Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications

T. Tony Cai

University of Pennsylvania, Philadelphia, USA

Zijian Guo

Rutgers University, Piscataway, USA

Summary. This paper considers statistical inference for the explained variance $\beta^\intercal\Sigma\beta$ under the high-dimensional linear model $Y = X\beta + \epsilon$ in the semi-supervised setting, where β is the regression vector and Σ is the design covariance matrix. A calibrated estimator, which efficiently integrates both labelled and unlabelled data, is proposed. It is shown that the estimator achieves the minimax optimal rate of convergence in the general semi-supervised framework. The optimality result characterizes how the unlabelled data contributes to the estimation accuracy. Moreover, the limiting distribution for the proposed estimator is established and the unlabelled data has also proven useful in reducing the length of the confidence interval for the explained variance. The proposed method is extended to the semi-supervised inference for the unweighted quadratic functional, $\|\beta\|_2^2$. The obtained inference results are then applied to a range of high-dimensional statistical problems, including signal detection and global testing, prediction accuracy evaluation, and confidence ball construction. The numerical improvement of incorporating the unlabelled data is demonstrated through simulation studies and an analysis of estimating heritability for a yeast segregant data set with multiple traits.

Keywords: Unlabelled Data, Confidence set, Heritability, Prediction Accuracy, Signal Detection, Minimaxity.

1. Introduction

High-dimensional linear models are ubiquitous in contemporary statistical modeling with a wide range of applications in many scientific fields. The early focus has been mainly on developing methods for the recovery of the whole regression vector via penalized or constrained ℓ_1 minimization approaches. Examples include the Lasso [Tibshirani, 1996], Dantzig Selector [Candès and Tao, 2007], MCP [Zhang, 2010], square-root Lasso [Belloni et al., 2011], and scaled Lasso [Sun and Zhang, 2012]. There have been significant recent interests in statistical inference for low-dimensional functionals, including confidence intervals and hypothesis testing for individual regression coefficients [Zhang and Zhang, 2014, van de Geer et al., 2014, Javanmard and Montanari, 2014a,b], minimaxity and adaptivity of confidence intervals for general linear functionals [Cai and Guo, 2017c], estimation of the signal-to-noise-ratio [Verzelen and Gassiat, 2016, Janson et al., 2015], inference for the ℓ_q accuracy of a given estimator [Cai and Guo, 2017a], and estimation of quadratic functionals [Janson et al., 2015, Guo et al., 2017]. Motivated by a range of applications, the present paper considers the semi-supervised inference problem in

high dimensions, where the main statistical goal is to integrate both the labelled and unlabelled data, and propose efficient point and interval estimators.

1.1. Problem Formulation and Motivations

We consider the high-dimensional linear model with a random design,

$$y_i = X_{i\cdot}^\intercal \beta + \epsilon_i, \quad \text{for } 1 \le i \le n \tag{1}$$

where $y_i \in \mathbb{R}$ and $X_{i\cdot} \in \mathbb{R}^p$ denote respectively the outcome and the measured covariates of the i-th observation, ϵ_i denotes the error and $\beta \in \mathbb{R}^p$ denotes the high-dimensional regression vector. The covariates $X_{i\cdot}$ are i.i.d. p-dimensional random vectors with mean 0 and covariance matrix Σ and the errors $\{\epsilon_i\}_{1\le i\le n}$ are i.i.d random variables with mean 0 and variance σ^2 and independent of $\{X_{i\cdot}\}_{1\le i\le n}$. The explained variance under the regression model (1) is represented by $\mathrm{Q} = \mathrm{Var}(X_{i\cdot}^\intercal \beta) = \beta^\intercal \Sigma \beta$. We focus on the semi-supervised setting, where the data is a combination of the labelled data $\{y_i, X_{i\cdot}\}_{1\le i\le n}$ in the regression model (1) and the unlabelled data $\{X_{i\cdot}\}_{n+1\le i\le n+N}$. Here the measured covariates of both the labelled and unlabelled data are assumed to be independent and follow the same distribution. The more conventional supervised setting is treated as a special case with no additional unlabelled data.

The setting of semi-supervised learning is commonly seen in applications where the outcomes are more expensive to collect than the covariates. For example, in the analysis of Electronic Health Records (EHR) databases, the covariates are easy to be automatically extracted while labelling of the outcomes is costly and time-consuming [Chakrabortty and Cai, 2017, Gronsbell and Cai, 2017]. In addition, semi-supervised learning naturally arises in the integrative analysis of multiple (genetics) data sets where the covariates are the same across all data sets but the outcomes measured vary from study to study due to the specific purposes of individual studies [van Iperen et al., 2017]. This can be naturally formulated as semi-supervised learning, where the pre-specified outcome is only measured over one or several (but not all) data sets while the covariates are measured across all data sets.

The construction of the optimal estimator and confidence intervals for $\mathrm{Q} = \beta^\intercal \Sigma \beta$ in the semi-supervised and high-dimensional setting is not only of significant interest on its own right, but is also closely connected to several other important statistical problems.

(a) **Heritability.** Heritability is among the most important genetics concepts. Under the model (1) with the outcome normalized to have unit variance, $\beta^\intercal \Sigma \beta$ is a measure of heritability, which quantifies the total variance explained by genetic variants [Owen, 2012, Guo et al., 2017, Janson et al., 2015, Verzelen and Gassiat, 2016].

(b) **Signal-to-Noise Ratio (SNR) and Proportion-of-Variance Explained (PVE).** SNR and PVE are important statistics concepts and are defined respectively as $\beta^\intercal \Sigma \beta / (\beta^\intercal \Sigma \beta + \sigma^2)$ and $\beta^\intercal \Sigma \beta / \sigma^2$ under model (1). Together with a good estimator of σ^2 [Sun and Zhang, 2012, Belloni et al., 2011], the results for $\beta^\intercal \Sigma \beta$ established in this paper are useful for inference of SNR and PVE.

(c) **Signal Detection and Global Testing.** Inference for the explained variance can be applied to testing the global hypothesis $H_0 : \beta = \beta^{\mathrm{null}}$ for $\beta^{\mathrm{null}} \in \mathbb{R}^p$,

which includes signal detection as a special case with $\beta^{\text{null}} = 0$. The connection is revealed in the adjusted linear model, $y_i - X_{i\cdot}^\intercal \beta^{\text{null}} = X_{i\cdot}^\intercal \left(\beta - \beta^{\text{null}}\right) + \epsilon_i$ for $1 \le i \le n$, where testing for $H_0 : \beta = \beta^{\text{null}}$ is recast as testing the hypotheses $H_0 : \left(\beta - \beta^{\text{null}}\right)^\intercal \Sigma \left(\beta - \beta^{\text{null}}\right) = 0$ versus $H_1 : \left(\beta - \beta^{\text{null}}\right)^\intercal \Sigma \left(\beta - \beta^{\text{null}}\right) > 0$.

(d) **Prediction Accuracy Assessment.** Accuracy assessment is of significant importance in applications. Let $\check{\beta}$ denote a given estimator based on the training data. We define the out-of-sample prediction accuracy for a given observation x_{new} as $\mathbb{E}_{x_{\text{new}}} \left(x_{\text{new}}^\intercal(\check{\beta} - \beta)\right)^2 = (\check{\beta} - \beta)^\intercal \Sigma (\check{\beta} - \beta)$. We introduce the following adjusted linear model for the independent test data $\{X_{i\cdot}, y_i\}_{1 \le i \le n}$,

$$y_i - X_{i\cdot}^\intercal \check{\beta} = X_{i\cdot}^\intercal \left(\beta - \check{\beta}\right) + \epsilon_i \quad \text{for } 1 \le i \le n. \tag{2}$$

Inference results developed for the explained variance can be applied to (2) to obtain the corresponding results for the prediction accuracy $\mathbb{E}_{x_{\text{new}}} \left(x_{\text{new}}^\intercal(\check{\beta} - \beta)\right)^2$.

(e) **Confidence Ball for β.** Construction of confidence balls for β is another important application. Based on (2), a confidence interval $(L(Z), U(Z))$ for $(\check{\beta}-\beta)^\intercal \Sigma (\check{\beta}-\beta)$ leads to a confidence ball for β centering at $\check{\beta}$, $\left\{\beta : \|\beta - \check{\beta}\|_2^2 \le U(Z)/\lambda_{\min}(\Sigma)\right\}$, where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ.

More detailed discussions about these statistical applications are present in Section 5.

1.2. Results and Contributions

A central question in semi-supervised learning is *how to efficiently use both labelled and unlabelled data* [Chakrabortty and Cai, 2017, Gronsbell and Cai, 2017]. We introduce a novel two-step estimator, Calibrated High-dimensional Inference for Variance Explained (CHIVE), where the first step is to plug in the estimators of β and Σ, denoted by $\widehat{\beta}$ and $\widehat{\Sigma}$, respectively, and the second step is to calibrate this plug-in estimator $\widehat{\beta}^\intercal \widehat{\Sigma} \widehat{\beta}$ through estimating a dominating term in its error decomposition. The second step is to rebalance the bias and variance and improve the estimation accuracy. Different forms of $\widehat{\beta}$ and $\widehat{\Sigma}$ can be taken as inputs of CHIVE method and this flexibility is useful in integrating the unlabelled data to estimate Σ more accurately. This idea is then extended to semi-supervised inference for the unweighted quadratic functional $\|\beta\|_2^2$, where the additional unlabelled data facilitates the estimation of Σ^{-1}.

Another important question is whether the unlabelled data has been efficiently utilized in semi-supervised learning. We address this question by establishing the minimax optimal rate of convergence for estimating $\beta^\intercal \Sigma \beta$, where the optimal rate is $M/\sqrt{n} + M^2/\sqrt{N+n} + k \log p/n$, with p, n, N, k, and M denoting respectively the dimension, the size of the labelled data, the size of the unlabelled data, the sparsity, and the ℓ_2 norm of β. The proposed CHIVE estimator achieves this optimal rate, which justifies the efficient use of the unlabelled data. The optimal rate is not just achieved for the case where there is a large amount of unlabelled data, but is also for any given amount of unlabelled data. The minimax optimal rate characterizes the fundamental difficulty of the inference problem in the semi-supervised setting and is independent of specific

procedures. This minimax rate also reveals that the unlabelled data is most effective when the signal strength $\|\beta\|_2$ is large.

We establish the limiting distribution of the CHIVE estimator and construct data-driven confidence intervals for $\beta^\intercal\Sigma\beta$ based on this estimator. The limiting distribution is normal and its variance is scaled to the proportion of the labelled data, which is unique to the semi-supervised setting. A larger amount of unlabelled data leads to a smaller proportion of the labelled data and hence a smaller asymptotic variance, which leads to a shorter confidence interval for $\beta^\intercal\Sigma\beta$. The effect of the unlabelled data is also demonstrated in the numerical studies. Specifically, in comparison with the estimators based only on the labelled data, the RMSE for estimation and the length of confidence intervals can be reduced by as much as 70%. See details in Section 6.

The improvement in semi-supervised inference for $\|\beta\|_2^2$ is similar to that for $\beta^\intercal\Sigma\beta$ at a high level but different in technical details. Specifically, the estimation accuracy is significantly improved in the strong signal regime, and the improvement is limited if the signal strength $\|\beta\|_2^2$ is not large enough. Construction of confidence intervals for $\|\beta\|_2^2$ also gets easier in the sense that the condition for sample size and model complexity is weakened by making use of the unlabelled data.

The inference results obtained in this paper are applied to (i) signal detection and global testing, (ii) prediction accuracy evaluation, and (iii) confidence ball construction. For signal detection, we control the type I error and characterize the type II error by establishing the power function under a local alternative. The results can be easily extended to the general global testing problem. For evaluation of out-of-sample prediction accuracy of a given sparse estimator of β, both the point and interval estimators are developed. We establish the estimation error bound for the point estimator of the prediction accuracy and control the length of the corresponding confidence interval. A confidence ball for the regression vector β with controlled radius is also constructed. We stress that these procedures are data-driven and do not require a priori knowledge of the design covariance matrix Σ or the noise level σ. See more details in Section 5.

1.3. Related Work

Estimation and inference for quadratic functionals have been studied in the literature in a range of settings. In particular, minimax and adaptive estimation of quadratic functionals plays an important role in nonparametric inference and has been well studied in density estimation, nonparametric regression, and white noise with drift model. See, for example, Bickel and Ritov [1988], Donoho and Nussbaum [1990], Efromovich and Low [1996], Laurent and Massart [2000], Cai and Low [2005, 2006], Collier et al. [2015].

The most related works to the current paper are Verzelen and Gassiat [2016] and Guo et al. [2017], which considered estimation of $\beta^\intercal\Sigma\beta/\sigma^2$ and $\|\beta\|_2^2$, respectively, in high-dimensional linear regression. The main difference between the current paper and these two related works are two-fold: (a) Verzelen and Gassiat [2016] and Guo et al. [2017] only considered the supervised setting instead of the semi-supervised setting. As demonstrated in both theoretical and numerical justifications, a careful integration of the unlabelled data proves useful in improving the estimation accuracy and reducing the length of constructed confidence intervals. (b) The focus of Verzelen and Gassiat [2016] and Guo et al. [2017] is about point estimation while the current paper studies

the more challenging problem of uncertainty quantification and also related hypothesis testing, in addition to point estimation. As is well known, uncertainty quantification in high dimensions is significantly different from and more involved than point estimation [Nickl and van de Geer, 2013, Cai and Guo, 2017c].

Another related paper, Janson et al. [2015], studied the construction of confidence intervals for $\|\beta\|_2^2$ in the setting of $\Sigma = \mathrm{I}$, moderate dimension where $n/p \to \xi \in (0, 1)$ and no sparsity assumption on β. The inference problem considered in the current paper is significantly different from the setting considered in Janson et al. [2015], mainly due to the complicated geometry induced by the sparsity structure and the unknown design covariance matrix Σ. Other works related to quadratic functional inference include construction of confidence intervals for the ℓ_2 loss of the estimator considered in Cai and Guo [2017a]. In addition, Javanmard and Lee [2017], Zhu and Bradic [2017] considered hypothesis testing for high-dimensional linear regression. As another significant difference, the current paper studies how to efficiently integrate the labelled and unlabelled data in the general semi-supervised setting while all the aforementioned works solely focused on the supervised regression.

The statistical applications studied in this paper have also been considered separately in the literature. Signal detection was studied in Ingster et al. [2010], Arias-Castro et al. [2011] under the linear model (1) in a special setting where the design covariance matrix Σ is equal to or closed to the identity matrix. In this setting, Ingster et al. [2010], Arias-Castro et al. [2011] established optimal signal detection method and theory. The results established in the present paper enable the study of signal detection under a general setting where the design covariance matrix Σ is unknown. The confidence ball construction for the whole regression vector was considered in Nickl and van de Geer [2013] in the case of known σ and the optimal size and possibility of adaptive confidence balls was also established. The results obtained in the current paper lead to a confidence ball construction for β in the case of unknown σ. A problem related to prediction accuracy is inference for the estimation accuracy, which was considered in Cai and Guo [2017a], Janson et al. [2015]. However, inference for the prediction accuracy and that for the estimation accuracy are different problems.

1.4. Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we introduce in detail the CHIVE estimator and establish its minimax rate optimality in the semi-supervised setting. Section 3 focuses on quantifying the uncertainty of the CHIVE estimator and construction of the confidence intervals for $\beta^\intercal \Sigma \beta$. In Section 4, we extend the methodology to the semi-supervised inference for $\|\beta\|_2^2$. We apply in Section 5 the developed procedures to tackle three important problems, signal detection and global testing, prediction accuracy evaluation and confidence ball construction. Simulation results are given in Section 6 to illustrate the numerical improvement through incorporating the unlabelled data. An analysis of a yeast data set is presented in Section 7. A discussion is provided in Section 8. The proofs and the additional simulation results are presented in the appendix.

2. Semi-supervised Estimation of $\beta^\intercal\Sigma\beta$

In this section, we first introduce the calibration methodology for estimating the explained variance in the general semi-supervised framework and then establish the minimax convergence rate of estimating $\beta^\intercal\Sigma\beta$. A significant statistical gain is obtained by carefully integrating the unlabelled data and the proposed estimator is shown to achieve the optimal rate in the semi-supervised setting. The supervised setting and the setting with known design covariance matrix are then discussed as special cases. We begin with the notation that will be used in the rest of the paper.

For a matrix A, $A_{i\cdot}$, $A_{\cdot j}$, and $A_{i,j}$ denote respectively the i-th row, j-th column, and (i,j) entry of the matrix A. The spectral norm of A is $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$ and the matrix ℓ_1 norm is $\|A\|_{L_1} = \sup_{1\le j\le p}\sum_{i=1}^p |A_{ij}|$. For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalue of A. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, $\mathrm{supp}(x)$ denotes the support of x and the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{i=1}^p |x_i|^q)^{\frac{1}{q}}$ for $q \ge 0$ with $\|x\|_0 = |\mathrm{supp}(x)|$ and $\|x\|_\infty = \max_{1\le j\le p}|x_j|$. For $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We use c and C to denote generic positive constants that may vary from place to place. For a sequence of random variables X_n indexed by n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For a sequence of random variables X_n and numbers a_n, we define $X_n = o_p(a_n)$ if X_n/a_n converges to zero in probability. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \le Cb_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\overline{\lim}_{n\to\infty}\frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$. We define the signal-to-noise ratio (SNR) in the context of model (1) as $\mathrm{SNR} = \frac{1}{\sigma}\sqrt{\beta^\intercal\Sigma\beta}$.

2.1. Calibration of Plug-in Estimators

In semi-supervised learning, we observe the labelled data $(X_{1\cdot}, y_1), \cdots, (X_{n\cdot}, y_n)$ and the unlabelled data $X_{n+1\cdot},, \cdots, X_{n+N\cdot}$, where $X_{1\cdot}, \cdots, X_{n\cdot}, X_{n+1\cdot},, \cdots, X_{n+N\cdot}$ are i.i.d realizations of p-dimensional covariates. We use $\widehat{\beta}$ and $\widehat{\Sigma}$ to denote some estimators of β and Σ, which will be specified later. A preliminary estimator of the quadratic functional $\mathrm{Q} = \beta^\intercal\Sigma\beta$ is the plug-in estimator $\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$, which has the following error decomposition,

$$\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta} - \beta^\intercal\Sigma\beta = 2\widehat{\beta}^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta) - (\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta) + \beta^\intercal(\widehat{\Sigma}-\Sigma)\beta. \tag{3}$$

Since the first term $2\widehat{\beta}^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta)$ on the right hand side can be estimated in a data-dependent way, the corresponding estimation error of the preliminary estimator $\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$ can be further reduced. We estimate the term $2\widehat{\beta}^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta)$ by $-2\widehat{\beta}^\intercal\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta})$ and propose the following calibrated estimator,

$$\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}) = \widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta} + 2\widehat{\beta}^\intercal\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta}). \tag{4}$$

This estimator is referred to as the Calibrated High-dimensional Inference for Variance Explained (CHIVE) estimator. The calibration step in (4) is essentially to improve the plug-in estimator $\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$ through re-balancing the bias and variance.

The CHIVE estimator requires three inputs, the initial estimators $\widehat{\beta}$ and $\widehat{\Sigma}$ and the data (X, y). With this machinery, we have the flexibility of choosing the initial estimators $\widehat{\beta}$ (and also $\widehat{\sigma}^2$) and $\widehat{\Sigma}$ based on the observed data. We begin with the estimator for β and σ^2 and then move on to the estimator for Σ. Throughout the paper, we assume that the estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ satisfy the following conditions.

(B1) With probability larger than $1-\gamma(n)$ where $\gamma(n) \to 0$, the estimator $\widehat{\beta}$ satisfies

$$\max\{\frac{1}{n}\sum_{i=1}^{n}[X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)]^2, \|\widehat{\beta}-\beta\|_2^2\} \lesssim \frac{k\log p}{n}\sigma,\ \|(\widehat{\beta}-\beta)_{S^c}\|_1 \le C_0\|(\widehat{\beta}-\beta)_S\|_1$$

where $S = \mathrm{supp}(\beta)$ and $C_0 > 0$ is some positive constant.

(B2) $\widehat{\sigma}^2$ is a consistent estimator of σ^2, that is, $\left|\widehat{\sigma}^2/\sigma^2 - 1\right| \xrightarrow{p} 0$.

One of the key assumptions for the general penalized estimators satisfying (B1) and (B2) is the following restricted eigenvalue condition on the population covariance matrix Σ,

$$\kappa(k, C_0, \Sigma) = \min_{S\in\{1,\cdots,p\},\,|S|\le k}\ \min_{v\neq 0, \|v_{S^c}\|_1\le C_0\|v_S\|_1} \frac{\|\Sigma^{\frac{1}{2}}v\|_2}{\|v_S\|_2} \ge c,$$

for some positive constant $c > 0$. This population version restricted eigenvalue condition implies the sample version restricted eigenvalue condition introduced in Bickel et al. [2009], under the assumption that the covariates $X_{i\cdot}$ are in a certain broad family of sub-gaussian random vectors and the sparsity k satisfies $k \lesssim n/\log p$; See Zhou [2009], Raskutti et al. [2010] for the exact statement.

Estimators satisfying (B1) **and** (B2)**.** The scaled lasso estimator $\{\widehat{\beta}, \widehat{\sigma}\}$ defined by

$$\{\widehat{\beta}, \widehat{\sigma}\} = \arg\min_{\beta\in\mathbb{R}^p, \sigma\in\mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sqrt{\frac{2.01\log p}{n}}\sum_{j=1}^{p}\frac{\|X_{\cdot j}\|_2}{\sqrt{n}}|\beta_j| \tag{5}$$

has been shown in Sun and Zhang [2012] to satisfy (B1) and (B2) under regularity conditions. See also Lemma 1 in Guo et al. [2017] for more details. Since the square root lasso estimator [Belloni et al., 2011] is numerically the same with the scaled Lasso estimator, the square root lasso estimators of β and σ also satisfy (B1) and (B2). In addition, with a prior knowledge of σ, the Lasso estimator of β and other variants are also shown to satisfy the above condition (B1); see Candès and Tao [2007], Zhang [2010], Ye and Zhang [2010] for more details.

Now, we turn to the estimators of Σ. This is exactly the place where we make use of the unlabelled data. Specifically, we pool the information contained in both the labelled and unlabelled data and estimate Σ by $\widehat{\Sigma}^S = \frac{1}{n+N}\sum_{i=1}^{n+N} X_{i\cdot}X_{i\cdot}^{\intercal}$. Then we use $\widehat{\beta}$ and $\widehat{\Sigma}^S$ as inputs and utilize the calibration idea introduced in (4),

$$\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^S) = \widehat{\beta}^{\intercal}\widehat{\Sigma}^S\widehat{\beta} + 2\widehat{\beta}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}(y_i - X_{i\cdot}^{\intercal}\widehat{\beta}). \tag{6}$$

When there is no confusion, we use $\widehat{\mathrm{Q}}$ to denote the estimator proposed in (6). We introduce the following regularity conditions and then establish the convergence rate of the proposed estimator in (6) in Theorem 1.

(A1) The regression vector β is assumed to be k-sparse; The errors $\{\epsilon_i\}_{1\le i\le n}$ are independent of $\{X_{i\cdot}\}_{1\le i\le n+N}$ and follow i.i.d sub-gaussian random variable with mean zero and variance σ^2; The rows $X_{i\cdot}$ are i.i.d. p-dimensional random vectors and can be expressed in the form of $X_{i\cdot} = \Sigma^{\frac{1}{2}} Z_{i\cdot}$ where $Z_{i\cdot} \in \mathbf{R}^p$ is a subgaussian random vector of mean 0 and identity covariance matrix and Σ has a bounded restricted largest eigenvalue $\rho_{\max}(k,\Sigma)$, which is defined as $\rho_{\max}(k,\Sigma) = \max_{\|v\|_2=1,\,\|v\|_0\le k} v^\intercal \Sigma v$.

(A2) $\sqrt{\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal\Sigma\beta\right)^2} \ge c_0 \beta^\intercal\Sigma\beta$. for some positive constant $c_0 > 0$.

Assumption (A1) requires that the restricted largest eigenvalue $\rho_{\max}(k,\Sigma)$ is upper bounded, where the "restricted" here means that the maximum in the definition of $\rho_{\max}(k,\Sigma)$ is taken with respect to k-sparse vectors. Note that the restricted (smallest) eigenvalue condition is not required for the theoretical analysis of the proposed estimator $\widehat{\mathrm{Q}}$ as long as the estimator $\widehat{\beta}$ of β satisfies the condition (B1). Define $U = X_{i\cdot}^\intercal\beta/\sqrt{\beta^\intercal\Sigma\beta}$, where $\mathbb{E}(U) = 0$ and $\mathbb{E}(U^2) = 1$. Assumption (A2) is placed on this random variable U such that $\mathrm{Var}(U^2)$ is not vanishing. This assumption is imposed such that $\mathrm{Var}(U^2)$ can be well estimated and this type of assumption has been introduced in covariance matrix estimation literature [Cai and Liu, 2011] for the same purpose.

