Is PRS Good for Predicting Complex Polygenic Traits?

Bingxin Zhao[1] Fei Zou[1,2*]

Abstract

Polygenic risk score (PRS) is extremely popular for complex traits prediction with single-nucleotide polymorphisms (SNPs) data. The PRS, as its name suggests, is designed for polygenic traits to aggregate information from a large number of potentially causal SNPs. It has been viewed as a powerful method for polygenic traits within the genetics community. However, disappointing to many researchers, the prediction accuracy of PRS in practice is low, even for highly heritable polygenic traits. To solve the puzzle, we investigate PRS both empirically and theoretically. We show how the performance of PRS is influenced by the triplet (n, p, m), where n, p, m are the sample size, the number of SNPs studied, and the number of true causal SNPs, respectively. For a given heritability, we find that i) when PRS is constructed with all p SNPs (referred as GWAS-PRS), its prediction accuracy is controlled by the p/n ratio; while ii) when PRS is built with the top-ranked SNPs that pass a pre-specified P-value threshold (referred as threshold-PRS), its accuracy can vary dramatically depending on how sparse the 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 $m >> n$, or genetic signals are not sparse, which is often the case for complex polygenic traits, threshold-PRS can fail badly. Our results demystify the poor performance of PRS and demonstrate the practical paradox associated with PRS for polygenic trait prediction.

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
[2]Version: November 13, 2018. ***Correspondence: feizou@email.unc.edu**

Introduction

With the rapid development in biomedical technologies, various types of large-scale genetics and genomics data have been collected for better understanding of genetic etiologies underlying complex human diseases and traits. Genome-wide association studies (GWAS) aim to examine association between complex traits and single-nucleotide polymorphisms (SNPs). To detect SNPs that are associated with a given phenotype, GWAS commonly perform single SNP analysis to estimate and test the association between the phenotype and each candidate SNP one at a time, while adjusting for the effects of non-genetic factors and population substructures.[1] Tens of thousands of statistically significant SNPs have been detected for hundreds of human diseases/traits through GWAS.[2,3] 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,[4] resulting in a so called "missing heritability" phenomenon.[5,6]

One explanation for the missing heritability is that most complex traits are polygenic, affected by many genes whose individual effects are small.[7] The polygenicity of traits has long been hypothesized[8–10] and is supported by increasing empirical evidence.[11–18] Given the polygenicity nature of complex traits, it has been hypothesized that though many SNPs are not statistically significant, they can be causal and should be used for polygenic trait prediction. For these reasons, Purcell et al.[19] propose a polygenic risk score (PRS), which is a weighted sum of top-ranked candidate SNPs where each SNP is weighted by its estimated marginal effect from a discovery GWAS. As its name suggests, PRS aims to aggregate genetic effects of polygenes. Thus, it is expected to be powerful for polygenic traits, and would be more true for omnigenic traits.[20] The omnigenic model is a newly emerging model for traits affected by the majority (if not all) of candidate SNPs.

Although PRS has been widely used in neuropsychiatric diseases/disorders, such as bipolar and schizophrenia,[21,22] the prediction power of PRS remains disappointingly

low with little clinical utility, even for traits with known high heritability.[23–25] Two legitimate reasons for such poor performance include 1) low quality SNP arrays with low coverage of causal SNPs; and 2) low quality top-ranked SNPs in tagging causal SNPs.[26,27] 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 few studies,[28–32] little 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 with a hope to clear some misperceptions on PRS and to provide some practical guidelines on PRS. Since PRS is built upon the marginal SNP effect estimates, we start our investigation from the statistical properties of the estimates. Note for polygenic traits, the single SNP analysis is always misspecified since the effects of the other causal SNPs are not modeled but lumped into the random error. When all causal SNPs are independent of each other, such model misspecification is in general fine for traits with sparse genetic signals, but it can fail badly for traits with dense genetic signals. For a given SNP, the omitted causal SNPs can greatly increase the uncertainty in its marginal effect estimate and make its estimate unreliable. As will be illustrated later, even for a fully heritable trait with a 100% genetic heritability, the estimated genetic effects of causal and null SNPs can be totally mixed and nonseparable from each other. 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.[33], 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,[33–39] but are done mostly out of the genetics field. For complex polygenic traits, spurious correlation makes the separation of causal and null SNPs difficult, leading to a doomed failure of PRS, which contradicts the original motive of PRS completely (Figure 1).

In recognition of the relationship between the GWAS marginal screening and PRS,

we show how the asymptotic prediction accuracy of PRS is affected by the triplet of (n, p, m), where n is the sample size of the training dataset, p is the total number of candidate SNPs, and m is the number of causal SNPs. Our investigation on PRS starts from the GWAS-PRS and is generalized to the threshold-PRS. Extensive simulations are carried out to evaluate PRS empirically and to evaluate our theoretical results under finite sample settings.

Material and Methods

Single SNP Analysis

For a training dataset, let $\boldsymbol{y}$ be an $n \times 1$ phenotypic vector, $\boldsymbol{X}_{(1)}$ denote an $n \times m$ matrix of the causal SNPs and $\boldsymbol{X}_{(2)}$ donate an $n \times (p-m)$ matrix of the null SNPs, resulting in an $n \times p$ matrix of all SNPs donated as $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m, \boldsymbol{x}_{m+1}, \cdots, \boldsymbol{x}_p]$. Columns of $\boldsymbol{X}$ are assumed to be independent for simplification. Furthermore, column-wise normalization is performed on $\boldsymbol{X}$ such that each SNP has mean zero and variance one. Define the following condition:

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

The polygenic model assumes the following relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$:

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m, \beta_{m+1}, \cdots, \beta_p)^T$ 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 $\boldsymbol{\beta}_{(1)} = (\beta_1, \cdots, \beta_m)^T$ and $\boldsymbol{\beta}_{(2)}$ be a $(p-m) \times 1$ vector with all elements being zero, and $\boldsymbol{\epsilon}$ represents the random error vector. For simplicity and without loss of generality, we assume that there exist no other covariates. According to the above model, the overall

genetic heritability h^2 of $\boldsymbol{y}$ is

$$h^2 = \frac{\mathrm{Var}(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)})}{\mathrm{Var}(\boldsymbol{y})} = \frac{\mathrm{Var}(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)})}{\mathrm{Var}(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}) + \mathrm{Var}(\boldsymbol{\epsilon})}.$$

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

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i = \sum_{i=1}^{m} \boldsymbol{x}_i \beta_i = \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}, \tag{2}$$

the most optimistic situation in predicting phenotypes.

Though for a typical large-scale discovery GWAS, the sample size n is often not small (e.g., $n \sim 10,000$ or $100,000$), the number of candidate SNPs p is typically even larger (e.g., $p \sim 500,000$ or $10,000,000$). On the other hand, depending on their underlying genetic architectures, the number of causal SNPs can vary dramatically from one trait to another. To cover most of modern GWAS data, we therefore assume $n, p \to \infty$ and that

$$\frac{p}{n} = \gamma \to \gamma_0 \quad \text{and} \quad \frac{m}{p} = \omega \to \omega_0, \quad \text{where} \quad 0 < \gamma_0 \leq \infty, \quad 0 \leq \omega_0 \leq 1.$$

For continuous traits in absence of any covariates, a typical single SNP analysis employs the following simple linear regression model

$$\boldsymbol{y} = \boldsymbol{1_n}\mu + \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon}^*$$

for a given SNP i, where $\boldsymbol{1_n}$ is an $n \times 1$ vector of ones, β_i is its effect ($i = 1, \cdots, p$). When both $\boldsymbol{y}$ and $\boldsymbol{x}_i$ are normalized and $n \to \infty$, under Condition (1) and polygenic model (2), the maximum likelihood estimate (MLE) of μ, $\hat{\mu} \equiv 0$, and the MLE of the genetic effect β_i equals

$$\widehat{\beta}_i = (\boldsymbol{x}_i^T \boldsymbol{x}_i)^{-1} \boldsymbol{x}_i^T \boldsymbol{y} = \frac{1}{n} \boldsymbol{x}_i^T \boldsymbol{y} = \sum_{j=1}^{m} r_{ij} \beta_j,$$

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

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

Given that SNPs in X are independent of each other, or the correlation ρ_{ij} for any ijth SNP pair $(i \ \& \ j)(i \neq j)$ is zero, it can be shown that asymptotically $\widehat{\beta}_i$ is an unbiased estimator of β_i such that

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

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

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}_i) &= \mathrm{Var}\big(\sum_{j \neq i}^{m} r_{ij} \beta_j\big) = \sum_{j \neq i}^{m} \beta_j^2 \cdot \mathrm{Var}(r_{ij}) = \frac{1}{n^2} \sum_{j \neq i}^{m} \beta_j^2 \cdot \mathrm{E}\big(\sum_{k_1=1}^{n} \sum_{k_2=1}^{n} x_{ik_1} x_{jk_1} x_{ik_2} x_{jk_2}\big) \\ &= \frac{1}{n^2} \sum_{j \neq i}^{m} \beta_j^2 \cdot \mathrm{E}\big(\sum_{k_1=1}^{n} x_{ik_1}^2 x_{jk_1}^2\big) = \frac{\sum_{j \neq i}^{m} \beta_j^2}{n} = \mathcal{O}(\frac{m}{n}) = \mathcal{O}(\gamma \cdot \omega). \end{aligned}$$

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

$$\mathrm{Var}(\widehat{\beta}_i) = \mathrm{Var}\big(\sum_{j=1}^{m} r_{ij} \beta_j\big) = \frac{\sum_{j=1}^{m} \beta_j^2}{n} = \mathcal{O}(\frac{m}{n}) = \mathcal{O}(\gamma \cdot \omega).$$

Therefore, $\mathrm{Var}(\widehat{\beta}_i) = \mathcal{O}(m/n) = \mathcal{O}(\gamma \cdot \omega)$ for all i $(i = 1, \cdots, p)$, suggesting that the associated variances of all $\widehat{\beta}_i$s are in the same scale regardless of whether their corresponding SNPs are causal or not. When $m/n = \gamma \cdot \omega \to \gamma_0 \cdot \omega_0$ is large, $\widehat{\beta}_i$ is no longer a reliable estimate of β_i. Moreover, the $\hat{\beta}_i$s of the causal and null SNPs become well mixed and cannot be easily separated when m/n is large, which raises two important concerns include 1) how does this affect the SNP selection; and 2) how does this affect the weights in PRS which ultimately affect the performance of PRS?

We address these two concerns below.

PRS

For a testing dataset with n_z samples, define its $n_z \times p$ SNP matrix as $\boldsymbol{Z} = [\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}]$ with $\boldsymbol{Z}_{(1)} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_m]$ and $\boldsymbol{Z}_{(2)} = [\boldsymbol{z}_{m+1}, \cdots, \boldsymbol{z}_p]$. Then $\boldsymbol{y}_z = \sum_{i=1}^{p} \boldsymbol{z}_i \beta_i = \boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}$ and the PRS equals

$$\widehat{\boldsymbol{y}}_P = \sum_{i=1}^{p} \boldsymbol{z}_i \widehat{d}_i = \boldsymbol{Z}\widehat{\boldsymbol{d}} = \boldsymbol{Z}_{(1)}\widehat{\boldsymbol{d}}_{(1)} + \boldsymbol{Z}_{(2)}\widehat{\boldsymbol{d}}_{(2)},$$

where $\widehat{\boldsymbol{d}} = (\widehat{d}_1, \cdots, \widehat{d}_m, \widehat{d}_{m+1}, \cdots, \widehat{d}_p)^T = \big(\widehat{\boldsymbol{d}}_{(1)}^T, \widehat{\boldsymbol{d}}_{(2)}^T\big)$, $\widehat{d}_i = \widehat{\beta}_i \cdot \mathrm{I}(|\widehat{\beta}_i| > c)$, $\mathrm{I}(\cdot)$ 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 typically measured by

$$A_P = \frac{\boldsymbol{y}_z^T \widehat{\boldsymbol{y}}_P}{\|\boldsymbol{y}_z\| \cdot \|\widehat{\boldsymbol{y}}_P\|} = \frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\big)^T \big(\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{d}}_{(1)} + \boldsymbol{Z}_{(2)}\widehat{\boldsymbol{d}}_{(2)}\big)}{\|\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\| \cdot \|\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{d}}_{(1)} + \boldsymbol{Z}_{(2)}\widehat{\boldsymbol{d}}_{(2)}\|},$$

where $\|\cdot\|$ is the l_2 norm of a vector.

GWAS-PRS

In GWAS-PRS, $\widehat{\boldsymbol{d}}_{(1)} = \widehat{\boldsymbol{\beta}}_{(1)}$ and $\widehat{\boldsymbol{d}}_{(2)} = \widehat{\boldsymbol{\beta}}_{(2)}$. For simplification, in rest of the paper, we set $n_z = n$ and our general conclusions should remain the same when the two are different. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_m, \widehat{\beta}_{m+1}, \cdots, \widehat{\beta}_p)^T = \big(\widehat{\boldsymbol{\beta}}_{(1)}^T, \widehat{\boldsymbol{\beta}}_{(2)}^T\big)$, where $\widehat{\boldsymbol{\beta}}_{(1)} = \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}/n$, and $\widehat{\boldsymbol{\beta}}_{(2)} = \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}/n$. Then

$$A_P = \frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \big(\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{\beta}}_{(1)} + \boldsymbol{Z}_{(2)}\widehat{\boldsymbol{\beta}}_{(2)}\big)}{\|\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\| \cdot \|\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{\beta}}_{(1)} + \boldsymbol{Z}_{(2)}\widehat{\boldsymbol{\beta}}_{(2)}\|} = \frac{C_1}{\{\mathrm{Var}_1\}^{1/2}\{\mathrm{Var}_2\}^{1/2}}, \text{ where}$$

$$C_1 = \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)},$$

$$\mathrm{Var}_1 = \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}, \quad \text{and}$$

$$\mathrm{Var}_2 = \big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big).$$

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

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

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

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

as $n \to \infty$.

