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

Yanyan ZHAO¹,²,³ and Lei SUN³,⁴*

¹Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, 250012, Shandong, China
²National Institute of Health Data Science of China, Jinan, 250003, Shandong, China
³Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada M5S 3G3
⁴Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada M5T 3M7

Key words and phrases: Adaptive; high-dimensional regression; polygenic risk score; robustness; sample splitting.

MSC 2020: Primary 62F03; secondary 62F40.

Abstract: Focusing on polygenic signal detection in high-dimensional genetic association studies of complex traits, we develop a stable and adaptive test for generalized linear models to accommodate different alternatives. To facilitate valid post-selection inference for high-dimensional data, our study here adheres to the original sample-splitting principle but does so repeatedly to increase stability of the inference. We show the asymptotic null distribution of the proposed test for both fixed and diverging numbers 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 two applications. The proposed procedure is computationally efficient and has been implemented as the R package DoubleCauchy.

Résumé: Nous nous concentrons sur la détection du signal polygénique dans les études d’association génétique à haute dimension pour les traits complexes. Nous proposons un test stable et adaptatif pour les modèles linéaires généralisés afin de gérer différentes alternatives. Pour assurer une inférence valide pour les données à haute dimension, notre étude adopte le principe de division de l’échantillon de manière répétée pour augmenter la stabilité de l’inference. Nous déterminons la loi asymptotique du test proposé sous l’hypothèse nulle pour différents nombres de variants. Nous analysons également les propriétés asymptotiques du test proposé sous des contre-hypothèses locales, ce qui nous permet de comprendre comment la sélection et le poids des variables peuvent compenser la perte d’efficacité due à la division de l’échantillon. Nous confirmons nos conclusions analytiques par des études de simulation intensives et deux applications pratiques. La méthode proposée est efficace au plan calcul et est disponible en tant que paquet R DoubleCauchy.

Additional Supporting Information may be found in the online version of this article at the publisher’s website.

*Corresponding author: sun@utstat.toronto.edu

© 2023 Statistical Society of Canada / Société statistique du Canada
1. INTRODUCTION

Polygenic signal detection can improve the power of genetic association studies of complex traits by aggregating weak signals across a large number of genetic variants that do not, individually, achieve statistical significance. The general concept of set-based testing has been well examined in settings such as gene-based association studies (Derkach, Lawless & Sun, 2014; Lee et al., 2014; Barnett, Mukherjee & Lin, 2017; Zhao & Sun, 2021) or multiple-phenotype analyses (Liu & Lin, 2018), but applications of existing methods to high-dimensional genetic data, when both the number of variants and the sample size are diverging, require additional considerations (Zhong & Chen, 2015; Sur, Chen & Candès, 2017).

For simultaneously testing regression coefficients in high-dimensional, generalized linear models (GLMs), Goeman, Van Houwelingen & Finos (2011) proposed a feasible test statistic for the scenario in which the number of variants is fixed but can be larger than the sample size. Guo & Chen (2016) first investigated the asymptotic properties of the test statistic of Goeman, Van Houwelingen & Finos (2011) for a diverging number of variants, and then proposed a U-statistic for GLMs with unbounded link functions. The $P$-value calculation based on asymptotical normal approximation, however, is not accurate for stringent significance levels, and the test is not adaptive to different alternatives.

Recently, Wu, Xu & Pan (2019) proposed an adaptive method where the test statistic is based on different functions of variant-specific score statistics, with different functions targeting different alternatives. However, accurate inference requires the parametric bootstrap when the covariance matrix of genetic variants is unknown, which can be computationally expensive for large-scale studies or stringent significance levels; see Section 2.3 for detailed illustrations and Table S1 in the Supplementary Material for computational time comparison. Furthermore, it is of interest to understand the power trade-off between using the full sample for testing but without variable selection (Wu, Xu & Pan, 2019) and using a sub-sample for variable selection and the remaining data for testing.

In this article, we focus on polygenic signal detection in high-dimensional GLMs and propose to use sample splitting, repeatedly, for valid and stable post-variable-selection inference. One sample splitting produces two independent sub-samples, of which one is used for variable selection, and the other for valid association testing without the need for correcting for variable-selection bias. This general principle has been used in many study settings, but the inherent instability has been noted, including in variable selection (Meinshausen, Meier & Bühlmann, 2009; Wasserman & Roedern, 2009; Meinshausen & Bühlmann, 2010; Guo et al., 2021), change-point detection (Zou, Wang & Li, 2020), and more recently selective inference (Barber & Candès, 2019; Rinaldo, Wasserman & G’Sell, 2019; Dai et al., 2022). Repeated sample splitting is an intuitive remedy (Meinshausen & Bühlmann, 2010; Rinaldo, Wasserman & G’Sell, 2019; Dai et al., 2022), but it is unclear how to aggregate information across multiple, correlated sample splits to derive a valid and efficient test.

In the context of polygenic signal detection, the first polygenic risk score (PRS) method (Purcell et al., 2009) used the one-time-only sample-splitting strategy. Specifically, the method divides the data into a training sample and a testing sample, and then performs a two-stage analysis. Stage 1 applies a variable selection procedure to the training sample to obtain a set of potentially associated variants and their corresponding weights. Using the independent testing sample, stage 2 first constructs a polygenic risk score for each individual by calculating a weighted sum of the frequencies of the risk allele across the selected variants, and then evaluates the aggregated score for association with the trait of interest. The original polygenic method has since been extended (Vilhjálmsson et al., 2015; Shi et al., 2016; Li et al., 2020), but the stability of the inference after sample splitting has not been well studied.

In this article, we investigate the stability of polygenic association testing by leveraging the strategy of repeated sample splitting. We combine the concepts of repeated sample splitting
and adaptive testing to develop a robust polygenic association test for testing high-dimensional regression coefficients in GLMs. In Section 2, we first review the classical polygenic association test, based on a weighted sum of the numbers of the risk allele across the selected variants, and we note its equivalence to a weighted sum of the variant-specific score statistics. We then consider different weighting factors for the score vector, where the different weights are tailored to different alternatives. To aggregate information across the different weighting factors, we use the recent Cauchy method of Liu & Xie (2020), which has since been used for genetic studies in different settings (Tang et al., 2015; Liu et al., 2019). We also discuss the connection of our adaptive method with that of Wu, Xu & Pan (2019). To improve the stability of our inference, we then introduce the combination procedure that aggregates information across multiple, correlated sample splits. Finally, we derive the asymptotic null distributions of the proposed test for both fixed and diverging numbers of variants, and we study its asymptotic properties under local alternatives. In Section 3 we present extensive simulation results for method evaluation and comparison, including additional simulation studies using the real genetic data from two applications combined with simulated outcome data. In Section 4 we provide the results from the two applications. We conclude with a discussion in Section 5, which includes information for DoubleCauchy, an R package that implements the proposed test.

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 \( j' \), \( i \in \{1, \ldots, n\} \), \( j \in \{1, \ldots, J\} \), and \( j' \in \{1, \ldots, q\} \). 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 conditionally on \((G_i, X_i)\), \( Y_i \) follows a distribution with density function

\[
 f(y_i) = \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\}
\]

for some specific functions \( a(\cdot), b(\cdot), \) and \( c(\cdot) \), where \( \theta_i \) is the canonical parameter, \( \phi \) the dispersion parameter, \( \var(Y_i|G_i, X_i) = a(\phi)v(\mu_i) \), and \( v(\mu_i) \) the variance function. We consider the GLM 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 \( G(\cdot), G(\mu_i) = G^T \beta + X_i^T \beta_x \), where \( \beta \) and \( \beta_x \) 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 \( M^* \) and \( |M^*| \) to denote, respectively, the set and number of those that are truly associated with the response. 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 unit variance.

