On set-based association tests: insights from a regression using summary statistics

Yanyan Zhao[1] and Lei Sun[1,2]

[1]Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada,

yanyan.zhao@utoronto.ca

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada,

sun@utstat.toronto.edu

March 13, 2020

Abstract

Motivated by, but not limited to, association analyses of multiple genetic variants, we propose here a summary statistics-based regression framework. The proposed method requires only variant-specific summary statistics, and it unifies earlier methods based on individual-level data as special cases. The resulting score test statistic, derived from a linear mixed-effect regression model, inherently transforms the variant-specific statistics using the precision matrix to improve power for detecting sparse alternatives. Furthermore, the proposed method can incorporate additional variant-specific information with ease, facilitating omic-data integration. We study the asymptotic properties of the proposed tests under the null and alternatives, and we investigate efficient p-value calculation in finite samples. Finally, we provide supporting empirical evidence from extensive simulation studies and two applications.

Keywords: Correlation; Summary statistics; Regression; Set-based tests; Sparse alternatives.

Set-based joint analyses of multiple variables are increasingly important in many current scientific studies. For example, in modern genome-wide association studies (GWAS), one might be interested in jointly analyzing multiple (rare) genetic variants influencing a complex, and heritable trait (also known as gene- or pathway-based association studies), identifying one genetic variant influencing multiple traits (also known as pleitrotropy studies), or combining evidence from multiple studies as in the classical meta-analyses.

Without loss of generality, let us focus on set-based analyses of multiple rare genetic variants. In this setting, myriad statistical tests have been proposed and they fit into three general categories (e.g. Derkach, Lawless & Sun, 2014; Lee et al., 2014): the linear or burden tests (e.g. Madsen & Browning, 2009), the quadratic or variance-component tests (e.g. Pan, 2009; Wu & Lin, 2011), and the hybrid tests combining evidence from the linear and quadratic tests (e.g. Lee, Wu & Lin, 2012; Derkach, Lawless & Sun, 2013).

These earlier methods have been studied extensively, but they can be improved in several aspects. First, these methods may not perform well in the sparse-signal setting where only a small proportion of the variants are truly associated (Xu et al., 2016). Second, individual/subject-level data may not be available in practice, so it is useful to develop summary statistics-based association tests, and ideally the tests can also incorporate additional information available for each variant. Finally, earlier work have shown that the performance of a simple minimum-p value approach (Derkach, Lawless & Sun, 2013) is comparable with that of the optimal sequence kernel association test, SKAT-O (Lee, Wu & Lin, 2012). This suggests that a grid search for the 'optimal' weighting factor may not be necessary, and it is helpful to develop new robust hybrid test statistics with model-driven weights.

To this end, we propose a flexible and unifying linear mixed-effect regression model that requires only variant-specific summary statistics, and we show that earlier methods based on individual-level data are special cases of the proposed testing framework. The set-based association test statistic derived from the regression model inherently transforms the variant-specific

summary statistics using the precision matrix to improve power for detecting sparse alternatives; see Fan, Jin & Yao (2013) and Cai, Liu & Xia (2014) for the use of precision matrix in other high-dimension analytical settings. Furthermore, the proposed method can incorporate additional variant-specific information as a covariate, e.g. the functional importance of the variants to be analyzed (Ionita-Laza et al., 2016). Through simulation and application studies, we show that the proposed method can have substantial power gain when the included covariate contains useful information, while power loss is minimal when the covariate is uninformative.

Although the proposed set-based regression approach is motivated by joint analyses of multiple rare genetic variants, the method can be used in other settings, for example, meta-analyses of multiple studies (Han & Eskin, 2011), data integration of GWAS and gene-expression data (Gamazon et al., 2015), pleiotropy association studies of multiple phenotypes (Liu & Lin, 2018b), as well as polygenic risk score analyses of multiple common variants (Purcell et al., 2009). We will discuss later the differences and connections between the proposed method and these earlier work.

The remainder of the paper is organized as follows. Section 1 reviews existing association methods for jointly analyzing a set of rare genetic variants based on individual-level data. Section 2 first outlines the proposed regression framework based on summary statistics, from which we derive a catalog of association test statistics from fixed-, random- or mixed-effect models. We then demonstrate the analytical equivalence between some of the newly proposed statistics and the existing ones for rare variants analyses. This section also investigates efficient p-value calculation in finite samples, studies the asymptotic properties of the proposed tests, and considers covariate adjustments. Section 3 provides supporting empirical evidence from extensive simulation studies, and Section 4 presents results from two application studies. Section 5 concludes with remarks and discussion, and the Supplementary Material provides theoretical proofs and additional numerical results.

1 Existing association tests for jointly analyzing a set of rare genetic variants

1.1 Regression set-up using individual-level data

Consider a sample of n independent individuals and a set of J genetic variants of interest, let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ denote the phenotype variable and $\boldsymbol{G}$ be a $n \times J$ matrix for the corresponding genotypes with elements G_{ij}, $i = 1, \cdots, n$ and $j = 1, \cdots, J$, $\boldsymbol{G}_i = (G_{i1}, G_{i2}, \cdots, G_{iJ})^T$ and $\boldsymbol{G}_j = (G_{1j}, G_{2j}, \cdots, G_{nj})^T$. Assume that y_i given $\boldsymbol{G}_i$ follows an exponential family distribution with mean μ_i and consider using the canonical link function $g(\cdot)$ and

$$g(\mu_i) = g\left(E(y_i|\boldsymbol{G}_i)\right) = \beta_0 + \beta_1 G_{i1} + \ldots + \beta_J G_{iJ}, \tag{1}$$

to model the phenotype-genotype association, where β_j is the regression coefficient for genetic variant j. Individual-level covariates information such as age and sex, if available, should be added to the model but are omitted for the moment for notation simplicity.

We are interested in testing the hypothesis that

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T = \boldsymbol{0} \text{ v.s. } H_1 : \beta_j \neq 0, \text{ for any } j, j = 1, \ldots, J.$$

The corresponding score vector is

$$\boldsymbol{s} = (s_1, \ldots, s_J)^T = \boldsymbol{G}^T(\boldsymbol{y} - \bar{\mu}_y \boldsymbol{1_n}),$$

where $\boldsymbol{1_n}$ is a $n \times 1$ unit vector, $\bar{\mu}_y = \frac{1}{n}\sum_{i=1}^n y_i$ and $s_j = \sum_{i=1}^n (y_i - \bar{\mu}_y) G_{ij}$ capturing the linear

relationship between phenotype $\boldsymbol{y}$ and genotype $\boldsymbol{G}_j$. The variance-covariance matrix of $\boldsymbol{s}$ is

$$\boldsymbol{\Sigma_0} = g_1^{-1}(\bar{\mu}_y)\boldsymbol{G}^T(I_n - \boldsymbol{1_n}\boldsymbol{1_n}^T/n)\boldsymbol{G}, \tag{2}$$

where $g_1(\cdot)$ denotes the first derivative of the link function $g(\cdot)$, and I_n is a identity matrix of size n.

1.2 Existing methods based on the score vector s

Although it may not be obvious, most test statistics developed for rare variants analyses are functions of $\boldsymbol{s}$. For example, the original burden test (Madsen & Browning, 2009) first constructs a 'super-allele', $G^* = \sum_{j=1}^{J} w_j G_{ij}$, where $\boldsymbol{w} = (w_1, w_2, \cdots, w_J)^T$ are pre-specified weights often associated with the minor allele frequency (MAF) of the variants. The burden test then considers regression model, $g(\mu_i) = \beta_0^* + \beta^* G^*$ and test $H_0 : \beta^* = 0,$ v.s. $H_1 : \beta^* \neq 0$. However, it is not difficult to show that the score test statistic derived from the regression using G^* is proportional to

$$T_1 = (\boldsymbol{w}^T\boldsymbol{s})^2,$$

where under the null of no association $(\boldsymbol{w}'\boldsymbol{\Sigma_0}\boldsymbol{w})^{-1} T_1$ follows a central chi-square distribution with 1 degrees of freedom (d.f.), χ_1^2, for a fixed J but assuming that $\boldsymbol{s}$ is multivariate normal asymptotically with respect to n (Derkach, Lawless & Sun, 2014).

