Methods for High-Dimensional Inference in Genetic Association Studies

Methods for High-Dimensional Inference in Genetic Association Studies

A dissertation presented

by

Ryan Sun

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2017

Dissertation Advisor: Professor Xihong Lin Ryan Sun

Methods for high-dimensional inference in genetic association studies

Abstract

Genetic association studies are frequently characterized by high-dimensional datasets containing rare and weak signals. To detect these signals, it is important to choose inference methods that are both robust and powerful under such challenging settings. In this work we study the theoretical properties of popular existing techniques, and we propose new methods which aim to increase the accuracy and detection ability of genetic association testing.

In chapter 1, we discuss improper inference in Genome-Wide Environment Interaction Studies (GWEIS). Modeling gene-environment (GxE) interactions is often challenged by the unknown functional form of the environment term in the true data-generating mechanism. We study the impact of misspecification of the environmental exposure effect on inference for the GxE interaction term in linear and logistic regression models. We first examine the asymptotic bias of the GxE interaction regression coefficient, allowing for confounders as well as arbitrary misspecification of the exposure and confounder effects. For linear regression, we show that under gene-environment independence and some confounder-dependent conditions, when the environment effect is misspecified, the regression coefficient of the GxE interaction can be unbiased. However, inference on the GxE interaction is still often incorrect. In logistic regression, we show that the regression coefficient is generally biased if the genetic factor is associated with the outcome directly or indirectly. Further we show that the standard robust sandwich variance estimator for the GxE interaction does not perform well in practical GxE studies, and we provide an alternative testing procedure that has better finite sample properties.

In chapter 2, we propose a new set-based test for genetic association studies. Study-

ing the effects of groups of Single Nucleotide Polymorphisms (SNPs), as in a gene, genetic pathway, or network, can provide novel insight into complex diseases, above that which can be gleaned from studying SNPs individually. Common challenges in set-based genetic association testing include weak effect sizes, correlation between SNPs in a SNP-set, and scarcity of signals, with single-SNP effects often ranging from moderately sparse to extremely sparse in number. Motivated by these challenges, we propose the Generalized Berk-Jones (GBJ) test for the association between a SNP-set and outcome. The GBJ extends the Berk-Jones (BJ) statistic by accounting for correlation among SNPs, and it provides advantages over the Generalized Higher Criticism (GHC) test when signals in a SNP-set are moderately sparse. We also provide an analytic p-value calculation procedure for SNP-sets of any finite size. Using this p-value calculation, we illustrate that the rejection region for GBJ can be described as a compromise of those for BJ and GHC. We develop an omnibus statistic as well, and we show that this omnibus test is robust to the degree of signal sparsity. An additional advantage of our methods is the ability to conduct inference using individual SNP summary statistics from a Genome Wide Association Study (GWAS). We evaluate the finite sample performance of the GBJ though simulation studies, and we apply the method to gene-level association analysis of breast cancer risk using data from the Cancer Genetic Markers of Susceptibility GWAS.

In chapter 3, we investigate the power of different set-based tests for genetic association studies. It has become increasingly popular to perform set-based inference with a class of methods, popularized by the Higher Criticism statistic, which has asymptotic optimality properties in detecting sparse alternatives. However the choice of which test to use is not always clear. A key distinction between these methods is the manner they account for correlation among features in a set - either through a transformation to decorrelate the data, as in the innovated Higher Criticism (iHC), or by building the correlation into the test statistic, as in the Generalized Higher Criticism (GHC). In this paper we show that, depending on the correlation structure of the features, the decorrelation step in innovation-based methods can greatly increase power when testing for associations between one explanatory variable and a set of multiple outcomes, which we term the multiple phenotype setting. However when testing the association between one outcome

and a set of explanatory variables, which we term the SNP-set setting, the same advantages are no longer present. We validate our findings through simulation and application to both a methylation quantitative trait loci study of lung cancer patients and a GWAS of breast cancer risk.

Contents

Acknowledgments

I would like to thank my advisor, Professor Xihong Lin, for your unending patience in working with me for the past five years. I do not know that any other faculty member could have tolerated advising my younger self, and I cannot imagine pursuing this journey with another mentor.

I would also like to thank my committee members, Professor David Christiani, Professor Samuel Kou, Professor Peter Kraft, and Professor Eric Tchetgen Tchetgen, for your willingness to share your time and advice with me throughout the years. Your generous assistance has been invaluable in helping me overcome some of my largest challenges.

I would like to thank my fiance, Huili Zhu, for being by my side every day of this journey. You have been a better partner than I deserve, and you constantly motivate me to be a better person.

Finally, I would like to thank my parents and my brother, for a lifetime of unconditional support and encouragement. None of this was possible without you.

Testing for gene-environment interaction under exposure misspecification

Ryan Sun

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Raymond J. Carroll

Department of Statistics

Texas A&M University

School of Mathematical and Physical Sciences

University of Technology Sydney

David C. Christiani

Department of Environmental Health

Harvard T.H. Chan School of Public Health

Xihong Lin

Department of Biostatistics

Harvard T.H. Chan School of Public Health

1.1 Introduction

Many human diseases possess an etiology which is characterized by complex relationships between genetic and environmental risk factors. Studying gene-environment (GxE) interactions can help us understand biological mechanisms that cause these complex diseases (Thomas, 2010). There has been a dramatic increase in the number of Genome Wide Environmental Interaction Studies (GWEIS) over the past decade, yet remarkably, the number of replicable GxE interactions in the literature is only a handful (Aschard et al., 2012; Hutter et al., 2013). The lack of success in identifying GxE interactions is often attributed to study design issues, such as inadequate sample size and population heterogeneity (Thomas, 2010), but it also suggests limitations with current statistical methodology.

The standard approach to GWEIS performs single-marker analysis over a large number of Single Nucleotide Polymorphisms (SNPs) across the genome, repeatedly fitting a gene-environment interaction generalized linear model - e.g. linear regression for continuous traits and logistic regression for binary traits. In these models, the effect of the environmental exposure is often modeled parametrically. However we generally do not know the correct functional form of the environment covariate in the true data-generating mechanism (Aschard et al., 2012). Therefore the exposure effects can be misspecified, resulting in invalid model-based inference, as in the case of Cornelis et al. (2012). Environment misspecification may also cause the appearance of heteroscedasticity with respect to the exposure, which can similarly invalidate inference (Almli et al., 2014).

Our work is motivated by a GWEIS from the Harvard School of Public Health Superfund Research Project. One of the main goals of the Superfund program is to study how toxic metal exposures and genetic variants interact to affect neurodevelopment outcomes, such as Bayley Scales of Infant Development (BSID) scores, among infants. Data are available on approximately 500 infants in Bangladesh and 400 infants in Mexico. About 500,000 SNPs are used in an initial analysis, and a standard GxE interaction linear model is fit for each SNP. The QQ-plots of p-values generated by testing for GxE interaction show large departures from uniformity across many different exposures, multiple out-

comes, both cohorts, and even in meta-analyses of the two cohorts together. See Figure 1.2 for an example. However, tests for the main effects of SNPs (G) while adjusting for exposure ($\mathcal{E}$) produce a very uniform distribution of p-values. These diagnostics suggest misspecification of the GxE mean model, as described and explained in detail by Voorman et al. (2011). Surprisingly, GxE inference which utilizes the Huber-White 'sandwich' variance estimator, a commonly-proposed remedy for incorrect inference in GWEIS (Voorman et al., 2011; Tchetgen and Kraft, 2011; Almli et al., 2014) often produces inflated p-values that show a larger departure from uniformity than p-values calculated from model-based standard errors. Again, see Figure 1.2 for an example. Here inflation means there is an excess of small p-values, while deflation refers to the opposite.

The impact of performing inference with misspecified models for the effects of covariates has been investigated by many authors, primarily in main effects models. The GWEIS setting is unique, because we are interested in testing a possibly misspecified interaction term and not a main effect. Additionally, we allow for confounders in the model, and these confounders may be arbitrarily misspecified. In contrast, past work primarily focuses on main effects models and often assumes the term of interest is completely independent of other covariates in the model. Relevant literature includes Gail, Weiand and Piantadosi (1984), Lagakos (1988), and Begg and Lagakos (1992).

For interaction models, Vansteelandt et al. (2008) derived a set of multiply robust estimators for interaction effects, but these estimators require specification of a distribution for each SNP conditional on the other covariates in the model, a complicated task with hundreds of thousands of SNPs. Rosenblum and van der Laan (2009) and Tchetgen and Kraft (2011) studied misspecification constrained to the setting when G is completely independent of all other terms in the fitted model, including the outcome. In this scenario, they showed that the estimated GxE interaction effect will be asymptotically unbiased under the null, even under environment misspecification. While important, these findings are heavily constrained by the independence assumption. For example, under the infinitesimal model of genetic contribution to disease (Gibson, 2012), a large number of G terms are associated with an outcome, so a large proportion of tests in GWEIS will violate the assumption. Furthermore, adjusting for population stratification with genotype

principal components is important in genetic association studies. The principal components will introduce regression covariates that are associated with G, another example of common practice that violates the above assumption. Our work considers arbitrary dependence among all covariates in the model and can hence incorporate confounders like principal components. We allow for misspecification of these confounders as well.

There are two main objectives to this paper. First, we provide conditions for valid GxE interaction inference under the null hypothesis of no interaction effect when the environmental exposure, and possibly other covariates, are misspecified in the generalized linear model. We perform asymptotic bias analysis to show that for a linear regression model, the estimated interaction coefficient is asymptotically unbiased under the null if the genetic factor is independent of the environment and all additional covariates in both the true and fitted model are independent of either gene or environment. However, standard inference on the GxE interaction is incorrect even under these conditions. In addition, we show that for a logistic regression model, the asymptotic estimate of the interaction coefficient will generally be biased under environmental misspecification when the genetic factor is associated directly or indirectly with the outcome, even under gene-environment independence. For both models we confirm that bias in the model-based standard error estimate can lead to inflated and deflated QQ-plots.

Secondly, we describe why the often-proposed sandwich variance estimator may not be a panacea for inference in practical GWEIS with moderate sample sizes. Specifically, we show that the sandwich estimate can be plagued by high variability under environmental misspecification. We propose an estimator that has better finite sample properties and illustrate its utility through both simulation and application to the Superfund dataset.

1.2 Exposure misspecification in GxE inference

1.2.1 Assumptions and standard approach

Suppose that the outcome Y_i is related to covariates G_i, $\mathcal{E}_i$, $\mathbf{Z}_i$, and M_i by the generalized linear model (McCullagh and Nelder, 1989)

$$g(\mu_i) = \beta_0 + \beta_G G_i + \beta_{\mathcal{E}} f(\mathcal{E}_i) + \beta_I G_i h(\mathcal{E}_i) + \mathbf{Z}_i^T \boldsymbol{\beta}_Z + \mathrm{M}_i^T \boldsymbol{\beta}_M, \tag{1.1}$$

where $\mu_i = E(Y_i|G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{M}_i)$. For binary outcomes, $g(\cdot)$ is the logistic link. For continuous outcomes, $g(\cdot)$ is the identity link and $\text{var}(Y_i|G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{M}_i) = \sigma^2$. Let G_i denote a discrete genetic marker and $\mathcal{E}_i$ an environmental exposure variable. The additional covariates $\mathbf{Z}_i^T = (Z_{1i}, ..., Z_{pi})$ are correctly modeled, and the covariates $\mathbf{M}_i^T = (M_{1i}, ..., M_{qi})$ are subject to mismodeling. Take $f(\cdot)$ and $h(\cdot)$ to be known, possibly nonlinear functions of $\mathcal{E}$. In vector notation, we have $\boldsymbol{\beta} = (\beta_0, \beta_G, \beta_\mathcal{E}, \beta_I, \boldsymbol{\beta}_Z^T, \boldsymbol{\beta}_M^T)^T$ and $\mathbf{X}_i = \left\{1, G_i, f(\mathcal{E}_i), G_i h(\mathcal{E}_i), \mathbf{Z}_i^T, \mathbf{M}_i^T\right\}^T$ so that $g(\mu_i) = \mathbf{X}_i^T\boldsymbol{\beta}$. In the context of GWEIS, which are hypothesis-generating procedures, we are most interested in inference about whether $\beta_I = 0$.

Suppose that the observed data consist of n independent and identically distributed random vectors $(Y_i, G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{W}_i)$ for $i = 1, ..., n$, where additional observed covariates $\mathbf{W}_i^T = (W_{1i}, ..., W_{ri})$ are a possibly misspecified version of $\mathbf{M}_i^T$. The only restriction we place on $\mathbf{W}_i$ is that $E(Y_i|G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{M}_i, \mathbf{W}_i) = E(Y_i|G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{M}_i)$, or in other words, the misspecified covariates do not add information about Y_i above that given by $(G_i, \mathcal{E}_i, \mathbf{Z}_i, \mathbf{M}_i)$.

The standard test for gene-environment interaction fits the misspecified model

$$g(\widetilde{\mu}_i) = \alpha_0 + \alpha_G G_i + \alpha_\mathcal{E}\mathcal{E}_i + \alpha_I G_i\mathcal{E}_i + \mathbf{Z}_i^T\boldsymbol{\alpha}_Z + \mathbf{W}_i^T\boldsymbol{\alpha}_W \tag{1.2}$$

and performs inference on $H_0 : \alpha_I = 0$. We use $\widetilde{\mu}_i$ to denote that this is a misspecified model and not the true conditional mean.

Let $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_G, \alpha_\mathcal{E}, \alpha_I, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T\right)^T$ and $\widetilde{\mathbf{X}}_i = \left(1, G_i, \mathcal{E}_i, G_i\mathcal{E}_i, \mathbf{Z}_i^T, \mathbf{W}_i^T\right)^T$ so that $g(\widetilde{\mu}_i) = \widetilde{\mathbf{X}}_i^T\boldsymbol{\alpha}$. We denote $\boldsymbol{\alpha}$ to be the large sample limiting value of the parameter in the fitted model and let $\widehat{\boldsymbol{\alpha}}$ represent the data estimate of $\boldsymbol{\alpha}$.

1.2.2 Misspecification of the exposure effect may appear as heteroscedasticity

Almli et al. (2014) studied GxE interaction models in a post-traumatic stress disorder dataset and reported that the presence of heteroscedasticity was invalidating their inference. The authors found that the residual variance was a function of the environment, which led a QQ-plot to show heavily inflated p-values when performing genome-wide

interaction testing with the standard GxE model. We show in this section that for linear regression, misspecification of the exposure effect may cause the appearance of heteroscedasticity in the environment as reported by Almli et al. (2014).

Suppose the true linear GxE interaction model is given by

$$Y_i = \beta_0 + \beta_G G_i + \beta_{\mathcal{E}} f(\mathcal{E}_i) + \beta_I G_i h(\mathcal{E}_i) + \epsilon_i,$$

where $f(\cdot)$ and $h(\cdot)$ are known non-linear functions of $\mathcal{E}_i$, and $\epsilon_i \sim N(0, \sigma^2)$. Assume we fit the misspecified GxE interaction model with a linear effect of $\mathcal{E}_i$:

$$Y_i = \alpha_0 + \alpha_G G_i + \alpha_{\mathcal{E}} \mathcal{E}_i + \alpha_I G_i \mathcal{E}_i + e_i. \tag{1.3}$$

One can easily show that under the misspecified model (1.3),

$$E(e_i^2 | G_i, \mathcal{E}_i) = \sigma^2 + d(G_i, \mathcal{E}_i),$$

where $d(G_i, \mathcal{E}_i) = \{\mu_i(G_i, \mathcal{E}_i) - \mu_{i,mis}(G_i, \mathcal{E}_i)\}^2$, and

$$\begin{aligned} \mu_i(G_i, \mathcal{E}_i) &= \beta_0 + \beta_G G_i + \beta_{\mathcal{E}} f(\mathcal{E}_i) + \beta_I G_i h(\mathcal{E}_i) \\ \mu_{i,mis}(G_i, \mathcal{E}_i) &= \alpha_0 + \alpha_G G_i + \alpha_{\mathcal{E}} \mathcal{E}_i + \alpha_I G_i \mathcal{E}_i. \end{aligned}$$

If $f(\cdot)$ is not linear in $\mathcal{E}_i$, then the function $d(G_i, \mathcal{E}_i)$ is generally not 0 even under the null hypothesis $\beta_I = 0$. Thus there will appear to be heteroscedasticity with respect to the effect of the environment. This example suggests that it is possible GxE studies of continuous outcomes may misdiagnose exposure misspecification as heteroscedasticity; such studies may also find the following results relevant to their work.

1.3 Inference in the misspecified model

1.3.1 Asymptotic bias of fitted coefficients for identity link

Although our primary concern lies in testing α_I, it is often also of interest to estimate the parameters $(\alpha_0, \alpha_G, \alpha_{\mathcal{E}}, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T)$ for interpretability reasons or joint tests such as the 2-df test of $H_0 : \alpha_G = \alpha_I = 0$ proposed by Kraft et al. (2007) and utilized by Almli et al. (2014).

When $g(\cdot)$ is the identity link, the $p + r + 4$ score equations for estimating $(\alpha_0, \alpha_G, \alpha_{\mathcal{E}}, \alpha_I, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T)$ under the fitted model (1.2) are

$$\frac{1}{n\sigma^2}\sum_{i=1}^{n}(1, G_i, \mathcal{E}_i, G_i\mathcal{E}_i, \mathbf{Z}_i^T, \mathbf{W}_i^T)^T \left[Y_i - \widetilde{\mathbf{X}}_i^T\boldsymbol{\alpha}\right] = \mathbf{0}_{(p+r+4)\times 1}.$$

The asymptotic limit $\boldsymbol{\alpha}$ of the MLE $\widehat{\boldsymbol{\alpha}}$ is the value such that

$$E\left[(1, G, \mathcal{E}, G\mathcal{E}, \mathbf{Z}^T, \mathbf{W}^T)^T \left(\mathbf{X}\boldsymbol{\beta} - \widetilde{\mathbf{X}}\boldsymbol{\alpha}\right)\right] = \mathbf{0}_{(p+r+4)\times 1}. \tag{1.4}$$

Under distributional assumptions, it is possible to solve equation (1.4) in closed form and find the asymptotic bias of each fitted covariate. In derivations for this section, we will assume without loss of generality that the covariates G and $\mathcal{E}$ are centered at 0. Also, subscripts on μ will denote the expectation of those subscripts, so that $\mu_{G\mathcal{E}} = E(G\mathcal{E}) = \text{Cov}(G, \mathcal{E})$.

The asymptotic value of α_I takes the general form:

$$\alpha_I = \beta_{\mathcal{E}} * C_1 + \beta_I * C_2 + \mathbf{C}_3^T\boldsymbol{\beta}_M,$$

where C_1, C_2 denote constants and $\mathbf{C}_3$ denotes a $q \times 1$ vector of constants. These constants depend on the form of misspecification as well as the marginal and joint distribution of the covariates. α_G is similarly a complicated function of the true effect sizes. The full expansions are unwieldy and difficult to examine, so we leave them to Appendix 1.8.1. Under the null, we can perform valid inference on α_I if $C_1 = C_2 = 0$ and $\mathbf{C}_3 = \mathbf{0}_{q\times 1}$. The same is true for the constants relating to α_G.

In the following paragraphs, we briefly highlight some of the most interpretable consequences of the equations and describe the implications on GWEIS study design. For an arbitrary set of covariates and dependence structures, we offer an `R` package `GEint` that is able to calculate the exact magnitude of bias in fitted coefficients, given some inputs on the true model. This software offers a very flexible platform for users to analyze bias on a case-by-case basis, and it can also be used, for example, to perform thorough sensitivity analysis on GWEIS models.

- Consider first a simple testing case where only the environment term is misspecified in the fitted model, that is, $\mathbf{W} = \mathbf{M} = 0$. Under $H_0 : \beta_I = 0$, sufficient conditions for

$\alpha_I = 0$ are gene-environment independence combined with

$$\mu_{G\mathcal{E}Z_1} = ... = \mu_{G\mathcal{E}Z_p} = 0. \tag{1.5}$$

The sufficient conditions are achieved under gene-environment independence and if at least one of G or $\mathcal{E}$ is independent of each Z_j for all $j = 1, ..., p$.

- Additionally, under the joint null $H_0 : \beta_G = \beta_I = 0$, sufficient conditions for $\alpha_G = \alpha_I = 0$ are gene-environment independence combined with

$$\mu_{GZ_1} = ... = \mu_{GZ_p} = 0. \tag{1.6}$$

The sufficient condition (1.6) is achieved if G is independent of each Z_j for all $j = 1, ...p$, which is much more stringent. This result suggests the joint test is much more susceptible to issues of bias due to model misspecification.

- Next consider the case where other covariates are also misspecified, so that $\mathbf{W} \neq \mathbf{M}$. Under the null $H_0 : \beta_I = 0$, two sufficient conditions for $\alpha_I = 0$ are gene-environment independence combined with

$$\mu_{G\mathcal{E}Z_1} = ... = \mu_{G\mathcal{E}Z_p} = \mu_{G\mathcal{E}M_1} = ... = \mu_{G\mathcal{E}M_q} = \mu_{G\mathcal{E}W_1} = ... = \mu_{G\mathcal{E}W_r} = 0. \tag{1.7}$$

The sufficient condition (1.7) is achieved if at least one of $\mathcal{E}$ or G is independent of each $Z_1...Z_p$, each $M_1...M_q$, and each $W_1...W_r$ (in addition to gene-environment independence). The result of Rosenblum and van der Laan (2009) is a special case of this result.

- Under the joint null $H_0 : \beta_G = \beta_I = 0$, two sufficient conditions for $\alpha_G = \alpha_I = 0$ are gene-environment independence combined with

$$\mu_{GZ_1} = ... = \mu_{GZ_p} = \mu_{GM_1} = ... = \mu_{GM_q} = \mu_{GW_1} = ... = \mu_{GW_r} = 0. \tag{1.8}$$

The sufficient condition (1.8) is achieved if G is independent of each $Z_1, ..., Z_p$, each $M_1, ..., M_q$, and each $W_1, ..., W_r$.

The scenarios discussed above suggest that when genetic and environmental covariates are independent, GWEIS inference is likely to be more robust to model misspecification. When G and $\mathcal{E}$ are dependent and the effect of $\mathcal{E}$ is misspecified, the estimate of the interaction term will often be asymptotically biased. In addition, the results suggest that introducing many precision covariates into the model, for instance to reduce the standard error of estimated coefficients, is likely to increase the chance of model misspecification and cause biased inference on GxE interactions. It is desirable to have more parsimonious models for GxE studies, as they are likely to contain fewer conditions that must be satisfied for valid inference on the interaction term.

1.3.2 Controlling for the confounding effects of population stratification

Population stratification due to heterogeneous populations is common in genome-wide association studies and is routinely adjusted for by introducing genetic principal components into the model as covariates. In the presence of population stratification and use of principal components, the results in (1.7) suggest that if the environmental exposure varies with sub-populations, misspecification of the exposure effects is likely to result in biased inference on the GxE interaction. For instance, if a study cohort is composed of Northern and Southern Europeans, and the exposure of interest is differentiated between these two sub-population groups, neither the environmental term nor the genetic term will be independent of the principal components covariates. Then inference on α_I is likely to be sensitive to misspecification of the exposure effects.

1.3.3 Asymptotic bias of fitted coefficients for logistic link

For binary outcomes and a logistic regression model, the score equations become:

$$0 \;=\; \frac{1}{n}\sum_{i=1}^{n}(1, G_i, \mathcal{E}_i, G_i\mathcal{E}_i, \mathbf{Z}_i^T, \mathbf{W}_i^T)^T \left\{Y_i - \mu_i(\widetilde{\mathbf{X}}_i^T\boldsymbol{\alpha})\right\}, \tag{1.9}$$

where $\mu(x) = g^{-1}(x) = \exp(x)/\{1+\exp(x)\}$. The asymptotic limit $\boldsymbol{\alpha}$ is the value such that

$$E\left[(1, G, \mathcal{E}, G\mathcal{E}, \mathbf{Z}^T, \mathbf{W}^T)^T \left\{\mathbf{Y} - \mu(\widetilde{\mathbf{X}}\boldsymbol{\alpha})\right\}\right] = \mathbf{0}. \tag{1.10}$$

These equations generally do not have closed forms. Rosenblum and van der Laan (2009) and Tchetgen and Kraft (2011) studied the case where $\beta_G = \beta_I = 0$ and G is independent of all other terms in the true model (1.1). The authors showed that in this setting, $\alpha_G = \alpha_I = 0$ even under environment misspecification.

Here we focus on situations where the independence assumptions do not hold. We perform asymptotic bias calculations by numerically solving (1.10) for specific cases to demonstrate that when the Rosenblum and van der Laan conditions are not met, α_I will likely be biased. That is, if G has some association with Y and the effect of the environment is misspecified, then α_I is generally biased away from 0 under the null.

Figure 1.1 illustrates the asymptotic bias in the interaction term under four different misspecification scenarios. To be completely clear, we assume in these scenarios and for the rest of this section that the main effect of the exposure exists in the true model and has been misspecified in the fitted model. In all cases, we solve the asymptotic score equations (1.10) using numerical methods. For each setting we assume G has a Binomial(2,0.3) distribution, and it is correlated with the underlined variable by an amount given on the x-axis. The variable correlated with G is assumed to be a mixture of normal random variables, with mean conditional on G, and it has marginal mean 0 and variance 1. In scenarios 2, 3, and 4, the environment term is independently generated as a standard normal random variable. In scenario 3, $M = W^2$ provides additional misspecification.

We see that in scenario 1, the interaction coefficient is biased because G is associated with $\mathcal{E}$, which is in the true model. Thus when there is no gene-environment independence, the interaction coefficient will be biased. In scenarios 2 and 3, the interaction coefficient is biased because G is indirectly associated with Y through correlation with Z and W respectively. Thus if the true model includes principal components to control for population stratification, α_I will be biased. In scenario 4, G is correlated with W, but W has no association with terms in the true model, so there is no bias.

These four scenarios cover a wide range of possibilities, and they show that the estimate of the interaction term is generally biased under environment misspecification if the genetic term is directly or indirectly associated with the outcome. In Appendix 1.8.2 we are able to provide some more intuition on how bias arises in the simplest situations

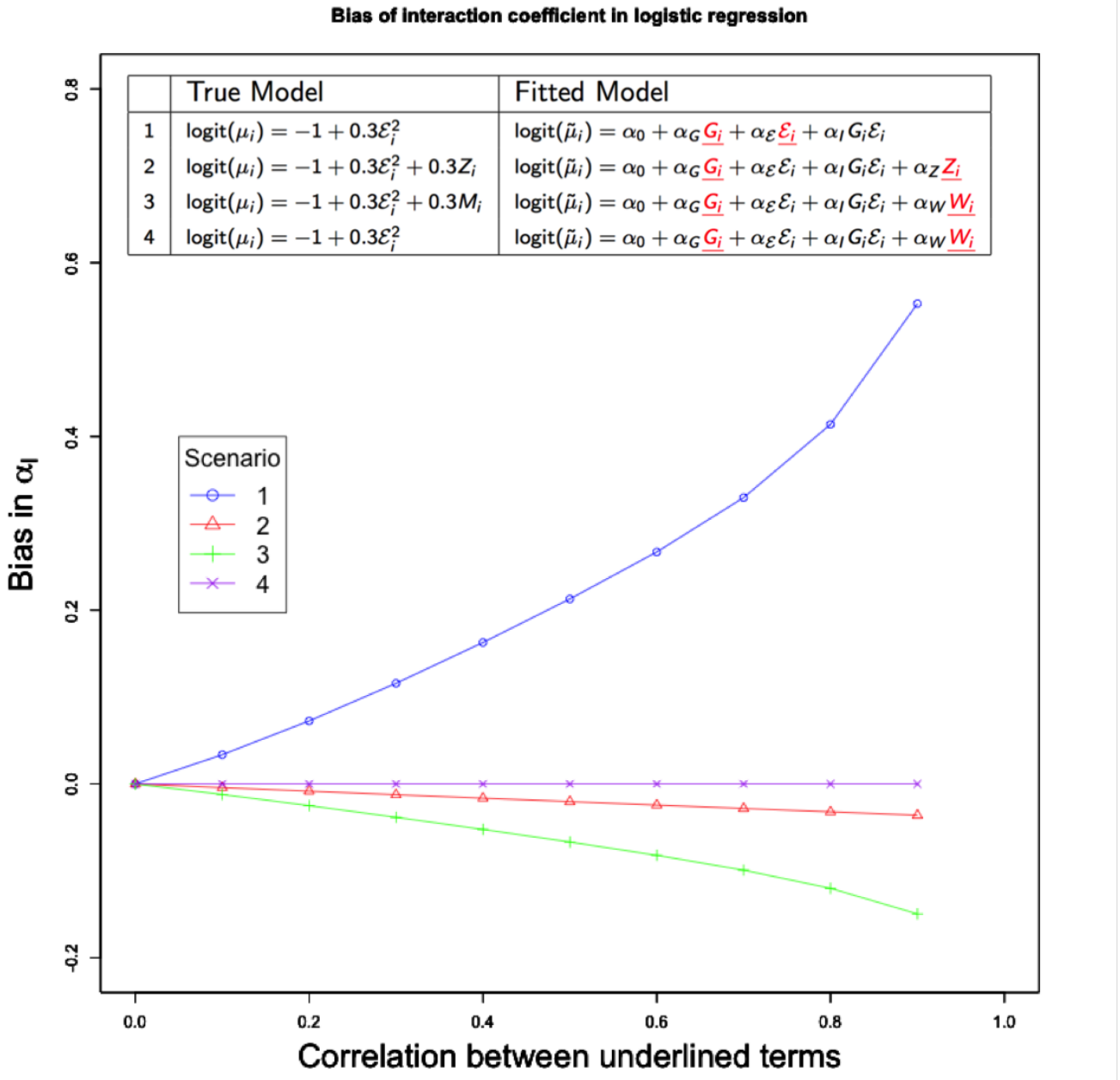

Figure 1.1: Bias of the fitted interaction coefficient in logistic regression over four different misspecification settings. The underlined terms are correlated, with the magnitude of correlation given on the x-axis. At each data point we solved the score equations numerically and confirmed these results through simulation.

where there are no additional covariates. If G is not directly or indirectly associated with the outcome Y through correlation with other terms in the true model, then the interaction regression coefficient will be asymptotically unbiased under the null.

