Is Polygenic Risk Scores Prediction Good?

Bingxin Zhao and Fei Zou

University of North Carolina at Chapel Hill

October 19, 2018

Abstract

Polygenic risk scores (PRS) is one of the most popular prediction methods for complex traits and diseases with high-dimensional genome-wide association (GWAS) data where sample size n is typically much smaller than the number of SNPs p. PRS is a weighted sum of candidate SNPs in a testing data where each SNP is weighted by its estimated marginal effect from a training data. The motivations behind PRS are that 1) only summary statistics are needed for constructing PRS rather than raw data which may not be readily available due to privacy concerns; 2) most complex traits are affected by many genes with small effects, or follow a polygenic (or newly emerging omnigenic) model. PRS aggregates information from all potential causal SNPs and thus as its name suggested, is expected to be powerful for ploygenic and omnigenic traits. However, disappointing to many researchers, the prediction accuracy of PRS in practice is low, even for traits with known high heritability. To solve this perplex, in this paper we investigate PRS both empirically and theoretically. We show in addition to heritability, how the performance of PRS is influenced by the triplet (n, p, m), where m is the number of true causal SNPs. Our major findings are that 1) when PRS is constructed with all p SNPs (referred as GWAS-PRS), its prediction accuracy is solely determined by the p/n ratio; 2) when PRS is built with a list of top-ranked SNPs that pass a pre-specified P-value threshold (referred as threshold-PRS), its accuracy can vary dramatically depending on how sparse true genetic signals are. Only when m is magnitude smaller than n, or genetic signals are sparse, can threshold-PRS perform well. In contrast, if m is much larger than n, or genetic signals are not sparse, which is often the case for complex polygenic traits, threshold-PRS is expected to fail. Our results demystify the poor performance of PRS and demonstrate that the original purpose of PRS to aggregate effects from a large number of causal SNPs for polygenic traits is wishful and can lead us to a practical paradox for polygenic/omnigenic traits. Our results, as turned out, are closely related to the "spurious correlation" problem of Fan et al. [2012], which has been gaining more and more attention in the statistics community.

Keywords and phrases: polygenic risk scores; GWAS; high-dimensional prediction; phenotypic prediction; spurious correlation; polygenic; omnigenic; dense signals

1 Introduction

With the rapid development in biomedical technologies, various types of large-scale genetics and genomics data, including genome-wide association studies (GWAS) data have been collected for better understanding of genetic etiologies underlying complex human diseases and traits. GWAS study association between complex traits and genome-wide single-nucleotide polymorphisms (SNPs), one of the most common types of genetic variants. To detect SNPs that are associated with a given phenotype, single SNP analysis is commonly performed to estimate and test an association between the phenotype and each candidate SNP one at a time, while effects of non-genetic factors and population substructures are adjusted for [Price et al., 2006]. Tens of thousands of statistically significant SNPs have been detected for hundreds of human diseases/traits through GWAS [MacArthur et al., 2016, Visscher et al., 2017]. However, most of the identified SNPs have very low marginal genetic effects, explaining only a very small portion of the phenotypic variation even for traits with known high heritability [Visscher et al., 2012], resulting in a so called "missing heritability" phenomenon [Manolio et al., 2009, Zuk et al., 2012]. One explanation for the missing heritability is that most complex traits are polygenic, affecting by many genes whose individual effect is small [Timpson et al., 2018]. The polygenicity has long been hypothesized [Fisher, 1919, Gottesman and Shields, 1967, Penrose, 1953] and supported by increasing empirical evidence [Dudbridge, 2016, Ge et al., 2017, Kemp et al., 2017, Lee et al., 2012, Shi et al., 2016, Wray et al., 2018, Yang et al., 2015, 2010].

One of the ultimate goals of GWAS is to build a genetic risk model for accurate phenotype prediction. For polygenic traits, Purcell et al. [2009] propose a polygenic risk score (PRS), which is a weighted sum of top ranked candidate SNPs in a testing data where each SNP is weighted by its estimated marginal effect from a training data. As its name suggested, PRS aims to aggregate genetic effects of polygenes, and is thus expected to be powerful for polygenic traits, and more true for omnigenic traits. The omnigenic model is a newly emerging model Boyle et al. [2017] assuming that a trait is affected by majority (if not all) of candidate SNPs.

Though PRS has been widely used in neuropsychiatric diseases/disorders, such as bipolar and schizophrenia [Bogdan et al., 2018, Ripke et al., 2014], the prediction power of PRS remains disappointedly low with little clinical utility, even for traits with known high heritability [Márquez-Luna et al., 2017, Torkamani et al., 2018, Zheutlin and Ross, 2018]. Two legitimate reasons for the poor performance of PRS include 1) poor SNP arrays with low coverage of causal SNPs; and 2) low quality top-ranked SNPs in tagging causal SNPs [Chatterjee et al., 2016, Wray et al., 2013]. However, as will be shown by the paper, even in the absence of the above two reasons, PRS can still perform poorly. Thus far, except some experimental studies [Chatterjee et al., 2013, Daetwyler et al., 2008, Dudbridge, 2013, Pasaniuc and Price, 2017, Vilhjálmsson et al.,

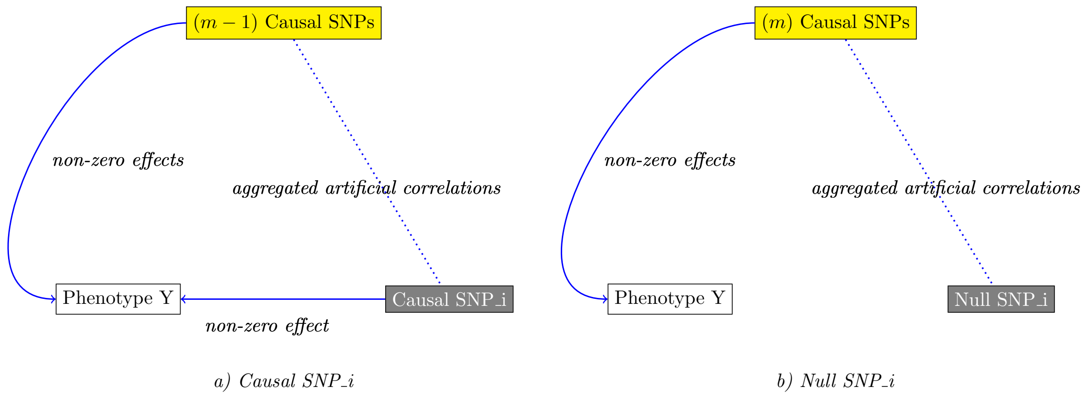


Fig. 1: Impact of artificial correlation on the marginal SNP effect estimate of SNP i $(i = 1, \cdots, p)$.

2015], few research has been seriously done to study the asymptotic properties of PRS for polygenic and omnigenic traits.

We aim to fill the gap by empirically and theoretically studying PRS in hoping to clear some misperceptions on PRS and to provide some practical guidelines on PRS. Since PRS is built upon marginal SNP effects, we start our investigation on the statistical properties of marginal SNP effect estimates. Note for polygenic traits, the single SNP analysis is always misspecified since the effects of many other SNPs are ignored. When all causal SNPs are independent of each other, such model misspecification is in general fine for traits with sparse genetic signals, but can fail badly for traits with dense genetic signals. For a given SNP, the omitted SNPs can greatly influence the uncertainty in its marginal effect estimate, and make the estimate unreliable. As will be illustrated later, even for a fully heritable phenotype with genetic heritability of one, the estimated genetic effects of causal and non causal SNPs can be totally mixed and nonseparable from each other, and the prediction power of PRS can go as low as zero.

It turns out, our theoretical investigation on the marginal genetic effect estimates is highly relevant to the *spurious correlation* problem of Fan et al. [2012], which provides another perspective on PRS. Under high-dimensional settings, the negative influences of (maximum) artificial/spurious correlation have been characterized in the context of variable selection, covariance structure testing, and variance estimation [Cai et al., 2013, 2011, Chen et al., 2018, Fan et al., 2012, 2018, Fan and Zhou, 2016, Su, 2018], but is mainly out of the genetics field. For complex polygenic traits, spurious correlation makes the estimation of marginal effects unreliable and the separation of causal and null SNPs difficult, leading to a doomed failure of PRS which contradicts the original motive of PRS (Figure 1) completely.

In recognition of the relationship between the GWAS marginal screening and PRS, we prove that the asymptotic prediction accuracy of PRS is largely affected by the triplet of (n, p, m). Our investigation on PRS starts with GWAS-PRS, and ends with the threshold-PRS. Extensive simulation will be performed to evaluate PRS empirically and to evaluate our theoretic results under finite sample settings.

Single SNP Analysis

For a training data, let y be an $n \times 1$ phenotypic vector. Let $X_{(1)}$ denote an $n \times m$ matrix of the causal SNPs, $X_{(2)}$ donate an $n \times (p-m)$ matrix of the null SNPs which results in the $n \times p$ matrix of all SNPs, $X = [X_{(1)}, X_{(2)}] = [x_1, \cdots, x_m, x_{m+1}, \cdots, x_p]$. Columns of X are assumed to be independent for simplification. Further, column-wise normalization on X is often performed such that each SNP has sample mean zero and sample variance one. Define the following condition:

Condition 1. Entries of $X = [X_{(1)}, X_{(2)}]$ are real-value independent random variables with mean zero, variance one and a finite eighth order moment.

The polygenic model assumes the following relationship between y and X:

$$y = \sum_{i=1}^{p} x_i \beta_i + \epsilon = X\beta + \epsilon \tag{1}$$

where $\beta = (\beta_1, \cdots, \beta_m, \beta_{m+1}, \cdots, \beta_p)$ is the vector of SNP effects such that the β_is are i.i.d and follow $N(0, \sigma_\beta^2)$ for $i = 1, \cdots, m$ and $\beta_i = 0$ for $i > m$. Let $\beta_{(1)} = (\beta_1, \cdots, \beta_m)$ and $\beta_{(2)}$ be an $(p-m) \times 1$ vector with all elements being zero, and ϵ represents the random error vector. For simplicity and without loss of generality, we assume that there exists no other fixed covariate effects. According to the above model, the overall genetic heritability h^2 of y is therefore

$$h^2 = \frac{Var[X_{(1)}\beta_{(1)}]}{Var(y)} = \frac{Var[X_{(1)}\beta_{(1)}]}{Var[X_{(1)}\beta_{(1)}] + Var(\epsilon)}. \tag{2}$$

For the rest of the paper, we set $h^2 = 1$, reducing the above model to the following deterministic model

$$y = \sum_{i=1}^{p} x_i \beta_i = \sum_{i=1}^{m} x_i \beta_i = X_{(1)}\beta_{(1)}, \tag{3}$$

the most optimistic situation in predicting phenotypes.

Note for a typical large-scale GWAS, the sample size n is often not small (e.g., $n \sim 1000$ or 10000), but the number of candidate SNPs p is usually even larger (e.g., $p \sim 500000$). On the other hand, depending on their underlying genetic architectures, the number of causal SNPs, m can vary dramatically from one trait to another. We therefore assume $n, p \to \infty$ and that

$$\frac{p}{n} = \gamma \to \gamma_0, \quad \frac{m}{p} = \omega \to \omega_0, \quad \text{where} \quad 0 < \gamma_0 \le \infty, \quad 0 \le \omega_0 \le 1 \tag{4}$$

to cover the most of modern GWAS data.

For continuous traits, a typical single SNP analysis employs the following linear regression model

$$y = 1_n\mu + x_i\beta_i + \epsilon^* \tag{5}$$

for a given SNP i, where β_i is its effect ($i = 1, \cdots, p$). When both y and x_i are normalized and $n \to \infty$, under Condition (1) and polygenic model (3), the maximum likelihood estimate (MLE) of μ, $\hat{\mu} \equiv 0$, and the MLE of the genetic effect β_i equals

$$\widehat{\beta}_i = (x_i^T x_i)^{-1} x_i^T y = \frac{1}{n} x_i^T y = \sum_{j=1}^{m} r_{ij}\beta_j \tag{6}$$

where $r_{ij} = \frac{1}{n} x_i^T x_j = \frac{1}{n}\sum_{k=1}^{n} x_{ik}x_{jk}$ is the sample correlation between x_i and x_j, $j = 1, \cdots, p$. Specifically, for SNP i, $i = 1, \cdots, p$, we have

$$\widehat{\beta}_i = \begin{cases} \beta_i + \sum_{j \neq i}^{m} r_{ij}\beta_j, & \text{if} \quad i \in [1, m] \\ \sum_{j=1}^{m} r_{ij}\beta_j, & \text{if} \quad i \in [m+1, p]. \end{cases} \tag{7}$$

Given that SNPs in X are independent of each other, or correlation ρ_{ij}=0 for all SNP pairs $(i \;\&\; j)(i \neq j)$, it is easy to show that asymptotically, $\widehat{\beta}_i$ is an unbiased estimator of β_i

$$E(\widehat{\beta}_i) = \begin{cases} \beta_i, & \text{if} \quad i \in [1, m] \\ 0, & \text{if} \quad i \in [m+1, p] \end{cases} \tag{8}$$

given $n \to \infty$. The associated variance of $\widehat{\beta}_i$ grows linearly with m since for any causal SNP i ($1 \leq i \leq m$)

$$Var(\sum_{j \neq i}^{m} r_{ij}\beta_j) = \sum_{j \neq i}^{m} \beta_j^2 \cdot Var(r_{ij}) = \frac{1}{n^2}\sum_{j \neq i}^{m} \beta_j^2 \cdot E(\sum_{k_1=1}^{n}\sum_{k_2=1}^{n} x_{ik_1}x_{jk_1}x_{ik_2}x_{jk_2}) \tag{9}$$

$$= \frac{1}{n^2}\sum_{j \neq i}^{m} \beta_j^2 \cdot E(\sum_{k_1=k_2=1}^{n} x_{ik_1}^2 x_{jk_1}^2) = \frac{\sum_{j \neq i}^{m} \beta_j^2}{n} = O(\frac{m}{n}) = O(\gamma \cdot \omega). \tag{10}$$

Similarly, for any null SNP i ($m < i \leq p$)

$$Var(\sum_{j \neq i}^{m} r_{ij}\beta_j) = \frac{\sum_{j=1}^{m} \beta_j^2}{n} = O(\frac{m}{n}) = O(\gamma \cdot \omega). \tag{11}$$

It follows that

$$Var(\widehat{\beta}_i) = \begin{cases} \frac{\sum_{j \neq i}^{m} \beta_j^2}{n} = O(\frac{m}{n}) = O(\gamma \cdot \omega), & \text{if} \quad i \in [1, m] \\ \frac{\sum_{j=1}^{m} \beta_j^2}{n} = O(\frac{m}{n}) = O(\gamma \cdot \omega), & \text{if} \quad i \in [m+1, p]. \end{cases} \tag{12}$$

Therefore, the (other) causal SNPs can significantly impact the effect estimate of SNP i by inflating its variance. That is, when $m/n = \gamma \cdot \omega \rightarrow \gamma_0 \cdot \omega_0$ is large, $\widehat{\beta}_i$ is no longer a reliable estimate of β_i. Also the associated variances of the $\widehat{\beta}_i$s $(i = 1, \cdots, p)$ are all in the same scale regardless whether their corresponding SNPs are causal or not. Thus the $\hat{\beta}_i$s corresponding to the causal SNP set and the null set become well mixed and cannot be separated easily when m/n is large, raising two important concerns that we address in Section 2: 1) how will this affect the SNP selection; and 2) how will this affect the weights in PRS which ultimately affect the performance of PRS?

