Group Inference in High Dimensions with Applications to Hierarchical Testing *

Zijian Guo, Claude Renaux, Peter Bühlmann and T. Tony Cai

September 3, 2019

Abstract

Group inference has been a long-standing question in statistics and the development of high-dimensional group inference is an essential part of statistical methods for analyzing complex data sets, including hierarchical testing, tests of interaction, detection of heterogeneous treatment effects and local heritability. Group inference in regression models can be measured with respect to a weighted quadratic functional of the regression sub-vector corresponding to the group. Asymptotically unbiased estimators of these weighted quadratic functionals are constructed and a procedure using these estimator for inference is proposed. We derive its asymptotic Gaussian distribution which allows to construct asymptotically valid confidence intervals and tests which perform well in terms of length or power. The results simultaneously address four challenges encountered in the literature: controlling coverage or type I error even when the variables inside the group are highly correlated, achieving a good power when there are many small coefficients inside the group, computational efficiency even for a large group, and no requirements on the group size. We apply the methodology to several interesting statistical problems and demonstrate its strength and usefulness on simulated and real data.

Key words: Hierarchical Testing; Interaction Test; Detection of Heterogeneous Treatment Effects; Local Heritability.

*The research of Z. Guo was supported in part by the NSF DMS 1811857; Z. Guo also acknowledges financial support for visiting the Institute of Mathematical Research (FIM) at ETH Zurich. The research of P. Bühlmann was supported in part by the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG); The research of T. Cai was supported in part by NSF Grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056.

1 Introduction

High-dimensional linear regression has found many applications in modern data analysis through extracting useful information from complex data. Statistical inference in the high-dimensional sparse linear regression is an important but also challenging problem. In the past few years, there has been a fast growing literature on this topic. The current paper addresses an important and long-standing statistical inference problem, namely inference or testing significance of groups of covariates. Specifically, we consider the following high-dimensional linear regression

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

where $X_{i\cdot} \in \mathbf{R}^p$ are i.i.d. p-dimensional random vectors with $\Sigma = \mathbf{E}X_{i\cdot}X_{i\cdot}^{\intercal}$ and ϵ_i are i.i.d. Gaussian errors, independent of $X_{i\cdot}$, with mean zero and variance σ^2. For a given set $G \subset \{1, 2, \cdots, p\}$, the group significance test is for the hypothesis

$$H_0 : \beta_G = 0, \tag{2}$$

where $\beta_G = \{\beta_j;\ j \in G\}$. More generally, for inference we will consider certain weighted functionals of β_G. In many applications, identification of group inference or significance is as important as that of individual significance, especially in the scenario when the covariates are likely to affect the outcome jointly or the covariates are highly correlated with each other. In the corresponding low-dimensional setting, the (partial) F-test is the classical procedure for testing this hypothesis. However, there is still a lack of methods for conducting group inference or testing (2) in high dimensions, especially when the group G has large size.

In the following, we shall provide a series of motivations for group inference or significance.

1. **Hierarchical Testing.** As written above, when the variables are highly correlated, it is often too ambitious to detect significant single variables and groups of correlated variables are considered instead. Hierarchical testing is a multiple testing procedure which uses p-values for group significance as input. It is a hybrid between sequential testing and Bonferroni-type correction to control the familywise error rate [23].

   A hierarchy is provided in terms of a tree which is typical the output of hierarchical clustering of the covariates: each node in the tree is a group of variables and at every level of the tree, the groups build a partition of $\{1, \ldots, p\}$. Hierarchical testing then proceeds by testing in a top-down manner according to the tree by using group significance testing multiple times: at the beginning, the upper part of the tree, the groups are large and they become smaller when moving downward the tree.

Hierarchical testing determines the resolution level in the tree in a fully data-driven way, depending on the signal strength and the correlation structure among the variables. For strong signals with large regression coefficients in absolute values and when the variables are not too highly correlated, the method will discover small groups or even single variables; and vice-versa when the signal is not so strong or the correlation among the variables is high. Thus, hierarchical testing addresses very elegantly the trade-off between signal strength and high correlation and it is perhaps among the most natural ways to deal with large-scale high-dimensional testing problems in real applications [6, 19]. More details are given in Section 4.1.

2. **Interaction Test and Detection of Effect Heterogeneity.** Group significance test can be used to test the existence of interaction or detection of heterogeneous treatment effects. With a slight modification of (1), we write the following model with interaction terms,

$$y_i = X_{i\cdot}^\intercal \beta + D_i \, (X_{i\cdot}^\intercal \gamma) + \epsilon_i, \quad \text{for } 1 \le i \le n. \tag{3}$$

Here, D_i is the treatment or exposure variable and can be a binary variable in many applications. The interaction test in the above model (3) is formulated as testing $H_0 : \gamma = 0$. If D_i is an indicator whether the i-th observation receives a treatment or not, then the test of $H_0 : \gamma = 0$ can be viewed as detecting the existence of heterogeneous treatment effect. Define $W_i = (D_i X_{i\cdot}^\intercal, X_{i\cdot}^\intercal)^\intercal$ and $\eta = (\gamma^\intercal, \beta^\intercal)^\intercal$ and then the model (3) can be expressed as $y_i = W_{i\cdot}^\intercal \eta + \epsilon_i$ and the test $H_0 : \gamma = 0$ can be formulated in the form of (2) as $H_0 : \eta_G = 0$ with $G = \{1, 2, \cdots, p\}$. For a detailed discussion, see Section 4.2 and [9, 30].

3. **Local Heritability in Genetics.** Heritability measures how much of the phenotype (response) variance can be explained by the genotypes (covariates). Local heritability is among the most important heritability measures [28] as it represents the proportion of variance explained by a subset of genotypes indexed by the group G. In applications, the group G can be naturally formulated, for example, the set of SNPs located in on the same chromosome. The group significance test is motivated from studying whether the joint effect of a group of genotypes with the index set G is significant. Additionally, it is also of great interest to perform statistical inference for local heritability in terms of confidence intervals, which is closely related to the group significance test problem. See more discussion in Section 4.3.

1.1 Problem Formulation

There exist different ways of conducting the testing problem (2), including the the F-test in low-dimensional settings or the maximum test where one considers the maximum of absolute values of single normalized components of an estimated regression vector. The latter is implicit in the work of [34] for the debiased Lasso. Here, we consider testing (2) through weighted quadratic functionals. For a given index set $G \subseteq \{1, 2, \cdots, p\}$, we introduce the following null hypothesis,

$$H_{0,A} : \ \beta_G^\intercal A \beta_G = 0, \tag{4}$$

for some positive definite matrix $A \in \mathbb{R}^{|G| \times |G|}$ with $|G|$ denoting the cardinality of G. So as long as A is positive definite, the null hypothesis (4) is equivalent to (2). We shall highlight the following most interesting cases of (4). The first one is to replace A with $\Sigma_{G,G}$, where $\Sigma = \mathbf{E} X_{i\cdot} X_{i\cdot}^\intercal$ is the second order moment matrix,

$$H_{0,\Sigma} : \ \beta_G^\intercal \Sigma_{G,G} \beta_G = 0. \tag{5}$$

The other one is to replace the weight matrix A with the identity matrix,

$$H_{0,\mathrm{I}} : \ \beta_G^\intercal \beta_G = 0. \tag{6}$$

The quantity $\beta_G^\intercal \Sigma_{G,G} \beta_G$ in (5) is naturally used for group significance testing as the quantity itself measures the variance explained by the set of variables $X_{i,G}$, $\beta_G^\intercal \Sigma_{G,G} \beta_G = \mathbf{E} |X_{i,G}^\intercal \beta_G|^2$.

We shall remark that the testing problem (5) can be conducted in both cases where the matrix in the middle $\Sigma_{G,G}$ is known or not. If $\Sigma_{G,G}$ is known, then it can be simply treated as a special case of (4). However, in a more practical setting where the positive definite matrix $\Sigma_{G,G}$ is unknown, we need to estimate $\Sigma_{G,G}$ in the construction of the test and also need to quantify the additional uncertainty of estimating this matrix from data.

Throughout the paper, when there is no confusion of the definition of the index set G, we omit the dependence on G in the definitions and introduce the following notations throughout the paper,

$$\mathrm{Q}_\Sigma = \mathbf{E} |X_{i,G}^\intercal \beta_G|^2 = \beta_G^\intercal \Sigma_{G,G} \beta_G \quad \text{and} \quad \mathrm{Q}_A = \beta_G^\intercal A \beta_G.$$

Though the focus of the current paper is mainly about testing group significance, the inference problem for Q_Σ and Q_A in terms of confidence regions is of independent interest, in particular for the applications in genetics about local heritability.

1.2 Results and Contribution

Recently, statistical inference in high dimensional models has been carefully studied in both statistics and econometrics [7,11,18,31,34], with a focus on confidence interval construction

and hypothesis testing related to single regression coefficients. Together with a careful use of bootstrap methods, certain maximum tests have been developed in [11,13,35] to conduct the hypothesis testing problem (2). Although the developed tests are successful in controlling the type I error under different conditions, it is known that the coordinate-based maximum test suffers from the following problems in both statistical inference and computational efficiency as well. Statistically speaking, when the variables inside the group G are highly correlated, there is no guarantee that the coordinate-based maximum test controls the type I error; additionally, if β_G is not zero but contains many small regression coefficients, the maximum test is expected to have a relatively low power. Both of these phenomena are observed in our simulation studies. Computationally speaking, if the group G is of large size, then the coordinate-based maximum test requires implementation of $|G| + 1$ number of high-dimensional penalized/constrained optimization problems, where each optimization problem involves the p-dimensional parameter. This is time-consuming, especially when the size of G is large or when conducting several group significance tests.

A major goal of the current paper is to do the hypothesis testing (2) via an estimation procedure for the quadratic form in (4) and (5). At the same time, we would like to address the statistical and computational challenges for the coordinate-based maximum test. The testing procedure proposed in the current paper is to make inference for $\beta_G^\intercal A\beta_G$ or $\beta_G^\intercal \Sigma_{G,G}\beta_G$, where both of them can be viewed as measures of the joint group effects. One of the advantages for testing the group effect jointly is that even if the variables inside the group G are highly correlated, a joint test based on $\beta_G^\intercal \Sigma_{G,G}\beta_G$ is more reliable as the individual effects inside G are nearly non-identified.

The main methodology in the current paper is to carefully calibrate certain reasonably good initial estimators for $\mathrm{Q}_A = \beta_G^\intercal A\beta_G$ or $\mathrm{Q}_\Sigma = \beta_G^\intercal \Sigma_{G,G}\beta_G$. We take the inference problem for Q_Σ as an example and briefly describe the main idea for our proposed procedure. Denote by $\widehat{\beta}$ and $\widehat{\Sigma}$ some reasonably good estimators of β and Σ, respectively. For example, $\widehat{\beta}$ can be taken as a penalized estimator β with a proper tuning parameter and $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal$ is the sample second order moment matrix. A natural estimator for Q_Σ is $\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G}\widehat{\beta}_G$ and it is known that such an initial estimator is not proper for conducting statistical inference due to the fact that the bias of the penalized estimator $\widehat{\beta}$ will be carried over to the plug-in estimator $\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G}\widehat{\beta}_G$ and results in a large error. We calibrate this plug-in estimator through constructing a projection direction for correcting the bias. Although the idea of using projection directions for bias correction has been developed for the inference for each regression coefficient [34], we need to develop a new way of constructing the projection direction for our specific purpose of group inference. This ensures that our proposed group significance test is statistically accurate and also computationally efficient.

We shall consider a most challenging case where the index set G has a very large size

and highlight how this new construction of projection direction will enable us to achieve both nice statistical and computational properties simultaneously. Intuitively, the new construction of projection direction is to correct the whole bias of $\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$ all at once, instead of correcting the bias of $\widehat{\beta}$ coordinate by coordinate. Such a direct correction has the computational advantage of only implementing an optimization problem with p-dimensional parameters twice, no matter how large the group size is. More fundamentally, a direction calibration of $\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$ carefully rebalances the bias and variance, in the sense that the bias is well controlled and the variance is dominating the bias.

To sum up, the proposed group significance test has the following three main advantages.

1. **High-correlation inside G.** The proposed testing procedure is effective even if there exists high-correlation for variables inside G. This is particularly useful for hierarchical testing and also the inference for local heritability, where the variables belonging to the same group tend to be highly correlated. See the numerical illustration in Section 5.2.

2. **Detection of dense but weak signals.** While the maximum test is effective in the presence of a spike regression coefficients, our proposed test is more effective if the regression vector has many non-zero entries but each is of small order. In the application of interaction testing and detection of heterogeneous treatment effect, the interaction term tends to be dense but of a small order of magnitude. In Section 5.1, we have observed that the proposed method is more powerful in such a setting.

3. **Computationally efficient for a large set G.** It is of interest to conduct significance testing for a large group of variables. As examples, we mention the starting levels for the hierarchical testing problem (see Sections 4.1) and also the interaction test (see Section 4.2). If a unit of computational cost is defined as implementing a penalized or constrained optimization problem involved with p variables, then the proposed testing procedure requires two units of computational costs while the coordinate-based maximum test needs $|G| + 1$ units. For the case of large $|G|$, the computational cost is significantly reduced with using our proposed test. At the same time, the statistical accuracy is guaranteed with this computationally efficient testing procedure.

