STATISTICAL INFERENCE FOR HIGH DIMENSIONAL REGRESSION WITH PROXY DATA

Sai Li[1], T. Tony Cai[2] and Hongzhe Li[3]

Institute of Statistics and Big Data[1]

Renmin University of China

Department of Statistics and Data Science[2]

Department of Biostatistics, Epidemiology and Informatics[3]

University of Pennsylvania

Abstract: Existing high-dimensional statistical methods are largely developed for analyzing individual-level data. In this work, we study estimation and inference for high-dimensional linear models when only "proxy data" is available. These proxies encompass marginal statistics and sample covariance matrices computed from distinct sets of individuals. We develop a rate optimal method for estimation and inference for the regression coefficient vector and its linear functionals based on the proxy data. We show the intrinsic limitations in the proxy-data based inference: the minimax optimal rate for estimation is slower than that in the conventional case where individual data are observed. These interesting findings are illustrated through simulation studies and an analysis of a dataset concerning the genetic associations of hindlimb muscle weights in a mouse population.

1. Introduction

Large-scale genome-wide association studies (GWAS) provide opportunities for developing genetic risk prediction models that have the potential to improve disease prevention, intervention, and treatment. In epidemiology and genetics, there is a growing interest in utilizing the published summary statistics, especially those from GWAS, for disease risk prediction. The abundant summary data can enhance the power in signal detection using the framework of meta-analysis. Comparing with the individual-level data, the summary data are less privacy-sensitive and are more communication efficient for data sharing. However, statistical properties of learning based on the summary data remain largely unknown.

1.1 Problem formulation

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the genetic variants measurements in n independent individuals whose i-th row is $\boldsymbol{x}_i^\mathsf{T}$, where p is the dimension of genetic variants. Let $y \in \mathbb{R}^n$ denote the mean-adjusted response vector in this sample. In the linear

model for the association between the outcome and the covariates,

$$y_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i, \tag{1.1}$$

where $\mathbb{E}[\epsilon_i|\boldsymbol{x}_i] = 0$ and $\mathbb{E}[\epsilon_i^2|\boldsymbol{x}_i] = \sigma^2$, the goal is to estimate and infer the effect size vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and its functionals using only the summary data but not the individual-level data.

GWAS reports the marginal statistics

$$\widehat{S}_j = \boldsymbol{X}_{.,j}^{\mathsf{T}}\boldsymbol{y}/n, \ j = 1, \ldots, p$$

and their estimated standard errors. Besides the marginal statistics $\widehat{\boldsymbol{S}}$, an estimator of the covariance matrix of $\boldsymbol{x}_i$ is often needed for estimation and inference. One challenge is that the empirical covariance matrix for the samples involved in $\widehat{\boldsymbol{S}}$ is often not available because this genomic data set is too large or privacy-sensitive to share. That is, we do not observe $\widehat{\Sigma} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n$. A common practice is to obtain an estimate of covariance matrix $\Sigma = \mathbb{E}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n]$ from some external genome panel, such as the 1000 genome project (Consortium et al., 2015). Let $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^p, i = 1, \ldots, \tilde{n}$, denote the genotype of the observations in the exter-

nal data and define

$$\widetilde{\Sigma} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{x}_i \tilde{x}_i^\intercal.$$

We call $\tilde{x}_i$, $i = 1, \ldots, \tilde{n}$, the proxy data and $\widetilde{\Sigma}$ the proxy covariance matrix. In this work, we assume that $\mathbb{E}[\tilde{x}_i \tilde{x}_i^\intercal] = \Sigma$. That is, the proxy covariates have the same covariance structure as the covariates in computing the marginal statistics $\widehat{S}$. In genetic applications, the number of SNPs can be much larger than the sample sizes. Hence, we focus on the regime that p is larger or much larger than $\max\{n, \tilde{n}\}$. We assume that β is sparse with $\|\beta\|_0 = s$, i.e., β has s nonzero elements and $s \ll p$.

1.2 Motivating applications with the proxy data

Learning the linear model (1.1) with summary statistics has important applications in genomic studies. Polygenic risk score (PRS) regression concerns predicting a certain health-related outcome using the associated single nucleotide polymorphisms (SNPs), which can be formulated as a high-dimensional regression problem (Chen et al., 2021). Vilhjálmsson et al. (2015) introduces an Bayesian approach for PRS based on summary data. Mak et al. (2017) and Chen et al. (2021) both consider shrinkage methods as extensions of the Lasso. The method in Mak et al. (2017) is for linear models and the method in Chen et al. (2021) can deal with binary outcomes based on approximations. However, the

statistical guarantees and the choice of tuning parameters are largely unknown. In addition, inference for the linear functionals of high-dimensional regression coefficients based on proxy data has not been studied in literature. It is crucial in PRS prediction to provide confidence intervals for $\boldsymbol{x}_*^\intercal \boldsymbol{\beta}$ given a new individual's genomic information $\boldsymbol{x}_* \in \mathbb{R}^p$.

Besides, summary data provide the opportunity to combine multiple studies (e.g., cohorts) into one large study to increase the sample size (Albiñana et al., 2021), which is the goal of meta-analysis (Deelen et al., 2019). Hence, it is also of statistical interest to estimate and make inference of $\boldsymbol{\beta}$ with the proxy data.

Another application of the proxy-data based inference is distributed inference, where the whole data set contains *i.i.d.* observations but the data are distributed at multiple remote machines. Distributed algorithms estimate the target parameter by communicating some summary information across machines. To reduce the communication costs, the gradient vectors are communicated but not the high-dimensional Hessian matrix. The overall Hessian matrix can be approximated by a local matrix or by subsampling. See, for example, Jordan et al. (2018) and Wang and Zhang (2019). This type of distributed inference also falls in the category of proxy-data based inference.

1.3   Main results and our contribution

Motivated by aforementioned applications, we study proxy-data based statistical inference in high-dimensional linear models. When individual level data are available, estimation and inference for the regression coefficients have been extensively studied in high-dimensional linear models. Many penalized methods have been proposed for prediction, estimation, and variable selection in high-dimensional linear models, including Tibshirani (1996); Fan and Li (2001); Zou (2006); Candes and Tao (2007); Meinshausen and Bühlmann (2010); Zhang (2010). For statistical inference of individual coefficient, Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014) consider debiased estimators in linear models and generalized linear models. The minimaxity and adaptivity of confidence intervals have been studied in Cai and Guo (2017). Cai et al. (2021) and Javanmard and Lee (2020) propose methods for inference for a linear functional of the regression coefficients in linear models.

When only summary statistics or proxy data are available, statistical inference for high-dimensional linear regression models has not been addressed in the literature. The key challenge is that the samples used to compute the observed covariance matrix $\widetilde{\Sigma}$ and summary statistics $\widehat{S}$ are different. We highlight some of our key contributions.

Methodology-wise, we consider proxy-data based ℓ_1-penalized regression and prove that it is minimax optimal under typical regularity conditions. We further propose debiased estimators to make inference for β_j, $1 \leq j \leq p$ and $\boldsymbol{x}_*^\mathsf{T}\boldsymbol{\beta}$ with the proxy data, respectively. The debiased estimators are asymptotically normal and can be used to construct confidence intervals and for multiple testing under certain conditions. We also demonstrate that confidence interval length given by the debiased estimator of β_j has minimax optimal length under certain conditions.

