Statistical Inference and Large-scale Multiple Testing for High-dimensional Regression Models

T. Tony Cai[1], Zijian Guo[2] and Yin Xia[3]

[1]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA.
[2]Department of Statistics, Rutgers University, Piscataway, USA.
[3]Department of Statistics and Data Science, Fudan University, Shanghai, China.

Abstract

This paper presents a selective survey of recent developments in statistical inference and multiple testing for high-dimensional regression models, including linear and logistic regression. We examine the construction of confidence intervals and hypothesis tests for various low-dimensional objectives such as regression coefficients and linear and quadratic functionals. The key technique is to generate debiased and desparsified estimators for the targeted low-dimensional objectives and estimate their uncertainty. In addition to covering the motivations for and intuitions behind these statistical methods, we also discuss their optimality and adaptivity in the context of high-dimensional inference. In addition, we review the recent development of statistical inference based on multiple regression models and the advancement of large-scale multiple testing for high-dimensional regression. The R package `SIHR` has implemented some of the high-dimensional inference methods discussed in this paper.

Keywords: confidence interval, debiasing, false discovery rate, hypothesis testing, linear functionals, quadratic functionals, simultaneous inference

1 Introduction

High-dimensional data analysis has become a vital part of scientific research in many fields. While the abundance of high-dimensional data offers numerous opportunities for statistical analysis, it also presents significant technical

challenges, as the number of variables can be much larger than the number of observations. In high-dimensional settings, most of the classical inferential procedures, such as the maximum likelihood, are no longer applicable. In recent years, there has been significant progress in developing new theories and methods for parameter estimation, hypothesis testing, confidence interval construction, and large-scale simultaneous inference in the context of high-dimensional data analysis.

In this paper, we provide a selective survey of recent advances in statistical inference and multiple testing for high-dimensional regression models, commonly used in modern data analysis across various fields such as genetics, metabolomics, finance, health, and economics. Much progress has been made in estimation and prediction under the high-dimensional linear and generalized linear models (GLMs); see, for example, [1–20]. Theoretical properties have been established in different settings, including the minimax estimation rate and the rate of convergence for the estimation and prediction errors of several penalized procedures.

Uncertainty quantification is at the heart of many critical scientific applications and is more challenging than point estimation. For high-dimensional regression models, although the Lasso and other penalized estimators have been shown to achieve the optimal rates of convergence for estimation, these estimators suffer from non-negligible bias that makes them unsuitable to be directly used for statistical inference, as noted in several studies [21–23]. To overcome this issue, debiased inference methods have been developed in [21–23] that correct the bias of penalized estimators and allow for statistical inference based on the debiased estimators. The development of debiased inference methods has led to an increase in research on statistical inference for a wide range of low-dimensional objectives in different high-dimensional models; see, for example, [24–47]. In particular, [29] studied the minimaxity and adaptivity of confidence intervals for general linear functionals of a high-dimensional regression vector and found significant differences between the cases of sparse and dense loading vectors. Another important inference tool, known as Neyman's orthogonalization or double machine learning, has been proposed in econometrics to enable inference for low-dimensional objectives with high-dimensional nuisance parameters; see, for example, [48–52].

In the single regression model (one-sample) setting, we observe data $\{Y_k, X_{k,\cdot}\}_{1 \leq k \leq n}$, where $Y_k \in \mathbb{R}$ and $X_{k,\cdot} \in \mathbb{R}^{p+1}$ denote the outcome and the high-dimensional covariates respectively, generated independently from the high-dimensional GLM,

$$Y_k = h(X_{k,\cdot}^\mathsf{T} \beta) + \epsilon_k, \quad \text{for} \quad 1 \leq k \leq n \tag{1}$$

with $\mathbb{E}(\epsilon_k | X_{k,\cdot}) = 0$ and the high-dimensional regression vector $\beta \in \mathbb{R}^{p+1}$. Throughout the paper, we use β_1 to denote the intercept and set $X_{k,1} = 1$. We assume that the high-dimensional covariates $X_{k,-1} \in \mathbb{R}^p$ are centered and sub-Gaussian and the matrix $\Sigma \equiv \mathbb{E} X_{k,\cdot} X_{k,\cdot}^\mathsf{T} \in \mathbb{R}^{(p+1) \times (p+1)}$ is well-conditioned. We focus on the linear and logistic regression models with the link function $h(z) = z$ and $h(z) = \exp(z)/[1 + \exp(z)]$ respectively. The regression vector

β is assumed to be sparse, and its sparsity level is denoted by $\|\beta\|_0$. The high-dimensional covariates $X_{k,\cdot}$ might come from a large number of measured covariates or the basis transformations of the baseline covariates. For the linear model, we further assume the error ϵ_k is sub-Gaussian with homoscedastic regression error $\sigma_\epsilon^2 = \mathbb{E}(\epsilon_k^2|X_{k,\cdot})$. Note that, the sub-Gaussian condition can be furthered relaxed to polynomial conditions [e.g., 53, 59] and heteroscedastic regression errors are allowed [e.g., 48, 55, 81].

In addition to the one-sample setting, we examine the statistical inference methods for the two-sample high-dimensional regression models. For $d = 1, 2$, we assume that the data $\{Y_k^{(d)}, X_{k,\cdot}^{(d)}\}_{1 \le k \le n_d}$ are i.i.d. generated, following

$$Y_k^{(d)} = h([X_{k,\cdot}^{(d)}]^\mathsf{T}\beta^{(d)}) + \epsilon_k^{(d)}, \quad \text{for} \quad 1 \le k \le n_d, \tag{2}$$

where $\mathbb{E}(\epsilon_k^{(d)}|X_{k,\cdot}^{(d)}) = 0$ and $h(\cdot)$ is the pre-specified link function.

Based on the models above, this paper first addresses the challenges of making statistical inferences for low-dimensional objectives (e.g., linear and quadratic functionals) in high-dimensional regression, both in one- and two-sample settings. Specifically, the following quantities are of particular interest.

1. **Linear functional** $x_{\text{new}}^\mathsf{T}\beta$ with $x_{\text{new}} \in \mathbb{R}^{p+1}$ in one-sample setting. The linear functional $x_{\text{new}}^\mathsf{T}\beta$ includes as special cases the single regression coefficient β_j [21–23, 29] when x_{new} is the jth canonical unit vector and the conditional mean of the outcome under (1) when x_{new} is a future observation's covariates. When x_{new} denotes the average of the covariates observations for a group, $x_{\text{new}}^\mathsf{T}\beta$ is closely related to average treatment effect [30]. In logistic regression, the linear functional $x_{\text{new}}^\mathsf{T}\beta$ is closely related to the case probability [24].

2. **Quadratic functionals** $\beta_G^\mathsf{T} A \beta_G$ **and** $\beta_G^\mathsf{T} \Sigma_{G,G} \beta_G$ with $G \subset \{1, 2, \cdots, p+1\}$ and $A \in \mathbb{R}^{|G| \times |G|}$ in one-sample setting. For a subset G, these quadratic functionals measure the total effect of variables in G. Statistical inference for quadratic functionals can be motivated from the group significance test, and the (local) genetic heritability estimation [25, 26]. The inference method can be generalized to handle heterogeneous effect tests, hierarchical testing, prediction loss evaluation, and confidence ball construction [26, 27].

3. **Difference between linear functionals** $h(x_{\text{new}}^\mathsf{T}\beta^{(2)}) - h(x_{\text{new}}^\mathsf{T}\beta^{(1)})$ with $x_{\text{new}} \in \mathbb{R}^{p+1}$ in two-sample setting. This difference measures the discrepancy between the conditional means, which is closely related to individual treatment selection for the new observation $x_{\text{new}} \in \mathbb{R}^{p+1}$ [53].

4. **Inner products of regression vectors** $[\beta^{(1)}]^\mathsf{T}\beta^{(2)}$ **and** $[\beta^{(1)}]^\mathsf{T} A \beta^{(2)}$ with the weighting matrix $A \in \mathbb{R}^{(p+1) \times (p+1)}$ in two-sample setting. The inner product of regression vectors or its weighted version measures the similarity between the two regression vectors. In genetic studies, such inner products can be used as the genetic relatedness measure when the covariates are genetic variants, and outcome variables are different phenotypes [25, 28].

We examine statistical inference procedures for linear and quadratic functionals from both methodological and theoretical perspectives. We also discuss the optimality results for the corresponding estimation and inference problems.

4   *High-dimensional Regression*

A user-friendly R package SIHR [54] has been developed to implement the statistical inference methods for the low-dimensional objectives mentioned above. This package provides a convenient way to apply the discussed statistical inference methods.

Beyond the aforementioned inference for a single coordinate of the regression vector or other one-dimensional functionals, we also discuss the simultaneous inference of high-dimensional regression models. This includes using global methods with maximum-type statistics to test the entire regression coefficient vector [55–57], as well as component-wise simultaneous inference methods that control the false discovery rate (FDR). Specifically, we examine the one-sample testing of high-dimensional linear regression coefficients [58], the comparison of two high-dimensional linear regression models [59], and the joint testing of regression coefficients across multiple responses [60]. We also extend our discussion to logistic regression models. We discuss these large-scale multiple testing problems focusing on controlling the asymptotic FDR.

While error rate control is important for simultaneous inference, statistical power is also crucial. However, many existing testing methods for high-dimensional linear models do not consider auxiliary information, such as model sparsity and heteroscedasticity, that could improve statistical power. While there has been a significant amount of research on methods to enhance power in multiple testing in general [61–69, among many others], recent efforts have also focused on simultaneous inference methods that incorporate auxiliary information to assist power improvement of high-dimensional regression analysis. For example, [60] achieved power gains by leveraging similarities across multivariate responses; [70] explored the sparsity information hidden in the data structures and improved the power through p-value weighting mechanisms; [71] obtained power enhancement through integrating heterogeneous linear models. In the current paper, we primarily focus on power enhancement in a two-sample inference setting where the high-dimensional objects of interest are individually sparse. We will discuss methods for controlling FDR with and without power enhancement and the related theoretical aspects.

The rest of the paper is organized as follows. We finish this section by introducing the notation. Section 2 discusses the debiased inference idea for the regression coefficients, and Section 3 presents the debiased methods for linear and quadratic functionals in one- and two-sample settings. Section 4 focuses on simultaneous inference for high-dimensional regression models. We conclude the paper by discussing other related works in Section 5.

Notation. For an event E, denote by $\mathbb{I}\{E\}$ its indicator function. For an index set $J \subset \{1, 2, \cdots, p\}$ and a vector $x \in \mathbb{R}^p$, x_J is the sub-vector of x with indices in J and x_{-J} is the sub-vector with indices in J^c. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = \left(\sum_{l=1}^p |x_l|^q \right)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1 \leq l \leq p} |x_l|$. Denote by e_j the jth canonical unit vector and $I_p \in \mathbb{R}^{n \times p}$ the identity matrix. For a symmetric matrix A, denote by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ its maximum and minimum eigenvalues, respectively. For a matrix

$A \in \mathbb{R}^{n \times p}$, $A_{\cdot,j}$ and $A_{i,\cdot}$ respectively denote the jth column and i^{th} row of A, $A_{i,j}$ denotes the $(i,j)^{\text{th}}$ entry of A, $A_{i,-j}$ denotes the ith row of A with its jth entry removed, $A_{-i,j}$ denotes the jth column of A with its ith entry removed, $A_{i,-\{j_1,j_2\}}$ denotes the ith row of A with its j_1^{th} and j_2^{th} entries both removed and $A_{-i,-j} \in \mathbb{R}^{(n-1)\times(p-1)}$ denotes the submatrix of A with its ith row and j^{th} column removed. We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all n; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\overline{\lim}_{n\to\infty} a_n/b_n = 0$. Let $o_{\mathbb{P}}\{a_n\}$ and $O_{\mathbb{P}}\{a_n\}$ respectively represent the sequences that grow in a smaller and equal/smaller rate of the sequence a_n with probability approaching 1 as $n \to \infty$.

2 Inference for Regression Coefficients

In this section, we begin with a discussion in Section 2.1 on several commonly used penalized estimators for the high-dimensional GLMs. We review in Section 2.2 the debiased methods for linear models introduced in [21–23] and discuss its extensions to high-dimensional logistic regression in Section 2.3. We also present the optimality of the confidence interval construction in both linear and logistic high-dimensional regression models.

2.1 Estimation in high-dimensional regression

For the high-dimensional linear model (1), a commonly used estimator of the regression vector β is the Lasso estimator [1], defined as

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^{p+1}} \frac{\|Y - X\beta\|_2^2}{2n} + \lambda_0 \sum_{j=2}^{p+1} \frac{\|X_{\cdot,j}\|_2}{\sqrt{n}} |\beta_j|, \tag{3}$$

with the tuning parameter $\lambda_0 = A\sigma_\epsilon\sqrt{\log p/n}$ for some positive constant $A > 2$. In the penalized regression (3), we do not penalize the intercept β_1. The tuning parameter λ_0 is typically chosen by cross-validation, as implemented in the R package `glmnet` [18]. With the Lasso estimator $\widehat{\beta}$, the variance σ_ϵ^2 can be estimated by $\widehat{\sigma}_\epsilon^2 = \frac{1}{n}\|Y - X\widehat{\beta}\|_2^2$.

The tuning parameter λ_0 in (3) depends on the noise level σ_ϵ. Alternative estimators have been proposed such that the tuning parameter does not depend on the unknown noise [12, 13, e.g.]. Particularly, [13] proposed the scaled Lasso estimator

$$\{\widehat{\beta}, \widehat{\sigma}_\epsilon\} = \arg\min_{\beta \in \mathbb{R}^{p+1}, \sigma_\epsilon \in \mathbb{R}^+} \frac{\|y - X\beta\|_2^2}{2n\sigma_\epsilon} + \frac{\sigma_\epsilon}{2} + \lambda_0 \sum_{j=2}^{p+1} \frac{\|X_{\cdot,j}\|_2}{\sqrt{n}} |\beta_j| \tag{4}$$

with the tuning parameter $\lambda_0 = A\sqrt{\log p/n}$ for some positive constant $A > 2$. Beyond the estimators mentioned above, a wide collection of estimators of high-dimensional regression vectors have been proposed [e.g., 3, 4, 6, 7, 9–12].

For the high-dimensional logistic model (1), the penalized methods have also been well developed to estimate $\beta \in \mathbb{R}^{p+1}$ [e.g., 6, 7, 9, 14–16]. In this

paper, we use the penalized log-likelihood estimator $\widehat{\beta}$ in [6], defined as

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{k=1}^{n} \{\log[1 + \exp(X_{k,\cdot}^{\mathsf{T}}\beta)] - Y_k(X_{k,\cdot}^{\mathsf{T}}\beta)\} + \lambda_0 \sum_{j=2}^{p+1} \frac{\|X_{\cdot,j}\|_2}{\sqrt{n}} |\beta_j|, \quad (5)$$

with a positive tuning parameter $\lambda_0 \asymp \sqrt{\log p / n}$.

2.2 Debiased or desparsified estimators in linear models

The penalized estimators introduced in Section 2.1 have been shown to achieve the optimal convergence rate and satisfy desirable variable selection properties [5, 72–74]. However, [21–23] highlighted that the Lasso and other penalized estimators are not ready for statistical inference due to the non-negligible estimation bias. They further proposed correcting the penalized estimators' bias and then making inferences based on the bias-corrected estimators.

In the following, we present the main idea of the bias correction method. To illustrate the main idea, we fix the parameter index $2 \leq j \leq p+1$ and focus on the confidence interval construction for β_j in the model (1). With $\widehat{\beta}$ denoting the Lasso estimator in (3), the main idea of the method proposed in [22, 23] is to estimate the error of the plug-in estimator $\widehat{\beta}_j - \beta_j$. The approximation of the error $\widehat{\beta}_j - \beta_j$ can be motivated by the following decomposition: for any vector $u \in \mathbb{R}^{p+1}$,

$$u^{\mathsf{T}} \frac{1}{n} \sum_{k=1}^{n} X_{k,\cdot}(Y_k - X_{k,\cdot}^{\mathsf{T}}\widehat{\beta}) - (\beta_j - \widehat{\beta}_j) = u^{\mathsf{T}} \frac{1}{n} X^{\mathsf{T}}\epsilon + \left(\widehat{\Sigma}u - e_j\right)^{\mathsf{T}} (\beta - \widehat{\beta}), \quad (6)$$

with $\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} X_{k,\cdot} X_{k,\cdot}^{\mathsf{T}}$. We explain how to construct the vector u by balancing the two terms on the right-hand side of the decomposition in (6). The first term on the right-hand side of (6) has the conditional variance $\sigma_\epsilon^2 / n \cdot u^{\mathsf{T}}\widehat{\Sigma}u$ while the second term can be further upper-bounded by

$$\left|\left(\widehat{\Sigma}u - e_j\right)^{\mathsf{T}} (\beta - \widehat{\beta})\right| \leq \|\widehat{\Sigma}u - e_j\|_\infty \|\beta - \widehat{\beta}\|_1. \quad (7)$$

An algorithmic method of constructing u is to constrain the bias and minimize the variance. Particularly, [22, 23] proposed the following construction of the projection direction,

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^{p+1}} u^{\mathsf{T}}\widehat{\Sigma}u \quad \text{subject to} \quad \|\widehat{\Sigma}u - e_j\|_\infty \leq \lambda \quad (8)$$

with $\lambda \asymp \sqrt{\log p / n}$ denoting a positive tuning parameter. The construction in (8) is designed to minimize the conditional variance $u^{\mathsf{T}}\widehat{\Sigma}u$ of the "asymp normal" term and constrain $\|\widehat{\Sigma}u - e_j\|_\infty$, which further controls the "remaining bias" term as in (7).

