A stable and adaptive polygenic signal detection method based on repeated sample splitting

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

August 7, 2020

Abstract

Using polygenic risk score for trait association analyses and disease prediction are paramount for genetic studies of complex traits. Valid inference relies on sample splitting, or more recently external data, to obtain a set of potentially associated genetic variants, along with their weights, for polygenic risk score construction. The use of external data has been popular, but recent work increasingly calls its use into question due to adverse effects of potential data heterogeneity between different samples. Our study here adheres to the original sampling-splitting principle but does so, repeatedly, to increase stability of our inference. To accommodate different polygenic structures, we develop an adaptive test for generalized linear models. We provide the asymptotic null distributions of the proposed test for both fixed and diverging number of variants. We also show the asymptotic properties of the proposed test under local alternatives, providing insights on why power gain attributed to variable selection and weighting can compensate for efficiency loss due to sample splitting. We support our analytical findings through extensive simulation studies and an application.

Keywords: Adaptive; Polygenic risk score; Sample splitting; Signal detection; Stability.

1 Introduction

The polygenic risk score concept has recently been successfully applied to study many complex and heritable traits, by aggregating weak genetic effects across a large number of variants that do not, individually, achieve statistical significance (Purcell et al., 2009; Fritsche et al., 2018). A polygenic study has two stages, where stage one applies a variable selection procedure to a training sample to obtain a set of potentially associated variants and their corresponding weights. Using an independent testing sample, stage two first constructs a polygenic risk score for each individual by calculating the weighted sum of the risk alleles across the selected variants, and stage two then studies the aggregated score for its association with the trait of interest. A significant association implies that genetic signals are present among the selected variants, and the corresponding polygenic score can be used for disease prediction. The use of external data for variable selection and weight estimation has been popular, but recent work has shown that this approach can be extremely sensitive to sample mis-match beyond population stratification (Mostafavi et al., 2020). Thus, our work here focuses on methods that do not rely on the use of external data.

Focusing on association studies, the original polygenic method of Purcell et al. (2009) has been improved (Dudbridge, 2013; Zhou et al., 2013; Vilhjalmsson et al., 2015; Wu et al., 2019). For example, Wu et al. (2019) proposed an adaptive method, where the test statistic is based on different functions of variant-specific score statistics, derived from the *whole* sample across *all* variants, where different functions target different alternatives. However, accurate inference of the method of Wu et al. (2019) requires simulation or bootstrap, and the lack of variable selection can adversely affect power despite its use of the whole sample for association testing.

Sample splitting allows variable selection while preserving validity of an inference, and this principle has been used in many other study settings. However, variability inherent in one-time only sample splitting has been noted, including in variable selection (Wasserman and Roedern, 2009; Meinshausen et al., 2009), estimation (Kravitz et al., 2018), and more recently selective inference (Rinaldo et al., 2019; Romano and DiCiccio, 2019; Barber and Candes, 2019). Repeated sample splitting is a natural remedy, but it is not obvious how to aggregate information across multiple correlated sample splits to derive a valid and efficient test.

In this paper, we develop an adaptive and robust polygenic association test for testing high dimensional regression coefficients for generalized linear models. In section 2, we first review the classical polygenic association test, based on weighted sum of the risk alleles, and note its equivalency with weighted sum of variant-specific score statistics. We then consider different weighting factors for the score vector, where different weights are tailored to different alternatives while still leveraging information from the training sample. To aggregate information across different weight-

ing factors, we use the recent Cauchy method of Liu and Xie (2020), and we discuss the connection of our adaptive method with that of Wu et al. (2019). To improve stability of our inference, we then consider repeated sample splitting and utilize the Cauchy method again to aggregate information across multiple sample splits without explicitly modelling the inherent correlation. Finally, we derive the asymptotic null distributions of the proposed test for both fixed and diverging number of variants, and we study its asymptotic properties under local alternatives. In section 3 we present extensive simulation results for method evaluation and comparison, and in section 4 we provide an application. We conclude with discussion in section 5.

2 Methods

2.1 Notations

Let $Y \in \mathbb{R}^{n\times 1}$ be the outcome variable of interest, $G \in \mathbb{R}^{n\times J}$ the genotype matrix, and $X \in \mathbb{R}^{n\times q}$ the covariate matrix for a sample of size n with J genetic variants and q covariates. For clarity, let y_i be the response for individual i, g_{ij} the genotype for individual i and variant j, and $x_{ij'}$ the covariate value for individual i and covariate j', $i = 1,\ldots,n$, $j = 1,\ldots,J$, and $j' = 1,\ldots,q$. Further, let $G_i \in \mathbb{R}^{J\times 1}$ be the genotype vector for individual i, $G_j \in \mathbb{R}^{n\times 1}$ the genotype vector for variant j, $X_i \in \mathbb{R}^{q\times 1}$ the covariate vector for individual i, and $X_{j'} \in \mathbb{R}^{n\times 1}$ the vector for covariate j'. We assume that conditional on (G_i, X_i), y_i follows a distribution with density function, $f(y_i) = exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i,\phi)\}$ for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$, where θ is the canonical parameter, ϕ the dispersion parameter, $\text{var}(y_i|G_i,X_i) = a_i(\phi)v(\mu_i)$, and $v(\mu_i)$ the variance function (MacCullagh and Nelder, 1989). We consider the generalized linear model that models $\mu_i = b'(\theta_i) = E(y_i|G_i,X_i)$ for different types of response variables in the exponential family by a monotone and differentiable link function $\mathscr{G}(\cdot)$,

$$\mathscr{G}(\mu_i) = G_i^T\beta + X_i^T\beta_x,$$

where β and β_x are, respectively, J- and q-dimensional vectors of regression coefficients; q is fixed but J may vary depending on the study setting. Among the J genetic variables, we use $\mathscr{M}^*$ and $|\mathscr{M}^*|$ to denote, respectively, the set and number of truly associated ones. For simplicity but without loss of generality, we also assume that G_i and X_i have been mean centred at zero and standardized to have variance one.

