Statistical Inference for High-Dimensional Generalized Linear Models With Binary Outcomes

Eurekaimer

Nankai University
Department of Statistics and Data Science

December 4, 2025

Table of Contents

Table of Contents

Background: High-Dimensional Statistical Inference

Generalized linear models (GLMs) with binary outcomes are ubiquitous in modern data-driven scientific research including genetics, metabolomics, finance, and econometrics, and many observational studies.

Main High-Dimensional Statistical Inference Challenges:

- The number of variables(usually denoted as p) can be much larger than the sample size(usually denoted as n), $p \gg n$.
- Most of the classical Inferential procedures (e.g., maximum likelihood) are no longer valid.
- There is a need for new inferential procedures that can handle high-dimensional data.

Table of Contents

Notation

ℓ_p norm

The ℓ_p norm of a vector $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$, we define the ℓ_p norm

$$\|a\|_p = \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p}$$

so we can get the ℓ_0 norm $\|a\|_0 = \sum_{i=1}^{n} 1\{a_i \neq 0\}$, ℓ_∞ norm $\|a\|_\infty = \max_{1 \leqslant i \leqslant n} |a_i|$, and $a_{-j} \in \mathbb{R}^{n-1}$ stand for the subvector of a without the $j-th$ component.

Asymptotic notation

For positive sequences $\{a_n\}, \{b_n\}$, we write $a_n = o(b_n), a_n \ll b_n$ if $\lim_{n \to \infty} \frac{a_n}{b_n} = 0$ and $a_n = O(b_n), a_n \precsim b_n / b_n \succsim a_n$ if there exists a C such that $a_n \leqslant C b_n$ for all n. If $a_n \precsim b_n, b_n \precsim a_n \implies a_n \asymp b_n$

Problem Formulation: High-Dimensional Binary GLMs

Assumptions

Data(Observations): $(X_i, y_i) \in \mathbb{R}^p \times \{0,1\}, i = 1, \ldots, n$ are independently generated from

$$y_i | X_i \sim Bernoulli(f(X_i^T \beta)), X_i \sim P_X$$

- High-dimensional covariates: $p \gg n$.
- k is the sparsity of β, i.e., $\|\beta\|_0 \leqslant k$.
- $\beta \in \mathbb{R}^p$ is a high-dimensional sparse regression vector with sparsity k.
- $f : \mathbb{R} \to (0,1)$ is a known link function.
- P_X is some probability distribution on $\mathbb{R}^p$(unknown)

More rigorous definition

Definition: Link Function

The link function g provides the relationship between the linear predictor and the mean of the distribution function.($g(\mu) = \eta = X^T\beta = \mathbb{E}[Y]$)

Example

- Logistic link function: $f(x) = \frac{\exp(x)}{1+\exp(x)}$
- Probit link function: the link function is the standard Gaussian cumulative distribution function, $f(x) = \Phi(x)$
- Latent variable model: Consider an auxiliary random variable $y_i^* = X_i^T\beta + \epsilon_i$ with $\epsilon_i \sim P_\epsilon$ for $1 \leq i \leq n$. Then the observed binary outcome variable $y_i = 1(y_i^* \geqslant 0)$ can be reformulated as a binary GLM with $y_i|X_i \sim Bernoulli(f(X_i^T\beta))$, where $f(\cdot)$ is the cdf of $-\epsilon_i$.

Motivations

Existing works

- Parameter Estimation & Support Recovery of β [**?**] [**?**]
- Hypothesis Testing & Confidence Intervals(CIs) of β [**?**]
- Statistical inference for high-dimensional linear regression [**?**] [**?**]

Our Goal

- Construct optimal CIs for the individual components of β
- Conduct simultaneous hypothesis testing for the individual components of β
- Establish the minimum sample size requirement for constructing CIs which are adaptive to the sparsity level k of β

Main Results and Contributions

We propose a unified two - step procedure for constructing CIs and performing statistical tests for the regression coefficients in the high - dimensional binary GLM.

The proposed LSW method is effective for a general class of link functions and the general unknown sub - Gaussian design.

Our proposed CIs:

