\right)^{-1}M^T\hat{\Sigma}^{-1} = C^T C - C^T\widetilde{M}\left(\widetilde{M}^T\widetilde{M}\right)^{-1}\widetilde{M}C$$

and

$$\left(G_j \circ Z\right)^T P_M Y = \widetilde{\left(G_j \circ Z\right)}^T\left(I - \widetilde{M}\left(\widetilde{M}^T\widetilde{M}\right)^{-1}\widetilde{M}\right)\widetilde{Y} = \widetilde{\left(G_j \circ Z\right)}^T P_{\widetilde{M}}\widetilde{Y}$$

Where $P_{\widetilde{M}} = I - \widetilde{M}\left(\widetilde{M}^T\widetilde{M}\right)^{-1}\widetilde{M}$.

Note $P_{\widetilde{M}}$ is the matrix that regresses out $\widetilde{M}$ in the simple linear model, so we can compute it using the above iterative method:

$$\begin{aligned} \widetilde{H} = P_{0\prime} &= I - \widetilde{1}\left(\widetilde{1}^T\widetilde{1}\right)^{-1}\widetilde{1}^T \\ P_1' &= P_0' - \frac{P_{0\prime}\widetilde{G_j}\widetilde{G_j}^T P_{0\prime}}{\widetilde{G_j}^T P_0'\widetilde{G_j}} \\ P_{\widetilde{M}} &= P_1' - \frac{P_1'\widetilde{Z}\widetilde{Z}^T P_{1\prime}}{\widetilde{Z}^T P_{1\prime}\widetilde{Z}} \end{aligned} \tag{3.7}$$

Therefore, we can iteratively compute an explicit expression for $P_{\widetilde{M}}$:

$$P_1' = \widetilde{H} - \frac{\left(\widetilde{H}\widetilde{G_j}\right)\left(\widetilde{H}\widetilde{G_j}\right)^T}{\widetilde{G_j}^T\widetilde{H}\widetilde{G_j}},$$

and

$$P_{\widetilde{M}} = \widetilde{H} - \frac{\left(\widetilde{H}\widetilde{G_j}\right)\left(\widetilde{H}\widetilde{G_j}\right)^T}{\widetilde{G_j}^T\widetilde{H}\widetilde{G_j}} - \frac{(\widetilde{H}\widetilde{Z} - \frac{\left(\widetilde{H}\widetilde{G_j}\right)\left(\widetilde{H}\widetilde{G_j}\right)^T\left(\widetilde{H}\widetilde{Z}\right)}{\widetilde{G_j}^T\widetilde{H}\widetilde{G_j}})(\widetilde{Z}^T\widetilde{H} - \frac{(\widetilde{Z}^T\widetilde{H})\left(\widetilde{H}\widetilde{G_j}\right)\left(\widetilde{H}\widetilde{G_j}\right)^T}{\widetilde{G_j}^T\widetilde{H}\widetilde{G_j}})}{\widetilde{Z}^T\widetilde{H}\widetilde{Z} - \frac{(\left(\widetilde{H}\widetilde{Z}\right)^T\left(\widetilde{H}\widetilde{G_j}\right))^2}{\widetilde{G_j}^T\widetilde{H}\widetilde{G_j}}}$$

Let $S_{ab} = \left(\widetilde{H}a\right)^T\left(\widetilde{H}b\right)$ for any vectors $a,\ b$, then

$$\widetilde{(G_j \circ Z)}^T P_{\widetilde{M}}\widetilde{Y} = S_{\widetilde{(G_j\circ Z)}\widetilde{Y}} - \frac{S_{\widetilde{(G_j\circ Z)}\widetilde{G_j}}\ S_{\widetilde{G_j}\widetilde{Y}}\ S_{\widetilde{Z}\widetilde{Z}} + S_{\widetilde{(G_j\circ Z)}\widetilde{Z}}\ S_{\widetilde{Z}\widetilde{Y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{(G_j\circ Z)}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Y}} - S_{\widetilde{(G_j\circ Z)}\widetilde{G_j}}\ S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{Y}\widetilde{Z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{G_j}}} \tag{3.8}$$

Note this formula is the same as the formula for non-GRM case, except that S_{ab} is the simple inner product in that case. We can then transform these variables back and get an

expression in terms of the original variables:

$$
\begin{aligned}
S_{\widetilde{G_j}\widetilde{Y}} &= \widetilde{G_j}^T \widetilde{H}\widetilde{Y} = G_j^T C^T \left(I - \widetilde{1}\left(\widetilde{1}^T\widetilde{1}\right)^{-1}\widetilde{1}^T \right) CY \\
&= G_j^T \left(\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}1\left(1^T\hat{\Sigma}^{-1}1\right)^{-1}1\hat{\Sigma}^{-1} \right) Y \\
&= G_j^T \hat{H} Y
\end{aligned}
$$

Where $\hat{H} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}1\left(1^T\hat{\Sigma}^{-1}1\right)^{-1}1\hat{\Sigma}^{-1}$. Similarly for other S_{ab} terms. Then we reorganize these terms and get

$$
\widetilde{(G_j \circ Z)}^T P_{\widetilde{M}}\widetilde{Y} = (G_j \circ Z)^T \hat{H}\left(Y - \frac{S_{\widetilde{Z}\widetilde{Y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Y}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}} Z \right) - \frac{S_{\widetilde{G_j}\widetilde{Y}}\ S_{\widetilde{Z}\widetilde{Z}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{Y}\widetilde{Z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}\ (G_j \circ Z)^T \hat{H} G_j \tag{3.9}
$$

It is worth noting that the 2 fraction terms, $\frac{S_{\widetilde{Z}\widetilde{Y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Y}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}$ and $\frac{S_{\widetilde{G_j}\widetilde{Y}}\ S_{\widetilde{Z}\widetilde{Z}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{Y}\widetilde{Z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}$, are actually the estimated coefficients from the marginal effect model

$$
Y \sim N(\alpha + \beta G_j + \gamma Z, \sigma^2\Sigma) \tag{3.10}
$$

where

$$
\begin{aligned}
\hat{\gamma} &= \frac{S_{\widetilde{Z}\widetilde{Y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Y}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}} \\
\hat{\beta} &= \frac{S_{\widetilde{G_j}\widetilde{Y}}\ S_{\widetilde{Z}\widetilde{Z}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{Y}\widetilde{Z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}
\end{aligned} \tag{3.11}
$$

(see proof in Proof of equation 3.11).

Then we have

$$
\widetilde{\left(G_j \circ Z\right)}^T P_{\widetilde{M}}\widetilde{Y} = (G_j \circ Z)^T \hat{H}(Y - \hat{\gamma}Z - \hat{\beta}G_j) \tag{3.12}
$$

which verifies our calculation.

We simplify terms $\frac{S_{\widetilde{Z}\widetilde{Y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{G_j}\widetilde{Y}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}$ and $\frac{S_{\widetilde{G_j}\widetilde{Y}}\ S_{\widetilde{Z}\widetilde{Z}} - S_{\widetilde{G_j}\widetilde{Z}}\ S_{\widetilde{Y}\widetilde{Z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{Z}\widetilde{Z}} - S^2_{\widetilde{G_j}\widetilde{Z}}}$ by dividing both numerator and denominator by n^2 and approximating $\frac{S_{ab}}{n}$ terms by their asymptotic means

under the null conditional distribution of G_j given $(Y = y, Z = z)$, where the lower case y, z denote the observed values of Y, Z.

Assume G_j has the conditional mean and variance

$$\begin{aligned} E\left(G_j|(Y=y,\ Z=z)\right) &= \mu_2, \\ \text{Var}\left(G_j|(Y=y,\ Z=z)\right) &= V_2 \end{aligned} \tag{3.13}$$

, where $\mu_2,\ V_2$ are some functions of y, z and independent of G_j.

We have, under some regularity conditions, that for large n, by law of large number:

$$\begin{aligned} \frac{S_{\widetilde{G_j}\widetilde{G_j}}}{n} = \frac{G_j^T \hat{H} G_j}{n} &\sim \frac{1}{n}\left(\mu_2^T \hat{H} \mu_2\right) + \frac{1}{n} tr\left(\hat{H} V_2\right) \\ \frac{S_{\widetilde{G_j}\widetilde{z}}}{n} = \frac{G_j^T \hat{H} z}{n} &\sim \frac{1}{n}\left(\mu_2^T \hat{H} z\right) \\ \frac{S_{\widetilde{G_j}\widetilde{y}}}{n} = \frac{G_j^T \hat{H} y}{n} &\sim \frac{1}{n}\left(\mu_2^T \hat{H} y\right) \end{aligned}$$

Then we could replace the "S" terms involving G_j by their asymptotic approximates and get simplified versions of $\frac{S_{\widetilde{z}\widetilde{y}}\ S_{\widetilde{G_j}\widetilde{G_j}} - S_{\widetilde{G_j}\widetilde{z}}\ S_{\widetilde{G_j}\widetilde{y}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{z}\widetilde{z}} - S^2_{\widetilde{G_j}\widetilde{z}}}$ and $\frac{S_{\widetilde{G_j}\widetilde{y}}\ S_{\widetilde{z}\widetilde{z}} - S_{\widetilde{G_j}\widetilde{z}}\ S_{\widetilde{y}\widetilde{z}}}{S_{\widetilde{G_j}\widetilde{G_j}}\ S_{\widetilde{z}\widetilde{z}} - S^2_{\widetilde{G_j}\widetilde{z}}}$, denoted by $\alpha_1,\ \alpha_2$, respectively.

Note $G_j \circ z = D_{Hz} H G_j$, where D_{Hz} is the diagonal matrix whose diagonal entries are Hz. Then we obtain an approximation to the numerator of the t statistic that is a quadratic function of G_j:

$$\begin{aligned} N_j|(Y=y, Z=z) \approx \tilde{N}_j &= G_j^T H D_{Hz} \hat{H}\left(y - \alpha_1 z\right) - \alpha_2 G_j^T H D_{Hz} \hat{H} G_j \\ &= G_j^T B G_j + b^T G_j \end{aligned} \tag{3.14}$$

where

$$
\begin{aligned}
H &= I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T \\
\hat{H} &= \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}1\left(1^T\hat{\Sigma}^{-1}1\right)^{-1}1\hat{\Sigma}^{-1} \\
B &= -\alpha_2 H D_{Hz}\hat{H} \\
b &= H D_{Hz}\hat{H}(y - \alpha_1 z) \\
D_{Hz} &= \mathrm{Diag}(Hz) = \mathrm{Diag}(z_1 - \bar{z}, ..., z_n - \bar{z})
\end{aligned}
\tag{3.15}
$$

Case with population structure and covariates Next, we consider a more general scenario in which there are additional covariates in the model. In the case, to derive the asymptotic approximation, we could first project out the covariates and the rest will be the same.

Suppose $Y \sim N(A\alpha + \beta_1 G_j + \gamma Z + \delta\left(G_j \circ Z\right),\ \sigma_T^2\Sigma)$, where $A \in \mathbb{R}^{n\times k}$ is the covariate matrix, k<n. Here we assume A contains the intercept term.

We want to first eliminate the covariate terms. Let $A^c \in \mathbb{R}^{(n-k)\times n}$. Rows of A^c are linearly independent vectors in the orthogonal complement of the column space of A. Therefore, $A^cA = 0_{(n-k)\times k}$.

In practice, we can get A^c by SVD : $A = U\Lambda V,\ U \in \mathbb{R}^{n\times n},\ \Lambda \in \mathbb{R}^{n\times k},\ V \in \mathbb{R}^{k\times k}$. Then $U[,\ (k+1) : n]^T$ can be our A^c because U is an orthogonal matrix. After getting A^c, we multiply Y by it:

$$
A^cY := Y^r \sim N(\beta_1 {G_j}^r + \gamma Z^r + \delta(G_j \circ Z)^r, \sigma_T^2 A^c\Sigma A^{cT}),
$$

where $w^r = A^cw$ for any vector w.
Note that A^c has full row rank, so $A^c\Sigma A^{cT}$ is positive definite and a valid variance matrix.

Now the model is very similar to the no-covariate part with variables $Y^r,\ {G_j}^r,\ Z^r,\ \left(G_j \circ Z\right)^r$ and variance matrix being $A^c\Sigma A^{cT}$, except that there is no more intercept term. Therefore, in the first step of forming $P_{\widetilde{U}}$, we do not need to regress out the 1 term and hence $\widetilde{H} = I$.

Then

$$S_{\widetilde{G_j^r}\widetilde{Y^r}} = \left(\widetilde{H}\widetilde{G_j}^r\right)^T \left(\widetilde{H}\widetilde{Y}^r\right) \quad = \widetilde{G_j}^{rT}\widetilde{Y}^r = G_j^{rT}C^TCY^r = G_j^{rT}\left(\widehat{A^c\Sigma A}\right)^{-1}Y^r$$
$$= G_j^TA^{cT}\left(\widehat{A^c\Sigma A^{cT}}\right)^{-1}A^cY := G_j^T\hat{H}Y \tag{3.16}$$

Where $\hat{H} = A^{cT}\left(\widehat{A^c\Sigma A^{cT}}\right)^{-1}A^c$.

After getting the new $\hat{H}$, the calculation of conditional mean and variance follows the same procedure.

3.2.2 Step 2: calculate $E_0(\tilde{N}_j|Z,Y)$ and$Var(\tilde{N}_j|Z,Y)$ as functions of $E_0(G_j|Z,Y)$ and $Var(G_j|Z,Y)$.

Assume that G_j has the conditional mean and variance:

$$\begin{aligned} E\left(G_j|(Y=y,\ Z=z)\right) &= \mu_2, \\ \text{Var}\left(G_j|(Y=y,\ Z=z)\right) &= V_2 \end{aligned} \tag{3.17}$$

where $\mu_2 \in \mathbb{R}^{n\times 1}$ and $V_2 \in \mathbb{R}^{n\times n}$. From equation 3.14, we have

$$\tilde{N}_j = G_j^TBG_j + b^TG_j. \tag{3.18}$$

Then the conditional expectation of $\tilde{N}_j$ is straightforward:

$$\begin{aligned} E(\tilde{N}_j|Y=y,Z=z) &= E(G_j^TBG_j + b^TG_j|Y=y,\ Z=z) \\ &= \mu_2^TB\mu_2 + tr\left(BV_2\right) + b^T\mu_2 \end{aligned} \tag{3.19}$$

The variance is a bit more complicated, and we have 2 approaches to calculate it: approximation approach and exact approach.

Approximation approach

We treat $G_j|(Y = y,\ Z = z)$ as if it follows a multivariate normal distribution and we apply the properties of the quadratic form of the multivariate normal variable to obtain the variance. This approach is exact for the case when $G_j|Z, Y$ has a normal distribution and is otherwise approximate.

Note that the quadratic form of multivariate normal variable requires the matrix $B = -\alpha_2 H D_{Hz} \hat{H}$ to be symmetric. In the case of independent individuals, we have $\hat{H} = H$ and B is indeed symmetric. In the case of non-independent individuals, we can re-write the quadratic term as

$$G_j^T B G_j = G_j^T B_s G_j \tag{3.20}$$

where

$$B_s = \frac{B + B^T}{2}$$

is symmetric. Then we can get the (approximated) conditional variance of $\tilde{N}_j$:

$$\begin{aligned} \text{var}\left(G_j^T B_s G_j + b^T G_j \middle| Y = y,\ Z = z\right) = 2tr\left(B_s V_2 B_s V_2\right) + 4\mu_2^T B_s V_2 B_s \mu_2 + b^T V_2 b \\ +2cov(G_j^T B_s G_j,\ b^T G_j | Y = y,\ Z = z) \end{aligned} \tag{3.21}$$

Note that when $w \sim N\left(0,\ I\right),\ cov\left(w^T B_s w,\ b^T w\right) = 0$. Write $G_j = \mu_2 + Vw$, where $w|y,\ z \sim N\left(0,\ I\right),\ VV^T = V_2$,

$$cov\left(G_j^T B_s G_j,\ b^T G_j \middle| y,\ z\right) = 2\mu_2^T B_s V V^T b = 2\mu_2^T B_s V_2 b$$

Therefore,

$$\begin{aligned} var\left(G_j^T B_s G_j + b^T G_j \middle| y,\ z\right) = 2tr\left(B_s V_2 B_s V_2\right) + \\ 4\mu_2^T B_s V_2 B_s \mu_2 + b^T V_2 b + 4\mu_2^T B_s V_2 b \end{aligned} \tag{3.22}$$

Exact approach (Bernoulli case)

In the above approach, we estimated the variance by treating $G|Y, Z$ as Gaussian distributed and applying the distribution properties of quadratic forms of multivariate normal. However, this approximation may not be accurate when G follows a Bernoulli distribution. In particular, Gaussian and Bernoulli random variables have different 3rd and 4th moments. Here we give the calculation result that uses the Bernoulli property of $G|Y, Z$ to compute the conditional variance of $\tilde{N}$.

We assume that conditional on $Y = y, Z = z$, G_j follows a Bernoulli distribution with mean vector μ and variance matrix V, where the vector of diagonal entries of V is $\mu(1-\mu)$.

Firstly, we define array $(M)_{ijk} \in \mathbb{R}^{n\times n\times n}$ and $(M)_{ijkl} \in \mathbb{R}^{n\times n\times n\times n}$ to be the 3rd and 4th central moments of $G|Y, Z$:

$$\begin{aligned} M_{ijk} &= E[(G_i - \mu_i)(G_j - \mu_j)(G_k - \mu_k)|Y, Z] \\ M_{ijkl} &= E[(G_i - \mu_i)(G_j - \mu_j)(G_k - \mu_k)(G_l - \mu_l)|Y, Z] \end{aligned} \tag{3.23}$$

We can compute the special cases in which there are at most 2 distinct indices:

$$\begin{aligned} M_{iij} &= (1 - 2\mu_i)V_{ij} \\ M_{iii} &= V_{ii}(1 - 2\mu_i) \\ M_{iijj} &= (1 - 2\mu_i)(1 - 2\mu_j)V_{ij} + V_{ii}V_{jj} \\ M_{iiij} &= (1 - 3\mu_i - 3\mu_i^2)V_{ij} \\ M_{iiii} &= (1 - 3\mu_i - 3\mu_i^2)V_{ii} \end{aligned} \tag{3.24}$$

For other cases where there are more than 2 distinct indices (i.e., individuals), we approximate M_{ijk}, M_{ijkl} by 0.

To take advantage of the fact that for a Bernoulli random variable, we have $G_i^2 = G_i$. It

is helpful to directly expand the matrix product

$$\begin{gathered}
\text{var}(G^T BG + b^T G|Y, Z) = \text{var}\left(\sum_{i,j} B_{ij}G_iG_j + \sum_k b_k G_k|Y, Z\right) \\
= \text{var}(\sum_{i,j} B_{ij}G_iG_j|Y, Z) + \text{var}(\sum_k b_k G_k|Y, Z) + 2\text{cov}(\sum_{i,j} B_{ij}G_iG_j, \sum_k b_k G_k|Y, Z) \\
= ... \\
= 4\mu^T BVB\mu - \text{tr}(BV)^2 + 4\mu^T BVb + b^T Vb \\
+4\sum_{ijkl} \mu_j B_{ij}B_{kl}M_{ikl} + \sum_{ijkl} B_{ij}B_{kl}M_{ijkl} + 2\sum_{ijk} b_k B_{ij}M_{ijk}
\end{gathered} \tag{3.25}$$

Independent individuals When we assume unrelated individuals, V is a diagonal matrix. In this case, we can simplify equation 3.25 to

$$\begin{gathered}
\text{var}(G^T BG + b^T G|Y, Z) = 4\mu^T BVB\mu - tr(BV)^2 + 4\mu^T BVb + b^T Vb \\
+\sum_j B_{jj}M_{jjj}\left(4\sum_i \mu_i B_{ij} + 2b_j\right) + \sum_{ij} M_{iijj}\left(2B_{ij}^2 + B_{ii}B_{jj}\right) - 2\sum_i B_{ii}^2 M_{iiii} \\
= 4\mu^T BVB\mu - tr(BV)^2 + 4\mu^T BVb + b^T Vb+ \\
\sum_j B_{jj}V_{jj}(1 - 2\mu_j)\left(4(\mu^T B)_j + 2b_j\right) + \sum_i B_{ii}^2 V_{ii}(1 - 6\mu_i + 6\mu_i^2) \\
+\sum_{i,j} V_{ii}V_{jj}(2B_{ij}^2 + B_{ii}B_{jj})
\end{gathered} \tag{3.26}$$

Related individuals If the individuals are related, then V is not diagonal and we have a more complicated result

$$\begin{gathered}
\text{var}(G^T BG + b^T G|Y, Z) = 4\mu^T BVB\mu - tr(BV)^2 + 4\mu^T BVb + b^T Vb \\
= +8\sum_j (BV)_{jj}(1 - 2\mu_j)(\mu^T B)_j + 4\sum_k B_{kk}(1 - 2\mu_k)(VB\mu)_k \\
-8\sum_j B_{jj}V_{jj}(1 - 2\mu_j)(B\mu)_j + \sum_{ij}(2B_{ij}^2 + B_{ii}B_{ij})[(1 - 2\mu_i)(1 - 2\mu_j)]V_{ij} + V_{ii}V_{jj}] \\
+4\sum_i B_{ii}(1 - 3V_{ii})(BV)_{ii} - 6\sum_i B_{ii}^2 V_{ii}(1 - 3V_{ii}) \\
+2\sum_i b_i(1 - 2\mu_i)(BV)_{ii} + 2\sum_i B_{ii}(1 - 2\mu_i)(Vb)_i - 4\sum_i B_{ii}(1 - 2\mu_i)b_i V_{ii}
\end{gathered} \tag{3.27}$$

3.2.3 Steps 3-4: Estimation of the conditional distribution $G_j|Y,Z$

For steps 3-4, we have 2 ways to compute the conditional distribution $G_j|Y,Z$:

1. Gaussian approximation approach

   We assume a normal regression model for $G_j|Z$, i.e., we take $G_j = a1_n + bZ + \eta$, where $\eta \sim N_n(0, \sigma_j^2 I_n)$, or, more generally, where $\tilde{U}$ consists of the intercept and any confounding covariates that are in U, we take $G_j = a\tilde{U} + bZ + \eta$, with a, b, and σ_j^2 unknown. We also assume that $(G_j, Y)|Z$ follows a multivariate normal distribution. Then, $E_0(G_j|Z,Y)$ and $\mathrm{Var}(G_j|Z,Y)$ can be easily computed using standard properties of multivariate normal.

   The basic idea uses the

$$Y = \alpha + \gamma Z + \beta G_j + \delta(G_j - m_{G_j})(Z - m_Z) + \epsilon$$

$$\begin{pmatrix} G_j \\ Y \end{pmatrix} |z \sim N\left(\begin{pmatrix} \mu_{G_j|z} \\ \mu_{y|z} \end{pmatrix}, \begin{pmatrix} \sigma^2_{G_j|z} & (\beta + \delta z_c)\sigma^2_{G_j|z} \\ (\beta + \delta z_c)\sigma^2_{G_j|z} & v_{y|z} \end{pmatrix} \right) \tag{3.28}$$

   where

$$\mu_{G_j|z} = a + bz$$

$$\sigma^2_{G_j|z} = \sigma_j^2$$

   Then we can compute the the conditional mean and variance of $G_j|Y,Z$ using the multivariate normal approximation:

$$G_j|\,(y,\ z) \sim\ N(\mu_{G_j|z} + \frac{(\beta + \delta z_c)\,\sigma^2_{G_j|z}}{v_{y|z}}(y - \mu_{y|z}),\ \sigma^2_{G_j|z} - \frac{(\beta + \delta z_c)^2\,\sigma^4_{G_j|z}}{v_{y|z}}) \tag{3.29}$$

   The parameters such as $a, b, \sigma_j^2, \beta, \delta, v_{y|z}, \mu_{y|z}$ are estimated by fitting the associated

linear and/or linear mixed models. Details are in Chapter 4 "Detailed steps for parameter estimation".

2. Approach that uses a discrete model for G_j (e.g. Bernoulli)

   Alternatively, we can use a discrete model for $G_j|Z$, where we assume that conditional on Z, the n entries of the vector G_j, call them $G_{1j}, \ldots, G_{nj}$, are independent with $P(G_{ij} = k|Z_i = z) = p_{k|z}$ for all choices of (i, k, z), where these may also depend on $\tilde{U}$ as needed. Since G_j is a genotype, we will have $k \in \{0, 1, 2\}$ when the genotypes are from a diploid organism or $k \in \{0, 1\}$ when the genotypes are from a haploid organism or inbred line. For the latter case, we can use a logistic regression model for $G_j|Z$, and for the former case a generalized linear model.

   We can apply the Bayes rule to obtain the discrete distribution of $\Pr(G_j|Z, Y)$. For example, if we assume unrelated individuals, then

$$P(G_{ij} = k|Z, Y) = P(G_{ij} = k|Z_i, Y_i) = \frac{P\left(Y_i\middle|G_{ij} = k,\ Z_i\right) * p_{k|Z_i}}{\sum_l P\left(Y_i\middle|G_{ij} = l,\ Z_i\right) * p_{l|Z_i}}, \tag{3.30}$$

   $P(Y|G_j, Z)$ can be estimate by fitting a model of Y on (G_j, Z)

   For the case where there is some population structure, we can still estimate the mean and variance for each individual by 3.30 and estimate the conditional covariance $\mathrm{Cov}(G_j|Y, Z)$ by incorporating the GRM. Details are in Chapter 4 " Detailed steps for parameter estimation".