With the projection direction $\widehat{u}$ in (8), [22, 23] introduced the debiased estimator,

$$\widehat{\beta}_j^{\text{Deb}} = \widehat{\beta}_j + \widehat{u}^{\mathsf{T}} \frac{1}{n} \sum_{k=1}^{n} X_{k,\cdot}(Y_k - X_{k,\cdot}^{\mathsf{T}}\widehat{\beta}), \quad \text{for} \quad 2 \leq j \leq p+1. \quad (9)$$

Following from (6), we obtain the following error decomposition of $\widehat{\beta}_j^{\mathrm{Deb}}$,

$$\widehat{\beta}_j^{\mathrm{Deb}} - \beta_j = \widehat{u}^\intercal \frac{1}{n} \sum_{k=1}^n X_{k,.} \epsilon_k + (\widehat{\Sigma}\widehat{u} - e_j)^\intercal (\beta - \widehat{\beta}). \tag{10}$$

The first term on the right-hand side of (10) is asymptotically normal with the asymptotic variance $(\sigma_\epsilon^2/n) \cdot \widehat{u}^\intercal \widehat{\Sigma}\widehat{u}$. Implied by (7), we show that the projection direction $\widehat{u}$ in (8) constrains the second term on the right-hand side of (10) by an upper bound $\lambda\|\beta - \widehat{\beta}\|_1$, with the tuning parameter λ specified in (8). This upper bound can be established to be of rate $\|\beta\|_0 \log p / n$ and the rate of convergence of the estimating error $\widehat{\beta}_j^{\mathrm{Deb}} - \beta_j$ is $\frac{1}{\sqrt{n}} + \|\beta\|_0 \frac{\log p}{n}$, which is shown to be the minimax optimal rate of estimating the regression coefficient β_j [31, 33]. More importantly, [22, 23] have shown that the debiased estimator $\widehat{\beta}_j^{\mathrm{Deb}}$ in (9) is approximately unbiased and asymptotically normal when $\|\beta\|_0 \ll \sqrt{n}/\log p$. Based on the asymptotic normality, [22, 23] constructed the following confidence interval

$$\mathrm{CI} = \left(\widehat{\beta}_j^{\mathrm{Deb}} - z_{\alpha/2}\sqrt{\widehat{V}}, \quad \widehat{\beta}_j^{\mathrm{Deb}} + z_{\alpha/2}\sqrt{\widehat{V}}\right) \quad \text{with} \quad \widehat{V} = \frac{\widehat{\sigma}_\epsilon^2}{n}\widehat{u}^\intercal\widehat{\Sigma}\widehat{u}, \tag{11}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard normal distribution.

Remark 1. In the low-dimensional setting, we may set $\lambda = 0$ as in (8) and obtain $\widehat{u} = \widehat{\Sigma}^{-1}e_j$. This choice of $\widehat{u}$ reduces the debiased estimator in (9) to the OLS estimator. The debiased estimator is also referred to as the "desparsified" estimator [21, 23] since $\widehat{\beta}_j^{\mathrm{Deb}}$ is generally not zero even if the true β_j is zero. Hence, even if β is a sparse vector, the vector $\widehat{\beta}^{\mathrm{Deb}} = (\widehat{\beta}_1^{\mathrm{Deb}}, \widehat{\beta}_2^{\mathrm{Deb}}, \cdots, \widehat{\beta}_{p+1}^{\mathrm{Deb}})^\intercal$ is dense. Consequently, the vector $\widehat{\beta}^{\mathrm{Deb}}$ does not estimate β well in general even though $\widehat{\beta}_j^{\mathrm{Deb}}$ is an optimal estimator of β_j for every $2 \leq j \leq p + 1$.

2.2.1 Optimality of statistical inference

In high-dimensional linear model, the paper [31] established the minimax expected length of the confidence interval over the parameter space

$$\Theta(s) = \{\theta = (\beta, \Sigma, \sigma_\epsilon) : \|\beta\|_0 \leq s, c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0, 0 < \sigma_\epsilon \leq C_1\}, \tag{12}$$

where $C_0 \geq c_0 > 0$ and $C_1 > 0$ are positive constants. The space $\Theta(s)$ contains all regression vectors of less than s non-zero elements. As established in [31], the minimax expected length of confidence intervals over $\Theta(s)$ for the regime $s \lesssim n/\log p$ is $\frac{1}{\sqrt{n}} + s\frac{\log p}{n}$. When $s \lesssim \frac{\sqrt{n}}{\log p}$, the optimal length $1/\sqrt{n}$ can be achieved by the confidence interval in (11). Over the regime $\frac{\sqrt{n}}{\log p} \ll s \lesssim n/\log p$, [31] proposed a confidence interval attaining the optimal rate $s \log p/n$, where the construction requires the prior knowledge of the sparsity level s. We illustrate the minimax expected length in Figure 1.
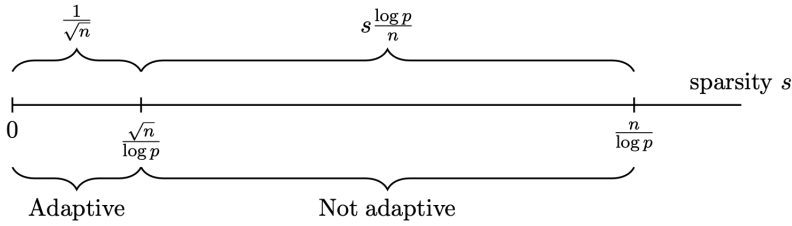
Fig. 1 An illustration of the minimax optimality and adaptivity of the confidence intervals concerning the sparsity s of β for the unknown design setting. On the top of the figure, we report the minimax expected lengths of the confidence intervals. At the bottom of the figure, the possibility of being adaptive to the sparsity s is presented.

More importantly, [29] studied the possibility of constructing a rate-optimal adaptive confidence interval. Here, an adaptive confidence interval means that, even without knowing about the sparsity level s, the length of the constructed confidence interval is automatically adapting to s. Since the sparsity information of the regression vector β is generally unknown, it is desirable to construct an adaptive confidence interval. The work [29] established the following adaptivity results: if the design covariance matrix Σ is unknown, it is possible to construct an adaptive confidence interval only for the ultra-sparse regime $s \lesssim \sqrt{n}/\log p$. That is, if $\sqrt{n}/\log p \ll s \lesssim n/\log p$, it is impossible to construct a rate-optimal confidence interval that is adaptive to the sparsity level. This phase transition about the possibility of constructing adaptive confidence intervals is presented in Figure 1.

The information of Σ is critical for constructing optimal confidence intervals. If Σ is known, the minimax expected length is $1/\sqrt{n}$ over the entire sparse regime $s \lesssim n/\log p$; see [29, 75] for details. This contrasts sharply with the optimality results in Figure 1.

Remark 2. The above results regarding the exact sparse regression vector can be extended to approximate sparse settings. Specifically, the statistical inference and optimality results can be generalized to the capped-ℓ_1 sparsity scenario, which is defined as

$$\sum_{j=1}^{p} \min\{\beta_j/(\lambda_0\sigma), 1\} \le s, \quad \text{with } \lambda_0 = \sqrt{2\log p/n}. \tag{13}$$

Such capped-ℓ_1 condition holds if $\|\beta\|_0 \le s$ or $\|\beta\|_q^q/(\sigma\lambda_0)^q \le s$ for $0 < q \le 1$ [23]. The work [53] has considered the following parameter space that contains the capped-ℓ_1 sparse regression vectors:

$$\Theta'(s) = \{\theta = (\beta, \Sigma, \sigma_\epsilon) : \sum_{j=1}^{p} \min\{\beta_j/(\lambda_0\sigma), 1\} \le s,$$

$$c_0 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_0, 0 < \sigma_\epsilon \le C_1\},$$

and established the same results as in Figure 1.

2.2.2 Another viewpoint: debiasing with decorrelation

In this subsection, we detour to introduce a slightly different view of debiased Lasso estimator proposed in [21, 23], which is of the following form

$$\widetilde{\beta}_j^{\text{Deb}} = \frac{Z_{\cdot,j}^{\mathsf{T}}(Y - X_{\cdot,-j}\widehat{\beta}_{-j})}{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,j}}, \tag{14}$$

where $Z_{\cdot,j} \in \mathbb{R}^n$ is a decorrelation vector to be specified. [21, 23] proposed to construct the vector $Z_{\cdot,j} \in \mathbb{R}^n$ as the residual $Z_{\cdot,j} = X_{\cdot,j} - X_{\cdot,-j}\widehat{\gamma}$, with the Lasso estimator $\widehat{\gamma} = \frac{1}{2n}\|X_{\cdot,j} - X_{\cdot,-j}\gamma\|_2^2 + \lambda_\gamma \sum_{l\neq j} \frac{\|X_{\cdot,l}\|_2}{\sqrt{n}}|\gamma_l|$, where $\lambda_\gamma > 0$ is a positive tuning parameter. The KKT condition ensures that the residual $Z_{\cdot,j} = X_{\cdot,j} - X_{\cdot,-j}\widehat{\gamma}$ is nearly orthogonal to all columns of $X_{\cdot,-j}$, via

$$\frac{1}{n}\|Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,-j}\|_\infty \leq \lambda_\gamma \cdot \max_{l\neq j} \frac{\|X_{\cdot,l}\|_2}{\sqrt{n}}.$$

To see the effectiveness of the estimator in (14), we examine its estimation error

$$\widetilde{\beta}_j^{\text{Deb}} - \beta_j = \frac{Z_{\cdot,j}^{\mathsf{T}}\epsilon}{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,j}} + \frac{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,-j}(\beta_{-j} - \widehat{\beta}_{-j})}{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,j}}. \tag{15}$$

In the above expression, the first term on the right-hand side can be shown to be asymptotically normal under regularity conditions, while the second term is constrained as

$$\left| \frac{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,-j}(\beta_{-j} - \widehat{\beta}_{-j})}{Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,j}} \right| \leq \frac{1}{n}\|Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,-j}\|_\infty \cdot \frac{\|\beta_{-j} - \widehat{\beta}_{-j}\|_1}{\frac{1}{n}|Z_{\cdot,j}^{\mathsf{T}}X_{\cdot,j}|} \leq C\lambda_\gamma\|\widehat{\beta} - \beta\|_1,$$

where the last inequality holds with a high probability for some positive constant $C > 0$. With the above argument, [21, 23] have shown that the first term in the decomposition (15) is the dominating term for the ultra-sparse regime $\|\beta\|_0 \ll \sqrt{n}/\log p$. This leads to the asymptotic normality of the estimator $\widetilde{\beta}_j^{\text{Deb}}$. [21, 23] further constructed the confidence interval based on asymptotic normality. In the current paper, we focus on methods generalizing the debiased estimator in (9), instead of the decorrelation form in (14). However, the decorrelation idea has been extended to handle other statistical inference problems, including high-dimensional generalized linear model [21, 32], gaussian graphical model [32], high-dimensional confounding model [76, 77], and high-dimensional additive model [78].

2.3 Debiasing in binary outcome models

We generalize the debiased estimator in (9) to high-dimensional GLM with a binary outcome. Similarly to the Lasso estimator, the penalized logistic regression estimator $\widehat{\beta}$ in (5) suffers from the bias due to the ℓ_1 penalty. The bias-corrected estimator is proposed as the following generic form,

$$\widehat{\beta}_j^{\text{Deb}} = \widehat{\beta}_j + \widehat{u}^{\mathsf{T}}\frac{1}{n}\sum_{k=1}^n W_k X_{k,\cdot}(Y_k - h(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})), \tag{16}$$

where $W_k \in \mathbb{R}$, for $1 \le k \le n$, are the weights to be specified and $\widehat{u} \in \mathbb{R}^{p+1}$ is the projection direction to be specified. We apply the Taylor expansion of the h function and obtain

$$Y_k - h(X_{k,.}^{\mathsf{T}}\widehat{\beta}) = h(X_{k,.}^{\mathsf{T}}\beta) - h(X_{k,.}^{\mathsf{T}}\widehat{\beta}) + \epsilon_k = h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}^{\mathsf{T}}(\beta - \widehat{\beta}) + R_k + \epsilon_k \quad (17)$$

with the approximation error $R_k = \int_0^1 (1-t)h''(X_{k,.}^{\mathsf{T}}\widehat{\beta} + tX_{k,.}^{\mathsf{T}}(\beta - \widehat{\beta}))dt \cdot (X_{k,.}^{\mathsf{T}}(\widehat{\beta} - \beta))^2$. We plug in the Taylor expansion (17) into the weighted bias-correction estimator in (16), leading to the error decomposition of $\widehat{\beta}_j^{\text{Deb}} - \beta_j$ as

$$\underbrace{\frac{1}{n}\sum_{k=1}^n \widehat{u}^{\mathsf{T}}X_{k,.}W_k\epsilon_k}_{\text{asymp normal}} + \underbrace{\frac{1}{n}\sum_{k=1}^n \left(W_k h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}X_{k,.}^{\mathsf{T}}\widehat{u} - e_j\right)^{\mathsf{T}}(\beta - \widehat{\beta})}_{\text{remaining bias}}$$

$$+ \underbrace{\frac{1}{n}\sum_{k=1}^n \widehat{u}^{\mathsf{T}}X_{k,.}W_k R_k}_{\text{nonlinearity bias}}. \tag{18}$$

In the following, we describe two ways of specifying the weights W_k.

1. **Linearization weighting.** [24] proposed to construct the weight $W_k = 1/h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})$. Then the "remaining bias" term in (18) is reduced to $\frac{1}{n}\sum_{k=1}^n \left(X_{k,.}X_{k,.}^{\mathsf{T}}\widehat{u} - e_j\right)^{\mathsf{T}}(\beta - \widehat{\beta})$, which is the same as the corresponding term in the linear regression. This enables us to directly adopt the projection direction $\widehat{u}$ constructed in (8). This connection reveals the advantage of the weight $W_k = 1/h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})$, that is, the bias-correction developed under the linear regression model can be directly extended to the logistic regression.
2. **Link-specific weighting.** For a general link function $h(\cdot)$, [31] constructed the weight $W_k = h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})/[h(X_{k,.}^{\mathsf{T}}\widehat{\beta})(1-h(X_{k,.}^{\mathsf{T}}\widehat{\beta}))]$. If h is the logistic link, we have $h'(\cdot) = h(\cdot)(1 - h(\cdot))$ and obtain the constant weight $W_k = 1$. Such a link-specific weighting can be generalized to other binary outcome models (e.g., the probit model) with various link functions $h(\cdot)$; see details in [31].

After specifying the weights, the projection direction can be constructed as

$$\widehat{u} = \operatorname*{arg\,min}_{u \in \mathbb{R}^{p+1}} u^{\mathsf{T}}\left(\frac{1}{n}\sum_{k=1}^n W_k h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}X_{k,.}^{\mathsf{T}}\right)u$$

$$\text{subject to} \quad \left\|\left(\frac{1}{n}\sum_{k=1}^n W_k h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}X_{k,.}^{\mathsf{T}}\right)u - e_j\right\|_\infty \le \lambda, \quad \|Xu\|_\infty \le \tau \tag{19}$$

with the positive tuning parameters $\lambda \asymp \sqrt{\log p/n}$ and $\tau \asymp \sqrt{\log n}$. The construction in (19) can be motivated from a similar view as (8). The constraint $\left\|\left(\frac{1}{n}\sum_{k=1}^n W_k h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}X_{k,.}^{\mathsf{T}}\right)u - e_j\right\|_\infty \le \lambda$ is imposed to constrain the "remaining bias" term in (18) and $\|Xu\|_\infty \le \tau$ is imposed to constrain the "nonlinearity bias" in (18). For the linearization weighting, the conditional variance of $\frac{1}{n}\sum_{k=1}^n u^{\mathsf{T}}X_{k,.}W_k\epsilon_k$ is $u^{\mathsf{T}}\left(\frac{1}{n}\sum_{k=1}^n W_k^2 h'(X_{k,.}^{\mathsf{T}}\widehat{\beta})X_{k,.}X_{k,.}^{\mathsf{T}}\right)u$, which

is of the same order as the objective function in (19) for a bounded W_k. So, instead of minimizing the exact variance, we minimize a scaled conditional variance in (19), which has the advantage of leading to almost the same optimization as in (8) for the linear model.

