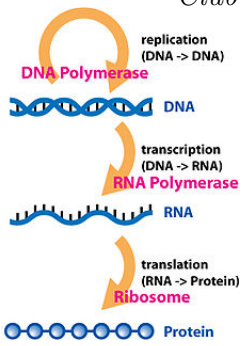
Honglang Wang
Depart. of Stat. & Prob.
wangho16@msu.edu

Omics Data Integration

Statistical Genetics/Genomics Journal Club



Summary and discussion of "Joint Analysis of SNP and Gene Expression Data in Genetic Association Studies of Complex Diseases"

Omics data integration @ Statistical Genetics/Genomics Journal Club @ MSU

Abstract *Technological platforms have advanced to a stage where many biological entities, e.g., genes, transcripts, and proteins, can be measured on the whole genome scale, yielding massive high-throughput 'omics data, such as genetic, genomic, epigenetic, protemic, and metabolomic data. The 'omics era provides and unprecedented promise of understanding common complex diseases, developing strategies for disease risk assessment, early detection, and prevention and intervention, and personalized therapies. In genetical genomics studies, it is important to jointly analyze gene expression data and genetic variants in exploring their associations with complex traits. This is for the discussion of the paper (Huang et al., 2014).*

Contents

1 Preliminary

1.1 Variance component test in GLM

In this section, we will recall some useful results for the variance component testing in generalized linear model with mixed effects (Lin, 1997).

Consider the following generalized linear mixed model

$$\begin{aligned}
&\mathbb{E}(Y|\mathbf{b}) = \boldsymbol{\mu}^b, \mathrm{Var}(Y|\mathbf{b}) = \mathrm{diag}\{\phi a_i^{-1} v(\mu_i^b)\};\\
&g(\boldsymbol{\mu}^b) = \boldsymbol{\eta}^b = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b},\\
&\mathbf{b} \sim F(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})),
\end{aligned} \tag{1}$$

where $g(\cdot)$ is a monotonic differentiable link function, and the covariance matrix of the random effects has the property that $\boldsymbol{\theta} = \mathbf{0}$ implies $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{0}$.

1. First of all by GLM theory we have the conditional log-quasilikelihood of $\boldsymbol{\alpha}$ given $\mathbf{b}$

$$l_i(\boldsymbol{\alpha}; \mathbf{b}) \propto \int_{y_i}^{\mu_i^b} \frac{a_i(Y_i - u)}{\phi v(u)} du. \tag{2}$$

   Note that for the logistic mixed model, we have the conditional log-likelihood, so we could just use it here.

2. Then we have the marginal (integrated) quasilikelihood of $(\boldsymbol{\alpha}, \boldsymbol{\theta})$

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \int \prod_i \exp(l_i(\boldsymbol{\alpha}; \mathbf{b})) dF(\mathbf{b}; \boldsymbol{\theta}) = \int \exp(\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{b})) dF(\mathbf{b}; \boldsymbol{\theta}). \tag{3}$$

3. Approximate the above likelihood (3) by using Laplace method since its hard to calculate the integral:

$$\begin{aligned}
\exp(\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{b})) \approx &\exp\Big\{\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{0})\Big\}\Big\{1 + \sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i} \mathbf{Z}_i^\intercal \mathbf{b}\\
&+ \frac{1}{2}\mathbf{b}^\intercal\big[\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i\}\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i^\intercal\} + \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i^2}\mathbf{Z}_i\mathbf{Z}_i^\intercal\big]\mathbf{b}\Big\},\\
L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = &\mathbb{E}_{\mathbf{b}}\big\{\exp(\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{b}))\big\} \approx \exp\Big\{\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{0})\Big\}\\
&\Big\{1 + \frac{1}{2}\mathrm{tr}\Big(\big[\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i\}\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i^\intercal\} + \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i^2}\mathbf{Z}_i\mathbf{Z}_i^\intercal\big]\mathbf{D}(\boldsymbol{\theta})\Big)\Big\},\\
l(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \log L(\boldsymbol{\alpha}, \boldsymbol{\theta}) \approx &\sum_{i=1}^n l_i(\boldsymbol{\alpha}; \mathbf{0})\\
&+ \frac{1}{2}\mathrm{tr}\Big(\big[\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i\}\{\sum_{i=1}^n \frac{\partial l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i}\mathbf{Z}_i^\intercal\} + \sum_{i=1}^n \frac{\partial^2 l_i(\boldsymbol{\alpha}; \mathbf{0})}{\partial \eta_i^2}\mathbf{Z}_i\mathbf{Z}_i^\intercal\big]\mathbf{D}(\boldsymbol{\theta})\Big).
\end{aligned} \tag{4}$$

4. A global score statistic for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is constructed as follows:

$$\chi^2_G := U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}})^\intercal \tilde{I}(\hat{\boldsymbol{\alpha}})^{-1} U_{\boldsymbol{\theta}}(\hat{\boldsymbol{\alpha}}) \tag{5}$$

where $\hat{\boldsymbol{\alpha}}$ is the MLE estimator under the null hypothesis (which is the usual generalized linear model), $U_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is the gradient vector $\partial l(\boldsymbol{\alpha}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$, and $\tilde{I}$ is the information matrix of $\boldsymbol{\theta}$ under the null, which takes the form

$$\tilde{I} = I_{\boldsymbol{\theta}\boldsymbol{\theta}} - I^\intercal_{\boldsymbol{\alpha}\boldsymbol{\theta}} I^{-1}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} I_{\boldsymbol{\alpha}\boldsymbol{\theta}},$$

with $I_{\boldsymbol{\alpha}\boldsymbol{\theta}} = \mathbb{E}(\frac{\partial l}{\partial \boldsymbol{\alpha}} \frac{\partial l}{\partial \boldsymbol{\theta}^\intercal})$, $I_{\boldsymbol{\theta}\boldsymbol{\theta}} = \mathbb{E}(\frac{\partial l}{\partial \boldsymbol{\theta}} \frac{\partial l}{\partial \boldsymbol{\theta}^\intercal})$ and $I_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \mathbb{E}(\frac{\partial l}{\partial \boldsymbol{\alpha}} \frac{\partial l}{\partial \boldsymbol{\alpha}^\intercal})$.

1.2 Mediation Analysis

Mediation analysis (VanderWeele and Vansteelandt, 2010) is just one type of causal inference in statistics (Rubin, 1990).

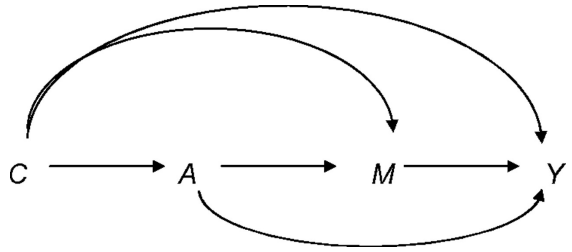


Figure 1: Example of mediation with exposure A, mediator M, outcome Y, and covariates C.

We will let A denote an exposure of interest, Y a dichotomous outcome, and M a potential mediator. We let C denote a set of baseline covariates not affected by the exposure. The relations among these variables are depicted in Figure (1). For example, A may denote genetic variants such as SNP, M m-RNA expression level, and Y cardiovascular disease. A question of interest may then be the extent to which the effect of genetic variants A on cardiovascular disease Y is mediated through m-RNA expression level M and the extent to which it is through other ways.

Let $Y(a, m)$ be the potential outcome that would have been observed if $A = a$ and $M = m$, and $M(a)$ be the potential outcome of M had the A been set to a.