1.3.4 Asymptotic standard error of fitted coefficients

Most earlier works on model misspecification (Rosenblum and van der Laan, 2009; Voorman et al., 2011; Tchetgen and Kraft, 2011; Almli et al., 2014) advocate that using a robust sandwich standard error estimate will provide asymptotically correct Type I error when α_I is unbiased under the null. The same theory holds for the models we study, because the asymptotic covariance matrix of $\widehat{\alpha}$ is given by:

$$\begin{aligned} V_{\widehat{\alpha}} &= B(\boldsymbol{\alpha})^{-1} A(\boldsymbol{\alpha}) \left\{ B(\boldsymbol{\alpha})^{-1}) \right\}^T; \\ B(\boldsymbol{\alpha}) &= E\left\{ \frac{\partial \psi(\widetilde{\mathbf{X}}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} \right\}, A(\boldsymbol{\alpha}) = E\left\{ \psi(\widetilde{\mathbf{X}}, \boldsymbol{\alpha}) \psi(\widetilde{\mathbf{X}}, \alpha)^T \right\} \end{aligned}$$

where $\psi(\widetilde{\mathbf{X}}, \boldsymbol{\alpha})$ are the $p+r+4$ score equations from above. The model-based variance estimator assumes that $B(\boldsymbol{\alpha}) = -A(\boldsymbol{\alpha})$, which is incorrect under exposure misspecification and will invalidate the inference even if the regression coefficient estimate is unbiased.

Denote by $\widehat{m}_{\widehat{\alpha}_I}$ the model-based standard error estimate of $\widehat{\alpha}_I$. We show in Appendix 1.8.3 that the model-based Wald statistic for testing the interaction term $T_{mod} = (\widehat{\alpha}_I / \widehat{m}_{\widehat{\alpha}_I})^2$ asymptotically follows a scaled chi-square distribution $c\chi^2_1$, where the expressions of c for linear and logistic regression are given in that appendix. If $c > 1$ for many SNPs across the genome, then the QQ-plot for GxE interactions using model-based standard errors will be inflated. If $c < 1$ for many SNPs, then the QQ-plot will be deflated. The value of c is determined both by the true model and the design matrix of fitted coefficients.

Using a sandwich estimator with $\widehat{\boldsymbol{\alpha}}$ instead of $\boldsymbol{\alpha}$ in the above expression will give a consistent variance estimate. However, when we utilized this strategy in simulation and on our actual dataset, the p-values calculated with the robust standard error actually appeared less uniform than those p-values calculated with the model-based standard error (as shown in Figure 1.2).

1.4 Alternative standard error estimates

1.4.1 Inflation caused by the sandwich estimator

Even though many studies suggest using the robust sandwich variance estimator, the Superfund data ($n \approx 400$ and $n \approx 500$), the study of Almli et al. ($n > 3000$), and the analysis of Cornelius et al. ($n > 5000$) are a few examples where inference conducted with the sandwich estimator appears to return an excess of highly significant p-values. It is known that the sandwich estimator is often biased downwards and is more variable than model-based estimators (Kauermann and Carroll, 2001) even when the model is not misspecified, and these characteristics can cause inflated Type I error in hypothesis testing (Kauermann and Carroll, 2001).

Exposure misspecification can exacerbate the variability of the sandwich estimator in linear regression. This occurs because the sandwich estimator is a linear combination of the squared regression residuals, and the squared regression residuals have more variance under exposure misspecification. We demonstrate in detail in Appendix 1.8.4 how the variance of the sandwich estimator can be much larger under model misspecification than when the model is correctly specified. The natural downward bias of the sandwich estimator as well the additional variability caused by exposure misspecification provide intuition for the heavily inflated sandwich p-values seen in the Superfund data.

A similar derivation incorporating residual variability in logistic regression is complicated by the difficulty of specifying a distribution for the residuals. However, in our simulations, we find that testing for binary outcomes with the sandwich standard error can have slightly incorrect size as well. Thus it is of interest to find variance estimators which can better protect the level of the test when performing inference under exposure misspecification.

1.4.2 Bootstrap Inference with Corrected Sandwich

As an alternative to the model-based and sandwich variance estimators, we propose a resampling-based method. The proposed method can be thought of as a finite sample correction to the sandwich. Denote by $T_{sand} = (\widehat{\alpha}_I / \widehat{s}_{\widehat{\alpha}_I})^2$ a test statistic for the interac-

tion effect calculated using the sandwich standard error estimate $\widehat{s}_{\widehat{\alpha}_I}$. This test statistic should asymptotically have a χ^2_1 distribution under the null. If the sandwich estimator is biased in finite samples, then the bias will cause the test statistic to instead have an approximately scaled chi-square distribution: $T_{sand} \approx c\chi^2_1$. We can approximate the $c\chi^2_1$ distribution by resampling the test statistic and matching the moments of its sampling distribution with a Satterthwaite-type idea as folows:

Fit model (2) on the observed data to find the estimated interaction coefficient $\widehat{\alpha}_I^{(init)}$ and sandwich test statistic $T_{sand}^{(init)}$. For each of $b = 1, 2, ..., B$, say $B = 1000$, bootstrap iterations, perform a nonparametric bootstrap by sampling $(Y_i, \widetilde{\mathbf{X}}_i)$ from the original data n times with replacement. Fit model (2) on the new sample. Calculate the squared, centered bootstrap test statistics $T_{sand}^{(b)} = \{(\widehat{\alpha}_I^{(b)} - \widehat{\alpha}_I^{(init)})/\widehat{s}_{\widehat{\alpha}_I}^{(b)}\}^2$ where $\widehat{\alpha}_I^{(b)}$ and $\widehat{s}_{\widehat{\alpha}_I}^{(b)}$ are the regression coefficient and the sandwich standard error estimate for the interaction term based on the bth bootstrap sample. Match the mean and variance of $\mathbf{T} = \left(T_{sand}^{(1)}, ..., T_{sand}^{(B)}\right)$ to the moments of a $k\chi^2_a$ distribution, where we solve for k, a using the equations $k = \text{Var}(\mathbf{T})/\{2 * \text{Mean}(\mathbf{T})\}$ and $a = \text{Mean}(\mathbf{T})/k$. Find the p-value of the original test statistic $T_{sand}^{(init)}$ using $k\chi^2_a$ as the reference distribution. We will refer to this method as the Bootstrap Inference with Corrected Sandwich (BICS) procedure. We would also like to note that a natural alternative, using the empirical standard error of $\widehat{\alpha}_I^{(1)}, ..., \widehat{\alpha}_I^{(B)}$ instead of the sandwich standard error, does not work well.

1.5 Simulation studies

We conduct a variety of simulations to evaluate control of Type I error rate in GWEIS for different testing procedures over a range of misspecification scenarios. All misspecified models we consider are generated under the null of $\beta_I = 0$, and all satisfy the conditions for valid inference discussed previously, that is, $\alpha_I = 0$ asymptotically. The Type I error rate of the tests should be controlled at the nominal size of 0.05 with an unbiased standard error estimator. In all simulations we fit the model (1.2) with $\mathbf{Z}_i = \mathbf{W}_i = 0$. We use a Wald t-test to generate p-values with the naive and sandwich standard errors. Each misspecified model is tested at sample sizes of 400, 800, and 1600 to reflect the finite

sample problem which affects the Superfund study. We perform 50000 replications of the simulation at each parameter setting and report the percentage of times that each testing procedure rejects the null.

We first describe the misspecification for continuous outcomes. Simulation A has outcome Y generated from the model $Y_i = \beta_{\mathcal{E}}\mathcal{E}_i^2 + \epsilon_i, \epsilon_i \sim N(0,1)$ where $\beta_{\mathcal{E}}$ is chosen such that $\mathcal{E}$ explains 10% of the variance in Y. In Simulation B we increase the degree of misspecification by taking the true model to be $Y_i = \beta_{\mathcal{E}}\mathcal{E}_i^3 + \epsilon_i, \epsilon_i \sim N(0,1)$, where $\beta_{\mathcal{E}}$ is again chosen such that $\mathcal{E}$ explains 10% of the variance in Y. For both Simulations A and B, we generate $\mathcal{E}_i \sim N(1,1)$. Simulations C and D have the same true model as A and B, except we generate $\mathcal{E}_i \sim$ Beta(2,5) to introduce skewness into the exposure variable. We also adjust $\beta_{\mathcal{E}}$ so that $\mathcal{E}$ continues to explain 10% of the variance in Y. Finally, Simulation E differs from the previous four in that we generate the outcome as $Y_i = \beta_G G_i + \beta_{\mathcal{E}}\mathcal{E}_i^2 + \epsilon_i$, with $\mathcal{E}_i$ and ϵ_i again as they were in Simulation A. This situation mimics testing for interaction with a SNP that has a marginal effect but no interaction effect. The values of β_G and $\beta_{\mathcal{E}}$ are chosen such that $\mathcal{E}$ and G would explain 10% and 1% of the variance in Y respectively if G had minor allele frequency 0.3. For all scenarios above, G is simulated by using HAPGEN2 to generate the number of minor alleles at a random SNP on chromosome 1 (HapMap3 CEU population used as reference), thus G and $\mathcal{E}$ are always independent.

We see from Table 1 that the sandwich estimator constantly produces inflated Type I error rates. BICS performs very well, protecting the size almost exactly in every single situation. Of course, the sandwich estimator performs progressively better as the sample size increases. In contrast, BICS does not appear to show a trend in n and increases its relative superiority over the sandwich estimator at the smallest sample sizes. The naive estimator is always biased and shows the most inflation. These results closely reflect the trends in our data example, where QQ-plots of p-values calculated with the sandwich and naive estimators show very early departures from the 45-degree line, indicating lack of uniformity.

Next we consider binary outcomes. Simulations F,G,H,I, and J are conducted in the same spirit as the previous five. The outcome Y_i in simulation F is generated from the

Table 1.1: Type I error rate at level $\alpha = 0.05$ when testing $H_0 : \alpha_I = 0$ with naive model-based standard error (N), sandwich standard error (S), and Bootstrap Inference with Corrected Sandwich (BICS). Simulation parameters A-E correspond to continuous outcomes and F-J to binary outcomes. Inference with the sandwich estimator shows inflated Type I error rates for continuous outcomes. Inference with the naive estimator shows inflated Type I error rates for continuous outcomes and deflated Type I error rates for binary outcomes. The BICS procedure protects the Type I error rate in all simulations.

Simulation (n)	**Continuous Outcome**			**Binary Outcome**		
	N	S	BICS	N	S	BICS
A/F						
(400)	0.073	0.063	0.052	0.037	0.047	0.050
(800)	0.077	0.060	0.053	0.039	0.048	0.051
(1600)	0.074	0.053	0.049	0.038	0.049	0.051
B/G						
(400)	0.086	0.065	0.052	0.037	0.049	0.052
(800)	0.088	0.058	0.050	0.037	0.048	0.050
(1600)	0.089	0.055	0.050	0.037	0.050	0.051
C/H						
(400)	0.051	0.059	0.050	0.049	0.050	0.053
(800)	0.052	0.055	0.050	0.049	0.050	0.052
(1600)	0.054	0.053	0.051	0.050	0.051	0.052
D/I						
(400)	0.065	0.063	0.052	0.048	0.049	0.052
(800)	0.064	0.056	0.050	0.049	0.050	0.052
(1600)	0.065	0.055	0.051	0.048	0.048	0.050
E/J						
(400)	0.074	0.065	0.053	0.035	0.048	0.051
(800)	0.073	0.056	0.049	0.037	0.049	0.051
(1600)	0.076	0.054	0.051	0.036	0.049	0.051

model:

$$Y_i \sim \text{Bernoulli}(\pi_i);\ \pi_i = \frac{\exp(0.4\mathcal{E}_i^2)}{1+\exp(0.4\mathcal{E}_i^2)}.$$

The parameter $\beta_0 = 0$ is chosen to give a subject with $\mathcal{E} = 0$ a disease probability of 0.5. Simulation G is conducted under a higher degree of misspecification as we take the true probability of disease to be $\pi_i = \exp(0.2\mathcal{E}_i^3) / \{1 + \exp(0.2\mathcal{E}_i^3)\}$. For simulations F and G we generate $\mathcal{E}_i \sim N(0,1)$. Simulations H and I have $\pi_i = \exp(\mathcal{E}_i^2) / \{1 + \exp(\mathcal{E}_i^2)\}$ and $\pi_i = \exp(\mathcal{E}_i^3) / \{1 + \exp(\mathcal{E}_i^3)\}$ respectively with $\mathcal{E}_i \sim$ Beta(2,5). Finally, in Simulation J each SNP has a marginal effect with $\pi_i = \exp(0.1G_i + 0.4\mathcal{E}_i^2) / \{1 + \exp(0.1G_i + 0.4\mathcal{E}_i^2)\}$, and $\mathcal{E}_i \sim N(0,1)$ again.

In these logistic regression simulations we see that the sandwich estimator actually performs fairly well, with the correct size in most situations. It can be slightly conservative when $n = 400$. BICS similarly performs well, although it appears to be slightly less conservative than using the sandwich estimate. In absolute terms, BICS and the sandwich estimator both appear to deviate a similar amount from the expected size. Once again the naive standard error estimate is biased and produces tests at the incorrect size.

Based on the results of our simulation study, we recommend that our resampling methods be used quite widely in linear regression GWEIS of moderate sample sizes, as exposure misspecification is likely to occur to some degree when testing for GxE interaction. A simple and fast implementation is available through `GEint`. When n is large or logistic regression is used, we agree with previous suggestions that the sandwich estimator should be employed for its speed and simplicity, however BICS can be used as an alternative if diagnostic QQ-plots appear worrisome. In a practical setting, the environment term will remain constant for each SNP, while in our simulation the environment term is newly generated with each different SNP. This choice was made to present the fairest possible comparison in simulation. When the environment term is held constant for each SNP, the difference between BICS and the sandwich estimator can be even more drastic (again see Figure 1.2).

1.6 Application to Superfund data

One major goal of the Superfund Research Program is to study the interplay of genes and toxic metal exposures on childhood neurological outcomes. The metal exposure of interest is lead concentration in the umbilical cord blood. The neurological outcome is Mental Development Index (MDI) score from the BSID.

There exists evidence that high levels of exposure to certain metals during the prenatal period can seriously impair the cognitive development of infants (Claus Henn et al., 2012), but to date we are unaware of any previous gene-environment interaction studies covering toxic metal exposures and neurodevelopment outcomes.

The participants enrolled in the study come from two cohorts. Recruitment in Mexico was described in Burris et al. (2013), with 389 of the recruited mother-infant pairs having complete genetic and covariate data. Recruitment in Bangladesh was described in Kile et al. (2014), with 497 of these mother-infant pairs having complete genetic and covariate data. Briefly, women were enrolled during hospital visits in the early weeks of their pregnancy, and covariate information was collected upon subsequent visits to the hospital. Genotyping was performed using the Illumina OmniExpressExome-8 in the Bangladesh cohort and the Illumina HumanOmni1-Quad Beadchip in the Mexico cohort. About 500,000 SNPs common to both cohorts remained after quality control.

We conducted a standard GWEIS by repeatedly fitting the model

$$Y_i = \alpha_0 + \alpha_G G_i + \alpha_{\mathcal{E}} \mathcal{E}_i + \alpha_I G_i \mathcal{E}_i + \mathbf{Z}_i^T \boldsymbol{\alpha}_Z + \epsilon_i, \tag{1.11}$$

where $\mathbf{Z}_i$ is an 8×1 vector of additional covariates including sex, birthweight, gestational age, education of mother (binary, 1 if primary school or greater), household environmental smoke (binary), child's age at time of assessment, and the first two genotype principal component vectors. Here $\mathcal{E}$ is the logarithm of umbilical cord blood lead concentration. Two distinct genome-wide scans were conducted, one for each cohort. A meta-analysis was then performed to pool the data, and the meta-analysis was examined for SNPs with highly significant p-values.

The initial analyses implemented with a model-based standard error produced QQ-

plots of highly non-uniform p-values (Figure 1.2 and Figure 1.3). We conjectured that a major cause of the non-uniformity was misspecification of the effect of the environmental covariate. To investigate possible misspecification, we repeated the initial analyses but introduced a spline term for the environment instead of modeling it linearly. QQ-plots produced after this modification improved somewhat but still showed some non-uniformity. We also performed a standard GWAS by removing the interaction term from the fitted model and only testing for the marginal effect of G. QQ-plots for the GWAS seemed relatively uniform.

Under model misspecification, the theoretical results derived in Section 1.3 suggest that we can have robust tests of the null hypothesis under some independence conditions, which we believe are reasonable to assume here. However, as shown in Figure 1.2, the QQ-plot based on sandwich standard errors is inflated.

We next re-analyze the data by fitting model (1.11) and using the resampling-based methods to generate p-values for the cohort-specific GWEIS. These resampling-based p-values are much more uniform than p-values calculated using the naive or sandwich variance estimate. It appears that our assumptions about independence mostly hold, as there is little inflation using BICS. The corrected p-values seem to reflect that $\alpha_I = 0$ throughout much of the genome, and inflation seen from using the sandwich estimate can likely be attributed to the drawbacks discussed in Section 1.4. For the meta-analysis, we use METAL (Willer et al., 2010) to perform a sample size based analysis on the corrected p-values, and again the corrected meta-analysis p-values seem uniformly distributed.

After applying BICS and accounting for multiple testing, we do not find any SNPs to be significant at the genome-wide level in either of the cohorts. No SNPs reach genome-wide significance in the meta-analysis either. However, the meta-analysis does suggest a promising region for future study. Two of the top SNPs identified in the meta-analysis are rs9642758 and rs10503970 (p-values of 8.79×10^{-6} and 2.57×10^{-5} respectively), which are both located on chromosome 8 in the region of the gene UNC5D. UNC5D encodes a receptor for netrin, which may be involved in axon guidance and could plausibly affect infant neurodevelopment through interaction with toxic metals. We believe the interaction between UNC5D, exposure to lead, and neurodevelopment outcomes is a promising

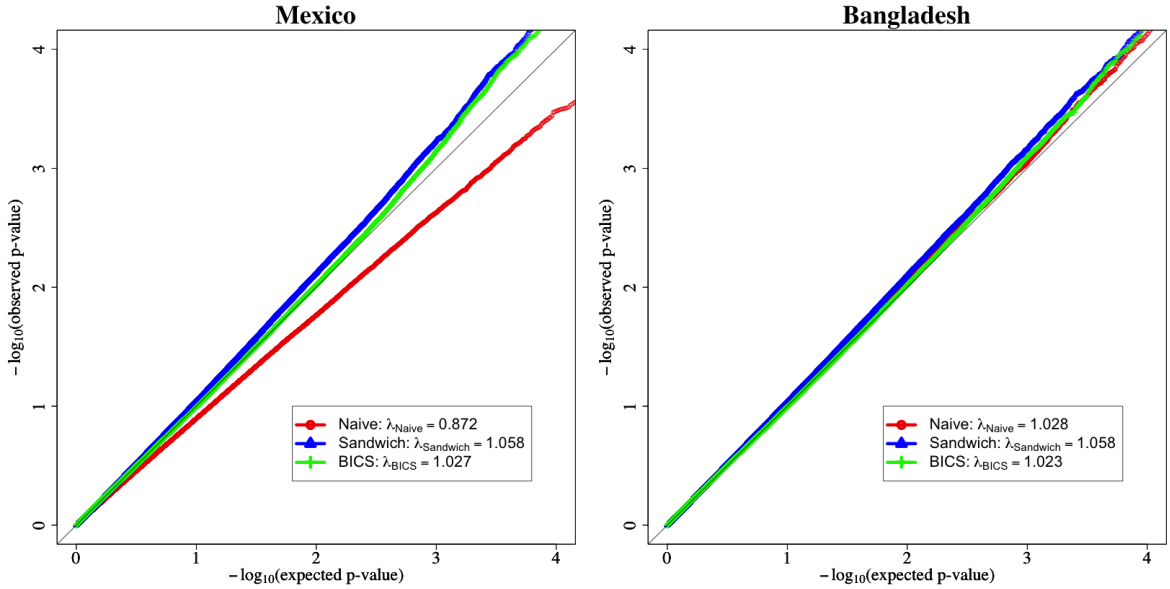

Figure 1.2: QQ-plots of p-values generated by testing for interaction effect in Model (1.11) with naive model-based variance, sandwich variance, and BICS procedure. The outcome is BSID Mental Development Index score. The exposure is logarithm of umbilical cord blood lead concentration. On the left side is the Mexico cohort and on the right side is the Bangladesh cohort. A uniform distribution of p-values would adhere very closely to the 45-degree line; we expect this outcome as we believe an overwhelming majority of our tests should be conducted under the null hypothesis. However in both cohorts the sandwich and naive p-values show very early departures from the line. Such behavior is worrisome because it indicates the inference procedure is not producing uniform p-values under the null, and thus all inferences we make may be invalidated. A quantitative measure of the departure from uniformity is given by the genomic inflation factor, provided in the legend. This factor is defined as the ratio of the median of the empirically observed test statistics to the expected median of a chi-squared distribution with one degree of freedom.

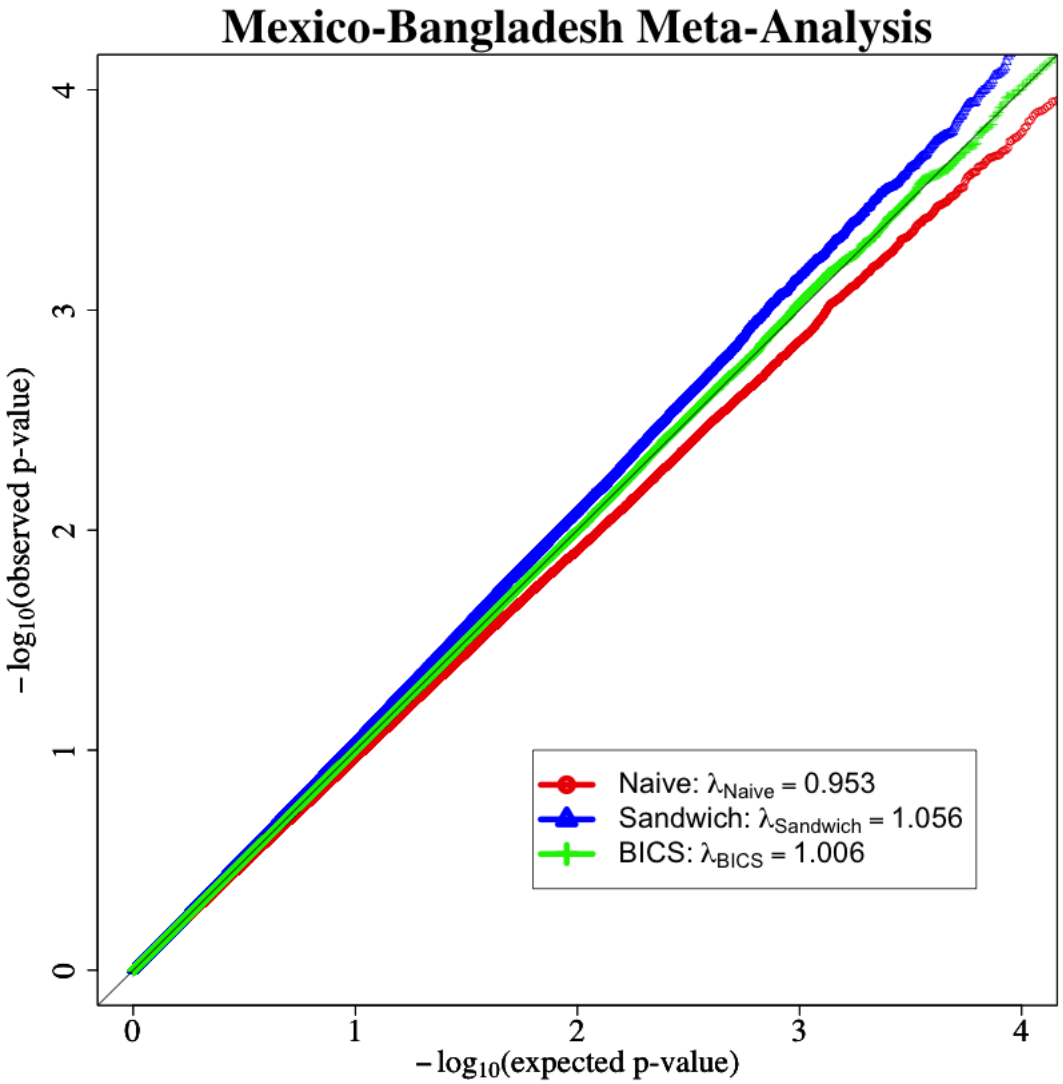

Figure 1.3: QQplots of p-values generated from the meta-analysis of the p-values shown in Figure 1.2. We use METAL to perform the meta-analysis by combining weighted p-values from the Bangladesh and Mexico cohorts. The default genomic control option in METAL is turned off. Again we see the p-values calculated using the sandwich and naive variance estimates depart from the 45-degree line very early.

candidate for further study.

1.7 Discussion

It is often the case in the standard GWEIS approach that the environment covariate is likely to be misspecified. We have demonstrated conditions under which inference for the interaction effect is still valid under model misspecification. These results provide guidance on fitting GxE interaction models. We show that for linear regression models, the estimate of the interaction effect will be asymptotically unbiased under gene-environment independence and if either the genetic or environment term is independent of each other true and fitted coefficient in the model. For logistic regression models, the estimate of the interaction effect will generally only be asymptotically unbiased if the genetic term is neither directly nor indirectly associated with the outcome.

When the conditions for valid inference on GxE interactions are met, hypothesis testing may still be difficult to conduct because the model-based estimate of standard error is biased under environment misspecification, and the Huber-White sandwich estimator can lead to excess Type I error. We provide a resampling-based method of obtaining p-values and show its advantages both in simulation and through application to the Superfund dataset. We recommend BICS be widely used in linear regression GWEIS with moderate sample sizes. After reanalysis of the Superfund data, we have identified UNC5D as a strong candidate gene for further study in how lead exposure can affect infant neurodevelopment.

While our resampling method can be computed rather quickly and has been found to work well in practical studies of moderate sample size, it is still a minor drawback to perform a bootstrap procedure for every SNP across the genome. It is of future research interest to develop more computationally efficient and robust inference methods for testing GxE interactions in GWEIS. In addition, as more and more GWEIS are conducted, it will be necessary to develop semiparametric gene-environment interaction models that are more robust to model misspecification in both the exposure and confounder covariates.

1.8 Appendix

1.8.1 Solutions to asymptotic score equations for linear regression from Section 1.3.1

Here we present the unsimplified versions of the asymptotic values of $(\alpha_0, \alpha_G, \alpha_\mathcal{E}, \alpha_I, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T)$.

First, a few comments on our notation. Remember that $\mathbf{Z} = (Z_1, Z_2, ..., Z_p)$ and $\mathbf{W} = (W_1, W_2, ..., W_r)$ are vectors, so when we write an expression of the form $\mu_{G\mathbf{Z}}$, we mean the $p \times 1$ vector $\left(\mu_{GZ_1}, \mu_{GZ_2}, ..., \mu_{GZ_p}\right)^T$. Sometimes we will also use expressions of the form $\mu_{\mathbf{Z}^T\mathbf{Z}}$, which refers to the matrix of expectations where the (1,1) element is $E(Z_1^2)$, the (1,2) and (2,1) elements are $E(Z_1Z_2)$, and so on. This is distinct from $\mu_\mathbf{Z}^T\mu_\mathbf{Z}$, which corresponds to taking the expectations of the vectors first, then multiplying to form a matrix - i.e. the (1,1) element is $E(Z_1)^2$. Finally, whenever we add a scalar term to a vector, what we mean is to add that scalar term to each element of the vector.

For sake of presentation, we provide the solutions in nested form, so that the solution to α_0 depends on $(\alpha_G, \alpha_\mathcal{E}, \alpha_I, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T)$, the solution to α_G depends on $(\alpha_\mathcal{E}, \alpha_I, \boldsymbol{\alpha}_Z^T, \boldsymbol{\alpha}_W^T)$, and so on. Recall that we assumed the genetic and environment covariates are centered at 0, so μ_G and $\mu_\mathcal{E}$ will not appear. We use f for $f(\mathcal{E})$ and h for $h(\mathcal{E})$ so that $\mu_{Gf} = E\{Gf(\mathcal{E})\}$ and $\mu_{Gh} = E\{Gh(\mathcal{E})\}$.