2 PRS

For a testing data with n_z samples, define its $n_z \times p$ SNP matrix as $Z = [Z_{(1)}, Z_{(2)}]$ with $Z_{(1)} = [z_1, \cdots, z_m]$ and $Z_{(2)} = [z_{m+1}, \cdots, z_p]$. The polygenic model assumes the following relationship between y_z and Z

$$y_z = \sum_{i=1}^{p} z_i \beta_i = Z_{(1)} \beta_{(1)}. \tag{13}$$

Then PRS is defined as

$$\widehat{y}_P = \sum_{i=1}^{p} z_i \widehat{d}_i = Z\widehat{d} = Z_{(1)}\widehat{d}_{(1)} + Z_{(2)}\widehat{d}_{(2)} \tag{14}$$

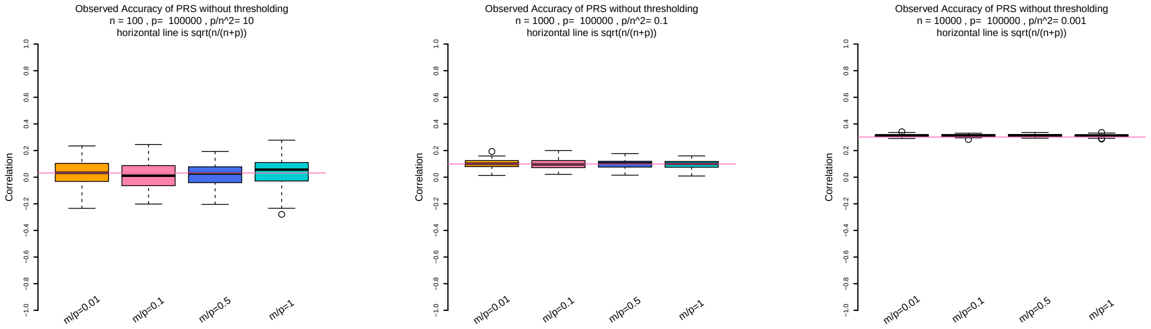
where $\widehat{d} = (\widehat{d}_1, \cdots, \widehat{d}_m, \widehat{d}_{m+1}, \cdots, \widehat{d}_p) = [\widehat{d}_{(1)}, \widehat{d}_{(2)}]$, $\widehat{d}_i = \widehat{\beta}_i \cdot I(|\widehat{\beta}_i| > c)$, $I(\cdot)$ is the indicator function and c is a given threshold for screening SNPs. When $c = 0$, all candidate SNPs are used, leading to GWAS-PRS. The prediction accuracy of PRS is measured by

$$\begin{aligned} A_P =& \frac{y_z^T \widehat{y}_P}{||y_z|| \, ||\widehat{y}_P||} \qquad (15) \\ =& \frac{(Z_{(1)}\beta_{(1)})^T (Z_{(1)}\widehat{d}_{(1)} + Z_{(2)}\widehat{d}_{(2)})}{||Z_{(1)}\beta_{(1)}|| \, ||Z_{(1)}\widehat{d}_{(1)} + Z_{(2)}\widehat{d}_{(2)}||} = \frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)}\widehat{d}_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)}\widehat{d}_{(2)}}{||Z_{(1)}\beta_{(1)}|| \, ||Z_{(1)}\widehat{d}_{(1)} + Z_{(2)}\widehat{d}_{(2)}||}. \qquad (16) \end{aligned}$$

2.1 GWAS-PRS

For GWAS-PRS, $\widehat{d}_{(1)} = \widehat{\beta}_{(1)}$, and $\widehat{d}_{(2)} = \widehat{\beta}_{(2)}$. For simplification, for rest of the paper, we set $n_z = n$ and our general conclusions remain the same when the two are different. Let $\widehat{\beta} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_m, \widehat{\beta}_{m+1}, \cdots, \widehat{\beta}_p) = [\widehat{\beta}_{(1)}, \widehat{\beta}_{(2)}]$, then

$$\widehat{\beta}_{(1)} = \frac{1}{n} X_{(1)}^T X_{(1)} \beta_{(1)}, \qquad \widehat{\beta}_{(2)} = \frac{1}{n} X_{(2)}^T X_{(1)} \beta_{(1)}, \text{ and} \tag{17}$$



(a) n=100, p=100000 **(b)** n=1000, p=100000 **(c)** n=10000, p=100,000

Fig. 2: Prediction accuracy (A_P) of GWAS-PRS across different m/p ratios when $c = 0$ (i.e., all SNPs are selected). We set p=100000, n=100, 1000 and 10000, respectively.

$$A_P = \frac{\beta_{(1)}^T Z_{(1)}^T [Z_{(1)}\widehat{\beta}_{(1)} + Z_{(2)}\widehat{\beta}_{(2)}]}{||Z_{(1)}\beta_{(1)}||||Z_{(1)}\widehat{\beta}_{(1)} + Z_{(2)}\widehat{\beta}_{(2)}||} = \frac{C_1}{\{VAR_1\}^{1/2}\{VAR_2\}^{1/2}} \tag{18}$$

where

$$C_1 = \beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \tag{19}$$

$$VAR_1 = \beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)} \tag{20}$$

$$VAR_2 = [\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]. \tag{21}$$

Theorem 1. Under the polygenic model (3) and Condition (1), if $m \to \infty$, $p \to \infty$ and $p/n^2 \to 0$ as $n \to \infty$, then

$$A_P^2/(\frac{n}{n+p}) = A_P^2/(\frac{1}{1+\gamma}) = 1 + o_p(1). \tag{22}$$

If further we assume that $p = c \cdot n^\alpha$ for some constant $c \in (0, \infty)$, $\alpha \in (0, \infty]$, then

$$A_P^2 = \begin{cases} 1 + o_p(1), & \text{if} \quad 0 < \alpha < 1 \\ 1/(1+c) + o_p(1), & \text{if} \quad \alpha = 1 \\ o_p(1), & \text{if} \quad 1 < \alpha \end{cases} \tag{23}$$

as $n \to \infty$.

Remark 1. A_P^2 has nonzero asymptotic limit provided that $\alpha \in (0, 1]$. As illustrated in Figure 2, A_P converges to zero if $\alpha \in [2, \infty]$, indicating the null prediction power of GWAS-PRS even for traits that are fully heritable.

Remark 2. Since causal SNPs are not known as a prior but estimated, poorly selected SNPs are often used to explain the poor performance of PRS. However, our theorem suggests that when m is large, even in the oracle situation where the selected SNP list contains all and only all of the causal SNPs, PRS performance is limited by the ratio of m/n. For complex traits underlying the omnigenic model [Boyle et al., 2017], the expected prediction power of PRS is essentially zero regardless.

Remark 3. In our investigation, we set $h^2 = 1$ to reflect the most optimistic situation for phenotype prediction. If $h^2 < 1$, as demonstrated by the simulation study, the prediction power of GWAS-PRS gets further decreased.

2.2 Illustration of asymptotic limits

We numerically evaluate the analytical results above and the performance of A_P^2 with $p = 100000$, and $n = n_z = 100$, 1000 and 10000, respectively. Each entry of X and Z is independently generated from $N(0, 1)$. We also vary the ratio of causal SNPs m/p from 0.01 to 1 to reflect a wide range of SNP signals, from very sparse to very dense situations, respectively. The non-zero SNP effects of $\beta_{(1)}$ are independently generated from $N(0, 1)$. The phenotypes y and y_z are generated from Model (3) and Model (13), respectively. A total of 100 replications are conducted for each simulation set up.

Figure 2 shows the distributions of the 100 A_P values across different simulation set ups. As expected, the mean of A_P remains nearly constant regardless of m, and is close to $\sqrt{n/(n+p)}$. For small n, A_P is close to zero with a large variance.

2.3 Threshold-PRS

As shown in Theorem 1, the asymptotic limit of A_P^2 associated with GWAS-PRS does not depend on m, the number of causal SNPs, but n, the sample size of the training data. For polygenic and omnigenic traits where sample size is surely smaller than the number of candidate SNPs, GWAS-PRS is doomed to fail and therefore should be avoided. It is thus natural to turn our attention to threshold-PRS and investigate whether with a properly selected threshold c, the performance of PRS can be rescued.

2.3.1 General Setup

For a given threshold $c > 0$, let's define $q = pa$ ($a \in (0, 1]$) where q is the number of selected SNPs, among which q_1 is the number of true causal SNPs and the remaining q_2 is the number of null SNPs. Therefore, $q = q_1 + q_2$. Let $Z_{(1)} = [Z_{(11)}, Z_{(12)}]$, $Z_{(2)} = [Z_{(21)}, Z_{(22)}]$; $X_{(1)} = [X_{(11)}, X_{(12)}]$, $X_{(2)} = [X_{(21)}, X_{(22)}]$; $\widehat{\beta}_{(1)} = [\widehat{\beta}_{(11)}, \widehat{\beta}_{(12)}]$, and $\widehat{\beta}_{(2)} = [\widehat{\beta}_{(21)}, \widehat{\beta}_{(22)}]$, where $Z_{(11)}$, $X_{(11)}$, $\widehat{\beta}_{(11)}$ correspond to the selected q_1 causal SNPs, and $Z_{(21)}$, $X_{(21)}$, $\widehat{\beta}_{(21)}$ correspond to the selected q_2 null SNPs. The prediction accuracy of threshold-PRS is measured by

$$A_P = \frac{\beta_{(1)}^T Z_{(1)}^T [Z_{(11)}\widehat{\beta}_{(11)} + Z_{(21)}\widehat{\beta}_{(21)}]}{||Z_{(1)}\beta_{(1)}||||Z_{(11)}\widehat{\beta}_{(11)} + Z_{(21)}\widehat{\beta}_{(21)}||} = \frac{C_1}{\{VAR_1\}^{1/2}\{VAR_2\}^{1/2}} \qquad (24)$$

where

$$C_1 = \beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)} \tag{25}$$

$$VAR_1 = \beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)} \tag{26}$$

$$VAR_2 = [\beta_{(1)}^T X_{(1)}^T X_{(11)} Z_{(11)}^T + \beta_{(1)}^T X_{(1)}^T X_{(21)} Z_{(21)}^T][Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}]. \tag{27}$$

Theorem 2. Under the polygenic model (3) and Condition (1), if $m, q_1, q_2 \to \infty$ when $n, p \to \infty$, and further if $[m^2(q_1 + q_2)]/(n^2 q_1^2) \to 0$, we have

$$A_P^2 / [\frac{nq_1^2}{nmq_1 + qm^2}] = 1 + o_p(1). \tag{28}$$

However, if $[m^2(q_1 + q_2)]/(n^2 q_1^2) \nrightarrow 0$, then

$$A_P^2 = O_p(\frac{1}{n}) = o_p(1). \tag{29}$$

Theorem 2 shows that given n and m, A_P is determined by q_1, the number of selected causal SNPs, and q, the number of selected SNPs. Expressing q_1 as a function of q such that $q_1 = \phi(q)$, we have A_P expressed as a function of q:

$$A_P^2(q) = \frac{n\phi(q)^2}{nm\phi(q) + qm^2}. \tag{30}$$

2.3.2 Role of $\phi(q)$

Function ϕ is non-decreasing with q and plays an important role in determining the asymptotic distribution of A_P. The exact form of ϕ is trait dependent and not easy to obtain. But in the following two special examples, we can demonstrate the impact of ϕ on A_P straightforwardly. To begin with, we first study the marginal distribution of the $\hat{\beta}_i$s, which is a mixture of two distributions, one corresponding to the causal SNP set and one to the null SNP set. Let $\widehat{\beta} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_m, \widehat{\beta}_{m+1}, \cdots, \widehat{\beta}_p) = [\widehat{\beta}_{(1)}, \widehat{\beta}_{(2)}]$. Given that $\beta_{(1)} \sim MNV_m(0_m, \sigma_\beta^2 \cdot I_m)$, and the remaining ones in $\beta_{(2)}$ are all 0, according to the central limit theorem, we have asymptotically

$$\widehat{\beta}_c \sim \begin{cases} N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}), & \text{if} \quad c \in [1, m] \\ N(0, \sigma_\beta^2 \cdot \frac{m}{n}), & \text{if} \quad c \in [m+1, p]. \end{cases} \tag{31}$$

When $m/n = \gamma \cdot \omega \to \gamma_0 \cdot \omega_0 = 0$, the spread of the marginal distribution of the causal SNPs is much wider than that of the marginal distribution of the null SNPs, making the two distributions separable and single SNP analysis powerful. However, as the genetic signal gets denser and denser (or m increases), the difference between the two distributions gets smaller and smaller, leading to two well mixed distributions and poorly performed single SNP analysis. To see how the ratio of m/n impacts single

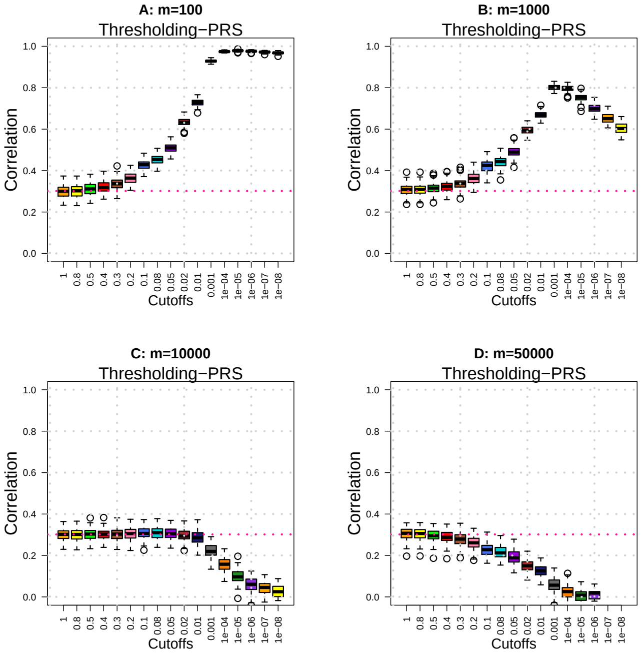


Fig. 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set p=100000, n = 10000 in training data and n=1000 in testing data.