The direct effect of SNPs is the effect of the SNPs on the disease outcome that is not through gene expression, whereas the indirect effect is the effect of the SNPs on the disease outcome that is through the gene expression. We can define the direct effect (DE), the indirect effect (IE) and the total effect (TE) of the SNPs, respectively, on the log odds ratio (OR) scale as:

1. the total effect (TE), conditional on $C = c$, comparing exposure level a with a^*, is defined by

$$\log OR^{TE}_{a,a^*|c} = \text{logit}(\mathbb{P}(Y(a, M(a)) = 1)) - \text{logit}(\mathbb{P}(Y(a^*, M(a^*)) = 1)).$$

2. the direct effect (DE), conditional on $C = c$ and $M = M(a)$, comparing exposure level a with a^*, is defined by

$$\log OR^{DE}_{a,a^*|c}(a) = \text{logit}(\mathbb{P}(Y(a, M(a)) = 1)) - \text{logit}(\mathbb{P}(Y(a^*, M(a)) = 1)).$$

3. the indirect effect (IE), conditional on $C = c$ and $A = a$, comparing exposure level a with a^*, is defined by

$$\log OR^{IE}_{a,a^*|c}(a) = \text{logit}(\mathbb{P}(Y(a, M(a)) = 1)) - \text{logit}(\mathbb{P}(Y(a, M(a^*)) = 1)).$$

Thus we have that

$$\log OR^{TE}_{a,a^*|c} = \log OR^{DE}_{a,a^*|c}(a) + \log OR^{IE}_{a,a^*|c}(a^*).$$

2 Summary of the paper

2.1 Methodology

In this paper, the authors proposed to jointly model a set of SNPs within a gene, a gene expression, and disease status, where a logistic model is used to model the dependence of disease status on the SNP set and the gene expression, and a linear model is used for the dependence of the gene expression on the SNP-set, both adjusting for covariates. We are primarily interested in testing whether a gene, whose effects are captured by SNPs and/or gene expression, is associated with a disease phenotype.

A SNP-set and gene expression pair can be defined in multiple ways. For example, the SNP-set is the SNPs in a gene and the expression of the gene. Alternatively, one can choose the SNP-set as the eQTLs of the corresponding gene expression.

Consider the following model:

$$\text{logit}\{\mathbb{P}(Y_i = 1|\mathbf{S}_i, G_i, \mathbf{X}_i; \boldsymbol{\beta}_\mathbf{S}, \boldsymbol{\gamma})\} = \mathbf{X}_i^\intercal \boldsymbol{\alpha} + \mathbf{S}_i^\intercal \boldsymbol{\beta}_\mathbf{S} + G_i \beta_G + \mathbf{C}_i^\intercal \boldsymbol{\gamma}, i = 1, 2, \cdots, n \quad (6)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^q, \boldsymbol{\beta}_\mathbf{S} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathbb{R}^p, \beta_G \in \mathbb{R}$, $\mathbf{X}_i$ is the $q-$dim vector of covariates, $\mathbf{S}_i$ is the $p-$dim vector of SNPs, $G_i \in \mathbb{R}$ is the expression level, $\mathbf{C}_i = G_i \mathbf{S}_i$ is the interaction, and Y_i is the dichotomous response. And since the SNPs in a gene might be large and some might be highly correlated (due to linkage disequilibrium), we assume $\beta_{\mathbf{S},j} \overset{IID}{\sim} (0, \tau_S)$ and $\gamma_j \overset{IID}{\sim} (0, \tau_I)$. We are interested in the following global test problem

$$H_0 : \tau_S = \tau_I = \beta_G = 0. \quad (7)$$

Remark. How to understand our hypothesis testing in terms of the mediation analysis? In order to do this, we have to introduce the model which states the relationship between G_i and $(\mathbf{X}_i, \mathbf{S}_i)$:

$$G_i = \mathbf{X}_i^\intercal \boldsymbol{\phi} + \mathbf{S}_i^\intercal \boldsymbol{\delta} + \varepsilon_i, \quad (8)$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma_G^2)$. Now we are ready to calculate the TE, DE and IE by the following formula

$$\begin{aligned}
\log OR^{TE}_{\mathbf{s},\mathbf{s}^*|\mathbf{x}} =& (\mathbf{s}-\mathbf{s}^*)^\intercal\big\{\boldsymbol{\beta}_S + \beta_G\boldsymbol{\delta} + \boldsymbol{\gamma}(\mathbf{x}^\intercal\boldsymbol{\phi} + \mathbf{s}^{*\intercal}\boldsymbol{\delta} + \beta_G\sigma_G^2) + \boldsymbol{\delta}\mathbf{s}^\intercal\boldsymbol{\gamma}\big\} \\
&+ \frac{1}{2}\sigma_G^2(\mathbf{s}+\mathbf{s}^*)^\intercal\boldsymbol{\gamma}(\mathbf{s}-\mathbf{s}^*)^\intercal\boldsymbol{\gamma}, \\
\log OR^{DE}_{\mathbf{s},\mathbf{s}^*|\mathbf{x}}(\mathbf{s}) =& (\mathbf{s}-\mathbf{s}^*)^\intercal\big\{\boldsymbol{\beta}_S + \boldsymbol{\gamma}(\mathbf{x}^\intercal\boldsymbol{\phi} + \mathbf{s}^{*\intercal}\boldsymbol{\delta} + \beta_G\sigma_G^2)\big\} \\
&+ \frac{1}{2}\sigma_G^2(\mathbf{s}+\mathbf{s}^*)^\intercal\boldsymbol{\gamma}(\mathbf{s}-\mathbf{s}^*)^\intercal\boldsymbol{\gamma}, \\
\log OR^{IE}_{\mathbf{s},\mathbf{s}^*|\mathbf{x}}(\mathbf{s}^*) =& (\mathbf{s}-\mathbf{s}^*)^\intercal\big\{\boldsymbol{\beta}_S + \boldsymbol{\delta}\mathbf{s}^\intercal\boldsymbol{\gamma}\big\}.
\end{aligned} \tag{9}$$

Then we have the following explanation:

1. If $\boldsymbol{\delta} \neq \mathbf{0}$, i.e. the gene expression is associated with the SNPs, then $H_0 : \boldsymbol{\beta}_\mathbf{S} = \mathbf{0}, \beta_G = 0, \boldsymbol{\gamma} = \mathbf{0} \Leftrightarrow H_0 : DE = 0, IE = 0$. The test for the joint effects of SNPs in a SNP set and a gene expression on disease, i.e. the total effect of a gene, is equivalent to a test for the total SNP effect on the disease.

2. If $\boldsymbol{\delta} = \mathbf{0}$, i.e. the gene expression is not associated with the SNPs, then $H_0 : \boldsymbol{\beta}_\mathbf{S} = \mathbf{0}, \beta_G = 0, \boldsymbol{\gamma} = \mathbf{0}$ simply ties to evaluate the joint effect of SNP-set, gene expression and the interaction on disease.

By the above general methodology for the variance component testing, we have the following score statistics for τ_S, τ_I, β_G

$$U_S = (\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\mathbf{S}\mathbf{S}^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0), U_I = (\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\mathbf{C}\mathbf{C}^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0), U_G = (\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\mathbf{G}\mathbf{G}^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0), \tag{10}$$

where $\hat{\boldsymbol{\mu}}_0$ is the MEL of the mean of $\mathbf{Y}$ under the null hypothesis H_0 by using the standard method in GLM.

The authors proposed the following weighted sum of three scores as the test statistic for the null hypothesis:

$$Q = \frac{1}{n}\big\{w_1U_S + w_2U_G + w_3U_I\big\}, \tag{11}$$

where the weights w_1, w_2, w_3 are chosen to be the inverse of the standard deviation of the U_S, U_G, U_I, which can be calculated explicitly (Lin, 1997) as follows

$$1/w_1^2 = \mathbf{1}^\intercal(\mathbf{S}\mathbf{S}^\intercal \cdot \mathbf{K} \cdot \mathbf{S}\mathbf{S}^\intercal)\mathbf{1}, 1/w_2^2 = \mathbf{1}^\intercal(\mathbf{G}\mathbf{G}^\intercal \cdot \mathbf{K} \cdot \mathbf{G}\mathbf{G}^\intercal)\mathbf{1}, 1/w_3^2 = \mathbf{1}^\intercal(\mathbf{C}\mathbf{C}^\intercal \cdot \mathbf{K} \cdot \mathbf{C}\mathbf{C}^\intercal)\mathbf{1}, \tag{12}$$

where $\mathbf{K} = ((K_{ij}))$ with $K_{ii} = -4\hat{\mu}_{0i}^4 + 8\hat{\mu}_{0i}^3 - 5\hat{\mu}_{0i}^2 + \hat{\mu}_{0i}$ and for $i \neq j$, $K_{ij} = 2[\hat{\mu}_{0i}(1-\hat{\mu}_{0i})][\hat{\mu}_{0j}(1-\hat{\mu}_{0j})]$. Here $\mathbf{A} \cdot \mathbf{B}$ denotes the component-wise multiplication.