The solution to α_0 is:

$$\alpha_0 = \beta_0 + \beta_\mathcal{E}\mu_f + \beta_I\mu_{Gh} - \alpha_I\mu_{G\mathcal{E}} + \left(\boldsymbol{\beta}_\mathbf{Z} - \boldsymbol{\alpha}_\mathbf{Z}\right)^T\mu_\mathbf{Z} + \boldsymbol{\beta}_\mathbf{M}^T\mu_\mathbf{M} - \boldsymbol{\alpha}_\mathbf{W}^T\mu_\mathbf{W}.$$

The solution to α_G is:

$$\alpha_G = \beta_G - \left(\alpha_\mathcal{E}\mu_{G\mathcal{E}} - \beta_\mathcal{E}\mu_{Gf} + \alpha_I\mu_{G^2\mathcal{E}} - \beta_I\mu_{G^2h} + (\boldsymbol{\alpha}_\mathbf{Z} - \boldsymbol{\beta}_\mathbf{Z})^T\mu_{G\mathbf{Z}} + \boldsymbol{\alpha}_\mathbf{W}^T\mu_{G\mathbf{W}} - \boldsymbol{\beta}_\mathbf{M}^T\mu_{G\mathbf{M}}\right)/\mu_{G^2}.$$

The solution to $\alpha_\mathcal{E}$ is:

$$\begin{aligned}\alpha_\mathcal{E} \;=\; & \{\beta_\mathcal{E}\left(\mu_{\mathcal{E}f} - \mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right) + \beta_I\left(\mu_{G\mathcal{E}h} - \mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right) + \alpha_I\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{G\mathcal{E}^2}\right) \\ & + (\boldsymbol{\beta}_\mathbf{Z} - \boldsymbol{\alpha}_\mathbf{Z})^T\left(\mu_{\mathcal{E}\mathbf{Z}} - \mu_{G\mathbf{Z}}\mu_{G\mathcal{E}}/\mu_{G^2}\right) + \boldsymbol{\beta}_\mathbf{M}^T\left(\mu_{\mathcal{E}\mathbf{M}} - \mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right) + \boldsymbol{\alpha}_\mathbf{W}^T\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{\mathcal{E}\mathbf{W}}\right)\} \\ & /(\mu_{\mathcal{E}^2} - \mu_{G\mathcal{E}}^2/\mu_{G^2}).\end{aligned}$$

Next we want to solve for $\boldsymbol{\alpha}_Z$. Since $\boldsymbol{\alpha}_Z$ is a vector, to solve for it we have p equations of the form

$$\begin{pmatrix} P_1 \\ . \\ . \\ . \\ P_p \end{pmatrix} = \begin{pmatrix} O_{11} & . & . & . & O_{1p} \\ . & & & & \\ . & & & & \\ . & & & & \\ O_{p1} & . & . & . & O_{pp} \end{pmatrix} \begin{pmatrix} \beta_{Z_1} - \alpha_{Z_1} \\ . \\ . \\ . \\ \beta_{Z_p} - \alpha_{Z_p} \end{pmatrix}$$

where we will designate the vector of constants on the left as $\mathbf{P}$ and the $p \times p$ matrix as $\mathbf{O}$. Then to solve the equations we have $\boldsymbol{\alpha}_Z = \boldsymbol{\beta}_Z - \mathbf{O}^{-1}\mathbf{P}$. We will assume that $\mathbf{P}$ is invertible - clearly if it is singular then we cannot solve for $\boldsymbol{\alpha}_Z$. The forms of $\mathbf{O}$ and $\mathbf{P}$ are:

$$\begin{aligned}
\mathbf{P} &= \beta_{\mathcal{E}} \left\{ -\mu_f \mu_{\mathbf{Z}} - \mu_{Gf}\mu_{G\mathbf{Z}}/\mu_{G^2} + \mu_{f\mathbf{Z}} + \mathbf{A}\left(\mu_{\mathcal{E}f} - \mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \beta_I \left\{ \mu_{Gh\mathbf{Z}} - \mu_{Gh}\mu_{\mathbf{Z}} - \mu_{G^2h}\mu_{G\mathbf{Z}}/\mu_{G^2} + \mathbf{A}\left(\mu_{G\mathcal{E}h} - \mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \alpha_I \left\{ \mu_{G\mathcal{E}}\mu_{\mathbf{Z}} + \mu_{G^2\mathcal{E}}\mu_{G\mathbf{Z}}/\mu_{G^2} - \mu_{G\mathcal{E}\mathbf{Z}} + \mathbf{A}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{G\mathcal{E}^2}\right)\right\} \\
&\quad + \left\{ \mathbf{A}\left(\mu_{\mathcal{E}\mathbf{M}} - \mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T - \mu_{\mathbf{Z}}\mu_{\mathbf{M}}^T - \mu_{G\mathbf{Z}}\mu_{G\mathbf{M}}^T/\mu_{G^2} + \mu_{\mathbf{Z}\mathbf{M}^T} \right\} \boldsymbol{\beta}_{\mathbf{M}} \\
&\quad + \left\{ \mathbf{A}\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{\mathcal{E}\mathbf{W}}\right)^T + \mu_{\mathbf{Z}}\mu_{\mathbf{W}}^T + \mu_{G\mathbf{Z}}\mu_{G\mathbf{W}}^T/\mu_{G^2} - \mu_{\mathbf{Z}\mathbf{W}^T} \right\} \boldsymbol{\alpha}_{\mathbf{W}} \\
\mathbf{O} &= \left\{ \left(\mu_{\mathbf{Z}}\mu_{\mathbf{Z}}^T + \mu_{G\mathbf{Z}}\mu_{G\mathbf{Z}}^T/\mu_{G^2} - \mu_{\mathbf{Z}\mathbf{Z}^T}\right) - \mathbf{A}\left(\mu_{\mathcal{E}\mathbf{Z}} - \mu_{G\mathbf{Z}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T \right\} \\
\mathbf{A} &= \left(\mu_{G\mathcal{E}}\mu_{G\mathbf{Z}}/\mu_{G^2} - \mu_{\mathcal{E}\mathbf{Z}}\right)/\left(\mu_{\mathcal{E}^2} - \mu_{G\mathcal{E}}^2/\mu_{G^2}\right).
\end{aligned}$$

The solution to $\boldsymbol{\alpha}_W$ similarly involves inverting a matrix:

$$\begin{aligned}
\boldsymbol{\alpha}_W &= -\mathbf{Q}^{-1}\mathbf{R} \\
\mathbf{R} &= \beta_{\mathcal{E}} \left\{ -\mu_f \mu_{\mathbf{W}} - \mu_{Gf}\mu_{G\mathbf{W}}/\mu_{G^2} + \mu_{f\mathbf{W}} + \mathbf{B}\left(\mu_{\mathcal{E}f} - \mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \mathbf{C}\beta_{\mathcal{E}} \left\{ -\mu_f \mu_{\mathbf{Z}} - \mu_{Gf}\mu_{G\mathbf{Z}}/\mu_{G^2} + \mu_{f\mathbf{Z}} + \mathbf{A}\left(\mu_{\mathcal{E}f} - \mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \beta_I \left\{ -\mu_{Gh}\mu_{\mathbf{W}} - \mu_{G^2h}\mu_{G\mathbf{W}}/\mu_{G^2} + \mu_{Gh\mathbf{W}} + \mathbf{B}\left(\mu_{G\mathcal{E}h} - \mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \mathbf{C}\beta_I \left\{ \mu_{Gh\mathbf{Z}} - \mu_{Gh}\mu_{\mathbf{Z}} - \mu_{G^2h}\mu_{G\mathbf{Z}}/\mu_{G^2} + \mathbf{A}\left(\mu_{G\mathcal{E}h} - \mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\} \\
&\quad + \left\{ -\mu_{\mathbf{W}}\mu_{\mathbf{M}}^T - \mu_{G\mathbf{W}}\mu_{G\mathbf{M}}^T/\mu_{G^2} + \mu_{\mathbf{W}\mathbf{M}^T} + \mathbf{B}\left(\mu_{\mathcal{E}\mathbf{M}} - \mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T \right\} \boldsymbol{\beta}_{\mathbf{M}} \\
&\quad + \mathbf{C}\left\{ \mathbf{A}\left(\mu_{\mathcal{E}\mathbf{M}} - \mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T - \mu_{\mathbf{Z}}\mu_{\mathbf{M}}^T - \mu_{G\mathbf{Z}}\mu_{G\mathbf{M}}^T/\mu_{G^2} + \mu_{\mathbf{Z}\mathbf{M}^T} \right\} \boldsymbol{\beta}_{\mathbf{M}} \\
&\quad + \alpha_I \left\{ \mu_{G\mathcal{E}}\mu_{\mathbf{W}} + \mu_{G^2\mathcal{E}}\mu_{G\mathbf{W}}/\mu_{G^2} - \mu_{G\mathcal{E}\mathbf{W}} + \mathbf{B}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{G\mathcal{E}^2}\right)\right\} \\
&\quad + \mathbf{C}\alpha_I \left\{ \mu_{G\mathcal{E}}\mu_{\mathbf{Z}} + \mu_{G^2\mathcal{E}}\mu_{G\mathbf{Z}}/\mu_{G^2} - \mu_{G\mathcal{E}\mathbf{Z}} + \mathbf{A}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{G\mathcal{E}^2}\right)\right\} \\
\mathbf{Q} &= \left\{ \mu_{\mathbf{W}}\mu_{\mathbf{W}}^T + \mu_{G\mathbf{W}}\mu_{G\mathbf{W}}^T/\mu_{G^2} - \mu_{\mathbf{W}\mathbf{W}^T} + \mathbf{B}\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2} - \mu_{\mathcal{E}\mathbf{W}}\right)^T \right\}
\end{aligned}$$

$$
\begin{aligned}
&+\mathbf{C}\left\{\mathbf{A}\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{\mathcal{E}\mathbf{W}}\right)^T+\mu_{\mathbf{Z}}\mu_{\mathbf{W}}^T+\mu_{G\mathbf{Z}}\mu_{G\mathbf{W}}^T/\mu_{G^2}-\mu_{\mathbf{Z}\mathbf{W}^T}\right\}\\
\mathbf{B} &= \left(\mu_{G\mathcal{E}}\mu_{G\mathbf{W}}/\mu_{G^2}-\mu_{\mathcal{E}\mathbf{W}}\right)/\left(\mu_{\mathcal{E}^2}-\mu_{G\mathcal{E}}^2/\mu_{G^2}\right)\\
\mathbf{C} &= \left\{-\mu_{\mathbf{W}}\mu_{\mathbf{Z}}^T-\mu_{G\mathbf{W}}\mu_{G\mathbf{Z}}^T/\mu_{G^2}+\mu_{\mathbf{W}\mathbf{Z}^T}+\mathbf{B}\left(\mu_{\mathcal{E}\mathbf{Z}}-\mu_{G\mathbf{Z}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T\right\}\times\mathbf{O}^{-1}.
\end{aligned}
$$

Finally the solution to α_I is:

$$
\begin{aligned}
\alpha_I &= S/T\\
S &= \beta_{\mathcal{E}}\left\{-\mu_f\mu_{G\mathcal{E}}-\mu_{Gf}\mu_{G^2\mathcal{E}}/\mu_{G^2}+\mu_{G\mathcal{E}f}+D\left(\mu_{\mathcal{E}f}-\mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad+\mathbf{E}\beta_{\mathcal{E}}\left\{-\mu_f\mu_{\mathbf{Z}}-\mu_{Gf}\mu_{G\mathbf{Z}}/\mu_{G^2}+\mu_{f\mathbf{Z}}+\mathbf{A}\left(\mu_{\mathcal{E}f}-\mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad+\beta_I\left\{-\mu_{Gh}\mu_{G\mathcal{E}}-\mu_{G^2h}\mu_{G^2\mathcal{E}}/\mu_{G^2}+\mu_{G^2\mathcal{E}h}+D\left(\mu_{G\mathcal{E}h}-\mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad+\mathbf{E}\beta_I\left\{\mu_{Gh\mathbf{Z}}-\mu_{Gh}\mu_{\mathbf{Z}}-\mu_{G^2h}\mu_{G\mathbf{Z}}/\mu_{G^2}+\mathbf{A}\left(\mu_{G\mathcal{E}h}-\mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad+\left\{\mu_{G\mathcal{E}\mathbf{M}}-\mu_{\mathbf{M}}\mu_{G\mathcal{E}}-\mu_{G\mathbf{M}}\mu_{G^2\mathcal{E}}/\mu_{G^2}+D\left(\mu_{\mathcal{E}\mathbf{M}}-\mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}^T\boldsymbol{\beta}_{\mathbf{M}}\\
&\quad+\mathbf{E}\left\{\mathbf{A}\left(\mu_{\mathcal{E}\mathbf{M}}-\mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T-\mu_{\mathbf{Z}}\mu_{\mathbf{M}}^T-\mu_{G\mathbf{Z}}\mu_{G\mathbf{M}}^T/\mu_{G^2}+\mu_{\mathbf{Z}\mathbf{M}^T}\right\}\boldsymbol{\beta}_{\mathbf{M}}\\
&\quad-\mathbf{F}\beta_{\mathcal{E}}\left\{-\mu_f\mu_{\mathbf{W}}-\mu_{Gf}\mu_{G\mathbf{W}}/\mu_{G^2}+\mu_{f\mathbf{W}}+\mathbf{B}\left(\mu_{\mathcal{E}f}-\mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad-\mathbf{FC}\beta_{\mathcal{E}}\left\{-\mu_f\mu_{\mathbf{Z}}-\mu_{Gf}\mu_{G\mathbf{Z}}/\mu_{G^2}+\mu_{f\mathbf{Z}}+\mathbf{A}\left(\mu_{\mathcal{E}f}-\mu_{Gf}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad-\mathbf{F}\beta_I\left\{-\mu_{Gh}\mu_{\mathbf{W}}-\mu_{G^2h}\mu_{G\mathbf{W}}/\mu_{G^2}+\mu_{Gh\mathbf{W}}+\mathbf{B}\left(\mu_{G\mathcal{E}h}-\mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad-\mathbf{FC}\beta_I\left\{\mu_{Gh\mathbf{Z}}-\mu_{Gh}\mu_{\mathbf{Z}}-\mu_{G^2h}\mu_{G\mathbf{Z}}/\mu_{G^2}+\mathbf{A}\left(\mu_{G\mathcal{E}h}-\mu_{G^2h}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}\\
&\quad-\mathbf{F}\left\{-\mu_{\mathbf{W}}\mu_{\mathbf{M}}^T-\mu_{G\mathbf{W}}\mu_{G\mathbf{M}}^T/\mu_{G^2}+\mu_{\mathbf{W}\mathbf{M}^T}+\mathbf{B}\left(\mu_{\mathcal{E}\mathbf{M}}-\mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T\right\}\boldsymbol{\beta}_{\mathbf{M}}\\
&\quad-\mathbf{FC}\left\{\mathbf{A}\left(\mu_{\mathcal{E}\mathbf{M}}-\mu_{G\mathbf{M}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)^T-\mu_{\mathbf{Z}}\mu_{\mathbf{M}}^T-\mu_{G\mathbf{Z}}\mu_{G\mathbf{M}}^T/\mu_{G^2}+\mu_{\mathbf{Z}\mathbf{M}^T}\right\}\boldsymbol{\beta}_{\mathbf{M}}\\
T &= \mathbf{F}\left\{\mu_{G\mathcal{E}}\mu_{\mathbf{W}}+\mu_{G^2\mathcal{E}}\mu_{G\mathbf{W}}/\mu_{G^2}-\mu_{G\mathcal{E}\mathbf{W}}+\mathbf{B}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}^2}\right)\right\}\\
&\quad+\mathbf{FC}\left\{\mu_{G\mathcal{E}}\mu_{\mathbf{Z}}+\mu_{G^2\mathcal{E}}\mu_{G\mathbf{Z}}/\mu_{G^2}-\mu_{G\mathcal{E}\mathbf{Z}}+\mathbf{A}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}^2}\right)\right\}\\
&\quad-\left\{\mu_{G\mathcal{E}}^2+\mu_{G^2\mathcal{E}}^2/\mu_{G^2}-\mu_{G^2\mathcal{E}^2}+D\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}^2}\right)\right\}\\
&\quad-\mathbf{E}\left\{\mu_{G\mathcal{E}}\mu_{\mathbf{Z}}+\mu_{G^2\mathcal{E}}\mu_{G\mathbf{Z}}/\mu_{G^2}-\mu_{G\mathcal{E}\mathbf{Z}}+\mathbf{A}\left(\mu_{G^2\mathcal{E}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}^2}\right)\right\}\\
D &= \left(\mu_{G\mathcal{E}}\mu_{G^2\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}^2}\right)/\left(\mu_{\mathcal{E}^2}-\mu_{G\mathcal{E}}^2/\mu_{G^2}\right)\\
\mathbf{E} &= \left\{\mu_{G\mathcal{E}\mathbf{Z}}-\mu_{\mathbf{Z}}\mu_{G\mathcal{E}}-\mu_{G\mathbf{Z}}\mu_{G^2\mathcal{E}}/\mu_{G^2}+D\left(\mu_{\mathcal{E}\mathbf{Z}}-\mu_{G\mathbf{Z}}\mu_{G\mathcal{E}}/\mu_{G^2}\right)\right\}^T\times\mathbf{O}^{-1}\\
\mathbf{F} &= \left[\left\{\mu_{\mathbf{W}}\mu_{G\mathcal{E}}+\mu_{G\mathbf{W}}\mu_{G^2\mathcal{E}}/\mu_{G^2}-\mu_{G\mathcal{E}\mathbf{W}}+D\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{\mathcal{E}\mathbf{W}}\right)\right\}^T\right.\\
&\quad\left.+\mathbf{E}\left\{\mathbf{A}\left(\mu_{G\mathbf{W}}\mu_{G\mathcal{E}}/\mu_{G^2}-\mu_{\mathcal{E}\mathbf{W}}\right)^T+\mu_{\mathbf{Z}}\mu_{\mathbf{W}}^T+\mu_{G\mathbf{Z}}\mu_{G\mathbf{W}}^T/\mu_{G^2}-\mu_{\mathbf{Z}\mathbf{W}^T}\right\}\right]\times\mathbf{Q}^{-1}.
\end{aligned}
$$

1.8.2 Intuition on how bias arises in logistic regression models from Section 1.3.3

To develop a sense for how the bias arises in the simplest models, assume for now that G is binary, as in a dominant genetic model. Also assume that we have gene-environment independence and that $\mathbf{Z} = \mathrm{M} = \mathbf{W} = 0$. The asymptotic limit of the four score equations is then:

$$E\left[(1, G, \mathcal{E}, G\mathcal{E})^T \left\{Y - g^{-1}(\alpha_0 + \alpha_G G + \alpha_\mathcal{E}\mathcal{E} + \alpha_I G\mathcal{E})\right\}\right] = 0.$$

Taking the expectation with respect to G and performing some simple algebra gives:

$$E\left\{g^{-1}\left(\beta_0 + \beta_\mathcal{E} f(\mathcal{E})\right) - g^{-1}(\alpha_0 + \alpha_\mathcal{E}\mathcal{E})\right\} = 0 \quad (1.12)$$

$$E\left\{\mathcal{E}g^{-1}\left(\beta_0 + \beta_\mathcal{E} f(\mathcal{E})\right) - \mathcal{E}g^{-1}(\alpha_0 + \alpha_\mathcal{E}\mathcal{E})\right\} = 0. \quad (1.13)$$

See that α_0 and $\alpha_\mathcal{E}$ are determined wholly by β_0 and $\beta_\mathcal{E}$ in these two equations. Thus α_G and α_I are entirely determined by β_G and β_I in the other two equations.

Now suppose that $\beta_0, \beta_\mathcal{E}, \beta_I$ are fixed. If we substitute in different values of β_G and attempt to solve the system of 4 equations, then α_0 and $\alpha_\mathcal{E}$ will stay constant (since (1.12) and (1.13) are constant as β_G varies), and α_G and α_I will generally both vary. Thus α_I will rarely be unbiased for a specific value of β_I except in special cases, such as when $\beta_G = \beta_I = 0$. Intuitively, this argument shows that when there is a relationship between G and Y (captured by $\beta_G \neq 0$), then the interaction coefficient will be asymptotically biased.

1.8.3 Explanation of biased model-based variance estimates from Section 1.3.4

Even if the conditions for valid inference from Section 3.2 are met, we showed in Section 3.3 that the model-based standard error estimates are incorrect under exposure misspecification. Test statistics calculated using the naive model-based standard error estimates will not have the assumed χ^2_1 distribution under the null hypothesis. Here we show how the naive inference leads to non-uniform p-values, and we explain the effects of the design matrix and the form of misspecification.

Denote by $\widehat{m}_{\widehat{\alpha}_I}$ a model-based standard error estimate for $\widehat{\alpha}_I$. Assume that the conditions for valid inference are met, so that $\alpha_I = 0$ asymptotically. By a modification of Theorem 1 from Rotnitzky and Jewell (1990), we can show that the test statistic calculated using the model-based standard error $T_{mod} = (\widehat{\alpha}_I/\widehat{m}_{\widehat{\alpha}_I})^2$ will have an asymptotic scaled chi-square distribution: $T_{mod} = c\chi^2_1 + o_p(1)$. For the identity link, the scale factor is given by:

$$c_{lin} = \frac{\sum_{i=1}^n \widetilde{D}_i^2 \left(Y_i - \widetilde{\mathbf{X}}_i^T \boldsymbol{\alpha}\right)^2}{E\left\{\left(Y_i - \widetilde{\mathbf{X}}_i^T \boldsymbol{\alpha}\right)^2\right\} \sum_{i=1}^n \widetilde{D}_i^2}$$

where $\widetilde{D}_i = D_i^{(1)} - D_i^{(2)} \left(\sum_{j=1}^n D_j^{(2)T} D_j^{(2)}\right)^{-1} \left(\sum_{j=1}^n D_j^{(2)T} D_j^{(1)}\right)$ with $D_i^{(1)} = G_i \mathcal{E}_i$ and $D_i^{(2)} = \left(1, G_i, \mathcal{E}_i, \mathbf{Z}_i^T, \mathbf{W}_i^T\right)$.

For the logistic link, the scale factor is given by:

$$c_{log} = \frac{\sum_{i=1}^n \widetilde{D}_i^2 \widetilde{V}_i^{-2} \left\{Y_i - g^{-1}(\widetilde{\mathbf{X}}_i^T \boldsymbol{\alpha})\right\}^2}{\sum_{i=1}^n \widetilde{D}_i^2 \widetilde{V}_i^{-1}}$$

with $\widetilde{D}_i = D_i^{(1)} - D_i^{(2)} \left(\sum_{i=1}^n D_i^{(2)T} \widetilde{V}_i^{-1} D_i^{(2)}\right)^{-1} \left(\sum_{i=1}^n D_i^{(2)T} \widetilde{V}_i^{-1} D_i^{(1)}\right)$ where $D_i^{(1)} = G_i \mathcal{E}_i \widetilde{\mu}_i (1 - \widetilde{\mu}_i)$, $D_i^{(2)} = \left(1, G_i, \mathcal{E}_i, \mathbf{Z}_i^T, \mathbf{W}_i^T\right) \times \widetilde{\mu}_i (1 - \widetilde{\mu}_i)$, $\widetilde{V}_i = \widetilde{\mu}_i (1 - \widetilde{\mu}_i)$, and $\widetilde{\mu}_i = g^{-1}(\widetilde{\mathbf{X}}_i^T \boldsymbol{\alpha})$.

If $c > 1$ throughout the genome, then our model-based test statistics will be inflated, and vice versa. These expressions highlight the importance of the design matrix in determining the distribution of naive test statistics calculated with the model-based standard error. Specifically, in the identity link case, $\widetilde{\mathbf{D}}$ has an interpretation as the projection of the interaction vector onto the orthogonal complement of the column space of the other vectors in the design matrix. If the residuals are often larger when $\widetilde{D}_i$ is larger, then we can expect inflated test statistics, and vice versa.

1.8.4 Variance of the sandwich estimator under misspecification from Section 1.4.1

Suppose that we fit the model (2) from Section 2.1. Then the sandwich estimator of the variance of $\widehat{\alpha}_I$ is:

$$\widehat{s}_{\widehat{\alpha}_I} = \sum_{i=1}^n a_i^2 (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^n a_i^2 \widehat{\epsilon}_i^2$$

$$
\begin{aligned}
a_i &= l^T(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}_i \\
l &= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix}^T_{1\times(q+r+4)}.
\end{aligned}
$$

Assume the true linear model is still given by (1):

$$
Y_i = \beta_0 + \beta_G G_i + \beta_{\mathcal{E}} f(\mathcal{E}_i) + \beta_I G_i h(\mathcal{E}_i) + \mathbf{Z}_i^T\boldsymbol{\beta}_Z + \mathbf{M}_i^T\boldsymbol{\beta}_M + \epsilon_i, \epsilon_i \sim N(0,\sigma^2).
$$

Then we can find the distribution of the $n\times 1$ residual vector conditional on $\widetilde{\mathbf{X}}$. This is (we omit the conditioning statement in the following derivations for sake of presentation):

$$
\begin{aligned}
\widehat{\epsilon} &= \left(\mathbf{I} - \widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T\right)\mathbf{Y} \\
&= \left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)\mathbf{Y} \\
&\sim MVN\left\{\left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)E\left(\mathbf{Y}|\widetilde{\mathbf{X}}\right), \left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)\sigma^2\mathbf{I}\left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)^T\right\} \\
&= MVN\left\{\boldsymbol{\theta}, \left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)\sigma^2\right\} \\
\boldsymbol{\theta} &= \left(\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}\right)E\left(\mathbf{Y}|\widetilde{\mathbf{X}}\right).
\end{aligned}
$$

Due to the misspecification, $\boldsymbol{\theta}$ will generally not equal $\mathbf{0}$ by a similar argument to Section 2.2. If we had correctly specified the model then it would be true that $\boldsymbol{\theta} = \mathbf{0}$.

