mvgS Summary

Irene Zhang

April 25, 2020

1 Introduction

Conceptually, complex traits are thought to result from genetic variation at multiple genes or their regulators, and how this genetic profile interacts with behavioral and environmental factors. In case-control or cohort studies, testing for gene-environment interactions is often performed to identify genetic variants whose impact seems to be modified by different environmental exposures [**thomas2010gene**], although the power of such tests is often low [**cordell2009detecting**]. Traditionally, the phenotypic effect of a gene-environment interaction is estimated by fitting a regression model for the outcome of interest(Y) including both the main effects of a genetic variant(G) and an environmental factor (E), as well as the product between G and E. For a known environmental exposure, genome wide association studies (GWAS) for interaction effects usually scan through millions of single nucleotide polymorphisms (SNPs) using this model; this approach requires adjustment for multiple comparison [**he2017set**]. However, results from such single SNP analyses have rarely been shown to be reproducible. Statistical power is known to be low for tests of interaction. Also, the environmental exposure is often measured with error or misclassified, which can reduce power[**thomas2010gene**, **tchetgen2011robustness**, **cornelis2012gene**, **boonstra2016tests**]. To alleviate penalties from multiple testing, filtering strategies have been adopted to reduce the number of candidate SNPs considered for interaction effects [**kooperberg2008increasing**, **thomas2010gene**] by setting a threshold p-value for significance of the main effects and only considering significant subset of SNPs for interaction analysis.

These $G \times E$ approaches with pre-filtering do not explicitly consider or use any dependence among genetic variants and can lose power from not capturing those variants whose effects are only manifest through interaction. In an attempt to improve power, it has become popular to aggregate or jointly modeling the effects of all SNPs in

a pre-defined set, such as a gene, pathway or specific genomic region. Some proposed methods for testing $G \times E$ interaction effects are extension of joint tests of multiple-SNP main effects, for example, by aggregating SNPs in the same gene and testing the interaction between the environment and the aggregated measures [**jiao2013sberia**, **wang2015powerful**, **basu2011comparison**]. These collapsing approaches assume, in general, that a large proportion of the genetic variants is causal, and that the causal effects have the same direction. The power of association analysis decreases if these assumptions do not hold. Another popular set of analysis methods are variance component derived tests, developed from the concept of kernel models. Variance component tests for interaction such as the Gene-Environment Set Association test (GESAT)[**lin2013test**], and the Interaction Sequence Kernel Association Test (iSKAT)[**lin2016test**] extend the sequence kernel association test (SKAT)[**wu2011rare**] developed for estimating the main effects of SNPs.

All these direct tests assume that the environmental factor is observed and measured without error. However, this is not always the case. Robust interaction tests are needed to cope with inaccurately-measured exposure as well as the case where the environment factor is not observed at all. Pare et al. [**pare2010use**] noticed that when a SNP is involved in an interaction, with either another SNP or environmental exposure, quantitative trait variance can differ across the SNP genotypes even then the interacting partner is not observed. They therefore proposed a SNP prioritization method based on the variances, using Levene's test to assess the variance heterogeneity. SNPs with insignificant heterogeneity signals are filtered out of further interaction analysis to reduce the number of tests performed. For a single SNP analysis, Levene's test can be directly applied to the categorical genotype data, i.e. comparing distributions of the outcome across genotypes AA, AB, or BB. However, implementing of this simple concept when jointly testing multiple SNPs is not feasible, since simply enumerating haplotype pairs would lead to an exponential increase in the number of categories, and Levene's test would have little power.

Soave and Sun (2017) proposed a two-stage regression framework that generalized the classic Levene's test to allow for non-categorical genotypes,such as those found, for example, when missing genotypes are imputed. Their method is called the generalized Levene's Scale test (gS). The gS test first fits a main effect model and obtains the residuals, then tests if the variance of the residuals differ between genotype groups; this latter stage is achieved by modeling the relationship between the absolute residuals and the genotype. To be precise, a gS test that uses an ordinary least squares in the first stage and an F statistic in the second stage is equivalent to a classic Levene's test;

and a gS implementation that uses least absolute deviances in the first stage (and an F test in the second stage) corresponds to the Brown-Forsythe test. Although there are some other popular variance tests, such as the likelihood ratio test(LRT) ([**cao2014versatile**]), they have been shown to be less robust than Levene's test to what??. Here, here we generalize the gS test to a multi-SNP version, to create a region based test to detect changes in variances across multiple genotypes. As for the gS test, our multi-SNP generalized Levene's scale test (mvgS) does not require a measured environmental factor and hence is robust to missing or poorly measured environmental factors. Also, mvgS can be used for imputed (non-categorical) genotype data.