2.2 The classical polygenic risk score for association tests

Suppose we have $2n$ independent observations, the classical polygenic risk score-based association analysis (Purcell et al., 2009) first randomly splits the sample into two equal subsets, $D_{n,1}$ and $D_{n,2}$; the corresponding data and parameter estimates such as y, X, G and $\hat{\beta}$ will carry superscripts $^{(1)}$ and $^{(2)}$, respectively for the two subsets, unless specified otherwise. A variable selection procedure is then applied to the training sample $D_{n,1}$ to select a subset of candidate genetic variables, $\mathscr{M}$, where we define $J_2 = |\mathscr{M}|$. In the testing sample $D_{n,2}$, a polygenic risk score G_i^* is constructed by aggregating the J_2 selected variables using $G_i^{(2)}$, but weighted by the effect estimates $\hat{\beta}_j^{(1)}$ obtained from $D_{n,1}$, $j \in \mathscr{M}$. That is, $G_i^* = \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} g_{ij}^{(2)}$ $(i = 1, \ldots, n)$. The inference is then based on the generalized linear regression model applied to $D_{n,2}$,

$$\mathscr{G}\{E(y_i^{(2)}|G_i^*, X_i^{(2)})\} = G_i^* \beta^* + X_i^{(2)T} \beta_x, \tag{1}$$

and testing

$$H_0 : \beta^* = 0 \text{ versus } H_1 : \beta^* \neq 0. \tag{2}$$

The corresponding score statistic is, $T_1 = \sum_{i=1}^{n} (y_i^{(2)} - \hat{\mu}_i^{(2)}) G_i^*$, where $\hat{\mu}_i^{(2)} = \mathscr{G}^{-1}(X_i^{(2)T} \hat{\beta}_x)$ and $\hat{\beta}_x$ is the maximum likelihood estimate of β_x under H_0. The distribution of standardized T_1 can be approximated by χ_1^2, and the p-value of a test based on T_1 will be denoted as p_1.

This classical polygenic association testing has since been improved on several fronts, including modelling dependency structure (i.e. linkage disequilibrium) between genetic variables (Vilhjalmsson et al., 2015) and better estimation of $\beta_j^{(1)}$ (Shi et al., 2016), among others (Lloyd-Jones et al., 2019). However, additional work are needed. To facilitate our discussion, it is instructive to re-formulate T_1 as the following,

$$\begin{aligned} T_1 &= \sum_{i=1}^{n} (y_i^{(2)} - \hat{\mu}_i^{(2)}) G_i^* = \sum_{i=1}^{n} (y_i^{(2)} - \hat{\mu}_i^{(2)}) \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} g_{ij}^{(2)} \\ &= \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} \sum_{i=1}^{n} (y_i^{(2)} - \hat{\mu}_i^{(2)}) g_{ij}^{(2)} = n \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} S_j, \end{aligned}$$

where $S_j = n^{-1} \sum_{i=1}^{n} (y_i^{(2)} - \hat{\mu}_i^{(2)}) g_{ij}^{(2)}$. Thus, T_1 constructed based on the aggregated risk score G_i^* is analytically equivalent to a *linear* weighted average of the score statistics, S_j's, across the J_2 selected genetic variants.

Tests based on T_1 are sub-optimal when signs of $\hat{\beta}_j^{(1)}$ and S_j differ. For example, recent work in association tests for rare variants have shown that T_1 type of tests are only powerful when a

large proportion of the variants being tested are causal *and* their genetic effects are in the same direction (Derkach et al., 2014). Further, the direct use of $\hat{\beta}_j^{(1)}$'s as weights may not be robust to different alternatives. Finally, when the signal-to-noise ratio is low as often the case in practice, the one-time only sample splitting approach may not be reliable (Meinshausen et al., 2009). Figure 1 is an illustration of the p-value lottery phenomenon associated with T_1, when it is applied to a real dataset with $2n = 1409$ and $J = 3754$; see Section 4 for details of the application data.

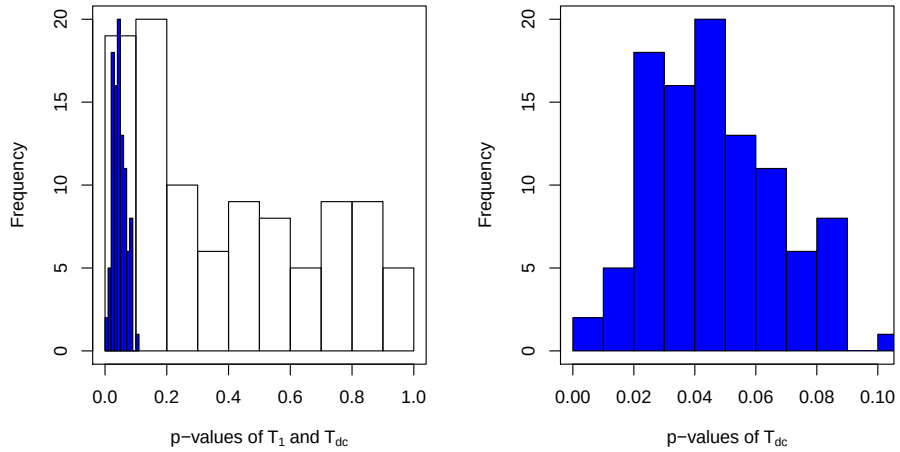

Figure 1: Histograms of p-values of T_1 (white) and the proposed T_{dc} (blue) based on randomly splitting a real dataset to training and testing samples, independently 100 times. The right figure is a zoom-in plot for the proposed T_{dc}.

2.3 An adaptive procedure for polygenic signal detection

Here we develop a robust method that is adaptive to different alternatives. We first propose new tests by considering different weighting schemes, given a particular sample split. We then improve the stability of our inference through repeated sample splitting.

Recall that testing (2) in (1) can be reformulated as testing

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0, \tag{3}$$

in

$$\mathscr{G}\{E(y_i^{(2)}|G_i^{(2)}, X_i^{(2)})\} = G_i^{(2)T}\beta + X_i^{(2)T}\beta_x, \tag{4}$$

where $\beta = (\beta_1, \ldots, \beta_{J_2})^T$. The proposed new test statistics have the following form,

$$T_\gamma = n \sum_{j=1}^{J_2} w_j^{\gamma-2} S_j^2,\ \gamma \in \Gamma = \{2,4,6,\ldots\}, \tag{5}$$

where w_j depends on $\hat{\beta}_j^{(1)}$ obtained from $D_{n,1}$, and γ is an even integer to avoid signal cancellation between variants.

Let $S = (S_1, \ldots, S_{J_2})^T$ and $R = \text{diag}\{r_j\} = \text{diag}\{w_j^{\gamma-2}\}$, $j = 1, \ldots, J_2$, we have

$$T_\gamma = nS^T RS. \tag{6}$$

We can easily modify R to include off-diagonal elements to reflect potential linkage disequilibrium between genetic variables, and we will study the asymptotic null distribution of T_γ in Theorems 1–3 for both fixed and diverging J_2.