Denote by h_{ii} the (i,i) element of $\mathbf{I} - \mathbf{H}_{\widetilde{\mathbf{X}}}$. Now we can calculate Var $(\widehat{\epsilon}_i^2)$:

$$
\begin{aligned}
\text{Var}\left(\widehat{\epsilon}_i^2\right) &= E\left(\widehat{\epsilon}_i^4\right) - \left\{E\left(\widehat{\epsilon}_i^2\right)\right\}^2 \\
&= \theta_i^4 + 6\theta_i^2(1-h_{ii})\sigma^2 + 3(1-h_{ii})^2\sigma^4 - \left\{\theta_i^2 + (1-h_{ii})\sigma^2\right\}^2 \\
&= \theta_i^4 + 6\theta_i^2(1-h_{ii})\sigma^2 + 3(1-h_{ii})^2\sigma^4 - \theta_i^4 - 2\theta_i^2(1-h_{ii})\sigma^2 - (1-h_{ii})^2\sigma^4 \\
&= 4\theta_i^2(1-h_{ii})\sigma^2 + 2(1-h_{ii})^2\sigma^4.
\end{aligned}
$$

The next step is to calculate Cov$(\widehat{\epsilon}_i^2, \widehat{\epsilon}_j^2)$:

$$
\begin{aligned}
\text{Cov}(\widehat{\epsilon}_i^2, \widehat{\epsilon}_j^2) &= E\left(\widehat{\epsilon}_i^2\widehat{\epsilon}_j^2\right) - E\left(\widehat{\epsilon}_i^2\right)E\left(\widehat{\epsilon}_j^2\right) \\
&= E\left\{\widehat{\epsilon}_j^2 E\left(\widehat{\epsilon}_i^2|\widehat{\epsilon}_j\right)\right\} - \left\{\theta_i^2 + (1-h_{ii})\sigma^2\right\}\left\{\theta_j^2 + (1-h_{jj})\sigma^2\right\}.
\end{aligned}
$$

Recall that if (X,Y) have a bivariate normal distribution with correlation ρ, then the conditional distribution $X|Y$ is:

$$
X|Y \sim N\left\{\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y-\mu_Y), (1-\rho^2)\sigma_X^2\right\}.
$$

So we have

$$
\begin{aligned}
E\left(\widehat{\epsilon}_i^2\widehat{\epsilon}_j^2\right) &= E\left\{\widehat{\epsilon}_j^2 E\left(\widehat{\epsilon}_i^2|\epsilon_j\right)\right\} \\
&= E\left[\widehat{\epsilon}_j^2\left\{\theta_i^2 + 2\rho\theta_i\frac{\sigma_i}{\sigma_j}(\widehat{\epsilon}_j - \theta_j) + \rho^2\frac{\sigma_i^2}{\sigma_j^2}(\widehat{\epsilon}_j^2 - 2\widehat{\epsilon}_j\theta_j + \theta_j^2) + (1-\rho^2)\sigma_i^2\right\}\right] \\
&= E\left\{\widehat{\epsilon}_j^4\left(\rho^2\frac{\sigma_i^2}{\sigma_j^2}\right) + \widehat{\epsilon}_j^3\left(2\rho\theta_i\frac{\sigma_i}{\sigma_j} - 2\rho^2\theta_j\frac{\sigma_i^2}{\sigma_j^2}\right)\right\} \\
&\quad + E\left\{\widehat{\epsilon}_j^2\left(\theta_i^2 - 2\rho\frac{\sigma_i}{\sigma_j}\theta_i\theta_j + \rho^2\frac{\sigma_i^2}{\sigma_j^2}\theta_j^2 + (1-\rho^2)\sigma_i^2\right)\right\} \\
&= \left(\theta_j^4 + 6\theta_j^2\sigma_j^2 + 3\sigma_j^4\right)\left(\rho^2\frac{\sigma_i^2}{\sigma_j^2}\right) + \left(\theta_j^3 + 3\theta_j\sigma_j^2\right)\left(2\rho\theta_i\frac{\sigma_i}{\sigma_j} - 2\rho^2\theta_j\frac{\sigma_i^2}{\sigma_j^2}\right) \\
&\quad + \left(\theta_j^2 + \sigma_j^2\right)\left(\theta_i^2 - 2\rho\frac{\sigma_i}{\sigma_j}\theta_i\theta_j + \rho^2\frac{\sigma_i^2}{\sigma_j^2}\theta_j^2 + (1-\rho^2)\sigma_i^2\right) \\
&= \theta_j^4\rho^2\frac{\sigma_i^2}{\sigma_j^2} + 6\theta_j^2\rho^2\sigma_i^2 + 3\rho^2\sigma_i^2\sigma_j^2 + 2\theta_i\theta_j^3\rho\frac{\sigma_i}{\sigma_j} - 2\theta_j^4\rho^2\frac{\sigma_i^2}{\sigma_j^2} + 6\theta_i\theta_j\rho\sigma_i\sigma_j \\
&\quad - 6\theta_j^2\rho^2\sigma_i^2 + \theta_i^2\theta_j^2 - 2\theta_i\theta_j^3\frac{\sigma_i}{\sigma_j} + \theta_j^4\rho^2\frac{\sigma_i^2}{\sigma_j^2} + \theta_j^2(1-\rho^2)\sigma_i^2 + \theta_i^2\sigma_j^2 \\
&\quad - 2\theta_i\theta_j\rho\sigma_i\sigma_j + \theta_j^2\rho^2\sigma_i^2 + (1-\rho^2)\sigma_i^2\sigma_j^2 \\
&= 2\rho^2\sigma_i^2\sigma_j^2 + 4\theta_i\theta_j\rho\sigma_i\sigma_j + \theta_i^2\theta_j^2 + \theta_j^2\sigma_i^2 + \theta_i^2\sigma_j^2 + \sigma_i^2\sigma_j^2 \\
&= 2\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j)^2 + 4\theta_i\theta_j\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j) + \theta_i^2\theta_j^2 + \theta_i^2\sigma_j^2 + \theta_j^2\sigma_i^2 + \sigma_i^2\sigma_j^2 \\
&= 2\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j)^2 + 4\theta_i\theta_j\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j) + \left(\theta_i^2 + (1-h_{ii})\sigma^2\right)\left(\theta_j^2 + (1-h_{jj})\sigma^2\right),
\end{aligned}
$$

and finally the covariance is:

$$
\begin{aligned}
\mathrm{Cov}(\widehat{\epsilon}_i^2, \widehat{\epsilon}_j^2) &= E\left(\widehat{\epsilon}_i^2\widehat{\epsilon}_j^2\right) - E\left(\widehat{\epsilon}_i^2\right)E\left(\widehat{\epsilon}_j^2\right) \\
&= E\left\{\widehat{\epsilon}_j^2 E\left(\widehat{\epsilon}_i^2|\widehat{\epsilon}_j\right)\right\} - \left\{\theta_i^2 + (1-h_{ii})\sigma^2\right\}\left\{\theta_j^2 + (1-h_{jj})\sigma^2\right\} \\
&= 2\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j)^2 + 4\theta_i\theta_j\mathrm{Cov}(\widehat{\epsilon}_i, \widehat{\epsilon}_j).
\end{aligned}
$$

The variance of the sandwich estimator for the interaction term is then

$$
\mathrm{Var}\left(\widehat{s}_{\widehat{\alpha}_I}\right) = \sum_{i=1}^{n} a_i^4\left\{4\theta_i^2(1-h_{ii})\sigma^2 + 2(1-h_{ii})^2\sigma^4\right\} + \sum_{i\neq j} a_i^2 a_j^2\left(2h_{ij}^2\sigma^4 - 4\theta_i\theta_j h_{ij}\sigma^2\right),
$$

which is larger than it would be if the model were correctly specified and $\boldsymbol{\theta} = \mathbf{0}$.

Set-based tests using the Generalized Berk-Jones statistic in genetic association studies

Ryan Sun

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Xihong Lin

Department of Biostatistics

Harvard T.H. Chan School of Public Health

2.1 Introduction

Genome-Wide Association Studies (GWAS) have been successful in identifying the association between thousands of Single Nucleotide Polymorphisms (SNPs) and hundreds of complex traits (Manolio et al., 2009). A traditional GWAS analysis tests for the effect of each individual SNP, and this approach has shown that single-SNP effects are often weak across the genome (Visscher et al., 2012). Recently, set-based tests which jointly analyze a group of SNPs - e.g. SNPs in a gene, pathway, or network - have become increasingly popular as complementary tools which can boost analysis power in GWAS (Wu et al., 2010). These tests are also standard for rare variant analysis in whole genome sequencing studies (Lee et al., 2014).

SNPs can be aggregated into sets based on a variety of genomic features. For example, they can be grouped by physical position, such as location in a gene or Linkage Disequilibrium (LD) block, or similar biological functions, such as membership in a genetic pathway or protein network. Set-based analyses then allow for some natural advantages over individual SNP methods. Besides reducing the number of multiple comparisons across the genome, SNP-set methods can increase power by pooling sparse and weak effects into a stronger aggregated signal, as well as by incorporating biological information into the test (Wu et al., 2011). In addition, set-based interpretations of association may be more meaningful than their single-marker counterparts, such as in a gene-level or pathway-level analysis.

A number of set-based tests for genetic association studies have been developed in recent years, including burden tests (Li and Leal, 2008), the Sequence Kernel Association Test (SKAT) (Wu et al., 2011), the minimum p-value test (MinP) (Conneely and Boehnke, 2007), and most recently the Generalized Higher Criticism (GHC) (Barnett, Mukherjee and Lin, 2016). SKAT and burden tests are examples of methods more suitable for detecting dense signals. If the signals reside in only a few SNPs that are not correlated with noise SNPs, then the power of SKAT and burden tests will suffer.

While certain SNP-sets may contain a large number of signals, it is more common that genomic constructs formed with GWAS data will have only a few signal SNPs. In-

terestingly, within tests designed for this sparse alternative setting, there are still subtle differences in performance. Under extreme sparsity, as in the case of only one or very few signals in the entire set, the minimum p-value test and the GHC have good power for detecting a SNP-set effect. However GHC and MinP can lose power under moderate sparsity settings, which are relatively common in gene and pathway level analyses of GWAS data. For example, in the Cancer Genetic Markers of Susceptibility (CGEMS) GWAS for breast cancer risk (Hunter et al., 2007), four out of 42 SNPs in the FGFR2 gene, a known breast cancer risk loci, showed strong evidence of association without reaching genome-wide significance. It is hence of substantial interest to develop testing procedures that can reliably detect associations across a range of alternatives in the sparse signal regime.

When factors in a set are independent, several goodness-of-fit type methods have been proposed to perform set-based tests in the presence of sparse signals. (Donoho and Jin, 2004; Jager and Wellner, 2007; Walther, 2013). These methods test for the effect of a set by aggregating evidence from marginal test statistics, and they have been shown to possess attractive asymptotic properties when the size of a set goes to infinity. Specifically, they reach a so-called detection boundary when signals are sparse. In a certain sense, they are able to detect the weakest signals detectable by any statistical procedure under the sparse alternative. The class of tests with this ability includes the Higher Criticism (HC) (Donoho and Jin, 2004) and the Berk-Jones (BJ) (Berk and Jones, 1979). Compared to the HC and BJ tests, the minimum p-value test, for example, is known to attain the detection boundary over a smaller portion of the sparse regime. In terms of finite sample performance, it has been demonstrated through simulation that Berk-Jones outperforms Higher Criticism over a range of moderately sparse alternatives when marginal test statistics are independent (Walther, 2013; Li and Siegmund, 2015). Donoho and Jin (2004) provide an explanation for this result by showing that HC disproportionately weights evidence from the most extreme observed marginal test statistic, at the cost of losing sensitivity to detect signals in other locations.

However, a direct application of BJ to SNP-set testing is not desirable, as standard p-value calculations for BJ rely on independence of the observations (Moscovich-Eiger and Nadler, 2017). This assumption is violated by LD-induced correlation between neigh-

boring SNPs in a SNP-set. In addition, we will see that even if we correct the inference procedure of the Berk-Jones, the power of BJ can fall dramatically in high LD settings.

To overcome the challenges posed by correlated SNPs, this paper proposes the Generalized Berk-Jones (GBJ) statistic for testing the association between a SNP-set and outcome. GBJ accounts for LD among SNPs in a set while still preserving the ability to detect moderately sparse and weak signals in finite samples. In fact, GBJ reduces to the Berk-Jones statistic when all SNPs in a set are independent. GBJ can also be applied to SNP-set tests using both individual-level genotype data or GWAS summary statistics from single SNP analysis. To facilitate use, we additionally provide an analytic p-value calculation for GBJ. Our method is more computationally efficient than permutation and is shown to be accurate even at the extremely small levels required for genome-wide significance.

Additional insight into the strengths and weaknesses of GBJ is provided by studying the rejection regions of SNP-set tests developed for sparse alternatives. The rejection regions allow us to quantitively describe how the power of each test is susceptible to changes in parameters such as the amount of correlation between SNPs or the size of the SNP-set. Since in practice we never have knowledge of the type of alternative, we also propose an omnibus test that combines GBJ, GHC, MinP, and SKAT for added robustness to different degrees of sparsity. An extensive simulation study demonstrates that GBJ outperforms alternative methods in testing SNP-set effects when signals are weak and moderately sparse, and we also show that the omnibus test is robust to a wide range of sparsity levels.

The remainder of the paper is organized as follows. Section 2.2 discusses the SNP-set testing framework using both individual-level data and GWAS summary statistics. In Section 2.3 we propose the Generalized Berk-Jones statistic for testing the association between a SNP-set and outcome. We also provide an analytic p-value calculation for GBJ and develop the omnibus test. Section 2.4 compares the rejection regions of GBJ and other tests designed for the sparse regime. In Section 2.5 we demonstrate the finite sample performance of GBJ through simulation. Finally Section 2.6 presents an analysis of the CGEMS data, and we conclude with a discussion in Section 2.7.

2.2 The SNP-set testing framework

2.2.1 Individual-level genotype data

We begin by describing the SNP-set testing framework using individual-level data on genotype, outcome, and other covariates for n total subjects. Suppose for subject i, $i = 1, ..., n$, we observe the outcome Y_i, a genotype vector $\mathbf{G}_i = (G_{i1}, ..., G_{id})^T$ of d SNPs in a SNP-set, and a vector of q covariates $\mathbf{X}_i = (X_{i1}, ..., X_{iq})^T$. Assume that Y_i conditional on $(\mathbf{G}_i, \mathbf{X}_i)$ follows a distribution in the exponential family (McCullagh and Nelder, 1989) with the density function $f(Y_i) = \exp\{(Y_i\theta_i - b(\theta_i))/a_i(\phi) + c(Y_i, \phi)\}$, where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$, are known functions, θ_i is a canonical parameter, and ϕ is a dispersion parameter. Let $\mu_i = E(Y_i|\mathbf{G}_i, \mathbf{X}_i)$ denote the conditional mean of Y_i and assume it follows the Generalized Linear Model (GLM)

$$g(\mu_i) = \mathbf{X}_i^T\boldsymbol{\alpha} + \mathbf{G}_i^T\boldsymbol{\beta}.$$

where $g(\cdot)$ is a differentiable monotone link function. We here only consider canonical link functions for simplicity. In matrix notation, the data take the form $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{G}_{n\times d} = [\ \mathbf{G}_1, ..., \mathbf{G}_n\ ]^T$, and $\mathbf{X}_{n\times q} = [\ \mathbf{X}_1, ..., \mathbf{X}_n\ ]^T$.

The null hypothesis of no association between a SNP-set and outcome, after controlling for covariates, is given by $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Both the dimension of d and the sparsity of signals can vary greatly between sets, i.e. from gene to gene, and the number of nonzero β_j is unknown. Our aim is to develop a test suitable for different levels of sparsity while also accounting for the correlation among individual SNP test statistics.

The marginal score statistic for β_j, $j = 1, ..., d$, is

$$Z_j = \frac{\mathbf{G}_{\cdot j}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)}{\sqrt{\mathbf{G}_{\cdot j}^T\mathbf{P}\mathbf{G}_{\cdot j}}},$$

where $\mathbf{G}_{\cdot j}$ denotes the jth column vector of $\mathbf{G}$, $\mathbf{P} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$ is the projection matrix, $\mathbf{W} = \text{diag}\{a_1\phi v(\hat{\mu}_{01}), ..., a_n\phi v(\hat{\mu}_{0n})\}$, $\hat{\mu}_{0i} = g(\mathbf{X}_i^T\hat{\boldsymbol{\alpha}}_0)$, $\hat{\boldsymbol{\alpha}}_0$ is the MLE of $\boldsymbol{\alpha}$ under the null hypothesis, and $v(\mu_i) = b''(\theta_i)$ is the variance function.

The d marginal score test statistics have an asymptotic multivariate normal distribu-

tion

$$\mathbf{Z} = (Z_1, \cdots, Z_d)^T \sim N(\mathbf{0}_{d\times 1}, \mathbf{\Sigma}_{d\times d})$$

under the null, where $\Sigma_{jj} = 1$ for all j, and for $j \neq k$ we can estimate

$$\hat{\Sigma}_{jk} = \frac{\mathbf{G}_{.j}^T \mathbf{P} \mathbf{G}_{.k}}{\sqrt{\mathbf{G}_{.j}^T \mathbf{P} \mathbf{G}_{.j}} \sqrt{\mathbf{G}_{.k}^T \mathbf{P} \mathbf{G}_{.k}}}. \tag{2.1}$$

2.2.2 GWAS summary statistics

Many GWAS may not release individual-level data due to logistical challenges or data confidentiality agreements. Instead it is much more likely that a marginal test statistic for association with the outcome is available for each individual SNP (Pasaniuc and Price, 2016). It is hence of great interest to be able to perform SNP-set testing using precomputed Z_j from across the genome. To test a set of precomputed Z_j with GBJ, we require estimation of their correlation matrix $\mathbf{\Sigma}$ using external information.

Assume we have a panel of reference genotypes from n_r subjects of the same ethnicity as those used to construct the summary statistics. For example, this could come from the publicly available 1000 Genomes dataset (1000 Genomes Project Consortium, 2015). We estimate $\mathbf{\Sigma}$ using equation (2.1) but replace $\mathbf{G}_{.j}$ and $\mathbf{X}$ with $\mathbf{G}_{.j}^{(r)}$ and $\mathbf{X}^{(r)}$, where $\mathbf{G}_{.j}^{(r)}$ is the $n_r \times 1$ genotype vector of SNP j from the reference panel, $\mathbf{X}^{(r)} = (\mathbf{1}, \mathbf{PC}_1, ..., \mathbf{PC}_m)$, $\mathbf{PC}_1, ..., \mathbf{PC}_m$ are the first m principal component vectors calculated from the reference panel, and m is the same number of principal components as was used to control for population stratification (Price et al., 2006) in the original GWAS analysis of the data. Additionally for each subject we estimate $v(\hat{\mu}_{0i})$ by setting $\hat{\mu}_{0i}$ equal to the sample mean of the outcome. For a normally distributed outcome, this is exact as $v(\cdot) = 1$. For a binary outcome, since population stratification is the primary confounder of the SNP-outcome relationship in GWAS, and because μ_{i0} generally varies slowly with the principal components, this approximation is practically reasonable. Ultimately we are approximating $\mathbf{G}_{.j}^T \mathbf{P} \mathbf{G}_{.k}$ in (2.1) by $\mathbf{G}_{.j}^{(r)T} \mathbf{G}_{.k}^{(r)} - \mathbf{G}_{.j}^{(r)T} \mathbf{X}^{(r)} \{\mathbf{X}^{(r)T} \mathbf{X}^{(r)}\}^{-1} \mathbf{X}^{(r)T} \mathbf{G}_{.j}^{(r)}$ up to a scale parameter, with the scale parameter eventually cancelled out in $\hat{\Sigma}_{jk}$.

2.3 The Generalized Berk-Jones test for SNP-set effects

2.3.1 The Berk-Jones statistic

We briefly review the Berk-Jones statistic in this section to help introduce the Generalized Berk-Jones statistic in Section 2.3.2. The BJ statistic is designed to test for $H_0 : \boldsymbol{\beta} = 0$ against the alternative that a nonempty subset of the β_j are nonzero, assuming the marginal test statistics are independent. Let $\bar{\Phi}(t) = 1 - \Phi(t)$ denote the survival function of a standard normal random variable and $\Phi^{-1}(t)$ denote its inverse. Let $|Z|_{(j)}$ denote the order statistics of the vector that results from applying the absolute value operator to each element of $\mathbf{Z}$, so that $|Z|_{(1)}$ is the smallest value of $\mathbf{Z}$ in magnitude.

Set $S(t) = \sum_{j=1}^{d} \mathbf{1}\left(|Z_j| \geq t\right)$, which is the number of marginal test statistics with a magnitude greater than some threshold t. For a fixed $t \geq 0$, and if $Z_j \overset{\text{iid}}{\sim} N(0,1)$ for all j, then $S(t)$ has a binomial distribution with the size d and the mean parameter $\pi = 2\bar{\Phi}(t)$. This viewpoint motivates the Berk-Jones statistic for independent observations (Donoho and Jin, 2004) as:

$$\begin{aligned} BJ_d &= \max_{t > |Z|_{(d/2)}} \left[S(t) \log\left\{ \frac{S(t)}{2d\bar{\Phi}(t)} \right\} + \{d - S(t)\} \log\left\{ \frac{1 - S(t)/d}{1 - 2\bar{\Phi}(t)} \right\} \right] \mathbf{I}\left\{ 2\bar{\Phi}(t) < \frac{S(t)}{d} \right\} \qquad (2.2) \\ &= \max_{1 \leq j \leq d/2} \log \left[\frac{\Pr\left\{ S(|Z|_{(d-j+1)}) = j \middle| \pi = j/d \right\}}{\Pr\left\{ S(|Z|_{(d-j+1)}) = j \middle| \pi = 2\bar{\Phi}\left(|Z|_{(d-j+1)}\right) \right\}} \right] \mathbf{I}\left\{ 2\bar{\Phi}\left(|Z|_{(d-j+1)}\right) < \frac{j}{d} \right\}, \end{aligned}$$

where the second line uses the characterization of $S(t) \sim Bin(d, \pi)$. We see that BJ can roughly be explained as the maximum of a one-sided likelihood ratio test on the mean parameter of $S(t)$, performed over the larger half of observed test statistic magnitudes.

Implicit in this interpretation are the assumptions that the test statistics Z_j are independent and have a common mean. Specifically, under the binomial likelihood null, the common mean of the Z_j is 0, and under the binomial likelihood alternative the common mean at $t = |Z|_{(d-j+1)}$ is $\hat{\mu}_{j,d}$, where $\hat{\mu}_{j,d} > 0$ solves the equation

$$j/d = 1 - \left\{ \Phi(|Z|_{(d-j+1)} - \hat{\mu}_{j,d}) - \Phi(-|Z|_{(d-j+1)} - \hat{\mu}_{j,d}) \right\}. \qquad (2.3)$$

We say binomial likelihood null and alternative to make clear that we are talking about

an interpretation of the Berk-Jones statistic and to distinguish from the actual set-based null and alternative hypotheses being tested.

Note that the Higher Criticism test differs from the Berk-Jones by replacing the likelihood ratio statistic in (2.2) with the Pearson Chi-square statistic $\{S(t)-2\bar{\Phi}(t)\}^2/\{2\bar{\Phi}(t)(1-\bar{\Phi}(t)\}$. Let k be the number of causal SNPs in a set. The sparse regime is designated as $k < d^{1/2}$, and we call $d^{1/4} < k < d^{1/2}$ moderately sparse, with $k \leq d^{1/4}$ referred to as extremely sparse. Donoho and Jin (2004) showed that, when the Z_j are all mutually independent, both HC and BJ are able to reach the detection boundary over the entire sparse signal regime as $d \to \infty$. Walther (2013) and Li and Siegmund (2015) showed that the BJ statistic generally has better power than HC when the size of the set d is finite and the signals are moderately sparse.

If $Z_1, ..., Z_d$ are correlated, as they will be for test statistics arising from neighboring SNPs in a gene, then $S(t)$ no longer has a binomial distribution under the null. In this case, the standard Berk-Jones statistic no longer has a meaningful interpretation, and we may expect it to lose efficiency. In fact, we will show later that the rejection region of the Berk-Jones has a less desirable shape under various correlation structures, leading to a significant loss in power when the test statistics are not independent. Therefore we are interested in developing a modified BJ statistic that can account for correlation among the marginal test statistics in a set and thus possesses rejection regions which are more robust to arbitrary correlation structures.

2.3.2 The Generalized Berk-Jones statistic

We now propose the Generalized Berk-Jones statistic for testing the association between a SNP-set and outcome. Following the spirit of Berk-Jones, GBJ considers a likelihood ratio type statistic on the mean parameter of $S(t)$, but the key difference is GBJ explicitly accounts for the correlation structure of $Z_1, ..., Z_d$. More precisely, we define the GBJ statistic as:

$$GBJ_d = \max_{1\leq j\leq d/2} \log \left[\frac{\Pr\left\{ S\left(|Z|_{(d-j+1)}\right) = j \middle| \pi = j/d, \mathrm{cov}(\mathbf{Z}) = \mathbf{\Sigma} \right\}}{\Pr\left\{ S\left(|Z|_{(d-j+1)}\right) = j \middle| \pi = 2\bar{\Phi}\left(|Z|_{(d-j+1)}\right), \mathrm{cov}(\mathbf{Z}) = \mathbf{\Sigma} \right\}} \right]$$

$$\cdot \mathbf{I}\left\{2\bar{\Phi}\left(|Z|_{(d-j+1)}\right) < \frac{j}{d}\right\}$$

When the Z_j are correlated, $S(t)$ follows either an underdispersed or overdispersed binomial distribution instead of the standard binomial. However finding the exact distribution of $S(t)$ when $\mathbf{\Sigma} \neq \mathbf{I}$ is difficult. For a general $\mathbf{\Sigma}$, computing $\Pr\{S(t) = m\}$ requires iterating through d choose m terms and is very time consuming. In special cases, such as when $\mathbf{\Sigma}$ has an exchangeable correlation structure with $\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T$, the calculation is much easier. However these scenarios occur rarely, if ever, in practice.

We propose to approximate the full distribution of $S(t)$ using an Extended Beta-Binomial (EBB) distribution (Prentice, 1986). The Extended Beta-Binomial is a reparameterization and extension of the standard Beta-Binomial(α, β) distribution with the standard Beta-Binomial being a special case of the EBB. A random variable $V \sim \text{EBB}\,(d, \lambda, \gamma)$ has probability mass function

$$\Pr\left(V = v; d, \lambda, \gamma\right) = \left(\begin{array}{c} d \\ v \end{array}\right) \prod_{k=0}^{v-1}(\lambda + \gamma k) \prod_{k=0}^{d-v-1}(1 - \lambda + \gamma k) \Big/ \prod_{k=0}^{d-1}(1 + \gamma i), \tag{2.4}$$

where we follow the convention $\prod_{k=0}^{a} c_k = 1$ for $a < 0$. The mean of V is given by $E(V) = d\lambda$ and the variance is $\text{Var}(V) = d\lambda(1-\lambda)\left\{1 + (d-1)\gamma(1+\gamma)^{-1}\right\}$.

The Extended Beta-Binomial distribution reduces to the Beta-Binomial distribution if we set $\lambda = \alpha/(\alpha+\beta)$ and $\gamma = (\alpha+\beta+1)^{-1} / \left\{1 - (\alpha+\beta+1)^{-1}\right\}$ for $\alpha, \beta > 0$. Because the standard Beta-Binomial distribution requires $\alpha, \beta > 0$, it cannot accommodate underdispersion and never reduces to the binomial distribution. In contrast, the EBB allows for both overdispersion and underdispersion, and it reduces exactly to the binomial distribution when $\gamma = 0$. This mechanism allows our GBJ statistic to reduce to the Berk-Jones when there is no correlation among the observations.

2.3.3 Calculation of the Generalized Berk-Jones statistic

We now describe more precisely the mechanics of calculating the GBJ statistic. To begin, check if the condition $\mathbf{I}\left\{2\bar{\Phi}\left(|Z|_{(d-j+1)}\right) < \frac{j}{d}\right\}$ is satisfied at any $j \leq d/2$. If the condition is never satisfied, then the observed value of GBJ is 0 and we do not need to perform any

more computation. The following steps should only be taken on indices $j \leq d/2$ where the condition is satisfied.

At each qualifying j, we approximate the distribution of $S(|Z|_{(d-j+1)})$ by an Extended Beta-Binomial random variable under both the binomial likelihood null and the binomial likelihood alternative. Denote these two variables by $V_0^{(j)} \sim EBB(\lambda_0^{(j)}, \gamma_0^{(j)})$ and $V_a^{(j)} \sim EBB(\lambda_a^{(j)}, \gamma_a^{(j)})$. We solve for $(\lambda_0^{(j)}, \gamma_0^{(j)})$ through moment matching equations

$$\begin{aligned} \lambda_0^{(j)} &= E_0\left\{S(|Z|_{(d-j+1)})\right\}/d, \\ \frac{\gamma_0^{(j)}}{1+\gamma_0^{(j)}} &= \frac{\text{Var}_0\left\{S(|Z|_{(d-j+1)})\right\} - d\lambda_0^{(j)}(1-\lambda_0^{(j)})}{d(d-1)\lambda_0^{(j)}(1-\lambda_0^{(j)})}, \end{aligned}$$

where E_0 and Var_0 denote the expectation and variance conditional on $\mathbf{Z} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. Similarly, we solve for $(\lambda_a^{(j)}, \gamma_a^{(j)})$ using the same equations except with E_a and Var_a, which are the expectation and variance conditional on $\mathbf{Z} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Analogous to the BJ statistic, $\boldsymbol{\mu} = (\hat{\mu}_{j,d}, ..., \hat{\mu}_{j,d})_{d\times 1}^T$ where $\hat{\mu}_{j,d} > 0$ is again the root of (2.3).

The first moment matching equation is simple to solve, since clearly $E_0\left\{S(|Z|_{(d-j+1)})\right\} = 2d\bar{\Phi}(|Z|_{(d-j+1)})$ and $E_a\left\{S(|Z|_{(d-j+1)})\right\} = j$. The variance term in the second equation is more difficult. We can use Theorem 1 of Barnett et al. (2016) for Var_0. For Var_a, we need the following theorem:

Theorem 1: Define $\bar{r^i} = \frac{2}{d(d-1)}\sum_{1\leq k<l\leq d}\left(\sum_{kl}\right)^i$ and let $\mathcal{H}_i(t)$ be the probabilists' Hermite polynomials. Under the binomial likelihood alternative $\mathbf{Z} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\hat{\mu}_{j,d}, ..., \hat{\mu}_{j,d})_{d\times 1}^T$, the variance of $S(|Z|_{(d-j+1)})$ is given by

$$\begin{aligned} & \text{Var}_a\left\{S(|Z|_{(d-j+1)})\right\} \\ = \; & d(d-1)\left\{\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \phi(t-\hat{\mu}_{j,d})^2 \cdot \sum_{i=1}^{\infty}\frac{H_+^2\bar{r^i}}{i!}\right\} \\ & +d(d-1)\left\{1 - 2\bar{\Phi}(-t-\hat{\mu}_{j,d}) + \bar{\Phi}(-t-\hat{\mu}_{j,d})^2 + \phi(-t-\hat{\mu}_{j,d})^2 \cdot \sum_{i=1}^{\infty}\frac{H_-^2\bar{r^i}}{i!}\right\} \\ & -2d(d-1)\left\{\phi(t-\hat{\mu}_{j,d})\phi(-t-\hat{\mu}_{j,d}) \cdot \sum_{i=1}^{\infty}\frac{H_+H_-\bar{r^i}}{i!}\right\} \\ & +2d(d-1)\left\{\bar{\Phi}(t-\hat{\mu}_{j,d}) - \bar{\Phi}(t-\hat{\mu}_{j,d})\bar{\Phi}(-t-\hat{\mu}_{j,d})\right\} + j\left(1-\frac{j}{d}\right) - d(d-1)\left(\frac{j}{d}\right)^2 \\ H_+ &= \mathcal{H}_{i-1}(t-\hat{\mu}_{j,d}) \\ H_- &= \mathcal{H}_{i-1}(-t-\hat{\mu}_{j,d}). \end{aligned}$$

The proof of this theorem is given in the Supplementary Materials. The terms in the infinite sum shrink very quickly, and in practice we see good accuracy using only the first seven.

After matching all four parameters $(\lambda_0^{(j)}, \gamma_0^{(j)}\lambda_a^{(j)}, \gamma_a^{(j)})$, we calculate

$$GBJ_d^{(j)} = \log\left\{\frac{\Pr\left(V_a^{(j)} = j; d, \lambda_a^{(j)}, \gamma_a^{(j)}\right)}{\Pr\left(V_0^{(j)} = j; d, \lambda_0^{(j)}, \gamma_0^{(j)}\right)}\right\}.$$

The maximum value of $GBJ_d^{(j)}$ among all qualifying j is then the observed Generalized Berk-Jones statistic.