3.2.4 Proof of equation 3.11

For simplicity, we assume a simple linear model

$$Y \sim N(\alpha + \beta G_j + \gamma Z, \sigma^2 I) \tag{3.31}$$

Let $U = (1, G_j),\ P_U = I - U(U^TU)^{-1}U^T$, then by properties of partial regression, we have

$$\hat{\gamma} = \frac{Z^T P_U Y}{Z^T P_U Z} \tag{3.32}$$

By equations 3.6, we have

$$\begin{aligned} H = P_0 &= I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T \\ P_U &= P_0 - \frac{P_0 G_j G_j^T P_0}{G_j^T P_0 G_j} \end{aligned} \tag{3.33}$$

Plugging in the expression 3.32 for $\hat{\gamma}$, we get

$$\begin{aligned} Z^T P_U Y &= Z^T H Y - \frac{Z^T H G_j G_j^T H Y}{G_j^T H G_j} = S_{ZY} - \frac{S_{ZG_j} S_{G_j Y}}{S_{G_j G_j}} \\ Z^T P_U Z &= Z^T H Z - \frac{Z^T H G_j G_j^T H Z}{G_j^T H G_j} = S_{ZZ} - \frac{S_{ZG_j} S_{G_j Z}}{S_{G_j G_j}} \\ \hat{\gamma} &= \frac{S_{ZY} S_{G_j G_j} - S_{ZG_j} S_{G_j Y}}{S_{ZZ} S_{G_j G_j} - S_{G_j Z}^2} \end{aligned} \tag{3.34}$$

Similarly for $\hat{\beta}$.

Then suppose there is some covariance structure

$$Y \sim N(\alpha + \beta G_j + \gamma Z, \sigma^2 \Sigma) \tag{3.35}$$

then let $C^T C = \Sigma^{-1}$, then

$$C^{-1} C^{-T} = \Sigma,\ C \Sigma C^T = I \tag{3.36}$$

and

$$\begin{aligned} CY &\sim N(\alpha C\mathbf{1} + \beta C G_j + \gamma C Z, \sigma^2 I) \\ \widetilde{Y} &\sim N(\alpha \tilde{\mathbf{1}} + \beta \widetilde{G_j} + \gamma \widetilde{Z}, \sigma^2 I) \end{aligned} \tag{3.37}$$

Then the result follow the same argument as the independent individual case by replacing $\mathbf{1}, G_j, Z, Y$ with $\tilde{\mathbf{1}}, \widetilde{G_j}, \widetilde{Z}, \widetilde{Y}$.

□

3.3 Adjustments and extensions

3.3.1 Heteroscadasticity correction

In an interaction GWAS, it can potentially be important to consider a specific type of heteroscedasticity that arises naturally in a model in which Z interacts with some other variable in a linear model or LMM for Y, even if it does not interact with G_j [46; 43; 44; 45]. That is, suppose the true model is

$$Y = U\alpha + G_j\beta + Z\gamma + X\zeta + (X \circ Z)\theta + \epsilon, \tag{3.38}$$

where Y, U, α, G_j, β, Z, γ, and ϵ are as before, $(X \circ Z) = (X - \mu_X)(Z - \mu_Z)$, ζ and θ are unknown scalar coefficients, and X is some additional variable that is not included in the fitted model (and that might or might not even be observed), is independent of (G_j, Z), and that interacts with Z. Then from the point of view of testing for interaction between Z and G_j using fitted model 3.1, the null hypothesis of no interaction is true. However, when Z is interacting with some other known or unknown X, the conditional variance of $Y|Z$ is

$$Var(Y|Z) = \beta^2\sigma^2_{G_j} + (\zeta + \theta(Z - \mu_Z))^2\,\sigma^2_X + \sigma^2_\epsilon$$

which is a quadratic function of Z. This heteroscedasticity tends to lead to inflated type I error [45] if it is not accounted for in the fitted model.

In TINGA, for each G_j we correct for heteroscedasticity by first regressing G_j out of Y to get the residual $r_{Y|G_j}$, and we replace Y by $r_{Y|G_j}$ in all rest steps of the calculation and fit a heteroscedastic model of Y what allows Y to have different variances for different Z values. The details are in Appendix: Heteroscedasticity correction strategy.

3.3.2 Compute variance under alternative model

When estimating $G_j|Y,Z$, for both the normal approximation approach (equation 3.29) and the discrete model approach (equation 3.30), we need to fit a linear model of $Y|G_j,Z$ to get estimates of the parameters. Then we have 2 options for this: one is fitting the null model

$$Y|Z,G_j \sim N(\alpha+\beta G_j+\gamma Z,\sigma^2 I) \tag{3.39}$$

and one is fitting the alternative model

$$Y|Z,G_j \sim N(\alpha+\beta G_j+\gamma Z+\delta(G_j\circ Z),\sigma^2 I). \tag{3.40}$$

Consider the regular t-statistic in testing for the interaction term $(G_j\circ Z)$ in the model

$$Y|Z,G_j \sim N(\alpha+\beta G_j+\gamma Z+\delta(G_j\circ Z),\sigma^2 I)=N(Xb,\sigma^2 I)$$

where $X=(1,G_j,Z,(G_j\circ Z))$:

$$t=\frac{\hat{\delta}-0}{\sqrt{\hat{\sigma}^2(XX^T)^{-1}_{\delta,\delta}}}=\frac{\hat{\delta}-E\left[\hat{\delta}|null\right]}{\sqrt{var(\hat{\delta}|alt)}}, \tag{3.41}$$

the mean is under the null ($\delta=0$), but the variance $\hat{\sigma}^2$ is estimated under the alternative model 3.40 so that we gain more power. In analogous to t-statistic, we could also estimate the conditional distribution of G_j under alternative model for $Var(\tilde{N}_j|Y,Z)$.

3.3.3 Summary of different versions

Firstly, we have 2 ways for estimating the conditional distribution of $G_j|Y,Z$, we denote them as

1. Bernoulli approach: directly estimate the Bernoulli distribution using equation 3.30

2. Gaussian approach: estimate a conditional multivariate normal distribution by equation 3.29

Depending on whether to apply heteroscedasticity correction and whether to fit alternative model of Y for $\text{Var}(\tilde{N}_j|Z,Y)$, our adjustment method has 4 versions:

1. Method1: No heteroscedasticity correction and estimate $var(\tilde{N}_j|Y,Z)$ using the null model

2. Method2: Conduct heteroscedasticity correction and estimate $var(\tilde{N}_j|Y,Z)$ using the null model

3. Method3: Conduct heteroscedasticity correction and estimate $var(\tilde{N}_j|Y,Z)$ using the alternative model

4. Method4: No heteroscedasticity correction and estimate $var(\tilde{N}_j|Y,Z)$ using the alternative model

They can be summarized in Table 3.1

	$\widehat{\text{Var}}(N_j\|Z,Y)_{null}$	$\widehat{\text{Var}}(N_j\|Z,Y)_{alt}$
No heteroscedasticity correction	Method1	Method4
Heteroscedasticity correction	Method2	Method3

Table 3.1: 4 Methods bases on model of Y

See details in Section Detailed steps for parameter estimation.

3.3.4 Appendix: Heteroscedasticity correction strategy

Non-GRM case

Suppose we have $Y, Z, G_1, ..., G_m \in \mathbb{R}^n$ and we want to test for the interaction in the model

$$Y = \alpha + G_j\beta + Z\gamma + (G_j \circ Z)\delta + \epsilon \tag{3.42}$$

where we allow Y to have different variances for different values of Z.

For each G_j, we first test for its marginal association with Y in the model

$$Y = \alpha + G_j\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I) \tag{3.43}$$

If the p-value for testing $H_{0j} : \beta = 0$ is smaller than 10^{-3}, we let

$$r_{Y|G_j} = Y - \hat{Y}$$

where $\hat{Y} = \hat{\alpha} + G_j\hat{\beta}$ is the fitted value of model 3.43, so $r_{Y|G_j}$ is the residual of Y after regressing out G_j. Then we replace Y by $r_{Y|G_j}$ in all steps of our adjustment method with heteroscedasticity correction applied. That is to say, we are testing the interaction term in the heteroscedastic model

$$r_{Y|G_j} = \alpha + G_j\beta + Z\gamma + (G_j \circ Z)\delta + \epsilon,\ \epsilon \sim N(0, \sigma_0^2\mathbb{I}_{Z=0} + \sigma_1^2\mathbb{I}_{Z=1})$$

and apply our adjustment method.

If the p-value is larger than 10^{-3}, we use the original Y with heteroscedasticity correction applied.

GRM case

The the case where the individuals are related by a GRM K, the heteroscedasticity correction is similar to the non-GRM case, the only difference is that we fit LMMs instead of linear models.

For each G_j, we first test for its marginal association with Y in the model

$$Y = \alpha + G_j\beta + \epsilon,\ \epsilon \sim N(0, \sigma_e^2 I + \sigma_g^2 K) \tag{3.44}$$

If the p-value for the Wald-test on $H_{0j} : \beta = 0$ is smaller than 10^{-3}, we let

$$r_{Y|G_j} = Y - \hat{Y}$$

where $\hat{Y} = \hat{\alpha} + G_j\hat{\beta}$ is the fitted value of model 3.44. Again we replace Y by $r_{Y|G_j}$ in all steps of our adjustment method with heteroscedasticity correction applied.

If the p-value is larger than 10^{-3}, we use the original Y with heteroscedasticity correction applied.

3.4 Additional methodological considerations

In the special case when at least one of Z and G_j is discrete, it is natural to place certain constraints on when one would or would not perform any sort of interaction test. For example, if both Z and G_j are binary and are perfectly correlated, then there would typically be zero information in the data on interaction between them as a predictor of Y, and if they are almost perfectly correlated, then the amount of information available on interaction would be quite low. In the case when Z and G_j are both binary, we can think of constructing a 2×2 table of counts of the four possible observed values of (Z, G_j) in the data as in Table 3.2: We require the minimum cell count (MCC), i.e., the smallest of the counts of the four

Table 3.2: **Cell counts**

	$Z = 0$	$zZ = 1$
$G_j = 0$	$\#(Z, G_j) = (0, 0)$	$\#(Z, G_j) = (1, 0)$
$G_j = 1$	$\#(Z, G_j) = (0, 1)$	$\#(Z, G_j) = (1, 1)$

Minimum cell count (MCC) is the smallest count in above table. If it is too small, it means there are very few samples with a certain combination of Z and G_j value. This may make the estimation inaccurate.

possible observed values, to be at least 5 in order to perform the interaction t-test.

Step 4 of the TINGA method requires some additional parameter estimation compared

to the interaction t-test. If all variables were continuous, then with typical GWAS sample sizes, the estimation of a handful of additional parameters would pose little problem for the inference. When G_j and Z are both binary, however, then we require the MCC not be too small.

Furthermore, for Method 3 and 4, the calculation of $\widehat{\text{Var}}(N_j|Y,Z)_{\text{alt}}$ involves fitting an alternative model in Y:

$$Y|Z, G_j \sim N(\alpha + \beta G_j + \gamma Z + \delta(G_j \circ Z), \sigma^2 I). \tag{3.45}$$

when both G_j and Z are binary genotypes and their MAF are not very large, the estimation of δ may not be very accurate and is variational.

For Bernoulli methods, the estimated conditional probability $\hat{p}_{alt} := \hat{P}(G_j = 1|Y,Z)_{alt}$ and the conditional variance $\widehat{\text{Var}}(G_j|Y,Z)_{\text{alt}} = \hat{p}_{alt}(1-\hat{p}_{alt})$ is somehow more sensitive to the estimation of δ than the Gaussian methods. We observed that when estimated $\hat{\delta}$ is relatively large, for the individuals who get the minor allele of Z, $\hat{p}_{alt}$ tends to be more spread out, so there are more entries in $\hat{p}_{alt}$ that are close to 0 or 1, making the conditional variance of G_j smaller. Based on above reasoning, we explore additional ways to estimate $\widehat{\text{Var}}(N_j|Y,Z)_{\text{alt}}$ as following:

1. The shrinkage method:

   We apply the following shrinkage only to the individuals whose SNP Z have the minor allele:

   $$\hat{p}_{alt}^{new} = c\hat{p}_{alt} + (1-c)\hat{p}_{null}, \tag{3.46}$$

   with $c = 0.7$

2. The Lasso method:

We fit the alternative model by Lasso regression:

$$\min_{\alpha,\beta,\gamma,\delta}\|Y-\alpha-G_j\beta-Z\gamma-(G_j\circ Z)\delta\|^2+\lambda\|(\alpha,\beta,\gamma,\delta)\|_1 \tag{3.47}$$

and choose the optimal λ with cross validation.

In addition to the MAF and MCC for the SNPs, we also require the correlation between G_j and Z to be relatively small. Specifically for the problem of epistasis detection, it has been noted that in the presence of an untyped causal variant, two typed variants in strong linkage disequilibrium that form a haplotype that tags the untyped variant could exhibit false epistasis [38]. Therefore, in detection of epistasis, we only test for epistasis between variants G_j and Z if their sample correlation is close to 0. (In our data analysis we use a cut-off of .1 for absolute value of correlation.)

CHAPTER 4

DETAILED STEPS FOR PARAMETER ESTIMATION

In this section, we give the detailed steps of how to estimate the parameters in the different versions of our methods for a given dataset.

4.1 Gaussian approaches

4.1.1 Independent individuals

We first consider the case where the individuals are independent, and the G_j, Y, Z denotes the variables for one particular individual.

Recall that for the Gaussian approximation approach, we get the conditional distribution of $G_j|Y, Z$ by

$$G_j|\,(y,\ z) \sim\ N(\mu_{G_j|z} + \frac{(\beta + \delta z_c)\,\sigma^2_{G_j|z}}{v_{y|z}}(y - \mu_{y|z}),\ \sigma^2_{G_j|z} - \frac{(\beta + \delta z_c)^2\,\sigma^4_{G_j|z}}{v_{y|z}}) \tag{4.1}$$

Method1: estimate Var$(N_j|Y, Z)$ under null and no heteroscedasticity correction

For Method1, we fit the homoscedastic model of $Y|Z$ and fit the null (non-interaction) model of $Y|G_j, Z$.

We estimate $\mu_{y|z},\ v_{y|z}$ by fitting the ordinary linear regression

$$\begin{aligned} Y|Z &\sim N(\alpha + \gamma Z, \sigma^2 I) \\ \mu_{y|z} &= \hat{\alpha} + \hat{\gamma} z \\ v_{y|z} &= \hat{\sigma}^2 \end{aligned} \tag{4.2}$$

For the parameters β and δ, we assume $\delta = 0$ and estimate β the parameters by

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j, \sigma^2 I)$$

Method2: estimate Var$(N_j|Y,Z)$ under null and conduct heteroscedasticity correction

We first regress G_j out of Y and replace Y by the residuals (see section 3.3.4). Then we fit the heteroscedastic model in $Y|Z$ as a linear mixed model:

$$Y|Z \sim N(\alpha + \gamma Z, \sigma_2^2 diag(Z^2) + \sigma_1^2 diag(Z) + \sigma_0^2 I). \tag{4.3}$$

We estimate $\mu_{y|z},\ v_{y|z}$ by the fitted LMM 4.3.

In the case where Z is binary, we estimate $\mu_{y|z},\ v_{y|z}$ by simply taking the sample means and variances of Y given $Z = 1$ and $Z = 0$.

We then estimate β and δ by letting $\delta = 0$, and fitting the non-interaction model

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j, \sigma^2 W) \tag{4.4}$$

where $W = diag(w_1, ..., w_n),\ w_i = \hat{\sigma}_2^2 z_i^2 + \hat{\sigma}_1^2 z_i + \hat{\sigma}_0^2$ for the fitted model 4.3.

Method3: estimate Var$(N_j|Y,Z)$ under alternative and conduct heteroscedasticity correction

For Method3, the parameters for computing $E(N_j|Y,Z)$ are the same as those for Method2.

For the parameters for computing Var$(N_j|Y,Z)$, $\mu_{y|z},\ v_{y|z}$ are those got from fitted model 4.3. For β, δ, we fit the interaction model of $Y|G_j, Z$:

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j + \delta(Z \circ G_j), \sigma^2 W) \tag{4.5}$$

Method4: estimate Var$(N_j|Y,Z)$ under alternative and no heteroscedasticity correction

For Method4, the parameters for computing $E(N_j|Y,Z)$ are the same as those for Method1.

For the parameters for computing Var$(N_j|Y,Z)$, $\mu_{y|z},\ v_{y|z}$ are those got from fitted model

4.3 as in Method1. For β, δ, we fit the interaction model of $Y|G_j, Z$:

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j + \delta(Z \circ G_j), \sigma^2 I) \tag{4.6}$$

4.1.2 GRM case

In the case where the individuals are non-independent, and has some population structure such as a GRM K, we still assume a joint normal distribution of $G_j, Y|Z$. In this part, $G_j, Y, Z \in \mathbb{R}^n$ are in vector form.

$$\begin{pmatrix} G_j \\ Y \end{pmatrix} |z \sim N\left(\begin{pmatrix} \mu_{G_j|z} \\ \mu_{y|z} \end{pmatrix}, \begin{pmatrix} \sigma^2_{G_j|z} I_n & \text{diag}(\beta + \delta z_c)\sigma^2_{G_j|z} \\ \text{diag}(\beta + \delta z_c)\sigma^2_{G_j|z} & V_{y|z} \end{pmatrix} \right) \tag{4.7}$$

where $\text{diag}(\beta + \delta z_c)\sigma^2_{G_j|z} := D_j$ is an $n \times n$ diagonal matrix with the i-th diagonal entry being $(\beta + \delta(z_i - \bar{z}))\sigma^2_{G_j|z}$.

Then we can get the conditional distribution of $G_j|Y = y, Z = z$ by

$$G_j|\,(y,\ z) \sim\ N(\mu_{G_j|z} + D_j V^{-1}_{y|z}(y - \mu_{y|z}),\ \sigma^2_{G_j|z} I_n - D_j V^{-1}_{y|z} D_j) \tag{4.8}$$

We estimate $\mu_{y|z}, V_{y|z}$ bu fitting the LMM

$$\begin{gathered} Y|Z \sim N(\alpha + \gamma Z, \sigma^2_g K + \sigma^2_0 I) \quad \text{(homoscedastic model)} \\ Y|Z \sim N(\alpha + \gamma Z, \sigma^2_2 diag(Z^2) + \sigma^2_1 diag(Z) + \sigma^2_g K + \sigma^2_0 I) \quad \text{(heteroscedastic model)} \end{gathered} \tag{4.9}$$

The parameters β, δ are estimated by fitting a LMM:

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j, \sigma^2_g K + \sigma^2_0 I) \quad \text{(homoscedastic null model)} \tag{4.10}$$

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j + \delta(Z \circ G_j), \sigma_g^2 K + \sigma_0^2 I) \quad \text{(homoscedastic alternative model)} \tag{4.11}$$

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j, \sigma_b^2 \hat{\Sigma} + \sigma_e^2 I) \quad \text{(heteroscedastic null model)} \tag{4.12}$$

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j + \delta(Z \circ G_j), \sigma_b^2 \hat{\Sigma} + \sigma_e^2 I) \quad \text{(heteroscedastic alternative model)} \tag{4.13}$$

where $\hat{\Sigma} = \hat{\sigma}_2^2 diag(Z^2) + \hat{\sigma}_1^2 diag(Z) + \hat{\sigma}_g^2 K$ is the fitted variance components except the noise term from the heteroscedastic model in 4.9. We use $\sigma_b^2 \hat{\Sigma} + \sigma_e^2 I$ to approximate $\text{Var}(Y|G_j, Z)$ so that we could use a faster LMM program that takes only one variance component other than I and avoid fitting a LMM with 4 variance components for every G_j.

We could also fit more precise heteroscedastic models for Method 2 and 3:
heteroscedastic null model (Method2):

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j, \sigma_2^2 diag(Z^2) + \sigma_1^2 diag(Z) + \sigma_g^2 K + \sigma_0^2 I) \tag{4.14}$$

heteroscedastic alternative model (Method3):

$$Y|Z, G_j \sim N(\alpha + \gamma Z + \beta G_j + \delta(Z \circ G_j), \sigma_2^2 diag(Z^2) + \sigma_1^2 diag(Z) + \sigma_g^2 K + \sigma_0^2 I) \tag{4.15}$$

4.2 Bernoulli approaches

Here we design the methods that make use of the Bernoulli feature of G. Similar as above sections, we denote y, z as the observed values of Y, Z.

When G_j only takes values in $0, 1$, we can estimate its conditional probability by Bayes formula

$$P(G_{ij} = k|Z, Y) = P(G_{ij} = k|Z_i, Y_i) = \frac{P\left(Y_i \middle| G_{ij} = k,\ Z_i\right) * p_{k|Z_i}}{\sum_l P\left(Y_i \middle| G_{ij} = l,\ Z_i\right) * p_{l|Z_i}}, k = 0, 1 \tag{4.16}$$

Here Z can be either discrete (Binomial, Bernoulli, etc.) or continuous. When Z is discrete, $p_{k|Z}$ can be easily estimated by taking the empirical conditional mean of G_j on Z; when Z is continuous, it can be estimated by fitting a logistic model.

4.2.1 Independent individuals

Method1: estimate Var$(N_j|Y,Z)$ under null and no heteroscedasticity correction

When we do not consider the heteroscedasticity in Y , we can estimate the conditional distribution $G_j|y,z$ by fitting a logistic model

$$G_j|y,z \sim Ber\left(p\right),\ p = logit(\mu + \alpha y + \beta z)$$

and use the fitted distribution to compute the conditional mean and variance of G_j : $G_j|y,z \sim (\mu_2,\ V_2)$, where $\mu_2 = \hat{p},\ V_2 = diag(\hat{p}\left(1-\hat{p}\right))$

Method2: estimate Var$(N_j|Y,Z)$ under null and conduct heteroscedasticity correction

Here we allow Y to have different variance for different values of Z. Compute the conditional distribution of $G_j|y,z$ by Bayes rule:

$$p_{G_j|y,z} := p\left(G_j = 1\middle|y,z\right) = \frac{p\left(y\middle|G_j=1,\ z\right)p(G_j=1|z)}{p(Y|z)} \tag{4.17}$$

$$= \frac{\left(2\pi\sigma_z^2\right)^{-\frac{1}{2}} e^{\frac{1}{2\sigma_z^2}(y-\alpha-\beta-\gamma z-\delta(1-m_x)(z-m_z))^2} p(G_j=1|z)}{p\left(y\middle|G_j=1,\ z\right)p\left(G_j=1\middle|z\right) + p\left(y\middle|G_j=0,\ z\right)p(G_j=0|z)} \tag{4.18}$$

We estimate $p(G_j|z)$ by taking the sample means of G_j given Z when Z has a discrete distribution; when Z takes continuous values, we fit a logistic model of $G_j|Z$.

For $p\left(Y\middle|G_j,\ Z\right)$, for Method 2 our estimations are all under the null, so $\delta = 0$.

We hope to estimate the parameters in the heteroscedastic model:

$$Y|z, G_j \sim N\left(\alpha + \beta G_j + \gamma z,\ \sigma_0^2 \mathbb{I}_{z=0} + \sigma_1^2 \mathbb{I}_{z=1}\right) \tag{4.19}$$

Let

$$\hat{v}_0^2 = \widehat{\mathrm{Var}}(Y|Z=0);\ \hat{v}_1^2 = \widehat{\mathrm{Var}}(Y|Z=1) \tag{4.20}$$

We first fit the weighted linear model

$$Y|z, G_j \sim N\left(\alpha + \beta G_j + \gamma z,\ \sigma^2 W\right), \tag{4.21}$$

here W is a diagonal matrix with i-th diagonal entry being $\hat{v}_0^2 \mathbb{I}_{z_i=0} + \hat{v}_1^2 \mathbb{I}_{z_i=1}$ and get estimated $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$.

Then we estimate $var(Y|G_j,\ z)$ by $\hat{\sigma}^2 W$, where

$$\hat{\sigma}^2 = \frac{1}{n-3} \sum_i \left(y_i - \hat{\alpha} - \hat{\beta} G_{ij}\right)^2 W_{ii}^{-1} \tag{4.22}$$

or

$$\hat{\sigma}_0^2 = \frac{1}{n_0 - 1.5} \Sigma_{i:z_i=0} \left(y_i - \hat{\alpha} - \hat{\beta} G_{ij}\right)^2,\ \hat{\sigma}_1^2 = \frac{1}{n_1 - 1.5} \Sigma_{i:z_i=1} \left(y_i - \hat{\alpha} - \hat{\beta} G_{ij} - \hat{\gamma}\right)^2, \tag{4.23}$$

where n_k is the number of z equaling k.

Then we plug the estimated parameters in above Bayes formula to get $\hat{p}_{G_j|y,z}$. We estimate the conditional mean and variance of G_j by

$$EG_j|y, z \approx \hat{p}_{G_j|y,z},\quad var\left(G_j|y, z\right) \approx \hat{p}_{G_j|y,z}(1 - \hat{p}_{G_j|y,z}) \tag{4.24}$$

and use this to compute $E(N_j|y, z)$ and $var(N_j|y, z)$.

Method3: estimate Var$(N_j|Y,Z)$ under alternative and conduct heteroscedasticity correction

We compute var$(N_j|y,z)$ under alternative model: for $E(N_j|y,z)$, we use the conditional mean and variance of G_j under the null as in method2; for var$(N_j|y,z)$, we fit the interaction model.

For $E(T|y,z)$:

$p(y_i|G_{ij},z_i)$ in $E(G_j|y,z)$ and $var(G_j|y,z)$ are given by fitting the null heteroscedastic model

$$Y|Z,\ G_j \sim N(\alpha+\beta G_j+\gamma Z,\ \sigma_0^2\mathbb{I}_{z=0}+\sigma_1^2\mathbb{I}_{z=1})$$

as in method 2.

For $var(T|y,z)$:

$p(y_i|G_{ij},z_i)$ is given by fitting the interaction model

$$Y|G_j,Z \sim N(\alpha+\beta G_j+\gamma Z+\delta(G_j\circ Z),\sigma_0^2\mathbb{I}_{z=0}+\sigma_1^2\mathbb{I}_{z=1})$$

When Z is not Bernoulli, we estimate $var(Y|G_j,Z)$ by $\sigma^2 W$, where $W=\hat{var}(Y|Z)$.

The rest steps are the same as method 1 & 2.

Method4: estimate Var$(N_j|Y,Z)$ under alternative and no heteroscedasticity correction

For Method4, we assume a homoscedastic model of Y, so the steps are the same as Method3 in the homoscedastic Y scenario.

4.2.2 GRM case

Method1:

Let K be the GRM. Let i denote the individual index.

