Journal of Machine Learning Research xx (2020) x-xx Submitted a/bc; Revised d/ef; Published g/hi

Estimation and Inference for High Dimensional Generalized Linear Models: A Splitting and Smoothing Approach

Zhe Fei FEIZ@UCLA.EDU
Department of Biostatistics
UCLA
Los Angeles, California, 90025

Yi Li YILI@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, Michigan, 48109

Editor: Boaz Nadler

Abstract

The focus of modern biomedical studies has gradually shifted to explanation and estimation of joint effects of high dimensional predictors on disease risks. Quantifying uncertainty in these estimates may provide valuable insight into prevention strategies or treatment decisions for both patients and physicians. High dimensional inference, including confidence intervals and hypothesis testing, has sparked much interest. While much work has been done in the linear regression setting, there is lack of literature on inference for high dimensional generalized linear models. We propose a novel and computationally feasible method, which accommodates a variety of outcome types, including normal, binomial, and Poisson data. We use a "splitting and smoothing" approach, which splits samples into two parts, performs variable selection using one part and conducts partial regression with the other part. Averaging the estimates over multiple random splits, we obtain the smoothed estimates, which are numerically stable. We show that the estimates are consistent, asymptotically normal, and construct confidence intervals with proper coverage probabilities for all predictors. We examine the finite sample performance of our method by comparing it with the existing methods and applying it to analyze a lung cancer cohort study.

Keywords: Confidence intervals; Dimension reduction; High dimensional inference for GLMs; Sparsity; Sure Screening.

1. Introduction

In the big data era, high dimensional regression has been widely used to address questions arising from many scientific fields, ranging from genomics to sociology (Hastie et al., 2009; Fan and Lv, 2010). For example, modern biomedical research has gradually shifted to understanding joint effects of high dimensional predictors on disease outcomes (e.g. molecular biomarkers on the onset of lung cancer) (Vaske et al., 2010; Chen and Yan, 2014, among others). A motivating clinical study is the Boston Lung Cancer Survivor Cohort (BLCSC), one of the largest comprehensive lung cancer survivor cohorts, which investigates the molecular mechanisms underlying lung cancer (Christiani, 2017). Using a target gene approach

©2020 Zhe Fei and Yi Li.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/vxx/xx-xxx.html.

(Moon et al., 2003; Garrigos et al., 2018; Ho et al., 2019), we analyzed a subset of 708 lung cancer patients and 751 controls, with 6,800 single nucleotide polymorphisms (SNPs) from 15 cancer related genes, in addition to demographic variables such as age, gender, race, education level, and smoking status. Our objective was to determine which covariates were predictive in distinguishing cases from controls. As smoking is known to play a significant role in the development of lung cancer, we were interested in estimating and testing the interaction between smoking status (never versus ever smoked) and each SNP, in addition to the main effect of the SNP. Quantifying uncertainty of the estimated effects helps inform prevention strategies or treatment decisions for patients and physicians (Minnier et al., 2011).

Considerable progress has been made in drawing inferences based on penalized linear models (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Bühlmann et al., 2014; Dezeure et al., 2015). While techniques for variable selection and estimation in high dimensional settings have been extended to generalized linear models (GLMs) and beyond (Van de Geer, 2008; Fan et al., 2009; Witten and Tibshirani, 2009), high dimensional inference in these settings is still at its infancy stage. For example, Bühlmann et al. (2014) generalized the de-sparsified LASSO to high dimensional GLMs, while Ning and Liu (2017) proposed a de-correlated score test for penalized M-estimators. In the presence of high dimensional control variables, Belloni et al. (2014, 2016) proposed a post-double selection procedure for estimation and inference of a single treatment effect and Lee et al. (2016) characterized the distribution of a post-LASSO-selection estimator conditional on the *selected variables*, but only for the linear regression.

However, the performance of these methods may depend heavily on tuning parameters, whose choices are often determined by computationally intensive cross-validation. Also, these methods may require inverting a $p \times p$ information matrix (where p is the number of predictors) or estimating a $p \times p$ precision matrix, with extensive computation and stringent technical conditions. For example, the sparse precision matrix assumption may be violated in GLMs, resulting in biased estimates (Xia et al., 2020).