Now the asymptotic distribution of our test statistic Q under the null hypothesis can be derived as follows.

$$\begin{aligned}
Q &= \frac{1}{n}\big\{w_1 U_S + w_2 U_G + w_3 U_I\big\} = \frac{1}{n}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\big\{w_1\mathbf{S}\mathbf{S}^\intercal + w_2\mathbf{G}\mathbf{G}^\intercal + w_3\mathbf{C}\mathbf{C}^\intercal\big\}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)\\
&= \frac{1}{n}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\big(\sqrt{w_1}\mathbf{S}, \sqrt{w_2}\mathbf{G}, \sqrt{w_3}\mathbf{C}\big)\big(\sqrt{w_1}\mathbf{S}, \sqrt{w_2}\mathbf{G}, \sqrt{w_3}\mathbf{C}\big)^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)\\
&= \frac{1}{n}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)^\intercal\mathbf{V}\mathbf{V}^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0) = \|\frac{1}{\sqrt{n}}\mathbf{V}^\intercal(\mathbf{Y}-\hat{\boldsymbol{\mu}}_0)\|_2^2 = \|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{V}_i(Y_i-\hat{\mu}_{0i})\|_2^2
\end{aligned}$$

where $\mathbf{V}_i = (\sqrt{w_1}\mathbf{S}_i^\intercal, \sqrt{w_2}G_i, \sqrt{w_3}\mathbf{C}_i^\intercal)^\intercal \in \mathbb{R}^{2p+1}$.

Note that for $S_U(\boldsymbol{\theta}) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{U}_i(Y_i-\mu_i)$ with $\mathbf{U}_i = (\mathbf{X}_i^\intercal, \mathbf{V}_i^\intercal) \in \mathbb{R}^{q+2p+1}$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\beta}_\mathbf{S}^\intercal, \beta_G, \boldsymbol{\gamma}^\intercal)^\intercal$ with $\boldsymbol{\theta}^0 = (\boldsymbol{\alpha}^{0\intercal}, \mathbf{0}^\intercal, 0, \mathbf{0}^\intercal)^\intercal$, from the GLM theory, we have that (not rigorous)

$$S_U(\boldsymbol{\theta}^0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{D}) \tag{13}$$

where $\mathbf{D} = \begin{pmatrix}\mathbf{D}_{XX} & \mathbf{D}_{XV}\\ \mathbf{D}_{VX} & \mathbf{D}_{VV}\end{pmatrix} = n^{-1}\mathbf{U}^\intercal\mathbf{W}\mathbf{U}$ with $\mathbf{W} = \mathrm{diag}\{\mu_i(1-\mu_i)\}$.

By Taylor expansion, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{V}_i(Y_i-\hat{\mu}_{0i}) \approx \mathbf{A}S_U(\boldsymbol{\theta}^0)$$

where $\mathbf{A} = (-\mathbf{D}_{XV}^\intercal\mathbf{D}_{XX}^{-1}, \mathbf{I}_{2p+1})$, and then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{V}_i(Y_i-\hat{\mu}_{0i}) \xrightarrow{d|H_0} \mathrm{N}(\mathbf{0}, \mathbf{A}\mathbf{D}\mathbf{A}^\intercal), \tag{14}$$

which implies the following mixture of χ^2 asymptotic distribution

$$Q = \|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{V}_i(Y_i-\hat{\mu}_{0i})\|_2^2 \xrightarrow{d|H_0} \sum_{l=1}^{2p+1}(\mathbf{A}_l^\intercal\boldsymbol{\epsilon})^2 := Q_0 \tag{15}$$

where $\mathbf{A}_l$ is the $l-$th row of $\mathbf{A}$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$.

And people use a scaled χ^2 distribution by matching the first two moments to approximate the mixture of χ^2 distribution, i.e. $Q_0 \approx \kappa\chi_\nu^2$, with $\kappa = \mathrm{Var}(Q_0)/(2\mathbb{E}(Q_0))$ and $\nu = 2[\mathbb{E}(Q_0)]^2/\mathrm{Var}(Q_0)$.

The paper want to demonstrate the following points:

1. For eQTL SNPs, i.e. $\boldsymbol{\delta} \neq \mathbf{0}$, the null hypothesis is equivalent to no total genetic effect, which is defined above.

2. For non eQTL SNPs, i.e. $\boldsymbol{\delta} = \mathbf{0}$, then there is no such equivalence.

3. For traditional SNP only genetic analysis, we use the following model

$$\text{logit}\{\mathbb{P}(Y_i = 1|\mathbf{S}_i, \mathbf{X}_i)\} = \mathbf{X}_i^\intercal \boldsymbol{\alpha}^* + \mathbf{S}_i^\intercal \boldsymbol{\beta}_\mathbf{S}^*, i = 1, 2, \cdots, n, \tag{16}$$

for testing $H_0 : \boldsymbol{\beta}_\mathbf{S}^* = \mathbf{0}$. If our integrative model is true without interaction, then we have the following random intercept true model

$$\text{logit}\{\mathbb{P}(Y_i = 1|\mathbf{S}_i, \mathbf{X}_i, \epsilon_i)\} = \mathbf{X}_i^\intercal(\boldsymbol{\alpha} + \beta_G \boldsymbol{\phi}) + \mathbf{S}_i^\intercal(\boldsymbol{\beta}_\mathbf{S} + \beta_G \boldsymbol{\delta}) + \beta_G \epsilon_i, i = 1, 2, \cdots, n,$$

which leads to by integrating the random intercept out

$$\text{logit}\{\mathbb{P}(Y_i = 1|\mathbf{S}_i, \mathbf{X}_i)\} \approx c\Big\{\mathbf{X}_i^\intercal(\boldsymbol{\alpha} + \beta_G \boldsymbol{\phi}) + \mathbf{S}_i^\intercal(\boldsymbol{\beta}_\mathbf{S} + \beta_G \boldsymbol{\delta})\Big\}, i = 1, 2, \cdots, n.$$

Thus testing $H_0 : \boldsymbol{\beta}_\mathbf{S}^* = \mathbf{0}$ is approximately equivalent to testing for no total effect of the SNPs.

And if with interaction, it can be shown that the difference is that traditional approach will lose power.

2.2 Implementation

In order to raise the power for different alternatives such as only SNPs, (SNPs,gene expression), and (SNPs, gene expression, interaction), which correspond to the following three test statistics

$$\begin{aligned} Q_{SGC} &= \frac{1}{n}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)^\intercal \big\{w_1 \mathbf{S}\mathbf{S}^\intercal + w_2 \mathbf{G}\mathbf{G}^\intercal + w_3 \mathbf{C}\mathbf{C}^\intercal\big\}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0), \\ Q_{SG} &= \frac{1}{n}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)^\intercal \big\{w_1 \mathbf{S}\mathbf{S}^\intercal + w_2 \mathbf{G}\mathbf{G}^\intercal\big\}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0), \\ Q_S &= \frac{1}{n}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)^\intercal \big\{w_1 \mathbf{S}\mathbf{S}^\intercal\big\}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0), \end{aligned} \tag{17}$$

the authors proposed to use the minimum of the three p-values as the new test statistic. But the null distribution of the minimum p-values is hard to derive. So we resort to the score-based wild bootstrapping (Kline and Santos, 2012).

1. Generate the bootstrapped score:

$$\boldsymbol{\epsilon}^{*b} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}_i (Y_i - \hat{\mu}_{0i}) \mathcal{N}_i, b = 1, 2, \cdots, B,$$

where $\mathcal{N}_i \overset{IID}{\sim} \text{N}(0, 1)$.