THEOREM 1. *Suppose that Condition* (A1) *holds and* $k \le cn/\log p$ *for some constant* $c > 0$. *For any estimator* $\widehat{\beta}$ *satisfying Condition* (B1), *with probability at least* $1-\gamma(n)-C(p^{-c}+\exp(-cN)+e^{-ct^2})$, *the estimator* $\widehat{\mathrm{Q}} = \widehat{\mathrm{Q}}(\widehat{\beta},\widehat{\Sigma}^S)$ *defined in* (6) *satisfies*

$$\left|\widehat{\mathrm{Q}} - \mathrm{Q}\right| \lesssim t\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}} + t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}} + (1+\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}\frac{N}{n+N})\frac{k\log p}{n}\sigma^2. \tag{7}$$

Under the additional assumptions $k \ll \sqrt{n}/\log p$ *and* $\mathrm{SNR} \gg k\log p/\sqrt{n}$,

$$\frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \mathrm{Q}\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1) \tag{8}$$

where $\rho = \lim_{n\to\infty}\frac{n}{N+n}$.

As a remark, the probability $1-\gamma(n)-C(p^{-c}+\exp(-cN)+e^{-ct^2})$ holds for the finite sample n and finite dimension p and also any non-negative constant $t \ge 0$. However, the established result is more interesting over the regime $\min\{p,n\}\to\infty$ and $t\to\infty$ as in this scenario, the corresponding probability $1-\gamma(n)-C(p^{-c}+\exp(-cN)+e^{-ct^2})$ approaches 1. Since $\mathrm{Q} \ge 0$, the convergence rate (7) also holds for $\widehat{\mathrm{Q}}_+$, the positive part of $\widehat{\mathrm{Q}}$. To keep the notation simpler, we only present the results for $\widehat{\mathrm{Q}}$ in this paper.

The rate of convergence in (7) reveals the effect of the unlabelled data. The sample size of the unlabelled data, N, appears only in the second term $t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}}$. An interesting observation is that the usefulness of the unlabelled data varies across different signal strengths. If the signal is strong in the sense that $\mathrm{SNR} \gtrsim \max\{1, k\log p/\sqrt{n}\}$, in which case the term $t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}}$ is dominant in (7), then the additional unlabelled data

reduces the estimation error significantly; if the signal is weak in the sense that SNR $\ll \max\{1, k\log p/\sqrt{n}\}$, then the impact of the additional unlabelled data is limited.

To demonstrate the effect of calibration, we note that an upper bound for the term $\widehat{\beta}^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta)$ in (3) is at the order of magnitude $\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2\sqrt{k\log p/n}$ while the remaining error after the calibration step is $t\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}}+(1+\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}\frac{N}{n+N})\frac{k\log p}{n}\sigma^2$, as shown in (7). By comparing these upper bounds, we note that the calibration step is useful in reducing the upper bound for the rate of convergence. This reduction of estimation error has also been numerically demonstrated in Section 6.2. The terms $t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}}+\frac{k\log p}{n}\sigma^2$ in (7) capture the convergence rate of the last two terms in (3).

The distributional result in (8) is established under the additional assumptions $k \ll \sqrt{n}/\log p$ and SNR $\gg k\log p/\sqrt{n}$. These additional assumptions are imposed to ensure that the variance component $t\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}}+t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}}$, captured by the normal limiting distribution after re-scaling, dominates the bias component $(1+\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}\frac{N}{n+N})\frac{k\log p}{n}\sigma^2$. Since the bias term is hard to characterize, we impose these sufficient conditions such that the variance term is the dominating term. The normal limiting distribution in (8) can be used in Section 3 to construct confidence intervals for $\beta^\intercal\Sigma\beta$.

Another interesting phenomenon is that the limiting distribution established in (8) depends on the proportion of the labelled data, which is unique in the semi-supervised inference problem. If the amount of unlabelled data dominates that of labelled data (that is, $\rho = 0$), then the limiting distribution in (8) is simplified to $\frac{\sqrt{n}(\widehat{\mathrm{Q}}-\mathrm{Q})}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta}} \xrightarrow{d} N(0,1)$. Theorem 1 demonstrates that the CHIVE estimator integrating the unlabelled data improves the rate of convergence in estimating the explained variance. The lower bound given in the next subsection shows that CHIVE is optimal in terms of the rate of convergence.

2.2. Optimal Estimation in the Semi-supervised Setting

In this section, we further investigate the fundamental limit for estimating $\mathrm{Q} = \beta^\intercal\Sigma\beta$ in the general semi-supervised setting over the following specific parameter space,

$$\Theta(k, M) = \left\{\theta = (\beta, \Sigma, \sigma) : \|\beta\|_0 \le k,\ M/2 \le \|\beta\|_2 \le M,\ \frac{1}{M_1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M_1,\ \sigma \le M_2\right\},$$

where $M_1 \ge 1$ and $M_2 > 0$ are positive constants. Here k quantifies the sparsity of β and M quantifies the signal strength of the true signal β in terms of its ℓ_2 norm. Both k and M are allowed to grow with n and p. The other conditions $1/M_1 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M_1$ and $\sigma \le M_2$ are regularity conditions. The following theorem establishes the minimax lower bounds for estimating Q over the parameter space $\Theta(k, M)$.

THEOREM 2. *Suppose $k \le c\min\{n/\log p, p^\nu\}$ for some constants $c > 0$ and $0 \le \nu < \frac{1}{2}$. Then*

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\theta\in\Theta(k,M)} \mathbb{P}\left(\left|\widetilde{\mathrm{Q}} - \mathrm{Q}\right| \gtrsim \frac{M^2}{\sqrt{N+n}} + \min\left\{\frac{M}{\sqrt{n}} + \frac{k\log p}{n}, M^2\right\}\right) \ge \frac{1}{4}. \tag{9}$$

One interesting observation of the above theorem is that only the first term in the lower bound is involved with the amount of the unlabelled data. Theorems 1 and 2 together

show that the estimator proposed in Section 2.1 is minimax rate optimal under regularity conditions.

COROLLARY 1. *Suppose that Condition* (A1) *holds and* $k \le c\min\{n/\log p, p^\nu\}$ *for some constants* $c > 0$ *and* $0 \le \nu < \frac{1}{2}$. *For any estimator* $\widehat{\beta}$ *satisfying Condition* (B1), *the estimator* $\widehat{\mathrm{Q}}$ *defined in* (6) *is minimax rate optimal over* $\Theta(k, M)$ *where* $\sqrt{k\log p/n} \lesssim M \le C$ *for some constant* $C > 0$, *that is,*

$$\sup_{\theta\in\Theta(k,M)} \mathbb{P}\left(\left|\widehat{\mathrm{Q}} - \mathrm{Q}\right| \gtrsim t\frac{M^2}{\sqrt{n+N}} + \frac{M}{\sqrt{n}} + \frac{k\log p}{n}\right) \le C(p^{-c} + \exp(-cN) + e^{-ct^2}) + \gamma(n) \tag{10}$$

The CHIVE estimator attains the optimal convergence rate when the ℓ_2 norm of β is relatively strong, that is, M is bounded away from zero by $\sqrt{k\log p/n}$. As shown in Theorem 2, for the case where $M \ll \sqrt{k\log p/n}$, the lower bound of estimating $\beta^\intercal\Sigma\beta$ is M^2. This optimal convergence rate can be achieved by a trivial estimator 0.

In Corollary 1, the lower bound (9) is only matched for the regime where $M \le C$ for some constant $C > 0$. For theoretical interest, we will modify the proposed estimator $\widehat{\mathrm{Q}}$ defined in (6) such that the modified version achieves the lower bound (9) over the regime $M \gtrsim \sqrt{k\log p/n}$. We randomly split the data (y, X) into two subsamples $\left(y^{(1)}, X^{(1)}\right)$ with sample size n_1 and $\left(y^{(2)}, X^{(2)}\right)$ with sample size n_2, where $n_1 \asymp n_2$. Let $\widehat{\beta}$ denote an estimator which is produced by the first sub-sample $\left(y^{(1)}, X^{(1)}\right)$ and satisfies Condition (A1). One example of such an estimator is the scaled Lasso estimator (5) applied to the subsample $\left(y^{(1)}, X^{(1)}\right)$. We propose the following estimator of Q,

$$\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)}) = \widehat{\beta}^\intercal\widehat{\Sigma}^{(2)}\widehat{\beta} + 2\widehat{\beta}^\intercal\frac{1}{n_2}\sum_{i=n_1+1}^{n} X_{i\cdot}^\intercal(y_i - X_{i\cdot}\widehat{\beta}), \tag{11}$$

where $\widehat{\Sigma}^{(2)} = \frac{1}{n+N-n_1}\sum_{i=n_1+1}^{n+N} X_{i\cdot}X_{i\cdot}^\intercal$. The following theorem establishes the convergence rate of $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)})$ and shows that this estimator achieves the optimal convergence rate of estimating Q for $M \gtrsim \sqrt{k\log p/n}$.

THEOREM 3. *Suppose that Condition* (B1) *holds and* $k \le cn/\log p$ *for some constant* $c > 0$. *Let* $\widehat{\beta}$ *be an estimator depending on the first half sample* $\left(y^{(1)}, X^{(1)}\right)$ *and satisfying Condition* (A1). *Then with probability larger than* $1 - \gamma(n) - C(p^{-c} + \exp(-cN) + e^{-ct^2})$, *the estimator* $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)})$ *defined in* (11) *satisfies*

$$\left|\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)}) - \mathrm{Q}\right| \lesssim (t+1)\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}} + t\frac{\beta^\intercal\Sigma\beta}{\sqrt{N+n}} + \frac{k\log p}{n}\sigma^2. \tag{12}$$

Hence, the estimator $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)})$ *defined in* (11) *achieves the optimal estimation rate over* $\Theta(k, M)$ *in the sense of* (10) *over the regime* $k \le c\min\{n/\log p, p^\nu\}$ *for some constants* $c > 0$ *and* $0 \le \nu < \frac{1}{2}$ *and* $M \gtrsim \sqrt{k\log p/n}$.

2.3. Two Special Cases

We now turn to two important special cases, the inference in the supervised setting and the setting with known design covariance matrix.

2.3.1. Case I: Supervised Inference

In the supervised setting without any additional unlabelled data, Σ is estimated by $\widehat{\Sigma}^L = \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal$. The following corollary establishes the convergence rate of the estimator $\widehat{\mathrm{Q}} = \widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$, which is a special case of the estimator (6) with $N = 0$.

COROLLARY 2. *Suppose that Condition* (A1) *holds and* $k \le cn/\log p$ *for some constant* $c > 0$. *For any estimator* $\widehat{\beta}$ *satisfying* (B1), *with probability larger than* $1 - \gamma(n) - C(p^{-c} + \exp(-ct^2))$, $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$ *proposed in* (4) *with* $\widehat{\Sigma}^L = \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal$ *satisfies*

$$\left|\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L) - \mathrm{Q}\right| \lesssim t\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2 + \beta^\intercal\Sigma\beta}{\sqrt{n}} + \frac{k\log p}{n}\sigma^2. \tag{13}$$

Under the additional assumption (A2) *and* $\mathrm{SNR} \gg \min\left\{k\log p/\sqrt{n}, (k\log p/\sqrt{n})^{1/2}\right\}$,

$$\frac{\sqrt{n}\left(\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L) - \mathrm{Q}\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1) \tag{14}$$

Corollary 2 basically follows from Theorem 1 with $N = 0$ except for some technical difference. By comparing Corollary 2 with Theorems 1 and 3, we observe that the unlabelled data leads to a faster convergence rate by reducing $\beta^\intercal\Sigma\beta/\sqrt{n}$ in (13) to $\beta^\intercal\Sigma\beta/\sqrt{N+n}$ in (7) and (12); the unlabelled data does not affect other terms in the convergence rate. The effect of the unlabelled data is also revealed in the limiting distribution in (14), where the exact variance level is reduced from $[4\sigma^2\beta^\intercal\Sigma\beta + \mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2]/n$ in (14) to $[4\sigma^2\beta^\intercal\Sigma\beta + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2]/n$ in (8) for $\rho = \lim_{n\to\infty}\frac{n}{N+n} \in [0,1]$. The following corollary further establishes the minimax rate for estimating $\beta^\intercal\Sigma\beta$ in the supervised setting.

COROLLARY 3. *Suppose that Condition* (A1) *holds and* $k \le c\min\{n/\log p, p^\nu\}$ *for some constants* $c > 0$ *and* $0 \le \nu < \frac{1}{2}$. *For any estimator* $\widehat{\beta}$ *satisfying Condition* (B1), *the estimator* $\widehat{\mathrm{Q}} = \widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$ *defined in* (4) *with* $\widehat{\Sigma}^L = \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal$ *achieves the optimal estimation rate over* $\Theta(k, M)$ *for* $M \gtrsim \sqrt{k\log p/n}$, *that is,* $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$ *satisfies*

$$\sup_{\theta\in\Theta(k,M)} \mathbb{P}\left(\left|\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L) - \mathrm{Q}\right| \gtrsim t\frac{M^2}{\sqrt{n}} + \frac{M}{\sqrt{n}} + \frac{k\log p}{n}\right) \le C(p^{-c} + e^{-ct^2}) + \gamma(n). \tag{15}$$

REMARK 1. In the supervised setting, [Guo et al., 2017] established that the optimal rate of estimating $\|\beta\|_2^2$ over $\Theta(k, M)$ for $M \gtrsim \sqrt{k\log p/n}$ is $M/\sqrt{n} + (M+1)k\log p/n$. In contrast to (15), we can see that neither of these two problems is easier than the other, where there is an additional term $M^2/\sqrt{n}$ in (15) and an additional term $Mk\log p/n$ in the optimal convergence rate of estimating $\|\beta\|_2^2$.

Inference for $\beta^\intercal\Sigma\beta$ in the supervised setting is closely connected to Sun and Zhang [2012], Verzelen and Gassiat [2016], where Sun and Zhang [2012] studied the inference problem for σ^2 and Verzelen and Gassiat [2016] studied the estimation of $\beta^\intercal\Sigma\beta/\sigma^2$. In particular, Sun and Zhang [2012] proposed the scaled lasso estimator $\widehat{\sigma}^2$ in (5) to estimate σ^2 and Verzelen and Gassiat [2016] proposed to estimate $\beta^\intercal\Sigma\beta$ by $\left(\frac{1}{n}\|y\|_2^2 - \widehat{\sigma}^2\right)_+$ as an intermediate step of estimating $\beta^\intercal\Sigma\beta/\sigma^2$. For the estimator $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$ defined in (4), if $\widehat{\beta}$ is taken as the scaled Lasso estimator, then $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^L)$ is reduced to being the same as the estimator proposed in Verzelen and Gassiat [2016], where the equivalence is shown by the following expression,

$$\widehat{\beta}^\intercal\widehat{\Sigma}^L\widehat{\beta} + 2\widehat{\beta}^\intercal\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}\widehat{\beta}) = \frac{1}{n}\left(\|y\|_2^2 - \|y - X\widehat{\beta}\|_2^2\right) = \frac{1}{n}\|y\|_2^2 - \widehat{\sigma}^2. \tag{16}$$

As a remark, in the supervised setting, the calibration idea in (4) provides a completely new perspective on estimation of $\beta^\intercal\Sigma\beta$, where instead of using the expression $\mathrm{Q} = \mathbb{E}(y_i^2) - \sigma^2$ and estimating σ^2 first, we estimate Q directly by calibrating the plug-in estimator. This new perspective establishes a general machinery taking reasonable good initial estimators of β and Σ as inputs. As shown in (6), the flexibility of the calibrated estimator has proven useful in efficiently pooling additional information on Σ while the estimation method introduced in Verzelen and Gassiat [2016] cannot be directly extended to integrating the unlabelled data in the semi-supervised setting.

In the numerical studies, we have demonstrated that the effect of including unlabelled data is of great practical significance, where in the case of dense Σ, the RMSE of the new proposed CHIVE estimator is 60% to 70% smaller than the estimators in (16) without using the unlabelled data. See Table 1 in Section 6 for details.

Additionally, Verzelen and Gassiat [2016] focused on the estimation problem instead of confidence interval construction and hypothesis testing problems. In terms of technical details on estimation optimality, the results in Verzelen and Gassiat [2016] allowed for a more general regime $k \geq \sqrt{p}$ than Corollary 3 but did not handle the optimality in the semi-supervised setting and did not allow the signal strength M to grow with n, p.

2.4. Case II: Known Σ

The general semi-supervised results also shed light on another interesting setting where the design covariance Σ is known. In the semi-supervised setting, the unlabelled data is used for estimating Σ, so the case of known Σ is an extreme case of the semi-supervised setting with N taken as infinity. The estimator (11) can be modified as $\widehat{\mathrm{Q}}(\widehat{\beta}, \Sigma, Z^{(2)}) = \widehat{\beta}^\intercal\Sigma\widehat{\beta} + 2\widehat{\beta}^\intercal\frac{1}{n_2}\sum_{i=n_1+1}^n X_i^\intercal(y_i - X_{i\cdot}\widehat{\beta})$. Similarly, the estimator proposed in (6) is changed to $\widehat{\mathrm{Q}}(\widehat{\beta}, \Sigma) = \widehat{\beta}^\intercal\Sigma\widehat{\beta} + 2\widehat{\beta}^\intercal\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - X_{i\cdot}^\intercal\widehat{\beta})$.

COROLLARY 4. *Suppose that Condition* (A1) *holds and $k \leq cn/\log p$ for some constant $c > 0$.*

(a) For any estimator $\widehat{\beta}$ depending on the first half sample $\left(y^{(1)}, X^{(1)}\right)$ and satisfying Condition (B1)*, then with probability larger than $1 - \gamma(n) - C(p^{-c} + \exp(-ct^2))$,*

$$\left|\widehat{\mathrm{Q}}(\widehat{\beta}, \Sigma, Z^{(2)}) - \mathrm{Q}\right| \lesssim (t+1)\frac{\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}} + \frac{k\log p}{n}\sigma^2. \tag{17}$$

(b) For any estimator $\widehat{\beta}$ satisfying Condition (B1)*, then with probability larger than* $1-\gamma(n)-C(p^{-c}+\exp(-ct^2))$*,*

$$\left|\widehat{\mathrm{Q}}(\widehat{\beta},\Sigma)-\mathrm{Q}\right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}}+(\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}+1)\frac{k\log p}{n}\sigma^2. \tag{18}$$

Through comparing (17) with (12) and (18) with (7), the uncertainty of estimating the design covariance matrix leads to the additional term $\beta^\intercal\Sigma\beta/\sqrt{N+n}$. By applying Theorem 2, it can be shown that the upper bound in (17) leads to the optimal convergence rate $M/\sqrt{n}+k\log p/n$. The term $M^2/\sqrt{N+n}$ disappears due to the known design covariance matrix Σ.

3. Semi-supervised Confidence Intervals for $\beta^\intercal\Sigma\beta$

In this section, we quantify the uncertainty of the CHIVE estimator proposed in Section 2 and then construct confidence intervals for $\beta^\intercal\Sigma\beta$ in the semi-supervised setting.

3.1. Confidence Interval Construction

The main next step of confidence interval construction for Q is to consistently estimate the standard error $\sqrt{4\sigma^2\beta^\intercal\Sigma\beta+\rho\mathbb{E}\left(\beta^\intercal X_1.X_1.^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2}/\sqrt{n}$ of the limiting distribution established in (8). Specifically, we estimate $4\sigma^2\beta^\intercal\Sigma\beta$ by $\widehat{\phi}_1$, ρ by $\widehat{\rho}=n/(N+n)$ and $\mathbb{E}\left(\beta^\intercal X_1.X_1.^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2$ by $\widehat{\phi}_2$, where $\widehat{\phi}_1=\widehat{\sigma}^2\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}$ and $\widehat{\phi}_2=\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_i.X_i.^\intercal\widehat{\beta}-\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}\right)^2$, with $\widehat{\Sigma}^S$ defined in (6). Then we propose the following confidence interval centered at $\widehat{\mathrm{Q}}$,

$$\mathrm{CI}(Z)=[(\widehat{\mathrm{Q}}-z_{\alpha/2}\widehat{\phi})_+,\ \widehat{\mathrm{Q}}+z_{\alpha/2}\widehat{\phi}],\ \text{where}\ \widehat{\phi}=\sqrt{(4\widehat{\phi}_1+\widehat{\rho}\widehat{\phi}_2)/n}, \tag{19}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution. The following theorem establishes the coverage and precision properties of $\mathrm{CI}(Z)$, where the length of the interval $\mathrm{CI}(Z)=(L(Z),U(Z))$ is defined as $\mathbf{L}(\mathrm{CI}(Z))=U(Z)-L(Z)$.

THEOREM 4. *Suppose that Conditions* (A1) *and* (A2) *hold,* $k\ll\min\{n/(\log(N+n)\log p),\sqrt{n}/\log p\}$ *and* $\mathrm{SNR}\gg k\log p/\sqrt{n}$*. For $\widehat{\beta}$ and $\widehat{\sigma}^2$ satisfying Conditions* (B1) *and* (B2)*, respectively, the confidence interval given in* (19) *satisfies,*

$$\underline{\lim}_{n\to\infty}\ \mathbb{P}\left(\beta^\intercal\Sigma\beta\in\mathrm{CI}(Z)\right)\geq 1-\alpha \tag{20}$$

$$\lim_{n\to\infty}\mathbb{P}\left(\mathbf{L}(\mathrm{CI}(Z))\geq(1+\delta_0)\sqrt{4\sigma^2\beta^\intercal\Sigma\beta/n+\mathbb{E}\left(\beta^\intercal X_1.X_1.^\intercal\beta-\beta^\intercal\Sigma\beta\right)^2/(N+n)}\right)=0 \tag{21}$$

for any positive constant $\delta_0>0$.

The effect of the unlabelled data on the length of confidence interval is carefully characterized in (21), where the unlabelled data shrinks part of the length of confidence interval, $\mathbb{E}\left(\beta^\intercal X_1. X_1^\intercal.\beta - \beta^\intercal\Sigma\beta\right)^2/(N+n)$. This term corresponds to the uncertainty of estimating $\beta^\intercal\Sigma\beta$ in the oracle setting of known β. The most effective regime of integrating the unlabelled data is when the ratio $\frac{\mathbb{E}(\beta^\intercal X_1. X_1^\intercal.\beta-\beta^\intercal\Sigma\beta)^2}{\sigma^2\beta^\intercal\Sigma\beta}$ is not vanishing to zero. Otherwise, the dominating term in the length of (21) is $4\sigma^2\beta^\intercal\Sigma\beta/n$ and the additional unlabelled data is not helpful in this regime. In the numerical studies, we investigate how much shorter confidence intervals can be after integrating the unlabelled data. The lengths of CIs in the semi-supervised setting can be reduced to being as short as 30% to 40% of those in the supervised setting. See Table 1 for details.

The upper bound for CI length established in (21) is further upper bounded by $\sigma\|\Sigma^{\frac{1}{2}}\beta\|_2/\sqrt{n} + \beta^\intercal\Sigma\beta/\sqrt{N+n}$, which matches the optimal convergence rate of estimation $M/\sqrt{n} + M^2/\sqrt{N+n}$ over the parameter space $\Theta(k, M)$ for $k \ll \sqrt{n}/\log p$ and $M \gg k\log p/\sqrt{n}$.

As shown in Theorem 4, the validity of the proposed confidence interval (19) requires the condition that SNR is bounded away from zero by $k\log p/\sqrt{n}$. Although $k\log p/\sqrt{n}$ converges to zero over the extreme sparse regime $k \ll \sqrt{n}/\log p$, it reveals the difficulty of constructing stable confidence intervals for $\beta^\intercal\Sigma\beta$ when SNR is at a local neighborhood of zero. The next section will address the inference problem when SNR is at a local neighborhood of 0.

3.2. Inference for Weak Signals

As discussed in the introduction, uncertainty quantification of $\mathrm{Q} = \beta^\intercal\Sigma\beta$ is closely connected to other important statistical problems, including (1) signal detection and global testing; (2) prediction accuracy evaluation and (3) confidence ball construction. These applications provide a strong motivation for studying the inference problem for the explained variance under the settings of weak signals (that is, $\mathrm{SNR} \lesssim k\log p/\sqrt{n}$). The main goal of this section is to discuss extensions of the proposed procedure to conduct statistical inference uniformly over different levels of signal strength, measured by SNR.