SNP analysis, we first approximate the density of $\widehat{\beta}_c$ by

$$f(b) \approx \frac{m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}) + \frac{p-m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{m}{n}) \tag{32}$$

$$= \frac{m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(-\frac{b^2}{2\sigma_1^2}) + \frac{p-m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{b^2}{2\sigma_2^2}) \tag{33}$$

with CDF

$$F(b) \approx \frac{m}{p} \cdot \Phi(\frac{b}{\sigma_1}) + \frac{p-m}{p} \cdot \Phi(\frac{b}{\sigma_2}) \tag{34}$$

where $\sigma_1^2 = \sigma_\beta^2 \cdot \frac{n+m}{n}$, $\sigma_2^2 = \sigma_\beta^2 \cdot \frac{m}{n}$, and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2) dt$ is the CDF of the standard normal random variable. Since the mixture distribution is symmetric about zero, without loss of generality, in the following we consider one-sided test and SNPs with the largest $(100 \times a)\%$ estimated genetic effects $(0 < a < 1/2)$ are selected. For a causal SNP, its selection probability κ_1 equals

$$Pr[b > F^{-1}(1-a) | b \sim N(0, \sigma_1^2)] = 1 - Pr[b \leq F^{-1}(1-a) | b \sim N(0, \sigma_1^2)] \tag{35}$$

$$= 1 - \Phi[\frac{F^{-1}(1-a)}{\sigma_1}] = 1 - \Phi[\frac{F^{-1}(1-a)}{\sigma_\beta} \sqrt{\frac{n}{n+m}}]. \tag{36}$$

Similarly for a given null SNP, its selection probability κ_2 is

$$Pr[b > F^{-1}(1-a)|b \sim N(0,\sigma_2^2)] = 1 - Pr[b \leq F^{-1}(1-a)|b \sim N(0,\sigma_2^2)] \tag{37}$$

$$= 1 - \Phi\Big[\frac{F^{-1}(1-a)}{\sigma_2}\Big] = 1 - \Phi\Big[\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{m}}\Big]. \tag{38}$$

Therefore, among $q = pa = q_1 + q_2$ selected SNPs, we expect

$$m \cdot \kappa_1 = m \cdot \big[1 - \Phi(\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{n+m}})\big] \quad \text{and} \tag{39}$$

$$(p-m) \cdot \kappa_2 = (p-m) \cdot \big[1 - \Phi(\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{m}})\big], \tag{40}$$

causal and null SNPs, respectively. For a given a or equivalently c, $\frac{F^{-1}(1-a)}{\sigma_\beta}$ is the same for both causal and null SNPs. Therefore the quality of top-ranked SNP list is largely determined by m/n. The next Remark discusses the upper bounds of A_p under two extreme cases.

Remark 4. When $n/m = o(1)$, it is easy to see $\kappa_1 = \kappa_2 \cdot (1 + o(1))$. Thus when q_1, q_2 are large, $q_1/q \approx m/p$, and thus $q_1 = \phi(q) \approx \frac{m}{p} \cdot q$. It follows that

$$A_P^2(q) = \frac{n\phi(q)^2}{nm\phi(q) + qm^2} \approx \frac{n}{np + p^2} \cdot q. \tag{41}$$

Therefore A_P reaches its upper bound $\sqrt{n/(n+p)}$ at $q = p$, suggesting that the best performing PRS is the one that constructed *without SNP selection or GWAS-PRS* when the genetic signals are dense. On the other hand, when $m/n = o(1)$, κ_1 becomes much larger than κ_2. Thus causal SNPs can be relatively easy to detect by single SNP analysis. As a increases, q_1 eventually gets saturated at m, and threshold-PRS reaches its upper performance limit $\sqrt{n/(n+m)}$ with $q = q_1 = \phi(q) = m$, which is the oracle case described in Remark 2.

In conclusion, the above analysis provides guidelines on constructing PRS: 1) SNP screening should be avoided for highly polygenic/omnigenic traits with a large m/n ratio. 2) For monogenic and oligogenic traits [Timpson et al., 2018] with a small m/n ratio, threshold-PRS is preferred.

3 More simulation studies

To illustrate the finite sample performance of threshold-PRS, we simulate $p = 100000$ uncorrelated SNPs. Again as in Figure 2, we (naively) generate each entry of the SNPs from $N(0,1)$. To study effect of m/p, we vary the number of causal SNPs m and set it to 100, 1000, 10000 and 50000. The nonzero SNP effects β_is are independently generated from $N(0,1)$. The linear polygenic model in Model (1) is used to generate phenotype y. The sample size is set to 1000 and 10000 for training data, and 1000 for testing data. For threshold-PRS, as in Márquez-Luna et al. [2017], we consider a series of P-value thresholds $\{1, 0.8, 0.5,\ 0.4, 0.3, 0.2,\ 0.1, 0.08, 0.05,\ 0.02, 0.01, 10^{-3}$, $10^{-4}, 10^{-5},\ 10^{-6}, 10^{-7}, 10^{-8}\}$. We name this simulation setting Case 1. A total of 100 replications are conducted for each simulation condition.

Figure 3 and Supplementary Figure 1 display the performance of threshold-PRS across a series of m/p ratios in Case 1. As expected, the performance of GWAS-PRS (i.e. at P-value threshold of 1) is nearly constant around $\sqrt{n/(n+p)}$ regardless of the m/p ratios, which is about 0.3 (as shown in Figure 3) for $n = 10000$ and 0.1 (shown in Supplementary Figure 1) for $n = 1000$. The performance of threshold-PRS varies with the m/p (or m/n) ratio. When m is small compared to n, threshold-PRS performs significantly better than GWAS-PRS provided a reasonable c is chosen which in general is small as shown by Supplementary Figure 1. In this figure, when $m = 100$ and $n = 1000$, threshold-PRS achieves its best performance at $c = 10^{-5}$, with A_P of 0.75, in contrast to its oracle performance which is about 0.95. Figure 3 shows that when m gets close to n or larger than n, the performance of threshold-PRS drops significantly regardless of c. When m is close to n, its performance remains similar for a wide range of c values; and when m gets much larger than n, its performance improves as c increases, and eventually reaches the same performance level of GWAS-PRS.

In addition, we vary Case 1 settings to check the sensitivity of our results. In Case 2, we generate actual SNP genotype data where the minor allele frequency (MAF) of each SNP, f, is independently generated from Uniform $[0.05, 0.45]$ and SNP genotypes are independently sampled from $\{0, 1, 2\}$ with probabilities $\{(1-f)^2, 2f(1-f), f^2\}$, respectively according to the Hardy-Weinberg equilibrium principle. In Case 3, we simulate mixed samples from five subpopulations. The overall MAF of each SNP in mixed samples is independently generated from Uniform $[0.05, 0.45]$, and the F_{st} values are independently generated from Uniform $[0.01, 0.04]$ [Lee et al., 2011] based on which the MAF of each sub-population is generated according to the Balding-Nichols model [Balding and Nichols, 1995]. We set the sample size of each sub-population the same at 200 and 2000. The population substructures are estimated with the PCA analysis of Price et al. [2006] and the top 4 PCs are included as covariates in the single SNP analysis. Case 4 allows larger variability in the causal SNP effects such that β_is are independently generated from $N(0, \sigma_i^2)$, where σ_i^{-2} follows a gamma distribution with $\alpha = 10$ and $\beta = 9$.

The results of Case 2 are displayed in Supplementary Figures 2 - 3, which are similar to those of Case 1. Supplementary Figures 4 - 5 display the oracle performance of PRS under varying m/p ratios in Case 2. Clearly A_p is around $\sqrt{n/(n+m)}$, confirming the poor performance of PRS even in the oracle case when genetic signals are dense. The results of Case 3 are displayed in Supplementary Figures 6 - 7. In the presence of sub-population structures, if they are properly adjusted, the main pattern of threshold-PRS remains unchanged and the performance of GWAS-PRS agrees well with the theoretical results. The results of Case 4 are displayed in Supplementary Figures 8 - 9, which are also similar to those of Case 1, indicating that our asymptotic results are not sensitive to the distribution of nonzero SNP effects β_is.

4 Discussion

PRS is one of the most popular prediction methods for GWAS data and has been widely used for predicting a phenotype of the same type as the training GWAS or other pleiotropy phenotypes [Choi et al., 2018]. Motivated by the poor practice performance of PRS, we empirically and theoretically study the properties of PRS for complex polygenic traits generated from modern GWAS. For GWAS-PRS, our asymptotic results align well with those of Daetwyler et al. [2008] and Chatterjee et al. [2013], Dudbridge [2013], but more statistically rigorous. In addition, for threshold-PRS, we illustrate how genetic sparseness affects its prediction performance and recognize its distinct behaviors under dense and sparse genetic signal scenarios. It turns out, the performance of PRS is closely related to the increasingly recognized spurious correlation problem [Fan et al., 2012] associated with marginal screenings such as single SNP analysis. For polygenic traits, models used by single SNP analysis are always misspecified where effects of a large number of causal SNPs are absorbed into the error term, leading to spurious correlation, which can profoundly affect the performances of GWAS-PRS and threshold-PRS, an issue that can be safely ignored for non-polygenic traits. In our study, we set h^2=1 and assume all causal SNPs are observed, which is the most optimistic situation for phenotypic prediction. We can easily extend our results to $h^2 < 1$ cases. For example, the asymptotic prediction accuracy of GWAS-PRS becomes to $\sqrt{nh^2/(n+p/h^2)}$. The performance of PRS under Case 2 but for traits with h^2=0.5 is presented in Supplementary Figures 10 - 11. Compared to Supplementary Figures 2 - 3 where h^2=1, though the prediction accuracy of PRS is reduced, the general conclusions remain the same. Besides phenotypic prediction, our research also illustrates for the first time how and why commonly used marginal screening approaches for GWAS data may fail in preserving the rank of genetic signals.

In summary, our investigation clears up some misconceptions on PRS, and demonstrates that PRS is not as useful as its name suggested, and also not as powerful as the genetics community expected for polygenic trait prediction. We hope this research will serve as a wake up call to the genetics community in recognizing the real challenges in analyzing and predicting complex polygenic traits. As such, for complex polygenic traits, more devoted efforts are needed for developing better experiments and statistical methods.

Appendix A: Proofs

In this appendix, we highlight the key steps and important intermediate results to prove our main results in Section 2. More technical details can be found in the supplementary file.

Proposition A 1. Under the polygenic model (3) and Condition (1), if $m \to \infty$ when $n, p \to \infty$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} = 1 + o_p(1) \tag{42}$$

$$\frac{[\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]}{[n^2 m(p-m) + n^2 m(m+n)] \cdot \sigma_\beta^2} = 1 + o_p(1). \tag{43}$$

Further if $p/n^2 \to 0$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2} = 1 + o_p(1). \tag{44}$$

By continuous mapping theorem, we have

$$A_P^2 / (\frac{n}{n+p}) = A_P^2 / (\frac{1}{1+\gamma}) = 1 + o_p(1). \tag{45}$$

It follows that Theorem 1 is proved for $\alpha \in (0, 2)$. Now consider the case that $p/n^2 \not\to o(1)$, i.e., $\alpha \in [2, \infty]$. Note that

$$\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \tag{46}$$

$$= \frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)} - n^2 m \cdot \sigma_\beta^2}{\sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)}} \tag{47}$$

$$\cdot \sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)} + n^2 m \cdot \sigma_\beta^2 \tag{48}$$

$$= O_p(\sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)}) + n^2 m \cdot \sigma_\beta^2. \tag{49}$$

It follows that

$$A_P^2 = \frac{O_p[(n^2 m^2 p + n^4 m^2) \cdot \sigma_\beta^4]}{[nm \cdot \sigma_\beta^2 \cdot (1 + o(1))] \cdot [n^2 m(n+p) \cdot \sigma_\beta^2 \cdot (1 + o(1))]} \tag{50}$$

$$= O_p(\frac{n^2 m^2 p + n^4 m^2}{n^3 m^2 p + n^4 m^2}) = O_p(\frac{n^2 + c \cdot n^\alpha}{n^2 + c \cdot n^{1+\alpha}}) = O_p(\frac{1}{n}) \tag{51}$$

when $\alpha \in [2, \infty]$. Thus Theorem 1 is proved for $\alpha \in (0, \infty]$.

Proposition A 2. Under the polygenic model (3) and Condition (1), if $m, q_1, q_2 \to \infty$

when $n, p \to \infty$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} = 1 + o_p(1) \tag{52}$$

$$\frac{[\beta_{(1)}^T X_{(1)}^T X_{(11)} Z_{(11)}^T + \beta_{(1)}^T X_{(1)}^T X_{(21)} Z_{(21)}^T][Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}]}{[n^2 m q_2 + n^2 q_1 (m+n)] \cdot \sigma_\beta^2} = 1 + o_p(1). \tag{53}$$

Further if $[m^2(q_1 + q_2)]/(n^2 q_1^2) \to 0$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}}{n^2 q_1 \cdot \sigma_\beta^2} = 1 + o_p(1). \tag{54}$$

By continuous mapping theorem, we have

$$A_P^2 / [\frac{n q_1^2}{nm q_1 + q m^2}] = 1 + o_p(1). \tag{55}$$

Note that

$$\beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)} \tag{56}$$

$$= O_p[\sqrt{n^2 q_1^3 + 2n^2 q_1^2 (m - q_1) + 2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2}] + n^2 q_1. \tag{57}$$

Then if $[m^2(q_1 + q_2)]/(n^2 q_1^2) \not\to 0$, we have

$$A_P^2 = \frac{O_p(n^2 q_1^3 + 2n^2 q_1^2 (m - q_1) + 2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2 + n^4 q_1^2)}{[nm \cdot (1 + o_p(1))] \cdot [(n^3 q_1 + n^2 m q_1 + n^2 m q_2) \cdot (1 + o_p(1))}] \tag{58}$$

$$= O_p[\frac{m^2(q_1 + q_2) + 2n(m - q_1) q_1 + n^2 q_1^2}{nm^2 (q_1 + q_2) + n^2 m q_1}] = O_p[\frac{m^2(q_1 + q_2) + n^2 q_1^2}{nm^2 (q_1 + q_2) + n^2 m q_1}] \tag{59}$$

$$= O_p(\frac{1}{n}) = o_p(1). \tag{60}$$

Thus Theorem 2 is proved.

Acknowledgement

The authors would like to thank Ziliang Zhu, Jingwen Zhang, Hongtu Zhu and Xinlei Mi for helpful discussions and suggestions.

References

Balding, D. J. and Nichols, R. A. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Bogdan, R., Baranger, D. A. and Agrawal, A. (2018) Polygenic risk scores in clinical psychology: bridging genomic risk to individual differences. *Annual review of clinical psychology*, **14**, 119–157.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.

Cai, T., Fan, J. and Jiang, T. (2013) Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research*, **14**, 1837–1864.

Cai, T. T., Jiang, T. et al. (2011) Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, **39**, 1496–1525.

Chatterjee, N., Shi, J. and García-Closas, M. (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, **17**, 392.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J. and Park, J.-H. (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, **45**, 400.

Chen, Z., Fan, J. and Li, R. (2018) Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, **113**, 315–327.