2. Produce three p-values: first notice that the three different $Q_{0,SGC}, Q_{0,SG}, Q_{0,S}$ correspond to the three different matrix $\mathbf{A}$ with different $\mathbf{V}_i = (\sqrt{w_1}\mathbf{S}_i^\intercal, \sqrt{w_2}G_i,$

$\sqrt{w_3}\mathbf{C}_i^\intercal)^\intercal, (\sqrt{w_1}\mathbf{S}_i^\intercal, \sqrt{w_2}G_i)^\intercal$ and $\sqrt{w_1}\mathbf{S}_i$ respectively. And we denote these three different $\mathbf{A}$ by $\mathbf{A}_{SGC}, \mathbf{A}_{SG}, \mathbf{A}_S$. Then we have

$$\begin{aligned} Q_{0,SGC}^{*b} &= \|\mathbf{A}_{SGC}\boldsymbol{\epsilon}^{*b}\|_2^2, \\ Q_{0,SG}^{*b} &= \|\mathbf{A}_{SG}\boldsymbol{\epsilon}^{*b}\|_2^2, \\ Q_{0,S}^{*b} &= \|\mathbf{A}_{S}\boldsymbol{\epsilon}^{*b}\|_2^2, \end{aligned} \tag{18}$$

which leads to the three different p-values

$$\hat{P}_{SGC} = \mathbb{P}^*(Q_{0,SGC}^* > Q_{SGC}), \hat{P}_{SG} = \mathbb{P}^*(Q_{0,SG}^* > Q_{SG}), \hat{P}_S = \mathbb{P}^*(Q_{0,S}^* > Q_S), \tag{19}$$

which can be estimated by $\sum_{b=1}^B \mathbb{1}(Q_{0,SGC}^{*b} > Q_{SGC})/B, \sum_{b=1}^B \mathbb{1}(Q_{0,SG}^{*b} > Q_{SG})/B, \sum_{b=1}^B \mathbb{1}(Q_{0,S}^{*b} > Q_S)/B$.

3. The minimum p-value statistic is now $\hat{P}_{\min} = \min\{\hat{P}_{SGC}, \hat{P}_{SG}, \hat{P}_S\}$, whose null distribution can be approximated by the empirical distribution of

$$\{\hat{P}_{\min}^b = \min\{\mathbb{P}^*(Q_{0,SGC}^* > Q_{0,SGC}^{*b}), \mathbb{P}^*(Q_{0,SG}^* > Q_{0,SG}^{*b}), \mathbb{P}^*(Q_{0,S}^* > Q_{0,S}^{*b})\}\}_{b=1}^B,$$

where the items such as $\mathbb{P}^*(Q_{0,SGC}^* > Q_{0,SGC}^{*b})$ can be approximated by $\sum_{b'=1}^B \mathbb{1}(Q_{0,SGC}^{*b'} > Q_{0,SGC}^{*b})/B$.

4. The p-value for the minimum p-value statistic can be calculated via

$$\sum_{b=1}^B \mathbb{1}(\hat{P}_{\min} < \hat{P}_{\min}^b)/B.$$

3 Simulation

3.1 Linkage Disequilibrium

There are four types of Haplotype for two loci in Table 1. And the data structure is usually given by the Table 2.

	B	b	Total
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
Total	p_B	p_b	1

Table 1: **Haplotype Frequencies.** For two loci, locus A and locus B, there are four haplotypes: AB, Ab, aB, ab.

	$BB(P_{BB})$	$Bb(P_{Bb})$	$bb(P_{bb})$
$AA(P_{AA})$	$n_{22}(p_{AB}^2)$	$n_{21}(2p_{AB}p_{Ab})$	$n_{20}(p_{Ab}^2)$
$Aa(P_{Aa})$	$n_{12}(2p_{AB}p_{aB})$	$n_{11}(2p_{AB}p_{ab} + 2p_{Ab}p_{aB})$	$n_{10}(2p_{Ab}p_{ab})$
$aa(P_{aa})$	$n_{02}(p_{aB}^2)$	$n_{01}(2p_{aB}p_{ab})$	$n_{00}(p_{ab}^2)$

Table 2: **Data Structure and expected genotype frequencies (assuming a random mating).** For two loci, locus A and locus B, there are four haplotypes: AB, Ab, aB, ab. $n = \sum_{i,j=0,1,2} n_{ij}$.

Linkage Equilibrium (Expected for Distant Loci) means two loci A and B are independent, which is equivalent to the follows from the contingency table

$$p_{AB} = p_A p_B,$$

which implies the following three

$$p_{Ab} = p_A p_b,\ p_{aB} = p_a p_B,\ p_{ab} = p_a p_b.$$

Then Linkage Disequilibrium (Expected for Nearby Loci) means

$$p_{AB} \neq p_A p_B.$$

And we have the coefficient of linkage disequilibrium (LD) between the two loci in the population,

$$D = p_{AB} - p_A p_B \tag{20}$$

which reflects the degree of LD. The larger D, the stronger LD.

Since D corresponds to the covariance between the loci, we have the following standardized version

$$r = D/\sqrt{p_A(1-p_A)p_B(1-p_B)}. \tag{21}$$

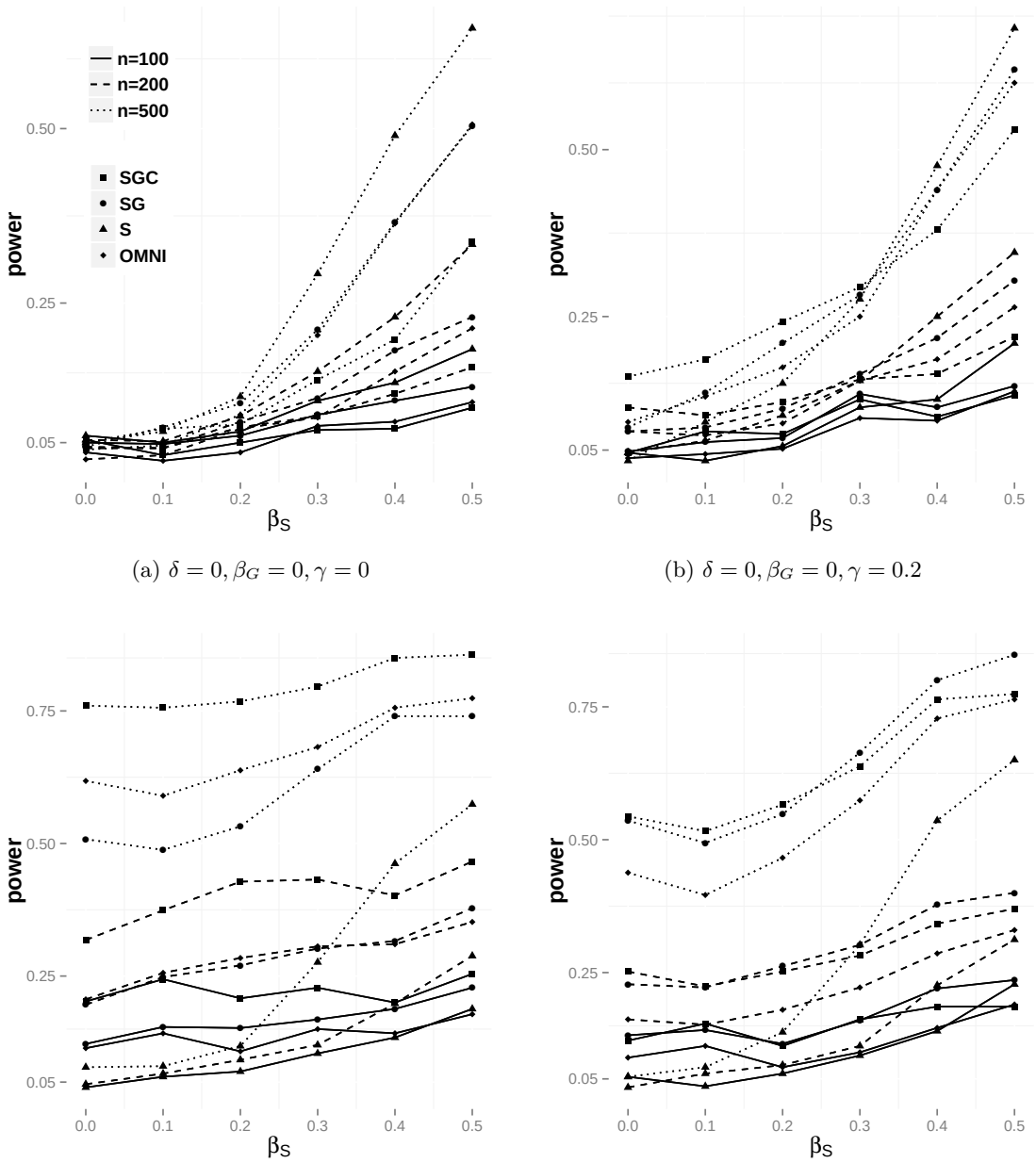
In summary, if we are given p_A, p_B and r, then we can calculate D, and then the haplotype frequencies such as p_{AB}. Then we will have the genotype frequencies such as p_{AaBB}. And then by the conditional arguments, we can simulate the data such as