Compute the conditional mean of $G_j:\ EG_{ij}|y,z = p(G_{ij}=1|y,z)$ by Bayes rule:

$$p\left(G_{ij}=1\middle|y,z\right) \approx p\left(G_{ij}=1\middle|y_i,z_i\right) = \frac{p\left(y_i\middle|G_{ij}=1,\ z_i\right)p(G_{ij}=1|z_i)}{p(y_i|z_i)} \tag{4.25}$$

$$= \frac{p\left(y_i\middle|G_{ij}=1,\ z_i\right)p(G_{ij}=1|z_i)}{p\left(y_i\middle|G_{ij}=1,\ z_i\right)p\left(G_{ij}=1\middle|z_i\right)+p\left(y_i\middle|G_{ij}=0,\ z_i\right)p(G_{ij}=0|z_i)} \tag{4.26}$$

where

$$Y|G_j,Z \sim N(\alpha+\beta G_j+\gamma Z,\ \sigma_g^2K+\sigma_e^2I)$$

$$Y_i|G_{ij},\ Z_i \sim N\left(\alpha+\beta G_{ij}+\gamma Z_i,\ \sigma_g^2K+\sigma_e^2I\right) \approx N(\hat{\alpha}+\hat{\beta}G_{ij}+\hat{\gamma}Z_i,\ \hat{\sigma}_g^2K_{ii}+\hat{\sigma}_e^2)$$

Once get estimated $p\left(G_{ij}=1\middle|y,z\right) := \hat{p}$, we can estimate $cov\left(G_j|y,z\right)$ by

$$\widetilde{K} = cov2cor\left(K+cI\right),\ c=10^{-7}$$

$$cov\left(G_j|y,z\right)_{ij} \approx \widetilde{K}_{ij}\sqrt{\hat{p}_i\left(1-\hat{p}_i\right)\hat{p}_j(1-\hat{p}_j)}$$

The computation of approximated $T,\ E\left(T|y,z\right),\ var(T|y,z)$ is the same as non-GRM case except that $H=I-\frac{1}{n}\mathbf{1}\mathbf{1}^T,\ \hat{H}=\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}1\left(1^T\hat{\Sigma}^{-1}1\right)^{-1}1\hat{\Sigma}^{-1}$, where $\hat{\Sigma}=\hat{\sigma}_g^2K+\hat{\sigma}_e^2I$ is the estimated covariance matrix in the LMM model

$$Y \sim N(\alpha+\beta G_j+\gamma Z+\delta(G_j\circ Z),\ \sigma_g^2K+\sigma_e^2I)$$

Method2:

We fit the heteroscedastic model by `regress`:

$$Y|Z \sim N(\alpha+\gamma Z,\ \sigma_h^2diag\left(Z\right)+\sigma_g^2K+\sigma_e^2I)$$

Let $\hat{\Sigma} = \hat{\sigma}_h^2 diag\,(z) + \hat{\sigma}_g^2 K$, we estimate $Y|G_j, Z$ by

$$Y|G_j, Z \sim N(\alpha + \beta G_j + \gamma z,\ \sigma_b^2 \hat{\Sigma} + \sigma_e^2 I)$$

$$Y_i|G_{ij},\ Z_i \approx N(\hat{\alpha} + \hat{\beta} G_{ij} + \hat{\gamma} Z_i,\ \hat{\sigma}_b^2 \hat{\Sigma}_{ii} + \hat{\sigma}_e^2)$$

where

$$\hat{\Sigma} = \hat{\sigma}_g^2 K + \hat{\sigma}_h^2 \text{diag}(Z) \tag{4.27}$$

Method3:

For $E(T|y, z)$:

$p(y_i|G_{ij}, z_i)$ in $E(G_j|y, z)$ and $var(G_j|y, z)$ are given by fitting the null model

$$Y|G_j, Z \sim N(\alpha + \beta G_j + \gamma Z,\ \sigma_b^2 K + \sigma_e^2 I)$$

For $var(T|y, z)$:

$p(y_i|G_{ij}, z_i)$ is given by fitting the interaction model

$$Y|G_j, Z \sim N(\alpha + \beta G_j + \gamma Z + \delta(G_j \circ Z),\ \sigma_b^2 K + \sigma_e^2 I)$$

Other parts are the same as method1 & 2.

Method4

Sames steps as Method3 in the homoscedasticity scenario.

CHAPTER 5

RESULTS OF SIMULATIONS

In the simulations, we simulate $Y \in \mathbb{R}^n$ as the phenotype; $Z \in \mathbb{R}^n$ as the fixed SNP/environmental factor; $G_1, ..., G_m \in \mathbb{R}^n$ as the SNPs in the genome. In this section, we first show the improvement of our methods on the uniformity of null p-values within one GWAS and on the distribution of genomic control inflation factor. We then show the simulation results for the Type I error rates and power across multiple GWAS's and show that our methods have desired type I error and better power performance than the regular methods. For the simulations, we particularly focus on the case where both Z, G_j are Bernoulli distributed and apply the Bernoulli version of our methods.

In this section, we have simulation experiments to access the performance of our methods in 2 aspects: (1). Fixing the "feast or famine" effect in one interaction GWAS; (2). Type I error rates and power across many GWAS's.

5.1 Simulation under null: check p-values within a GWAS

In this part, we focus on the performance of the m p-values within one GWAS in 2 aspects: uniformity and genomic control inflation factor. For each replicate, the sample size is 1000. We simulate a fixed Z and $m = 5000$ G_j's independently. Y is simulated under null . Then for each G_j, we test for its interaction with Z and get a p-value. Finally, we test for the uniformity of the resulting $m = 5000$ p-values using ELL [1] at level 0.05. We also got the GCIF for each GWAS by taking the median of the $m = 5000$ χ^2 scores and divide it by 0.456.

5.1.1 Gaussian Methods

In this section, we compare the performance of Gaussian Method1-4.

We simulate interaction GWAS as follows:

- $n = 1000,\ m = 5000$
- $Z \sim Bernoulli(m_z),\ m_z \sim Unif(0.2, 0.8)$
- $G_j \overset{\text{indep.}}{\sim} Bernoulli(m_j),\ m_j \overset{\text{iid}}{\sim} Unif(0.2, 0.8), cor(Z, G_j) = 0,\ j = 1, 2, ..., m$
- $Y = \alpha + \gamma Z + \sum_{j=1}^{m} \beta_j G_j + \epsilon,\ \alpha \sim Unif(-10, 10),\ \epsilon \sim N(0, 1)$
- $\gamma = \sqrt{\frac{0.025}{\sigma_z^2}}$, $\beta = \begin{cases} \sqrt{\frac{0.025}{\sigma_j^2}} & j = 1, 2, ..., 50 \\ 0 & j = 51, ..., 5000 \end{cases}$

We let 50 out of 5000 SNPs have marginal association with Y to represent the situation where there are a few marginal association signals.

When simulate each G_j, we check the following 2 conditions and keep re-generating G_j until both of the 2 conditions are satisfied:

1. $|cor(G_j, Z)| \leq 0.1$

2. $MCC(G_j, Z) \geq 5$

For each $j = 1, 2, ..., 5000$, we test $H_{0j} : \delta_i = 0$ in

$$Y = \alpha_j + G_j\beta_j + Z\gamma_j + (G_j \circ Z)\delta_j + \epsilon_j$$

and get $m = 5000$ interaction p-values.

We did 3000 replicates. For each replicate, we test for the uniformity of the 5000 p-values using the method of Equal Local Level (ELL)[1]. We then got 3000 p-values for uniformity (for simplicity, we denote them as "ELL p-values"). If the null hypothesis that the $m = 5000$ p-values from an interaction GWAS are iid unifom is true, then the 3000 ELL p-values are expected to be uniformly distributed. Figure 5.1 depicts the (differenced) QQ-plots of the 1000 ELL p-values against standard uniform: we take the -log10 scaled p-values and plot the difference between the observed quantiles and the theoretical quantiles (quantile of uniform$(0, 1)$). As we can see, with the regular t-test, the p-values in a null GWAS tend to be significantly non-uniform; while they are close to uniform for TINGA.

Figure 5.2 compares the genomic control inflation factors between the regular t-test and the 4 Gaussian methods. As we can see, TINGA methods make the GCIF more concentrated at 1, meaning there is less systematic inflation or deflation in the testing statistics.

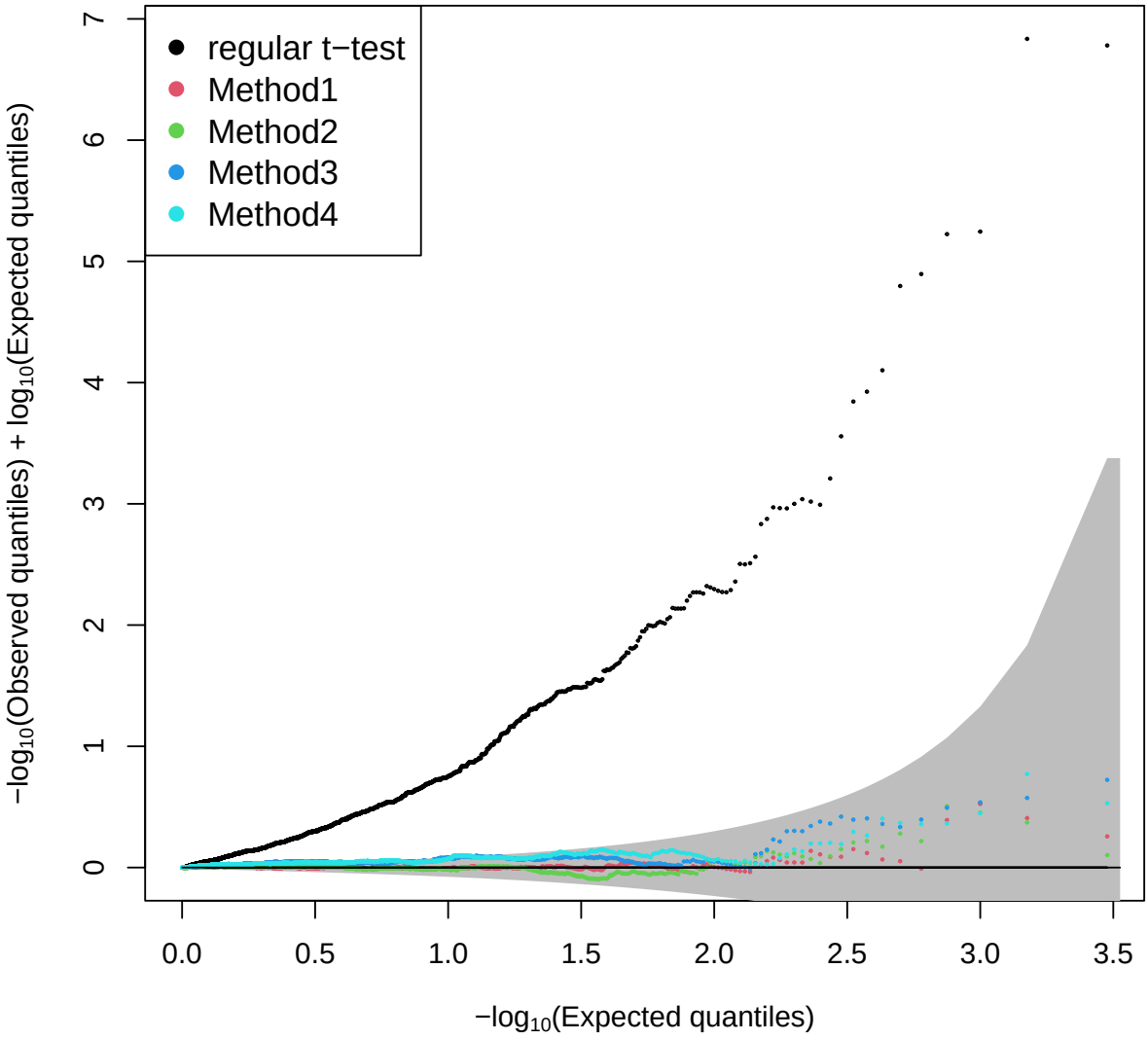


Figure 5.1: **QQ-plots of ELL p-values** 3000 points. Each point represents a simulated interaction GWAS of $m = 5000$ interaction test. For each GWAS, test whether the $m = 5000$ p-values are iid uniformly distributed using ELL [1]. The shaded region is the 95% confidence region by ELL

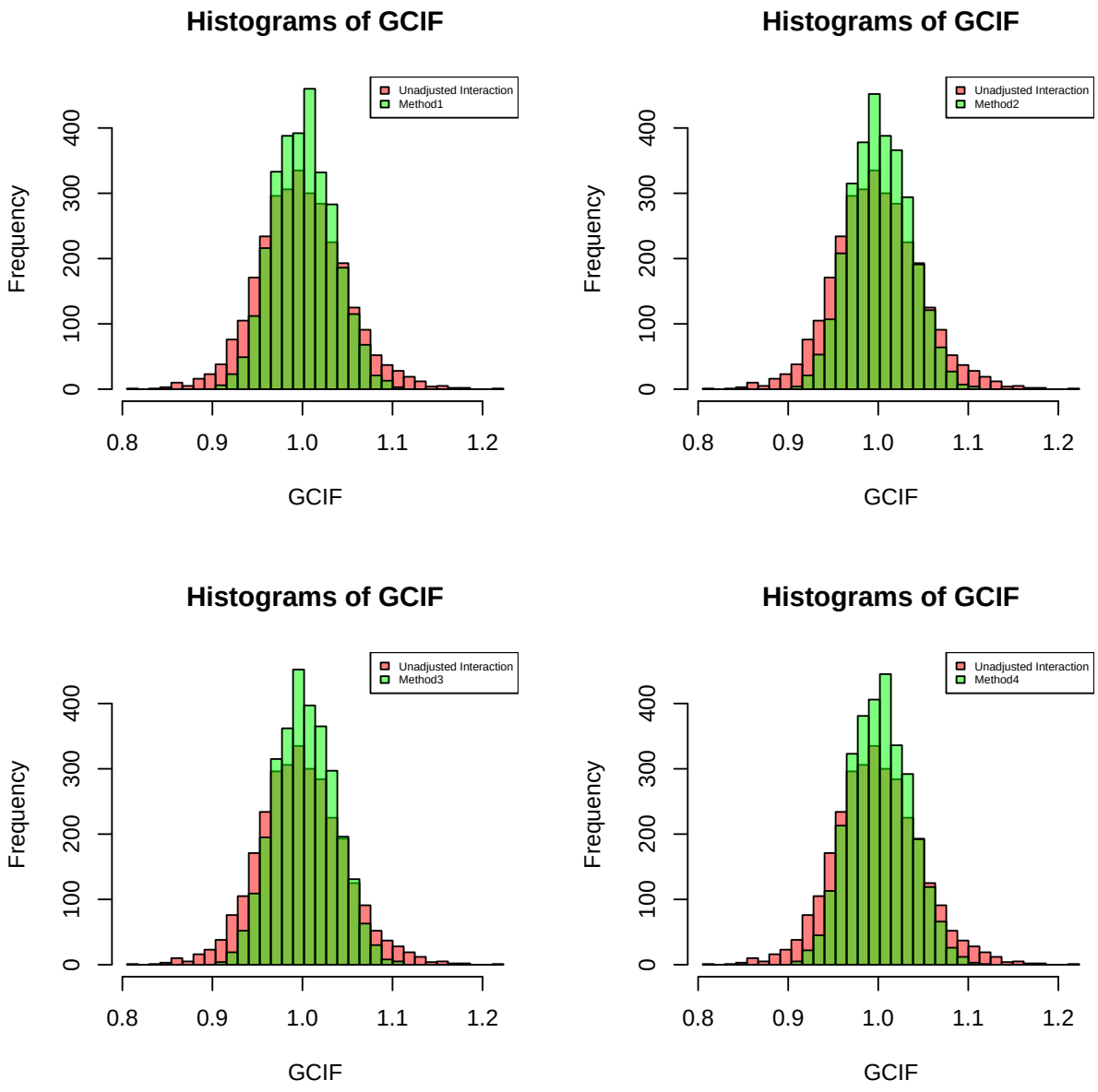


Figure 5.2: **Genomic Control Inflation Factors** 4 penels represent the GCIFs from the same 3000 replicates of simulation. Red bars: GCIF from regular t-test; Green: GCIF from 4 Gaussian versions of TINGA method

5.1.2 Bernoulli Methods

In this section, we assess the performances of the following 4 versions of Bernoulli methods:

1. Method1
2. Method2
3. Method3, shrink (See equation 3.46 in section 3.4)
4. Method3, Lasso (See equation 3.47 in section 3.4)

We use the same simulation setting as above section 5.1.1 and we did 1000 replicates. We again get the 1000 ELL p-values and plot the (differenced) QQ-plots in Figure 5.3.

Figure 5.4 compares the distributions of genomic control inflation factors.

We got similar conclusions as in section 5.1.1 for the Gaussian methods. In summary, in the simulations, our TINGA method succeeds in eliminating or effectively reducing the "Feast or Famine" effect. It makes the null interaction GWAS p-values approximately uniform as we expect.

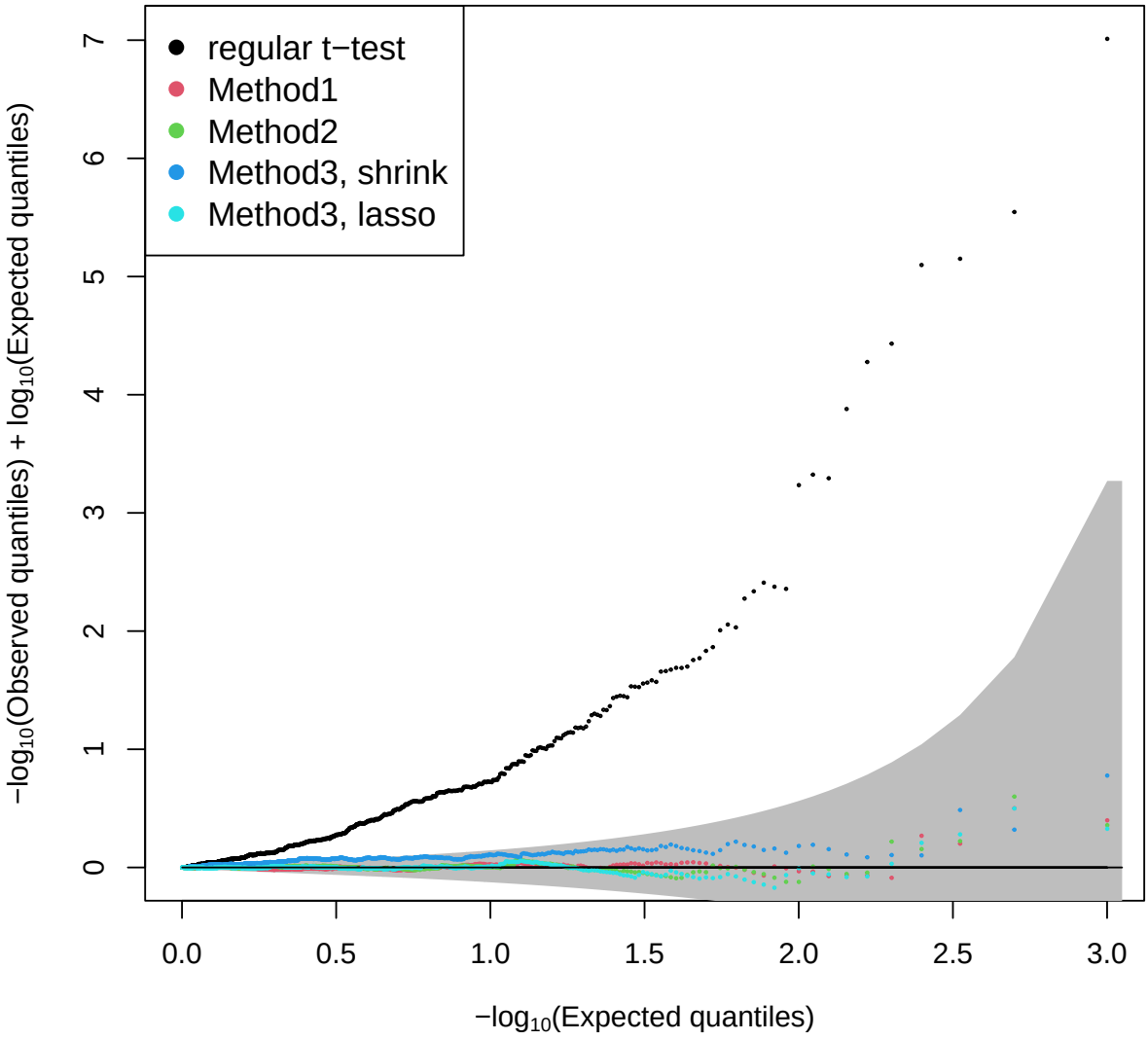


Figure 5.3: **QQ-plots of ELL p-values** 1000 points. Each point represents a simulated interaction GWAS of $m = 5000$ interaction test. For each GWAS, test whether the $m = 5000$ p-values are iid uniformly distributed using ELL [1]. The shaded region is the 95% confidence region by ELL

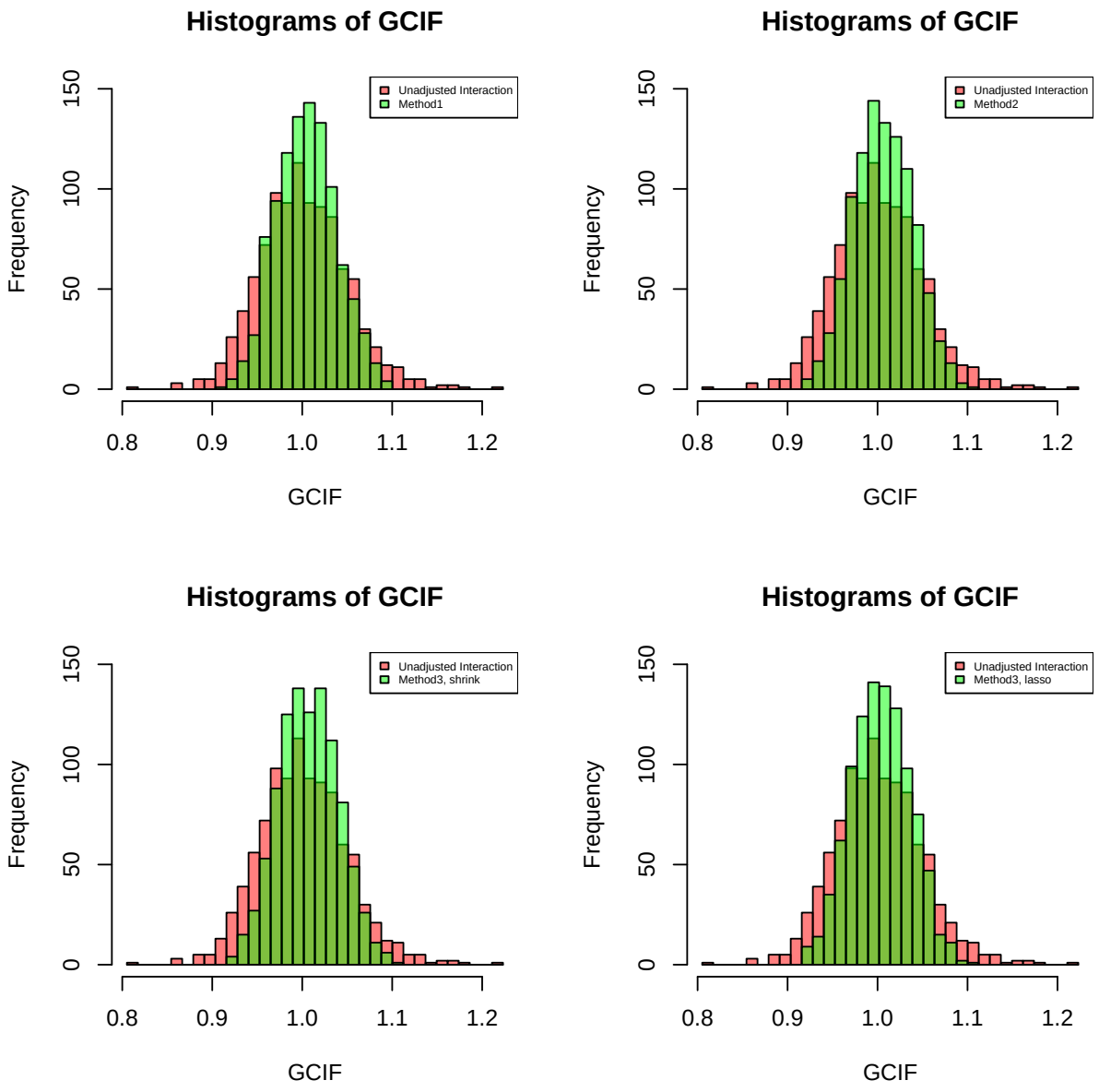


Figure 5.4: **Genomic Control Inflation Factors** 4 penels represent the GCIFs from the same 1000 replicates of simulation. Red bars: GCIF from regular t-test; Green: GCIF from 4 Bernoulli versions of TINGA method

5.1.3 More situations and comparisons

In this section, we explore more simulation settings.