To begin with, we recall the reasoning for the non-uniformity assumption $\mathrm{SNR} \gg k\log p/\sqrt{n}$. This assumption is imposed such that the variance component of the CHIVE estimator dominates the bias component and in this case an asymptotic limiting distribution for the variance component is used to construct confidence interval for the explained variance. Specifically, we discuss two possible solutions to remove this stringent assumption, a) to enlarge the confidence interval by an upper bound for the bias in Section 3.2.1; b) to increase the variance level by randomized calibration in Section 3.2.2.

3.2.1. Bound the bias term

One way to construct confidence intervals uniformly over all SNR is to enlarge the estimated variance level defined in (19) to

$$\widehat{\phi}^E = \widehat{\phi}^E(y, X, \tau_0) = \sqrt{\frac{1}{n}4\widehat{\sigma}^2\left(\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} + \tau_0^2\right) + \frac{1}{(n+N)^2}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_i. X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}\right)^2}, \tag{22}$$

for some positive constant $\tau_0 > 0$. Then we construct the confidence interval as

$$\mathrm{CI}^E(Z) = [(\widehat{\mathrm{Q}} - z_{\alpha/2}\widehat{\phi}^E)_+, \quad \widehat{\mathrm{Q}} + z_{\alpha/2}\widehat{\phi}^E], \tag{23}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution. The reason for adding the term $\frac{1}{n}4\widehat{\sigma}^2\tau_0^2$ in the width (22) is that this additional term is an upper bound for the bias term in the regime $k \ll \sqrt{n}/\log p$. The following corollary establishes the coverage and the precision property of the enlarged confidence interval, $\mathrm{CI}^E(Z)$.

COROLLARY 5. *Suppose that Conditions* (A1) *and* (A2) *hold,* $k \ll \min\{n/(\log(N+n)\log p), \sqrt{n}/\log p\}$ *and* $\tau_0 > 0$ *is a positive constant. For* $\widehat{\beta}$ *and* $\widehat{\sigma}^2$ *satisfying Conditions* (B1) *and* (B2)*, respectively, then the confidence interval defined in* (23) *satisfies,*

$$\varliminf_{n\to\infty} \mathbb{P}\left(\beta^\intercal\Sigma\beta \in \mathrm{CI}^E(Z)\right) \geq 1-\alpha \tag{24}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbf{L}(\mathrm{CI}^E(Z)) \geq (1+\delta_0)\sqrt{4\sigma^2\left(\beta^\intercal\Sigma\beta+\tau_0^2\right)/n + \mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2/(N+n)}\right) = 0 \tag{25}$$

for any positive constant $\delta_0 > 0$.

In contrast to the length of confidence interval in (21), the length in (25) is enlarged by the exact amount $4\sigma^2\tau_0^2/n$. In contrast to Theorem 4, the inference is uniform over all levels of SNR at the expense of a slightly longer confidence interval.

3.2.2. Randomized Calibration

The construction in (23) still uses the CHIVE estimator as the center and enlarges the constructed confidence interval. We introduce a randomized version of the CHIVE estimator as the new center, where the main intuition is to increase the variance level through randomization such that the variance of this randomized estimator dominates its bias level. We generate random variables $u_i \overset{iid}{\sim} N(0,\tau_0^2)$ for $1 \leq i \leq n$, independent of the observed data Z and propose the following randomized calibrated estimator,

$$\widehat{\mathrm{Q}}^R = \widehat{\mathrm{Q}}^R(\widehat{\beta},\widehat{\Sigma}^S,u) = \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} + 2\frac{1}{n}\sum_{i=1}^n(X_{i\cdot}^\intercal\widehat{\beta}+u_i)(y_i - X_{i\cdot}^\intercal\widehat{\beta}). \tag{26}$$

When there is no confusion, we use $\widehat{\mathrm{Q}}^R$ to denote the estimator proposed in (26). In contrast to (6), the calibration step in (26) is involved with an additional term $2\frac{1}{n}\sum_{i=1}^n u_i(y_i - X_{i\cdot}^\intercal\widehat{\beta})$. If u_i is zero instead of being generated as normal random variables in (26), the estimator $\widehat{\mathrm{Q}}^R(\widehat{\beta},\widehat{\Sigma}^S,0)$ is reduced to being exactly the same as $\widehat{\mathrm{Q}}(\widehat{\beta},\widehat{\Sigma}^S)$ defined in (6). Since u_i in (26) is randomly generated normal random variables, this additional term approximately follows a normal distribution with mean zero and variance $4\sigma^2\tau_0^2/n$. Even in the presence of weak signals, this additional term further enlarges the variance level of the calibrated estimator such that the bias level of the calibrated estimator is dominated by the corresponding variance level. The following theorem establishes the limiting distribution of the estimator $\widehat{\mathrm{Q}}^R$ after randomized calibration.

THEOREM 5. *Suppose that Condition* (A1) *holds,* $k \ll \sqrt{n}/\log p$ *and* $\tau_0 > 0$ *is a positive constant. For any estimator* $\widehat{\beta}$ *satisfying Condition* (B1)*, then*

$$\sqrt{n}\frac{\widehat{\mathrm{Q}}^R - \mathrm{Q}}{\sqrt{4\sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right) + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1) \tag{27}$$

where $\rho = \lim \frac{n}{n+N}$.

In comparison to the limiting distribution (8) in Theorem 1, Theorem 5 requires no condition on SNR to establish the asymptotic limiting distribution while the variance level of the established normal distribution is enlarged by the amount $4\sigma^2\tau_0^2/n$. This additional variance term is a side effect of the randomized calibration. However, it enables a uniform inference procedure over all levels of SNR. Then we propose the following confidence interval, $\mathrm{CI}^R(Z) = \left[(\widehat{\mathrm{Q}}^R - z_{\alpha/2}\widehat{\phi}^E)_+, \quad \widehat{\mathrm{Q}}^R + z_{\alpha/2}\widehat{\phi}^E\right]$, where $\widehat{\phi}^E$ is defined in (22). This confidence interval has the same length as that of (23) but different centers. The proposed estimator $\widehat{\mathrm{Q}}^R$ enjoys the advantage of having an asymptotic normal distribution but it suffers from the disadvantage as all randomized procedure, where the output is random even given the same data set. The following corollary characterizes the coverage and precision properties of $\mathrm{CI}^R(Z)$.

COROLLARY 6. *Under the same conditions as Corollary 5, the coverage property in* (24) *and precision property in* (25) *hold for the confidence interval* $\mathrm{CI}^R(Z)$.

Algorithm 1 summarizes the uncertainty quantification methods for $\beta^\intercal\Sigma\beta$.

Algorithm 1: Semi-supervised Uncertainty Quantification for $\beta^\intercal\Sigma\beta$

Input : Labelled data $\{y_i, X_{i\cdot}\}_{1\le i\le n}$ and unlabelled data $\{X_{i\cdot}\}_{n+1\le i\le n+N}$; $\tau_0 > 0$

Output: Point estimator $\widehat{\mathrm{Q}} = \widehat{\mathrm{Q}}(y, X)$, $\widehat{\mathrm{Q}}^R = \widehat{\mathrm{Q}}^R(y, X, \tau_0)$ and variance estimator $\widehat{\phi}^E = \widehat{\phi}^E(y, X, \tau_0)$

1 Initialization: Construct point estimator $\widehat{\beta}$ and $\widehat{\sigma}^2$ satisfying (B1) and (B2); Estimate Σ by $\widehat{\Sigma}^S$ defined in (6);

2 Calibration: Estimate Q by the CHIVE estimator $\widehat{\mathrm{Q}}$ in (4) or its randomized version $\widehat{\mathrm{Q}}^R$ in (26).

3 Uncertainty Quantification: Quantify the error of the proposed estimator by $\widehat{\phi}^E$ defined in (22).

We conclude this section with some additional comments. Compared to point estimation, construction of confidence intervals for the explained variance is a more challenging problem, mainly due to the fact that one needs to characterize the uncertainty of the proposed estimator. Specifically, accurate estimation of Q can be conducted uniformly over all levels of SNR while construction of confidence intervals uniformly over all levels of SNR requires much more efforts. Another interesting observation is that inference for explained variance is different from that for linear functional [Zhang and Zhang, 2014, van de Geer et al., 2014, Javanmard and Montanari, 2014a,b, Cai and Guo, 2017c], where the valid inference results for the latter do not depend on the magnitude of SNR.

4. Related Semi-supervised Inference Problem

The improvement due to integrating the unlabelled data is not just limited to the inference problem for $\beta^\intercal\Sigma\beta$, but can also be obtained in the semi-supervised inference for $\|\beta\|_2^2$. This unweighted quadratic functional is different from $\beta^\intercal\Sigma\beta$ as the covariance matrix Σ does not appear in the expression. Hence, it is even unclear whether the unlabelled data can be of any help. We introduce in this section a procedure integrating the unlabelled data and also carefully quantify the improvement with making use of the additional unlabelled data in the semi-supervised setting.

The estimation of $\|\beta\|_2^2$ in the supervised setting was studied in Guo et al. [2017], where the error decomposition of the plug-in estimator $\|\widehat{\beta}\|_2^2$ was established as $\|\widehat{\beta}\|_2^2 - \|\beta\|_2^2 = 2\widehat{\beta}^\intercal(\widehat{\beta}-\beta) - (\widehat{\beta}-\beta)^\intercal(\widehat{\beta}-\beta)$. In Guo et al. [2017], the bias term $2\widehat{\beta}^\intercal(\widehat{\beta}-\beta)$ in the decomposition was estimated and hence the plug-in estimator $\|\widehat{\beta}\|_2^2$ was corrected.

We illustrate here how the additional unlabelled data facilitates the bias-correction step. We randomly split the labelled data (y, X) into two subsamples $\left(y^{(1)}, X^{(1)}\right)$ with sample size n_1 and $\left(y^{(2)}, X^{(2)}\right)$ with sample size n_2, where $n_1 \asymp n_2$. Let $\widehat{\beta}$ denote an estimator of β produced by the first sub-sample $\left(y^{(1)}, X^{(1)}\right)$ satisfying Condition (B1), where one example is the scaled Lasso estimator (5) applied to $\left(y^{(1)}, X^{(1)}\right)$. Then we construct a projection direction $\widehat{u} \in \mathbb{R}^p$ and propose the estimator $\|\widehat{\beta}\|_2^2$ as

$$\widehat{\|\beta\|_2^2} = \|\widehat{\beta}\|_2^2 + 2\widehat{u}^\intercal \frac{1}{n_2} \sum_{i=n_1+1}^{n} X_{i\cdot}\left(y_i - X_{i\cdot}^\intercal\widehat{\beta}\right). \tag{28}$$

The unlabelled data is particularly useful in estimating the projection direction $\widehat{u} \in \mathbb{R}^p$. The projection direction $\widehat{u}$ is constructed as $\widehat{u} = \widehat{\Omega}\widehat{\beta} = \sum_{l\in\mathrm{supp}(\widehat{\beta})} \widehat{\Omega}_{\cdot l}\widehat{\beta}_l$ where $\widehat{\Omega}_{\cdot l}$ is the CLIME estimator defined as

$$\widehat{\Omega}_{\cdot l} = \arg\min \|m\|_1 \quad \text{subject to} \quad \|\widetilde{\Sigma}m - e_l\|_\infty \le \lambda_S \tag{29}$$

with $\widetilde{\Sigma} = \frac{1}{N+n_1}(\sum_{i=1}^{n_1} X_{i\cdot}X_{i\cdot}^\intercal + \sum_{i=n+1}^{n+N} X_{i\cdot}X_{i\cdot}^\intercal)$ and $\lambda_S \asymp \sqrt{\log p/(n_1+N)}$. The additional unlabelled data plays a role in constructing the sample covariance matrix $\widetilde{\Sigma}$ in (29) and hence constructing the projection direction $\widehat{u}$. The specific way of including the unlabelled data to improve the estimation accuracy of $\|\beta\|_2^2$ is different from that of $\beta^\intercal\Sigma\beta$, where the additional unlabelled data is used to estimate Σ directly in estimating $\beta^\intercal\Sigma\beta$ while the additional unlabelled is used to estimate Σ^{-1} in estimating $\|\beta\|_2^2$. However, the high level idea is the same, that is, making use of the flexibility of calibrated estimator and properly incorporating the information about Σ contained in the unlabelled data.

Precision matrix estimation has been studied in the literature; see Cai et al. [2011] and the reference in the paper. We restrict the attention to $\widehat{\Omega}$ satisfying the following condition.

(B3) The estimator $\widehat{\Omega}$ satisfies $\mathbb{P}\left(\|\widehat{\Omega}-\Omega\|_2 \gtrsim C_\Omega s\sqrt{\log p/(N+n)}\right) \ge 1-\gamma_1(N+n)$ where $\gamma_1(N+n) \to 0$, $s = \max_{1\le l\le p}\|\Omega_{l\cdot}\|_0$ and C_Ω is a constant depending on $\|\Omega\|_{L_1}$.

The CLIME estimator $\widehat{\Omega} = \begin{pmatrix} \widehat{\Omega}_{\cdot 1} & \widehat{\Omega}_{\cdot 2} & \cdots \widehat{\Omega}_{\cdot p} \end{pmatrix}$ with $\widehat{\Omega}_{l\cdot}$ constructed in (29) is shown to satisfy the condition (B3) under certain regularity conditions. See the exact statement in Cai et al. [2011]. We show in the following theorem that, with a sufficiently large amount of unlabelled data, the inference results for the semi-supervised setting are distinguished from those in the supervised data.

THEOREM 6. *Suppose that Condition* (A1) *holds,* $k \leq cn/\log p$ *for some constant* $c > 0$ *and* $c_0 \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq C_0$ *for some positive constants* $C_0 \geq c_0 > 0$. *Suppose that* $\widehat{\beta}$ *satisfies* (B1) *and* $\widehat{\Omega}$ *satisfies* (B3). *Under the sample size condition* $N+n \gg C_\Omega^2 k (s \log p)^2$, *then with probability larger than* $1-\gamma(n)-C(p^{-c}+\exp(-ct^2))-\gamma_1(N+n)$,

$$\left| \widehat{\|\beta\|_2^2} - \|\beta\|_2^2 \right| \lesssim \sigma \frac{\|\beta\|_2}{\sqrt{n}} + k\frac{\log p}{n}\sigma^2. \tag{30}$$

In addition, if $\frac{1}{\sigma}\|\beta\|_2 \gg k \log p/\sqrt{n}$ *and* ϵ_i *are i.i.d Gaussian random variables, then*

$$\sqrt{n/(\sigma^2 \mathrm{V})} \left(\widehat{\|\beta\|_2^2} - \|\beta\|_2^2 \right) \xrightarrow{d} N(0,1), \quad \textit{with} \quad \mathrm{V} = 4 \sum_{i=n_1+1}^{n} (\widehat{u}^\intercal X_{i\cdot})^2 / n_2^2 \tag{31}$$

The limiting distribution in (31) leads to the confidence interval construction

$$\mathrm{CI}_{\|\beta\|_2^2} = \left(\widehat{\|\beta\|_2^2} - z_{\alpha/2}\widehat{\sigma}\sqrt{\mathrm{V}}, \widehat{\|\beta\|_2^2} + z_{\alpha/2}\widehat{\sigma}\sqrt{\mathrm{V}} \right)$$

where $\widehat{\|\beta\|_2^2}$ is defined in (28), V is defined as (31) and $\widehat{u} = \sum_{l \in \mathrm{supp}(\widehat{\beta})} \widehat{\Omega}_{\cdot l}\widehat{\beta}_l$.

A few remarks are in order for the semi-supervised inference for $\|\beta\|_2^2$. The results established in Guo et al. [2017] showed that the optimal rate for estimating $\|\beta\|_2^2$ in the supervised setting is $\frac{\sigma\|\beta\|_2}{\sqrt{n}} + (1+\|\beta\|_2)k\frac{\log p}{n}\sigma^2$. In contrast, the term $\|\beta\|_2 \cdot k\frac{\log p}{n}\sigma$ disappears in the rate of convergence (37) by efficiently incorporating the unlabelled data. The improvement varies across different signal strengths, where the reduction in RMSE is limited if the signal strength $\|\beta\|_2$ is small but is significant if $\|\beta\|_2$ is large. While integrating the unlabelled data is useful in reducing the RMSE for estimating both $\beta^\intercal\Sigma\beta$ and $\|\beta\|_2^2$, it is interesting to observe that the improvement by incorporating the unlabelled data is different, where for estimating $\beta^\intercal\Sigma\beta$, part of the variance component is reduced but for estimating $\|\beta\|_2^2$, the bias component is reduced by $\|\beta\|_2 \cdot k\frac{\log p}{n}\sigma$. More interestingly, when the size of the unlabelled data is large enough and the spectrum of Σ is bounded away from zero and infinity, the rate of estimating $\|\beta\|_2^2$ in (30) coincides with that of estimating $\beta^\intercal\Sigma\beta$ in (17).

Theorem 6 requires the additional sample size condition for the unlabelled data, $N+n \gg C_\Omega^2 k (s\log p)^2$. The general results for any $N \geq 0$ are given in Section A in the supplementary material.

The additional unlabelled data is not just useful in improving the estimation accuracy, but is also useful in confidence interval construction. The specific effect is different from that for $\beta^\intercal\Sigma\beta$; the confidence interval for $\beta^\intercal\Sigma\beta$ is shortened as in (21) while the length of confidence interval $\mathrm{CI}_{\|\beta\|_2^2}$ is not shortened in terms of order of magnitude. However,

the additional unlabelled data significantly weakens the model complexity and sample size condition for establishing the limiting distribution, where the sufficient condition for the supervised setting is $\frac{1}{\sigma}\|\beta\|_2 \gg k\log p/\sqrt{n}$ and $k \ll \sqrt{n}/\log p$. Corollary 6 has shown that the condition $k \ll \sqrt{n}/\log p$ is not needed if there is sufficient amount of unlabelled data.

5. Statistical Applications

In this section, we apply the inference procedure related to the CHIVE estimator to tackle several important statistical problems.

5.1. Application 1: Signal Detection and Global Testing

Signal detection is of great importance in statistics and related scientific applications and the detection problem in high-dimensional linear regression was studied in Arias-Castro et al. [2011], Ingster et al. [2010]. The inference procedure stated in Algorithm 1 has profound implications on signal detection and the general global testing in high-dimensional linear regression. We consider the global hypothesis testing problem $H_0 : (\beta-\beta^{\text{null}})^\intercal\Sigma(\beta-\beta^{\text{null}}) = 0$ v.s. $H_1 : (\beta-\beta^{\text{null}})^\intercal\Sigma(\beta-\beta^{\text{null}}) > 0$. , which includes signal detection as a special case with $\beta^{\text{null}} = 0$. We apply Algorithm 1 with a given $\tau_0 > 0$ and obtain the point estimator $\widehat{\mathrm{Q}}^R(y-X\beta^{\text{null}}, X, \tau_0)$ and its standard error estimator $\widehat{\phi}^E(y-X\beta^{\text{null}}, X, \tau_0)$. Then we propose the detection procedure, with Type I error controlled at $\alpha \in (0,1)$ as $D(\tau_0) = \mathbf{1}\left(\widehat{\mathrm{Q}}^R(y-X\beta^{\text{null}}, X, \tau_0) \geq \widehat{\phi}^E(y-X\beta^{\text{null}}, X, \tau_0)z_\alpha\right)$. Define the null parameter space $\mathcal{H}_0 = \left\{\theta = \left(\beta^{\text{null}}, \Sigma, \sigma\right) : \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1,\ \sigma \leq M_2\right\}$ and the local alternative parameter space as

$$\mathcal{H}_1(\Delta) = \left\{\theta = (\beta, \Sigma, \sigma) : (\beta-\beta^{\text{null}})^\intercal\Sigma(\beta-\beta^{\text{null}}) = \frac{\Delta}{\sqrt{n}},\ \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1,\ \sigma \leq M_2\right\}.$$

The following corollary establishes that $D(\tau_0)$ controls the type I error asymptotically and also establishes the asymptotic power function of the proposed test.

COROLLARY 7. *Suppose that Conditions* (A1) *and* (A2) *hold,* $\tau_0 > 0$ *is a positive constant and the vector* $\delta = \beta-\beta^{\text{null}}$ *satisfies the conditions that* $\|\delta\|_0 \ll \min\{n/(\log(N+n)\log p), \sqrt{n}/\log p\}$ *and* $\sqrt{\mathbb{E}\left(\delta^\intercal X_1.X_1^\intercal\delta - \delta^\intercal\Sigma\delta\right)^2} \geq c_0\delta^\intercal\Sigma\delta$ *for some positive constant* c_0. *Then for any* $\theta \in \mathcal{H}_0$, *the type I error is controlled,* $\overline{\lim}_{n\to\infty}\mathbb{P}_\theta(D(\tau_0)=1) \leq \alpha$. *For* $\rho > 0$ *and any* $\theta \in \mathcal{H}_1(\Delta)$ *with some positive constant* $\Delta > 0$, *then*

$$\lim_{n\to\infty}\mathbb{P}_\theta(D(\tau_0)=1) = 1-\Phi^{-1}\left(z_\alpha - \frac{\Delta}{\sqrt{4\sigma^2\left(\delta^\intercal\Sigma\delta+\tau_0^2\right)+\rho\mathbb{E}\left(\delta^\intercal X_1.X_1^\intercal\delta-\delta^\intercal\Sigma\delta\right)^2}}\right). \quad (32)$$

The assumptions of Corollary 7 are the same as those of Corollary 6 from the perspective that the conditions imposed on β in Corollary 6 are now imposed on the difference vector $\delta = \beta - \beta^{\text{null}}$. One sufficient condition for the difference vector δ being sparse is that both the true signal β and the null hypothesis β^{null} are sparse. Corollary 7 shows that for

any positive constant τ_0, $D(\tau_0)$ controls the type I error asymptotically. The asymptotic power of the proposed test is established in (32), where the additional unlabelled data proves useful in improving the power. See Section D.2 for the improvement in the numerical studies. For the finite sample performance, we have investigated how to choose the randomization level τ_0 in the simulation section. See Section D.3 for the numerical performance.

5.2. Application 2: Prediction Accuracy Assessment

Inference for explained variance has important applications to evaluating the out-of-sample prediction for a given sparse estimator $\check{\beta}$. To keep the notation consistent, we assume $\check{\beta}$ is estimated based on a training data set (X^0, y^0) and (X, y) is an independent test data to evaluate its prediction accuracy. We start with computing the residual on the test data set $y - X\check{\beta} = X(\beta - \check{\beta}) + \epsilon$. The out-of-sample prediction accuracy is defined as $\mathrm{PA}(\check{\beta}) = \mathbb{E}_{x_{\mathrm{new}}}\left(x_{\mathrm{new}}^{\intercal}(\check{\beta} - \beta)\right)^2 = (\check{\beta} - \beta)^{\intercal}\Sigma(\check{\beta} - \beta)$ and it is reduced to the explained variance for the residual model with outcome $r = y - X\check{\beta}$ and covariates X. Let $\widehat{\mathrm{Q}}^R(r, X, \tau_0)$ and $\widehat{\phi}^E(r, X, \tau_0)$ denote the outputs of Algorithm 1 with the labeled data $\{(r_i, X_{i\cdot})\}_{1\le i\le n}$ and unlabelled data $\{X_{i\cdot}\}_{n+1\le i\le n+N}$ as inputs. Then we propose the point estimator of $\mathrm{PA}(\check{\beta})$ as $\widehat{\mathrm{Q}}^R(r, X, \tau_0)$ and the interval estimator for $\mathrm{PA}(\check{\beta})$ as

$$\mathrm{CI}_{\mathrm{PA}(\check{\beta})} = [(\widehat{\mathrm{Q}}^R(r, X, \tau_0) - z_{\alpha/2}\widehat{\phi}^E(r, X, \tau_0))_+, \widehat{\mathrm{Q}}^R(r, X, \tau_0) + z_{\alpha/2}\widehat{\phi}^E(r, X, \tau_0)] \quad (33)$$

The following corollary establishes the convergence rate for the point estimator and the coverage and precision properties of the interval estimator.