2.3.4 Analytic p-value calculation for the Generalized Berk-Jones statistic

Let G_d be a general supremum-based global statistic such as the GBJ statistic. Suppose G_d is constructed using independent marginal test statistics $Z_1, ..., Z_d$. Denote the observed value of this statistic by g, where higher values of g indicate more evidence for the alternative. As noted by Moscovich-Eiger and Nadler (2017), the p-value for g can often be written

$$\Pr\left(G_d \geq g\right) = 1 - \Pr\left\{\forall j = 1, 2, ..., d : |Z|_{(j)} \leq b_j \middle| Z_j \overset{iid}{\sim} N(0,1)\right\},$$

where $0 \leq b_1 \leq b_2 \leq ... \leq b_d$ are 'boundary points' that come from inversion of the test statistic. Moscovich-Eiger and Nadler (2017) proposed a method that can calculate the p-value of G_d very quickly if $Z_1, ..., Z_d$ are independent. However when $Z_1, ..., Z_d$ are correlated, their techniques for a fast calculation are not applicable.

An exact p-value for GBJ, and for any global test applied to correlated observations, must take into account the covariance structure of $\mathbf{Z}$. The p-value for these tests is then

$$\Pr\left(G_d \geq g\right) \quad = \quad 1 - \Pr\left\{\forall j = 1, 2, ..., d : |Z|_{(j)} \leq b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}. \tag{2.5}$$

We are unaware of any computationally feasible expressions to calculate the distribution of the order statistics $|Z|_{(1)}, ..., |Z|_{(d)}$ when d is moderate or large. The Supplementary Materials provides a procedure to compute this probability by partitioning the region into

$d!$ separate sections, but the method is very computationally expensive and not feasible for use with d greater than 10.

However, an alternative way to write the rejection region of equation (2.5) is

$$\Pr\left\{\forall j : |Z|_{(j)} \leq b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\} = \Pr\left\{\forall j : S(b_j) \leq (d-j) \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}. \quad (2.6)$$

The right hand side of (2.6) suggests that the quantity can be calculated recursively. Indeed, define $b_0 = 0$ and $q_{j,a} = \Pr\left\{S(b_j) = a, \bigcap_{k=1}^{j-1} S(b_k) \leq d-k\right\}$. The quantity in (2.6) is just $q_{d,0}$ and can be calculated recursively as

$$\begin{aligned} q_{j,a} &= \sum_{m=a}^{d-j+1} \Pr\left\{S(b_j) = a, S(b_{j-1}) = m, \bigcap_{k=1}^{j-2} S(b_k) \leq d-k\right\} \\ &= \sum_{m=a}^{d-j+1} \Pr\left\{S(b_j) = a \middle| S(b_{j-1}) = m, \bigcap_{k=1}^{j-2} S(b_k) \leq d-k\right\} q_{j-1,m} \\ &\approx \sum_{m=a}^{d-j+1} \Pr\left\{S(b_j) = a \middle| S(b_{j-1}) = m\right\} q_{j-1,m}. \end{aligned} \quad (2.7)$$

We use an EBB approximation to calculate the first probability in equation (2.7), with the equations

$$\begin{aligned} \lambda_j &= \frac{\bar{\Phi}(b_j)}{\bar{\Phi}(b_{j-1})} \\ \frac{\gamma_j}{1+\gamma_j} &= \frac{\sum_{u<v}\left[\frac{\Pr\{|Z_u|,|Z_v| \geq b_j\}}{\Pr\{|Z_u|,|Z_v| \geq b_{j-1}\}} - \left\{\frac{\bar{\Phi}(b_j)}{\bar{\Phi}(b_{j-1})}\right\}^2\right]}{d(d-1)\lambda_l(1-\lambda_l)} \end{aligned}$$

to match the moments. Finally, set $\Pr\{S(b_j) = a | S(b_{j-1}) = m\} := \Pr(V_j = a)$ where $V_j \sim EBB(m, \lambda_j, \gamma_j)$. Evaluation of $\Pr(|Z_j|, |Z_k| \geq b_j)$ follows from steps similar to the proof of Theorem 1.

Note that we can generalize the scheme described above to calculate p-values for many different supremum-based global tests by adopting the general approach of Moscovich-Eiger and Nadler (2017). As long as the test statistic can be inverted to create the bounds $b_1, ..., b_d$, we can use equation (2.7) to calculate its p-value when applied to correlated observations. In particular we can use this procedure to perform p-value calculations for the HC, GHC, BJ, and GBJ statistics. Both the calculation and the Gen-

eralized Berk-Jones test are implemented in the R package `GBJ`, freely available on the `CRAN` repository.

2.3.5 The omnibus test

While we will show that the Generalized Berk-Jones test possesses an attractive finite sample rejection region when signals are moderately sparse, GBJ may also lose power in the presence of very sparse or dense signals. As SNP-set inference involves testing for a composite alternative $H_1 : \boldsymbol{\beta} \neq 0$, there is no uniformly optimal test for both sparse and dense alternatives. As signal sparsity varies between genes, the best test will also change from gene to gene, but it is unknown prior to scanning the genome. Thus we propose an omnibus test that offers robust power over a range of different sparsity levels.

The omnibus test is constructed by combining the SKAT, GBJ, GHC, and minimum p-value statistics, which have been described above. The motivation for choosing these four methods is to combine tests that are known to have good power when signals are dense, moderately sparse, very sparse, and the sparsest possible, respectively. The MinP method uses the set's largest marginal test statistic in magnitude $|Z|_{(d)}$ as a test statistic. When the Z_j are independent, Donoho and Jin (2004) showed that MinP asymptotically reaches the same detection boundary as HC and BJ in the very sparse regime $k \leq d^{1/4}$, but not the moderately sparse regime $d^{1/4} < k < d^{1/2}$. In finite samples, MinP can have better power than the other three tests when there are only one or two causal SNPs. In contrast, SKAT is known to have high power when signals in a SNP-set are dense.

The omnibus test first applies each of the four tests to the same SNP-set, and then it carries forward the smallest p-value from the four tests as a test statistic. Specifically, the omnibus test statistic is defined as:

$$OMNI = \min\left(p_{GBJ}, p_{GHC}, p_{SKAT}, p_{MinP}\right),$$

where $p_{GBJ}, p_{GHC}, p_{SKAT}$, and p_{MinP} denote the p-values of the four tests applied on the same SNP-set. As these tests are applied to the same data, the four p-values will be correlated.

Calculations of the p-value for OMNI must again account for the correlation be-

tween tests. We employ a Gaussian copula approximation for the joint distribution of the inverse-normal transformed p-values:

$$p_{OMNI} = 1 - \Phi_M \left[\left\{ \Phi^{-1}(OMNI), \Phi^{-1}(OMNI), \Phi^{-1}(OMNI), \Phi^{-1}(OMNI) \right\}_{4\times 1}; \mathbf{R}_{4\times 4} \right],$$

where $\Phi_M(\cdot; \mathbf{R})$ denotes the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and correlation matrix $\mathbf{R}$. The correlation matrix $\mathbf{R}$ of the four component test statistics is estimated through parametric bootstrap under the null. For each subject i in the study, we simulate a new outcome based on the null mean $\hat{\mu}_{0i}$. When individual-level data is not available, we take $\hat{\mu}_{0i}$ to be the same constant for all subjects as an approximation. Then each of the four tests are applied with the simulated outcome instead of the original one. The original design matrix, or approximated design matrix if working with summary statistics, is used each time. Each of the four p-values is then inverse-normal transformed; under the null hypothesis the four transformed values have marginal normal distributions with mean zero. As we only need to estimate the correlation matrix $\mathbf{R}$, only a small number of parametric bootstrap samples are needed. In practice, this procedure is repeated 100 times, and then we set $\mathbf{R}$ equal to the sample correlations of the inverse-normal transformed statistics. We will see that this omnibus test performs well across a variety of settings.

2.4 Rejection region analysis of different SNP-set tests

We study in this section the finite sample rejection regions for the BJ, GBJ, HC, and GHC tests, and we advocate for viewing these statistics as boundary-defining algorithms. By using the p-value calculation from Section 2.3.4, we can employ standard root-finding routines to find the observed value g which would result in a p-value of α for a given SNP-set size, correlation structure, and global test statistic. Then by inverting g to find the boundary points $b_1, ..., b_d$, we can define the rejection region as bounds on the observed test statistic magnitudes. Plotting these bounds for different tests shows us exactly what types of signals a given test is well-powered to detect.

To numerically compare the rejection boundaries, consider a simplified model of SNP-set correlation structure where the set is partitioned into only two sections. Let one

section be the independence section, where all SNPs in this portion are completely independent of all other SNPs in the set. Let the other section be the correlated section, where all SNPs in this portion have common pairwise correlation ρ with other SNPs in the section. For our numerical study, $\rho = 0.3$ for the correlated section. We investigate SNP-sets of size $d = 20$ and 100, correlated sections which contain 25%, 50%, and 75% of the SNPs, and tests at size $\alpha = 0.01$. These parameters are chosen to represent reasonable boundaries on the correlation structures seen in common GWAS data; Dawson et al. (2002) estimated that the average r^2 between SNPs separated by 100kb is around 0.1.

The rejection regions for all each SNP-set are plotted in Figure 2.1. At the jth coordinate on the x-axis, if the observed $|Z|_{(j)}$ lies above the boundary of a particular test at that coordinate, then we would reject the null for that test at level $\alpha = 0.01$. The lines on the graph are added to aid in visualization, but there should be no interpretation of interpolation between the points. It does not make sense to think of the boundary at $|Z|_{(2.5)}$, for example. While standard methods for inference on HC and BJ are invalid in the presence of correlation, valid p-values for these tests can be computed with the same ideas we have introduced for GBJ inference, specifically following equation (2.7). Thus we can show that HC and BJ sometimes have much less desirable rejection regions when SNPs in a set are correlated.

One of the clearest trends from Figure 2.1 is that the HC and GHC boundaries are lower for a small region around $|Z|_{(d)}$, and then the BJ and GBJ boundaries quickly become smaller as we move left. This behavior indicates that HC and GHC are better at detecting the sparsest alternatives with only one or two signals, as those signals would almost always manifest as the test statistics with the largest magnitude. In contrast, the plots demonstrate that BJ and GBJ can have more power to detect weaker, less sparse signals which may be more easily found by examining the test statistic which is, say, fifth or tenth largest in magnitude. The boundaries of HC and GHC can drop below BJ and GBJ again for the smallest observed magnitudes, but signals would only be found in these observations if they are particularly dense, a setting which is not the focus of our efforts. The intuition we can glean from this figure is closely aligned with the theoretical development of Donoho and Jin (2004) and the simulations of Li and Siegmund (2015) when

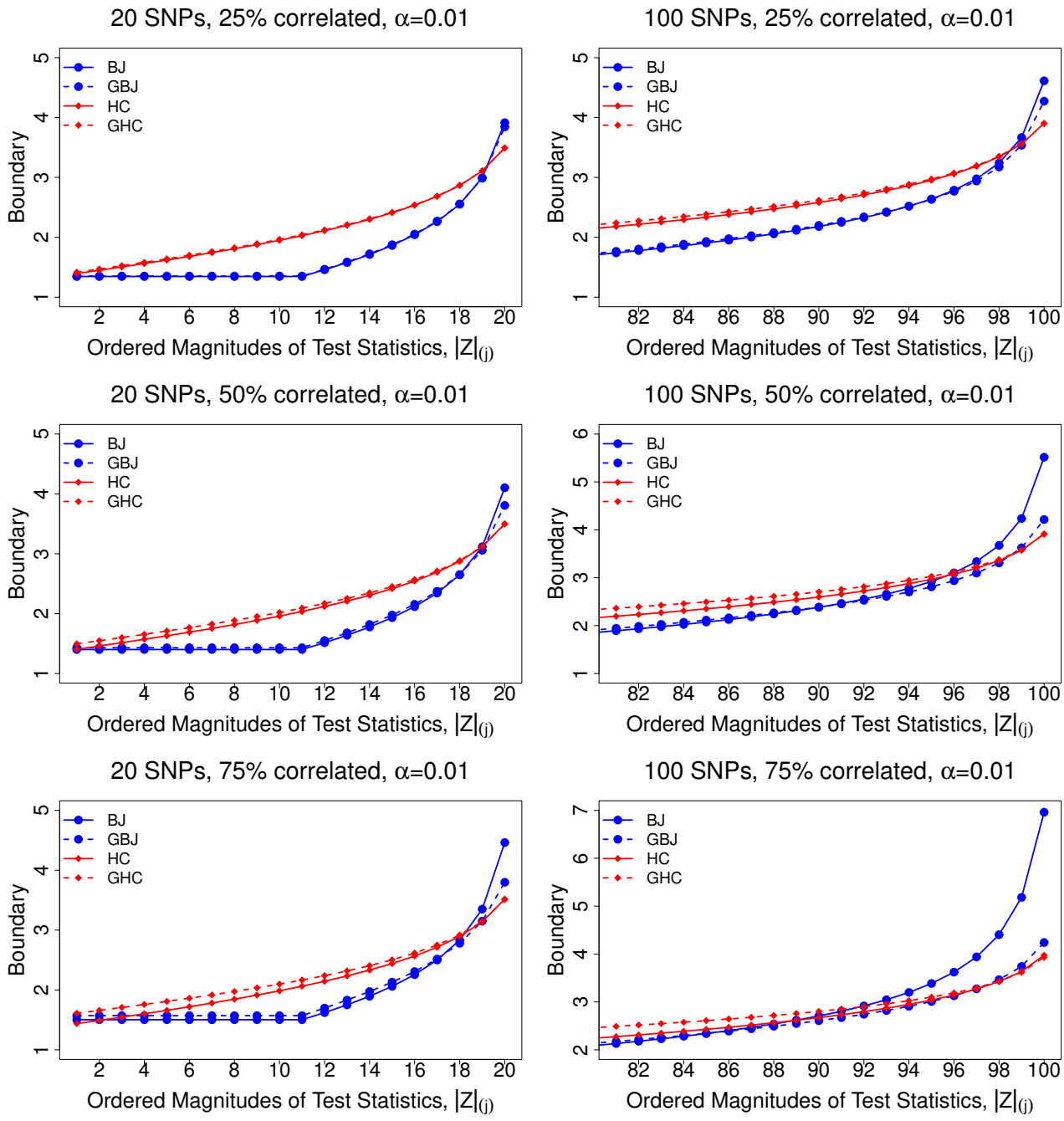

Figure 2.1: Rejection region of Berk-Jones, Generalized Berk-Jones, Higher Criticism, and Generalized Higher Criticism tests, plotted according to the order statistics of the absolute values of the test statistics. At each point j on the x-axis, if the jth smallest test statistic in magnitude is greater than the boundary point for a specific test at j, then we would reject the null using that test at level $\alpha = 0.01$. The difference between BJ and GBJ becomes much more pronounced as both the size of the set and the amount of correlation increase.

the marginal test statistics Z_j are independent. These authors showed that HC is attuned to detect sparse signals arising at the very tail of the observed distribution, while BJ has more power as the number of signals rises.

These results also show why BJ is likely to have low power for detecting sparse signals when the level of correlation is high. When 75% of the SNPs are correlated, the rejection boundary for BJ at the largest few observations is the highest by multiple orders of magnitude on the p-value scale. It would not be desirable to apply BJ in these types of settings, as the test loses an extremely large amount of sensitivity to detect signals in the most outlying values. BJ is still likely to be suitable for detecting dense signals in these situations. Here, GBJ acts as a compromise between BJ and GHC under high correlation. GBJ provides a much lower boundary than BJ at the tail in exchange for slightly higher boundaries near the middle. Thus, GBJ can detect both sparse and dense signals in this example. On the other hand, GBJ provides a slightly higher boundary than GHC at the tail in exchange for lower boundaries past the tail, so it trades some power in the extremely sparse regime for more power to detect moderately sparse signals.

We see that choosing a different statistic is essentially choosing a different boundary-setting algorithm, and this choice should ideally be informed by parameters such as the amount of correlation and estimated sparsity level. Ultimately these plots illustrate that there is no single best global test for all types of alternatives. A genome-wide analysis strategy using the omnibus test will be likely to have robust power across different sparsity settings, correlation structures, and SNP-set sizes.

2.5 Simulation results

2.5.1 Type I error of the Generalized Berk-Jones test

We first illustrate that our p-value calculations for the GBJ and omnibus tests are accurate enough to control the Type I error rate at levels required to declare genome-wide significance of a SNP-set. To replicate the setting of traditional GWAS data, we perform the size simulation on a high-LD subset of the FGFR2 gene and also a low-LD subset of the FGFR2 gene. All SNP-sets are simulated with HAPGEN2 (Su, Marchini and Donnelly,

2011) using the CEU population from HapMap3 as a reference. We choose FGFR2 because it contains both high and low LD regions, and because it will later be the most significant gene in our analysis of the CGEMS data.

In all simulations the outcome is generated as $Y \sim N(0,1)$, and we fit the linear regression model (1) with $\boldsymbol{\beta} = 0$ and $\mathbf{X}_i = 1$. For our SNP-sets, we generate eight pre-determined SNPs from FGFR2 which are known to be in high LD and then eight pre-determined SNPs which are known to be in low LD. Each simulation is repeated 20 million times, and we report the Type I error down to 10^{-5}. Table 2.1 shows that our GBJ p-value calculation is accurate and protects the correct size for correlation structures seen in actual data. The p-value calculation for the omnibus test is similarly accurate at the most stringent significance levels, but it is somewhat conservative at larger significance levels.

Table 2.1: Type I error of GBJ and GHC tests computed over 20 million simulations. The strong LD setting refers to eight SNPs from FGFR2 which are highly correlated. The weak LD setting refers to eight SNPs from FGFR2 which demonstrate only a small amount of correlation with each other.

Significance Level	GBJ, Strong LD	GBJ, Weak LD	OMNI, Strong LD	OMNI, Weak LD
$1 \cdot 10^{-2}$	$8.50 \cdot 10^{-3}$	$9.65 \cdot 10^{-3}$	$6.89 \cdot 10^{-3}$	$7.22 \cdot 10^{-3}$
$1 \cdot 10^{-3}$	$9.18 \cdot 10^{-4}$	$9.67 \cdot 10^{-4}$	$7.16 \cdot 10^{-4}$	$6.95 \cdot 10^{-4}$
$1 \cdot 10^{-4}$	$9.82 \cdot 10^{-5}$	$9.74 \cdot 10^{-5}$	$9.01 \cdot 10^{-5}$	$7.61 \cdot 10^{-5}$
$1 \cdot 10^{-5}$	$1.12 \cdot 10^{-5}$	$9.40 \cdot 10^{-5}$	$1.35 \cdot 10^{-5}$	$9.50 \cdot 10^{-5}$

2.5.2 Power of the Generalized Berk-Jones test under varying sparsity and correlation structures

To study the power of the GBJ, we conduct simulations under a variety of correlation structures and sparsity settings. The performance of the GBJ is compared to GHC, SKAT, the minimum p-value test, and the omnibus test described in Section 2.3.5. For GBJ and GHC, we calculate the p-value through the method described in Section 2.3.4. The MinP test p-value is calculated by casting MinP as a boundary-defining test with $b_j = |Z|_{(d)}$ for all j. For SKAT, we use the corresponding R package.

To study how power is impacted by different correlation structures between the

SNPs, we utilize block correlation structures which are slightly more complex than those used for the rejection region analysis in Section 2.4. Specifically, consider a set of causal SNPs that are correlated amongst themselves with common pairwise correlation ρ_1. All other SNPs are then non-causal, and we allow half of them to have an exchangeable correlation structure with correlation ρ_3; the other half of the non-causal SNPs are completely independent of all other non-causal SNPs. Finally the pairwise correlation between a causal SNP and a non-causal SNP is set at ρ_2. The three correlations ρ_1, ρ_2, ρ_3 will vary between 0 and 0.3. All SNPs are generated to have minor allele frequency of 0.3.

We demonstrate the effects of signal sparsity by using a large SNP-set of d=200 SNPs and varying the number of causal SNPs from $k = 1$ to $k = 15$. This allows us to examine power profiles in the very sparse regime (one to four causal SNPs), in the moderately sparse regime (four to 14 causal SNPs), and at the edge of the dense regime (greater than 14 causal SNPs). Within each set of simulations, we hold constant the percentage of variance in the outcome explained by the causal SNPs. That is, we lower the per-SNP effect size as the number of causal SNPs increases, which can cause the power to fall even as the number of signals grows in Figures 2.2-2.3. Hence in these power curves the main comparison should be made vertically across different methods at the same sparsity, as opposed to horizontally for one method across different sparsity levels. The percentage of variance explained by causal SNPs is set at $R^2 = 0.01$ when the correlation between causal SNPs is $\rho_1 = 0.3$, and it is set at $R^2 = 0.02$ when the correlation between causal SNPs is $\rho_1 = 0$.

The true disease model is

$$Y_i = \sum_{j=1}^{k} \beta_j G_{ij} + \epsilon_i, \epsilon_i \sim N(0, 1), \tag{2.8}$$

where all the β_j are the same and depend on the number of causal SNPs k. We perform 500 simulations at each different value of k and test at $\alpha = 0.01$. Figure 2.2 considers the case where the noise SNPs are independent and Figure 2.3 considers the case where the noise SNPs are correlated. All the power curves are smoothed to show empirical power.

The first significant trend appearing in Figure 2.2 is the effect of sparsity on power. We see that GHC and MinP perform well when the number of causal SNPs is low, as these

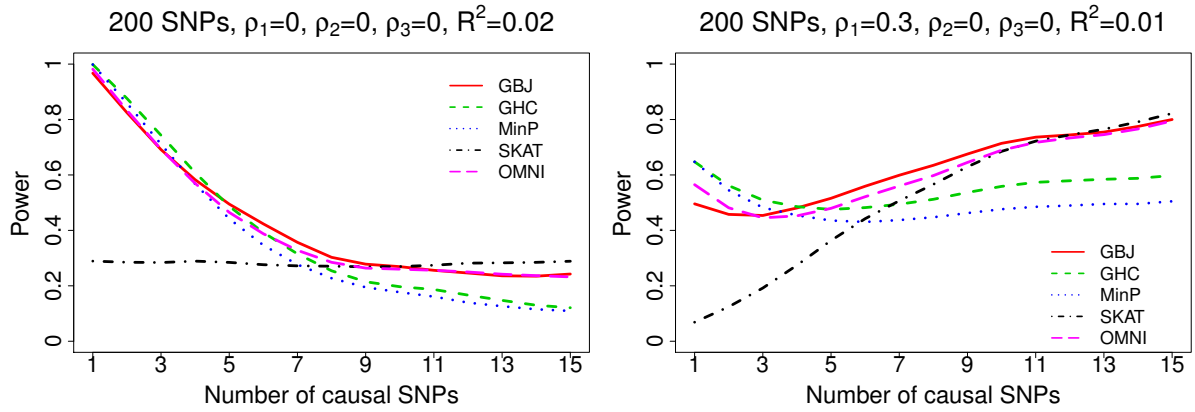

Figure 2.2: Power of set-based tests when noise SNPs are independent. On the left, all SNPs are completely independent of each other, and causal SNPs collectively explain 2% of the variance in the outcome. On the right, causal SNPs are correlated within themselves at $\rho_1 = 0.3$, and they collectively explain 1% of the variance in the outcome. As the number of causal SNPs increases, we decrease the per-SNP effect size so that the percentage of variance explained is constant within each figure, thus power can fall even as the number of causal SNPs increases.

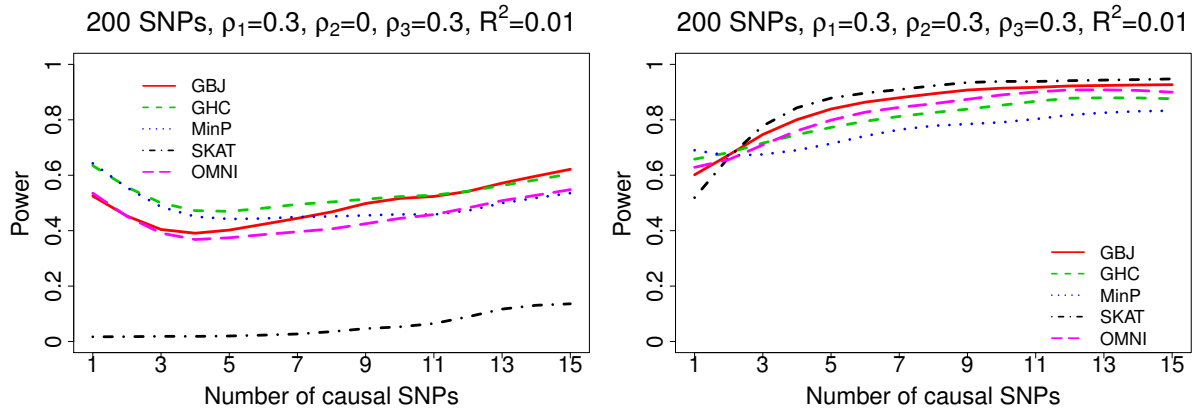

Figure 2.3: Power of set-based tests when there is correlation within noise SNPs (left) and across all SNPs (right). The correlation structure on the right is slightly simpler than the previous three structures, as we switch to an exchangeable correlation matrix in order to accommodate $\rho_2 = 0.3$ while keeping the correlation matrix positive definite. As the number of causal SNPs increases, we decrease the per-SNP effect size so that the percentage of variance explained is constant within each figure, thus the power can fall even as the number of causal SNPs increases.

tests often have the most power in the very sparse regime. In both panels of Figure 2.2, the transition to GBJ having the most power occurs in the moderately sparse regime. Then as the number of causal SNPs increases into the dense regime, SKAT begins to catch up and eventually becomes the most powerful test. This behavior matches our intuition as well as previously published simulation results. GHC and minimum p-value place excess weight on the most outlying observations, so they are well-tuned to detect the very sparse signals. The rejection region of GBJ is better-suited to find moderately sparse signals, and SKAT is known to perform well with dense signals.

The relationship between sparsity and power can be modified by the total amount of correlation. In the left panel of Figure 2.3 we set $\rho_1 = \rho_3 = 0.3$, and MinP and GHC become the top-performing tests for a larger range of sparsity settings, with GBJ losing some of its advantage in the moderately sparse regime. SKAT has almost no power in these situations, as the signals are sparse and there is no correlation between causal and non-causal SNPs. It appears that a large amount of correlation between the non-causal SNPs is detrimental to the performance of GBJ. An explanation for this behavior can be found in the rejection region analysis of Figure 2.1. We see that the bounds of GBJ appear less favorable compared to GHC when the amount of correlation is high. Since over half of the SNPs in Figure 2.3 are correlated with $\rho_1 = \rho_3 = 0.3$, these settings represent a much larger amount of total correlation than was present in Figure 2.2. In the right panel of Figure 2.3 we investigate the setting of $\rho_2 \neq 0$ by using an exchangeable correlation structure, and SKAT dominates as the most powerful test across almost all sparsity levels. Here we break slightly from the above framework by using exchangeable correlation to accommodate $\rho_2 = 0.3$ while still allowing the correlation matrix to be positive definite. GBJ is a close second to SKAT under most sparsity settings with these parameters. SKAT is known to have good performance in the presence of LD between causal SNPs and noise SNPs, which makes signals appear to be dense. The increased density of signals also buoys the performance of GBJ compared to GHC and minimum p-value, which perform the worst under exchangeable correlation.

As the second or third best test across all settings, the omnibus test appears to be robust to LD structure and sparsity. Although it is never the most powerful test, similar

to GBJ, it never loses too much power compared to the best test. This behavior is expected as OMNI integrates information from tests which perform well across multiple sparsity settings.

Overall, GBJ demonstrates good power in a variety of situations, and it is the best-performing test when the level of sparsity is moderate and there is weak correlation among the noise SNPs. The correlation between causal and non-causal SNPs can also be an important driver of performance, as when ρ_2 is nonzero it makes signals appear more dense, so SKAT shows the most power with GBJ as a close second. However unlike SKAT, which can have almost no power in certain situations, GBJ demonstrates more robustness with respect to signal sparsity and between-SNP correlation These results suggest that GBJ is a good choice to use when the signal sparsity is unknown. The omnibus test is also robust to different degrees of sparsity. GHC and MinP outperform when the signal is very sparse, or when there is excess correlation among the noise SNPs. Power for the standard BJ is not plotted in the interest of space, but it behaves like a dense test, similar to SKAT, under correlation. This behavior again matches Figure 2.1, which showed that BJ is more suited to detect dense signals as the amount of correlation increases.

2.5.3 Power of GBJ under actual chromosome 5 correlation structures

We conduct one final simulation to investigate the power of Generalized Berk-Jones under the unstructured LD patterns found in real GWAS data. Blocks of 40 SNPs are chosen at random locations on chromosome 5, and then genotype data are generated using HAPGEN2 (Su et al., 2011), with the entire HapMap3 CEU population used as reference. We choose 40 because this is approximately the median size of genes in our CGEMS analysis below. A total of 160 different blocks on the chromosome are selected, and then 20 are assigned to each sparsity level from one to eight causal genes. We perform 100 simulations for each block, for a total of 2000 at each sparsity setting. We set constant $\beta_j = 0.07$ in equation (2.8). Testing is performed at $\alpha = 1 \cdot 10^{-5}$.