$$P(Bb|aa) = \frac{P(aaBb)}{P(aa)} = \frac{2p_{aB}p_{ab}}{p_{aa}}.$$

3.2 Simulation Results

We tried the simulation for the following date generation:

```
We have q=1, i.e. with X only the intercept and p=10 SNPs
p=10;n=c(100,200,500)

pAs=c(0.1,0.15,0.1,0.2,0.25,0.1,0.2,0.05,0.4,0.3)#MAF
rs=c(0.1,0.8,0.5,0.4,-0.3,-0.2,0.2,0.15,0.4)#r
k0=4#the fourth one is the only causal SNP

delta=c(0, 0.5, 1)
beta_S=c(0, 0.1, 0.2, 0.3, 0.4, 0.5))
beta_G=c(0, 0.2, 0.5)
gamma=c(0, 0.2, 0.5)

generate SNP-set with LD structure
FSNP=sample(2:0,n,replace=TRUE,c(pAs[1]^2,2*pAs[1]*(1-pAs[1]),(1-pAs[1])^2))
SNP_data=FSNP
for(j in 2:p)
  NSNP=genSNP(FSNP, rs[j-1], pAs[j-1],pAs[j])  FSNP=NSNP  SNP_data=cbind(SNP_data,FSNP)

generate m-RNA expression data by the linear model without covariates
G=matrix(delta*SNP_data[,k0]+rnorm(n),ncol=1)

generate the response
eta=-0.2+beta_S*SNP_data[,k0]+beta_G*G+gamma*SNP_data[,k0]*G
p_response=apply(matrix(eta,ncol=1),1,function(x) exp(x)/(1+exp(x)))
Y=apply(matrix(p_response,ncol=1),1,function(x) rbinom(1,1,x))
```



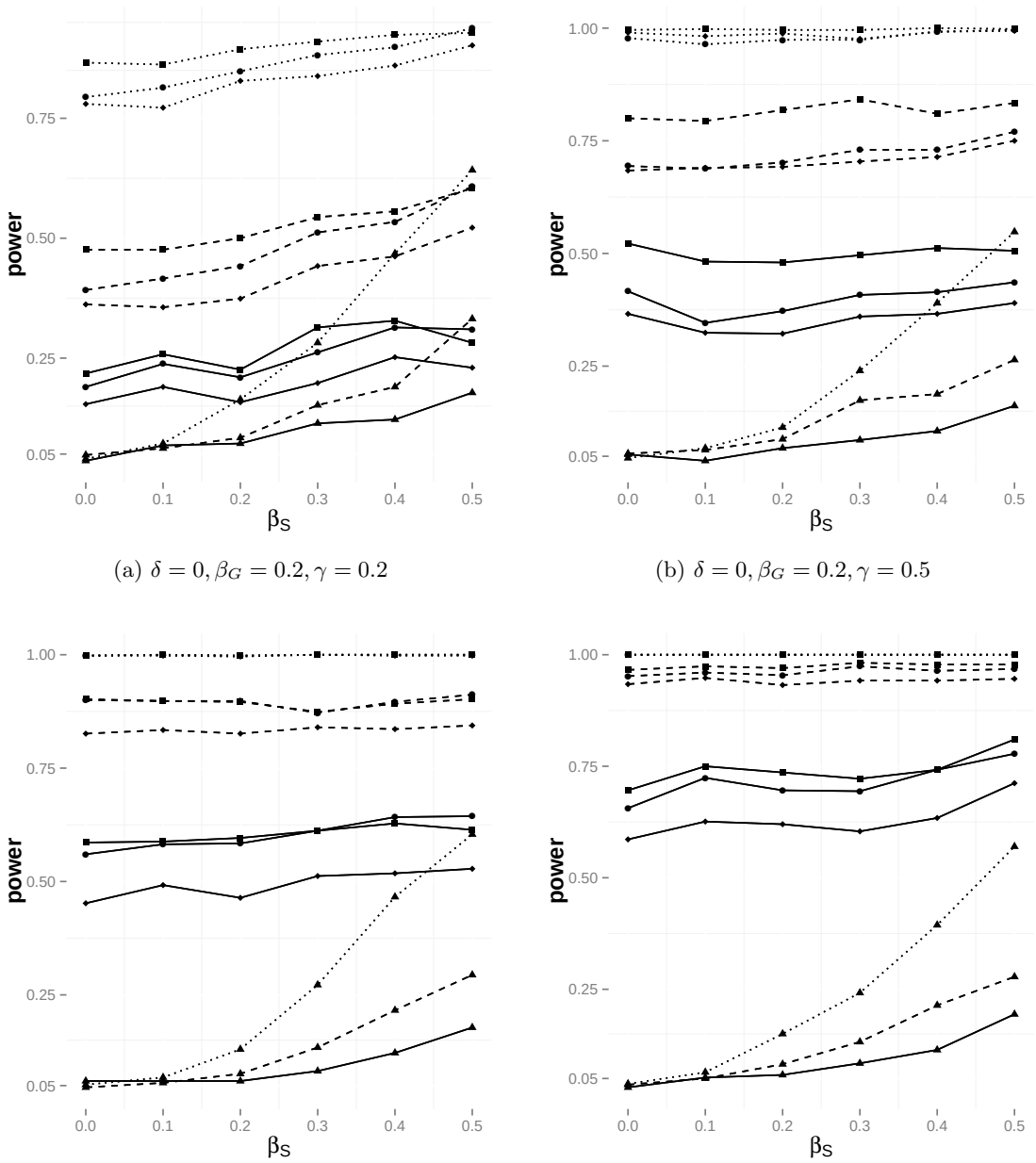
(a) $\delta = 0, \beta_G = 0, \gamma = 0$

(b) $\delta = 0, \beta_G = 0, \gamma = 0.2$

(c) $\delta = 0, \beta_G = 0, \gamma = 0.5$

(d) $\delta = 0, \beta_G = 0.2, \gamma = 0$

Figure 2: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



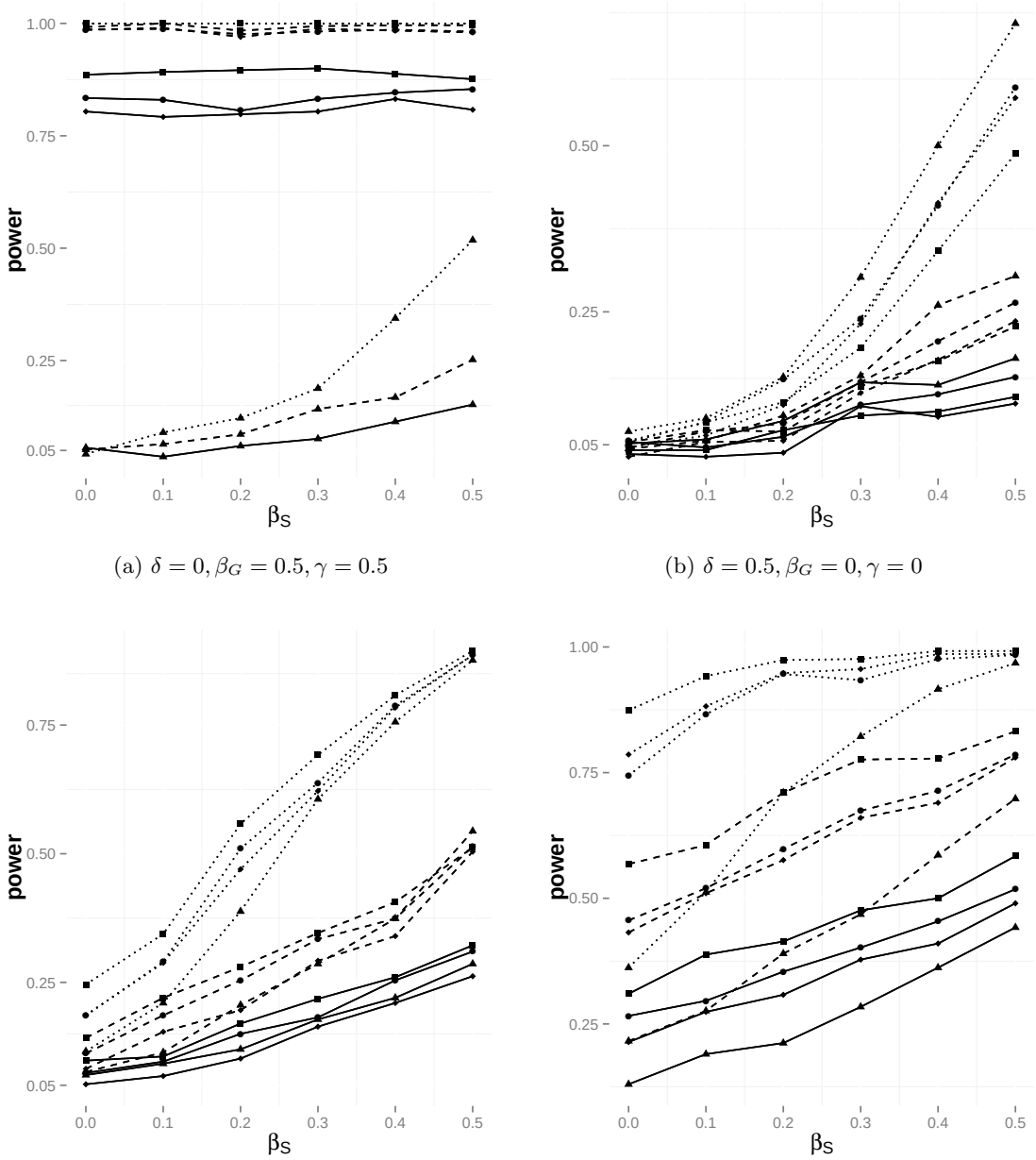
(a) $\delta = 0, \beta_G = 0.2, \gamma = 0.2$