We propose a new approach for drawing inference with high dimensional GLMs. The idea is to randomly split the samples into two sub-samples (Meinshausen et al., 2009), use the first sub-sample to select a subset of important predictors and achieve dimension reduction, and use the remaining samples to sequentially fit low dimensional GLMs by appending each predictor to the selected set, one at a time, to obtain the estimated coefficient for each predictor, regardless of being selected or not. As with other methods for high dimensional regression (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Bühlmann et al., 2014), one key assumption is that the number of non-zero components of $\boldsymbol{\beta}^*$ is small relative to the sample size, where $\boldsymbol{\beta}^*$ are the true values underlying the parameter vector, $\boldsymbol{\beta}$, in a regression model. The sparsity condition is reasonable in some biomedical applications. For example, in the context of cancer genomics, it is likely that a certain type of cancer is related to only a handful of oncogenes and tumor suppressor genes (Lee and Muller, 2010; Goossens et al., 2015). Under this sparsity condition, we show that our proposed estimates are consistent and asymptotically normal. However, these estimates can be highly variable due to both the random splitting of data and the variation incurred through selection. To stabilize the estimation and account for the variation induced by variable selection, we repeat the random splitting a number of times and average the resulting estimates to ob-

tain the smoothed estimates. These smoothed estimators are consistent and asymptotically normal, with improved efficiency.

Our approach, termed Splitting and Smoothing for GLM (SSGLM), aligns with multi-sample splitting (Meinshausen et al., 2009; Wang et al., 2020) and bagging (Bühlmann and Yu, 2002; Friedman and Hall, 2007; Efron, 2014) and differs from the existing methods based on penalized regression (Zhang and Zhang, 2014; Bühlmann et al., 2014; Ning and Liu, 2017; Javanmard and Montanari, 2018). The procedure has several novelties. First, it addresses the high dimensional estimation problem through the aggregation of low dimensional estimations and presents computational advantages over other existing methods. For example, de-biased methods (Bühlmann et al., 2014; Javanmard and Montanari, 2018) require well-estimated high dimensional precision matrices for proper inference (e.g. correct coverage probabilities), which is statistically and computationally challenging. Complicated procedures that involve choosing a large number of tuning parameters are needed to strike a balance between estimation accuracy and model complexity; see Bühlmann et al. (2014) and Javanmard and Montanari (2014). In contrast, our algorithm is more straightforward as it avoids estimating a high dimensional precision matrix by adopting a "split and select" strategy with minimal tuning. Second, we derived the variance estimator using the infinitesimal jackknife method adapted to the splitting and smoothing procedure (Efron, 2014). This is free of parametric assumptions and leads to confidence intervals with correct coverage probabilities. Third, we have relaxed the stringent "selection consistency" assumption on variable selection, which is required in Fei et al. (2019). Our procedure is valid with a mild "sure screening" assumption for the selection method. Finally, our framework facilitates hypothesis testing and drawing inference on predetermined contrasts in the presence of high dimensional nuisance parameters.

The rest of the paper is organized as follows. Section 2 describes the SSGLM procedure and Section 3 introduces its theoretical properties. Section 4 describes the inferential procedure and Section 5 extends it to accommodate any sub-vectors of parameters of interest. Section 6 provides simulations and comparisons with the existing methods. Section 7 reports our analysis of the BLCSC data. We conclude the paper with a brief discussion in Section 8.

2. Method

2.1 Notation

We assume the observed data $(Y_i, \mathbf{x}_i) = (Y_i, x_{i1}, x_{i2}, \ldots, x_{ip}), i = 1, \ldots, n$, are i.i.d. copies of $(Y, \mathbf{x}) = (Y, x_1, x_2, \ldots, x_p)$. Without loss of generality, we assume that the predictors are centered with $\mathbf{E}(x_j) = 0,\ j = 1, \ldots, p$. In the matrix form, we denote the n samples of observed data by $\mathbf{D}^{(n)} = (\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$. Here, $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ for $j = 1, \ldots, p$. In addition, $\overline{\mathbf{X}} = (\mathbf{1}, \mathbf{X})$ includes an $n \times 1$ column vector of 1's. To accommodate non-Gaussian outcomes, we assume the outcome variable belongs to the linear exponential distribution family, which includes the normal, Bernoulli, Poisson, and negative-binomial distributions. That is, given $\mathbf{x}$, the conditional density function for Y is

$$f(Y|\theta) = \exp\{Y\theta - A(\theta) + c(Y)\}, \tag{1}$$

where $A(\cdot)$ is a specified function that links the mean of Y to $\mathbf{x}$ through θ. We assume the second derivative of $A(\theta)$ is continuous and positive. We consider the canonical mean parameter, $\theta = \overline{\mathbf{x}}\boldsymbol{\beta}$, where $\overline{\mathbf{x}} = (1, \mathbf{x})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ include an intercept term. Specifically, denote $\mu = \mathbf{E}(Y|\mathbf{x}) = A'(\theta) = g^{-1}(\overline{\mathbf{x}}\boldsymbol{\beta})$, and $\mathbf{V}(Y|\mathbf{x}) = A''(\theta) = \nu(\mu)$, where $g(\cdot)$ and $\nu(\cdot)$ are the link and variance functions, respectively.