COROLLARY 8. *Suppose that Conditions* (A1) *and* (A2) *hold,* $\tau_0 > 0$ *is a positive constant and* $c_0 \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le C_0$, $\sigma \le M_2$ *for some positive constants* $C_0 \ge c_0 > 0$ *and* $M_2 > 0$. *For any sparse estimator satisfying* $\|\check{\beta}\|_0 \le C\|\beta\|_0$ *and* $C > 0$,

(a) If $k \le cn/\log p$ *for some positive constant* $c > 0$, *then with probability larger than* $1 - \gamma(n) - C(p^{-c} + \exp(-cN) + e^{-ct^2})$,

$$\left|\widehat{\mathrm{Q}}^R(r, X, \tau_0) - \mathrm{Q}\right| \lesssim t\frac{\|\check{\beta} - \beta\|_2 + \tau_0}{\sqrt{n}} + t\frac{\|\check{\beta} - \beta\|_2^2}{\sqrt{N+n}} + \left(\|\check{\beta} - \beta\|_2 + 1\right)\frac{k\log p}{n} \quad (34)$$

(b) If $k \ll \min\{n/(\log(N+n)\log p), \sqrt{n}/\log p\}$ *and* $\sqrt{\mathbb{E}\left(\delta^{\intercal}X_{1\cdot}X_{1\cdot}^{\intercal}\delta - \delta^{\intercal}\Sigma\delta\right)^2} \ge c_0\delta^{\intercal}\Sigma\delta$ *for* $\delta = \beta - \check{\beta}$ *and some positive constant* c_0, *then the confidence interval defined in* (33) *satisfies the coverage property* $\underline{\lim}_{n\to\infty}\mathbb{P}\left(\mathrm{PA}(\check{\beta}) \in \mathrm{CI}_{\mathrm{PA}(\check{\beta})}\right) \ge 1 - \alpha$ *and*

$$\lim_{n\to\infty}\mathbb{P}\left(\mathbf{L}(\mathrm{CI}_{\mathrm{PA}(\check{\beta})}) \ge C\left(\frac{\|\check{\beta} - \beta\|_2 + \tau_0}{\sqrt{n}} + \frac{\|\check{\beta} - \beta\|_2^2}{\sqrt{N+n}}\right)\right) = 0 \quad (35)$$

for some constant $C > 0$.

The above corollary has shown that the precision of confidence interval for the prediction accuracy is not just related to the sample sizes n, N, the sparsity k and the dimension p,

but also related to the accuracy of the evaluated estimator $\|\check{\beta}-\beta\|_2$. As characterized in (34) and (35), the integration of the unlabelled data is useful in improving the estimation accuracy and confidence interval precision. See Sections 6.1 and D.4 for the numerical performance.

5.3. Application 3: Confidence Ball Construction

The prediction accuracy evaluation established in (33) can be used to construct confidence ball for β. For the setting where $\lambda_{\min}(\Sigma)$ is known, then we have $\lambda_{\min}(\Sigma)\|\check{\beta}-\beta\|_2^2 \leq (\check{\beta}-\beta)^\intercal\Sigma(\check{\beta}-\beta)$ and construct the confidence ball for β as

$$\mathrm{CB}(\check{\beta}) = \left\{\beta : \|\beta-\check{\beta}\|_2^2 \leq z_{\alpha/2}\frac{1}{\lambda_{\min}(\Sigma)}\widehat{\phi}^E(r,X,\tau_0)\right\} \tag{36}$$

As shown in (35), the radius of the confidence ball $\mathrm{CB}(\check{\beta})$ is upper bounded by $\frac{\|\check{\beta}-\beta\|_2+\tau_0}{\sqrt{n}}+\frac{\|\check{\beta}-\beta\|_2^2}{\sqrt{N+n}}$. To minimize the radius, we need to select the center $\check{\beta}$ for the confidence ball in (36) such that $\check{\beta}$ is sparse and $\|\check{\beta}-\beta\|_2$ is small. In the high-dimensional literature, several penalized estimators are shown to satisfy such properties, such as Lasso, scaled Lasso and Dantzig Selector.

6. Simulation Study

We carry out simulation studies in this section to demonstrate the numerical performance of the CHIVE estimator. Specifically, we illustrate the numerical improvement of pooling over the unlabelled data in Section 6.1; we compare the performance of the CHIVE estimator with the plug-in estimator in Section 6.2. Additional simulation results are postponed to Section D in the supplementary material.

We first introduce the general simulation setting up used for this section. We generate the high-dimensional linear regression (1) with the dimension $p = 800$ and the labelled data with sample size n and unlabelled data with sample size N. For the linear model (1), the covariates $\{X_{i\cdot}\}_{1\leq i\leq n}$ for the labelled data and also $\{X_{i\cdot}\}_{n+1\leq i\leq n+N}$ for the unlabelled data are generated in i.i.d. fashion to follow multivariate normal distribution with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{800\times 800}$ and the errors $\{\epsilon_i\}_{1\leq i\leq n}$ are generated as i.i.d standard normal distribution.

6.1. Effect of Pooling-over Additional Unsupervised Data

The focus of this section is to illustrate the improvement after integrating the unlabelled data in the semi-supervised setting. We first consider the inference problem for $\beta^\intercal\Sigma\beta$ and then the out-of-sample prediction loss evaluation.

Inference for $\beta^\intercal\Sigma\beta$ We fix the labelled data sample size as $n = 400$ and vary the unlabelled data sample size N across $\{2,000, 6,000, 20,000\}$. We consider the following settings for the design covariance matrix Σ and high-dimensional regression vector β,

- Across Settings 1,2 and 3, the regression coefficients are generated as $\beta_i = i/10$ for $1 \leq i \leq 0$ and $\beta_i = 0$ for $i \geq 11$; The covariance matrix Σ is generated as follows,

  - Setting 1: $\Sigma_{ij} = 0.5^{|i-j|}$;
  - Setting 2: $\Sigma_{ij} = 0.35$ for $1 \le i \ne j \le p$ and $\Sigma_{ii} = 1$ for $1 \le i \le p$;
  - Setting 3: $\Sigma_{ij} = 0.7$ for $1 \le i \ne j \le p$ and $\Sigma_{ii} = 1$ for $1 \le i \le p$.

- Across Settings 4,5 and 6, the regression coefficients are generated as $\beta_i = 1.5 \cdot 0.8^i$ for $1 \le i \le 800$. The covariance matrix Σ is generated as follows,
  - Setting 4: $\Sigma_{ij} = 0.5^{|i-j|}$;
  - Setting 5: $\Sigma_{ij} = 0.35$ for $1 \le i \ne j \le p$ and $\Sigma_{ii} = 1$ for $1 \le i \le p$;
  - Setting 6: $\Sigma_{ij} = 0.7$ for $1 \le i \ne j \le p$ and $\Sigma_{ii} = 1$ for $1 \le i \le p$.

		RMSE			Coverage		Length		
Setting	N	Semi-S	S	Ratio	Semi-S	S	Semi-S	S	Ratio
1	2000	0.420	0.733	57.2%	0.950	0.942	1.598	2.769	57.7%
	6000	0.370	0.751	49.2%	0.950	0.949	1.388	2.791	49.7%
	20000	0.341	0.732	46.6%	0.933	0.940	1.291	2.777	46.5%
2	2000	0.554	1.026	54.0%	0.933	0.928	2.077	3.834	54.2%
	6000	0.421	0.949	44.3%	0.951	0.957	1.741	3.855	45.2%
	20000	0.407	0.994	41.0%	0.940	0.948	1.581	3.838	41.2%
3	2000	0.813	1.612	50.4%	0.950	0.958	3.213	6.539	49.1%
	6000	0.642	1.654	38.8%	0.960	0.946	2.510	6.530	38.4%
	20000	0.559	1.597	35.0%	0.942	0.956	2.148	6.509	33.0%
4	2000	0.415	0.745	55.7%	0.938	0.939	1.591	2.740	58.1%
	6000	0.361	0.742	48.6%	0.932	0.935	1.383	2.742	50.4%
	20000	0.324	0.738	43.9%	0.955	0.949	1.290	2.748	46.9%
5	2000	0.589	1.088	54.2%	0.953	0.968	2.329	4.462	52.2%
	6000	0.496	1.149	43.2%	0.934	0.939	1.909	4.447	42.9%
	20000	0.465	1.181	39.4%	0.936	0.935	1.713	4.441	38.6%
6	2000	0.924	2.013	45.9%	0.962	0.949	3.698	7.689	48.1%
	6000	0.724	1.914	37.8%	0.945	0.951	2.822	7.692	36.7%
	20000	0.632	1.894	33.3%	0.935	0.959	2.371	7.696	30.8%

Table 1: Inference for $\beta^\intercal \Sigma \beta$ with $n = 400$ and $N = 2000, 6000, 20000$

Settings 1 to 3 correspond to the exact sparse case while Settings 4 to 6 correspond to the approximate sparse case. Settings 1 and 4 correspond to the case of approximated banded covariance matrix while Settings 2,3,5 and 6 are about denser covariance matrices. The simulations are replicated over 1,000 simulations. The Root Mean Squared Error (RMSE) and the coverage and length of confidence intervals are present in Table 1, where the columns under "Semi-S" correspond to the semi-supervised method and the columns under "S" correspond to the supervised method. Regarding RMSE, we observe that incorporation of unlabelled data reduces the RMSE significantly. The column under "Ratio" reports the ratio of RMSE of the semi-supervised method to that of the supervised method and RMSE of the semi-supervised method is reduced to 33% to 57% of that of the supervised method, depending on the amount of the unlabelled data and also

the structure on Σ. Since $X_{i\cdot}$ follows multivariate Gaussian, the variance component depending on the unlabelled data is expressed as $\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2/(N+n) = 2(\beta^\intercal \Sigma \beta)^2/(N+n)$. From setting 1 to setting 3, the value $\beta^\intercal \Sigma \beta$ increases as Σ becomes denser and this explains why the effect of using the unlabelled data becomes more significant; the same phenomenon holds for settings 4 to 6.

In terms of constructed confidence intervals, both confidence intervals constructed in the semi-supervised setting and supervised setting have near 95% coverage while the confidence interval constructed using the unlabelled data have much shorter lengths. Specifically, we use "Ratio" to measure the ratio of the length of CI in the semi-supervised setting to that in the supervised setting and observe that the length of confidence intervals can be reduced by as much as 70%.

The unlabelled data is not just useful in inference for $\beta^\intercal \Sigma \beta$, but also useful in prediction loss evaluation, which will be illustrated in the following.

Prediction Loss Evaluation We generate $\beta_i = i/5$ for $1 \le i \le 0$ and $\beta_i = 0$ for $i \ge 11$ and $\Sigma_{ij} = 0.5^{|i-j|}$. We fix the labelled data sample size as $n = 400$ and vary the unlabelled data sample size N across $\{2,000, 6,000, 20,000\}$. We use this generated data (both labelled and unlabelled) to evaluate the out-of-sample prediction accuracy $(\widehat{\beta}(\lambda) - \beta)^\intercal \Sigma (\widehat{\beta}(\lambda) - \beta)$, where $\widehat{\beta}(\lambda)$ is the Lasso estimator based on an independent training data $\left(X^{(0)}, y^{(0)}\right)$ with sample size 300 with the tuning parameter λ, $\widehat{\beta}\,(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \frac{\|y^{(0)} - X^{(0)}\beta\|_2^2}{2n_0} + \lambda \sum_{j=1}^p \frac{\|X_j^{(0)}\|_2}{\sqrt{n_0}} |\beta_j|$. Note that $\left(X^{(0)}, y^{(0)}\right)$ is an independent copy of the labelled data (X, y). Specifically, we consider three estimators $\widehat{\beta}(\lambda_0), \widehat{\beta}(6\lambda_0)$ and $\widehat{\beta}(10\lambda_0)$ with $\lambda_0 = \sqrt{\frac{z_{1-1/(10p)}}{n_0}}$ and use the randomization level $\tau_0 = 2$ in terms of estimating this out-of-sample prediction accuracy.

			RMSE			Coverage		Length		
Estimator	Loss	N	Semi-S	S	Ratio	Semi-S	S	Semi-S	S	Ratio
$\widehat{\beta}(\lambda_0)$	0.145	2000	0.269	0.279	96.3%	0.910	0.898	0.896	0.898	99.8%
		6000	0.270	0.281	96.3%	0.918	0.902	0.895	0.897	99.8%
		10000	0.262	0.273	96.0%	0.924	0.910	0.895	0.897	99.8%
$\widehat{\beta}(6\lambda_0)$	1.818	2000	0.294	0.363	81.2%	0.924	0.896	1.046	1.148	91.1%
		6000	0.308	0.373	82.6%	0.918	0.888	1.042	1.148	90.8%
		10000	0.299	0.368	81.1%	0.926	0.892	1.038	1.148	90.3%
$\widehat{\beta}(10\lambda_0)$	4.679	2000	0.365	0.548	66.5%	0.930	0.928	1.318	1.841	71.6%
		6000	0.378	0.553	68.3%	0.934	0.902	1.291	1.839	70.2%
		10000	0.362	0.551	65.8%	0.920	0.916	1.267	1.841	68.8%

Table 2: Inference for the out-of-sample prediction accuracy $(\widehat{\beta} - \beta)^\intercal \Sigma (\widehat{\beta} - \beta)$.

The simulations are replicated over 1,000 simulations and we report the numerical performance of both point and interval estimators of the corresponding prediction accuracy in Table 2. The observation is consistent with that for $\beta^\intercal \Sigma \beta$, where CIs in both semi-supervised and supervised settings have coverage but the semi-supervised estimators are uniformly better than the supervised estimators in terms of both RMSE and the length of CI. As observed in Table 2, across the three estimators $\widehat{\beta}(\lambda_0), \widehat{\beta}(6\lambda_0)$ and $\widehat{\beta}(10\lambda_0)$, the effect of unlabelled data is different. The effect of unlabelled data for es-

timating $\widehat{\beta}(\lambda_0)$ is marginal while the effect of unlabelled data $\widehat{\beta}(10\lambda_0)$ is much more significant, where RMSE and length of CI can be reduced by 30%. This matches with the theory, where in the simulation setting of Gaussian design, the unlabelled data reduces the term $((\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta))^2/(N+n)$ and $(\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta)$ is pretty small (0.145) for $\widehat{\beta}=\widehat{\beta}(\lambda_0)$ and is much larger (4.679) for $\widehat{\beta}=\widehat{\beta}(10\lambda_0)$.

The semi-supervised data is also useful for improving the power for signal detection. Since the detection power is only improved around 5%, we defer the detailed results to the supplementary material Section D.2.

6.2. Comparison with Other Estimators

In the following, we compare the CHIVE estimator with the plug-in estimator. We fix the size of unlabelled data at $N = 2,000$ and vary the labelled data sample size n across $\{200, 400, 600, 800, 1,000\}$. The simulations are replicated over 500 simulations. We generate the design covariance matrix as $\Sigma_{ij} = 0.5^{|i-j|}$ and the high-dimensional regression vector β across the following three settings,

a. **Setting a:** β is generated with sparsity 10 where $\beta_j = j/10$ for $1 \le j \le 10$ and $\beta_j = 0$ for $j \ge 11$;

b. **Setting b:** β is generated with sparsity 50 where $\beta_j = j/50$ for $1 \le j \le 50$ and $\beta_j = 0$ for $j \ge 51$;

c. **Setting c:** β is generated as approximate sparse vector with $\beta_j = (0.5)^{p-1}$.

We compare four different estimators, where "CHIVE" and "CHIVE.semi" stand for the CHIVE estimator in the supervised setting and semi-supervised setting, respectively; "Plugin" and "Plugin.semi" stand for the plug-in estimator $\widehat{\beta}^\intercal\widehat{\Sigma}\widehat{\beta}$ in the supervised setting and semi-supervised setting, respectively. The numerical comparison has been reported in Figure 1. Across all three settings, it is observed that the proposed CHIVE estimator has achieved uniformly much better estimation accuracy than the plug-in estimators, in both supervised and semi-supervised settings. This numerical observation demonstrates that the calibration step is useful in improving the estimation accuracy.

We shall also point out that the unlabelled data is useful only if it is incooporated in a proper way. "Plugin.semi" is another estimator also using the unlabelled data to estimate Σ, but it is only slightly better than the "Plugin" estimator. In contrast, together with the calibration machinery, "CHIVE.semi" uses the additional data in an efficient way and the corresponding RMSE is significantly reduced in comparison to the "CHIVE" estimator.

7. Real Data Application

In this section, we analyze a yeast data set reported in Bloom et al. [2013] and study how the genetic variants explain the colony sizes under different growth media. The goal is to estimate the heritability measures of colony sizes under different growth media, which represent the variance of the colony sizes explained by the genetic variants.

Bloom et al. [2013] investigated a large scale genome-wide association study of 46 quantitative traits based on 1,008 *Saccharomyces cerevisiae* segregants crossbred from a

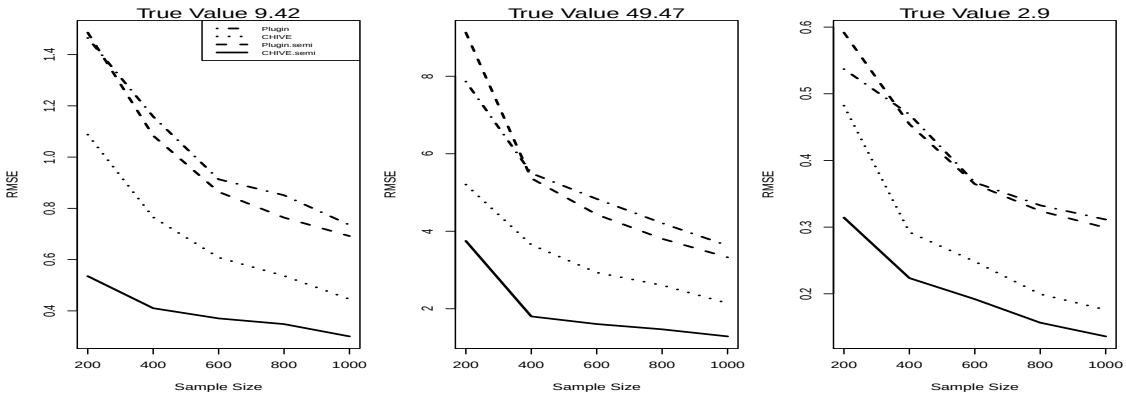

Fig. 1: Root Mean Squared Error (RMSE) of different estimators of $\beta^\intercal\Sigma\beta$. The x-axis stands for the sample size and y-axis stands for the RMSE of corresponding estimators. The dotted line and the solid line represent the corresponding RMSEs of the CHIVE estimator in the supervised setting and semi-supervised setting, respectively; The dashed line and the dotted-dashed solid line represent the corresponding RMSE of the plug-in estimator in the supervised setting and semi-supervised setting, respectively. From the leftmost to the rightmost, the figures correspond to setting a to c, with the corresponding values for $\beta^\intercal\Sigma\beta$ as 9.42, 49.47 and 2.9.

laboratory strain and a wine strain. These quantitative traits are measures of end-point colony size under 46 different growth media, including Hydrogen Peroxide, Cadmium Chloride, Calcium Chloride, Lactose, Raffinose, Sorbitol, Yeast Nitrogen Base (YNB) and Yeast Peptone Dextrose (YPD). The genetic maker genotypes are coded as 1 or -1, according to which strain it comes from. A set of 11,623 unique genotype markers of the 1,008 segregants is measured. Since many of these markers are highly correlated and the corresponding codes are only different in serval samples, Bloom et al. [2013] further selected a set of $4,410$ markers that are weakly dependent based on the linkage disequilibrium information. All traits are normalized to have unit variance and hence the explained variance is a measure for heritability. Bloom et al. [2013] showed that the genetic variants are associated with many of such trait values and highlighted the importance of addressing *missing heritability*. Bloom et al. [2013] pointed out one key reason for missing heritability as "the undiscovered factors could have effects that are too small to be detected with current sample sizes, or even too small to ever be individually detected with statistical significance". We demonstrate that the CHIVE estimator has exactly addressed this concern of missing heritability. As reported in Table 3, we choose 6 traits out of the total 46 traits and observe that the CHIVE estimates are always larger than the corresponding plug-in estimates. This means that the calibration step adds back the missing heritability due to plugging in the Lasso estimator, where the Lasso estimator tends to ignore the genetic markers with small effects. The results for all 46 traits is reported in Section E in the supplementary material.

We also construct confidence intervals for heritability of all 46 traits and report part of the results in Table 3. Note that a proportion of the outcome variables for different growth media have missing values, with the proportion of missing ranging from 0.2% to

40.58%. This forms the semi-supervised type data naturally (note that the unlabelled data is of a smaller size than the labelled data in this specific example). After applying the proposed methods to analyzing the corresponding outcomes, we have the following interesting observations, 1) the heritability measures of the colony sizes under different growth media range from 0.3 to 0.8 and all of the confidence interval estimators do not contain zero. This means the colony sizes under different growth media are strongly genetically heritable; 2) The integration of the unlabelled data has shortened the length of the constructed confidence intervals. For example the length is shorten by around 3% for Sorbitol (with 40.58% outcome missing), around 2% for Raffinose (with 34.33% outcome missing) and around 1% for Hydrogen Peroxide (with 23.71% outcome missing).

	Supervised			Semi-Supervised			
Media	Plug	CHIVE	CI	Plug	CHIVE	CI	Missing
Cadmium Chloride	0.6240	0.7682 (0.0308)	[0.7077, 0.8286]	0.6215	0.7657 (0.0306)	[0.7058, 0.8256]	20.73%
Calcium Chloride	0.1807	0.3701 (0.0323)	[0.3068, 0.4333]	0.1785	0.3679 (0.0321)	[0.3050, 0.4308]	5.85%
Hydrogen Peroxide	0.2909	0.4835 (0.0380)	[0.4090, 0.5581]	0.2879	0.4806 (0.0375)	[0.4071, 0.5540]	23.71%
Raffinose	0.3168	0.5105 (0.0410)	[0.4300, 0.5909]	0.3105	0.5041 (0.0399)	[0.4259, 0.5824]	34.33%
Sorbitol	0.2968	0.4893 (0.0431)	[0.4049, 0.5737]	0.2864	0.4789 (0.0417)	[0.3972, 0.5606]	40.58%
YPD	0.3754	0.5960 (0.0349)	[0.5275,0.6645]	0.3761	0.5966 (0.0349)	[0.5282, 0.6651]	0.20%

Table 3: Confidence intervals for heritability. The column indexed with " Media" represents the growth media for the yeast segragents; The three columns under "Supervised" corresponds to the case of only using the labelled data, where the column indexed with "Plug" represents the plug-in estimator, indexed with "CHIVE" represents the CHIVE estimator, and indexed with "CI" represents the constructed confidence interval; Similarly, the three columns under "Semi-Supervised" corresponds to analyzing the semi-supervised type data, that is also using the observations with missing outcome variables. The numbers inside the parenthesis represent the standard errors of the proposed CHIVE estimators. The column indexed with "Missing" represents the proportion of missing outcome for the corresponding media.

8. Discussions

This paper studies statistical inference for the explained variance $\beta^\intercal \Sigma \beta$ in the semi-supervised setting, which includes the supervised setting as a special case. By comparing the theoretical as well as the numerical results for the semi-supervised and supervised settings, it is easy to see the significant contributions of the unlabelled data to the inference accuracy. In addition, the constructed confidence interval, using the idea of calibration, has been shown to be useful in tackling other important statistical applications, including signal detection and global testing, prediction accuracy evaluation and confidence ball construction. There remain a few open questions for future research.

Although the CHIVE estimator has been shown to achieve the optimal rates over the

whole sparse regime $k \lesssim n/\log p$, construction of confidence intervals for $\beta^\intercal\Sigma\beta$ is only considered over the ultra-sparse regime $k \ll \sqrt{n}/\log p$. Since both point and interval estimator do not require the prior knowledge of the exact sparsity level, they are referred to as adaptive estimation and adaptive confidence interval, respectively. However, it remains open whether it is possible to construct adaptive confidence intervals over the moderate sparse regime $\sqrt{n}/\log p \lesssim k \lesssim n/\log p$. The possibility of adaptive confidence interval for the general linear functional $\eta^\intercal\beta$ for $\eta \in \mathbb{R}^p$ has been studied in Cai and Guo [2017c] and the technical tools developed in Cai and Guo [2017c] can be useful to study the adaptive confidence intervals for $\beta^\intercal\Sigma\beta$.

Due to the emerging semi-supervised data sets, it is of significant importance to propose procedures incorporating the unlabelled data efficiently and study how the unlabelled data affects the statistical accuracy. This paper has studied both methodological and theoretical perspectives of the semi-supervised statistical inference for the explained variance $\beta^\intercal\Sigma\beta$ and the unweighted quadratic functional $\|\beta\|_2^2$. However, it is largely unknown how these unlabelled data can facilitate the statistical inference problem for other quantities of interests, such as the general linear functional $\eta^\intercal\beta$ for some given $\eta \in \mathbb{R}^p$ and the variance level σ^2. These are interesting problems left for future research.