Theoretically, we discover some interesting and new phenomena with the proxy-data based learning. First, the minimax rates for estimation are slower than the corresponding rates with individual data, even if $\tilde{n} \to \infty$. The relative loss gets larger when the signal-to-noise ratio gets larger. Second, comparing with the debiased Lasso based on individual data, the debiased Lasso estimator of β_j based on proxy data has larger bias and variance. Hence, its asymptotic normality requires different sample size conditions and simply treating the proxy summary data as individual-level data can lead to invalid inference.

1.4 Organization and notation

In Section 2, we describe the ℓ_1-regularized method for estimating $\boldsymbol{\beta}$ with high-dimensional proxy data and study its convergence rate and minimax optimality.

In Section 3, we construct the debiased estimator of β_j based on proxy data and study its limiting distribution In Section 4, we construct confidence interval for $\boldsymbol{x}_*^{\mathsf{T}}\boldsymbol{\beta}$ and provide theoretical guarantees. In Section 5, we study the empirical performance of our proposals via extensive numerical experiments. In Section 6, we apply the individual-level data Lasso and summary statistics Lasso to a GWAS in an outbred mice population. In Section 7, we discuss some other summary data motivated problems for future research. The proofs and other supplementary information are provided in the Supplementary Materials (Li et al., 2022).

Notation. For real-valued sequences $\{a_n\}, \{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \leq cb_n$ for some universal constant $c \in (0, \infty)$, and $a_n \gtrsim b_n$ if $a_n \geq c'b_n$ for some universal constant $c' \in (0, \infty)$. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. $c, C, c_0, c_1, c_2, \cdots$, and so on refer to universal constants in the paper, with their specific values possibly varying from place to place. For a vector $\boldsymbol{v} \in \mathbb{R}^d$ and a subset $S \subseteq [d]$, we use $\boldsymbol{v}_S$ to denote the restriction of vector $\boldsymbol{v}$ to the index set S. For a matrix $A \in \mathbb{R}^{n_1 \times n_2}$, let $\Lambda_{\max}(A)$ denote the largest singular value of A, $\Lambda_{\min}(A)$ denote the smallest singular value of A, and $\|A\|_{\infty,\infty}$ denote $\max_{i \leq n_1, j \leq n_2} |A_{i,j}|$. For a random variable $u \in \mathbb{R}$, define its sub-Gaussian norm as $\|u\|_{\psi_2} = \sup_{l \geq 1} l^{-1/2}\mathbb{E}^{1/l}[|u|^l]$. For a random vector $\boldsymbol{U} \in \mathbb{R}^n$, define its sub-Gaussian norm as $\|\boldsymbol{U}\|_{\psi_2} = \sup_{\|\boldsymbol{v}\|_2=1, \boldsymbol{v}\in\mathbb{R}^n} \|\langle \boldsymbol{U}, \boldsymbol{v}\rangle\|_{\psi_2}$. Let SNR $=$

$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2/\sigma^2$ denote the signal-to-noise ratio. Let τ_q denote the q-th quantile of standard normal distribution.

2. Estimation and prediction based on proxy data

In this section, we introduce our proposed estimators for prediction and estimation based on proxy data in Section 2.1. We study its theoretical properties and minimax optimality in Section 2.2.

2.1 Lasso based on two-sample summary data

For the estimation and prediction tasks, the methods for proxy data resemble high-dimensional regression methods with individual-level data. The Lasso estimator (Tibshirani, 1996) provides a rate optimal estimator of $\boldsymbol{\beta}$ in the conventional setting. Decomposing the empirical loss $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2$ as $\|\boldsymbol{y}\|_2^2 - 2\boldsymbol{b}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y} + \|\boldsymbol{X}\boldsymbol{b}\|_2^2$ and removing the constant term, the Lasso estimator can be written as

$$\widehat{\boldsymbol{\beta}}^{(os)} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\boldsymbol{b}^\intercal\widehat{\boldsymbol{\Sigma}}\boldsymbol{b} - \boldsymbol{b}^\intercal\widehat{\boldsymbol{S}} + \lambda^{(os)}\|\boldsymbol{b}\|_1\right\}$$

with some tuning parameter $\lambda^{(os)} > 0$ and the superscript "os" is short for "one-sample" or individual-level data. In fact, we have seen that the Lasso can be equivalently performed based on one-sample summary data $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{S}}$. Hence,

we will refer to $\widehat{\boldsymbol{\beta}}^{(os)}$ as one-sample Lasso for simplicity in the sequel.

With proxy data, it is natural to consider the following estimator

$$\widehat{\boldsymbol{\beta}}^{(ts)} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p}\{\frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{b} - \boldsymbol{b}^{\mathsf{T}}\widehat{\boldsymbol{S}} + \lambda^{(ts)}\|\boldsymbol{b}\|_1\}, \tag{2.1}$$

where we replace the unknown $\widehat{\boldsymbol{\Sigma}}$ with its proxy $\widetilde{\boldsymbol{\Sigma}}$ and consider a possibly different tuning parameter $\lambda^{(ts)}$. The superscript "ts" is short for "two-sample" or summary statistics data.

The summary statistics-based Lasso estimator, or two-sample Lasso estimator, $\widehat{\boldsymbol{\beta}}^{(ts)}$ has been considered in Mak et al. (2017) and Chen et al. (2021). However, the choice of $\lambda^{(ts)}$, the convergence rate, and minimax optimality have not been established. We provide the theoretical guarantees for $\widehat{\boldsymbol{\beta}}^{(ts)}$ in the next subsection.

2.2 Convergence rates for estimation and prediction

We assume the following conditions for theoretical analysis.

Condition 2.1 (Gaussian designs)**.** Each row of $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are i.i.d. Gaussian with mean zero and positive definite covariance $\boldsymbol{\Sigma}$ such that $c_1 \leq \Lambda_{\min}(\boldsymbol{\Sigma}) \leq \Lambda_{\max}(\boldsymbol{\Sigma}) \leq c_2$ for some positive constants $c_1 \leq c_2$.

Condition 2.2 (sub-Gaussian noises)**.** The random noises ϵ_i, $i = 1, \ldots, n$, are

i.i.d. with mean zero and variance $\sigma^2 > 0$. ϵ_i and $\boldsymbol{x}_i$ are independent for $i = 1, \ldots, n$. The sub-Gaussian norms of ϵ_i are upper bounded by a constant.

For estimation and prediction, it suffices to relax Conditions 2.1 and 2.2 to assume independent sub-Gaussian designs and independent sub-Gaussian noises. Here we assume slightly stronger regularity conditions, Gaussian designs and *i.i.d.* noises, which ensure that the asymptotic variance of the debiased estimators only depends on the first two moments of the observations. With individual samples, this assumption is not necessary because one can estimate the variance based on the empirical noises in heteroskedastic settings (Dezeure et al., 2017). In lack of the individual-level data, we cannot estimate the asymptotic variance empirically and have to rely on the properties of higher moments, which makes Gaussian distribution a natural assumption.

We first derive the rate of convergence for $\widehat{\boldsymbol{\beta}}^{(ts)}$ in the two-sample summary setting. Let $\mathbb{E}[y_i^2] = M$ and

$$\gamma_{n,\tilde{n}} = \sigma^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 (\frac{n}{\tilde{n}} + 1) = M + \frac{n}{\tilde{n}}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}. \tag{2.2}$$

Loosely speaking, $\gamma_{n,\tilde{n}}$ is the variance of the random noises based on proxy data. In Theorem 1, we establish the convergence rate of $\widehat{\boldsymbol{\beta}}^{(ts)}$ under mild conditions.