In section 2, we first review the two direct interaction tests mentioned above (GESAT and iSKAT) together with the original SKAT that underpins both. Then we introduce the concept of a multi-SNP generalized Levene's scale test (mvgS), together with the relevant model assumptions, choice of residuals, and the choice of second stage test statistics. Simulations are performed in section 3 to compare the performance of variants of mvgS under different assumptions. First we estimate type 1 error of all versions of the mvgS tests under a range of error distributions, using a permutation-based empirical simulation design. Then we compare the power of mvgS test versions in different settings and select the best combinations of residual type and second stage test statistic. Finally, we compare the power of the best mvgS versus GESAT and iSKAT, where the environmental factor is binary or continuous, and with or without measurement error.

Results of this type 1 error simulation study are presented in section 4, and based on the results, we decided to use (xxx residual and xxx as second stage test) in mvgS. We then applied the mvgS test to UK10K data (need more details) and results are presented in section 5. Overall, our proposed multi-variable generalized Levene's scale test can be used as a filter of regions demonstrating possible interaction effects when the interacting factor is not observed; The test has been shown to be robust to measurement error when compared to existing direct interaction tests; and exhibits good power when using the xxx second stage approach.

2 Methods

2.1 Review of Direct Interaction Tests

In most studies of gene-environment interaction effects, the environmental factor E is assumed to be available and accurately collected. The products of E and a set of SNPs $G_1, G_2 \cdots, G_J$ are included in a linear or generalized linear model with interaction terms:

$$g(\mathbb{E}(Y_i)) = \beta_0 + \sum_{j=1}^{J} \beta_{1j} G_{ij} + \beta_2 E_i + \sum_{j=1}^{J} \beta_{3j} G_{ij} E_i \tag{1}$$

and statistical tests are performed on the null hypothesis $\mathcal{H}_0 : \beta_{3j} = 0$ for all $j = 1, \cdots, J$, searching for significant signals. We call these methods, i.e. those that are applicable when E is available, *direct interaction tests*. As addressed in the introduction, when the directions of the interaction effects and/or when magnitudes of the interaction effects differ across the variants, variance component type tests, in contrast to burden type tests, have stable type 1 error control and better power. Before introducing our proposed test, we will first review two variance component-based direct interaction tests, GESAT and iSKAT. Both these tests are extensions of SKAT, which was originally developed by Wu et al. [**wu2011rare**] to test the main effects of a set of SNPs. When testing for interaction effects, GESAT assumes that the interaction SNP effects $\{\beta_{3j}\}$ are independent of each other $j = 1, \cdots, J$, and follow an arbitrary but common distribution with mean 0 and variance τ^2. Hence, the GESAT null hypothesis can be written as $\mathcal{H}_0 : \tau^2 = 0$. Following the idea implemented in SKAT, the test statistics for the variance component τ^2 is:

$$Q_s = (\mathbf{Y} - \widehat{\mu})^{\mathbf{T}} \mathbf{S}\mathbf{S}^{\mathbf{T}} (\mathbf{Y} - \widehat{\mu}) \tag{2}$$

where $S_i = (G_{i1}E_i, G_{i2}E_i, \cdots, G_{iJ}E_i)^T$ and the matrix $\mathbf{S} = (S_1, \cdots, S_n)^T$ contains the gene-environment interaction terms. The term $\widehat{\mu}$ is estimated from a the relevant null model without interaction terms via ridge regression; this penalized model is selected to protect against overfitting in case the number of SNPs J is large or the set of SNPs are in high linkage disequilibrium.