2.2. The Classical Polygenic Risk Score for High-Dimensional Association Test

Suppose we have \( 2n \) independent observations. The classical polygenic risk score-based high-dimensional association testing method (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, \( M \), where we define \( J_2 = |M| \). For variable selection under GLMs, we considered distance correlation sure independence screening DCSIS (Li, Zhong & Zhu, 2012) and sure independence screening (SIS; Fan & Lv, 2008), because these methods require fewer assumptions than, e.g., ElasticNet (Zou & Hastie, 2005) for the property

DOI: 10.1002/cjs.11768
of sure screening to hold (Bühlmann, Kalisch & Meier, 2014). We also evaluated other methods such as ElasticNet, but the results were similar especially for weak signals.

To test the \( J_2 \) variants simultaneously in the testing sample \( D_{n,2} \), one single polygenic risk score \( G^*_i \) is constructed by aggregating the \( J_2 \) selected variables using \( G_i^{(2)} \) weighted by the effect estimate \( \hat{\beta}_j^{(1)} \) obtained from \( D_{n,1} \), \( j \in M \). That is, \( G^*_i = \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} g_{ij}^{(2)}, i \in \{1, \ldots, n\} \). The inference is then based on the generalized linear regression model applied to \( D_{n,2} \)

\[
G \left\{ E \left( Y_i^{(2)} | G_i^*, X_i^{(2)} \right) \right\} = G_i^* \beta^* + X_i^{(2)T} \beta_x, \tag{1}
\]

and testing

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

The corresponding score statistic is \( T_1 = \sum_{i=1}^{n} \left( y_i^{(2)} - \hat{\mu}_i^{(2)} \right) G_i^* \), where \( \hat{\mu}_i^{(2)} = G^{-1} \left( X_i^{(2)T} \hat{\beta}_x \right) \) and \( \hat{\beta}_x \) is the maximum likelihood estimate of \( \beta_x \) under \( H_0 \). The distribution of the standardized \( T_1 \) can be approximated by \( \chi^2_1 \), 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 (Vilhjálmsdóttir et al., 2015) and better estimation of \( \hat{\beta}_j^{(1)} \) (Shi et al., 2016), among others (Li et al., 2020). However, additional work is needed. To facilitate our discussion, first it is instructive to re-formulate \( T_1 \) as the following:

\[
T_1 = \sum_{i=1}^{n} \left( y_i^{(2)} - \hat{\mu}_i^{(2)} \right) G_i^* = \sum_{i=1}^{n} \left( y_i^{(2)} - \hat{\mu}_i^{(2)} \right) \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} g_{ij}^{(2)}
\]

\[
= \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} \sum_{i=1}^{n} \left( y_i^{(2)} - \hat{\mu}_i^{(2)} \right) g_{ij}^{(2)} = n \sum_{j=1}^{J_2} \hat{\beta}_j^{(1)} S_j,
\]

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

Tests based on \( T_1 \) are sub-optimal when signs of \( \hat{\beta}_j^{(1)} \) and \( S_j \) differ. When the effect size \( \beta_j \) is large, it is likely to obtain sign-consistent results between \( \hat{\beta}_j^{(1)} \) from the training sample and \( S_j \) from the testing sample. This will prevent \( S_j \)'s of variants with opposite direction of effect being cancelled out. However, for weak signals there is no theoretical guarantee for obtaining sign-consistent \( \hat{\beta}_j^{(1)} \) and \( S_j \) (Jin, Zhang & Zhang, 2014), and so it is better to develop a test that is robust to this assumption. Recent work in association tests for rare variants have also shown that \( T_1 \)-type tests are powerful only when a large proportion of the variants being tested are causal, in addition to their genetic effects being in the same direction (Derkach, Lawless & Sun, 2014; Lee et al., 2014). Further, the direct use of \( \hat{\beta}_j^{(1)} \)'s as weights may not be most efficient under different alternatives. Finally, when the signal-to-noise ratio (SNR) is low, as is often the case in practice, the one-time-only sample-splitting approach may be unreliable (Meinshausen, Meier & Bühlmann, 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 data.
2.3. A Robust and 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 model (1) can be reformulated as testing

\[ \begin{align*}
H_0 : \beta = 0 & \text{ versus } H_1 : \beta \neq 0
\end{align*} \]  

in

\[ G \left\{ \mathbb{E} \left( Y^{(2)}_i | G^{(2)}_i, X^{(2)}_i \right) \right\} = G^{(2)T}_i \beta + X^{(2)T}_i \beta \times, \]  

where \( \beta = (\beta_1, \ldots, \beta_{J_2})^T \). Denote \( w_j \) as a function of \( \beta_j \). The proposed new test statistics have the following form:

\[ T_\gamma = n \sum_{j=1}^{J_2} \hat{w}_j^{\gamma-2} S_j^2, \quad \gamma \in \Gamma = \{2, 4, 6, \ldots\}, \]  

where \( \hat{w}_j \) is a consistent estimator of \( w_j \) that depends on \( \hat{\beta}_j^{(1)} \) obtained from \( D_{n,1} \), and \( \gamma \) is an even integer to avoid signal cancellation between variants; see Section S2 in the Supplementary Material for additional justifications on the choice of \( \gamma \).

Define \( S = (S_1, \ldots, S_{J_2})^T \) and \( R = \text{diag}(r_j) = \text{diag}(w_j^{\gamma-2}) \), and let \( \hat{R} = \text{diag}(\hat{r}_j) = \text{diag}(\hat{w}_j^{\gamma-2}) \), \( j \in \{1, \ldots, J_2\} \), be a consistent estimator of \( R \). Hence, we have

\[ T_\gamma = n S^T \hat{R} S. \]  

We can easily modify \( R \) to include off-diagonal elements to reflect potential linkage disequilibrium between genetic variables. If we choose \( w_j = \beta_j \), then \( \hat{R} \) is a consistent estimator of
R = \text{diag}\{r_j\} = \text{diag}\{\beta_j^{w-2}\}, \ j \in \{1, \ldots, J_2\}. However, we note that test accuracy of T_γ does not rely on the consistency assumption, because the data used for estimating R are independent of those used for calculating S. Nevertheless, consistent estimation of R can improve the power of T_γ.

The different γ values in Equation (5) adapt to different signal sparsities. To obtain an accurate yet computationally efficient adaptive test, we propose to aggregate the p_γ’s, the P-values of T_γ, γ ∈ Γ = \{2, 4, 6, \ldots\}, using the Cauchy combination method recently proposed by Liu & Xie (2020). The Cauchy method can accommodate complex dependency structure among P-values without explicitly modelling it. As T_1 can work well if a large proportion of the variants being tested are causal, in addition to their genetic effects being in the same direction (Figure S2 in the Supplementary Material), we include p_1, the P-value of T_1, in the P-value aggregation. Thus, the proposed test statistic 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 - (\text{arctan} \ t_c)/\pi, where t_c is the observed value of T_c.