This T_1 test is also termed as the sum test by Pan (2009), and it belongs to the linear class of tests (Derkach, Lawless & Sun, 2014). Because T_1 is based on the weighted average of s_j, and s_j can be positive or negative depending on the direction of effect (i.e. sign of β_j in model (1)), T_1 is only powerful when a large proportion of variants are causal *and* effects are in the same direction.

Variance-component tests, such as SSU (Pan, 2009) and SKAT (Wu & Lin, 2011), are alternatives that belong to the quadratic class of tests; see Table 1 of Derkach, Lawless & Sun (2014)

for a summary. Again, although most of these tests started with regression model (1), they can be formulated as

$$T_2 = \boldsymbol{s}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{s},$$

where $\boldsymbol{A} = \text{diag}\{w_1^2, \cdots, w_J^2\}$ and w_j depends on the MAF of variant $j, j = 1, \cdots, J$. Under the null and the multivariate normality assumption for $\boldsymbol{s}$, T_2 follows a weighted sum of χ_1^2 distribution for a fixed J (Derkach, Lawless & Sun, 2014). T_2 is robust to heterogenous effect directions, but it is less powerful than T_1 when most variants are causal and with the same direction of effects.

Since the true genetic model is unknown in practice, hybrid tests combining T_1 and T_2 have been proposed. For example, SKAT-O of Lee, Wu & Lin (2012) considers

$$Q_\rho = \rho(\boldsymbol{w}^T \boldsymbol{s})^2 + (1-\rho)\boldsymbol{s}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{s} = \rho T_1 + (1-\rho) T_2.$$

It then performs a grid search for the 'optimal' ρ, $0 = \rho_1 < \rho_2 < \cdots < \rho_m = 1$. Let p_ρ be the p-value corresponding to each Q_ρ, the final SKAT-O test statistic is

$$T_{skato} = \min\{p_{\rho_1}, \cdots, p_{\rho_m}\}.$$

The asymptotic p-value of T_{skato} can be calculated with one-dimensional numerical integration (Lee, Wu & Lin, 2012).

Instead of considering data-driven 'optimal' ρ and adjusting for the inherent selection bias, Derkach, Lawless & Sun (2013) proposed two simpler yet competitive hybrid test statistics, T_{Fisher} and T_{Minp}. Let p_{T_1} and p_{T_2} be the p-values corresponding to T_1 and T_2, respectively, the Fisher and Minp statistics take the form of

$$T_{Fisher} = -2\log(p_{T_1}) - 2\log(p_{T_2}) \text{ and } T_{Minp} = \min\{p_{T_1}, p_{T_2}\}.$$

Under the null of no association, p_{T_1} and p_{T_2} are asymptotically independent of each other, thus T_{Fisher} is χ^2_4 distributed and T_{Minp} is $Beta(1,2)$ distributed.

Derkach, Lawless & Sun (2013) have also shown that T_{Minp} and T_{skato} perform similarly, and they are slightly more powerful than T_{Fisher} when T_1 has no power (e.g. less than 20%). On the other hand, when both T_1 and T_2 have some power T_{Fisher} is more powerful than T_{Minp} and T_{skato}. However, we expect all three hybrid tests to have diminished power of detecting sparse alternatives (Donoho & Jin, 2004; Barnett, Mukherjee & Lin, 2017) when only a small proportion of the variants in a set are truly associated.

To improve performance, we first note that if variants are correlated with each other, we can consider for example $\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s}$ instead of $\boldsymbol{s}$ to construct a more powerful test for detecting sparse alternatives. Second, it is clear now that we only need variant-specific summary statistics $\boldsymbol{s} = (s_1, \cdots, s_J)^T$ to joint analyze the J variants of interest, provided $\boldsymbol{\Sigma_0}$ can be accurately estimated. Further, the fact that T_{Minp} and T_{skato} having similar performance suggests that a grid search for ρ may not be necessary, and an easy-to-compute yet theoretically justified 'optimal' ρ may exist. Lastly, when additional variant-specific information z_j (e.g. variant j being non-synonymous or not) is available, we can improve power by incorporating $\boldsymbol{z} = (z_1, \cdots, z_J)^T$. Intuitively we can achieve this by modifying w_j to incorporate z_j in additional to the MAF, but a less add-hoc approach is desirable. In the section, we develop a flexible and unifying regression framework that (i) requires only s_j and $\boldsymbol{\Sigma_0}$, (ii) yields T_1 and T_2 as special cases and a hybrid statistic similar to T_{skato}, and (iii) provides new test statistics that incorporate the precision matrix $\boldsymbol{\Sigma_0}^{-1}$ and account for covariate information $\boldsymbol{z}$.

2 A regression framework based on summary statistics

Here we assume that $\boldsymbol{s} = (s_1, \cdots, s_J)^T$ is available, summarizing the association relationship between the phenotype of interest and a set of J genetic variants as detailed in Section 1. We also

assume that $\boldsymbol{\Sigma_0}$ is known or estimated accurately from a reference panel (Cheng et al., 2019; Fan, Liao & Liu, 2016). Let $\boldsymbol{z} = (z_1, \cdots, z_J)^T$ represent variant-specific information available.

2.1 Fixed-effect and random-effect models

We first consider a fixed-effect (FE) model that models the common effect μ present among $s_j, j = 1, \cdots, J$,

$$\boldsymbol{s} = \mu \boldsymbol{w} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{w} = (w_1, \cdots, w_J)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_J)^T$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_0})$. Based on this model, we test

$$H_0 : \mu = 0 \ \text{ v.s. } \ H_1 : \mu \neq 0, \tag{4}$$

and the corresponding score test statistic is

$$T_{FE} = \left(\boldsymbol{w}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{w}\right)^{-1} (\boldsymbol{w}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{s})^2. \tag{5}$$

If we let w in T_{FE} to be $\boldsymbol{\Sigma_0 w}$, the analytical equivalence between T_{FE} and T_1 in Section 1 is easy to show because $T_1 \propto \left(\boldsymbol{w}^T \boldsymbol{\Sigma_0} \boldsymbol{w}\right)^{-1} (\boldsymbol{w}^T \boldsymbol{s})^2$ (Table 1).

Alternatively, we can consider the following random-effect (RE) model,

$$\boldsymbol{s} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{\eta} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{R})$, $\boldsymbol{R}$ is a $J \times J$ pre-defined positive or semi-definite symmetric matrix, and $\boldsymbol{\varepsilon}$ as defined above. If we test

$$H_0 : \tau^2 = 0 \ \text{ v.s. } \ H_1 : \tau^2 \neq 0, \tag{7}$$

the corresponding score test statistic is

$$T_{RE} = \left(2tr\left(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R}\right)^2\right)^{-1/2}\left(\boldsymbol{s}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R}\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s} - tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})\right) = c_2^{-1/2}(Q(\boldsymbol{s}) - c_1), \tag{8}$$

where $Q(\boldsymbol{s}) = \boldsymbol{s}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R}\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s}$, and $c_1 = tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})$ and $c_2 = 2tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})^2$ are respectively the mean and variance of $Q(\boldsymbol{s})$. If we let $\boldsymbol{R} = \boldsymbol{\Sigma_0}\boldsymbol{A}\boldsymbol{\Sigma_0}$, the analytical equivalence between T_{RE} and $T_2 = \boldsymbol{s}^T\boldsymbol{A}\boldsymbol{s}$ is also apparent because $T_{RE} \propto Q(\boldsymbol{s})$ (Table 1). In addition, if we let $\boldsymbol{R} = \boldsymbol{\Sigma_0}\boldsymbol{W}\boldsymbol{R}_\rho\boldsymbol{W}\boldsymbol{\Sigma_0}$, where $\boldsymbol{W} = \text{diag}\{w_j\}$ and $\boldsymbol{R}_\rho = \rho\boldsymbol{1_J}\boldsymbol{1_J}^T + (1-\rho)\boldsymbol{I_J}$, then $Q(\boldsymbol{s}) = \rho T_1 + (1-\rho)T_2 = Q_\rho$, the key element of T_{skato}.