(b) $\delta = 0, \beta_G = 0.2, \gamma = 0.5$

(c) $\delta = 0, \beta_G = 0.5, \gamma = 0$

(d) $\delta = 0, \beta_G = 0.5, \gamma = 0.2$

Figure 3: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



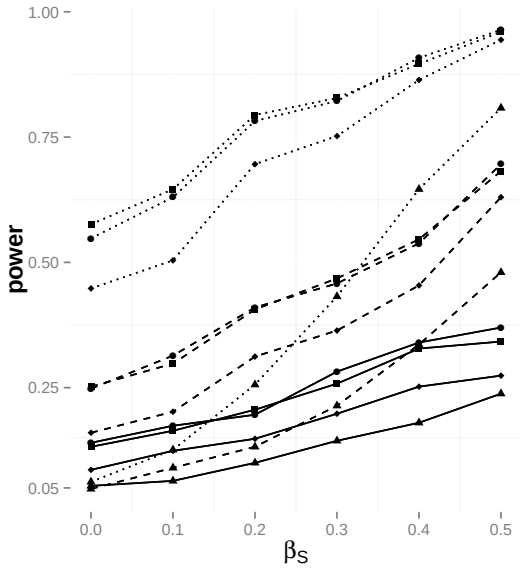
(a) $\delta = 0, \beta_G = 0.5, \gamma = 0.5$

(b) $\delta = 0.5, \beta_G = 0, \gamma = 0$

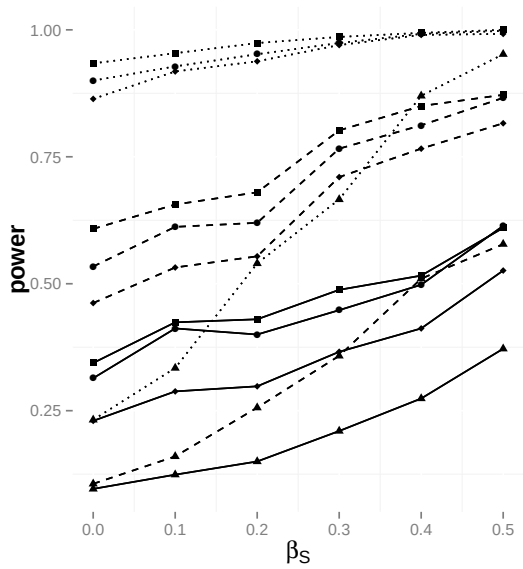
(c) $\delta = 0.5, \beta_G = 0, \gamma = 0.2$

(d) $\delta = 0.5, \beta_G = 0, \gamma = 0.5$

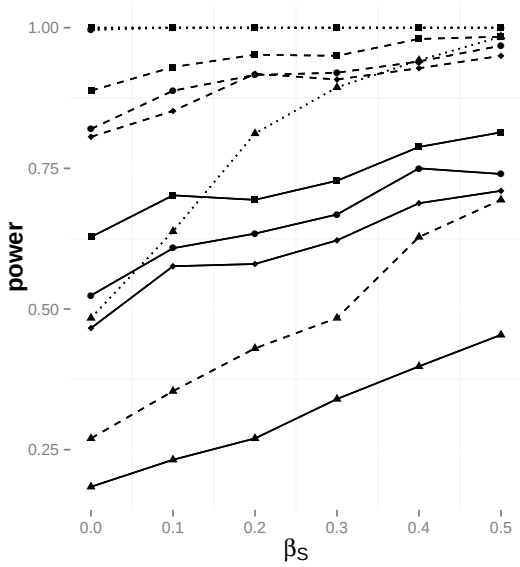
Figure 4: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



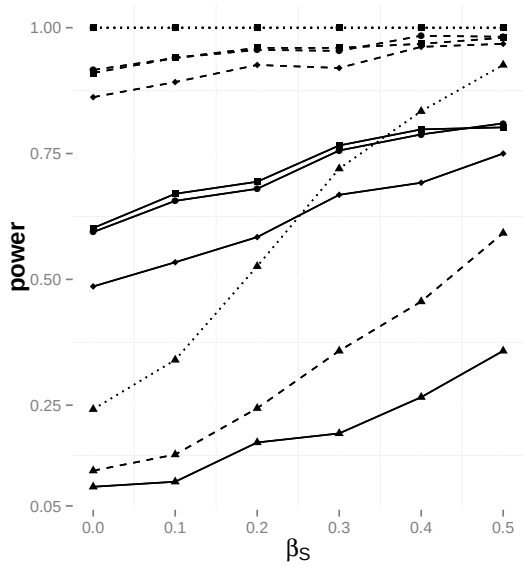
(a) $\delta = 0.5, \beta_G = 0.2, \gamma = 0$