1.3 Literature Comparison

There is a rich literature about testing group significance and related statistical inference problem. We have reviewed some of the existing work based on the maximum test and shall mention a few other related work to the group significance test. Inference for the quadratic

functionals is closely related to the group significance test. Statistical inference for $\beta^\intercal\Sigma\beta$ and $\|\beta\|_2^2$ has been carefully investigated in [8, 14, 33] and the inference methods developed in [8] have been extended to conduct the signal detection problem, which is a special case of (1) by setting $G = \{1, 2, \cdots, p\}$. We shall emphasize that the group significance test problem is much more challenging than the signal detection problem, mainly due to the fact that the group significance test requires a decoupling between variables inside G and other variables inside G^c. This decoupling step between G and G^c is essentially requiring a novel construction of the projection direction. The same comments can be made to differentiate the current paper with the signal detection problem considered in [1, 17].

Along another line, the works [25,32] extended the F-test or χ^2 test for fixed dimension to the high-dimensional setting and this extended test can also be used for group significance testing. However, such a test is expected to be less powerful than the proposed test in the case of a large group G. To see this, the standard deviation level of the F-test or χ^2 test is at the scale of $\sqrt{|G|/n}$; in contrast, the standard deviation level of our proposed test statistics is at the scale of $\sqrt{\|\beta_G\|_2^2/n}$. If $|G|$ is large, then the significance test proposed in the current paper is more powerful as long as $\|\beta_G\|_2^2$ is smaller than $|G|$, which is typically true if β satisfies a certain sparsity structure. Additionally, [25] requires that the group size is smaller than the sample size n while the current paper imposes no condition on the group size G. Another related work [23] considered the group significance test without any compatibility condition on the design.

Paper Organization. The remaining part of the paper is organized as follows. In Section 2, we introduce the methodology for group significance testing. In Section 3, we provide theoretical justification for the proposed method. In Section 4, we apply the general procedure to several interesting problems, including hierarchical testing, test of interaction and detection of the effect heterogeneity and inference for local heritability; In Section 5, we conduct a large set of numerical studies to support our claims and theory. In Section 6, we apply our method to two real data sets from genomics and genetics. Proofs and additional numerical results are presented in the supplementary material.

2 Methodology for Testing Group Significance

In this section, we propose statistical inference procedures for both $\beta_G^\intercal\Sigma_{G,G}\beta_G$ and $\beta_G^\intercal A\beta_G$.

2.1 Inference for Q_Σ

For estimating the quantity $\mathrm{Q}_\Sigma = \beta_G^\intercal\Sigma_{G,G}\beta_G$, we start with a plug-in estimator $\widehat{\beta}_G^\intercal\widehat{\Sigma}_{G,G}\widehat{\beta}_G$, where, throughout the paper, we use $\widehat{\beta}$ to denote a reasonably good estimator of β and use

the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}$ as the estimator of Σ. To simplify the discussion, we also assume the index set G is of the form $\{1, 2, \cdots, |G|\}$.

We decompose the error of the plug-in estimator $\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$,

$$\begin{aligned}\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G - \beta_G^{\intercal}\Sigma_{G,G}\beta_G = &-2\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}(\beta_G - \widehat{\beta}_G) + \beta_G^{\intercal}(\widehat{\Sigma}_{G,G} - \Sigma_{G,G})\beta_G \\ &- (\widehat{\beta}_G - \beta_G)^{\intercal}\widehat{\Sigma}_{G,G}(\widehat{\beta}_G - \beta_G)\end{aligned}$$

Based on this decomposition, we need to estimate $2\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}(\beta_G - \widehat{\beta}_G)$ as this is one of the dominant terms and further calibrate the plug-in estimator $\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$. The calibration can be done through identifying a projection direction $u \in \mathbb{R}^p$ for the following expression

$$u^{\intercal}\frac{1}{n}X^{\intercal}(y - X\widehat{\beta}) = \frac{1}{n}u^{\intercal}X^{\intercal}\epsilon + u^{\intercal}\widehat{\Sigma}(\beta - \widehat{\beta}).$$

The challenging part is how to identify such a projection direction $u \in \mathbf{R}^p$ such that $u^{\intercal}\widehat{\Sigma}(\beta - \widehat{\beta})$ is a good estimator of $\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}(\beta_G - \widehat{\beta}_G)$. The construction of the projection direction can be motivated from the decomposition

$$u^{\intercal}\widehat{\Sigma}(\beta - \widehat{\beta}) - \widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}(\beta_G - \widehat{\beta}_G) = \left[\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right]^{\intercal}(\beta - \widehat{\beta}).$$

for $u \in \mathbf{R}^p$. Note that $\left|\left[\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right]^{\intercal}(\beta - \widehat{\beta})\right| \leq \|\beta - \widehat{\beta}\|_1 \left\|\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right\|_{\infty}$. As long as $\widehat{\beta}$ is a reasonably good estimator with a small $\|\beta - \widehat{\beta}\|_1$, it remains to construct the projection direction $u \in \mathbb{R}^p$ such that $\left\|\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right\|_{\infty}$ is upper bounded by a small value. From a geometric perspective, this constraint is to ensure that the projection of the approximation vector $\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}$ to all Euclidean basis $\{e_j\}_{1\leq j\leq p}$ is small. However, this idea is only useful for the case that the group size $|G|$ is small. If the group size $|G|$ is quite large, then this intuition is not guaranteed to work due to the fact that $\left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}$ can be quite dense. We need to constrain the difference term $\widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}$ from one additional direction $\left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}$. That is, the projection direction u is constructed such that the projection $\left\langle w, \widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right\rangle$ is small for all $w \in \mathcal{C}$ where

$$\mathcal{C} = \left\{e_1, \cdots, e_p, \frac{1}{\|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2}\left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right\}.$$

We introduce the following projection direction,

$$\begin{aligned}\widehat{u} = \arg\min \ & u^{\intercal}\widehat{\Sigma}u \\ \text{s.t.} \ & \max_{w\in\mathcal{C}}\left\langle w, \widehat{\Sigma}u - \left(\widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^{\intercal}\right\rangle \leq \|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2\lambda_n\end{aligned}$$

where $\lambda_n = C\sqrt{\log p/n}$. Then we propose the final estimator of $\beta_G^{\intercal}\Sigma_{G,G}\beta_G$ as

$$\widehat{\mathrm{Q}}_{\Sigma} = \widehat{\beta}_G^{\intercal}\widehat{\Sigma}_{G,G}\widehat{\beta}_G + \frac{2}{n}\widehat{u}^{\intercal}X^{\intercal}(y - X\widehat{\beta}). \tag{7}$$

Let $\widehat{\sigma}^2$ denote an reasonable estimator of σ^2. We then estimate the variance of the proposed estimator $\widehat{\mathrm{Q}}_\Sigma$ by

$$\widehat{\mathrm{V}}_\Sigma(\tau) = \frac{4\widehat{\sigma}^2}{n}\widehat{u}^\intercal\widehat{\Sigma}\widehat{u} + \frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{\beta}_G^\intercal X_{iG}X_{iG}^\intercal\widehat{\beta}_G - \widehat{\beta}_G^\intercal\widehat{\Sigma}_{G,G}\widehat{\beta}_G\right)^2 + \frac{\tau}{n}, \tag{8}$$

for some positive constant $\tau > 0$. We can combine the proposed point estimator and the asymptotic variance to construct a confidence interval for Q_Σ and conduct the group significance testing. More details will be provided in Section 2.3.

Two remarks about the proposed testing procedure are in order. First, the computational cost of the proposed algorithm is independent of the group size $|G|$. No matter whether the group size is large or small, the construction of the projection direction $\widehat{u}$ involves a constrained optimization problem with a p-dimensional parameter or equivalently in its dual problem a penalized optimization problem with a p-dimensional parameter. Second, we correct the bias of the plug-in estimator all at once, instead of conducting coordinate-wise bias correction. Such a direct correction is not only leading to a computationally efficient algorithm, but also providing a good balance between the bias and variance, especially for a large group size $|G|$.

2.2 Inference for Q_A

The main idea of estimating Q_A is similar to that of estimating Q_Σ though the problem itself is slightly easier due to the fact that the matrix A is known. We start with the error decomposition of the plug-in estimator $\widehat{\beta}_G^\intercal A\widehat{\beta}_G$,

$$\widehat{\beta}_G^\intercal A\widehat{\beta}_G - \beta_G^\intercal A\beta_G = -2\widehat{\beta}_G^\intercal A(\beta_G - \widehat{\beta}_G) - (\widehat{\beta}_G - \beta_G)^\intercal A(\widehat{\beta}_G - \beta_G).$$

Similarly, we can construct the projection direction $\widehat{u}_A$ as

$$\begin{aligned}\widehat{u}_A = \arg\min\ & u^\intercal\widehat{\Sigma}u \\ \text{s.t.}\ & \max_{w\in\mathcal{C}}\left\langle w, \widehat{\Sigma}u - \begin{pmatrix}\widehat{\beta}_G^\intercal A & \mathbf{0}\end{pmatrix}^\intercal\right\rangle \le \|A\widehat{\beta}_G\|_2\lambda_n\end{aligned} \tag{9}$$

where $\lambda_n = C\sqrt{\log p/n}$ and $\mathcal{C}_A = \left\{e_1, \cdots, e_p, \frac{1}{\|A\widehat{\beta}_G\|_2}\begin{pmatrix}\widehat{\beta}_G^\intercal A & \mathbf{0}\end{pmatrix}^\intercal\right\}$. Then we propose the final estimator of Q_A as

$$\widehat{\mathrm{Q}}_A = \widehat{\beta}_G^\intercal A\widehat{\beta}_G + \frac{2}{n}\widehat{u}_A^\intercal X^\intercal(y - X\widehat{\beta}) \tag{10}$$

We then estimate the variance of $\widehat{\mathrm{Q}}_\Sigma$ by $\widehat{\mathrm{V}}_A(\tau)$ with

$$\widehat{\mathrm{V}}_A(\tau) = \frac{4\widehat{\sigma}^2}{n}\widehat{u}_A^\intercal\widehat{\Sigma}\widehat{u}_A + \frac{\tau}{n} \tag{11}$$

for some positive constant $\tau > 0$. In the following, we provide some discussion on comparing $\widehat{\mathrm{Q}}_\Sigma$ in (7) and $\widehat{\mathrm{Q}}_A$ in (10) and then we consider the special case $\|\beta_G\|_2^2$.

2.2.1 Comparison to inference procedure for Q_Σ

We compare now the general inference results for Q_A with those for Q_Σ. The connection is that Q_Σ is equal to Q_A if the positive definite matrix A is taken as $\Sigma_{G,G}$. However, in most applications, the covariance submatrix $\Sigma_{G,G}$ is typically unknown and this leads to a more challenging problem for conducting statistical inference for Q_Σ, that is, we have to estimate $\Sigma_{G,G}$ by the data. As a result, we also need to quantify the uncertainty of estimating $\Sigma_{G,G}$. Through comparing the variance of $\widehat{\mathrm{Q}}_\Sigma$ in (8) and the variance of $\widehat{\mathrm{Q}}_A$ in (11), we observe that there is one additional term $\frac{1}{n^2}\sum_{i=1}^n \left(\widehat{\beta}_G^\intercal X_{iG} X_{iG}^\intercal \widehat{\beta}_G - \widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \widehat{\beta}_G\right)^2$ in the variance level for $\widehat{\mathrm{Q}}_\Sigma$, which captures the additional uncertainty of estimating $\Sigma_{G,G}$. Beyond the additional complexity of dealing with the unknown matrix $\widehat{\Sigma}_{G,G}$ in Q_Σ, the quantity with the true covariance matrix in the middle has its own advantage. To illustrate the advantage, we consider a simpler setting where the random vector $X_{i,G}$ is independent of the random vector X_{i,G^c}, in such a setting, we can simply estimate Σ by $\begin{pmatrix} \widehat{\Sigma}_{G,G} & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_{G^c,G^c} \end{pmatrix}$ and provide an alternative way of constructing the projection defined in (3) by $\widehat{u} = \left(\widehat{\beta}_G^\intercal \quad \mathbf{0}\right)^\intercal$. If we further assume that $|G| \ll n$, we can provide an alternative way of constructing the projection defined in (9) by $\widehat{u}_A = \left(\widehat{\beta}_G^\intercal A \widehat{\Sigma}_{G,G}^{-1} \quad \mathbf{0}\right)^\intercal$. Even in such a simplified setting where $X_{i,G}$ is independent of the random vector X_{i,G^c} and $|G| \ll n$, we observe that the constructed projection direction $\widehat{u}$ is easier than $\widehat{u}_A$ as the construction of $\widehat{u}$ is free of inverting the matrix $\widehat{\Sigma}_{G,G}$. In the case that the covariates in $X_{i,G}$ are highly correlated or the matrix $\widehat{\Sigma}_{G,G}$ is close to singular, the inference procedure depending on Q_Σ is more reliable due to this observation. Additionally, $\mathrm{Q}_\Sigma = \mathbf{E}|X_{i,G}^\intercal \beta_G|^2$ amounts to estimating the regression surface with $\tilde{y}_i = y_i - X_{i,G^c}^\intercal \beta_{G^c}$. Therefore, Q_Σ is identifiable even if some of the components of $X_{i,G}$ exhibit correlation with absolute values being closed to 1. See Section 5.2 for the numerical results.