Theorem 1 (Convergence rates for $\widehat{\boldsymbol{\beta}}^{(ts)}$)**.** *Assume Conditions 2.1 and 2.2 and*

$Ms \log p \ll \min\{n, \tilde{n}\}$. *For* $\lambda^{(ts)} = c_1 \sqrt{\gamma_{n,\tilde{n}} \log p / n}$, *with large enough constant* c_1, *it holds that*

$$\|\widetilde{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta})\|_2^2 \vee \|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 \leq C \frac{\gamma_{n,\tilde{n}} s \log p}{n}$$

$$\|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_1 \leq Cs \sqrt{\frac{\gamma_{n,\tilde{n}} \log p}{n}}$$

with probability at least $1 - \exp(-c_2 \log p) - \exp(-c_3 \tilde{n})$ *for some positive constants* c_2 *and* c_3.

The one-sample optimal rates in squared ℓ_2-norm, which is $\sigma^2 s \log p / n$. We can see that the ratio of two rates (two-sample over one-sample) is

$$1 + \text{SNR}(\frac{n}{\tilde{n}} + 1), \tag{2.3}$$

where $\text{SNR} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 / \sigma^2$. This implies that the estimation error rate in the two-sample case is strictly worse than the one-sample case as long as $\text{SNR} > 0$. Larger n and larger SNR lead to larger relative loss with $\widehat{\boldsymbol{\beta}}^{(ts)}$ relative to $\widehat{\boldsymbol{\beta}}^{(os)}$. In contrast, larger $\tilde{n}$ implies smaller relative loss. As a result, the condition for consistency is no weaker in the proxy setting than that with one-sample data. The two-sample tuning parameter $\lambda^{(ts)} \asymp \sqrt{\gamma_{n,\tilde{n}} \log p / n}$, whose order is always no smaller than its one-sample counterpart. The choice of $\lambda^{(ts)}$ is crucial in practice

and cross validation cannot be performed without individual-level data. We will discuss some practical choices in Section 5 and Section C.5.

To better understand the unique challenges with proxy data, we consider a special scenario where $\tilde{n} \to \infty$, or equivalently, Σ is known.

Remark 1 (The scenario of $\tilde{n} \to \infty$). If $\tilde{n} \to \infty$, which is equivalent to observing $(\widehat{S}, \Sigma)$, then

$$\|\widehat{\boldsymbol{\beta}}^{(ts)} - \boldsymbol{\beta}\|_2^2 = O_P \left(\frac{Ms \log p}{n} \right).$$

Remark 1 shows that even if $\tilde{n} \to \infty$, the convergence rate of $\widehat{\boldsymbol{\beta}}^{(ts)}$ is still inflated when SNR > 0 in comparison to having one-sample data. This comparison implies that, without the in-sample $\widehat{\Sigma}$, any estimator of Σ, even the oracle one, can lead to dramatic loss in estimation accuracy. Comparing Remark 1 with Theorem 1, we see that the error caused by finite external data is of order $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} s \log p / \tilde{n}$.

We now show that the convergence rate of $\widehat{\boldsymbol{\beta}}^{(ts)}$ is minimax optimal in ℓ_2-norm. Consider the parameter space

$$\Xi(s, M_0, \sigma_0^2) = \big\{ \|\boldsymbol{\beta}\|_0 \leq s, \boldsymbol{\beta}^\mathsf{T} \Sigma \boldsymbol{\beta} \leq M_0, \ 0 < \sigma^2 \leq \sigma_0^2,$$

$$0 < 1/C_1 \leq \Lambda_{\min}(\Sigma) \leq \Lambda_{\max}(\Sigma) \leq C_1 < \infty \big\} \qquad (2.4)$$

for some constant $C_1 > 1$ and σ_0^2 can be any positive constant. Let $\mathcal{Z} = \{\widehat{S}, \widetilde{\Sigma}\}$

denote the observed data and $\mathcal{F}(\mathcal{Z})$ denote functions based on the summary data $\mathcal{Z}$.

Theorem 2 (Lower bound for estimating $\boldsymbol{\beta}$). *Consider the parameter space* $\Xi(s, M_0, \sigma_0^2)$ *in (2.4) with* $s \geq 2$. *Suppose that* $Ms \log p \ll n$, *and* $(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \vee 1)s \log p \ll \tilde{n}$. *Then there exists some constant* c_1 *that*

$$\min_{\widehat{\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \Xi(s, M_0, \sigma_0^2)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \geq \frac{c_1(M_0 + \sigma_0^2)s \log p}{n} + \frac{c_1 M_0 s \log p}{\tilde{n}} \right) \geq 1/2.$$

In the parameter space $\Xi(s, M_0, \sigma_0^2)$, it holds that $M = \mathbb{E}[y_i^2] \leq M_0 + \sigma_0^2$. Hence, the lower bound in Theorem 2 matches the ℓ_2-upper bound in Theorem 1 in terms of rates. We mention that the sample size condition in Theorem 2 essentially restricts us to a class of $\widehat{\boldsymbol{S}}$ with distributional regularity, i.e., its distribution conditioning on $\boldsymbol{y}$ has positive definite covariance matrix. As far as we know, this is the first lower bound result based on summary data and the proof is based on some novel analysis of the distribution of the marginal correlation statistics.

3. Inference for individual coefficient based on proxy data

In this section, we consider statistical inference, such as hypothesis testing and constructing confidence intervals for β_j with some fixed $1 \leq j \leq p$. It is known

that the ℓ_1-regularized estimates are biased and cannot be directly used for inference. For inference based on proxy data, we follow a similar idea as the debiased methods, which have been proposed based on individual-level data. Specifically, the debiased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014) can be written as

$$\hat{\beta}_j^{(os-db)} = \hat{\beta}_j^{(os)} + \frac{(\boldsymbol{X}\widehat{\boldsymbol{w}}_j)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(os)})}{n} = \hat{\beta}_j^{(os)} + \widehat{\boldsymbol{w}}_j^{\mathsf{T}}(\widehat{\boldsymbol{S}} - \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(os)}), \quad (3.1)$$

where $\widehat{\boldsymbol{\beta}}^{(os)}$ is the one-sample Lasso estimator and $\widehat{\boldsymbol{w}}_j \in \mathbb{R}^p$ is a correction score vector that can be computed based on $\widehat{\boldsymbol{\Sigma}}$. We see that the debiased Lasso for β_j can also be realized based on the summary statistics $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{S}}$. Hence, we refer to the estimate in (3.1) as one-sample debiased Lasso (os-db) in the sequel. This similarly motivates its counterpart with two-sample summary data:

$$\hat{\beta}_j^{(ts-db)} = \hat{\beta}_j^{(ts)} + \tilde{\boldsymbol{w}}_j^{\mathsf{T}}(\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}), \qquad (3.2)$$

where $\widehat{\boldsymbol{\beta}}^{(ts)}$ is computed in (2.1) and $\tilde{\boldsymbol{w}}_j \in \mathbb{R}^p$ is a correction score vector computed based on $\widetilde{\boldsymbol{\Sigma}}$. Specifically, we consider

$$\tilde{\boldsymbol{w}}_j = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg \min} \|\boldsymbol{w}\|_1 \tag{3.3}$$

$$\text{subject to } \|\widetilde{\boldsymbol{\Sigma}}\boldsymbol{w} - \boldsymbol{e}_j\|_\infty \leq \lambda_j,$$

where $\lambda_j = c_j\sqrt{\log p/\tilde{n}}$ for some positive constant c_j. The realization of $\tilde{\boldsymbol{w}}_j$ is via a Dantzig selector optimization (Candes and Tao, 2007), which induces a sparse solution of the j-th column of the inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Some existing one-sample methods, such as Javanmard and Montanari (2014), do not look for a sparse estimate $\tilde{\boldsymbol{w}}_j$ but choose a different objective function in (3.3). In the proxy setting, however, the sparsity of $\tilde{\boldsymbol{w}}_j$ plays a crucial role in the analysis. Those non-sparse methods for one-sample setting cannot be directly generalized for the current purpose as we will further discuss in Section 3.1.