We consider the following 4 cases:

1. Both Z and G_j's are Bernoulli, and Y is simulated under a linear model:

   - $n = 1000,\ m = 5000$
   - $Z \sim Bernoulli(m_Z),\ m_Z \sim Unif(0.1, 0.9)$
   - $G_j \overset{\text{indep.}}{\sim} Bernoulli(m_j),\ m_j \overset{\text{iid}}{\sim} Unif(0.1, 0.9), cor(Z, G_j) = 0,\ j = 1, 2, ..., m$
   - $Y = \alpha + \epsilon,\ \alpha \sim Unif(-10, 10),\ \epsilon \sim N(0, 1)$

   When simulate each G_j, we check the following 2 conditions and keep re-generating G_j until both of the 2 conditions are satisfied:

   (a) $|cor(G_j, Z)| \leq 0.1$

   (b) $MCC(G_j, Z) \geq 5$

2. Both Z and G_j's are Binomial(2), and Y is simulated under a linear model:

   - $n = 1000,\ m = 5000$
   - $z \sim Binom(2, m_z),\ m_z \sim Unif(0.1, 0.9)$
   - $G_j \overset{\text{indep.}}{\sim} Binom(2, m_j),\ m_j \overset{\text{iid}}{\sim} Unif(0.1, 0.9), cor(z, G_j) = 0,\ j = 1, 2, ..., m$
   - $y = \alpha + \epsilon,\ \alpha \sim Unif(-10, 10),\ \epsilon \sim N(0, 1)$

3. Both Z and the G_j's are normal, and Y is simulated under a linear model:

   - $n = 1000,\ m = 5000$
   - $z \sim Normal(m_z, 1),\ m_z \sim Unif(-10, 10)$
   - $G_j \overset{\text{indep.}}{\sim} Normal(m_j, 1),\ m_j \overset{\text{iid}}{\sim} Unif(-10, 10), cor(z, G_j) = 0,\ j = 1, 2, ..., m$

- $y = \alpha + \epsilon,\ \alpha \sim Unif(-10, 10),\ \epsilon \sim N(0, 1)$

4. Both Z and G_j's are Bernoulli, and Y is simulated under a LMM:

   - $n = 1000,\ m = 5000$
   - $z \sim Bernoulli(m_z),\ m_z \sim Unif(0.1, 0.9)$
   - $G_j \overset{\text{indep.}}{\sim} Bernoulli(m_j),\ m_j \overset{\text{iid}}{\sim} Unif(0.1, 0.9), cor(z, G_j) = 0,\ j = 1, 2, ..., m$
   - GRM K is calculated from a simulated genotype matrix G, which is independent of z and x:

$$G = (g_1, ..., g_{10000}),\ g_i \overset{indep.}{\sim} Ber(f_i),\ f_i \overset{iid}{\sim} Unif(0.1, 0.9)$$
$$\tilde{G} = (\tilde{g_1}, ...\tilde{g}_{10000}),\ \tilde{g_i} = \frac{g_i - \bar{g_i}}{\sqrt{\bar{g_i}(1 - \bar{g_i})}}$$
$$K = \frac{1}{10000}\tilde{G}\tilde{G}^T$$

   - $\alpha \sim Unif(-10, 10),\ h^2 = 0.3,\ \sigma_T^2 = 1$
   - $y = \alpha + \epsilon,\ \epsilon \sim N(0,\ \sigma_T^2\left(h^2K + (1 - h^2)I\right)$

   When simulate each G_j, we check the following 2 conditions and keep re-generating G_j until both of the 2 conditions are satisfied:

   (a) $|cor(G_j, z)| \leq 0.1$

   (b) $MCC(G_j, z) \geq 5$

Table 5.1, 5.2 compare the rates of rejection of uniformity of p-values for regular t/Wald test and the Bernoulli versions of our correction methods. As we can see, the regular t-test gives large rejection rates, meaning the resulting p-values are not uniformly distributed, even if Y is under completely null.

This problem exists in general cases. G_j's can be Binomial or Bernoulli, representing genotypes. Z can be either normal (representing environmental factors) or Binomial/Bernoulli (genotypes).

Table 5.1: **Rejection rates in non-GRM case (1000 replicates)**

Uncorrected	Method1	Method2	Method3	MCC20
286	59	54	129	131

Number of times that the uniformity of the resulting 5000 p-values is rejected, out of 1000 replicates. Both Z, G_j are Bernoulli and independent across individuals; Y is simulated under the null. Methods are the Bernoulli version.

Table 5.2: **Rejection rates in GRM case (200 replicates)**

Wald	Method1	Method2	Method3	TINGA
58	14	14	27	31

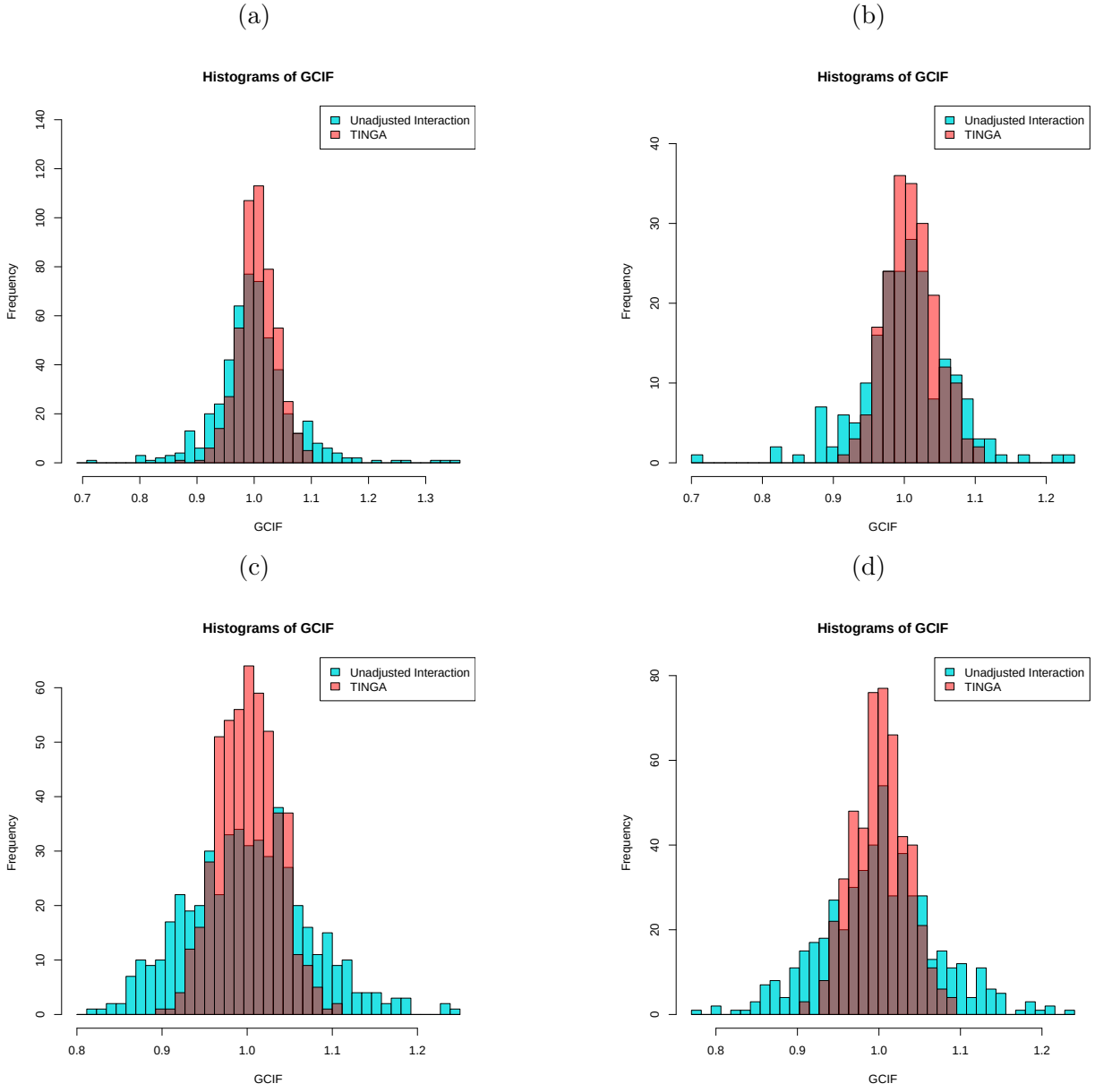
Number of times that the uniformity of the resulting 5000 p-values is rejected, out of 1000 replicates. Both Z, G_j are Bernoulli and independent across individuals; Y is simulated with a GRM. Methods are the Bernoulli version.

Fig 5.5 are the histograms of genomic control inflation factors for different simulation settings. From this we believe that our methods make the null p-values more "uniform" and make the genomic control inflation factor more concentrate around 1.

Figure 5.6, 5.7 are histograms of GCIF for the case where both of G_j, Z are Normal and both of G_j, Z are Binomial. In addition to the comparison between the uncorrected interaction test and the TINGAinteraction test, we also compare them with the ordinary marginal association test, it shows that the distribution of GCIF from TINGAis close to that of the ordinary GWAS, which has no problem with the "feast or famine" effect.

Table 5.3 contains the rejection counts for different distributions for Z and G_j.

Figure 5.5: **Uncorrecred vs. corrected GCIF under the null**



Genomic control inflation factors of interaction tests between $m = 5000$ G_j's and Y; Both Z and G_j's are Bernoulli distributed and independent across individuals (a). Z, G_j Bernoulli. Non-GRM case, 500 replicates; (b). Z, G_j Bernoulli. GRM case, 200 replicates. (c). Z, G_j Gaussian. Non-GRM case, 500 replicates. (d). Z, G_j Binomial$(2, *)$. Non-GRM case, 500 replicates.

Figure 5.6: $Z,\ G_j$ **both normal**

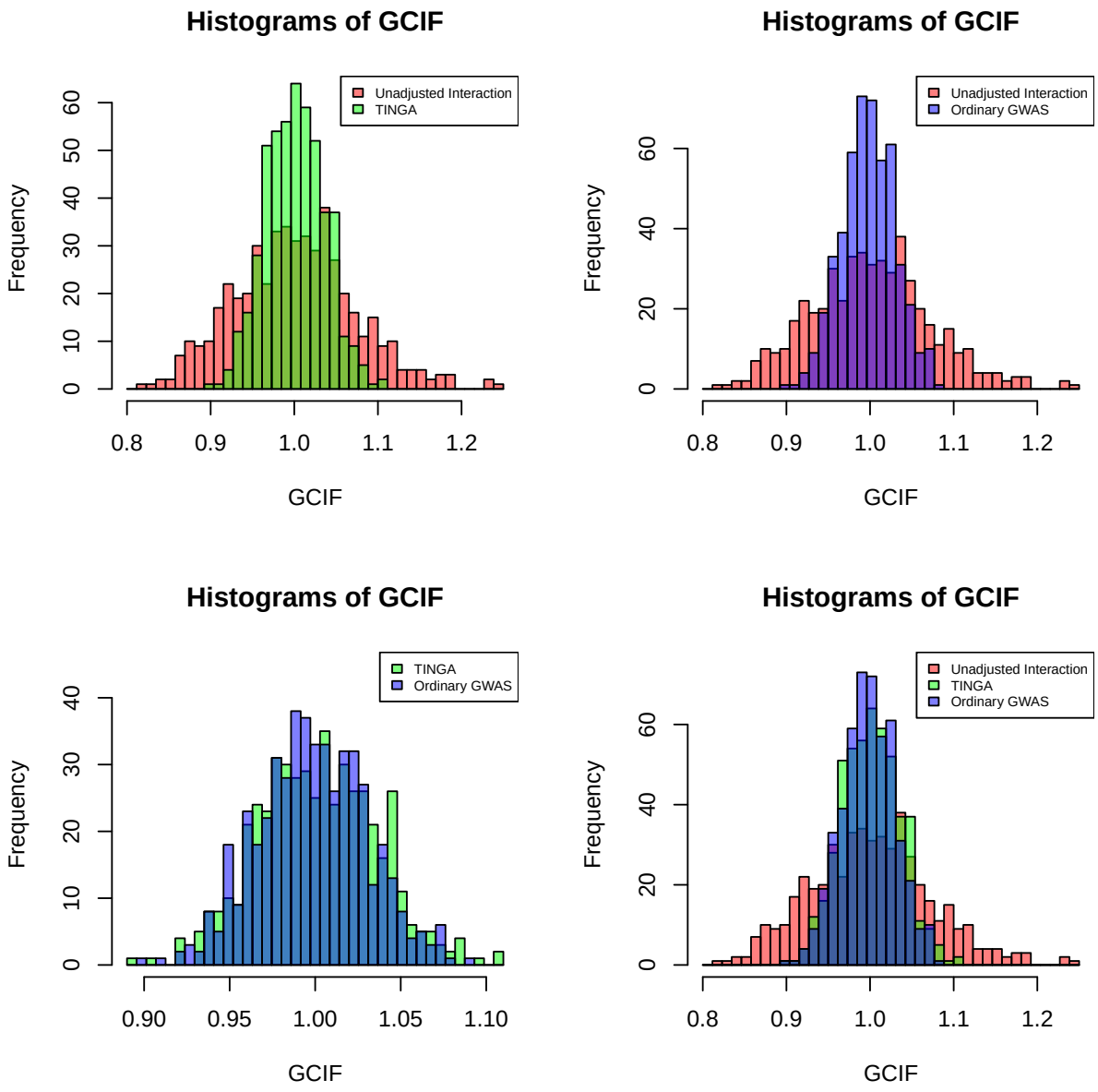


Figure 5.7: Z, G_j **both binomial**

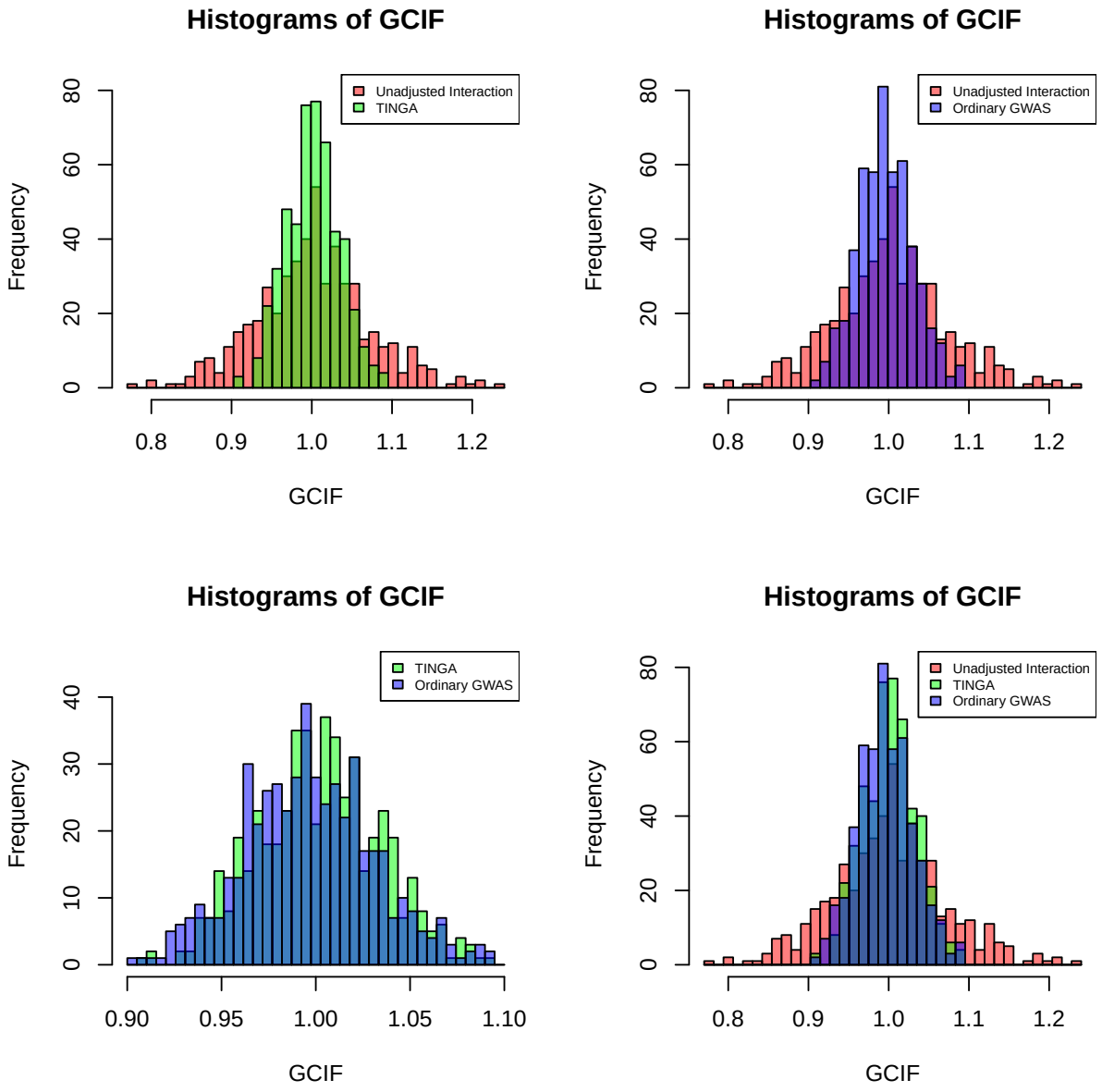


Table 5.3: **Rejection rates** 100 replicates, using normal approximation methods

Z	G_j	$cor(Z, G_j)$	Y	Uncorrected	Corrected
$N(m_z, 1)$	$Bin(2, p)$	0	$N(m_y, 1)$	46	6
$N(m_z, 1)$	$Ber(p)$	0	$N(m_y, 1)$	36	5
$Bin(2, p_1)$	$Bin(2, p_2)$	0	$N(m_y, 1)$	39	8
$Ber(p_1)$	$Ber(p_2)$	0	$N(m_y, 1)$	35	9

The conditional distribution $G_j|Y, Z$ is computed by normal approximation.

5.2 Type I error rates and power across GWAS's

As we mentioned in previous sections, the t-statistic for the interaction test is indeed $\mathcal{T}_{n-4}$ distributed both marginally and conditioning on (G_j, Z). It means if we take the t-statistics from many interaction GWAS's (so many different Y, Z pairs, they follow a t-distribution and so the p-values from t-test are actually uniform. We also want to check if the testing statistic from our TINGAmethod follows (approximately) standard normal distribution as we expect. In this part, we run a simulation multiple times independently to mimic multiple independent GWAS's. Then we look at the Type I error rates and power across GWAS's.

5.2.1 Gaussian Methods

In this section, we assess the performance of Gaussian Method2 and Method3 by looking at the type 1 error and power across many independent GWAS's.

We simulate each replicate under the null as following:

- $n = 1000, m = 5$
- $Z \sim$ Bernoulli(m_Z)
- $G_1 \sim$ Bernoulli(m_1)
- $m_z, m_1 \overset{\text{iid}}{\sim}$ Unif$(0.15, 0.85)$

- $G_j \overset{\text{indep.}}{\sim} \text{Bernoulli}(m_j),\ m_j \overset{\text{iid}}{\sim} \text{Unif}(0.2, 0.5),\ j = 2, 3, 4, 5$
- $\alpha \sim \text{Unif}(-10, 10)$
- $Y = \alpha + \gamma Z + \beta_1 G_1 + \sum_{j=2}^{5} \beta_j G_j + \epsilon,\ \epsilon \sim N(0, \sigma_e^2 I)$
- $\sigma_e^2 = 1 - \gamma^2 \sigma_Z^2 - \sum_{j=1}^{5} \beta_j^2 \sigma_{G_j}^2$

When simulate each G_j, we check the following 2 conditions and keep re-generating j until both of the 2 conditions are satisfied:

1. $|cor(G_j, Z)| \leq 0.1$
2. $MCC(G_j, Z) \geq 5$

We did $\approx 10^5$ replicates for the following 7 cases:

1. $\gamma = 0,\ (\beta_1, \beta_2, \beta_2, \beta_4, \beta_5) = (0, 0, 0, 0, 0)$
2. $\gamma = 0,\ \beta_1 = 0,\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$
3. $\gamma = \sqrt{\frac{0.01}{\sigma_Z^2}},\ \beta_1 = 0,\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$
4. $\gamma = 0,\ \beta_1 = \sqrt{\frac{0.01}{\sigma_{G_1}^2}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$
5. $\gamma = \sqrt{\frac{0.01}{\sigma_Z^2}},\ \beta_1 = \sqrt{\frac{0.01}{\sigma_{G_1}^2}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$
6. $\gamma = \sqrt{\frac{0.01}{\sigma_Z^2}},\ \beta_1 = \sqrt{\frac{0.04}{\sigma_{G_1}^2}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$
7. $\gamma = \sqrt{\frac{0.04}{\sigma_Z^2}},\ \beta_1 = \sqrt{\frac{0.01}{\sigma_{G_1}^2}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}})$

We simulate G_2, G_3, G_4, G_5 to mimic the situation where there are other positively or negatively associated SNPs.

For each replicate in each of the 7 cases, we look at the p-values for testing the interaction between Z and G_1 in

$$Y|Z, G_1 \sim N(0, \alpha + \gamma Z + \beta G_1 + \delta(G_1 \circ Z), \sigma^2 I). \tag{5.1}$$

Note that the fitted model does not contain G_2, G_3, G_4, G_5, so there is model mis-specification except for case 1.

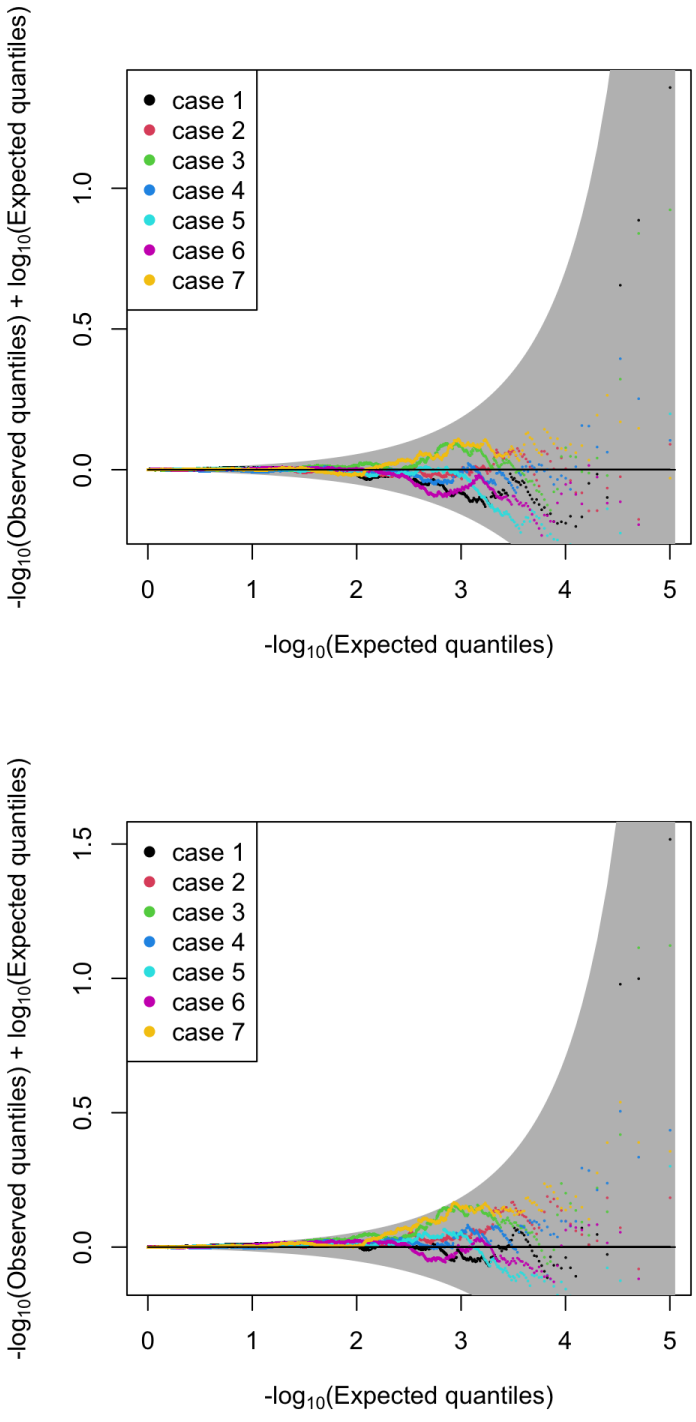
Figure 5.8 are the (differenced) QQ-plots of -log10 p-values for the cases. The grey regions are the 99% ELL confidence regions [1]. As we can see, Gaussian Method 2 and 3 have acceptable type 1 error performance.

We simulate the following alternative case to compare the power of regular t-test and Gaussian Method 2 and 3:

$$\begin{gathered}
Y = \alpha + \gamma Z + \textstyle\sum_{j=1}^{5} \beta_j G_j + \sum_{j=1}^{5} \delta_j (G_j \circ Z) + \epsilon, \\
\epsilon \sim N(0, \sigma_e^2 I) \\
\sigma_e^2 = 1 - \gamma^2 \sigma_Z^2 - \textstyle\sum_{j=1}^{5} \beta_j^2 \sigma_{G_j}^2 - \sum_{j=1}^{5} \delta_j^2 \sigma_{G_j \circ Z}^2 \\
\gamma = \sqrt{\frac{0.02}{\sigma_Z^2}}, \ \beta_1 = \sqrt{\frac{0.02}{\sigma_{G_1}^2}} \\
(\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}}) \\
(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5) = (\sqrt{\frac{0.02}{\sigma_{G_1 \circ Z}^2}}, 0, 0, 0, 0)
\end{gathered} \tag{5.2}$$

Figure 5.9 draws the power curves for the 3 testing approaches. The x-axis is for -log10 scaled type 1 error for testing $G_2 \circ Z$. The y-axis is for power for testing $G_1 \circ Z$. As we can see, in our simulation setting, power of Gaussian Method 3 > regular t-test > Gaussian Method 2.