We see in Figure 2.4 that GBJ, GHC, SKAT, and the omnibus test all have very similar power curves in this setting, while minP lags slightly behind. As the number of causal SNPs increases, GBJ demonstrates the best power by a small amount. These results are

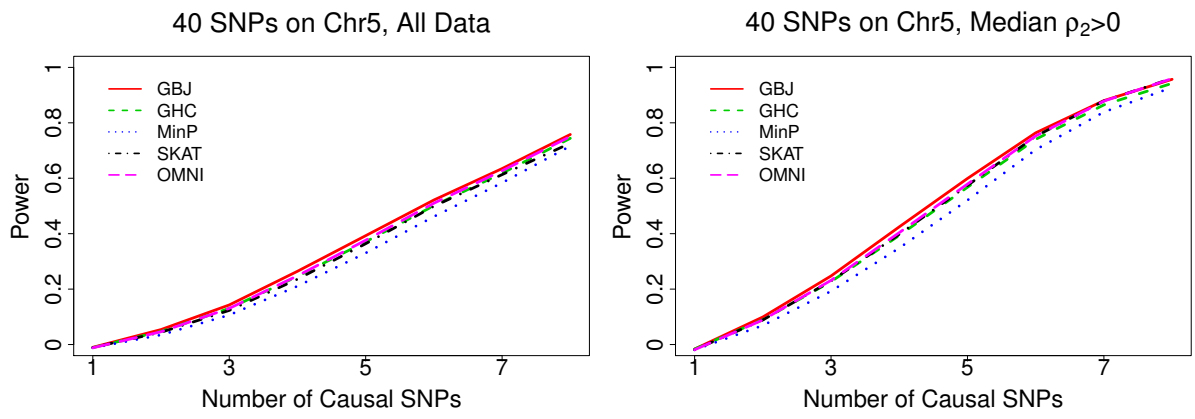

Figure 2.4: Power of set-based tests with correlation structures found in actual chromosome 5 data. On the left, we show all 2000 simulations at each sparsity setting, and on the right we only use simulations where the median value of ρ_2 is greater than 0; this corresponds to roughly half of the data. The effect size is kept constant at $\beta_j = 0.07$, so the power continues to grow as the number of causal SNPs increases. In both panels GBJ is the best-performing test by a small amount as the number of causal SNPs increases.

rather homogenous because sparsity levels are more coarse, and because the parameters are a mix of the values defined in Figures 2.2 and 2.3. When restricting our analysis to the blocks which have median $\rho_2 > 0$, we see that power is higher across the board, but the relative performance of all tests remains approximately unchanged. Median ρ_2 is not a perfect summary measure, as it cannot single-handedly capture the large number of parameters in an unstructured correlation matrix. Further parsing of the data would be necessary to see larger differences in performance. In a practical setting, we might switch between tests based on certain SNP-set characteristics, such as applying GBJ when the set is large and likely to have moderately sparse signals. These results do again demonstrate the robustness and power of GBJ across multiple situations, as it provides the most power across a large portion of the sparse regime.

2.6 Gene-level analysis of the CGEMS GWAS data

The CGEMS breast cancer dataset contains a case-control sample of 1145 breast cancer cases, all postmenopausal women with European ancestry, and 1142 controls recruited from the Nurses' Health Study. These women were genotyped at 528143 SNPs with the Il-

lumina HumanHap500 array. The dataset was originally analyzed by Hunter et al. (2007) in the single-marker GWAS approach. The authors did not find any individual SNPs to reach the genome-wide significance level of 5×10^{-8}, but they highlighted FGFR2 as a strong candidate for future studies based on four SNPs in the gene that showed suggestive association with breast cancer. Such a situation succinctly illustrates the burden of adjusting for multiple comparisons when testing individual SNPs. Gene-level analysis provides an attractive alternative strategy that can reduce the number of comparisons and also aggregate evidence of signals across multiple SNPs in a gene. Here we perform a gene-level analysis to study the association between genes and breast cancer risk.

Since individual-level genotype data were available for this study, we first calculated the marginal test statistics for each SNP using the model in Section 2.2.1. Specifically, we fit a logistic regression model with covariates age and the first three genotype principal components. Then, for each of 14991 genes, we collected the marginal test statistics for all SNPs located within the region defined by that gene. Each gene with more than one marginal SNP test statistic was analyzed with GBJ, GHC, SKAT, MinP, and the omnibus test.

In Table 2.2, we rank the top ten genes according to the smallest p-value produced by any of the five tests. In this sample, GBJ provides the strongest evidence of association for the top four genes, and five of the top ten. Most of these genes are ranked highly by multiple other methods, however no other method provides the lowest p-value for more than two of the top ten genes. In fact, GHC and MinP produce the smallest p-value only once between the two of them. One possible explanation for the underperformance of GHC and MinP is that there may be multiple tagged SNPs surrounding the true causal loci for each of these genes, which could create a lack of extremely sparse alternatives.

The lowest p-value for any gene over all five tests is produced by testing FGFR2 with GBJ, supporting the conclusions of Hunter et al. (2007). Since FGFR2 appears to have signals coming from at least four different SNPs and contains 35 SNPs in total, it would seem to fall into the category of moderate signal sparsity, where GBJ has good performance. Thus we may have expected beforehand that GBJ would be the most powerful test for this gene. FGFR2 has been further validated as a breast cancer associated locus in

Table 2.2: Top significant genes in gene-level analysis of CGEMS breast cancer GWAS data, ranked by minimum p-value produced by any of the five tests. The test which produces the smallest p-value for each gene is highlighted in red.

Gene	GHC	GBJ	MinP	SKAT	OMNI	d
FGFR2	$2.84 \cdot 10^{-5}$	$4.58 \cdot 10^{-6}$	$8.20 \cdot 10^{-5}$	$3.32 \cdot 10^{-5}$	$2.58 \cdot 10^{-5}$	35
CNGA3	$3.00 \cdot 10^{-4}$	$4.04 \cdot 10^{-5}$	$1.75 \cdot 10^{-3}$	$8.34 \cdot 10^{-5}$	$1.84 \cdot 10^{-4}$	26
PTCD3	$1.21 \cdot 10^{-4}$	$5.50 \cdot 10^{-5}$	$3.16 \cdot 10^{-4}$	$1.87 \cdot 10^{-4}$	$6.83 \cdot 10^{-5}$	12
POLR1A	$9.58 \cdot 10^{-5}$	$6.19 \cdot 10^{-5}$	$4.62 \cdot 10^{-4}$	$4.23 \cdot 10^{-4}$	$3.87 \cdot 10^{-4}$	17
ZNF263	$4.89 \cdot 10^{-4}$	$3.90 \cdot 10^{-4}$	$8.09 \cdot 10^{-4}$	$1.26 \cdot 10^{-3}$	$6.84 \cdot 10^{-5}$	3
VWA3B	$4.20 \cdot 10^{-4}$	$2.32 \cdot 10^{-4}$	$1.43 \cdot 10^{-3}$	$1.48 \cdot 10^{-4}$	$4.87 \cdot 10^{-4}$	51
TBK1	$7.04 \cdot 10^{-4}$	$3.35 \cdot 10^{-4}$	$1.27 \cdot 10^{-3}$	$1.48 \cdot 10^{-4}$	$6.05 \cdot 10^{-4}$	11
ABCA1	$3.74 \cdot 10^{-3}$	$1.65 \cdot 10^{-4}$	$7.92 \cdot 10^{-3}$	$4.99 \cdot 10^{-4}$	$2.26 \cdot 10^{-4}$	63
MMRN1	$2.31 \cdot 10^{-4}$	$5.51 \cdot 10^{-4}$	$1.72 \cdot 10^{-4}$	$3.34 \cdot 10^{-2}$	$7.73 \cdot 10^{-4}$	10
TIGD7	$5.79 \cdot 10^{-4}$	$3.78 \cdot 10^{-4}$	$1.32 \cdot 10^{-3}$	$1.33 \cdot 10^{-3}$	$2.05 \cdot 10^{-4}$	4

multiple follow-up studies (Meyer et al., 2008; Liang et al., 2008).

Besides FGFR2, genes such as PTCD3 and POLR1A have also been implicated as risk loci in independent investigations (Boehm et al., 2007; Jia et al., 2011). The overlap of our findings with other studies and other statistics provides a level of reassurance that GBJ performs well in identifying truly significant genes and not simply spurious associations. Alternately, ABCA1 is an example of a gene that may not have received further scrutiny if we were not utilizing the GBJ test. ABCA1 expression has been linked with breast cancer risk (Smith and Land, 2012), but MinP and GHC do not provide the same strength of evidence that GBJ does. It seems likely that there are more than a few signal SNPs in ABCA1, especially since ABCA1 contains a relatively large number of SNPs compared to the other genes in this dataset.

Perhaps due to the limited sample size, no test produces a p-value low enough to be declared significant after Bonferroni correction for 14991 genes. Still, this analysis highlights the advantages of Generalized Berk-Jones compared to alternative tests. The GBJ p-value for FGFR2 does come very close to the Bonferroni-corrected level (3.34×10^{-6}), and it certainly provides more evidence of association than the single SNP statistics. Additionally, GBJ often gives the highest measure of significance, and never the lowest, in the genes displayed, demonstrating its robustness across different set sizes and LD patterns.

2.7 Discussion

We have proposed the Generalized Berk-Jones statistic to test for association between a SNP-set and an outcome. Our GBJ generalizes the standard Berk-Jones by modifying the BJ statistic to directly account for the correlation between individual SNPs. This modification results in a test that is more powerful when SNPs are in LD. We also provide an analytic p-value calculation for GBJ and generalize it to a class of supremum-based global tests, allowing valid inference for HC, GHC, BJ, and other methods when these procedures are applied as SNP-set tests using correlated marginal test statistics. Rejection region analysis demonstrates that GBJ can be described as a compromise between Berk-Jones and Higher Criticism-type tests in terms of finite sample performance.

For example, while our numerical analysis shows situations where GBJ does not set the lowest boundary at either $|Z|_{(d)}$ or $|Z|_{(d/2)}$, GBJ generally comes very close to the lowest boundary at both locations, which affords it both robustness to signal sparsity and power to detect moderately sparse signals. GHC and HC often set the lowest boundary around $|Z|_{(d)}$, but in return they concede a large amount of volume past the first few most extreme observations, which lowers power in the moderately sparse regime. BJ frequently sets the lowest boundary past the tail, but its tail boundary can be orders of magnitude larger than that of GBJ, HC, and GHC. Bounds in the expected signal regions must be viewed holistically, so slightly lower bounds at a few locations are not necessarily desirable if the price is much higher bounds in other signal locations, as in the case of BJ. Thus GBJ offers good power to detect moderately sparse effects without losing too much power when single-SNP signals are extremely sparse.

Simulation results reinforce the conclusions we find from examining the rejection regions of GHC and GBJ. Additionally we see that the MinP test performance is quite good when signals are very sparse, similar to GHC, but MinP does not perform as well as GHC when signals become more dense. SKAT has a unique power profile, as it can be very powerful when signals are dense or there is correlation between causal and non-causal SNPs, but it is also not robust to different correlation structures and will often have very little power when there is no correlation between causal and non-causal SNPs. The om-

nibus test is another test which offers robust power across different sparsity levels, and while it is never the best test, it also never has the worst power. When applied to data from the CGEMS study, we see that GBJ often produces the most significant p-values, perhaps owing to its versatility across different parameter settings.

In demonstrating that the BJ statistic can be adapted for increased robustness to correlation, we have also demonstrated that these types of boundary-defining algorithms can be modified to increase finite sample power under specific set-level parameters. It would be of interest to develop different boundary-defining methods that offer more favorable rejection regions in narrow but well-defined settings. For example, we see that GBJ's advantage over GHC in the moderately sparse regime tends to decrease as the level of correlation increases. It would be convenient to have another test which dominates both GHC and GBJ when the set is large and the level of correlation is extremely high, for the rare occasion that we find such a set in the data. While this kind of test may possess poor boundaries in the majority of common situations, it could be very useful in a toolbox-type approach that matches each SNP-set with a specially tuned test.

In a similar vein, it would be interesting to understand the boundary shapes for other previously proposed boundary algorithms (Jager and Wellner, 2007) in the class of Berk-Jones and Higher Criticism. While many of these algorithms share the same asymptotic guarantees of BJ and HC, little is known about their comparative finite sample performance, especially when observations in a set are correlated. These other methods might also have great value as part of the toolbox technique mentioned above.

As genomic data collection techniques continue to evolve, it may be necessary to adapt the GBJ as well. In particular, the rise of whole genome sequencing and fine mapping studies is leading to the discovery of more SNPs with extremely rare minor alleles. Marginal test statistics generated from these SNPs are known to be non-Gaussian in finite samples, and thus they will not have the distribution we assume for GBJ. GBJ will need to account for null distributions that are not standard normal before SNP-sets containing these rare variants can be tested.

2.8 Appendix

2.8.1 Proof of Theorem 1 from Section 2.3.3

We are interested in the variance of

$$\begin{aligned} S(t) &= \sum_{j=1}^{d} \mathbf{1}\left(|Z_j| \geq t\right), \\ \mathbf{Z} &\sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned}$$

Assume that all the Z_j have a common mean with $\boldsymbol{\mu} = (\hat{\mu}_{,j,d}, ..., \hat{\mu}_{j,d})$. Then the variance is given by

$$\begin{aligned} \text{Var}\left\{S(t)\right\} &= n\pi(1-\pi) + 2\sum_{1\leq j<k\leq d} \text{Cov}(Z_j, Z_k), \\ \pi &= 1 - \left\{\bar{\Phi}(t-\hat{\mu}_{j,d}) - \bar{\Phi}(-t-\hat{\mu}_{j,d})\right\}, \end{aligned}$$

and term involving the covariances can be rewritten as

$$\begin{aligned} 2\sum_{1\leq j<k\leq d} \text{Cov}(Z_j, Z_k) &= 2\sum_{1\leq j<k\leq d} \left\{\Pr\left(|Z_j|, |Z_k| \geq t\right) - \pi^2\right\} \\ &= -d(d-1)\pi^2 + 2\sum_{1\leq j<k\leq d} \Pr\left(Z_j, Z_k \geq t\right) + 2\sum_{1\leq j<k\leq d} \Pr\left(Z_j, Z_k \leq -t\right) \\ &\quad +2\sum_{1\leq j<k\leq d} \Pr\left(Z_j \leq -t, Z_k \geq t\right) + 2\sum_{1\leq j<k\leq d} \Pr\left(Z_j \geq t, Z_k \leq -t\right). \end{aligned}$$

Each of the four types of probabilities can be reexpressed using the standard Mehler kernel for the bivariate normal distribution:

$$\begin{aligned} &\sum_{1\leq j<k\leq d} \Pr\left(Z_j, Z_k \geq t\right) \\ = &\sum_{1\leq j<k\leq d} \int_t^\infty \int_t^\infty \phi_2\left\{(z_j - \hat{\mu}_{j,d}), (z_k - \hat{\mu}_{j,d}); \rho_{jk}\right\} dz_j dz_k \\ = &\sum_{1\leq j<k\leq d} \int_t^\infty \int_t^\infty \phi(z_j - \hat{\mu}_{j,d})\phi(z_k - \hat{\mu}_{j,d}) \sum_{r=0}^{\infty} \frac{\rho_{jk}^r}{r!} \mathcal{H}_r(z_j - \hat{\mu}_{j,d})\mathcal{H}_r(z_k - \hat{\mu}_{j,d}) dz_j dz_k \\ = &\sum_{1\leq j<k\leq d} \left\{\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \phi(t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})^2}{r!} \rho_{jk}^r\right\} \\ = &\frac{d(d-1)}{2}\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \phi(t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})^2}{r!} \sum_{1\leq j<k\leq d} \rho_{jk}^r \\ = &\frac{d(d-1)}{2}\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \frac{d(d-1)}{2}\phi(t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})^2}{r!} \bar{\rho^r}. \end{aligned}$$

Here $\phi_2(x, y; \rho)$ represents the standard bivariate normal distribution with mean 0, unit variances, and correlation parameter ρ. Also, $\bar{\rho^r} = 2\sum_{1\le j<k\le d} \rho_{jk}^r / \{d(d-1)\}$. We skip the similar derivation for the other three probabilities and give only the final expressions:

$$\begin{aligned}
& 2\sum_{1\le j<k\le d} \Pr\left(Z_j, Z_k \ge t\right) \\
= \;& d(d-1)\left\{\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \phi(t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})^2}{r!}\bar{\rho^r}\right\}, \\
& 2\sum_{1\le j<k\le d} \Pr\left(Z_j, Z_k \le -t\right) \\
= \;& d(d-1)\left\{1 - 2\bar{\Phi}(-t-\hat{\mu}_{j,d}) + \bar{\Phi}(-t-\hat{\mu}_{j,d})^2 + \phi(-t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(-t-\hat{\mu}_{j,d})^2}{r!}\bar{\rho^r}\right\},
\end{aligned}$$

and

$$\begin{aligned}
& 2\sum_{1\le j<k\le d} \Pr\left(Z_j \le -t, Z_k \ge t\right) \\
= \;& d(d-1) \times \left\{\bar{\Phi}(t-\hat{\mu}_{j,d}) - \bar{\Phi}(t-\hat{\mu}_{j,d})\bar{\Phi}(-t-\hat{\mu}_{j,d}) - \phi(t-\hat{\mu}_{j,d})\phi(-t-\hat{\mu}_{j,d}) \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})\mathcal{H}_{r-1}(-t-\hat{\mu}_{j,d})}{r!}\bar{\rho^r}\right\}.
\end{aligned}$$

So in total we have:

$$\begin{aligned}
& 2\sum_{1\le j<k\le d} \mathrm{Cov}(Z_j, Z_k) \\
= \;& -d(d-1)\pi^2 + d(d-1)\left[\bar{\Phi}(t-\hat{\mu}_{j,d})^2 + \phi(t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})^2}{r!}\bar{\rho^r}\right] \\
& + d(d-1)\left(1 - 2\bar{\Phi}(-t-\hat{\mu}_{j,d}) + \bar{\Phi}(-t-\hat{\mu}_{j,d})^2 + \phi(-t-\hat{\mu}_{j,d})^2 \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(-t-\hat{\mu}_{j,d})^2}{r!}\bar{\rho^r}\right) \\
& + 2d(d-1) \times \\
& \left\{\bar{\Phi}(t-\hat{\mu}_{j,d}) - \bar{\Phi}(t-\hat{\mu}_{j,d})\bar{\Phi}(-t-\hat{\mu}_{j,d}) - \phi(t-\hat{\mu}_{j,d})\phi(-t-\hat{\mu}_{j,d}) \sum_{r=1}^{\infty} \frac{\mathcal{H}_{r-1}(t-\hat{\mu}_{j,d})\mathcal{H}_{r-1}(-t-\hat{\mu}_{j,d})}{r!}\bar{\rho^r}\right\}.
\end{aligned}$$

Put it all back together for the result given in the theorem.

2.8.2 Exact p-value calculation using equation 2.5 from Section 2.3.4

We are interested in calculating the probability

$$\Pr\left(G_d \ge g\right) \;=\; 1 - \Pr\left\{\forall j = 1, 2, ..., d : |Z|_{(j)} \le b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}.$$

Using the law of total probability, our quantity of interest is

$$\begin{aligned}
& \Pr\left\{\forall j : |Z|_{(j)} \le b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\} \\
= \;& \sum_{a\in\mathcal{A}} \Pr\left\{\forall j : |Z|_{(j)} \le b_j \text{ and } |Z|_{(j)} = |Z_{a_j}| \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\},
\end{aligned}$$

$$\begin{aligned}\mathcal{A} &= \{(1,2,...,d-1,d),(1,2,...,d,d-1),....,(d,d-1,...,2,1)\}\\ &\quad \text{(All } d! \text{ possible permutations of the integers from 1 to } d).\end{aligned}$$

Thus the p-value can be expressed as

$$\begin{aligned}&\Pr(G \geq g)\\ =\ & 1-\sum_{a\in\mathcal{A}}\Pr\left\{0\leq|Z_{a_1}|\leq b_1,|Z_{a_1}|\leq|Z_{a_2}|\leq b_2,....,|Z_{a_{d-1}}|\leq|Z_{a_d}|\leq b_d|\mathbf{Z}\sim MVN(\mathbf{0},\mathbf{\Sigma})\right\}.\end{aligned}$$

At this point it is apparent that we will need some sort of distribution function for $\mathbf{Y} = |\mathbf{Z}|$, where $\mathbf{Y}$ is the result of applying the absolute value operator on every element of $\mathbf{Z}$. $\mathbf{Y}$ is also known as the multivariate half-normal distribution. If $\mathbf{Z} \sim MVN(\mathbf{0},\mathbf{\Sigma})$, then the probability density function of $\mathbf{Y}$ can be written as

$$\begin{aligned}f_{\mathbf{Y}}(\mathbf{y}) &= \sum_{s\in S}(2\pi)^{-\frac{d}{2}}|\mathbf{\Sigma}_s|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}_s^{-1}\mathbf{y}\right\}, \qquad (2.9)\\ S &= \{(\delta_1,...,\delta_d):\delta_j=\pm1\forall j\},\\ \Lambda_s &= \{\text{diag}(s)\},\\ \mathbf{\Sigma}_s^{-1} &= \Lambda_s\mathbf{\Sigma}^{-1}\Lambda_s.\end{aligned}$$

Note that there are 2^d elements in S. With the use of (2.9), the p-value can be expressed as a d-dimensional integral:

$$\Pr(G\geq g)=1-\sum_{a\in\mathcal{A}}\sum_{s\in S}\int_0^{b_1}\int_{Y_1}^{b_2}...\int_{Y_{d-1}}^{b_d}(2\pi)^{-\frac{d}{2}}|\mathbf{\Sigma}_s^{(a)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(\mathbf{\Sigma}_s^{(a)}\right)^{-1}\mathbf{y}\right\}dY_d...dY_1. \qquad (2.10)$$

By the use of $\Sigma_s^{(a)}$ we mean the variance matrix that is permuted to account for the ordering a. It can be defined as:

$$\begin{aligned}\Sigma_s^{(a)} &= P^{(a)}\mathbf{\Sigma}_s P^{(a)T}\\ P^{(a)} &= \begin{pmatrix} e_{a_1}^T\\ e_{a_2}^T\\ .\\ .\\ e_{a_d}^T\end{pmatrix}\\ P^{(d,d-1,...,3,2,1)} &= \begin{pmatrix} & & & & 1\\ & & & 1 & \\ & & . & & \\ & . & & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix}\end{aligned}$$

where e_i denotes the $d \times 1$ vector with a 1 in the ith position and 0 everywhere else. Although equation (2.10) appears to be calculable through many calls to a multivariate normal distribution solver, the lower bounds are functions of variables in the integration, which is not a feature supported by many statistical computing packages. To put the expression into a form more accessible for computation, we can reinterpret the d-dimensional integral:

$$\begin{aligned}
&\int_0^{a_1}\int_{Y_1}^{a_2}\cdots\int_{Y_{d-1}}^{a_d}(2\pi)^{-\frac{d}{2}}|\mathbf{\Sigma}_s^{(a)}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(\mathbf{\Sigma}_s^{(a)}\right)^{-1}\mathbf{y}\right\}dY_d,...,dY_1 \\
= \;& \Pr\left\{0\le Y_1\le b_1, Y_1\le Y_2\le b_2,...,Y_{d-1}\le Y_d\le b_d; \mathbf{Y}\sim MVN(\mathbf{0},\mathbf{\Sigma}_s^{(a)})\right\} \\
= \;& \Pr\left\{0\le Y_1\le b_1, Y_2\le b_2, Y_3\le b_3,...,Y_d\le b_d, Y_2-Y_1\ge 0,...,Y_d-Y_{d-1}\ge 0; \mathbf{Y}\sim MVN(\mathbf{0},\mathbf{\Sigma}_s^{(a)})\right\}.
\end{aligned}$$

Now we can easily find the distribution of $(Y_1, Y_2, ..., Y_d, Y_2-Y_1, ..., Y_{d-1}-Y_d)$ so that the last line above becomes

$$\begin{aligned}
&\Pr\left\{0\le Y_1\le b_1, Y_2\le b_2, Y_3\le b_3,...,Y_d\le b_d, Y_2-Y_1\ge 0,...,Y_d-Y_{d-1}\ge 0; \mathbf{Y}\sim MVN(\mathbf{0},\mathbf{\Sigma}_s^{(a)})\right\} \\
= \;& \Pr\left(0\le T_1\le b_1, T_2\le b_2, T_3\le b_3,...,T_d\le a_d, T_{d+1}\ge 0,...,T_{2d-1}\ge 0\right), \\
& T\sim MVN\left(\mathbf{0}_{(2d-1)\times 1}, \Delta_d\Sigma_s^{(a)}\Delta_d^T\right), \\
& \Delta_d = \begin{pmatrix}\mathbf{I}_{d\times d}\\ \mathbf{D}\end{pmatrix}_{(2d-1)\times d}, \mathbf{D} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & . & . & \\ & & . & . & \\ & & & -1 & 1\end{pmatrix}_{(d-1)\times d}.
\end{aligned}$$

And the final p-value is given by

$$\begin{aligned}
\Pr\left(G\ge g\right) &= 1-\sum_{a\in\mathcal{A}}\sum_{s\in S}\Pr\left(\mathbf{L}\le T_{a,s}\le\mathbf{U}\right), \qquad (2.11)\\
T_{a,s} &\sim MVN\left(\mathbf{0}_{(2d-1)\times 1}, \Delta_d\Sigma_s^{(a)}\Delta_d^T\right), \\
\mathbf{L} &= (0, \underbrace{-\infty,...,-\infty}_{d-1}, \underbrace{0,...,0}_{d-1}), \\
\mathbf{U} &= (b_1, b_2, ..., b_d, \underbrace{\infty,...,\infty}_{d-1}).
\end{aligned}$$

Equation (2.11) gives us the integral bounds as constants, at a cost of increasing the dimension of the multivariate normal distribution of interest from d to $2d-1$. This final expression can be used in any number of computing packages to produce the desired probability.

Set-based inference for sparse alternatives in genetic association studies

Ryan Sun

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Xihong Lin

Department of Biostatistics

Harvard T.H. Chan School of Public Health

3.1 Introduction

Genetic association studies commonly attempt to perform inference on the association between a single variable of interest and a large number of possibly related features. However, signals among the features are often rare and weak, and the high number of tests creates a large multiple testing burden, so power to detect associations at the individual feature level is low. Common examples of such analyses include Genome Wide Association Studies (GWAS), where only a small number of Single Nucleotide Polymorphisms (SNPs) will show association with an outcome of interest (Manolio et al., 2009), and Phenome Wide Association Studies (PheWAS), which invert the GWAS design by testing many phenotypes for their association with one SNP (Denny et al., 2013).

Because of the challenges associated with testing features individually, it has become increasingly popular to group features into sets and then perform set-based association tests (Lee et al., 2014). For example, in the GWAS setting, SNPs can be grouped into sets by their location in the same gene or pathway (Wu et al., 2011). In the PheWAS framework, diseases can be categorized by the similarity of their symptoms or the biological networks they affect. Set-based tests then offer the ability to increase power by pooling signals to make them stronger (Lee et al., 2012) or by reducing the multiple testing burden. Set-based interpretations may also be more useful to clinicians and other researchers.

As set-based inference has become more prevalent, the number of methods designed to performed this type of inference in rare-weak genetics settings has also grown. Many of these methods drawn inspiration from the Higher Criticism (HC) statistic (Donoho and Jin, 2004), which is notable for developing the idea of a rare-weak detection boundary. In brief, the Higher Criticism, Berk-Jones (BJ), and a class of related tests are asymptotically able to detect the sparsest and weakest signals detectable by any statistical test when features in a set are independent. This optimality property has made it a very attractive choice for genomics applications, where true signal features are expected to be rare and have small effects. In finite samples HC is known to excel at detecting extremely sparse signals, while BJ has more power in moderately sparse settings (Walther, 2013; Li and Siegmund, 2015). However because this class of tests was designed to aggregate inde-

pendent marginal test statistics, the tests cannot be directly applied to genomic features, which are often not independent and will thus produce correlated marginal test statistics.

To adapt the Higher Criticism for use with correlated data, Hall and Jin (2010) proposed the innovated Higher Criticism (iHC), which first decorrelates marginal test statistics through a linear transformation. The original Higher Criticism statistic can then be applied to the decorrelated data. This principle can be applied to other tests in the same family, producing, for example, the innovated Berk-Jones (iBJ). A drawback of the innovation idea is that the transformation has been observed to reduce the magnitude of signals. For example, in the Cancer Genetic Markers of Susceptibility (CGEMS) GWAS, four SNPs in the FGFR2 gene show strong evidence of association with breast cancer risk, producing marginal p-values less than $1.2 \cdot 10^{-5}$. However, after decorrelation, the smallest marginal p-value is only $1.2 \cdot 10^{-4}$, leading to an insignificant iHC p-value for the association between FGFR2 and breast cancer risk. As FGFR2 has been validated to be breast cancer risk gene in multiple follow-up studies (Stevens et al., 2006; Eliassen et al., 2007), this is a disappointing result for iHC and demonstrates the disadvantages of the decorrelation step.