Acknowledgments

The research of Tony Cai was supported in part by NSF Grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056. The research of Zijian Guo was supported in part by NSF DMS 1811857. The authors are grateful for the constructive and helpful comments from the Editor, the Associate Editor and three referees.

References

Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, pages 2533–2556, 2011.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A*, 50(3):381–393, 1988.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Joshua S Bloom, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494 (7436):234–237, 2013.

T. Tony Cai and Zijian Guo. Accuracy assessment for high-dimensional linear regression. *Annals of Statistics*, to appear, 2017a.

T. Tony Cai and Zijian Guo. Supplement to "accuracy assessment for high-dimensional linear regression". *Annals of Statistics*, to appear, 2017b.

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

T. Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

T. Tony Cai and Mark G Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005.

T. Tony Cai and Mark G Low. Optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 34(5):2298–2325, 2006.

T. Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

T. Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Emmanuel Candès and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *arXiv preprint arXiv:1701.04889*, 2017.

Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, To appear, 2015.

David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

Sam Efromovich and Mark Low. On optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 24(3):1106–1125, 1996.

Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

Zijian Guo, Wanjie Wang, T. Tony Cai, and Hongzhe Li. Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, To appear, 2017.

Yuri I Ingster, Alexandre B Tsybakov, and Nicolas Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.

Lucas Janson, Rina Foygel Barber, and Emmanuel Candès. Eigenprism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*, 2015.

Adel Javanmard and Jason D Lee. A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971*, 2017.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *Information Theory, IEEE Transactions on*, 60(10):6522–6554, 2014a.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1): 2869–2909, 2014b.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876, 2013.

A Owen. Quasi-regression for heritability. 2012. URL `http://statweb.stanford.edu/ owen/reports/herit.pdf`.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2): 269–284, 2012.

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

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

EPA van Iperen, GK Hovingh, FW Asselbergs, and AH Zwinderman. Extending the use of gwas data by combining data from different genetic platforms. *PloS one*, 12(2): e0172082, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1602.08006*, 2016.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Cun-Hui Zhang and Stephanie S Zhang. 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, 2014.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

Yinchu Zhu and Jelena Bradic. A projection pursuit framework for testing general high-dimensional hypothesis. *arXiv preprint arXiv:1705.01024*, 2017.

A. Additional Results on Semi-supervised Inference for $\|\beta\|_2^2$

The following theorem establishes the rate of convergence of the estimator defined in (28) for general $N \geq 0$, which is more general than the results presented in Theorem 6.

THEOREM 7. *Suppose that Condition* (A1) *holds,* $k \leq cn/\log p$ *for some constant* $c > 0$ *and* $c_0 \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq C_0$ *for some positive constants* $C_0 \geq c_0 > 0$. *For any estimator* $\widehat{\beta}$ *satisfying* (B1) *and* $\widehat{\Omega}$ *satisfying* (B3), *with probability larger than* $1 - \gamma(n) - C(p^{-c} + \exp(-ct^2)) - \gamma_1(N+n)$, *then* $\widehat{\|\beta\|_2^2}$ *proposed in* (28) *satisfies*

$$\left|\widehat{\|\beta\|_2^2} - \|\beta\|_2^2\right| \lesssim \sigma \frac{\|\beta\|_2}{\sqrt{n}} \left(1 + C_\Omega \frac{s\sqrt{k}\log p}{\sqrt{N+n}}\right) + k\frac{\log p}{n}\sigma^2 \left(1 + C_\Omega \frac{s\sqrt{\log p}}{\sqrt{N+n}}\right). \tag{37}$$

The above theorem illustrates the usefulness of the unlabelled data, where the amount of the unlabelled data N plays a role in (37).

B. Proof

In this section, we will prove theorems and corollaries in the main paper and the proofs of lemmas are present in Section C. The proofs of Theorem 1 and Corollary 2 are present in Section B.1; The proof of Theorem 5 is present in Section B.2; The proof of Theorem 4 is present in Section B.3; The proof of Theorem 3 is present in Section B.4; The proof of Theorem 2 is present in Section B.5; The proof of Corollary 4 is present in Section B.6; Thr proofs of Corollaries 5 and 6 are present in Section B.7; The proofs of Theorem 7 and Corollary 6 are present in Section B.8; The proofs of Corollaries 7 and 8 are present in Sections B.9 and B.10, respectively.

B.1. Proofs of Theorem 1 and Corollary 2

To establish Theorem 1 and Corollary 2, we first decompose the difference between the calibrated estimator $\widehat{\mathrm{Q}} = \widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^S)$ and $\mathrm{Q} = \beta^\intercal \Sigma \beta$,

$$\begin{aligned}
\widehat{\mathrm{Q}} - \mathrm{Q} =& \frac{2}{n}\widehat{\beta}^\intercal X^\intercal \epsilon + \beta^\intercal \left(\widehat{\Sigma}^S - \Sigma\right)\beta - (\widehat{\beta}-\beta)^\intercal \widehat{\Sigma}^S(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta) \\
=& \frac{2}{n}\beta^\intercal X^\intercal \epsilon + \beta^\intercal \left(\widehat{\Sigma}^S - \Sigma\right)\beta + \frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal \epsilon - (\widehat{\beta}-\beta)^\intercal \widehat{\Sigma}^S(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta).
\end{aligned} \tag{38}$$

Lemma 1 characterizes the convergence rates of the last three terms in (38). The proof can be found in Section C.2 in the appendix.

LEMMA 1. *Suppose that Condition* (A1) *holds and* $k \leq cn/\log p$ *for some constant* $c > 0$. *For any estimator* $\widehat{\beta}$ *satisfying Condition* (B1), *then with probability larger than* $1 - \gamma(n) - cp^{-c} - \exp(-cN) - e^{-ct^2}$,

$$\left|\frac{1}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal \epsilon\right| \leq \|\widehat{\beta}-\beta\|_1 \left\|\frac{1}{n}X^\intercal\epsilon\right\|_\infty \lesssim \frac{k\log p}{n}\sigma^2; \tag{39}$$

$$\left|(\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\right| = \frac{1}{N+n}\sum_{i=1}^{N+n}\left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2 \lesssim \frac{k\log p}{n}\sigma^2. \tag{40}$$

$$\left|\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta)\right| \lesssim k\frac{\log p}{n}\sigma^2 + \|\Sigma^{\frac{1}{2}}\beta\|_2\sigma\left(t\frac{\sqrt{N}}{n+N}\sqrt{\frac{k\log p}{n}} + \frac{N}{n+N}\frac{k\log p}{n}\right) \tag{41}$$

For the first two terms in (38), the following lemma establishes their convergence rate and also the limiting distribution. The proof can be found in Section C.3 in the appendix.

LEMMA 2. *Suppose that Condition* (A1) *holds and* $k \le cn/\log p$ *for some constant* $c > 0$. *Then with probability larger than* $1 - e^{-ct^2}$,

$$\left|\frac{2}{n}\beta^\intercal X^\intercal\epsilon\right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}}\sigma, \quad \left|\beta^\intercal\left(\widehat{\Sigma}^S-\Sigma\right)\beta\right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2^2}{\sqrt{N+n}} \tag{42}$$

In addition, we establish the limiting distribution

$$\sqrt{n}\frac{\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \beta^\intercal\left(\widehat{\Sigma}^S-\Sigma\right)\beta}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1) \tag{43}$$

Proof of Theorem 1. In proving Theorem 1, the convergence rate in (7) follows from the decomposition (38), Lemma 1, (42) in Lemma 2 and the fact that $\frac{\sqrt{N}}{n+N}\sqrt{\frac{k\log p}{n}} \ll \frac{1}{\sqrt{n}}$. Under the additional assumptions $k \ll \sqrt{n}/\log p$ and $\mathrm{SNR} = \frac{1}{\sigma}\|\Sigma^{\frac{1}{2}}\beta\|_2 \gg k\log p/\sqrt{n}$, it follows from Lemma 1 that

$$\frac{\sqrt{n}\left(\frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal\epsilon - (\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal\left(\widehat{\Sigma}-\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^\intercal\right)(\widehat{\beta}-\beta)\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{p} 0.$$

Combined with (43) in Lemma 2, we establish the limiting distribution (8).

Proof of Corollary 2. The proof of Corollary 2 is similar to that of Theorem 1. The main change is that $\widehat{\Sigma}^S$ in (38) is replaced by $\widehat{\Sigma}^L = \frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^\intercal$ and hence the last term $2\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta)$ in the decomposition (38) becomes zero in this case. Hence, the convergence rate in (13) follows from the decomposition (38) and Lemma 1 and (42) in Lemma 2. Under the additional assumptions $\mathrm{SNR} = \frac{1}{\sigma}\|\Sigma^{\frac{1}{2}}\beta\|_2 \gg \min\left\{k\log p/\sqrt{n}, (k\log p/\sqrt{n})^{1/2}\right\}$ and the condition (A2) it follows from Lemma 1 that $\frac{\sqrt{n}\left(\frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal\epsilon - (\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}(\widehat{\beta}-\beta)\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \mathbb{E}(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta)^2}} \xrightarrow{p} 0$. Combined with (43) in Lemma 2, we establish the limiting distribution (14).

B.2. Proof of Theorem 5

Following from (38), we establish the error decomposition of $\widehat{\mathrm{Q}}^R - \mathrm{Q}$,

$$\begin{aligned}\widehat{\mathrm{Q}}^R - \mathrm{Q} = &\frac{2}{n}\beta^\intercal X^\intercal \epsilon + \frac{2}{n}u^\intercal \epsilon + \beta^\intercal\left(\widehat{\Sigma}^S - \Sigma\right)\beta + \frac{2}{n}u^\intercal X^\intercal\left(\beta - \widehat{\beta}\right) \\ &+ \frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal\epsilon - (\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta).\end{aligned} \tag{44}$$

The theorem follows from Lemma 1, the decomposition (44) and the following lemma, whose proof is postponed to Section C.4 in the appendix.

LEMMA 3. *Under the same assumptions as Theorem 5, we have*

$$\sqrt{n}\frac{\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \frac{2}{n}u^\intercal\epsilon + \beta^\intercal\left(\widehat{\Sigma}^S - \Sigma\right)\beta + \frac{2}{n}u^\intercal X^\intercal\left(\beta - \widehat{\beta}\right)}{\sqrt{4\sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right) + \rho\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1) \tag{45}$$

B.3. Proof of Theorem 4

Define $\phi_1 = \sigma^2\beta^\intercal\Sigma\beta$ and $\phi_2 = \mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2$. Recall the definitions

$$\widehat{\phi}_1 = \widehat{\sigma}^2\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} \quad \text{and} \quad \widehat{\phi}_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}\right)^2.$$

The coverage property (20) follows from the following observation,

$$\mathbb{P}\left(\beta^\intercal\Sigma\beta \in \mathrm{CI}(Z)\right) = \mathbb{P}\left(-z_{\frac{\alpha}{2}} \le \frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \beta^\intercal\Sigma\beta\right)}{\sqrt{4\phi_1 + \rho\phi_2}} \cdot \sqrt{\frac{4\phi_1 + \rho\phi_2}{4\widehat{\phi}_1 + \widehat{\rho}\widehat{\phi}_2}} \le z_{\frac{\alpha}{2}}\right) \tag{46}$$

The precision property of the constructed confidence intervals require the following lemma and the proof can be found in Section C.7 in the appendix.

LEMMA 4. *Under the same assumptions as Theorem 4, then*

$$\left|\frac{\widehat{\phi}_1 - \phi_1}{\phi_1}\right| \xrightarrow{p} 0 \tag{47}$$

$$\left|\frac{\widehat{\phi}_2 - \phi_2}{4\phi_1 + \rho\phi_2}\right| \xrightarrow{p} 0 \ \ \textit{for} \ \ \rho > 0 \tag{48}$$

$$\frac{\frac{1}{\sqrt{n}}\sqrt{4\widehat{\phi}_1 + \widehat{\rho}\widehat{\phi}_2}}{\sqrt{4\phi_1/n + \phi_2/(N+n)}} \xrightarrow{p} 1 \tag{49}$$

To establish the coverage property (20), we consider the following two cases,

(a) For the case $\rho = 0$, we have $\widehat{\rho}\widehat{\phi}_2 \geq \rho\phi_2$ and hence

$$\left|\frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \beta^\intercal\Sigma\beta\right)}{\sqrt{4\phi_1 + \rho\phi_2}} \cdot \sqrt{\frac{4\phi_1 + \rho\phi_2}{4\widehat{\phi}_1 + \widehat{\rho}\widehat{\phi}_2}}\right| \leq \left|\frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \beta^\intercal\Sigma\beta\right)}{\sqrt{4\phi_1 + \rho\phi_2}}\right| \cdot \sqrt{\frac{\phi_1}{\widehat{\phi}_1}}. \tag{50}$$

Together with (47), we establish the coverage property (20).

(b) For the case $\rho > 0$, by Lemma 4, we have $\frac{4\phi_1+\rho\phi_2}{4\widehat{\phi}_1+\widehat{\rho}\widehat{\phi}_2} \xrightarrow{p} 1$ and hence

$$\frac{\sqrt{n}\left(\widehat{\mathrm{Q}} - \beta^\intercal\Sigma\beta\right)}{\sqrt{4\phi_1 + \rho\phi_2}} \cdot \sqrt{\frac{4\phi_1 + \rho\phi_2}{4\widehat{\phi}_1 + \widehat{\rho}\widehat{\phi}_2}} \xrightarrow{d} N(0,1), \tag{51}$$

which leads to the coverage property (20).

The precision property (21) follows from (49).

B.4. Proof of Theorem 3

The error $\widehat{\mathrm{Q}}(\widehat{\beta}, \widehat{\Sigma}^{(2)}, Z^{(2)}) - \mathrm{Q}$ is decomposed as follows,

$$\frac{2}{n_2}\widehat{\beta}^\intercal(X^{(2)})^\intercal\epsilon^{(2)} + \widehat{\beta}^\intercal\left(\widehat{\Sigma}^{(2)} - \Sigma\right)\widehat{\beta} - (\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal\left(\Sigma - \frac{1}{n_2}(X^{(2)})^\intercal X^{(2)}\right)(\widehat{\beta}-\beta) \tag{52}$$

The following Lemma controls the terms involved in the above decomposition and the corresponding proof is present in Section C.5.

LEMMA 5. *With probability larger than* $1 - p^{-c_1} - e^{-c_1 t^2} - \gamma(n)$,

$$(\widehat{\beta} - \beta)^\intercal\Sigma(\widehat{\beta} - \beta) \lesssim \frac{k\log p}{n}\sigma^2, \tag{53}$$

$$\left|\widehat{\beta}^\intercal\left(\Sigma - \frac{1}{n_2}(X^{(2)})^\intercal X^{(2)}\right)(\widehat{\beta} - \beta)\right| \lesssim \frac{\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2}{\sqrt{n}}\|\Sigma^{\frac{1}{2}}(\widehat{\beta} - \beta)\|_2 \tag{54}$$

$$\left|\frac{2}{n_2}\widehat{\beta}^\intercal\left(X^{(2)}\right)^\intercal\epsilon^{(2)}\right| \lesssim t\sigma\frac{\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2}{\sqrt{n}}, \quad \left|\widehat{\beta}^\intercal\left(\widehat{\Sigma}^{(2)} - \Sigma\right)\widehat{\beta}\right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2^2}{\sqrt{N+n}} \tag{55}$$

The proof of (12) follows from the error decomposition (52), the separate error bounds in Lemma 5 and the fact $k \lesssim n/\log p$ and the upper bounds $\left|\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2 - \|\Sigma^{\frac{1}{2}}\beta\|_2\right| \leq \|\Sigma^{\frac{1}{2}}(\widehat{\beta} - \beta)\|_2$ and $\left|\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2^2 - \|\Sigma^{\frac{1}{2}}\beta\|_2^2\right| \leq \|\Sigma^{\frac{1}{2}}(\widehat{\beta} - \beta)\|_2^2 + 2\|\Sigma^{\frac{1}{2}}\beta\|_2\|\Sigma^{\frac{1}{2}}(\widehat{\beta} - \beta)\|_2$.

B.5. Proof of Theorem 2

We start with introducing two definitions. Define the χ^2 distance between two distributions $f_1(z)$ and $f_0(z)$ as $\chi^2(f_1, f_0) = \int \frac{(f_1(z)-f_0(z))^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1$ and the total variation distance as $L_1(f_1, f_0) = \int |f_1(z) - f_0(z)|\, dz$. It is well known that

$$L_1(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}. \tag{56}$$

Part of the lower bound in Theorem 2 follows from the lower bounds established in Guo et al. [2017], where, using the the current paper's terminology, equation (29) of Theorem 3 in Guo et al. [2017] is expressed as

$$\inf_{\widetilde{\|\beta\|_2^2}} \sup_{\theta \in \Theta(k,M)} \mathbb{P}\left(\left| \widetilde{\|\beta\|_2^2} - \|\beta\|_2^2 \right| \gtrsim \min\left\{ M/\sqrt{n} + k \log p/n, M^2 \right\} \right) \geq \frac{1}{4}. \tag{57}$$

The constructed least favorable null and alternative hypotheses in the proof of (57) belong to the subspace $\theta \in \Theta\,(k, M) \cap \{\Sigma = \mathrm{I}\}$. For $\Sigma = \mathrm{I}$, $\mathrm{Q} = \beta^\intercal \Sigma \beta$ is reduced to $\|\beta\|_2^2$ and (57) implies the following lower bound,

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\theta \in \Theta(k,M)} \mathbb{P}\left(\left| \widetilde{\mathrm{Q}} - \mathrm{Q} \right| \gtrsim \min\left\{ M/\sqrt{n} + k \log p/n, M^2 \right\} \right) \geq \frac{1}{4}. \tag{58}$$

It remains to establish the additional term of the lower bound $M^2/\sqrt{N+n}$, whose proof is based on the following version of Le Cam's Lemma (stated as Lemma 4 in Guo et al. [2017]; See also LeCam [1973], Yu [1997], Ren et al. [2015]).

LEMMA 6. *Let* $\mathrm{T}\,(\theta)$ *denote a functional on* θ. *Suppose that* $\theta_0, \theta_1 \in \Theta$, $\mathcal{H}_0 = \{\theta_0\}$ *and* $\mathcal{H}_1 = \{\theta_1\}$ *and* $d = |\mathrm{T}\,(\theta_1) - \mathrm{T}\,(\theta_0)|$. *Then we have*

$$\inf_{\widehat{\mathrm{T}}} \sup_{\theta \in \mathcal{H}_0 \cup \mathcal{H}_1} \mathbb{P}_\theta \left(\left| \widehat{\mathrm{T}} - \mathrm{T}\,(\theta) \right| \geq \frac{d}{2} \right) \geq \frac{1 - L_1\,(f_{\theta_1}, f_{\theta_0})}{2}. \tag{59}$$

To establish the lower bound $M^2/\sqrt{N+n}$, we need to perturb the design covariance matrix and introduce the following null and alternative parameter spaces,

$$\begin{aligned} \mathcal{H}_0 &= \{\theta_0 = (\beta, \mathrm{I}, \sigma_0)\} \\ \mathcal{H}_1 &= \left\{ \theta_1 = \left(\beta, \mathrm{I} + \frac{c}{\sqrt{N+n}\|\beta\|_2^2} \beta\beta^\intercal, \sigma_0 \right) \right\}, \end{aligned} \tag{60}$$

where $\beta \in \mathbb{R}^p$ satisfies $\|\beta\|_0 \leq k$ and $\|\beta\|_2 = M$ and $c = \min\left\{ \sqrt{\log\left(1 + \left(\frac{1}{4} - \frac{\alpha}{2}\right)^2\right)}, M_1 - 1 \right\}$. Note that $\mathcal{H}_0, \mathcal{H}_1 \in \Theta(k, M)$. Since the conditional distribution $f(y|X)$ is the same under both θ_0 and θ_1, then we have the decompositions $f_{\theta_0}(y, X) = f(y|X) f_{\theta_0}(X)$ and $f_{\theta_1}(y, X) = f(y|X) f_{\theta_1}(X)$ and hence

$$\begin{aligned} &\int\int |f_{\theta_0}(y, X) - f_{\theta_1}(y, X)|\, dX dy = \int\int f(y|X)\, |f_{\theta_1}(X) - f_{\theta_0}(X)|\, dX dy \\ &= \int \left(\int f(y|X) dy \right) |f_{\theta_1}(X) - f_{\theta_0}(X)|\, dX = L_1(f_{\theta_1}(X), f_{\theta_0}(X)). \end{aligned} \tag{61}$$

Hence, it is sufficient to control the L_1 or χ^2 distance between $f_{\theta_1}(X)$ and $f_{\theta_0}(X)$. To control the distance, we introduce the following Lemma, which was established in Cai and Zhou [2012], Ren et al. [2015] and stated as Lemma 3 in Cai and Guo [2017b].

LEMMA 7. *Let g_i be the density function of $N(0,\Sigma_i)$ for $i=0,1,2$, respectively. Then*

$$\int \frac{g_1 g_2}{g_0} = \left(\det\left(\mathrm{I} - \Sigma_0^{-1}\left(\Sigma_1 - \Sigma_0\right)\Sigma_0^{-1}\left(\Sigma_2 - \Sigma_0\right)\right)\right)^{-\frac{1}{2}}.$$

Note that $\chi^2(f_{\theta_1}(X), f_{\theta_0}(X)) + 1 = \prod_{i=1}^n \int \frac{f_{\theta_1}^2(X_{i\cdot})}{f_{\theta_0}(X_{i\cdot})}$. By applying Lemma 7 with $\Sigma_0 = \mathrm{I}$ and $\Sigma_1 = \Sigma_2 = \mathrm{I} + \frac{c_0}{\sqrt{N+n}\|\beta\|_2^2}\beta\beta^\intercal$, we have

$$\chi^2(f_{\theta_1}(X), f_{\theta_0}(X)) + 1 = \left(\det\left(\mathrm{I} - \frac{c_0^2}{(N+n)\|\beta\|_2^2}\beta\beta^\intercal\right)\right)^{-\frac{N+n}{2}} = \left(1 - \frac{c_0^2}{N+n}\right)^{-\frac{N+n}{2}}.$$

For a sufficient small c such that $\frac{c^2}{N+n} < \frac{\log 2}{2}$, we have $\left(1-\frac{c^2}{N+n}\right)^{-\frac{N+n}{2}} \le \exp\left(c^2\right) \le 1 + \left(\frac{1}{4} - \frac{\alpha}{2}\right)^2$, where the first inequality follows from the inequality $\frac{1}{1-x} \le \exp(2x)$ for $x \in [0, \frac{\log 2}{2})$ and the second inequality follows from the definition of c. By (56), we have $L_1(f_{\theta_1}(X), f_{\theta_0}(X)) \le \frac{1}{4} - \frac{\alpha}{2}$. To apply Lemma 6, we consider the functional $\mathrm{Q}(\theta) = \beta^\intercal\Sigma\beta$ and calculate

$$|\mathrm{Q}(\theta_1) - \mathrm{Q}(\theta_0)| = \left|\beta^\intercal\beta - \beta^\intercal\left(\mathrm{I} + \frac{c}{\sqrt{N+n}\|\beta\|_2^2}\beta\beta^\intercal\right)\beta\right| = c\frac{\|\beta\|_2^2}{\sqrt{N+n}} = c\frac{M^2}{\sqrt{N+n}}.$$

By applying Lemma 6, we establish

$$\inf_{\widetilde{\mathrm{Q}}} \sup_{\theta\in\Theta(k,M)} \mathbb{P}\left(\left|\widetilde{\mathrm{Q}} - \mathrm{Q}\right| \ge \frac{c}{2}\frac{M^2}{\sqrt{N+n}}\right) \ge \frac{1}{4} + \frac{\alpha}{2}. \tag{62}$$

Combining (58) and (62), we establish the theorem.

B.6. Proof of Corollary 4

To establish (17), we decompose the error $\widehat{\mathrm{Q}}(\widehat{\beta}, \Sigma, Z^{(2)}) - \mathrm{Q}$ as follows,

$$\frac{2}{n_2}\widehat{\beta}^\intercal(X^{(2)})^\intercal\epsilon^{(2)} - (\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal\left(\Sigma - \frac{1}{n_2}(X^{(2)})^\intercal X^{(2)}\right)(\widehat{\beta}-\beta).$$

Then (17) follows from the above decomposition and Lemma 5. To establish (18), the error $\widehat{\mathrm{Q}}(\widehat{\beta}, \Sigma, Z) - \mathrm{Q}$ is decomposed as

$$\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal\epsilon - (\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta) + 2\widehat{\beta}^\intercal(\Sigma - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta) \tag{63}$$

By (39), (42) and (53), we have

$$\left|\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \frac{2}{n}(\widehat{\beta}-\beta)^\intercal X^\intercal\epsilon - (\widehat{\beta}-\beta)^\intercal\Sigma(\widehat{\beta}-\beta)\right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sqrt{n}}\sigma + \frac{k\log p}{n}\sigma^2. \tag{64}$$

By the similar argument as the proof of (41), we establish $\left|2\widehat{\beta}^\intercal(\Sigma - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta)\right| \lesssim \frac{k\log p}{n}\sigma^2 + \|\Sigma^{\frac{1}{2}}\beta\|_2\frac{k\log p}{n}\sigma$. Together with (63) and (64), we establish (18).

