Introduction to Methodologies of SIHR

by Zhenyu(Zach) Wang, Prabrisha Rakshit, T. Tony Cai, and Zijian Guo

April 30, 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 submatrix 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 $\beta_{\mathrm{G}}^{\intercal}A\beta_{\mathrm{G}}$ and $\beta_{\mathrm{G}}^{\intercal}\Sigma_{\mathrm{G,G}}\beta_{\mathrm{G}}$ 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 $\beta_{\mathrm{G}}^{(1)\intercal}A\beta_{\mathrm{G}}^{(2)}$ and $\beta_{\mathrm{G}}^{(1)\intercal}\Sigma_{\mathrm{G,G}}\beta_{\mathrm{G}}^{(2)}$ in section 2.5 and $\gamma_{\mathrm{G}}^{\intercal}A\gamma_{\mathrm{G}}$ and $\gamma_{\mathrm{G}}^{\intercal}\Sigma_{\mathrm{G,G}}\gamma_{\mathrm{G}}$ 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 main idea, we start with the linear functional for the linear model, which will be extended to a unified version in the section Linear functional for GLM. 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 $x_{\text{new}} \in \mathbb{R}^p$, a natural idea for the point estimator is to use the plug-in estimator $x_{\text{new}}^{\intercal}\widehat{\beta}$ with the initial 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

$$\widehat{u} = \arg\min_{u\in\mathbb{R}^p} u^{\intercal}\widehat{\Sigma}u \quad \text{subject to: } \|\widehat{\Sigma}u - x_{\text{new}}\|_\infty \le \|x_{\text{new}}\|_2\mu_0 \tag{6}$$

$$\left|x_{\text{new}}^{\intercal}\widehat{\Sigma}u - \|x_{\text{new}}\|_2^2\right| \le \|x_{\text{new}}\|_2^2\mu_0, \tag{7}$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}$ and $\mu_0 \asymp \sqrt{\log p/n}$. 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 constrained optimization problem in (6) and (7) is designed to minimize the error on the right-hand side of the above equation: the first constraint in (6) controls the "remaining bias" term in the above equation while the objective function in (6) is used to minimize the variance of the "asymp. normal" term. Importantly, the second constraint in (7) ensures the standard error 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 Linear functional for linear model to GLM in (1). Given the initial estimator $\widehat{\beta}$ defined in (3), the key step is to estimate the bias $x_{\text{new}}^{\intercal}(\widehat{\beta} - \beta)$. We can propose a generalized version of the bias-corrected estimator for $x_{\text{new}}^{\intercal}\beta$ as

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

where the projection direction $\widehat{u}$ is defined in the following (9) and $\omega : \mathbb{R} \to \mathbb{R}$ denotes a weight function specified in the following Table 1 associated with different link functions.

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.

In Table 1, we consider different GLM models and present the link function $f(\cdot)$, its derivative $f'(\cdot)$, and the corresponding weight function $\omega(\cdot)$. Note that there are two ways of specifying the weights $w(z)$ for logistic regression, where the linearization weighting was proposed in Guo et al. [7] for logistic regression while the link-specific weighting function was proposed in Cai et al. [4] for general link function $f(\cdot)$. The projection direction $\widehat{u} \in \mathbb{R}^p$ in (8) is constructed as follows:

$$\begin{aligned}
\widehat{u} = \arg\min_{u\in\mathbb{R}^p} u^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}\omega(X_{i\cdot}^{\intercal}\widehat{\beta})f'(X_{i\cdot}^{\intercal}\widehat{\beta})X_{i\cdot}X_{i\cdot}^{\intercal}\right]u \quad \text{subject to:}& \\
\left\|\frac{1}{n}\sum_{i=1}^{n}\omega(X_{i\cdot}^{\intercal}\widehat{\beta})f'(X_{i\cdot}^{\intercal}\widehat{\beta})X_{i\cdot}X_{i\cdot}^{\intercal}u - x_{\text{new}}\right\|_{\infty} \leq \|x_{\text{new}}\|_2\mu_0& \\
\left|x_{\text{new}}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\omega(X_{i\cdot}^{\intercal}\widehat{\beta})f'(X_{i\cdot}^{\intercal}\widehat{\beta})X_{i\cdot}X_{i\cdot}^{\intercal}u - \|x_{\text{new}}\|_2^2\right| \leq \|x_{\text{new}}\|_2^2\mu_0.&
\end{aligned} \tag{9}$$

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]. The variance of $\widehat{x_{\text{new}}^{\intercal}\beta}$ can be estimated by $\widehat{\text{V}}$, defined as