The GESAT test, based on variance-type arguments, has advantages over the collapsing type or burden tests when the set of SNPs contains many non-causal variants or when causal variants have different directions of association. However, if the target region has a high proportion of causal variants with the effects in the same direction, burden tests can be more powerful than GESAT. Since such knowledge is likely to be unknown prior to performing the test,

iSKAT was then proposed as an optimal combination of a burden type test and a variance component test. For a parameter $\rho \in [0, 1]$, which weights the contributions of these two types of tests, iSKAT defines a combined statistic:

$$Q_\rho = (\mathbf{Y} - \widehat{\mu})^{\mathbf{T}}\mathbf{S}\mathbf{W}\big((1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^{\mathbf{T}}\big)\mathbf{W}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}}(\mathbf{Y} - \widehat{\mu}) \tag{3}$$

where $\mathbf{W} = diag(w_1, w_2, \cdots, w_j)$ is a diagonal weight matrix. The weights w_i are usually chosen as a function of the minor allele frequencies of the SNPs. As the best value of ρ is unknown, iSKAT search for the minimum p-value of Q_ρ over the range $(0, 1)$ and the optimal statistic is defined as:

$$Q_{iSKAT} = \min_{0 \le \rho \le 1} p_\rho \tag{4}$$

This unified test can be considered as equivalent to a generalized GESAT test that no longer assumes independent interaction SNP effects β_{3j}s. The iSKAT test assumes that β_{3j} follows an arbitrary distribution with mean 0, variance τ and pairwise correlation ρ. Theoretically, iSKAT with identical weights w_j, and $\rho = 0$ is equivalent to GESAT.

2.2 Proposed Multi-variable Generalized Levene's Scale Test(mvgS)

Assuming that the true data generating model involves J SNPs with either main effects, interaction effects, or both, we can write an equation for the outomce Y_i as:

$$Y_i = \beta_0 + \sum_{j=1}^{J}\beta_{1j}X_{ij,1} + \sum_{j=1}^{J}\beta_{2j}X_{ij,2} + \beta_3 E_i + \sum_{j=1}^{J}\beta_{4j}X_{ij,1} \times E_i + \sum_{j=1}^{J}\beta_{5j}X_{ij,2} \times E_i + \varepsilon_i \tag{5}$$

where $X_{ij,k} = \mathbb{1}(G_{ij} = k)$ for $k = 0, 1, 2$. The model above can be naturally simplified to the assumption of additive genetic effects, where I HAVE ADDED THE NUMBER 2 HERE. CHECK ASSUMPTIONS/ACCURACY $2\beta_{1j} = \beta_{2j}$ and $2\beta_{3j} = \beta_{4j}$ for all $j = 1, \cdots, J$. We assume that the error term ε_i follows a normal distribution, and these errors are independent from the genotypes $\{G_{ij}\}_{j=1}^{J}$.. Gene environment independence is also assumed. Hence the conditional expectation and variance for this model are:

Do you want to briefly mention violations of the normal assumption for ε here since this leads to the empirical type 1 error considerations later?

$$\begin{aligned}\mathbb{E}\big[Y_i|\{G_{ij}\}\big] &= \beta_0 + \beta_3\mathbb{E}(E_i) + \sum_{j=1}^{J}\gamma_{1j}X_{ij,1} + \sum_{j=1}^{J}\gamma_{2}X_{ij,2} \\ Var\big[Y_i|\{G_{ij}\}\big] &= \left(\beta_3 + \sum_{j=1}^{J}\beta_{4j}X_{ij,1} + \sum_{j=1}^{J}\beta_{5j}X_{ij,2}\right)^2 \cdot \sigma_E^2 + \sigma^2\end{aligned} \tag{6}$$

where $\gamma_{1j} = \beta_{1j} + \beta_{4j}\mathbb{E}(E_i)$, $\gamma_{2j} = \beta_{2j} + \beta_{5j}\mathbb{E}(E_i)$.

As described in the introduction, the gS test contains two analysis stages, where the first stage calculates the residuals after removing the genotype main effects, and the second stage looks at the variance of these residuals. For implementation of the gS concept in a region of SNPs, we then propose to proceed as follows:

First stage: Fit a regression model containing only the main effects of SNPs:

$$Y_i = b_0 + \sum_{j=1}^{J} b_{1j}X_{ij,1} + \sum_{j=1}^{J} b_{2j}X_{ij,2} \tag{7}$$

using a least absolute deviance(LAD) estimation procedure, and obtain the residuals $e_i = Y_i - \widehat{Y}_i = Y_i - X_i^T\widehat{\beta}$. Under the alternative model (5), the residuals will follow $e_i \sim \mathcal{N}(0, \sigma_i^2)$ where $\sigma_i^2 = Var[Y_i|\{G_{ij}\}]$.