Choi, S. W., Mak, T. S. H. and O'Reilly, P. (2018) A guide to performing polygenic risk score analyses. *bioRxiv*, 416545.

Daetwyler, H. D., Villanueva, B. and Woolliams, J. A. (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS one*, **3**, e3395.

Dudbridge, F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, **9**, e1003348.

— (2016) Polygenic epidemiology. *Genetic epidemiology*, **40**, 268–272.

Fan, J., Guo, S. and Hao, N. (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 37–65.

Fan, J., Shao, Q.-M., Zhou, W.-X. et al. (2018) Are discoveries spurious? distributions of maximum spurious correlations and their applications. *The Annals of Statistics*, **46**, 989–1017.

Fan, J. and Zhou, W.-X. (2016) Guarding against spurious discoveries in high dimensions. *Journal of Machine Learning Research*, **17**, 1–34.

Fisher, R. A. (1919) Xv.the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. and Smoller, J. W. (2017) Phenome-wide heritability analysis of the uk biobank. *PLoS genetics*, **13**, e1006711.

Gottesman, I. and Shields, J. (1967) A polygenic theory of schizophrenia. *Proceedings of the National Academy of Sciences*, **58**, 199–205.

Kemp, J. P., Morris, J. A., Medina-Gomez, C., Forgetta, V., Warrington, N. M., Youlten, S. E., Zheng, J., Gregson, C. L., Grundberg, E., Trajanoska, K. et al. (2017) Identification of 153 new loci associated with heel bone mineral density and functional involvement of gpc6 in osteoporosis. *Nature genetics*, **49**, 1468.

Lee, S., Wright, F. A. and Zou, F. (2011) Control of population stratification by correlation-selected principal components. *Biometrics*, **67**, 967–974.

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., Wray, N. R., Consortium, S. P. G.-W. A. S. et al. (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature genetics*, **44**, 247.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. et al. (2016) The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, **45**, D896–D901.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A. et al. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747.

Márquez-Luna, C., Loh, P.-R. and Price, A. L. (2017) Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, **41**, 811–823.

Pasaniuc, B. and Price, A. L. (2017) Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, **18**, 117.

Penrose, L. (1953) The genetical background of common diseases. *Human Heredity*, **4**, 257–265.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**, 904–909.

Purcell, S. M., Wray, R., Stone, L., Visscher, M., O'Donovan, C., Sullivan, F., Sklar, P., Ruderfer, M., McQuillin, A., Morris, W. et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H. et al. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421.

Shi, H., Kichaev, G. and Pasaniuc, B. (2016) Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, **99**, 139–153.

Su, W. J. (2018) When is the first spurious variable selected by sequential regression procedures? *Biometrika*, **105**, 517–527.

Timpson, N. J., Greenwood, C. M., Soranzo, N., Lawson, D. J. and Richards, J. B. (2018) Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, **19**, 110.

Torkamani, A., Wineinger, N. E. and Topol, E. J. (2018) The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 1.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R. et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, **97**, 576–592.

Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. (2012) Five years of gwas discovery. *The American Journal of Human Genetics*, **90**, 7–24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017) 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**, 5–22.

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. and Visscher, P. M. (2018) Common disease is more complex than implied by the core gene omnigenic model. *Cell*, **173**, 1573–1580.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. and Visscher, P. M. (2013) Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, **14**, 507.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V. et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, **47**, 1114.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010) Common snps explain a large proportion of the heritability for human height. *Nature genetics*, **42**, 565.

Zheutlin, A. B. and Ross, D. A. (2018) Polygenic risk scores: What are they good for? *Biological psychiatry*, **83**, e51–e53.

Zuk, O., Hechter, E., Sunyaev, S. R. and Lander, E. S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, **109**, 1193–1198.

5 Supplementary Material

5.1 Intermediate results

Proposition S 1. Under Condition (1), if $m \to \infty$ when $n, p \to \infty$, then

$$\begin{align}
& E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) = n^2 m \cdot \sigma_\beta^2 \tag{61}\\
& Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{62}\\
& = [(n^2 m^3 + 4n^3 m^2 + n^2 m^2 (p-m)) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)] \cdot (1 + o(1)) \tag{63}\\
& \tag{64}\\
& E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = nm \cdot \sigma_\beta^2 \tag{65}\\
& Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = o(n^2 m^2 \cdot \sigma_\beta^4) \tag{66}\\
& \tag{67}\\
& E([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]) \tag{68}\\
& = n^2 m(n+m) \cdot \sigma_\beta^2 \cdot (1 + o(1)) + n^2 m(p-m) \cdot \sigma_\beta^2 \tag{69}\\
& Var([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]) \\
& \tag{70}\\
& = o[n^4 m^2 (n+p)^2 \cdot \sigma_\beta^4] \tag{71}
\end{align}$$

where b_4 is the forth moment of β.

Proposition S 1 quantifies the scale of the three terms in A_P. Particularly, for the two variance terms $\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}$ and

$$[\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}],$$

the expected values can respectively dominate the corresponding standard error for any ratios among p, m, n. However, for the covariance term

$$\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)},$$

its standard error may or may be dominated by its expected value depending on p/n. Following Proposition S 1, by Markov's inequality, for any constant $k > 0$, we have

$$P_r(|\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} - 1| \geq k) \leq \frac{Var(\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2})}{k^2} = \frac{Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)})}{n^2 m^2 \cdot \sigma_\beta^4 k^2} = o(1), \tag{72}$$

and

$$P_r(|\frac{[\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]}{n^2 m(n+p)\cdot\sigma_\beta^2} - 1| \geq k) \tag{73}$$

$$\leq \frac{Var(\frac{[\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]}{n^2 m(n+p)\cdot\sigma_\beta^2})}{k^2} \tag{74}$$

$$= \frac{Var([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}])}{n^4 m^2 (n+p)^2 \cdot \sigma_\beta^4 k^2} \tag{75}$$

$$= o(1), \tag{76}$$

and

$$P_r(|\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2} - 1| \geq k) \tag{77}$$

$$\leq \frac{Var(\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2})}{k^2} \tag{78}$$

$$= \frac{(n^2 m^3 + n^2 m^2 (p-m)) \cdot (1+o(1))}{n^4 m^2 \cdot \sigma_\beta^4 k^2} = \frac{p}{n^2 k^2} \cdot (1 + o(1)). \tag{79}$$

Thus Proposition A 1 is proved. More generally, if training and testing data have different sample sizes, we have

Proposition S 2. Under Condition (1), if $m \to \infty$ when $n, n_z, p \to \infty$, then

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) = n n_z m \cdot \sigma_\beta^2 \tag{80}$$

$$Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) = \tag{81}$$

$$[(n n_z m^2 p + 2n^2 n_z m^2 + 2n n_z^2 m^2) \cdot \sigma_\beta^4 + n^2 n_z^2 m \cdot (b_4 - \sigma_\beta^4)] \cdot (1 + o(1)) \tag{82}$$

$$\tag{83}$$

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = n_z m \cdot \sigma_\beta^2 \tag{84}$$

$$Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = o(n_z^2 m^2 \cdot \sigma_\beta^4) \tag{85}$$

$$\tag{86}$$

$$E([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]) \tag{87}$$

$$= n n_z m (n+m) \cdot \sigma_\beta^2 \cdot (1 + o(1)) + n n_z m (p - m) \cdot \sigma_\beta^2 \tag{88}$$

$$Var([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]) \tag{89}$$

$$= o[n^2 n_z^2 m^2 (n+p)^2 \cdot \sigma_\beta^4]. \tag{90}$$

By Markov's inequality and continuous mapping theorem again,

Proposition S 3. Under Condition (1), if $m \to \infty$ when $n, n_z, p \to \infty$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}}{n_z m \cdot \sigma_\beta^2} = 1 + o_p(1) \tag{91}$$

$$\frac{[\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]}{[n n_z m(p-m) + n n_z m(n+m)] \cdot \sigma_\beta^2} = 1 + o_p(1). \tag{92}$$

If we further have $p/(n n_z) \to 0$, then

$$\frac{\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}}{n n_z m \cdot \sigma_\beta^2} = 1 + o_p(1) \tag{93}$$

and thus

$$A_P^2 / (\frac{n}{n+p}) = 1 + o_p(1). \tag{94}$$

When $p/(n n_z) \not\to 0$, i.e., $\alpha \in [1, \infty]$, note that

$$\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \tag{95}$$

$$= O_p(\sqrt{n n_z m^2 p + 2n^2 n_z m^2 + 2n n_z^2 m^2 + n^2 n_z^2 m (b_4 - \sigma_\beta^4)}) + n n_z m \cdot \sigma_\beta^2 \tag{96}$$

$$= O_p[(n^{1/2} n_z^{1/2} m p^{1/2} + n n_z m) \cdot \sigma_\beta^2]. \tag{97}$$

It follows that

$$A_P^2 = \frac{O_p(n n_z m^2 p + n^2 n_z^2 m^2)}{[n_z m \cdot (1 + o(1))] \cdot [n n_z m (n+p) \cdot (1 + o(1))]} = O_p(\frac{p + n n_z}{n_z p + n n_z}) = O_p(\frac{1}{n_z}). \tag{98}$$

The results of threshold-PRS can also be derived in a similar way. Without loss of generality, we set $\sigma_\beta^2 = 1$ below.

Proposition S 4. Under Conditions (1), if $m \to \infty$, $p - m \to \infty$ when n increase to ∞, for any $q_1, q_2 \geq 0$, then

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}) = n^2 q_1 \tag{99}$$

$$Var[\beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}] = [n^2 q_1^3 + 4n^3 q_1^2 \tag{100}$$

$$+ n^4 q_1 (b_4 - 1) + 2n^2 q_1^2 (m - q_1) + 2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2] \cdot (1 + o(1)) \tag{101}$$

$$E(VAR_2) = [n^2 q_1 (n+m) + n^2 q_2 m] \cdot (1 + o(1)) \tag{102}$$

$$Var(VAR_2) = o([n^2 q_1 (n+m) + n^2 q_2 m]^2). \tag{103}$$

By Markov's inequality and continuous mapping theorem again, Proposition A 2 is proved.

5.2 Technical details

The following technical details are useful to prove our theoretical results. Most of them involve in calculating the asymptotic expectation of the trace of the product of multiple large random matrices. To our knowledge, there is no easy way to calculate the asymptotic trace of the product of multiple general random matrices. Instead, we use the definition of matrix trace and apply the combination theory to calculate the total variations. The results provided below may also benefit other research questions involving the similar calculations.

5.2.1 GWAS-PRS

First moment of covariance term

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)}) \tag{104}$$

$$= \sigma_\beta^2 \cdot E[tr(Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)})] \tag{105}$$

$$= \sigma_\beta^2 \cdot E(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_1=1}^m \sum_{k_2=1}^m Z_{ik_1} X_{jk_1} Z_{ik_2} X_{jk_2}) \tag{106}$$

$$= \sigma_\beta^2 \cdot E(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_1=k_2=1}^m Z_{ik_1}^2 X_{jk_1}^2) = \sigma_\beta^2 \cdot n^2 m \tag{107}$$

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{108}$$

$$= \sigma_\beta^2 \cdot E[tr(Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)})] \tag{109}$$

$$= \sigma_\beta^2 \cdot E(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_2=1}^m \sum_{k_1=1}^m Z_{(2)ik_2} X_{(2)jk_2} X_{(1)jk_1} Z_{(1)ik_1}) = 0 \tag{110}$$

Thus

$$E(\beta_{(1)}^T Z_{(1)}^T Z X^T X_{(1)} \beta_{(1)}) = \sigma_\beta^2 \cdot n^2 m \tag{111}$$

First moment of variance term I

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) \tag{112}$$

$$= E[E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)} | Z)] = E[tr(Z_{(1)}^T Z_{(1)} \cdot I_m \cdot \sigma_\beta^2) + 0] \tag{113}$$

$$= \sigma_\beta^2 \cdot E[tr(Z_{(1)} Z_{(1)}^T)] = \sigma_\beta^2 \cdot E(\sum_{i=1}^n \sum_{j=1}^m Z_{ij}^2) = \sigma_\beta^2 \cdot nm \tag{114}$$

First moment of variance term II

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)}) \tag{115}$$

$$= \sigma_\beta^2 \cdot E[tr(X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)})] \tag{116}$$

$$= \sigma_\beta^2 \cdot E(\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k_1=1}^{m} \sum_{l_1=1}^{n} \sum_{k_2=1}^{m} \sum_{l_2=1}^{n} Z_{ik_1} Z_{ik_2} X_{l_1k_1} X_{l_1j} X_{l_2k_2} X_{l_2j}) \tag{117}$$

$$= \sigma_\beta^2 \cdot E(\sum_{i=1}^{n} \sum_{k_1=k_2=j}^{m} \sum_{l_1 \neq l_2}^{n(n-1)} Z_{ik}^2 X_{l_1k}^2 X_{l_2k}^2 + \sum_{i=1}^{n} \sum_{k_1=k_2\neq j}^{m(m-1)} \sum_{l_1=l_2}^{n} Z_{ik}^2 X_{lk}^2 X_{lj}^2 \tag{118}$$

$$+ \sum_{i=1}^{n} \sum_{k_1=k_2=j}^{m} \sum_{l_1=l_2}^{n} Z_{ik}^2 X_{lk}^4) \tag{119}$$

$$= \sigma_\beta^2 \cdot n^2 m(n+m+c_4-2) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot (1+o(1)) \tag{120}$$

where $c_4 = E(X_{11}^4) < \infty$.

$$E(\beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{121}$$

$$= \sigma_\beta^2 \cdot E[tr(X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)})] \tag{122}$$

$$= \sigma_\beta^2 \cdot E(\sum_{c=1}^{n} \sum_{i=1}^{m} \sum_{k=1}^{p-m} \sum_{l=1}^{n} \sum_{q=1}^{p-m} \sum_{r=1}^{n} Z_{(2)ck} Z_{(2)cq} X_{(2)lk} X_{(2)rq} X_{(1)li} X_{(1)ri}) \tag{123}$$

$$= \sigma_\beta^2 \cdot n^2 m(p-m) \tag{124}$$

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) = 0 \tag{125}$$

Thus

$$E(\beta_{(1)}^T X_{(1)}^T X Z^T Z X^T X_{(1)} \beta_{(1)}) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot (1+o(1)) + \sigma_\beta^2 \cdot n^2 m(p-m) \tag{126}$$

Second moment of covariance term

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)}) \tag{127}$$

$$= E[tr(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)})] \tag{128}$$

$$= E(\sum_{c=1}^{n} \sum_{d=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{n} \sum_{q=1}^{m} \sum_{r=1}^{m} \sum_{s=1}^{n} \sum_{t=1}^{m} Z_{ck} Z_{dq} Z_{dr} Z_{ct} X_{lk} X_{li} X_{sr} X_{sj} \beta_q \beta_i \beta_t \beta_j) \tag{129}$$

$$= [(n^4 m^2 + 4n^3 m^2 + n^2 m^3) \cdot \sigma_\beta^4 + mn^4 \cdot (b_4 - \sigma_\beta^4)] \cdot (1+o(1)) \tag{130}$$

