Introduction to Methodologies of SIHR

Zhenyu(Zach) Wang

April 11, 2024

The package `SIHR` aims to perform statistical inference in high-dimensional generalized linear models with continuous and binary outcomes. It provides tools for constructing confidence intervals and performing hypothesis tests for low-dimensional objectives in both one-sample and two-sample regression settings.

1 Introduction

We consider the high-dimensional GLMs: for $1 \leq i \leq n$,

$$\mathbb{E}(y_i \mid X_{i\cdot}) = f(X_{i\cdot}^\intercal \beta), \quad \text{with } f(z) = \begin{cases} z & \text{for linear model;} \\ \exp(z)/\left[1+\exp(z)\right] & \text{for logistic model;} \end{cases} \tag{1}$$

where $\beta \in \mathbb{R}^p$ denotes the high-dimensional regression vector, $y_i \in \mathbb{R}$ and $X_{i\cdot} \in \mathbb{R}^p$ denote respectively the outcome and the measured covariates of the i-th observation. Throughout the paper, define $\Sigma = \mathbb{E}X_{i\cdot}X_{i\cdot}^\intercal$ and assume β to be a sparse vector with its sparsity level denoted as $\|\beta\|_0$. In addition to the one-sample setting, we examine the statistical inference methods for the two-sample regression models. Particularly, we generalize the regression model in (1) and consider:

$$\mathbb{E}(y_i^{(k)} \mid X_{i\cdot}^{(k)}) = f(X_{i\cdot}^{(k)\intercal}\beta^{(k)}) \quad \text{with } k = 1, 2 \text{ and } 1 \leq i \leq n_k, \tag{2}$$

where $f(\cdot)$ is the pre-specified link function defined as (1), $\beta^{(k)} \in \mathbb{R}^p$ denotes the high-dimensional regression vector in k-th sample, $y_i^{(k)} \in \mathbb{R}$ and $X_{i\cdot}^{(k)} \in \mathbb{R}^p$ denote respectively the outcome and the measured covariates in the k-th sample.

1.1 Package Components

This package consists of five main functions `LF`, `QF`, `CATE`, `InnProd`, and `Dist` implementing the statistical inferences for five different quantities, under the one-sample model (1) or two-sample model (2).

1. `LF`, abbreviated for linear functional, implements the inference approach for $x_{\text{new}}^\intercal\beta$, with $x_{\text{new}} \in \mathbb{R}^p$ denoting a loading vector. With $x_{\text{new}} = e_j$ as a special case, `LF` infers the regression coefficient β_j.

2. `QF`, abbreviated for quadratic functional, makes inferences for $\beta^\intercal A\beta$. A is either a pre-specified sub-matrix or the unknown covariance matrix Σ.

3. `CATE`, abbreviated for conditional average treatment effect, is to make inference for $f(x_{\text{new}}^\intercal\beta^{(2)}) - f(x_{\text{new}}^\intercal\beta^{(1)})$. This difference measures the discrepancy between conditional means, closely related to the conditional average treatment effect for the new observation with covariates x_{new}.

4. `InnProd`, abbreviated for inner products, implements the statistical inference for $\beta^{(1)\intercal}A\beta^{(2)}$. The inner products measure the similarity between the high-dimensional vectors $\beta^{(1)}$ and $\beta^{(2)}$, which is useful in capturing the genetic relatedness in the GWAS applications.

5. `Dist`, short-handed for distance, makes inferences for the weighted distances $\gamma^\intercal A\gamma$ with $\gamma = \beta^{(2)} - \beta^{(1)}$. The distance measure is useful in comparing different high-dimensional regression vectors.

1.2 Outlines

In section 2.2, we propose a unified inference method for $x_{\text{new}}^\intercal \beta$ under linear and logistic outcome models. We also discuss inferences for quadratic functionals $\beta_{\text{G}}^\intercal A \beta_{\text{G}}$ and $\beta_{\text{G}}^\intercal \Sigma_{\text{G,G}} \beta_{\text{G}}$ in section 2.3. In the case of the two-sample high-dimensional regression model (2), we develop the inference method for conditional treatment effect $\Delta(x_{\text{new}}) = f(x_{\text{new}}^\intercal \beta^{(2)}) - f(x_{\text{new}}^\intercal \beta^{(1)})$ in section 2.4; we consider inference for $\beta_{\text{G}}^{(1)\intercal} A \beta_{\text{G}}^{(2)}$ and $\beta_{\text{G}}^{(1)\intercal} \Sigma_{\text{G,G}} \beta_{\text{G}}^{(2)}$ in section 2.5 and $\gamma_{\text{G}}^\intercal A \gamma_{\text{G}}$ and $\gamma_{\text{G}}^\intercal \Sigma_{\text{G,G}} \gamma_{\text{G}}$ with $\gamma = \beta^{(2)} - \beta^{(1)}$ in section 2.6.