2.2 Mixed-effect model

The fixed-effect model (3) captures the common underlying effect, while the random-effect model (6) accounts for potential effect heterogeneity. A logical next step is to consider a mixed-effect (ME) modelling framework that includes models (3) and (6) as special cases,

$$\boldsymbol{s} = \mu\boldsymbol{w} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\boldsymbol{w}$, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are defined as before. If we test

$$H_0 : \mu = 0, \tau^2 = 0 \text{ v.s. } H_1 : \mu \neq 0 \text{ or } \tau^2 \neq 0, \tag{10}$$

the corresponding score vector is $\left(\boldsymbol{w}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s},\ \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R}\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s} - \frac{1}{2}tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})\right)^T$, and the test statistic is

$$T_{ME} = \left(\boldsymbol{w}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{w}\right)^{-1}(\boldsymbol{w}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s})^2 + c_2^{-1}\left(\boldsymbol{s}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R}\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s} - c_1\right)^2, \tag{11}$$

where $c_1 = tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})$ and $c_2 = 2tr(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{R})^2$, the same as for T_{RE}.

Note that T_{ME} uses $\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s}$ to account for the correlation between the tested variants. Transforming $\boldsymbol{s}$ by the precision matrix has been considered previously in other settings. For example, Cai, Liu & Xia (2014) showed that the transformation can improve power of a two-sample high-dimensional means test for detecting sparse alternatives in the presence of high correlation.

Following the construction of T_{ME}, intuitively we can also consider the following weighted average of $T_1 = (\boldsymbol{w}^T\boldsymbol{s})^2$ and $T_2 = \boldsymbol{s}^T\boldsymbol{A}\boldsymbol{s}$ (and $\boldsymbol{A} = \boldsymbol{R}$),

$$T_{12} = \left(\boldsymbol{w}^T\boldsymbol{\Sigma_0}\boldsymbol{w}\right)^{-1}(\boldsymbol{w}^T\boldsymbol{s})^2 + \check{c}_2^{-1}\left(\boldsymbol{s}^T\boldsymbol{R}\boldsymbol{s} - \check{c}_1\right)^2, \tag{12}$$

where $\check{c}_1 = tr(\boldsymbol{\Sigma_0}\boldsymbol{R})$ and $\check{c}_2 = 2tr(\boldsymbol{\Sigma_0}\boldsymbol{R})^2$ are respectively the mean and variance of T_2 under the H_0 of (10). The connection between T_{12} with T_{skato} is also immediate. However, there are two key differences. First, given a ρ, T_{skato} relies on $Q_\rho = \rho T_1 + (1-\rho)T_2$. In contrast, T_{12} combines T_1 and the *square* of *centralized* T_2 (not T_2 itself). Secondly, T_{skato} searches for the 'optimal' ρ that minimizes the p-value associated with Q_ρ and then adjusts for selection bias. In contrast, T_{12} uses the variances of T_1 and T_2 as weights.

2.3 Covariate adjustment

Given $\boldsymbol{z} = (z_1, \cdots, z_J)^T$, one could consider directly modifying $\boldsymbol{w}$ to reflect the additional information available for the J variants of interest, but a principled approach is lacking. The proposed regression framework, however, can naturally incorporate $\boldsymbol{z}$ as a covariate into the regression,

$$\boldsymbol{s} = \mu\boldsymbol{w} + \theta\boldsymbol{z} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{w}$, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are defined as before. If we are interested in testing,

$$H_0 : \mu = 0, \theta = 0, \tau^2 = 0 \text{ v.s. } H_1 : \mu \neq 0 \text{ or } \theta \neq 0 \text{ or } \tau^2 \neq 0,$$

the corresponding score test statistic is

$$T_{ME,cov} = \boldsymbol{s}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{u} (\boldsymbol{u}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{u})^{-1} \boldsymbol{u}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{s} + c_2^{-1} \left(\boldsymbol{s}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{R} \boldsymbol{\Sigma_0}^{-1} \boldsymbol{s} - c_1 \right)^2, \tag{14}$$

where $\boldsymbol{u} = (\boldsymbol{w}, \boldsymbol{z})$, and c_1 and c_2 are defined as before. Table 1 summarizes all the tests discussed so far.

Table 1: **Summary of different test statistics for analyzing a set of genetic variants.** [a]The equivalence can be analytical or numerical; see corresponding sessions for details. See Table 1 of Derkach, Lawless & Sun (2014) for a more detailed summary of the previous methods.

Proposed Approach		Equivalence[a] with previous methods based on individual-level data	
Regression Models	Test statistics	Methods	Test statistics, e.g.
Fixed-effect Model, Section 2.1		Linear T_1	
$\boldsymbol{s} = \mu \boldsymbol{w} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_0})$ (3)	T_{FE} (5)	let $\boldsymbol{w}$ be $\boldsymbol{\Sigma_0 w}$	Sum-test/burden, Pan (2009)
$H_0 : \mu = 0$			Weighted-sum, Madsen & Browning (2009)
Random-effect Model, Section 2.1		Quadratic T_2	
$\boldsymbol{s} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \boldsymbol{\eta} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{R})$ (6)	T_{RE} (8)	let $\boldsymbol{R}$ be $\boldsymbol{\Sigma_0 A \Sigma_0}$	SSU, Pan (2009)
$H_0 : \tau^2 = 0$			SKAT, Wu & Lin (2011)
Mixed-effect Model, Section 2.2		Hybrid T_1 '+' T_2	
$\boldsymbol{s} = \mu \boldsymbol{w} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ (9)	T_{ME} (11)		
$H_0 : \mu = 0, \tau^2 = 0$	T_{12} (12)		T_{skato}, Lee, Wu & Lin (2012)
			T_{Fisher}, T_{Minp}, Derkach, Lawless & Sun (2013)
Covariate Adjustment, Section 2.3			
$\boldsymbol{s} = \mu \boldsymbol{w} + \theta \boldsymbol{z} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ (13)	$T_{ME,cov}$ (14)		
$H_0 : \mu = 0, \theta = 0, \tau^2 = 0$			

2.4 Asymptotic distributions of the proposed test statistics

The asymptotic distributions of the existing tests, T_1, T_2, T_{skato}, T_{Fisher}, and T_{Minp}, have been previously established. Here, we establish the asymptotic distribution of T_{ME} and T_{12} when $J \to \infty$. In the next section, we will study finite sample behaviour of the tests. We begin with some notations and mild conditions needed for Theorem 1. Let $\boldsymbol{C}$ be a matrix such that $\boldsymbol{C \Sigma_0 C}^T = \boldsymbol{I_J}$, and let

$\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ be the minimum and maximum eigenvalues of a matrix, respectively. For two sequences of real numbers $\{a_{1J}\}$ and $\{a_{2J}\}$, denote $a_{1J} = o(a_{2J})$ if $\lim_{J\to\infty}(a_{1J}/a_{2J}) = 0$.

Condition 1. $c_0^{-1} \leq \lambda_{min}(\boldsymbol{\Sigma_0}) \leq \lambda_{max}(\boldsymbol{\Sigma_0}) \leq c_0$ *for some constant* c_0.

Condition 2. $c^{-1} \leq \lambda_{min}(\boldsymbol{R}) \leq \lambda_{max}(\boldsymbol{R}) \leq c$ *for some constant* c.