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{131}$$

$$= E[tr(\beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)})] \tag{132}$$

$$= E(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{p-m}\sum_{l=1}^{n}\sum_{q=1}^{m}\sum_{r=1}^{p-m}\sum_{s=1}^{n}\sum_{t=1}^{m} \tag{133}$$

$$Z_{(2)ck} Z_{(2)dr} Z_{(1)dq} Z_{(1)ct} X_{(2)lk} X_{(2)sr} X_{(1)li} X_{(1)sj} \beta_q \beta_i \beta_t \beta_j) \tag{134}$$

$$= n^2(p-m)[m^2 + (\sigma_\beta^4 - 1)\cdot m]\cdot \sigma_\beta^4 = n^2 m^2 (p-m)\cdot \sigma_\beta^4 \cdot (1+o(1)) \tag{135}$$

and

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) = 0 \tag{136}$$

It follows that

$$Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{137}$$

$$= [(n^2m^3 + 4n^3m^2 + n^2m^2(p-m))\cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)]\cdot(1+o(1)) \tag{138}$$

Second moment of variance term I

$$E(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = E[tr(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)})] \tag{139}$$

$$= E(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} X_{ci}X_{dj}X_{dk}X_{cl}\beta_i\beta_j\beta_k\beta_l) \tag{140}$$

$$= E[\sum_{c=d}^{n}(\sum_{i=j\neq k=l}^{m(m-1)} X_{ci}^2 X_{ck}^2 \beta_i^2 \beta_k^2 + \sum_{i=k\neq j=l}^{m(m-1)} X_{ci}^2 X_{cj}^2 \beta_i^2 \beta_j^2 + \sum_{i=l\neq j=k}^{m(m-1)} X_{ci}^2 X_{cj}^2 \beta_i^2 \beta_j^2 \tag{141}$$

$$+ \sum_{i=l=j=k}^{m} X_{ci}^4 \beta_i^4) + \sum_{c\neq d}^{n(n-1)} (\sum_{i=l\neq k=j}^{m(m-1)} X_{ci}^2 X_{dj}^2 \beta_i^2 \beta_j^2 + \sum_{i=l=k=j}^{m} X_{ci}^2 X_{dj}^2 \beta_i^4)] \tag{142}$$

$$= \sigma_\beta^4 \cdot n^2 m^2 + nm\cdot[2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4 b_4 - 2\sigma_\beta^4 - b_4] \tag{143}$$

where $b_4 = E(\beta_1^4) < \infty$. It follows that

$$Var(\beta_{(1)}^T Z_{(1)}^T Z_{(1)} \beta_{(1)}) = nm\cdot[2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4 b_4 - 2\sigma_\beta^4 - b_4] = o(n^2m^2\cdot\sigma_\beta^4) \tag{144}$$

Second moment of variance term II

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)}) \tag{145}$$

$$= E[tr(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)})] \tag{146}$$

$$= E(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{h=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{i=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{n}\sum_{s=1}^{m}\sum_{t=1}^{m}\sum_{u=1}^{n}\sum_{w=1}^{m}\sum_{a=1}^{m}\sum_{b=1}^{n} \tag{147}$$

$$Z_{ck} Z_{dq} Z_{dt} Z_{ca} X_{lk} X_{lh} X_{rq} X_{ri} X_{ut} X_{us} X_{ba} X_{bw} \beta_h \beta_i \beta_s \beta_w) \tag{148}$$

To have non-zero means, we need the possible combinations of the following:

- either $Z_{\cdot\cdot}^2 Z_{\cdot\cdot}^2$ or $Z_{\cdot\cdot}^4$
- one of $X_{\cdot\cdot}^2 X_{\cdot\cdot}^2 X_{\cdot\cdot}^2 X_{\cdot\cdot}^2$, $X_{\cdot\cdot}^2 X_{\cdot\cdot}^2 X_{\cdot\cdot}^4$, $X_{\cdot\cdot}^4 X_{\cdot\cdot}^4$, $X_{\cdot\cdot}^6 X_{\cdot\cdot}^2$, and $X_{\cdot\cdot}^8$
- either $\beta_{\cdot}^2 \beta_{\cdot}^2$ or $\beta_{\cdot}^4$.

After tedious calculations, we have

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)}) \tag{149}$$

$$= \sigma_\beta^4 \cdot [n^4 m^2 (n+m)^2] \cdot (1 + o(1)) \tag{150}$$

Next

$$E(\beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{151}$$

$$= E[tr(\beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)})] \tag{152}$$

$$= E(\sum_{c=1}^{n} \sum_{d=1}^{n} \sum_{h=1}^{m} \sum_{k=1}^{p-m} \sum_{l=1}^{n} \sum_{i=1}^{m} \sum_{q=1}^{p-m} \sum_{r=1}^{n} \sum_{s=1}^{m} \sum_{t=1}^{p-m} \sum_{u=1}^{n} \sum_{w=1}^{m} \sum_{a=1}^{p-m} \sum_{b=1}^{n} \tag{153}$$

$$Z_{(2)ck} Z_{(2)dq} Z_{(2)dt} Z_{(2)ca} X_{(2)lk} X_{(2)rq} X_{(2)ut} X_{(2)ba} X_{(1)lh} X_{(1)ri} X_{(1)us} X_{(1)bw} \beta_h \beta_i \beta_s \beta_w) \tag{154}$$

$$= \sigma_\beta^4 \cdot [n^4 m^2 (p-m)^2] \cdot (1 + o(1)) \tag{155}$$

and

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{156}$$

$$= E[tr(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)})] \tag{157}$$

$$= E(\sum_{c=1}^{n} \sum_{d=1}^{n} \sum_{h=1}^{m} \sum_{k=1}^{p-m} \sum_{l=1}^{n} \sum_{i=1}^{m} \sum_{q=1}^{m} \sum_{r=1}^{n} \sum_{s=1}^{m} \sum_{t=1}^{p-m} \sum_{u=1}^{n} \sum_{w=1}^{m} \sum_{a=1}^{m} \sum_{b=1}^{n} \tag{158}$$

$$Z_{(2)ck} Z_{(1)dq} Z_{(2)dt} Z_{(1)ca} X_{(2)lk} X_{(2)ut} X_{(1)lh} X_{(1)rq} X_{(1)ri} X_{(1)us} X_{(1)ba} X_{(1)bw} \beta_h \beta_i \beta_s \beta_w) \tag{159}$$

$$= \sigma_\beta^4 \cdot n(p-m)[(m^3 - 3m^2 + 2m)(n^2 + 2n) + (m^2 - m)(n^3 - n^2 + 2(c_4 - 1)n + \tag{160}$$

$$4n^2 + 4(c_4 - 1)n + b_4 n^2) + b_4 m n^3] \tag{161}$$

$$= \sigma_\beta^4 \cdot [n^3 m^2 (p-m)(m+n)] \cdot (1 + o(1)) \tag{162}$$

Similarly,

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{163}$$

$$= o[\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2] \tag{164}$$

Also

$$E(\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{165}$$

$$= E(\beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}) \tag{166}$$

$$= 0 \tag{167}$$

It follows that

$$Var([\beta_{(1)}^T X_{(1)}^T X_{(1)} Z_{(1)}^T + \beta_{(1)}^T X_{(1)}^T X_{(2)} Z_{(2)}^T][Z_{(1)} X_{(1)}^T X_{(1)} \beta_{(1)} + Z_{(2)} X_{(2)}^T X_{(1)} \beta_{(1)}]) \tag{168}$$

$$= o[\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2] \tag{169}$$

The results of different sample sizes can be similarly derived and are ignored. Without loss of generality, we set $\sigma_\beta^2 = 1$ for simplicity in later steps.

5.2.2 Threshold-PRS

First moment of covariance term

$$C_1 = \beta_{(1)}^T Z_{(1)}^T [Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}] \tag{170}$$

$$= \beta_{(1)}^T Z_{(1)}^T Z_{(11)} X_{(11)}^T X_{(1)} \beta_{(1)} + \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)} \tag{171}$$

$$= \beta_{(1)}^T Z_{(1)}^T [Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)} + Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)} + Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}] \tag{172}$$

$$= \beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)} + \beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)} \tag{173}$$

$$+ \beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)} + \beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)} \tag{174}$$

$$+ \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)} \tag{175}$$

$$= C_{11} + C_{12} + C_{13} + C_{14} + C_{15} \tag{176}$$

$$\text{Thus } E(C_1) = E(C_{11}) = E(\beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)}) = n^2 q_1 \tag{177}$$

First moment of variance term II

$$VAR_2 = \tag{178}$$

$$[\beta^T_{(11)}X^T_{(11)}X_{(11)}Z^T_{(11)} + \beta^T_{(12)}X^T_{(12)}X_{(11)}Z^T_{(11)} + \beta^T_{(11)}X^T_{(11)}X_{(21)}Z^T_{(21)} + \beta^T_{(12)}X^T_{(12)}X_{(21)}Z^T_{(21)}]\cdot \tag{179}$$

$$[Z_{(11)}X^T_{(11)}X_{(11)}\beta_{(11)} + Z_{(11)}X^T_{(11)}X_{(12)}\beta_{(12)} + Z_{(21)}X^T_{(21)}X_{(11)}\beta_{(11)} + Z_{(21)}X^T_{(21)}X_{(12)}\beta_{(12)}] \tag{180}$$

$$= \beta^T_{(11)}X^T_{(11)}X_{(11)}Z^T_{(11)}Z_{(11)}X^T_{(11)}X_{(11)}\beta_{(11)} + \beta^T_{(11)}X^T_{(11)}X_{(11)}Z^T_{(11)}Z_{(11)}X^T_{(11)}X_{(12)}\beta_{(12)} \tag{181}$$

$$+ \beta^T_{(11)}X^T_{(11)}X_{(11)}Z^T_{(11)}Z_{(21)}X^T_{(21)}X_{(11)}\beta_{(11)} + \beta^T_{(11)}X^T_{(11)}X_{(11)}Z^T_{(11)}Z_{(21)}X^T_{(21)}X_{(12)}\beta_{(12)} \tag{182}$$

$$+ \beta^T_{(12)}X^T_{(12)}X_{(11)}Z^T_{(11)}Z_{(11)}X^T_{(11)}X_{(11)}\beta_{(11)} + \beta^T_{(12)}X^T_{(12)}X_{(11)}Z^T_{(11)}Z_{(11)}X^T_{(11)}X_{(12)}\beta_{(12)} \tag{183}$$

$$+ \beta^T_{(12)}X^T_{(12)}X_{(11)}Z^T_{(11)}Z_{(21)}X^T_{(21)}X_{(11)}\beta_{(11)} + \beta^T_{(12)}X^T_{(12)}X_{(11)}Z^T_{(11)}Z_{(21)}X^T_{(21)}X_{(12)}\beta_{(12)} \tag{184}$$

$$+ \beta^T_{(11)}X^T_{(11)}X_{(21)}Z^T_{(21)}Z_{(11)}X^T_{(11)}X_{(11)}\beta_{(11)} + \beta^T_{(11)}X^T_{(11)}X_{(21)}Z^T_{(21)}Z_{(11)}X^T_{(11)}X_{(12)}\beta_{(12)} \tag{185}$$

$$+ \beta^T_{(11)}X^T_{(11)}X_{(21)}Z^T_{(21)}Z_{(21)}X^T_{(21)}X_{(11)}\beta_{(11)} + \beta^T_{(11)}X^T_{(11)}X_{(21)}Z^T_{(21)}Z_{(21)}X^T_{(21)}X_{(12)}\beta_{(12)} \tag{186}$$

$$+ \beta^T_{(12)}X^T_{(12)}X_{(21)}Z^T_{(21)}Z_{(11)}X^T_{(11)}X_{(11)}\beta_{(11)} + \beta^T_{(12)}X^T_{(12)}X_{(21)}Z^T_{(21)}Z_{(11)}X^T_{(11)}X_{(12)}\beta_{(12)} \tag{187}$$

$$+ \beta^T_{(12)}X^T_{(12)}X_{(21)}Z^T_{(21)}Z_{(21)}X^T_{(21)}X_{(11)}\beta_{(11)} + \beta^T_{(12)}X^T_{(12)}X_{(21)}Z^T_{(21)}Z_{(21)}X^T_{(21)}X_{(12)}\beta_{(12)} \tag{188}$$

$$= a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2 \tag{189}$$

$$E(VAR_2) \tag{190}$$

$$= E(a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2) \tag{191}$$

$$= E(a^2 + b^2 + c^2 + d^2) \tag{192}$$

$$= [n^2q_1(n + q_1)] \cdot (1 + o(1)) + n^2(m - q_1)q_1 + n^2q_1q_2 + n^2(m - q_1)q_2 \tag{193}$$

$$= [n^2q_1(n + m + q_2) + n^2q_2(m - q_1)] \cdot (1 + o(1)) \tag{194}$$

$$= [n^2q_1(n + m) + n^2q_2m] \cdot (1 + o(1)) \tag{195}$$

Second moment of covariance term

$$E(C_1C_1) = E[(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})] \tag{196}$$
$$= E[C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2] \tag{197}$$

$$E(C_{11}^2) = E(\beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)} \beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)}) \tag{198}$$
$$= n^4 q_1^2 + [(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1 (b_4 - 1)] \cdot (1 + o(1)) \tag{199}$$

$$E(C_{12}^2) = E(\beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)} \beta_{(11)}^T Z_{(11)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)}) \tag{200}$$
$$= E(\sum_{c=1}^{n} \sum_{d=1}^{n} \sum_{i=1}^{m-q_1} \sum_{j=1}^{m-q_1} \sum_{k=1}^{q_1} \sum_{l=1}^{n} \sum_{q=1}^{q_1} \sum_{r=1}^{q_1} \sum_{s=1}^{n} \sum_{t=1}^{q_1} \tag{201}$$
$$Z_{(11)ck} Z_{(11)dq} Z_{(11)dr} Z_{(11)ct} X_{(11)lk} X_{(11)sr} X_{(12)li} X_{(12)sj} \beta_{(11)q} \beta_{(11)t} \beta_{(12)i} \beta_{(12)j}) \tag{202}$$
$$= [n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)] \cdot (1 + o(1)) \tag{203}$$

Similarly

$$E(C_{13}^2) = E(\beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)} \beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(11)} \beta_{(11)}) \tag{204}$$
$$= [n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)] \cdot (1 + o(1)) = E(C_{12}^2) \tag{205}$$

$$E(C_{14}^2) = E(\beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)} \beta_{(12)}^T Z_{(12)}^T Z_{(11)} X_{(11)}^T X_{(12)} \beta_{(12)}) \tag{206}$$
$$= n^2 (m - q_1)^2 q_1 \tag{207}$$

$$E(C_{15}^2) = E(\beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)} \beta_{(1)}^T Z_{(1)}^T Z_{(21)} X_{(21)}^T X_{(1)} \beta_{(1)}) = n^2 m^2 q_2 \tag{208}$$