Here we acknowledge that T_γ is related to the SPU-type test statistics proposed by Wu, Xu & Pan (2019). For an integer γ ≥ 1 SPU(γ) = \sum_j S_{j}^γ, where S_j is obtained from the whole sample. If we omit the sample-splitting step in our approach (i.e., J_2 = J) and let \hat{w}_j = S_j, we have T_γ ∝ SPU(γ) for all γ > 1. The authors of SPU(γ) have noted that for an even integer γ → ∞, SPU(γ) ∝ (\sum_j |S_j|^γ)^{1/γ} → \max_j |S_j|, defined as SPU(∞); this suggests that larger γ is more powerful for sparse alternatives. To make SPU robust to different alternatives, the authors then proposed an adaptive SPU, aSPU = \min_{\gamma \in \Gamma_{aSPU}} \{p_{\text{spu}(\gamma)}\}, where the recommended Γ_{aSPU} = \{1, 2, 3, 4, 5, 6, \infty\}, and p_{\text{spu}(\gamma)} is the P-value of SPU(γ). The asymptotic p_{\text{spu}(\gamma)} can be obtained with mild conditions imposed on the moments of the S_j’s and their correlation structure (Wu, Xu & Pan, 2019), but the asymptotic approximation is not accurate for stringent significance levels (Table S6 in the Supplementary Material). The authors then proposed to calculate p_{\text{spu}(\gamma)}, and subsequently p_{\text{aSPU}}, based on the parametric bootstrap, which is computationally expensive (Table S1 in the Supplementary Material).

The distinction between aSPU and the proposed T_c is four-fold. First, although T_c includes evidence from T_1, the building block of T_1 is T_γ, where γ is an even integer, which facilitates studying the asymptotic properties of the proposed tests; see Theorems 1–4 for details. Second, tests using different γ values are correlated with each other. Thus, the minimum-p approach of aSPU makes the inference more difficult than that of T_c, which is based on the easy-to-implement Cauchy method. We note that although the Cauchy method could also be used to aggregate the individual p_{\text{spu}(\gamma)}’s to obtain aSPU, accurate p_{\text{spu}(\gamma)}’s are difficult to obtain and require the bootstrap. Third, although aSPU uses the whole sample for association testing, it aggregates information across all J genetic variants, many of which may be from the null. In contrast, although the proposed T_c uses a sub-sample for testing, it can benefit from variable selection, which we investigate. Lastly, the flexible structure of \hat{w}_j in T_γ can incorporate other information available for each variant j, such as the functional importance of a genetic variant.

To further increase the robustness of T_c against sampling variation inherent in the one-time-only sample-splitting approach, we then consider m-times-repeated sample splitting. For the s-th 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
The proposed double Cauchy combination test statistic is

\[ T_{dc} = m^{-1} \sum_{s=1}^{m} \tan \left\{ (0.5 - p_{c,s}) \pi \right\}. \]  

(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.

**Remark 1.** To combine dependent \( P \)-values over multiple splits, Meinshausen, Meier & Bühlmann (2009) proposed a quantile-based method of combining \( P \)-values while giving asymptotic control of the family-wise error rate (FWER) and false discovery rate (FDR) for high-dimensional variable selection. Recently, this method has been adopted by Guo et al. (2022) for their conditional global test of ultra-high-dimensional linear regression coefficients under a sparsity condition. A multiple-sample-splitting approach was applied in their paper to get rid of noise variables. However, as mentioned by Meinshausen, Meier & Bühlmann (2009), the quantile of multiple dependent \( P \)-values of sample splitting is an added parameter that needs to be selected. To select a suitable quantile, a data-driven adaptive method was presented. In our study, to avoid the quantile selection, we prefer to use the Cauchy combination method.

### 2.4. Asymptotic Properties of \( T_\gamma \)

To make the dependency of \( T_\gamma \) on \( n \) and \( J_2 \) explicit, we use \( T_{n,J_2,\gamma} \) to denote \( T_\gamma \) 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 notational simplicity, we now omit the 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 variants to be tested.

Recall that \( T_{n,J_2,\gamma} = n \sum_{\gamma} S^T \hat{R} S \), where \( S = (S_1, \ldots, S_{J_2})^T \) is the score vector, \( \hat{R} = \text{diag}\{\hat{r}_j\} \), and \( \gamma \) is an even integer. The covariance matrix of \( n^{1/2} S \) is \( \Sigma_s = E \{ a_i(\phi)v(\mu_s)G_iG_i^T \} \), where \( G_i \in \mathbb{R}^{J_2 \times 1} \) is the genotype vector for individual \( i \), \( e = (e_1, \ldots, e_n)^T = Y - G^{-1}(G\beta + X\beta_x) \), and \( e_0 = (e_{01}, \ldots, e_{0n})^T = Y - G^{-1}(X\beta_x) \).

The following theorem gives the asymptotic null distribution of \( T_{n,J_2,\gamma} \), provided that the same regularity conditions hold as those required for the convergence of \( S \) to a multivariate normal random variable (Goeman, Van Houwelingen & Finos, 2011). In addition, we ignore the nuisance parameters \( a_i(\phi) \) and \( \beta_x \) 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 \( \gamma \), \( T_{n,J_2,\gamma} \sim T_{J_2,\gamma} \) 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, \( \chi^2_{1j} \)'s are independent variables with the central chi-square distribution with one degree of freedom (denoted by \( \chi^2_{1} \)), \( \lambda_{J_2,1} \geq \ldots \geq \lambda_{J_2,J_2} \) are the eigenvalues of \( C_s^T R C_s \), and \( \Sigma_s = C_s C_s^T \).

When \( Y \) is normally distributed, that \( T_{n,J_2,\gamma} \) and \( T_{J_2,\gamma} \) are 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 \( \text{E}(Z_i) = 0 \) and \( \text{cov}(Z_i) = I_{J_2} \). Assume \( z_{ij} \) has finite eighth moment.
and E\(z_{ij}^4\) = 3 + \(\Delta < \infty\), \(\forall j\), where \(\Delta\) is a constant and \(\Delta > -3\), and E\((\Pi_j z_{ij}^\prime) = \Pi_j E\((z_{ij}^\prime)\), where \(\sum_j v_j \leq 8\) and all \(v_j\)'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\((e|G) = 0\) and E\((e^3|G) = 0\), and there are positive constants \(K_1\) and \(K_2\) such that E\((e^2|G) > K_1\) and E\((e^4|G) < K_2\) almost everywhere for \(g \in D(f_g)\).

**Assumption 3.** There exist real numbers \(\rho_{\infty,j}\)'s such that \(\lim_{j \to \infty} \rho_{j,2} = \rho_{\infty,j}\) uniformly \(\forall j\), and \(\lim_{j \to \infty} \sum_{j=1}^{J_2} \rho_{j,2,j} = \sum_{j=1}^{\infty} \rho_{\infty,j} < \infty\), where \(\rho_{j,2,j} = \lambda_{j,2,j}/\sqrt{\text{tr}(R \Sigma z_2^2)}\), \(j \in \{1, \ldots, J_2\}\), which are the eigenvalues of \(C_s^T R C_s / \sqrt{\text{tr}(R \Sigma z_2^2)}\) in descending order.