Theoretical properties of the debiased estimators (16) have been established for the logistic outcome model. With the weights $W_k = 1/h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})$ for the linearization weighting, [24] established the asymptotic normality of $\widehat{\beta}_j^{\mathrm{Deb}}$ in (16). [31] established a similar theoretical property for $\widehat{\beta}_j^{\mathrm{Deb}}$ in (16) with the weights $W_k = 1$ for the link-specific weighting. The asymptotic normality results in both works require the ultra-sparse condition $\|\beta\|_0 \ll \sqrt{n}/[\log p \log n]$. The use of the weights $W_k = 1$ in [31] leads to a smaller standard error than using the weights $W_k = 1/h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})$ proposed in [24]. The theoretical justification in [31] requires a sample splitting such that the initial estimator $\widehat{\beta}$ is constructed from an independent sample. Such sample splitting is not required in the analysis of [24], which is part of the benefit of linearization weighting.

Based on the asymptotic normality, we construct the confidence interval

$$\mathrm{CI} = \left(\widehat{\beta}_j^{\mathrm{Deb}} - z_{\alpha/2}\sqrt{\widehat{V}}, \quad \widehat{\beta}_j^{\mathrm{Deb}} + z_{\alpha/2}\sqrt{\widehat{V}}\right) \quad \text{with} \quad \widehat{V} = \frac{1}{n}\widehat{u}^{\mathsf{T}}\widehat{\Sigma}^G\widehat{u}, \qquad (20)$$

where $\widehat{\Sigma}^G = \frac{1}{n}\sum_{k=1}^n W_k^2 h(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})(1 - h(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta}))X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}$.

The optimality of confidence interval construction in high-dimensional logistic regression was studied in [31]. The minimax expected length and the possible regime of constructing adaptive confidence intervals are similar to those in Figure 1, up to a polynomial order of $\log n$; see the results in [31].

Remark 3 (Assumptions on $X_{k,\cdot}^{\mathsf{T}}\beta$ and Σ^{-1}). To analyze the inference properties on high-dimensional GLMs, many existing works [e.g., 21, 24, 30, 32, 57, 79, 80] have made the assumption of bounded $X_{k,\cdot}^{\mathsf{T}}\beta$ for $1 \leq k \leq n$ with a high probability. This assumption ensures that the function h and its derivatives remain bounded from above and below when evaluated on the observed data $X_{k,\cdot}^{\mathsf{T}}\beta_{1 \leq k \leq n}$. Condition (iv) of Theorem 3.3 in [21] and Condition (A2) in [24] provide examples of such assumptions.

However, this boundedness assumption can be restrictive in practical applications. For binary outcomes, the recent work by [31] relaxes this assumption to $c \leq \mathbb{P}(Y_i = 1) \leq 1 - c$, where $c > 0$ is a small constant. The discussion after Theorem 1 in [31] provides further details on this relaxation. Additionally, in the context of logistic regression, both [24] and [31] have conducted numerical studies to test the robustness of the inference results under violations of the boundedness assumption. However, it remains an open question on how to relax this assumption for Poisson regression.

Another strong assumption made for inference on high-dimensional GLMs is the sparsity condition on the inverse of the Hessian matrix $\mathbb{E}\left(h(X_{k,\cdot}^{\mathsf{T}}\beta)\left(1 - h(X_{k,\cdot}^{\mathsf{T}}\beta)\right)X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}\right)$ or the precision matrix Σ^{-1}. Several works, including [21], [32], and [57], have incorporated this sparsity condition. However, [24] and [31] have employed different debiased methods to relax these sparsity conditions.

3 Linear and Quadratic Functionals Inference

We consider in this section statistical inference for linear and quadratic transformations of the regression vectors under high-dimensional linear and logistic regressions. We investigate both one- and two-sample regression models. In Section 3.7, we discuss the R package SIHR [54] that implements these methods.

3.1 Linear functionals for linear regression

For a given vector $x_{\text{new}} \in \mathbb{R}^{p+1}$, we present the construction of the point estimator and confidence interval for $x_{\text{new}}^{\mathsf{T}}\beta$ under the high-dimensional linear model (1). Similar to inference for β_j, the plug-in estimator $x_{\text{new}}^{\mathsf{T}}\widehat{\beta}$ suffers from the estimation bias by directly plugging in the Lasso estimator $\widehat{\beta}$ in (3). The work [53] proposed the following bias-corrected estimator,

$$\widehat{x_{\text{new}}^{\mathsf{T}}\beta} = x_{\text{new}}^{\mathsf{T}}\widehat{\beta} + \widehat{u}^{\mathsf{T}}\frac{1}{n}\sum_{k=1}^{n} X_{k,\cdot}(Y_k - X_{k,\cdot}^{\mathsf{T}}\widehat{\beta}), \tag{21}$$

with the projection direction $\widehat{u}$ defined as

$$\widehat{u} = \operatorname*{arg\,min}_{u \in \mathbb{R}^{p+1}} u^{\mathsf{T}}\widehat{\Sigma}u \quad \text{subject to} \quad \left\|\widehat{\Sigma}u - x_{\text{new}}\right\|_{\infty} \leq \|x_{\text{new}}\|_2\lambda \tag{22}$$

$$\left|x_{\text{new}}^{\mathsf{T}}\widehat{\Sigma}u - \|x_{\text{new}}\|_2^2\right| \leq \|x_{\text{new}}\|_2^2\lambda, \tag{23}$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}$ and $\lambda \asymp \sqrt{\log p/n}$ is a positive tuning parameter.

The debiased estimator in (21) satisfies the following error decomposition,

$$\widehat{x_{\text{new}}^{\mathsf{T}}\beta} - x_{\text{new}}^{\mathsf{T}}\beta = \underbrace{\widehat{u}^{\mathsf{T}}\frac{1}{n}X^{\mathsf{T}}\epsilon}_{\text{asymp normal}} + \underbrace{\left(\widehat{\Sigma}\widehat{u} - x_{\text{new}}\right)^{\mathsf{T}}(\beta - \widehat{\beta})}_{\text{remaining bias}}. \tag{24}$$

The construction in (22), without the additional constraint (23), can be viewed as a direct generalization of (8) by replacing e_j with the general loading x_{new}. Specifically, (22) minimizes the conditional variance of the "asymp normal" term in (24) and provides a control of the "remaining bias" term by the inequality $|(\widehat{\Sigma}u - x_{\text{new}})^{\mathsf{T}}(\beta - \widehat{\beta})| \leq \|\widehat{\Sigma}u - x_{\text{new}}\|_{\infty}\|\beta - \widehat{\beta}\|_1$. Such a construction of the projection direction for linear functional has been proposed in [29, 30]. However, such a direct generalization is not universally effective for all loadings. As established in Proposition 2 in [53], the projection direction, constructed without the additional constraint (23), does not correct the bias for a wide class of dense loading vectors.

We shall emphasize that the new constraint in (23) is crucial to ensuring the asymptotic normality of $\widehat{x_{\text{new}}^{\mathsf{T}}\beta} - x_{\text{new}}^{\mathsf{T}}\beta$ for any loading vector x_{new}. This constraint is imposed such that the variance of the "asymp normal" term in (24) always dominates the "remaining bias" term in (24). With this additional constraint, the projection direction $\widehat{u}$ constructed in (22) and (23) enables the effective bias correction for any loading vector x_{new}, no matter it is sparse or dense. The work [53] established the asymptotic normality of the estimator

$\widehat{x_{\text{new}}^{\mathsf{T}}\beta}$ in (21) for any loading vector $x_{\text{new}} \in \mathbb{R}^{p+1}$. Based on the asymptotic normality, we construct a confidence interval for $x_{\text{new}}^{\mathsf{T}}\beta$ as

$$\text{CI} = \left(\widehat{x_{\text{new}}^{\mathsf{T}}\beta} - z_{\alpha/2}\sqrt{\widehat{V}}, \quad \widehat{x_{\text{new}}^{\mathsf{T}}\beta} + z_{\alpha/2}\sqrt{\widehat{V}} \right) \quad \text{with} \quad \widehat{V} = \frac{\widehat{\sigma}_{\epsilon}^2}{n}\widehat{u}^{\mathsf{T}}\widehat{\Sigma}\widehat{u}. \tag{25}$$

Remark 4 (Choices of $\widehat{\beta}$)**.** Throughout the paper, we adopt the Lasso estimator in Equation (3) as the default initial estimator. However, the debiasing approach is compatible with a wide range of initial estimators. In general, the results we present hold if the Lasso estimator is replaced by any estimator $\widehat{\beta}$ that satisfies the condition

$$\|\widehat{\beta} - \beta\|_1 \leq Cs\sqrt{\log p/n}. \tag{26}$$

with a high probability. Condition (B1) in [53] provides an example of such a condition. It is worth noting that various initial estimators $\widehat{\beta}$ satisfy Equation (26), even in cases where β is approximately sparse, such as the capped-ℓ_1 sparsity in Equation (13) [e.g., 5, 6, 12, 13].

3.2 Linear functionals for logistic regression

We now consider the high-dimensional logistic model and present the inference procedures for $x_{\text{new}}^{\mathsf{T}}\beta$ or $h(x_{\text{new}}^{\mathsf{T}}\beta)$ proposed in [24]. In particular, [24] proposed the following debiased estimator,

$$\widehat{x_{\text{new}}^{\mathsf{T}}\beta} = x_{\text{new}}^{\mathsf{T}}\widehat{\beta} + \widehat{u}^{\mathsf{T}}\frac{1}{n}\sum_{k=1}^{n} W_k X_{k,\cdot}\left(Y_k - h(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})\right), \tag{27}$$

where $\widehat{\beta}$ is defined in (5), $W_k = 1/h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})$ for $1 \leq k \leq n$, and the projection direction $\widehat{u} \in \mathbb{R}^{p+1}$ is defined as

$$\widehat{u} = \underset{u \in \mathbb{R}^{p+1}}{\arg\min}\, u^{\mathsf{T}}\left(\frac{1}{n}\sum_{k=1}^{n} W_k h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}\right)u$$

$$\text{subject to} \quad \left\|\left(\frac{1}{n}\sum_{k=1}^{n} W_k h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}\right)u - x_{\text{new}}\right\|_{\infty} \leq \|x_{\text{new}}\|_2\lambda, \quad \|Xu\|_{\infty} \leq \tau$$

$$\left|x_{\text{new}}^{\mathsf{T}}\left(\frac{1}{n}\sum_{k=1}^{n} W_k h'(X_{k,\cdot}^{\mathsf{T}}\widehat{\beta})X_{k,\cdot}X_{k,\cdot}^{\mathsf{T}}\right)u - \|x_{\text{new}}\|_2^2\right| \leq \|x_{\text{new}}\|_2^2\lambda. \tag{28}$$

The bias-corrected estimator in (27) can be viewed as a generalization of those in (22) and (23) by incorporating the weighted bias-correction in (16).

It has been established in [24] that $\widehat{x_{\text{new}}^{\mathsf{T}}\beta}$ in (27) is asymptotically unbiased and normal. Based on the asymptotic normality, we construct the following confidence interval

$$\text{CI} = \left(\widehat{x_{\text{new}}^{\mathsf{T}}\beta} - z_{\alpha/2}\sqrt{\widehat{V}}, \quad \widehat{x_{\text{new}}^{\mathsf{T}}\beta} + z_{\alpha/2}\sqrt{\widehat{V}} \right) \quad \text{with} \quad \widehat{V} = \frac{1}{n}\widehat{u}^{\mathsf{T}}\widehat{\Sigma}^G\widehat{u}, \tag{29}$$

where $\widehat{\Sigma}^G = \frac{1}{n} \sum_{k=1}^{n} W_k^2 h(X_{k,\cdot}^{\mathsf{T}} \widehat{\beta})(1 - h(X_{k,\cdot}^{\mathsf{T}} \widehat{\beta})) X_{k,\cdot} X_{k,\cdot}^{\mathsf{T}}$. We estimate the case probability $\mathbb{P}(Y_k = 1 | X_{k,\cdot} = x_{\text{new}})$ by $h(\widehat{x_{\text{new}}^{\mathsf{T}}\beta})$ and construct the confidence interval for $h(x_{\text{new}}^{\mathsf{T}}\beta)$ as

$$\text{CI} = \left[h\left(\widehat{x_{\text{new}}^{\mathsf{T}}\beta} - z_{\alpha/2}\sqrt{\widehat{V}} \right), h\left(\widehat{x_{\text{new}}^{\mathsf{T}}\beta} + z_{\alpha/2}\sqrt{\widehat{V}} \right) \right]. \tag{30}$$

3.3 Conditional average treatment effects

The inference methods proposed in Sections 3.1 and 3.2 can be generalized to make inferences for conditional average treatment effects, which can be expressed as the difference between two linear functionals. For $1 \leq k \leq n$, let $A_k \in \{1, 2\}$ denote the treatment assignment for the kth observation, where $A_k = 1$ and $A_k = 2$ represent the subject receiving the control or the treatment assignment, respectively. Moreover, in the context of comparing the treatment effectiveness, $A_k = 1$, and $A_k = 2$ may stand for the subject receiving the first or the second treatment assignment, respectively. As a special case of (2), we consider the following conditional outcome models $\mathbb{E}(Y_k | X_{k,\cdot}, A_k = 1) = X_{k,\cdot}^{\mathsf{T}} \beta^{(1)}$ and $\mathbb{E}(Y_k | X_{k,\cdot}, A_k = 2) = X_{k,\cdot}^{\mathsf{T}} \beta^{(2)}$. For an individual with $X_{k,\cdot} = x_{\text{new}}$, we define $\Delta(x_{\text{new}}) = \mathbb{E}(Y_k | X_{k,\cdot} = x_{\text{new}}, A_k = 2) - \mathbb{E}(Y_k | X_{k,\cdot} = x_{\text{new}}, A_k = 1) = x_{\text{new}}^{\mathsf{T}}(\beta^{(2)} - \beta^{(1)})$, which measures the change of the conditional mean from being untreated to treated for individuals with the covariates x_{new}.

By generalizing (21), we construct the debiased estimators $\widehat{x_{\text{new}}^{\mathsf{T}}\beta^{(1)}}$ and $\widehat{x_{\text{new}}^{\mathsf{T}}\beta^{(2)}}$, together with their corresponding variance estimators $\widehat{V}_{\beta^{(1)}}$ and $\widehat{V}_{\beta^{(2)}}$. The paper [53] proposed to estimate $\Delta(x_{\text{new}})$ by $\widehat{\Delta}(x_{\text{new}}) = \widehat{x_{\text{new}}^{\mathsf{T}}\beta^{(2)}} - \widehat{x_{\text{new}}^{\mathsf{T}}\beta^{(1)}}$ and construct the confidence interval for $\Delta(x_{\text{new}})$ as

$$\text{CI} = \left(\widehat{\Delta}(x_{\text{new}}) - z_{\frac{\alpha}{2}}\sqrt{\widehat{V}_{\beta^{(1)}} + \widehat{V}_{\beta^{(2)}}}, \quad \widehat{\Delta}(x_{\text{new}}) + z_{\frac{\alpha}{2}}\sqrt{\widehat{V}_{\beta^{(1)}} + \widehat{V}_{\beta^{(2)}}} \right). \tag{31}$$

Regarding the hypothesis testing problem $H_0 : x_{\text{new}}^{\mathsf{T}}(\beta^{(2)} - \beta^{(1)}) \leq 0$ versus $H_1 : x_{\text{new}}^{\mathsf{T}}(\beta^{(2)} - \beta^{(1)}) > 0$, the paper [53] proposed the following test procedure,

$$\phi_\alpha(x_{\text{new}}) = \mathbb{I}\left\{ \widehat{\Delta}(x_{\text{new}}) - z_\alpha \sqrt{\widehat{V}_{\beta^{(1)}} + \widehat{V}_{\beta^{(2)}}} \geq 0 \right\}. \tag{32}$$

As a direct generalization, we may consider the logistic version, $\mathbb{E}(Y_k | X_{k,\cdot}, A_k = 1) = h(X_{k,\cdot}^{\mathsf{T}} \beta^{(1)})$ and $\mathbb{E}(Y_k | X_{k,\cdot}, A_k = 2) = h(X_{k,\cdot}^{\mathsf{T}} \beta^{(2)})$. For an individual with $X_{k,\cdot} = x_{\text{new}}$, our inference target becomes $\Delta(x_{\text{new}}) = h(x_{\text{new}}^{\mathsf{T}}\beta^{(2)}) - h(x_{\text{new}}^{\mathsf{T}}\beta^{(1)})$. The methods in Section 3.2 can be applied here to make inferences for $\Delta(x_{\text{new}})$.