2.2.2 A special case with $A = \mathrm{I}$

In this part, we consider a commonly used special example by setting $A = \mathrm{I}$ and decompose the error of the plug-in estimator as, $\|\widehat{\beta}_G\|_2^2 - \|\beta_G\|_2^2 = 2\langle \widehat{\beta}_G, \widehat{\beta}_G - \beta_G \rangle - \|\widehat{\beta}_G - \beta_G\|_2^2$. For this special case, the projection direction can actually be identified via the following optimization algorithm,

$$\widehat{u}_{\mathrm{I}} = \arg\min \quad u^\intercal \widehat{\Sigma} u \quad \text{s.t.} \quad \left\| \widehat{\Sigma} u - \left(\widehat{\beta}_G^\intercal \quad \mathbf{0}\right)^\intercal \right\|_\infty \le \|\widehat{\beta}_G\|_2 \lambda_n.$$

Note that $\left\| \widehat{\Sigma} u - \left(\widehat{\beta}_G^\intercal \quad \mathbf{0}\right)^\intercal \right\|_\infty$ can be viewed as $\max_{w \in \mathcal{C}_0} \left\langle w, \widehat{\Sigma} u - \left(\widehat{\beta}_G^\intercal \quad \mathbf{0}\right)^\intercal \right\rangle$ where $\mathcal{C}_0 = \{e_1, \cdots, e_p\}$. In contrast to $\widehat{u}$ and $\widehat{u}_{\mathrm{A}}$, this algorithm for constructing $\widehat{u}_{\mathrm{I}}$ is simpler since the constraint set $\mathcal{C}_0$ is smaller than $\mathcal{C}$, that is, we do not need to constraint the difference from

the additional projection $\frac{1}{\|\widehat{\beta}_G\|_2}\left(\widehat{\beta}_G^{\mathsf{T}} \quad \mathbf{0}\right)^{\mathsf{T}}$. The reason is that $\widehat{\beta}_G$ is closed to β_G, which is a sparse vector no matter how large the set G is.

2.3 Testing procedures and confidence intervals

Having introduced the point estimators and the quantification of the variance, we propose the following two α-level significance tests,

$$\phi_\Sigma(\tau) = \mathbf{1}\left(\widehat{\mathrm{Q}}_\Sigma \geq z_{1-\alpha}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)}\right) \quad \text{and} \quad \phi_A(\tau) = \mathbf{1}\left(\widehat{\mathrm{Q}}_A \geq z_{1-\alpha}\sqrt{\widehat{\mathrm{V}}_A(\tau)}\right), \tag{12}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal random variable. As side products, we can construct confidence intervals for Q_A and Q_Σ as follows:

$$\begin{aligned} \mathrm{CI}_\Sigma(\tau) &= \left(\widehat{\mathrm{Q}}_\Sigma - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)},\ \widehat{\mathrm{Q}}_\Sigma + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)}\right) \\ \mathrm{CI}_A(\tau) &= \left(\widehat{\mathrm{Q}}_A - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{V}}_A(\tau)},\ \widehat{\mathrm{Q}}_A + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{V}}_A(\tau)}\right). \end{aligned} \tag{13}$$

3 Theoretical Justification

In this section, we establish the theoretical properties of the estimators and inference procedures proposed in the methodological section. We first introduce the following regularity conditions.

(A) In model (1), the rows $X_{i,\cdot}$ are i.i.d. p-dimensional sub-Gaussian random vectors with mean $\mu = \mathbf{E}X_{i\cdot}$ and the second order moment $\Sigma = \mathbf{E}(X_{i,\cdot}X_{i,\cdot}^{\mathsf{T}})$ where $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$ for positive constants $C_0 > c_0 > 0$. The errors $\epsilon_1, ..., \epsilon_n$ are i.i.d centered Gaussian random variables with variance σ^2 and are assumed to be independent of X.

Under the sub-Gaussian condition on the covariates X, Condition (A) implies the restricted eigenvalue condition introduced in [3]; see [26,36] for the exact statement. The Gaussianity of the error is only imposed for simplifying the technical analysis and such an assumption can be weakened to sub-Gaussianity using a more refined analysis.

Beyond the model assumption, we also introduce the following conditions for the initial estimators and give an example right after to demonstrate the existence of initial estimators $\widehat{\beta}$ and $\widehat{\sigma}$ satisfying the following conditions (B1) and (B2).

(B1) With probability larger than $1 - g(n)$ where $g(n) \to 0$, the initial estimator $\widehat{\beta}$ and $\widehat{\sigma}^2$ satisfy,

$$\|\widehat{\beta} - \beta\|_2 \lesssim \sqrt{\frac{\|\beta\|_0 \log p}{n}},\ \|\widehat{\beta} - \beta\|_1 \lesssim \|\beta\|_0\sqrt{\frac{\log p}{n}},\ \left|\frac{\widehat{\sigma}^2}{\sigma^2} - 1\right| \lesssim \frac{1}{\sqrt{n}} + \frac{\|\beta\|_0 \log p}{n}.$$

(B2) The initial estimator $\widehat{\beta}$ is independent of the data (X, y) used in the construction of (7) and (10).

Most of the high-dimensional estimators proposed in the literature are shown to satisfy the above condition (B1) of estimating the regression vector and the variance of the regression error under various conditions. See [2,3,29] and the references therein for more details. The condition (B2) is imposed for technical analysis. With such an independence assumption, the asymptotic normality of the proposed estimators is easier to establish. However, such a condition is believed to be only technical and not necessary for the proposed method. As shown in the simulation study, we demonstrate that the proposed method, even not satisfying the independence assumption imposed in the condition (B2), still works well numerically. On the other hand, there exist scenarios where we can construct estimators satisfying the condition (B2). If we have historical data, then we can estimate the regression vector using the historical data and conduct the correction as proposed in (7) and (10) using the current data. Even in the case where we have only access to a single data set (X, y), we can use sample splitting e to create the independence. We randomly split the data into two subsamples $(X^{(1)}, y^{(1)})$ with sample size $n_1 = \lfloor \frac{n}{2} \rfloor$ and $(X^{(2)}, y^{(2)})$ with sample size $n_2 = n - n_1$ and estimate $\widehat{\beta}$ based on the data $(X^{(1)}, y^{(1)})$ and conduct the correction in (7) in the following form,

$$\widehat{\mathrm{Q}}_{\Sigma} = \widehat{\beta}_G^{\intercal} \widehat{\Sigma}^{(2)}_{G,G} \widehat{\beta}_G + \frac{2}{n} \widehat{u}^{\intercal} (X^{(2)})^{\intercal} (y^{(2)} - X^{(2)} \widehat{\beta})$$

with $\widehat{\Sigma}^{(2)} = \frac{1}{n_2} (X^{(2)})^{\intercal} X^{(2)}$ and

$$\begin{aligned} \widehat{u} = \arg\min \ & u^{\intercal} \widehat{\Sigma}^{(2)} u \\ \text{s.t.} \ & \max_{w \in \mathcal{C}} \left\langle w, \widehat{\Sigma}^{(2)} u - \begin{pmatrix} \widehat{\beta}_G^{\intercal} \widehat{\Sigma}^{(2)} & \mathbf{0} \end{pmatrix}^{\intercal} \right\rangle \leq \| \widehat{\Sigma}^{(2)}_{G,G} \widehat{\beta}_G \|_2 \lambda_n. \end{aligned}$$

A similar argument can be used to reconstruct the estimator $\widehat{\mathrm{Q}}_A$ defined in (10). As a result, the estimator using sampling-splitting is less efficient due to the fact that only half of the data is used in constructing the initial estimator and also correcting the bias. Multiple sample splitting and aggregation [24], single sample splitting and cross-fitting [10] or data-swapping [15] are commonly used sample splitting techniques.

The following theorem characterizes the behavior of the proposed estimator $\widehat{\mathrm{Q}}_{\Sigma}$. The estimation error is decomposed into two components, where a stochastic error M_{Σ} has an asymptotic normal limit and the remaining error B_{Σ} is controlled.

Theorem 1 *Suppose that condition* (A) *holds and the initial estimator* $\widehat{\beta}$ *satisfies the conditions* (B1) *and* (B2)*, then the proposed estimator* $\widehat{\mathrm{Q}}_{\Sigma}$ *satisfies* $\widehat{\mathrm{Q}}_{\Sigma} - \mathrm{Q}_{\Sigma} = \mathrm{M}_{\Sigma} + \mathrm{B}_{\Sigma}$ *where*

the main error component M_Σ *and remaining error component* B_Σ *satisfy*

$$\mathrm{M}_\Sigma/\sqrt{\mathrm{V}_\Sigma^0} \to N(0,1) \quad \textit{with} \quad \mathrm{V}_\Sigma^0 = \frac{4\sigma^2}{n}\widehat{u}^\intercal\widehat{\Sigma}\widehat{u} + \frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{\beta}_G^\intercal X_{iG}X_{iG}^\intercal\widehat{\beta}_G - \widehat{\beta}_G^\intercal\widehat{\Sigma}_{G,G}\widehat{\beta}_G\right)^2 \tag{14}$$

$$\mathbf{P}\left(|\mathrm{B}_\Sigma| \gtrsim \left(\|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2 + \|\Sigma_{G,G}\|_2\right)\frac{k\log p}{n}\right) \le p^{-c} \tag{15}$$

As a consequence, under the additional condition $k \ll \sqrt{n}/\log p$*, we have*

$$\limsup_{n,p\to\infty}\mathbf{P}\left(\left|\widehat{\mathrm{Q}}_\Sigma - \mathrm{Q}_\Sigma\right| \ge z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{V}_\Sigma}\right) \le \alpha \quad \textit{with}\ \mathrm{V}_\Sigma = \mathrm{V}_\Sigma^0 + \frac{\tau}{n} \tag{16}$$

for some positive constant $\tau > 0$*.*

The above theorem establishes that the main error component M_Σ of the error decomposition has an asymptotic normal limit and the remaining part B_Σ is well controlled in terms of the convergence rate in (15). Such a decomposition is useful from the inference perspective, where (16) establishes that if the sparsity level is small enough, then the $\alpha/2$ quantile of the standardized difference $(\widehat{\mathrm{Q}}_\Sigma - \mathrm{Q}_\Sigma)/\sqrt{\mathrm{V}_\Sigma}$ is similar to that of standard normal. In this case, we show that B_Σ is negligible in comparison to $\sqrt{\mathrm{V}_\Sigma} = \sqrt{\mathrm{V}_\Sigma^0 + \frac{\tau}{n}}$, for any given positive constant $\tau > 0$. The variance level V_Σ is slightly enlarged from V_Σ^0 to $\mathrm{V}_\Sigma^0 + \frac{\tau}{n}$ to quantify the uncertainty of $\mathrm{M}_\Sigma + \mathrm{B}_\Sigma$. Since there is no distributional result for B_Σ, an upper bound for B_Σ would be one (conservative) alternative to quantify the uncertainty of B_Σ.

The enlargement of the uncertainty quantification from V_Σ^0 to $\mathrm{V}_\Sigma^0 + \frac{\tau}{n}$ is closely related to "super-efficiency". Consider the case of bounded spectral norm $\|\Sigma\|_2$. The variance level V_Σ^0 in (14) is of the order $(\|\beta_G\|_2 + \|\beta_G\|_2^2)/\sqrt{n}$ and hence the variance level near the null hypothesis $\mathrm{Q}_\Sigma = 0$ is much faster than the parametric rate, which corresponds to the "super-efficiency" phenomenon. In this case, the worst upper bound for B_Σ is $(1 + \|\beta_G\|_2)\cdot\frac{k\log p}{n}$, which can dominate $\sqrt{\mathrm{V}_\Sigma^0}$ for $\|\beta_G\|_2$ being close to zero, even if $k \ll \sqrt{n}/\log p$. To overcome the challenges posed by "super-efficiency", we enlarge the variance a bit by the level $\frac{\tau}{n}$ such that it always dominates the worst upper bound for B_Σ.

Additionally, the statistical accuracy of the test statistics depends only weakly on the group size $|G|$, in the sense that, the standard deviation of the test statistic depends on $\|\beta_G\|_2$, at the order of magnitude $(\|\beta_G\|_2 + \|\beta_G\|_2^2)/\sqrt{n}$; but since $\|\beta_G\|_2 \le \|\beta\|_2$ and β is a sparse vector, then the standard deviation level is not always strictly increasing with a growing set G and this phenomenon explains the statistical efficiency of the proposed test, especially when the test size G is large. In contrast, the F-test or χ^2 test proposed in [25,32] have the standard deviation at the scale of $\sqrt{|G|/n}$, which is strictly increasing with the testing size $|G|$ and also explains the condition on the group size $|G| \ll n$.

An important feature of the proposed method is that the current proposed testing procedure imposes no conditions on the test size G. It works for a small group but also for the group size as large as $|G| = p$.

In parallel to Theorem 1, the following theorem characterizes the error decomposition of the estimator $\widehat{\mathrm{Q}}_A$.