Condition 3. $w_{max} = o(\sqrt{\boldsymbol{w}'\boldsymbol{w}})$ *where* $w_{max} = \max_j w_j$.

Theorem 1. *Theorem 1.Under the null hypothesis of (10) and assume Conditions 1-3 hold,*

$$T_{ME} \xrightarrow{d} \chi_2^2 \ \ \textit{and} \ \ T_{12} \xrightarrow{d} \chi_2^2, \ \textit{as} \ J \to \infty.$$

See the Supplementary Material for the proofs of Theorem 1 and all other theorems. When J is small, significance evaluation based on the asymptotic distributions may not be adequate. Theorem 2 provides an approximation for the finite-sample distribution of T_{ME}; the result for T_{12} is similar.

Theorem 2. *Theorem 2.Let* λ_j *be the jth eigenvalue of* $\boldsymbol{CRC}^T$, $j = 1, \cdots, J$, *then*

$$T_{ME} \overset{\mathscr{D}}{=} u_1^2 + \left(2\sum_{j=1}^{J}\lambda_j^2\right)^{-1}\left(\sum_{j=1}^{J}\lambda_j(v_j^2 - 1)\right)^2,$$

where u_1 *and* v_j, $j = 1, \cdots, r$, *are independent* $N(0,1)$ *random variables, and* $\overset{\mathscr{D}}{=}$ *denotes equality in distribution.*

We note that the above finite and asymptotic results are with respect to J. The validities of T_{ME} and T_{12} do not require $n \to \infty$ explicitly, as long as $\boldsymbol{s}$ is multivariate normal.

2.5 Power comparison

In this section, we first establish the asymptotic distributions of T_{ME} and T_{12} (as $J \to \infty$) under alternatives.

Theorem 3. *Theorem 3.Under the alternative hypothesis* $H_1 : \mu = \mu_1, \tau^2 = \tau_1^2$,

(a) Assume Conditions 1-3 hold, then

$$T_{ME} \xrightarrow{d} \pi_0^2 \chi_1^2 \left(\varphi_0^2/\pi_0^2\right) + \pi_1^2 \chi_1^2 \left((\varphi_1+\varphi_2)^2/\pi_1^2\right), \text{ as } J \to \infty,$$

where $\varphi_0 = \mu_1 \sqrt{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{w}}$, $\varphi_1 = 2^{-1/2} \tau_1^2 \sqrt{tr(\boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R})^2}$,
$\varphi_2 = \mu_1^2 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{w} \Big/ \sqrt{2tr(\boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R})^2}$,
$\pi_0^2 = 1 + \tau_1^2 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{w} \big/ \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{w}$, *and*
$\pi_1^2 = 1 + \left(\tau_1^4 tr(\boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R})^4 + 2\tau_1^2 tr(\boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R})^3\right) \big/ tr(\boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{R})^2$.

(b) Assume Conditions 1-3 hold, then

$$T_{12} \xrightarrow{d} \check{\pi}_0^2 \chi_1^2 \left(\check{\varphi}_0^2/\check{\pi}_0^2\right) + \check{\pi}_1^2 \chi_1^2 \left((\check{\varphi}_1+\check{\varphi}_2)^2/\check{\pi}_1^2\right), \text{ as } J \to \infty,$$

where $\check{\varphi}_0 = \mu_1 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w} \big/ \sqrt{\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{w}}$, $\check{\varphi}_1 = \tau_1^2 tr(\boldsymbol{R}^2) \big/ \sqrt{2tr(\boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{R})^2}$, $\check{\varphi}_2 = \mu_1^2 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{R} \boldsymbol{w} \big/ \sqrt{2tr(\boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{R})^2}$,
$\check{\pi}_0^2 = 1 + \tau_1^2 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{R} \boldsymbol{w} \big/ \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{w}$, *and* $\check{\pi}_1^2 = 1 + \left(\tau_1^4 tr(\boldsymbol{R}^4) + 2\tau_1^2 tr(\boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{R}^{\boldsymbol{3}})\right) \big/ tr(\boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{R})^2$.

To compare the asymptotic power between T_{ME} and T_{12}, first let us consider the simple case of no random effect, i.e. $\tau_1^2 = 0$. In that case, $\pi_0 = \check{\pi}_0 = \pi_1 = \check{\pi}_1 = 1$, and we can show that T_{ME} is at least as powerful as T_{12} provided that $\varphi_0^2/\check{\varphi}_0^2 = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}}^{-1} \boldsymbol{w} \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{w} \big/ (\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w})^2 \geqslant 1$. In fact, this inequality holds as long as $\boldsymbol{\Sigma}_{\boldsymbol{0}}$ is a positive definite symmetric matrix. This suggests that if the true underlying model for $\boldsymbol{s}$ is a fixed-effect model, then T_{ME} is more powerful than T_{12}. Our analytical conclusion here is consistent with that observed by Liu & Lin (2018a) for joint analyses of multiple phenotypes and by Brockwell & Gordon (2001) for meta-analyses.

We next consider a local alternative assuming

$$H_{1J} : \mu_1^2 \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{R} \boldsymbol{w} = o\left(\sqrt{2tr(\boldsymbol{\Sigma}_{\boldsymbol{0}} \boldsymbol{R})^2}\right), \ \tau_1^2 = O(J^{-1/2}).$$

In this case, $\varphi_2 = o(1)$, $\check{\varphi}_2 = o(1)$, $\pi_0 = 1 + o(1), \check{\pi}_0 = 1 + o(1), \pi_1 = 1 + o(1)$, and $\check{\pi}_1 = 1 + o(1)$ based on Conditions 1-2. As a result, T_{ME} is at least as powerful as T_{12}, provided that $\varphi_0^2/\check{\varphi}_0^2 = \boldsymbol{w^T}\boldsymbol{\Sigma_0^{-1}}\boldsymbol{w}\boldsymbol{w^T}\boldsymbol{\Sigma_0}\boldsymbol{w}/(\boldsymbol{w^T}\boldsymbol{w})^2 \geqslant 1$ and $\varphi_1^2/\check{\varphi}_1^2 = tr(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{R})^2 tr(\boldsymbol{\Sigma_0}\boldsymbol{R})^2/tr^2(\boldsymbol{R}^2) \geqslant 1$. These two inequalities always hold as long as $\boldsymbol{R}$ and $\boldsymbol{\Sigma_0}$ are positive definite symmetric matrices.

3 Simulation studies

To compare the finite-sample performance of T_{ME} and T_{12}, with T_{skato}, T_{Minp} and T_{Fisher}, we conduct extensive simulation studies, examining the effects of different correlation structures and signal sparsities on power.

To obtain the summary statistics, $\boldsymbol{s}$ and $\boldsymbol{\Sigma_0}$, for association studies of rare variants, similar to Derkach, Lawless & Sun (2014) we assume that the underlying individual-level model is $E(y_i|\boldsymbol{G}_i) = \beta_0 + \beta_1 G_{i1} + \cdots + \beta_J G_{iJ}$, where y_i is normally distributed with variance $\sigma^2 = 1$, G_{ij} is Bernoulli with $Pr(G_{ij} = 1) = p_j$ where p_j is approximately twice the minor allele frequency of variant j, and $i = 1, \cdots, n$ and $j = 1, \cdots, J$. Given this set-up, $\boldsymbol{s} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma_0})$, where $\boldsymbol{\mu} = (np_1(1-p_1)\beta_1, \cdots, np_J(1-p_J)\beta_J)'$, $\boldsymbol{\Sigma}_0 = \{\sigma_{jk}^2\}_{J\times J}$, $\sigma_{jk}^2 = np_j(1-p_j)$ for $j = k$ and $\sigma_{jk}^2 = n(p_{jk} - p_j p_k)$ for $j \neq k$, and $p_{jk} = P(G_{ij} = 1, G_{ik} = 1)$.