The different γ values in (5) adapt to different signal sparsities. To obtain an accurate yet computationally efficient adaptive test, we propose to aggregate p_γ's, the p-values of $T_\gamma, \gamma \in \Gamma = \{2,4,6,\ldots\}$, and p_1, the p-value of T_1, using the Cauchy combination method recently proposed by Liu and Xie (2020). The Cauchy method can accommodate complex dependency structure among p-values without explicitly modelling it. In our setting, the proposed test statistics is

$$T_c = (|\Gamma| + 1)^{-1} \sum_{\gamma \in \Gamma \cup 1} \tan\{(0.5 - p_\gamma)\pi\}. \tag{7}$$

The tail of the null distribution of T_c can be well approximated by the standard Cauchy distribution, as long as the individual p_γ's are accurate, which we study in sections 2.4 and 3. The final p-value of T_c is $p_c = 1/2 - (\arctan t_c)/\pi$, where t_c is the observed value for T_c.

Here we acknowledge that T_γ is related to SPU type of test statistics proposed by Xu et al. (2016, 2017); Wu et al. (2019). For an integer $\gamma \geq 1$, $T_{\text{spu}(\gamma)} = \sum_{j=1}^{J} S_j^\gamma$, where S_j is obtained from the *whole* sample. If we omit the sample splitting step in our approach, $J_2 = J$, and let $w_j = S_j$, we have $T_\gamma \propto T_{\text{spu}(\gamma)}$ for all $\gamma > 1$. The authors of $T_{\text{spu}(\gamma)}$ have noted that for an even integer $\gamma \to \infty$, $T_{\text{spu}(\gamma)} \propto (\sum_j |S_j|^\gamma)^{1/\gamma} \to \max_j |S_j|$, defined as $T_{\text{spu}(\infty)}$. This suggests that larger γ is more powerful for sparse alternatives and smaller γ is the opposite. To make the SPU test robust to different alternatives, the authors then proposed an adaptive SPU test based on $T_{\text{aSPU}} = \min_{\gamma \in \Gamma_{\text{aSPU}}} \{p_{\text{spu}(\gamma)}\}$, where the recommended $\Gamma_{\text{aSPU}} = \{1,2,3,4,5,6,\infty\}$, and $p_{\text{spu}(\gamma)}$ is the p-value associated with each $T_{\text{spu}(\gamma)}$. The asymptotic $p_{\text{spu}(\gamma)}$ for $\gamma = 1$ and 2 can be obtained with mild conditions imposed on moments of S_j's and their correlation structure (Xu et al., 2017; Wu et al., 2019). However, the

asymptotic $p_{\text{spu}(\gamma)}$ for $\gamma > 2$ are not accurate for stringent significant levels. The authors then proposed to calculate $p_{\text{spu}(\gamma)}$, and subsequently p_{aSPU}, based on bootstrap or simulation, which is computational expensive.

The distinction between T_{aSPU} and the proposed T_c is four-fold. Firstly, the building block of T_c is T_γ, where γ is an even integer, which facilitates studies of asymptotic properties of the proposed tests; see Theorems 1–4 for details. Secondly, tests using different γ values are correlated with each other. Thus, even if the individual p-value estimation is accurate, the minimum-p approach of T_{aSPU} makes the inference more difficult than that of T_c, which is based on the easy-to-implement Cauchy method. Thirdly, although T_{aSPU} uses the whole sample for association testing, it aggregates information across all J genetic variants, many of which may be from the null leading to reduced power as compared to T_c, which benefits from variable selection. Lastly, the flexible structure of w_j in T_γ allows incorporation of additional information available for each genetic variant, such as its functional importance (Ionita-Laza et al., 2016).

To further robustify T_c against sampling variation inherent in the one-time only sample splitting approach, we then consider repeated sample splitting of m times. For the sth sample split, $s = 1, \ldots, m$, we obtain $T_{c,s}$ and its corresponding p-value, $p_{c,s}$. To combine the $p_{c,s}$'s while not explicitly modelling the correlation, we again utilize the Cauchy method of Liu and Xie (2020). The proposed double Cauchy combination test statistic is

$$T_{dc} = m^{-1} \sum_{s=1}^{m} \tan\{(0.5 - p_{c,s})\pi\}. \tag{8}$$

Similar to inference based on T_c, the tail of the null distribution of T_{dc} can be well approximated by the standard Cauchy distribution, as long as the individual p-values to be combined are accurate which we study next.

2.4 Asymptotic properties of T_γ

To make the dependency of T_γ on n and J_2 explicit, we use $T_{n,J_2,\gamma}$ to denote T_γ in this section. We study the asymptotic properties of $T_{n,J_2,\gamma}$ for both fixed and diverging J_2, under the null or local alternatives. For notation simplicity, we now omit superscript $^{(2)}$ from $Y \in \mathbb{R}^{n\times 1}$, $G \in \mathbb{R}^{n\times J_2}$ and $X \in \mathbb{R}^{n\times q}$, representing, respectively, the outcome, genotype and covariate data in the testing sample $D_{n,2}$, where J_2 is the number of SNPs to be tested. Recall that $T_{n,J_2,\gamma} = nS^T RS$, where $S = (S_1, \ldots, S_{J_2})^T$ is the score vector, $R = \text{diag}\{r_j\}$, and γ is an even integer. The covariance matrix of $n^{1/2}S$ is $\Sigma_s = E\{a_i(\phi)v(\mu_i)G_i G_i^T\}$, where $G_i \in \mathbb{R}^{J_2\times 1}$, the genotype vector for individual i, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T = Y - \mathscr{G}^{-1}(G\beta + X\beta_x)$, and $\varepsilon_0 = (\varepsilon_{01}, \ldots, \varepsilon_{0n})^T = Y - \mathscr{G}^{-1}(X\beta_x)$.

The following theorem gives the asymptotic null distribution of $T_{n,J_2,\gamma}$, provided that the same regularity conditions, required for the convergence of S to a multivariate normal random variable, hold (Goeman et al., 2011; Vandekar et al., 2019). In addition, we ignore the nuisance parameters $a_i(\phi)$ and β_x for now and discuss how to include them in section 5. We provide all proofs in the Supplementary Material.

Theorem 1. *Under the null hypothesis H_0 in (3), for any fixed finite J_2 and γ,*

$$T_{n,J_2,\gamma} \to T_{J_2,\gamma}$$

in distribution as $n \to \infty$, where $T_{J_2,\gamma}$ and $\sum_{j=1}^{J_2} \lambda_{J_2,j}\chi^2_{1j}$ are equivalent in distribution, χ^2_{1j}'s are independent variables with central chi-square distribution with 1 degrees of freedom, χ^2_1, $\lambda_{J_2,1} \geq \ldots, \geq \lambda_{J_2,J_2}$ are the eigenvalues of $C_s^T R C_s$, and $\Sigma_s = C_s C_s^T$.