The Generalized Higher Criticism (GHC) (Barnett et al., 2016) and Generalized Berk-Jones (GBJ) (Sun and Lin, 2017) were recently proposed as alternatives to apply the principles of Higher Criticism and Berk-Jones without needing to first transform the data. These tests adapt the HC and BJ test statistics to directly incorporate the correlation among test statistics in a set. Thus the GHC and GBJ provide better finite sample rejection regions than HC and BJ when features are correlated (Sun and Lin, 2017). Applying GHC and GBJ to FGFR2 in the CGEMS data produces gene-based p-values that are approximately two orders of magnitude smaller than iHC, in large part because these tests can be applied directly to the four original single-SNP signals.

In this paper we seek to derive analytically an understanding of the factors which determine performance of set-based tests for sparse alternatives. While development of the GHC and GBJ was originally motivated by dissatisfaction with the attenuation of signals in innovated tests, as in the FGFR2 example above, there are also situations where the decorrelation transformation may increase the magnitude of test statistics. Additionally,

both iHC and GHC are known to provide certain asymptotic guarantees in the same vein as Higher Criticism (Hall and Jin, 2010; Barnett et al., 2016). Thus it is often not clear which test a researcher should choose for any given dataset. In particular, little is known about the performance of innovated methods relative to non-innovated tests. Our goal is to provide some guidance regarding the power of these procedures so that statisticians may perform more informed inference.

In our work we uncover a strong relationship between the correlation structure of features in a set and the resulting performance of innovated tests, but this behavior occurs only in models which test the association between one explanatory variable and a set of outcomes, which we term multiple phenotype testing for its similarity to PheWAS. When the model is designed to test the association between one outcome and a set of explanatory variables, which we term SNP-set testing for its connection to GWAS, the correlation structure of the features produces less of an impact on the power of innovated tests. Based on these findings, we recommend that innovated tests be used in multiple phenotype settings, especially those where the correlation between features is high, while non-transformation methods be used in SNP-set settings. Other elements, for example the strength and direction of individual signals, clearly may also play large roles in performance, but the correlation structure and class of test are generally the only parameters known to the researcher in advance, which motivates our focus.

We further use simulation to demonstrate how correlation, and specifically the correlation between signal and noise features, leads to crucial differences in performance between transformation-based and non-transformation tests in the multiple phenotype setting but not the SNP-set setting. As a case study of how these properties are represented in actual genomic data, we investigate the multiple phenotype setting by testing how individual SNPs affect DNA methylation at various locations along the genome. In the SNP-set setting, we use GWAS data to perform an analysis of the genes that most significantly affect breast cancer. For the remainder of the manuscript we will interchangeably use 'K:1' to denote the multiple phenotype setting with K outcomes and 1 explanatory variable and '1:K' to denote the SNP-set setting with 1 outcome and K explanatory variables.

The rest of the paper is organized as follows. Section 3.2 introduces the models for SNP-set and multiple phenotype association testing. Section 3.2 explains how decorrelation can be cast as an eigendecomposition and demonstrates how correlation structures can upweight or downweight signals differently in the 1:K and K:1 settings. Section 3.4 explores how differences in rejection region and signal sparsity can affect performance of innovated and non-innovated tests. Section 3.5 validates our analytic results through simulation. Section 3.6 presents our applications to searching for regions methylated by lung cancer risk SNPs and discovering genes associated with breast cancer. We conclude with a discussion in Section 3.7.

3.2 Frameworks for SNP-set and multiple phenotype testing

3.2.1 The K:1 multiple phenotype testing framework

In the K:1 multiple phenotype setting, each feature is an outcome, and we are interested in aggregating the results from K different regression models. A typical motivating example for this framework is the search for pleiotropic effects between one SNP and multiple diseases. To simplify the development of our results on the efficacy of decorrelation, we consider the following standardized regression model.

Suppose the data consist of n individuals, and for each individual i we observe $(G_i^{(P)}, \mathbf{Y}_i^{(P)})$ where $G_i^{(P)}$ is a scalar explanatory variable of interest, i.e. the number of minor alleles at a certain SNP, and $\mathbf{Y}_i^{(P)} = \left(Y_{i1}^{(P)}, ..., Y_{iK}^{(P)}\right)$ is a vector of K possibly correlated outcomes. Assume that each $\mathbf{Y}_{.k}^{(P)} = \left(Y_{1k}^{(P)}, ..., Y_{nk}^{(P)}\right)^T$, $k = 1, ..., K$ is centered and scaled to have mean 0 and variance 1, and also assume that $\mathbf{G}^{(P)} = \left(G_1^{(P)}, ..., G_n^{(P)}\right)^T$ is centered with mean 0 and scaled so that $||\mathbf{G}^{(P)}||^2 = 1$. Let the model for the kth phenotype be

$$Y_{ik}^{(P)} = \beta_k^{(P)} G_i^{(P)} + \epsilon_{ik}^{(P)} \tag{3.1}$$

where $\boldsymbol{\epsilon}_i^{(P)} = \left\{\epsilon_{i1}^{(P)}, ..., \epsilon_{iK}^{(P)}\right\}^T \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{K\times K}^{(P)}\right)$ describes the correlation among the phenotypes. Covariates can be incorporated into this model by first regressing the outcomes on

the covariates and then using the regression residuals as the outcomes in equation (3.1).

The global null hypothesis $H_0 : \boldsymbol{\beta}^{(P)} = 0$ corresponds to the situation where no outcomes in the set are associated with our explanatory variable. Under this null, we can calculate a marginal score statistic for each outcome as $\mathbf{Z}_k^{(P)} = \left(\mathbf{G}^{(P)}\right)^T \mathbf{Y}_{.k}^{(P)}$, and we then have

$$\mathbf{Z}^{(P)} \overset{H_0}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{(P)}).$$

Under the alternative, $E\left(\mathbf{Z}^{(P)}\right) = \boldsymbol{\beta}^{(P)}$, which we assume to be a sparse vector.

At this point in the analysis, the choice of set-based test will dictate the next step. Either GHC or GBJ can be applied directly to $\mathbf{Z}^{(P)}$, or the innovated tests can be applied if we first decorrelate $\mathbf{Z}^{(P)}$ through multiplication by a matrix $\mathbf{A}$ which satisfies $\mathbf{A}\boldsymbol{\Sigma}^{(P)}\mathbf{A}^T = \mathbf{I}$. In general $\mathbf{A}$ is not unique, because if $\mathbf{A}\boldsymbol{\Sigma}^{(P)}\mathbf{A}^T = \mathbf{I}$ then $(\mathbf{QA})\boldsymbol{\Sigma}^{(P)}(\mathbf{QA})^T = \mathbf{I}$ also, where $\mathbf{Q}$ is any orthogonal matrix. Hall and Jin (2010) choose $\mathbf{A}$ to be the inverse of the Cholesky decomposition of $\mathbf{Z}^{(P)}$, so that $\mathbf{A} = \left(\mathbf{L}^{(P)}\right)^{-1}$ is the inverse of the unique lower triangular matrix $\mathbf{L}^{(P)}$ such that $\mathbf{L}^{(P)}\left(\mathbf{L}^{(P)}\right)^T = \boldsymbol{\Sigma}^{(P)}$.

In our work, we will find it more convenient to work with the eigendecomposition of $\mathbf{Z}^{(P)}$. Let $\boldsymbol{\Sigma}^{(P)}$ be a symmetric, positive definite covariance matrix. Then it can be decomposed as $\boldsymbol{\Sigma}^{(P)} = \mathbf{U}^{(P)}\boldsymbol{\Lambda}^{(P)}\left(\mathbf{U}^{(P)}\right)^T$ where $\boldsymbol{\Lambda}^{(P)}$ is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}^{(P)}$, sorted so that $\boldsymbol{\Sigma}_{1,1}^{(P)}$ contains the largest eigenvalue and $\boldsymbol{\Sigma}_{K,K}^{(P)}$ contains the smallest. $\mathbf{U}^{(P)}$ is then an orthogonal matrix where the kth column is the eigenvector corresponding to the eigenvalue at $\boldsymbol{\Sigma}_{k,k}^{(P)}$. We can see that

$$\left\{\left(\boldsymbol{\Lambda}^{(P)}\right)^{-1/2}\left(\mathbf{U}^{(P)}\right)^T\right\}\boldsymbol{\Sigma}^{(P)}\left\{\left(\boldsymbol{\Lambda}^{(P)}\right)^{-1/2}\left(\mathbf{U}^{(P)}\right)^T\right\}^T = \mathbf{I}$$

as well. As noted previously, $\left(\boldsymbol{\Lambda}^{(P)}\right)^{-1/2}\left(\mathbf{U}^{(P)}\right)^T = \mathbf{Q}\left(L^{(P)}\right)^{-1}$ where $\mathbf{Q}$ is the orthogonal matrix with columns that are eigenvectors of $\left(\mathbf{L}^{(P)}\right)^T\mathbf{L}^{(P)}$.

Denote by $\mathbf{V}^{(P)}$ the decorrelated test statistics in the multiple phenotype setting. Then under the null hypothesis, $\mathbf{V}^{(P)} \overset{H_0}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$, and under the alternative,

$$E_a\left(\mathbf{V}^{(P)}\right) = \left(\boldsymbol{\Lambda}^{(P)}\right)^{-1/2}\left(\mathbf{U}^{(P)}\right)^T\boldsymbol{\beta}^{(P)}.$$

Note that the inverses of the square roots of the eigenvalues acts as weights on the signal in $\left(\mathbf{U}^{(P)}\right)^T\boldsymbol{\beta}^{(P)}$.

3.2.2 The 1:K SNP-set testing framework

In the 1:K SNP-set setting, each feature is an explanatory variable. A typical motivating example for this framework is testing for the association between one phenotype and all the SNPs in a genetic construct, for instance a gene or a pathway. A naive approach is to fit all K features in the same regression model and perform a K degree of freedom test, but when K is large this test can lose much of its power.

Again assume that we have data for n individuals, and for each individual i we observe $(\mathbf{G}_i^{(S)}, Y_i^{(S)})$, where $Y_i^{(S)}$ is the scalar outcome of interest, and $\mathbf{G}_i^{(S)} = \left(G_{i1}^{(S)}, ..., G_{iK}^{(S)}\right)$ is a vector of K possibly correlated explanatory variables, as in the SNPs located in a gene. Assume that $\mathbf{Y}^{(S)} = \left(Y_1^{(S)}, ..., Y_n^{(S)}\right)^T$ is centered and scaled to have mean 0 and variance 1, and also assume that each $\mathbf{G}_{.k}^{(S)} = \left(\mathbf{G}_{1k}^{(S)}, ..., G_{nk}^{(S)}\right)^T$ is centered with mean 0 and scaled so that $||\mathbf{G}_{.j}^{(S)}||^2 = 1$. Denote by $\mathbf{G}^{(S)} = \left(\mathbf{G}_{.1}^{(S)}, ..., \mathbf{G}_{.K}^{(S)}\right)$ the entire $n \times K$ feature matrix. The correlation matrix of the features is given by $\mathbf{\Sigma}^{(S)} = \left(\mathbf{G}^{(S)}\right)^T \mathbf{G}^{(S)}$.

Let the model for $Y_i^{(S)}$ be

$$Y_i^{(S)} = \beta_1^{(S)} G_{i1}^{(S)} + ... + \beta_K^{(S)} G_{iK}^{(S} + \epsilon_i^{(S)} \tag{3.2}$$

with $\epsilon_i^{(S)} \sim N\left\{0, (\sigma^2)^{(S)}\right\}$. Under the null hypothesis $H_0 : \boldsymbol{\beta}^{(S)} = \mathbf{0}$, a marginal score statistic for each β_k is given by $Z_k^{(S)} = \left(\mathbf{G}_{.k}^{(S)}\right)^T \mathbf{Y}^{(S)}$, and the distribution of the entire vector of test statistics is

$$\mathbf{Z}^{(S)} \overset{H_0}{\sim} N(\mathbf{0}, \mathbf{\Sigma}^{(S)}).$$

Under the alternative, $E\left(\mathbf{Z}^{(S)}\right) = \mathbf{\Sigma}^{(S)} \boldsymbol{\beta}^{(S)}$.

Again, at this point we can either apply GHC or GBJ to $\mathbf{Z}^{(S)}$, or we can choose to decorrelate the $\mathbf{Z}^{(S)}$. When $\mathbf{\Sigma}^{(S)}$ is a symmetric, positive definite covariance matrix, it can be decomposed as $\mathbf{\Sigma}^{(S)} = \mathbf{U}^{(S)} \mathbf{\Lambda}^{(S)} \left(\mathbf{U}^{(S)}\right)^T$. Denote by $\mathbf{V}^{(S)}$ the decorrelated test statistics in the SNP-set setting. Then under the null hypothesis, $\mathbf{V}^{(S)} \overset{H_0}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$, and under the alternative,

$$E_a\left(\mathbf{V}^{(S)}\right) = \left(\mathbf{\Lambda}^{(S)}\right)^{-1/2} \left(\mathbf{U}^{(S)}\right)^T \mathbf{\Sigma}^{(S)} \boldsymbol{\beta}^{(S)} = \left(\mathbf{\Lambda}^{(S)}\right)^{1/2} \left(\mathbf{U}^{(S)}\right)^T \boldsymbol{\beta}^{(S)}.$$

Note that the square roots of the eigenvalues acts as weights on the signal $\left(\mathbf{U}^{(S)}\right)^T \boldsymbol{\beta}^{(S)}$.

3.3 Effect of correlation on signal strength

3.3.1 Eigendecomposition of block correlation matrices

We see that a key advantage of using eigendecomposition to perform the innovation step is that the eigenvalues of the covariance matrix can be viewed as signal weights. To model how correlation structure impacts the power of our set-based tests, we will utilize block matrix structures of the form

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{\Sigma}_2 \\ \mathbf{\Sigma}_2^T & \mathbf{\Sigma}_3 \end{bmatrix} \tag{3.3}$$

where $\mathbf{\Sigma}_1 = \rho_1 \mathbf{1}_{K_1 \times 1} \mathbf{1}_{K_1 \times 1}^T + (1-\rho_1)\mathbf{I}_{K_1 \times K_1}$, $\mathbf{\Sigma}_2 = \rho_2 \mathbf{1}_{K_1 \times 1} \mathbf{1}_{K_0 \times 1}^T$, and $\mathbf{\Sigma}_3 = \rho_3 \mathbf{1}_{K_0 \times 1} \mathbf{1}_{K_0 \times 1}^T + (1-\rho_3)\mathbf{I}_{K_0 \times K_0}$. Here $\mathbf{\Sigma}_1$ represents the correlation between the K_1 causal features, which we model with an exchangeable structure described by ρ_1. Then $\mathbf{\Sigma}_2$ represents the correlations between causal features and noise features, which we set to always be ρ_2. Finally, $\mathbf{\Sigma}_3$ characterizes the correlation between the K_0 noise features, which we also take to be exchangeable with common pairwise correlation ρ_3. In our work we are interested in sparse alternatives, so generally $K_0 \gg K_1$.

Clearly the signal strengths after decorrelation are also dependent on $\left(\mathbf{U}^{(P)}\right)^T \boldsymbol{\beta}^{(P)}$ and $\left(\mathbf{U}^{(S)}\right)^T \boldsymbol{\beta}^{(S)}$, however in practice we never have knowledge of $\boldsymbol{\beta}^{(P)}$ or $\boldsymbol{\beta}^{(S)}$. Thus we restrict ourselves to analysis of the eigenvalues. We do make the one assumption that $\rho_2 \neq 0$ in equation (3.3), so the signal features are not completely independent of the noise. In our simplified block correlation model, $\rho_2 = 0$ means that only the eigenvalues corresponding to $\mathbf{\Sigma}_1$ will be relevant. While this situation can be interesting in its own right, assuming $\rho_2 \neq 0$ will allow for a broader range of results that naturally also cover the more restrictive case.

3.3.2 Eigenvalues as signal weights

Under the structure in equation (3.3), $\mathbf{\Sigma}$ has four distinct eigenvalues. These eigenvalues are

$$\begin{aligned} \lambda_1, \lambda_2 &= \frac{g + \sqrt{g^2 - 4h}}{2}, \frac{g - \sqrt{g^2 - 4h}}{2} \\ g &= 2 + (K_1 - 1)\rho_1 + (K_0 - 1)\rho_3, \end{aligned} \tag{3.4}$$

$$h \;=\; 1 + (K_1 - 1)\rho_1 + (K_0 - 1)\rho_3 + \rho_1\rho_3(K_1 - 1)(K_0 - 1) - \rho_2^2 K_0 K_1,$$

with $\lambda_3 = 1 - \rho_1$ and $\lambda_4 = 1 - \rho_3$. These four will not always fall in the same order, for instance sometimes λ_2 will be the smallest eigenvalue, but we will continue to refer to them with the above labels to prevent confusion. The eigenvalue λ_3 has multiplicity $K_1 - 1$ and λ_4 has multiplicity $K_0 - 1$, for a total of $K = K_0 + K_1$ eigenvalues.

One immediate conclusion we can draw is that a higher level of correlation among the signals, the noise, or both will all lead to larger post-decorrelation signal weights in the multiple phenotype setting and smaller post-decorrelation signal weights in the SNP-set setting. As ρ_1 and ρ_3 increase, λ_3 and λ_4 will continue to fall. Since $\left(\mathbf{\Lambda}^{(P)}\right)^{-1/2}$ is the weight matrix in the K:1 setting while $\left(\mathbf{\Lambda}^{(S)}\right)^{1/2}$ is the weight matrix in the 1:K setting, clearly a larger value of ρ_1 and ρ_3 helps the case for innovation in the K:1 setting, and vice versa in the 1:K setting.

In a certain sense, the increase in ρ_1 and ρ_3 helps the K:1 setting more than it hurts the 1:K setting. Let $w_3^{(P)} = (1 - \rho_1)^{-1/2}$ be the weight corresponding to λ_3 in the K:1 setting and let $w_3^{(S)} = (1 - \rho_3)^{1/2}$ be the weight corresponding to λ_3 in the 1:K setting. Then $\partial w_3^{(P)}/\partial\rho_1 = (1 - \rho_1)^{-3/2}/2$ while $\partial w_3^{(S)}/\partial\rho_1 = -(1 - \rho_3)^{-1/2}/2$, so an increase in ρ_1 causes a sharper increase in $w_3^{(P)}$ than the corresponding decline in $w_3^{(S)}$.

Another important factor in the performance of innovated tests is the value of ρ_2. We can see from equation (3.4) that ρ_2 only impacts two eigenvalues, λ_1 and λ_2 , but these are arguably the most important eigenvalues because they are often either the two largest, or the largest and the smallest, and thus they have outsized effects. As ρ_2 increases, λ_1 increases and λ_2 decreases by the same amount. By an elementary linear algebra fact, the sum of the eigenvalues of $\mathbf{\Sigma}$ is equal to the trace of $\mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ in our work is a correlation matrix, this means that the sum of the eigenvalues is equal to K, and it follows that, for a fixed ρ_1 and ρ_3, an upper bound on λ_1 is $2 + (K_1 - 1)\rho_1 + (K_0 - 1)\rho_3$ while a lower bound on λ_2 is 0.

As ρ_2 increases, λ_1 and λ_2 will separate more and more, and this separation can greatly increase the power of innovated tests in the multiple phenotype setting. To understand why, note that both λ_1 and λ_2 are greater than or equal to 1 when $\rho_2 = 0$. Thus they

are downweighting the signal in the multiple phenotype setting. As λ_1 grows, it continues to provide a poor weight for the K:1 setting, but as λ_2 nears zero, it can dramatically upweight its corresponding signal. More precisely, we have $\partial w_2^{(P)}/\partial \rho_2 = O(\rho_2^{-3/2})$, which can be very large when ρ_2 is near 0. Thus as the correlation between causal and noise features increases, we can expect the decorrelation step to provide more and more power for iHC and iBJ. In fact, the value of ρ_2 matters most when $(K_0 - 1)\rho_3 = (K_1 - 1)\rho_1$. When this occurs, $\partial w_2^{(P)}/\partial \rho_2$ reaches its maximum, and an increase in ρ_2 causes the largest corresponding increase in $w_2^{(P)}$. Under the sparse signal assumption, we have $(K_0 - 1)\rho_3 = (K_1 - 1)\rho_1$ when ρ_1 is much larger than ρ_3.

However ρ_2 does not have the same outsized effect in the SNP-set setting. In the K:1 setting ρ_2 was driving a change from two poor weights to one poor weight and one very good weight, but in the 1:K setting, ρ_2 drives a change from two good weights to one poor weight and one very good weight. The increase in $w_1^{(S)}$ happens at a rate of $\partial w_1^{(S)}/\partial \rho_2 = O(\rho_2)$, and the decrease in $w_2^{(S)}$ occurs at the same rate. Therefore an increase in the correlation between causal and noise features does not provide the same increase in power for innovated tests in the 1:K setting as it did in the K:1 setting.

3.4 Effect of correlation on rejection region and signal sparsity

3.4.1 Varying rejection regions

The GHC and GBJ methods are designed to test whether a vector of marginal test statistics arises from a multivariate normal distribution with mean zero and covariance $\mathbf{\Sigma}$. That is, the null hypothesis changes with the correlation of the features. Therefore the rejection region also changes with $\mathbf{\Sigma}$. However the innovated tests always apply Higher Criticism and Berk-Jones on data that has been decorrelated. Thus for a given K, the rejection regions of iHC and iBJ are constant in $\mathbf{\Sigma}$. This distinction can cause the rejection regions of GHC and GBJ to become less favorable compared to the rejection regions of the innovated tests as the total amount of correlation among the features increases.

Unfortunately we are unaware of any tractable closed-form expression for the rejec-

tion regions corresponding to any of these tests. Instead, following Sun and Lin (2017), we can invert the p-value calculation for each test to numerically determine rejection regions under different correlation structures. Recall the block correlation matrix of Equation (3.3). Although there is no notion of causality in the present discussion, it will be useful to appropriate the structure. Let $\rho_1 = 0$, $\rho_2 = 0$, and $\rho_3 = 0.3$, let $d = 50$, and suppose we let K_0 take the values $K_0 = 13, 25, 38, 50$. As K_0 increases, we are allowing a larger proportion of the SNPs in the set to be correlated, thus raising the total amount of correlation.

Figure 3.1 depicts the rejection regions for iHC and GHC at $\alpha = 0.01$ as K_0 becomes 25%, 50%, 75%, and 100% of the set size. At each point j on the x-axis, if the jth smallest test statistic in magnitude crosses the boundary for a particular setting, then we can reject the null hypothesis for that setting. Although the lines are added to in visualization, we note that the rejection region is only defined at the whole numbers from 1 to K.

We can see that as K_0 increases, the boundaries for GHC also increase slightly for each ordered test statistic magnitude. This is a general trend across GHC and GBJ. On the other hand, the iHC rejection region is constant in K_0, because the test statistics are always assumed to be independent before HC is applied. However the relative superiority of the iHC rejection region as K_0 increases does not necessarily lead to a direct increase in power, as K_0 will alter the form of $\mathbf{\Lambda}$ and $\mathbf{U}$, which are used to decorrelate the data. Rather, the more forgiving region can simply provide a slight advantage to innovated tests as the total amount of correlation increases.

3.4.2 Signal sparsity after transformation

As noted in Section 3.3.1, while we never have knowledge of the product $\mathbf{U}^T\boldsymbol{\beta}$, the resulting vector will often be more dense than the original signal vector $\boldsymbol{\beta}$. The increased density of the signal plays a large role in increasing power for the innovated tests in the multiple phenotype setting. However, in a situation similar to Section 3.3.2, the effect on the SNP-set setting is much less muted.

Recall $E\left(\mathbf{Z}^{(P)}\right) = \boldsymbol{\beta}^{(P)}$ and $E\left(\mathbf{V}^{(P)}\right) = \left(\mathbf{\Lambda}^{(P)}\right)^{-1/2}\left(\mathbf{U}^{(P)}\right)^T\boldsymbol{\beta}^{(P)}$ in the K:1 setting. Thus, before decorrelation the mean vector is only sparsely populated with signals, but

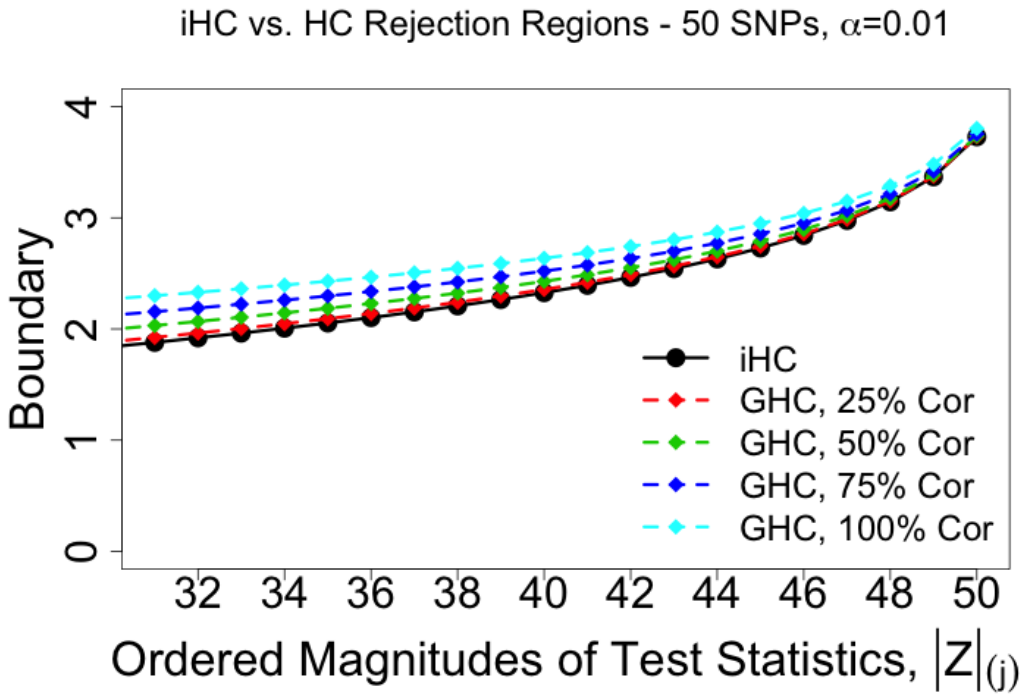

Figure 3.1: Rejection regions of the GHC under four different correlation matrices $\mathbf{\Sigma}$. For each of these correlation patterns, the iHC rejection region is the same. The rejection region of the GHC becomes less and less favorable compared to the iHC as the total amount of correlation rises.

after transformation the mean vector often has a high density of signals. Since the density of signals is intimately connected to the power of these sparse tests (Donoho and Jin, 2004), the decorrelation transformation in the K:1 setting has the potential to give innovated tests distinctly improved performance over the GHC and GBJ. In contrast, $E\left(\mathbf{Z}^{(S)}\right) = \mathbf{\Sigma}^{(S)}\boldsymbol{\beta}^{(S)} = \mathbf{U}^{(S)}\mathbf{\Lambda}\left(\mathbf{U}^{(S)}\right)^T$ and $E\left(\mathbf{V}^{(S)}\right) = \left(\mathbf{\Lambda}^{(S)}\right)^{1/2}\left(\mathbf{U}^{(S)}\right)^T\boldsymbol{\beta}^{(S)}$ in the 1:K setting, so the mean vector is often dense both before and after transformation. An exception is when $\rho_2 = 0$ in our block correlation structure, as noted in Section 3.3.1. However, in general, we would not expect the innovation to have a dramatic impact on sparsity in the SNP-set setting.

3.5 Simulation

We conduct a variety of simulation studies to illustrate how innovation can improve power in the multiple phenotype setting but not the SNP-set setting. Test statistics are

simulated again using the block correlation structure from Section 3.3. We vary the correlation parameters ρ_1, ρ_2, and ρ_3 to demonstrate how the correlation plays a large role in the performance of each type of test. The number of features is fixed at $K = 40$ and the number of causal features is set at $K_1 = 4$ so that effects are moderately sparse. Additionally we let the mean vector of the signal be $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, 0, ...0) / \sqrt{\sum_{j=1}^{4} \mu_j^2 / M}$, where $\mu_j \sim U(-1, 1)$ for $j = 1, 2, 3, 4$ and M is a constant chosen for each setting that allows power to vary widely between 0 and 1. To interpret this mean vector in terms of n and the effect size of each explanatory variable, we can calculate that $R_j^2 \approx \mu_j^2 / n$, where R_j^2 is the percent of variance in the outcome explained by the covariate corresponding to β_j.

Power is calculated empirically from 500 simulations at each 0.01 increment of the correlation parameters, and we test at $\alpha = 0.01$ always. For comparison, we also include a variance-component type test which is known to be powerful for detecting dense signals (Liu and Lin, 2017). This test corresponds to SKAT (Wu et al., 2011) in the 1:K setting. The variance component test can be seen as providing a benchmark for the point at which signals are dense and strong enough that the tests designed for rare-weak regimes lose their superiority.