2 Methodologies

We briefly review the penalized maximum likelihood estimator of β in the high-dimensional GLM (1), defined as:

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

with $X_{\cdot j}$ denoting the j-th column of X, the first column of X set as the constant 1, and

$$\ell(\beta) = \begin{cases} \frac{1}{n} \sum_{i=1} \left(y_i - X_{i\cdot}^\intercal \beta\right)^2 & \text{for linear model} \\ -\frac{1}{n} \sum_{i=1}^n y_i \log\left[\frac{f(X_{i\cdot}^\intercal \beta)}{1-f(X_{i\cdot}^\intercal \beta)}\right] - \frac{1}{n} \sum_{i=1}^n \log\left(1 - f(X_{i\cdot}^\intercal \beta)\right) & \text{for GLM with binary outcome.} \end{cases}. \tag{4}$$

The tuning parameter $\lambda_0 \asymp \sqrt{\log p/n}$ is chosen by cross-validation. In the penalized regression (3), we do not penalize the intercept coefficient β_1. The penalized estimators have been shown to achieve the optimal convergence rates and satisfy desirable variable selection properties [10, 1, 14, 12]. However, these estimators are not ready for statistical inference due to the non-negligible estimation bias induced by the penalty term [11, 8, 13].

2.1 Linear functional for linear model

To illustrate the idea of constructing the inference method, we start with the linear functional for the linear model, which will be extended to a unified version in the section 2.2. For the linear model in (1), we define $\epsilon_i = y_i - X_{i\cdot}^\intercal \beta$ and rewrite the model as $y_i = X_{i\cdot}^\intercal \beta + \epsilon_i$ for $1 \le i \le n$. Given the vector $x_{\text{new}} \in \mathbb{R}^p$, we construct the point estimator and the CI for $x_{\text{new}}^\intercal \beta$.

A natural idea for the point estimator is to use the plug-in estimator $x_{\text{new}}^\intercal \widehat{\beta}$ with the penalized estimator $\widehat{\beta}$ defined in (3). However, the bias $x_{\text{new}}^\intercal(\widehat{\beta} - \beta)$ is not negligible. The work Cai et al. [3] proposed the bias-corrected estimator as,

$$\widehat{x_{\text{new}}^\intercal \beta} = x_{\text{new}}^\intercal \widehat{\beta} + \widehat{u}^\intercal \frac{1}{n} \sum_{i=1}^n X_{i\cdot} \left(y_i - X_{i\cdot}^\intercal \widehat{\beta}\right), \tag{5}$$

where the second term on the right hand side in (5) is the estimate of negative bias $-x_{\text{new}}^\intercal(\widehat{\beta} - \beta)$, and the projection direction $\widehat{u}$ is defined as

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^\intercal \widehat{\Sigma} u \quad \text{subject to:} \quad \|\widehat{\Sigma} u - x_{\text{new}}\|_\infty \le \|x_{\text{new}}\|_2 \lambda \tag{6}$$

$$\left| x_{\text{new}}^\intercal \widehat{\Sigma} u - \|x_{\text{new}}\|_2^2 \right| \le \|x_{\text{new}}\|_2^2 \lambda, \tag{7}$$

where $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_{i\cdot} X_{i\cdot}^\intercal$ and $\lambda \asymp \sqrt{\log p/n}$. The bias-corrected estimator $\widehat{x_{\text{new}}^\intercal \beta}$ satisfies the following error decomposition,

$$\widehat{x_{\text{new}}^\intercal \beta} - x_{\text{new}}^\intercal \beta = \underbrace{\widehat{u}^\intercal \frac{1}{n} \sum_{i=1}^n X_{i\cdot}^\intercal \epsilon_i}_{\text{asymp. normal}} + \underbrace{\left(\widehat{\Sigma}\widehat{u} - x_{\text{new}}\right)^\intercal (\beta - \widehat{\beta})}_{\text{remaining bias}}.$$

The first constraint in (6) controls the remaining bias term in the above equation while the second constraint in (7) is crucial to ensuring the asymptotic normality of $\widehat{x_{\text{new}}^\intercal \beta} - x_{\text{new}}^\intercal \beta$ for any vector x_{new} such that the

variance of the "asymp. normal" term always dominates the "remaining bias" term. Based on the asymptotic normality, we construct the CI for $x_{\text{new}}^\intercal \beta$ as

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

where $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - X_{i\cdot}^\intercal \widehat{\beta})^2$ and $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile for the standard normal distribution.