In stage 2 of the original gS test, dispersion associated with the absolute residuals was examined, since the absolute residuals are more robust to non-normality than squared residuals. We follow this recommendation again here; the distribution of $d_i = |e_i|$ is a folded normal with expectation associated with G_{ij}:

$$\mathbb{E}[d_i] \propto \sqrt{Var[e_i]} = \sqrt{\left(\beta_3 + \sum_{j=1}^{J}\beta_{4j}X_{ij,1} + \sum_{j=1}^{J}\beta_{5j}X_{ij,2}\right)^2 \cdot \sigma_E^2 + \sigma^2} \tag{8}$$

If we are willing to assume $\sigma^2 \ll \left(\beta_3 + \sum_{j=1}^{J}\beta_{4j}X_{ij,1} + \sum_{j=1}^{J}\beta_{5j}X_{ij,2}\right)^2 \cdot \sigma_E^2$, an approximate linear relationship is obtained :

$$\mathbb{E}[d_i] \propto \beta_3\sigma_E + \sum_{j=1}^{J}\beta_{4j}\sigma_E X_{ij,1} + \sum_{j=1}^{J}\beta_{5j}\sigma_E X_{ij,2} \tag{9}$$

However, when the assumption of small residual variance σ^2 is not realistic, we can then consider the squared residual $d_i = e_i^2$ with:

$$\begin{aligned}\mathbb{E}(e_i^2) = Var(e_i) =& \Big(\beta_3 + \sum_{j=1}^{J} \beta_{4j} X_{ij,1} + \sum_{j=1}^{J} \beta_{5j} X_{ij,2}\Big)^2 \cdot \sigma_E^2 + \sigma^2 \\ \text{(simplified to be)} =& \gamma_0 + \sum_{j=1}^{J} \gamma_{1j} X_{ij,1}^2 + \sum_{j=1}^{J} \gamma_{2j} X_{ij,2}^2 \\ &+ \sum_{j<k} \theta_{jk}^{1,1} X_{ij,1} X_{ik,1} + \theta^{2,2} X_{ij,2} X_{ik,2} + \sum_{j \neq k} \theta_{j,k}^{1,2} X_{ij,1} X_{ik,2}\end{aligned} \tag{10}$$

where $\gamma_{1j} = \beta_{4j}^2$ and $\gamma_{2j} = \beta_{5j}^2$. And $\theta_{j,k}^{1,1} = \beta_{4j} \cdot \beta_{4k}, \theta_{j,k}^{2,2} = \beta_{5j} \cdot \beta_{5k}, \theta_{j,k}^{1,2} = \beta_{4j} \cdot \beta_{5k}$ for $j, k = 1, \cdots, J$. If all genotypes are categorical, naturally $X_{ij,1} \cdot X_{ij,2} = \mathbb{1}(G_{ij} = 1) \cdot \mathbb{1}(G_{ij} = 2) = 0$ and $X_{ij,1}^2 = X_{ij,1}$, $X_{ij,2}^2 = X_{ij,2}$. Hence we can safely drop the terms $\sum_{j,k} \theta_{j,k}^{1,2} X_{ij,1} X_{ik,2}$ and thereby reduce the expectation to:

$$\mathbb{E}(e_i^2) \propto \gamma_0 + \sum_j \gamma_{1j} X_{ij,1} + \sum_j \gamma_{2j} X_{ij,2} + \sum_{j<k} \theta_{j,k}^{1,1} X_{ij,1} X_{ik,1} + \sum_{j<l} \theta_{j,k}^{2,2} X_{ij,2} X_{ik,2} \tag{11}$$

Second stage: We now propose two versions of the mvgS test corresponding to these two transformations of the residuals: absolute and squared. Let d_i be the transformed residuals. In the second stage, we propose to fit the regression model:

$$d_i = c_0 + \sum_{j=1}^{J} c_{1j} X_{ij,1} + \sum_{j=1}^{J} c_{2j} X_{ij,2} \tag{12}$$

and jointly test $\mathcal{H}_0 : c_{1j} = c_{2j} = 0$, for all $j = 1, \cdots, J$. For the absolute residuals, this null hypothesis is equivalent to assuming that no non-zero values of β_4 or β_5 exist for any SNP in the set : $\mathcal{H}_0 : \beta_{4j} = \beta_{5j} = 0 \; \forall j \in [J]$. For the squared residual, the null hypothesis is equivalent to $\mathcal{H}_0 : \beta_{4j}^2 = \beta_{5j}^2 = 0 \; \forall j \in [J]$, which is the same null hypothesis.