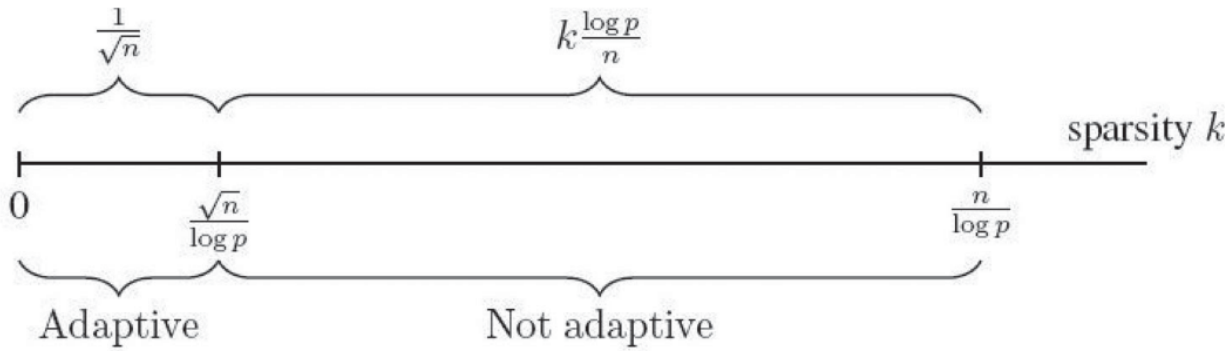

Figure 1. An illustration of the optimality and adaptivity of the CIs with respect to the sparsity k of β for the unknown design setting. On the top of the figure, we report the minimax expected lengths of the CIs, which can be attained by our proposed LSW method (up to a $\log n$ factor for the rate $k\frac{\log p}{n}$). On the bottom of the figure, the possibility of being adaptive to the sparsity k is presented.

Table of Contents

LSW: Inference of β

- Recall

$$y_i|X_i \sim Bernoulli(f(X_i^T\beta)), X_i \sim P_X$$

So we have $P(y_i|X_i^T;\beta) = [f(X_i^T\beta)]^{y_i}[1-f(X_i^T\beta)]^{1-y_i}$

$$\ell_f(\beta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log\left[\frac{f(X_i^T\beta)}{1-f(X_i^T\beta)}\right] - \frac{1}{n}\sum_{i=1}^{n}\log\left[1-f(X_i^T\beta)\right]$$

where $\ell_f(\beta)$ denote the negative log-likelihood function($\frac{1}{n}$ is used to normalize)

- Penalized negative log-likelihood estimator:

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\ell_f(\beta) + \lambda\|\beta\|_1\right\} \tag{1}$$

with $\lambda \asymp \sqrt{\log p/n}$.

Example

Logistic LASSO

$$\hat{\beta} = \underset{\beta}{argmin} \left\{ -\frac{1}{n}\sum_{i=1}^{n} y_i \beta^T X_i + \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(\beta^T X_i)) + \lambda \|\beta\|_1 \right\}$$

Why $\lambda \asymp \sqrt{\log p/n}$?

To ensure Lasso can shrink the coefficients of unimportant features to zero, a key condition must be met [**?**]:

$$\left\| \frac{1}{n}(X^T(y - X\hat{\beta}))_j \right\| \leq \lambda$$

we want to make $\lambda \asymp \max \left| \frac{1}{n} X_j^T \varepsilon \right|$

Sample splitting

For technical reasons, we split the samples such that the initial estimation step and the bias-correction step are conducted on independent datasets.

WLOG, we assume there are $2n$ samples $D = \{(X_i, y_i)\}_{i=1}^{2n}$, divided into two disjoint subsets $D_1 = \{(X_i, y_i)\}_{i=1}^{n}$ and $D_2 = \{(X_i, y_i)\}_{i=n+1}^{2n}$. The initial estimator $\hat{\beta}$ is obtained by applying **??** to D_2 while the bias-correction step is based on $\hat{\beta}$ and the samples in D_1.

Remark

Importantly, the sample splitting procedure is used only to facilitate the theoretical analysis, which does not make it a restriction for practical applications.

Idea of bias correction

Fundamentally, the bias of $\hat{\beta}$ dominates the variance of the estimator, which leads to a suboptimal confidence interval. [**?**] [**?**] [**?**]

Outline of the bias correction procedure:

- Analysis of $\hat{\beta}$ and its bias;
- Construct de-biased estimator $\tilde{\beta}_j = \hat{\beta}_j + \Delta_j$;
- Inference procedure based on $\tilde{\beta}_j$(Asymptotic normality).

Definition: $\tilde{\beta}_j$

For $j = 1, \ldots, p$, define:

$$\tilde{\beta}_j = \hat{\beta}_j + \frac{1}{n} \sum_{i=1}^{n} W_i (y_i - f(X_i^T \hat{\beta})) u^T X_j$$

where $\hat{\beta}$ is defined in **??**, $W_i \in \mathbb{R}$ for $1 \leq i \leq n$ and $u \in \mathbb{R}^p$ denote respectively the data-dependent weights and projection direction to be constructed.

We will construct the link-specific weights $\{W_i\}_{i=1}^n$ and a projection vector $u \in \mathbb{R}^p$ such that $u^T \frac{1}{n} \sum_{i=1}^n W_i \cdot X_i (y_i - f(X_i^T \hat{\beta}))$ is an accurate estimator of the bias $\hat{\beta}_j - \beta_j$.