Figure 5.8: **QQ-plots for Gaussian Methods** Top: Gaussian Method 2; Bottom: Gaussian Method3. The 7 cases are described in section 5.2.1



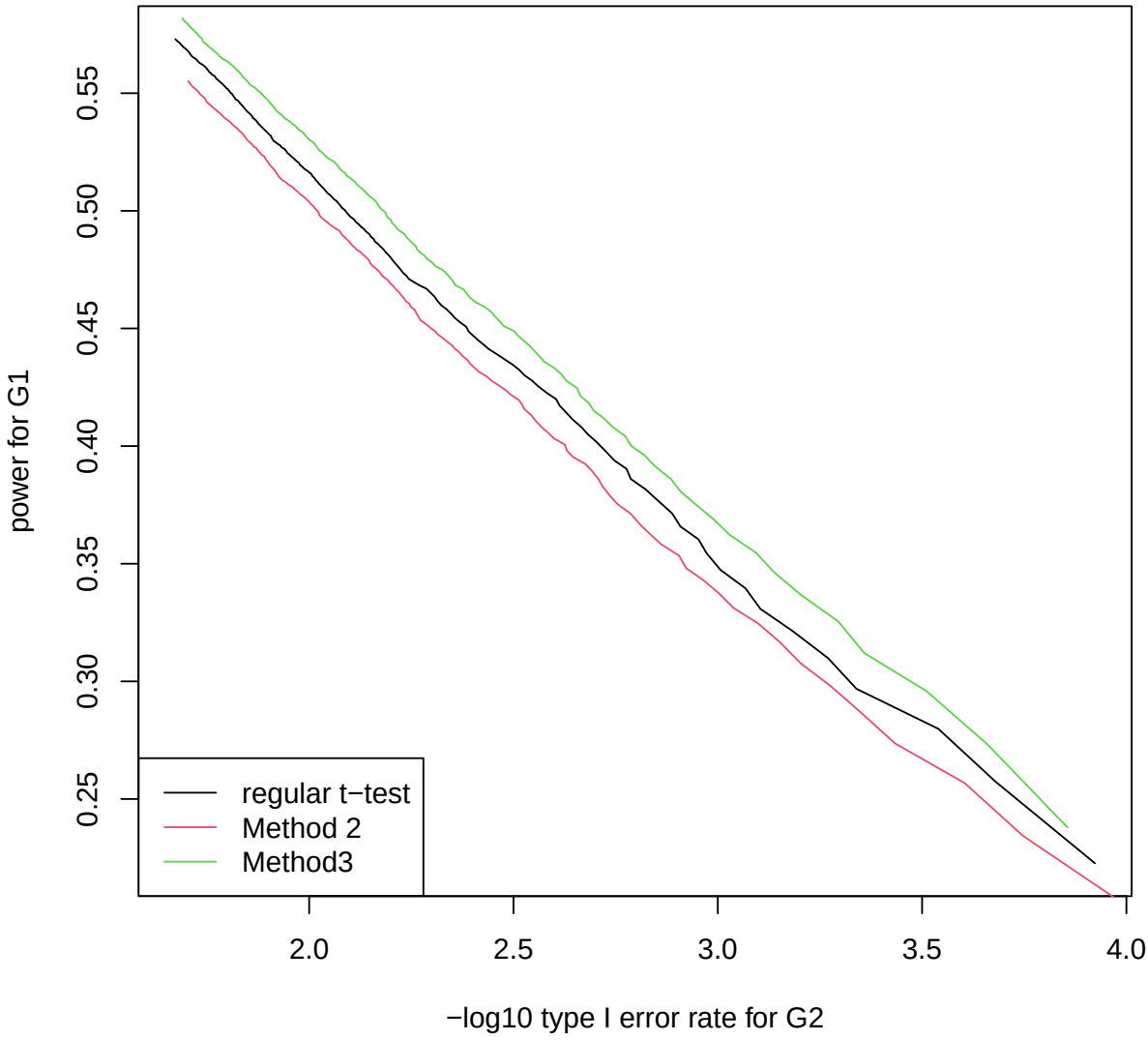


Figure 5.9: **Power curves for Gaussian Methods** x-axis: -log10 scaled type 1 error for testing $G_2 \circ Z$; y-axis: power for testing $G_1 \circ Z$

5.2.2 Bernoulli Methods

In this section, we assess the performance of the following 3 Bernoulli versions of TINGA by looking at the type 1 error and power across many independent GWAS's:

1. Method2
2. Method3, shrink (see equation 3.46 in section 3.4)
3. Method3, Lasso (see equation 3.47 in section 3.4)

We use the same simulation setting as in above section 5.2.1 except that we set the lower bound for MAF to 0.2 instead of 0.15:

$$\begin{aligned} m_z, m_1 &\overset{\text{iid}}{\sim} \text{Unif}(0.2, 0.8) \\ m_2, m_3, m_4, m_5 &\overset{\text{iid}}{\sim} \text{Unif}(0.2, 0.5) \end{aligned} \tag{5.3}$$

Figure 5.10 depicts the (differenced) QQ-plots of -log10 p-values for the null cases. Figure 5.11 compares the power curves of regular t-test and Bernoulli methods in the alternative case.

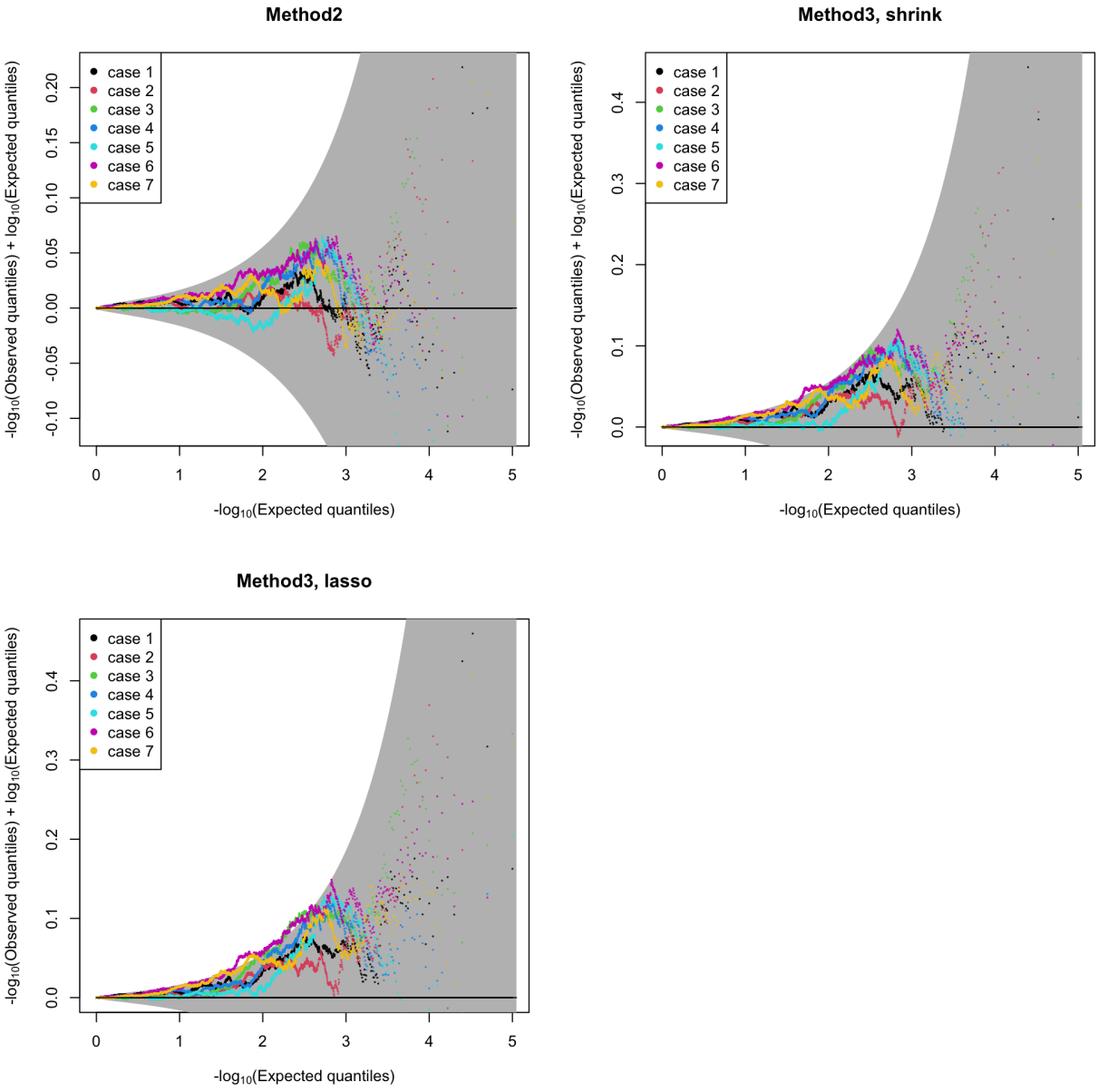


Figure 5.10: **QQ-plots for Bernoulli Methods** Top left: Bernoulli Method 2; Top right: Bernoulli Method3, shrinkl Bottom: Bernoulli Method3, lasso. The 7 cases are described in section 5.2.1

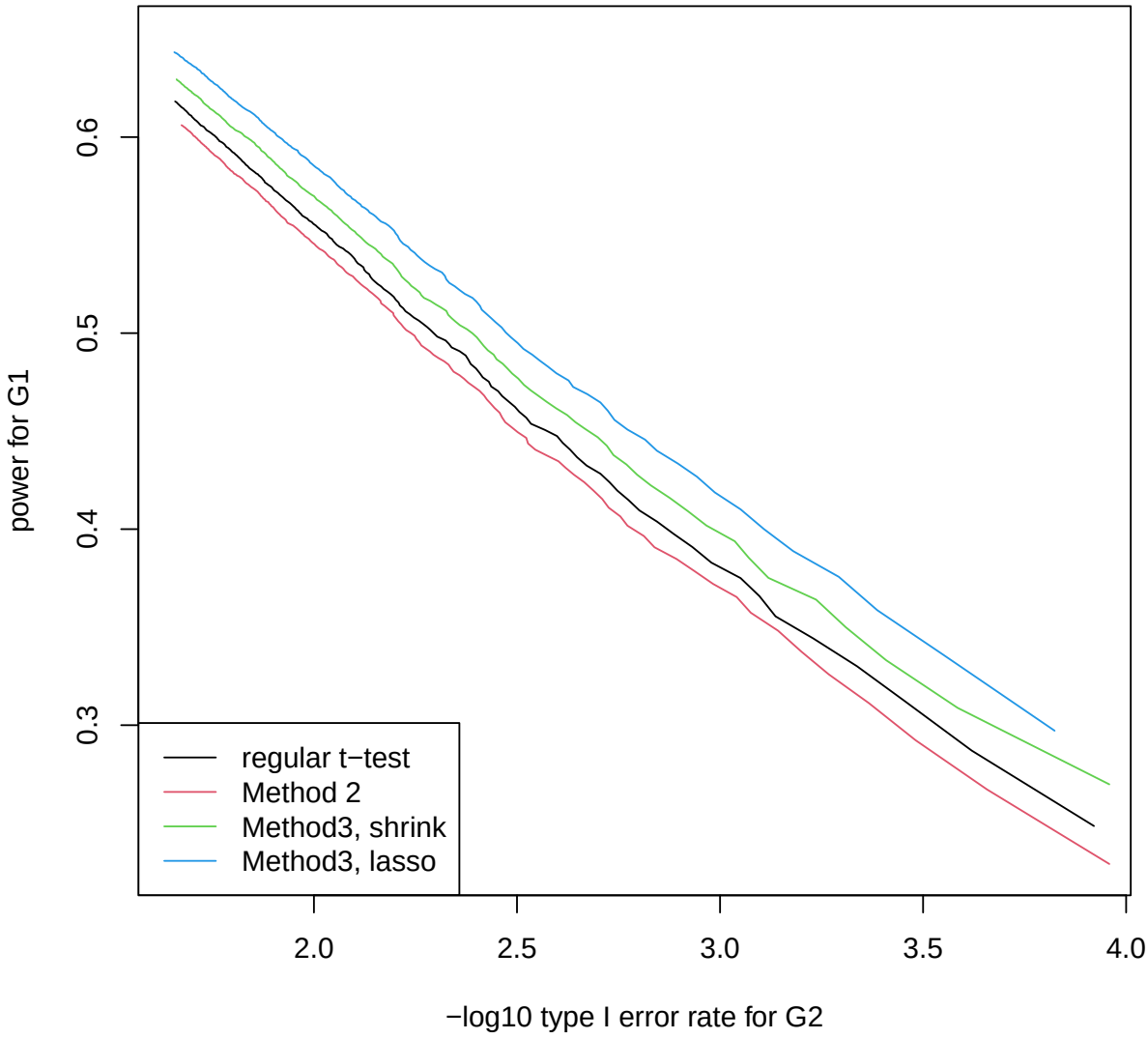


Figure 5.11: **Power curves for Bernoulli Methods** x-axis: -log10 scaled type 1 error for testing $G_2 \circ Z$; y-axis: power for testing $G_1 \circ Z$

5.2.3 Other simulation settings

We compare the performances of the conventional methods (t/Wald test) and our methods in 3 simulation settings: Non-GRM case, GRM case 1 and GRM case 2.

Non-GRM case In each replicate, we simulate a Bernoulli Z and $m = 4$ Bernoulli $G_j's$ independently for $n = 1000$ independent individuals, where the Bernoulli frequencies are generated independently from Unif(0.1, 0.5):

- $n = 1000,\ m = 4$
- $z \sim Ber(m_z),\ m_z \sim Unif(0.2, 0.8)$
- $x_j \overset{\text{indep.}}{\sim} Ber(m_j),\ m_j \overset{\text{iid}}{\sim} Unif(0.2, 0.8),\ cor(x_j, z) = 0,\ j = 1, 2, 3, 4$

When simulate each G_j, we check the following 2 conditions and keep re-generating G_j until both of the 2 conditions are satisfied:

1. $|cor(G_j, Z)| \leq 0.1$
2. $MCC(G_j, Z) \geq 5$

We simulate Y under the alternative model 5.4

$$Y = \alpha + \gamma Z + \sum_{j=1}^{m} \beta_j G_j + \sum_{i=j}^{m} \delta_j (Z - m_Z)(G_j - m_j) + \epsilon,\ \epsilon \sim N(0, I_n) \tag{5.4}$$

We set

$$\begin{aligned}
\alpha &\sim \text{Unif}(-10, 10) \\
(\beta_1, \beta_2, \beta_3, \beta_4) &= \left(0, \sqrt{\frac{0.025}{\sigma_{G_2^2}}}, -\sqrt{\frac{0.025}{\sigma_{G_3^2}}}, \sqrt{\frac{0.05}{\sigma_{G_4^2}}}\right), \\
\gamma &= \sqrt{\frac{0.025}{\sigma_Z^2}}, \\
(\delta_1, \delta_2, \delta_3) &= (0, 0, 0) \\
\delta_4 &= 0 \text{ (Null case)} \\
\delta_4 &= \sqrt{\frac{0.025}{\sigma^2_{(G_4 \circ Z)}}} \text{ (Alternative case)}
\end{aligned} \tag{5.5}$$

so that Z, G_2, G_3, G_4 have marginal effects on Y and only G_4 has interactive effect with Z on Y.

We run 10^5 replicates for the non-GRM case. Figure 5.12 are (differenced) QQ-plots of -log10 p-values for Gaussian Method1-4 under the null case ($\delta_4 = 0$).

Figure 5.13 are QQ-plots of of -log10 p-values for the following Bernoulli versions of TINGA: (1). Method1. (2). Method2. (3). Method3, shrink. (4). Method3, lasso. (5). Method4, shrink. (6). Method4, lasso.

The results are from the same simulation replicates as in Figure 5.12 for the Gaussian Methods.

For the alternative case, we let

$$\delta_4 = \sqrt{\frac{0.025}{\sigma^2_{(G_4 \circ Z)}}} \tag{5.6}$$

Figure 5.14 compares the power curves for the regular t-test with TINGA methods. The top panel compares regular t-test and Gaussian Method 1-4. The bottom panel compares regular t-test and the 6 different Bernoulli methods as those applied in above type 1 error part. The top and bottom panels are from the same 10^5 replicates. The x-axis is -log10 of the type 1 error for testing $G_2 \circ Z$. The y-axis is the power for testing $G_4 \circ Z$.

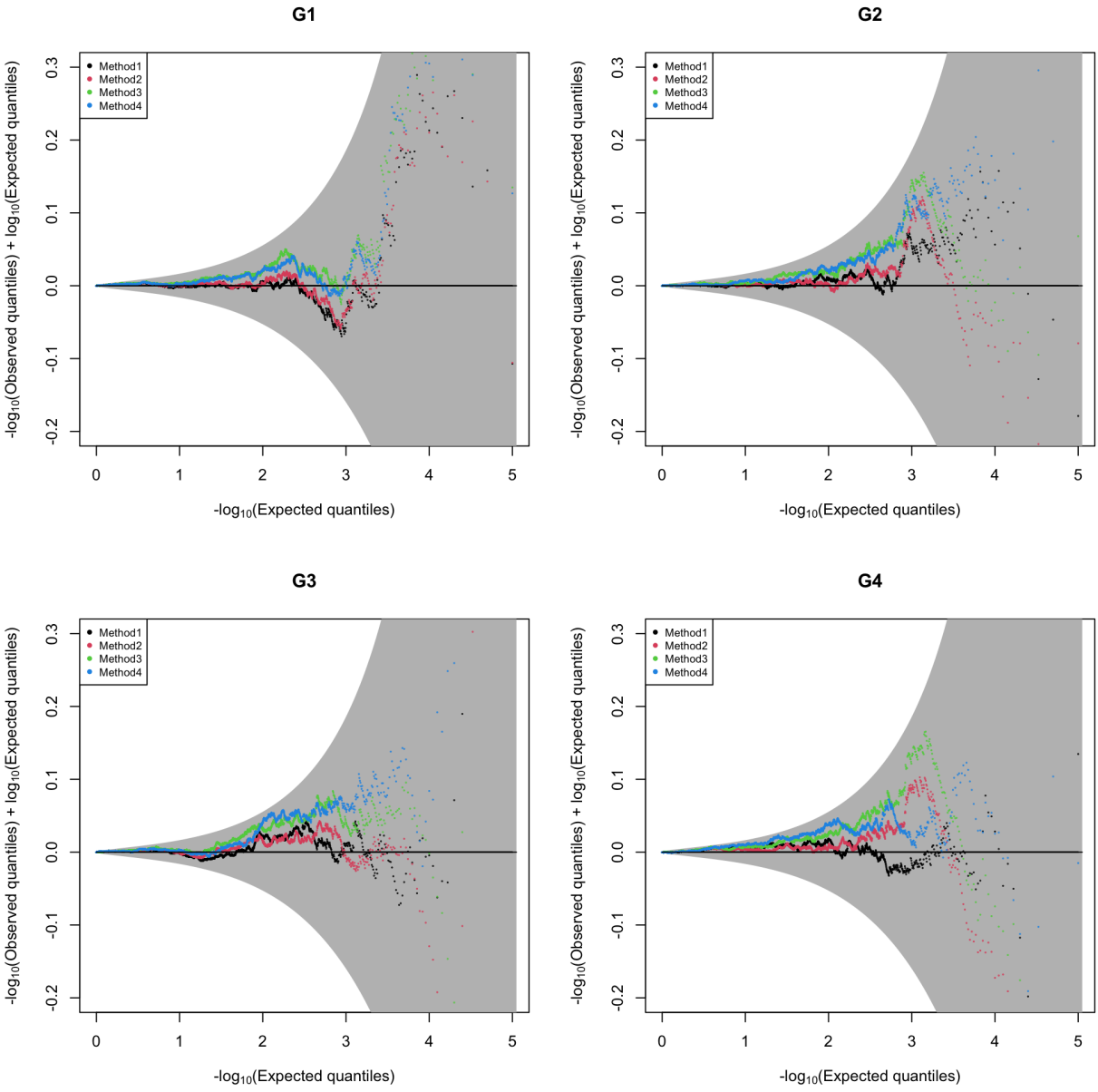


Figure 5.12: **QQ-plots for Gaussian Methods** 10^5 replicates under the null case of no interaction. 4 panels represents tests for interaction between G_1, G_2, G_3, G_4 and Z

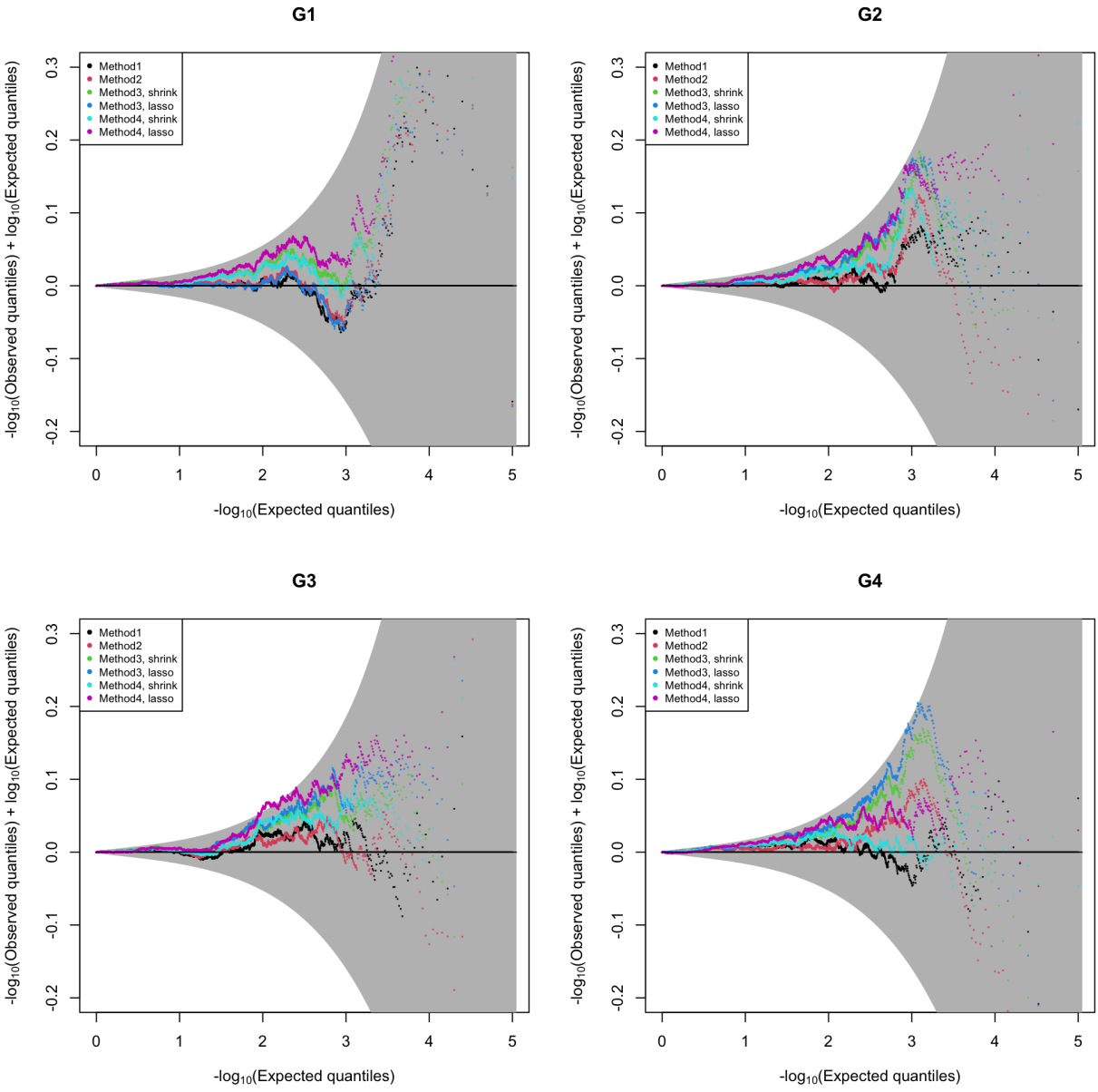


Figure 5.13: **QQ-plots for Bernoulli Methods** Results from the same 10^5 replicates as in Figure 5.12. 4 panels represents tests for interaction between G_1, G_2, G_3, G_4 and Z

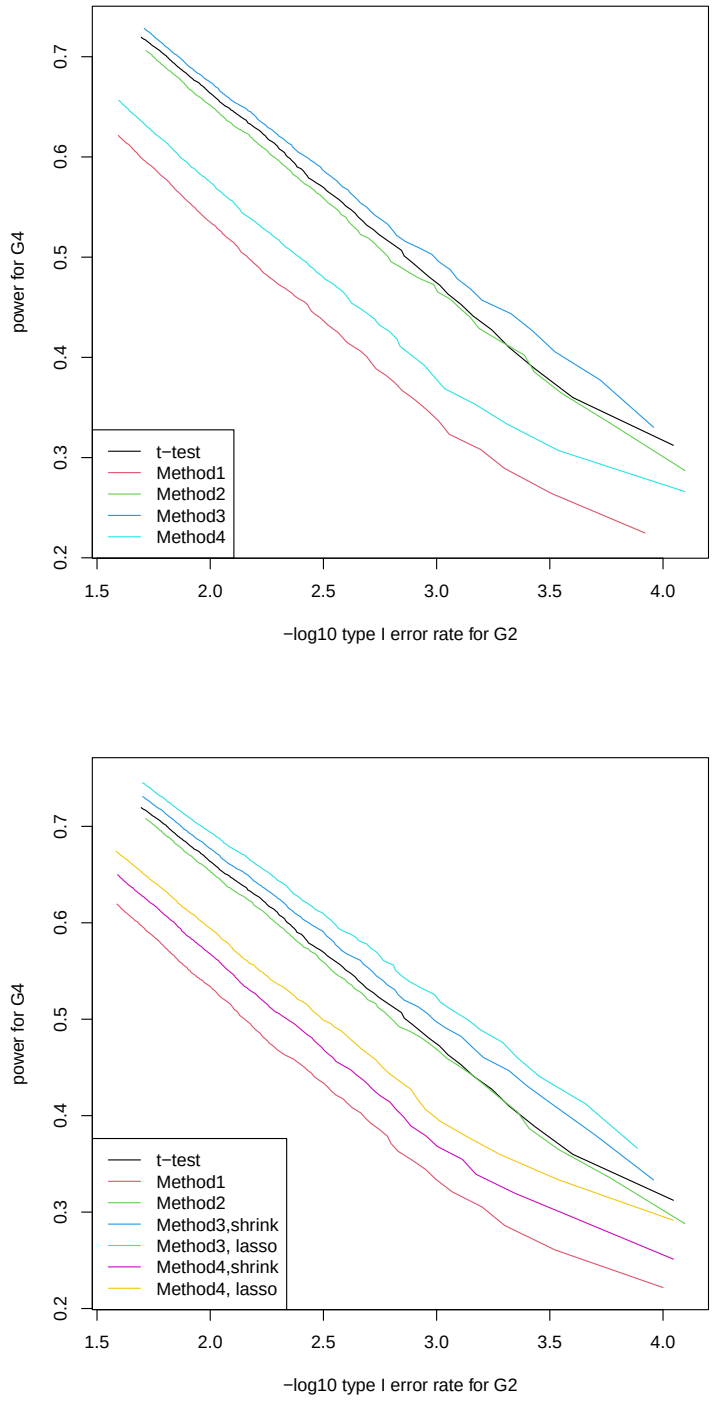


Figure 5.14: **Power curves** x-axis: -log10 scaled type 1 error for testing $G_2 \circ Z$; y-axis: power for testing $G_4 \circ Z$. Top: Gaussian Methods; Bottom: Bernoulli Methods. Top and bottom panels are results from the same 10^5 replicates under alternative case ($\delta_4 = \sqrt{\frac{0.025}{\sigma^2_{(G_4 \circ Z)}}}$)

GRM case 1: independent individuals In this case, Z, G_j's are simulated in the same way as GRM case 1. Y is simulated with the same model 5.4 and the same effect parameters as in 5.5, except a GRM as an extra variance component 5.7

$$\epsilon \sim N(0,\ \sigma_T^2 \left(h^2 K + (1-h^2) I\right), \tag{5.7}$$

where K is the GRM computed from 10^4 independently simulated Bernoulli SNPs, $h^2 = 0.3, \sigma_T^2 = 1$:

$$G = (g_1, ..., g_{10000}),\ g_i \overset{indep.}{\sim} Ber(f_i),\ f_i \overset{iid}{\sim} Unif(0.1, 0.9)$$
$$\tilde{G} = (\tilde{g_1}, ...\tilde{g}_{10000}),\ \tilde{g_i} = \frac{g_i - \bar{g_i}}{\sqrt{\bar{g_i}(1-\bar{g_i})}}$$
$$K = \frac{1}{10000}\tilde{G}\tilde{G}^T$$

.