In the gS paper (Soave and Sun 2017), the same two options for the first stage regression–least absolute deviance (LAD) regression and least squares regression– were explored theoretically, and their equivalence to Levene's test or the Brown-Forsythe test was shown. We follow the recommendation of (Soave and Sun 2017) and select LAD regression for the first stage. With this choice and when $J = 1$, the mvgS test reduces to the original gS test. However, when $J > 1$, additional considerations and choices are required for the multivariate generalized scale test(mvgS):

(1) *Choice of residuals in the second stage: squared residuals or absolute values.* The approximate linearity of equation 9 for the absolute residuals rests on the assumption of a very small or negligible error variance σ^2. In the simulation section, we will study how violations of this assumption affect performance of mvgS with absolute

residual regarding its type 1 error and power. On the contrary, Levene's test with squared residuals is based on a negligible value of σ^2, but this choice is sensitive to the violation of the normality of the error distribution in equation (5). Loh[**loh1987some**] compared several modifications of Levene's test, among which, they showed that a second-order power transformation on the residuals may lead to unstable type 1 error when $\varepsilon_i \sim t_{(df)}$, i.e. with heavier tailed distributions. The same type 1 error problem is anticipated when using squared residuals in our proposed method. Overall, it is unknown whether there is a uniformly better choice for the transformations of the residuals, and therefore we will conduct comprehensive simulations in section 3 to examine the advantages and disadvantages associated with each choice.

(2) *Choice of test statistic in the second stage*: After completing stage 1 with a chosen form for residual transformations, we would like to choose a stage 2 test that can capture the increase in variability associated with the genotype at J SNPs. No matter which type of residual the test uses in stage 2, it requires a joint statistical test for $2J$ coefficients. Candidate test statistics include the ANOVA F-test, as adopted in the original gS test, sequence kernel association test (SKAT) and the optimal version of SKAT (SKAT-O).

(3) *Choice of genotype codings*: the original gS test used a genotype coding, so that the possible values at a SNP were represented by 3 values, AA, AB or BB. For the proposed multi-SNP test, we propose to investigate both genotype coding for each SNP (i.e. a maximum of 2 degrees of freedom times J SNPs) or an additive coding for the count of minor alleles at each of J SNPs. The latter choice is much more parsimonious, and may therefore be more powerful if the additive assumption is correct. If an additive coding is assumed, equations (7) and (12) can be simplified, and the stage 1 and stage 2 regressions become:

$$\begin{aligned} \text{stage 1: } Y_i &= b_0 + \sum_{j=1}^{J} b_{1j} G_{ij} \\ \text{stage 2: } d_i &= c_0 + \sum_{j=1}^{J} c_{1j} G_{ij} \end{aligned} \tag{13}$$

Hence we consider both coding schemes for the genotype and examine the power differences in section 3.

From the simulation results that compare the combinations of transformations of residuals in stage 1, the test statistics in stage 2, and the genotype coding schemes, we then finalize the proposed multivariate generalized Levene's scale test(mvgS) with by selecting the default settings as follows:

- In stage 1, we fit $Y_i = b0 + \sum_{j=1}^{J} b_{1j} G_{ij}$, and obtain the absolute residuals $d_i = |e_i| = |Y_i - \widehat{Y}_i|$.
- Then in stage 2, we fit $di = c0 + \sum_{j=1}^{J} c_{1j} G_{ij}$ and test $\mathcal{H}_0 : c_{1j} = 0$ for $j = 1, \cdots, J$ by xxx .

3 Simulation Study

3.1 Data Generating Model

[UK10K data description of the genotypes here.] Our simulation is based on genotyping data taken from the UK10K study. Then we simulate an environmental factor independently from the genetic data, $E_i \sim Bernoulli(0.3)$. A continuous environmental exposure with the same mean and variance is also considered in the additional power simulation settings.

Assuming an additive effect, the phenotpye is generated from:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_{1j} G_{ij} + \beta_2 E_i + \sum_{j=1}^{J} \beta_{3j} G_{ij} \times E_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

We use the same set of simulation coefficients as were chosen in the iSKAT paper, where $\beta_0 = 3.6$, $\beta_2 = 0.015$ and $\sigma^2 = 0.27$. The main effects of β_{1j} and the interaction effects β_{3j} for 11 SNPs, selected XXX are summarized in Table 2.