When Y is normally distributed, $T_{n,J_2,\gamma}$ and $T_{J_2,\gamma}$ equivalent in distribution always holds for any n (and finite J_2); when both n and J_2 are diverging, additional assumptions are required.

Assumption 1. *Assume $G_i = C_g Z_i, \forall i$, where C_g is a $J_2 \times J_2$ matrix and $C_g C_g^T = \Sigma_g$, and $Z_i = (z_{i1}, \ldots, z_{iJ_2})^T$ with $E(Z_i) = 0$ and $cov(Z_i) = I_{J_2}$. Assume z_{ij} has finite eighth moment and $E(z_{ij}^4) = 3 + \Delta < \infty$, $\forall j$, where Δ is a constant and $\Delta > -3$, and $E(\Pi_j z_{ij}^{\nu_j}) = \Pi_j E(z_{ij}^{\nu_j})$, where $\sum_j \nu_j \leq 8$ and all ν_j's are non-negative integers.*

Assumption 2. *Let f_g be the probability density of G and $D(f_g)$ be its support. Assume $E(\varepsilon \mid G) = 0$ and $E(\varepsilon^3 \mid G) = 0$, and there are positive constants K_1 and K_2 such that $E(\varepsilon^2 \mid G) > K_1$ and $E(\varepsilon^4 \mid G) < K_2$ almost everywhere for $g \in D(f_g)$.*

Assumption 3. *There exist real numbers $\rho_{\infty,j}$'s such that $\lim_{J_2 \to \infty} \rho_{J_2,j} = \rho_{\infty,j}$ uniformly $\forall j$, and $\lim_{J_2 \to \infty} \sum_{j=1}^{J_2} \rho_{J_2,j} = \sum_{j=1}^{\infty} \rho_{\infty,j} < \infty$, where $\rho_{J_2,j} = \lambda_{J_2,j}/\sqrt{tr(R\Sigma_s)^2}$, $j = 1, \ldots, J_2$, which are the eigenvalues of $C_s^T R C_s/\sqrt{tr(R\Sigma_s)^2}$ in descending order.*

Assumption 4. *$n\{tr(R\Sigma_g)^2/tr^2(R\Sigma_g)\} \to \infty$ as n and $J_2 \to \infty$.*

Assumptions 1–3 are standard in studying high-dimensional testing (Guo and Chen, 2016; Zhang et al., 2019). Assumption 4 specifies a relationship between n and J_2. Because $tr(R\Sigma_g) = \sum_{j=1}^{J_2} \lambda_{J_2,j}$ and $tr(R\Sigma_g)^2 = \sum_{j=1}^{J_2} \lambda_{J_2,j}^2$, we have $n\{tr(R\Sigma_g)^2/tr^2(R\Sigma_g)\} = n\{\sum_{j=1}^{J_2} \lambda_{J_2,j}^2/(\sum_{j=1}^{J_2} \lambda_{J_2,j})^2\}$. Thus, Assumption 4 holds for any diverging n and J_2 if $\lambda_{J_2,j}$'s are dominated by first few larger ones. When all $\lambda_{J_2,j}$'s are similar in magnitude, Assumption 4 is equivalent to requiring sample size n grows to infinity at a rate faster than J_2. The following theorem generalizes Theorem 1 from finite to infinite J_2.

Theorem 2. *Under the null hypothesis H_0 in (3) and assume Assumptions 1–4 hold,*

$$\sigma_{n,0}^{-1}\{T_{n,J_2,\gamma} - tr(R\Sigma_s)\} \to \zeta \text{ and } \{2tr(R\Sigma_s)^2\}^{-1/2}\{T_{J_2,\gamma} - tr(R\Sigma_s)\} \to \zeta$$

in distribution as n and $J_2 \to \infty$, where ζ and $\sum_{j=1}^{\infty} \rho_{\infty,j}(\chi_{1j}^2 - 1)/\sqrt{2}$ are equivalent in distribution, $\sigma_{n,0}^2 = 2tr(R\Sigma_s)^2 + \delta$, and $\delta = n^{-1}\left\{\sum_{j=1}^{J_2}\sum_{k=1}^{J_2} r_j r_k E(g_{ij}^2 g_{ik}^2 \varepsilon_{0i}^4) - tr^2(R\Sigma_s) - 2tr(R\Sigma_s)^2\right\} = o\{tr(R\Sigma_s)^2\}$. Therefore, as n and $J_2 \to \infty$,

$$sup_x |pr(T_{n,J_2,\gamma} \le x) - pr(T_{J_2,\gamma} \le x)| \to 0.$$

Theorems 1 and 2 show that we can use $\sum_{j=1}^{J_2} \lambda_{J_2,j}\chi_{1j}^2$ to approximate the asymptotic null distribution of $T_{n,J_2,\gamma}$ for both fixed and diverging J_2. The corresponding p-value can be calculated using the method of Imhof (1961) or Davies (1980); we use the method of Davies (1980) in our implementation below for both simulation and application studies.

To show the asymptotic normality of $T_{n,J_2,\gamma}$ under the null, we need to impose the following assumption, which substitutes for specifying an explicit relationship between J_2 and n.

Assumption 5. *$tr^2(R\Sigma_g)^2/tr(R\Sigma_g)^4 \to \infty$ and $tr(R\Sigma_g)^2 \to \infty$ as n and $J_2 \to \infty$.*

Theorem 3. *Under the null hypothesis H_0 in (3) and assume Assumptions 1–5 hold,*

$$\sigma_{n,0}^{-1}\{T_{n,J_2,\gamma} - tr(R\Sigma_s)\} \to N(0,1),$$

in distribution as n and $J_2 \to \infty$.

We now study the adverse effect of sample splitting on power, while considering the beneficial effect of variable selection afforded by sample splitting, under the local alternative $\mathscr{L}_\beta$,

$$\mathscr{L}_\beta = \left\{\Delta_\beta^T R\Sigma_g R\Delta_\beta = o\{n^{-1}tr(R\Sigma_g)^2\} \text{ and } \{\mathscr{G}^{-1}(G_i^T\beta)\}^2 = O(1)\right\},$$

where $\Delta_\beta = E\{\mathscr{G}^{-1}(G_i^T\beta)G_i\}$.

Theorem 4. *Under the local alternative $\mathscr{L}_\beta$ and assume Assumptions 1–5 hold,*

$$\sigma_{n,1}^{-1}\{T_{n,J_2,\gamma} - tr(R\Sigma_s) - \mu_{n,\beta}\} \to N(0,1),$$

in distribution as n and $J_2 \to \infty$, where $\mu_{n,\beta} = tr(R\Xi_\beta) + (n-1)\Delta_\beta^T R\Delta_\beta$, $\sigma_{n,1}^2 = \{2tr(R\Sigma_s + R\Xi_\beta)^2\}\{1 + o(1)\}$, and $\Xi_\beta = E\left[\{\mathscr{G}^{-1}(G_i^T\beta)\}^2 G_i G_i^T\right]$.