We run 10^4 replicates for GRM case 1. Figure 5.15 are (differenced) QQ-plots of -log10 p-values from different methods under GRM case 1. Figure 5.16 are power curves of different methods.

GRM case 2: population structure with 3 sub-populations We also tried the setting with population structure in which there are 3 sub-populations following the Balding-Nichols model [47]. Both Z and G_j's are still Bernoulli distributed, and simulated with the 3 sub-populations. Z, G_2, G_3, G_4 and $(Z \circ G_4)$ have effects on Y. Y also has indicators of the sub-populations as covariates:

We assign 1/3 of the total population to each sub-population. In our case, $n = 1000$, so the sample sizes for sub-population 1, 2, 3 are $333, 333, 334$ respectively. Let the fixation index $F = 0.1$.

For each SNP s, the ancestral allele frequency p_s is drawn $\overset{\text{iid}}{\sim} Unif(0.2, 0.8)$ (iid across

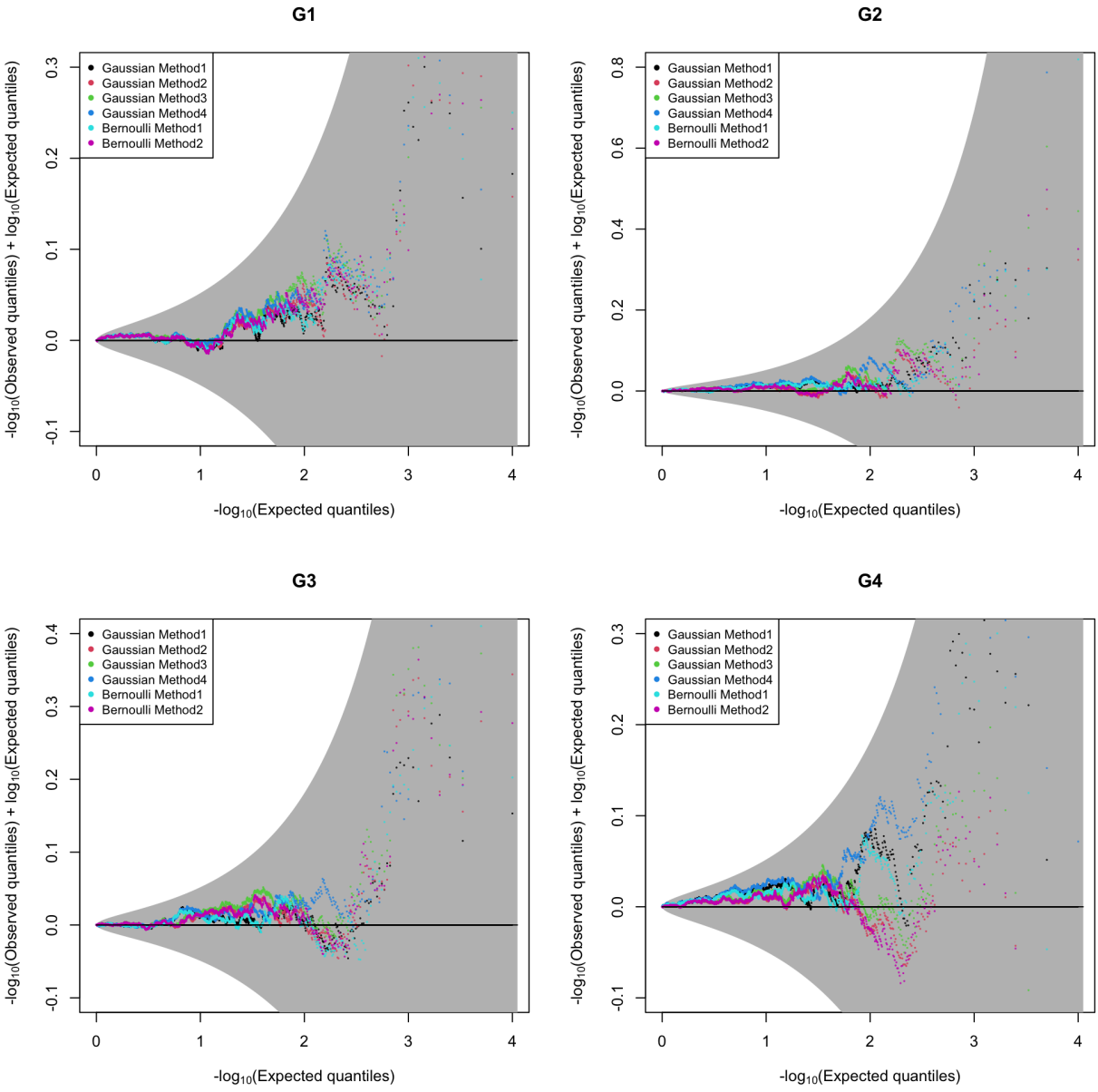


Figure 5.15: **QQ-plots for GRM case 1** 10^4 replicates. 4 panels represents tests for interaction between G_1, G_2, G_3, G_4 and Z

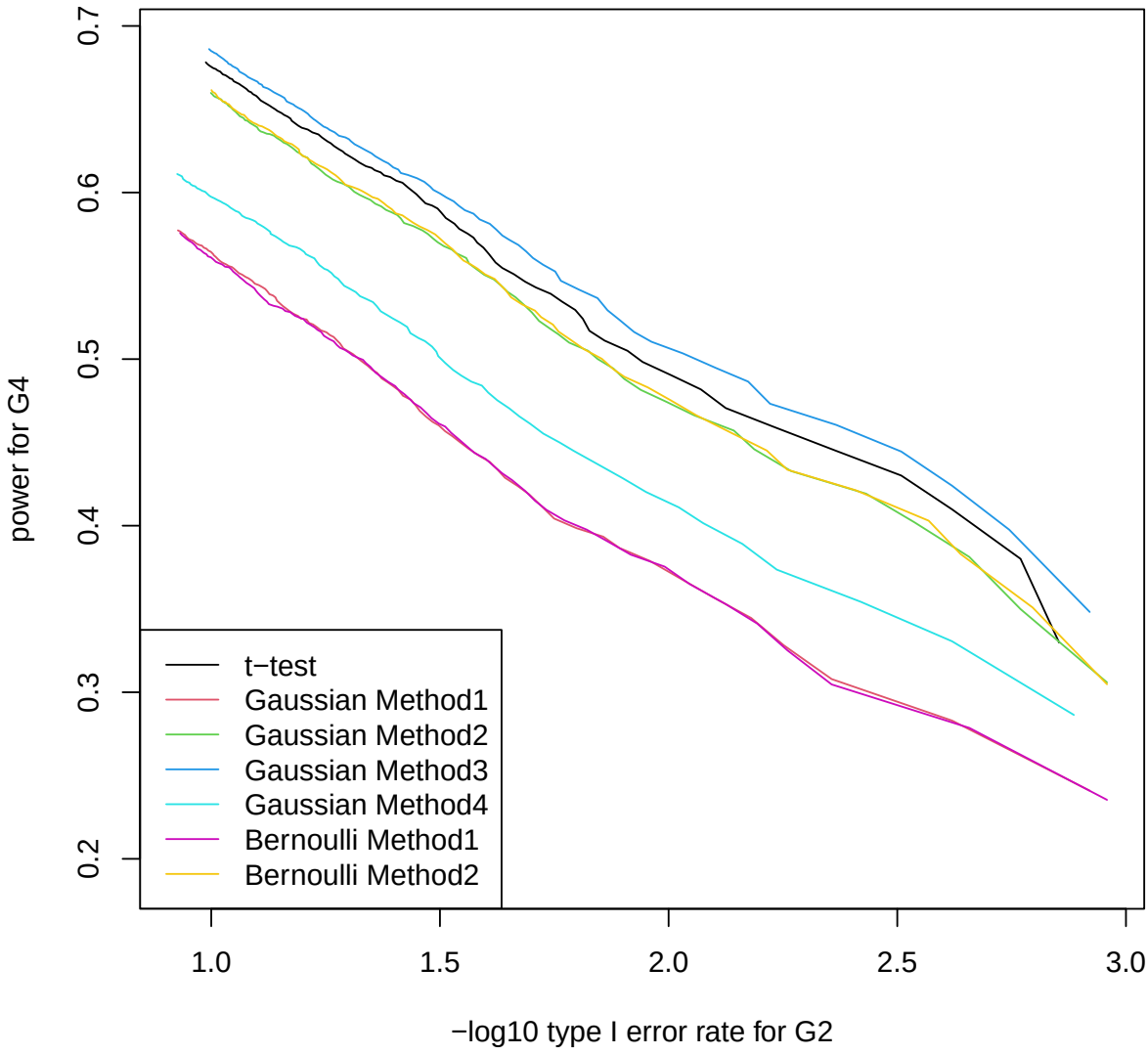


Figure 5.16: **Power curves for GRM case 1** 10^4 replicates. x-axis: -log10 scaled type 1 error for testing $G_2 \circ Z$; y-axis: power for testing $G_4 \circ Z$

SNPs). For each sub-population $k = 1, 2, 3$, the allele frequency p_k is drawn independently from $Beta(\frac{p_s(1-F)}{F}, \frac{(1-p_s)(1-F)}{F})$. Then for an individual assigned to sub-population k, the genotype is drawn iid from $Ber(p_k)$. We only keep the SNPs with MAF ≥ 0.05.

We use the above strategy to simulate one Z and $m = 4$ G_j's, independently. When simulate each G_j, we check the following 2 conditions and keep re-generating G_j until both of the 2 conditions are satisfied:

1. $|cor(G_j, Z)| \leq 0.1$
2. $MCC(G_j, Z) \geq 5$

For the GRM, we simulate 10^5 independent SNPs, these SNPs are also independent from Z and G_j's.

We simulate y by the following model:

$$Y = \alpha + \gamma Z + \sum_{j=1}^{m} \beta_j G_j + \sum_{i=j}^{m} \delta_j (Z - m_Z)(G_j - m_j) + \epsilon$$

where

- $(\beta_1, \beta_2, \beta_3, \beta_4) = \left(0, \sqrt{\frac{0.025}{\sigma_{G_2^2}}}, -\sqrt{\frac{0.025}{\sigma_{G_3^2}}}, \sqrt{\frac{0.05}{\sigma_{G_4^2}}}\right)$
- $\gamma = \sqrt{\frac{0.025}{\sigma_Z^2}}$
- $(\delta_1, \delta_2, \delta_3) = (0, 0, 0)$
- $\delta_4 = 0$ (Null case); $\delta_4 = \sqrt{\frac{0.025}{\sigma^2_{(G_4 \circ Z)}}}$ (Alternative case)
- $\epsilon \sim N(0,\ \sigma_T^2 \left(h^2 K + (1 - h^2) I\right)$
- $h^2 = 0.3$

We run 10^4 replicates for GRM case 2. Figure 5.17 are (differenced) QQ-plots of -log10 p-values from different methods under GRM case 2. Figure 5.18 are power curves of different methods.

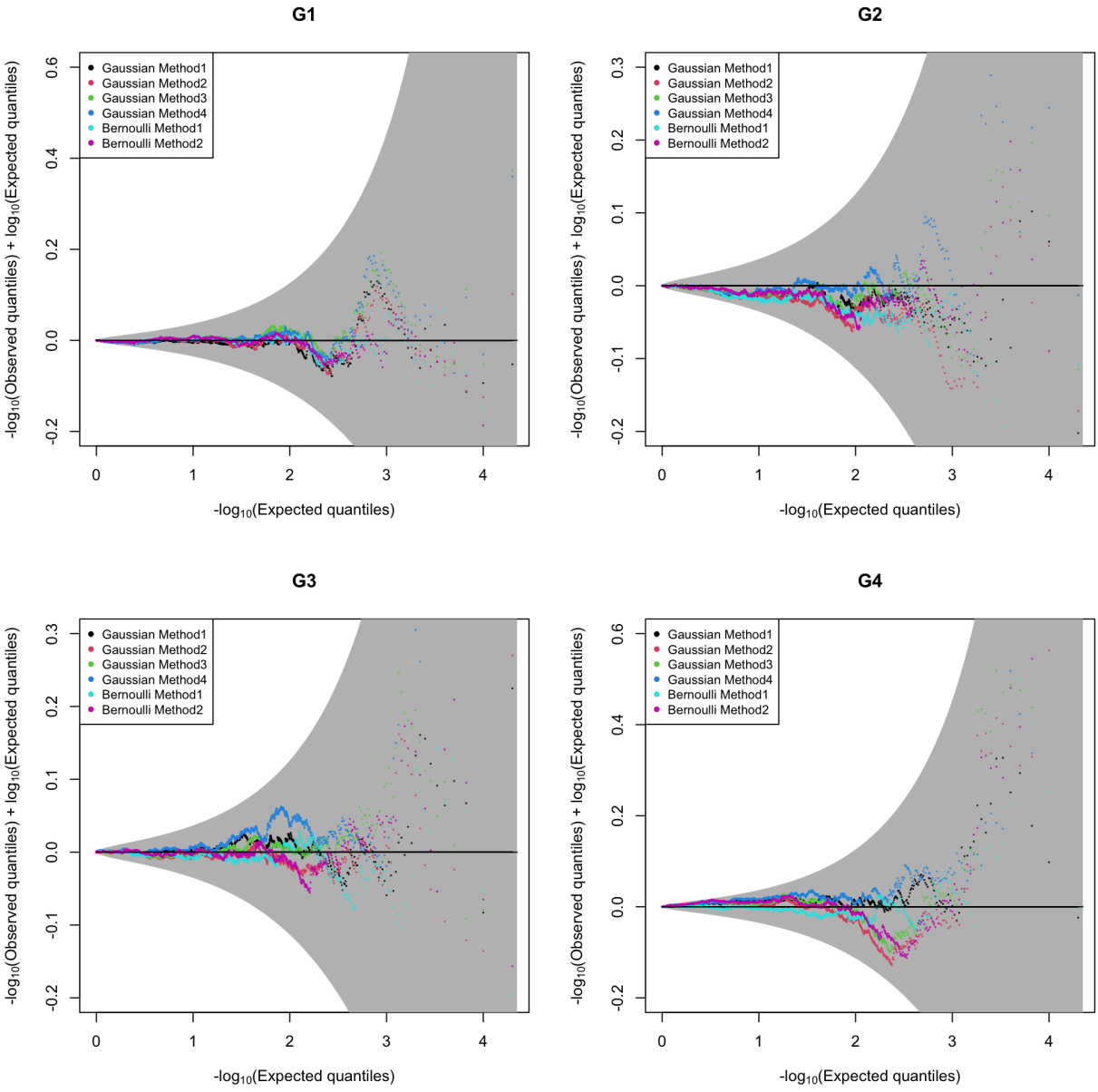


Figure 5.17: **QQ-plots for GRM case 2** 10^4 replicates. 4 panels represents tests for interaction between G_1, G_2, G_3, G_4 and Z

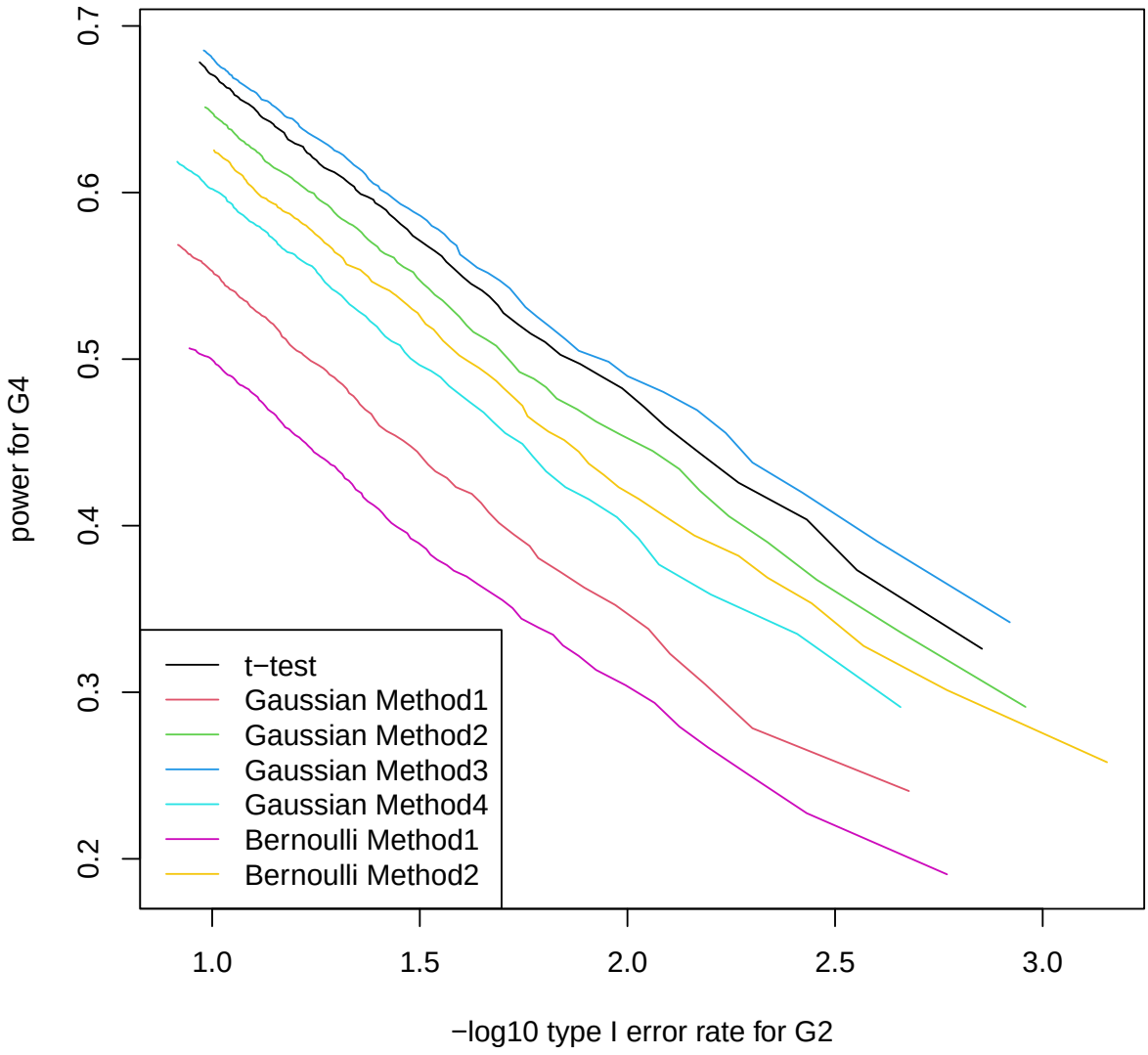


Figure 5.18: **Power curves for GRM case 2** 10^4 replicates. x-axis: -log10 scaled type 1 error for testing $G_2 \circ Z$; y-axis: power for testing $G_4 \circ Z$

5.3 Effect of different minor allele frequencies of Z and G_j on type 1 error

In our TINGA method, we intentionally denote the 2 SNPs by different letters Z and G, to emphasis the situation where Z is fixed and we test the interaction between Z and other SNPs in the genome, and our TINGA method is conditioning on "Z" and is not symmetric between Z and G. However, when searching for pairwise interaction among all possible pairs of SNPs in a GWAS, the 2 SNPs, say, G_k and G_j are actually symmetric. Therefore, there could be two possible ways to apply TINGA:

1. Let G_k be the "Z", condition on G_k, Y and treat N_j as a function in random variable G_j
2. Switch the role: Let G_j be the "Z", condition on G_j, Y and treat N_k as a function in random variable G_k

In this section, we investigate the condition under which our method is applicable and which SNP to condition on in a pairwise search for interaction signals. We focus on Gaussian approach Method 3.

We have the following simulation setting:

- $Z \sim \text{Bernoulli}(m_Z)$
- $G_1 \sim \text{Bernoulli}(m_1)$
- $G_j \overset{\text{indep.}}{\sim} \text{Bernoulli}(m_j),\ m_j \sim \text{Unif}(0.2, 0.5),\ j = 2, 3, 4, 5$
- $\alpha \sim \text{Unif}(-10, 10)$
- $Y = \alpha + \gamma Z + \beta_1 G_1 + \sum_{j=2}^{5} \beta_j G_j + \epsilon,\ \epsilon \sim N(0, \sigma_e^2 I)$
- $\sigma_e^2 = 1 - \gamma^2 \sigma_Z^2 - \sum_{j=1}^{5} \beta_j^2 \sigma_{G_j}^2$

For the allele frequencies of Z and G_1, we use a 317×317 fixed grid: we pick 317 equally distanced points on the interval $[0.05, 0.95]$ and iterate m_Z and m_1 over all possible pairs of frequencies among the 317 points, resulting to $317^2 = 100489$ replicates.

We look at the p-values for testing the interaction between Z and G_1 in the same 7 cases as in section 5.2.1:

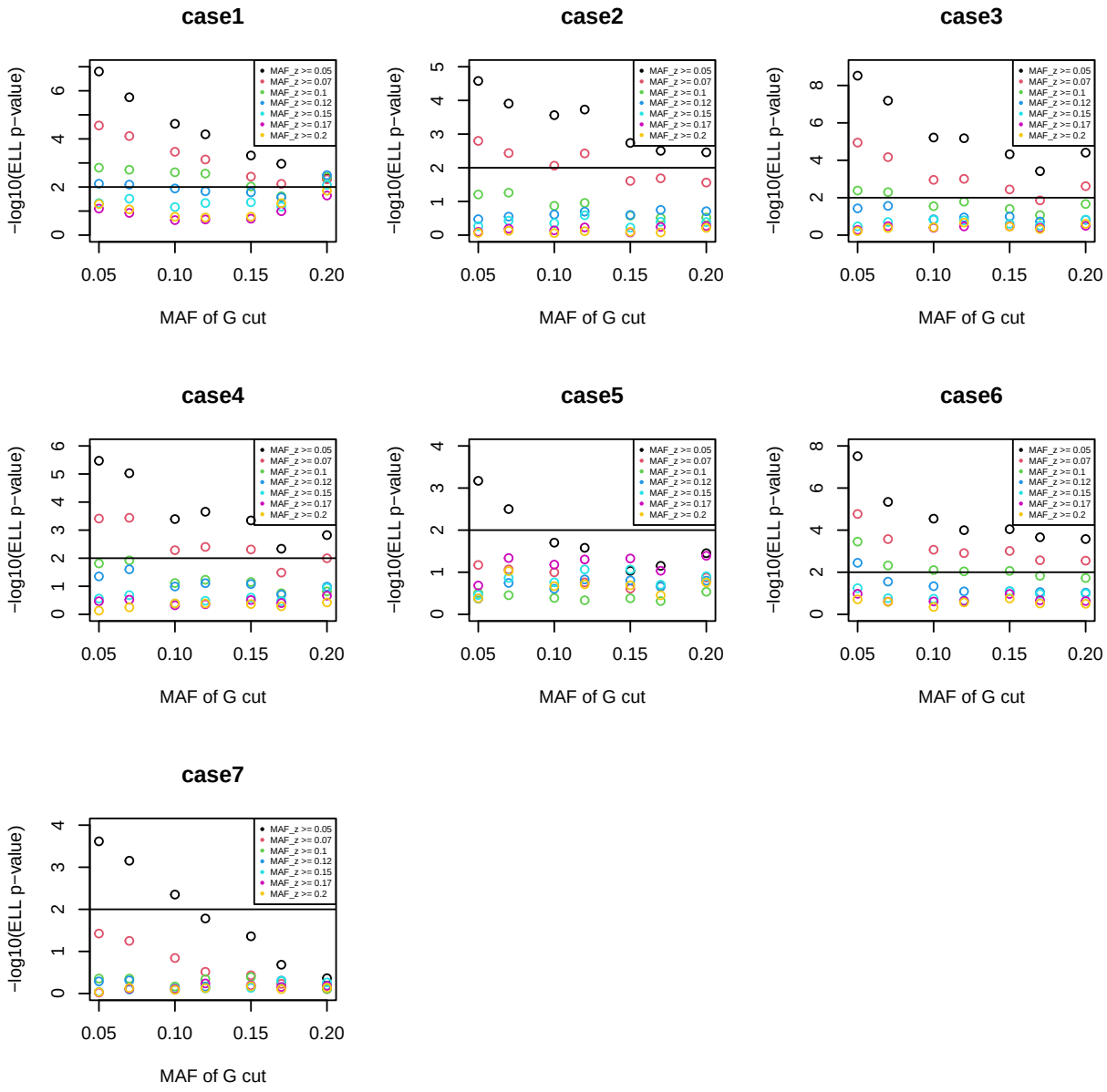
1. $\gamma = 0,\ (\beta_1, \beta_2, \beta_2, \beta_4, \beta_5) = (0, 0, 0, 0, 0)$
2. $\gamma = 0,\ \beta_1 = 0,\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$
3. $\gamma = \sqrt{\frac{0.01}{\sigma^2_Z}},\ \beta_1 = 0,\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$
4. $\gamma = 0,\ \beta_1 = \sqrt{\frac{0.01}{\sigma^2_{G_1}}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$
5. $\gamma = \sqrt{\frac{0.01}{\sigma^2_Z}},\ \beta_1 = \sqrt{\frac{0.01}{\sigma^2_{G_1}}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$
6. $\gamma = \sqrt{\frac{0.01}{\sigma^2_Z}},\ \beta_1 = \sqrt{\frac{0.04}{\sigma^2_{G_1}}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$
7. $\gamma = \sqrt{\frac{0.04}{\sigma^2_Z}},\ \beta_1 = \sqrt{\frac{0.01}{\sigma^2_{G_1}}},\ (\beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma^2_{G_2}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_3}}}, \sqrt{\frac{0.01}{\sigma^2_{G_4}}}, -\sqrt{\frac{0.01}{\sigma^2_{G_5}}})$

We simulate G_2, G_3, G_4, G_5 to mimic the situation where there are other positively or negatively associated SNPs. We assign different values to γ and β_1 to see how the cutoff for MAF of Z and cutoff for MAF of G_1 is related to their effect coefficients.