We begin by demonstrating the large effect that ρ_2 has on innovated tests in the K:1 setting. In Figure 3.2, $\rho_1 = \rho_3 = 0.4$ are fixed, and we vary ρ_2 along the x-axis. As the correlation between signal and noise features increases, we can see that there is a large increase in power of iHC and iBJ in the multiple phenotype framework. In particular, the increase in power begins around the point of $\rho_2 = 0.35$, which is when the eigenvalue λ_2 approaches 1. In other words, at $\rho_2 \approx 0.35$, λ_2 begins to upweight the signal, whereas previously λ_2 was attenuating the signal. Thus the signals in the mean vector of iHC and iBJ become particularly strong as ρ_2 passes this point. In contrast, the power of GHC and GBJ stay constant as ρ_2 varies in the K:1 setting; the signals of the non-innovated tests are not affected by ρ_2 in the same manner. In the SNP-set setting, we see that the powers of all sparse tests remain reasonably constant. As explained in Section 3.3.2, the innovated tests do not receive the same upweighting benefits in the 1:K setting, and following Section 3.4.2, we know that the mean vector is already relatively dense in the original $\mathbf{Z}^{(S)}$, so the

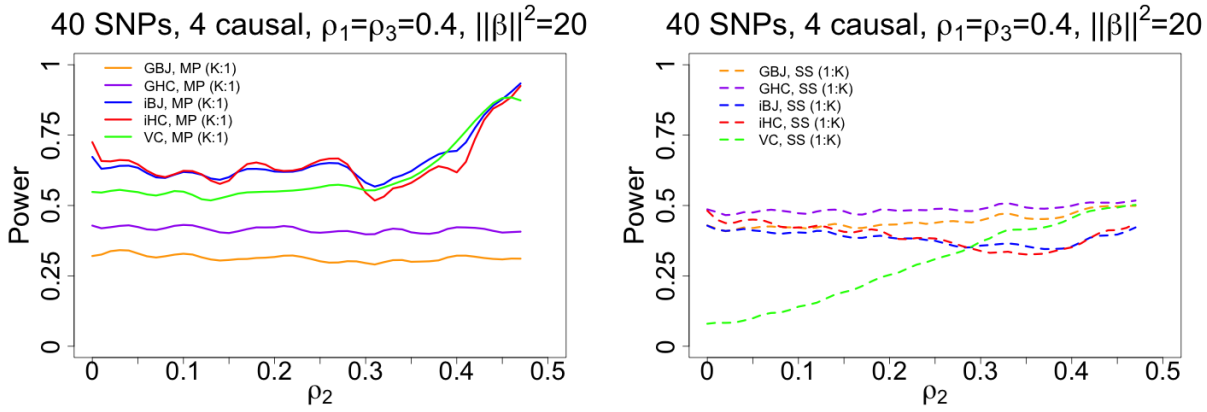

Figure 3.2: Power of iBJ, iHC, VC as ρ_2 increases, for each 0.01 increment of ρ_2 that admits a positive definite correlation matrix. See that the powers of iBJ and iHC greatly increase in the multiple phenotype setting as ρ_2 increases. The powers of GBJ and GHC show almost no change in the K:1 setting because the expectation of the untransformed test statistics does not change in ρ_2. In the 1:K setting, the cumulative change in $w_1^{(S)}$ and $w_2^{(S)}$ is not as favorable, which explains why iBJ and iHC do not show a power increase. Both the expectation of the untransformed test statistics and the rejection region for GBJ and GHC change with ρ_2 in the 1:K setting, and the combined effect appears to give it more power than the innovated tests.

innovation transformation does not create many more signals.

In Figure 3.3, we build upon the previous simulation and consider the situation where ρ_2 has the most weight, which occurs when $(K_0 - 1)\rho_3 = (K_1 - 1)\rho_1$. We set $\rho_1 = 0.7/3$, $\rho_3 = 0.02$, and we again vary ρ_2. Although ρ_2 can only take values up to 0.14, we see that it has the ability to dramatically increase the power of innovated tests in the K:1 setting. Similar to Figure 3.2, the power of non-innovated tests in the K:1 setting and the power of all tests in the 1:K setting is mostly constant as ρ_2 increases. Note that the power of the variance component test in the K:1 setting increases with the innovated tests, suggesting that it benefits from upweighted signals as ρ_2 increases as well.

Finally in Figure 3.4 we investigate varying ρ_1 and ρ_3 as well, utilizing an exchangeable correlation structure with $\rho_1 = \rho_2 = \rho_3 = \rho$ and varying them all at the same time. In the K:1 setting, the power of the innovated tests again increases quite dramatically, starting at almost 0 and increasing to 1 as $\rho = 0.9$. We see that signal strength and sparsity can be overshadowed by correlation strength as the driving parameters of performance,

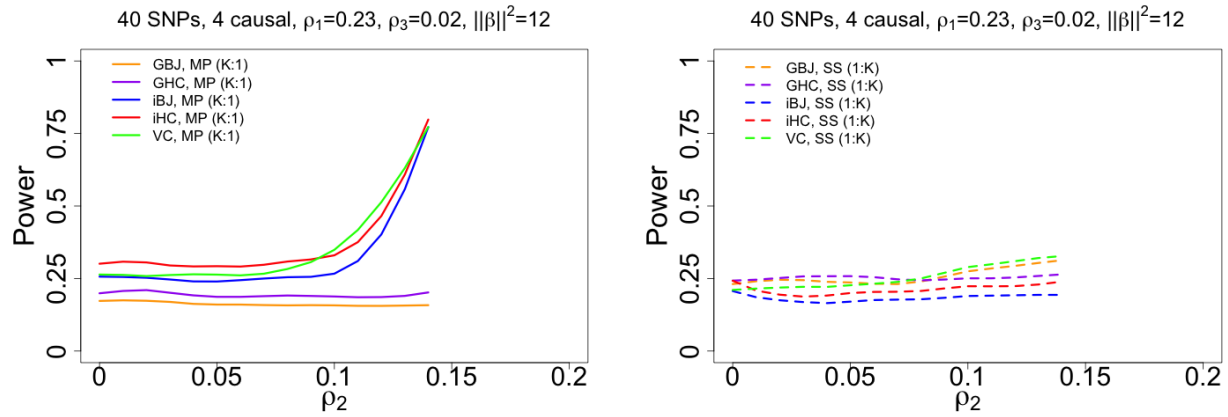

Figure 3.3: Power of iBJ, iHC, VC as ρ_2 increases, for each 0.01 increment of ρ_2 that admits a positive definite correlation matrix. Notice that the powers of iBJ and iHC in the multiple phenotype setting increase even faster than in Figure 3.2. The powers of GBJ and GHC show little change in both the 1:K and K:1 settings.

as the same signal vector can go from almost never detected to almost always detected depending on the correlation between features. This occurs because innovated tests in the K:1 setting have $K_0 - 1$ weights of the form $1/\sqrt{\rho_3}$ and $K_1 - 1$ weights of the form $1/\sqrt{\rho_1}$, so obviously when $\rho_1 = \rho_3 = 0.9$, almost all of the weights are extremely large. However the non-innovated tests do not receive the same benefits and show constant power as the correlation matrix changes. In the 1:K setting, the innovated tests have $K_1 - 1$ weights of the form $\sqrt{\rho_1}$ and $K_0 - 1$ weights of the form $\sqrt{\rho_3}$, which become very small as ρ increases, but there is also one weight of the form $\sqrt{\lambda_1}$, which becomes incredibly large as ρ increases. In fact, this weight becomes so large that the single signal it affects appears to make up for the downweighting of all the other signals. However, performance of the innovated tests still falls below that of the non-innovated tests.

3.6 Application to breast cancer and lung cancer datasets

In this section, we apply innovated and non-innovated tests to case studies in both the K:1 and 1:K settings. For the K:1 setting, we study how one particular lung cancer risk SNP affects DNA methylation across the genome. Sets of DNA methylation probes are grouped according their nearest gene transcription factor start site, and then we test if

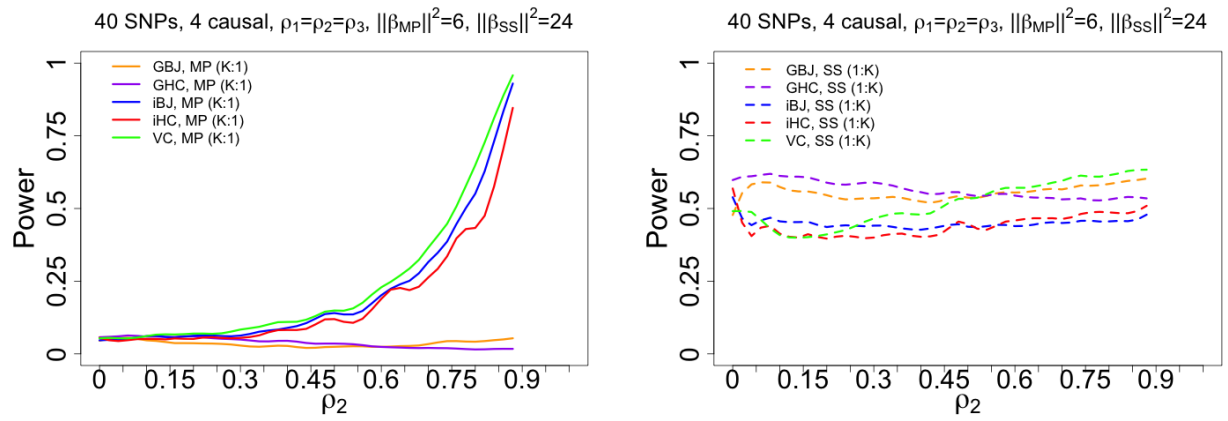

Figure 3.4: Power of iBJ, iHC, VC as ρ_2 increases, for each 0.01 increment of ρ_2 that admits a positive definite correlation matrix. The power for iBJ and iHC in the multiple phenotype setting rises from near 0 to 1 as we change only the exchangeable correlation parameter $\rho_1 = \rho_2 = \rho_3 = \rho$, demonstrating the strong influence of correlation on these tests. The power for GHC and GBJ in the K:1 setting change very little, and the power for all tests in the 1:K setting do not vary much at all for reasons discussed above.

methylation at each group of probes is affected by our SNP of interest. The objective is to discover if the known risk SNP is possibly inducing disease processes through its effect on the methylation patterns of other genes. In the 1:K setting, we investigate which genes are most associated with breast cancer risk. Using GWAS data, we group SNPs into sets according to their location in a gene, and then we test the association between breast cancer disease status and the entire set of SNPs.

3.6.1 Analysis of lung cancer methylation data

Lung cancer was once thought to be a disease with largely environmental causes, in particular, cigarette smoke. However, epidemiological studies as well as multiple GWAS have demonstrated that there is a significant genetic component to the disease as well (Tokuhata and Lilienfeld, 1963; Lan et al., 2012). In particular, many studies have shown that the SNP rs1051730, located in the the nicotine receptor gene CHRNA3, is significantly associated with risk of the disease (Amos et al., 2008), even in subjects who do not smoke. Because rs1051730 is associated with disease risk even in non-smokers, it could possibly play a role in pathways outside of nicotine sensitivity as well. Here we investigate if the

SNP is increasing disease risk through methylation of other genes that may be important to the biological processes involved in lung cancer. Specifically, we study a sample of 77 lung cancer patients recruited from the TCGA Lung Adenocarcinoma (TCGA-LUAD) cohort, TCGA Lung Squamous Cell Carcinoma (TCGA-LUSC) cohort, and Massachusetts General Hospital in Boston, all of whom provide genotype information on rs1051730 as well as methylation data at over 450000 probes on the Illumina Infinium HumanMethylation450 BeadChip. The TCGA-LUAD and TCGA-LUSC cohorts are two programs from The Cancer Genome Atlas (TCGA) project, a government initiative to catalogue the multiple genomic dimensions of over 30 different cancers (The Cancer Genome Atlas Research Network et al., 2013). The cohort from Massachusetts General Hospital is described in Wang et al. (2017).

For each probe, we first search for the gene transcription start site that is nearest to the probe. All probes nearest the same gene transcription start site are treated as a set, and the methylation level at each probe is treated as a separate outcome. For all outcomes we fit the multiple phenotype model in equation (3.1), where the number of minor alleles at rs1051730 is the single explanatory variable in each model. The correlation between the outcomes is estimated by taking the sample correlations across the probes. We then apply the GHC, GBJ, iHC, and iBJ to each gene.

The results in Table 3.1 show that iHC provides the most evidence of significance for the seven of the top ten significant genes in our study, while GBJ provides the lowest p-value for the other three. This outcome would seem to support the conclusion that innovated tests will generally have more power in the K:1 setting. Using a Bonferroni correction for the 14991 genes studied, our nominal level of significance is set a 3.34×10^{-6}, which none of the genes meet, possibly due to our smaller sample size. However, it should be noted that GBJ provides the lowest p-value across all tests for any gene. iHC and iBJ do appear to find evidence of significance in some genes that GBJ and GHC rank very lowly, which shows that the transformation and weighting of signals can have a large impact. On the other hand, for the three genes that GBJ provides the lowest p-value, the iHC and iBJ also show relatively strong evidence of association.

Table 3.1: P-values of genes for which methylation is most associated with the lung cancer risk SNP rs1051730. Most significant p-value for any gene is highlighted in red.

Gene	GBJ	GHC	iBJ	iHC	K
B3GALT6	$9.06 \cdot 10^{-4}$	$1.53 \cdot 10^{-3}$	$6.63 \cdot 10^{-2}$	$2.14 \cdot 10^{-2}$	13
RAD18	$8.60 \cdot 10^{-1}$	$8.09 \cdot 10^{-1}$	$6.25 \cdot 10^{-3}$	$1.07 \cdot 10^{-3}$	37
SNRK-AS1	$8.16 \cdot 10^{-1}$	$8.04 \cdot 10^{-1}$	$5.46 \cdot 10^{-3}$	$1.09 \cdot 10^{-3}$	23
UROS	$3.64 \cdot 10^{-1}$	$2.89 \cdot 10^{-1}$	$2.88 \cdot 10^{-3}$	$1.17 \cdot 10^{-3}$	6
SLC44A2	1	$9.81 \cdot 10^{-1}$	$6.88 \cdot 10^{-3}$	$1.25 \cdot 10^{-3}$	31
GRN	$1.59 \cdot 10^{-3}$	$1.67 \cdot 10^{-3}$	$4.78 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$	24
POLR2A	1	$8.03 \cdot 10^{-1}$	$8.62 \cdot 10^{-3}$	$1.64 \cdot 10^{-3}$	30
USP2	$7.21 \cdot 10^{-1}$	$8.44 \cdot 10^{-1}$	$9.83 \cdot 10^{-3}$	$1.83 \cdot 10^{-3}$	31
MIR4469	$1.86 \cdot 10^{-3}$	$2.14 \cdot 10^{-3}$	$1.76 \cdot 10^{-2}$	$1.38 \cdot 10^{-2}$	4
TMEM97	$5.77 \cdot 10^{-1}$	$8.18 \cdot 10^{-1}$	$9.87 \cdot 10^{-3}$	$2.12 \cdot 10^{-3}$	21

3.6.2 Analysis of breast cancer GWAS data

Breast cancer is known to possess a complex genetic etiology, and it has also been extensively analyzed in the GWAS format. The Cancer Genetic Markers of Susceptibility (CGEMS) is one such GWAS dataset, with a case-control sample of 1145 European ancestry cases and 1142 controls recruited from the Nurses' Health Study. The Illumina HumanHap500 array was used collect genotype information, and 528143 SNPs remained after quality control. Hunter et al. (2007) first analyzed this dataset and did not find any SNPs to reach genome-wide significance, but they did identify FGFR2 as a gene of interest based on four SNPs in the gene which showed suggestive evidence of association. Follow-up set-based approaches using GHC and GBJ (Barnett et al., 2016; Sun and Lin, 2017) have appeared to confirm that FGFR2 is indeed a gene contributing to breast cancer risk.

We fit a slightly more complicated 1:K model than given in equation (3.2), using a logistic regression model with additional covariates for age and the first three genotype principal components to control for population stratification. The correlation between test statistics was estimated as in Section 2.1 of Sun and Lin (2017). Then for each gene, we collected the marginal test statistics for all SNPs located with the gene boundaries and analyzed the set with each of our four tests.

As we can see from Table 3.2, the Generalized Berk-Jones and Generalized Higher

Table 3.2: P-values of genes which are most associated with breast cancer. Most significant p-value for any gene highlighted in red.

Gene	GBJ	GHC	iBJ	iHC	K
FGFR2	$4.58 \cdot 10^{-6}$	$2.84 \cdot 10^{-5}$	$2.19 \cdot 10^{-2}$	$4.21 \cdot 10^{-3}$	35
CNGA3	$4.04 \cdot 10^{-5}$	$3.00 \cdot 10^{-4}$	$1.56 \cdot 10^{-2}$	$3.27 \cdot 10^{-3}$	26
PTCD3	$5.50 \cdot 10^{-5}$	$1.21 \cdot 10^{-4}$	$2.30 \cdot 10^{-3}$	$6.12 \cdot 10^{-4}$	12
POLR1A	$6.19 \cdot 10^{-5}$	$9.58 \cdot 10^{-5}$	$5.13 \cdot 10^{-3}$	$1.15 \cdot 10^{-3}$	17
ZNF263	$3.90 \cdot 10^{-4}$	$4.89 \cdot 10^{-4}$	$8.32 \cdot 10^{-5}$	$2.83 \cdot 10^{-3}$	3
LOC643923	$4.55 \cdot 10^{-1}$	$7.70 \cdot 10^{-1}$	$2.98 \cdot 10^{-4}$	$8.42 \cdot 10^{-5}$	10
ELMOD1	$3.98 \cdot 10^{-1}$	$6.05 \cdot 10^{-1}$	$8.72 \cdot 10^{-4}$	$1.62 \cdot 10^{-4}$	24
ABCA1	$1.65 \cdot 10^{-4}$	$3.74 \cdot 10^{-3}$	$6.38 \cdot 10^{-3}$	$4.26 \cdot 10^{-2}$	63
PDE8B	$6.89 \cdot 10^{-1}$	$3.97 \cdot 10^{-1}$	$1.77 \cdot 10^{-4}$	$4.24 \cdot 10^{-2}$	74
MMRN1	$5.51 \cdot 10^{-4}$	$2.31 \cdot 10^{-4}$	$2.88 \cdot 10^{-2}$	$1.51 \cdot 10^{-2}$	10

Criticism provide the strongest evidence of association for six of the top ten genes in the entire set, including all of the top four. This result agrees with our simulations, which suggest that the non-innovated tests are slightly more powerful in the SNP-set setting. Interestingly, the p-values for innovated and non-innovated tests are often extremely different for the same gene. The discrepancy suggests that inference can take completely opposite directions depending on the choice of test. None of the four most significant genes according to Generalized Berk-Jones and Generalized Higher Criticism is seen as highly associated with breast cancer according to iHC and iBJ, demonstrating again that the decorrelation transformation can greatly alter signals in the original test statistics. Similarly, the genes which iBJ and iHC find significant are assigned relatively large p-values by GBJ and GHC.

3.7 Discussion

We have studied the differences between innovated and non-innovated set-based tests for sparse outcomes in the multiple phenotype and SNP-set settings. In particular, we have demonstrated the importance of the correlation between signal and noise features in driving the power of innovated tests under the K:1 setting. When in the multiple phenotype setting, power can be gained as the total amount of correlation increases by using iHC and iBJ, since weights on the signals become much more favorable. The decorrela-

tion procedure is also able to increase the density of signals and offers a slightly larger rejection region. Thus when there appears to be a high degree of correlation between outcomes in a set, we should prefer the innovated tests. When the amount of correlation is low, the difference between innovated and non-innovated tests is not large.

In the 1:K SNP-set setting, the advantages of high correlation for innovated tests falls away. The difference is due to the form of the signal weights, which are the inverse of those in the K:1 setting. Additionally, the decorrelation transformation does not make the signal any more dense, and it has the potential to attenuate some of the originally large test statistics, as was seen in the analysis of CGEMS data. Simulation results seem to show that the GHC and GBJ are preferred in this setting under most correlation patterns.

An interesting remaining question is whether the advantages of innovation can be harnessed in the 1:K setting as they are in the K:1 setting. We see in our application to the CGEMS study that the decorrelation transformation often reduces the magnitude of marginal test statistics from each factor, but it seems plausible that a more suitable transformation could be tailored for specific situations. More specifically, since the decorrelation is still valid after we multiply the eigen-transformation matrix by any orthogonal matrix, it seems that there would be some orthogonal matrix which strengthens the existing signals rather than attenuating them.

Another interesting area of further research is whether power gains can be achieved by removing some subset of factors and applying a set-based test with smaller sets. If many factors receive very small weights, then it could be counterproductive to introduce them into the set as they will largely provide more noise. It is possible that a strategy which only makes use of highly upweighted factors could provide more finite sample power.

References

1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

ALMLI, L., DUNCAN, R., FENG, H., GHOSH, D., BINDER, E., BRADLEY, B., RESSLER, K., CONNEELY, K. and EPSTEIN, M. (2014). Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiatry* **71** 1392–1399.

AMOS, C. I., WU, X., BRODERICK, P., GORLOV, I. P., GU, J., EISEN, T., DONG, Q., ZHANG, Q., GU, X., VIJAYAKRISHNAN, J., SULLIVAN, K., MATAKIDOU, A., WANG, Y. and ET AL., G. M. (2008). Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40** 616–622.

ASCHARD, H., LUTZ, S., MAUS, B., DUELL, E., FINGERLIN, T., CHATTERJEE, N., KRAFT, P. and VAN STEEN, K. (2012). Challenges and opportunities in genome-wide environmental interaction (gweis) studies. *Human Genetics* **131** 1591–1613.

BARNETT, I., MUKHERJEE, R. and LIN, X. (2016). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association* Doi:10.1080/01621459.2016.1192039.

BEGG, M. and LAGAKOS, S. (1992). Effects of mismodeling on tests of association based on logistic regression models. *The Annals of Statistics* **20** 1929–1952.

BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47.

BOEHM, J. S., ZHAO, J. J., YAO, J., KIM, S. Y., FIRESTEIN, R., DUNN, I. F., SJOSTROM, S. K., GARRAWAY, L., WEREMOWICZ, S. and RICHARDSON, A. (2007). Integrative genomic approaches identify ikbke as a breast cancer oncogene. *Cell* **129** 1065.

BURRIS, H., BRAUN, J., BYUN, H., TARANTINI, L., MERCADO, A., WRIGHT, R., SCHNAAS, L., BACCARELLI, A., WRIGHT, R. and TELLEZ-ROJO, M. (2013). Association between birth weight and dna methylation of igf2, glucocorticoid receptor and repetitive elements line-1 and alu. *Epigenomics* **5** 271–281.

CLAUS HENN, B., SCHNASS, L., ETTINGER, A., SCHWARTZ, J., LAMADRID-FIGUEROA, H., HERNNDEZ-AVILA, M., AMARASIRIWARDENA, C., HU, H., BELLINGER, D., R., W. and TLLEZ-ROJO, M. (2012). Associations of early childhood manganese and lead coexposure with neurodevelopment. *Environmental Health Perspectives* **120** 126–131.

CONNEELY, K. and BOEHNKE, M. (2007). So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics* **81** 1158.

CORNELIS, M., TCHETGEN, E., LIANG, L., QI, L., CHATTERJEE, N., HU, F. and KRAFT, P. (2012). Gene-environment interactions in genome wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology* **120** 191–202.

DAWSON, E., ABECASIS, G. R., BUMPSTEAD, S., CHEN, Y., HUNT, S., BEARE, D. M., PABIAL, J., DIBLING, T., TINSLEY, E., KIRBY, S., CARTER, D., PAPASPYRIDONOS, M., LIVINGSTONE, S., GANSKE, R., LHMUSSAAR, E., ZERNANT, J., TONISSON, N., REMM, M., MGI, R., PUURAND, T., VILO, J., KURG, A., RICE, K., DELOUKAS, P., MOTT, R., METSPALU, A., BENTLEY, D. R., CARDON, L. R. and DUNHAM, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418** 544–548.

DENNY, J. C., BASTARACHE, L., RITCHIE, M. D., CARROLL, R. J., ZINK, R., MOSLEY, J. D., FIELD, J. R., PULLEY, J. M., RAMIREZ, A. H., BOWTON, E. and MELISSA A BASFORD, E. A. (2013). Systematic comparison of phenome-wide association study of elec-

tronic medical record data and genome-wide association study data. *Nature Biotechnology* **31** 1102–1111.

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994.

ELIASSEN, A. H., TWOROGER, S. S., MANTZOROS, C. S., POLLAK, M. N. and HANKINSON, S. E. (2007). Circulating insulin and c-peptide levels and risk of breast cancer among predominately premenopausal women. *Cancer Epidemiology Biomarkers and Prevention* **16** 161–164.

GAIL, M., WEIAND, S. and PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71** 431–444.

GIBSON, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13** 135–145.

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38** 1686–1732.

HUNTER, D., KRAFT, P., JACOBS, K., COX, D., YEAGER, M., HANKINSON, S., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A. and WANG, J. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* **39** 870–874.

HUTTER, C., MECHANIC, L., CHATTERJEE, N., KRAFT, P. and GILLANDERS, E. (2013). Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report. *Genetic Epidemiology* **37** 643–657.

JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics* **35** 2018–2053.

JIA, P., ZHENG, S., LONG, J., ZHENG, W. and ZHAO, Z. (2011). dmgwas: dense module

searching for genome-wide association studies in proteinprotein interaction networks. *Bioinformatics* **27** 95–102.

KAUERMANN, G. and CARROLL, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96** 1387–1396.

KILE, M., RODRIGUES, E., MAZUMDAR, M., DOBSON, C., DIAO, N., GOLAM, M., QUAMRUZZAMAN, Q., RAHMAN, M. and CHRISTIANI, D. (2014). A prospective cohort study of the association between drinking water arsenic exposure and self-reported maternal health symptoms during pregnancy in bangladesh. *Environmental Health* **13** 29.

KRAFT, P., YEN, Y., STRAM, D., MORRISON, J. and GAUDERMAN, W. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63** 111–119.

LAGAKOS, S. (1988). Effects of mismodelling and and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine* **7** 257–274.

LAN, Q., HSIUNG, C. A., MATSUO, K., HONG, Y.-C., SEOW, A., WANG, Z., II, H. D. H., CHEN, K., WANG, J.-C., CHATTERJEE, N., HU, W. and ET AL., M. P. W. (2012). Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in asia. *Nature Genetics* **44** 1330–1335.

LEE, S., ABECASIS, G. R., BOEHNKE, M. and LIN, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95** 5–23.

LEE, S., WU, M. C. and LIN, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83** 311.

LI, J. and SIEGMUND, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics* **43** 1323–1350.

LIANG, J., CHEN, P., HU, Z., ZHOU, X., CHEN, L., LI, M., WANG, Y., TANG, J., WANG, H. and SHEN, H. (2008). Genetic variants in fibroblast growth factor receptor 2 (fgfr2) contribute to susceptibility of breast cancer in chinese women. *Carcinogenesis* **29** 2341–2346.

LIU, Z. and LIN, X. (2017). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *Submitted* .

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J. and MCCARTHY, M. I. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. CRC press.

MEYER, K. B., MAIA, A.-T., O'REILLY, M., TESCHENDORFF, A. E., CHIN, S.-F., CALDAS, C. and PONDER, B. A. (2008). Allele-specific up-regulation of fgfr2 increases susceptibility to breast cancer. *PLoS Biology* **1** e108.

MOSCOVICH-EIGER, A. and NADLER, B. (2017). Fast calculation of boundary crossing probabilities for poisson processes. *Statistics & Probability Letters* **123** 177–182.

PASANIUC, B. and PRICE, A. L. (2016). Dissecting the genetics of complex traits using summary association statistics. *Nature Genetics Reviews* **18** 117–127.

PRENTICE, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* **81** 321–327.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904–909.

ROSENBLUM, M. and VAN DER LAAN, M. (2009). Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* **65** 937–945.

SMITH, B. and LAND, H. (2012). Anticancer activity of the cholesterol exporter abca1 gene. *Cell Reports* **2.3** 580–590.

STEVENS, V., RODRIGUEZ, C., PAVLUCK, A., THUN, M. and CALLE, E. (2006). Association of polymorphisms in the paraoxonase 1 gene with breast cancer incidence in the cps-ii nutrition cohort. *Cancer Epidemiology Biomarkers and Prevention* **15** 1226–1228.

SU, Z., MARCHINI, J. and DONNELLY, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics* **27** 2304–2305.

SUN, R. and LIN, X. (2017). Set-based tests using the generalized berk-jones statistic in genetic association studies [ph.d. thesis]. *Harvard University* .

TCHETGEN, E. and KRAFT, P. (2011). On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology* **22** 257–261.

THE CANCER GENOME ATLAS RESEARCH NETWORK, WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C. and STUART, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45** 1113–1120.

THOMAS, D. (2010). Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics* **11** 259–272.

TOKUHATA, G. K. and LILIENFELD, A. (1963). Familial aggregation of lung cancer in humans. *Journal of the National Cancer Institute* **30** 289–312.

VANSTEELANDT, S., VANDERWEELE, T., TCHETGEN, E. and ROBINS, J. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association* **103** 1693–1704.

VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. and YANG, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics* **90** 7–24.

VOORMAN, A., LUMLEY, T., MCKNIGHT, B. and RICE, K. (2011). Behavior of qq-plots and genomic control in studies of gene-environment interaction. *PLOS ONE* **6** e19416.

WALTHER, G. (2013). The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*, vol. 9. IMS, Beachwood, OH.

WANG, Z., WEI, Y., ZHANG, R., SU, L., MCKAY, J., BRENNAN, P., STILP, A., LAURIE, C., DOHENY, K., PUGH, E., XIAO, X., PIKIELNY, C., HUNG, R. J., AMOS, C. I., LIN, X. and CHRISTIANI, D. C. (2017). Integrative analysis of multi-omics data reveal a network of hypoxia-inducible factors family and a hub gene epas1 associated with lung adenocarcinoma. *Submitted* .

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86** 929–942.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.