B.7. Proof of Corollaries 5 and 6

Corollary 6 follows from Theorem 5 and the consistency of the standard deviation estimator $\widehat{\phi}^R$. Define $\phi_3 = \sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right)$ and $\widehat{\phi}_3 = \widehat{\sigma}^2\left(\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} + \tau_0^2\right)$. Using the same proof of Lemma 4, we can establish the following lemma.

LEMMA 8. *Under the same assumptions as Corollary 6, then*

$$\left|\frac{\widehat{\phi}_2 - \phi_2}{4\phi_3 + \rho\phi_2}\right| \xrightarrow{p} 0 \ \textit{for}\ \rho > 0 \ \textit{and}\ \left|\frac{\widehat{\phi}_3 - \phi_3}{\phi_3}\right| \xrightarrow{p} 0 \tag{65}$$

$$\frac{\frac{1}{\sqrt{n}}\sqrt{4\widehat{\phi}_3 + \widehat{\rho}\widehat{\phi}_2}}{\sqrt{4\phi_3/n + \phi_2/(N+n)}} \xrightarrow{p} 1 \tag{66}$$

Applying the same argument as (50) and (51), we establish (24) for CI^R; By (66), we establish (25) for CI^R. The proof of corollary 5 follows from the same argument as the proof of Theorem 4 and the fact that

$$(1 + \frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}\frac{N}{n+N})\frac{k\log p}{n}\sigma^2 \ll \sqrt{\frac{1}{n}4\sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right)} \ \text{ if } \ k \ll \frac{\sqrt{n}}{\log p}.$$

Together with Lemma 8, the bias term $(1 + \frac{\|\Sigma^{\frac{1}{2}}\beta\|_2}{\sigma}\frac{N}{n+N})\frac{k\log p}{n}\sigma^2$ in (7) is upper bounded by the width of the enlarged confidence interval.

B.8. Proof of Theorem 7 and Corollary 6

The proofs rely on the error decomposition of the proposed estimator $\widehat{\|\beta\|_2^2}$

$$\widehat{\|\beta\|_2^2} - \|\beta\|_2^2 = \frac{2}{n_2}\widehat{\beta}^\intercal\widehat{\Omega}X_{i\cdot}^\intercal\epsilon_i + 2\widehat{\beta}^\intercal\left(\widehat{\Omega}\frac{1}{n_2}\sum_{i=n_1+1}^n X_{i\cdot}X_{i\cdot}^\intercal - \mathrm{I}\right)\left(\beta - \widehat{\beta}\right) - (\widehat{\beta}-\beta)^\intercal(\widehat{\beta}-\beta), \tag{67}$$

and the decomposition of the second term on the right hand side of (67),

$$\begin{aligned}
&2\widehat{\beta}^\intercal\left(\widehat{\Omega}\frac{1}{n_2}\sum_{i=n_1+1}^n X_{i\cdot}X_{i\cdot}^\intercal - \mathrm{I}\right)\left(\beta - \widehat{\beta}\right) \\
&= 2\widehat{\beta}^\intercal\left(\widehat{\Omega} - \Omega\right)\frac{1}{n_2}\sum_{i=n_1+1}^n X_{i\cdot}X_{i\cdot}^\intercal\left(\beta - \widehat{\beta}\right) + 2\widehat{\beta}^\intercal\Omega\left(\frac{1}{n_2}\sum_{i=n_1+1}^n X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\left(\beta - \widehat{\beta}\right).
\end{aligned} \tag{68}$$

To establish the rate of convergence for estimating $\|\beta\|_2^2$, we introduce the following lemma, whose proof is deferred to Section C.6.

LEMMA 9. *Under the assumption of Theorem 7, then with probability larger than* $1 - p^{-c} - \gamma(n) - e^{-ct^2}$,

$$\left| \frac{2}{n_2} \sum_{i=n_1+1}^{n} \widehat{\beta}^\intercal \widehat{\Omega} X_{i\cdot}^\intercal \epsilon_i \right| \lesssim \sigma \sqrt{\frac{1}{n_2} \widehat{\beta}^\intercal \widehat{\Omega} \Sigma \widehat{\Omega} \widehat{\beta}} \lesssim \|\widehat{\Omega}\widehat{\beta}\|_2 \frac{\sigma}{\sqrt{n_2}} \tag{69}$$

$$\left| 2\widehat{\beta}^\intercal \left(\widehat{\Omega} - \Omega\right) \frac{1}{n_2} \sum_{i=n_1+1}^{n} X_{i\cdot} X_{i\cdot}^\intercal \left(\beta - \widehat{\beta}\right) \right| \lesssim \| \left(\widehat{\Omega} - \Omega\right) \widehat{\beta} \|_2 \|\Sigma^{\frac{1}{2}} (\widehat{\beta} - \beta)\|_2 \tag{70}$$

$$\left| 2\widehat{\beta}^\intercal \Omega \left(\frac{1}{n_2} \sum_{i=n_1+1}^{n} X_{i\cdot} X_{i\cdot}^\intercal - \Sigma \right) \left(\beta - \widehat{\beta}\right) \right| \lesssim \frac{1}{\sqrt{n_2}} \|\Omega\widehat{\beta}\|_2 \|\Sigma^{\frac{1}{2}} (\widehat{\beta} - \beta)\|_2 \tag{71}$$

By the error decomposition (67), (68) and Lemma 9, we have

$$\left| \widehat{\|\beta\|_2^2} - \|\beta\|_2^2 \right| \lesssim \sigma \frac{\|\beta\|_2}{\sqrt{n}} + k \frac{\log p}{n} \sigma^2 + \|(\widehat{\Omega} - \Omega)\widehat{\beta}\|_2 \sqrt{\frac{k \log p}{n}} \sigma. \tag{72}$$

The rate of convergence in (37) follows from (72), Condition (B3) and the following inequality,

$$\|(\widehat{\Omega} - \Omega)\widehat{\beta}\|_2 \leq \|(\widehat{\Omega} - \Omega)\|_2 \left(\|\beta\|_2 + \|\beta - \widehat{\beta}\|_2 \right).$$

The rate of convergence in (30) follows from (37) and (**??**). Note that

$$\frac{\frac{2}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \widehat{\beta}^\intercal \widehat{\Omega} X_{i\cdot}^\intercal \epsilon_i}{\sqrt{\frac{4\sigma^2}{n_2} \widehat{u}^\intercal \sum_{i=n_1+1}^{n} X_{i\cdot}^\intercal X_{i\cdot} \widehat{u}}} \xrightarrow{d} N(0,1) \tag{73}$$

and

$$\frac{\sqrt{n_2} k \frac{\log p}{n} \sigma^2}{\sqrt{\frac{4\sigma^2}{n_2} \widehat{u}^\intercal \sum_{i=n_1+1}^{n} X_{i\cdot}^\intercal X_{i\cdot} \widehat{u}}} \asymp \frac{k \log p / \sqrt{n}}{\|\widehat{\Omega}\widehat{\beta}\|_2} \tag{74}$$

Since

$$\|\widehat{\Omega}\widehat{\beta} - \Omega\beta\|_2 \leq \|(\widehat{\Omega} - \Omega)\|_2 \left(\|\beta\|_2 + \|\beta - \widehat{\beta}\|_2 \right) + \|\Omega(\widehat{\beta} - \beta)\|_2, \tag{75}$$

we have

$$\begin{aligned} \|\widehat{\Omega}\widehat{\beta}\|_2 &\geq \|\Omega\beta\|_2 - \|\widehat{\Omega}\widehat{\beta} - \Omega\beta\|_2 \\ &\geq \|\Omega\beta\|_2 - \|(\widehat{\Omega} - \Omega)\|_2 \left(\|\beta\|_2 + \|\beta - \widehat{\beta}\|_2 \right) - \|\Omega(\widehat{\beta} - \beta)\|_2 \end{aligned} \tag{76}$$

Under condition (B3) and the sample size condition (**??**), we have $\|\widehat{\Omega}\widehat{\beta}\|_2 \gtrsim \|\beta\|_2 - \sqrt{\frac{k \log p}{n}} \sigma$. Under the assumption $\frac{1}{\sigma}\|\beta\|_2 \gg k \log p / \sqrt{n}$, we have $\frac{\sqrt{n_2} k \frac{\log p}{n} \sigma^2}{\sqrt{\frac{4\sigma^2}{n_2} \widehat{u}^\intercal \sum_{i=n_1+1}^{n} X_{i\cdot}^\intercal X_{i\cdot} \widehat{u}}} \xrightarrow{p} 0$. Combined with (73), we establish (31).

B.9. Proof of Corollary 7

By applying Corollary 6 to the following linear model,

$$y - X\beta^{\text{null}} = X(\beta - \beta^{\text{null}}) + \epsilon, \tag{77}$$

we establish the following limiting distribution,

$$\sqrt{n}\frac{\widehat{\mathrm{Q}}^R(y - X\beta^{\text{null}}, X, \tau_0) - (\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}})}{\text{SE}} \to N(0,1), \tag{78}$$

where $\text{SE} = \sqrt{4\sigma^2\left(\delta^\intercal\Sigma\delta + \tau_0^2\right) + \rho\mathbb{E}\left(\delta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\delta - \delta^\intercal\Sigma\delta\right)^2}$ with $\delta = \beta - \beta^{\text{null}}$. Hence $\mathbb{P}\left(D(\tau_0) = 1\right)$ can be expressed as

$$\mathbb{P}\left(\frac{\widehat{\mathrm{Q}}^R(y - X\beta^{\text{null}}, X, \tau_0) - (\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}})}{\text{SE}} \geq \frac{\widehat{\phi}^E(y - X\beta^{\text{null}}, X, \tau_0)z_\alpha - (\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}})}{\text{SE}}\right)$$

Note that $\lim_{n\to\infty}\frac{\widehat{\phi}^E(y - X\beta^{\text{null}}, X, \tau_0)}{\text{SE}/\sqrt{n}} \geq 1$, where the equality holds as long as $\rho > 0$. By the limiting distribution (78), we show that

$$\lim_{n\to\infty}\mathbb{P}\left(D(\tau_0) = 1\right) \leq \Phi^{-1}\left(z_\alpha - \frac{\sqrt{n}(\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}})}{\text{SE}}\right), \tag{79}$$

where the equality holds as long as $\rho > 0$. By applying (79) with $(\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}}) = 0$, we controls the type I error; For the case $\rho > 0$, by applying (79) with $(\beta - \beta^{\text{null}})^\intercal\Sigma(\beta - \beta^{\text{null}}) = \frac{\Delta}{\sqrt{n}}$, we establish (32).

B.10. Proof of Corollary 8

The estimation bound (34) follows from the argument of Theorem 1 and the decomposition of (44). Note that the additional randomization term can be controlled as in (45).

The proof of the coverage and precision properties follows from the application of Corollary 6 to the following linear model,

$$y - X\check{\beta} = X(\beta - \check{\beta}) + \epsilon. \tag{80}$$

Note that the precision property also relies on the following observation,

$$\mathbb{E}\left(\delta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\delta - \delta^\intercal\Sigma\delta\right)^2 \leq 4\|X_{1\cdot}\|_{\psi_2}^2\|\delta\|_2^4,$$

which follows from Lemma 10 and the definition of sub-exponential random variable.

C. Proof of Lemmas

To establish the technical lemmas, we introduce the following definitions. For a random variable U, its sub-gaussian norm is defined as $\|U\|_{\psi_2} = \sup_{q\geq 1}\frac{1}{\sqrt{q}}\left(\mathbb{E}|U|^q\right)^{\frac{1}{q}}$, and its sub-exponential norm is defined as $\|U\|_{\psi_1} = \sup_{q\geq 1}\frac{1}{q}\left(\mathbb{E}|U|^q\right)^{\frac{1}{q}}$. For a random vector

$U \in \mathbb{R}^p$, its sub-gaussian norm is defined as $\|U\|_{\psi_2} = \sup_{v\in S^{p-1}} \|\langle v, U\rangle\|_{\psi_2}$ and sub-exponential norm is defined as $\|U\|_{\psi_1} = \sup_{v\in S^{p-1}} \|\langle v, U\rangle\|_{\psi_1}$, where S^{p-1} is the unit sphere in $\mathbb{R}^p$. The following lemma shows that the product of two sub-gaussian variables is a sub-exponential variable, whose proof is present in Section C.1.

LEMMA 10. *Suppose that U and V are sub-gaussian random variables, then*

$$\|UV\|_{\psi_1} \le 2\|U\|_{\psi_2}\|V\|_{\psi_2} \quad \textit{and} \quad \|UV - \mathbb{E}UV\|_{\psi_1} \le 4\|U\|_{\psi_2}\|V\|_{\psi_2} \tag{81}$$

We introduce the following events to facilitate the proofs,

$$\begin{aligned} G_1 &= \left\{ \max\left\{ \|\widehat{\beta}-\beta\|_2^2, \frac{1}{n}\sum_{i=1}^n \left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2 \right\} \lesssim \frac{k\log p}{n}\sigma^2 \right\},\ G_2 = \left\{ \|\widehat{\beta}-\beta\|_1 \lesssim k\sqrt{\frac{\log p}{n}}\sigma \right\}, \\ G_3 &= \left\{ \|\frac{1}{n}\sum_{i=1}^n X_{i\cdot}\epsilon_i\|_\infty \lesssim C\sqrt{\frac{\log p}{n}}\sigma \right\},\ G_4 = \left\{ \frac{1}{N}\sum_{i=n+1}^{n+N} \left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2 \lesssim \frac{k\log p}{n}\sigma^2 \right\}, \end{aligned} \tag{82}$$

and

$$\begin{aligned} G_5(w,t) &= \left\{ \left| \frac{1}{n} w^\intercal X^\intercal \epsilon \right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}w\|_2}{\sqrt{n}}\sigma \right\}, \\ G_6(w,v,t) &= \left\{ \left| w^\intercal \left(\frac{1}{m}\sum_{i=1}^m X_i X_i^\intercal \right) v - w^\intercal \Sigma v \right| \lesssim t\frac{\|\Sigma^{\frac{1}{2}}w\|_2\|\Sigma^{\frac{1}{2}}v\|_2}{\sqrt{m}} \right\} \end{aligned} \tag{83}$$

for $w, v \in \mathbb{R}^p$. Define $G = \cap_{i=1}^4 G_i$. The following Lemma demonstrates that the above events happen with high probability and the corresponding proof is present in Section C.1.

LEMMA 11. *For any estimator $\widehat{\beta}$ satisfying* (B1)*, then*

$$\mathbb{P}(G) \ge 1 - \gamma(n) - cp^{-c} - \exp(-cN). \tag{84}$$

For given $w, v \in \mathbb{R}^p$ and $t > 0$, then

$$\mathbb{P}(G_5(w,t)) \ge 1 - 2\exp(-ct^2) \quad \textit{and} \quad \mathbb{P}(G_6(w,v,t)) \ge 1 - 2\exp(-ct^2). \tag{85}$$

LEMMA 12. *Suppose that the condition holds for Σ, then we have*

$$\max_{\|v_{S^c}\|_1 \le C_0\|v_S\|_1} \|\Sigma^{\frac{1}{2}}v\|_2 \le \rho_{\max}(k,\Sigma)(2+C_0)\|v\|_2. \tag{86}$$

The proof of the above lemmas is present in next subsection.

C.1. Proof of Lemmas 10, 11 and 12

Proof of Lemma 10 The proof for $\|UV\|_{\psi_1}$ follows from the following inequality

$$\|UV\|_{\psi_1} = \sup_{q\geq 1} \frac{1}{q} \left(\mathbb{E}|UV|^q\right)^{\frac{1}{q}} \leq 2\frac{1}{\sqrt{2q}} \left(\mathbb{E}|U|^{2q}\right)^{\frac{1}{2q}} \frac{1}{\sqrt{2q}} \left(\mathbb{E}|V|^{2q}\right)^{\frac{1}{2q}} \leq 2\|U\|_{\psi_2}\|V\|_{\psi_2},$$

where the first inequality is by Cauchy-Schwarz and the second inequality follows from the definition of sub-gaussian norm. The proof of the centered part $\|UV - \mathbb{E}UV\|_{\psi_1}$ follows from the upper bound for $\|UV\|_{\psi_1}$ and the remark 5.18 in Vershynin [2012].

Proof of Lemma 11 The control of the events G_1 and G_2 follows from the definition of (B1) and the following inequality,

$$\|\widehat{\beta} - \beta\|_1 \leq (1 + C_0)\|(\widehat{\beta} - \beta)_S\|_1 \leq (1 + C_0)\sqrt{k}\|(\widehat{\beta} - \beta)_S\|_2.$$

In the following, we first establish (85) and then come back to the control of events G_3 and G_4. By Lemma 10, $w^\intercal \left(X_i X_i^\intercal - \Sigma\right) v = w^\intercal \left(\Sigma^{\frac{1}{2}} Z_i Z_i^\intercal \Sigma^{\frac{1}{2}} - \Sigma\right) v = (\Sigma^{\frac{1}{2}} w)^\intercal \left(Z_i Z_i^\intercal - \mathrm{I}\right) \Sigma^{\frac{1}{2}} v$ is centered random variable with sub-exponential norm

$$\|w^\intercal \left(X_i X_i^\intercal - \Sigma\right) v\|_{\psi_1} \leq 2\|\Sigma^{\frac{1}{2}} w\|_2 \|\Sigma^{\frac{1}{2}} v\|_2 \|Z_{i\cdot}\|_{\psi_2}^2 = K_1 \|\Sigma^{\frac{1}{2}} w\|_2 \|\Sigma^{\frac{1}{2}} v\|_2$$

where $K_1 = 2\|Z_{i\cdot}\|_{\psi_2}^2$. Similarly, $w^\intercal X_{i\cdot} \epsilon_i = (\Sigma^{\frac{1}{2}} w)^\intercal Z_{i\cdot} \epsilon_i$ is centered sub-exponential random variable with sub-exponential norm $\|w^\intercal X_{i\cdot} \epsilon_i\|_{\psi_1} \leq \|\Sigma^{\frac{1}{2}} w\|_2 \|Z_{i\cdot}\|_{\psi_2} \|\epsilon_i\|_{\psi_2} \leq K_2 \|\Sigma^{\frac{1}{2}} w\|_2 \sigma$ where $K_2 = \|Z_{i\cdot}\|_{\psi_2} \|\epsilon_i/\sigma\|_{\psi_2}$. By applying Corollary 5.17 in Vershynin [2012], we have for $t \leq \sqrt{n}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n w^\intercal X_{i\cdot} \epsilon_i\right| \geq \frac{t}{\sqrt{n}} \cdot K_2 \|\Sigma^{\frac{1}{2}} w\|_2 \sigma\right) \leq 2\exp(-ct^2)$$

and for $t \leq \sqrt{m}$,

$$\mathbb{P}\left(\left|w^\intercal \left(\frac{1}{m}\sum_{i=1}^m X_i X_i^\intercal\right) v - w^\intercal \Sigma v\right| \geq \frac{t}{\sqrt{m}} \cdot K_1 \|\Sigma^{\frac{1}{2}} w\|_2 \|\Sigma^{\frac{1}{2}} v\|_2\right) \leq 2\exp(-ct^2)$$

Then (85) follows from the above two concentration inequality. Note that, on the event $\cap_{i=1}^p G_5(e_i, \sqrt{\log p})$, the event G_3 holds and hence we have $\mathbb{P}(G_3) \geq 1 - 2p\exp(-c(\sqrt{\log p})^2)$; on the event $G_6(\widehat{\beta} - \beta, \widehat{\beta} - \beta, t)$,

$$\frac{1}{N}\sum_{i=n+1}^{n+N} \left(X_{i\cdot}^\intercal (\widehat{\beta} - \beta)\right)^2 \leq \left(1 + \frac{t}{\sqrt{N}}\right) (\widehat{\beta} - \beta)^\intercal \Sigma (\widehat{\beta} - \beta) \lesssim \left(1 + \frac{t}{\sqrt{N}}\right) \frac{k \log p}{n} \sigma^2.$$

By taking $t = \sqrt{N}$, we have $\mathbb{P}(G_4) \geq 1 - 2\exp(-cN)$.

Proof of Lemma 12 For a given set S and vector v, we divide S^c into disjoint sets, $S^c = \cup_{j=1}^M T_j$ where $|T_1| = |T_2| = \cdots = |T_{M-1}| = k$ and $|T_M| \leq k$ and also

$$\min_{i \in T_j} |v_i| \geq \max_{l \in T_{j+1}} |v_l| \tag{87}$$

For any given unit vector $\delta \in \mathbb{R}^p$, we have

$$\left|\langle \delta, \Sigma^{\frac{1}{2}} v\rangle\right| \leq \left|\langle \delta, \Sigma^{\frac{1}{2}} v_S\rangle\right| + \sum_{j=1}^{M} \left|\langle \delta, \Sigma^{\frac{1}{2}} v_{T_j}\rangle\right| \leq \rho_{\max}(k, \Sigma) \left(\|v_S\|_2 + \sum_{j=1}^{M} \|v_{T_j}\|_2 \right). \quad (88)$$

Due to the property (87), we have $\|v_{T_j}\|_2 \leq \frac{1}{\sqrt{k}} \|v_{T_{j-1}}\|_1$ for $j \geq 2$ and hence

$$\sum_{j=1}^{M} \|v_{T_j}\|_2 \leq \|v_{T_1}\|_2 + \frac{1}{\sqrt{k}} \sum_{j=1}^{M-1} \|v_{T_j}\|_1 \leq \|v_{T_1}\|_2 + \frac{1}{\sqrt{k}} C_0 \|v_S\|_1 \leq \|v_{T_1}\|_2 + C_0 \|v_S\|_2 \leq (1+C_0)\|v\|_2.$$

Then we have $\|v_S\|_2 + \sum_{j=1}^{M} \|v_{T_j}\|_2 \leq (2 + C_0)\|v\|_2$. By (88), we have $\left|\langle \delta, \Sigma^{\frac{1}{2}} v\rangle\right| \leq \rho_{\max}(k, \Sigma)(2 + C_0)\|v\|_2$. Taking the maximum over the unit vector $\delta \in \mathbb{R}^p$, we have $\|\Sigma^{\frac{1}{2}} v\|_2 \leq \rho_{\max}(k, \Sigma)(2 + C_0)\|v\|_2$.