For each case, we run the 317^2-replicate simulation 3 times, resulting $3 \times 317^2 = 301467$ replicates in total.

We plot the -log10 scaled ELL p-value [1] for testing the uniformity of the resulting 301467 p-values for different ranges of MAFs of Z and G_1 in Figure 5.19. We can see that when MAF of $Z \geq 0.15$ (or 0.12) and MAF of $G_1 \geq 0.05$, the ELL p-values are all larger than 0.01.

Figure 5.19: **ELL p-values for different ranges of minor allele frequencies for Z and G_1**



Each point is the -log10 scaled ELL p-values [1] for a sub-vector of the 301467 p-values. The sub-vector is given by taking the replicates where MAF of $Z \geq a$ and MAF of $G_1 \geq b$. The cutoff a for Z is denoted by different colors; the cutoff b for G_1 is the x-axis.

We then look at the power performance under the condition that (1). MAF of $Z \geq 0.15$ and MAF of $G_1 \geq 0.05$. (2). both MAF of Z and MAF of $G_1 \geq 0.12$ We still use the model in the type 1 error simulation:

$$\begin{gathered} Y = \alpha + \gamma Z + \beta_1 G_1 + \sum_{j=2}^{5} \beta_j G_j + \sum_{j=1}^{5} \delta_j (G_j \circ Z) + \epsilon,\ \epsilon \sim N(0, \sigma_e^2 I) \\ \sigma_e^2 = 1 - \gamma^2 \sigma_Z^2 - \sum_{j=1}^{5} \beta_j^2 \sigma_{G_j}^2 - \sum_{j=1}^{5} \delta_j^2 \sigma_{G_j \circ Z}^2 \\ \gamma = \sqrt{\frac{0.02}{\sigma_Z^2}} \\ (\beta_1, \beta_2, \beta_2, \beta_4, \beta_5) = (\sqrt{\frac{0.01}{\sigma_{G_1}^2}}, \sqrt{\frac{0.02}{\sigma_{G_2}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_3}^2}}, \sqrt{\frac{0.01}{\sigma_{G_4}^2}}, -\sqrt{\frac{0.01}{\sigma_{G_5}^2}}) \\ (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5) = (0, \sqrt{\frac{0.02}{\sigma_{G_2 \circ Z}^2}}, 0, 0, 0) \end{gathered} \tag{5.8}$$

In this case, Z is interacting with G_2 only. We look at the QQ-plot for the p-values for testing $G_1 \circ Z$ in Figure 5.20 (a), where we restrict the MAF of G_1 to be greater than or equal to 0.05 and MAF of Z to be greater than or equal to 0.15 and 0.12, ending up with about 95,000 replicates and in Figure 5.20 (b), where we let both Z and G_1 have MAF ≥ 0.12, ending up with about 81,000 replicates.

We compare the power for detecting $G_2 \circ Z$ of the regular t-test, heteroscedasticity corrected t-test and Gaussian approach method 3. Here the heteroscedasticity corrected t-test is performed by fitting a weighted least square model, where the weights are $\frac{1}{\widehat{\text{Var}}(Y|Z=0)}$ for the individual with $Z_i = 0$ and $\frac{1}{\widehat{\text{Var}}(Y|Z=1)}$ for the individual with $Z_i = 1$. Figure 5.20 (c) are the power curves when we restrict the MAF of G_1, G_2 to be greater than or equal to 0.05 and MAF of Z to be greater than or equal to 0.15 and 0.12, ending up with about 87,651 replicates. Figure 5.20 (d) are the power curves when we restrict the MAF of G_1, G_2, Z to be greater than or equal to 0.12, ending up with about 81,000 replicates.

We conclude when testing the interaction $G_j \circ Z$ where Z is the SNP that we choose to condition on, having (1). MAF of $Z \geq 0.15$ and MAF of $G_1 \geq 0.05$ or (2). both MAF of Z and MAF of $G_1 \geq 0.12$ is needed for Gaussian approach Method 3 to have good type 1

error performance.

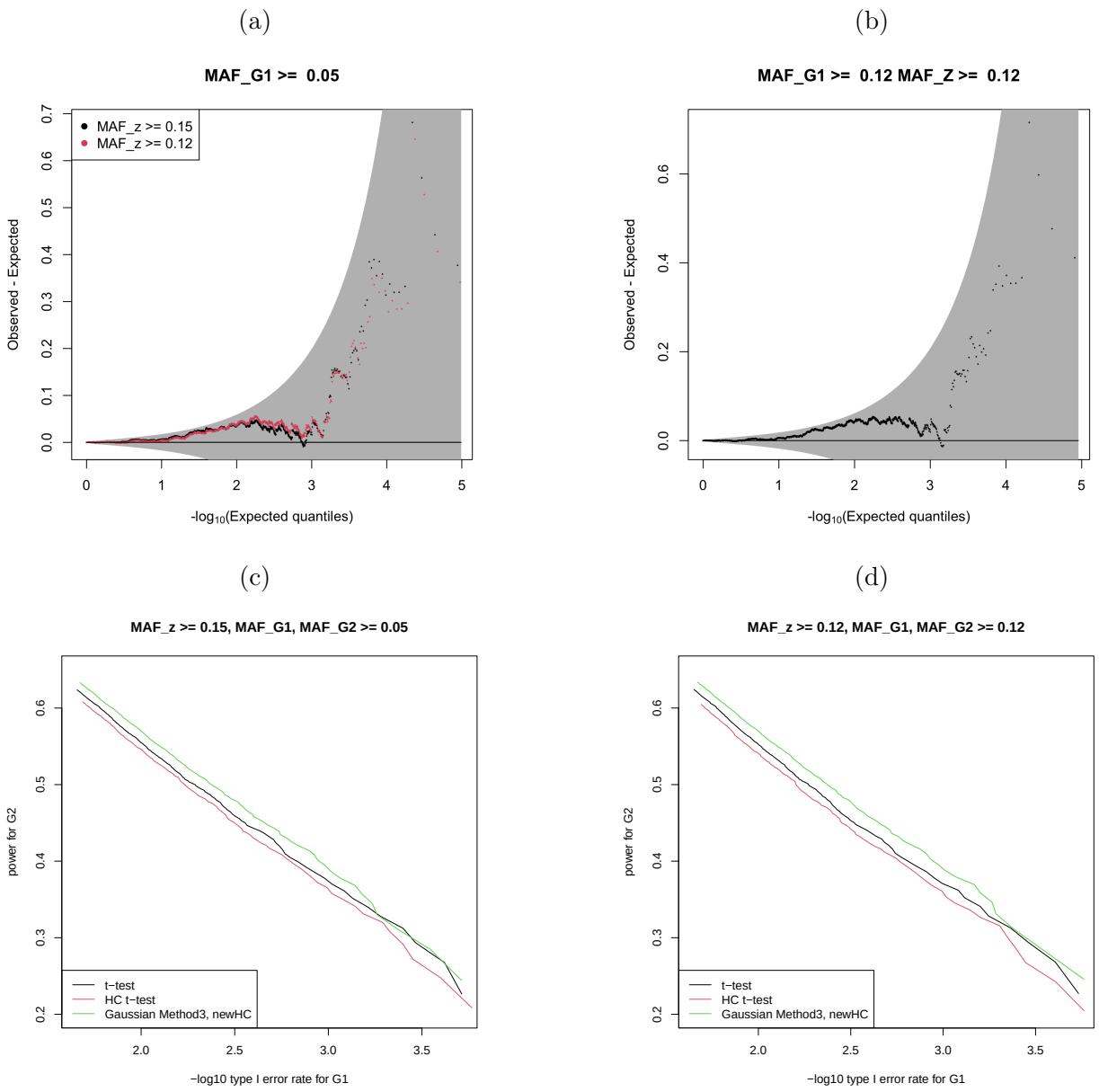


Figure 5.20: **QQ-plots and power curves for different MAF lower bounds**

(a). (Differenced) QQ-plot of -log10 p-values when MAF of $G_1 \geq 0.05$ and MAF of $Z \geq 0.15$ (black), 0.12 (red). (b). (Differenced) QQ-plot of -log10 p-values when MAF of $G_1 \geq 0.12$ and MAF of $Z \geq 0.12$. (c) Power curves when MAF of $G_1, G_2 \geq 0.05$ and MAF of $Z \geq 0.15$. x-axis is the -log10 scaled type 1 error for G_1, y-axis is the power for $(G_2 \circ Z)$. "HC t-test" means heteroscedasticity corrected t-test. (d). Power curves when MAF of $G_1, G_2 \geq 0.12$ and MAF of $Z \geq 0.12$

5.4 Effect of different minor allele frequencies of Z and G_j on power

From above section, we conclude that in a scenario where we perform a pairwise search among all possible pairs of SNPs in a genome, then if one of the SNP has MAF ≥ 0.15 and another has MAF $\in [0.05, 0.15)$, we should let Z be the one that has MAF ≥ 0.15. That is to say, we condition one the SNP that has larger MAF. When both SNPs have MAF ≥ 0.15, conditioning on either of them will lead to acceptable type 1 error performance.

In this section, we design an experiment where the MAFs for both Z and G_j are ≥ 0.15, and we assign a relatively small MAF for G_j and a relatively large MAF for Z, to see if the power performance of our method changes when the frequencies change.

For each choice of coefficients, we use the following setting to simulate $N = 10^4$ pairs of (Z, G_j) independently:

- $n = 1000$
- $G_j \sim Ber(0.15),\ Z \sim Ber(0.35),\ G_j, Z$ independent
- $Y = \alpha + bG_j + rZ + \delta(G_j - \mu_j)(Z - \mu_Z) + \epsilon,\ \epsilon \sim\ N(0,\ 1)$
- $\alpha = 1,\ \delta = \sqrt{\frac{0.025}{\sigma^2_{G_j}\sigma^2_Z}} = 0.928$
- (G_j, Z) are filtered by criterion of minimum cell count ≥ 20

We tried the following 3 cases:

1. $b = r = 0$
2. $b = \sqrt{\frac{0.025}{\sigma^2_G}} = 0.44, r = 0$
3. $b = 0,\ r = \sqrt{\frac{0.025}{\sigma^2_z}} = 0.33$

We then test interaction between the two variants using Gaussian Method3 with (1) Z being the variant with smaller MAF and G_j being the variant with larger MAF and (2) the reverse (Z being the variant with larger MAF and G_j being the variant with smaller MAF). Fig 5.21 are QQ-plots of the resulting p-values on the -log10 scale. We can see that Method 3 has slightly larger power for the reversed pair, where Z is chosen to be the one that has smaller MAF. However, the difference is too small. For example, at p-value cutoff 10^{-5}, the power for the original pair is

$$p_o = 0.7064,$$

the power for the reversed pair is

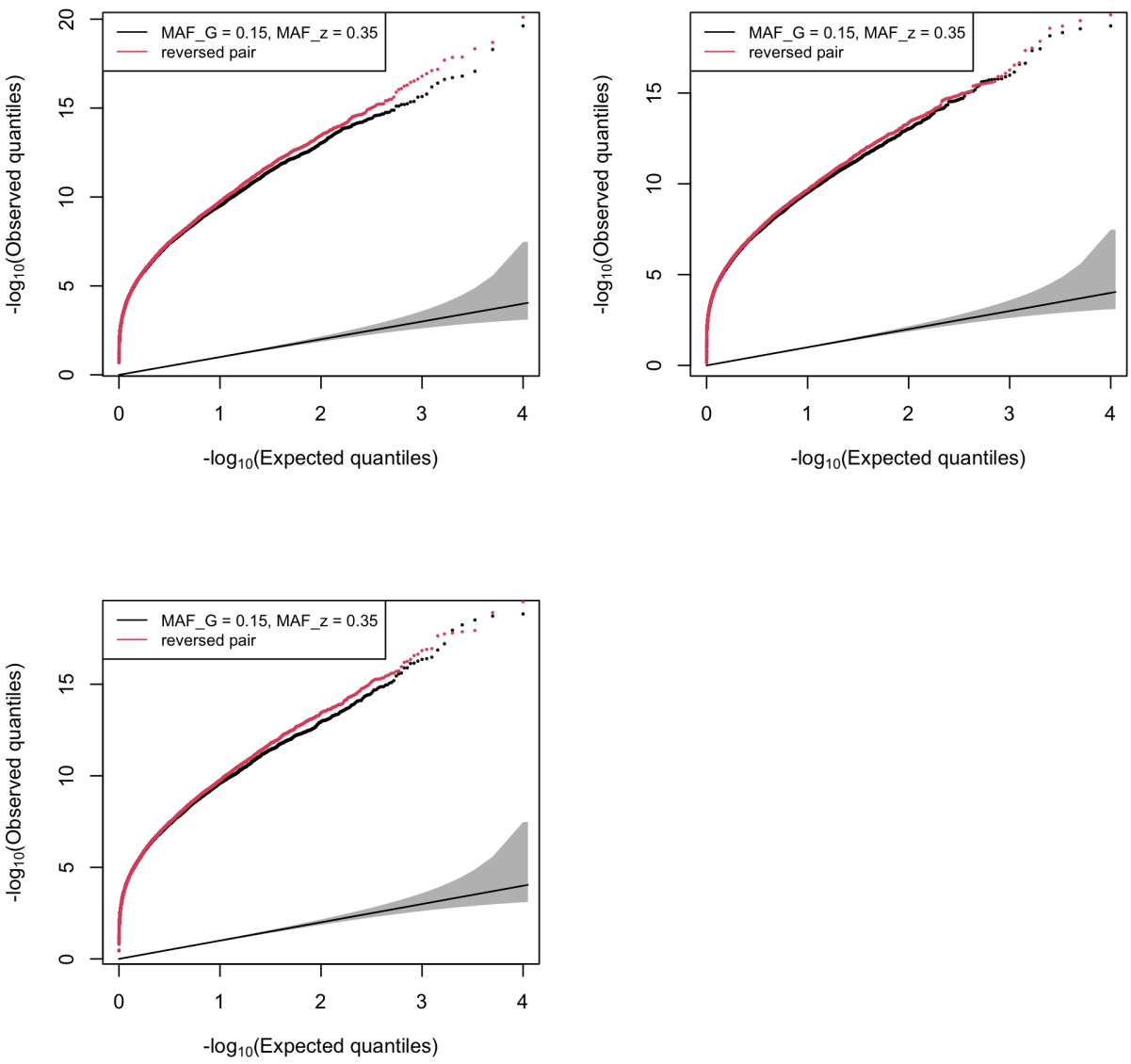
$$p_r = 0.711$$

The z-score for testing the significance of the difference is

$$z = \frac{p_r - p_o}{\sqrt{\frac{p_r(1-p_r)}{N} + \frac{p_o(1-p_o)}{N}}} = 0.716 \tag{5.9}$$

and the p-value is 0.24, so the difference between the power of the original pair and the power for the reversed pair is not significant. Therefore, we may conclude that when the MAFs of both 2 SNPs to be tested for interaction are ≥ 0.15, then whether to condition on one or another does not affect the power or type 1 error performance. In this case, we may choose which one to condition on based on correction of the "feast or famine" effect: we may choose the one that potentially has worse "feast or famine" effect to be Z. By doing so, we can get larger correction.

Figure 5.21: $-log10$ **scaled p-values from TINGA of reversed against original pairs of** $Z,\ G_j$



x-axis is for the original pair, y-axis is for the reversed pair. Both $Z,\ G_j$ are Bernoulli distributed. For the original pair, frequency of G_j is 0.07, frequency for Z is 0.25. Y is generated by $Y = \alpha + bG_j + rZ + \delta(G_j - \mu_G)(Z - \mu_Z) + \epsilon,\ \alpha = 1,\ \delta = \sqrt{\frac{0.025}{\sigma_G^2 \sigma_Z^2}}$. (a). $b = r = 0$. (b). $b = \sqrt{\frac{0.025}{\sigma_G^2}},\ \ r = 0$. (c). $b = 0,\ r = \sqrt{\frac{0.025}{\sigma_z^2}}$

CHAPTER 6

ANALYSIS OF FLOWERING TIME IN *A. THALIANA*

In this chapter we show the application of our methods on the *A. thaliana* dataset, for which the genotypes are binary.

6.1 Data Description

We study the genetic data from Arabidopsis thaliana and use its flowering time, the log scaled number of days between germination and flowering at 10°C , as phenotype [48]. We include 931 selected *A. thaliana* accessions (inbred lines) from difference regions. Therefore, the genotypes are binary (has 0/1 values only). The SNPs were filtered based on minor allele frequency (MAF) ≥ 0.03 [49]. LD pruning was done to remove loci in pairwise LD of $r^2 > 0.99$ [49]. After filtering, there are 865,350 SNPs remained.

It is worth noting that the flowering time used as our trait is obtained by taking average over 10 identical accessions: 10 "individuals" of *A. thaliana* that have exactly the same genetic data. In this way, environmental noise is reduced. The estimated heritability is about 88%. Before taking the average, it was about 44%.

In our analyses with this dataset, we use a LMM for the phenotype, where the GRM is computed by all SNPs with allele frequency ≥ 0.05.

6.2 Performance for one particular SNP

Firstly, we pick some special SNPs as Z and do an interaction GWAS to have an rough idea of the performance of different methods.

We pick SNP Chr5_18593622 as our first Z because it has relatively small marginal p-value. G_j's are all SNPs in the genome that has correlation with Z between -0.1 to 0.1.

- frequency of Z: 0.28

- number of G_j's: 696396

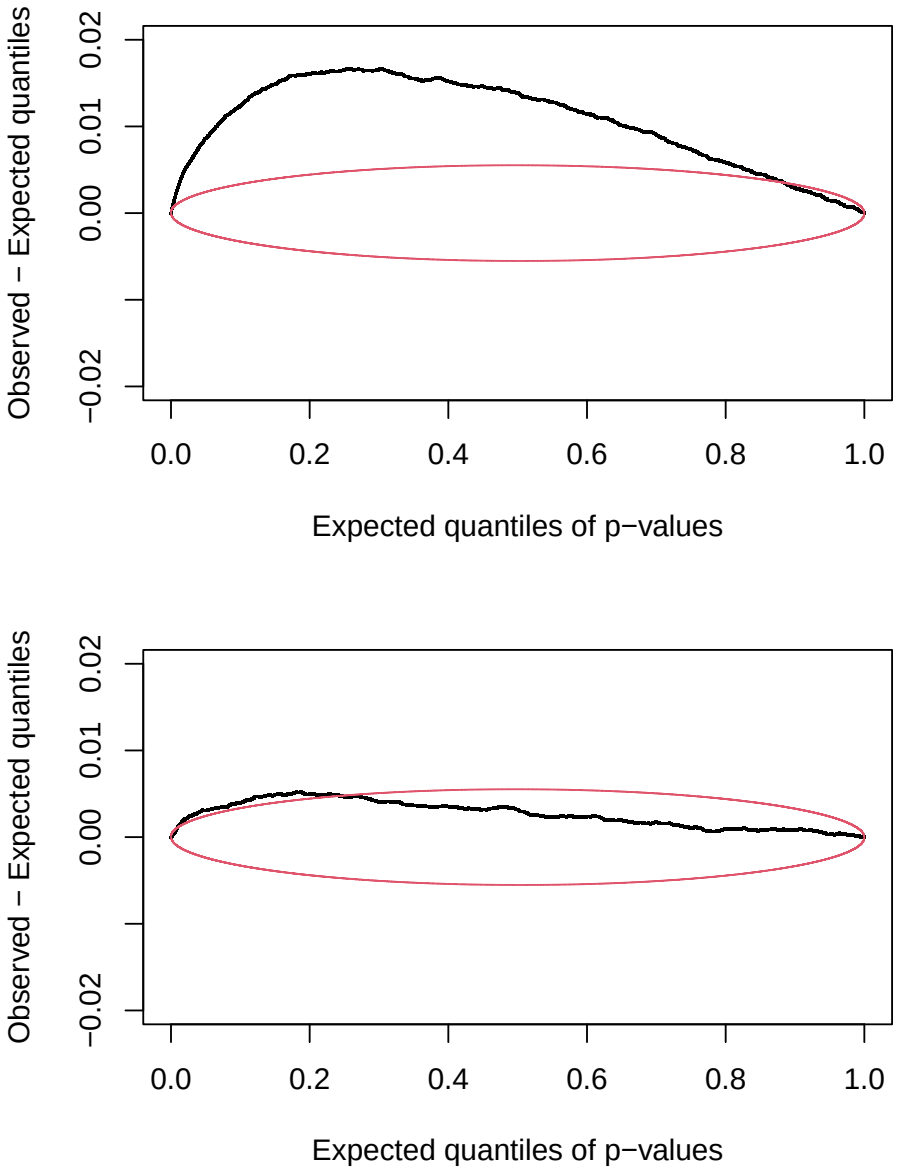
We test for interaction between selected Z and G_j's using Wald test and TINGA.

Fig 6.1 are (differenced) QQ-plots of p-values from different methods, with simultaneous 95% ELL acceptance regions for i.i.d. uniform p-values outlined in red, where these use the method of [1]. (In a differenced QQ-plot, the y-axis depicts the difference between observed and expected p-values, which is particularly helpful for creating a useful visualization when the plot contains a large number of points.) We can see that for this particular SNP, the distribution of p-values is much closer to uniform after TINGA adjustment.

6.3 "feast or famine" problem persists for simulated G_j's

As we mentioned in previous section, in an interaction GWAS, the true null distribution of the testing statistic is some distribution that depends on (Y, Z). Therefore, we may expect to get similar pattern of the "feast or famine" effect for the same pair of Y, Z but with different set of G_j's. To illustrate the persistence of "feast or famine" effect, we pick one SNP in the *A. thaliana* dataset that has small GCIF for genome-wide interaction tests, denote it as our Z. We apply MAF filtering and LD pruning using PLINK to the *A. thaliana* dataset and get a set of 8183 SNPs with MAF ≥ 0.1 and pairwise LD $r^2 < 0.075$. We first conduct interaction tests between Z and each of the 8183 SNPs, and the trait Y is still the flowering time we use for the real data analysis. The QQ-plots and GC inflation factors of the interaction p-values are in Fig 6.2 (a). Then we keep the (Y, Z) pair the same, and simulate 8183 SNPs from Bernoulli distribution. These simulated SNPs are independent of (Y, Z) and independent of each other. We conduct the interaction tests again and get the QQ-plots in Fig 6.2 (b). As we can see, for both the original SNPs and simulated SNPs, the -log10 p-values of Wald test are severely deflated, and our methods have good performance in fixing this issue. Similar to the results in Fig 2.1 and Table 2.1, this experiment also gives us an idea of how the "feast

Figure 6.1: **QQ-plots of p-values**



Fix SNP Chr5_18593622 as Z and test its interaction with other SNPs in the genome. The theoretical uniform quantiles are subtracted from both x and y coordinates, so that it has a horizontal view. The red lines are the 95% ELL null confidence interval assuming there are 100,000 effective independent SNPs. Top: Unadjusted interaction p-values from Wald test. Bottom: p-values from TINGA.

or famine" effect is related to the particular value of (Y, Z) and presumably can be predicted by it.

6.4 Strategy for detecting epistasis

Due to the large number of SNPs (totally 865,350), we are not able to do a pairwise search over all possible pairs of SNPs for epistasis. Therefore, we first apply some filtering procedures to narrow down the set of SNP pairs that we apply TINGA to. Then for the 2 SNPs in a pair, we decide which SNP to condition on based on their MAFs and a diagnostic value for the severity of the "feast or famine" effect (see Appendix 6.7).

Following are detailed steps:

Step1: select 865 Z's with smallest marginal p-values

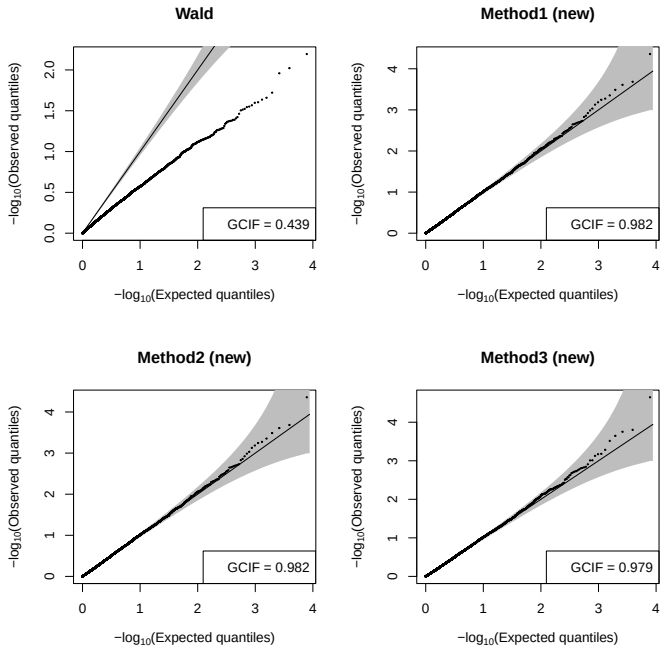
We start by only picking the SNPs with marginal p-values (from Wald test) in the smallest 0.1% range, that's to say, 865 SNPs with smallest marginal p-values, and use them as Z, the SNPs we want to test for interaction.