3.4 Quadratic functionals

We now focus on inference for the quadratic functionals $Q_A = \beta_G^{\mathsf{T}} A \beta_G$ and $Q_\Sigma = \beta_G^{\mathsf{T}} \Sigma_{G,G} \beta_G$, where $G \subset \{1, \cdots, p+1\}$ and $A \in \mathbb{R}^{|G| \times |G|}$ denotes a pre-specified matrix. Without loss of generality, we set $G = \{2, \cdots, |G| + 1\}$. In the following, we mainly discuss the main idea under the high-dimensional linear regression, which can be generalized to the high-dimensional logistic regression. Let $\widehat{\beta}$ denote the Lasso estimator in (3).

We start with the error decomposition of the plug-in estimator $\widehat{\beta}_G^\mathsf{T} A \widehat{\beta}_G$,

$$\widehat{\beta}_G^\mathsf{T} A \widehat{\beta}_G - \beta_G^\mathsf{T} A \beta_G = 2\widehat{\beta}_G^\mathsf{T} A(\widehat{\beta}_G - \beta_G) - (\widehat{\beta}_G - \beta_G)^\mathsf{T} A(\widehat{\beta}_G - \beta_G). \tag{33}$$

In consideration of high-dimensional linear models, [25, 26] proposed to construct the bias-corrected estimator through estimating the error component $2\widehat{\beta}_G^\mathsf{T} A(\widehat{\beta}_G - \beta_G)$ on the right-hand side of (33). Since $\widehat{\beta}_G^\mathsf{T} A(\widehat{\beta}_G - \beta_G)$ can be expressed as $x_{\text{new}}^\mathsf{T}(\widehat{\beta} - \beta)$ with $x_{\text{new}} = \left(0 \ \ \widehat{\beta}_G^\mathsf{T} A \ \ \mathbf{0}^\mathsf{T} \right)^\mathsf{T}$, the techniques of estimating the error component for the linear functional can be directly applied to approximate $\widehat{\beta}_G^\mathsf{T} A(\widehat{\beta}_G - \beta_G)$. Particularly, [25, 26] proposed the following estimator of Q_A,

$$\widehat{Q}_A = \widehat{\beta}_G^\mathsf{T} A \widehat{\beta}_G + 2\widehat{u}_A^\mathsf{T} X^\mathsf{T}(Y - X\widehat{\beta})/n, \tag{34}$$

where $\widehat{u}_A$ denotes the solution of (22) and (23) with $x_{\text{new}} = \left(0 \ \ \widehat{\beta}_G^\mathsf{T} A \ \ \mathbf{0}^\mathsf{T} \right)^\mathsf{T}$. No bias correction is required for the last term on the right-hand side of (33) since it is a higher-order error term under regular conditions.

We now turn to the estimation of $Q_\Sigma = \beta_G^\mathsf{T} \Sigma_{G,G} \beta_G$, where the matrix $\Sigma_{G,G}$ has to be estimated from the data. With $\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^n X_{k,\cdot} X_{k,\cdot}^\mathsf{T}$, we decompose the estimation error of the plug-in estimator $\widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G} \widehat{\beta}_G$ as

$$\begin{aligned}\widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G} \widehat{\beta}_G - \beta_G^\mathsf{T} \Sigma_{G,G} \beta_G =& 2\widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G}(\widehat{\beta}_G - \beta_G) + \beta_G^\mathsf{T}(\widehat{\Sigma}_{G,G} - \Sigma_{G,G})\beta_G \\ & - (\widehat{\beta}_G - \beta_G)^\mathsf{T} \widehat{\Sigma}_{G,G}(\widehat{\beta}_G - \beta_G).\end{aligned} \tag{35}$$

By a similar approach as (34), [26] proposed the following estimator of Q_Σ,

$$\widehat{Q}_\Sigma = \widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G} \widehat{\beta}_G + 2\widehat{u}_\Sigma^\mathsf{T} X^\mathsf{T}(Y - X\widehat{\beta})/n, \tag{36}$$

where $\widehat{u}_\Sigma$ denotes the solution of (22) and (23) with $x_{\text{new}} = \left(0 \ \ \widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G} \ \ \mathbf{0}^\mathsf{T} \right)^\mathsf{T}$. As a special case, [27] considered $G = \{1, \cdots, p+1\}$ and constructed $\widehat{u}_\Sigma = \widehat{\beta}$.

The works [25, 26] established the asymptotic properties for a sample-splitting version of $\widehat{Q}_A$ and $\widehat{Q}_\Sigma$, where the initial estimator $\widehat{\beta}$ is constructed from a subsample independent from the samples $\{X_{k,\cdot}, Y_k\}_{1 \le k \le n}$ used in (34) and (36). Based on the asymptotic properties, [26] estimated the variance of $\widehat{Q}_A$ by $\widehat{V}_A(\tau) = 4\widehat{\sigma}_\epsilon^2 \widehat{u}_A^\mathsf{T} \widehat{\Sigma} \widehat{u}_A/n + \tau/n$ for some positive constant $\tau > 0$, where the term τ/n is introduced as an upper bound for the term $(\widehat{\beta}_G - \beta_G)^\mathsf{T} A(\widehat{\beta}_G - \beta_G)$ in (33). [26] estimated the variance of $\widehat{Q}_\Sigma$ by

$$\widehat{V}_\Sigma(\tau) = \frac{4\widehat{\sigma}_\epsilon^2}{n} \widehat{u}_\Sigma^\mathsf{T} \widehat{\Sigma} \widehat{u}_\Sigma + \frac{1}{n^2} \sum_{k=1}^n \left(\widehat{\beta}_G^\mathsf{T} X_{k,G} X_{k,G}^\mathsf{T} \widehat{\beta}_G - \widehat{\beta}_G^\mathsf{T} \widehat{\Sigma}_{G,G} \widehat{\beta}_G \right)^2 + \frac{\tau}{n},$$

and further constructed the following confidence intervals for Q_A and Q_Σ,

$$\begin{aligned}\text{CI}_A(\tau) &= \left(\widehat{Q}_A - z_{\alpha/2} \cdot \sqrt{\widehat{V}_A(\tau)}, \ \widehat{Q}_A + z_{\alpha/2} \cdot \sqrt{\widehat{V}_A(\tau)} \right), \\ \text{CI}_\Sigma(\tau) &= \left(\widehat{Q}_\Sigma - z_{\alpha/2} \cdot \sqrt{\widehat{V}_\Sigma(\tau)}, \ \widehat{Q}_\Sigma + z_{\alpha/2} \cdot \sqrt{\widehat{V}_\Sigma(\tau)} \right).\end{aligned} \tag{37}$$

For a positive definite matrix A, the significance test $H_0 : \beta_G = 0$ can be recast as $H_{0,A} : \beta_G^\intercal A \beta_G = 0$ or $H_{0,\Sigma} : \beta_G^\intercal \Sigma_{G,G} \beta_G = 0$. The following α-level significance tests of $H_{0,A}$ and $H_{0,\Sigma}$ have been respectively proposed in [26],

$$\phi_A(\tau) = \mathbb{I}\left\{\widehat{Q}_A \geq z_\alpha \cdot \sqrt{\widehat{V}_A(\tau)}\right\}, \quad \phi_\Sigma(\tau) = \mathbb{I}\left\{\widehat{Q}_\Sigma \geq z_\alpha \cdot \sqrt{\widehat{V}_\Sigma(\tau)}\right\}. \qquad (38)$$

The inference for Q_A and Q_Σ can be generalized to the high-dimensional logistic regression together with the weighted bias correction detailed in (28). The paper [28] studied how to construct the debiased estimators of Q_Σ under the high-dimensional logistic regression.

3.5 Semi-supervised inference

We summarize the optimal estimation of $\|\beta\|_2^2$ and $\beta^\intercal \Sigma \beta$ in a general semi-supervised setting, with the labelled data $(X_{1,\cdot}, Y_1), \cdots, (X_{n,\cdot}, Y_n)$ and the unlabelled data $X_{n+1,\cdot}, \cdots, X_{n+N,\cdot}$, where the covariates $X_{1,\cdot}, \cdots, X_{n+N,\cdot}$ are assumed to be identically distributed.

The work [27] proposed the following semi-supervised estimator of $\beta^\intercal \Sigma \beta$,

$$\widehat{Q}(\widehat{\beta}, \widehat{\Sigma}^S) = \widehat{\beta}^\intercal \widehat{\Sigma}^S \widehat{\beta} + 2\widehat{\beta}^\intercal \frac{1}{n} \sum_{k=1}^{n} X_{k,\cdot}(Y_k - X_{k,\cdot}^\intercal \widehat{\beta}), \widehat{\Sigma}^S = \frac{1}{n+N} \sum_{k=1}^{n+N} X_{k,\cdot} X_{k,\cdot}^\intercal \quad (39)$$

The matrix estimator $\widehat{\Sigma}^S$ in (39) utilizes both the labeled and unlabelled data. The work [27] constructed the following confidence interval for $\beta^\intercal \Sigma \beta$ in the semi-supervised setting,

$$\mathrm{CI}(Z) = \left(\left(\widehat{Q}(\widehat{\beta}, \widehat{\Sigma}^S) - z_{\alpha/2}\sqrt{\widehat{V}}\right)_+, \ \widehat{Q}(\widehat{\beta}, \widehat{\Sigma}^S) - z_{\alpha/2}\sqrt{\widehat{V}}\right),$$

where $\widehat{V} = \frac{1}{n}\left(4\widehat{\sigma}_\epsilon^2 \widehat{\beta}^\intercal \widehat{\Sigma}^S \widehat{\beta} + \frac{\widehat{\rho}}{n+N} \sum_{k=1}^{n+N} \left(\widehat{\beta}^\intercal X_{k,\cdot} X_{k,\cdot}^\intercal \widehat{\beta} - \widehat{\beta}^\intercal \widehat{\Sigma}^S \widehat{\beta}\right)^2 + \tau\right)$ with $\widehat{\rho} = n/(N+n)$ and $\tau > 0$ being a user-specific tuning parameter adjusting for the higher order estimation error. The above confidence interval construction demonstrates the usefulness of integrating the unlabelled data, significantly reducing the ratio $\widehat{\rho}$ and the interval length.

Define $\Theta(s, M) = \left\{(\beta, \Sigma, \sigma_\epsilon) \in \Theta(s) : \frac{M}{2} \leq \|\beta\|_2 \leq M\right\}$, as a subspace of $\Theta(s)$ in (12) with the constraint on $\|\beta\|_2$. In the semi-supervised setting, it was established in [27] that the minimax rate for estimating $\beta^\intercal \Sigma \beta$ over $\Theta(s, M)$ is

$$\frac{M^2}{\sqrt{N+n}} + \min\left\{\frac{M}{\sqrt{n}} + \frac{s\log p}{n}, M^2\right\}. \qquad (40)$$

The estimator $\widehat{Q}(\widehat{\beta}, \widehat{\Sigma}^S)$ in (39) is shown to achieve the optimal rate in (40) for $M \geq C\sqrt{s\log p/n}$. When $M \leq C\sqrt{s\log p/n}$, then estimating Q by zero achieves the optimal rate in (40). The optimal convergence rate in (39) characterizes its dependence on the amount of unlabelled data. With a larger size N of the unlabelled data, the convergence rate $\frac{M^2}{\sqrt{N+n}}$ decreases. In the semi-supervised setting, the unlabelled data can also be used to construct a more accurate estimator of $\|\beta\|_2^2$; see Section 4 of [27] for details.

In the following, we consider the two extremes under semi-supervised learning: the supervised learning setting with $N = 0$ and the other extreme setting of knowing $\Sigma = I$, which can be viewed as a special case of the semi-supervised setting with $N \to \infty$. In Table 1, we summarize the optimal convergence rate of estimating $\|\beta\|_2^2$ and $\beta^\intercal \Sigma \beta$ over both $\Theta(s, M)$ and $\Theta_0(s, M)$, where $\Theta_0(s, M) = \{(\beta, \Sigma, \sigma_\epsilon) \in \Theta(s, M) : \Sigma = I\}$. With leveraging the information of Σ, Table 1 shows that the optimal rates of estimating $\|\beta\|_2^2$ and $\beta^\intercal \Sigma \beta$ are reduced by $M \frac{s \log p}{n}$ and $M^2 \frac{1}{\sqrt{n}}$, respectively. The improvement can be substantial when the parameter M, characterizing the ℓ_2 norm $\|\beta\|_2$, is relatively large. The optimal rate of estimating $\|\beta\|_2^2$ was established in [82] under the sequence model $Y_j = \beta_j + \frac{1}{\sqrt{n}} \epsilon_j$ for $1 \leq j \leq p$. When we know $\Sigma = I$ and focus on the regime $s \leq c \min\{n/\log p, p^\nu\}$ for $0 \leq \nu < 1/2$, the optimal rate of estimating $\|\beta\|_2^2$ in the high-dimensional linear model matches that in [82]. However, when Σ is unknown, estimating $\|\beta\|_2^2$ is much harder in the high-dimensional linear model than in the sequence model.

Target	Optimal Rate over $\Theta(s, M)$	Optimal Rate over $\Theta_0(s, M)$
$\|\beta\|_2^2$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{s\log p}{n} + M\frac{s\log p}{n}, M^2\right\}$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{s\log p}{n}, M^2\right\}$
$\beta^\intercal \Sigma \beta$	$\min\left\{M\frac{1}{\sqrt{n}} + \frac{s\log p}{n} + M^2\frac{1}{\sqrt{n}}, M^2\right\}$	

Table 1 The minimax optimal rate of estimating $\|\beta\|_2^2$ over $\Theta(s, M)$ was established in [25]; the remaining optimal rates are implied by (40), which was established in [27]. The focus is on the regime $s \leq c \min\{n/\log p, p^\nu\}$ for $0 \leq \nu < 1/2$.

3.6 Inner products of regression vectors

We next consider the two-sample regression models in (2). When the high-dimensional covariates are genetic variants and the outcome variables measure different phenotypes, then the inner product $[\beta^{(1)}]^\intercal \beta^{(2)}$ can be interpreted as the genetic relatedness [25], measuring the similarity between the two association vectors $\beta^{(1)}$ and $\beta^{(2)}$. We present the debiased estimator of $[\beta^{(1)}]^\intercal \beta^{(2)}$ proposed in [25]. For $d = 1, 2$, let $\widehat{\beta}^{(d)}$ be the Lasso estimator of $\beta^{(d)}$. Similar to the decomposition in (33), we decompose the error of the plug-in estimator $[\widehat{\beta}^{(1)}]^\intercal \widehat{\beta}^{(2)}$,

$$\beta \widehat{\beta}^{(1)} t^\intercal \widehat{\beta}^{(2)} - [\beta^{(1)}]^\intercal \beta^{(2)} = [\widehat{\beta}^{(1)}]^\intercal (\widehat{\beta}^{(2)} - \beta^{(2)}) + [\widehat{\beta}^{(2)}]^\intercal (\widehat{\beta}^{(1)} - \beta^{(1)}) \qquad (41)$$
$$- (\widehat{\beta}^{(2)} - \beta^{(2)})^\intercal (\widehat{\beta}^{(1)} - \beta^{(1)}).$$

The key is to estimate $[\widehat{\beta}^{(1)}]^\intercal (\widehat{\beta}^{(2)} - \beta^{(2)})$ and $[\widehat{\beta}^{(2)}]^\intercal (\widehat{\beta}^{(1)} - \beta^{(1)})$ separately in the above decomposition, which can be viewed as a projection of $\widehat{\beta}^{(2)} - \beta^{(2)}$ and $\widehat{\beta}^{(1)} - \beta^{(1)}$, respectively.

The work [25] proposed the following bias-corrected estimator of $[\beta^{(1)}]^{\mathsf{T}}\beta^{(2)}$,

$$
\begin{aligned}
\widehat{[\beta^{(1)}]^{\mathsf{T}}\beta^{(2)}} =& [\widehat{\beta}^{(1)}]^{\mathsf{T}}\widehat{\beta}^{(2)} + \widehat{u}_1^{\mathsf{T}}\frac{1}{n_2}\sum_{k=1}^{n_2}[X_{k,\cdot}^{(2)}]^{\mathsf{T}}(Y_k^{(2)} - [X_{k,\cdot}^{(2)}]^{\mathsf{T}}\widehat{\beta}^{(2)}) \\
&+ \widehat{u}_2^{\mathsf{T}}\frac{1}{n_1}\sum_{k=1}^{n_1}[X_{k,\cdot}^{(1)}]^{\mathsf{T}}(Y_k^{(1)} - [X_{k,\cdot}^{(1)}]^{\mathsf{T}}\widehat{\beta}^{(1)})
\end{aligned}
\tag{42}
$$

where the projection direction vectors are constructed as

$$
\widehat{u}_1 = \operatorname*{arg\,min}_{u\in\mathbb{R}^{p+1}} u^{\mathsf{T}}\widehat{\Sigma}^{(2)}u \quad \text{subject to } \left\|\widehat{\Sigma}^{(2)}u - \widehat{\beta}^{(1)}\right\|_\infty \le \|\widehat{\beta}^{(1)}\|_2\lambda_2
$$

$$
\widehat{u}_2 = \operatorname*{arg\,min}_{u\in\mathbb{R}^{p+1}} u^{\mathsf{T}}\widehat{\Sigma}^{(1)}u \quad \text{subject to } \left\|\widehat{\Sigma}^{(1)}u - \widehat{\beta}^{(2)}\right\|_\infty \le \|\widehat{\beta}^{(2)}\|_2\lambda_1
$$

with $\widehat{\Sigma}^{(d)} = \frac{1}{n}\sum_{k=1}^{n_d} X_{k,\cdot}^{(d)}[X_{k,\cdot}^{(d)}]^{\mathsf{T}}$ and $\lambda_d \asymp \sqrt{\log p/n_d}$ for $d = 1, 2$.