Proof of Theorem 1 is given in Appendix A.

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, a poorly selected SNP list can be a reason for the poor performance of PRS. However, Theorem 1 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, its performance is limited by the ratio of m/n. For complex traits underlying the omnigenic model,[20] the expected prediction power of the oracle PRS is essentially zero.

Remark 3. In our investigation, we set $h^2 = 1$ to reflect the most optimistic situation for phenotype prediction. If $h^2 < 1$, the prediction power of GWAS-PRS is expected to get further reduced.

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 dataset. 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 for us to turn around and ask with a properly selected threshold c, whether the performance of PRS can be improved or not.

General Setup

For a given threshold $c > 0$, let $q = p \cdot a$ ($a \in (0, 1]$), where q is the number of selected SNPs, among which there are q_1 true causal SNPs and the remaining q_2 are null SNPs, that is, $q = q_1 + q_2$. Let $\boldsymbol{Z}_{(1)} = [\boldsymbol{Z}_{(11)}, \boldsymbol{Z}_{(12)}]$, $\boldsymbol{Z}_{(2)} = [\boldsymbol{Z}_{(21)}, \boldsymbol{Z}_{(22)}]$, $\boldsymbol{X}_{(1)} = [\boldsymbol{X}_{(11)}, \boldsymbol{X}_{(12)}]$, $\boldsymbol{X}_{(2)} = [\boldsymbol{X}_{(21)}, \boldsymbol{X}_{(22)}]$, $\widehat{\boldsymbol{\beta}}_{(1)}^T = \big(\widehat{\boldsymbol{\beta}}_{(11)}^T, \widehat{\boldsymbol{\beta}}_{(12)}^T\big)$, and $\widehat{\boldsymbol{\beta}}_{(2)}^T = \big(\widehat{\boldsymbol{\beta}}_{(21)}^T, \widehat{\boldsymbol{\beta}}_{(22)}^T\big)$, where $\boldsymbol{Z}_{(11)}$, $\boldsymbol{X}_{(11)}$, and $\widehat{\boldsymbol{\beta}}_{(11)}$ correspond to the selected q_1 causal SNPs, and $\boldsymbol{Z}_{(21)}$, $\boldsymbol{X}_{(21)}$, and $\widehat{\boldsymbol{\beta}}_{(21)}$ correspond to the selected q_2 null SNPs. The prediction accuracy of threshold-PRS is therefore

$$A_P = \frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \big(\boldsymbol{Z}_{(11)}\widehat{\boldsymbol{\beta}}_{(11)} + \boldsymbol{Z}_{(21)}\widehat{\boldsymbol{\beta}}_{(21)}\big)}{\big\|\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\big\| \cdot \big\|\boldsymbol{Z}_{(11)}\widehat{\boldsymbol{\beta}}_{(11)} + \boldsymbol{Z}_{(21)}\widehat{\boldsymbol{\beta}}_{(21)}\big\|} = \frac{C_1}{\{\mathrm{Var}_1\}^{1/2}\{\mathrm{Var}_2\}^{1/2}},$$

where

$$C_1 = \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)},$$

$$\mathrm{Var}_1 = \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}, \quad \text{and}$$

$$\mathrm{Var}_2 = \big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T\big)\big(\boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big).$$

Theorem 2. Under the polygenic model (2) and Condition (1), if $m, q_1, q_2 \to \infty$ as $\min(n, p) \to \infty$, and further if $\{m^2(q_1 + q_2)\}/(n^2 q_1^2) \to 0$, then we have

$$A_P^2 / \Big(\frac{n q_1^2}{n m q_1 + q m^2}\Big) = 1 + o_p(1).$$

However, if $\{m^2(q_1+q_2)\}/(n^2q_1^2) \not\to 0$ as $\min(n,p) \to \infty$, then

$$A_P^2 = \mathcal{O}_p(n^{-1}) = o_p(1).$$

Proof of Theorem 2 is given in Appendix A. Theorem 2 shows that given n and m, A_P is determined by q_1, the number of selected causal SNPs, and q, the total number of selected SNPs. Expressing q_1 as a function of q, or $q_1 = \phi(q)$, then A_P can be re-expressed as a function of q:

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

Role of $\phi(q)$

The function $\phi(q)$ is a non-decreasing function of q and plays an important role in determining A_P. The exact form of $\phi(q)$ is trait-dependent and not easy to obtain. But for the following two special examples, we can evaluate $\phi(q)$ straightforwardly and investigate its impact on A_P. 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 $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_m, \widehat{\beta}_{m+1}, \cdots, \widehat{\beta}_p)^T = \left(\widehat{\boldsymbol{\beta}}_{(1)}^T, \widehat{\boldsymbol{\beta}}_{(2)}^T\right)$. Given that β_is in $\boldsymbol{\beta}_{(1)}$ are i.i.d and follow $N(0, \sigma_\beta^2)$, and the remaining ones in $\boldsymbol{\beta}_{(2)}$ are all zeros, it follows from the central limit theorem that

$$\widehat{\beta} \sim \begin{cases} N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}), & \text{for causal SNPs;} \\ N(0, \sigma_\beta^2 \cdot \frac{m}{n}), & \text{for non-causal SNPs.} \end{cases}$$

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 as 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

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

$$\begin{aligned} f(b) &\approx \frac{m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}) + \frac{p-m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{m}{n}) \\ &= \frac{m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(-\frac{b^2}{2\sigma_1^2}) + \frac{p-m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{b^2}{2\sigma_2^2}) \end{aligned}$$

with cumulative distribution function (CDF) given by

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

where $\sigma_1^2 = \sigma_\beta^2 \cdot (1 + m/n)$, $\sigma_2^2 = \sigma_\beta^2 \cdot m/n$, and $\Phi(x) = \int_{-\infty}^{x} \exp(-t^2/2)dt/\sqrt{2\pi}$ is the CDF of the standard normal random variable. Since the mixture distribution is symmetric about zero, without loss of generality, 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 to

$$\begin{aligned} &\Pr[b > F^{-1}(1-a) | b \sim N(0, \sigma_1^2)] = 1 - \Pr[b \leq F^{-1}(1-a) | b \sim N(0, \sigma_1^2)] \\ &= 1 - \Phi\big[\frac{F^{-1}(1-a)}{\sigma_1}\big] = 1 - \Phi\big[\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{n+m}}\big]. \end{aligned}$$

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

$$\begin{aligned} &\Pr[b > F^{-1}(1-a) | b \sim N(0, \sigma_2^2)] = 1 - \Pr[b \leq F^{-1}(1-a) | b \sim N(0, \sigma_2^2)] \\ &= 1 - \Phi\big[\frac{F^{-1}(1-a)}{\sigma_2}\big] = 1 - \Phi\big[\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{m}}\big]. \end{aligned}$$

Therefore, among $q = p \cdot a = q_1 + q_2$ selected SNPs, we have

$$\begin{aligned} m \cdot \kappa_1 &= m \cdot \big[1 - \Phi(\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{n+m}})\big] \quad \text{and} \\ (p-m) \cdot \kappa_2 &= (p-m) \cdot \big[1 - \Phi(\frac{F^{-1}(1-a)}{\sigma_\beta}\sqrt{\frac{n}{m}})\big] \end{aligned}$$

causal and null SNPs, respectively. For a given a or equivalently c, $F^{-1}(1-a)/\sigma_\beta$ is the same for both causal and null SNPs. Therefore, the quality of top-ranked SNP list

is largely determined by m/n. Remark 4 discusses the upper bounds of A_p under two extreme cases.

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

$$A_P^2(q) = \frac{n\phi(q)^2}{nm\phi(q) + qm^2} \approx \frac{n}{np + p^2} \cdot q. \tag{3}$$

Therefore, A_P reaches its upper bound $\sqrt{n/(n+p)}$ at $q = p$, suggesting that the best performing PRS is the one *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 using 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 $q = q_1 = \phi(q) = m$, the oracle case described in Remark 2.

In conclusion, the above analysis provides some practical guidelines on the use of PRS: 1) SNP screening should be avoided for highly polygenic/omnigenic traits where the m/n ratio is large; and 2) for monogenic and oligogenic traits[7] with a small m/n ratio, threshold-PRS should be used.

Results

Illustration of asymptotic limits of GWAS-PRS

We numerically evaluate the analytical results above and the performance of A_P^2 with $p = 100,000$, and $n = n_z = 100$, 1000 or $10,000$. Each entry of $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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 effects of the causal SNPs $\boldsymbol{\beta}_{(1)} \sim MVN(0, \boldsymbol{I}_m)$, based on which the phenotypes $\boldsymbol{y}$ and $\boldsymbol{y}_z$ are generated from model (2), where $\boldsymbol{I}_m$ is an $m \times m$ identity

matrix. A total of 100 replications are conducted for each simulation set up.

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

Simulation study on threshold-PRS

To illustrate the finite sample performance of threshold-PRS, we simulate $p = 100,000$ uncorrelated SNPs. As in Figure 2, we (naively) generate SNPs from $N(0,1)$. To study the effect of m/p, we vary the number of causal SNPs m and set it to 100, 1000, 10,000 and 50,000. The casual SNP effects $\boldsymbol{\beta}_{(1)} \sim MVN(0, \boldsymbol{I}_m)$. The linear polygenic model (2) is used to generate phenotype $\boldsymbol{y}$. The sample size is set to 1000 and 10,000 for training data, and 1000 for testing data. For threshold-PRS, as in Márquez-Luna et al.[24], 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. Again, a total of 100 replications are conducted for each simulation condition.

Figure 3 and Figure S1 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) stays nearly constant around $\sqrt{n/(n+p)}$ regardless of the m/p ratios, which is about 0.3 (shown in Figure 3) for $n = 10,000$ and 0.1 (shown in Figure S1) for $n = 1000$. In contrast, the performance of threshold-PRS varies with m. 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 Figure S1. For example, 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 gets 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 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 then correspondingly generated from Uniform $[0.05, 0.45]$, and the F_{st} values are independently generated from Uniform $[0.01, 0.04]$,[40] based on which the MAF of each sub-population is generated according to the Balding-Nichols model.[41] We let the sample size of sub-population be the same, and set them at either 200 or 2000. The population substructures are estimated with the principal component analysis of Price et al.[1] and the top 4 principal components 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 Figures S2 and S3, which are similar to those of Case 1. Figures S4 and S5 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 Figures S6 and S7. In the presence of population substructures, 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 Figures S8 and S9, which are also similar to those of Case 1, indicating that our asymptotic results are not sensitive to the distribution of the causal SNP effects, or $\boldsymbol{\beta}_{(1)}$.

Discussion

The PRS is widely used in GWAS for predicting a phenotype of the same type or for studying the co-heritability of a pleiotropy phenotype with the one in the training GWAS.[19,26,42–44] The purpose of PRS is to aggregate effects from a large number of causal SNPs, each of which has small contribution to the phenotype. However, recent years have witnessed widespread not so satisfying performances of PRS in real data applications.[22–25,45–48] Though previous studies[28–30] have found that the prediction accuracy of PRS is related to (n, p), the asymptotic properties of PRS is largely unknown. More rigorous investigation done in this paper would greatly help researchers gain better insights on PRS and appreciate the limitations of PRS to avoid over or under interpreting their findings.

We empirically and theoretically study the properties of PRS for complex polygenic traits prediction. To begin with, we illustrate for the first time how and why the commonly used marginal screening approaches for polygenic traits may fail in preserving the rank of GWAS signals. The properties of single SNP analysis is closely related to the increasingly recognized spurious correlation problem[33–39] in the statistics community. For polygenic/omnigenic traits, single SNP analysis is always misspecified since the effects of a large number of causal SNPs are not modeled but absorbed into the error term, which can profoundly affect the accuracy of marginal screening, an issue that can be safely ignored for traits with sparse genetic signals.

For GWAS-PRS, our asymptotic results align well with those of Daetwyler et al.[28], Dudbridge[29] and Chatterjee et al.[30], but are more statistically rigorous and general. In addition, we illustrate in Figure 2 that the variation of the PRS performance A_P increases with the increase of the p/n ratio as well. When the sample size of training dataset is small, or p/n is large, the prediction accuracy of GWAS-PRS fluctuates around zero with a large variation. Thus, the results of PRS should be interpreted with caution. We further generalize our results to threshold-PRS. We highlight how the number of causal SNPs influences the performance of threshold-PRS, and recognize

its distinct behaviors under dense and sparse genetic signal scenarios. For traits with highly polygenic or omnigenic genetic architecture, the general rule is to avoid the use of threshold-PRS but GWAS-PRS. However, the performance of the later is solely determined by the n/p ratio and is doomed to perform poorly when n is not large, leading to a practical paradox.

In our study, we set h^2=1, the most optimistic situation for phenotypic prediction. Our results can be easily extended to $h^2 < 1$. For example, the asymptotic prediction accuracy of GWAS-PRS becomes $\sqrt{nh^2/(n + p/h^2)}$. The performance of threshold-PRS under Case 2 for traits with h^2=0.5 is presented in Figures S10 and S11. Compared to Figures S2 and S3 where h^2=1, the prediction accuracy of threshold-PRS decreases, but the general conclusions remain the same. Further, in our investigation, we assume all typed SNPs are independent of each other, plus all causal SNPs are observed, an assumption that would never hold for any real GWAS. For real GWAS data with SNPs in linkage disequilibrium (LD) and partially genotyped causal SNPs, GWAS-PRS and threshold-PRS would perform even worse.[26,27,30,49]

Last, it is worth mentioning that the low performance observed for PRS is not specific to PRS. Though regularization-based methods such as LASSO[50] or Elastic-Net[51] fit multiple SNPs simultaneously, they are not immune to the spurious correlation problem,[37,39] and are not able to separate causal and null SNPs as well when the genetic signals are not sparse.[52] Thus, they suffer equally as PRS in general. The best linear unbiased prediction methods[53] such as the one implemented in the popular GCTA (genome-wide complex trait analysis) toolset[54] cannot improve much over PRS either according to Pasaniuc and Price[32] and our simulation results.