Step2: perform (Appendix: Fast approximate Wald test) to selected Z and all possible G_j's in the genome

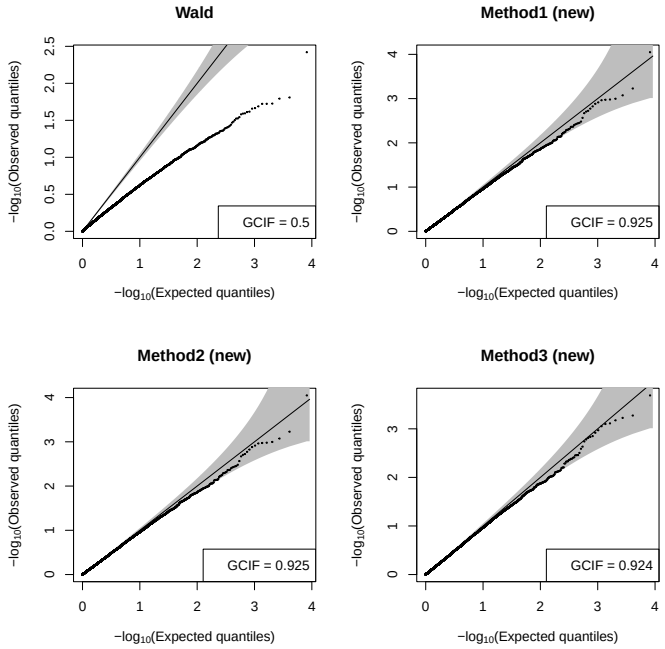
Even with the number of tests reduced by a factor of more than 500, we still need a fast computation strategy because we are performing interaction tests based on an LMM. We take a two-stage approach, where we first apply a fast, approximate Wald test. Then we only perform more time-consuming and accurate calculations for p-values that are small based on the fast, approximate Wald test, and we content ourselves with the coarser approximation for the p-values that are large. The key idea of the fast approximate Wald test is to regress out all variables aside from the interaction term step by step using matrix operations, so that we can avoid looping over the SNPs. We adopted this method to linear mixed model and got approximate p-values for the interaction GWAS.

Figure 6.2: **QQ-plots of −log10 p-values for interaction tests**

(a)



(b)



(a). Interaction tests between Z and 8183 original SNPs from the *A. thaliana* dataset. (b). Interaction tests between Z and 8183 independently simulated Bernoulli SNPs

Step3: Perform more accurate p-value calculation only for those pairs with fast approximate Wald p-value $< 10^{-4}$ and minimum cell count (MCC) ≥ 5

We use above fast approximate p-values for further filtering:

Both the p-value for interaction in a LMM and the TINGA method will be applied only to those pairs with fast, approximate Wald p-value $< 10^{-4}$. Furthermore, for some pairs, interaction was not tested at all because informativeness constraints were not met (we required MCC ≥ 5) or our constraint on correlation was not met (we required $r^2 < .01$).

After these filtering steps (based on MCC, r^2 and fast approximate Wald p-value), there are 57,149 pairs of SNPs remaining, with 728 of the originally chosen SNPs having at least one pair, and these 57,129 are the pairs for which we calculate the interaction t-statistic and TINGA statistic.

For the 57,129 pairs of SNPs, we first apply Gaussian Method 4 (fit homoscedastic model of Y under alternative (interaction)) with the following strategy:

Let M_1, M_2 be the minor allele frequencies of the 2 SNPs in the pair.

1. if $\min(M_1, M_2) \geq 0.15$, or both $M_1, M_2 \in [0.12, 0.15)$, condition on the SNP with worse diagnostic ratio (see Appendix 6.7)

2. if $\max(M_1, M_2) \geq 0.15$ and $\min(M_1, M_2) \in [0.05, 0.15)$, condition on the SNP with larger MAF

3. otherwise, do not apply TINGA

Then we pick the top 200 significant pairs from Gaussian Method 4 and apply Gaussian Method 3 to them, where we fit the heteroscedastic model by

$$Y \sim N(\alpha + \beta G_j + \gamma Z + \delta(G_j \circ Z),\ \sigma_h^2 \text{diag}(Z) + \sigma_g^2 K + \sigma_e^2 I). \tag{6.1}$$

Step 4: look for interesting pairs We look for interesting pairs for which

1. regular Wald test gives relatively significant result while TINGA gives much larger p-values: could be due to correction of "feast or famine" effect or

2. TINGA gives more significant result than Wald test

6.5 Findings

Among the top 10 significant pairs from Wald test, several of them involve a common SNP Chr1_27657389. It has large diagnostic ratio: 2.03. For the interaction GWAS where SNP Chr1_27657389 is fixed, the GCIF is 1.67. That is to say, with Wald test, there is a systematic deflation of p-values in the interaction GWAS. Table 6.1 compares the results from Wald test and TINGA. The last column shows the heteroscedasticity p-values by testing for heteroscedasticity in Y due to Z. i.e., testing $H_0 : \sigma_2^2 = 0$ in

$$Y = \alpha + Z + G_j + (Z \circ G_j) + \epsilon, \epsilon \sim N(0, \sigma_0^2 I + \sigma_1^2 K + \sigma_2^2 \text{diag}(Z))$$

As we can see, all 3 pairs have quite large heteroscedasticity p-values, indicating that the difference between Wald and TINGA results are most likely due to correction for "feast or famine" effect, instead of correction for heteroscedasticity.

SNP Z	SNP G_j	Wald p-value	TINGA	Hetero. p-value
Chr1_27657389	Chr2_5319468	2.2×10^{-9}	2.9×10^{-5}	0.49
Chr1_27657389	Chr1_5422500	1.5×10^{-8}	4.9×10^{-5}	0.35
Chr1_27657389	Chr2_12389533	3.3×10^{-8}	5.3×10^{-4}	0.14

Table 6.1: Comparison between Wald test and TINGA

SNP Z has diagnostic ratio 2.03 and GCIF 1.67. "Hetero. p-value": p-value for test heteroscedasticity in Y due to Z

Table 6.2 shows some pairs that TINGA gives more significant results. As we can see, for the first 2 pairs, the GCIFs and diagnostic ratios are > 1. It means that after correcting for

the systematic deflation of interaction p-values, TINGA still gives more significant result. For the other 2 pairs, the GCIFs and diagnostic ratios are < 1. After correcting for the systematic inflation of interaction p-values, TINGA gives more significant result, which is as expected.

SNP Z	GCIF	R	SNP G_j	Wald	TINGA
Chr2_9312347	1.23	1.38	Chr1_30240701	2.5×10^{-7}	2.7×10^{-9}
Chr5_12971212	1.30	1.30	Chr3_22936958	6.8×10^{-6}	4.0×10^{-8}
Chr1_24265614	0.88	0.87	Chr4_2614033	2.1×10^{-5}	2.0×10^{-7}
Chr3_21971020	0.93	0.91	Chr1_22559639	3.9×10^{-5}	1.4×10^{-7}

Table 6.2: Pairs for which TINGA gives more significant results

GCIF: genomic control inflation factor for interaction GWAS with SNP Z fixed
R: diagnostic ratio for SNP Z

6.6 Appendix: Fast approximate Wald test

Fast approximate t-test

The basic idea is to regress out everything else except y and the interaction term. Then the t-statistics for m x's can be computed at once via matrix multiplication.

Suppose we have phenotype y, SNPs z, $X = (x_1, x_2, ..., x_m)$ and covariates u. Let z_c, X_c be the centered genotypes. Let $W = X_c \circ z_c = (x_{1c} \circ z_c, ..., x_{mc} \circ z_c)$ be the matrix of interactions. We want the t-statistics for each of the epistasis by fitting m linear models

$$y \sim N(u\alpha_i + \gamma_i z_c + \beta_i x_{ic} + \delta_i (x_{ic} \circ z_c), \sigma_i^2 I)$$

Step 1: regress u, z_c out of y, $X_c, Xc \circ z_c$ Let P be the matrix that projects to the subspace spanned by (u, z_c). We let

$$y_r = y - Py$$

$$X_r = X_c - PXc$$

$$W_r = X_c \circ z_c - P(X_c \circ z_c)$$

Step 2: regress each column of X_r out of y_r and each column of W_r Since when the subspace only has 1 dimension, the projection matrix can be directly written as $\frac{1}{\|x_r\|^2} x_r x_r^T$ and the resulting variables can be computed by matrix operations in R. Let the results be y_{rx} and W_{rx}

Step 3: get the interaction t-statistics by regress each column of W_{rx} on corresponding column of y_{rx} Again, since each projection subspace has dimension 1, the result can be got by matrix operations in R

This fast method gets all m t-statistics without running a loop of m iterations.

Fast approximate Wald test

When fitting a linear mixed model, we can modify above method to get approximated test statistics. Suppose we want to fit the model

$$y \sim N(\alpha_i + \gamma_i z + \beta_i x_i + \delta_i (x_i \circ z), \Omega)$$

where $\Omega = \sigma_g^2 K + \sigma_e^2 I$. We could approximate it by a linear model by pre-multiply everything by $\Omega^{-1/2}$:

$$\Omega^{-1/2} y \sim N(\alpha_i (\Omega^{-1/2} \mathbf{1}) + \gamma_i (\Omega^{-1/2} z) + \beta_i (\Omega^{-1/2} x_i) + \delta_i \Omega^{-1/2} (x_i \circ z), I)$$

We could let $u = \Omega^{-1/2} \mathbf{1}$ be the new covariate and apply the fast t-statistics method to the new variables.

6.7 Appendix: A diagnostic for "Feast or Famine" effect

As mentioned in Chapter 2, for a given (Y, Z), the true null conditional distribution $F_{Y,Z}$ depends on value of (Y, Z).

- For some value of Y, Z, we consistently have a "feast" problem;
- For some value of Y, Z, we consistently have a "famine" problem

Therefore, we could find some diagnostic as a function of Y, Z that predicts the FoF problem.

Recall that the regular t-statistic for testing interaction is

$$T_j = \frac{\sqrt{n-k-3}\,(G_j \circ Z)^T P_M Y}{\sqrt{(G_j \circ Z)^T P_M (G_j \circ Z)\ \cdot Y^T P_M Y - ((G_j \circ Z)^T P_M Y)^2}} := \sqrt{n-k-3}\frac{N_j}{D_j}, \tag{6.2}$$

For a given (Y, Z), the true variance for the numerator N_j should be $\mathrm{Var}(N_j|Y, Z)$, but the regular t-test is using

$$D_j^2 = \big(G_j \circ Z\big)^T P_M \big(G_j \circ Z\big)\ \cdot Y^T P_M Y - (\big(G_j \circ Z\big)^T P_M Y)^2$$

Therefore, the ratio $\frac{\mathrm{Var}(N_j|Y,Z)}{D_j^2}$ represents how the true variance is different from the variance used in t-test

For D_j^2, take asymptotic approximation in $G_j|Y, Z$, we get a function $D_{Y,Z}^2$ in Y, Z that represents the asymptotic value of the (square of) denominator used in t-test given (Y, Z)

We get the diagnostic ratio by

$$R = \frac{\mathrm{Var}(N_j|Y, Z)}{D_{Y,Z}^2} \tag{6.3}$$

In the simplest case where there is no covariates other than intercept, and the noise is $N(0, \sigma^2 I)$,

$$R = \frac{nS_{zzrr}}{S_{zz}S_{rr}}, \tag{6.4}$$

where r is the residual of Y after regressing out $(1, Z)$ and $S_{ab} = \sum_i (a_i - \bar{a})(b_i - \bar{b}),\ S_{abcd} = \sum_i (a_i - \bar{a})(b_i - \bar{b})(c_i - \bar{c})(d_i - \bar{d})$

Extending it to a GRM case,

$$R = \frac{n \sum_{i,j} (\widehat{\Sigma}^{-1} r)_i (\widehat{\Sigma}^{-1} r)_j K_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{(r^T \widehat{\Sigma}^{-1} r) \sum_{i,j} (Z_i - \bar{Z})(Z_j - \bar{Z}) P_{ij} K_{ij}}, \tag{6.5}$$

where

$$\begin{aligned} \widehat{\Sigma} &= \hat{\sigma}_e^2 I + \hat{\sigma}_g^2 K \\ P &= \widehat{\Sigma}^{-1} - \widehat{\Sigma}^{-1} \mathbf{1} (\mathbf{1} \widehat{\Sigma}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \widehat{\Sigma}^{-1} \\ r &= Y - U(U^T \widehat{\Sigma}^{-1} U)^{-1} U \widehat{\Sigma}^{-1} Y,\ U = (1, Z) \end{aligned} \tag{6.6}$$

Equation 6.4, 6.5 are derived by McPeek, M. S. (McPeek, M. S., personal communication, March 25, 2024).

CHAPTER 7

DISCUSSION

Identifying interaction, either $G \times G$ or $G \times E$, can give insight into both genetic effects on a complex trait and underlying biological mechanisms, and it can also help to clarify the role of environment in the case of $G \times E$ testing. For testing interaction in a GWAS context, we have identified and described the "feast or famine" effect, in which different interaction GWASs have fundamentally different null distributions. We show that the "feast or famine" effect applies for different types of variables, including normal, binomial or binary, and for standard testing methods such as the t-test, F-test or likelihood ratio test for interaction. We show that it affects only interaction GWAS, not ordinary association GWAS. If we consider GWASs in which there is no interaction under the null hypothesis (so heteroscedasticity is not present), then on average over different GWASs standard methods have correct type 1 error overall, but false positives are overly concentrated in certain GWASs ("feast" GWASs) and false negatives are overly concentrated in certain other GWASs ("famine" GWASs). The "feast or famine effect" can lead to excess type 1 error, reduced power, inconsistent results across studies, and failure to replicate true signal. Furthermore, we show that whether a given GWAS will be a "feast or famine" GWAS is a reproducible property, and that it can be corrected for.

We develop the TINGA method which corrects the t-statistic for interaction by choosing different conditioning variables that are more appropriate for a GWAS than the standard choice. TINGA also allows for covariates and population structure through a LMM, and it accounts for heteroscedasticity. In simulations we show that TINGA can greatly reduce the "feast or famine" effect while preserving the overall type 1 error, which we show can result in higher power.

We apply TINGA to a GWAS for flowering time in *A. thaliana*. Using TINGA we detect 5 significant interactions after Bonferroni correction, where all the detected interactions involve

loci identified in previous studies as associated with flowering time. This demonstrates the potential of the TINGA method for detecting interaction in a GWAS.

For epistasis detection in a GWAS, there is a computational challenge in testing epistasis for all possible pairs of variants. When the model for Y is a LMM, as in our data analysis, this computational challenge is made much greater, even for the usual LMM-based t-test for interaction without any correction. We have developed a fast approximate version of the LMM-based t-test for interaction, and we use it as part of an adaptive approach to genome-wide testing, where more accurate but time-consuming methods are applied only if the approximate p-value is sufficiently small. In other words, our strategy is to spend more computational time on small p-values and to be content with coarse approximations to large p-values. In future work, there could be further scope for making faster algorithms for all aspects of interaction testing with a LMM in a GWAS context.

In epistasis detection, the situation of fixing one SNP and test its interaction with other SNPs in a genome is related to another concept of marginal epistasis [26], which test the null hypothesis that the fixed SNP has no interaction with any other SNP in the genome. The "feast or famine" effect could be expected to have a huge impact on the testing of marginal epistasis, making it all but impossible to reliably perform valid tests of marginal epistasis without adjusting for the effect in some way. This could be an avenue of possible future work. It is promising to apply the idea of conditioning on (Z, Y) to the existing methods for marginal epistasis and improve their performance.

Since detecting interactions in a GWAS setting might involve fixing Y and searching through all possible SNP pairs, it makes sense to consider conditioning on Y only, which could be another direction of future work.

CHAPTER 8

SUPPLEMENTAL INFORMATION

8.1 S1 R script to calculate p-values for the two-sided equal local levels test for i.i.d. uniformity

A two-sided equal local levels (ELL) test for i.i.d. uniformity a set of variables $X_1, \ldots, X_n$ is described in [1], who created the R package qqconf, which is available on CRAN. The primary purpose of qqconf is to generate appropriate simultaneous testing bands for a QQ-plot, but in addition, the functions available in qqconf can be used to generate p-values for the two-sided ELL test for i.i.d. uniformity.

In the R code below, suppose $x \in (0,1)^n$. The code obtains a p-value for the deviation of x from i.i.d. uniform(0,1). The test is a QQ-plot based ELL test. It answers the question: what is the largest level α for the acceptance region for the qq-plot that would result in non-rejection of x, where the acceptance region is based on 2-sided ELL.

```
library(qqconf)

qqpvu <- function(x){
n = length(x)
tmp1 = sort(x)
tmp2 = pbeta(tmp1,c(1:n),c(n:1))
tmp3 = min(min(tmp2),1-max(tmp2))*2
lb = qbeta(tmp3/2,c(1:n),c(n:1))
ub = qbeta(1-tmp3/2,c(1:n),c(n:1))
get_level_from_bounds_two_sided(lb,ub)
}
```

REFERENCES

1. Weine E, McPeek MS, Abney M. Application of equal local levels to improve Q-Q plot testing bands with R package qqconf. J Stat Softw. 2023;106(10):1–31. doi:10.18637/jss.v106.i10.

2. Glass D, Viñuela A, Davies MN, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biol. 2013;14(7):R75. doi:10.1186/gb-2013-14-7-r75.

3. Sleiman BM, Roy S, AW G, et al. Sex- and age-dependent genetics of longevity in a heterogeneous mouse population. Science. 2022;377(6614):eabo3191. doi:doi:10.1126/science.abo3191.

4. Myers R, Scott N, Gauderman W, et al. Genome-wide interaction studies reveal sex-specific asthma risk alleles. Hum Mol Genet. 2014;23(19):5251–5259. doi:doi:10.1093/hmg/ddu222.

5. Mitra I, Tsang K, Ladd-Acosta C, Croen L, Aldinger K, Hendren R, et al. Pleiotropic Mechanisms Indicated for Sex Differences in Autism. PLoS Genet. 2016;12(11):e1006425. doi:https://doi.org/10.1371/journal.pgen.1006425.

6. Small K, Todorčević M, Civelek M, et al. Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. Nat Genet. 2018;50(4):572–580. doi:doi:10.1038/s41588-018-0088-x.

7. Leite J, Soler J, Horimoto A, Alvim R, Pereira A. Heritability and Sex-Specific Genetic Effects of Self-Reported Physical Activity in a Brazilian Highly Admixed Population. Hum Hered. 2019;84(3):151–158. doi:doi:10.1159/000506007.

8. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, et al. The impact of sex on gene expression across human tissues. Science. 2020;369(6509):eaba3066. doi:doi:10.1126/science.aba3066.

9. Laville V, Majarian T, Sung Y, et al. Gene-lifestyle interactions in the genomics of human complex traits. Eur J Hum Genet. 2022;30(6):730–739. doi:doi:10.1038/s41431-022-01045-6.

10. Yazar S, Alquicira-Hernandez J, Wing K, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science. 2022;376(6589):eabf3041. doi:doi:10.1126/science.abf3041.

11. Carbone M, Arron ST, Beutler B, et al. Tumour predisposition and cancer syndromes as models to study gene–environment interactions. Nat Rev Cancer. 2020;20(9):533–549. doi:https://doi.org/10.1038/s41568-020-0265-y.

12. Evans D, Spencer C, Pointon J, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet. 2011;43(8):761–767. doi:doi:10.1038/ng.873.

13. Moutsianas L, Jostins L, Beecham A, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet. 2015;47(10):1107–1113. doi:doi:10.1038/ng.3395.

14. Wang M, Roux F, Bartoli C, Huard-Chauveau C, Meyer C, Lee H, et al. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. Proc Natl Acad Sci U S A. 2018;115(24):E5440–E5449. doi:doi:10.1073/pnas.1710980115.

15. Clark M, Chazara O, Sobel E, et al. Human Birth Weight and Reproductive Immunology: Testing for Interactions between Maternal and Offspring KIR and HLA-C Genes. Hum Hered. 2016;81(4):181–193. doi:doi:10.1159/000456033.

16. Evans L, Arehart C, Grotzinger A, Mize T, Brasher M, Stitzel J, et al. Transcriptome-wide gene-gene interaction associations elucidate pathways and functional enrichment of complex traits. PLoS Genet. 2023;19(5):e1010693. doi:https://doi.org/10.1371/journal.pgen.1010693.

17. Vasseur F, Exposito-Alonso M, Ayala-Garay OJ, Wang G, Enquist BJ, Vile D, et al. Adaptive diversification of growth allometry in the plant Arabidopsis thaliana. Proceedings of the National Academy of Sciences. 2018;115(13):3416–3421. doi:10.1073/pnas.1709141115.

18. Visscher P, Brown M, McCarthy M, J Y. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24. doi:doi:10.1016/j.ajhg.2011.11.029.

19. Eichler E, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–450. doi:doi:10.1038/nrg2809.

20. Robinson M, English G, Moser G, et al. Genotype-covariate interaction effects and the heritability of adult body mass index. Nat Genet. 2017;49(8):1174–1181. doi:doi:10.1038/ng.3912.

21. Mackay T. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014;15(1):22–33. doi:doi:10.1038/nrg3627.

22. Roth C, Murray D, Scott A, Fu C, Averette A, Sun S, et al. Pleiotropy and epistasis within and between signaling pathways defines the genetic architecture of fungal virulence. PLoS Genet. 2021;17(1):e1009313. doi:https://doi.org/10.1371/journal.pgen.1009313.

23. Ritz B, Chatterjee N, Garcia-Closas M, et al. Lessons Learned From Past Gene-Environment Interaction Successes. Am J Epidemiol. 2017;186(7):778–786. doi:doi:10.1093/aje/kwx230.

24. Lopez-Cruz M, Aguate F, Washburn J, et al. Leveraging data from the Genomes-to-Fields Initiative to investigate genotype-by-environment interactions in maize in North America. Nat Commun. 2023;14(1):6904. doi:doi:10.1038/s41467-023-42687-4.

25. Alipour H, Abdi H, Rahimi Y, Bihamta M. Dissection of the genetic basis of genotype-by-environment interactions for grain yield and main agronomic traits in Iranian bread wheat landraces and cultivars. Sci Rep. 2021;11(1):17742. doi:doi:10.1038/s41598-021-96576-1.

26. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 2017;13(7):e1006869. doi:https://doi.org/10.1371/journal.pgen.1006869.

27. Dahl A, Nguyen K, Cai N, Gandal M, Flint J, N Z. A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. Am J Hum Genet. 2020;106(1):71–91. doi:doi:10.1016/j.ajhg.2019.11.015.

28. Tang D, Freudenberg J, Dahl A. actorizing polygenic epistasis improves prediction and uncovers biological pathways in complex traits. Am J Hum Genet. 2023;110(11):1875–1887. doi:doi:10.1016/j.ajhg.2023.10.002.

29. Greene CS, Penrod NM, Kiralis J, et al. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Mining. 2009;2(1):5. doi:https://doi.org/10.1186/1756-0381-2-5.

30. Emily M, Mailund T, Hein J, Schauser L, Schierup M. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet. 2009;17(10):1231–1240. doi:doi:10.1038/ejhg.2009.15.

31. Lippert C, Listgarten J, Davidson R, et al. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. Sci Rep. 2013;3:1099. doi:doi:10.1038/srep01099.

32. Moore R, Casale FP, Jan Bonder M, et al. A linear mixed-model approach to study multivariate gene–environment interactions. Nat Genet. 2019;51(1):180–186. doi:doi:10.1038/s41588-018-0271-0.

33. Dudbridge F, Koeleman B. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet. 2004;75(3):424–435. doi:doi:10.1086/423738.

34. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb). 2005;95(3):221–227. doi:doi:10.1038/sj.hdy.6800717.

35. Evans D, Marchini J, Morris A, Cardon L. Two-Stage Two-Locus Models in Genome-Wide Association. PLoS Genet. 2006;2(9):e157. doi:https://doi.org/10.1371/journal.pgen.0020157.

36. Marderstein AR, Davenport ER, Kulm S, Van Hout CV, Elemento O, Clark AG. Leveraging phenotypic variability to identify genetic interactions in human phenotypes. Am J Hum Genet. 2021;108(1):49–67. doi:10.1016/j.ajhg.2020.11.016.

37. Westerman KE, Majarian TD, Giulianini F, et al. Variance-quantitative trait loci enable systematic discovery of gene-environment interactions for cardiometabolic serum biomarkers. Nat Commun. 2022;13(1):3993. doi:https://doi.org/10.1038/s41467-022-31625-5.

38. Wei W, Hemani G, Haley C. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15(11):722–733. doi:doi:10.1038/nrg3747.

39. Ahmad S, Varga T, Franks P. Gene × environment interactions in obesity: the state of the evidence. Hum Hered. 2013;75(2-4):106–115. doi:doi:10.1159/000351070.

40. McAllister K, Mechanic L, Amos C, et al. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. Am J Epidemiol. 2017;186(7):753–761. doi:doi:10.1093/aje/kwx227.

41. Wood A, Tuke M, Nallsm M, et al. Another explanation for apparent epistasis. Nature. 2014;514(7520):E3–E5. doi:doi:10.1038/nature13691.

42. Hemani G, Powell J, Wang H, et al. Phantom epistasis between unlinked loci. Nature. 2021;596(7871):E1–E3. doi:doi:10.1038/s41586-021-03765-z.

43. Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-Plots and Genomic Control in Studies of Gene-Environment Interaction. PLoS ONE. 2011;6(5):e19416. doi:https://doi.org/10.1371/journal.pone.0019416.

44. Rao T, Province M. A Framework for Interpreting Type I Error Rates from a Product-Term Model of Interaction Applied to Quantitative Traits. Genet Epidemiol. 2016;40(2):144–153. doi:doi:10.1002/gepi.21944.

45. Zhang T, Sun L. Beyond the traditional simulation design for evaluating type 1 error control: From the "theoretical" null to "empirical" null. Genet Epidemiol. 2019;43(2):166–179. doi:doi:10.1002/gepi.22172.

46. Pare G, Cook NR, Ridker PM, Chasman DI. On the Use of Variance per Genotype as a Tool to Identify Quantitative Trait Interaction Effects: A Report from the Women's Genome Health Study. PLoS Genetics. 2010;6(6):e1000981. doi:10.1371/journal.pgen.1000981.

47. Balding D, Nichols R. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995;96(1-2):3–12. doi:doi:10.1007/BF01441146.

48. Consortium TG. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. J Cell. 2016;doi:https://doi.org/10.1016/j.cell.2016.05.063.

49. Zan Y, Carlborg O. A Polygenic Genetic Architecture of Flowering Time in the Worldwide *Arabidopsis thaliana* Population. Molecular Biology and Evolution. 2019;36(1):141–154. doi:https://doi.org/10.1093/molbev/msy203.