2.2 Linear functional for GLM

In this subsection, we generalize the inference method specifically for the linear model in Section 2.1 to GLM in (1). Given the initial estimator $\widehat{\beta}$, the key step is to estimate the bias $x_{\text{new}}^\intercal (\widehat{\beta} - \beta)$. We can propose a unified version of the bias-corrected estimator for $x_{\text{new}}^\intercal \beta$ as

$$\widehat{x_{\text{new}}^\intercal \beta} = x_{\text{new}}^\intercal \widehat{\beta} + \widehat{u}^\intercal \frac{1}{n} \sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta}) \left(y_i - f(X_{i\cdot}^\intercal \widehat{\beta}) \right) X_{i\cdot}, \tag{8}$$

with the second term on the right hand side of (8) being the estimate of $-x_{\text{new}}^\intercal (\widehat{\beta} - \beta)$. In consideration of different link functions $f(\cdot)$ in (1), we shall specify in the following how to construct the projection direction $\hat{u}$ and the weight function $\omega : \mathbb{R} \mapsto \mathbb{R}$ in (8). In Table 1, we consider different GLM models and present

Model	Outcome Type	$f(z)$	$f'(z)$	$\omega(z)$	Weighting
linear	Continuous	z	1	1	
logistic	Binary	$\frac{e^z}{1+e^z}$	$\frac{e^z}{(1+e^z)^2}$	$\frac{(1+e^z)^2}{e^z}$	Linearization
logistic_alter	Binary	$\frac{e^z}{1+e^z}$	$\frac{e^z}{(1+e^z)^2}$	1	Link-specific

Table 1: Definitions of the functions ω and f for different GLMs.

the corresponding functions $f(\cdot)$ and $\omega(\cdot)$, together with the derivative $f'(\cdot)$. Note that there are two ways of specifying the weights $w(z)$ for logistic regression. The linearization weighting is proposed in Guo et al. [7] specifically for logistic regression; while Cai et al. [4] constructed the link-specific weighting method for general link function $f(\cdot)$. The projection direction $\widehat{u} \in \mathbb{R}^p$ in (8) is constructed as follows:

$$\begin{aligned} \widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^\intercal \left[\frac{1}{n} \sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta}) f'(X_{i\cdot}^\intercal \widehat{\beta}) X_{i\cdot} X_{i\cdot}^\intercal \right] u \quad & \text{subject to:} \\ \left\| \frac{1}{n} \sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta}) f'(X_{i\cdot}^\intercal \widehat{\beta}) X_{i\cdot} X_{i\cdot}^\intercal u - x_{\text{new}} \right\|_\infty & \le \|x_{\text{new}}\|_2 \lambda \\ \left| x_{\text{new}}^\intercal \frac{1}{n} \sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta}) f'(X_{i\cdot}^\intercal \widehat{\beta}) X_{i\cdot} X_{i\cdot}^\intercal u - \|x_{\text{new}}\|_2^2 \right| & \le \|x_{\text{new}}\|_2^2 \lambda. \end{aligned} \tag{9}$$

It has been established that $\widehat{x_{\text{new}}^\intercal \beta}$ in (8) is asymptotically unbiased and normal for the linear model [3], the logistic model [6, 4], and the probit model [4]. The variance of $\widehat{x_{\text{new}}^\intercal \beta}$ can be estimated by $\widehat{\text{V}}$, defined as

$$\widehat{\text{V}} = \widehat{u}^\intercal \left[\frac{1}{n^2} \sum_{i=1}^n \left(\omega(X_{i\cdot}^\intercal \widehat{\beta}) \right)^2 \widehat{\sigma}_i^2 X_{i\cdot} X_{i\cdot}^\intercal \right] \widehat{u} \quad \text{with :} \tag{10}$$

$$\widehat{\sigma}_i^2 = \begin{cases} \frac{1}{n} \sum_{j=1}^n \left(y_j - X_{j\cdot}^\intercal \widehat{\beta} \right)^2, & \text{for linear model} \\ f(X_{i\cdot}^\intercal \widehat{\beta})(1 - f(X_{i\cdot}^\intercal \widehat{\beta})), & \text{for GLM with binary outcome.} \end{cases} . \tag{11}$$

Based on the asymptotic normality, the CI for $x_{\text{new}}^\intercal \beta$ is:

$$\text{CI} = \left(\widehat{x_{\text{new}}^\intercal \beta} - z_{\alpha/2} \sqrt{\widehat{\text{V}}}, \quad \widehat{x_{\text{new}}^\intercal \beta} + z_{\alpha/2} \sqrt{\widehat{\text{V}}} \right).$$

Subsequently, for the binary outcome case, we estimate the case probability $\mathbb{P}(y_i = 1 \mid X_{i\cdot} = x_{\text{new}})$ by $f(\widehat{x_{\text{new}}^\intercal \beta})$ and construct the CI for $f(x_{\text{new}}^\intercal \beta)$ as:

$$\text{CI} = \left(f\left(\widehat{x_{\text{new}}^\intercal \beta} - z_{\alpha/2}\sqrt{\widehat{\text{V}}}\right), f\left(\widehat{x_{\text{new}}^\intercal \beta} + z_{\alpha/2}\sqrt{\widehat{\text{V}}}\right)\right).$$

2.3 Quadratic functional for GLM

We now move our focus to inference for the quadratic functional $\text{Q}_A = \beta_{\text{G}}^\intercal A \beta_{\text{G}}$, where $G \subset \{1, ..., p\}$ and $A \in \mathbb{R}^{|G|\times|G|}$ denotes a pre-specified matrix of interest. Without loss of generality, we set $G = \{1, 2, \cdots, |G|\}$. In the following, we propose a unified version of the point estimator and CI under the GLM (1). With the initial estimator $\widehat{\beta}$ defined in (3), the plug-in estimator $\widehat{\beta}_{\text{G}}^\intercal A \widehat{\beta}_{\text{G}}$ suffers from the following error,

$$\widehat{\beta}_{\text{G}}^\intercal A \widehat{\beta}_{\text{G}} - \beta_{\text{G}}^\intercal A \beta_{\text{G}} = 2\widehat{\beta}_{\text{G}}^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}}) - (\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}}).$$