Theorem 4 reveals that power of $T_{n,J_2,\gamma}$ under $\mathscr{L}_\beta$ is determined by $\text{SNR}_n(\beta) = \mu_{n,\beta}/\sigma_{n,1}$, where $\text{SNR}_n(\beta)$ can be interpreted as signal-to-noise ratio following Guo and Chen (2016). As detailed in the Supplementary Material,

$$\mu_{n,\beta} = \sum_{j=1}^{J_2} r_j E\{\mathscr{G}^{-1}(G_i^T\beta)g_{ij}\}^2 + (n-1)\sum_{j=1}^{J_2} r_j E^2\{\mathscr{G}^{-1}(G_i^T\beta)g_{ij}\},$$
$$\sigma_{n,1}^2 = \{\sigma_{n,0}^2 + 2tr(R\Xi_\beta)^2 + 4tr(R\Sigma_s R\Xi_\beta)\}\{1+o(1)\},$$

where $\sigma_{n,0}^2 = 2\sum_{j=1}^{J_2}\sum_{k=1}^{J_2} r_j r_k E^2(g_{ij}g_{ik}\varepsilon_i^2)\{1+o(1)\}$, $tr(R\Xi_\beta)^2 = \sum_{j=1}^{J_2}\sum_{k=1}^{J_2} r_j r_k E^2\left[g_{ij}g_{ik}\{\mathscr{G}^{-1}(G_i^T\beta)\}^2\right]$, and $tr(R\Sigma R\Xi_\beta) = \sum_{j=1}^{J_2}\sum_{k=1}^{J_2} r_j r_k E(g_{ij}g_{ik}\varepsilon_i^2)E\left[g_{ij}g_{ik}\{\mathscr{G}^{-1}(G_i^T\beta)\}^2\right]$.

Now define $T_{2n,J} = 2n\sum_{j=1}^{J} S_j^2$ as the test statistic calculated based on the *whole* sample of size $2n$ but *without* variable selection and assuming $R = I$. In this case, $G_j \in \mathbb{R}^{2n\times 1}$ and $\beta \in \mathbb{R}^{J\times 1}$ for calculating S_j. The signal-to-noise ratio corresponding to $T_{2n,J}$ is $\text{SNR}_{2n}(\beta) = \mu_{2n,\beta}/\sigma_{2n,1}$.

$$\mu_{2n,\beta} = \sum_{j=1}^{J} E\{\mathscr{G}^{-1}(G_i^T\beta)g_{ij}\}^2 + (2n-1)\sum_{j=1}^{J} E^2\{\mathscr{G}^{-1}(G_i^T\beta)g_{ij}\},$$
$$\sigma_{2n,1}^2 = \{\sigma_{2n,0}^2 + 2tr(\Xi_{\beta,J})^2 + 4tr(\Sigma_{s,J}\Xi_{\beta,J})\}\{1+o(1)\},$$

where $\sigma_{2n,0}^2 = 2\sum_{j=1}^{J}\sum_{k=1}^{J} E^2(g_{ij}g_{ik}\varepsilon_i^2)\{1+o(1)\}$, $tr(\Xi_{\beta,J})^2 = \sum_{j=1}^{J}\sum_{k=1}^{J} E^2\left[g_{ij}g_{ik}\{\mathscr{G}^{-1}(G_i^T\beta)\}^2\right]$, and $tr(\Sigma_{s,J}\Xi_{\beta,J}) = \sum_{j=1}^{J}\sum_{k=1}^{J} E(g_{ij}g_{ik}\varepsilon_i^2)E\left[g_{ij}g_{ik}\{\mathscr{G}^{-1}(G_i^T\beta)\}^2\right]$.

To provide additional insights, assume $pr(\mathscr{M} \supset \mathscr{M}^*) \to 1$ as $n \to \infty$; this assumption can be fulfilled by various variable selection algorithms (Fan and Lv, 2008; Li et al., 2012; Jin et al., 2014; Zhang, 2017). Comparing $\mu_{n,\beta}$ with $\mu_{2n,\beta}$, it is not surprising that sample size reduction is the primary cause of power loss for a sample splitting-based method. On the other hand, the characterizations of $\sigma_{n,1}^2$ and $\sigma_{2n,1}^2$ indicate noise-filtering in the training sample $D_{n,1}$ can reduce variance of the test statistic calculated in the testing sample $D_{n,2}$, leading to increased power for the sample splitting-based method. The use of weights derived from $D_{n,1}$ can further compensate efficient loss due to reduced sample size in $D_{n,2}$. Thus, $\text{SNR}_n(\beta)$ can be larger than $\text{SNR}_{2n}(\beta)$, and tests based on $T_{n,J_2,\gamma}$ can be more powerful than $T_{2n,J}$. Simulation studies in the next section show that, even if sure-screening fails in $D_{n,1}$, the sample splitting approach can have comparable power with the methods of Wu et al. (2019); Guo and Chen (2016) applied to the full-sample without variable selection.

3 Simulation studies

3.1 Simulation designs

To evaluate the performance of T_{dc} and compare it to tests proposed by Guo and Chen (2016) and Wu et al. (2019), we consider two simulation designs. Design one simulates G, where we can explicitly specify and study effects of different model parameters, while design two builds upon a set of real genotypes. Design one considers sample size of $2n = 200$ or 1500 and dimension $J \in \{10, 50, 200, 400, 1000, 4000\}$. It generates G based on a multivariate normal distribution with mean vector 0 and correlation matrix $\Sigma_g = \{\rho^{|i-j|}\}_{J\times J}$, where $\rho = 0.2, 0.5, 0.8$, and i and $j = 1, \ldots, J$. That is, we use the autoregressive model for correlation between the J variants. For design two, G comes from an application sample of $2n = 1409$ individuals and a set of $J = 3754$ genetic variants; for a more streamlined presentation, we present simulation results of design two in section 4, along with application results.

To implement T_{dc}, we let $r_j = \hat{\beta}_j^{\gamma-2}$ $(j = 1, \ldots, J_2)$ and $\Gamma = \{2, 4, 6, 42\}$ to first obtain T_c of (7), we then use $m = 10$ or 50 to derive the more stable T_{dc} of (8). For completeness, we also study the performance of the individual T_γ's but present the results in the Supplementary Material. The numbers of simulation replicates are 10^6 for evaluating type I error control and 500 for power, and additional simulation design details are provided below when appropriate.