For the estimator (42), the bias-correction terms $\widehat{u}_1^{\mathsf{T}}\frac{1}{n_2}\sum_{k=1}^{n_2}[X_{k,\cdot}^{(2)}]^{\mathsf{T}}(Y_k^{(2)} - [X_{k,\cdot}^{(2)}]^{\mathsf{T}}\widehat{\beta}^{(2)})$ and $\widehat{u}_2^{\mathsf{T}}\frac{1}{n_1}\sum_{k=1}^{n_1}[X_{k,\cdot}^{(1)}]^{\mathsf{T}}(Y_k^{(1)} - [X_{k,\cdot}^{(1)}]^{\mathsf{T}}\widehat{\beta}^{(1)})$ are constructed to estimate $[\widehat{\beta}^{(1)}]^{\mathsf{T}}(\beta^{(2)} - \widehat{\beta}^{(2)})$ and $[\widehat{\beta}^{(2)}]^{\mathsf{T}}(\beta^{(1)} - \widehat{\beta}^{(1)})$, respectively. The construction of the projection directions $\widehat{u}_1$ and $\widehat{u}_2$ can be viewed as extensions of that in (22) with $x_{\text{new}} = \widehat{\beta}^{(1)}$ and $x_{\text{new}} = \widehat{\beta}^{(2)}$, respectively. An additional constraint as in (23) is not needed here since both $x_{\text{new}} = \widehat{\beta}^{(1)}$ and $x_{\text{new}} = \widehat{\beta}^{(2)}$ are sufficiently sparse. The paper [25] has established the convergence rate of the debiased estimator proposed in (42). The analysis can be extended to establish the asymptotic normal distribution of $\widehat{[\beta^{(1)}]^{\mathsf{T}}\beta^{(2)}}$ under suitable conditions.

The quantity $[\beta^{(1)}]^{\mathsf{T}}\Sigma\beta^{(2)}$ is another genetic relatedness measure if $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$ with $\Sigma^{(d)} = \mathbb{E}X_{k,\cdot}^{(d)}[X_{k,\cdot}^{(d)}]^{\mathsf{T}}$ for $d = 1, 2$. We can propose the debiased estimator $\widehat{[\beta^{(1)}]^{\mathsf{T}}\Sigma\beta^{(2)}}$, defined as $[\widehat{\beta}^{(1)}]^{\mathsf{T}}\widehat{\Sigma}\widehat{\beta}^{(2)} + [\widehat{\beta}^{(1)}]^{\mathsf{T}}\frac{1}{n_2}\sum_{k=1}^{n_2}[X_{k,\cdot}^{(2)}]^{\mathsf{T}}(Y_k^{(2)} - [X_{k,\cdot}^{(2)}]^{\mathsf{T}}\widehat{\beta}^{(2)}) + [\widehat{\beta}^{(2)}]^{\mathsf{T}}\frac{1}{n_1}\sum_{k=1}^{n_1}[X_{k,\cdot}^{(1)}]^{\mathsf{T}}(Y_k^{(1)} - [X_{k,\cdot}^{(1)}]^{\mathsf{T}}\widehat{\beta}^{(1)})$. The above results have been extended to the logistic regression models, where the quantity $[\beta^{(1)}]^{\mathsf{T}}\Sigma\beta^{(2)}$ still captures an interpretation of genetic relatedness. The paper [28] has carefully investigated the confidence interval construction for $[\beta^{(1)}]^{\mathsf{T}}\Sigma\beta^{(2)}$ in consideration of several high-dimensional logistic regression models. Moreover, we might need to estimate and make inferences for $[\beta^{(1)}]^{\mathsf{T}}A\beta^{(2)}$ with A denoting a general matrix. The idea in (42) can be generalized to make inference for $[\beta^{(1)}]^{\mathsf{T}}A\beta^{(2)}$. The work [83] applied such a generalized debiased estimator of $[\beta^{(1)}]^{\mathsf{T}}A\beta^{(2)}$ in an intermediate step to determine the optimal aggregation weight of multiple regression models.

3.7 R package SIHR

The methods reviewed in Sections 2 and 3 have been implemented in the R package SIHR [54], which is available from CRAN. The SIHR package contains the main functions: LF, QF, and CATE. The LF function implements the confidence interval for $x_{\text{new}}^{\mathsf{T}}\beta$ in (25) under the high-dimensional linear regression with specifying model="linear" and the confidence intervals for $h(x_{\text{new}}^{\mathsf{T}}\beta)$ in (30) under the high-dimensional logistic regression with specifying model="logistic" or model="logisticalter", corresponding to the

linearization weighting and link-specific weighting introduced in Section 2.3, respectively.

The `QF` function implements the confidence interval construction in (37) and hypothesis testing in (38) with different choices of the weighting matrix A. The `CATE` function implements the confidence interval in (31) and hypothesis testing in (32). Both `QF` and `CATE` functions can be applied to the logistic regression setting by specifying `model="logistic"` or `model="logisticalter"`. The detailed usage of the `SIHR` package can be found in the paper [54].

4 Multiple Testing

In the previous sections, we have examined statistical inference for individual regression coefficients and related one-dimensional functionals. However, in many applications, such as genomics, it is necessary to perform simultaneous inference for multiple regression coefficients while controlling for the false discovery rate (FDR) and false discovery proportion (FDP). In this section, we will explore several large-scale multiple testing procedures for high-dimensional regression models. We start with the linear models in Section 4.1 and extend the discussion to logistic models in Section 4.2. The testing procedures are unified in Section 4.3 and the power enhancement methods are discussed next.

4.1 Simultaneous inference for linear regression

One-sample simultaneous inference for high-dimensional linear regression coefficients is closely related to the problem of variable selection. Common approaches for variable selection include regularization methods, such as Lasso [1], SCAD [10], Adaptive Lasso [4] and MCP [11], which simultaneously estimate parameters and select features, and stepwise feature selection techniques like LARS [19] and FoBa [84], which prioritize variable selection. See the discussions in [58] and references therein. However, both of these approaches aim to find the model that is closest to the truth, which may not be achievable in practice. Alternatively, [58] approached the problem from a multiple testing perspective and focused on controlling false discoveries rather than achieving perfect selection results. Specifically, for the high-dimensional regression model (1) with link function $h(z) = z$, i.e., $Y = X\beta + \epsilon$, where $\beta = (\beta_1, \ldots, \beta_{p+1})^\mathsf{T} \in \mathbb{R}^{p+1}$, $X = (X_{1,\cdot}^\mathsf{T}, \ldots, X_{n,\cdot}^\mathsf{T})^\mathsf{T}$, $Y = (Y_1, \ldots, Y_n)^\mathsf{T}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}$, with $\{\epsilon_k\}$ being independent and identically distributed (i.i.d) random variables with mean zero and variance σ_ϵ^2 and independent of $X_{k,\cdot}$, $k = 1, \ldots, n$, [58] considered the following multiple testing problem,

$$H_{0,i} : \beta_i = 0 \text{ versus } H_{1,i} : \beta_i \neq 0, \quad i = 2, \ldots, p+1, \tag{43}$$

with the control of FDR and FDP.

In some fields, one-sample inference may not be sufficient, particularly for detecting interactions. For example, as demonstrated in [85], many complex diseases are the result of interactions between genes and the environment. Therefore, it is important to thoroughly examine the effects of the environment

and its interactions with genetic predispositions on disease phenotypes. When the environmental factor is a binary variable, such as smoking status or gender, interaction detection can be approached under a two-sample high-dimensional regression framework. Specifically, interactions can be identified by comparing two high-dimensional regression models as introduced in (2) with identity link function, i.e., $Y^{(d)} = X^{(d)}\beta^{(d)} + \epsilon^{(d)}$, for $d = 1, 2$, and recovering the nonzero components of $\beta_{-1}^{(1)} - \beta_{-1}^{(2)}$, where $\beta^{(d)} = (\beta_1^{(d)}, \ldots, \beta_{p+1}^{(d)})^\mathsf{T} \in \mathbb{R}^{p+1}$, $X^{(d)} = (X_{1,\cdot}^{(d)\mathsf{T}}, \ldots, X_{n_d,\cdot}^{(d)\mathsf{T}})^\mathsf{T}$, $Y^{(d)} = (Y_1^{(d)}, \ldots, Y_{n_d}^{(d)})^\mathsf{T}$, and $\epsilon^{(d)} = (\epsilon_1^{(d)}, \ldots, \epsilon_{n_d}^{(d)})^\mathsf{T}$, with $\{\epsilon_k^{(d)}\}$ being i.i.d random variables with mean zero and variance $\sigma_{\epsilon^{(d)}}^2$ and independent of $X_{k,\cdot}^{(d)}$, $k = 1, \ldots, n_d$. Assume that $n_1 \asymp n_2$ and let $n = \max\{n_1, n_2\}$. Then [59] investigated simultaneous testing of the hypotheses

$$H_{0,i} : \beta_i^{(1)} = \beta_i^{(2)} \text{ versus } H_{1,i} : \beta_i^{(1)} \neq \beta_i^{(2)}, \quad i = 2, \ldots, p + 1, \qquad (44)$$

with FDR and FDP control.

In genetic association studies, it is common to measure multiple correlated phenotypes on the same individuals. To detect associations between high-dimensional genetic variants and these phenotypes, one can individually assess the relationship between each response and each covariate, as in (43), and then adjust for multiplicity in the comparisons. However, as noted by [86] and [87], jointly analyzing these phenotypic measurements may increase the power to detect causal genetic variants. Therefore, motivated by the potential to enhance power by leveraging the similarity across multivariate responses, [60] used high-dimensional multivariate regression models to address applications in which D correlated responses are measured on n independent individuals:

$$Y_{n \times D} = X_{n \times (p+1)} B_{(p+1) \times D} + \Upsilon_{n \times D}, \qquad (45)$$

where $Y = (Y_{.,1}, \ldots, Y_{.,D}) \in \mathbb{R}^{n \times D}$, with $Y_{.,d} = (Y_{1,d}, \ldots, Y_{n,d})^\mathsf{T}$, denotes D responses with D fixed, and $X = (X_{1,\cdot}^\mathsf{T}, \ldots, X_{n,\cdot}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{n \times (p+1)}$ is the covariate matrix. In (45), $B = (B_{.,1}, \ldots, B_{.,D}) \in \mathbb{R}^{(p+1) \times D}$, with $B_{.,d} = (B_{1,d}, \ldots, B_{p+1,d})^\mathsf{T} \in \mathbb{R}^{p+1}$, represents the regression coefficient matrix, where $B_{i,\cdot}$ represents the regression coefficients of the ith covariate; $\Upsilon = (\epsilon_{.,1}, \ldots, \epsilon_{.,D}) \in \mathbb{R}^{n \times D}$, where $\epsilon_{.,d} = (\epsilon_{1,d}, \ldots, \epsilon_{n,d})^\mathsf{T}$, and $\{\epsilon_{k,d}\}$ are i.i.d random variables with mean zero and variance σ_ϵ^2 and independent of X. To examine whether the ith covariate is associated with any of the D responses, [60] simultaneously tested

$$H_{0,i} : B_{i,\cdot} = 0 \text{ versus } H_{1,i} : B_{i,\cdot} \neq 0, \quad i = 2, \ldots, p + 1, \qquad (46)$$

while controlling FDR and FDP. Because the effect of the ith variable on each of the D responses may share strong similarities, namely, if $B_{i,d} \neq 0$, then the rest of the entries in this row are more likely to be nonzero, a row-wise testing method using the group-wise information is more favorable than testing the significance of the matrix B column by column as in testing problem (43).

4.1.1 Bias corrections via inverse regression

In the multiple testing problems discussed in Section 4.1, our goal is to simultaneously infer the regression coefficients while controlling for error rates.

Therefore, it is crucial to begin with an asymptotically unbiased estimator for each regression component. The debiasing techniques outlined in Section 2.2 can be used to attain nearly unbiased estimates, however, as noted in [58], the constraints or tuning parameters utilized in debiasing can significantly affect test accuracy. Additionally, the asymptotic distribution of these debiased estimators is conditional, making it challenging to characterize the dependence structure among the test statistics, which is vital for error rate control in simultaneous inference. As an alternative, we will explore an inverse regression approach in this section, which establishes the unconditional asymptotic distribution of bias corrected statistics and allows for explicit characterization of the correlation structure. Alternatively, the debiasing method (14) in [21, 23] can be used as long as the decorrelation vectors $Z_{\cdot,j}$'s introduced in Section 2.2.2 are close enough to their population counterparts. This approach will be illustrated in Section 4.2.

Recall that $X_{k,1}^{(d)} = 1$. To achieve bias correction, by taking testing problem (44) as an example, we consider the inverse regression models obtained by regressing $X_{k,i}^{(d)}$ on $(Y_k^{(d)}, \widetilde{X}_{k,-i}^{(d)})$, where $\widetilde{X}_{k,-i}^{(d)} = X_{k,-\{1,i\}}^{(d)}$, for $i = 2, \ldots, p+1$:

$$X_{k,i}^{(1)} = \alpha_i^{(1)} + (Y_k^{(1)}, \widetilde{X}_{k,-i}^{(1)})\gamma_i^{(1)} + \eta_{k,i}^{(1)}, \quad (k = 1, \ldots, n_1)$$
$$X_{k,i}^{(2)} = \alpha_i^{(2)} + (Y_k^{(2)}, \widetilde{X}_{k,-i}^{(2)})\gamma_i^{(2)} + \eta_{k,i}^{(2)}, \quad (k = 1, \ldots, n_2)$$

where for $d = 1, 2$, $\eta_{k,i}^{(d)}$ has mean zero and variance $\sigma_{i,d}^2$ and is uncorrelated with $(Y_k^{(d)}, \widetilde{X}_{k,-i}^{(d)})$, and the first component of $\gamma_i^{(d)} = (\gamma_{i,1}^{(d)}, \ldots, \gamma_{i,p}^{(d)})^\intercal$ satisfies

$$\gamma_{i,1}^{(d)} = \sigma_{i,d}^2 \beta_i^{(d)} / \sigma_{\epsilon^{(d)}}^2, \quad i = 2, \ldots, p+1, \tag{47}$$

where $\sigma_{i,d}^2 = [\{\beta_i^{(d)}\}^2/\sigma_{\epsilon^{(d)}}^2 + \omega_{i-1,i-1}^{(d)}]^{-1}$ with $\{\mathsf{Cov}(X_{k,-1}^{(d)})\}^{-1} = \Omega_d = (\omega_{i,j}^{(d)})$.