Theorem 2 *Suppose that condition* (A) *holds and the initial estimator* $\widehat{\beta}$ *satisfies the conditions* (B1) *and* (B2)*, then the proposed estimator* $\widehat{\mathrm{Q}}_A$ *satisfies* $\widehat{\mathrm{Q}}_A - \mathrm{Q}_A = \mathrm{M}_A + \mathrm{B}_A$ *where the main error component* M_A *and remaining error component* B_A *satisfy*

$$\mathrm{M}_A/\sqrt{\mathrm{V}_A^0} \to N(0,1) \quad \textit{with} \quad \mathrm{V}_A^0 = \frac{4\sigma^2}{n}\widehat{u}_A^\intercal \widehat{\Sigma}\widehat{u}_A \tag{17}$$

and

$$\mathbf{P}\left(|\mathrm{B}_A| \gtrsim \left(\|A\widehat{\beta}_G\|_2 + \|A\|_2\right)\frac{k\log p}{n}\right) \le p^{-c} \tag{18}$$

As a consequence, under the additional condition $k \ll \sqrt{n}/\log p$*, we have*

$$\limsup_{n,p\to\infty} \mathbf{P}\left(\left|\widehat{\mathrm{Q}}_A - \mathrm{Q}_A\right| \ge z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{V}_A}\right) \le \alpha \quad \textit{with} \quad \mathrm{V}_A = \mathrm{V}_A^0 + \frac{\tau}{n} \tag{19}$$

for some positive constant $\tau > 0$.

Theorem 2 is established in a parallel way to Theorem 1 with the main difference as the variance level of M_A is different from that of M_Σ. Specifically, the variance component of M_Σ consists of two components, the uncertainty of estimating β and Σ while the variance component of M_A only reflects the uncertainty of estimating β.

Three of the most interesting examples of A are $A = \mathrm{I}_{|G|\times|G|}$, $A = \mathrm{diag}\,(\Sigma_{G,G})$ and $A = \Sigma_{G,G}$, where $\mathrm{diag}\,(\Sigma_{G,G})$ denotes the diagonal matrix of $\Sigma_{G,G}$. For the case $A = \mathrm{I}_{|G|\times|G|}$, we treat all regression coefficients equally and focus on the magnitude of the regression coefficients instead of the magnitude of the covariate variance and the correlation between covariates. For the case $A = \mathrm{diag}\,(\Sigma_{G,G})$, the constructed significance test takes the variance level of each covariate into consideration but does not take the correlation between all covariates inside G into account. For the case $A = \Sigma_{G,G}$, we take into consideration both the covariate variance level and also the correlation between all covariates inside G. However, the constructed test, with assuming $A = \Sigma_{G,G}$ to be known a priori, is different from the test statistics constructed in (7), where the construction does not make use of any prior knowledge of $\Sigma_{G,G}$. The difference between (17) and (14) illustrates the additional uncertainty we need to account for due to the unknown weighted matrix $\Sigma_{G,G}$.

In the following, we control the type I error of the proposed testing procedure and establish the asymptotic power of the proposed estimator. We consider the following parameter

space for $\theta = (\beta, \Sigma, \sigma)$,

$$\Theta(k) = \left\{ (\beta, \Sigma, \sigma) : \|\beta\|_0 \leq k,\ \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1,\ \sigma_1 \leq M_2 \right\},$$

where $M_1 \geq 1$ and $M_2 > 0$ are positive constants. We define the null hypothesis parameter space as

$$\mathcal{H}_0 = \{(\beta, \Sigma, \sigma) \in \Theta(k) : \|\beta_G\|_2 = 0\}$$

for a fixed group G.

Corollary 1 *Suppose that condition* (A) *holds and the initial estimators* $(\widehat{\beta}, \widehat{\sigma})$ *satisfy the conditions* (B1) *and* (B2). *If* $k \ll \sqrt{n}/\log p$, *then, for any given positive constant* $\tau > 0$, *both proposed tests* $\phi_\Sigma(\tau)$ *and* $\phi_A(\tau)$ *defined in* (12) *control the type I error,*

$$\sup_{\theta \in \mathcal{H}_0} \liminf_{n,p \to \infty} \mathbf{P}_\theta(\phi = 1) \leq \alpha \ \text{ for } \ \phi = \phi_\Sigma(\tau) \text{ or } \phi_A(\tau).$$

To study the power, we define the local alternative hypothesis parameter space as

$$\mathcal{H}_{1,A}(\delta) = \left\{ (\beta, \Sigma, \sigma) \in \Theta(k) : \beta_G^{\mathrm{T}} A \beta_G = \delta \right\}$$

To facilitate the presentation, the local alternative parameter space is defined such that it depends on the positive-definite matrix A. The following corollaries present the power of the proposed testing procedures in (12).

Corollary 2 *Suppose that condition* (A) *holds and the initial estimators* $(\widehat{\beta}, \widehat{\sigma})$ *satisfy the conditions* (B1) *and* (B2). *If* $k \ll \sqrt{n}/\log p$, *then for any given positive constant* $\tau > 0$ *and any* $\theta \in \mathcal{H}_{1,\Sigma_{G,G}}(\delta(t))$ *with* $\delta(t) = (1.01 z_{1-\alpha} + t)\sqrt{\mathrm{V}_\Sigma}$, *the proposed test* $\phi_\Sigma(\tau)$ *in* (12) *has the asymptotic power,*

$$\liminf_{n,p \to \infty} \mathbf{P}_\theta(\phi_\Sigma(\tau) = 1) \geq 1 - \Phi(-t) \tag{20}$$

where V_Σ *is defined in* (16) *and* $\Phi(\cdot)$ *is the quantile function of standard normal distribution.*

As a remark, for the local alternative space defined in (3), the separation parameter of the indifference region $\delta(t) = (1.01 z_{1-\alpha} + t)\sqrt{\mathrm{V}_\Sigma}$ is of the order $\frac{1+t}{\sqrt{n}}(\sqrt{\tau} + \|\beta_G\|_2 + \|\beta_G\|_2^2)$ and hence the proposed test ϕ_Σ is of power converging to 1 as long as $t \to \infty$. Similarly, we control the asymptotic power for the proposed test ϕ_A in the following Corollary.

Corollary 3 *Suppose that condition* (A) *holds and the initial estimators* $(\widehat{\beta}, \widehat{\sigma})$ *satisfy the conditions* (B1) *and* (B2). *If* $k \ll \sqrt{n}/\log p$, *then for any given positive constant* $\tau > 0$ *and*

any $\theta \in \mathcal{H}_{1,A}(\delta)$ *with* $\delta(t) = (1.01 z_{1-\alpha} + t)\sqrt{\mathrm{V}_A}$, *the proposed test* $\phi_A(\tau)$ *in* (12) *has the asymptotic power,*

$$\liminf_{n,p\to\infty} \mathbf{P}_\theta \left(\phi_A(\tau) = 1\right) \geq 1 - \Phi(-t) \tag{21}$$

where V_A *is defined in* (19) *and* $\Phi(\cdot)$ *is the quantile function of standard normal distribution.*

4 Statistical Applications

In the following sections, we motivate three statistical applications of the group significance test. In Section 4.1, we explain how to apply hierarchical testing and its advantage of immensely reducing the number of possible group tests that could be performed. Section 4.2 covers interaction tests for detecting heterogeneous effects and finally Section 4.3 motivates the concept of local heritability in genetics measuring explained partial variance.

4.1 Hierarchical testing

In presence of many variables where some of them are highly correlated with each other, it is often too ambitious to detect significant single variables having regression coefficients being different from zero. An effect from a single variable after having adjusted for all others is often not sufficiently strong. This happens in particular in presence of high correlation or near collinearity among the variables. On the other hand, a group of variables is easier to be detected as significant, especially and again when the variables are highly correlated. The core issue then becomes which of the groups of variables to consider for testing.

Hierarchical testing is a powerful method to go through a sequence of groups to be tested, from larger groups to smaller ones depending on the strength of the signal and the amount of correlation among the variables in and between the groups. As such, it is a multiple testing scheme which controls the familywise error rate. The details are as follows.

The p covariates are structured into groups of variables in a hierarchical tree $\mathcal{T}$ such that at every level of the tree, the groups build a partition of $\{1, \ldots, p\}$. The groups at each level of the tree are such that the variables have high correlation within groups (and a tendency for low correlations between groups). The default choice for constructing such a tree is with hierarchical clustering of the variables [16, cf.], typically using $1 - \text{correlation}^2$ as dissimilarity measure and complete linkage.

We assume that $\mathcal{T}$ is deterministic, for example when conditioning on the covariates in the linear model and being the output of hierarchical clustering. Hierarchical testing with respect to $\mathcal{T}$ is then a sequential multiple testing adjustment procedure as follows.

Hierarchical testing procedure

1: **INPUT:** Hierarchical tree $\mathcal{T}$ with nodes corresponding to groups of variables; Group testing procedure returning p-values P_G for each group of variables G, e.g. as described in Section 2.3; Significance level α.

2: **OUTPUT:** Significant groups of variables such that the procedure controls the familywise error rate (FWER).

3: **repeat**

4: Go top-down the tree $\mathcal{T}$ and perform group significance testing for groups G. The raw p-value is corrected for multiplicity using

$$\tilde{P}_G = P_G \cdot p/|G|,$$
$$P_{G;\text{adjusted}} = \max_{G' \supseteq G} \tilde{P}_{G'},$$

where G' is any group in the tree $\mathcal{T}$. The second line enforces monotonicity of the adjusted p-values.

5: For each group G when going top-down in $\mathcal{T}$: if $P_{G;\text{adjusted}} \leq \alpha$, continue to consider the children of G for group testing.

6: **until** All the children of each group G (i.e., the finer partition of G) when going top-down in $\mathcal{T}$ are non-significant at level α.

A schematic illustration with a binary hierarchical tree is shown in Figure 1. The color encodes whether the null hypothesis of a group could be rejected or not.

There are a few interesting properties of hierarchical testing. First, we can think of it as a hybrid of a sequential procedure and Bonferroni correction: for every level in the tree, the p-value adjustment is a weighted Bonferroni correction (the standard Bonferroni correction if the groups have equal size) and across different levels it is a sequential procedure with no correction but a stopping criterion to not go further down the tree when no rejection happens. Indeed, the root node needs no adjustment at all and also for each level in the tree, the correction depends only on the partitioning on that level and not on how many tests have been done before. Secondly, there is no need to pre-define the level of resolution of the groups, i.e., to decide where the tree should be cut. How far one goes down the tree with testing significance of groups of variables is fully data-driven, based on the hierarchical testing procedure. Third, the hierarchical testing method is computationally attractive as no further tests are considered once a certain group does not exhibit any significance. It is shown in [22] that the procedure controls the familywise error rate (FWER). The hierarchical testing method has been used in the setting of high-dimensional linear models in [20] with a further refinement in [21] based on a multi sample-splitting testing method for the groups. The latter is justified with the strong and questionable assumption that

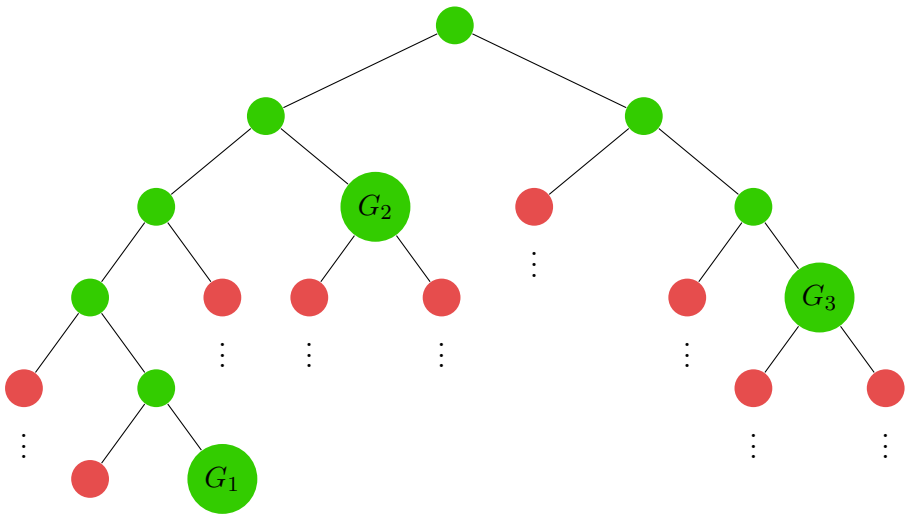


Figure 1: The hierarchical procedure returns the three groups G_1, G_2, and G_3. The green and red colors highlight significant groups and non-significant groups, respectively. The group G_1 is a leaf which means that it consists of one variable.

the Lasso is able to detect all the relevant variables and in this sense, this procedure is not fully reliable in terms of error control. Further details are provided in [27].

4.2 Testing interaction and detection of effect heterogeneity

The proposed significance test is useful in testing the existence of interaction, which itself is an important statistical problem. We focus on the interaction model (3) and re-express the model as $y_i = W_{i\cdot}^\intercal \eta + \epsilon_i$ where we define $W_i = (D_i X_{i\cdot}^\intercal, X_{i\cdot}^\intercal)^\intercal$ and $\eta = (\gamma^\intercal, \beta^\intercal)^\intercal$. Then the detection of interaction terms is to test whether $H_0 : \gamma = \eta_G = 0$ for $G = \{1, 2, \cdots, p\}$. We can then apply the group significance test developed here.