The last term in the above decomposition $(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$ is the higher-order approximation error under regular conditions; thus the bias mainly comes from the term $2\widehat{\beta}_{\text{G}}^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$, which can be expressed as $2\, x_{\text{new}}^\intercal(\widehat{\beta} - \beta)$ with $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^\intercal A,\ \mathbf{0}^\intercal)^\intercal$. Hence the term can be estimated directly by applying the linear functional approach in section 2.2. Utilizing this idea, Guo et al. [7, 5] proposed the following estimator of Q_A,

$$\widehat{\text{Q}}_A = \widehat{\beta}_{\text{G}}^\intercal A \widehat{\beta}_{\text{G}} + 2\,\widehat{u}_A^\intercal \left[\frac{1}{n}\sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta})\left(y_i - f(X_{i\cdot}^\intercal \widehat{\beta})\right) X_{i\cdot}\right],$$

with the second term being the estimate of $-2\widehat{\beta}_{\text{G}}^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$, where $\widehat{u}_A$ is the projection direction defined in (9) with $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^\intercal A,\ \mathbf{0}^\intercal)^\intercal$. Since Q_A is non-negative if A is positive semi-definite, we truncate $\widehat{\text{Q}}_A$ at 0 and define $\widehat{\text{Q}}_A = \max\left(\widehat{\text{Q}}_A,\ 0\right)$. We further estimate the variance of the $\widehat{\text{Q}}_A$ by

$$\widehat{\text{V}}_A(\tau) = 4\widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n \omega^2(X_{i\cdot}^\intercal \widehat{\beta})\widehat{\sigma}_i^2 X_{i\cdot} X_{i\cdot}^\intercal\right]\widehat{u} + \frac{\tau}{n}, \tag{12}$$

where the term τ/n with $\tau > 0$ (default value $\tau = 1$) is introduced as an upper bound for the term $(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$, and $\widehat{\sigma}_i^2$ is defined in (11). Then given a fixed value of τ, we construct the CI as $\text{CI}(\tau) = \left(\max\left(\widehat{\text{Q}}_A - z_{\alpha/2}\sqrt{\widehat{\text{V}}_A(\tau)},\ 0\right),\ \widehat{\text{Q}}_A + z_{\alpha/2}\sqrt{\widehat{\text{V}}_A(\tau)}\right)$.

Now we turn to the estimation of $\text{Q}_\Sigma = \beta_{\text{G}}^\intercal \Sigma_{\text{G,G}} \beta_{\text{G}}$ where the matrix $\Sigma_{\text{G,G}}$ is unknown and estimated by $\widehat{\Sigma}_{\text{G,G}} = \frac{1}{n}\sum_{i=1}^n X_{i\text{G}} X_{i\text{G}}^\intercal$. Decompose the error of the plug-in estimator $\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}} \widehat{\beta}$:

$$\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}} \widehat{\beta} - \beta_{\text{G}} \Sigma_{\text{G,G}} \beta_{\text{G}} = 2\,\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}}(\widehat{\beta}_{\text{G}} - \beta_{\text{G}}) + \beta_{\text{G}}^\intercal(\widehat{\Sigma}_{\text{G,G}} - \Sigma_{\text{G,G}})\beta_{\text{G}} - (\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal \widehat{\Sigma}_{\text{G,G}}(\widehat{\beta}_{\text{G}} - \beta_{\text{G}}).$$

The first term $\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}}(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$ is estimated by applying linear functional approach in Section 2.2 with $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}},\ \mathbf{0})^\intercal$; the second term $\beta_{\text{G}}^\intercal(\widehat{\Sigma}_{\text{G,G}} - \Sigma_{\text{G,G}})\beta_{\text{G}}$ can be controlled asymptotically by central limit theorem; and the last term $(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal \widehat{\Sigma}_{\text{G,G}}(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$ is negligible due to high-order bias. Guo et al. [7] proposed the following estimator of Q_Σ

$$\widehat{\text{Q}}_\Sigma = \widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}} \widehat{\beta}_{\text{G}} + 2\,\widehat{u}_\Sigma^\intercal \left[\frac{1}{n}\sum_{i=1}^n \omega(X_{i\cdot}^\intercal \widehat{\beta})\left(y_i - f(X_{i\cdot}^\intercal \widehat{\beta})\right) X_{i\cdot}\right],$$

where $\widehat{u}_\Sigma$ is the projection direction constructed in (9) with $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}},\ \mathbf{0})^\intercal$. We introduce the estimator $\widehat{\text{Q}}_\Sigma = \max(\widehat{\text{Q}}_\Sigma,\ 0)$ and estimate its variance as