We consider $n = 500$ and $J = 10,\ 50,\ 100,\ 500$, and 1000, and we draw p_j randomly from Unif $(0.005, 0.02)$, $j = 1, \cdots, J$. For $\text{diag}\{\sigma_{jj}^{-1}\}\boldsymbol{\Sigma_0}\,\text{diag}\{\sigma_{jj}^{-1}\}$, we consider an AR(1) pattern with correlation $\tilde{\rho}$, and $\tilde{\rho} = 0.2$, 0.5 and 0.8. For $\boldsymbol{w}$ and $\boldsymbol{A}$ in T_{skato}, T_{Minp}, T_{Fisher}, and T_{12}, we choose the commonly used $w_j = 1/\sqrt{p_j(1-p_j)}$ and $\boldsymbol{A} = \text{diag}\{w_j^2\}$. For $\boldsymbol{w}$ and $\boldsymbol{R}$ in T_{ME}, we choose the same w_j and let $\boldsymbol{R} = \boldsymbol{A}$ for a fair comparison.

3.1 Type I error

To examine the validity of the proposed tests, T_{ME} and T_{12}, we generate $\boldsymbol{s}$ from $N(\boldsymbol{0}, \boldsymbol{\Sigma_0})$, independently, 10^6 times for each J and $\tilde{\rho}$ combination. The results in Table 2 show that for small J,

in combination with stringent α level, the asymptotic p-values are not adequate. In that case, the approximate solution in Theorem 2 should be used.

For the existing methods, T_{skato}, T_{Minp} and T_{Fisher}, we observe in our simulation studies that T_{skato} is slightly conservative for the α levels considered regardless of the size of J. T_{Minp} is also slightly conservative for small J but has correct test size when $J > 50$, and T_{Fisher} has inflated type I error when the correlation among variants is strong (Supplementary Material). Our observations here are largely consistent with previous reports (e.g. Larson et al., 2017).

Table 2: **Type I error evaluation.** Empirical test sizes for $\alpha = 5\%$, 1%, 0.1%, and 0.01%, estimated based on 10^6 replications independently simulated under the null. $T_{.,asy}$ represents p-value calculation using the asymptotic distributions in Theorem 1, and $T_{.,apr}$ is for p-value estimation based on Theorem 2 with 10^7 independent u_1 and v_j standard normal variables generated for each of the 10^6 simulated null replicates. Results here are for $\tilde{\rho} = 0.5$; results for other $\tilde{\rho}$ values are provided in the Supplementary Material.

J	α	$T_{12,asy}$	$T_{me,asy}$	$T_{12,apr}$	$T_{me,apr}$
10	5%	6.1558	5.6556	5.7470	5.2226
	1%	3.1219	2.4281	1.7649	1.2969
	0.1%	1.4531	0.9427	0.2079	0.1316
	0.01%	0.7725	0.4294	0.0196	0.0141
20	5%	5.6845	5.5795	5.4308	5.2198
	1%	2.5849	2.4312	1.6450	1.3667
	0.1%	1.0219	0.9643	0.2212	0.1476
	0.01%	0.4703	0.4550	0.0246	0.0149
50	5%	5.1085	5.1317	5.0333	4.9875
	1%	1.9088	1.7082	1.4498	1.1475
	0.1%	0.5952	0.5002	0.2103	0.1307
	0.01%	0.2148	0.1794	0.0230	0.0132
100	5%	4.9508	5.0064	4.923	4.9664
	1%	1.5279	1.3854	1.2788	1.1030
	0.1%	0.3796	0.3082	0.1846	0.1273
	0.01%	0.1149	0.0831	0.025	0.0138
500	5%	4.9335	5.0020	4.9307	4.9948
	1%	1.1265	1.0917	1.0720	1.0238
	0.1%	0.1662	0.1547	0.1288	0.1130
	0.01%	0.0332	0.0281	0.0202	0.0139
1000	5%	4.9841	4.9552	4.9743	4.9436
	1%	1.0475	1.0254	1.0151	0.9864
	0.1%	0.1365	0.1190	0.1183	0.0951
	0.01%	0.0205	0.0162	0.0163	0.0109

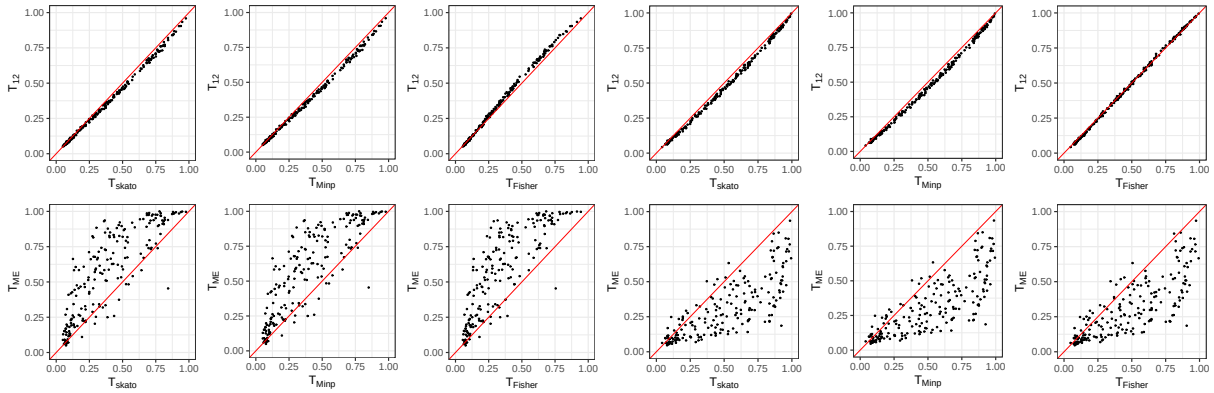
3.2 Power without covariates

We consider two different simulation designs to evaluate power. For both designs, P_c, the proportion of causal variants for a given set of J variants, is randomly drawn from $\text{Unif}(0.01, 0.1)$ for the case of sparse signals, and $P_c \sim \text{Unif}(0.1, 0.5)$ for the case of moderately sparse signals. Among the causal variants, p_d is the proportion of deleterious variants with $\beta_j > 0$ and $P_d \sim \text{Unif}(0.5, 0.75)$. We note that for each P_c and P_d combination, the locations of the signals ($\beta_j \neq 0$) are randomly drawn from $\{1, 2, \cdots, J\}$, without replacement. This randomness helps us to comprehensively explore the effect of different correlation structures between causal variants, between non-causal variants, as well as between causal and non-causal variants on power.

Design one follows the approach of Derkach, Lawless & Sun (2014). That is, $\boldsymbol{s} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ and $|\beta_j| \sim \text{Unif}(0.5, 1.5)$ for both the sparse and moderately sparse cases. Design two assumes that $\boldsymbol{s}$ is drawn from the mixed-effect model (9) with varying magnitudes of τ^2. Specifically, for a given J we first simulate $p_j \sim \text{Unif}(0.005, 0.02)$, $j = 1, \ldots J$, $P_c \sim \text{Unif}(0.1, 0.5)$ and $P_d \sim \text{Unif}(0.5, 0.75)$, and we draw the locations of the causal variants from $\{1, 2, \cdots, J\}$ as in design one. We then specify $\text{sign}(\beta_j)$ accordingly and obtain $\boldsymbol{w}^* = (np_1(1-p_1)\text{sign}(\beta_1), \cdots, np_J(1-p_J)\text{sign}(\beta_J))'$; for a null variant $\text{sign}(\beta_j) = 0$. Finally, we simulate $\boldsymbol{s} \sim N(\beta \boldsymbol{w}^*, \tau^2 \boldsymbol{I}_J + \boldsymbol{\Sigma}_0)$, where $\beta \sim \text{Unif}(0.5, 1.5)$ for both the sparse and moderately sparse cases, and $\tau^2 \sim \text{Unif}(1, 2)$.