In summary, our investigation clears up some misconceptions on PRS and demonstrates that PRS is not as useful as its name suggests, and as powerful as the genetics community expects neither for predicting polygenic traits. We hope this research will serve as a wake up call to the genetics community in appreciating more about the challenges in studying complex polygenic traits. As such, more resources and efforts can be devoted in performing better experiments and developing better statistical methods

for polygenic and omnigenic traits.

Appendix A

In this appendix, we highlight the key steps and important intermediate results to prove the two theorems in Material and Methods. More technical details can be found in the Supplemental Data.

Proposition A 1. Under the polygenic model (2) and Condition (1), if $m \to \infty$ when $n, p \to \infty$, then we have

$$\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}}{nm \cdot \sigma_\beta^2} = 1 + o_p(1)$$
$$\frac{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)}{\big\{n^2 m(p-m) + n^2 m(m+n)\big\} \cdot \sigma_\beta^2} = 1 + o_p(1).$$

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

$$\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}}{n^2 m \cdot \sigma_\beta^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

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

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

$$\begin{aligned}
&\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \\
&= \frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)} - n^2 m \cdot \sigma_\beta^2}{\sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)}} \\
&\cdot \sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)} + n^2 m \cdot \sigma_\beta^2 \\
&= \mathcal{O}_p\Big\{\sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)}\Big\} + n^2 m \cdot \sigma_\beta^2.
\end{aligned}$$

It follows that

$$A_P^2 = \frac{\mathcal{O}_p\{(n^2m^2p + n^4m^2)\cdot\sigma_\beta^4\}}{\big[nm\cdot\sigma_\beta^2\cdot\{1+o(1)\}\big]\cdot\big[n^2m(n+p)\cdot\sigma_\beta^2\cdot\{1+o(1)\}\big]}$$
$$= \mathcal{O}_p\Big(\frac{n^2m^2p+n^4m^2}{n^3m^2p+n^4m^2}\Big) = \mathcal{O}_p\Big(\frac{n^2+c\cdot n^\alpha}{n^2+c\cdot n^{1+\alpha}}\Big) = \mathcal{O}_p\big(\frac{1}{n}\big)$$

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

Proposition A 2. Under the polygenic model (2) and Condition (1), if $m, q_1, q_2 \to \infty$ when $n, p \to \infty$, then we have

$$\frac{\boldsymbol{\beta}_{(1)}^T\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}}{nm\cdot\sigma_\beta^2} = 1 + o_p(1)$$
$$\frac{\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(11)}\boldsymbol{Z}_{(11)}^T + \boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(21)}\boldsymbol{Z}_{(21)}^T\big)\big(\boldsymbol{Z}_{(11)}\boldsymbol{X}_{(11)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(21)}\boldsymbol{X}_{(21)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)}{\{n^2mq_2 + n^2q_1(m+n)\}\cdot\sigma_\beta^2} = 1 + o_p(1).$$

Further if $\{m^2(q_1+q_2)\}/(n^2q_1^2) \to 0$, then we have

$$\frac{\boldsymbol{\beta}_{(1)}^T\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(11)}\boldsymbol{X}_{(11)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(21)}\boldsymbol{X}_{(21)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}}{n^2q_1\cdot\sigma_\beta^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$A_P^2/\big(\frac{nq_1^2}{nmq_1+qm^2}\big) = 1 + o_p(1).$$

Without loss of generality and for simplicity, we let $\sigma_\beta^2 = 1$, and we note that

$$\boldsymbol{\beta}_{(1)}^T\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(11)}\boldsymbol{X}_{(11)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(21)}\boldsymbol{X}_{(21)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}$$
$$= \mathcal{O}_p\Big\{\sqrt{n^2q_1^3 + 2n^2q_1^2(m-q_1) + 2n^3q_1(m-q_1) + n^2(m-q_1)^2q_1 + n^2m^2q_2}\Big\} + n^2q_1.$$

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

$$
\begin{aligned}
A_P^2 &= \frac{\mathcal{O}_p\{n^2q_1^3 + 2n^2q_1^2(m-q_1) + 2n^3q_1(m-q_1) + n^2(m-q_1)^2q_1 + n^2m^2q_2 + n^4q_1^2\}}{\big[nm\cdot\{1+o_p(1)\}\big]\cdot\big[(n^3q_1+n^2mq_1+n^2mq_2)\cdot\{1+o_p(1)\}\big]} \\
&= \mathcal{O}_p\Big\{\frac{m^2(q_1+q_2)+2n(m-q_1)q_1+n^2q_1^2}{nm^2(q_1+q_2)+n^2mq_1}\Big\} = \mathcal{O}_p\Big\{\frac{m^2(q_1+q_2)+n^2q_1^2}{nm^2(q_1+q_2)+n^2mq_1}\Big\} \\
&= \mathcal{O}_p(\frac{1}{n}) = o_p(1).
\end{aligned}
$$

Thus, Theorem 2 is proved.

Supplemental Data

Supplemental Data include 11 Figures and Supplemental Methods and can be found with this article online at `http://www.ajhg.org`.

Declaration of Interests

The authors declare no competing interests.

Acknowledgement

We would like to thank Ziliang Zhu, Jingwen Zhang, and Hongtu Zhu for helpful discussions; and Xinlei Mi for independent technical verification. We are deeply grateful to Hongtu Zhu for providing extremely helpful comments and edits; and to Laura Zhou for careful proofreading and feedback. The manuscript has been greatly improved by their generous contributions.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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*, 747–753.

[6] 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.

[7] 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*, 110–124.

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

[9] 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.

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

[11] 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*, 1114–1120.

[12] 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*, 565–569.

[13] Dudbridge, F. (2016). Polygenic epidemiology. Genetic Gpidemiology *40*, 268–272.

[14] 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.

[15] 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.

[16] 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*, 247–250.

[17] 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*, 1468–1475.

[18] 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.

[19] 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.

[20] 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.

[21] 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*, 421–427.

[22] 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.

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

[24] 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.

[25] Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics *19*, 581–590.

[26] Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics *17*, 392–406.

[27] 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*, 507–515.

[28] 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.

[29] Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genetics *9*, e1003348.

[30] 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*, 400–405.

[31] 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.

[32] Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics *18*, 117–127.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

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

[41] 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.

[42] Power, R. A., Steinberg, S., Bjornsdottir, G., Rietveld, C. A., Abdellaoui, A., Nivard, M. M., Johannesson, M., Galesloot, T. E., Hottenga, J. J., Willemsen, G. *et al.* (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. Nature Neuroscience *18*, 953–955.

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

[44] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T. *et al.* (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics *50*, 1219–1224.

[45] Clarke, T., Lupton, M., Fernandez-Pujals, A., Starr, J., Davies, G., Cox, S., Pattie, A., Liewald, D., Hall, L., MacIntyre, D. *et al.* (2016). Common polygenic risk for autism spectrum disorder (asd) is associated with cognitive ability in the general population. Molecular Psychiatry *21*, 419–425.

[46] Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V., and Zammit, S. (2018). The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. Schizophrenia Research *197*, 2–8.

[47] Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V., and Zammit, S. (2018). The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: A systematic review. Journal of Affective Disorders *234*, 148–155.

[48] Socrates, A., Bond, T., Karhunen, V., Auvinen, J., Rietveld, C., Veijola, J., Jarvelin, M.-R., and O'Reilly, P. (2017). Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. BioRxiv , 203257.

[49] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2018). Polygenic prediction via bayesian regression and continuous shrinkage priors. BioRxiv , 416859.

[50] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) *58*, 267–288.

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

[52] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *70*, 849–911.

[53] Chen, C.-Y., Han, J., Hunter, D. J., Kraft, P., and Price, A. L. (2015). Explicit modeling of ancestry improves polygenic risk scores and blup prediction. Genetic Epidemiology *39*, 427–438.

[54] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for

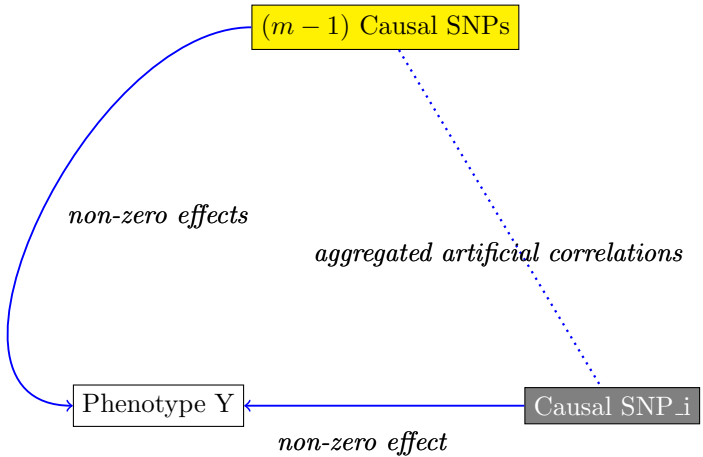
genome-wide complex trait analysis. The American Journal of Human Genetics *88*, 76–82.

Figure Titles and Legends

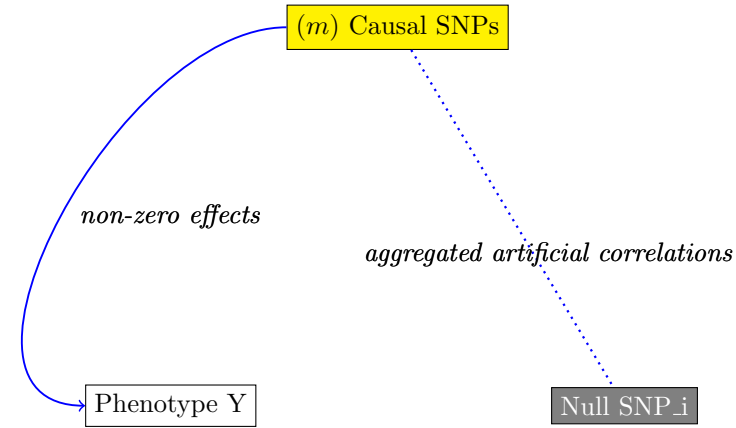
Figure 1: Impact of artificial correlation on the marginal SNP effect estimate of SNP i ($i = 1, \cdots, p$).

Figure 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=100,000 and n=100, 1000 and 10,000, respectively.

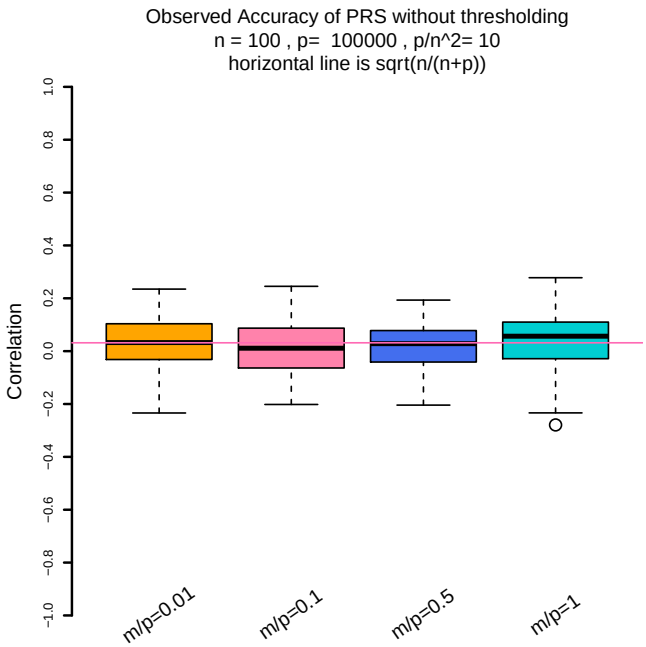
Figure 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set p=100,000 and $n = 10,000$ in training data and n=1000 in testing data.