Thus

$$E(C_1C_1) = E[C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2] \tag{209}$$
$$= n^4 q_1^2 + [(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1 (b_4 - 1)] \cdot (1 + o(1)) \tag{210}$$
$$+ 2 \cdot [n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)] \cdot (1 + o(1)) \tag{211}$$
$$+ n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2 \tag{212}$$

It follows that

$$Var(C_1) = [n^2 q_1^3 + 4n^3 q_1^2 + n^4 q_1 (b_4 - 1) + 2n^2 q_1^2 (m - q_1) + \tag{213}$$
$$2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2] \cdot (1 + o(1)) \tag{214}$$

Second moment of variance term II

$$E(VAR_2VAR_2) \tag{215}$$

$$= E[(a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2)^2] \tag{216}$$

$$= E[(a^4 + b^4 + c^4 + d^4 + 2a^2b^2 + 2a^2c^2 + 2a^2d^2 + 2b^2c^2 + 2b^2d^2 + 2c^2d^2] \tag{217}$$

$$= E(a^4) + E(b^4) + E(c^4) + E(d^4) + E(2a^2b^2 + 2a^2c^2 + 2a^2d^2 + 2b^2c^2 + 2b^2d^2 + 2c^2d^2) \tag{218}$$

$$= [n^2q_1(n + q_1)]^2 + o(n^4q_1^4 + n^5q_1^3 + n^6q_1^2) + [n^2(m - q_1)q_1]^2 \cdot (1 + o(1)) \tag{219}$$

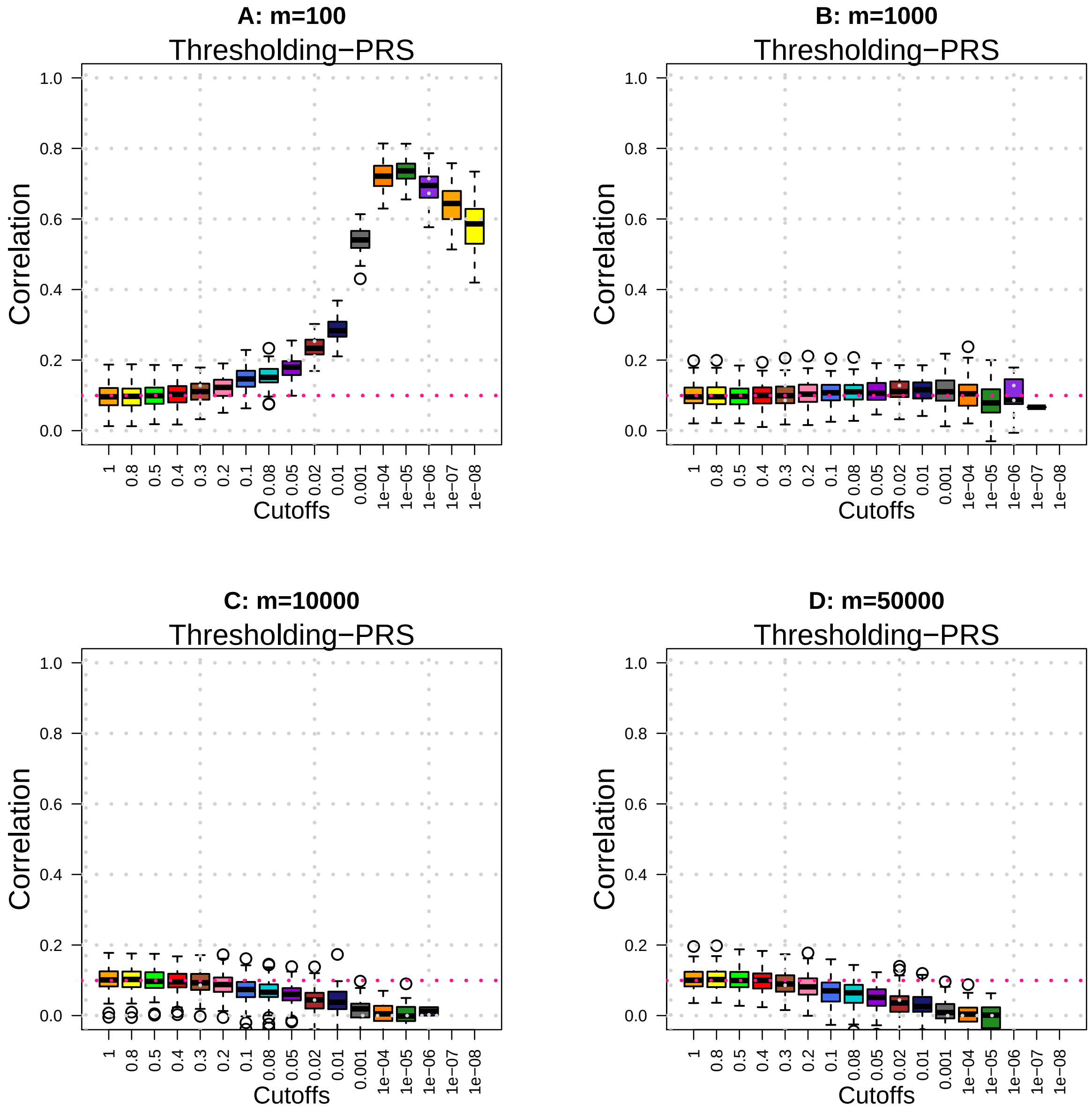
$$+ (n^2q_1q_2)^2 \cdot (1 + o(1)) + [n^2(m - q_1)q_2]^2 \cdot (1 + o(1)) \tag{220}$$