3.2 Type I error

Methods applied to binary outcomes often have worse performance than normally distributed traits. Thus, we generate Y based on a logistic regression with $\beta = 0$, and without loss of generality, intercept equals to one and with no other covariates. For evaluation of type I error control, the variable selection procedure is not critical as long as it is applied to a sample that is independent of the testing sample as the case here, and the size of J is also not critical. Thus, we choose $J_2 = J$ and regress the simulated Y on each of the J_2 simulated variants in $D_{n,1}$, and we obtain their corresponding $\hat{\beta}_j$ values for the polygenic association testing in $D_{n,2}$.

Table 1 shows the empirical test sizes of T_{dc} for $2n = 200$, $m = 10$, and nominal α values of 0.05, 0.01, and 10^{-3}, and Supplementary Material provides results for more stringent α values and $m = 50$; results for $2n = 1500$ are more accurate thus not shown. Here, the distributions of T_γ's are approximated by the weighted linear combination of independent χ^2_1 distributions as specified in Theorem 2, and the distributions of T_c and T_{dc} by the standard Cauchy distribution. It is clear that type I error of T_{dc} is controlled at or below the nominal levels, across the settings.

Table 1: Empirical test sizes of T_{dc}. Autoregressive model AR(1, ρ) for correlation between the J_2 variants, sample size $2n = 200$, and $m = 10$ for repeated sample splitting

J_2	10			50			200			400			1000		
ρ \ α	5%	1%	0.1%	5%	1%	0.1%	5%	1%	0.1%	5%	1%	0.1%	5%	1%	0.1%
0.2	5.285	0.981	0.105	5.330	0.938	0.071	5.341	0.928	0.072	5.403	0.943	0.074	5.434	0.941	0.081
0.5	5.265	0.961	0.095	5.340	0.936	0.069	5.372	0.941	0.071	5.369	0.938	0.072	5.409	0.941	0.076
0.8	5.347	0.983	0.084	5.330	0.942	0.070	5.343	0.937	0.068	5.338	0.940	0.073	5.326	0.923	0.076

Results in the Supplementary Material show that the normal approximation given in Theorem 3, however, is not adequate for T_γ when $J_2 \leq 1000$. Thus, we recommend the use of the $\sum_{j=1}^{J_2} \lambda_{J_2,j}\chi^2_{1j}$ approximation in practice. Consistent with previous reports, the empirical type I error rates of the methods of Wu et al. (2019) and Guo and Chen (2016), based on asymptotic approximations, are inflated (Supplementary Material). Thus, we use permutation with 10^4 replicates to evaluate power of these methods, as recommended by the authors.

3.3 Power

Similar to the type I error evaluation above, here we also focus on the more difficult case of analyzing binary outcomes than normally distributed traits. We generate Y based on logistic models with different proportions of nonzero regression coefficients, varying from 0.1%, 1%, 5%, to 10% for the J variants. We assume the indices of the nonzero β_j's to be uniformly distributed in $\{1,\ldots,J\}$, and we also randomly specify the nonzero β_j's to be half positive and half negative. Results below focus on power comparison between the proposed T_{dc} test and the methods of Guo and Chen (2016) and Wu et al. (2019), which are applied to the whole sample and without variable selection. Results of the original polygenic risk score test, T_1, are shown in the Supplementary Material.

To better delineate the factors influencing power, we consider three study scenarios. In all three scenarios, the weights inferred from the training sample $D_{n,1}$ are used to construct the T_{dc} test statistic using the testing sample $D_{n,2}$.

(I), Oracle: $\mathscr{M} = \mathscr{M}^*$. This is the 'best' case scenario for T_{dc}, where the selection step applied to $D_{n,1}$ identifies all and only truly associated variants; the estimated weights however may not be optimal. This study is to show power gain of T_{dc}, despite the reduction of sample size, as compared to the methods without variable selection.

(II), $J_2 = J$: $\mathscr{M} = \{G_1,\ldots,G_J\}$. This is the 'worst' case scenario for T_{dc}, where the selection step fails completely at filtering out non-signals; the estimated weights however may be infor-

mative. This scenario is tailored for studying power loss of T_{dc} due to sample size reduction as compared with methods without sample splitting, while also demonstrating the benefits of leveraging the weights inferred from $D_{n,1}$ to perform associate analysis using $D_{n,2}$.

(III), Variable Selection: $\mathscr{M}$ is estimated based on the variable selection method proposed in Li et al. (2012). This study investigates the impact of accuracy of variable selection on power of T_{dc} as compared to the methods without variable selection.

Figure 2 shows the results for $2n = 1500$ and $J = 4000$, reflecting the values observed in the real dataset studied in section 4; $\rho = 0.5$ for correlation between the J variants, and $m = 10$ for repeated 50%-50% sample splitting to construct T_{dc}. Supplementary Material provides additional simulation results.

For scenario (I), Fig. 2 (the first column) shows that the proposed T_{dc} test has substantial power gain, attributed to noise filtering despite the reduction in sample size for associate testing, as compared to the methods of Guo and Chen (2016) and Wu et al. (2019).

For scenario (II), Fig. 2 (the second column) shows that the anticipated power loss of T_{dc} due to sample splitting can be compensated by leveraging the weights inferred from $D_{n,1}$, as compared with the methods of Guo and Chen (2016) and Wu et al. (2019), which use the full sample. For the sparse alternative case, scenario (II) 4 signals, T_{dc} displays comparable power with the method of Wu et al. (2019), while both are substantially more powerful than the method of Guo and Chen (2016). For the other alternatives considered in this scenario, all three methods have comparable power with the method of Guo and Chen (2016) having slightly higher power. Overall, the proposed T_{dc} test is most robust to the different alternatives considered here, and it is computationally most efficient.

For the more realistic scenario (III), interestingly the results are similar to those of scenario (II). This suggests that while the variable selection step filters out noise, it also filters out some (weak) signals. The implementation of the method of Li et al. (2012) requires the specification of J_2. The results in Fig. 2 are for $J_2 = 1500$, the total sample size as recommended by the authors. In addition to the method of Li et al. (2012), we also evaluated other selection methods, such as ElasticNet (Zou and Hastie, 2005). However, results are similar especially for weak signals.