Determine W_i

We use $[1:p]$ to denote $1, \ldots, p$

- Talor expansion(for a given $j \in [1:p]$):

$$\frac{1}{n}\sum_{i=1}^{n} W_i X_i (y_i - f(X_i^T\hat{\beta})) = \frac{1}{n}\sum_{i=1}^{n} W_i X_i (f(X_i^T\beta) + \varepsilon_i - f(X_i^T\hat{\beta}))$$

$$= \frac{1}{n}\sum_{i=1}^{n} W_i X_i \varepsilon_i + \frac{1}{n}\sum_{i=1}^{n} W_i f'(X_i^T\hat{\beta}) X_i X_i^T(\beta - \hat{\beta}) + \frac{1}{n}\sum_{i=1}^{n} W_i X_i \Delta_i$$

$$\Delta_i = f''(X_i^T\hat{\beta} + tX_i^T(\beta - \hat{\beta})) \cdot [X_i^T(\hat{\beta} - \beta)]^2, t \in (0,1)$$

where $\epsilon_i = y_i - f(X_i^T\hat{\beta})$

Determine W_i

$$\tilde{\beta}j - \beta_j = \underbrace{\frac{1}{n}\sum_{i=1}^{n} W_i u^T X_i \varepsilon_i}_{Asymptotic\ normal} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} W_i f'(X_i^T\hat{\beta}) u^T X_i X_i^T - e_j^T\right)(\beta - \hat{\beta})}_{Remaining\ Bias} +$$

$$\underbrace{u^T \frac{1}{n}\sum_{i=1}^{n} W_i X_i \Delta_i}_{Approximation\ error}$$

- Bias-Variance tradeoff:

$$Var(\frac{1}{n}\sum_{1}^{n} W_i u^T X_i \epsilon_i | X_i) = \frac{1}{n^2}\sum_{1}^{n} W_i^2 f(X_i^T\beta)(1 - f(X_i^T\beta))(u^T X_i)^2$$

We construct the weights $\{W_i\}_{i=1}^n$ such that

$$W_i^2 \cdot f(X_i^T\beta)(1 - f(X_i^T\beta)) \approx W_i \cdot f'(X_i^T\beta)$$

So we get

$$w_i = \frac{f'(X_i^T\hat{\beta})}{f(X_i^T\beta)(1 - f(X_i^T\beta))}$$

The purpose of this construction is to make stochastic error is dominant.

Table 1. Examples of link functions and their corresponding weight functions

Link function	$f(x)$	Weight function $w(x)$
Logistic	$\frac{\exp(x)}{1+\exp(x)}$	1
Probit	$\Phi(x)$	$\frac{\phi(x)}{\Phi(x)(1-\Phi(x))}$
cdf of Student's t_ν	$1 - \frac{1}{2}I_{\frac{\nu}{x^2+\nu}}(\frac{\nu}{2}, \frac{1}{2}), \nu \in \mathbb{N}$	$\frac{2\Gamma(\frac{\nu+1}{2})(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})I_{\frac{\nu}{x^2+\nu}}(\frac{\nu}{2},\frac{1}{2})(1-\frac{1}{2}I_{x(t)}(\frac{\nu}{2},\frac{1}{2}))}, \nu \in \mathbb{N}$
Generalized logistic	$\frac{1}{2}\tanh^\gamma(\varphi x) + \frac{1}{2}, \varphi > 0, \gamma \geq 1$	$\frac{2\varphi\gamma\tanh^{\gamma-1}(\varphi x)\mathrm{sech}^2(\varphi x)}{1-\tanh^{2\gamma}(\varphi x)}, \varphi > 0, \gamma \geq 1$

Determine u

- Holder Inequality: the remaining bias

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{w}_i f'(X_i^T\hat{\beta})u^T X_i X_i^T - e_j^T\right)(\beta-\hat{\beta})\right| \leqslant \left\|\frac{1}{n}\sum_{i=1}^{n}\hat{w}_i f'(X_i^T\hat{\beta})u^T X_i X_i^T - e_j^T\right\|_\infty \|\beta-\hat{\beta}\|_1$$

- Minimize the asymptotic variance

$$\hat{u} = \underset{u\in\mathbb{R}^p}{arg\min}\ u^T\left[\frac{\sum_{i=1}^{n} w(X_i^T\hat{\beta})\cdot f'(X_i^T\hat{\beta})X_i X_i^T}{n}\right]u$$

subject to$(C_1, C_2 > 0)$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} w(X_i^T\hat{\beta})\cdot f'(X_i^T\hat{\beta})X_i X_i^T u - e_j\right\|_\infty \leqslant C_1\sqrt{\frac{\log p}{n}}$$

$$and\ \max_{1\leqslant i\leqslant n}|X_i^T u| \leqslant C_2\sqrt{\log n}$$

Link-specific bias-correction estimator

Hence

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{u}^T \frac{1}{n} \sum_{i=1}^{n} w(X_i^T \hat{\beta}) \cdot (y_i - f(X_i^T \hat{\beta})) X_i$$

Table of Contents

Consequences on Inference of β

- Asymptotic normality of $\tilde{\beta}_j$ for each individual j
- Confidence intervals($1-\alpha$-level): for each j

$$CI_\alpha(\beta_j) = [\tilde{\beta}_j - \tilde{\rho}_j, \tilde{\beta}_j + \tilde{\rho}_j]$$

and $\tilde{\rho}_j = \max\left\{\frac{z_{\alpha/2}\hat{\nu}_j^{\frac{1}{2}}}{\sqrt{n}}, C\frac{\tau_n k \log p}{n}\right\}$

where $\hat{v}_j = n^{-1}\sum_{i=1}^n \frac{[f'(X_i^T\hat{\beta})]^2(\hat{u}^T X_i)^2}{f(X_i^T\hat{\beta})(1-f(X_i^T\hat{\beta}))}$ and $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, $\tau_n = C_2\sqrt{\log n}$