$$\widehat{\text{V}}_\Sigma(\tau) = 4\widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n \omega^2(X_{i\cdot}^\intercal \widehat{\beta})\widehat{\sigma}_i^2 X_{i\cdot} X_{i\cdot}^\intercal\right]\widehat{u} + \frac{1}{n^2}\sum_{i=1}^n \left(\widehat{\beta}_{\text{G}}^\intercal X_{i,G} X_{i,G}^\intercal \widehat{\beta}_{\text{G}} - \widehat{\beta}_{\text{G}}^\intercal \widehat{\Sigma}_{\text{G,G}} \widehat{\beta}_{\text{G}}\right)^2 + \frac{\tau}{n}, \tag{13}$$

where $\tau > 0$, the term τ/n is introduced as an upper bound for the term $(\widehat{\beta}_{\mathrm{G}} - \beta_{\mathrm{G}})^{\intercal}\widehat{\Sigma}_{\mathrm{G,G}}(\widehat{\beta}_{\mathrm{G}} - \beta_{\mathrm{G}})$, and $\widehat{\sigma}_i^2$ is defined in (11). Then, thanks to the asymptotic normality, for a fixed value of τ, we can construct the CI as

$$\mathrm{CI}(\tau) = \left(\max\left(\widehat{\mathrm{Q}}_{\Sigma} - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Sigma}(\tau)},\ 0\right),\ \widehat{\mathrm{Q}}_{\Sigma} + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Sigma}(\tau)}\right).$$

2.4 Conditional average treatment effects

The inference methods proposed for one sample can be generalized to make inferences for conditional average treatment effects, which can be expressed as the difference between two linear functionals. Let $A_i \in \{1, 2\}$ denote the treatment assignment for i-th observation. Consider the two-sample GLMs as

$$\mathbb{E}(y_i|X_{i\cdot}, A_i = 1) = f(X_{i\cdot}^{\intercal}\beta^{(1)}) \quad \text{and} \quad \mathbb{E}(y_i|X_{i\cdot}, A_i = 2) = f(X_{i\cdot}^{\intercal}\beta^{(2)}),$$

where f is the link function listed in table 1. Then, for a future individual $X_{i\cdot} = x_{\mathrm{new}}$, we define $\Delta(x_{\mathrm{new}}) = \mathbb{E}(y_i|X_{i\cdot}, A_i = 2) - \mathbb{E}(y_i|X_{i\cdot}, A_i = 1)$, that measures the difference of the conditional mean of assignment of treatment for the individual with covariates x_{new}.

Following (8), we construct the bias-corrected point estimators of $\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(1)}}$ and $\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(2)}}$, together with their corresponding variance $\widehat{\mathrm{V}}_{(1)}$ and $\widehat{\mathrm{V}}_{(2)}$ as (10). The paper Cai et al. [3] proposed to estimate $\Delta(x_{\mathrm{new}})$ by $\widehat{\Delta}(x_{\mathrm{new}})$ as:

$$\widehat{\Delta}(x_{\mathrm{new}}) = f(\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(2)}}) - f(\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(1)}}).$$

Its variance can be estimated with delta method by:

$$\widehat{\mathrm{V}}_{\Delta} = \left(f'(\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(1)}})\right)^2 \widehat{\mathrm{V}}_{(1)} + \left(f'(\widehat{x_{\mathrm{new}}^{\intercal}\beta^{(2)}})\right)^2 \widehat{\mathrm{V}}_{(2)}.$$

Then we construct the CI as $\mathrm{CI} = \left(\widehat{\Delta}(x_{\mathrm{new}}) - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Delta}}, \widehat{\Delta}(x_{\mathrm{new}}) + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Delta}}\right)$.

2.5 Inner product of regression vectors

The paper Guo et al. [5], Ma et al. [9] have carefully investigated the CI construction for $\beta_{\mathrm{G}}^{(1)\intercal}A\beta_{\mathrm{G}}^{(2)}$, provided with a pre-specified submatrix $A \in \mathbb{R}^{|\mathrm{G}|\times|\mathrm{G}|}$ and the set of indices $\mathrm{G} \in \{1, ..., p\}$. Let $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ respectively be the initial estimators for their corresponding sample in (2), the plug-in but biased estimator is $\widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\widehat{\beta}_{\mathrm{G}}^{(2)}$. Its bias can be decomposed as:

$$\begin{aligned}\widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\widehat{\beta}_{\mathrm{G}}^{(2)} - \beta_{\mathrm{G}}^{(1)\intercal}A\beta_{\mathrm{G}}^{(2)} = \widehat{\beta}_{\mathrm{G}}^{(2)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(1)} - \beta_{\mathrm{G}}^{(1)}\right) + \widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(2)} - \beta_{\mathrm{G}}^{(2)}\right) \\ - \left(\widehat{\beta}_{\mathrm{G}}^{(1)} - \beta_{\mathrm{G}}^{(1)}\right)^{\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(2)} - \beta_{\mathrm{G}}^{(2)}\right).\end{aligned}$$