3.1 Asymptotic normality for debiased two-sample Lasso

We study the asymptotic property of $\hat{\beta}_j^{(ts-db)}$ defined in (3.2) and prove its asymptotic normality under certain conditions. Let $\boldsymbol{\Omega}_{.,j}$ denote the j-th column of $\boldsymbol{\Omega}$ and $s_j = \|\boldsymbol{\Omega}_{.,j}\|_0$.

Theorem 3 (Asymptotic normality of the debiased estimator)**.** *Assume that Con-*

dition 2.1 and Condition 2.2 hold, $n \gg \log p$, and $\tilde{n} \gg (s \vee s_j) \log p$. If $\lambda_j = c_j \sqrt{\log p / \tilde{n}}$ and $c_j \geq c_1 \sqrt{\Sigma_{j,j} \Omega_{j,j}}$ for some large enough constant $c_1 > 0$, then it holds that

$$\hat{\beta}_j^{(ts-db)} - \beta_j = z_j + O_P\left(\gamma_{n,\tilde{n}}^{1/2} \frac{(s+s_j)\log p}{\sqrt{n\tilde{n}}}\right), \tag{3.4}$$

where $\gamma_{n,\tilde{n}}$ is defined in (2.2) and

$$(V_j^{(ts)})^{-1/2} z_j \xrightarrow{D} N(0,1) \ for \ V_j^{(ts)} = \frac{\Omega_{j,j}\gamma_{n,\tilde{n}}}{n} + \frac{\beta_j^2}{n} + \frac{\beta_j^2}{\tilde{n}}.$$

Further assuming $(s \vee s_j)\log p \ll \sqrt{\tilde{n}}$, then $(V_j^{(ts)})^{-1/2}(\hat{\beta}_j^{(ts-db)} - \beta_j) \xrightarrow{D} N(0,1)$.

Theorem 3 establishes the asymptotic distribution of $\hat{\beta}_j^{(ts-db)}$ in (3.4) and provides the sample size condition for its asymptotic normality. The variance of $\hat{\beta}_j^{(ts-db)}$ is $V_j^{(ts)}$ and the remaining bias of $\hat{\beta}_j^{(ts-db)}$ is shown in the last term on the right hand side of (3.4).

We first bring some details into the magnitude of $V_j^{(ts)}$. The last two terms of $V_j^{(ts)}$, β_j^2/n and $\beta_j^2/\tilde{n}$, are dominated by the first term of $V_j^{(ts)}$ given the positive definiteness of Ω. Hence, when $\tilde{n} \gg n$, $V_j^{(ts)} \asymp \Omega_{j,j} M/n$; when $n \gg \tilde{n}$, $V_j^{(ts)} \asymp \Omega_{j,j}\|\Sigma^{1/2}\beta\|_2^2/\tilde{n}$. In comparison to its one-sample counterpart, $V_j^{(os)} =$

$\Omega_{j,j}\sigma^2/n$, the relative loss in efficiency is

$$\frac{V_j^{(ts)}}{V_j^{(os)}} \asymp 1 + \text{SNR}(\frac{n}{\tilde{n}} + 1),$$

which is identical to the relative loss in estimation (2.3). When $\tilde{n} \to \infty$, i.e. Σ is known, $V_j^{(ts)}$ is still larger than $V_j^{(os)}$. This shows the significant loss in efficiency for inference problems when the marginal statistics and covariance estimator are not based on the same set of samples. More importantly, the distinction between $V_j^{(ts)}$ and $V_j^{(os)}$ implies that simply applying the inference algorithms for individual level data to the proxy summary statistics data could be wrong.

The remaining bias of $\hat{\beta}_j^{(ts-db)}$ is of order $\gamma_{n,\tilde{n}}^{1/2}(s + s_j)\log p/\sqrt{n\tilde{n}}$. In one-sample setting, the remaining bias of debiased $\hat{\beta}_j^{(os-db)}$ in (3.1), is of order $s \log p/n$ (van de Geer et al., 2014; Cai and Guo, 2017). If $\tilde{n}/n$ is sufficiently large, the remaining bias of $\hat{\beta}_j^{(ts-db)}$ can be **smaller** than that of its one-sample counterpart. In view of the asymptotic bias and asymptotic variance in (3.4), it suffices to require $(s \vee s_j)\log p \ll \sqrt{\tilde{n}}$ for asymptotic normality. This condition implies that $\tilde{n}$ determines the range of sparsity such that valid inference can be established. In contrast, for one-sample debiased Lasso $\hat{\beta}_j^{(os-db)}$, n determines the range of sparsity for valid inference, which is $s \log p \ll \sqrt{n}$. This can be a blessing of proxy-data scenario. An extreme case is when Σ is known, or

equivalently $\tilde{n} \to \infty$ as in the following remark.

Remark 2 (The scenario of $\tilde{n} \to \infty$)**.** When $\boldsymbol{\Sigma}$ is known, $\tilde{\boldsymbol{w}}_j = \boldsymbol{\Omega}_{.,j}$ and $\hat{\beta}_j^{(ts-db)} = \boldsymbol{\Omega}_{j,.} \boldsymbol{X}^\intercal \boldsymbol{y}/n$, which is asymptotically normal with mean zero and variance $\Omega_{j,j} \gamma_{n,\infty}/n + \beta_j^2/n$.

It may be surprising to see that for fixed n and p, the remaining bias of $\hat{\beta}_j^{(ts-db)}$ vanishes when $\tilde{n} \to \infty$. However, many existing applications often have $\tilde{n} \lesssim n$. This can be due to the less cost of sharing GWAS statistics than sharing the LD matrix. Same pattern holds for distributed inference, in which case n is the total sample size and $\tilde{n}$ is the local sample size. This should raise some caution in applications with two-sample summary data.

We finally discuss the conditions on the sparsity s_j. In classical one-sample setting, inference for β_j may not require sparse $\boldsymbol{\Omega}_{.,j}$, see, for example, the analysis in Javanmard and Montanari (2014) for linear models. We mention that the condition on s_j cannot be removed using the same idea in our analysis. This comes from a unique challenge of proxy data, where $\widehat{\boldsymbol{S}}$ implicitly depends on $\widehat{\boldsymbol{\Sigma}}$, which is unobserved but approximated. While the condition on s_j can be avoided by sample splitting, the splitting scheme is not viable with summary data in most cases. Hence, we focus on the current procedure and the results without sample splitting.

In the next theorem, we establish the minimax lower bound for estimating

β_j.