(b) $\delta = 0.5, \beta_G = 0.2, \gamma = 0.2$

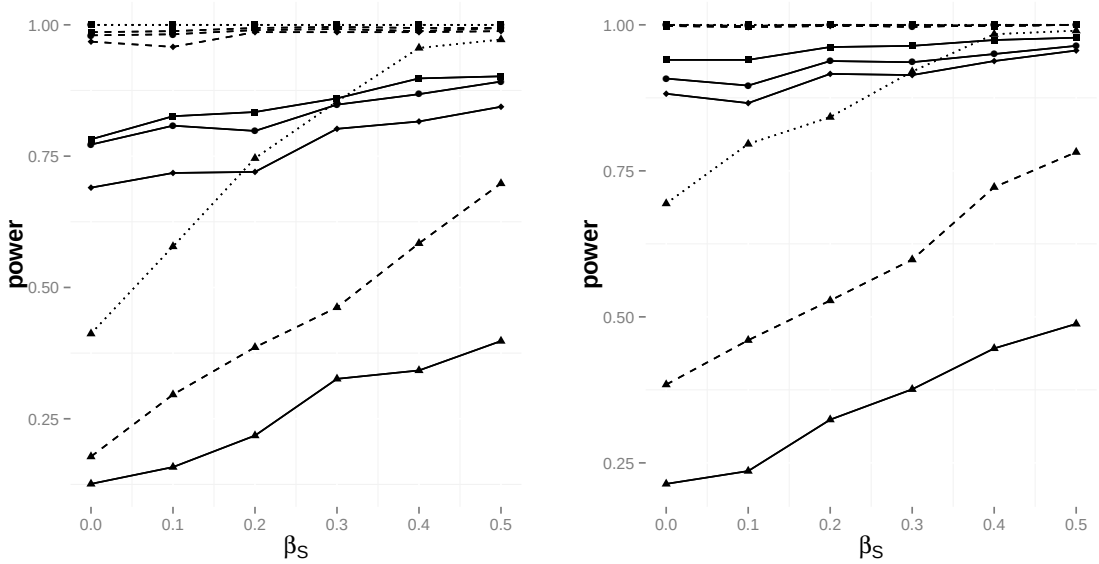


(c) $\delta = 0.5, \beta_G = 0.2, \gamma = 0.5$



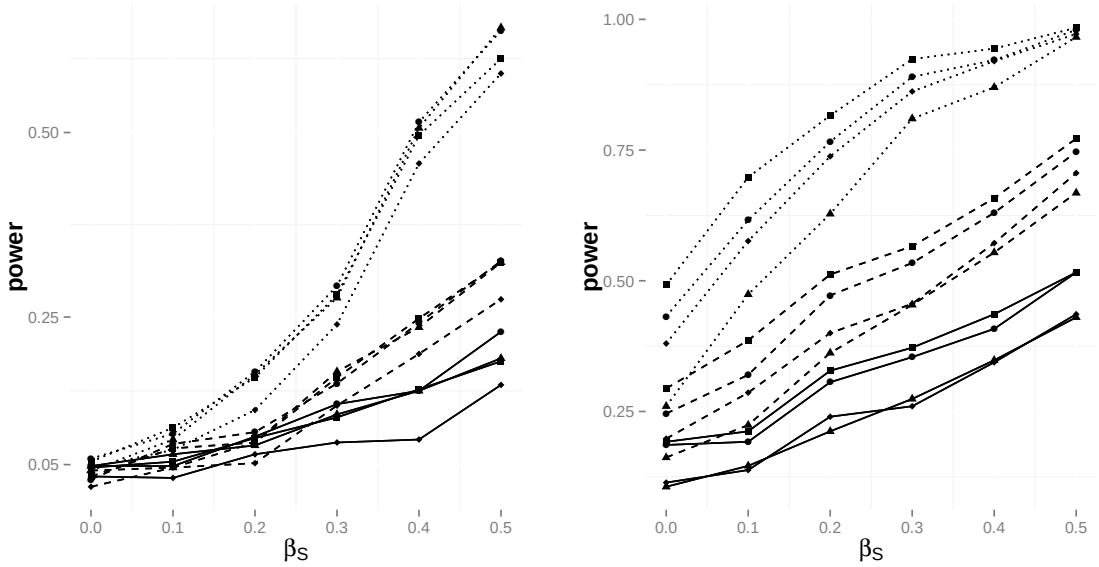
(d) $\delta = 0.5, \beta_G = 0.5, \gamma = 0$

Figure 5: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



(a) $\delta = 0.5, \beta_G = 0.5, \gamma = 0.2$

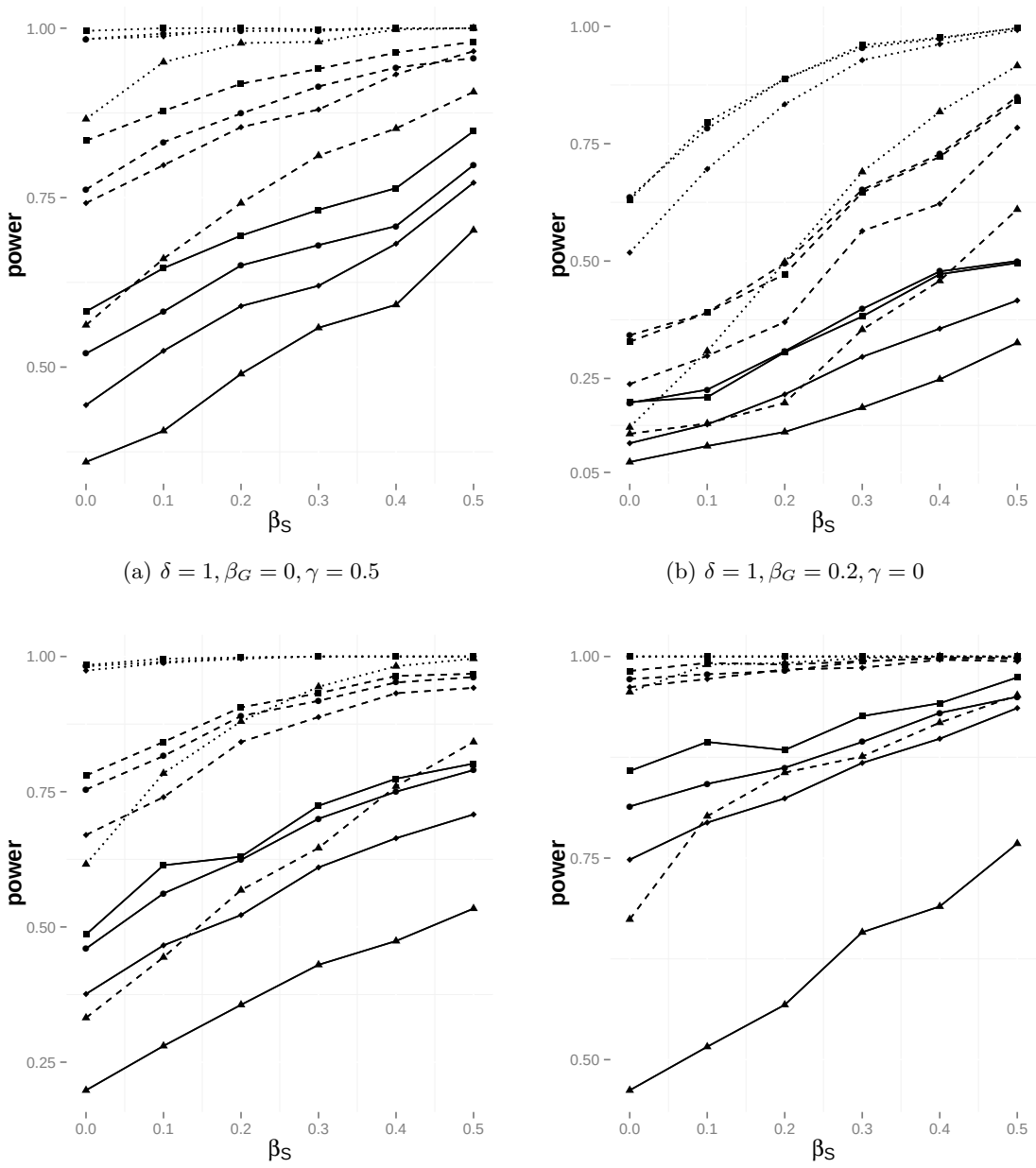
(b) $\delta = 0.5, \beta_G = 0.5, \gamma = 0.5$



(c) $\delta = 1, \beta_G = 0, \gamma = 0$

(d) $\delta = 1, \beta_G = 0, \gamma = 0.2$

Figure 6: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



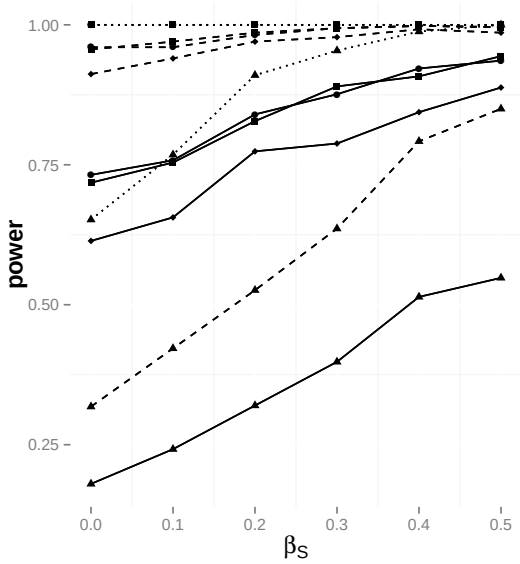
(a) $\delta = 1, \beta_G = 0, \gamma = 0.5$

(b) $\delta = 1, \beta_G = 0.2, \gamma = 0$

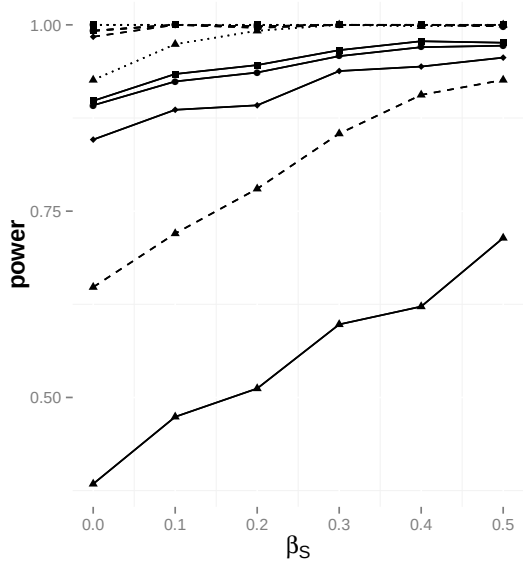
(c) $\delta = 1, \beta_G = 0.2, \gamma = 0.2$