Results so far focus on the 50%-50% sample splitting proportion. To exam the effects of different proportions on power, we also investigate 33%-67% and 67%-33% sample splitting (Supplementary Material). Results show that the overall power of T_{dc} is not extremely sensitive to the proportion. However, the 33%-67% sample splitting has slightly increased power for the scenarios considered here. This is consistent with the literature (Barber and Candes, 2019; Dudbridge, 2013), where it has been noted that uneven sample splitting, with more subjects assigned to the

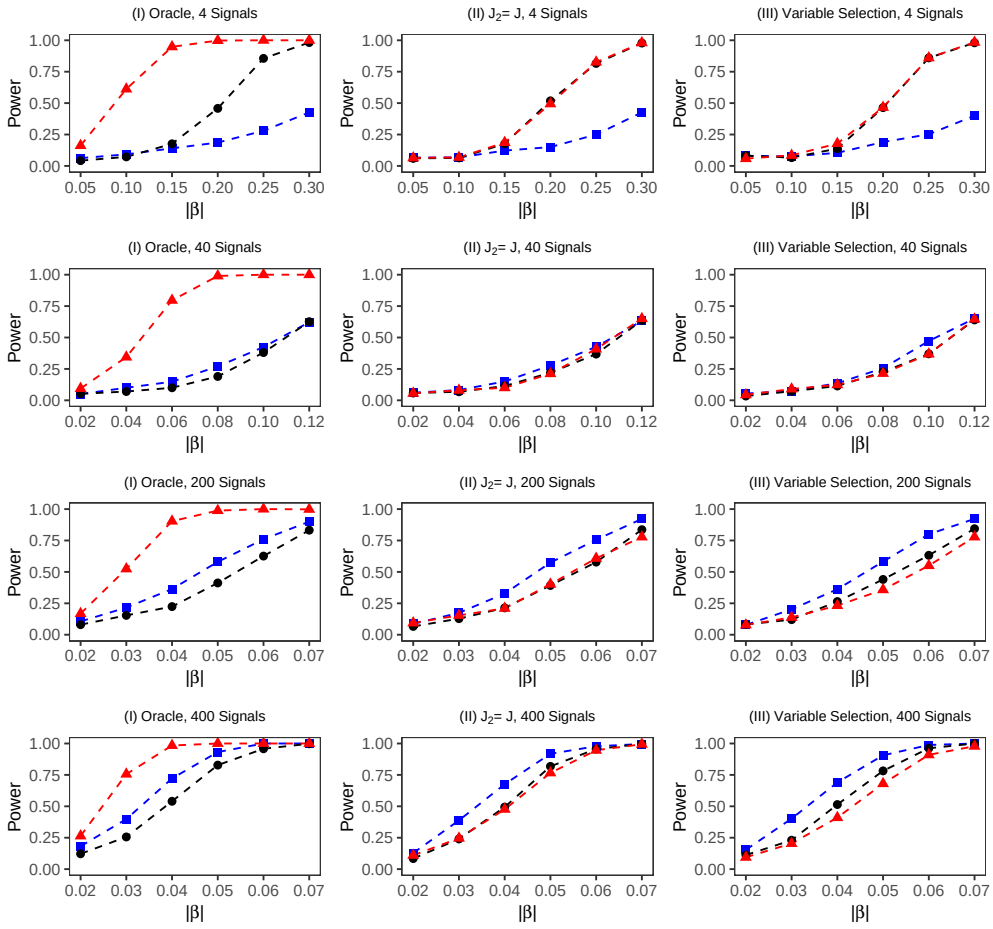

Figure 2: Power comparison of the proposed test T_{dc} (red triangle), the method of Guo and Chen (2016) (blue square), and the method of Wu et al. (2019) (black circle), for the three study scenarios (I), (II) and (III). Sample size $2n = 1500$, and the total variant $J = 4000$ among which 0.1% (row 1), 1% (row 2), 5% (row 3), and 10% (last row) are truly associated.

testing sample, can increase power as compared to even sample splitting.

4 Application and additional simulation studies

We apply the proposed T_{dc} test and the methods of Guo and Chen (2016) and Wu et al. (2019), as well as T_1, the original polygenic risk score test, to the cystic fibrosis data introduced in Soave et al. (2015). This dataset consists of $2n = 1409$ independent individuals from Canada with cystic fibrosis on whom lung functions have been measured. Of interest is the association between lung function and a set of $J = 3754$ genetic variants, which are the constituents of the apical plasma membrane.

To implement the proposed T_{dc} test, we first randomly divide the 1409 individuals into two subsets, D_{n_1} and D_{n_2}, where $n_1 = 409$ and $n_2 = 1000$, $n_1 = 705$ and $n_2 = 704$, or $n_1 = 1000$ and $n_2 = 409$. As in the simulation studies, we define $r_j = \hat{\beta}_j^{\gamma-2}$ $(j = 1, \ldots, J_2)$ and $\Gamma = \{2, 4, 6, 42\}$, and we apply the variable selection method of Li et al. (2012) and let $J_2 = n_2$, the sample size of the testing sample. Because the approximation of the asymptotic distribution of T_{dc} requires a positive definite matrix estimate of Σ_g in D_{n_2}, we use the algorithm proposed by Rothman (2012), with the tuning parameter selected by 5-fold cross-validations.

Using this application dataset, we first re-evaluate type I error control of T_{dc} by simulating Y independently of the observed G for the n_2 individuals in the D_{n_2} testing sample, where $y_i = 1 + \varepsilon_i$ $(i = 1, \ldots, n_2)$ and ε_i follows the standard normal distribution. The variable selection and estimation of weights, however, are based on the real data, both Y and G, of the D_{n_1} training sample. Results in Table S5 in the Supplementary Material show that the empirical type I error rates of T_{dc} are at the nominal level when $m = 10$ or 50 but are slightly inflated when $m = 100$. Thus, we recommend the use of $m = 10$ or 50 for constructing T_{dc}.

In the absence of knowledge of true association, application results focus on stability of tests based on sample splitting, and p-values of all methods. The empirical p-values are 0.0985 and 0.0727, respectively, for the methods of Guo and Chen (2016) and Wu et al. (2019), based on 10^4 bootstrap samples applied to the whole sample. For T_1, we randomly split the whole sample to the D_{n_1} and D_{n_2} subsets, independently 100 times, to obtain 100 different p_1's, the T_1-based p-values. The histogram of p_1's for $n_1 = 409$ and $n_2 = 1000$ is shown in Fig. 1. For T_{dc}, we also randomly split the whole sample to two subsets, but independently 100×50 times, and use a sequence of $m = 50$ repeated sample splits to obtain 100 p_{dc}'s, the T_{dc}-based p-values. The histogram of p_{dc}'s for $n_1 = 409$ and $n_2 = 1000$ is shown in Fig. 1 in blue. Results clearly show that the proposed repeated sample splitting strategy leads to a much more stable inference than the one-time only

Table 2: Summary of p-values of the proposed T_{dc} applied to the cystic fibrosis application data based on different n_1-n_2 sample splits, with $m = 50$ for constructing T_{dc} and repeatedly 100 times. The total sample size $2n = 1409$, the total genetic variants $J = 3754$, and $J_2 = n_2$, the sample of size of the testing sample

n_1	n_2	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
409	1000	0.003	0.030	0.045	0.046	0.061	0.101
705	704	0.020	0.067	0.079	0.086	0.105	0.195
1000	409	0.003	0.045	0.059	0.057	0.067	0.104

sample splitting approach: p_1 ranges from 0.0019 to 0.9446, while in contrast p_{dc} ranges from 0.003 to 0.101, with a mode of around 0.05.