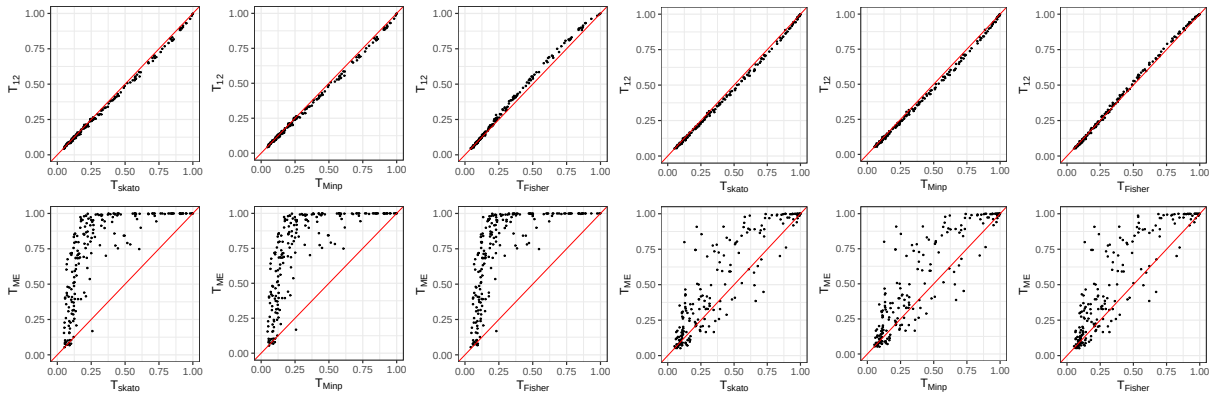
For both designs, results of power comparison focus on $J = 100$, $\alpha = 0.05$ and the sparse alternatives; results for the moderately sparse alternatives are characteristically similar (Supplementary Material). The empirical power for $\alpha = 0.05$ are estimated from 10^3 simulated replicates, and using the empirical critical values obtained from 10^4 null replicates.

Figures 1(a) and 1(b) show that when the correlation among variants is relatively strong (e.g. $\tilde{\rho} = 0.5$ in Figures 1(a) and $\tilde{\rho} = 0.8$ in the Supplementary Material), T_{ME} (based on $\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s}$) has better power than the other methods (based on $\boldsymbol{s}$) including T_{skato}. However, this approach may not be advantageous when there is only weak correlation in conjunction with sparse signal (e.g. $\tilde{\rho} = 0.2$ in Figures 1(b)) as discussed in Section 2.2. Interestingly, the new T_{12} test (inspired by T_{ME}



(a) Design ONE, $\tilde{\rho} = 0.5$

(b) Design ONE, $\tilde{\rho} = 0.2$



(c) Design TWO, $\tilde{\rho} = 0.5$

(d) Design TWO, $\tilde{\rho} = 0.2$

Figure 1: **Power comparison under sparse alternatives based on design ONE and TWO, comparing T_{12} and T_{ME} with T_{Fisher}, T_{Minp} and T_{skato}.** Empirical power of 200 alternative models are shown. For each model, $J = 100$ and the proportion of the causal variants varies from 1% to 10% randomly. See the main text for additional simulation details.

but based on $\boldsymbol{s}$) has comparable power with T_{skato}, but without the need to search for the 'optimal' ρ. Our simulation results here also confirm that the three existing hybrid tests, T_{skato}, T_{Minp} and T_{Fisher}, have similar performance, where T_{skato} and T_{Minp} perform more similarly with each other than with T_{Fisher}.

Figures 1(c) and 1(d) show that when $\boldsymbol{s}$ follows a mixed-effect model, the advantage of the proposed T_{ME} is enhanced. In this case, power of T_{ME} is considerably higher than the other tests even when $\tilde{\rho} = 0.2$ (Figures 1(d)).

3.3 Power with covariates

We now briefly study the effect of incorporating variant-specific additional information $\boldsymbol{z} = (z_1, \cdots, z_J)^T$. As discussed before, although one may revise w_j to be proportional to z_j in additional to MAF, it is not immediately clear how to choose an 'optimal' weighting function. Thus, we only study the proposed $T_{ME,cov}$, derived directly from the regression model (13), and we consider simulation design two only.

Without loss of generality, we assume z_j to be an indicator variable, for example indicating if the variant is non-synonymous ($z_j = 1$) or synonymous ($z_j = 0$). For causal variants we let $Pr(z_j = 1) = 0.5$, and for non-causal variants $Pr(z_j = 1) = 0$. We consider both the case of informative $\boldsymbol{z}$ ($\theta \neq 0$ in model (13) and $\theta \sim \text{Unif}(1,4)$) and the case of uninformative $\boldsymbol{z}$ ($\theta = 0$). Because of the additional information available from $\boldsymbol{z}$, we draw β from $\text{Unif}(0.1,1)$ and choose $\tau^2 = 0$. We also assumed $P_c = 0.1$ and $P_d \sim \text{Unif}(0.5,1)$ for this set of simulations.

Figure 2 show that, as expected, there can be substantial power gain when incorporating informative covariate information (the left plot), at the cost of slightly reduced power when $\boldsymbol{z}$ is uninformative (the right plot).

4 Applications

In this section, we examine eight test statistics including $T_{ME,cov}$, T_{ME} and T_{12}, and the existing methods, T_{skato}, T_{Minp}, T_{Fisher}, as well as T_1 and T_2 for completeness through two data applications. In the implementation of $T_{ME,cov}$, we use variants being non-synonymous or synonymous, annotated using the UCSC genome browser at https://genome.ucsc.edu/, as the variant-specific information.

The first application highlights the advantage of the proposed T_{ME} in the presence of high or moderately high correlation between variants, and it also demonstrates that the method is not limited to analyses of rare variants. The second application revisits the genetic analysis workshop 17

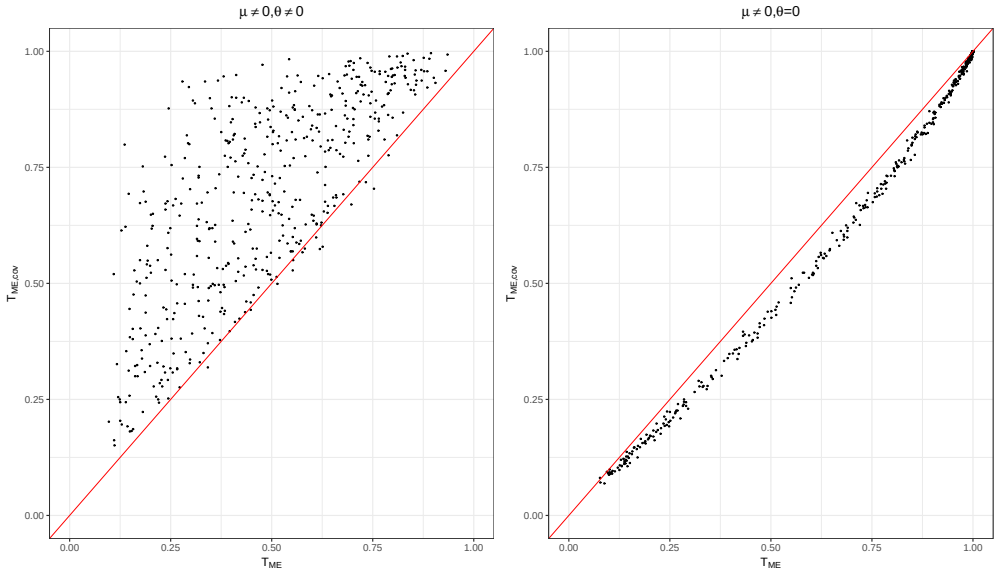


Figure 2: **Power comparison under sparse alternatives based on design TWO, comparing $T_{ME,cov}$ with T_{ME}.** The left plot is for the case of informative covariate ($\theta \neq 0$), and the right plot is for uninformative covariate ($\theta = 0$). Empirical power of 500 alternative models are shown. For each model, $J = 100$ and the proportion of the causal variants is 10%, and $\tilde{\rho} = 0.5$. See the main text for additional simulation details.