$$+ o[(n^2q_1(n + m) + n^2q_2m)^2] \tag{221}$$

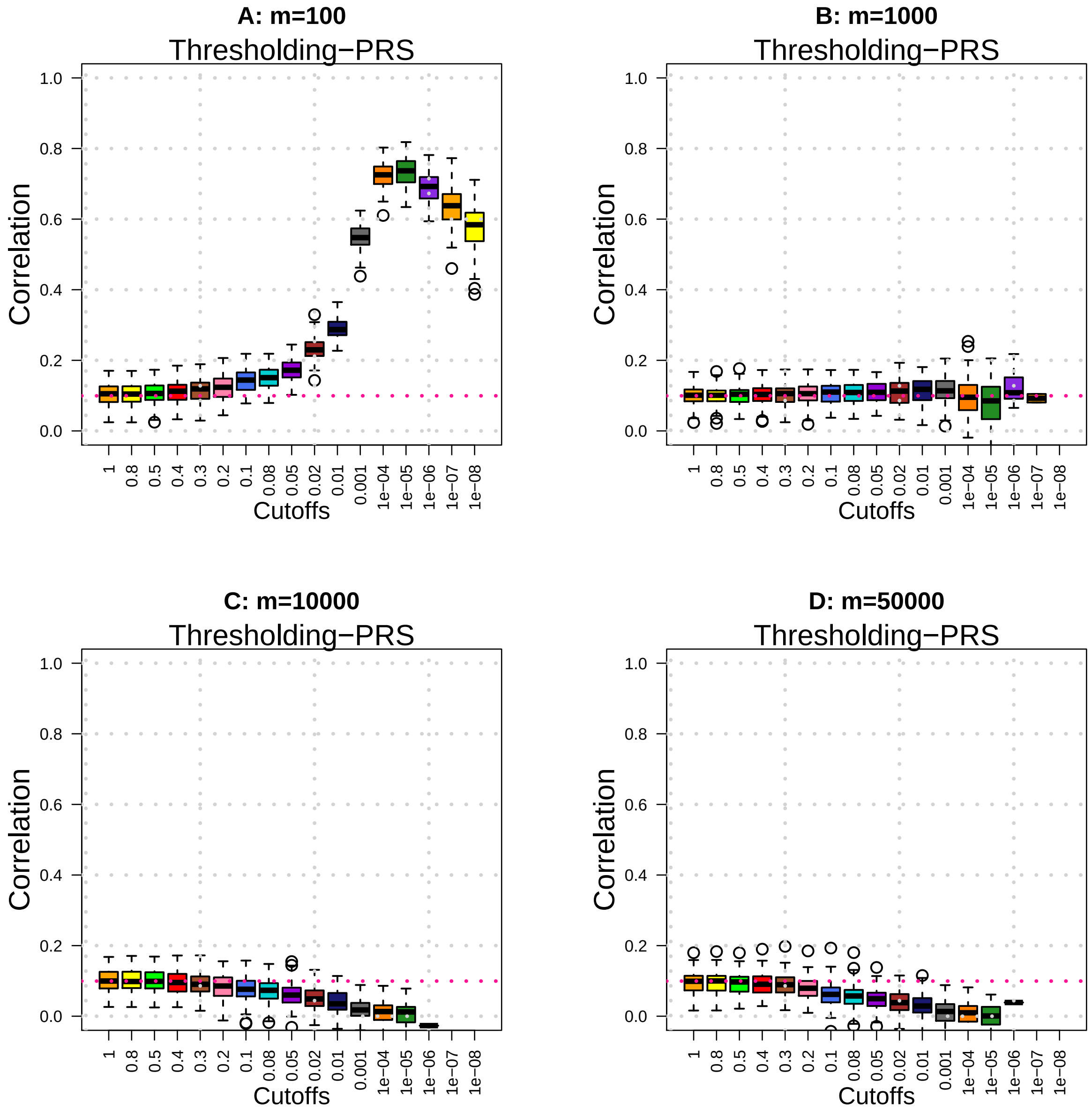
It follows that

$$Var(VAR_2) = o([n^2q_1(n + m) + n^2q_2m]^2) \tag{222}$$

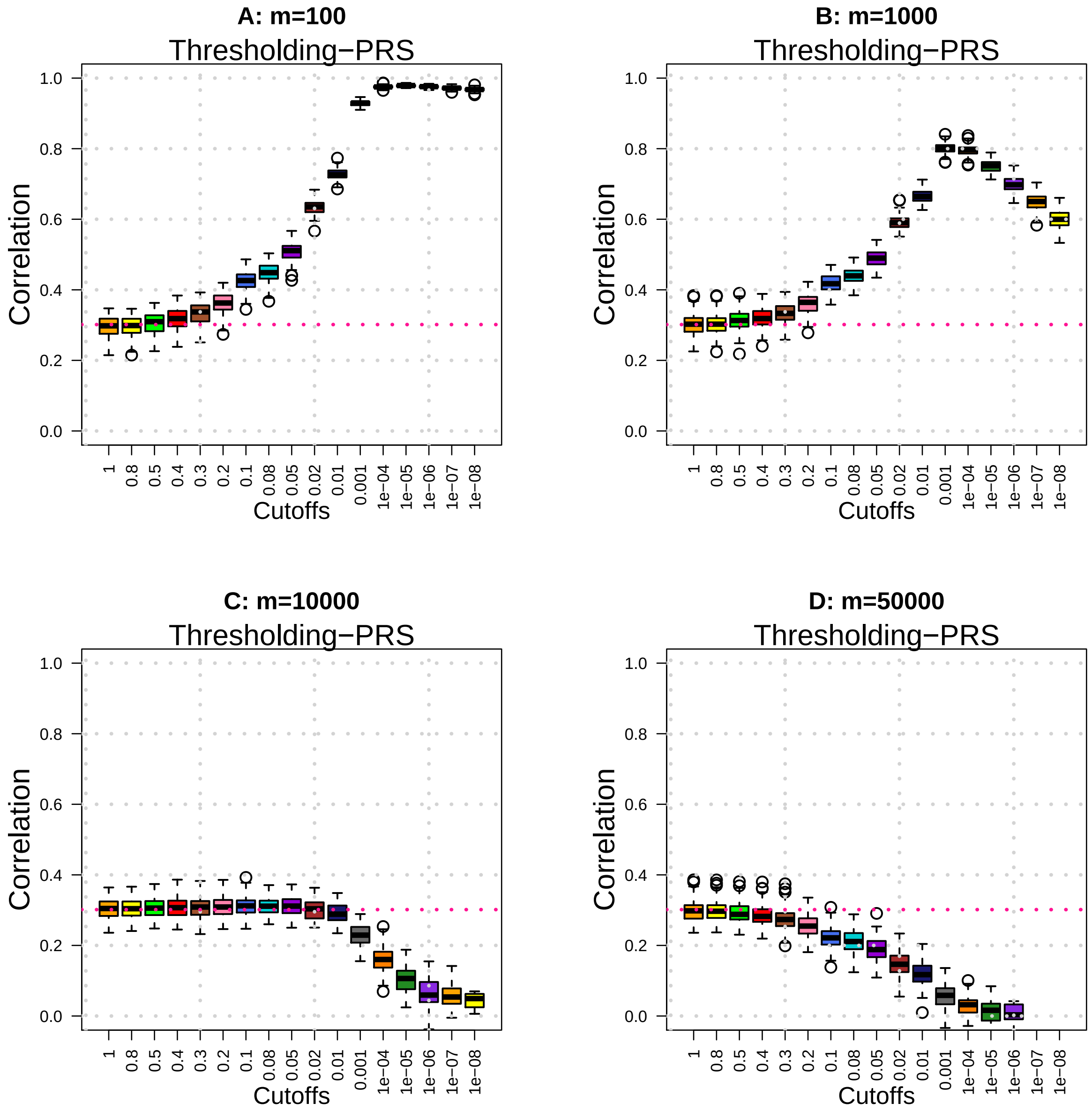
5.3 Supplementary figures



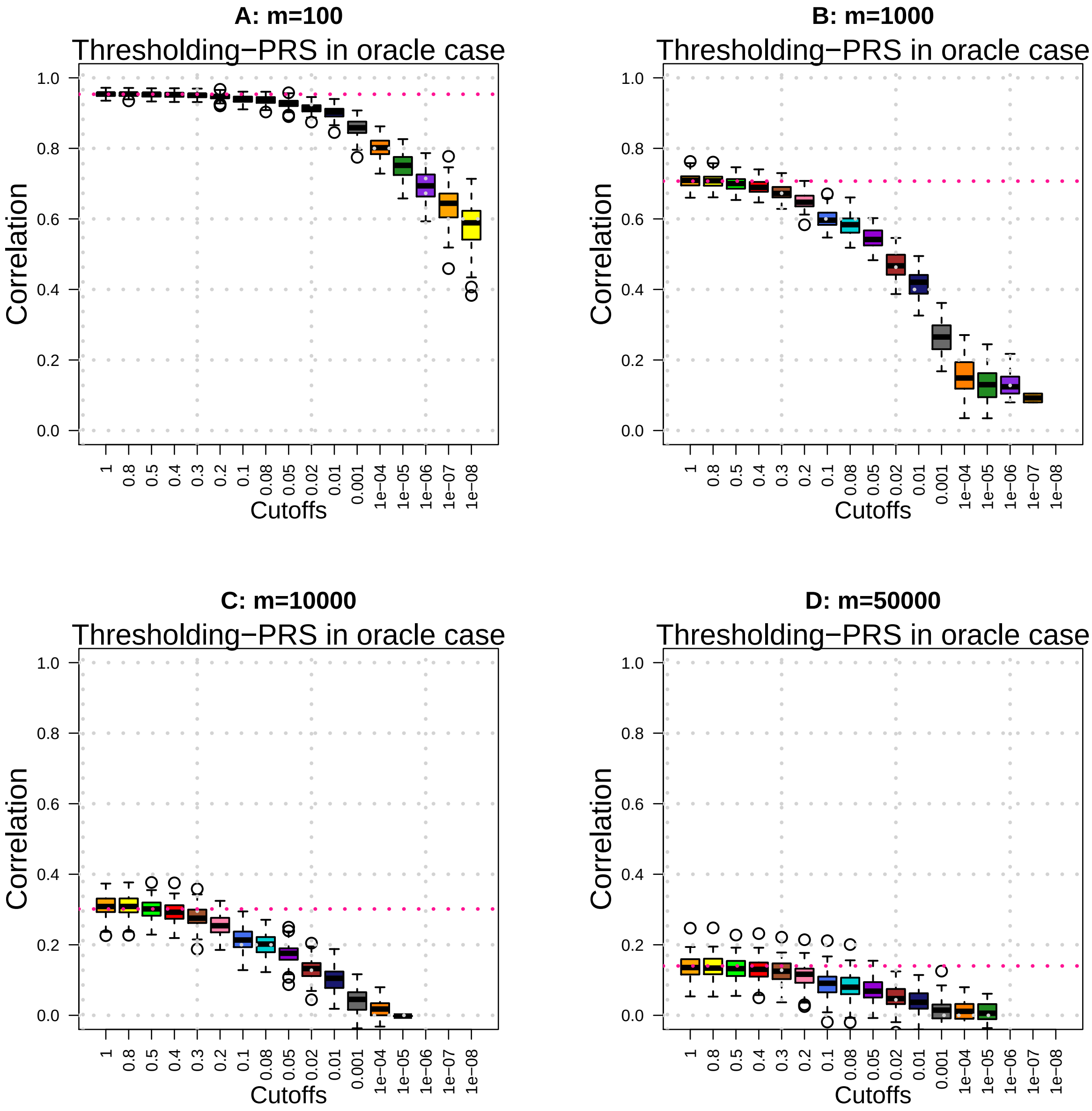
Supplementary Fig. 1: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set p=100000, n=1000 in both training and testing data.



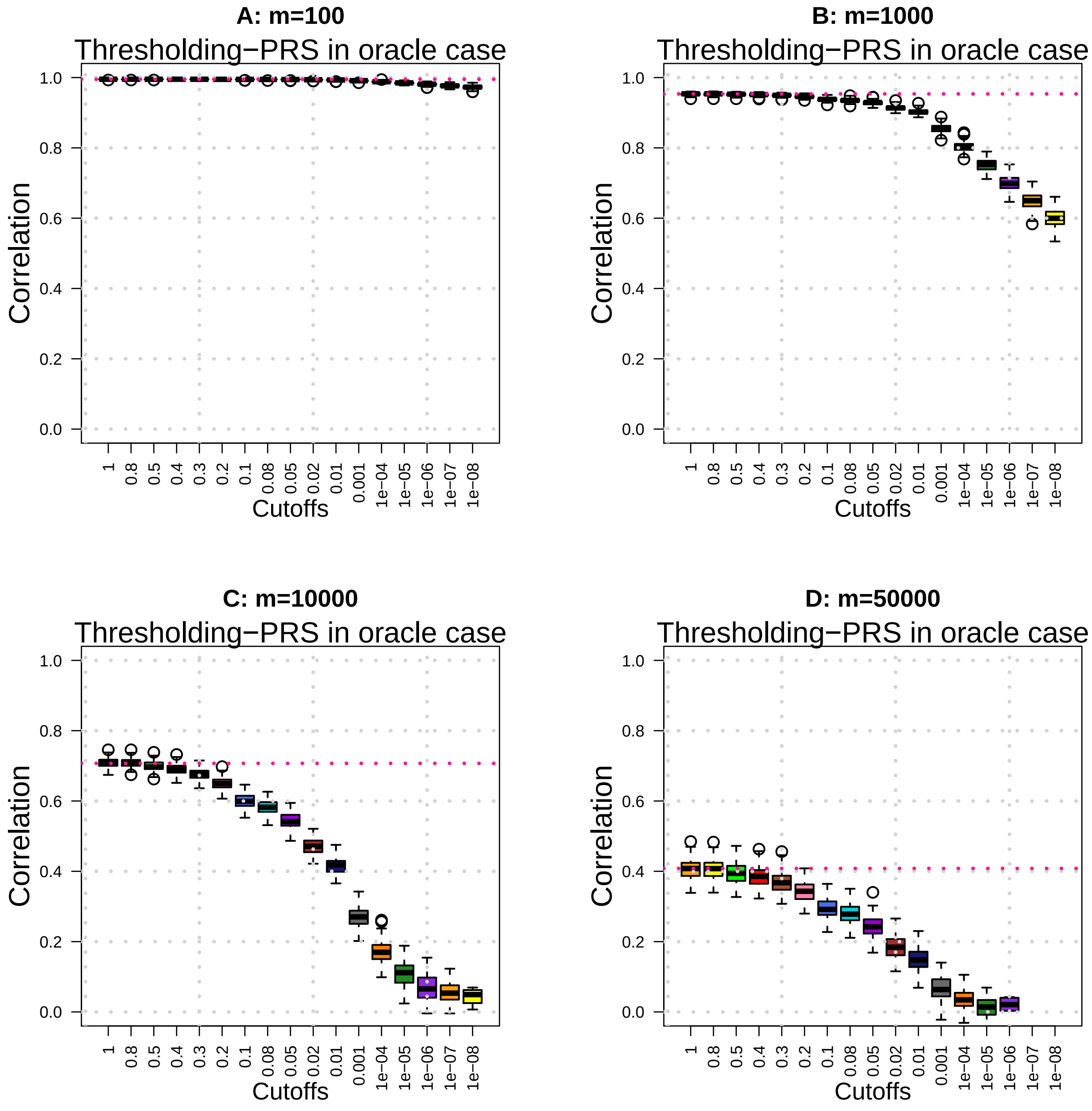
Supplementary Fig. 2: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=1000 in both training and testing data.



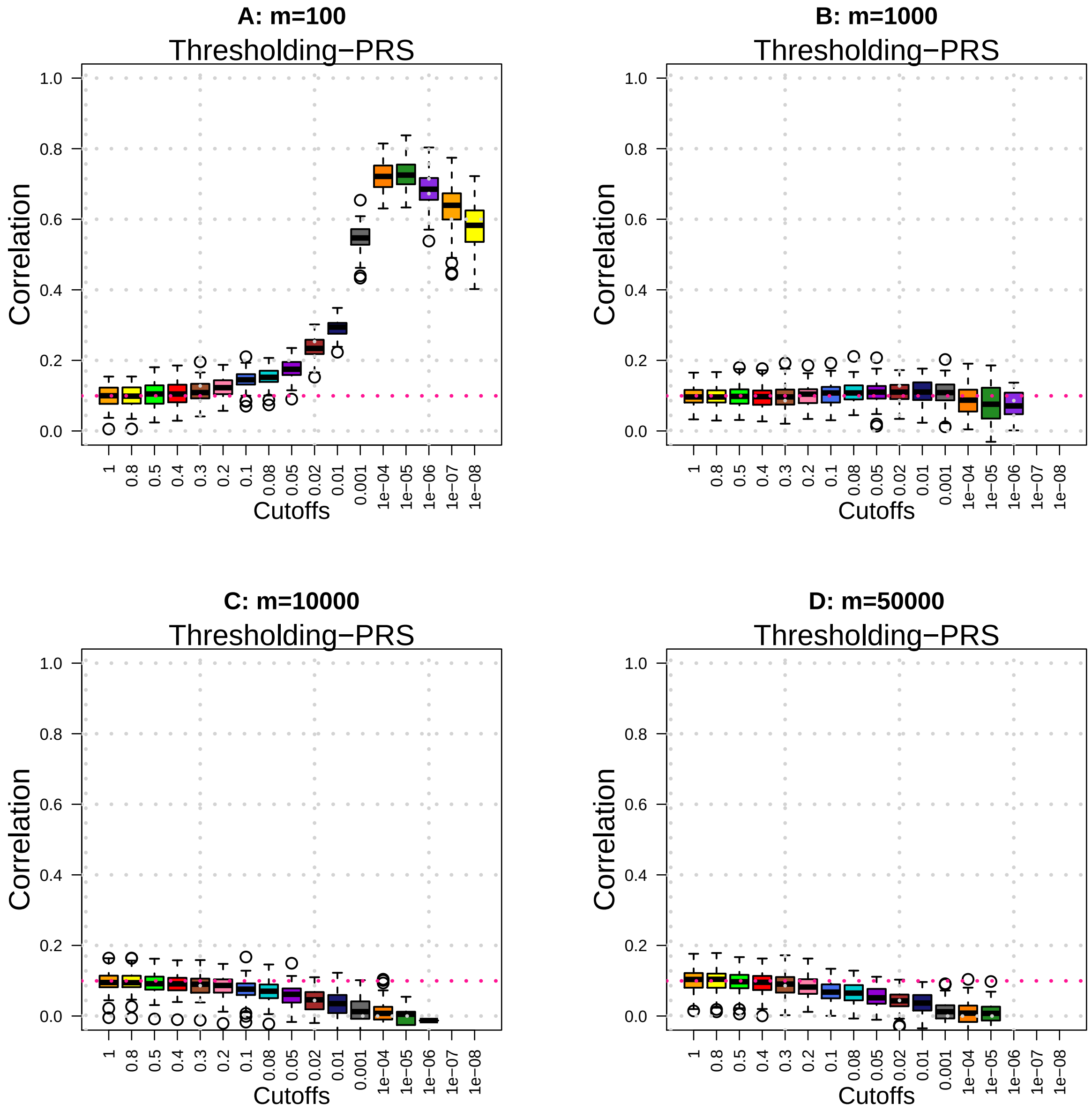
Supplementary Fig. 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=10000 in training data and $n = 1000$ in testing data.



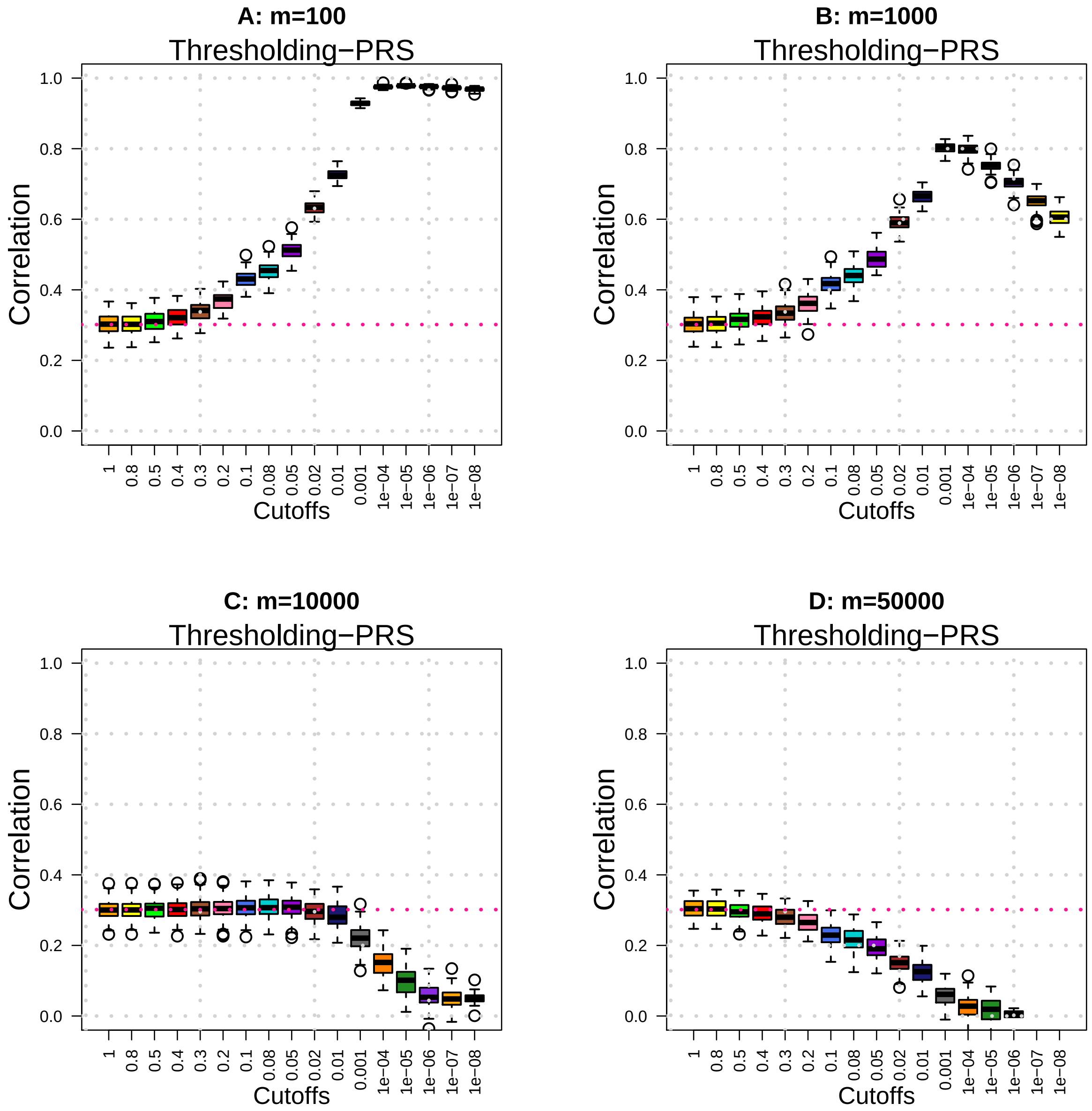
Supplementary Fig. 4: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=1000 in both training and testing data.