C.2. Proof of Lemma 1

The first inequality in (39) follows from the Holder's inequality while the second inequality holds under the event $G_2 \cap G_3$. On the event $G_1 \cap G_4$, the second error bound (40) follows from the following decomposition

$$\left|(\widehat{\beta} - \beta)^{\intercal} \widehat{\Sigma}^S (\widehat{\beta} - \beta)\right| = \frac{n}{N+n} \frac{1}{n} \sum_{i=1}^{n} \left(X_{i\cdot}^{\intercal}(\widehat{\beta} - \beta)\right)^2 + \frac{N}{N+n} \frac{1}{N} \sum_{i=n}^{n+N} \left(X_{i\cdot}^{\intercal}(\widehat{\beta} - \beta)\right)^2.$$

Together with (84), we establish (40). To establish (41), we start with the following decomposition,

$$\begin{aligned} &\widehat{\beta}^{\intercal}(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\intercal})(\widehat{\beta} - \beta) = (\widehat{\beta} - \beta)^{\intercal}(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\intercal})(\widehat{\beta} - \beta) \\ &+ \frac{N}{N+n}\left(\beta^{\intercal}(\frac{1}{N}\sum_{i=n+1}^{n+N} X_{i\cdot} X_{i\cdot}^{\intercal} - \Sigma)(\widehat{\beta} - \beta) - \beta^{\intercal}(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\intercal} - \Sigma)(\widehat{\beta} - \beta)\right) \end{aligned} \quad (89)$$

In the following, we are going to bound the terms separately in the above decomposition. On the event G_1, we have $(\widehat{\beta} - \beta)^{\intercal} \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\intercal} (\widehat{\beta} - \beta) \lesssim \frac{k \log p}{n} \sigma^2$; By (40), we have $(\widehat{\beta} - \beta)^{\intercal} \widehat{\Sigma}^S (\widehat{\beta} - \beta) \lesssim \frac{k \log p}{n} \sigma^2$ and hence

$$(\widehat{\beta} - \beta)^{\intercal}(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot} X_{i\cdot}^{\intercal})(\widehat{\beta} - \beta) \lesssim \frac{k \log p}{n} \sigma^2. \quad (90)$$

On the event $G_6(\beta, \widehat{\beta} - \beta, t)$, we have

$$\left|\beta^{\intercal}(\frac{1}{N}\sum_{i=n+1}^{n+N} X_{i\cdot} X_{i\cdot}^{\intercal} - \Sigma)(\widehat{\beta} - \beta)\right| \lesssim \frac{t}{\sqrt{N}} \|\Sigma^{\frac{1}{2}}\beta\|_2 \|\Sigma^{\frac{1}{2}}(\widehat{\beta} - \beta)\|_2 \lesssim \frac{t}{\sqrt{N}} \|\Sigma^{\frac{1}{2}}\beta\|_2 \sqrt{\frac{k \log p}{n}} \sigma, \quad (91)$$

where the last inequality follows from Lemma 12 and condition (B1). On the event $G_6(\beta, e_i, \sqrt{\log p})$, we have $\left|\beta^\intercal(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal - \Sigma)e_i\right| \lesssim \sqrt{\log p/n}\|\Sigma^{\frac{1}{2}}\beta\|_2$ and hence on the even $\cap_{i=1}^p G_6(\beta, e_i, \sqrt{\log p})$, we have $\|\beta^\intercal(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal - \Sigma)\|_\infty \lesssim \sqrt{\log p/n}\|\Sigma^{\frac{1}{2}}\beta\|_2$. By Holder's inequality, on the event $G_2 \cap \left(\cap_{i=1}^p G_6(\beta, e_i, \sqrt{\log p})\right)$, we have

$$\left|\beta^\intercal(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal - \Sigma)(\widehat{\beta} - \beta)\right| \lesssim \|\Sigma^{\frac{1}{2}}\beta\|_2 \frac{k\log p}{n}\sigma \tag{92}$$

By applying (90), (91) and (92) to the decomposition (89), we establish that with probability larger than $1 - p^{-c} - \gamma(n) - e^{-ct^2}$,

$$\left|\widehat{\beta}^\intercal(\widehat{\Sigma}^S - \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta} - \beta)\right| \lesssim k\frac{\log p}{n}\sigma^2 + \|\Sigma^{\frac{1}{2}}\beta\|_2\sigma\left(t\frac{\sqrt{N}}{n+N}\sqrt{\frac{k\log p}{n}} + \frac{N}{n+N}\frac{k\log p}{n}\right).$$

C.3. Proof of Lemma 2

On the event $G_5(\beta, t) \cap G_6(\beta, \beta, t)$, the inequality (42) holds. The probability control of (42) follows from (85) with taking $w = v = \beta$. Let ρ_n denote $n/(N+n)$ and hence $\rho_n \to \rho$ To establish (43), we start with the decomposition,

$$\begin{aligned}\sqrt{n}\left(\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \beta^\intercal\left(\widehat{\Sigma}^S - \Sigma\right)\beta\right) = &\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(2\beta^\intercal X_{i\cdot}\epsilon_i + \rho_n\beta^\intercal\left(X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\beta\right)\\ &+ \sqrt{\rho_n(1-\rho_n)}\frac{1}{\sqrt{N}}\sum_{i=n+1}^{N+n}\beta^\intercal\left(X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\beta\end{aligned} \tag{93}$$

Note that $\mathbb{E}\left(2\beta^\intercal X_{i\cdot}\epsilon_i + \rho_n\beta^\intercal\left(X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\beta\right)^2 = 4\sigma^2\beta^\intercal\Sigma\beta + \rho_n^2\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2$, then we have

$$\frac{\sqrt{n}\left(\frac{2}{n}\beta^\intercal X^\intercal\epsilon + \beta^\intercal\left(\widehat{\Sigma}^S - \Sigma\right)\beta\right)}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho_n^2\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1), \quad \frac{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho_n^2\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}}{\sqrt{4\sigma^2\beta^\intercal\Sigma\beta + \rho^2\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \to 1$$

$$\frac{\sqrt{\rho_n(1-\rho_n)}\frac{1}{\sqrt{N}}\sum_{i=n+1}^{N+n}\beta^\intercal\left(X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\beta}{\sqrt{\rho(1-\rho)\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2}} \xrightarrow{d} N(0,1)$$

By the above limiting distributions, together with the independence between the two terms on the right hand side of (93), we establish (43).

C.4. Proof of Lemma 3

The proof follows from that of Lemma 2. The main change is that

$$\mathbb{E}\left(2\left(\beta^\intercal X_{i\cdot} + u_i\right)\epsilon_i + \rho_n\beta^\intercal\left(X_{i\cdot}X_{i\cdot}^\intercal - \Sigma\right)\beta\right)^2 = 4\sigma^2\left(\beta^\intercal\Sigma\beta + \tau_0^2\right) + \rho_n^2\mathbb{E}\left(\beta^\intercal X_{1\cdot}X_{1\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2. \tag{94}$$

In addition, we need to show that $\sqrt{n}\frac{1}{n}u^\intercal X^\intercal(\beta-\widehat{\beta}) \xrightarrow{p} 0$. Since $\frac{u^\intercal X^\intercal(\beta-\widehat{\beta})}{\|X^\intercal(\beta-\widehat{\beta})\|}|Z=z \sim N(0,1)$, we have

$$\mathbb{P}\left(\left|\sqrt{n}\frac{1}{n}u^\intercal X^\intercal(\beta-\widehat{\beta})\right| \geq \delta_0\right) = \int \mathbb{P}\left(\left|\sqrt{n}\frac{1}{n}u^\intercal X^\intercal(\beta-\widehat{\beta})\right| \geq \delta_0|z\right) f(z)dz$$

$$= \int 2\Phi^{-1}\left(\frac{\delta_0}{\sqrt{n}\|X^\intercal(\beta-\widehat{\beta})\|}\right)dz = \int_{z\in G_1} 2\Phi^{-1}\left(\frac{\delta_0}{\sqrt{n}\|X^\intercal(\beta-\widehat{\beta})\|_2}\right)f(z)dz + P(G_1^c) \tag{95}$$

where Φ^{-1} denotes the inverse of quantile function for the standard normal random variable. By (84) and $\sqrt{n}\|X^\intercal(\beta-\widehat{\beta})\| \lesssim k\log p/\sqrt{n} \to 0$, we show that $\mathbb{P}\left(\left|\sqrt{n}\frac{1}{n}u^\intercal X^\intercal(\beta-\widehat{\beta})\right| \geq \delta_0\right) \to 0$ and hence $\sqrt{n}\frac{1}{n}u^\intercal X^\intercal(\beta-\widehat{\beta}) \xrightarrow{p} 0$.

C.5. Proof of Lemma 5

The proof of (53) and (55) in Lemma 5 follows the same arguments as those of Lemma 1 and 2. Conditioning on $\widehat{\beta}$, we have $\{\widehat{\beta}^\intercal(\Sigma - X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta)\}_{n_1+1\leq i\leq n+N}$ are i.i.d centered random variables and $\|\widehat{\beta}^\intercal(\Sigma - X_{i\cdot}X_{i\cdot}^\intercal)(\widehat{\beta}-\beta)\|_{\psi_1} \leq 2\|Z_i\|_{\psi_2}^2\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2\|\Sigma^{\frac{1}{2}}(\widehat{\beta}-\beta)\|_2$ for $n_1+1\leq i\leq n$. By applying Corollary 5.17 in Vershynin [2012], we have

$$\mathbb{P}\left(\left|\widehat{\beta}^\intercal\left(\Sigma - \frac{1}{n_2}(X^{(2)})^\intercal X^{(2)}\right)(\widehat{\beta}-\beta)\right| \geq \frac{t}{\sqrt{n_2}}\cdot 2\|Z_i\|_{\psi_2}^2\|\Sigma^{\frac{1}{2}}\widehat{\beta}\|_2\|\Sigma^{\frac{1}{2}}(\widehat{\beta}-\beta)\|_2\right) \leq 2\exp(-ct^2). \tag{96}$$

C.6. Proof of Lemma 9

The proof relies on the independence between $(\widehat{\Omega},\widehat{\beta})$ and (X_i,y_i) for $n_1+1\leq i\leq n$. On the event $G_5(\widehat{\Omega}\widehat{\beta},t)$, we have (69). On the event $G_6((\widehat{\Omega}-\Omega)\widehat{\beta},\widehat{\beta}-\beta,t)$, we establish (70). On the event $G_6(\Omega\widehat{\beta},\widehat{\beta}-\beta,t)$, we establish (71).

C.7. Proof of Lemma 4

We first establish (47) and then establish (48). Define $\Delta_1 = \widehat{\sigma}^2/\sigma^2 - 1$ and $\Delta_2 = \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}/\beta^\intercal\Sigma\beta - 1$. Then we have

$$\left|\frac{\widehat{\phi}_1}{\phi_1} - 1\right| \leq |\Delta_1| + |\Delta_2| + |\Delta_1|\cdot|\Delta_2|\,.$$

Note that

$$\Delta_2 = \frac{1}{\beta^\intercal\Sigma\beta}\left(2\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta) + (\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta) + \beta^\intercal\left(\widehat{\Sigma}^S-\Sigma\right)\beta\right). \tag{97}$$

The term $\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)$ is decomposed as

$$\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta) = \frac{1}{n+N}\sum_{i=1}^{n+N} X_{i\cdot}^\intercal(\widehat{\beta}-\beta)X_{i\cdot}^\intercal\beta \leq \sqrt{\frac{1}{n+N}\sum_{i=1}^{n+N}\left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2}\sqrt{\frac{1}{n+N}\sum_{i=1}^{n+N}(X_{i\cdot}^\intercal\beta)^2}, \tag{98}$$

where the inequality follows from the Cauchy-Schwarz inequality. Recall the definition of events in (82) and (83). On the event $G_1 \cap G_4$, then $(\widehat{\beta}-\beta)^\intercal \widehat{\Sigma}^S(\widehat{\beta}-\beta) \lesssim k\log p/n$; On the event $G_6(\beta,\beta,\sqrt{\log p})$, then $\frac{\beta^\intercal(\widehat{\Sigma}^S-\Sigma)\beta}{\beta^\intercal\Sigma\beta} \lesssim \sqrt{\frac{\log p}{n+N}}$. Together with (98), we show that on the event $G_1 \cap G_4 \cap G_6(\beta,\beta,\sqrt{\log p})$,

$$\frac{\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)}{\beta^\intercal\Sigma\beta} \lesssim \sqrt{\frac{k\log p/n}{\beta^\intercal\Sigma\beta}\cdot\left(1+\sqrt{\frac{\log p}{n+N}}\right)}.$$

Hence by the decomposition (97), we show that on the event $G_1 \cap G_4 \cap G_6(\beta,\beta,\sqrt{\log p})$,

$$|\Delta_2| \lesssim \frac{k\log p}{n} + \sqrt{\frac{\log p}{n}} + \sqrt{\frac{k\log p/n}{\beta^\intercal\Sigma\beta}\cdot\left(1+\sqrt{\frac{\log p}{n+N}}\right)}.$$

Together with the condition $\|\beta\|_2 \gg k\log p/\sqrt{n}$ and Condition (B2), we establish (47).

In the following, we present the proof of (48). Define $\bar{\phi}_2 = \frac{1}{(n+N)}\sum_{i=1}^{n+N}\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\widehat{\Sigma}^S\beta\right)^2$. Then

$$\frac{\widehat{\phi}_2-\phi_2}{4\phi_1+\rho\phi_2} = \frac{\widehat{\phi}_2-\bar{\phi}_2}{4\phi_1+\rho\phi_2} + \frac{\bar{\phi}_2-\phi_2}{4\phi_1+\rho\phi_2} \tag{99}$$

where

$$\widehat{\phi}_2-\bar{\phi}_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}\right)^2 - \left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\widehat{\Sigma}^S\beta\right)^2\right)$$

$$\bar{\phi}_2-\phi_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\widehat{\Sigma}^S\beta\right)^2 - \mathbb{E}\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\Sigma\beta\right)^2\right)$$

In the following, we will show

$$\mathbb{P}\left(\frac{1}{\phi_2}\left|\bar{\phi}_2-\phi_2\right| \geq C\frac{(\log(n+N))^{5/2}}{\sqrt{(n+N)}}\frac{(\beta^\intercal\Sigma\beta)^2}{\phi_2}\right) \lesssim (n+N)^{-c}, \tag{100}$$

$$\mathbb{P}\left(\frac{1}{4\phi_1+\rho\phi_2}\left|\widehat{\phi}_2-\bar{\phi}_2\right| \geq \sqrt{1+C\frac{(\log(n+N))^{5/2}}{\sqrt{(n+N)}}\frac{(\beta^\intercal\Sigma\beta)^2}{\phi_2}}\sqrt{\frac{\Lambda(n)}{4\phi_1+\rho\phi_2}}+\frac{\Lambda(n)}{4\phi_1+\rho\phi_2}\right) \lesssim (n+N)^{-c}+p^{-c}+\gamma(n), \tag{101}$$

where

$$\Lambda(n) = \frac{(k\log p)^2}{n+N} + \log(n+N)\frac{k\log p}{n}\left(\|\beta\|_2^2 + \frac{k\log p}{n}\right) \tag{102}$$

Since $4\phi_1+\rho\phi_2 \geq c(\|\beta\|_2^2+\rho\|\beta\|_2^4)$, under the regime $k \ll \sqrt{n}/\log p$, $\|\beta\|_2 \gg k\log p/\sqrt{n}$ and $\log(N+n)k\log p \ll n$, then (101) implies that $\frac{1}{4\phi_1+\rho\phi_2}\left|\widehat{\phi}_2-\bar{\phi}_2\right| \xrightarrow{p} 0$. Together with (100), we establish (48). The result (49) follows from (47) and (48) and the following decomposition,

$$\left|\frac{4\widehat{\phi}_1+\widehat{\rho}\widehat{\phi}_2}{4\phi_1+\widehat{\rho}\phi_2}-1\right| = \left|\frac{4(\widehat{\phi}_1-\phi_1)}{4\phi_1+\widehat{\rho}\phi_2}+\frac{\widehat{\rho}\left(\widehat{\phi}_2-\phi_2\right)}{4\phi_1+\widehat{\rho}\phi_2}\right| \leq \left|\frac{\widehat{\phi}_1-\phi_1}{\phi_1}\right| + \frac{\left|\widehat{\phi}_2-\phi_2\right|}{\max\{\phi_1,\phi_2\}}.$$

Proof of Equation (100). Define $A_i = X_{i,\cdot}^\intercal \beta / \sqrt{\beta^\intercal \Sigma \beta}$. Then we simply the expression of ϕ_2 and $\widehat{\phi}_2$ as

$$\frac{\phi_2}{(\beta^\intercal \Sigma \beta)^2} = \mathbb{E}\left(A_i^2 - \mathbb{E}A_i^2\right)^2 \quad \text{and} \quad \frac{\bar{\phi}_2}{(\beta^\intercal \Sigma \beta)^2} = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^2 - \frac{1}{n+N}\sum_{i=1}^{n+N} A_i^2\right)^2.$$

Define $\psi_2 = \frac{\phi_2}{(\beta^\intercal \Sigma \beta)^2}$ and $\bar{\psi}_2 = \frac{\bar{\phi}_2}{(\beta^\intercal \Sigma \beta)^2}$ and it is sufficient to show that $\left|\bar{\psi}_2 - \psi_2\right| \xrightarrow{p} 0$, which can be proved by applying Lemma 1 and 2 in Cai and Liu [2011]. To be self-contained, let's first re-state the Lemma 1 in Cai and Liu [2011] as Lemma 13.

LEMMA 13. *Let $\xi_1, \cdots, \xi_n$ be independent random variables with mean 0. Suppose that there exists some $\eta > 0$ and M_n such that $\sum_{i=1}^n \mathbb{E}\xi_i^2 \exp\left(\eta|\xi_i|\right) \leq M_n^2$. Then for $0 < t \leq M_n$,*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i \geq C_\eta M_n t\right) \leq \exp(-t^2), \tag{103}$$

where $C_\eta = \eta + \eta^{-1}$.

We bound $\bar{\psi}_2 - \psi_2$ based on the following decomposition,

$$\bar{\psi}_2 - \psi_2 = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^4 - \mathbb{E}A_i^4\right) + 2\mathbb{E}A_i^2 \cdot \frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^2 - \mathbb{E}A_i^2\right) - \left(\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^2 - \mathbb{E}A_i^2\right)\right)^2 \tag{104}$$

Since $\mathbb{E}A_i^2 = 1$, it is sufficient to establish upper bounds for $\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^2 - \mathbb{E}A_i^2\right)$ and $\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^4 - \mathbb{E}A_i^4\right)$. It follows from Lemma 10 that A_i^2 a sub-exponential random variable. By Remark 5.18 in Vershynin [2012], $A_i^2 - \mathbb{E}A_i^2$ is a sub-exponential random variable with sub-exponential norm smaller than $2M_1\|X_{i\cdot}\|_{\psi_2}^2$. By Corollary 5.17 in Vershynin [2012], we have

$$\mathbb{P}\left(\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^2 - \mathbb{E}A_i^2\right) \geq 2M_1\|X_{i\cdot}\|_{\psi_2}^2\sqrt{\frac{\log(n+N)}{n+N}}\right) \leq 2\exp\left(-c\log(n+N)\right) = 2(n+N)^{-c}. \tag{105}$$

Since A_i is a sub-gaussian random variable, there exists positive constants $C_1 > 0$ and $c > 2$ such that the following concentration inequality holds,

$$\sum_{i=1}^{n+N}\mathbb{P}\left(|A_i| \geq C_1\sqrt{\log(n+N)}\right) \leq (n+N)\max_{1\leq i\leq(n+N)}\mathbb{P}\left(|A_i| \geq C_1\sqrt{\log(n+N)}\right) \lesssim (n+N)^{-c} \tag{106}$$

Define $\bar{A}_i = A_i\mathbf{1}\left(|A_i| \leq C_1\sqrt{\log(n+N)}\right)$ and $\tilde{A}_i = A_i\mathbf{1}\left(|A_i| \geq C_1\sqrt{\log(n+N)}\right)$. Then we have

$$\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^4 - \mathbb{E}A_i^4\right) = \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right) + \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\tilde{A}_i^4 - \mathbb{E}\tilde{A}_i^4\right) \tag{107}$$

We control $\mathbb{E}\tilde{A}_i^4$ as follows,

$$\mathbb{E}\tilde{A}_i^4 \leq \sqrt{\mathbb{E}\left(A_i^8\right)\mathbb{P}\left(|A_i| \geq C_1\sqrt{\log(n+N)}\right)} \lesssim \mathbb{P}\left(|A_i| \geq C_1\sqrt{\log(n+N)}\right)^{1/2} \lesssim (n+N)^{-c/2}, \tag{108}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the fact that A_i is a sub-gaussian random variable and the last inequality follows from (106). Now we apply Lemma 13 to bound $\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right)$. By taking $\eta = c_1/\left(C_1\log(n+N)\right)^2$ for some small positive constant $c_1 > 0$, we have

$$\sum_{i=1}^{n+N}\mathbb{E}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right)^2\exp\left(\eta\left|\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right|\right) \leq C\sum_{i=1}^{n+N}\mathbb{E}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right)^2 \leq C_2(n+N).$$

By applying Lemma 13 with $M_n = \sqrt{C_2(n+N)}$, $\eta = c_1/\left(C_1\log(n+N)\right)^2$ and $t = \sqrt{\log(n+N)}$, then we have

$$\mathbb{P}\left(\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right) \geq C\frac{\left(\log(n+N)\right)^{5/2}}{\sqrt{n+N}}\right) \lesssim (n+N)^{-c}. \tag{109}$$

By (106), (107), (108) and (109), we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n+N}\sum_{i=1}^{n+N}\left(A_i^4 - \mathbb{E}A_i^4\right) \geq C\frac{\left(\log(n+N)\right)^{5/2}}{\sqrt{n+N}}\right) &\leq \sum_{i=1}^{n+N}\mathbb{P}\left(|A_i| \geq C\sqrt{\log(n+N)}\right)\\ + \mathbb{P}\left(\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\bar{A}_i^4 - \mathbb{E}\bar{A}_i^4\right) \geq C\frac{\left(\log(n+N)\right)^{5/2}}{\sqrt{n+N}}\right) &+ \mathbb{P}\left(\mathbb{E}\tilde{A}_i^4 \geq C\frac{\left(\log(n+N)\right)^{5/2}}{\sqrt{n+N}}\right) \lesssim (n+N)^{-c}.\end{aligned} \tag{110}$$

By (105) and (110), then there exisits a large constant C such that

$$\mathbb{P}\left(\left|\bar{\psi}_2 - \psi_2\right| \geq C\frac{\left(\log(n+N)\right)^{5/2}}{\sqrt{(n+N)}}\right) \lesssim (n+N)^{-c}, \tag{111}$$

for some $c > 0$. By the fact that $\left|\bar{\psi}_2 - \psi_2\right| = \frac{1}{(\beta^\intercal\Sigma\beta)^2}\left|\bar{\phi}_2 - \phi_2\right|$, this implies (100).

Proof of Equation (101). We start with the following decomposition,

$$\begin{aligned}&\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}\right)^2 - \left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\widehat{\Sigma}^S\beta\right)^2\right)\\ &= \frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} - \beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta + \beta^\intercal\widehat{\Sigma}^S\beta\right)^2\\ &+ \frac{1}{n+N}\sum_{i=1}^{n+N}2\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta - \beta^\intercal\widehat{\Sigma}^S\beta\right)\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta} - \widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta} - \beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta + \beta^\intercal\widehat{\Sigma}^S\beta\right)\end{aligned} \tag{112}$$

where the second term on the right hand side of (112) is further upper bounded by

$$\frac{2}{n+N}\sqrt{\sum_{i=1}^{n+N}\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta-\beta^\intercal\widehat{\Sigma}^S\beta\right)^2}\sqrt{\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta}-\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}-\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta+\beta^\intercal\widehat{\Sigma}^S\beta\right)^2} \tag{113}$$

Note that (111) implies

$$\mathbb{P}\left(\sqrt{\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta-\beta^\intercal\widehat{\Sigma}^S\beta\right)^2}\geq\sqrt{1+C\frac{(\log(n+N))^{5/2}}{\sqrt{(n+N)}}\frac{(\beta^\intercal\Sigma\beta)^2}{\phi_2}\sqrt{\phi_2}}\right)\lesssim(n+N)^{-c}, \tag{114}$$

Then it is sufficient to control $\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta}-\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}-\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta+\beta^\intercal\widehat{\Sigma}^S\beta\right)^2$, which is further decomposed as,

$$\begin{aligned}
&\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^\intercal X_{i\cdot}X_{i\cdot}^\intercal\widehat{\beta}-\widehat{\beta}^\intercal\widehat{\Sigma}^S\widehat{\beta}-\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal\beta+\beta^\intercal\widehat{\Sigma}^S\beta\right)^2\\
=&\frac{1}{n+N}\sum_{i=1}^{n+N}\left((\widehat{\beta}-\beta)^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)+2\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)-(\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)-2\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\right)^2\\
\leq&\frac{4}{n+N}\sum_{i=1}^{n+N}\left((\widehat{\beta}-\beta)^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2+4\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2+2\left((\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)-2\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\right)^2
\end{aligned} \tag{115}$$

Recall the definition of events in (82). On the event $G_1\cap G_4$, $(\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\lesssim k\log p/n$; On the event $G_1\cap G_4\cap G_6(\beta,\beta,\sqrt{\log p})$,

$$\left|\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\right|\leq\sqrt{\beta^\intercal\widehat{\Sigma}^S\beta}\sqrt{(\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)}\lesssim(1+\sqrt{\frac{\log p}{n+N}})\|\beta\|_2\sqrt{k\log p/n}.$$

Hence,

$$2\left((\widehat{\beta}-\beta)^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)-2\beta^\intercal\widehat{\Sigma}^S(\widehat{\beta}-\beta)\right)^2\lesssim\left(\frac{k\log p}{n}\right)^2+\|\beta\|_2^2\frac{k\log p}{n}. \tag{116}$$

It remains to control $\frac{4}{n+N}\sum_{i=1}^{n+N}\left((\widehat{\beta}-\beta)^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2+4\left(\beta^\intercal X_{i\cdot}X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2$ in the expression (115), which relies on the following fact. On the event G_1, $\sum_{i=1}^n\frac{\left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^2}{Ck\log p}\leq 1$ and hence

$$\frac{1}{n+N}\sum_{i=1}^n\left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^4=\frac{C^2(k\log p)^2}{n+N}\times\sum_{i=1}^n\frac{\left(X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right)^4}{C^2(k\log p)^2}\lesssim\frac{(k\log p)^2}{n+N}. \tag{117}$$

Define the event $\mathcal{B}_1$ as $\mathcal{B}_1=\left\{\max_{n+1\leq i\leq n+N}\left|X_{i\cdot}^\intercal(\widehat{\beta}-\beta)\right|\geq C\sqrt{\log(n+N)}\|\widehat{\beta}-\beta\|_2\right\}$ and the event $\mathcal{B}_2$ as $\mathcal{B}_2=\left\{\max_{1\leq i\leq n+N}|X_{i\cdot}^\intercal\beta|\geq C\sqrt{\log(n+N)}\|\beta\|_2\right\}$. Since $X_{i\cdot}$ is

sub-gaussian random variable and $\widehat{\beta} - \beta$ is independent of $X_{i\cdot}$ for $n+1 \le i \le n+N$, then

$$\max_{i=1,2} \mathbb{P}\left(\mathcal{B}_i\right) \lesssim (n+N)^{-c}. \tag{118}$$

On the event $\mathcal{B}_1 \cap G_6(\widehat{\beta} - \beta, \widehat{\beta} - \beta, \sqrt{\log p})$,

$$\frac{1}{N}\sum_{i=n+1}^{n+N}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^4 \le \frac{1}{N}\sum_{i=n+1}^{n+N}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \log(n+N)\|\widehat{\beta}-\beta\|_2^2 \lesssim \log(n+N)\left(k\log p/n\right)^2 \tag{119}$$

On the event $\mathcal{B}_2 \cap G_1 \cap G_4$, we have

$$\frac{4}{n+N}\sum_{i=1}^{n+N} 4\left(\beta^{\intercal}X_{i\cdot}X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \lesssim \log(n+N)\|\beta\|_2^2 \frac{4}{n+N}\sum_{i=1}^{n+N}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \lesssim \log(n+N)\|\beta\|_2^2\frac{k\log p}{n}$$

Combined with (116), (117) and (119), we show that on the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap G_6(\widehat{\beta} - \beta, \widehat{\beta} - \beta, \sqrt{\log p}) \cap G_1 \cap G_4$,

$$\left|\frac{1}{n+N}\sum_{i=1}^{n+N}\left(\widehat{\beta}^{\intercal}X_{i\cdot}X_{i\cdot}^{\intercal}\widehat{\beta} - \widehat{\beta}^{\intercal}\widehat{\Sigma}^S\widehat{\beta} - \beta^{\intercal}X_{i\cdot}X_{i\cdot}^{\intercal}\beta + \beta^{\intercal}\widehat{\Sigma}^S\beta\right)^2\right| \le \Lambda(n), \tag{120}$$

where $\Lambda(n) = \frac{(k\log p)^2}{n+N} + \log(n+N)\frac{k\log p}{n}\left(\|\beta\|_2^2 + \frac{k\log p}{n}\right)$. Together with (112), (114) and (113), we establish (101).