The key step is to estimate the error components $\widehat{\beta}_{\mathrm{G}}^{(2)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(1)} - \beta_{\mathrm{G}}^{(1)}\right)$ and $\widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(2)} - \beta_{\mathrm{G}}^{(2)}\right)$. Then the following procedures can be interpreted as applying Linear Functional twice on two independent samples. To be specific, we propose the following bias-corrected estimator for $\beta_{\mathrm{G}}^{(1)\intercal}A\beta_{\mathrm{G}}^{(2)}$

$$\begin{aligned}\widehat{\beta_{\mathrm{G}}^{(1)\intercal}A\beta_{\mathrm{G}}^{(2)}} = \widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\widehat{\beta}_{\mathrm{G}}^{(2)} + \widehat{u}_1^{\intercal}\frac{1}{n_1}\sum_{i=1}^{n_1}\omega(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)})\left(y_i^{(1)} - f(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)})\right)X_{i\cdot}^{(1)} \\ + \widehat{u}_2^{\intercal}\frac{1}{n_2}\sum_{i=1}^{n_2}\omega(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)})\left(y_i^{(2)} - f(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)})\right)X_{i\cdot}^{(2)},\end{aligned} \tag{14}$$

with the second term and the third term in right-hand-side of (14) estimating $-\widehat{\beta}_{\mathrm{G}}^{(2)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(1)} - \beta_{\mathrm{G}}^{(1)}\right)$ and $-\widehat{\beta}_{\mathrm{G}}^{(1)\intercal}A\left(\widehat{\beta}_{\mathrm{G}}^{(2)} - \beta_{\mathrm{G}}^{(2)}\right)$ respectively, where $\widehat{u}_1$ is the projection direction defined in (9) with $x_{\mathrm{new}} = (\widehat{\beta}_{\mathrm{G}}^{(2)\intercal}A,\ \mathbf{0})^{\intercal}$

and $\widehat{u}_2$ is the projection direction defined in (9) with $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^{(1)\intercal} A,\ \mathbf{0})^\intercal$. The corresponding variance of $\widehat{\beta_{\text{G}}^{(1)\intercal} A \beta_{\text{G}}^{(2)}}$, when A is a known positive definite matrix, is estimated as

$$\widehat{\text{V}}_A(\tau) = \widehat{\text{V}}^{(1)} + \widehat{\text{V}}^{(2)} + \frac{\tau}{\min(n_1, n_2)},$$

where $\widehat{V}^{(k)}$ is computed as (10) for the $k-$th regression model ($k = 1, 2$) in (2) and $\tau > 0$, the term $\tau/\min(n_1, n_2)$ is introduced as an upper bound for the term $(\widehat{\beta}_{\text{G}}^{(1)} - \beta_{\text{G}}^{(1)})^\intercal A(\widehat{\beta}_{\text{G}}^{(2)} - \beta_{\text{G}}^{(2)})$.

When A is not specified, we treat $A = \Sigma_{\text{G,G}}$, which is unknown. As a natural generalization, the quantity $\beta_{\text{G}}^{(1)\intercal}\Sigma_{\text{G,G}}\beta_{\text{G}}^{(2)}$ is well defined if the two regression models in (2) share the design covariance matrix $\Sigma = \mathbb{E}X_{i\cdot}^{(1)}X_{i\cdot}^{(1)\intercal} = \mathbb{E}X_{i\cdot}^{(2)}X_{i\cdot}^{(2)\intercal}$. We follow the above procedures replacing A by $\widehat{\Sigma}_{\text{G,G}} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_{i,\text{G}}X_{i,\text{G}}^\intercal$ where X is the row-combined matrix of $X^{(1)}$ and $X^{(2)}$. The variance of $\widehat{\beta_{\text{G}}^{(1)\intercal}\Sigma_{\text{G,G}}\beta_{\text{G}}^{(2)}}$ is now estimated as

$$\widehat{\text{V}}_\Sigma(\tau) = \widehat{\text{V}}^{(1)} + \widehat{\text{V}}^{(2)} + \frac{1}{(n_1+n_2)^2}\sum_{i=1}^{n_1+n_2}\left(\widehat{\beta}_{\text{G}}^{(1)\intercal}X_{i,\text{G}}X_{i,\text{G}}^\intercal\widehat{\beta}_{\text{G}}^{(2)} - \widehat{\beta}_{\text{G}}^{(1)\intercal}\widehat{\Sigma}_{\text{G,G}}\widehat{\beta}_{\text{G}}^{(2)}\right)^2 + \frac{\tau}{\min(n_1, n_2)}.$$