(d) $\delta = 1, \beta_G = 0.2, \gamma = 0.5$

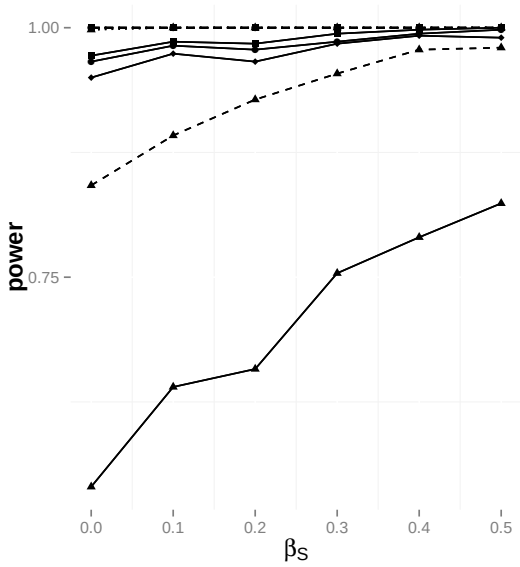
Figure 7: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.



(a) $\delta = 1, \beta_G = 0.5, \gamma = 0$



(b) $\delta = 1, \beta_G = 0.5, \gamma = 0.2$



(c) $\delta = 1, \beta_G = 0.5, \gamma = 0.5$

Figure 8: **Empirical Size and Power.** We used the score-based bootstrapping here. N=500, B=500.

From the simulation results, we could see that

1. For $\delta = 0$ case, in general, the model which is closest to the true model has the highest power.

2. For $\delta \neq 0$ case, it will have more power to detect some particular alternative hypothesis that the genetic effect is indirectly through expression level. Look at Figure 5 and compare the Plot (5a) and Plot (5d).

4 Discussion

4.1 Another way

For simplicity, we consider (note that our goal is testing the association between the outcome and a set of SNPs)

$$Y_i = \mathbf{G}_i^\intercal \boldsymbol{\beta}_G + \epsilon_{1i}, \tag{22}$$

$$\mathbf{G}_i^\intercal \boldsymbol{\beta}_G = \mathbf{S}_i^\intercal \boldsymbol{\alpha}_S + \epsilon_{2i} \tag{23}$$

and thus we have

$$Y_i = \mathbf{S}_i^\intercal \boldsymbol{\gamma}_S + \epsilon_i, \tag{24}$$

where $\boldsymbol{\gamma}_S = \boldsymbol{\alpha}_S, \epsilon_i = \epsilon_{1i} + \epsilon_{2i} \sim \mathrm{N}(0, \sigma_1^2 + \sigma_2^2)$. The null hypothesis:

$$H_0 : \boldsymbol{\alpha}_S = \mathbf{0}, i.e.\ \boldsymbol{\gamma}_S = \mathbf{0} \tag{25}$$

which is equivalent to no genetic effect. And this true model is reflecting our primary interest of alternative that $\mathbf{S}_i$ is associated with Y_i through regulation of the expression of $\mathbf{G}_i$ (Zhao et al., 2014).

If our integrative model is true, and the null hypothesis of no association between $\mathbf{S}_i$ and Y_i is equivalent to $\boldsymbol{\alpha}_S = \mathbf{0}$ in the integrative model and $\boldsymbol{\gamma}_S = \mathbf{0}$ in the usual linear model. Now the question is which one is more powerful.

1. Usual linear model: $\hat{\boldsymbol{\gamma}}_S = (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{Y}$ with $\mathrm{Var}(\hat{\boldsymbol{\gamma}}_S) = \boldsymbol{\Sigma}_{SS}^{-1}(\sigma_1^2 + \sigma_2^2)/n$ where $\boldsymbol{\Sigma}_{SS} = \mathbb{E}(\mathbf{S}_i\mathbf{S}_i^\intercal)$.

2. Integrative model: we first have $\hat{\boldsymbol{\beta}}_G = (\mathbf{G}^\intercal\mathbf{G})^{-1}\mathbf{G}^\intercal\mathbf{Y}$ and then we have $\hat{\boldsymbol{\alpha}}_S = (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\hat{\boldsymbol{\beta}}_G$. So $\mathrm{Var}(\hat{\boldsymbol{\alpha}}_S) = \mathrm{Var}(\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S) = \mathrm{Var}((\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\hat{\boldsymbol{\beta}}_G - \boldsymbol{\alpha}_S) = \mathrm{Var}((\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\hat{\boldsymbol{\beta}}_G - (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{S}\boldsymbol{\alpha}_S) = \mathrm{Var}((\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\hat{\boldsymbol{\beta}}_G - (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\boldsymbol{\beta}_G + (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\boldsymbol{\beta}_G - (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{S}\boldsymbol{\alpha}_S) = \mathrm{Var}((\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\hat{\boldsymbol{\beta}}_G - (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\boldsymbol{\beta}_G) + \mathrm{Var}((\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{G}\boldsymbol{\beta}_G - (\mathbf{S}^\intercal\mathbf{S})^{-1}\mathbf{S}^\intercal\mathbf{S}\boldsymbol{\alpha}_S) = \boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS}\boldsymbol{\Sigma}_{SS}^{-1}\sigma_1^2/n + \boldsymbol{\Sigma}_{SS}^{-1}\sigma_2^2/n$.

3. Now for $\mathbf{S}_i \in \mathbb{R}^p$, we take $p = 1$ for illustration. So if $\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS} < 1$ we will have $\mathrm{Var}(\hat{\boldsymbol{\alpha}}_S) < \mathrm{Var}(\hat{\boldsymbol{\gamma}}_S)$. If Gene expression are independent, then it's equivalent to $\|\mathrm{Corr}(S_i, \mathbf{G}_i)\|_2^2 < 1$. In other words, we gain the most power if the $\mathbf{G}_i$ are weakly correlated with $\mathbf{S}_i$. This is sensible, because otherwise the expression

data would add little additional information. In the extreme case where they are perfectly correlated, our integrative analysis would be no different from a standard analysis. On the other hand, while the integrative approach has more relative power for weak correlations, its absolute power can be low if the correlations are two low, as in the extreme case where the correlation is zero we have $\boldsymbol{\alpha}_S = 0$, i.e. null hypothesis holds. In the ideal setting, the correlations are weak but $\boldsymbol{\alpha}_S$ is still large, which only possible when $\mathbf{G}_i$ is highly associated with Y_i so that $\boldsymbol{\beta}_G$ is large.

4.2 Research Direction

1. How to generalize it to the nonlinear interaction situation?

2. Recently Haris et al. (2014) proposed how to fit the following interaction model

$$\text{logit}\{\mathbb{P}(Y_i = 1\} = \mathbf{S}_i^\intercal \boldsymbol{\beta}_\mathbf{S} + \mathbf{G}_i^\intercal \boldsymbol{\beta}_G + \mathbf{S}_i^\intercal \mathbf{B}\mathbf{G}_i, i = 1, 2, \cdots, n \tag{26}$$

under the strong heredity condition that if an interaction term is included in the model then both of the corresponding main effects must be present or weak heredity that if an interaction term is included in the model then at least one of the corresponding main effects must be present. That is variable selection with the constraints.

References

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*, 2014.

Yen-Tsung Huang, Tyler J VanderWeele, and Xihong Lin. Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *The annals of applied statistics*, 8(1):352, 2014.

Patrick Kline and Andres Santos. A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41, 2012.

Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

Donald B Rubin. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292, 1990.

Tyler J VanderWeele and Stijn Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172(12):1339–1348, 2010.

Sihai D Zhao, T Tony Cai, and Hongzhe Li. More powerful genetic association testing via a new statistical framework for integrative genomics. *Biometrics*, 64:1–21, 2014. ISSN 1541-0420. doi: 10.1111/biom.12206. URL `http://dx.doi.org/10.1111/biom.12206`.