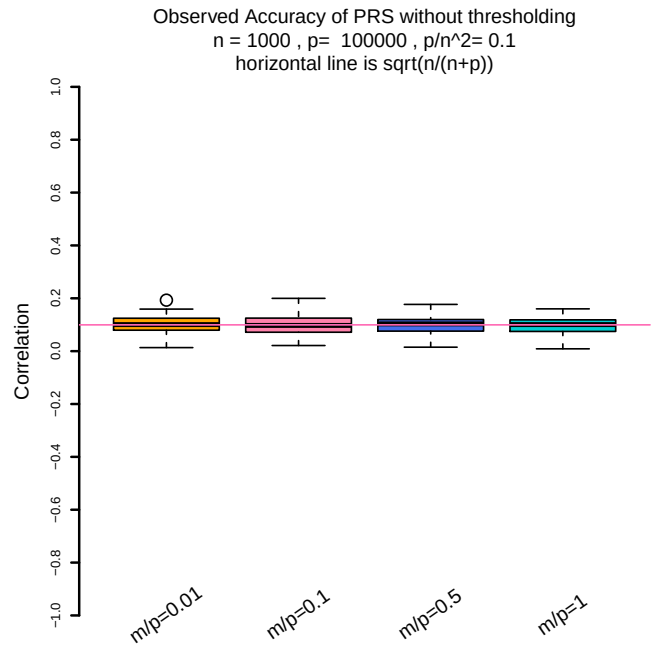
a) Causal SNP_i



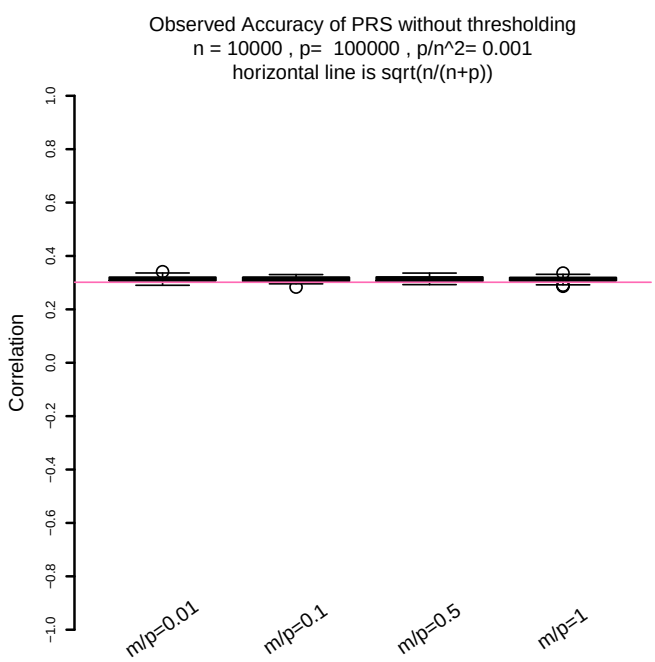
b) Null SNP_i



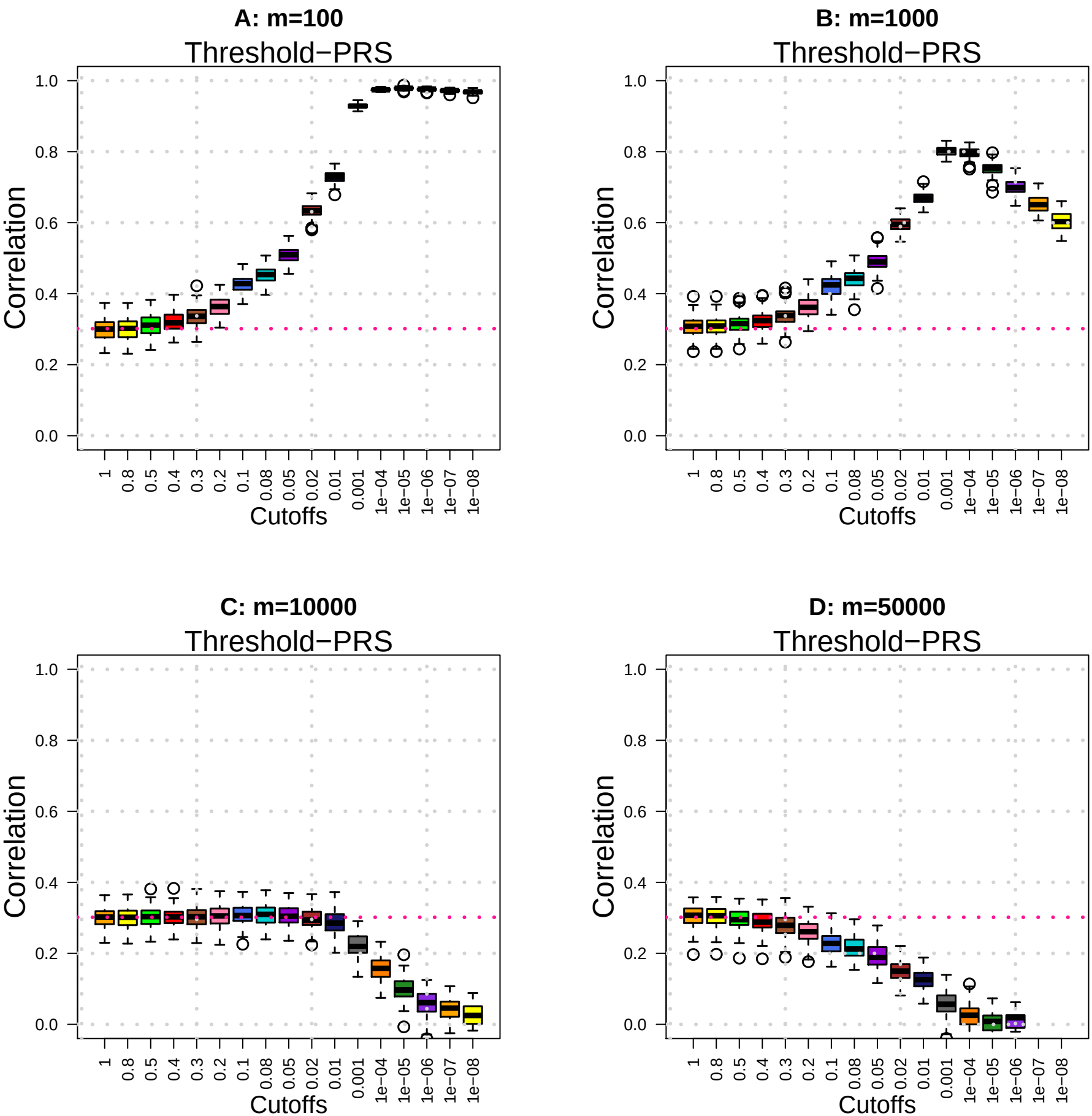
(a) (n,p)=(100, 100,000)



(b) (n,p)=(1000, 100,000)



(c) (n,p)=(10,000, 100,000)


A: m=100
Threshold−PRS
Correlation
Cutoffs
B: m=1000
Threshold−PRS
Correlation
Cutoffs
C: m=10000
Threshold−PRS
Correlation
Cutoffs
D: m=50000
Threshold−PRS
Correlation
Cutoffs
0.0
0.2
0.4
0.6
0.8
1.0
1
0.8
0.5
0.4
0.3
0.2
0.1
0.08
0.05
0.02
0.01
0.001
1e−04
1e−05
1e−06
1e−07
1e−08

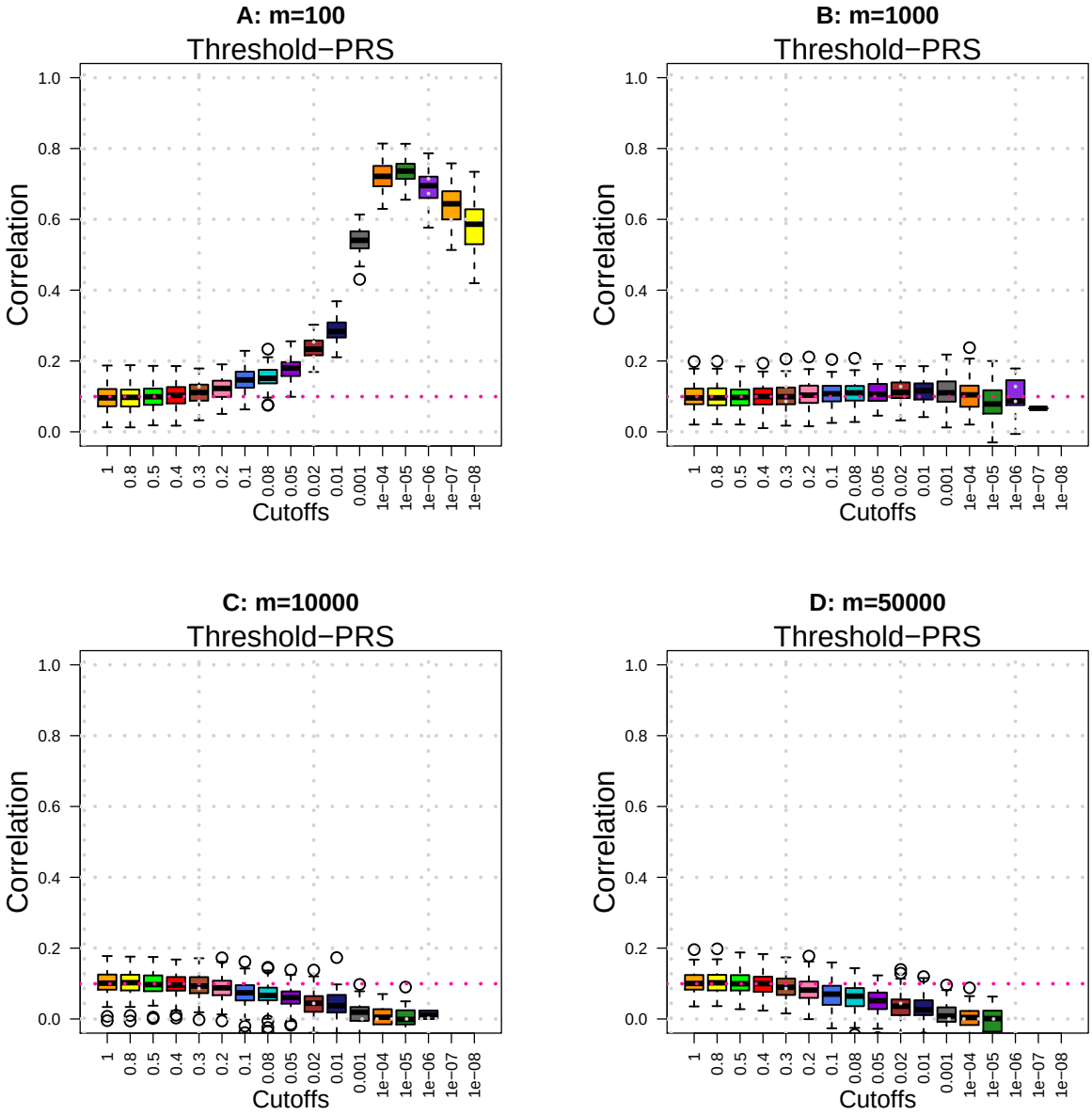


Figure S 1: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set p=100,000 and n=1000 in both training and testing data.

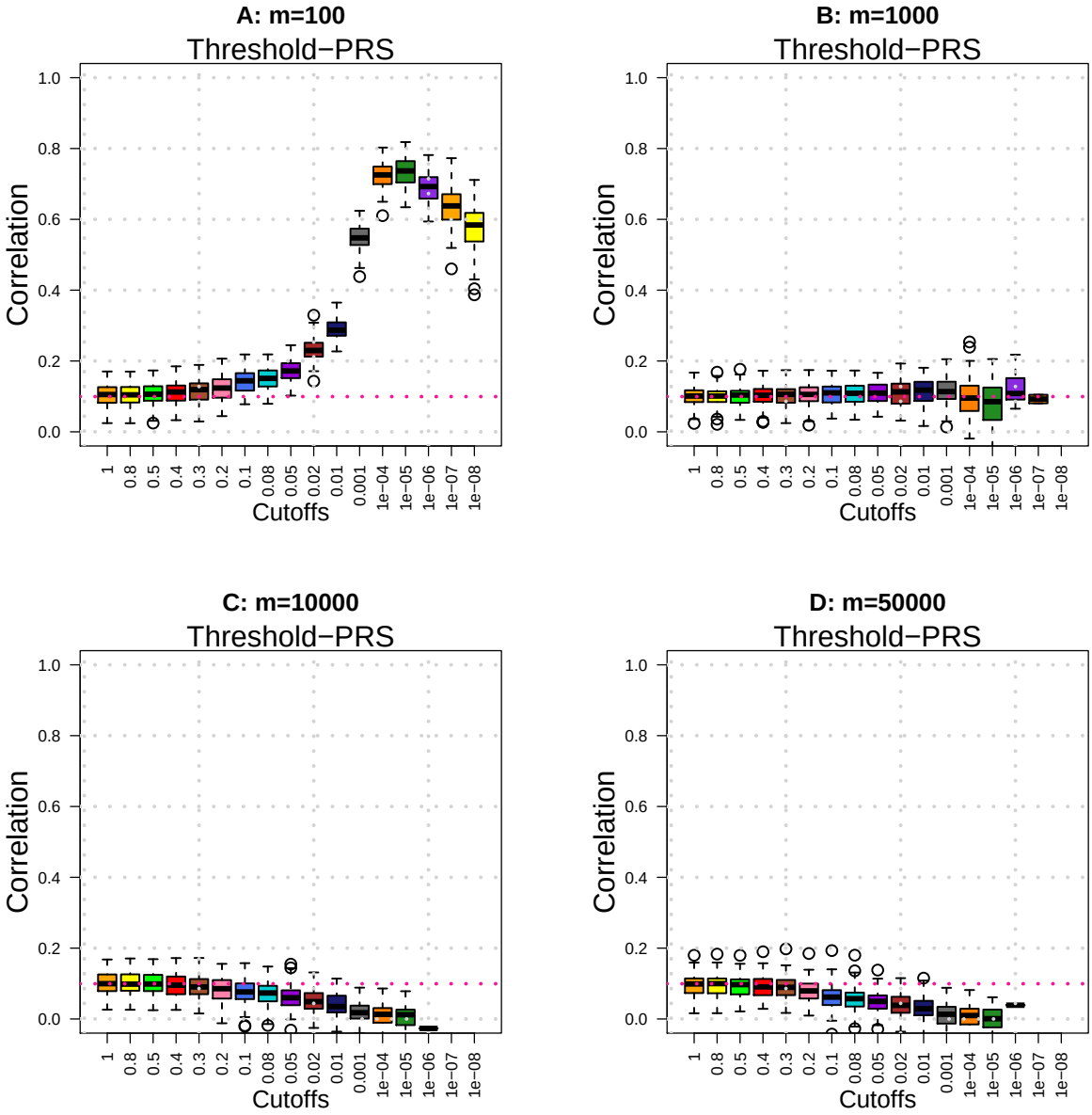


Figure S 2: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=1000 in both training and testing data.