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

$$\widehat{\sigma}_i^2 = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\left(y_j - X_{j\cdot}^{\intercal}\widehat{\beta}\right)^2, & \text{for linear model} \\ f(X_{i\cdot}^{\intercal}\widehat{\beta})(1 - f(X_{i\cdot}^{\intercal}\widehat{\beta})), & \text{for logistic regression with } f(z) = \exp(z)/[1+\exp(z)] \end{cases}. \tag{11}$$

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 $f(x_{\text{new}}^\intercal\beta)$, with $f(z) = \exp(z)/[1+\exp(z)]$, 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 $G = \{1, 2, \cdots, |G|\}$. With the initial estimator $\widehat{\beta}$ defined in (3), the plug-in estimator $\widehat{\beta}_{\text{G}}^\intercal A\widehat{\beta}_{\text{G}}$ has the following estimation 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 of $\widehat{\beta}_{\text{G}}^\intercal A\widehat{\beta}_{\text{G}}$ 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 $x_{\text{new}} = (\widehat{\beta}_{\text{G}}^\intercal A,\ \mathbf{0})^\intercal$. Hence the term can be estimated directly by applying the linear functional approach in section Linear functional for GLM. Utilizing this idea, Guo et al. [7, 5] proposed the following estimator of Q_A,

$$\widehat{\text{Q}}_A = \widehat{\beta}_{\text{G}}^\intercal A\widehat{\beta}_{\text{G}} + 2\,\widehat{u}_A^\intercal\left[\frac{1}{n}\sum_{i=1}^{n}\omega(X_{i\cdot}^\intercal\widehat{\beta})\left(y_i - f(X_{i\cdot}^\intercal\widehat{\beta})\right)X_{i\cdot}\right], \tag{12}$$

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

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

where $\widehat{\sigma}_i^2$ is defined in (11) and the term τ/n with $\tau > 0$ (default value $\tau = 1$) is introduced as an upper bound for the term $(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})^\intercal A(\widehat{\beta}_{\text{G}} - \beta_{\text{G}})$. Then given a fixed value of τ, we construct the CI for Q_A 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 Linear functional for GLM 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

$$\widehat{\text{V}}_\Sigma(\tau) = 4\widehat{u}_\Sigma^\intercal\left[\frac{1}{n^2}\sum_{i=1}^{n}\omega^2(X_{i\cdot}^\intercal\widehat{\beta})\widehat{\sigma}_i^2 X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u}_\Sigma + \frac{1}{n^2}\sum_{i=1}^{n}\left(\widehat{\beta}_{\text{G}}^\intercal X_{i,G}X_{i,G}^\intercal\widehat{\beta}_{\text{G}} - \widehat{\beta}_{\text{G}}^\intercal\widehat{\Sigma}_{\text{G,G}}\widehat{\beta}_{\text{G}}\right)^2 + \frac{\tau}{n}, \tag{14}$$

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

$$\mathrm{CI}(\tau) = \left(\max\left(\widehat{\mathrm{Q}}_\Sigma - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)},\ 0\right),\ \widehat{\mathrm{Q}}_\Sigma + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}_\Sigma(\tau)}\right). \tag{15}$$

2.4 Conditional average treatment effects

The inference methods developed for one sample can be generalized to make inferences for conditional average treatment effects (CATE). From a causality viewpoint, we consider the data set $\{(X_{i\cdot}, y_i, D_i)\}$ for $i = 1, ..., n$, where $D_i \in \{1, 2\}$ indicates the treatment assigned to the i-th observation. For a new observation with covariates $X_{i\cdot} = x_{\rm new}$, we define CATE as $\Delta(x_{\rm new}) = \mathbb{E}(y_i|X_{i\cdot}, D_i = 2) - \mathbb{E}(y_i|X_{i\cdot}, D_i = 1)$.

We group observations $\{i : D_i = k\}$ into the k-th data sample $\{(X_{i\cdot}^{(k)}, y_i^{(k)}\}$ for $k = 1, 2$, where $1 \le i \le n_k$ and $n_1 + n_2 = n$. Subsequently, we rewrite $\mathbb{E}(y_i|X_{i\cdot}, D_i = k)$ as $\mathbb{E}[y_i^{(k)}|X_i^{(k)} = x_{\rm new}]$ for $k = 1, 2$. Using the GLM model outlined in (2), the CATE can be formulated as

$$\Delta(x_{\rm new}) = \mathbb{E}[y_i^{(2)}|X_i^{(2)} = x_{\rm new}] - \mathbb{E}[y_i^{(1)}|X_i^{(1)} = x_{\rm new}] = f(x_{\rm new}^\intercal \beta^{(2)}) - f(x_{\rm new}^\intercal \beta^{(1)}).$$