For completeness, Table 2 provides the summary statistics of the 100 p_{dc}'s associated with different sample splitting proportions. Results show that although the proposed T_{dc} demonstrates much improved stability of inference as compared to T_1, variation remains in this application, suggesting the signals are too weak or sample size is not sufficient.

5 Discussion

In the theoretical study, we did not consider the impact of estimating nuisance parameters β_x and ϕ, as we expect that the results would be similar when we impose stringent conditions on the design matrix X and the relationship between n and q to ensure estimation accuracy of the nuisance parameters (Guo and Chen, 2016). In practice, we can estimate the nuisance parameters in the training sample and treat the estimates as known quantities to construct T_{dc} in the testing sample. This approach has been recommended by Chernozhukov et al. (2018) for another study setting where the sample splitting strategy is used.

Acknowledgement

We thank Dr. Lisa J. Strug and her lab for providing the cystic fibrosis data for application and helpful discussion. YZ is a trainee of the STAGE training program at the University of Toronto. This research is funded by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, and the University of Toronto McLaughlin Centre.

References

Barber, R., and E. Candes, 2019: A knockoff filter for high-dimensional selective inference. *Ann. Statist.*, **47**, 2504–2537.

Chernozhukov, V., D. Chetverikov, M. Demirer, C. H. Esther Duflo, W. Newey, and J. Robins, 2018: Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, **21**, C1–C68.

Davies, R., 1980: Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *J. Royal Stat. Soc.* C, **29**, 323–333.

Derkach, A., J. F. Lawless, and L. Sun, 2014: Pooled association tests for rare genetic variants: a review and some new results. *Stat. Sci.*, **29**, 302–321.

Dudbridge, F., 2013: Power and predictive accuracy of polygenic risk scores. *PLoS. Genet.*, **9**, e1003 348.

Fan, J., and J. Lv, 2008: Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B, **70**, 849–911.

Fritsche, L., S. Gruber, Z. Wu, ..., and B. Mukherjee, 2018: Association of polygenic risk scores for multiple cancers in a phenome-wide study: Results from the michigan genomics initiative. *Am. J. Hum. Genet.*, **102**, 1048–1061.

Goeman, J. J., H. C. Van Houwelingen, and L. Finos, 2011: Testing against a high-dimensional alternative in the generalized linear model: asymptotic type i error control. *Biometrika*, 381–390.

Guo, B., and S. Chen, 2016: Tests for high dimensional generalized linear models. *J. R. Statist. Soc.* B, **78**, 1079–1102.

Imhof, J., 1961: Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

Ionita-Laza, I., K. McCallum, B. Xu, and J. Buxhaum, 2016: A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.

Jin, J., C. Zhang, and Q. Zhang, 2014: Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.*, **15**, 2723–2772.

Kravitz, E., R. Carroll, and D. Ruppert, 2018: Sample splitting as an m-estimator with application to physical activity scoring. *Am. J. Hum. Genet.*, **102**, 1048–1061.

Li, R., W. Zhong, and L. Zhu, 2012: Feature screening via distance correlation learning. *J. Am. Stat. Assoc.*, **107**, 1129–1139.

Liu, Y. W., and J. Xie, 2020: Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.*, **115**, 393–402.

Lloyd-Jones, L. R., J. Zeng, J. Sidorenko, ..., and A. Metspalu, 2019: Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.*, **10**, 1–11.

MacCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models*. London: CRC Press.

Meinshausen, N., L. Meier, and P. Buhlmann, 2009: p-values for high-dimensional regression. *J. Am. Stat. Assoc.*, **104**, 1671–1681.

Mostafavi, H., A. Harpak, D. Conley, J. Pritchard, and M. Przeworski, 2020: Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, **9**, e48 376.

Purcell, S. M., N. R. Wray, J. L. Stone, ..., and I. S. Consortium., 2009: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460 (7256)**, 748–752.

Rinaldo, A., L. Wasserman, and M. G'Sell, 2019: Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.*, **47**, 3438–3469.

Romano, J., and C. DiCiccio, 2019: Multiple data splitting for testing. *Technical Report*.

Rothman, A. J., 2012: Positive definite estimators of large covariance matrices. *Biometrika*, **99 (3)**, 733–740.

Shi, J., J. Park, J. Duan, ..., and M. Garcia-Closas, 2016: Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS. Genet.*, **12**, e1006 493.

Soave, D., and Coauthors, 2015: A joint location-scale test improves power to detect associated snps, gene sets, and pathways. *Am. J. Hum. Genet.*, **97**, 125–138.

Vandekar, S., P. Reiss, and R. Shinohara, 2019: Interpretable high-dimensional inference via score projection with an application in neuroimaging. *J. Am. Stat. Assoc.*, **114**, 820–830.

Vilhjalmsson, B., J. Yang, H. Finucane, ..., and T. Hayeck, 2015: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, **97**, 576–592.

Wasserman, L., and K. Roedern, 2009: High-dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201.

Wu, C., G. Xu, and W. Pan, 2019: An adaptive test on high-dimensional parameters in generalized linear models. *Statistica Sinica*, **29**, 2163–2186.

Xu, G. J., L. F. Lin, P. Wei, and W. Pan, 2016: An adaptive two-sample test for high-dimensional. *Biometrika*, **103**, 609–624.

Xu, Z., C. Wu, P. Wei, and W. Pan, 2017: A powerful framework for integrating eqtl and gwas summary data. *Genetics*, **207**, 893–902.

Zhang, J., J. Guo, B. Zhou, and M. Y. Cheng, 2019: A simple two-sample test in high dimensions based on l2-norm. *J. Am. Statist. Assoc.*, **115**, 1011–1027.

Zhang, Y., 2017: Recovery of weak signal in high dimensional linear regression by data perturbation. *Electron. J. Stat.*, **11**, 3226–3250.

Zhou, X., P. Carbonetto, and M. Stephens, 2013: Polygenic modeling with bayesian sparse linear mixed models. *PLoS. Genet.*, **9**, e1003 264.

Zou, H., and T. Hastie, 2005: Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.