The detection of heterogeneous treatment effect can be viewed as a special case of testing the interaction term. If D_i in the interaction model (3) is taken as a binary variable, where $D_i = 0$ denotes that the subject belongs to the control group and $D_i = 1$ that it belongs to the treatment group. Then this specific test for interaction amounts to testing whether the treatment effect is heterogeneous. In a very similar way, if D_i takes two values where $D_i = 1$ and $D_i = 2$ represent the subject is receiving treatment 1 and 2, respectively, then the test of interaction is for testing whether the difference between two treatment effects is heterogeneous or not. The current developed method of detecting heterogeneous treatment effects is definitely not restricted to the case of a binary treatment. It can be applied to basically any type of treatment variables, such as count variables, categorical or continuous

variables.

4.3 Local heritability

Local heritability is defined as a measure of the partial variance explained by a given set of genetic variables. In contrast to the (global) heritability, the local heritability is more informative as it describes the variability explained by a pre-specified set of genetic variants and takes the global heritability as one special case. Assuming the regression model (1), the local heritability can be represented by the quantities, $\beta_G^{\intercal}\Sigma_{G,G}\beta_G$ and $\|\beta_G\|_2^2$, where G denotes the index set of interest. The following corollary establishes the coverage and precision properties of the proposed confidence intervals for two measures of local heritability, $\beta_G^{\intercal}\Sigma_{G,G}\beta_G$ and $\|\beta_G\|_2^2$.

Corollary 4 *Suppose that condition* (A) *holds and the initial estimators* $(\widehat{\beta}, \widehat{\sigma})$ *satisfy the conditions* (B1) *and* (B2)*, then the constructed confidence intervals defined in* (13) *satisfy*

$$\liminf_{n,p\to\infty} \mathbf{P}(\mathrm{Q}_\Sigma \in \mathrm{CI}_\Sigma(\tau)) \geq 1-\alpha, \quad \liminf_{n,p\to\infty} \mathbf{P}(\mathrm{Q}_A \in \mathrm{CI}_A(\tau)) \geq 1-\alpha,$$

where $\tau > 0$ *is any given positive constant.*

5 Simulation Studies

In this section, we investigate the finite sample performance of the proposed method over three different simulation scenarios. Throughout the simulation study, we generate the high dimensional linear model

$$y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i \quad \text{for } 1 \leq i \leq n.$$

with the dimension $p = 500$. We generate the covariates following $X_{i\cdot} \sim N(\mathbf{0}, \Sigma)$ and the error $\epsilon_i \sim N(0,1)$, both being i.i.d. over the indices i.

5.1 Dense alternatives

In this section, we consider the setting where the regression vector is relatively dense but of small non-zero coefficients, as this is a challenging scenario for detecting the signals. We generate the regression vector β as $\beta_j = \delta$ for $25 \leq j \leq 50$ and $\beta_j = 0$ otherwise and generate the covariance matrix $\Sigma_{ij} = 0.6^{|i-j|}$ for $1 \leq i, j \leq 500$. We consider the following group significance test,

$$H_{0,G} : \beta_i = 0 \text{ for } i \in G, \text{ where } G = \{30, 31, \cdots, 200\}.$$

We vary the signal strength parameter δ over $\{0, 0.02, 0.04, 0.06\}$ and the sample size n over $\{250, 350, 500, 800\}$.

We compare the proposed method with two alternative procedures, the maximum test based on the debiased estimator proposed in [18], shorthanded as FD (Fast Debiased) and the maximum test based on the debiased estimator proposed in [31], shorthanded as hdi. Specifically, we produce the FD debiased estimators $\{\widehat{\beta}_j^{\mathrm{FD}}\}_{1\le j\le p}$ by the online code on `https://web.stanford.edu/~montanar/sslasso/code.html` and the hdi estimator $\{\widehat{\beta}_j^{\mathrm{hdi}}\}_{1\le j\le p}$ by using the R package [12]. The additional products of these implemented algorithms include the corresponding variance and covariance matrix, denoted as $\mathrm{Cov}(\widehat{\beta}^{\mathrm{FD}}) \in \mathbb{R}^{p\times p}$ and $\mathrm{Cov}(\widehat{\beta}^{\mathrm{hdi}}) \in \mathbb{R}^{p\times p}$, respectively. Regarding the pre-specified group G, we sample i.i.d copies $Z_1, Z_2, \cdots, Z_{10,000} \in \mathbb{R}^{|G|}$ following $N(\mathbf{0}, \mathrm{Cov}(\widehat{\beta}^{\mathrm{FD}})_{G\times G})$ and calculate q_α^{FD} as the empirical top α quantile of $\max_{j\in G}|Z_{1,j}|, \max_{j\in G}|Z_{2,j}|, \cdots, \max_{j\in G}|Z_{10,000,j}|$. We can define q_α^{hdi} in the similar way with replacing $\mathrm{Cov}(\widehat{\beta}^{\mathrm{FD}})_{G\times G}$ by $\mathrm{Cov}(\widehat{\beta}^{\mathrm{hdi}})_{G\times G}$. Then we can define the following test of group significance as

$$\phi_{\mathrm{FD}} = \mathbf{1}\left(\max_{j\in G}|\widehat{\beta}_j^{\mathrm{FD}}| \ge q_\alpha^{\mathrm{FD}}\right) \quad \text{and} \quad \phi_{\mathrm{hdi}} = \mathbf{1}\left(\max_{j\in G}|\widehat{\beta}_j^{\mathrm{hdi}}| \ge q_\alpha^{\mathrm{hdi}}\right). \tag{22}$$

To implement the group significance test proposed in the current paper, we consider four specific tests, $\phi_{\mathrm{I}}(0), \phi_{\mathrm{I}}(1), \phi_\Sigma(0), \phi_\Sigma(1)$ where $\phi_\Sigma(0), \phi_\Sigma(1)$ are defined in (12) with $\tau = 0$ and $\tau = 1$, respectively and $\phi_{\mathrm{I}}(0), \phi_{\mathrm{I}}(1)$ are defined in (12) (by taking $A = \mathrm{I}$) with $\tau = 0$ and $\tau = 1$, respectively. We shall compare $\phi_{\mathrm{I}}(0), \phi_{\mathrm{I}}(1), \phi_\Sigma(0), \phi_\Sigma(1)$ and $\phi_{\mathrm{FD}}, \phi_{\mathrm{hdi}}$ in two settings, dense alternatives (the current section) and high correlation (Section 5.2).

Table 1 summarizes the hypothesis testing results for different methods. For $\delta = 0$, the empirical detection rate is an empirical measure of the type I error; For $\delta \neq 0$, the empirical detection rate is an empirical measure of the power. We have observed that, for $\tau = 0$, the testing procedures $\phi_{\mathrm{I}}(0)$ and $\phi_\Sigma(0)$ do not control the type I error. The reason is that the bias component dominates the variance and $\tau = 0$ only quantifies the uncertainty of the variance component but does not account for that from the bias. In contrast, as long as we set $\tau = 1$ thereby providing a conservative upper bound for the bias component, the proposed procedures $\phi_{\mathrm{I}}(1)$ and $\phi_\Sigma(1)$ control the type I error. As comparison, the maximum test ϕ_{hdi} controls the type I error while the other maximum test ϕ_{FD} does not reliably control the type I error. To compare the power, we focus on $\phi_{\mathrm{I}}(1)$, $\phi_\Sigma(1)$, ϕ_{hdi} and ϕ_{FD} and observe that $\phi_\Sigma(1)$ is the best in terms of power among all four tests and $\phi_{\mathrm{I}}(1)$ is the second best at most cases, especially when $\delta = 0.04, 0.06$. The power of both ϕ_{hdi} and ϕ_{FD} are much lower than our proposed two testing procedures, $\phi_\Sigma(1)$ and $\phi_{\mathrm{I}}(1)$. An interesting observation is that, although the proposed testing procedures $\phi_\Sigma(1)$ and $\phi_{\mathrm{I}}(1)$ control the type I error in a conservative sense, they still achieve a higher power than the existing maximum tests.

δ	n	$\phi_{\mathrm{I}}(0)$	$\phi_{\mathrm{I}}(1)$	$\phi_{\Sigma}(0)$	$\phi_{\Sigma}(1)$	ϕ_{FD}	ϕ_{hdi}
0	250	0.962	0.002	0.994	0.008	0.112	0.044
	350	0.992	0.000	0.998	0.004	0.086	0.042
	500	0.996	0.002	1.000	0.002	0.078	0.048
	800	0.980	0.000	1.000	0.000	0.064	0.038
0.02	250	0.970	0.020	1.000	0.058	0.138	0.064
	350	0.998	0.006	1.000	0.062	0.112	0.066
	500	0.994	0.004	1.000	0.084	0.102	0.068
	800	1.000	0.004	1.000	0.100	0.068	0.056
0.04	250	0.986	0.230	1.000	0.618	0.226	0.084
	350	1.000	0.188	1.000	0.854	0.184	0.106
	500	1.000	0.292	1.000	0.946	0.128	0.112
	800	1.000	0.374	1.000	1.000	0.128	0.180
0.06	250	0.998	0.746	1.000	0.986	0.344	0.162
	350	1.000	0.838	1.000	1.000	0.316	0.192
	500	1.000	0.948	1.000	1.000	0.252	0.272
	800	1.000	0.992	1.000	1.000	0.222	0.366

Table 1: Empirical Rejection Rate (ERR) for the Dense Alternative scenario (5% significance level). We report the ERR for six different tests $\phi_{\mathrm{I}}(0), \phi_{\mathrm{I}}(1), \phi_{\Sigma}(0), \phi_{\Sigma}(1), \phi_{\mathrm{FD}}$ and ϕ_{hdi}, where ERR denotes the proportion of rejected hypothesis among the total 500 simulations.

In Table 2, we report the coverage properties for the confidence intervals $\mathrm{CI}_{\mathrm{I}}(\tau = 0)$ and $\mathrm{CI}_{\mathrm{I}}(\tau = 1)$ for $\|\beta_G\|_2^2$ and $\mathrm{CI}_{\Sigma}(\tau = 0)$ and $\mathrm{CI}_{\Sigma}(\tau = 1)$ for $\beta_G^{\mathsf{T}}\Sigma_{G,G}\beta_G$. We observe that for $\tau = 0$, the constructed confidence intervals do not achieve 95% coverage while for $\tau = 1$, they do. This happens due to the reason that for small δ, the bias component of the proposed estimator cannot be simply ignored and we use a conservative upper bound to control the bias component. In the supplementary materials, we report in Table 8 the absolute value of the bias of the plug-in estimator and the proposed estimator.

5.2 Highly correlated covariates

Here, we consider the setting where the regression vector is relatively sparse but a few variables are highly correlated. We generate the regression vector β as $\beta_1 = \beta_3 = \delta$ and

δ	n	$\mathrm{CI}_{\mathrm{I}}(\tau=0)$	$\mathrm{CI}_{\mathrm{I}}(\tau=1)$	$\mathrm{CI}_{\Sigma}(\tau=0)$	$\mathrm{CI}_{\Sigma}(\tau=1)$
0	250	0.070	1.000	0.014	0.996
	350	0.020	1.000	0.006	1.000
	500	0.016	1.000	0.004	1.000
	800	0.038	1.000	0.002	1.000
0.02	250	0.148	0.998	0.436	0.996
	350	0.088	1.000	0.456	1.000
	500	0.102	1.000	0.478	1.000
	800	0.104	1.000	0.606	1.000
0.04	250	0.222	0.988	0.662	1.000
	350	0.158	0.996	0.724	0.994
	500	0.226	0.996	0.772	0.996
	800	0.244	1.000	0.788	0.990
0.06	250	0.276	0.954	0.592	0.922
	350	0.246	0.984	0.798	0.970
	500	0.246	0.972	0.802	0.966
	800	0.282	0.984	0.834	0.966

Table 2: Empirical Coverage for the Dense Alternative scenario (95% nominal coverage). We report the empirical coverage of $\mathrm{CI}_{\mathrm{I}}(\tau=0)$ and $\mathrm{CI}_{\mathrm{I}}(\tau=1)$ for $\|\beta_G\|_2^2$ and the empirical coverage of $\mathrm{CI}_{\Sigma}(\tau=0)$ and $\mathrm{CI}_{\Sigma}(\tau=1)$ for $\beta_G^{\intercal}\Sigma_{G,G}\beta_G$.

$\beta_j = 0$ for $j \neq 1, 3$ and we generate the covariance matrix as follows,

$$\Sigma_{ij} = \begin{cases} 0.8 & \text{if } 1 \leq i \neq j \leq 5 \\ 1 & \text{if } 1 \leq i = j \leq 5 \\ 0.6^{|i-j|} & \text{otherwise.} \end{cases}$$

There exists high correlations among the first five variables, where the pairwise correlation is 0.8 inside this group of five variables. In contrast to the previous simulation setting in Section 5.1, we do not generate a large number of non-zero entries in the regression coefficient but only assign the first and third coefficients to be possibly non-zero. We test the group hypothesis generated by the first five regression coefficients,

$$H_{0,G} : \beta_i = 0 \text{ for } i \in G, \text{ where } G = \{1, 2, \cdots, 5\}.$$

We vary the signal strength parameter δ over $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and the sample size n over $\{250, 350, 500\}$.