**Assumption 4.** \(n\{\text{tr}(R \Sigma_g)^2 / \text{tr}(R \Sigma_g)^2\} \to \infty\) as \(n\) and \(J_2 \to \infty\). Assumptions 1–3 are standard in studying high-dimensional testing (Guo & Chen, 2016; Zhang et al., 2019). Assumption 4 specifies a relationship between \(\lambda_{j,2,j}\)’s similar to Condition C1 in Li & Li (2021), and in general such an assumption can be fulfilled if all \(\lambda_{j,2,j}\)'s are similar in magnitude, corresponding to less correlation. Assumption 4 is equivalent to requiring that the sample size \(n\) grows to infinity at a rate faster than \(J_2\). Assumption 4 is similar to Condition C1 in Li & Li (2021), and in general such an assumption can be fulfilled much more easily by higher correlation in \(R \Sigma_g\).

The following theorem generalizes Theorem 1 from finite to infinite \(J_2\).

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

\[
\hat{\sigma}_{n,0}^{-1} \{T_{n,J_2,y} - \text{tr}(\hat{R} \Sigma_z)\} \Rightarrow \zeta \quad \text{and} \quad \{2\text{tr}(R \Sigma_z)^2\}^{-1/2} \{T_{J_2,y} - \text{tr}(R \Sigma_z)\} \Rightarrow \zeta
\]

as \(n\) and \(J_2 \to \infty\), where \(\zeta\) and \(\sum_{j=1}^{\infty} \rho_{\infty,j}(\chi_{1j}^2 - 1)/\sqrt{2}\) are equivalent in distribution, \(\sigma_{n,0}^2 = 2\text{tr}(R \Sigma_z)^2(1 + o(1))\), and \(\hat{\sigma}_{n,0}^2 = 2\text{tr}(\hat{R} \Sigma_z)^2 (1 + o_p(1))\). Therefore, as \(n\) and \(J_2 \to \infty\)

\[
sup_x |\text{pr}(T_{n,J_2,y} \leq x) - \text{pr}(T_{J_2,y} \leq x)| \to 0.
\]

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 \(T_{n,J_2,y}\) for both fixed and diverging \(J_2\). The corresponding \(P\)-value can be calculated using the method of Davies (1980). To show the asymptotic normality of \(T_{n,J_2,y}\) under the null, we need to impose the following assumption, which substitutes for specifying an explicit relationship between \(J_2\) and \(n\).

**Assumption 5.** \(\text{tr}(R \Sigma_g)^2 / \text{tr}(R \Sigma_g)^2 \to \infty\) and \(\text{tr}(R \Sigma_g)^2 \to \infty\) as \(n\) and \(J_2 \to \infty\).

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

\[
\hat{\sigma}_{n,0}^{-1} \{T_{n,J_2,y} - \text{tr}(\hat{R} \Sigma_z)\} \Rightarrow N(0, 1),
\]

as \(n\) and \(J_2 \to \infty\).
Theorem 4. The estimation of $R$ and $\Sigma_s$ will influence the approximations in Theorems 2 and 3. In Theorem 2, we used $\sum_{j=1}^{J_2} \lambda_{j,j} \chi_{1j}^2$ to approximate the asymptotic null distribution of $T_{n,J_2,J_2}$ for diverging $J_2$. Therefore, good estimates of the eigenvalues are required for accurate approximation. Based on Weyl’s theorem (Mohsen, 2013), we have $|\hat{\lambda}_{J_2,J_2} - \lambda_{J_2,J_2}| \leq \| \hat{\hat{\Sigma}}_s - R \Sigma_s \|$, where $\| \cdot \|$ is the operator or spectral norm of a matrix. Thus, a consistent estimation of $R \Sigma_s$ in operator norm can guarantee consistent estimation of eigenvalues. In Theorem 3, $\hat{R}$ and $\hat{\Sigma}_s$ regulate the normal approximation through the mean and variance of $T_{n,J_2,J_2}$. Because $\text{tr}(R \Sigma_s^2) = \sum_{j=1}^{J_2} \lambda_{j,j}^2$ and $\text{tr}(R \Sigma_s) = \sum_{j=1}^{J_2} \lambda_{j,j}$, we can obtain consistent estimators of $\text{tr}(R \Sigma_s^2)$ and $\sigma_{n,0}$, provided that $\hat{\hat{\Sigma}}_s$ is a consistent estimator of $R \Sigma_s$ in operator norm.

We now study the interplay between the adverse effect of reduced sample size on power and the beneficial effect of variable selection afforded by sample-splitting, under the local alternative $\mathcal{L}_0$:

$$\mathcal{L}_0 = \left\{ \Delta^T R \Sigma_s R \Delta = o \left\{ n^{-1} \text{tr}(R \Sigma_s^2) \right\} \text{ and } \left\{ G^{-1}(G^T \beta) \right\}_j^2 = O(1) \right\},$$

where $\Delta_0 = E \left\{ (G^{-1}(G^T \beta))^2 \right\}$. The SNR corresponding to $\mathcal{L}_0$ can be interpreted as the SNR following Guo & Chen (2016). As detailed in the Supplementary Material

$$\mu_{n,0} = \sum_{j=1}^{J_2} r_j E \left\{ G^{-1}(G^T \beta) g_{ij} \right\}^2 + (n - 1) \sum_{j=1}^{J_2} r_j E \left\{ G^{-1}(G^T \beta) g_{ij} \right\}$$

and $\sigma_{n,1}^2 = \left\{ \sigma_{n,0}^2 + 2 \text{tr}(R \Sigma_s) + 4 \text{tr}(R \Sigma_s R \Sigma_s) \right\} \{ 1 + o(1) \}$, where

$$\sigma_{n,0}^2 = 2 \sum_{j=1}^{J_2} \sum_{k=1}^{J_2} r_j r_k E \left[ g_{ij} g_{ik} e_i^2 \right] \{ 1 + o(1) \},$$

and

$$\text{tr}(R \Sigma_s) = \sum_{j=1}^{J_2} \sum_{k=1}^{J_2} r_j r_k E \left[ g_{ij} g_{ik} e_i^2 \right] \left\{ G^{-1}(G^T \beta) \right\}_j^2.$$
\[\sigma^2_{2n,0} = 2\sum_{j=1}^J \sum_{k=1}^J E^2\left(g_{ij}g_{ik}\epsilon_i^2\right) \{1 + o(1)\},\]

\[\text{tr}(\Xi_{\beta,j})^2 = \sum_{j=1}^J \sum_{k=1}^J E^2\left[g_{ij}g_{ik}\{G^{-1}(G_j^T\beta)\}^2\right],\]

and \(\text{tr}(\Sigma_x\Xi_{\beta,j}) = \sum_{j=1}^J \sum_{k=1}^J E\left\{g_{ij}g_{ik}\epsilon^2\right\}E\left[g_{ij}g_{ik}\{G^{-1}(G_j^T\beta)\}^2\right]\).