(GAW17) rare variants data previously studied by Derkach, Lawless & Sun (2014). This application reveals the benefit of incorporating covariate information using $T_{ME,cov}$. For both applications, we used method of Rothman (2012) to achieve a positive definite estimate for the covariance matrix $\mathbf{\Sigma}_0$.

4.1 Cystic Fibrosis (CF) data - common variants

Cystic Fibrosis is a life-limiting genetic condition for which lung function is a primary co-morbidity of interest. To *indirectly* study gene-environment interactions, Soave et al. (2015) proposed a joint location-scale (JLS) test and applied it to lung function measures in CF individuals, $n = 1,409$ from a Canadian sample and $n = 1,232$ from a French sample. They discovered and replicated the significance of the SLC9A3 complex set (35 common variants from four genes) based on the JLS test. However, the association evidence appears to come from the scale (interaction) component. For the traditional location (mean) test based on T_2, the SLC9A3 complex set was only significant

in the Canadian sample but not replicated in the French sample. Here we exam the performance of the eight tests applied to the French sample. For a fair comparison, we use $\boldsymbol{w} = \mathbf{1}$, $\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{R} = \boldsymbol{I}$ for all corresponding tests, and we obtain empirical p-values for all tests based on 10^4 permutation replicates.

Results in Table 3 show that only some of the genes appear to be truly associated with CF lung function. For SLC9A3, all tests have suggestive evidence with T_1 having p-value < 0.05. For SLC9A3R1, benefiting from the correlation structure (Supplementary Material) the proposed T_{ME} and $T_{ME,cov}$, which use $\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}\boldsymbol{s}$ instead of $\boldsymbol{s}$, are significant. When jointly analyzing all four genes in the SLC9A3 complex set, none of the tests is statistically significant but T_{ME} has the smallest p-value. A larger sample is needed to make a definitive conclusion of true association. The covariate information (non-synonymous vs. synonymous) appear not to be informative here, but the performance of $T_{ME,cov}$ is similar to that of T_{ME}.

Table 3: **Empirical p-values of the tests in the CF data application.**

		Empirical p-values							
Gene	J	T_1	T_2	T_{skato}	T_{Minp}	T_{Fisher}	T_{12}	T_{ME}	$T_{ME,cov}$
SLC9A3	7	**0.0443**	0.1198	0.079	0.0783	0.0541	0.0718	0.0886	0.0860
EZR	10	0.2192	0.6856	0.3241	0.3209	0.3897	0.239	0.3473	0.3474
SLC9A3R2	10	0.804	0.5951	0.6965	0.6984	0.709	0.9037	0.9683	0.8791
SLC9A3R1	8	0.0999	0.1103	0.1471	0.1656	0.0846	0.1042	**0.0243**	**0.0243**
4-gene jointly	35	0.8372	0.3079	0.4749	0.4738	0.5671	0.9261	0.1710	0.1709

4.2 The Genetic Analysis Workshop 17 (GAW17) data - rare variants

Here we apply the method to the GAW17 data provided by the 1000 Genomes Project (Consortium et al. 2010; Almasy et al. 2011), focusing on the simulated quantitative trait Q2. The phenotype Q2 is influenced by 72 variants in 13 genes but not by environmental factors, and the genotypes of these variants are obtained from a 'mini-exome' next-generation sequencing experiment (Consortium

et al. 2010). Available to us are 200 replicates, simulated based on a true phenotype-genotype association model determined by the GAW17 study group (Almasy et al. 2011) but blinded to this analysis. We consider $n = 321$ unrelated Asian samples and use only variants with MAF less than 0.05. The description of the variants is provided in Table 4. Among the 13 genes, GCKR is not analyzed since only one variant remained after variant screening. VNN1 does not have any causal rare variants but is kept for negative control.

For each of the 200 alternative replicates, we calculate the empirical p-values (based on 10^4 permutation replicates) for the eight test statistics. For each test, the power for $\alpha = 0.05$ is estimated as the proportion of the 200 replicates for which the empirical p-values ≤ 0.05. We separate the 11 genes into three categories based on power as in Derkach, Lawless & Sun (2014) (which examined T_1 and T_2) and Derkach, Lawless & Sun (2013) (which examined T_1, T_2, T_{Minp}, and T_{Fisher}), and we also jointly analyze all genes within each category.

In this application, because the correlation is weak among variants (Supplementary Material), we anticipate that methods relies on $\boldsymbol{s}$ will have better power than those based on $\boldsymbol{\Sigma_0}^{-1}\boldsymbol{s}$. Indeed, results in Table 4 show that T_{12} has better performance than T_{ME}. However, this application clearly demonstrates the potential of incorporating informative covariates. For example, the power of analyzing RARB, PLAT and VLDLR is higher using $T_{ME,cov}$, at the cost of slightly reduced performance for other genes when the included covariate is presumably not (detectably) informative. Interestingly, T_{Fisher} has comparable performance as T_{12}, and both outperform T_{skato} in almost all cases. Although the individual T_1 and T_2 tests may have the highest power for certain genes, the robustness of the hybrid tests is evident based on the overall performance exhibited in Table 4.

5 Discussion

In this paper, we considered a summary statistics-based regression framework to jointly analyze a set of J variants. As delineated in Table 1, the proposed approach is flexible and adaptive. The

Table 4: **Empirical power of the tests in the GAW17 data application.**

Gene	J_C	J_N	Empirical Power							
			T_1	T_2	T_{skato}	T_{Minp}	T_{Fisher}	T_{12}	T_{ME}	$T_{ME,cov}$
7 genes for which the maximum power is 10% or more										
SIRT1	4	7	0.44	0.385	0.455	0.43	0.495	**0.5**	0.285	0.285
BCHE	6	9	0.29	0.39	0.405	0.39	0.435	0.445	**0.46**	0.41
PDGFD	3	5	0.295	0.385	0.385	0.38	0.425	**0.45**	0.31	0.26
SREBF1	4	5	**0.495**	0.25	0.440	0.440	0.445	0.400	0.405	0.355
RARB	1	5	0.06	0.135	0.095	0.110	0.085	0.090	0.100	**0.215**
PLAT	4	7	0.13	0.125	0.105	0.100	0.135	0.12	0.13	**0.165**
VLDLR	4	3	0.11	0.085	0.100	0.095	0.115	0.115	0.110	**0.16**
7-gene jointly	26	41	0.935	0.765	0.94	0.935	**0.945**	0.94	0.765	0.765
4 genes for which the maximum power is 10% or less										
VNN3	2	2	0.035	0.04	0.035	0.035	0.04	0.04	0.04	0.035
INSIG1	3	1	0.05	0.03	0.03	0.03	0.035	0.035	0.025	0.015
LPL	1	3	0.03	0.065	0.035	0.035	0.04	0.035	0.035	0.050
VWF	1	3	0.025	0.01	0.01	0.01	0.015	0.01	0.04	0.045
4-gene jointly	7	9	0.075	0.01	0.06	0.06	0.055	0.045	0.055	0.045
1 gene for which there is no rare causal variants, used as a negative control										
VNN1	0	3	0.015	0.045	0.045	0.045	0.035	0.04	0.04	0.045

score test derived from the fixed-effect model, T_{FE}, unifies the linear class of tests (also known as the burden tests), T_1, while T_{RE} from the random-effect model connects the quadratic class or variance component tests, T_2. Further, the score test derived from the random-effect model offers a new hybrid test, T_{ME}, that naturally aggregates information from T_{FE} and T_{RE}.