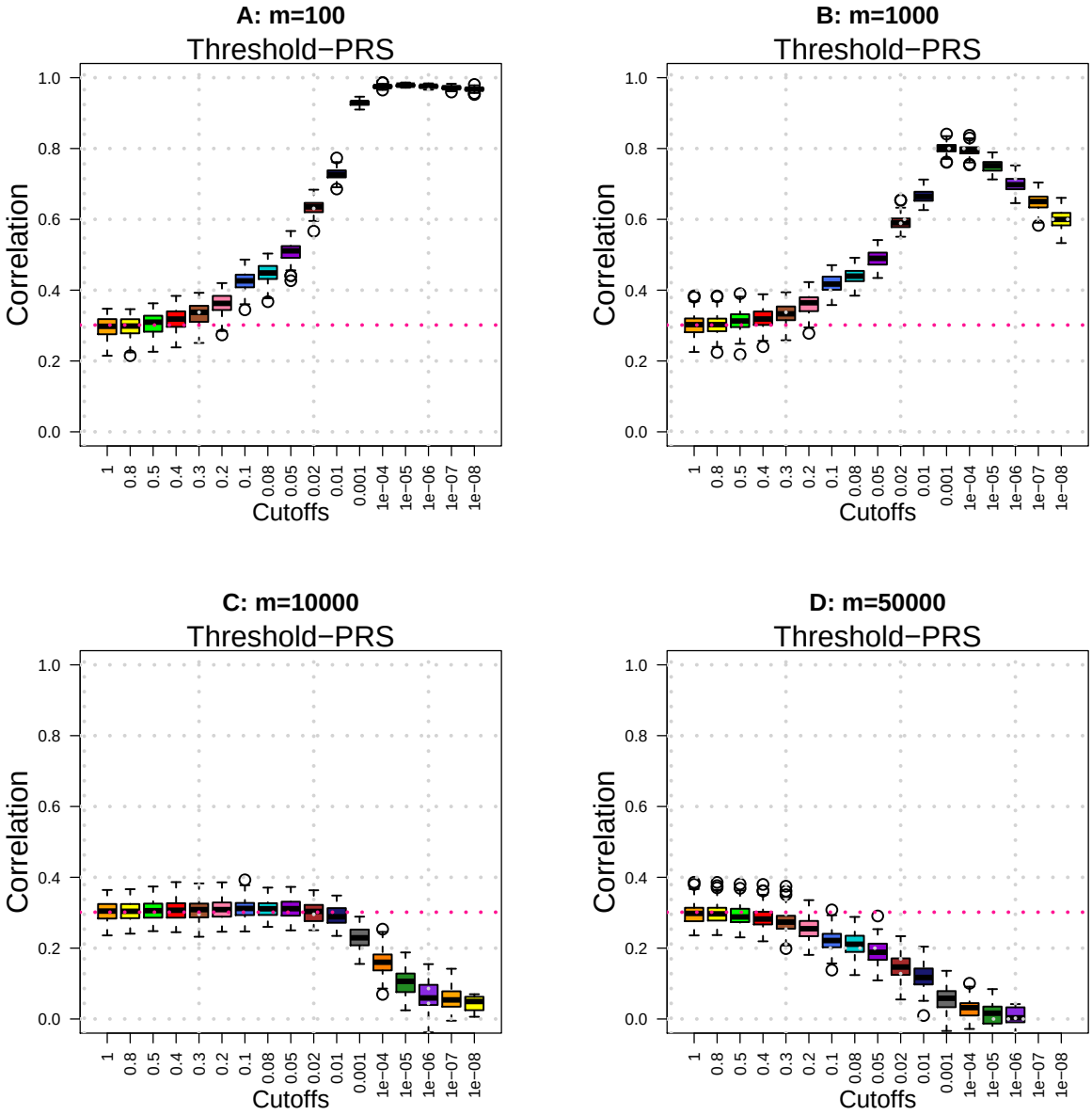


Figure S 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=10,000 in training data and $n = 1000$ in testing data.

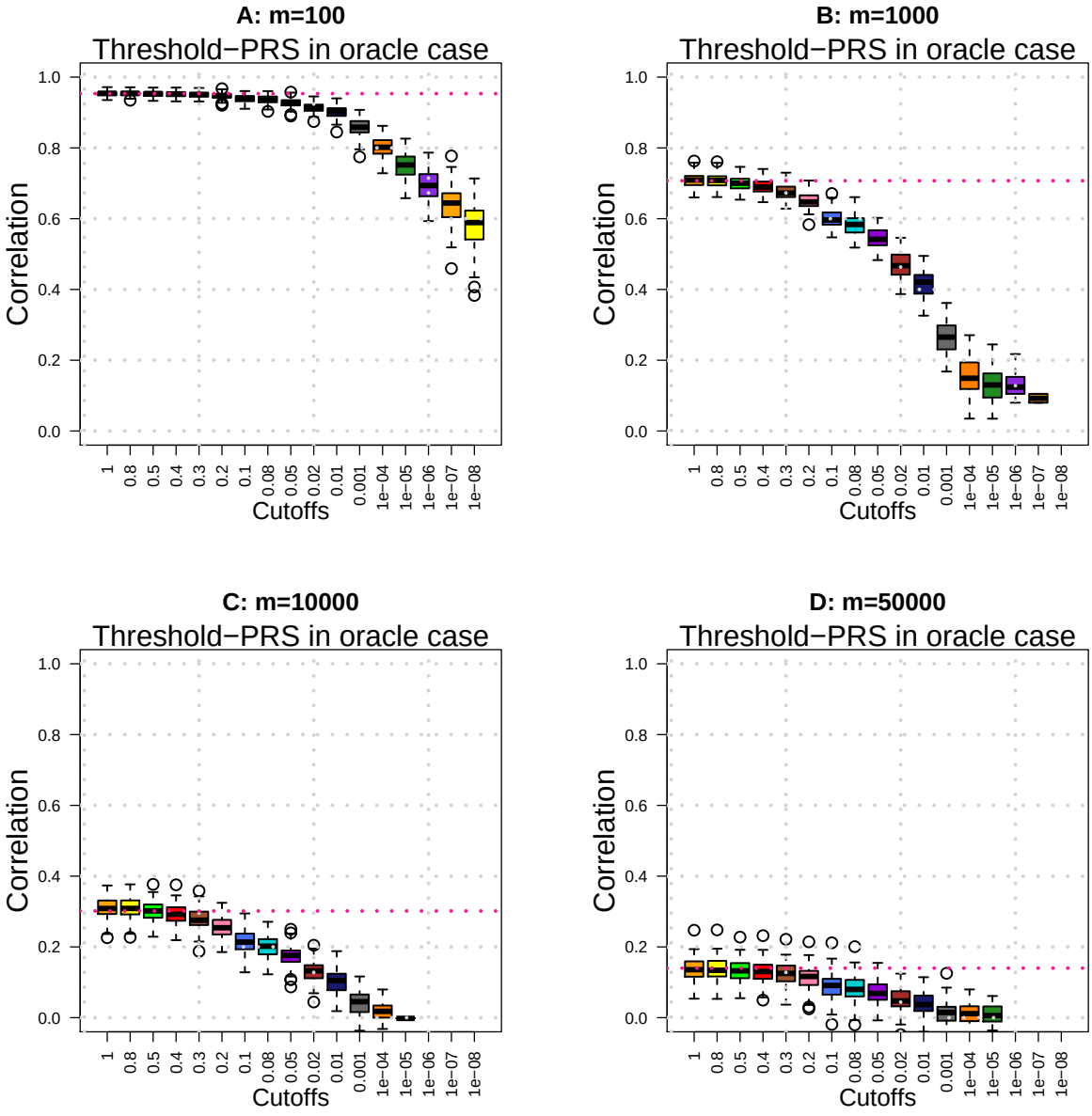


Figure S 4: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=1000 in both training and testing data.

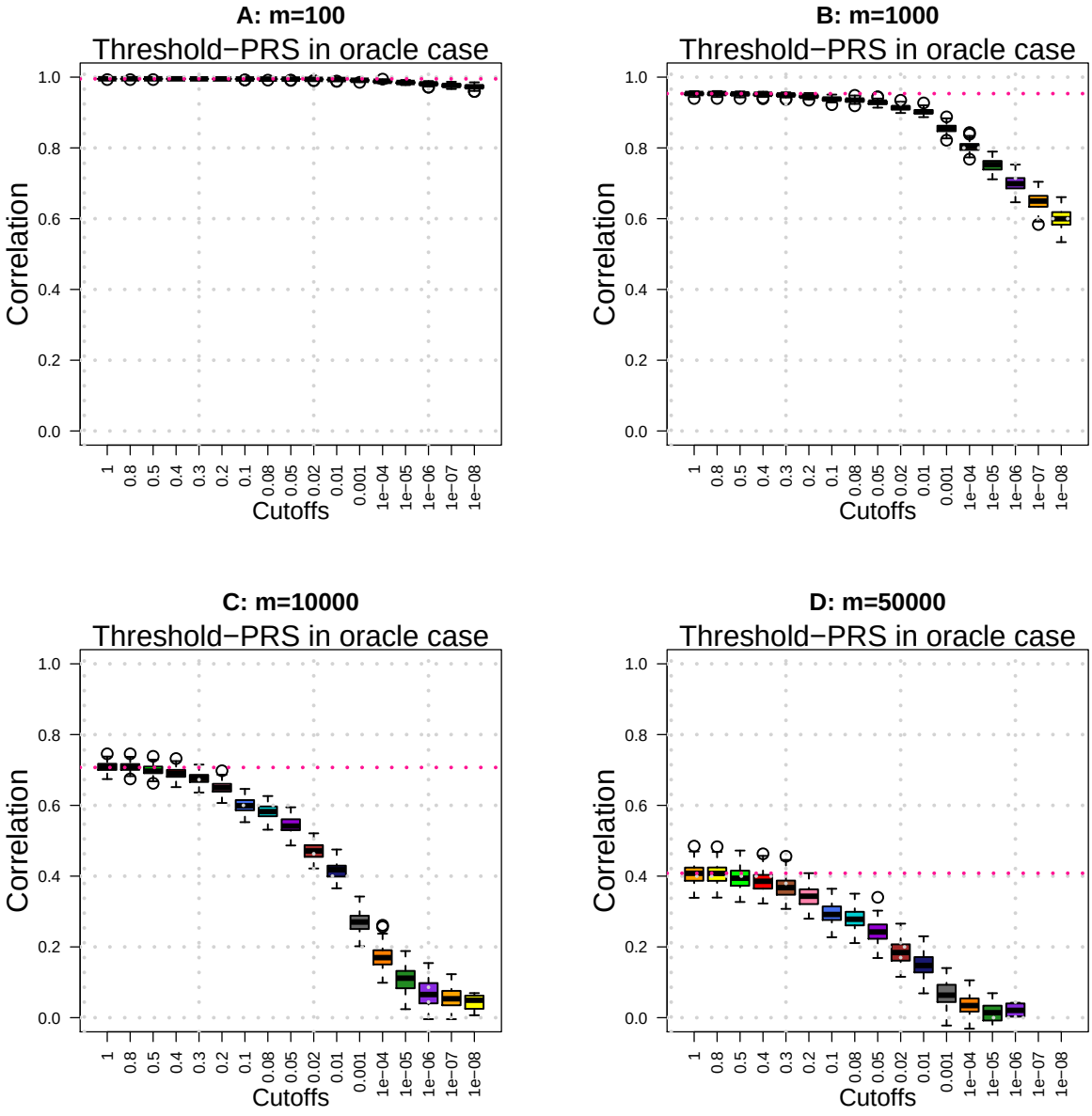


Figure S 5: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=10,000 in training data and $n = 1000$ in testing data.

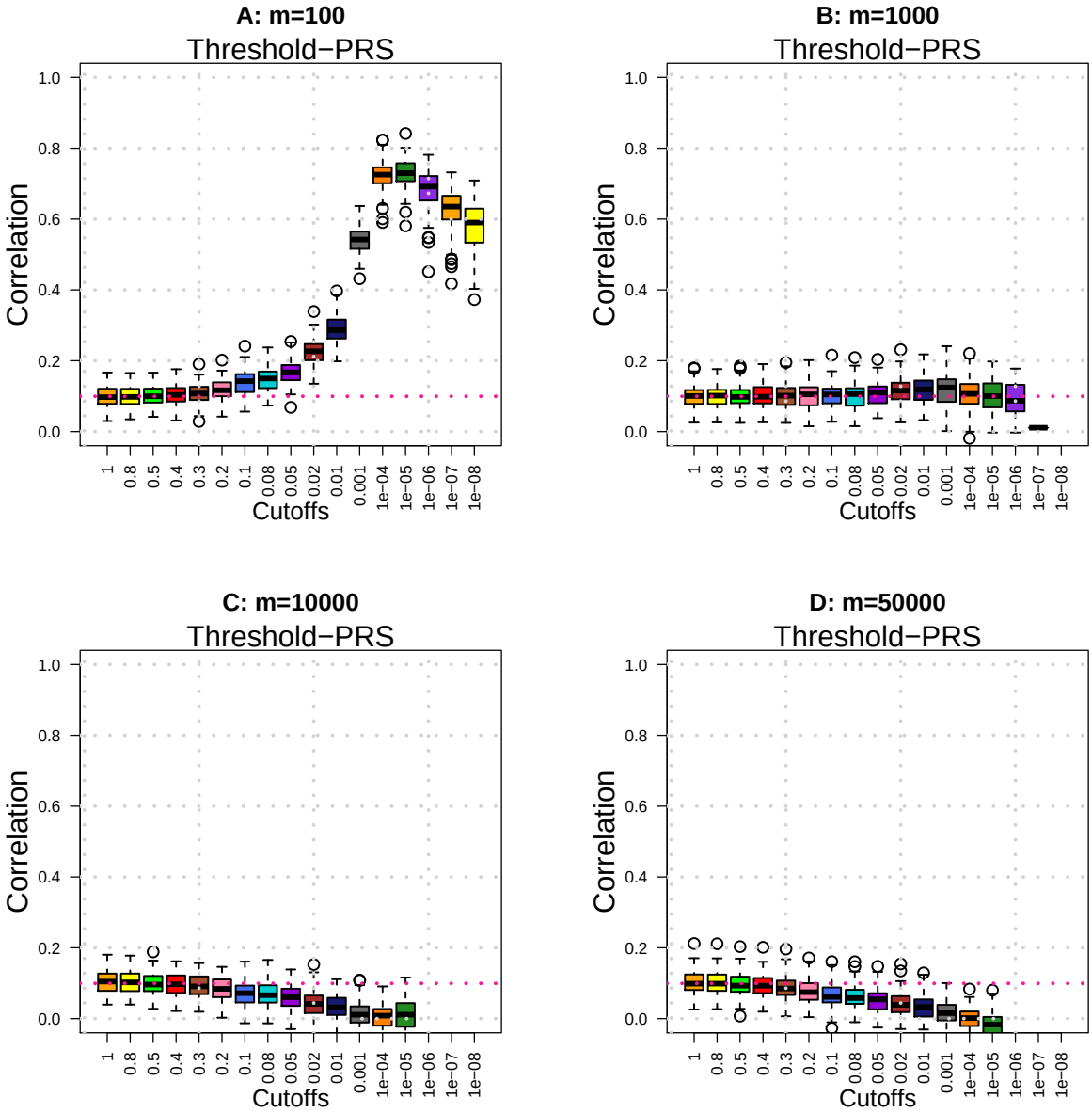


Figure S 6: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=1000 in both training and testing data.