Note that, $r_i^{(d)} = \mathsf{Cov}(\epsilon_k^{(d)}, \eta_{k,i}^{(d)})$ can be expressed as $-\gamma_{i,1}^{(d)}\mathsf{Cov}(\epsilon_k^{(d)}, Y_k^{(d)}) = -\gamma_{i,1}^{(d)}\sigma_{\epsilon^{(d)}}^2 = -\sigma_{i,d}^2\beta_i^{(d)}$, hence we can approach the debiasing of $\beta_i^{(d)}$ through the debiasing of $r_i^{(d)}$, and equivalently formulate the testing problem (44) as

$$H_{0,i} : r_i^{(1)}/\sigma_{i,1}^2 = r_i^{(2)}/\sigma_{i,2}^2, \quad i = 2, \ldots, p+1.$$

The most straightforward way to estimate $r_i^{(d)}$ is to use the sample covariance between the error terms, $n_d^{-1}\sum_{k=1}^{n_d} \epsilon_k^{(d)}\eta_{k,i}^{(d)}$. However, the error terms are unknown, so we first estimate them by

$$\widehat{\epsilon}_k^{(d)} = Y_k^{(d)} - X_{k,\cdot}^{(d)}\widehat{\beta}^{(d)}, \quad \widehat{\eta}_{k,i}^{(d)} = X_{k,i}^{(d)} - (Y_k^{(d)} - \bar{Y}^{(d)}, \widetilde{X}_{k,-i}^{(d)})\widehat{\gamma}_i^{(d)},$$

where $\widehat{\beta}^{(d)} = (\widehat{\beta}_1^{(d)}, \ldots, \widehat{\beta}_{p+1}^{(d)})$ and $\widehat{\gamma}_i^{(d)} = (\widehat{\gamma}_{i,1}^{(d)}, \ldots, \widehat{\gamma}_{i,p}^{(d)})$ are respectively the estimators of $\beta^{(d)}$ and $\gamma_i^{(d)}$ that satisfy

$$\max\{|\widehat{\beta}_{-1}^{(d)} - \beta_{-1}^{(d)}|_1, \max_{i=1,\ldots,p} |\widehat{\gamma}_i^{(d)} - \gamma_i^{(d)}|_1\} = O_\mathbb{P}(a_{n1}),$$
$$\max\{|\widehat{\beta}_{-1}^{(d)} - \beta_{-1}^{(d)}|_2, \max_{i=1,\ldots,p} |\widehat{\gamma}_i^{(d)} - \gamma_i^{(d)}|_2\} = O_\mathbb{P}(a_{n2}), \tag{48}$$

for some a_{n1} and a_{n2} such that

$$\max\{a_{n1}a_{n2}, a_{n2}^2\} = o\{(n\log p)^{-1/2}\}, \text{ and } a_{n1} = o(1/\log p). \tag{49}$$

As noted by [58, 59], estimators $\widehat{\beta}^{(d)}$ and $\widehat{\gamma}_i^{(d)}$ that satisfy (48) and (49) can be obtained easily via standard methods such as the Lasso and Danzig

selector. Following that, a natural estimator of $r_i^{(d)}$ can be constructed by $\widetilde{r}_i^{(d)} = n_d^{-1} \sum_{k=1}^{n_d} \widehat{\epsilon}_k^{(d)} \widehat{\eta}_{k,i}^{(d)}$. However, the bias of $\widetilde{r}_i^{(d)}$ exceeds the desired rate $(n_d \log p)^{-1/2}$ for the subsequent analysis. Hence, the difference of $\widetilde{r}_i^{(d)}$ and $n_d^{-1} \sum_{k=1}^{n_d} \epsilon_k^{(d)} \eta_{k,i}^{(d)}$ is calculated, and it is equal to $\widehat{\sigma}_{\epsilon^{(d)}}^2 \widehat{\gamma}_{i,1}^{(d)} + \widehat{\sigma}_{i,d}^2 \widehat{\beta}_i^{(d)}$ up to order $(n_d \log p)^{-1/2}$ under regularity conditions, where $\widehat{\sigma}_{\epsilon^{(d)}}^2 = n_d^{-1} \sum_{k=1}^{n_d} \{\widehat{\epsilon}_k^{(d)}\}^2$ and $\widehat{\sigma}_{i,d}^2 = n_d^{-1} \sum_{k=1}^{n_d} \{\widehat{\eta}_{k,i}^{(d)}\}^2$ are the sample variances. Hence, a bias-corrected estimator for $r_i^{(d)}$ is defined as

$$\widehat{r}_i^{(d)} = \widetilde{r}_i^{(d)} + \widehat{\sigma}_{\epsilon^{(d)}}^2 \widehat{\gamma}_{i,1}^{(d)} + \widehat{\sigma}_{i,d}^2 \widehat{\beta}_i^{(d)}. \tag{50}$$

For the other two testing problems, the bias corrections can be performed almost exactly the same via the inverse regression technique above that translates the debiasing of regression coefficients to the debiasing of residual covariances. Note that, through such an inverse regression approach, one can appropriately deal with the dependency of the component-wise debiased statistics, which is important for the following adjustment of multiplicity and the goal of FDR control.

4.1.2 Construction of test statistics

We next construct test statistics for each of the three problems discussed in Section 4.1, using the bias-corrected statistics as a starting point.

For problem (43), because testing whether $\beta_i = 0$ is equivalent as testing whether the residual covariance is equal to zero, the test statistics can be constructed directly based on the bias corrected $\widehat{r}_i$ (it can be obtained exactly the same as $\widehat{r}_i^{(d)}$ as shown in Section 4.1.1 where the superscript (d) is dropped since there is only one sample). Then the test statistics that standardize $\widehat{r}_i$'s are obtained by

$$W_i = \frac{\widehat{r}_i}{(\widehat{\sigma}_\epsilon^2 \widehat{\sigma}_i^2 / n)^{1/2}}, \quad i = 2, \ldots, p+1, \tag{51}$$

where $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_i^2$ are again the sample variances by respectively dropping the superscript (d) and subscript d in the one-sample case. As shown in [58], the statistics W_i's are asymptotically normal under the null.

The above construction cannot be directly applied to the problem (44), because $\beta_i^{(d)}$ is not necessary equal to 0 under the two-sample null and $r_i^{(1)}/\sigma_{i,1}^2 = r_i^{(2)}/\sigma_{i,2}^2$ is not equivalent to $r_i^{(1)} = r_i^{(2)}$. Thus, it is necessary to construct testing procedures based directly on estimators of $r_i^{(1)}/\sigma_{i,1}^2 - r_i^{(2)}/\sigma_{i,2}^2$. By the bias correction in Section 4.1.1, [59] proposed an estimator of $r_i^{(d)}/\sigma_{i,d}^2$:

$$T_i^{(d)} = \widehat{r}_i^{(d)}/\widehat{\sigma}_{i,d}^2, \quad i = 2, \ldots, p+1; \ d = 1, 2,$$

and tested (44) via the estimators $\{T_i^{(1)} - T_i^{(2)} : i = 2, \ldots, p+1\}$. Due to the heteroscedasticity, [59] considered a standardized version of $T_i^{(1)} - T_i^{(2)}$. Specifically, let

$$\widetilde{U}_i^{(d)} = (\beta_i^{(d)} + U_i^{(d)})/\sigma_{i,d}^2, \text{ with } U_i^{(d)} = n_d^{-1} \sum_{k=1}^{n_d} \{\epsilon_k^{(d)} \eta_{k,i}^{(d)} - \mathsf{E}(\epsilon_k^{(d)} \eta_{k,i}^{(d)})\}.$$

It was shown in [59] that $T_i^{(d)}$ is close to $\widetilde{U}_i^{(d)}$ asymptotically under regularity conditions. Because $\theta_i^{(d)} = \mathsf{Var}(\widetilde{U}_i^{(d)}) = \mathsf{Var}(\epsilon_k^{(d)}\eta_{k,i}^{(d)}/\sigma_{i,d}^2)/n_d = (\sigma_{\epsilon^{(d)}}^2/\sigma_{i,d}^2 + \{\beta_i^{(d)}\}^2)/n_d$, it can be estimated by $\widehat{\theta}_i^{(d)} = (\widehat{\sigma}_{\epsilon^{(d)}}^2/\widehat{\sigma}_{i,d}^2 + \{\widehat{\beta}_i^{(d)}\}^2)/n_d$, and the standardized statistics are defined by

$$W_i = \frac{T_i^{(1)} - T_i^{(2)}}{(\widehat{\theta}_i^{(1)} + \widehat{\theta}_i^{(2)})^{1/2}}, \quad i = 2, \ldots, p+1, \tag{52}$$

which are asymptotically normal under the null as studied in [59].

For the multivariate testing problem (46), by taking advantage of the similar effect of the ith variable on each of the responses, a group lasso penalty [88] can be imposed to obtain an estimator of B, such that

$$\max_{1 \le d \le D} |\widehat{B}_{-1,d} - B_{-1,d}|_1 = O_{\mathbb{P}}(a_{n1}), \max_{1 \le d \le D} |\widehat{B}_{-1,d} - B_{-1,d}|_2 = O_{\mathbb{P}}(a_{n2}),$$

for some a_{n1} and a_{n2} satisfying (49). By [89], the above rates can be fulfilled if the row sparsity of B satisfies $s(p) = o(n^{1/3}/\log p)$. Then following the same bias correction strategy as described in Section 4.1.1, a debiased estimator for $r_i^{(d)}$ can be obtained via $\widehat{r}_i^{(d)} = \widetilde{r}_i^{(d)} + \widehat{\sigma}_\epsilon^2\widehat{\gamma}_{i,1}^{(d)} + \widehat{\sigma}_{i,d}^2\widehat{B}_{i,d}$. Then similarly as the aforementioned two problems, the standardized statistic can be constructed by

$$\widetilde{T}_i^{(d)} = T_i^{(d)}/\{\widehat{\theta}_i^{(d)}\}^{1/2}, \quad i = 2, \ldots, p+1; \ d = 1, \ldots, D,$$

where $T_i^{(d)} = \widehat{r}_i^{(d)}/\widehat{\sigma}_{i,d}^2$ and $\widehat{\theta}_i^{(d)} = (\widehat{\sigma}_\epsilon^2/\widehat{\sigma}_{i,d}^2 + \widehat{B}_{i,d}^2)/n$. Finally, a sum-of-squares-type test statistic for testing the ith row of B is proposed:

$$W_i = \sum_{d=1}^{D}\{\widetilde{T}_i^{(d)}\}^2, \quad i = 2, \ldots, p+1, \tag{53}$$

which is asymptotically χ_D^2 distributed under the null as studied in [60]. We briefly mentioned that in the case where the errors in Model (45) are correlated within each individual (e.g., $\mathsf{Cov}(\epsilon_{i,\cdot}) = \Sigma_\epsilon \neq I_D$), an iterative procedure can be employed to estimate B and Σ_ϵ [e.g., 90]. Subsequently, based on inverse regression, the test statistics (53) can be constructed again. However, it is important to note that in this scenario, the null distribution of the test statistics may depend on Σ_ϵ and may not follow a χ_D^2 distribution.

4.2 Simultaneous inference for logistic regression

The principles of simultaneous inference for high-dimensional linear regression can also be applied to high-dimensional logistic regression models. In particular, [57] considered the regression model (1), i.e., $Y_k = h(X_{k,\cdot}^\mathsf{T}\beta) + \epsilon_k$, $k = 1, \ldots, n$, with the link function $h(z) = \exp(z)/[1+\exp(z)]$, and studied the simultaneous testing problem (43) as described in Section 4.1, namely testing

$$H_{0,i} : \beta_i = 0 \text{ versus } H_{1,i} : \beta_i \neq 0, \quad i = 2, \ldots, p+1,$$

with FDR control.

Based on the regularized estimator $\widehat{\beta}$ given in (5), [57] corrected the bias of $\widehat{\beta}$ via the Taylor expansion of $h(u_k)$ at $\widehat{u}_k$ for $u_k = X_{k,\cdot}^\mathsf{T}\beta$ and $\widehat{u}_k = X_{k,\cdot}^\mathsf{T}\widehat{\beta}$,

and obtained that

$$Y_k - h(\widehat{u}_k) + h'(\widehat{u}_k)X_{k,.}^\mathsf{T}\widehat{\beta} = h'(\widehat{u}_k)X_{k,.}^\mathsf{T}\beta + (R_k + \epsilon_k), \tag{54}$$

where R_k is the remainder term as specified in (17). Next, $Y_k - h(\widehat{u}_k) + h'(\widehat{u}_k)X_{k,.}^\mathsf{T}\widehat{\beta}$ can be treated as the new response, $h'(\widehat{u}_k)X_{k,.}$ as the new covariates, and $R_k + \epsilon_k$ as the new noise. Under such a formulation, testing $H_{0,i} : \beta_i = 0$ can be translated into the simultaneous inference for the regression coefficients of an approximate linear model. [57] applied the decorrelation method (14) and constructed the following debiased estimator $\widetilde{\beta}_i^{\text{Deb}}$,

$$\widetilde{\beta}_i^{\text{Deb}} = \widehat{\beta}_i + \frac{\sum_{k=1}^n Z_{k,i}(Y_k - h(X_{k,.}^\mathsf{T}\widehat{\beta}))}{\sum_{k=1}^n Z_{k,i}h'(X_{k,.}^\mathsf{T}\widehat{\beta})X_{k,i}}, \quad i = 2, \ldots, p+1, \tag{55}$$

where $Z_{.,i}$ is determined by the scaled residual that regresses $X_{.,i}$ on $X_{.,-i}$ through the linearization weighting $\widehat{W}_k = 1/h'(X_{k,.}^\mathsf{T}\widehat{\beta})$ introduced in Section 2.3; same strategy was also employed in [91, 92]. As summarized in Section 4.1.1, for the subsequent FDR control analysis, the decorrelation vectors $Z_{.,i}$'s were shown to be close to the true regression errors. Alternatively, the inverse regression technique in Section 4.1.1 can be similarly applied in the approximate linear model (54) for the bias correction.

Based on (55), [57] proposed the following standardized test statistic:

$$W_i = \frac{\widetilde{\beta}_i^{\text{Deb}}}{\{\sum_{k=1}^n h'(\widehat{u}_k)Z_{k,i}^2\}^{1/2}/\{\sum_{k=1}^n h'(\widehat{u}_k)Z_{k,i}X_{k,i}\}}, \quad i = 2, \ldots, p+1, \tag{56}$$

which is asymptotically normal under the null. Additionally, [57] extended the idea to the two-sample multiple testing (44). Similarly, multiple testing of (46) can also be approached in the logistic setting.

4.3 Multiple testing procedure

Using the test statistics $\{W_i : i = 2, \ldots, p+1\}$ in (51), (52), (53) and (56) as a foundation, we next examine a unified multiple testing procedure that guarantees error rates control.

Let $\mathcal{H} = \{2, \ldots, p+1\}$, $\mathcal{H}_0$ be the set of true null indices and $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$ be the set of true alternatives. We are interested in cases where most of the tests are nulls, that is, $|\mathcal{H}_1|$ is relatively small compared to $|\mathcal{H}|$. Let $\Psi(\cdot)$ be the asymptotic cumulative distribution function (cdf) of W_i under the null and let $\Phi(\cdot)$ be the standard normal cdf, then we develop a normal quantile transformation of W_i by $N_i = \Phi^{-1}\{1 - (1 - \Psi(W_i))/2\}$, which approximately has the same distribution as the absolute value of a standard normal random variable under the null $H_{0,i}$. Let t be the threshold level such that $H_{0,i}$ is rejected if $N_i \geq t$. For any given t, denote by $R_0(t) = \sum_{i \in \mathcal{H}_0} \mathbb{I}\{N_i \geq t\}$ and $R(t) = \sum_{i \in \mathcal{H}} \mathbb{I}\{N_i \geq t\}$ the total number of false positives and the total number of rejections, respectively. Then the FDP and FDR are defined as

$$\text{FDP}(t) = \frac{R_0(t)}{\max\{R(t), 1\}}, \quad \text{FDR}(t) = \mathbb{E}\{\text{FDP}(t)\}.$$

An ideal choice of t rejects as many true positives as possible while controlling the FDP at the pre-specified level α, i.e.,

$$t_0 = \inf\left\{0 \leq t \leq (2\log p)^{1/2} : \text{FDP}(t) \leq \alpha\right\}.$$

Since $R_0(t)$ can be estimated by $2\{1 - \Phi(t)\}|\mathcal{H}_0|$ and $|\mathcal{H}_0|$ is upper bounded by p, we conservatively estimate it by $2p\{1 - \Phi(t)\}$. Therefore, the following multiple testing algorithm is proposed in [e.g., 57–60].

Algorithm 1 The multiple testing procedure.

Step 1. Obtain the transformed statistics $N_i = \Phi^{-1}\{1 - (1 - \Psi(W_i))/2\}$ from the test statistics W_i, $i = 2, \ldots, p + 1$.

Step 2. For a given $0 \leq \alpha \leq 1$, calculate

$$\widehat{t} = \inf\left[0 \leq t \leq (2\log p - 2\log\log p)^{1/2} : \frac{2p\{1 - \Phi(t)\}}{\max\{R(t), 1\}} \leq \alpha\right]. \qquad (57)$$

If (57) does not exist, then set $\widehat{t} = (2\log p)^{1/2}$.

Step 3. For $i \in \mathcal{H}$, reject $H_{0,i}$ if $N_i \geq \widehat{t}$.

As noted in [93], the constraint $0 \leq t \leq (2\log p - 2\log\log p)^{1/2}$ in (57) is critical. When t exceeds the upper bound, $2p\{1 - \Phi(t)\} \to 0$ is not even a consistent estimate of $R_0(t)$. However, Benjamini-Hochberg (B-H) procedure [94] used $2p\{1 - \Phi(t)\}$ as an estimate of $R_0(t)$ for all $t \geq 0$ and hence may not be able to control the FDP. On the other hand, if t is not attained in the range, it is important to threshold N_i at $(2\log p)^{1/2}$, because thresholding N_i at $(2\log p - 2\log\log p)^{1/2}$ may cause too many false rejections. As a result, by applying the above multiple testing algorithm to each of the problems in Sections 4.1 and 4.2, under some regularity conditions as specified in [57–60], we reach both the FDP and FDR control at the pre-specified level α asymptotically, i.e., $\lim_{(n,p)\to\infty} \mathbb{P}\{\text{FDP}_{\text{Alg1}} \leq \alpha + \epsilon\} = 1$ for any $\epsilon > 0$ and $\overline{\lim}_{(n,p)\to\infty} \text{FDR}_{\text{Alg1}} \leq \alpha$.