Remark

When $k \ll \frac{\sqrt{n}}{\log p\sqrt{\log n}}$, the above $\tilde{\rho}_j = \frac{z_{\alpha/2}\hat{\nu}_j^{\frac{1}{2}}}{\sqrt{n}}$

Statistical test

Null hypothesis(for any given $j \in [1:p]$)

$$H_0 : \beta_j = \beta_j^0$$

As a direct consequence of the proposed CI, we can construct the test statistic $R_j = \sqrt{n}(\tilde{\beta}_j - \beta_j^0)/\hat{\nu}_j{}^{\frac{1}{2}}$ and define an α-level test as

$$T_\alpha(R_j) = 1\{|R_j| \geq z_{\frac{\alpha}{2}}\}$$

Table of Contents

Simultaneous Inference

Our proposed method can be extended to make simultaneous inference for a subset of regression coefficients such as testing multiple hypotheses with family-wise error rate (FWER) or false discovery rate (FDR) control.

Suppose we are interested in simultaneously testing the null hypotheses

$$H_{0j} : \beta_j = 0, j \in J$$

where $J \subseteq [1 : p]$

- Bonferroni correction

$$\begin{aligned} \text{FWER} &= P_{H_0}\left(\bigcup_{j \in J}\left\{|R_{0j}| \geq z_{\alpha/(2|J|)}\right\}\right) \\ &\leq |J| \cdot P_{H_0}\left(|R_{0j}| \geq z_{\alpha/(2|J|)}\right) \leq \alpha. \end{aligned}$$

As is well known, when $|J|$ is large, controlling FWER with Bonferroni correction is often too conservative and controlling for the FDR is more desirable.

To this end, one can apply the modified BH procedure, where one rejects the null hypothesis H_{0j} if $|R_{0j}| \geq t$ for a certain carefully chosen threshold t. A good choice for the threshold t can be seen as follows:

$$\mathrm{FDR}(t) = \mathbb{E}\left[\frac{\sum_{j\in J_0} \mathbf{1}\{|R_{0j}| \geq t\}}{\max\left\{\sum_{j\in J} \mathbf{1}\{|R_{0j}| \geq t\}, 1\right\}}\right],$$

the proposed threshold level $\hat{t}$ is defined by:

$$\hat{t} = \inf\left\{0 \leq t \leq \sqrt{2\log|J| - 2\log\log|J|} : \frac{|J|\{2-2\Phi(t)\}}{\max\left\{\sum_{j\in J} \mathbf{1}(|R_{0j}| \geq t), 1\right\}} \leq \alpha\right\}$$

Table of Contents

Regularity conditions

We begin with the regularity conditions for the general link function.

- (L1). The link function f is twice differentiable, monotonic increasing, Lipschitz on $\mathbb{R}$, and concave on $\mathbb{R}_+$; and for any $x \in \mathbb{R}$, it holds that $f(x) + f(-x) = 1$.
- (L2). There exist some constants $C_1, C_2 > 0$ such that, for all $x \geq 0$, $f(x) \leq \Phi(C_1 x)$ where $\Phi(x)$ is the standard Gaussian cdf, and $\max\left\{\frac{f'(x)}{x(1-f(x))}, x^2 f'(x)\right\} < C_2$.
- (L3). There exist some constants $c_1, c_2 > 0$ such that $\sup_{x\in\mathbb{R}} |x f''(x+\omega)| \leq c_1$ and $|\omega| < c_2$.
- (L4). For $\ell_f(\beta)$, there exists some constant $C > 1$ such that the Hessian matrix $\ell_f''(\beta)$ can be expressed as $\ell_f''(\beta) = \frac{1}{n}\sum_{i=1}^n h(\beta; y_i, X_i) X_i X_i^T$ for some $h(\beta; y_i, X_i) > 0$ satisfying

$$\max_{1\leq i\leq n} |\log h(\beta + b; y_i, X_i) - \log h(\beta; y_i, X_i)| \leq C(|X_i^T\beta|^2 + |X_i^T b|^2 + |X_i^T b|).$$

(A). $\{X_i\}_{1\leq i\leq 2n}$ are independent and identically distributed sub-Gaussian random vectors, that is, there exists a constant $c \in \mathbb{R}$ satisfying $\mathbb{E}\exp\{v^\top X\} \leq e^{\|v\|_2^2 c^2/2}$ for all $v \in \mathbb{R}^p$. Such a general characterization of the design covariates includes the special case where $X_{i1} = 1$ for all $1 \leq i \leq 2n$ so that β_1 represents the intercept.