Table 3 compares the empirical detection rate for our proposed method and also the maximum test. We have observed that, for $\tau = 0$, the proposed testing procedures $\phi_{\mathrm{I}}(0)$

and $\phi_\Sigma(0)$ do not control type I error while for $\tau = 1$, $\phi_{\mathrm{I}}(1)$ and $\phi_\Sigma(1)$ achieve control of the type I error below the significance level. The maximum test procedure ϕ_{hdi} controls the type I error while ϕ_{FD} barely controls it. Regarding the power, ϕ_{hdi} and ϕ_{FD} are better for $\delta = 0.2.0, 3$ while our proposed testing procedure $\phi_\Sigma(1)$ is comparable to ϕ_{hdi} and ϕ_{FD} when δ reaches 0.4.

Seemingly, our proposed procedures $\phi_\Sigma(1)$ and $\phi_\Sigma(0)$ do not perform better than the maximum test ϕ_{hdi} and ϕ_{FD}. We shall emphasize that the coverage properties related to the maximum test ϕ_{hdi} and ϕ_{FD} are not guaranteed although this is not visible in Table 3. Specifically, since we are testing $\beta_i = 0$ for $i = 1, 2, 3, 4, 5$, we can look at the coverage properties of these two proposed tests in terms of β_i. As reported in Table 5, for $\delta \neq 0$, we have observed that the coordinate-wise coverage properties are not guaranteed due to the high correlation among the first five variables. The reason for this phenomenon is that the coverage for an individual coordinate β_j requires a decoupling between the j-th and all other variables and if there exists high correlations, this decoupling step is difficult to be conducted accurately. In contrast, even though the first five variables are highly correlated, the constructed confidence intervals $\mathrm{CI}_{\mathrm{I}}(\tau = 1)$ and $\mathrm{CI}_\Sigma(\tau = 1)$ achieve the 95% coverage. This is reported in Table 4. As explained, our proposed testing procedure is more robust to high correlations inside the testing group as the whole group is tested as a unit instead of decoupling variables inside the testing group. In the supplementary materials, we report in Table 9 the absolute value of the bias of the plug-in estimator and the proposed estimator.

5.3 Hierarchical testing

We simulate data under two settings which differ in the set of active covariates and covariance matrix Σ used to generate $X_{i\cdot} \sim N(\mathbf{0}, \Sigma)$. The covariance matrix Σ is block diagonal in both cases. The set of indices of the active covariates is denoted by S_0. The $|S_0| = 10$ active covariates are chosen as the first variable in the first 10 blocks. The number of covariates p is kept fixed at $p = 500$ and we vary the number of observations n between $100, 200, 300, 500,$ and 800. The results are calculated based on 500 simulation runs.

In *setting 1*, the first 20 covariates are generated to have high correlation within small blocks of size 2. This means that the covariance matrix Σ has 1's on the diagonal, $\Sigma_{i,i+1} = \Sigma_{i+1,i} = 0.7$ on the first off-diagonals for $i = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$, and 0's otherwise. The set of active covariates is $S_0 = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$.

In *setting 2*, there are ten blocks corresponding each to 50 covariates that have a high correlation of 0.7. This means that the covariance matrix Σ has 1's on the diagonal, $\Sigma_{B_k} = 0.7$ for $B_k = \big\{(i, j) : i \neq j \text{ and } i, j \in \{k, k+1, \ldots, k+p/10\}\big\}$ for $k = 1, 51, 101, \ldots$, and 0's otherwise. The set of active covariates is $S_0 = \{1, 51, 101, 151, \ldots, 451\}$.

δ	n	$\phi_{\mathrm{I}}(0)$	$\phi_{\mathrm{I}}(1)$	$\phi_{\Sigma}(0)$	$\phi_{\Sigma}(1)$	ϕ_{FD}	ϕ_{hdi}
	250	0.034	0.000	0.072	0.000	0.070	0.036
0	350	0.058	0.000	0.104	0.000	0.082	0.062
	500	0.052	0.000	0.092	0.000	0.082	0.056
	250	0.314	0.002	0.686	0.016	0.836	0.370
0.1	350	0.450	0.006	0.920	0.006	0.948	0.462
	500	0.390	0.016	0.940	0.028	0.986	0.540
	250	0.682	0.134	0.982	0.590	0.998	0.936
0.2	350	0.704	0.138	1.000	0.822	1.000	0.972
	500	0.634	0.234	1.000	0.960	1.000	0.994
	250	0.844	0.518	0.998	0.982	1.000	1.000
0.3	350	0.850	0.654	1.000	1.000	1.000	1.000
	500	0.790	0.630	1.000	1.000	1.000	1.000
	250	0.950	0.906	1.000	1.000	1.000	1.000
0.4	350	0.922	0.896	1.000	1.000	1.000	1.000
	500	0.910	0.876	1.000	1.000	1.000	1.000
	250	0.966	0.952	1.000	1.000	1.000	1.000
0.5	350	0.962	0.960	1.000	1.000	1.000	1.000
	500	0.950	0.942	1.000	1.000	1.000	1.000

Table 3: Empirical Rejection Rate for the Highly Correlated scenario (5% significance level). We report the ERR for six different tests $\phi_{\mathrm{I}}(0), \phi_{\mathrm{I}}(1), \phi_{\Sigma}(0), \phi_{\Sigma}(1), \phi_{\mathrm{FD}}$ and ϕ_{hdi}, where ERR denotes the proportion of rejected hypothesis among the total 500 simulations.

We use a modified version of the power to measure the performance of the hierarchical procedure because groups of variable sizes are returned. The idea is that we weight the true findings by one over the group size because smaller significant groups are more informative than larger ones, i.e., single variables get a weight of one. The adaptive power is defined by

$$\mathrm{Power}_{\mathrm{adap}} = \frac{1}{|S_0|} \sum_{C \in \mathrm{MTD}} \frac{1}{|C|}$$

where MTD stands for Minimal True Detections, i.e., there is no significant subgroup ("Minimal"), the group has to be significant ("Detection"), and the group contains at least one active variable ("True"); see [20].

The results are illustrated in Figure 2 and Table 6. The hierarchical procedure performs very well for setting 1 since the adaptive power is close to 1 for all values of n. The setup in setting 2 is much harder because the 10 active covariates are each highly correlated with 49 other non-active covariates. It is difficult to distinguish the active variables from the

δ	n	$\mathrm{CI}_{\mathrm{I}}(\tau=0)$	$\mathrm{CI}_{\mathrm{I}}(\tau=1)$	$\mathrm{CI}_{\Sigma}(\tau=0)$	$\mathrm{CI}_{\Sigma}(\tau=1)$
	250	0.104	1.000	0.090	1.000
0	350	0.110	1.000	0.088	1.000
	500	0.094	1.000	0.070	1.000
	250	0.756	1.000	0.558	1.000
0.1	350	0.830	1.000	0.590	1.000
	500	0.910	0.998	0.692	1.000
	250	0.912	0.992	0.822	0.998
0.2	350	0.916	0.998	0.822	0.996
	500	0.924	0.994	0.842	0.996
	250	0.908	0.998	0.820	0.942
0.3	350	0.888	0.980	0.822	0.972
	500	0.896	0.972	0.872	0.968
	250	0.884	0.972	0.832	0.912
0.4	350	0.874	0.960	0.836	0.912
	500	0.918	0.956	0.890	0.956
	250	0.900	0.958	0.860	0.894
0.5	350	0.908	0.948	0.854	0.906
	500	0.926	0.960	0.902	0.946

Table 4: Empirical Coverage for the Highly Correlated scenario (95% nominal coverage). We report the empirical coverage of $\mathrm{CI}_{\mathrm{I}}(\tau=0)$ and $\mathrm{CI}_{\mathrm{I}}(\tau=1)$ for $\|\beta_G\|_2^2$ and the empirical coverage of $\mathrm{CI}_{\Sigma}(\tau=0)$ and $\mathrm{CI}_{\Sigma}(\tau=1)$ for $\beta_G^{\intercal}\Sigma_{G,G}\beta_G$.

correlated ones for the procedure and hence, it stops further up in the tree which results in larger significant groups compared to setting 1. This can be seen in the last three columns of Table 6. Thus, the adaptive power is smaller. The FWER is well controlled for both settings.

6 Real Data Analysis

In the following, we present the results of the hierarchical procedure for two real data sets. In Section 6.1, we compare the hierarchical procedure to testing single variables on a gene expression data set. In Section 6.2, the hierarchical procedure is analyzed for 46 traits (different responses) as linear functions of single nucleotide polymorphism (SNP) binary covariates.

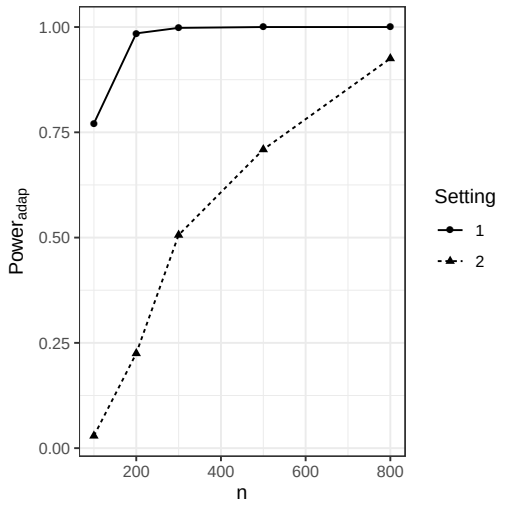
		CI_{FD}					CI_{hdi}				
δ	n	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
	250	0.972	0.968	0.970	0.976	0.976	0.952	0.950	0.944	0.950	0.946
0	350	0.968	0.972	0.962	0.970	0.968	0.942	0.942	0.932	0.966	0.948
	500	0.974	0.972	0.964	0.970	0.982	0.950	0.936	0.956	0.950	0.956
	250	0.766	0.824	0.778	0.808	0.820	0.948	0.894	0.940	0.892	0.692
0.1	350	0.590	0.756	0.610	0.782	0.800	0.950	0.904	0.936	0.906	0.640
	500	0.490	0.740	0.488	0.768	0.758	0.946	0.896	0.944	0.900	0.584
	250	0.400	0.714	0.418	0.720	0.758	0.864	0.798	0.910	0.828	0.268
0.2	350	0.464	0.696	0.414	0.722	0.680	0.910	0.822	0.922	0.844	0.234
	500	0.424	0.702	0.408	0.686	0.674	0.876	0.860	0.916	0.842	0.298
	250	0.430	0.724	0.426	0.720	0.740	0.870	0.808	0.890	0.818	0.218
0.3	350	0.386	0.732	0.432	0.682	0.720	0.832	0.836	0.904	0.836	0.258
	500	0.422	0.692	0.426	0.694	0.686	0.860	0.856	0.900	0.854	0.280
	250	0.422	0.708	0.432	0.694	0.742	0.832	0.832	0.878	0.782	0.272
0.4	350	0.432	0.704	0.454	0.728	0.704	0.848	0.844	0.882	0.854	0.256
	500	0.404	0.728	0.414	0.680	0.706	0.880	0.876	0.902	0.826	0.308
	250	0.416	0.712	0.418	0.720	0.680	0.852	0.810	0.856	0.816	0.234
0.5	350	0.398	0.708	0.424	0.662	0.692	0.854	0.844	0.878	0.810	0.270
	500	0.406	0.726	0.398	0.666	0.700	0.876	0.878	0.886	0.812	0.312

Table 5: Empirical Coverage for first five regression coefficients for Highly Correlated scenario (95% nominal coverage). The numbers under CI_{FD} represent the empirical coverage for β_j $(1 \leq j \leq 5)$ using the method proposed in [18] and the numbers under CI_{hdi} represent the empirical coverage for β_j $(1 \leq j \leq 5)$ using the method proposed in [31].

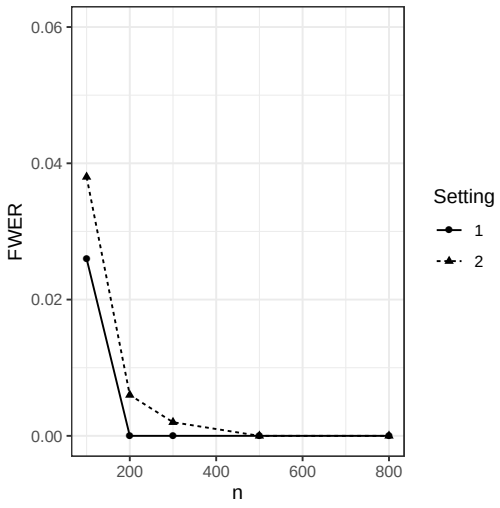
6.1 Riboflavin data set

We demonstrate the hierarchical procedure and compare it to testing of single variables on a data set about Riboflavin production with Bacillus Subtilis, made publicly available by [5]. It consists of $n = 71$ samples of strains of Bacillus Subtilis for which they measured the riboflavin (vitamin B_2) production rate. Engineered strains and strains grown under different fermentation conditions were hybridized multiple times during a fed-batch fermentation process. The log-expression level of $p = 4088$ genes is tested for association with the response.

The hierarchical procedure goes top-down through a hierarchical cluster tree which was estimated using $1 - (\text{empirical correlation})^2$ as dissimilarity measure and average linkage. The results of the hierarchical procedure are displayed in Table 7. We compare them



(a) Adaptive power.