Theorem 4 (Minimax lower bound for estimation of β_j). *Consider the parameter space $\Xi(s, M_0, \sigma_0^2)$ in (2.4). Suppose that $\max\{1, M_0 + \sigma_0^2\} \leq c_1 \min\{n, \tilde{n}\}$ for some constant $c_1 > 0$. Then for any fixed $1 \leq j \leq p$, there exists some constant c_2 that*

$$\inf_{\hat{\beta}_j \in \mathcal{F}(\mathcal{Z})} \sup_{\boldsymbol{\beta} \in \Xi(s, M_0, \sigma_0^2)} \mathbb{P}\left(|\hat{\beta}_j - \beta_j| \geq c_2 \sqrt{\frac{M_0 + \sigma_0^2}{n}} + c_2 \sqrt{\frac{M_0}{\tilde{n}}}\right) \geq \frac{1}{2}.$$

In Theorem 4, we show that the parametric part of the rate for $\hat{\beta}_j^{(ts-db)}$ is minimax optimal. That is, under the sample size condition $(s \vee s_j) \log p \ll \sqrt{\tilde{n}}$, the two-sample debiased estimator $\hat{\beta}_j^{(ts-db)}$ has rate optimal confidence interval length. Comparing with the minimax rate for one-sample inference, we see that the variance part are inflated with proxy data. For the nonparametric part, the proof based on summary statistics is much more involved. In the supplements (Theorem A.1), we provide the minimax lower bound for estimating β_j when $\boldsymbol{\Sigma}$ is known and the lower bound matches the upper bound derived in Remark 2.

3.2   Variance estimator and confidence intervals

In view of $V_j^{(ts)}$, we propose a variance estimator for $\hat{\beta}_j^{(ts-db)}$ as

$$\widehat{V}_j^{(ts)} = \tilde{\boldsymbol{w}}_j^{\intercal} \widetilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{w}}_j \big(\frac{\|\boldsymbol{y}\|_2^2}{n^2} + \frac{2(\widehat{\boldsymbol{\beta}}^{(ts)})^{\intercal} \widehat{\boldsymbol{S}} - (\widehat{\boldsymbol{\beta}}^{(ts)})^{\intercal} \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}^{(ts)}}{\tilde{n}} \big)$$
$$+ \frac{(\hat{\beta}_j^{(ts-db)})^2}{n} + \frac{(\hat{\beta}_j^{(ts-db)})^2}{\tilde{n}}. \tag{3.5}$$

Notice that $\widehat{V}_j^{(ts)}$ is **not** the two-sample analogy of variance estimator for the classical debiased Lasso. This is because the probabilistic limit of $\widehat{V}_j^{(ts)}$ is asymptotically larger than the asymptotic variance of the conventional debased Lasso. Hence, if we treat proxy data as one-sample summary data, correct coverages are not guaranteed. We propose the following $(1 - \alpha) \times 100\%$-confidence interval for β_j as

$$\hat{\beta}_j^{(ts-db)} \pm \tau_{\alpha/2} \sqrt{\widehat{V}_j^{(ts)}}. \tag{3.6}$$

Once the z-statistics $z_j^{(ts)} = \hat{\beta}_j^{(ts-db)} / \sqrt{\widehat{V}_j^{(ts)}}$ is obtained for $j = 1, \ldots, p$, we can perform multiple testing with FDR control using the procedure in Javanmard and Javadi (2019), which is a refined version based on Liu (2013).

In the next lemma, we prove the consistency of $\widehat{V}_j^{(ts)}$ defined in (3.5) and conclude the validness of the confidence interval (3.6).

Lemma 1 (A consistent variance estimator). *Under the conditions of Theorem*

3,

$$\frac{|\widehat{V}_j^{(ts)} - V_j^{(ts)}|}{V_j^{(ts)}} = O_P\left(\frac{\gamma_{n,\tilde{n}}}{\sqrt{n}} + \gamma_{n,\tilde{n}}\frac{s\log p}{n} + \gamma_{n,\tilde{n}}\frac{s_j\log p}{\tilde{n}} + \frac{1}{\sqrt{\tilde{n}}}\right).$$

To summarize, assuming Condition 2.1, Condition 2.2, and $(s \vee s_j)\log p \ll \sqrt{\tilde{n}}$,
then $(\widehat{V}_j^{(ts)})^{-1/2}(\hat{\beta}_j^{(ts-db)} - \beta_j) \xrightarrow{D} N(0,1)$.

In the Supplementary Materials (Section B), we also study the power function of two-sample test statistics and demonstrate its potential power loss compared with one-sample setting.

4. Inference for linear functionals

We now study statistical inference for the PRS $\mu_* = \boldsymbol{x}_*^\mathsf{T}\boldsymbol{\beta}$ given an individual's feature $\boldsymbol{x}_*$. We can use $\widehat{\mu}_* = \boldsymbol{x}_*^\mathsf{T}\widehat{\boldsymbol{\beta}}$ for prediction. Hence, we focus on constructing confidence intervals for μ_*. Inference for linear functionals of $\boldsymbol{\beta}$ have been studied in the classical setting. The minimax rate is established in Cai and Guo (2017) and various methods are established in Cai and Guo (2017), Cai et al. (2021), and Javanmard and Lee (2020). All the afore-mentioned methods consider the debiasing recipe: the correction scores are obtained by constrained minimizations, where the constraints can be directly used to upper bound the the remaining bias of the debiased estimator. Our problem is more challenging as

some uncertainty coming from the unobserved covariance matrix $\widehat{\Sigma}$ cannot be directly controlled based on the observed data.

Trading-off multiple sources of bias, we consider a different method. For $\tilde{\boldsymbol{w}}_j$ defined in (3.3), denote

$$\widetilde{\boldsymbol{\Omega}} = (\tilde{\boldsymbol{w}}_1, \ldots, \tilde{\boldsymbol{w}}_p) \in \mathbb{R}^{p \times p}. \tag{4.1}$$

In fact, $\widetilde{\boldsymbol{\Omega}}$ is an estimate of the inverse covariance matrix. Our estimated $\widetilde{\boldsymbol{\Omega}}$ is equivalent to the CLIME estimator (Cai et al., 2011), which can be expressed as

$$\widetilde{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}}{\arg \min} \|\boldsymbol{\Omega}\|_1$$

$$\text{subject to } \|\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - I_p\|_{\infty,\infty} \leq \tilde{\lambda},$$

where $\tilde{\lambda} = c_1 \sqrt{\log p / \tilde{n}}$ for some positive constant c_1. We then obtain an initial bias-correction score $\widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*$. Next, we refine $\widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*$ to reduce the bias in the direction of $\boldsymbol{x}_*$:

$$\tilde{\boldsymbol{w}}_* = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg \min} \|\boldsymbol{w} - \widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*\|_1 \tag{4.2}$$

$$\text{subject to } \|\widetilde{\boldsymbol{\Sigma}}\boldsymbol{w} - \boldsymbol{x}_*\|_\infty \leq \|\boldsymbol{x}_*\|_2 \tilde{\lambda}.$$

The optimization in (4.2) can be efficiently solved, because it equivalently computes the Dantzig selector (Candes and Tao, 2007) by treating $\boldsymbol{w} - \widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*$ as the target parameter.