To provide additional insights on power comparison, assume \(pr(M \supset M') \to 1\) as \(n \to \infty\); this assumption can be fulfilled by the existing variable selection algorithms (Fan & Lv, 2008; Li, Zhong & Zhu, 2012). Comparing \(\mu_{n,\beta}\) with \(\mu_{2n,\beta}\), it is clear that sample size reduction is the primary cause of power loss for a sample-splitting-based method. However, the expressions for \(\sigma^2_{n,1}\) and \(\sigma^2_{2n,1}\) show that the first two terms are non-negative, and each term is a summation over \(J_2\) and \(J\) variants, respectively, for \(\sigma^2_{n,1}\) and \(\sigma^2_{2n,1}\). Because \(J_2 \leq J\), noise filtering in the training sample \(D_{n,1}\) can reduce the variance of the test statistic calculated in the testing sample \(D_{n,2}\). Because SNR is the ratio of \(\mu\) to \(\sigma\), and SNR_{p,\beta} can be larger than SNR_{2p,\beta}, tests based on \(T_{n,J_2}\) can be more powerful than \(T_{2n,J}\). The use of weights derived from \(D_{n,1}\) can further compensate for the loss of efficiency due to reduced sample size in \(D_{n,2}\). Simulation studies in the next section show that the proposed method has significant power gain over existing methods when the sure screening property holds (Figures 2 and 3). Even if sure screening fails in \(D_{n,1}\), the sample-splitting approach can have comparable power with the methods of Wu, Xu & Pan (2019) and Guo & Chen (2016), applied to the full sample without variable selection (Figure 2).

3. SIMULATION STUDIES

3.1. Simulation designs

To evaluate the performance of \(T_{dc}\) and compare it with tests proposed by Guo & Chen (2016) and Wu, Xu & Pan (2019), we consider two simulation designs. Design 1 simulates \(G\), while design 2 builds upon real genetic data.

- **Design 1** considers a 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 (autoregressive) correlation matrix \(\Sigma_g = \{\rho|i-j|\}_{J \times J}\), where \(\rho = 0.2, 0.5, 0.8\), and \(i\) and \(j \in \{1, \ldots, J\}\).

- **Design 2** uses \(G\) from two genomic applications. The first dataset (cystic fibrosis data) consists of \(2n = 1409\) individuals and a set of \(J = 3754\) SNPs (single nucleotide polymorphisms). To test the performance of the proposed method in another scenario, we consider the gene expression data (riboflavin data), including \(2n = 71\) independent individuals and \(J = 4088\) gene expression levels. For a more streamlined presentation, we provide additional information on the two datasets in Section 4, along with the application results.

To implement \(T_{dc}\), we let \(r_j = \beta_j^{-2}\) (\(j \in \{1, \ldots, J_2\}\)) and \(\Gamma = \{2, 4, 6, 42\}\) to first obtain \(T_c\) of (7). For a fair method comparison, we choose \(\Gamma = \{2, 4, 6, 42\}\) to be aligned with \(\Gamma_{\text{aSPU}} = \{1, 2, 3, 4, 5, 6, \infty\}\), studied and recommended by the authors of the aSPU test (Wu, Xu & Pan, 2019); 42 in \(\Gamma\) is to mimic \(\infty\) in \(\Gamma_{\text{aSPU}}\). We then use \(m = 10, 50,\) or 100 to derive the more stable \(T_{dc}\) of (8), and also to study the effect of \(m\) on the performance of \(T_{dc}\) (Figures S4 and S8 in the Supplementary Material). We use \(n_1\) and \(n_2\) to denote the sample sizes of the training and testing data, respectively; when the sample split is even, we use \(n\) for both sub-samples.

For completeness, we also study the performance of the individual \(T_1\) and \(T_\gamma\)’s (\(\gamma = \{2, 4, 6, 42\}\)), but we present the corresponding results in the Supplementary Material.
FIGURE 2: Power comparison of the proposed test $T_{dc}$ (red triangle), the method of Guo & Chen (2016) (blue square), and the method of Wu, Xu & Pan (2019) (black circle), for the three study scenarios (I), (II), and (III).

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

For type I error evaluation, we consider two simulation settings. First, we generate $Y$ based on a logistic regression with $\beta = 0$, and without loss of generality, the intercept is equal to 1 and there are no other covariates. $G$ is generated based on design 1 as detailed above. Because $Y$ is simulated under the null hypothesis for type I error evaluation, the variable selection procedure and the value of $J$ are not critical. Thus, we choose $J_2 = J$, regress simulated $Y$ on each of the $J_2$ simulated variants in $D_{n,1}$, and obtain the corresponding $\hat{\beta}_j$. Based on even sample splitting ($n_1 = n_2 = n$), we then perform the high-dimensional polygenic association testing in $D_{n,2}$.

In the second setting, using the $G$ from the two application datasets, we re-evaluate the accuracy of $T_{dc}$ by simulating $Y$ independently of the observed $G$ for the $n_2$ individuals in $D_{n_2,2}$, where $y_i = 1 + \varepsilon_i$, $i \in \{1, \ldots, n_2\}$, and $\varepsilon_i$ follows the standard normal distribution. We consider three different sample-splitting proportions for the cystic fibrosis data, where $n_1:n_2 = 409:1000$. 

DOI: 10.1002/cjs.11768 The Canadian Journal of Statistics / La revue canadienne de statistique
Figure 3: Power comparison of the proposed test $T_{dc}$ (red triangle), the method of Guo & Chen (2016) (blue square), and the method of Wu, Xu & Pan (2019) (black circle) based on the real gene expression data of the riboflavin dataset.

$n_1:n_2 = 705:704$, and $n_1:n_2 = 1000:409$. Considering the total sample size of $2n = 71$ for the riboflavin data, the implementation of $T_{dc}$ in this study only used the 50%–50% sample-splitting proportion, where $n_1:n_2 = 35:36$. For variable selection and weight estimation, simulating $Y$ for the $n_1$ individuals in $D_{n1,1}$ is an obvious approach. However, to see how potentially non-random variable selection and weight estimation adversely affect the type I error control of the proposed method, we used the real data, both $Y$ and $G$, of the $D_{n1,1}$ training sample; note that $Y$ in the $D_{n2,2}$ testing sample is simulated. Choose $J_2 = n_2$ for variable selection.

In both simulation settings, the distributions of $T_{\gamma}$’s are approximated by $\sum_{j=1}^{J_2} \lambda J_2 \chi^2_1$ as specified in Theorem 2, and the distributions of $T_c$ and $T_{dc}$ by the standard Cauchy distribution. For setting 1, Table 1 shows the empirical test sizes of the $T_{\gamma}$’s, $T_c$, and $T_{dc}$ for $2n = 200$, $m = 10, 50, 100$, $\rho = 0.5$, and nominal $\alpha$ values of $0.05, 0.01, 10^{-3}$, and $10^{-4}$, and Tables S2 and S3 in the Supplementary Material show results for $\rho = 0.2$ and 0.8, respectively; results for $2n = 1500$ are more accurate and thus not shown.

Table 1 shows that the empirical type I error rate of $T_{dc}$ is controlled at or below the nominal $\alpha$ level when $m = 10$ or 50, considering Monte Carlo error. However, the empirical type I error rate is slightly inflated for larger $J_2$ and stringent $\alpha$ level ($\alpha = 10^{-4}$) when $m = 100$. To better understand this inflation problem, we provide the summary statistics of the empirical $\alpha$ from the $m = 100$ sample splits in Table S4 in the Supplementary Material when $J_2 = 400$ and 1000. Results show that the test size of $T_c$ is accurate and stable across the 100 sample splits. Thus, the inflation stems from the double Cauchy combination step.

The accuracy of the Cauchy approximation for large $m$ has been studied in Theorem 2 of Liu & Xie (2020). They showed that, to obtain accurate P-value approximation, $m$ should be bounded by $(t_\alpha)^3 \rho$, where $t_\alpha$ is the upper $\alpha$-quantile of the standard Cauchy distribution and $0 < c_0 < 1/2$. When $\alpha = 10^{-4}$, $t_\alpha = 3183$ and thus $t_\alpha^{1/2} = 56.4$ provides an upper bound on $m$. Although this theoretical result holds only under certain conditions (on the correlation matrix of the P-values to be combined) and is not accurate across all scenarios (Liu & Xie, 2020), it helps in explaining the approximation error when $m = 100$.