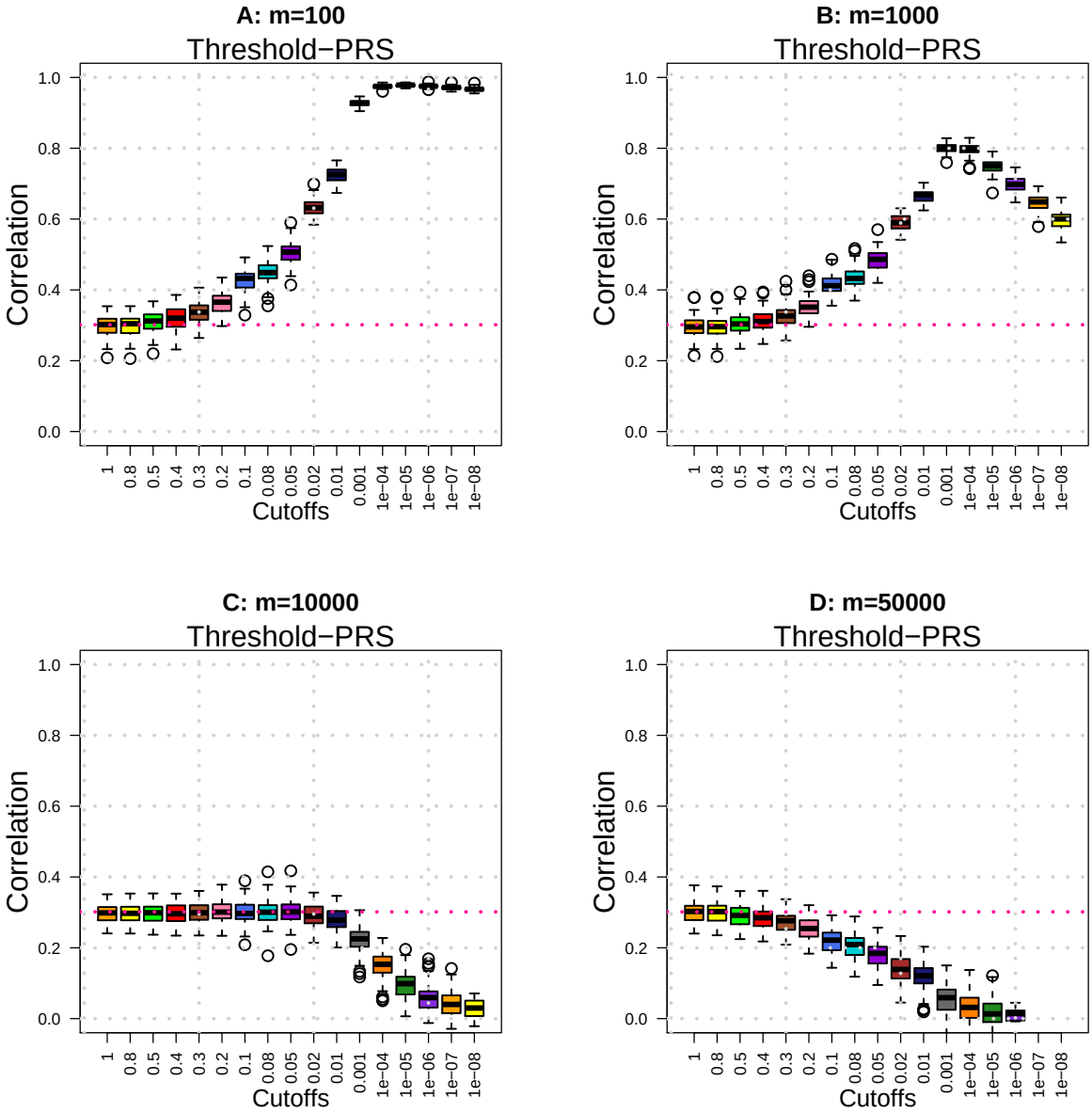


Figure S 7: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=10,000 in training data and $n = 1000$ in testing data.

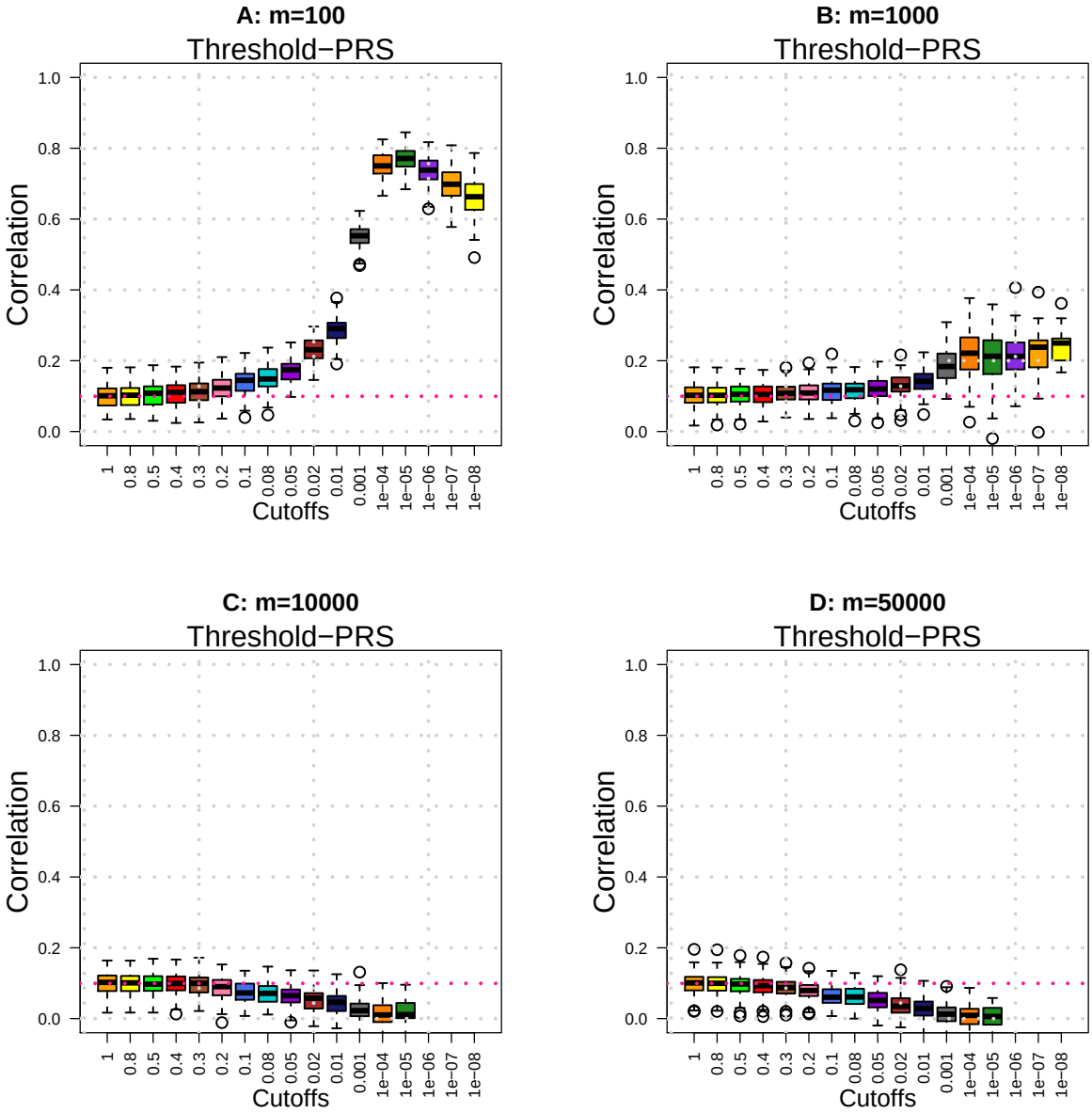


Figure S 8: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=1000 in both training and testing data.

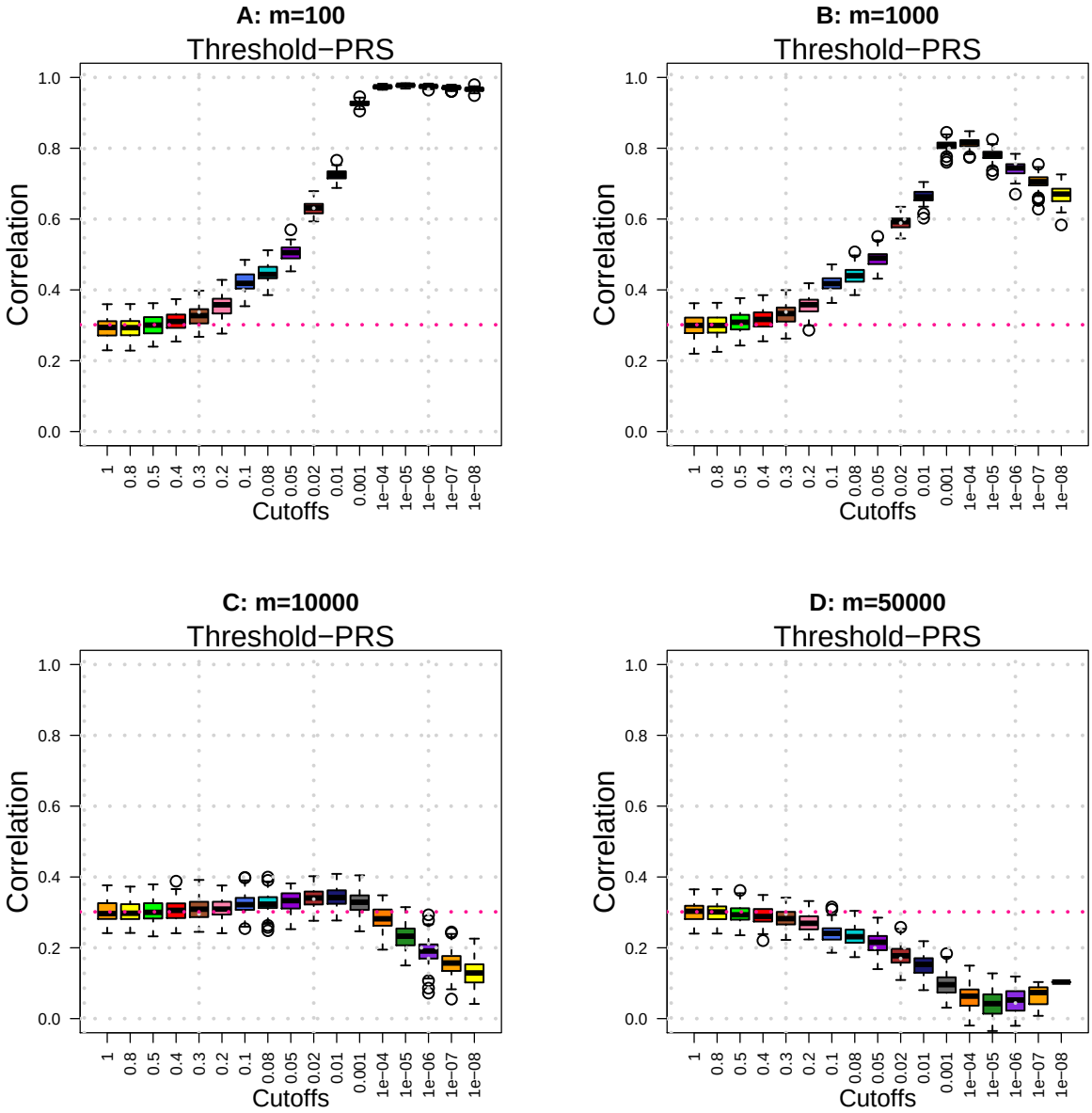


Figure S 9: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=10,000 in training data and $n = 1000$ in testing data.