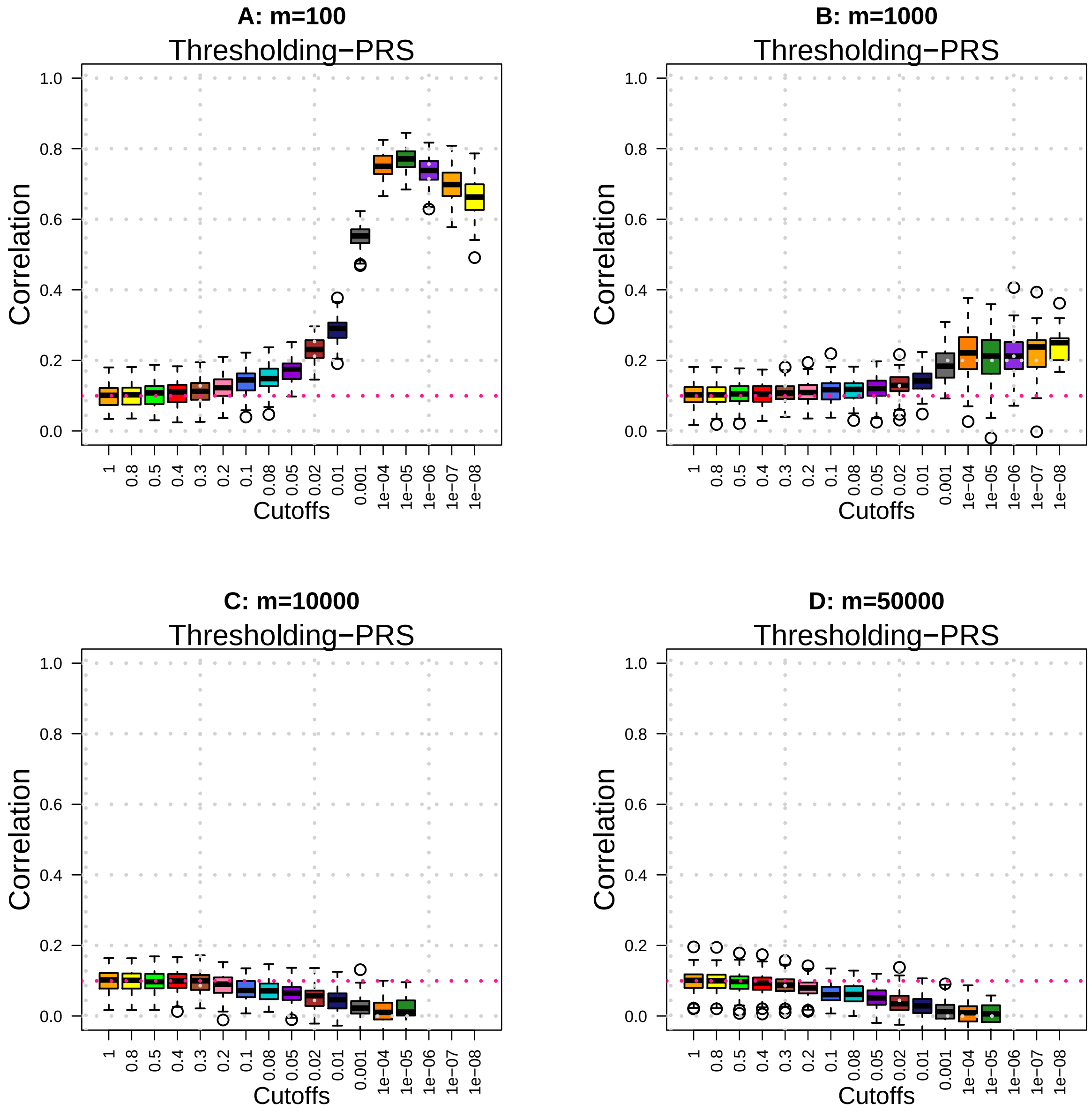
Supplementary Fig. 5: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=10000 in training data and $n = 1000$ in testing data.



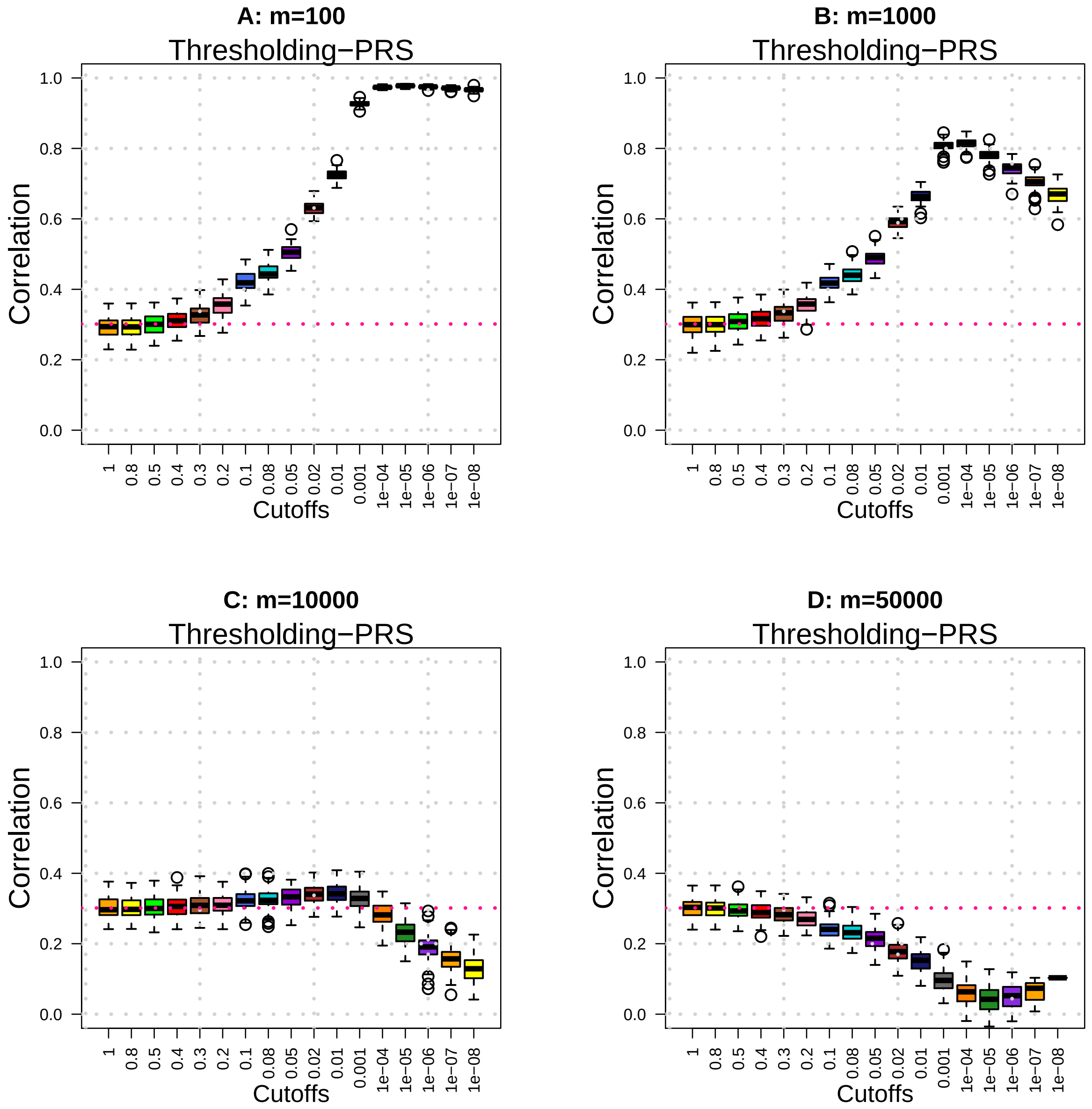
Supplementary Fig. 6: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=1000 in both training and testing data.



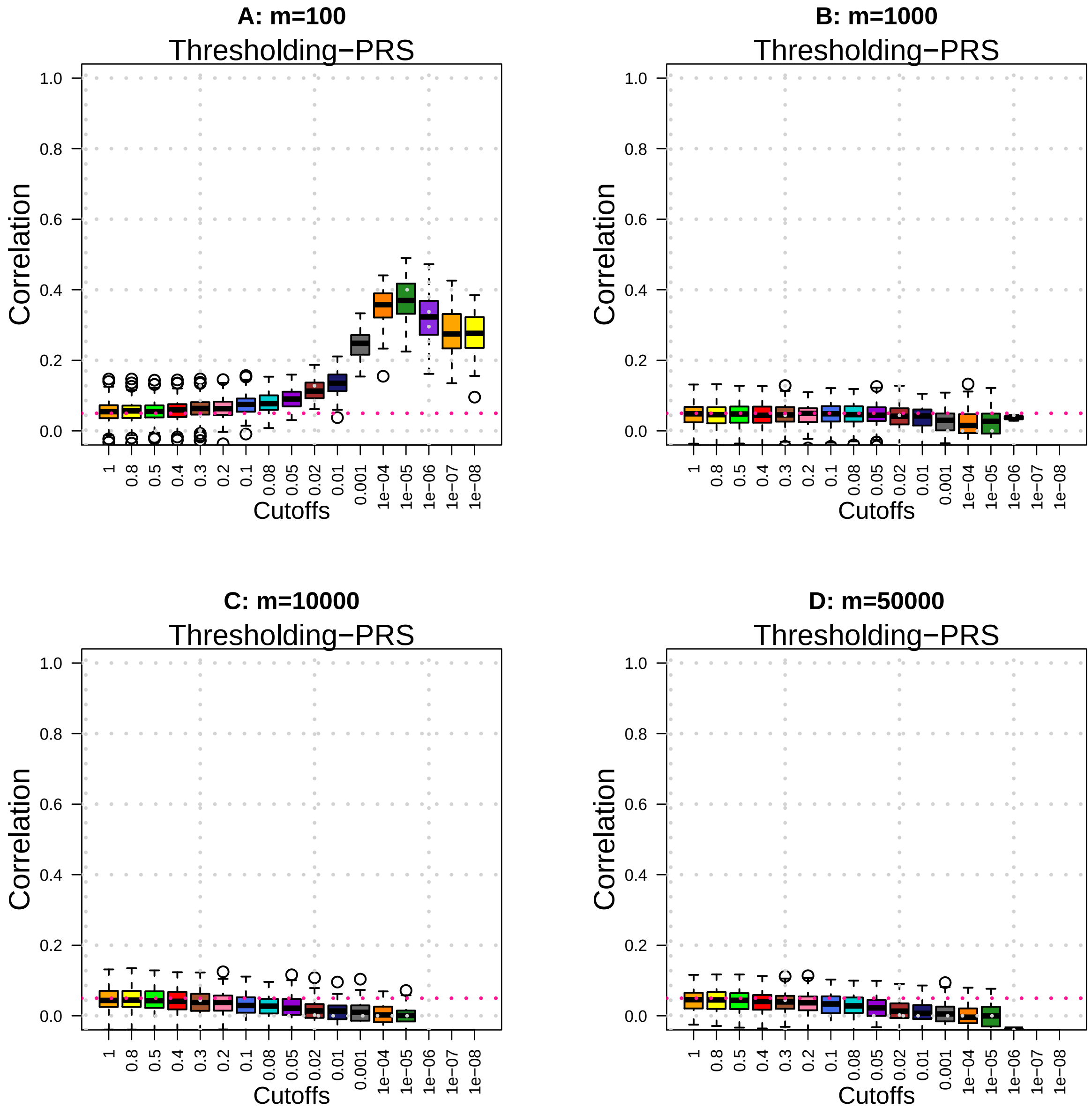
Supplementary Fig. 7: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=10000 in training data, and $n = 1000$ in testing data.



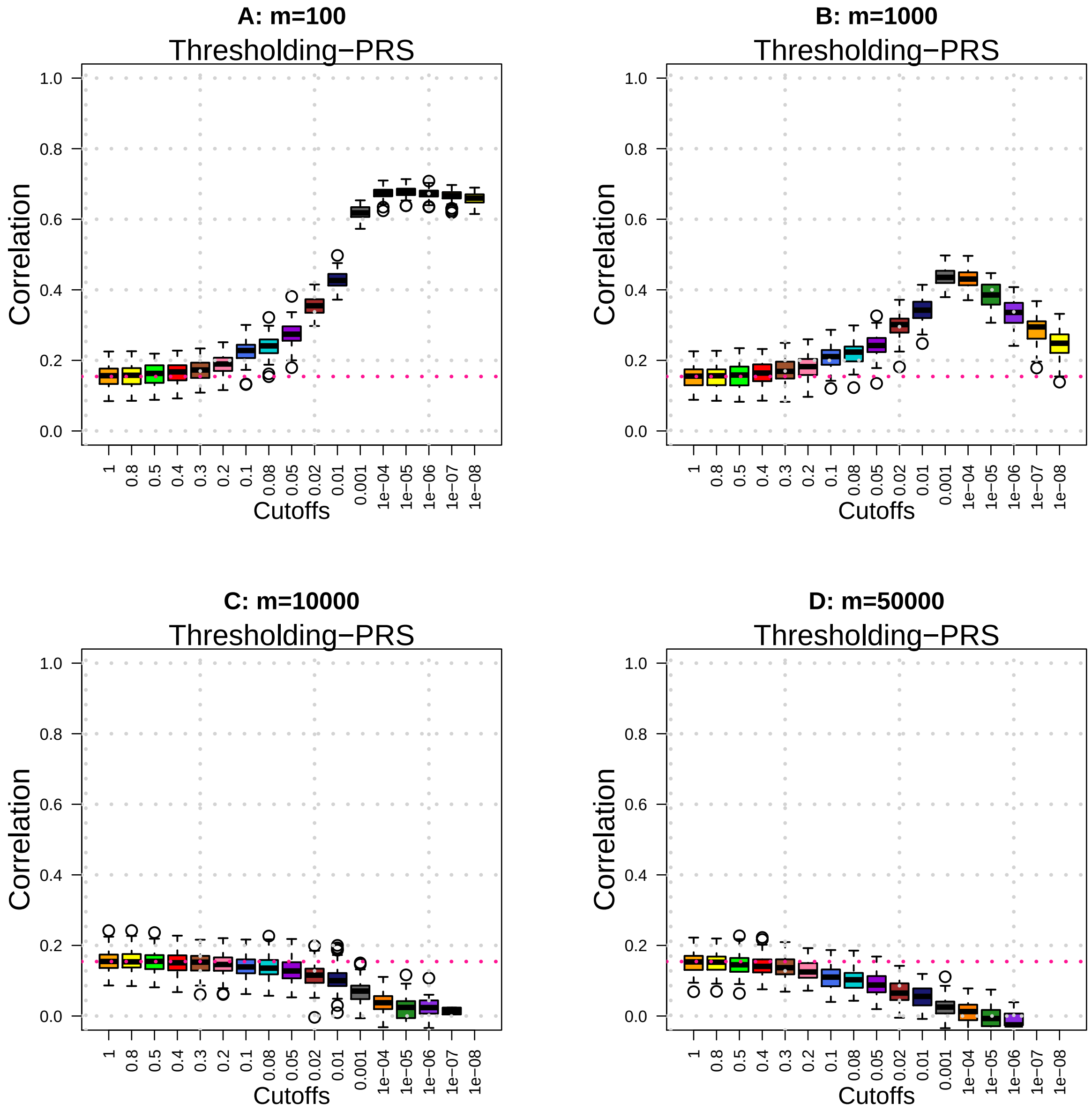
Supplementary Fig. 8: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=1000 in both training and testing data.



Supplementary Fig. 9: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=10000 in training data, and $n = 1000$ in testing data.



Supplementary Fig. 10: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the heritability h^2=0.5. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=1000 in both training and testing data.



Supplementary Fig. 11: Prediction accuracy (A_P) of threshold-PRS across different m/p when the heritability h^2=0.5. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100000, n=10000 in training data, and $n = 1000$ in testing data.