$j =$	1	2	3	4	5	6	7	8	9	10	11
β_{ij}	-0.030	-1.4	8.3	-4.1	2.2	0.005	-0.015	-0.0056	0.0069	-0.033	0.15
β_{3j}	-0.218	0	0	-0.476	0	0	-0.151	-0.845	0.0945	0	-0.133

Table 1: Main and interaction effects of 11 SNPs used in data generating model.

3.2 Empirical Design for the Type 1 Error Simulation

much more needed here. There is an extensive literature on how permutation tests should be performed when tests of interaction are of interest Permutation of Y versus both G and E simultaneously allows evaluation of the complete null hypothesis of no environmental or genetic effects, but in order to evaluate only the interaction, care must be taken to ensure valid type 1 error. (Anderson 2001 says that there is no exact permutation. Buzdoka recommends either

permutation of residuals after stage 1, or parametric bootstrap. However, since we need to take into account the correlation between genotypes across the genome, and therefore parametric bootstrap may not be appropriate). In GWAS studies of interaction effects, often a single environmental exposure E is considered at a time, and association tests are performed for phenotype Y versus a series of genetic regions across the genome using the same expoure. In this situation, permuting (Y, E) jointly does a good job of matching the GWAS analysis retaining correlation patterns between the test statistics (Budkova refs or others).

We evaluate performance of Type 1 error statistics by using a permutation based empirical simulation design. With the generated dataset $\{Y_i, E_i, G_{i1}, \cdots, G_{iJ}\}$, where Y_i is dependent on $\{G_{ij} \times E_i\}$ set, we first randomly permute (Y_i, E_i) pairs and obtain the new set $\{Y_i^{perm}, E_i^{perm}, G_{i1}, \cdots, G_{iJ}\}$ where $Y_i^{perm} \perp\!\!\!\perp G_{ij} \times E_i^{perm}$ for all $j = 1, \cdots, J$. Jointly permuting (Y_i, E_i) pair can maintain the association between Y_i and E_i, and retains the linkage pattern in $\{G_{ij}\}_{j=1}^J$, while breaking the association between Y_i and $\{G_{ij} \times E_i\}_{j=1}^J$. In each permutation, we evaluate mvgS with all sets of choices for the stage 1 and 2 choices, as well as the direct methods GESAT and iSKAT. Using 10^5 permutations, the empirical null distribution and type 1 error of each test statistic are obtained for the true model (14).

We added simulation scenarios to also study the robustness of each method to changes in minor allele frequencies, and to robustness to non-normal error distribution. Common variants are simulated from $G_{ij} \sim Bin(2, maf_j)$ with minor allele frequencies, maf_j, randomly chosen between 0.05 to 0.4; and rare variants with maf_j randomly chosen between 0.01 to 0.04. Error terms are simulated from t_5, χ_3^2 and a folded $\mathcal{N}(0, 1)$.

3.3 Power Evaluations

We continue simulations for power using the same assumed model (equation 14)) to generate a dataset $\{Y_i, E_i, G_{i1}, \cdots, G_{iJ}\}$ under the alternative hypothesis. For one generated dataset, we analyze the data using combinations of the two first stage residual transformations, the choices for second stage statistics and genotype codings and the obtain the final statistics. The p-values are then calculated from the estimated quantile of the test statistic with respect to its empirical null distribution. Then we repeat the whole process for 1000 times to obtain empirical power.

In addition to the situations described above, we also change settings consider changing settings to include more possible cases in the simulation study so to identify the best combination of residual type and stage II statistic and

genotype coding schemes within the proposed method.

(a) Power evaluation with β_{3j} simulated identically and independently from a density function

$$f(x) = 0.3\mathbb{1}(x = 0) + 0.7g(x)$$

where $g(x)$ is the normal density with mean μ and variance 1. μ ranges from 0.1 to 1.

(b) Power evaluation when $\varepsilon_i \sim \mathcal{N}(0, 2 * 0.27^2)$, to assess how sensitive mvgS (with absolute residual) is to the assumption of negligible σ^2.

(c) Power evaluation when $\beta_2 = 0.05$, to assess how sensitive mvgS is to the magnitude of unobserved environment effect.

Then to compare the proposed mvgS test with direct interaction test GESAT and iSKAT, we further consider two types of environment exposure variable: continuous and binary, matched with 5 different situations where we have no measurement error, and measurement errors at a specified error rate, symmetric or skewed.