Following (8), we construct the bias-corrected point estimators of $\widehat{x_{\rm new}^\intercal \beta^{(1)}}$ and $\widehat{x_{\rm new}^\intercal \beta^{(2)}}$, together with their corresponding variances $\widehat{\mathrm{V}}^{(1)}$ and $\widehat{\mathrm{V}}^{(2)}$ as (10). For the first sample $(X_i^{(1)}, y_i^{(1)})$, where $1 \le i \le n_1$, we use the methods described in equations (8) and (10) to compute the bias-corrected point estimator $\widehat{x_{\rm new}^\intercal \beta^{(1)}}$ and the variance estimator $\widehat{\mathrm{V}}^{(1)}$, respectively. Similarly, for the second sample $(X_i^{(2)}, y_i^{(2)})$, where $1 \le i \le n_2$, we apply the same procedures to derive the point estimator $\widehat{x_{\rm new}^\intercal \beta^{(2)}}$ and the variance estimator $\widehat{\mathrm{V}}^{(2)}$.

The paper Cai et al. [3] proposed to estimate $\Delta(x_{\rm new})$ by $\widehat{\Delta}(x_{\rm new})$ as follows,

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

Its variance can be estimated with delta method by:

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

Then we construct the CI for $\Delta(x_{\rm new})$ as

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

2.5 Inner product of regression vectors

The paper Guo et al. [5], Ma et al. [9] have investigated the CI construction for $\beta_{\rm G}^{(1)\intercal} A \beta_{\rm G}^{(2)}$, provided with a pre-specified submatrix $A \in \mathbb{R}^{|\mathrm{G}|\times|\mathrm{G}|}$ and the set of indices $\mathrm{G} \subset \{1, ..., p\}$. With $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ denoting the initial estimators fitted on first and second data sample via (3), respectively, the plug-in estimator $\widehat{\beta}_{\rm G}^{(1)\intercal} A \widehat{\beta}_{\rm G}^{(2)}$ admits the following bias,

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

The key step is to estimate the components $\widehat{\beta}_{\rm G}^{(2)\intercal} A \left(\widehat{\beta}_{\rm G}^{(1)} - \beta_{\rm G}^{(1)}\right)$ and $\widehat{\beta}_{\rm G}^{(1)\intercal} A \left(\widehat{\beta}_{\rm G}^{(2)} - \beta_{\rm G}^{(2)}\right)$, since the last term $(\widehat{\beta}_{\rm G}^{(1)} - \beta_{\rm G}^{(1)})^\intercal A (\widehat{\beta}_{\rm G}^{(2)} - \beta_{\rm G}^{(2)})$ is negligible due to high-order bias. We propose the following bias-corrected estimator for $\beta_{\rm G}^{(1)\intercal} A \beta_{\rm G}^{(2)}$

$$\begin{aligned}\widehat{\beta_{\rm G}^{(1)\intercal} A \beta_{\rm G}^{(2)}} = \widehat{\beta}_{\rm G}^{(1)\intercal} A \widehat{\beta}_{\rm G}^{(2)} + \widehat{u}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1} \omega(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)}) \left(y_i^{(1)} - f(X_{i\cdot}^{(1)\intercal}\widehat{\beta}^{(1)})\right) X_{i\cdot}^{(1)} \\ + \widehat{u}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2} \omega(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)}) \left(y_i^{(2)} - f(X_{i\cdot}^{(2)\intercal}\widehat{\beta}^{(2)})\right) X_{i\cdot}^{(2)}.\end{aligned} \tag{16}$$

Here $\widehat{u}_1$ represents the projection direction computed in (9), using the first sample data and $x_{\rm new} = (\widehat{\beta}_{\rm G}^{(2)\intercal} A,\ \mathbf{0})^\intercal$. Similarly, $\widehat{u}_2$ is the projection direction derived from the second sample data, using $x_{\rm new} = (\widehat{\beta}_{\rm G}^{(2)\intercal} A,\ \mathbf{0})^\intercal$. The corresponding variance of $\widehat{\beta_{\rm G}^{(1)\intercal} A \beta_{\rm G}^{(2)}}$, when A is a known positive definite matrix, is estimated as

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

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

We also consider the case of unknown $A = \Sigma_{\rm G,G}$. As a natural generalization, the quantity $\beta_{\rm G}^{(1)\intercal} \Sigma_{\rm G,G} \beta_{\rm G}^{(2)}$ 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 by replacing A with $\widehat{\Sigma}_{\rm G,G} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} X_{i,\rm G} X_{i,\rm G}^\intercal$ where X is the row-combined matrix of $X^{(1)}$ and $X^{(2)}$. The variance of $\widehat{\beta_{\rm G}^{(1)\intercal} \Sigma_{\rm G,G} \beta_{\rm G}^{(2)}}$ is now estimated as