Depending on whether the submatrix A is specified or not, the CI is

$$\text{CI}(\tau) = \begin{cases} \left(\widehat{\beta_{\text{G}}^{(1)\intercal}A\beta_{\text{G}}^{(2)}} - z_{\alpha/2}\widehat{\text{V}}_A(\tau),\ \ \widehat{\beta_{\text{G}}^{(1)\intercal}A\beta_{\text{G}}^{(2)}} + z_{\alpha/2}\widehat{\text{V}}_A(\tau)\right) & \text{if } A \text{ is specified} \\ \left(\widehat{\beta_{\text{G}}^{(1)\intercal}\Sigma_{\text{G,G}}\beta_{\text{G}}^{(2)}} - z_{\alpha/2}\widehat{\text{V}}_\Sigma(\tau),\ \ \widehat{\beta_{\text{G}}^{(1)\intercal}\Sigma_{\text{G,G}}\beta_{\text{G}}^{(2)}} + z_{\alpha/2}\widehat{\text{V}}_\Sigma(\tau)\right) & \text{otherwise.} \end{cases}$$

2.6 Distance of regression vectors

We denote $\gamma = \beta^{(2)} - \beta^{(1)}$ and its initial estimator $\widehat{\gamma} = \widehat{\beta}^{(2)} - \widehat{\beta}^{(1)}$. The quantity of interest is the distance between two regression vectors $\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}$, given a pre-specified submatrix $A \in \mathbb{R}^{|\text{G}|\times|\text{G}|}$ and the set of indices $\text{G} \in \{1, ..., p\}$. The bias of the plug-in estimator $\widehat{\gamma}_{\text{G}}^\intercal A\widehat{\gamma}_{\text{G}}$ is:

$$\widehat{\gamma}_{\text{G}}^\intercal A\widehat{\gamma}_{\text{G}} - \gamma_{\text{G}}^\intercal A\gamma_{\text{G}} = 2\,\widehat{\gamma}_{\text{G}}^\intercal A\left(\widehat{\beta}_{\text{G}}^{(2)} - \beta_{\text{G}}^{(2)}\right) - 2\,\widehat{\gamma}_{\text{G}}^\intercal A\left(\widehat{\beta}_{\text{G}}^{(1)} - \beta_{\text{G}}^{(1)}\right) - (\widehat{\gamma}_{\text{G}} - \gamma_{\text{G}})^\intercal A\,(\widehat{\gamma}_{\text{G}} - \gamma_{\text{G}})\,.$$

The key step is to estimate the error components $\widehat{\gamma}_{\text{G}}^\intercal A\left(\widehat{\beta}_{\text{G}}^{(1)} - \beta_{\text{G}}^{(1)}\right)$ and $\widehat{\gamma}_{\text{G}}^\intercal A\left(\widehat{\beta}_{\text{G}}^{(2)} - \beta_{\text{G}}^{(2)}\right)$ in the above decomposition. We apply linear functional techniques twice here, and propose the bias-corrected estimator:

$$\begin{aligned}\widehat{\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}} = \widehat{\gamma}_{\text{G}}^\intercal A\widehat{\gamma}_{\text{G}} &- 2\,\widehat{u}_1^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\omega(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)})\left(y_i^{(1)} - f(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)})\right)X_{i\cdot}^{(1)} \\ &+ 2\,\widehat{u}_2^\intercal\frac{1}{n_2}\sum_{i=1}^{n_2}\omega(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)})\left(y_i^{(2)} - f(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)})\right)X_{i\cdot}^{(2)}.\end{aligned} \tag{15}$$

Then by non-negative distance, we define $\widehat{\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}} = \max\left\{\widehat{\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}},\ 0\right\}$. The second term on right-hand-side of (15) is to estimate $-2\,x_{\text{new}}^\intercal(\widehat{\beta}_{\text{G}}^{(1)} - \beta_{\text{G}}^{(1)})$ with $x_{\text{new}} = (\widehat{\gamma}_{\text{G}}^\intercal A, \mathbf{0})^\intercal$; and the third term on right-hand-side of (15) is to estimate $-2\,x_{\text{new}}^\intercal(\widehat{\beta}_{\text{G}}^{(2)} - \beta_{\text{G}}^{(2)})$ with $x_{\text{new}} = (\widehat{\gamma}_{\text{G}}^\intercal A, \mathbf{0})^\intercal$ as well. The corresponding asymptotic variance for the bias-corrected estimator is

$$\widehat{\text{V}}_A(\tau) = 4\,\widehat{\text{V}}^{(1)} + 4\,\widehat{\text{V}}^{(2)} + \frac{\tau}{\min(n_1, n_2)},$$

where $\widehat{\text{V}}^{(k)}$ is computed as (10) for the k-th regression model ($k = 1, 2$) and $\tau > 0$, the term $\tau/\min(n_1, n_2)$ is introduced as an upper bound for the term $(\widehat{\gamma}_{\text{G}} - \gamma_{\text{G}})^\intercal A(\widehat{\gamma}_{\text{G}} - \gamma_{\text{G}})$. With asymptotic normality, we construct the CI

$$\text{CI}(\tau) = \left(\max\left(\widehat{\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}} - z_{\alpha/2}\sqrt{\widehat{\text{V}}_A(\tau)},\ 0\right),\ \widehat{\gamma_{\text{G}}^\intercal A\gamma_{\text{G}}} + z_{\alpha/2}\sqrt{\widehat{\text{V}}_A(\tau)}\right).$$