It is worth emphasizing three notable differences between the proposed T_{ME} and the well-known T_{skato} (Lee, Wu & Lin, 2012). First, the proposed framework aggregates evidence across J variants based on $\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}\boldsymbol{s}$, a precision matrix-based transformation of the score vector, that can increase power for detecting sparse alternatives (Fan, Jin & Yao, 2013; Cai, Liu & Xia, 2014). Secondly, unlike T_{skato}, T_{ME} naturally aggregates the information from the linear and quadratic types of tests, based on a regression model *without* searching for the 'optimal' weight. Lastly, when additional variant-specific information is available, it is straightforward to derive $T_{ME,cov}$ that accounts for covariate effects. We have demonstrated these features of the proposed method both analytically and empirically.

To exploit the assumption of signal sparsity, various supremum-type tests have been proposed including the generalized higher criticism (Wu et al., 2014; Barnett, Mukherjee & Lin, 2017) for sparse signals, and most recently the generalized Berk-Jones statistic (Sun & Lin, 2019) for moderately sparse signals. These methods, tailored for common variants, are not easy to adjust for additional variant-specific information.

The proposed set-based testing framework is a general one, and it can be used for other settings such as pleiotropy studies of multiple phenotypes, where the analtyical unit is each of the phenotypes. In that context, Liu & Lin (2018b) also proposed a summary statistics-based linear mixed-effect regression model, but they focused on the special case of $\boldsymbol{w} = \mathbf{1}$ and $\boldsymbol{R} = \boldsymbol{I}$. In addition, Liu & Lin (2018b) derived two score test statistics, respectively, for testing $\mu = 0$ and $\tau^2 = 0$ *separately*, then considered different ways to combine the evidence including SKAT-O type of statistics. In contrast, we derive T_{ME} from testing $\mu = 0$ and $\tau^2 = 0$ *jointly*, and the weighting factors are inherently justified based on the regression model. We also study the asymptotic properties

of the proposed tests under the alternatives as well as empirical property under sparse alternative.

The proposed method can also be used for the study of polygenic risk score (PRS) (Purcell et al., 2009), and the connection between PRS and T_1 has been noted by Pan, Chen & Wei (2015). In principle, T_{ME} can overcome the poor statistical efficiency of T_1, but accurate estimation of large precision matrices can be challenging and requires special considerations (Fan, Liao & Liu, 2016).

The link between T_1 and PrediXcan (Gamazon et al., 2015) and TWAS (Gusev et al., 2016) for association and tissue-specific gene-expression data integration has also been noted (Xu et al. 2017). The performance of T_{ME} in this setting and comparison with other concurrently developed newer methods are of our future research interest.

Fix-, random- and mixed-effect models for summary statistics have been studied for meta-analyses of GWAS (Han & Eskin, 2011). In that context, a likelihood ratio test was implemented for a mixed-effect model, and the resulting test is also known as the new RE meta-analysis. The original test of Han & Eskin (2011) was designed for meta-analyses of independent studies, and a modified procedure has since been developed by Lee, Eskin & Han (2017) to account for correlations between studies but without adjusting for covariate effects. Comparisons between the two approaches for meta-analyses and other studies warrant future investigations.

Acknowledgements

We thank Dr. Lisa J. Strug and her lab, and Dr. Harriet Corvol for providing the cystic fibrosis application data, and we thank the Genetic Analysis Workshop 17 (GAW17) committee and the 1000 Genomes Project for providing the GAW17 application data. We would also like to thank Dr. Andriy Derkach for helpful discussions. YZ is a trainee of the CIHR STAGE (Strategic Training in Advanced Genetic Epidemiology) training program at the University of Toronto. This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-04934 and RGPAS-522594), the Canadian Institutes of Health Research (CIHR, MOP-310732),

and the University of Toronto McLaughlin Centre Accelerator Grants in Genomic Medicine (MC-2019-15).

References

[1] Almasy, L., Dyer, T., Peralta, J., Kent, J., Charlesworth, J., Curran, J., & Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proceedings, 5, S2*.

[2] Barnett, I., Mukherjee, R., & Lin, X. (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association*, 112, 64–76.

[3] Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88, 586–598.

[4] Brockwell, S., & Gordon, I. (2011). A Comparison of Statistical Methods for Meta-analysis. *Statistics in Medicine.*, 20, 825–840.

[5] Cai, T., Liu, W., & Xia, Y.(2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B*, 76, 349–372.

[6] Cheng, Q., Yang, Y., Shi, X., Yeung, K., Yang, C., Peng, H., & Liu, J.(2019). MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting for linkage disequilibrium and horizontal pleiotropy. *https://www.biorxiv.org/content/early/2019/08/03/684746*.

[7] 1000 Genomes Project Consortium & others,(2010). A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061.

[8] Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boëlle, P., Durie, P., Paterson, A., Rommens, J., Strug, L., & Sun, L.(2015). A joint location-scale test improves power to detect

associated SNPs, gene sets, and pathways. *The American Journal of Human Genetics*, 97, 125–138.

[9] Derkach, A., Lawless, J., & Sun, L.(2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29, 302–321.

[10] Derkach, A., Lawless, J., & Sun, L.(2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, 37, 110–121.

[11] Donoho, D., & Jin, J.(2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*,32,962–994.

[12] Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19, 1–32.

[13] Fan, Y., Jin, J., & Yao, Z. (2013). Optimal classification in sparse gaussian graphic model. *The Annals of Statistics*, 41, 2537–2571.

[14] Gamazon, E., Wheeler, H., Shah, K., Mozaffari, S., Aquino-Michaels, K., Carroll, R., Eyler, A., Denny, J., GTEx, Consortium, Nicolae, D., Cox, N., & Im, H.(2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47, 1091–1098.

[15] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B., Jansen, R., De G., Eco J., Boomsma, D., Wright, F., & others.(2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48, 245–252.

[16] Larson, N., McDonnell, S., Cannon, A., Teerlink, C., Stanford, J., Ostrander, E., Isaacs, W., Xu, J., Cooney, K., Lange, E., & others.(2017). gsSKAT: Rapid gene set analysis and multiple

testing correction for rare-variant association studies using weighted linear kernels. *Genetic Epidemiology*, 41, 297–308.

[17] Lee, S., Wu, M., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13, 762–775.

[18] Lee, S., Abecasis, G., Boehnke, M., & Lin, X. (2014). Rare variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* , 95, 5–23.

[19] Lee, C., Eskin, E., & Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics*, 33, 379–388.

[20] Ionita-Laza, I., McCallum, K., Xu, B., & Buxhaum, J. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48, 214–220.

[21] Liu, Z., & Lin, X. (2018a). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *Journal of the American Statistical Association*, 114, 975–990.

[22] Liu, Z., & Lin, X. (2018b). Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics*, 74, 165–175.

[23] Madsen, B., & Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* , 5, e1000384.

[24] Rothman, A. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99, 733–740.

[25] Sun, R., & Lin, X. (2019) Genetic Variant Set-Based Tests Using the Generalized BerkJones Statistic With Application to a Genome-Wide Association Study of Breast Cancer. *Journal of the American Statistical Association*.

[26] Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium *Genetic Epidemiology.*, 33, 497–507.

[27] Pan, W., Chen, Y., & Wei, P. (2015). Testing for polygenic effects in genome-wide association studies. *Genetic Epidemiology.*, 39, 306–316.

[28] Purcell, S., Wray, N., Stone, J., Visscher, P., O'donovan, M., Sullivan, P., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460, 748–752.

[29] Xu, Z., Wu, C., Wei, P., & Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. *Genetics* , 207, 893–902.

[30] Xu, G., Lin, L., Wei, P., & Pan, W. (2016). An adaptive two-sample test for high-dimensional. *Biometrika*, 103, 609–624.

[31] Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., & 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.

[32] Wu, Z., Sun, Y., He, S., Cho, J., Zhao, H., & Jin, J. (2014). Detection boundary and Higher Criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics*, 8, 824–851.