For setting 2, Table S5 in the Supplementary Material shows that the empirical $\alpha$ level of $T_{dc}$ remains well controlled when $m = 10$ or 50, but it is slightly inflated when $m = 100$, which is consistent with the results based on simulated multivariate normal predictors. In general, a smaller $m$ leads to better type I error control, but a larger $m$ provides more inference stability. Thus, we will use $m = 50$ in our real data analyses.

We used the $\sum_{j=1}^{J_2} \lambda J_2 \chi^2_1$ approximation to obtain P-values in our simulation studies (and in the applications below) because the normal approximation given in Theorem 3 is not adequate for stringent $\alpha$ levels when $\gamma > 2$ (Table S6 in the Supplementary Material). Consistent with previous reports, the tests of Wu, Xu & Pan (2019) and Guo & Chen (2016) based on asymptotic approximations are not accurate (Table S6 in the Supplementary Material). Thus, we use the
TABLE 1: Empirical test sizes for seven test statistics with sample size \(2n = 200\), and autoregressive model \(AR(1, \rho)\) with \(\rho = 0.5\) for correlation between the \(J\) variants. One-time 50%–50% sample splitting for the first six methods, \(m = 10, 50, 100\) times repeated sample splitting for the proposed \(T_{dc}\).

| \(J\) | \(n\) | \(\alpha\) | \(T_1\) | \(T_2\) | \(T_4\) | \(T_6\) | \(T_{dc,10}\) | \(T_{dc,50}\) | \(T_{dc,100}\) |
|------|------|-------|------|------|------|------|---------|---------|---------|
| 10   | 5%   | 4.9975| 4.7951| 4.8761| 4.9097| 4.9130| 5.1774  | 5.2651  | 5.2281  | 4.8201  |
|      | 1%   | 0.9640| 0.8771| 0.8927| 0.8999| 0.9250| 0.9510  | 0.9614  | 1.0518  | 1.0082  |
|      | 0.1% | 0.0873| 0.0768| 0.0695| 0.0700| 0.0746| 0.0789  | 0.0951  | 0.0854  | 0.0730  |
|      | 0.01%| 0.0078| 0.0063| 0.0052| 0.0051| 0.0046| 0.0088  | 0.0081  | 0.0092  | 0.0096  |
| 50   | 5%   | 4.9503| 4.7245| 4.7715| 4.8447| 4.9615| 5.1745  | 5.3395  | 5.3038  | 5.2115  |
|      | 1%   | 0.9701| 0.8750| 0.8912| 0.9110| 0.9339| 0.9449  | 0.9360  | 0.9280  | 0.9033  |
|      | 0.1% | 0.0961| 0.0772| 0.0769| 0.0789| 0.0767| 0.0825  | 0.0698  | 0.0604  | 0.0598  |
|      | 0.01%| 0.0087| 0.0064| 0.0059| 0.0049| 0.0056| 0.0062  | 0.0041  | 0.0067  | 0.0110  |
| 200  | 5%   | 4.9820| 4.6781| 4.7274| 4.8011| 4.9341| 5.1217  | 5.3721  | 5.4933  | 5.5147  |
|      | 1%   | 0.9921| 0.8743| 0.8951| 0.8961| 0.9172| 0.9517  | 0.9408  | 0.9416  | 0.9374  |
|      | 0.1% | 0.0960| 0.0814| 0.0780| 0.0775| 0.0775| 0.0815  | 0.0705  | 0.0656  | 0.0679  |
|      | 0.01%| 0.0116| 0.0070| 0.0067| 0.0064| 0.0060| 0.0069  | 0.0042  | 0.0156  | 0.0263  |
| 400  | 5%   | 4.9763| 4.6932| 4.7571| 4.8307| 4.9550| 5.1717  | 5.3691  | 5.5359  | 5.6548  |
|      | 1%   | 0.9848| 0.8886| 0.8991| 0.9161| 0.9230| 0.9518  | 0.9375  | 0.9527  | 0.9679  |
|      | 0.1% | 0.0972| 0.0786| 0.0830| 0.0790| 0.0767| 0.0824  | 0.0724  | 0.0624  | 0.0740  |
|      | 0.01%| 0.0102| 0.0061| 0.0070| 0.0053| 0.0049| 0.0067  | 0.0053  | 0.0150  | 0.0245  |
| 1000 | 5%   | 4.9887| 4.6689| 4.7323| 4.7755| 4.9269| 5.1299  | 5.4092  | 5.7421  | 5.1065  |
|      | 1%   | 1.0117| 0.8923| 0.8829| 0.8865| 0.9227| 0.9312  | 0.9405  | 0.9734  | 0.8623  |
|      | 0.1% | 0.0967| 0.0754| 0.0808| 0.0755| 0.0798| 0.0821  | 0.0763  | 0.0604  | 0.0564  |
|      | 0.01%| 0.0100| 0.0073| 0.0082| 0.0060| 0.0059| 0.0074  | 0.0047  | 0.0087  | 0.0173  |

parametric bootstrap, as recommended by the authors, with \(10^3\) replicates to evaluate the power of these methods for a fair comparison.

3.3. Power Comparison Based on Design 1

Similar to the type I error evaluation above, here we consider two simulation designs. For design 1, we generate \(Y\) based on logistic models with different proportions of nonzero regression coefficients, varying from 0.1%, 1%, 5%, to 10% of the \(J\) variants. We assume the indices of the nonzero \(\beta_j\)'s to be uniformly distributed in \(\{1, \ldots, J\}\). We consider three different scenarios of the signs of the nonzero \(\beta_j\)'s: randomly assign half of them to be positive and half to be negative, assign all of them to be positive, or assign all of them to be negative.

Results below focus on power comparison between the proposed \(T_{dc}\) test and the methods of Guo & Chen (2016) and Wu, Xu & Pan (2019), which are applied to the whole sample without variable selection. Results of the original polygenic risk score test, \(T_1\), are shown in Figure S2 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 leveraged to construct the \(T_{dc}\) test statistic using the testing sample \(D_{n,2}\).

DOI: 10.1002/cjs.11768 The Canadian Journal of Statistics / La revue canadienne de statistique
(I) Oracle: \( M = M^* \). 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 scenario is used to show the power gain of \( T_{dc} \), despite the reduction of sample size, as compared to the methods without variable selection.

(II) \( J_2 = J: M = \{g_1, \ldots, g_J\} \). 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 informative. 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 \( D_{n, 1} \) for associate testing using \( D_{n, 2} \).

(III) Variable Selection: \( M \) is estimated after variable screening. This study investigates the impact of variable selection accuracy on the power of \( T_{dc} \) as compared to the methods without variable selection.

For variable selection, we use DCSIS (Li, Zhong & Zhu, 2012) and SIS (Fan & Lv 2008). The implementation of DCSIS and SIS also requires the specification of \( J_2 \). In our power simulation study, \( 2n = 1500 \) and \( J = 4000 \), and we choose \( J_2 = 2n \), which is more conservative than the value \( 2n/\log(2n) \) recommended by the authors.