D. Additional Simulation Results

D.1. Inference for $\beta^{\intercal}\Sigma\beta$

This section presents the additional inference results corresponding to Section 6.2. In addition to the significant improvement in terms of estimation, the CHIVE estimator serves as the center of confidence intervals for $\beta^{\intercal}\Sigma\beta$. The coverage and precision properties of the constructed confidence interval CI are reported in Table 4. With a larger sample size, the empirical coverage of the proposed confidence interval achieves 95% and the average lengths of the confidence intervals get shorter. The integration of the unlabelled data in the semi-supervised setting has shorten the lengths of confidence interval significantly.

D.2. Effect of Pooling over Unlabelled Data: Signal Detection

We generate the high-dimensional linear regression (1) with the dimension $p = 400$ and the labelled data with sample size $n = 100$ and unlabelled data with sample size $N = 3,000$. For the linear model (1), the covariates $\{X_{i\cdot}\}_{1\le i\le n}$ for the labelled data and also $\{X_{i\cdot}\}_{n+1\le i\le n+N}$ for the unlabelled data are generated in i.i.d. fashion to follow multivariate normal distribution with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ where $\Sigma_{ij} = 0.8$ for $1 \le i \ne j \le p$ and $\Sigma_{ii} = 1$ for $1 \le i \le p$. The errors $\{\epsilon_i\}_{1\le i\le n}$ are generated as i.i.d normal distribution with mean zero and standard deviation 0.2. For the detection problem, we generate β as $\beta_j = \delta$ for $1 \le j \le 40$ and $\beta_j = 0$ for $j \ge 41$

		Supervised		Semi-Supervised	
	n	Cov	Len	Cov	Len
Setting 1	200	0.922	3.750	0.896	1.796
	400	0.936	2.734	0.942	1.536
	600	0.946	2.293	0.950	1.393
	800	0.936	1.991	0.942	1.290
	1,000	0.966	1.800	0.960	1.215
Setting 2	200	0.906	18.218	0.510	6.217
	400	0.940	13.444	0.880	5.930
	600	0.946	11.045	0.920	5.643
	800	0.932	9.671	0.924	5.419
	1,000	0.966	8.757	0.956	5.247
Setting 3	200	0.864	1.342	0.866	0.903
	400	0.914	0.982	0.904	0.721
	600	0.934	0.828	0.906	0.624
	800	0.944	0.723	0.924	0.561
	1,000	0.942	0.650	0.950	0.516

Table 4: Coverage and precision properties of Proposed CIs. Different rows correspond to different settings (Setting 1,2,3) and different sample sizes ($n = 200, 400, 600, 800, 1000$) for the given setting. Each row reports empirical coverage (indexed with "Cov") and average lengths (indexed with "Len") of proposed CIs. The columns indexed with "Supervised" represent the results for the supervised setting and the columns indexed with "Semi-Supervised" represent the results for the semi-supervised setting. For example, in the first row of numbers $(0.922, 3.750, 0.896, 1.796)$, it corresponds to the setting 1 and sample size $n = 200$, in the supervised setting, CI has empirical coverage 0.922 and the average length is 3.750; in the semi-supervised setting, CI has empirical coverage 0.896 and the average length is 1.796.

and vary δ across $\frac{1}{100}\{0, 1.025, 1.075, 1.125, 1.175, 1.225, 1.275, 1.325, 1.375, 1.425\}$. We use the randomization level $\tau_0 = 2$ in conducting the hypothesis testing $H_0 : \beta = 0$.

The simulations are replicated over 500 simulations and the Empirical Rejection Rate (ERR) and the coverage and length of confidence intervals are present in Table 5,where the column under "Semi-S" corresponds to the semi-supervised method and the column under "S" corresponds to the supervised method. Regarding ERR, we observe that the incorporation of unlabelled data is of help in improving the detection rate though the improvement, reported under the column "Imp", is only at the level of 5%. This small improvement for β closed to 0 is actually predicted by the theoretical results. Since the design matrix is generated to follow multivariate Gaussian distribution in the simulation studies, the rate of convergence related to the unlabelled data is expressed as $\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2/(N+n) = 2(\beta^\intercal \Sigma \beta)^2/(N+n)$; For the hypothesis testing problem, the interesting but challenging regime is the local alternative near $\beta = 0$, that is, $\mathbb{E}\left(\beta^\intercal X_{1\cdot} X_{1\cdot}^\intercal \beta - \beta^\intercal \Sigma \beta\right)^2 = 2(\beta^\intercal \Sigma \beta)^2$ is near zero. This explains why improvement of integrating unlabelled data in testing $H_0 : \beta = 0$ is not as significant as the prediction loss evaluation and also inference for the explained variance $\beta^\intercal \Sigma \beta$ as presented in the main paper.

	ERR			Coverage		Length		
100δ	Semi-S	S	Imp	Semi-S	S	Semi-S	S	Ratio
0.000	0.018	0.018	0%	0.972	0.972	0.172	0.172	100%
1.025	0.662	0.628	5.4%	0.974	0.974	0.148	0.153	97.0%
1.075	0.702	0.652	7.7%	0.970	0.976	0.147	0.152	96.3%
1.125	0.760	0.734	3.5%	0.968	0.966	0.146	0.153	95.5%
1.175	0.834	0.804	3.7%	0.948	0.956	0.145	0.153	94.7%
1.225	0.882	0.838	5.3%	0.952	0.956	0.146	0.155	94.0%
1.275	0.946	0.904	4.6%	0.970	0.964	0.144	0.156	92.6%
1.325	0.982	0.946	3.8%	0.972	0.956	0.143	0.156	91.1%
1.375	0.982	0.960	2.3%	0.964	0.962	0.142	0.158	90.1%
1.425	0.984	0.968	1.6%	0.964	0.966	0.143	0.161	88.6%

Table 5: Signal detection $p = 400$, sample size $n = 100$ and $N = 3000$.

The constructed confidence intervals have coverage in both the supervised and semi-supervised settings and the confidence interval constructed using the unlabelled data has a shorter length. Under the column "Ratio", we report the ratio of the length of confidence intervals in the semi-supervised setting to that in the supervised setting and see the confidence intervals can be reduced by as much as12% in length. With the same reasoning as the ERR, the length of the confidence interval is not significantly reduced as the simulated setting mainly focuses on the case where β is close to zero.

D.3. More about Signal Detection

In this section, we investigate more on the signal detection problem. We switch the focus from how to integrate the unlabelled data in the detection problem to investigating the numerical effect of the randomization levels.

For the detection problem, we generate β as $\beta_j = \delta$ for $1 \leq j \leq 50$ and $\beta_j = 0$ for $51 \leq j \leq 800$ and vary δ across $\{0.00, 0.025, 0, 05, 0.075, 010, 0.125, 0.15\}$ and vary the sample size n across $\{600, 1200\}$. In Figure 2, we demonstrate the coverage and precision properties of the randomized confidence intervals across four methods, the non-randomized detector $D(0)$ and the three randomized detectors $D(2), D(4)$ and $D(6)$, where $D(\cdot)$ is defined in (**??**). The two plots on the top of Figure 2, corresponding to the supervised setting with $n = 600$ demonstrate the effect of randomization on the empirical coverage and average lengths, where the randomization leads to a interval estimator achieving the coverage properties at the expense of wider interval estimators. With the randomization level τ_0 reaching 2, the coverage property is guaranteed while the empirical coverage for the procedure without randomization ($\tau_0 = 0$) is much lower than 0.95, especially for weak signals with a small δ. The bottom two plots of Figure 2 corresponds to the supervised setting with $n = 1,200$ and the main observation is similar to the case of $n = 600$ but the confidence intervals are much shorter than the setting with $n = 600$.

The empirical detection rate is reported in Table 6, where the sample size n is generated across $n = 600$ and $n = 1,200$ and the explained variance $\beta^\intercal \Sigma \beta$ is controlled via

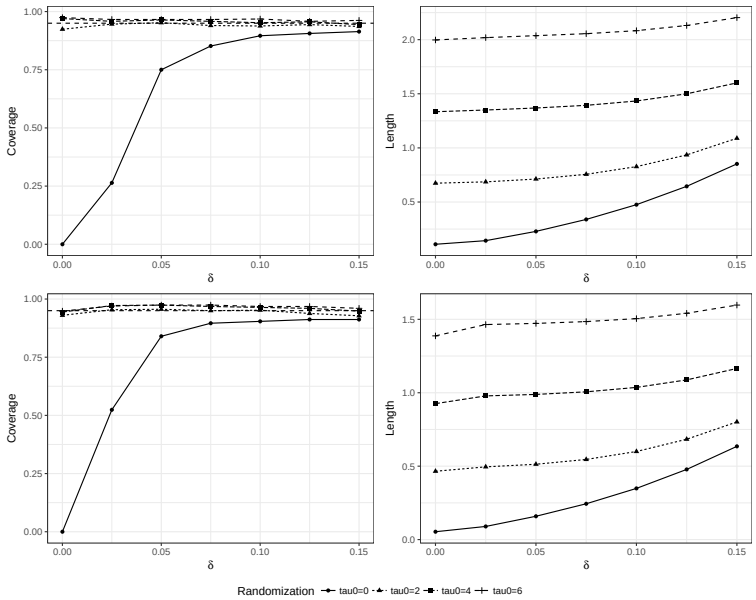

Fig. 2: Empirical coverage and average lengths of the proposed randomized confidence intervals in the supervised setting. The above two figures correspond to the sample size $n = 600$ and the bottom two figures correspond to $n = 1200$. The left hand side figures stand for the empirical coverage for different δ while the right hand side figures stand for the average lengths of CIs for different δ. Different type of the curves correspond to different randomization levels $\tau_0 \in \{0, 2, 4, 6\}$. The dashed horizontal lines on the left hand figures correspond to the targeted coverage level, 0.95.

the scaler δ. When $\delta = 0$, it corresponds to the null case and a proper detection procedure is expected to have type I error rate 0.05. As predicted by theory, the detection method without randomization $D(0)$ fails to give proper type I error due to presence of weak signals. With introducing the randomization procedure, the type I error rate gets closer to 0.05. When δ moves away from zero, the detection procedure is taken as a powerful procedure as the empirical detection rate approaches 1. For the detection procedure with randomization level $\tau_0 = 2$, the setting with $\delta = 0.025$ corresponds to an indistinguishable region, where it is challenging to detect the signal. However, as δ reaches 0.05, the detection rate reaches 0.800 for $n = 600$ and 0.944 for $n = 1200$. As characterized by theory, a larger randomization level requires a higher value of δ such that the signal can be detected, for example, for $\tau_0 = 4$, until δ reaches 0.075, the detection rate reaches 0.82 for $n = 600$ and 0.968 for $n = 1200$. The corresponding semi-supervised setting shows a similar phenomenon to the supervised setting but tends to be easier than the supervised setting due to the unlabelled data. The results are reported in the supplementary materials.

		$n = 600$				$n = 1,200$			
δ	$\beta^\intercal\Sigma\beta$	$D(0)$	$D(2)$	$D(4)$	$D(6)$	$D(0)$	$D(2)$	$D(4)$	$D(6)$
0.000	0.000	1.000	0.148	0.082	0.066	1.000	0.124	0.076	0.068
0.025	0.091	1.000	0.248	0.094	0.062	1.000	0.254	0.124	0.086
0.050	0.365	1.000	0.800	0.356	0.182	1.000	0.944	0.472	0.264
0.075	0.821	1.000	1.000	0.820	0.524	1.000	1.000	0.968	0.764
0.100	1.460	1.000	1.000	1.000	0.914	1.000	1.000	1.000	0.992
0.125	2.281	1.000	1.000	1.000	0.996	1.000	1.000	1.000	1.000
0.150	3.285	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6: Empirical detection rates in the supervised setting. The column indexed with δ represents the signal strength, where the signal is of sparsity 50 and of the form $\delta\cdot(1, 1, \cdots 1, 0, 0, \cdots, 0)$; the column indexed with $\beta^\intercal\Sigma\beta$ represents the value of $\beta^\intercal\Sigma\beta$; the columns under "n=600" and "n=1,200" correspond to sample size 600 and 1,200 respectively, where the column indexed with $D(\tau_0)$ report the empirical detection rates for the detector $D(\tau_0)$.

D.4. Prediction Loss Evaluation

In this subsection, we present additional results about prediction loss evaluation. We generate the high-dimensional regression vector β with sparsity 10 where $\beta_j = j/5$ for $1 \le j \le 10$ and $\beta_j = 0$ for $j \ge 11$. Let $\widehat{\beta}(\lambda)$ denote the Lasso estimator based on an independent training data $\left(X^{(0)}, y^{(0)}\right)$ with sample size $n_0 = 300$,

$$\widehat{\beta}\,(\lambda) = \arg\min_{\beta\in\mathbb{R}^p} \frac{\|y^{(0)} - X^{(0)}\beta\|_2^2}{2n_0} + \lambda\sum_{j=1}^{p} \frac{\|X_{\cdot j}^{(0)}\|_2}{\sqrt{n_0}}|\beta_j|.$$

We consider the inference problem for the out-of-sample prediction accuracy $(\widehat{\beta}(\lambda) - \beta)^\intercal\Sigma(\widehat{\beta}(\lambda) - \beta)$. Specifically, we consider three estimators $\widehat{\beta}(\lambda_0), \widehat{\beta}(5\lambda_0)$ and $\widehat{\beta}(10\lambda_0)$ with $\lambda_0 = \sqrt{\frac{Z_{(1-(0.1/p))}}{n_0}}$ and report the numerical performance of both point and interval estimators of the corresponding prediction accuracy. We consider the prediction accuracy problem across three different sample sizes, $\{600, 1200, 2400\}$ and introduce different randomization levels. We will use $\mathrm{PA}(\tau_0)$ to denote the procedure with randomization level τ_0.

Table 7 has reported the point and interval estimators of the prediction accuracy across different settings. In terms of point estimation, the sample averages get closer to the true accuracy with increasing sample sizes. Among the three estimators, the true prediction accuracy of $\widehat{\beta}(\lambda_0)$ is the smallest and also the most difficult to assess. The fundamental reason is that the small accuracy/error is hard to quantify. This phenomenon is connected to the theoretical results established in Cai and Guo [2017a], which showed that the estimation accuracy $\|\widehat{\beta} - \beta\|_2^2$ is hard to quantify for an accurate estimator $\widehat{\beta}$. The constructed confidence interval $\mathrm{PA}(0)$ without randomization has no coverage even for $n = 2400$. In such a scenario with weak signals, the evaluators involved with randomized calibration, $\mathrm{PA}(2)$ and $\mathrm{PA}(4)$ produce valid confidence intervals across different settings.

Not only the point and interval estimators are useful, the upper limit and lower limit

of the confidence intervals reported in Table 7 can also be informative in the prediction accuracy evaluation. For the estimator $\widehat{\beta}(\lambda_0)$, although the average of lower limit of confidence intervals for $(\widehat{\beta}(\lambda_0)-\beta)^\intercal\Sigma(\widehat{\beta}(\lambda_0)-\beta)$ is zero, the corresponding upper limits of confidence intervals are informative as they provide empirical guidance to practitioners with upper bounds for the prediction accuracy. For $\widehat{\beta}(5\lambda_0)$ and $\widehat{\beta}(10\lambda_0)$, both the upper and lower limits of the confidence intervals are informative on the size of the prediction accuracy.

		$\widehat{\beta}(\lambda_0)$			$\widehat{\beta}(5\lambda_0)$			$\widehat{\beta}(10\lambda_0)$		
	True Accuracy	0.065			0.636			2.310		
		PA(0)	PA(2)	PA(4)	PA(0)	PA(2)	PA(4)	PA(0)	PA(2)	PA(4)
Super, 600	Coverage	0.166	0.938	0.962	0.778	0.928	0.956	0.914	0.934	0.948
	Est Aver	0.157	0.158	0.160	0.739	0.743	0.746	2.417	2.421	2.424
	Lower Aver	0.090	0.000	0.000	0.584	0.373	0.058	2.062	1.932	1.665
	Upper Aver	0.223	0.502	0.837	0.895	1.112	1.434	2.772	2.909	3.183
Semi, 600	Coverage	0.180	0.938	0.962	0.800	0.936	0.962	0.930	0.944	0.958
	Est Aver	0.154	0.155	0.157	0.726	0.729	0.732	2.385	2.388	2.391
	Lower Aver	0.088	0.000	0.000	0.581	0.364	0.047	2.102	1.950	1.664
	Upper Aver	0.220	0.499	0.834	0.871	1.094	1.418	2.667	2.827	3.119
Super, 1200	Coverage	0.480	0.968	0.976	0.890	0.960	0.974	0.964	0.964	0.970
	Est Aver	0.107	0.108	0.109	0.684	0.686	0.688	2.356	2.358	2.360
	Lower Aver	0.067	0.000	0.000	0.575	0.420	0.190	2.099	2.004	1.810
	Upper Aver	0.146	0.355	0.599	0.793	0.952	1.185	2.613	2.712	2.909
Semi, 1200	Coverage	0.494	0.970	0.976	0.898	0.958	0.972	0.954	0.954	0.968
	Est Aver	0.106	0.107	0.108	0.680	0.682	0.684	2.348	2.350	2.352
	Lower Aver	0.066	0.000	0.000	0.576	0.418	0.187	2.133	2.026	1.821
	Upper Aver	0.145	0.354	0.598	0.783	0.946	1.180	2.563	2.675	2.883
Super, 2400	Coverage	0.738	0.972	0.978	0.916	0.954	0.972	0.948	0.944	0.958
	Est Aver	0.083	0.084	0.084	0.663	0.663	0.662	2.340	2.340	2.340
	Lower Aver	0.058	0.000	0.000	0.585	0.472	0.306	2.154	2.085	1.945
	Upper Aver	0.109	0.260	0.434	0.741	0.853	1.019	2.526	2.594	2.734
Semi, 2400	Coverage	0.742	0.972	0.978	0.912	0.962	0.976	0.950	0.950	0.960
	Est Aver	0.083	0.083	0.083	0.661	0.661	0.661	2.337	2.337	2.336
	Lower Aver	0.058	0.000	0.000	0.587	0.472	0.306	2.173	2.098	1.952
	Upper Aver	0.108	0.260	0.434	0.736	0.851	1.017	2.501	2.576	2.721

Table 7: Inference for prediction accuracy $(\widehat{\beta}(\lambda)-\beta)^\intercal\Sigma(\widehat{\beta}(\lambda)-\beta)$. The table reports six settings, corresponding to three different sample sizes (600,1200, 2400) and the supervised and semi-supervised setting. For example, "Super, 600" stands for the supervised setting with sample size $n = 600$ and "Semi, 600" stands for the semi-supervised setting with sample size $n = 600$. The true prediction accuracy of the three estimators $\widehat{\beta}(\lambda_0), \widehat{\beta}(5\lambda_0)$ and $\widehat{\beta}(\lambda_0)$ is reported as $0.065, 0.636$ and 2.310. Three prediction accuracy evaluators PA(0), PA(2) and PA(4) are reported, where PA(0) is the evaluator with no randomization, PA(2) is the evaluator with randomization level $\tau_0 = 2$ and PA(4) is the evaluator with randomization level $\tau_0 = 4$. For each setting, the row indexed with "Coverage" reports the empirical coverage of the corresponding confidence intervals over 500 simulations; the row indexed with "Est Aver" reports the sample average of the corresponding point estimators over 500 simulations; the rows indexed with "Lower Aver" and 'Upper Aver" report the sample averages of the lower and upper limits of interval estimators over 500 simulations.

E. Additional Real Data Analysis

In the following, we compare the plug-in estimator and the CHIVE estimator to demonstrate that the CHIVE estimator adds back the missing heritability across all 46 traits. We demonstrate this phenomenon in Figure 3 by comparing the CHIVE estimator and the plug-in estimators of heritability for all 46 traits. We shall stress that all points lie above the line $y = x$ and this means that the calibration step adds back the missing heritability due to simply plugging in the Lasso estimator, where the Lasso estimator tends to ignore the genetic markers with small effects.

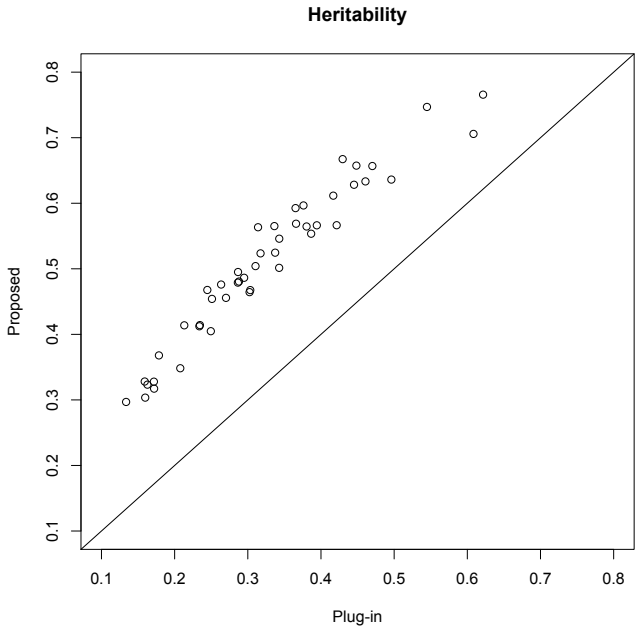

Fig. 3: Heritability for 46 traits. The x-axis represents the heritability estimated by the plug-in estimator and the y-axis represents the heritability by the proposed CHIVE estimator; the line represents $y = x$.