For $\hat{\boldsymbol{\beta}}^{(ts)}$ defined in (2.1) and $\tilde{\boldsymbol{w}}_*$ defined in (4.2), define the debiased estimator for $\mu_* = \boldsymbol{x}_*^{\mathsf{T}}\boldsymbol{\beta}$ as

$$\hat{\mu}_*^{(ts-db)} = \boldsymbol{x}_*^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(ts)} + \tilde{\boldsymbol{w}}_*^{\mathsf{T}}(\widehat{\boldsymbol{S}} - \widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}). \tag{4.3}$$

Some more comments on $\tilde{\boldsymbol{w}}_*$ defined in (4.2) are warranted. Our proposed $\tilde{\boldsymbol{w}}_*$ distinguishes from the constrained minimizations based on one-sample individual data, say, expressions (7) and (8) of Cai et al. (2021) or expression (12) of Javanmard and Lee (2020), directly control the bias in the direction of $\boldsymbol{x}_*$ based on the observed $\widehat{\boldsymbol{\Sigma}}$. Their correction scores have no sparse guarantees and their objective functions are quadratic. In the proxy setting, the analysis for debiasing has more remainder terms to control, which involve the discrepancy between the observed $\widetilde{\boldsymbol{\Sigma}}$ and the unobserved $\widehat{\boldsymbol{\Sigma}}$. To control the bias term involving $\widehat{\boldsymbol{\Sigma}}$, we rely on the sparsity of the precision matrix $\boldsymbol{\Omega}$. The estimator $\tilde{\boldsymbol{w}}_*$ in (4.2) can simultaneously leverage the sparsity structure of $\boldsymbol{\Omega}$ and control the bias in the direction of $\boldsymbol{x}_*$.

4.1   Asymptotic normality for summary statistics debiased μ_*

We study the theoretical properties of $\hat{\mu}_*^{(ts-db)}$ defined in (4.3). The theoretical analysis for a generic linear functional is more challenging, because $\boldsymbol{x}_*$ is non-sparse in general while canonical basis $\boldsymbol{e}_j$ has sparsity one. Let $s_\Omega = \max_{j \leq p} \|\boldsymbol{\Omega}_{.,j}\|_0$.

Theorem 5 (Asymptotic normality of $\hat{\mu}_*^{(ts-db)}$). *Assume Condition 2.1 and Condition 2.2 hold true, $n \gg \log p$ and $\tilde{n} \gg (s \vee s_\Omega) \log p$. It holds that*

$$\hat{\mu}_*^{(ts-db)} - \mu_* = z_* + O_P\left(\gamma_{n,\tilde{n}}^{1/2}\|\boldsymbol{x}_*\|_2 \frac{(s + s_\Omega^{3/2})\log p}{\sqrt{n\tilde{n}}}\right),$$

where $\gamma_{n,\tilde{n}}$ is defined in (2.2) and

$$(V_*^{(ts)})^{-1/2} z_* \xrightarrow{D} N(0,1) \ \text{for} \ V_*^{(ts)} = \frac{\boldsymbol{x}_*^\intercal \boldsymbol{\Omega} \boldsymbol{x}_* \gamma_{n,\tilde{n}}}{n} + \frac{\mu_*^2}{n} + \frac{\mu_*^2}{\tilde{n}}.$$

Hence, given that

$$s \log p \ll \sqrt{\tilde{n}} \ \text{and} \ s_\Omega^{3/2} \log p \ll \sqrt{\tilde{n}}, \tag{4.4}$$

$(V_*^{(ts)})^{-1/2}(\hat{\mu}_*^{(ts-db)} - \mu_*) \xrightarrow{D} N(0,1).$

Theorem 5 establishes the limiting distribution of $\hat{\mu}_*^{(ts-db)}$ and the asymp-

totic normality for $\hat{\mu}_*^{(ts-db)}$ given (4.4). The sparsity condition on s is the same as in Section 3.1 but the condition on s_Ω is stricter. This comes from the challenge of dealing with a non-sparse loading $\boldsymbol{x}_*$. We now connect Theorem 5 with the method in (4.2). We see that the remaining bias of $\hat{\mu}_*^{(ts-db)}$ depends on the sparsity of $\boldsymbol{\Omega}$, s_Ω. We leverage the sparsity of $\boldsymbol{\Omega}$ by first initialize $\widetilde{\boldsymbol{\Omega}}\boldsymbol{x}_*$ and compute $\tilde{\boldsymbol{w}}_*$ as its projection to the ℓ_∞-constrained space. Again, the number of proxy data $\tilde{n}$ determines the range of sparsity condition for constructing confidence intervals. The asymptotic variance $V_*^{(ts)}$ is determined by n and $\tilde{n}$ simultaneously.

We introduce the variance estimator of $\hat{\mu}_*^{(ts-db)}$, which is

$$\widehat{V}_*^{(ts)} = \tilde{\boldsymbol{w}}_*^\intercal\widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}_*\left(\frac{\|\boldsymbol{y}\|_2^2}{n^2} + \frac{2(\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal\widehat{\boldsymbol{S}} - (\widehat{\boldsymbol{\beta}}^{(ts)})^\intercal\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}^{(ts)}}{\tilde{n}}\right)$$
$$+ \frac{(\hat{\mu}_*^{(ts-db)})^2}{n} + \frac{(\hat{\mu}_*^{(ts-db)})^2}{\tilde{n}}. \tag{4.5}$$

We can similarly show that $\widehat{V}_*^{(ts)}$ defined in (4.5) is a consistent estimator of $V_*^{(ts)}$. Hence, we propose the following $(1 - \alpha) \times 100\%$-confidence interval for μ_* as

$$\hat{\mu}_*^{(ts-db)} \pm \tau_{\alpha/2}\sqrt{\widehat{V}_*^{(ts)}}. \tag{4.6}$$

Prior to the PRS prediction, an important question is whether the SNPs are predictive for the trait of interest or whether the trait is heritable. Therefore,

it is of interest to test $\mathcal{H}_0 : \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} = 0$ versus $\mathcal{H}_1 : \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} > 0$ based on two-sample summary data. This is equivalent to testing whether the covariance between the trait value y_i and the PRS $\boldsymbol{\beta}^T \boldsymbol{x}_i$, $\mathrm{Cov}(y_i, \boldsymbol{\beta}^T \boldsymbol{x}_i)$, is zero. We develop an asymptotically valid test in Section D of the supplements.

5. Numerical results

In this section, we evaluate the empirical performance of the procedures developed in previous sections. A practical issue is the choice of tuning parameter $\lambda^{(ts)}$. Without the individual data, cross-validation cannot be used. Alternative strategies include using some information criteria such as Bayesian information criterion (BIC). In the supplementary files (Section C.5), we compare three methods for choosing tuning parameters, including BIC, resampling-based criteria, and pseudo-validation (Mak et al., 2017). Results show that resampling-based tuning parameter selection gives to smallest prediction errors in almost all the scenarios which is adopted in all of our numerical studies. The variance of $\hat{\beta}_j^{(ts-db)}$ depends on M, the second moment of y_i. In fact, M can be approximated from the variance estimators of $\widehat{\boldsymbol{S}}$. Specifically, we can show that

$$\mathrm{Var}(\widehat{\boldsymbol{S}}_j) = \mathrm{Var}\left(\frac{\boldsymbol{X}_j^{\mathsf{T}} \boldsymbol{y}}{n \boldsymbol{\Sigma}_{j,j}} \right) (1 + o(1)) = \frac{\boldsymbol{\Sigma}_{j,j} M + (\boldsymbol{\Sigma}_{j,S} \boldsymbol{\beta}_S)^2}{n \boldsymbol{\Sigma}_{j,j}^2} (1 + o(1)).$$

When the correlation between $\boldsymbol{x}_{\cdot,j}$ $j \notin S$ and $\boldsymbol{X}_{\cdot,S}$ is sparse, say, $|\{j \leq p :$ $\boldsymbol{\Sigma}_{j,S} \neq 0\}| \ll p$, then a consistent estimate of M is $\widehat{M} = \frac{n}{p} \sum_{j=1}^{p} \widehat{var}(\widehat{\boldsymbol{S}}_j) \widetilde{\boldsymbol{\Sigma}}_{j,j}$.