Because all three methods—the proposed \( T_{dc} \) test and the methods of Guo & Chen (2016) and Wu, Xu & Pan (2019)—incorporate \( S_j^2 \)'s across variants, we expect them to be robust to the signs of \( \beta_j \)'s. This is confirmed by Figure S3 in the Supplementary Material, which also shows that the results are qualitatively similar between the two variable selection methods, DCSIS and SIS. Thus, below we only present the simulation results when the nonzero \( \beta_j \)'s are half positive and half negative, and DCSIS is the variable selection method. Figure 2 shows the results for \( 2n = 1500 \) and \( J = 4000 \), reflecting the values observed in the cystic fibrosis dataset studied in Section 4: \( \rho = 0.5 \) for correlation between the \( J \) variants, \( J_2 = 1500 \) for variable selection, and \( m = 10 \) for repeated 50%–50% sample splitting to construct \( T_{dc} \) (see Figure S4 in the Supplementary Material for \( m = 50 \)).

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

For scenario (II), the second column of Figure 2 shows that the anticipated power loss of \( T_{dc} \) due to sample splitting can be compensated for by leveraging the weights inferred from \( D_{n, 1} \), as compared with the methods of Guo & Chen (2016) and Wu, Xu & Pan (2019), which use the full sample; recall that \( J_2 = J \) for \( T_{dc} \), meaning the variable selection step completely failed at selecting relevant variants. For the sparse alternative case, scenario (II) 4 signals, \( T_{dc} \) displays comparable power with the method of Wu, Xu & Pan (2019), while both are substantially more powerful than the method of Guo & Chen (2016). For the other alternatives considered in this scenario, all three methods have comparable power, with the method of Guo & Chen (2016) having slightly higher power. Overall, the proposed \( T_{dc} \) test is most robust to the different alternatives considered here, and it is also computationally efficient, which we discuss in the Supplementary Material.

For 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 sure screening property requires that the nonzero regression coefficients must be sufficiently large (Fan & Lv, 2008; Bühlmann, Kalisch & Meier, 2014). In addition to DCSIS and SIS, we also evaluated other selection methods such as ElasticNet, but the results were similar, especially for weak signals. However, simulation results in Section 3.4, based on real \( G \) from the riboflavin data, show that the proposed \( T_{dc} \) method gains substantial power after variable selection.
Simulation results so far have focused on the 50%–50% sample-splitting proportion. We have also investigated 33%–67% (Figure S5 in the Supplementary Material) and 67%–33% (Figure S6 in the Supplementary Material) sample splitting. The results in Figures S5 and S6 show that the overall power of $T_{dc}$ is not very sensitive to the splitting proportion. However, the 33%–67% sample splitting has slightly increased power for the scenarios considered here. This is consistent with the literature (Barber & Candès, 2019), where it has been noted that uneven sample splitting, with more subjects assigned to the testing sample, can increase power as compared to even sample splitting.

3.4. Power Comparison Based on Design 2

For the more realistic design 2, we conduct a power simulation study based on real gene expression levels ($G$) of the 4088 genes in the riboflavin data. To simulate $Y$, we consider the number of nonzero regression coefficients to be 3, 40, or 200. To best mimic the presumed underlying signal structure (Shi et al., 2020), we assume $\beta_j, j = 1588, 3154, and 4004$ to be nonzero and all the $\beta_j$’s are positive with values ranging from 0.2 to 0.6 (Figure 3). For the scenarios where there are 40 and 200 signals in total, we further assume the first 37 and 197 $\beta_j$’s are nonzero, respectively; these $\beta_j$’s are generated from a normal distribution with mean 0 and variance 0.05, corresponding to weak signals. To implement $T_{dc}$, we use the same $n_1, n_2, and J_2$ values as those for the type I error evaluation above.

Results in Figure 3 show that the method of Guo & Chen (2016) (blue square) performs poorly when there are sparse strong signals (left plot) or sparse strong signals combined with some weak signals (middle plot). In either case, the power of the proposed $T_{dc}$ test is appreciably higher than that of the method of Wu, Xu & Pan (2019) (black circle). To be consistent with the power study shown in Figure 2, we note that the power in Figure 3 is for $m = 10$. Figure S8 in the Supplementary Material shows that the power of $T_{dc}$ can be slightly improved by using $m = 50$; this is consistent with the results in Figure S4 in the Supplementary Material.

4. Application

4.1. Cystic Fibrosis Data

We apply the proposed $T_{dc}$ test and the methods of Guo & Chen (2016) and Wu, Xu & Pan (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 function has 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. These are candidates for association with lung function in cystic fibrosis but selected in an unsupervised manner based on biological hypothesis alone (Sun et al., 2012).

To implement the proposed $T_{dc}$ test, as in the simulation studies, we define $r_j = \hat{\beta}_j^{-2}, j \in \{1, \ldots, J_2\}, and \Gamma = \{2, 4, 6, 42\}$, and we apply the variable selection method DCSIS (Li, Zhong & Zhu, 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 $\Sigma_g$ in $D_{n_2}$, we use the algorithm proposed by Rothman (2012), with the tuning parameter selected by five-fold cross-validation.

In the absence of oracle knowledge of the true association, the application results here focus on discussing the range of $P$-values of all methods. The empirical $P$-values are 0.0985 and 0.0727, respectively, for the methods of Guo & Chen (2016) and Wu, Xu & Pan (2019), based on $10^4$ bootstrap samples applied to the whole sample. For $T_1$, we randomly split the whole sample into 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 $n_1:n_2 = 409:1000$ is shown in Figure 1.
Table 2: Summary of $P$-values of the proposed $T_{dc}$ applied to two real datasets based on different sample splits, with $m = 50$ for constructing $T_{dc}$. $T_{dc}$ with $m = 50$ is constructed repeatedly 100 times. We choose $J_2 = n_2$, the sample size of the testing sample for all scenarios.

|                | $n_1$ | $n_2$ | Minimum | First quartile | Median | Mean | Third quartile | Maximum |
|----------------|-------|-------|---------|---------------|--------|------|---------------|---------|
| Cystic fibrosis| 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   |
| Riboflavin     | 35    | 36    | $5.551 \times 10^{-17}$ | $1.665 \times 10^{-16}$ | $1.665 \times 10^{-16}$ | $2.520 \times 10^{-16}$ | $2.359 \times 10^{-16}$ | $4.496 \times 10^{-15}$ |

For $T_{dc}$, we also randomly split the whole sample into two subsets, but independently $100 \times 50$ 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 $n_1:n_2 = 409:1000$ is shown in Figure 1 in blue. The results clearly show that the proposed repeated sample-splitting strategy leads to a much more stable inference than the one-time-only sample-splitting approach: $p_1$ ranges from 0.0019 to 0.9446, while $p_{dc}$ ranges from 0.003 to 0.101 with a mode of around 0.05. For completeness, we also provide the summary statistics of the 100 $p_{dc}$'s in Table 2.

To study $T_{dc}$'s robustness to different proportions of sample splitting, we consider 33%–67%, 50%–50%, and 67%–33%. However, a practical question arises: what should be the reported $P$-value for this application? This could be randomly drawn from the set of $P$-values, but we recommend the median of the 100 $P$-values from the 50%–50% sample split for a conservative estimate; 33%–67% can increase power as compared to 50%–50%, assuming sufficient total sample size (Barber & Candès, 2019). Overall, considerable variation remains, suggesting that the signals are too weak given the total sample size.