4.4 Power enhancement

In addition to controlling error rates, we consider strategies to improve the power of multiple testing procedures, focusing on enhancing the power for two-sample inference when the high-dimensional objects of interest are individually sparse. This is explored in [70] with an extension to a more general framework in [95]. It is worth noting that [70] improved the performance of Algorithm 1 by utilizing unknown sparsity in two-sample multiple testing. Additionally, power enhancement for the simultaneous inference of GLMs can be achieved through data integration [e.g., 71, 96] as well as other power boosting methods designed for general multiple testing problems as introduced in Section 1.

Recall that, in the two-sample problem studied in Section 4.1, we aim to make inference for $\delta_i = \mathbb{I}\{\beta_i^{(1)} \neq \beta_i^{(2)}\}$, $i = 2, \ldots, p + 1$. Following Algorithm 1,

we first summarize the data by a single vector of test statistics $\{N_2, \cdots, N_{p+1}\}$ and then choose a significance threshold to control the multiplicity. However, such approach ignores the important feature that both objects $\beta^{(1)}$ and $\beta^{(2)}$ are individually sparse. Let $\mathcal{I}_d = \{i = 2, \ldots, p + 1 : \beta_i^{(d)} \neq 0\}$ denote the support of $\beta^{(d)}$, $d = 1, 2$, and $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ the union support. Because the small cardinality of $\mathcal{I}$ implies that both $\beta^{(1)}$ and $\beta^{(2)}$ are sparse, the information on $\mathcal{I}$ can be potentially utilized to narrow down the alternatives via the logical relationship that $i \notin \mathcal{I}$ implies that $\delta_i = 0$.

The goal of [70] is to incorporate the sparsity information to improve the testing efficiency, and it is accomplished via the construction of an additional covariate sequence $\{S_i : i = 2, \ldots, p + 1\}$ to capture the information on $\mathcal{I}$. Note that S_i and N_i have different roles: N_i is the primary statistic to evaluate the significance of the test, while S_i is the auxiliary one that captures the sparsity information to assist the inference. It is also important to note that S_i should be asymptotically independent with N_i so that the null distribution of N_i would not be distorted by the incorporation of S_i. For the two-sample problem in Section 4.1, such auxiliary statistics can be constructed by

$$S_i = \frac{\widehat{r}_i^{(1)}/\widehat{\sigma}_{i,1}^2 + (\widehat{\theta}_i^{(1)}/\widehat{\theta}_i^{(2)})(\widehat{r}_i^{(2)}/\widehat{\sigma}_{i,2}^2)}{\{\widehat{\theta}_i^{(1)}(1 + \widehat{\theta}_i^{(1)}/\widehat{\theta}_i^{(2)})\}^{1/2}}, \quad i = 2, \ldots, p + 1. \tag{58}$$

Then based on the pairs of statistics $\{(N_i, S_i) : i = 2, \ldots, p+1\}$, the proposal in [70] operates in three steps: grouping, adjusting and pooling (GAP). The first step divides all tests into K groups based on S_i, which leads to heterogeneous groups with varied sparsity levels. The second step adjusts the p-values to incorporate the sparsity information. The final step combines the adjusted p-values and chooses a threshold to control the global FDR. Based on the p-values obtained by the test statistics N_i's, i.e., $p_i = 2\{1 - \Phi(N_i)\}$, the algorithm is summarized in Algorithm 2 and we refer to [70] for its detailed implementations such as the choices of the number of groups and the grid sets.

In addition, [70] provided some insights on why the GAP algorithm works. First, it adaptively chooses the group-wise FDR levels via adjusted p-values and effectively incorporates group-wise information. Intuitively, Algorithm 2 increases the overall power by assigning higher FDR levels to groups where signals are more common. It does not assume known groups and searches for the optimal grouping to maximize the power. Moreover, the construction of the weights in Algorithm 2 ensures that after all groups are combined, the weights are always "proper" in the sense of [63]. As a result, even inaccurate estimates of the non-null proportions would not affect the validity of overall FDR control.

Then, same as Algorithm 1, [70] provided the asymptotic error rates control results for Algorithm 2, namely, we have $\lim_{(n,p)\to\infty} \mathbb{P}\{\text{FDP}_{\text{Alg2}} \leq \alpha + \epsilon\} = 1$ for any $\epsilon > 0$ and $\overline{\lim}_{(n,p)\to\infty} \text{FDR}_{\text{Alg2}} \leq \alpha$. Moreover, due to the informative weights (59) that effectively incorporate the sparsity information through $\{S_i : i = 2, \ldots, p+1\}$, Algorithm 2 dominates Algorithm 1 in power asymptotically. Specifically, [70] provided the rigorous theoretical comparison that $\Psi_{\text{Alg2}} \geq \Psi_{\text{Alg1}} + o(1)$ as $p \to \infty$, where Ψ_{Alg1} and Ψ_{Alg2} represent the expectations

---

Algorithm 2 Multiple testing via grouping, adjusting and pooling (GAP).

Step 1 (Grouping). Divide hypotheses into K groups: $\mathcal{G}_l = \{i = 2, \ldots, p + 1 : \lambda_{l-1} < S_i \leq \lambda_l\}$, for $1 \leq l \leq K$. The optimal choice of grouping will be determined in Step 2.

Step 2 (Adjusting). Define $m_l = |\mathcal{G}_l|$. Calculate adjusted p-values $p_i^w = \min\{p_i/w_l^o, 1\}$ if $i \in \mathcal{G}_l$, $1 \leq l \leq K$, where $p_i = 2\{1 - \Phi(N_i)\}$ and w_l^o will be calculated as follows.

- *Initial adjusting.* For a given grouping $\{\mathcal{G}_l : 1 \leq l \leq K\}$, let $\widehat{\pi}_l$ be the estimated proportion of non-nulls in $\mathcal{G}_l$. Compute the group-wise weights

$$w_l = \left\{ \sum_{l=1}^{K} \frac{m_l \widehat{\pi}_l}{1 - \widehat{\pi}_l} \right\}^{-1} \frac{p \widehat{\pi}_l}{(1 - \widehat{\pi}_l)}, \ 1 \leq l \leq K. \tag{59}$$

  Define $p_i^w = \min\{p_i/w_l, 1\}$ for $i \in \mathcal{G}_l$.
- *Further refining.* For each $\Lambda = \{\lambda_l : 1 \leq l \leq K - 1\}$ (allowed to be empty), let

$$k = \max\{j : p_{(j)}^w \leq j\alpha/p\}, \tag{60}$$

  and reject the hypotheses corresponding to $\{p_{(1)}^w, \ldots, p_{(k)}^w\}$, where $p_{(1)}^w \leq \cdots \leq p_{(p)}^w$ are the re-ordered p-values. The weights w_l^o are computed using (59) based on the optimal grouping that yields most rejections.

Step 3 (Pooling). Combine p_i^w's computed from Step 2 based on the optimal grouping, apply (60) again and output the rejections.

---

of the proportions of correct rejections among all alternative hypotheses for Algorithms 1 and 2, respectively.

5 Discussion

In this expository paper, we provided a review of methods and theoretical results on statistical inference and multiple testing for high-dimensional regression models, including linear and logistic regression. Due to limited space, we were unable to discuss a number of related inference problems. In this section, we briefly mention a few of them.

Accuracy assessment. Accuracy assessment is a crucial part of high-dimensional uncertainty quantification. Its goal is to evaluate the estimation accuracy of an estimator $\widehat{\beta}$. For linear regression, [97] considered a collection of estimators $\widehat{\beta}$ and established the minimaxity and adaptivity of the point estimation and confidence interval for $\|\widehat{\beta} - \beta\|_q^q$ with $1 \leq q \leq 2$. Suppose that $\widehat{\beta}$ is independent of the data $\{X_{k,\cdot}, Y_k\}_{1 \leq k \leq n}$ (which can be achieved by sample splitting, for example). Define the residue $Re_k = Y_k - X_{k,\cdot}^\mathsf{T} \widehat{\beta} = X_{k,\cdot}^\mathsf{T}(\beta - \widehat{\beta}) + \epsilon_k$. The quadratic functional inference approach introduced in [27] can be applied to the data $\{Re_k, X_{k,\cdot}\}_{1 \leq k \leq n}$ to make inference for both $(\widehat{\beta} - \beta)^\mathsf{T} \Sigma (\widehat{\beta} - \beta)$

and $\|\widehat{\beta} - \beta\|_2^2$; see more details in Section 5.2 of [27]. [98] used an estimator of $\|\widehat{\beta} - \beta\|_2^2$ to construct a confidence set for β with $\widehat{\beta}$ denoting an accurate high-dimensional estimator. In the context of approximate message passing, [99, 100] considered a different framework with $n/(p+1) \to \delta \in (0,1)$ and independent Gaussian design and established the asymptotic limit of $\|\widehat{\beta} - \beta\|_2^2/(p+1)$ for the Lasso estimator $\widehat{\beta}$.

Semi-supervised Inference. The semi-supervised inference is well motivated by a wide range of modern applications, such as Electronic Health Record data analysis. In addition to the labeled data, additional covariate observations exist in the semi-supervised setting. It is critical to leverage the information in the unlabelled data and improve the inference efficiency. As reviewed in Section 3.5, the additional unlabelled data improves the accuracy of various high-dimensional inference procedures by facilitating the estimation of Σ or Σ^{-1} [27, 75]. Moreover, in a different context where the linear outcome model might be misspecified, one active research direction in semi-supervised learning is to construct a more complicated imputation model (e.g., by applying the classical non-parametric or machine learning methods) and conduct a following-up bias correction after outcome imputation [e.g., 101–105].

Applications of quadratic form inference. The statistical inference methods for quadratic functionals in Section 3.6 and inner products in Section 3.4 are useful in a wide range of statistical applications. Firstly, the group significance test of $\|\beta_G\|_2^2 = 0$ or $\beta_G^{\mathsf{T}} \Sigma_{G,G} \beta_G = 0$ forms an important basis for designing computationally efficient hierarchical testing approaches [26, 106]. Secondly, consider the high-dimensional interaction model $Y_k = A_k \eta + X_{k,\cdot}^{\mathsf{T}} \tau + A_k \cdot X_{k,\cdot}^{\mathsf{T}} \gamma$ with A_k denoting the variable of interest (e.g., the treatment) and $X_{k,\cdot}$ denoting a large number of other variables. In this model, testing the existence of the interaction term can be reduced to the inference for $\|\gamma\|_2^2$. Lastly, in the two-sample model (2), the methods proposed in Sections 3.6 and 3.4 have been applied in [107] to calculating the difference $\|\beta^{(1)} - \beta^{(2)}\|_2^2$, which has the expression of $\|\beta^{(1)}\|_2^2 + \|\beta^{(2)}\|_2^2 - 2[\beta^{(1)}]^{\mathsf{T}} \beta^{(2)}$.

Multiple heterogeneous regression models. It is essential to perform efficient integrative inference in various applications that combine multiple regression models. Examples include transfer learning, distributed learning, federated learning, and distributionally robust learning. Transfer learning provides a powerful tool for incorporating data from related studies to improve estimation and inference accuracy in a target study of direct interest. [108] studied transfer learning for high-dimensional linear regression. Minimax optimal convergence rates were established, and data-driven adaptive algorithms were proposed. [109, 110] explored transfer learning in high-dimensional GLMs and [111, 112] considered distributed learning for high-dimensional regression. Additionally, [71, 96] proposed integrative estimation and multiple testing procedures of cross-sites high-dimensional regression models that simultaneously accommodated between study heterogeneity and protected individual-level data. In a separate direction, [113] proposed the maximin effect as a robust prediction model for the target distribution being generated as a mixture of

multiple source populations. [83] further established that the maximin effect is a group distributionally robust model and studied the statistical inference for the maximin effect in high dimensions. Many questions in multiple heterogeneous regression models are open and warrant future research.

Other simultaneous inference problems. The principles of multiple testing procedures reviewed in Section 4 are also widely applicable to a range of simultaneous inference problems, including graph learning [e.g., 92, 93, 114, 115], differential network recovery [e.g., 116–118], as well as various structured regression analysis [e.g., 77, 119, 120]. For example, through similar regression-based bias-correction techniques as reviewed in Section 4.1.1, [114] studied the estimation of Gaussian Graphical Models with FDR control; [93] focused on the recovery of sub-networks in Gaussian graphs; [92] identified significant communities for compositional data. Additionally, [116, 117] proposed multiple testing procedures for differential network detections of vector-valued and matrix-valued observations, respectively. Multiple testing of structured regression models such as mixed-effects models and confounded linear models has also been extensively studied in the literature [77, 120].

Alternative multiple testing methods for regression models. Besides the bias-correction based multiple testing approaches reviewed in this article, there are a few alternative classes of methods that aim at finite sample FDR control for linear models. Examples include knockoff based methods, mirror statistics approaches and e-value proposals. In particular, [121] constructed a set of knockoff variables and selected those predictors that have considerably higher importance scores than the knockoff counterparts; [122] extended the work to a model-X knockoff framework that allowed unknown conditional distribution of the response. There are several generalizations along this line of research; see [123, 124] and many references therein. Inspired by the knockoff idea, [125] proposed a symmetrized data aggregation approach to build mirror statistics that incorporate data dependence structure; a general framework on mirror statistics of GLMs was studied in [126] and the references therein. The e-value based proposal [127] is another useful tool for multiple testing in general. [128] proposed an e-BH procedure that achieved FDR control under arbitrary dependence among the e-values, and the equivalence between the knockoffs and the e-BH was studied in [129].

References

[1] Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. B **58**(1), 267–288 (1996)

[2] Candes, E., Tao, T.: The dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. **35**(6), 2313–2351 (2007)

[3] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B **67**(2), 301–320 (2005)

[4] Zou, H.: The adaptive Lasso and its oracle properties. J. Am. Statist. Assoc. **101**(476), 1418–1429 (2006)

[5] Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and dantzig selector. Ann. Statist. **37**(4), 1705–1732 (2009)

[6] Bühlmann, P., van de Geer, S.: Statistics for High-dimensional Data: Methods, Theory and Applications. Springer, New York (2011)

[7] Negahban, S., Yu, B., Wainwright, M.J., Ravikumar, P.K.: A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In: Adv. Neural Inf. Process. Syst., pp. 1348–1356 (2009)

[8] van de Geer, S.A., Bühlmann, P.: On the conditions used to prove oracle results for the Lasso. Electron. J. Stat. **3**, 1360–1392 (2009)

[9] Huang, J., Zhang, C.-H.: Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. J. Mach. Learn. Res. **13**(Jun), 1839–1864 (2012)

[10] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Statist. Assoc. **96**(456), 1348–1360 (2001)

[11] Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. **38**(2), 894–942 (2010)

[12] Belloni, A., Chernozhukov, V., Wang, L.: Square-root Lasso: pivotal recovery of sparse signals via conic programming. Biometrika **98**(4), 791–806 (2011)

[13] Sun, T., Zhang, C.-H.: Scaled sparse linear regression. Biometrika **101**(2), 269–284 (2012)

[14] Bunea, F.: Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. Electron. J. Stat. **2**, 1153–1194 (2008)

[15] Bach, F.: Self-concordant analysis for logistic regression. Electron. J. Stat. **4**, 384–414 (2010)

[16] Meier, L., van de Geer, S., Bühlmann, P.: The group Lasso for logistic regression. J. R. Statist. Soc. B **70**(1), 53–71 (2008)

[17] Bellec, P.C., Lecué, G., Tsybakov, A.B.: Slope meets Lasso: improved oracle bounds and optimality. Ann. Statist. **46**(6B), 3603–3642 (2018)

[18] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1 (2010)

[19] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)

[20] Greenshtein, E., Ritov, Y.: Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli **10**(6), 971–988 (2004)

[21] van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R.: On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. **42**(3), 1166–1202 (2014)

[22] Javanmard, A., Montanari, A.: Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. **15**(1), 2869–2909 (2014)

[23] Zhang, C.-H., Zhang, S.S.: Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Statist. Soc. B **76**(1), 217–242 (2014)

[24] Guo, Z., Rakshit, P., Herman, D.S., Chen, J.: Inference for the case probability in high-dimensional logistic regression. J. Mach. Learn. Res. **22**(1), 11480–11533 (2021)

[25] Guo, Z., Wang, W., Cai, T.T., Li, H.: Optimal estimation of genetic relatedness in high-dimensional linear models. J. Am. Statist. Assoc. **114**(525), 358–369 (2019)

[26] Guo, Z., Renaux, C., Bühlmann, P., Cai, T.T.: Group inference in high dimensions with applications to hierarchical testing. Electron. J. Stat. **15**(2), 6633–6676 (2021)