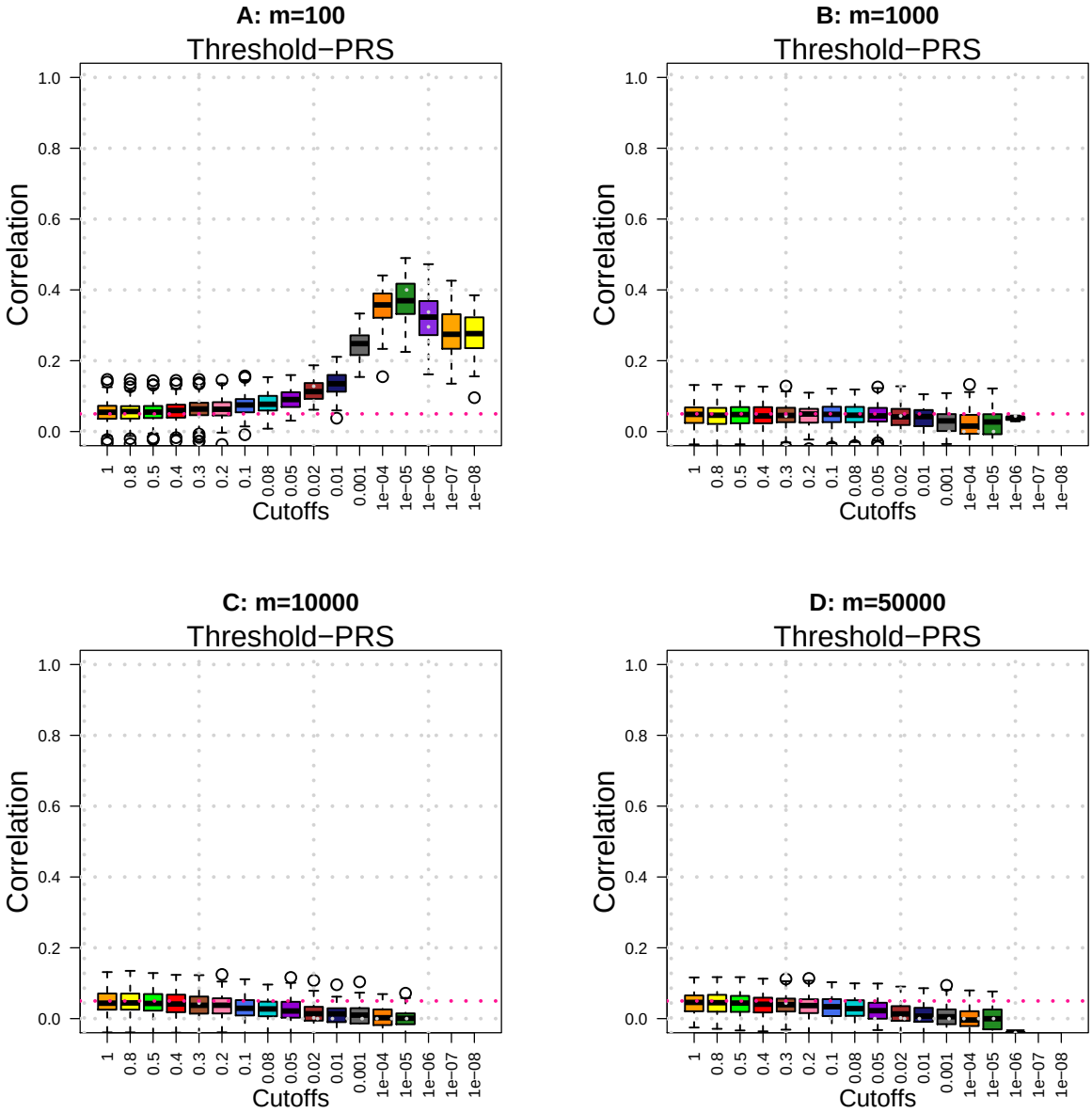


Figure S 10: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the heritability h^2=0.5. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=1000 in both training and testing data.

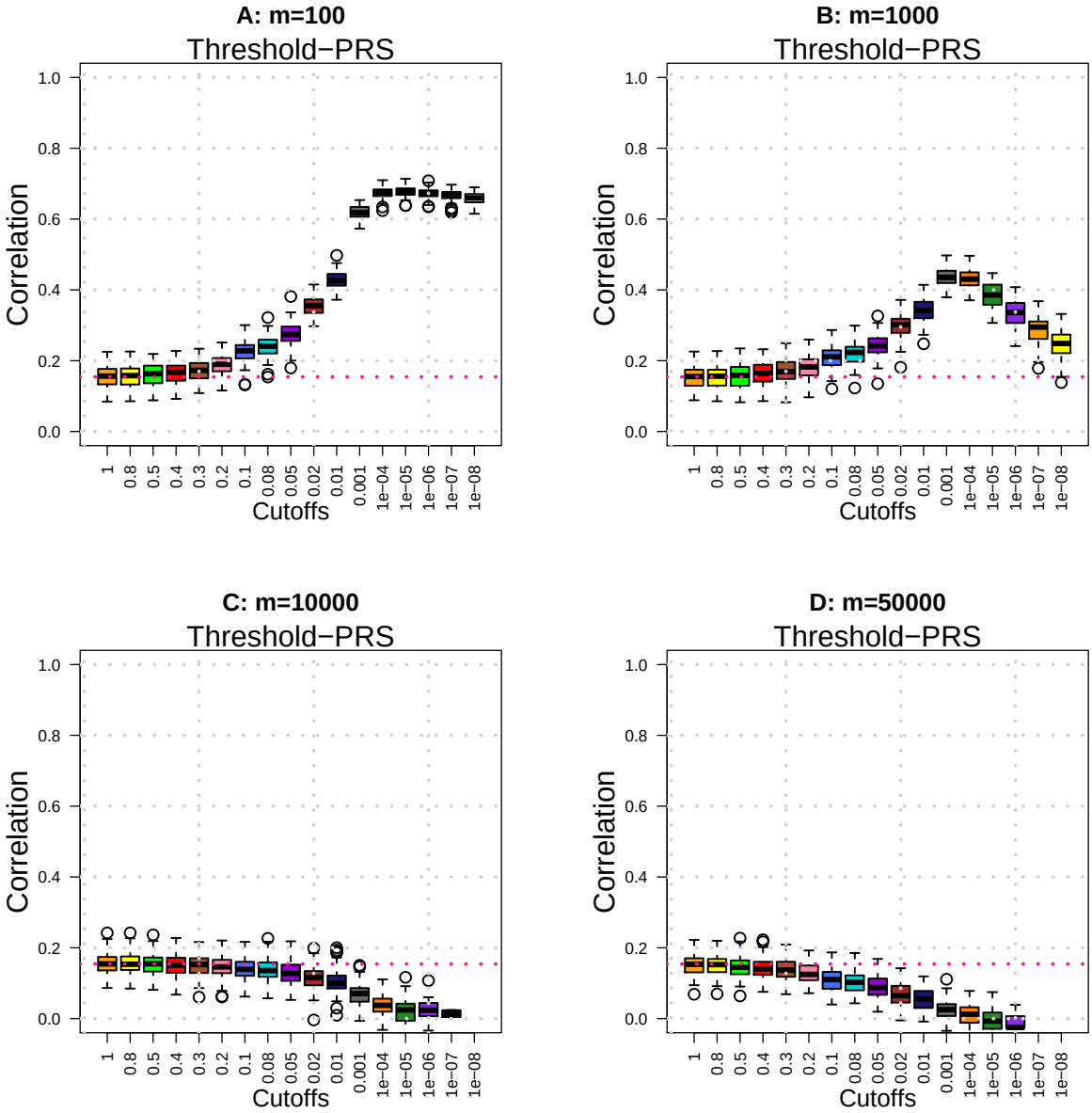


Figure S 11: Prediction accuracy (A_P) of threshold-PRS across different m/p when the heritability h^2=0.5. SNP data are independently sampled from $\{0, 1, 2\}$. We set p=100,000 and n=10,000 in training data and $n = 1000$ in testing data.

Supplemental Methods

Intermediate results

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

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = n^2 m \cdot \sigma_\beta^2 \\
&\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \{(n^2 m^3 + 4n^3 m^2 + n^2 m^2 (p-m)) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)\} \cdot \{1 + o(1)\}
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) = nm \cdot \sigma_\beta^2 \\
&\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) = o(n^2 m^2 \cdot \sigma_\beta^4)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\Big\{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)\Big\} \\
&= n^2 m(n+m) \cdot \sigma_\beta^2 \cdot \{1 + o(1)\} + n^2 m(p-m) \cdot \sigma_\beta^2 \\
&\mathrm{Var}\Big\{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)\Big\} \\
&= o\big\{n^4 m^2 (n+p)^2 \cdot \sigma_\beta^4\big\},
\end{aligned}
$$

where $b_4 < \infty$ 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 $\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}$ and

$$
\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big),
$$

the expected values can respectively dominate the corresponding standard error for any

ratios among (p, m, n). However, for the covariance term

$$\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)},$$

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

$$\Pr\Big(\Big|\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}}{nm \cdot \sigma_\beta^2} - 1\Big| \geq k\Big) \leq \frac{\mathrm{Var}\big(\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}}{nm \cdot \sigma_\beta^2}\big)}{k^2} = \frac{\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big)}{n^2 m^2 \cdot \sigma_\beta^4 k^2} = o(1),$$

and

$$\begin{aligned}
&\Pr\Big\{\Big|\frac{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)}{n^2 m(n+p) \cdot \sigma_\beta^2} - 1\Big| \geq k\Big\} \\
&\leq \mathrm{Var}\Big\{\frac{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)}{n^2 m(n+p) \cdot \sigma_\beta^2}\Big\} \cdot \frac{1}{k^2} \\
&= \frac{\mathrm{Var}\big\{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)\big\}}{n^4 m^2 (n+p)^2 \cdot \sigma_\beta^4 k^2} \\
&= o(1),
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\Big(\Big|\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}}{n^2 m \cdot \sigma_\beta^2} - 1\Big| \geq k\Big) \\
&\leq \mathrm{Var}\Big(\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}}{n^2 m \cdot \sigma_\beta^2}\Big) \cdot \frac{1}{k^2} \\
&= \frac{\{n^2 m^3 + n^2 m^2 (p-m)\} \cdot \{1 + o(1)\}}{n^4 m^2 \cdot \sigma_\beta^4 k^2} = \frac{p}{n^2 k^2} \cdot \{1 + o(1)\}.
\end{aligned}$$

In follows that Proposition A 1 is proved. More generally, if training and testing data have different sample sizes, donated as n and n_z, respectively, we have the following results.

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

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = nn_z m \cdot \sigma_\beta^2$$

$$\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) =$$

$$\big\{(nn_z m^2 p + 2n^2 n_z m^2 + 2nn_z^2 m^2) \cdot \sigma_\beta^4 + n^2 n_z^2 m \cdot (b_4 - \sigma_\beta^4)\big\} \cdot \{1 + o(1)\}$$

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) = n_z m \cdot \sigma_\beta^2$$

$$\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) = o(n_z^2 m^2 \cdot \sigma_\beta^4)$$

$$\mathrm{E}\Big\{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)\Big\}$$

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

$$\mathrm{Var}\Big\{\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)\Big\}$$

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

By Markov's inequality and continuous mapping theorem again, we have the following results.

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

$$\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}}{n_z m \cdot \sigma_\beta^2} = 1 + o_p(1)$$

$$\frac{\big(\beta_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big)}{[nn_z m(p-m) + nn_z m(n+m)] \cdot \sigma_\beta^2} = 1 + o_p(1).$$

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

$$\frac{\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}}{nn_z m \cdot \sigma_\beta^2} = 1 + o_p(1),$$

and it follows that

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

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

$$\begin{aligned}
&\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \\
&= \mathcal{O}_p\Big\{\sqrt{nn_z m^2 p + 2n^2 n_z m^2 + 2nn_z^2 m^2 + n^2 n_z^2 m(b_4 - \sigma_\beta^4)}\Big\} + nn_z m \cdot \sigma_\beta^2 \\
&= \mathcal{O}_p\big\{(n^{1/2} n_z^{1/2} m p^{1/2} + nn_z m) \cdot \sigma_\beta^2\big\}.
\end{aligned}$$

It follows that

$$A_P^2 = \frac{\mathcal{O}_p(nn_z m^2 p + n^2 n_z^2 m^2)}{\big[n_z m \cdot \{1 + o(1)\}\big] \cdot \big[nn_z m(n+p) \cdot \{1 + o(1)\}\big]} = \mathcal{O}_p(\frac{p + nn_z}{n_z p + nn_z}) = O_p(\frac{1}{n_z}).$$

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

Proposition S 4. Under Conditions (1), if $m, q_1, q_2 \to \infty$ as $n, p \to \infty$, then we have

$$\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = n^2 q_1 \\
&\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = \big\{n^2 q_1^3 + 4n^3 q_1^2 \\
&+ n^4 q_1 (b_4 - 1) + 2n^2 q_1^2 (m - q_1) + 2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2\big\} \cdot \{1 + o(1)\} \\
&\mathrm{E}(\mathrm{Var}_2) = \{n^2 q_1 (n + m) + n^2 q_2 m\} \cdot \{1 + o(1)\} \\
&\mathrm{Var}(\mathrm{Var}_2) = o\big[\{n^2 q_1 (n + m) + n^2 q_2 m\}^2\big].
\end{aligned}$$

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

Technical details

The following technical details are useful to prove our theoretical results. Most of them involve calculating the asymptotic expectation of the trace of the product of multiple large random matrices. 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 similar calculations.

GWAS-PRS

First moment of covariance term

$$
\begin{aligned}
&\boldsymbol{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\mathrm{tr}(\boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)})\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1=1}^{m}\sum_{k_2=1}^{m} Z_{ik_1} X_{jk_1} Z_{ik_2} X_{jk_2}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_1=1}^{m} Z_{ik_1}^2 X_{jk_1}^2\big) = \sigma_\beta^2 \cdot n^2 m
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\mathrm{tr}(\boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)})\big) \\
&= \sigma_\beta^2 \cdot E(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k_2=1}^{m}\sum_{k_1=1}^{m} Z_{(2)ik_2} X_{(2)jk_2} X_{(1)jk_1} Z_{(1)ik_1}) = 0
\end{aligned}
$$

Thus, we have $\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z} \boldsymbol{X}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = \sigma_\beta^2 \cdot n^2 m$.