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

We then construct the CI for $\beta_{\rm G}^{(1)\intercal} A \beta_{\rm G}^{(2)}$ as

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

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 $\gamma_{\rm G}^\intercal A \gamma_{\rm G}$, given a pre-specified submatrix $A \in \mathbb{R}^{|\rm G| \times |\rm G|}$ and the set of indices $\rm G \in \{1, ..., p\}$. The bias of the plug-in estimator $\widehat{\gamma}_{\rm G}^\intercal A \widehat{\gamma}_{\rm G}$ is:

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

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

$$\begin{aligned} \widehat{\gamma_{\rm G}^\intercal A \gamma_{\rm G}} = \widehat{\gamma}_{\rm G}^\intercal A \widehat{\gamma}_{\rm G} &- 2\, \widehat{u}_1^\intercal \frac{1}{n_1} \sum_{i=1}^{n_1} \omega(X_{i\cdot}^{(1)\intercal} \widehat{\beta}^{(1)}) \left(y_i^{(1)} - f(X_{i\cdot}^{(1)\intercal} \widehat{\beta}^{(1)}) \right) X_{i\cdot}^{(1)} \\ &+ 2\, \widehat{u}_2^\intercal \frac{1}{n_2} \sum_{i=1}^{n_2} \omega(X_{i\cdot}^{(2)\intercal} \widehat{\beta}^{(2)}) \left(y_i^{(2)} - f(X_{i\cdot}^{(2)\intercal} \widehat{\beta}^{(2)}) \right) X_{i\cdot}^{(2)}, \end{aligned} \tag{17}$$

where $\widehat{u}_1$ and $\widehat{u}_2$ are the projection directions defined in (9) with $x_{\rm new} = (\widehat{\gamma}_{\rm G}^\intercal A, \mathbf{0})^\intercal$ but on two different sample data respectively. The second term on right-hand-side of (17) is to estimate $-2\, x_{\rm new}^\intercal (\widehat{\beta}_{\rm G}^{(1)} - \beta_{\rm G}^{(1)})$ and the third term on right-hand-side of (17) is to estimate $-2\, x_{\rm new}^\intercal (\widehat{\beta}_{\rm G}^{(2)} - \beta_{\rm G}^{(2)})$.

To maintain non-negativity of distance, we define $\widehat{\gamma_{\rm G}^\intercal A \gamma_{\rm G}} = \max \left\{ \widehat{\gamma_{\rm G}^\intercal A \gamma_{\rm G}},\ 0 \right\}$ and estimate its corresponding asymptotic variance as

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

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

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

We further consider the unknown matrix $A = \Sigma_{\rm G,G}$ and construct the point estimator $\widehat{\gamma_{\rm G}^{\top}\Sigma_{\rm G,G}\gamma_{\rm G}}$ in a similar way as outlined in (17). In this case, the submatrix A is substituted with $\widehat{\Sigma}_{\rm G,G}$, where $\widehat{\Sigma}_{\rm G,G} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_{i,\rm G} X_{i,\rm G}^{\intercal}$ with X as the row-combined matrix of $X^{(1)}$ and $X^{(2)}$. Its corresponding asymptotic variance is

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

Next we present the CI for $\gamma_{\rm G}^{\intercal}\Sigma_{\rm G,G}\gamma_{\rm G}$.

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

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 $\widehat{u} = -\frac{1}{2}\left[\widehat{\mathbf{v}}_{-1} + \frac{x_{\rm new}}{\|x_{\rm new}\|_2}\widehat{\mathbf{v}}_1\right]$, where, $\widehat{\mathbf{v}} \in \mathbb{R}^{p+1}$ is defined as

$$\widehat{\mathbf{v}} = \underset{\mathbf{v}\in\mathbb{R}^{p+1}}{\arg\min}\left\{\frac{1}{4n}\mathbf{v}^{\intercal}\mathbf{H}^{\intercal}\mathrm{X}^{\intercal}\mathrm{Diag}(\mathbf{w})\mathrm{Diag}(\mathbf{f}')\mathrm{X}\mathbf{H}\mathbf{v} + x_{\rm new}^{\intercal}\mathbf{H}\mathbf{v} + \lambda_n\|x_{\rm new}\|_2\cdot\|\mathbf{v}\|_1\right\}, \tag{18}$$

with $\mathbf{H} = \left[\frac{x_{\rm new}}{\|x_{\rm new}\|_2}, \mathbf{I}_{p\times p}\right] \in \mathbb{R}^{p\times(p+1)}$, $\mathbf{w} = \left(\omega(X_1^{\intercal}\widehat{\beta}), ..., \omega(X_n^{\intercal}\widehat{\beta})\right)^{\intercal}$ 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 (18). 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.