(d) Power evaluation for continuous environment exposure $E_i \sim \mathcal{N}(0.3, 0.46)$

(i) no measurement error: $E_i^{obs} = E_i$

(ii) 10% measurement error: $E_i^{obs} = E_i + \delta_i, \delta_i \sim \mathcal{N}()$

(iii) 20% measurement error: $E_i^{obs} = E_i + \delta_i, \delta_i \sim \mathcal{N}()$

(iv) 10% measurement error, skewed: $E_i^{obs} = E_i + \delta_i, \delta_i \sim \chi^2()$

(v) 20% measurement error, skewed: $E_i^{obs} = E_i + \delta_i, \delta_i \sim \chi^2()$

(e) Power evaluation for binary $E_i \sim Bin(0.3)$

(i) no measurement error: $E_i^{obs} = E_i$

(ii) 10% measurement error: $E_i^{obs} = (1 - \alpha_i)E_i + \alpha_i(1 - E_i), \alpha_i \sim Ber(0.1)$

(iii) 20% measurement error: $E_i^{obs} = (1 - \alpha_i)E_i + \alpha_i(1 - E_i), \alpha_i \sim Ber(0.2)$

(iv) 10% measurement error, skewed: if $E_i = 0$, $E_i^{obs} \sim Ber(1/3)$, else $E_i^{obs} = E_i$

(v) 20% measurement error, skewed: if $E_i = 0$, $E_i^{obs} \sim Ber(2/3)$, else $E_i^{obs} = E_i$

4 Results

Table 2 shows the empirical type 1 error evaluation of mvgS test with 6 different combinations from 10^5 replicates. At a 5% nominal level, any empirical T1E within (0.0494, 0506) would be considered as correct control of type 1 error. Under a normal error model, using absolute residual with an F-test in the second stage is the only combination with type 1 error in this range. The absolute residuals combined with SKAT and the squared residuals with SKAT can be considered as having acceptable type 1 error, since the estimates are are slightly below the nominal level. Hence, these choices will give conservative conclusions when detecting true associations. However, SKATO demonstrates a significant T1E inflation with both residual types. Under a t- distributed error model, the absolute residuals with all three candidate test statistics lead to lower T1E than under a normal error model, while the squared residuals give higher T1E than under normal error model.

	absolute residual $d_i = \|e_i\|$			squared residual $d_i = e_i^2$		
	F-test	SKAT	SKATO	F	SKAT	SKATO
normal distributed error	0.0505	0.0478	0.1155	0.0200	0.0469	0.0521
t distributed error	0.0300	0.0476	0.0908	0.0277	0.0539	0.0588

Table 2: Empirical type 1 error simulation for 6 candidate combinations of mvgS test under normal error model and under setting (a) t distributed error.

Within mvgS test, we evaluate the power of each combinations under the genotypic regression model and the additive model. Comparing the performance of two differently coded mvgS, there is around 49% to 67% power loss due to model mis-specification(Table 3). In the same table when we increase the variance of error term in data generating model (setting(b)), the power of mvgS decrease due to the variance explained by observed variables are lessened. Among mvgS, those with absolute residual are affected more than mvgS with squared residual. In setting (d) where β_2 increase from 0.015 to 0.1, the power loss of mvgS with absolute residual varies from 13% to 17%, and from 5% to 9% for mvgS with squared residual.

Overall, the power of mvgS with squared residual is higher than the original mvgS with absolute residual when the error variance gets larger. As addressed in many previous literature ([**loh1987some**] and xxx), using squared error instead of absolute error in Levene's test gives unstable type 1 error control under non-normal error distribution. We

	absolute residual $d_i = \lvert e_i \rvert$			squared residual $d_i = e_i^2$		
	F-test	SKAT	SKATO	F	SKAT	SKATO
Original	0.259	0.212	0.250	0.232	0.193	0.211
Additive coding	0.673	0.413	0.499	0.695	0.445	0.442
(b) $\sigma^2 = 2 \times 0.27^2$	0.120	0.161	0.194	0.196	0.160	0.171
(b) Additive coding	0.661	0.376	0.464	0.694	0.434	0.444
(c) $\beta_2 = 0.1$	0.214	0.183	0.211	0.212	0.184	0.198
(c) Additive coding	0.553	0.285	0.383	0.572	0.331	0.327

Table 3: Within mvgS power evaluation of 6 combinations of different residual types and stage 2 statistics for original mvgS and for setting (b) additive coded mvgS, (c) larger variance of ϵ_i, and (d) larger environment effect β_2.

do not carry squared residuals in the following power comparison with direct tests, GESAT and iSKAT.