We consider $(n, \tilde{n}) \in \{(200, 400), (300, 300), (400, 200)\}$, which corresponds to $n/\tilde{n} \in \{0.5, 1, 2\}$. We consider $p = 2000$ and $s \in \{4, 8, 12\}$. For the design matrix, we consider two different settings. In the first setting, we simulate independent genotypes and allele frequency for each SNP is randomly chosen from $(0.1, 0.2, 0.3)$. In the second setting, we simulate correlated genotypes from outbred Carworth Farms White (CFW) mice population (Parker et al., 2016) which will be further used in the data analysis. Details of the data generation are given in the supplementary materials (Section C.1). We denote these two settings of the design matrices as "ind" and "cfw", respectively.

5.1 Estimation and prediction results

In Table 1, we report the estimation and prediction results with one-sample and two-sample Lasso. for a generic estimator $\boldsymbol{b} \in \mathbb{R}^p$, define its estimation error and test errors as $\|\boldsymbol{b} - \boldsymbol{\beta}\|_2^2$ and $\|\widetilde{X}(\boldsymbol{b} - \boldsymbol{\beta})\|_2^2 / \|\widetilde{X}\boldsymbol{\beta}\|_2^2$, respectively. As one-sample method only uses n individual samples, its estimation errors and test errors decrease as n increases for any given s. For the two-sample Lasso, the test and estimation errors are always larger than those with one-sample Lasso. When $n/\tilde{n}$ is fixed, both methods have errors increasing as the sparsity s increases.

Table 1: Sum of squared errors for estimating $\boldsymbol{\beta}$ and test errors based on one-sample Lasso (OS) and two-sample Lasso (TS) for $n/\tilde{n} \in \{1/2, 1, 2\}$, where the SNP genotypes are either independently simulated or sampled from the CFW mice genotypes (cfw).

SNPs	s	Method	Sum of Squared Errors			Test Errors		
			1/2	1	2	1/2	1	2
ind	4	OS	0.029	0.022	0.018	0.641	0.478	0.375
		TS	0.039	0.034	0.025	0.828	0.652	0.502
	8	OS	0.047	0.037	0.032	0.824	0.641	0.540
		TS	0.059	0.053	0.044	0.982	0.804	0.644
	12	OS	0.060	0.050	0.044	0.895	0.745	0.646
		TS	0.077	0.070	0.060	1.069	0.893	0.739
cfw	4	OS	0.055	0.036	0.028	0.404	0.265	0.207
		TS	0.066	0.066	0.077	0.632	0.612	0.707
	8	OS	0.154	0.116	0.093	0.594	0.429	0.340
		TS	0.207	0.181	0.134	0.936	0.750	0.542
	12	OS	0.266	0.211	0.177	0.721	0.540	0.440
		TS	0.334	0.301	0.226	1.008	0.827	0.605

The errors of two-sample Lasso increase more significantly. This is because, according to Theorem 1, as s increases, SNR increases which leads to larger estimation errors.

In Table 2, we evaluate the performance of statistical inference for the regression coefficients based on one-sample and two-sample debiased Lasso. We see that the one-sample and two-sample debiased Lasso have coverage probabilities close to the nominal level in various configurations. The standard deviations based on proxy data increase as s increases and are larger than those based on one-sample data.

Table 2: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso for $n/\tilde{n} \in \{1/2, 1, 2\}$, where the SNP genotypes are either independently simulated or sampled from the CFW mice genotypes (cfw).

SNPs	s	Method	Average Coverage			Average SD		
			1/2	1	2	1/2	1	2
ind	4	OS	0.964	0.965	0.965	0.03	0.03	0.02
		TS	0.979	0.983	0.988	0.04	0.04	0.04
	8	OS	0.971	0.972	0.972	0.04	0.03	0.03
		TS	0.982	0.986	0.988	0.05	0.04	0.04
	12	OS	0.976	0.977	0.977	0.04	0.04	0.03
		TS	0.985	0.988	0.989	0.06	0.05	0.05
cfw	4	OS	0.950	0.950	0.950	0.04	0.03	0.03
		TS	0.957	0.948	0.939	0.05	0.04	0.04
	8	OS	0.950	0.950	0.950	0.06	0.05	0.04
		TS	0.956	0.948	0.939	0.07	0.06	0.05
	12	OS	0.949	0.950	0.950	0.07	0.06	0.05
		TS	0.956	0.948	0.939	0.08	0.07	0.06

We report the inference results of the true PRS μ_* in Table 3. The coverage probabilities are close to the nominal level in different settings for the two-sample method. When $s = 12$, the one-sample debiased Lasso has low coverage when n is small with independent SNPs. This agrees with our analysis on the sample size conditions for asymptotic normality: when $\tilde{n}$ is larger than n, the two-sample debiased Lasso requires a weaker sample size condition for asymptotic normality.

Table 3: Coverage probabilities and standard deviations based on one-sample (OS) and two-sample (TS) debiased Lasso for μ_* for $n/\tilde{n} \in \{1/2, 1, 2\}$, where the SNP genotypes are either independently simulated or sampled from the CFW mice genotypes (cfw).

SNPs	s	Method	Average Coverage			Average SD		
			1/2	1	2	1/2	1	2
ind	4	OS	0.913	0.920	0.948	0.15	0.14	0.17
		TS	0.923	0.996	0.988	0.46	0.30	0.21
	8	OS	0.907	0.896	0.939	0.18	0.16	0.21
		TS	0.925	0.995	0.986	0.54	0.35	0.25
	12	OS	0.778	0.886	0.931	0.14	0.19	0.23
		TS	0.938	0.998	0.988	0.61	0.38	0.27
cfw	4	OS	0.970	0.965	0.953	0.24	0.21	0.17
		TS	0.951	0.995	0.997	0.27	0.33	0.43
	8	OS	0.968	0.960	0.937	0.34	0.31	0.24
		TS	0.953	0.996	0.996	0.38	0.48	0.61
	12	OS	0.964	0.962	0.929	0.42	0.38	0.30
		TS	0.950	0.995	0.996	0.46	0.60	0.75

6. Data analysis

We apply the one-sample and two-sample methods to a GWAS in the outbred Carworth Farms White (CFW) mice population (Parker et al., 2016). Parker et al. (2016) show no widespread population structure or cryptic relatedness in the CFW mice and therefore, we view these mice as independent of each other. The primary pre-processing of phenotypes and genotypes, including outliers removal and basic transformation, was conducted using the original paper's code. After the pre-processing, the data set consists of 1,038 mice with 79,824 genetic variants and 71 different phenotypes. We study the genetic associations for the weights of four hindlimb muscles: EDL (extensor digitorum longus), Gastroc (gastrocnemius), Soleus, and TA (transverse abdominal), respectively.