Define $\Sigma = \mathbb{E}[X_i X_i^\top] \in \mathbb{R}^{p\times p}$. We focus on the following parameter space indexed by the sparsity level k,

$$\Theta(k) = \left\{\theta = (\beta, \Sigma) : \|\beta\|_0 \leq k, \|\beta\|_2 \leq C, M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M\right\}$$

for constants $M > 1$ and $C > 0$ independent of n and p.

Asymptotic Normality

Theorem 1

Suppose that Conditions (L1)–(L4), and (A) hold, and $(\beta, \Sigma) \in \Theta(k)$. For any $j \in [1:p]$, if $k \ll \frac{n}{\log n \log p}$, then we have $\tilde{\beta}_j - \beta_j = A_n + B_n$, where conditioning on $\mathcal{D}_2 = \{(X_i, y_i)\}_{i=n+1}^{2n}$ and $\{X_i\}_{i=1}^n$, $\sqrt{n}A_n / v_j^{1/2} \to_d N(0,1)$ with v_j and $|B_n| \lesssim k \log p \sqrt{\log n}/n$ with probability at least $1 - p^{-c} - n^{-c}$ for some constant $c > 0$. Additionally, if $k \ll \frac{\sqrt{n}}{\log p \sqrt{\log n}}$, then $\sqrt{n}(\tilde{\beta}_j - \beta_j)/v_j^{1/2} \to_d N(0,1)$.

$$\tilde{\beta}_j - \beta_j = -\left(\frac{1}{n}\sum_{i=1}^n w(X_i^\top \hat{\beta}) f'(X_i^\top \hat{\beta}) \hat{u}^\top X_i X_i^\top - e_j^\top\right)(\hat{\beta} - \beta) + \frac{1}{n}\sum_{i=1}^n w(X_i^\top \hat{\beta}) \hat{u}^\top X_i \epsilon_i$$
$$- \frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top \hat{\beta}) \hat{u}^\top X_i, \tag{1.1}$$

where we denote $A_n = \frac{1}{n}\sum_{i=1}^n w(X_i^\top \hat{\beta}) \hat{u}^\top X_i \epsilon_i$ and $B_n = -\left(\frac{1}{n}\sum_{i=1}^n w(X_i^\top \hat{\beta}) f'(X_i^\top \hat{\beta}) \hat{u}^\top X_i X_i^\top - e_j^\top\right)(\hat{\beta} - \beta) - \frac{1}{n}\sum_{i=1}^n \Delta_i w(X_i^\top \hat{\beta}) \hat{u}^\top X_i$. In what follows, on the one hand, we show that, in

This theorem establishes the coverage and the upper bound of the expected length of the proposed confidence interval (CI).

Theorem 3

Suppose that Conditions (L1)–(L4), and (A) hold, and $\theta = (\beta, \Sigma) \in \Theta(k)$. If $k \ll \frac{n}{\log n \log p}$, then for any constant $0 < \alpha < 1$ and any $j \in [1 : p]$, the $\mathrm{CI}^*_\alpha(\beta_j, \mathcal{D})$ defined in (16) satisfies

$$\lim_{n,p\to\infty} \inf_{\theta\in\Theta(k)} P_\theta(\beta_j \in \mathrm{CI}^*_\alpha(\beta_j, \mathcal{D})) \geq 1 - \alpha$$

$$\sup_{\theta\in\Theta(k)} E_\theta L(\mathrm{CI}^*_\alpha(\beta_j, \mathcal{D})) \lesssim \frac{1}{\sqrt{n}} + \frac{k \log p \sqrt{\log n}}{n},$$

where $L(\mathrm{CI}^*_\alpha(\beta_j, \mathcal{D}))$ denotes the length of $\mathrm{CI}^*_\alpha(\beta_j, \mathcal{D})$.

Optimal length

Theorem 4

Suppose that the link function f satisfies Conditions (L1) and (L2), $\{X_i\}_{i=1}^{2n} \overset{\text{iid}}{\sim} N(0, \Sigma)$, $0 < \alpha < 1/2$ and $k \lesssim \min\left\{p^c, \frac{n}{\log p}\right\}$ for some $0 \le c < 1/2$. Then for any $j \in [1 : p]$,

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)} \sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \gtrsim \frac{1}{\sqrt{n}} + k\frac{\log p}{n}, \tag{23}$$

where $L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \in \mathbb{R}$ is the length of $\mathrm{CI}_\alpha(\beta_j, \mathcal{D})$.

This corollary further refines the properties of the CI in the ultra-sparse region ($k \ll \frac{\sqrt{n}}{\sqrt{\log n \log p}}$).