5 Discussion

- The mvgS with absolute residual and F test has correct type 1 error control, but the power is affected by the magnitude of error variance σ^2. The mvgS with squared residual has better power under normal error distribution, but its type 1 error is sensitive to non-normality. Despite of the power gain from using squared residual, the danger of unstable type 1 error is sufficiently a concern.

- Another powerful method: adaptive combination of Bayes factors method(ADABF) has been proposed in 2018 (Lin et al., 2018) that allows prior information in the test. Others propose mixed effect model instead of GLM in GxE interaction test.

- A summary of permutation testing in regression can be found in [**anderson2001permutation**]. The article summarises permutation testing in models with one and two main effects, and notes that in a model with two main effects and an interaction term there is no exact spermutation method for testing the interaction term. Consider the same data generating model equation

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 G_i E_i + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{15}$$

a simple permutation based method would fix G, E and permutes outcome Y to give Y^{perm}, independent of G and E. However, Y is not independent of G and E in model 16. In our simulation design, jointly permutated phenotype Y^{perm} is associated with E^{perm} but still independent from G. Anderson claims that unless the main effect β_1 equals to zero or its exact value being known, joint permutation only approximate a restrictive empirical null distribution. Comparing our design to the real GWAS setting, where we keep the environmental factor E and phenotype Y as fixed, and search through all candidate SNPs for the one with interaction effect. Most of the SNPs have no interaction effect, that is, under the null hypothesis, but with possible main effect. Our proposed permutation cannot honestly keep the main effect of G, thus is not perfect under the case. (need to rephrase this part to show: although not perfect, permuting YE pair is the closest to a real GWAS study.

- Contribution and limitation

6 Supplement Materials

6.1 Proof for Independence in the Permutation Type 1 Error Design

A short proof of the independence between Y_i^{perm} and $G_i \times E_i^{perm}$ is stated below:

proof:

Consider a single SNP case where the data generating model is

$$Y_i = G_i + E_i + G_i \times E_i + \varepsilon_i \tag{16}$$

Imaging we permute all variables (Y, E, G, ε) together, we can write the permuted phenotype as:

$$Y_i^{perm} = G_i^{perm} + E_i^{perm} + G_i^{perm} E_i^{perm} + \varepsilon_i^{perm}$$

Although G_i is never permuted and G_i^{perm} is not available in application, we use the representation of Y_i^{perm} for the independence derivation. Naturally, the permuted G_i^{perm} is independent of G_i, and ε_i^{perm} is independent of ε_i. From the model assumption, we further have $G_i \perp\!\!\!\perp E_i$, $G_i \perp\!\!\!\perp E_i^{perm}$ and $G_i^{perm} \perp\!\!\!\perp E_i^{perm}$. To evaluate the type 1 error of a particular test of interaction, we need to make sure we are under the null hypothesis that $Y_i^{perm} \perp\!\!\!\perp G_i \times E_i^{perm}$.

$$\begin{aligned}
&Cov(Y_i^{perm}, G_i \times E_i^{perm}) = Cov(G_i^{perm} + E_i^{perm} + G_i^{perm} E_i^{perm} + \varepsilon_i^{perm}, G_i E_i^{perm}) \\
=&Cov(G_i^{perm}, G_i E^{perm}) + Cov(E_i^{perm}, G_i E_i^{perm}) + Cov(G_i^{perm} E^{perm}, G_i E^{perm}) \\
=&\mathbb{E}(E_i^{perm} G_i E_i^{perm}) - \mathbb{E}(E_i^{perm})\mathbb{E}(G_i)\mathbb{E}(E_i^{perm}) \\
&+ \mathbb{E}(G_i^{perm} E^{perm} G E^{perm}) - \mathbb{E}(G_i^{perm})\mathbb{E}(E_i^{perm})\mathbb{E}(G_i)\mathbb{E}(E^{perm}) \\
=&\big(1 + \mathbb{E}(G_i)\big) \cdot \mathbb{E}(G_i) \cdot Var(E_i) \\
=&0 \quad \text{(given that } G \text{ is centered and scaled)}
\end{aligned} \tag{17}$$