We first predict each response based on constructed two-sample summary data and evaluate the out-of-sample prediction accuracy. Take the EDL response as an example. In each experiment, we randomly split the samples into two folds and use one fold to compute the GWAS statistic, $\widehat{S}$, and the other fold to compute the sample covariance matrix, $\widetilde{\Sigma}$. We consider different sample size ratios where the sample size for GWAS, $n \in \{250, 500, 750\}$, and $\tilde{n} = 1038 - n$, which gives $n/\tilde{n}$ is approximated one of $\{1/3, 1, 3\}$. For each sample size configuration, we repeat independent splittings and predictions 30 times. In order to build a PRS, we first perform marginal screening based on the $(\tilde{X}, \tilde{y})$ and keep the

covariates whose marginal correlation is among the top-3000 significant. This gives a subset of 3000 covariates, whose marginal correlations are strong.

We evaluate the PRS prediction using the samples that are used for estimating the sample covariate matrix. For a generic estimator $\boldsymbol{b} \in \mathbb{R}^p$, define its test errors as $\|\tilde{\boldsymbol{y}} - \widetilde{X}\hat{\boldsymbol{\beta}}^{(ts)}\|_2^2 / \|\tilde{\boldsymbol{y}}\|_2^2$. From Table 4, we see that the SNPs are predictive for the EDL and Soleus weights, but are not predictive for the TA and Gastroc weights. As $n/\tilde{n}$ increases, the test errors decrease significantly in the one-sample case but are less significant in the two-sample case. It aligns with our theoretical analysis that one-sample Lasso has more accurate predictions on average for different sample size ratios.

Table 4: Test errors using the samples that are used for estimating the sample covariate matrix based on one-sample (OS) and two-sample (TS) methods with $n/\tilde{n} \in \{1/3, 1, 3\}$ for four hindlimb muscles.

	EDL		Gastroc		Soleus		TA	
$n/\tilde{n}$	OS	TS	OS	TS	OS	TS	OS	TS
1/3	0.733	0.762	1.015	1.028	0.715	0.745	0.999	1.003
1	0.653	0.703	1.006	1.002	0.648	0.699	0.952	0.954
3	0.624	0.701	0.994	0.982	0.596	0.692	0.939	0.946

For the inference results of PRS, we consider constructing confidence intervals for y_* and its standard deviation is $\sqrt{V_*^{(ts)} + \sigma^2}$ as μ_* is unknown. We again evaluate the PRS inference results using the samples that are used for estimating the sample covariate matrix. We plug in $\tilde{\sigma}^2 = \widehat{M} - 2(\hat{\boldsymbol{\beta}}^{(ts)})^T \widehat{\boldsymbol{S}} + (\hat{\boldsymbol{\beta}}^{(ts)})^T \widetilde{\boldsymbol{\Sigma}} \hat{\boldsymbol{\beta}}^{(ts)}$

Table 5: Inference results for $\bar{\mu}_*$ using the samples that are used for estimating the sample covariate matrix based on one-sample (OS) and two-sample (TS) methods with $n/\tilde{n} \in \{1/3, 1, 3\}$ for four hindlimb muscles.

Method	Significance			Average Coverage			Average SD		
	1/3	1	3	1/3	1	3	1/3	1	3
				EDL					
OS	0.173	0.200	0.229	0.940	0.983	0.973	0.687	0.691	0.692
TS	0.023	0.037	0.007	0.977	0.993	0.997	1.808	1.123	0.851
				Gastroc					
OS	0.083	0.060	0.087	0.943	0.970	0.940	1.016	1.125	1.166
TS	0.003	0.000	0.000	0.997	0.977	0.967	1.907	1.182	0.954
				Soleus					
OS	0.187	0.267	0.273	0.910	0.923	0.973	0.630	0.667	0.691
TS	0.010	0.013	0.017	0.993	0.990	1.000	1.787	1.119	0.833
				TA					
OS	0.153	0.183	0.303	0.320	0.347	0.500	0.988	0.957	0.995
TS	0.030	0.020	0.107	0.870	0.597	0.317	1.920	1.193	0.928

as an estimate of σ^2. In Table 5, we report the proportion of significant PRS detected with 5% significance level. One-sample and two-sample debiased methods have average coverage close to the nominal level for outcomes EDL, Gastroc, and Soleus. In view of the lengths of the confidence intervals, we see that the one-sample confidence intervals are shorter than the two-sample confidence intervals and this observation agrees with our theoretical results.

7. Discussion

Statistical learning with summary data has attracted significant interests in genetic, epidemiology, and other health-related studies. In this work, we have provided statistical inference methods and theoretical guarantees with proxy data in high-dimensional linear models. Our theoretical analysis sheds light on the practical use of two-sample methods. Interesting future research directions including dealing with heteroscedastic noises and model misspecification. In these two cases, the asymptotic variance of the debiased Lasso need to be estimated empirically (Bühlmann et al., 2015; Dezeure et al., 2017), which cannot be directly applied with summary data.

Supplementary Materials

The Supplementary Material includes proofs of the main theorems and the technical lemmas, details on power analysis, and additional simulation results.

Acknowledgement

References

Albiñana, C., J. Grove, J. J. McGrath, E. Agerbo, N. R. Wray, T. Werge, A. D. Børglum, P. B. Mortensen, F. Privé, and B. J. Vilhjálmsson (2021). Leveraging both individual-level genetic data and gwas summary statistics increases poly-genic prediction. *The American Journal of Human Genetics 108*(6), 1001–1011.

Bühlmann, P., S. van de Geer, et al. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics 9*(1), 1449–1473.

Cai, T., T. T. Cai, and Z. Guo (2021). Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B 83*, 669–719.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity1. *The Annals of Statistics 45*(2), 615–646.

Cai, T. T., W. Liu, and X. Luo (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Chen, T.-H., N. Chatterjee, M. T. Landi, and J. Shi (2021). A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *116*(533), 133–143.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Deelen, J., D. S. Evans, D. E. Arking, N. Tesi, M. Nygaard, X. Liu, M. K. Wojczynski, M. L. Biggs, A. van Der Spek, G. Atzmon, et al. (2019). A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nature Communications 10*, 3669.

Dezeure, R., P. Bühlmann, and C.-H. Zhang (2017). High-dimensional simultaneous inference with the bootstrap. *TEST 26*(4), 685–719.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Javanmard, A. and H. Javadi (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics 13*(1), 1212–1253.

Javanmard, A. and J. D. Lee (2020). A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(3), 685–718.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*, 2869–2909.

Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Li, S., T. T. Cai, and H. Li (2022). Supplements to "estimation and inference with proxy data and its genetic applications".

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics 41*(6), 2948–2978.

Mak, T. S. H., R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology 41*(6), 469–480.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Parker, C. C., S. Gopalakrishnan, P. Carbonetto, N. M. Gonzales, E. Leung, Y. J. Park, E. Aryee, J. Davis, D. A. Blizard, C. L. Ackert-Bicknell, et al. (2016). Genome-wide association study of behavioral, physiological and gene expression traits in outbred cfw mice. *Nature Genetics 48*(8), 919–926.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Vilhjálmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics 97*(4), 576–592.

Wang, J. and T. Zhang (2019). Utilizing second order information in mini-batch stochastic variance reduced proximal iterations. *The Journal of Machine Learning Research 20*, 1578–1633.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Institute of Statistics and Big Data, Renmin University of China, Beijing, China, 100872

E-mail: (saili@ruc.edu.cn)

Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104

E-mail: (tcai@wharton.upenn.edu)

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

E-mail: (hongzhe@upenn.edu)