When the submatrix A is not specified, we treat $A = \Sigma_{\mathrm{G,G}}$, which is unknown. The point estimator $\widehat{\gamma_{\mathrm{G}}^{\top}\Sigma_{\mathrm{G,G}}\gamma_{\mathrm{G}}}$ can be computed similarly as outlined in (15). In this case, the submatrix A is substituted with $\widehat{\Sigma}_{\mathrm{G,G}}$ and the resulting value is truncated at 0, where $\widehat{\Sigma}_{\mathrm{G,G}} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_{i,\mathrm{G}}X_{i,\mathrm{G}}^{\intercal}$ with X as the row-combined matrix of $X^{(1)}$ and $X^{(2)}$. Its corresponding asymptotic variance is

$$\widehat{\mathrm{V}}_{\Sigma} = 4\,\widehat{\mathrm{V}}^{(1)} + 4\,\widehat{\mathrm{V}}^{(2)} + \frac{1}{(n_1+n_2)^2}\sum_{i=1}^{n_1+n_2}\left(\widehat{\gamma}_{\mathrm{G}}^{\intercal}X_{i,\mathrm{G}}X_{i,\mathrm{G}}^{\intercal}\widehat{\gamma}_{\mathrm{G}} - \widehat{\gamma}_{\mathrm{G}}^{\intercal}\widehat{\Sigma}_{\mathrm{G,G}}\widehat{\gamma}_{\mathrm{G}}\right)^2 + \frac{\tau}{\min(n_1,n_2)}.$$

Next we present its CI

$$\mathrm{CI}(\tau) = \left(\max\left(\widehat{\gamma_{\mathrm{G}}^{\intercal}\Sigma\gamma_{\mathrm{G}}} - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Sigma}(\tau)},\ 0\right),\ \widehat{\gamma_{\mathrm{G}}^{\intercal}\Sigma\gamma_{\mathrm{G}}} + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_{\Sigma}(\tau)}\right).$$

3 Others

3.1 Construction of Projection Direction

The construction of projection directions are key to the bias correction step, see (8). In the following, we introduce the equivalent dual problem of constructing the projection direction. The constrained optimizer $\widehat{u} \in \mathbb{R}^p$ can be computed in the form of $\widehat{u} = -\frac{1}{2}\left[\widehat{\mathbf{v}}_{-1} + \frac{x_*}{\|x_*\|_2}\widehat{\mathbf{v}}_1\right]$, where, $\widehat{\mathbf{v}} \in \mathbb{R}^{p+1}$ is defined as

$$\widehat{\mathbf{v}} = \underset{\mathbf{v}\in\mathbb{R}^{p+1}}{\arg\min}\left\{\frac{1}{4n}\mathbf{v}^{\intercal}\mathbf{H}^{\intercal}\mathrm{X}^{\intercal}\mathrm{Diag}(\mathbf{w})\mathrm{Diag}(\mathbf{f}')\mathrm{X}\mathbf{H}\mathbf{v} + x_{\mathrm{new}}^{\intercal}\mathbf{H}\mathbf{v} + \lambda_n\|x_{\mathrm{new}}\|_2\cdot\|\mathbf{v}\|_1\right\}, \tag{16}$$

with $\mathbf{H} = \left[\frac{x_{\mathrm{new}}}{\|x_{\mathrm{new}}\|_2}, \mathbf{I}_{p\times p}\right] \in \mathbb{R}^{p\times(p+1)}$, $\mathbf{w} = \left(\omega(X_1^{\intercal}\widehat{\beta}), ..., \omega(X_n^{\intercal}\widehat{\beta})\right)^{\intercal}$ and $\mathbf{f}' = \left(f'(X_1^{\intercal}\widehat{\beta}), ..., f'(X_n^{\intercal}\widehat{\beta})\right)^{\intercal}$. We refer to Proposition 2 in Cai et al. [2] for the detailed derivation of the dual problem (16). In this dual problem, when $\widehat{\Sigma}$ is singular and the tuning parameter $\lambda_n > 0$ gets sufficiently close to 0, the dual problem cannot be solved as the minimum value converges to negative infinity. Hence we choose the smallest $\lambda_n > 0$ such that the dual problem has a finite minimum value. Such selection of the tuning parameter dated at least back to Javanmard and Montanari [8].

References

[1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[2] T. Cai, T. Cai, and Z. Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

[3] T. Cai, T. Tony Cai, and Z. Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):669–719, 2021.

[4] T. T. Cai, Z. Guo, and R. Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, pages 1–14, 2021.

[5] Z. Guo, W. Wang, T. T. Cai, and H. Li. Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, 114:358–369, 2019.

[6] Z. Guo, P. Rakshit, D. S. Herman, and J. Chen. Inference for the case probability in high-dimensional logistic regression. *The Journal of Machine Learning Research*, 22(1):11480–11533, 2021.

[7] Z. Guo, C. Renaux, P. Bühlmann, and T. Cai. Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics*, 15(2):6633–6676, 2021.

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

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

[10] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

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

[12] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

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

[14] P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.