(b) FWER.

ht

Figure 2: Adaptive power and FWER for the simulation study of the hierarchical procedure (5% significance level). In setting 1, the 10 active covariates are each highly correlated with only one other covariate. In setting 2, the 10 active covariates are each highly correlated with 49 other covariates.

to testing all the single covariates using the group test, i.e., all groups of size one, and correcting for multiplicity using Bonferroni-Holm. Testing of all single covariates revels seven significant covariates. Ideally with the hierarchical procedure, we would find the same seven single variables plus some additional groups: indeed, this happens here. The seven single variables are highlighted in green in Table 7. The average pairwise correlation is 0.98 for distinct covariates within the significant group of size three from the hierarchical procedure. Because of high correlations, this suggests that it is not possible to refine this group further to significant subgroups.

6.2 Yeast colony growth data set

Bloom et al. [4] performed a genome-wide association study of 46 quantitative traits to investigate the sources of missing heritability. The authors crossbred 1'008 yeast Saccharomyces cerevisiae segregates from a laboratory strain and a wine strain and measured 11'623 genotype markers which they reduced to 4'410 markers that show less correlation. Bloom et al. [4] processed the data such that the covariates encode from which of the two strains a given genotype was passed on. This is encoded using the values 1 and -1. Each crossbred was exposed to 46 different conditions like different temperatures, pH values, carbon sources, additional metal ions, and small molecules. The traits of interest are the

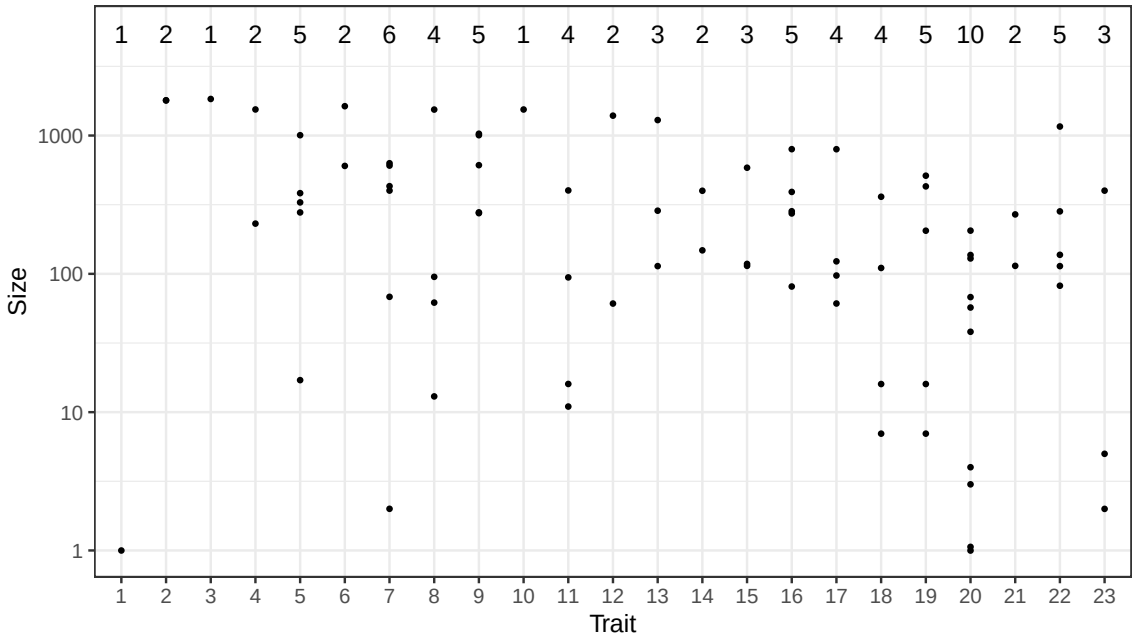
Setting	n	p	FWER	adaptive power	avg number	avg size	median size
1	100	500	0.026	0.771	9.4	3.6	1.0
1	200	500	0.000	0.985	10.0	1.0	1.0
1	300	500	0.000	0.998	10.0	1.0	1.0
1	500	500	0.000	1.000	10.0	1.0	1.0
1	800	500	0.000	1.000	10.0	1.0	1.0
2	100	500	0.038	0.029	7.3	49.2	47.5
2	200	500	0.006	0.225	10.0	31.6	40.5
2	300	500	0.002	0.506	10.0	19.2	13.5
2	500	500	0.000	0.709	10.0	11.0	1.0
2	800	500	0.000	0.925	10.0	3.4	1.0

Table 6: Results from the simulation study of the hierarchical procedure (5% significance level). The last three columns are average number, average size, and median size of the significant groups. The results are based on 500 simulation runs.

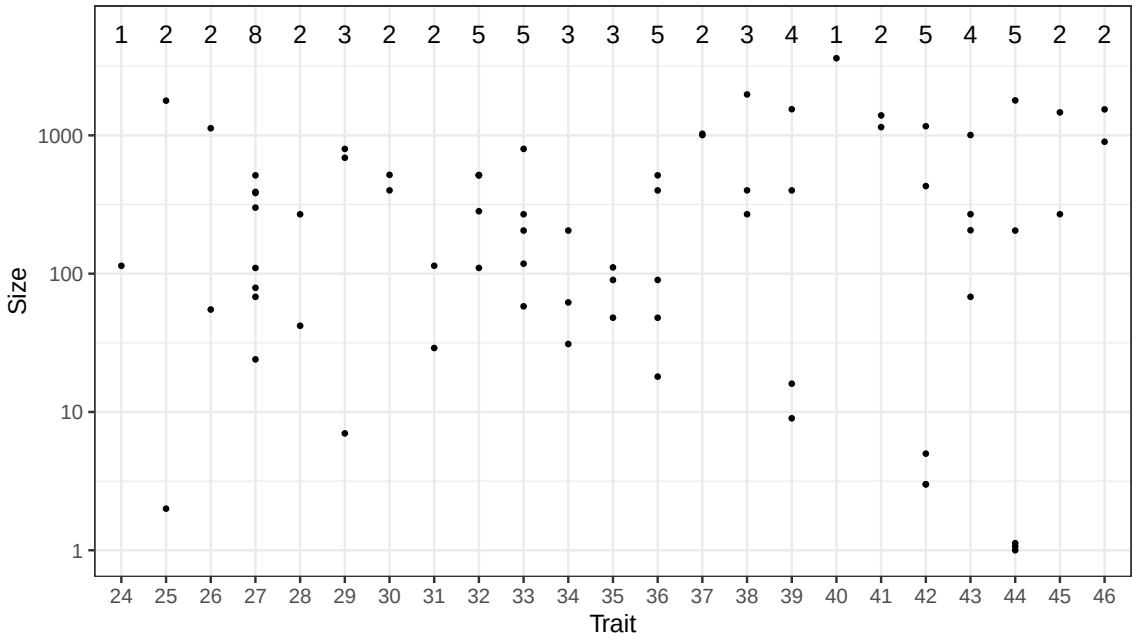
end-point colony size normalized by the colony size on control medium, see [4] for further details.

We use these data sets to illustrate the hierarchical procedure with our group testing method. We consider each trait separately and thus, we are considering 46 different regression problems. We only use complete observations without any missing values, leading to sample sizes between $n = 599$ and $n = 1007$ depending on the trait and number of covariates always being $p = 4410$.

The results across the 46 traits are given in Figure 3. The hierarchical procedure always finds some significant groups of SNP covariates. Some of the significant findings include single variable, namely in traits 1, 20, and 44; while some of them are large groups with cardinality bigger than 1000. It is plausible that one cannot find too many single variables but it is reasonable and convincing to see that the hierarchical method finds a substantial amount of significant groups. The amount of "signal", in terms of significant findings, varies quite a bit across the 46 traits. For example, we conclude that trait 20 exhibits most "signal": it leads to most significant groups, two of them being a single variable and some of them being small. Thus, this trait manifests most association between the colony size (response) and genetics in terms of SNP covariates.



(a) Traits 1 until 23.



(b) Traits 24 until 46.

Figure 3: Size of significant groups (p-values ≤ 0.05) by applying the hierarchical procedure to each of the 46 traits of the Yeast Colony Growth data set. The number of significant groups is displayed on the top. We added a small amount of jitter to all the groups of size one for traits 20 (two groups) and 44 (three groups) in order to prevent over-plotting.

p-value	significant cluster
1.631e-11	YEBC_at
$<$ 2.2e-16	LYSC_at
$<$ 2.2e-16	XTRA_at
$<$ 2.2e-16	XKDS_at
0.01420	YXLC_at, YXLD_at, YXLG_at
0.01420	YOAB_at
0.04544	BMR_at
0.01420	YCKE_at

Table 7: Results from applying the hierarchical procedure to the Riboflavin data set (5% significance level). The seven covariates which are found as well by testing for single variables (groups of size one only) are highlighted in green.

7 Conclusions

The current paper studies the group inference problem in high-dimensional linear models by constructing an approximately unbiased estimator of the weighted quadratic functionals of the regression sub-vector corresponding to the group and then quantifying its uncertainty. The proposed group inference procedure has been shown to enjoy nice theoretical and empirical properties, including, valid inference even when variables in the group are highly correlated, good power performance when there are many small coefficients in the group, computational efficiency even for a large group and assumption-free regarding the group size. The proposed method is then applied to hierarchical testing, the detection of heterogeneous treatment effects, tests of interaction and inference for local heritability.

Acknowledgment

Z. Guo is grateful for Dr. Cun-Hui Zhang's discussion on comparing the current proposed method with the high-dimensional F-test; Z. Guo is grateful for Dr. Hongzhe Li's discussion on local heritability and Dr. Dave Zhao's discussion on test of interaction.

References

[1] 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.

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

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

[4] 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.

[5] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, 1:255–278, 2014.

[6] L. Buzdugan, M. Kalisch, A. Navarro, D. Schunk, E. Fehr, and P. Bühlmann. Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics*, 32:1990–2000, 2016.

[7] 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, 2017.

[8] T Tony Cai and Zijian Guo. Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *arXiv preprint arXiv:1806.06179*, 2018.

[9] Tianxi Cai, Tony Cai, and Zijian Guo. Individualized treatment selection: An optimal hypothesis testing approach in high-dimensional models. *arXiv preprint arXiv:1904.12891*, 2019.

[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[11] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

[12] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30(4):533–558, 2015.

[13] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

[14] 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*, 114(525):358–369, 2019.

[15] Zijian Guo and Cun-Hui Zhang. Local inference in additive models with decorrelated local linear estimator. *arXiv preprint arXiv:1907.12732*, 2019.

[16] J. Hartigan. *Clustering Algorithms*. Wiley, 1975.

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

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

[19] Jonas R Klasen, Elke Barbez, Lukas Meier, Nicolai Meinshausen, Peter Bühlmann, Maarten Koornneef, Wolfgang Busch, and Korbinian Schneeberger. A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nature communications*, 7:13299, 2016.

[20] J. Mandozzi and P. Bühlmann. Hierarchical testing in the high-dimensional setting with correlated variables. *Journal of the American Statistical Association*, 111:331–343, 2016.

[21] J. Mandozzi and P. Bühlmann. A sequential rejection testing method for high-dimensional regression with correlated variables. *International Journal of Biostatistics*, 12:79–95, 2016.

[22] Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

[23] Nicolai Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945, 2015.

[24] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[25] Ritwik Mitra and Cun-Hui Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2):1829–1873, 2016.

[26] 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.

[27] Claude Renaux, Laura Buzdugan, Markus Kalisch, and Peter Bühlmann. Hierarchical inference for genome-wide association studies: a view on methodology with software. *arXiv preprint arXiv:1805.02988*, 2018.

[28] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, 2016.

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

[30] Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[31] 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.

[32] Sara van de Geer and Benjamin Stucky. χ 2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*, pages 279–306. Springer, 2016.

[33] Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

[34] 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.

[35] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

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

A Additional Simulation Results

Table 8 considers the dense alternative setting in Section 5.1 and summarizes the absolute values of bias of the proposed estimators $\widehat{\mathrm{Q}_{\mathrm{I}}}$ and $\widehat{\mathrm{Q}}_{\Sigma}$ and the plug-in estimators $\|\widehat{\beta}_G\|_2^2$ and $\widehat{\beta}_G^{\mathsf{T}}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$. Table 9 reports the same results for the highly correlated setting in Section 5.2. Through comparison, we have observed the effect of bias correction, that is, the absolute bias of the proposed estimators are lower than the corresponding plug-in estimators in most cases. However, when the true value of the estimand is close to zero, the plug-in estimator without correction can perform even better. Generally speaking, the plug-in estimator tends to be an under-estimator due to the shrinkage effect of penalization. For the special case that the true value is close to zero, the shrinkage effect leads to an estimator close to zero, which can actually be an accurate estimator in such a scenario.