4.2. Riboflavin Data

In this application, the outcome of interest ($Y$) is the standardized riboflavin (B2) production rate, measured on $2n = 71$ independent individuals, and the predictors ($G$) are standardized gene expression levels of $J = 4088$ genes. The dataset is freely available in the R package hdi (Bühlmann, Kalisch & Meier, 2014), and it has been used for studying variable selection and for constructing valid confidence intervals after model selection (Shi et al., 2020). We use a 50%–50% sample-splitting proportion, where $n_1:n_2 = 35:36$ and $J_2 = 36$ for variable selection using DCSIS, and then apply the three methods to the data for method comparison. The empirical $P$-values are 0.028 and $5 \times 10^{-5}$, respectively, for the methods of Guo & Chen (2016) and Wu, Xu & Pan (2019) based on 105 bootstrap samples. In contrast, the $P$-value of $T_{dc}$ is in the range $10^{-16}$ based on $m = 50$.

To show the stability of our inference, similar to Section 4.1, we also perform sample splitting independently 100 $\times 50$ times to obtain 100 $p_{dc}$'s. The summary statistics of the 100 $p_{dc}$'s are shown in Table 2. The maximum $p_{dc}$ is $4.496 \times 10^{-15}$, suggesting better performance of $T_{dc}$ as compared to the other two methods. For completeness, we also compare the performance between $T_{dc}$ and $T_1$. The results in Figure S7 in the Supplementary Material show that the proposed repeated sample splitting not only provides robustness but also improves power; the range of $p_1$ is $[7.8 \times 10^{-8}, 0.027]$ as compared with $[5.551 \times 10^{-17}, 4.496 \times 10^{-15}]$ of $p_{dc}$.

5. DISCUSSION

In this theoretical study, we did not consider the impact of estimating nuisance parameters $\beta_x$ and $\phi$, 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 & 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. The proposed $T_{dc}$ is computationally efficient (Section S1 in the Supplementary Material). The method has been implemented as an R package, DoubleCauchy, available at https://github.com/yanyan-zhao/DoubleCauchy.

In terms of the stability of the proposed $T_{dc}$ method, indeed, the randomness induced by (non-exhaustive) sample splitting may influence our inference. To remove randomness of repeated but non-exhaustive sample splitting, Wasserman, Ramdas & Balakrishnan (2020) conceptually considered all possible sample splits, but they also noted that this approach is computationally infeasible.

As part of our future work, we plan to leverage the findings of Liu, Yu & Li (2022): the $P$-values of their projection test, constructed from a multiple data-splitting procedure, are exchangeable under the setting of a high-dimensional one-sample mean test. It is then of future interest to investigate the exchangeability of the $P$-values of $\{T_{c,1}, \ldots, T_{c,m}\}$ to study the theoretical properties of $T_{dc}$.

ACKNOWLEDGEMENTS

We sincerely thank the editor, associate editor, and referee for their helpful and constructive comments. We thank Dr. Lisa J. Strug and her lab for providing the cystic fibrosis data for application and helpful discussion. This research was funded by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, and the University of Toronto McLaughlin Centre Accelerator Grants in Genomic Medicine. Zhao’s research was supported by the National Natural Science Foundation of China (12201351).

REFERENCES

Barber, R. & Candès, E. (2019). A knockoff filter for high-dimensional selective inference. The Annals of Statistics, 47, 2504–2537.

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.

Bühlmann, K., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application, 1, 255–278.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21, C1–C68.

Dai, G., Lin, B., Xing, X., & Liu, J. (2022). False discovery rate control via data splitting. Journal of the American Statistical Association, https://doi.org/10.1080/01621459.2022.2060113

Davies, R. (1980). Algorithm AS 155: The distribution of a linear combination of $\chi^2$ random variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29, 323–333.

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.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh-dimensional feature space. Journal of the Royal Statistical Society. Series B (Methodological), 70, 849–911.

Goeman, J. J., Van Houwelingen, H. C., & Finos, L. (2011). Testing against a high-dimensional alternative in the generalized linear model: Asymptotic type I error control. Biometrika, 98, 381–390.

Guo, B. & Chen, S. (2016). Tests for high-dimensional generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological), 78, 1079–1102.

Guo, W., Zhong, W., Duan, S., & Cui, H. (2022). Conditional test for ultrahigh-dimensional linear regression coefficients. Statistica Sinica, 32, 1–29.
Guo, X., Ren, H. J., Zou, C. L., & Li, R. Z. (2021). Threshold selection in feature screening for error rate control. *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2022.2063130

Jin, J., Zhang, C., & Zhang, Q. (2014). Optimality of graphlet screening in high-dimensional variable selection. *Journal of Machine Learning Research*, 15, 2723–2772.

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.

Li, C. C. & Li, R. Z. (2021). Linear hypothesis testing in linear models with high-dimensional responses. *Journal of the American Statistical Association*, 117(540), 1738–1750.

Li, R., Chen, Y., Ritchie, M., & Moore, J. (2020). Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics*, 21, 493–502.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107, 1129–1139.

Liu, W., Yu, X., & Li, R. (2022). Multiple-splitting project test for high-dimensional mean vectors. *Journal of Machine Learning Research*, 23(71), 1–27.

Liu, Y., Chen, S., Li, Z., Morrison, A., Boerwinkle, E., & Lin, X. (2019). ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104, 410–421.

Liu, Y. W. & Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115, 393–402.

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

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, 72, 417–473.

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104, 1671–1681.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*, New York, Wiley.

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.

Rinaldo, A., Wasserman, L., & G’Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47, 3438–3469.

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

Shi, C., Song, R., Lu, W., & Li, R. (2020). Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116, 1307–1318.

Shi, J., Park, J. D., Duan, J., Berndt, S. T., Moy, W., Yu, K., Song, L. et al. (2016). Winners curse correction and variable thresholding improve performance of polygenic risk modelling based on genome-wide association study summary-level data. *PLOS Genetics*, 12, e1006493.

Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boelle, P., Durie, P. et al. (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.

Sun, L., Rommens, J., Corvol, H., Li, W., Li, X., Chiang, T. A., Lin, F. et al. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature Genetics*, 44, 562–569.

Sur, P., Chen, Y., & Candès, E. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175, 487–558.

Tang, Y., Zhou, Y., Chen, L., Bao, Y., & Zhang, R. (2015). A powerful adaptive Cauchy variable combination method for rare-variant association analysis. *Russian Journal of Mathematical Physics*, 57, 238–245.

Vilhjálmsson, B. J., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., Genovese, G. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97, 576–592.
Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117, 16880–16890.

Wasserman, L. & Roedern, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, 37, 2178–2201.

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

Zhang, J., Guo, J., Zhou, B., & Cheng, M. Y. (2019). A simple two-sample test in high dimensions based on $L^2$-norm. *Journal of the American Statistical Association*, 115, 1011–1027.

Zhao, Y. & Sun, L. (2021). On set-based association tests: Insights from a regression using summary statistics. *The Canadian Journal of Statistics*, 49, 754–770.

Zhong, P. & Chen, S. (2015). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, 106, 260–274.

Zou, C. L., Wang, G. H., & Li, R. Z. (2020). Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics*, 48, 413–439.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67, 301–320.

---

Received 30 December 2021
Accepted 26 October 2022

DOI: 10.1002/cjs.11768