Corollary 1

Suppose that Conditions (L1)–(L4), and (A) hold, and $(\beta, \Sigma) \in \Theta(k)$. If $k \ll \frac{\sqrt{n}}{\sqrt{\log n \log p}}$, then for any constant $0 < \alpha < 1$ and any $j \in [1 : p]$, the $\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})$ defined in Equation (16) admits the expression $[\tilde{\beta}_j - z_{\alpha/2}\hat{v}_j^{1/2}/\sqrt{n}, \tilde{\beta}_j + z_{\alpha/2}\hat{v}_j^{1/2}/\sqrt{n}]$, and satisfies

$$\lim_{n,p\to\infty} \inf_{\theta\in\Theta(k)} P_\theta(\beta_j \in \mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \geq 1 - \alpha$$

$$\sup_{\theta\in\Theta(k)} E_\theta L(\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \lesssim 1/\sqrt{n}.$$

Compared with the CIs proposed by van de Geer et al. (2014) [?] and Belloni, Chernozhukov, and Wei (2016) [?], which only have guaranteed coverage when $k \ll \frac{\sqrt{n}}{\log p}$, the proposed CIs have guaranteed coverage for all $k \ll \frac{n}{\log n \log p}$, including the moderately sparse region $\frac{\sqrt{n}}{\log p} \lesssim k \ll \frac{n}{\log n \log p}$.

- Adaptivity (to sparsity): simultaneously optimal for various sparsity.
- Proposed CI optimal for all $k = O\left(\frac{\sqrt{n}}{\log p}\right)$, i.e., adaptive.
- What happens when $k \gtrsim \frac{\sqrt{n}}{\log p}$?
- Answer: in such cases, adaptive CI does not exist!(Reference: Theorem 5 [?])

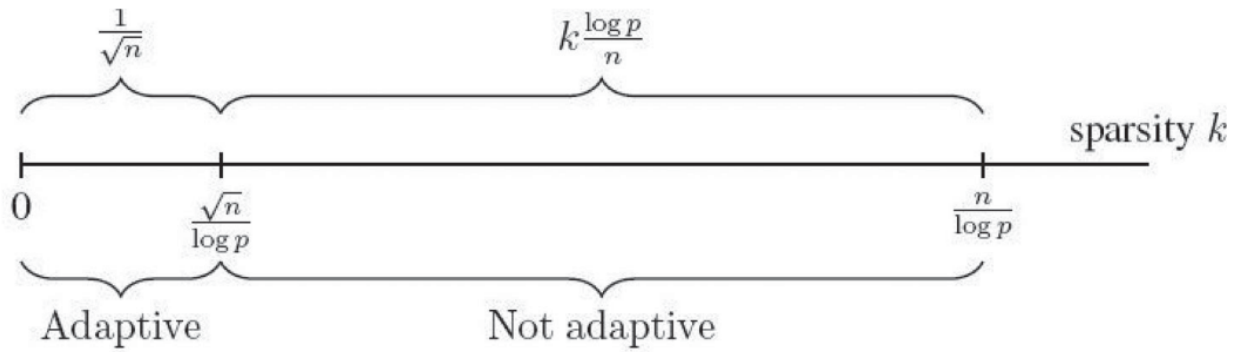

Figure 1. An illustration of the optimality and adaptivity of the CIs with respect to the sparsity k of β for the unknown design setting. On the top of the figure, we report the minimax expected lengths of the CIs, which can be attained by our proposed LSW method (up to a $\log n$ factor for the rate $k\frac{\log p}{n}$). On the bottom of the figure, the possibility of being adaptive to the sparsity k is presented.

Table of Contents

Simulations

We evaluate the empirical performance of our proposed method and compare it with some existing inference methods for high-dimensional binary outcome GLMs.

Table 2. Empirical performances of CIs for β_2 under $\mathbf{\Sigma} = 0.02 \cdot \mathrm{I}_p, \psi = 1, \alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
					$k = 20$					
400	95.2	94.6	96.8	75.1	80.7	2.79	2.77	2.81	1.60	1.99
700	95.2	97.1	97.7	91.1	82.0	2.61	2.77	2.79	2.24	1.99
1000	92.6	93.0	98.4	95.5	84.8	2.80	2.79	2.81	2.53	2.01
1300	95.3	95.3	97.1	94.6	79.1	2.70	2.78	2.80	2.65	2.00
					$k = 25$					
400	93.8	95.3	94.7	73.6	82.0	2.81	2.77	2.81	1.60	1.99
700	95.9	95.9	96.9	94.1	85.7	2.63	2.78	2.81	2.25	2.00
1000	95.9	95.6	97.8	92.8	83.2	2.79	2.77	2.80	2.52	2.00
1300	94.3	94.6	97.8	95.0	83.2	2.70	2.77	2.80	2.65	2.00
					$k = 35$					
400	95.5	94.2	94.7	73.6	82.9	2.81	2.77	2.81	1.60	1.99
700	94.4	95.5	96.8	87.5	80.6	2.63	2.78	2.80	2.24	1.99
1000	92.6	91.3	95.9	93.9	86.1	2.78	2.77	2.80	2.52	1.99
1300	95.9	96.3	96.8	94.7	79.1	2.70	2.77	2.80	2.65	2.00