δ	n	Q_{I}	Bias($\widehat{\mathrm{Q}_{\mathrm{I}}}$)	Bias($\|\widehat{\beta}_G\|_2^2$)	Q_{Σ}	Bias($\widehat{\mathrm{Q}}_{\Sigma}$)	Bias($\widehat{\beta}_G^{\mathsf{T}}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$)
0	250	0.000	0.049	0.009	0.000	0.044	0.010
	350	0.000	0.031	0.006	0.000	0.031	0.007
	500	0.000	0.026	0.004	0.000	0.024	0.005
	800	0.000	0.017	0.003	0.000	0.016	0.003
0.02	250	0.008	0.056	0.006	0.031	0.031	0.014
	350	0.008	0.039	0.004	0.031	0.023	0.016
	500	0.008	0.035	0.002	0.031	0.018	0.017
	800	0.008	0.026	0.001	0.031	0.011	0.018
0.04	250	0.034	0.075	0.002	0.122	0.003	0.071
	350	0.034	0.058	0.002	0.122	0.012	0.068
	500	0.034	0.049	0.000	0.122	0.010	0.067
	800	0.034	0.038	0.001	0.122	0.005	0.062
0.06	250	0.076	0.095	0.000	0.275	0.038	0.144
	350	0.076	0.074	0.001	0.275	0.002	0.135
	500	0.076	0.064	0.001	0.275	0.009	0.123
	800	0.076	0.046	0.003	0.275	0.004	0.109

Table 8: Absolute values of the bias for the Dense Alternative scenario. We report the empirical coverage absolute values of the bias of four estimators $\widehat{\mathrm{Q}}_{\mathrm{I}}, \widehat{\mathrm{Q}}_{\Sigma}, \|\widehat{\beta}_G\|_2^2$ and $\widehat{\beta}_G^{\mathsf{T}}\widehat{\Sigma}_{G,G}\widehat{\beta}_G$ and also the corresponding true values Q_{I} and Q_{Σ}.

δ	n	$\mathrm{Q_I}$	Bias($\widehat{\mathrm{Q_I}}$)	Bias($\|\widehat{\beta}_G\|_2^2$)	Q_Σ	Bias($\widehat{\mathrm{Q}}_\Sigma$)	Bias($\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \widehat{\beta}_G$)
	250	0.000	0.001	0.000	0.000	0.001	0.000
0	350	0.000	0.001	0.000	0.000	0.001	0.000
	500	0.000	0.001	0.000	0.000	0.001	0.000
	250	0.020	0.008	0.012	0.036	0.006	0.026
0.1	350	0.020	0.007	0.011	0.036	0.006	0.024
	500	0.020	0.008	0.010	0.036	0.002	0.022
	250	0.080	0.017	0.031	0.144	0.004	0.066
0.2	350	0.080	0.006	0.033	0.144	0.010	0.063
	500	0.080	0.006	0.030	0.144	0.005	0.055
	250	0.180	0.006	0.065	0.324	0.013	0.115
0.3	350	0.180	0.004	0.060	0.324	0.014	0.103
	500	0.180	0.005	0.059	0.324	0.006	0.090
	250	0.320	0.000	0.099	0.576	0.020	0.164
0.4	350	0.320	0.003	0.092	0.576	0.022	0.145
	500	0.320	0.000	0.083	0.576	0.007	0.123
	250	0.500	0.031	0.144	0.900	0.034	0.224
0.5	350	0.500	0.026	0.128	0.900	0.026	0.185
	500	0.500	0.018	0.112	0.900	0.010	0.155

Table 9: Absolute values of the bias for the Highly Correlated scenario. We report the empirical coverage absolute values of the bias of four estimators $\widehat{\mathrm{Q}}_\mathrm{I}, \widehat{\mathrm{Q}}_\Sigma, \|\widehat{\beta}_G\|_2^2$ and $\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \widehat{\beta}_G$ and also the corresponding true values $\mathrm{Q_I}$ and Q_Σ.

B Proofs

We present the proofs of Theorems 1 and 2 in the following and postpone the proofs of Corollaries 1, 2, 3 and 4 to Section C.2 of the supplementary materials.

This proposed estimator $\widehat{\mathrm{Q}}_\Sigma$ has the following error decomposition,

$$\begin{aligned}\widehat{\mathrm{Q}}_\Sigma - \mathrm{Q}_\Sigma = &\frac{2}{n}\widehat{u}^\intercal X^\intercal \epsilon + \beta_G^\intercal(\widehat{\Sigma}_{G,G} - \Sigma_{G,G})\beta_G \\ &+ 2\left[\widehat{\Sigma}\widehat{u} - \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^\intercal\right]^\intercal (\beta - \widehat{\beta}) - (\widehat{\beta}_G - \beta_G)^\intercal \widehat{\Sigma}_{G,G}(\widehat{\beta}_G - \beta_G).\end{aligned}$$

We define

$$\mathrm{M}_\Sigma = \frac{2}{n}\widehat{u}^\intercal X^\intercal \epsilon + \beta_G^\intercal(\widehat{\Sigma}_{G,G} - \Sigma_{G,G})\beta_G$$

and

$$\mathrm{B}_\Sigma = 2\left[\widehat{\Sigma}\widehat{u} - \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^\intercal\right]^\intercal (\beta - \widehat{\beta}) - (\widehat{\beta}_G - \beta_G)^\intercal \widehat{\Sigma}_{G,G}(\widehat{\beta}_G - \beta_G).$$

We establish (14) by the central limiting theorem and control (15) by the following lemma, whose proof is postponed to Section C.1 of the supplementary materials.

Lemma 1 *Suppose that Conditions* (A), (B1), (B2) *hold, then with probability larger than* $1 - p^{-c} - g(n) - \exp(-\sqrt{n})$,

$$\left| \left[\widehat{\Sigma}\widehat{u} - \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0} \right)^\intercal \right]^\intercal (\beta - \widehat{\beta}) \right| \lesssim \|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2 \frac{k \log p}{n} \tag{23}$$

$$\left| (\widehat{\beta}_G - \beta_G)^\intercal \widehat{\Sigma}_{G,G} (\widehat{\beta}_G - \beta_G) \right| \lesssim \|\Sigma_{G,G}\|_2 \frac{k \log p}{n} \tag{24}$$

$$\sqrt{\frac{1}{n} \widehat{u}^\intercal X^\intercal X \widehat{u}} \asymp \|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2 \tag{25}$$

Then the control of the reminder term (15) follows from (23) and (24). Then (16) follows from the fact that $\|\Sigma_{G,G}\|_2 \frac{k \log p}{n} \ll \frac{\tau}{\sqrt{n}}$ and $\|\widehat{\Sigma}_{G,G}\widehat{\beta}_G\|_2 \frac{k \log p}{n} \ll \sqrt{\frac{1}{n^2} \widehat{u}^\intercal X^\intercal X \widehat{u}} \leq \sqrt{\mathrm{V}_\Sigma^0}$, where the first bound is due to the condition $k \ll \sqrt{n}/\log p$ and the second bound is due to (25) and the condition $k \ll \sqrt{n}/\log p$.

The proof of Theorem 2 is similar to that of Theorem 1. We start with the decomposition

$$\widehat{\mathrm{Q}}_A - \mathrm{Q}_A = \frac{2}{n} \widehat{u}_A^\intercal X^\intercal \epsilon + 2 \left[\widehat{\Sigma}\widehat{u}_A - \left(\widehat{\beta}_G^\intercal A \quad \mathbf{0} \right)^\intercal \right]^\intercal (\beta - \widehat{\beta}) - (\widehat{\beta}_G - \beta_G)^\intercal A (\widehat{\beta}_G - \beta_G).$$

and define

$$\mathrm{M}_A = \frac{2}{n} \widehat{u}_A^\intercal X^\intercal \epsilon$$

and

$$\mathrm{B}_A = 2 \left[\widehat{\Sigma}\widehat{u}_A - \left(\widehat{\beta}_G^\intercal A \quad \mathbf{0} \right)^\intercal \right]^\intercal (\beta - \widehat{\beta}) - (\widehat{\beta}_G - \beta_G)^\intercal A (\widehat{\beta}_G - \beta_G)$$

We can establish a similar Lemma as Lemma 1 and present the corresponding proof in Section C.1 of the supplementary materials.

Lemma 2 *Suppose that Conditions* (A), (B1), (B2) *hold, then with probability larger than* $1 - p^{-c} - g(n)$,

$$\left| \left[\widehat{\Sigma}\widehat{u}_A - \left(\widehat{\beta}_G^\intercal A \quad \mathbf{0} \right)^\intercal \right]^\intercal (\beta - \widehat{\beta}) \right| \lesssim \|A\widehat{\beta}_G\|_2 \frac{k \log p}{n} \tag{26}$$

$$\left| (\widehat{\beta}_G - \beta_G)^\intercal A (\widehat{\beta}_G - \beta_G) \right| \lesssim \|A\|_2 \frac{k \log p}{n} \tag{27}$$

$$\sqrt{\frac{1}{n} \widehat{u}_A^\intercal X^\intercal X \widehat{u}_A} \asymp \|A\widehat{\beta}_G\|_2 \tag{28}$$

Then (17) follows from the fact that ϵ_i are normal random variables and (18) follows from (26) and (27). Then (19) follows from the fact that $\|A\|_2 \frac{k \log p}{n} \ll \frac{\tau}{\sqrt{n}}$ and $\|A\widehat{\beta}_G\|_2 \frac{k \log p}{n} \ll \sqrt{\frac{1}{n^2} \widehat{u}_A^\intercal X^\intercal X \widehat{u}_A} \leq \sqrt{\mathrm{V}_A^0}$, where the first bound is due to the condition $k \ll \sqrt{n}/\log p$ and the second bound is due to (28) and the condition $k \ll \sqrt{n}/\log p$.

C Additional Proofs

C.1 Proofs of Lemmas 1 and 2

The proof of (23) follows from

$$\left|\left[\widehat{\Sigma}\widehat{u} - \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^\intercal\right]^\intercal (\beta - \widehat{\beta})\right| \leq \|\widehat{\Sigma}\widehat{u} - \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^\intercal\|_\infty \|\beta - \widehat{\beta}\|_1$$

together with the constraint (3) and the condition (B1). The proof of (26) follows from

$$\left|\left[\widehat{\Sigma}\widehat{u}_A - \left(\widehat{\beta}_G^\intercal A \quad \mathbf{0}\right)^\intercal\right]^\intercal (\beta - \widehat{\beta})\right| \leq \|\widehat{\Sigma}\widehat{u}_A - \left(\widehat{\beta}_G^\intercal A \quad \mathbf{0}\right)^\intercal\|_\infty \|\beta - \widehat{\beta}\|_1$$

together with the constraint (9) and the condition (B1).

Under the independence assumption imposed in (B2), the control of (25) has been established in Lemma 1 of [9], where we specifically take $x_{\text{new}} = \left(\widehat{\beta}_G^\intercal \widehat{\Sigma}_{G,G} \quad \mathbf{0}\right)^\intercal$ and consider the simpler one sample case. Similarly, we can also establish (28).

The proof of (27) follows from $\left|(\widehat{\beta}_G - \beta_G)^\intercal A (\widehat{\beta}_G - \beta_G)\right| \leq \|A\|_2 \|\widehat{\beta}_G - \beta_G\|_2^2$ and Condition (B1). The proof of (24) follows from Lemma 10 of [8], specifically, the definition of event $\mathcal{G}_6(\widehat{\beta}_G - \beta_G, \widehat{\beta}_G - \beta_G, \sqrt{n})$ and hence with probability larger than $1 - \exp(-\sqrt{n})$,

$$\left|(\widehat{\beta}_G - \beta_G)^\intercal \widehat{\Sigma}_{G,G} (\widehat{\beta}_G - \beta_G)\right| \lesssim \left|(\widehat{\beta}_G - \beta_G)^\intercal \Sigma_{G,G} (\widehat{\beta}_G - \beta_G)\right| \leq \|\Sigma_{G,G}\|_2 \|\widehat{\beta}_G - \beta_G\|_2^2.$$

C.2 Proofs of Corollaries 1, 2, 3 and 4

By the condition (B1) and $k \log p \ll \sqrt{n}$, we have

$$d_n(\tau) = \frac{z_{1-\alpha}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)} - \mathrm{B}_\Sigma}{z_{1-\alpha}\sqrt{\mathrm{V}_\Sigma(\tau)}} - 1 = o_p(1).$$

Note that

$$\mathbf{P}_\theta\left(\phi_\Sigma(\tau) = 1\right) = \mathbf{P}_\theta\left(\widehat{\mathrm{Q}}_\Sigma \geq z_{1-\alpha}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)}\right) = \mathbf{P}_\theta\left(\mathrm{Q}_\Sigma + \mathrm{M}_\Sigma + \mathrm{B}_\Sigma \geq z_{1-\alpha}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)}\right).$$

Together with the definition of $d_n(\tau)$, we can further control the above probability by

$$\mathbf{P}_\theta\left(\mathrm{M}_\Sigma \geq (1 + d_n(\tau)) z_{1-\alpha}\sqrt{\mathrm{V}_\Sigma} - \mathrm{Q}_\Sigma\right) = \mathbf{P}_\theta\left(\frac{\mathrm{M}_\Sigma}{\sqrt{\mathrm{V}_\Sigma}} \geq (1 + d_n(\tau)) z_{1-\alpha} - \frac{\mathrm{Q}_\Sigma}{\sqrt{\mathrm{V}_\Sigma}}\right). \quad (29)$$

Then we control the type I error in Corollary 1, following from the limiting distribution established in (14). We can also establish the lower bound for the asymptotic power in Corollary 2 by (29) and the definition $\delta(t) = (1.01 z_{1-\alpha} + t)\sqrt{\mathrm{V}_\Sigma}$. We can use the same argument to control the type I error and the asymptotic power of $\phi_A(\tau)$ in Corollaries 1 and 3. The proof of Corollary 4 follows from (16) and (19) and the assumption (B1) that $\widehat{\sigma}^2$ is a consistent estimator of σ^2.