First moment of variance term I

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \mathrm{E}\big\{E(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)} | Z)\big\} = \mathrm{E}\big(\mathrm{tr}(\boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \cdot \boldsymbol{I}_m \cdot \sigma_\beta^2) + 0\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\mathrm{tr}(\boldsymbol{Z}_{(1)} \boldsymbol{Z}_{(1)}^T)\big) = \sigma_\beta^2 \cdot \mathrm{E}\big(\sum_{i=1}^n \sum_{j=1}^m Z_{ij}^2\big) = \sigma_\beta^2 \cdot nm
\end{aligned}
$$

First moment of variance term II

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\mathrm{tr}(\boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)})\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\sum_{i=1}^n \sum_{j=1}^m \sum_{k_1=1}^m \sum_{l_1=1}^n \sum_{k_2=1}^m \sum_{l_2=1}^n Z_{ik_1} Z_{ik_2} X_{l_1k_1} X_{l_1j} X_{l_2k_2} X_{l_2j}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\sum_{i=1}^n \sum_{k_1=k_2=j}^m \sum_{l_1 \neq l_2}^{n(n-1)} Z_{ik}^2 X_{l_1k}^2 X_{l_2k}^2 + \sum_{i=1}^n \sum_{k_1=k_2 \neq j}^{m(m-1)} \sum_{l_1=l_2}^n Z_{ik}^2 X_{lk}^2 X_{lj}^2 \\
&\qquad + \sum_{i=1}^n \sum_{k_1=k_2=j}^m \sum_{l_1=l_2}^n Z_{ik}^2 X_{lk}^4\big) \\
&= \sigma_\beta^2 \cdot n^2 m(n+m+c_4-2) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot \{1+o(1)\},
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \sigma_\beta^2 \cdot \mathrm{E}\big(\mathrm{tr}(\boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(2)} \boldsymbol{Z}_{(2)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)})\big) \\
&= \sigma_\beta^2 \cdot E(\sum_{c=1}^n \sum_{i=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{q=1}^{p-m} \sum_{r=1}^n Z_{(2)ck} Z_{(2)cq} X_{(2)lk} X_{(2)rq} X_{(1)li} X_{(1)ri}) \\
&= \sigma_\beta^2 \cdot n^2 m(p-m)
\end{aligned}
$$

$$
\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = 0
$$

Thus, we have

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X} \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{X}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot \{1+o(1)\} + \sigma_\beta^2 \cdot n^2 m(p-m)$$

Second moment of covariance term

$$\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)})\big) \\
&= \mathrm{E}\Big(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^n \sum_{q=1}^m \sum_{r=1}^m \sum_{s=1}^n \sum_{t=1}^m Z_{ck} Z_{dq} Z_{dr} Z_{ct} X_{lk} X_{li} X_{sr} X_{sj} \beta_q \beta_i \beta_t \beta_j\Big) \\
&= \{(n^4 m^2 + 4n^3 m^2 + n^2 m^3) \cdot \sigma_\beta^4 + mn^4 \cdot (b_4 - \sigma_\beta^4)\} \cdot \{1+o(1)\}
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)})\big) \\
&= \mathrm{E}\Big(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{q=1}^m \sum_{r=1}^{p-m} \sum_{s=1}^n \sum_{t=1}^m \\
&\qquad Z_{(2)ck} Z_{(2)dr} Z_{(1)dq} Z_{(1)ct} X_{(2)lk} X_{(2)sr} X_{(1)li} X_{(1)sj} \beta_q \beta_i \beta_t \beta_j\Big) \\
&= n^2(p-m)\{m^2 + (\sigma_\beta^4 - 1) \cdot m\} \cdot \sigma_\beta^4 = n^2 m^2 (p-m) \cdot \sigma_\beta^4 \cdot \{1+o(1)\}
\end{aligned}$$

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) = 0$$

It follows that

$$\begin{aligned}
&\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(2)} \boldsymbol{X}_{(2)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \big[\{n^2 m^3 + 4n^3 m^2 + n^2 m^2 (p-m)\} \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)\big] \cdot \{1+o(1)\}.
\end{aligned}$$

Second moment of variance term I

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)})\big) \\
&= \mathrm{E}\Big(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} X_{ci}X_{dj}X_{dk}X_{cl}\beta_i\beta_j\beta_k\beta_l\Big) \\
&= E\Big\{\sum_{c=d}^{n}\Big(\sum_{i=j\neq k=l}^{m(m-1)} X_{ci}^2X_{ck}^2\beta_i^2\beta_k^2 + \sum_{i=k\neq j=l}^{m(m-1)} X_{ci}^2X_{cj}^2\beta_i^2\beta_j^2 + \sum_{i=l\neq j=k}^{m(m-1)} X_{ci}^2X_{cj}^2\beta_i^2\beta_j^2 \\
&\qquad + \sum_{i=l=j=k}^{m} X_{ci}^4\beta_i^4\Big) + \sum_{c\neq d}^{n(n-1)}\Big(\sum_{i=l\neq k=j}^{m(m-1)} X_{ci}^2X_{dj}^2\beta_i^2\beta_j^2 + \sum_{i=l=k=j}^{m} X_{ci}^2X_{dj}^2\beta_i^4\Big)\Big\} \\
&= \sigma_\beta^4 \cdot n^2m^2 + nm \cdot \big\{2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4b_4 - 2\sigma_\beta^4 - b_4\big\},
\end{aligned}
$$

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

$$
\mathrm{Var}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{\beta}_{(1)}\big) = nm \cdot \{2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4b_4 - 2\sigma_\beta^4 - b_4\} = o(n^2m^2 \cdot \sigma_\beta^4).
$$

Second moment of variance term II

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(1)} \boldsymbol{X}_{(1)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)})\big) \\
&= \mathrm{E}\Big(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{h=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{i=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{n}\sum_{s=1}^{m}\sum_{t=1}^{m}\sum_{u=1}^{n}\sum_{w=1}^{m}\sum_{a=1}^{m}\sum_{b=1}^{n} \\
&\qquad Z_{ck}Z_{dq}Z_{dt}Z_{ca}X_{lk}X_{lh}X_{rq}X_{ri}X_{ut}X_{us}X_{ba}X_{bw}\beta_h\beta_i\beta_s\beta_w\Big)
\end{aligned}
$$

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

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

After tedious calculations, we have

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)\\
&= \sigma_\beta^4\cdot\{n^4m^2(n+m)^2\}\cdot\{1+o(1)\}.
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)\\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)})\big)\\
&= \mathrm{E}\Bigg(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{h=1}^{m}\sum_{k=1}^{p-m}\sum_{l=1}^{n}\sum_{i=1}^{m}\sum_{q=1}^{p-m}\sum_{r=1}^{n}\sum_{s=1}^{m}\sum_{t=1}^{p-m}\sum_{u=1}^{n}\sum_{w=1}^{m}\sum_{a=1}^{p-m}\sum_{b=1}^{n}\\
&\qquad Z_{(2)ck}Z_{(2)dq}Z_{(2)dt}Z_{(2)ca}X_{(2)lk}X_{(2)rq}X_{(2)ut}X_{(2)ba}X_{(1)lh}X_{(1)ri}X_{(1)us}X_{(1)bw}\beta_h\beta_i\beta_s\beta_w\Bigg)\\
&= \sigma_\beta^4\cdot\{n^4m^2(p-m)^2\}\cdot\{1+o(1)\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)\\
&= \mathrm{E}\big(\mathrm{tr}(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)})\big)\\
&= \mathrm{E}\Bigg(\sum_{c=1}^{n}\sum_{d=1}^{n}\sum_{h=1}^{m}\sum_{k=1}^{p-m}\sum_{l=1}^{n}\sum_{i=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{n}\sum_{s=1}^{m}\sum_{t=1}^{p-m}\sum_{u=1}^{n}\sum_{w=1}^{m}\sum_{a=1}^{m}\sum_{b=1}^{n}\\
&\qquad Z_{(2)ck}Z_{(1)dq}Z_{(2)dt}Z_{(1)ca}X_{(2)lk}X_{(2)ut}X_{(1)lh}X_{(1)rq}X_{(1)ri}X_{(1)us}X_{(1)ba}X_{(1)bw}\beta_h\beta_i\beta_s\beta_w\Bigg)\\
&= \sigma_\beta^4\cdot n(p-m)\big\{(m^3-3m^2+2m)(n^2+2n)+(m^2-m)(n^3-n^2+2(c_4-1)n+\\
&\qquad 4n^2+4(c_4-1)n+b_4n^2)+b_4mn^3\big\}\\
&= \sigma_\beta^4\cdot\{n^3m^2(p-m)(m+n)\}\cdot\{1+o(1)\}.
\end{aligned}
$$

Similarly, we have

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)$$
$$= o\{\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2\}$$

and

$$\mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)$$
$$= \mathrm{E}\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big) = 0.$$

It follows that

$$\mathrm{Var}\Big\{\big(\boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{Z}_{(1)}^T + \boldsymbol{\beta}_{(1)}^T\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(2)}\boldsymbol{Z}_{(2)}^T\big)\big(\boldsymbol{Z}_{(1)}\boldsymbol{X}_{(1)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)}\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}\big)\Big\}$$
$$= o\{\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2\}.$$

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 below for threshold-PRS.

Threshold-PRS

First moment of covariance term

$$
\begin{aligned}
C_1 &= \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \big(\boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \\
&= \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \big(\boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} + \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}\big) \\
&= \boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} = C_{11} + C_{12} + C_{13} + C_{14} + C_{15}
\end{aligned}
$$

Thus, $\mathrm{E}(C_1) = \mathrm{E}(C_{11}) = \mathrm{E}\big(\boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)}\big) = n^2 q_1$

First moment of variance term II

$$
\begin{aligned}
&\mathrm{Var}_2 = \\
&\big(\boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T + \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T\big) \cdot \\
&\big(\boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} + \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)}\big) \\
&= \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(11)} \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(11)}^T \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&+ \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} + \boldsymbol{\beta}_{(12)}^T \boldsymbol{X}_{(12)}^T \boldsymbol{X}_{(21)} \boldsymbol{Z}_{(21)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \\
&= a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2
\end{aligned}
$$

$\mathrm{E}(\mathrm{Var}_2)$

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

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

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

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

$= \{n^2 q_1(n + m) + n^2 q_2 m\} \cdot \{1 + o(1)\}$

Second moment of covariance term

$\mathrm{E}(C_1 C_1) = \mathrm{E}\big\{(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})\big\}$

$= \mathrm{E}\big(C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2\big)$

$\mathrm{E}(C_{11}^2) = \mathrm{E}\big(\boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} \boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)}\big)$

$= n^4 q_1^2 + \{(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1(b_4 - 1)\} \cdot \{1 + o(1)\}$

$\mathrm{E}(C_{12}^2) = \mathrm{E}\big(\boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \boldsymbol{\beta}_{(11)}^T \boldsymbol{Z}_{(11)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)}\big)$

$$= \mathrm{E}\Bigg(\sum_{c=1}^{n} \sum_{d=1}^{n} \sum_{i=1}^{m-q_1} \sum_{j=1}^{m-q_1} \sum_{k=1}^{q_1} \sum_{l=1}^{n} \sum_{q=1}^{q_1} \sum_{r=1}^{q_1} \sum_{s=1}^{n} \sum_{t=1}^{q_1} Z_{(11)ck} Z_{(11)dq} Z_{(11)dr} Z_{(11)ct} X_{(11)lk} X_{(11)sr} X_{(12)li} X_{(12)sj} \beta_{(11)q} \beta_{(11)t} \beta_{(12)i} \beta_{(12)j} \Bigg)$$

$= \{n^2 q_1^2(m - q_1) + n^3 q_1(m - q_1)\} \cdot \{1 + o(1)\}$

Similarly, we have

$\mathrm{E}(C_{13}^2) = \mathrm{E}\big(\boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)} \boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(11)} \boldsymbol{\beta}_{(11)}\big)$

$= \big\{n^2 q_1^2(m - q_1) + n^3 q_1(m - q_1)\big\} \cdot \{1 + o(1)\} = \mathrm{E}(C_{12}^2)$

$$\mathrm{E}(C_{14}^2) = \mathrm{E}\big(\boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)} \boldsymbol{\beta}_{(12)}^T \boldsymbol{Z}_{(12)}^T \boldsymbol{Z}_{(11)} \boldsymbol{X}_{(11)}^T \boldsymbol{X}_{(12)} \boldsymbol{\beta}_{(12)}\big)$$
$$= n^2(m-q_1)^2 q_1$$

$$\mathrm{E}(C_{15}^2) = \mathrm{E}(\boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} \boldsymbol{\beta}_{(1)}^T \boldsymbol{Z}_{(1)}^T \boldsymbol{Z}_{(21)} \boldsymbol{X}_{(21)}^T \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)}) = n^2 m^2 q_2.$$

Thus, we have

$$\mathrm{E}(C_1 C_1) = \mathrm{E}\big(C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2\big)$$
$$= n^4 q_1^2 + \{(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1 (b_4 - 1)\} \cdot \{1 + o(1)\}$$
$$+ 2 \cdot \{n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)\} \cdot \{1 + o(1)\} + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2.$$

It follows that

$$\mathrm{Var}(C_1) = \big\{n^2 q_1^3 + 4n^3 q_1^2 + n^4 q_1 (b_4 - 1) + 2n^2 q_1^2 (m - q_1) +$$
$$2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2\big\} \cdot \{1 + o(1)\}.$$

Second moment of variance term II

$$\mathrm{E}(\mathrm{Var}_2 \mathrm{Var}_2)$$
$$= \mathrm{E}\big\{(a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2)^2\big\}$$
$$= \mathrm{E}\big\{a^4 + b^4 + c^4 + d^4 + 2a^2b^2 + 2a^2c^2 + 2a^2d^2 + 2b^2c^2 + 2b^2d^2 + 2c^2d^2\big\}$$
$$= \mathrm{E}(a^4) + \mathrm{E}(b^4) + \mathrm{E}(c^4) + \mathrm{E}(d^4) + \mathrm{E}(2a^2b^2 + 2a^2c^2 + 2a^2d^2 + 2b^2c^2 + 2b^2d^2 + 2c^2d^2)$$
$$= \{n^2 q_1 (n + q_1)\}^2 + o(n^4 q_1^4 + n^5 q_1^3 + n^6 q_1^2) + \{n^2 (m - q_1) q_1\}^2 \cdot \{1 + o(1)\}$$
$$+ (n^2 q_1 q_2)^2 \cdot \{1 + o(1)\} + \{n^2 (m - q_1) q_2\}^2 \cdot \{1 + o(1)\} + o\{(n^2 q_1 (n + m) + n^2 q_2 m)^2\}$$

It follows that $\mathrm{Var}(\mathrm{Var}_2) = o\big[\{n^2 q_1 (n + m) + n^2 q_2 m\}^2\big]$.