Table 3. Empirical performances of CIs for β_2 under $\mathbf{\Sigma} = \mathbf{\Sigma}_M, \psi = 0.5, \alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
					$k = 20$					
400	92.9	98.4	57.1	92.6	97.9	1.13	1.56	0.53	0.98	1.04
700	94.6	97.2	45.4	90.3	94.6	1.12	1.32	0.53	0.86	1.08
1000	92 8	97.9	32.9	87.8	96.4	1.13	1.38	0.52	0.77	1.10
1300	94.6	97.2	27.9	81.7	95.8	0.96	1.27	0.53	0.73	1.12
					$k = 25$					
400	93.8	99.2	58.0	93.1	95.4	1.14	1.62	0.54	0.99	1.03
700	95.2	97.6	40.3	90.1	96.0	1.13	1.37	0.53	0.87	1.07
1000	92.8	97.0	31.3	84.0	96.2	1.16	1.43	0.53	0.77	1.09
1300	95.3	93.3	22.2	81.1	96.2	0.97	1.35	0.53	0.73	1.10
					$k = 35$					
400	94.1	97.6	51.8	91.5	94.7	1.15	1.52	0.54	1.00	1.01
700	96.5	99.3	37.3	89.4	92.5	1.14	1.57	0.53	0.87	1.05
1000	96.5	97.3	26.4	79.6	92.8	1.13	1.41	0.53	0.77	1.07
1300	94.7	93.8	22.7	76.7	92.8	0.98	1.30	0.53	0.73	1.09

Simulations

Most existing methods necessitate that the case/control probability be balanced, which means $P(y_i = 1|X_i)P(y_i = 0|X_i) > \delta$ holds for each $i = 1, 2, \ldots, n$.
This paper's method does not have this requirement.

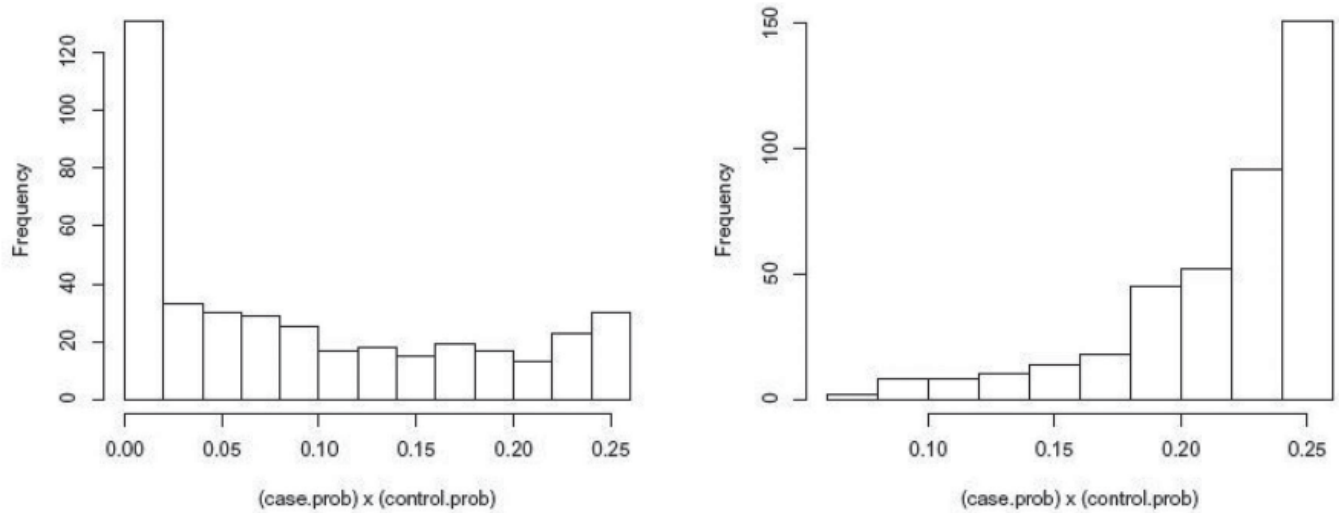

Figure 2. Histograms of $P(y_i = 1|X_i)(1 - P(y_i = 1|X_i))$ associated to the two settings corresponding to Table 2 (left) and Table 3 (right), with $p = 1000$, $n = 400$ and $k = 35$.

Real Data Analysis

You can use the R package SIHR [**?**] to have a try.