[27] Cai, T.T., Guo, Z.: Semisupervised inference for explained variance in high dimensional linear regression and its applications. J. R. Statist. Soc. B **82**(2), 391–419 (2020)

[28] Ma, R., Guo, Z., Cai, T.T., Li, H.: Statistical inference for genetic relatedness based on high-dimensional logistic regression. arXiv preprint arXiv:2202.10007 (2022)

[29] Cai, T.T., Guo, Z.: Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. Ann. Statist. **45**(2), 615–646 (2017)

[30] Athey, S., Imbens, G.W., Wager, S.: Approximate residual balancing: debiased inference of average treatment effects in high dimensions. J. R. Statist. Soc. B **80**(4), 597–623 (2018)

[31] Cai, T.T., Guo, Z., Ma, R.: Statistical inference for high-dimensional generalized linear models with binary outcomes. J. Am. Statist. Assoc. **116**, 1–14 (2021)

[32] Ning, Y., Liu, H.: A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Statist. **45**(1), 158–195 (2017)

[33] Ren, Z., Sun, T., Zhang, C.-H., Zhou, H.H.: Asymptotic normality and optimalities in estimation of large gaussian graphical models. Ann. Statist. **43**(3), 991–1026 (2015)

[34] Yu, Y., Bradic, J., Samworth, R.J.: Confidence intervals for high-dimensional cox models. arXiv preprint arXiv:1803.01150 (2018)

[35] Fang, E.X., Ning, Y., Li, R.: Test of significance for high-dimensional longitudinal data. Ann. Statist. **48**(5), 2622 (2020)

[36] Zhou, R.R., Wang, L., Zhao, S.D.: Estimation and inference for the indirect effect in high-dimensional linear mediation models. Biometrika **107**(3), 573–589 (2020)

[37] Zhao, T., Kolar, M., Liu, H.: A general framework for robust testing and confidence regions in high-dimensional quantile regression. arXiv preprint arXiv:1412.8724 (2014)

[38] Javanmard, A., Lee, J.D.: A flexible framework for hypothesis testing in high dimensions. J. R. Statist. Soc. B **82**(3), 685–718 (2020)

[39] Chen, Y., Fan, J., Ma, C., Yan, Y.: Inference and uncertainty quantification for noisy matrix completion. Proc. Natl. Acad. Sci. **116**(46), 22931–22937 (2019)

[40] Zhu, Y., Bradic, J.: Linear hypothesis testing in dense high-dimensional linear models. J. Am. Statist. Assoc. **113**(524), 1583–1600 (2018)

[41] Guo, Z., Kang, H., Cai, T.T., Small, D.S.: Testing endogeneity with high dimensional covariates. J. Econom. **207**(1), 175–187 (2018)

[42] Cai, T.T., Zhang, L.: High-dimensional gaussian copula regression: Adaptive estimation and statistical inference. Stat. Sin., 963–993 (2018)

[43] Eftekhari, H., Banerjee, M., Ritov, Y.: Inference in high-dimensional single-index models under symmetric designs. J. Mach. Learn. Res. **22**, 27–1 (2021)

[44] Deshpande, Y., Mackey, L., Syrgkanis, V., Taddy, M.: Accurate inference for adaptive linear models. In: International Conference on Machine Learning, pp. 1194–1203 (2018). PMLR

[45] Fang, E.X., Ning, Y., Liu, H.: Testing and confidence intervals for high dimensional proportional hazards models. J. R. Statist. Soc. B **79**(5), 1415–1437 (2017)

[46] Neykov, M., Ning, Y., Liu, J.S., Liu, H.: A unified theory of confidence regions and testing for high-dimensional estimating equations. Stat. Sci. **33**(3), 427–443 (2018)

[47] Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N.: High-dimensional inference: confidence intervals, p-values and R-software hdi. Stat. Sci., 533–558 (2015)

[48] Belloni, A., Chernozhukov, V., Hansen, C.: Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. **81**(2), 608–650 (2014)

[49] Chernozhukov, V., Hansen, C., Spindler, M.: Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econom. **7**(1), 649–688 (2015)

[50] Farrell, M.H.: Robust inference on average treatment effects with possibly more covariates than observations. J. Econom. **189**(1), 1–23 (2015)

[51] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. Econom. J. **21**(1), 1–68 (2018)

[52] Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C.: Program evaluation and causal inference with high-dimensional data. Econometrica **85**(1), 233–298 (2017)

[53] Cai, T., Tony Cai, T., Guo, Z.: Optimal statistical inference for individualized treatment effects in high-dimensional models. J. R. Statist. Soc. B **83**(4), 669–719 (2021)

[54] Rakshit, P., Cai, T.T., Guo, Z.: SIHR: An R package for statistical inference in high-dimensional linear and logistic regression models. arXiv preprint arXiv:2109.03365 (2021)

[55] Dezeure, R., Bühlmann, P., Zhang, C.-H.: High-dimensional simultaneous inference with the bootstrap. Test **26**(4), 685–719 (2017)

[56] Zhang, X., Cheng, G.: Simultaneous inference for high-dimensional linear models. J. Am. Statist. Assoc. **112**(518), 757–768 (2017)

[57] Ma, R., Tony Cai, T., Li, H.: Global and simultaneous hypothesis testing for high-dimensional logistic regression models. Journal of the American Statistical Association **116**(534), 984–998 (2021)

[58] Liu, W., Luo, S.: Hypothesis testing for high-dimensional regression models. Technical report (2014)

[59] Xia, Y., Cai, T., Tony Cai, T.: Two-sample tests for high-dimensional linear regression with an application to detecting interactions. Stat. Sin. **28**, 63–92 (2018)

[60] Xia, Y., Cai, T.T., Li, H.: Joint testing and false discovery rate control in high-dimensional multivariate regression. Biometrika **105**(2), 249–269 (2018)

[61] Benjamini, Y., Hochberg, Y.: Multiple hypotheses testing with weights. Scand. J. Stat. **24**(3), 407–418 (1997)

[62] Storey, J.D.: A direct approach to false discovery rates. J. Roy. Statist. Soc. B **64**(3), 479–498 (2002)

[63] Genovese, C.R., Roeder, K., Wasserman, L.: False discovery control with p-value weighting. Biometrika **93**(3), 509–524 (2006)

[64] Roeder, K., Wasserman, L.: Genome-wide significance levels and weighted hypothesis testing. Stat. Sci. **24**(4), 398 (2009)

[65] Ignatiadis, N., Klaus, B., Zaugg, J.B., Huber, W.: Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat. Methods **13**(7), 577–580 (2016)

[66] Lei, L., Fithian, W.: Adapt: an interactive procedure for multiple testing with side information. J. R. Statist. Soc. B **80**(4), 649–679 (2018)

[67] Li, A., Barber, R.F.: Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. J. R. Statist. Soc. B **81**(1), 45–74 (2019)

[68] Cai, T.T., Sun, W., Xia, Y.: LAWS: A locally adaptive weighting and screening approach to spatial multiple testing. J. Am. Statist. Assoc. **117**, 1370–1383 (2022)

[69] Fithian, W., Lei, L.: Conditional calibration for false discovery rate control under dependence. Ann. Statist. **50**(6), 3091–3118 (2022)

[70] Xia, Y., Cai, T.T., Sun, W.: GAP: A General Framework for Information Pooling in Two-Sample Sparse Inference. J. Am. Statist. Assoc. **115**(531), 1236–1250 (2020)

[71] Liu, M., Xia, Y., Cho, K., Cai, T.: Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. J. Mach. Learn. Res. **22**, 126–1 (2021)

[72] Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. Ann. Statist. **34**(3), 1436–1462 (2006)

[73] Zhao, P., Yu, B.: On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)

[74] Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inf. Theory **55**(5), 2183–2202 (2009)

[75] Javanmard, A., Montanari, A.: Debiasing the Lasso: Optimal sample size for gaussian designs. Ann. Statist. **46**(6A), 2593–2622 (2018)

[76] Guo, Z., Ćevid, D., Bühlmann, P.: Doubly debiased Lasso: High-dimensional inference under hidden confounding. Ann. Statist. **50**(3), 1320–1347 (2022)

[77] Sun, Y., Ma, L., Xia, Y.: A decorrelating and debiasing approach to simultaneous inference for high-dimensional confounded models. arXiv preprint arXiv:2208.08754 (2022)

[78] Guo, Z., Yuan, W., Zhang, C.-H.: Decorrelated local linear estimator: Inference for non-linear effects in high-dimensional additive models. arXiv preprint arXiv:1907.12732 (2019)

[79] Shi, C., Song, R., Lu, W., Li, R.: Statistical inference for high-dimensional models via recursive online-score estimation. Journal of the American Statistical Association **116**(535), 1307–1318 (2021)

[80] Luo, L., Han, R., Lin, Y., Huang, J.: Statistical inference in high-dimensional generalized linear models with streaming data. arXiv preprint arXiv:2108.04437 (2021)

[81] Fan, Q., Guo, Z., Mei, Z.: Testing overidentifying restrictions with high-dimensional data and heteroskedasticity. arXiv preprint arXiv:2205.00171 (2022)

[82] Collier, O., Comminges, L., Tsybakov, A.B.: Minimax estimation of linear and quadratic functionals on sparsity classes. Ann. Statist. **45**(3), 923–958 (2017)

[83] Guo, Z.: Inference for high-dimensional maximin effects in heterogeneous regression models using a sampling approach. arXiv preprint arXiv:2011.07568 (2020)

[84] Zhang, T.: Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE Trans. Inf. Theory **57**(7), 4689–4708 (2011)

[85] Hunter, D.J.: Gene–environment interactions in human diseases. Nat. Rev. Genet. **6**(4), 287–298 (2005)

[86] Zhou, J.J., Cho, M.H., Lange, C., Lutz, S., Silverman, E.K., Laird, N.M.: Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. Hum. Hered. **79**, 93–104 (2015)

[87] Schifano, L. E.D. Li, Christiani, D.C., Lin, X.: Genome-wide association analysis for multiple continuous secondary phenotypes. Am J Hum Genet., 744–759 (2013)

[88] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B **68**(1), 49–67 (2006)

[89] Lounici, K., Pontil, M., van de Geer, S., Tsybakov, A.B., *et al.*: Oracle inequalities and optimal inference under group sparsity. Ann. Statist. **39**(4), 2164–2204 (2011)

[90] Chen, S., Banerjee, A.: Alternating estimation for structured high-dimensional multi-response models. Advances in Neural Information Processing Systems **30** (2017)

[91] Ren, Z., Zhang, C.-H., Zhou, H.: Asymptotic normality in estimation of large ising graphical model. Unpublished Manuscript (2016)

[92] Cai, T.T., Li, H., Ma, J., Xia, Y.: Differential markov random field analysis with an application to detecting differential microbial community networks. Biometrika **106**(2), 401–416 (2019)

[93] Xia, Y., Cai, T., Tony Cai, T.: Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. J. Am. Statist. Assoc. **113**(521), 328–339 (2018)

[94] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B **57**, 289–300 (1995)

[95] Liang, Z., Cai, T.T., Sun, W., Xia, Y.: Locally adaptive transfer learning algorithms for large-scale multiple testing. arXiv preprint arXiv:2203.11461 (2022)

[96] Cai, T., Liu, M., Xia, Y.: Individual data protected integrative regression analysis of high-dimensional heterogeneous data. J. Am. Statist. Assoc. **116**, 1–15 (2021)

[97] Cai, T.T., Guo, Z.: Accuracy assessment for high-dimensional linear regression. Ann. Statist. **46**(4), 1807–1836 (2018)

[98] Nickl, R., van de Geer, S.: Confidence sets in sparse regression. Ann. Statist. **41**(6), 2852–2876 (2013)

[99] Donoho, D.L., Maleki, A., Montanari, A.: The noise-sensitivity phase transition in compressed sensing. IEEE Trans. Inf. Theory **57**(10), 6920–6941 (2011)

[100] Bayati, M., Montanari, A.: The Lasso risk for gaussian matrices. IEEE Trans. Inf. Theory **58**(4), 1997–2017 (2011)

[101] Zhang, A., Brown, L.D., Cai, T.T.: Semi-supervised inference: General theory and estimation of means. Ann. Statist. **47**(5), 2538–2566 (2019)

[102] Chakrabortty, A., Cai, T.: Efficient and adaptive linear regression in semi-supervised settings. Ann. Statist. **46**(4), 1541–1572 (2018)

[103] Zhang, Y., Chakrabortty, A., Bradic, J.: Double robust semi-supervised inference for the mean: Selection bias under mar labeling with decaying overlap. arXiv preprint arXiv:2104.06667 (2021)

[104] Deng, S., Ning, Y., Zhao, J., Zhang, H.: Optimal semi-supervised estimation and inference for high-dimensional linear regression. arXiv preprint arXiv:2011.14185 (2020)

[105] Hou, J., Guo, Z., Cai, T.: Surrogate assisted semi-supervised inference for high dimensional risk prediction. arXiv preprint arXiv:2105.01264 (2021)

[106] Mandozzi, J., Bühlmann, P.: Hierarchical testing in the high-dimensional setting with correlated variables. J. Am. Statist. Assoc. **111**(513), 331–343 (2016)

[107] Guo, Z., Li, X., Han, L., Cai, T.: Robust inference for federated meta-learning. arXiv preprint arXiv:2301.00718 (2023)

[108] Li, S., Cai, T.T., Li, H.: Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. J. R. Statist. Soc. B **84**(1), 149–173 (2021)

[109] Li, S., Zhang, L., Cai, T.T., Li, H.: Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Technical Report (2021)

[110] Tian, Y., Feng, Y.: Transfer learning under high-dimensional generalized linear models. J. Am. Statist. Assoc. **117**(just-accepted), 1–30 (2022)

[111] Battey, H., Fan, J., Liu, H., Lu, J., Zhu, Z.: Distributed testing and estimation under sparse high dimensional models. Ann. Statist. **46**(3), 1352 (2018)

[112] Lee, J.D., Liu, Q., Sun, Y., Taylor, J.E.: Communication-efficient sparse regression. J. Mach. Learn. Res. **18**(1), 115–144 (2017)

[113] Meinshausen, N., Bühlmann, P.: Maximin effects in inhomogeneous large-scale data. Ann. Statist. **43**(4), 1801–1830 (2015)

[114] Liu, W.: Gaussian graphical model estimation with false discovery rate control. Ann. Statist. **41**(6), 2948–2978 (2013)

[115] Xia, Y., Li, L.: Hypothesis testing of matrix graph model with application to brain connectivity analysis. Biometrics **73**(3), 780–791 (2017)

[116] Xia, Y., Cai, T., Tony Cai, T.: Testing differential networks with applications to the detection of gene-gene interactions. Biometrika **102**(2), 247–266 (2015)

[117] Xia, Y., Li, L.: Matrix graph hypothesis testing and application in brain connectivity alternation detection. Stat. Sin. **29**(1), 303–328 (2019)

[118] Kim, B., Liu, S., Kolar, M.: Two-sample inference for high-dimensional markov networks. J. R. Statist. Soc. B (2021)

[119] Zhang, L., Ma, R., Cai, T.T., Li, H.: Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression. arXiv preprint arXiv:2011.03598 (2020)

[120] Li, S., Cai, T.T., Li, H.: Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. J. Am. Statist. Assoc. **116**, 1–12 (2021)

[121] Barber, R.F., Candès, E.J.: Controlling the false discovery rate via knockoffs. Ann. Statist. **43**(5), 2055–2085 (2015)

[122] Candes, E., Fan, Y., Janson, L., Lv, J.: Panning for gold:'model-x' knockoffs for high dimensional controlled variable selection. J. R. Statist. Soc. B **80**(3), 551–577 (2018)

[123] Barber, R.F., Candès, E.J., Samworth, R.J.: Robust inference with knockoffs. Ann. Statist. **48**(3), 1409–1431 (2020)

[124] Ren, Z., Wei, Y., Candès, E.: Derandomizing knockoffs. J. Am. Statist. Assoc. **116**, 1–11 (2021)

[125] Du, L., Guo, X., Sun, W., Zou, C.: False discovery rate control under general dependence by symmetrized data aggregation. J. Am. Statist. Assoc. **116**, 1–15 (2021)

[126] Dai, C., Lin, B., Xing, X., Liu, J.S.: A scale-free approach for false discovery rate control in generalized linear models. J. Am. Statist. Assoc. (just-accepted), 1–31 (2023)

[127] Vovk, V., Wang, R.: E-values: Calibration, combination and applications. Ann. Statist. **49**(3), 1736–1754 (2021)

[128] Wang, R., Ramdas, A.: False discovery rate control with e-values. arXiv preprint arXiv:2009.02824 (2020)

[129] Ren, Z., Barber, R.F.: Derandomized knockoffs: leveraging e-values for false discovery rate control. arXiv preprint arXiv:2205.15461 (2022)