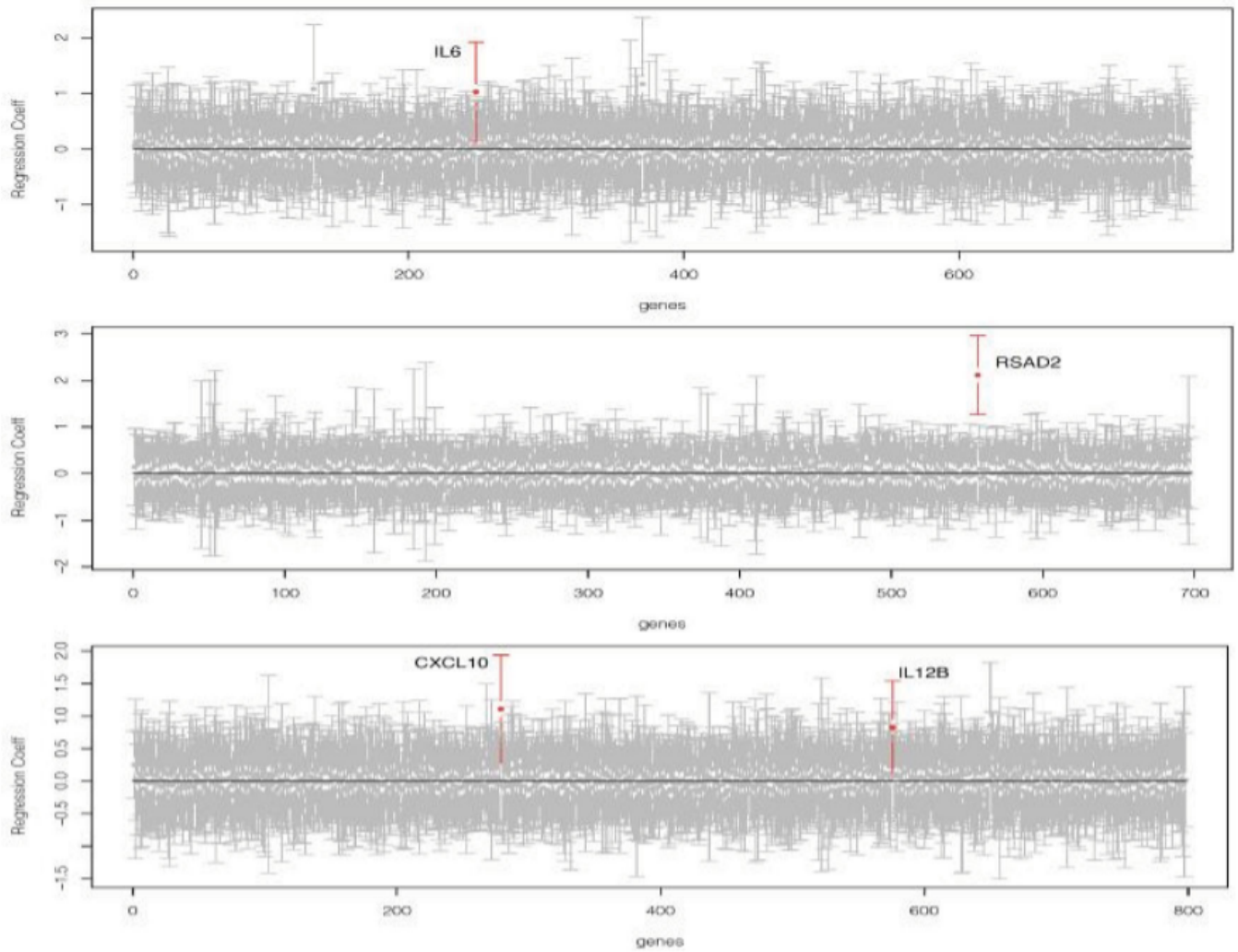

Figure 3. An illustration of the CIs for the high-dimensional logistic regressions corresponding to the stimulations by LPS (top), PIC (middle) and LPS (bottom), respectively. The CIs that do not cover zero are marked in red, with their gene names labeled.

Table of Contents

Summary

- Inference for β under high-dimensional binary GLMs.
- Unified debiasing method.
- Statistical limit, optimality and adaptivity for CIs.
- Practical advantage.

Thank you for listening!
Any questions?

References I

[1] Alexandre Belloni, Victor Chernozhukov, and Ying Wei, *Post-selection inference for generalized linear models with many controls*, Journal of Business & Economic Statistics **34** (2016), no. 4, 606–619.

[2] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov, *Simultaneous analysis of lasso and dantzig selector*, The Annals of Statistics **37** (2009), no. 4.

[3] T. Tony Cai and Zijian Guo, *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity*, The Annals of Statistics **45** (2017), no. 2, 615 – 646.

[4] T. Tony Cai, Zijian Guo, and Rong Ma, *Statistical inference for high-dimensional generalized linear models with binary outcomes*, Journal of the American Statistical Association **118** (2023), no. 542, 1319–1332, PMID: 37366472.

References II

[5] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, Journal of Machine Learning Research **15** (2014), no. 82, 2869–2909.

[6] Lukas Meier, Sara Van De Geer, and Peter Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society Series B: Statistical Methodology **70** (2008), no. 1, 53–71.

[7] Prabrisha Rakshit, Zhenyu Wang, T. Tony Cai, and Zijian Guo, *Sihr: Statistical inference in high-dimensional linear and logistic regression models*, 2023.

[8] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics **42** (2014), no. 3, 1166 – 1202.

References III

[9] Sara A. van de Geer, *High-dimensional generalized linear models and the lasso*, The Annals of Statistics **36** (2008), no. 2, 614 – 645.

[10] Cun-Hui Zhang and Stephanie S. Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society Series B: Statistical Methodology **76** (2013), no. 1, 217–242.