The forms of $A(\cdot)$, $g(\cdot)$, and $\nu(\cdot)$ depend on the data type of Y. For example, with the outcome in BLCSC being a binary indicator of lung cancer, $A(\theta) = \log\left(1 + e^{\theta}\right)$, $g(\mu) = \mathrm{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ and $\nu(\mu) = \mu(1-\mu)$, corresponding to the well known logistic regression. Based on $(\mathbf{Y}, \mathbf{X})$, the negative log-likelihood with model (1) is

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left\{A(\theta_i) - Y_i\theta_i\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{A(\overline{\mathbf{x}}_i\boldsymbol{\beta}) - Y_i(\overline{\mathbf{x}}_i\boldsymbol{\beta})\right\},$$

where $\theta_i = \overline{\mathbf{x}_i}\boldsymbol{\beta}$ and $\overline{\mathbf{x}_i} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$. The score and the observed information are

$$U(\boldsymbol{\beta}) = \frac{1}{n}\overline{\mathbf{X}}^{\mathrm{T}}\left\{A'(\overline{\mathbf{X}}\boldsymbol{\beta}) - \mathbf{Y}\right\} \text{ and } \widehat{I}(\boldsymbol{\beta}) = \frac{1}{n}\overline{\mathbf{X}}^{\mathrm{T}}\mathbf{V}\overline{\mathbf{X}},$$

which are a $(p+1) \times 1$ vector and a $(p+1) \times (p+1)$ matrix, respectively. Here, $\mathbf{V} = \mathrm{diag}\{\nu(\mu_1), \ldots, \nu(\mu_n)\}$ and $\mu_i = g^{-1}(\overline{\mathbf{x}_i}\boldsymbol{\beta})$ for $i = 1, \ldots, n$. When a univariate function such as $A'(\cdot)$ is applied to a vector, it operates component-wise and returns a vector of values.

We add an index set, $S \subset \{1, 2, \ldots, p\}$, to the subscripts of vectors and matrices to index subvectors $\mathbf{x}_{iS} = (x_{ij})_{j\in S}$ and $\overline{\mathbf{x}}_{iS} = (1, \mathbf{x}_{iS})$, and submatrices $\mathbf{X}_S = (\mathbf{X}_j)_{j\in S}$ and $\overline{\mathbf{X}}_S = (\mathbf{1}, \mathbf{X}_S)$. Moreover, we define $S_{+j} = \{j\} \cup S$ and $S_{-j} = S \setminus \{j\}$. As a convention, let $S_{+0} = S_{-0} = S$, where "0" corresponds to the intercept.

We write $\boldsymbol{\beta}_S = (\beta_0, \beta_j)_{j\in S}$, which always includes the intercept and is of length $1 + |S|$. The negative log-likelihood for model (1) that regresses $\mathbf{Y}$ on $\mathbf{X}_S$ (termed the partial regression) is

$$\ell_S(\boldsymbol{\beta}_S) = \ell(\boldsymbol{\beta}_S; \mathbf{Y}, \mathbf{X}_S) = \frac{1}{n}\sum_{i=1}^{n}\left\{A(\overline{\mathbf{x}}_{iS}\boldsymbol{\beta}_S) - Y_i\overline{\mathbf{x}}_{iS}\boldsymbol{\beta}_S\right\}. \tag{2}$$

Similarly, $U_S(\boldsymbol{\beta}_S) = n^{-1}{\overline{\mathbf{X}}_S}^{\mathrm{T}}\left(A'(\overline{\mathbf{X}}_S\boldsymbol{\beta}_S) - \mathbf{Y}\right)$ and $\widehat{I}_S(\boldsymbol{\beta}_S) = n^{-1}{\overline{\mathbf{X}}_S}^{\mathrm{T}}\mathbf{V}_S\overline{\mathbf{X}}_S$, where $\mathbf{V}_S = \mathrm{diag}\{A''(\overline{\mathbf{x}}_{1S}\boldsymbol{\beta}_S), \ldots, A''(\overline{\mathbf{x}}_{nS}\boldsymbol{\beta}_S)\}$. Let the true values of $\boldsymbol{\beta}$ be $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$. Define the expected information as $I^* = \mathbf{E}\{\widehat{I}(\boldsymbol{\beta}^*)\}$. Let $S^* = \left\{j \neq 0 : \beta_j^* \neq 0\right\}$ denote the active set, and let $s_0 = |S^*|$ be the number of nonzero and non-intercept elements in $\boldsymbol{\beta}^*$. When $S \supseteq S^*$, define the "observed" *sub-information* by $\widehat{I}_S = \widehat{I}_S(\boldsymbol{\beta}_S^*)$, and the "expected" sub-information by $I_S = \mathbf{E}\{\widehat{I}_S\}$. The latter is equal to the submatrix of I^* with rows and columns indexed by S, which is denoted by I_S^*.

2.2 Proposed SSGLM estimator

Under model (1), we assume a sparsity condition that s_0 is small relative to the sample size and will be detailed in Section 3. We randomly split the samples, $\mathbf{D}^{(n)}$, into two parts, $\mathbf{D}_1$ and $\mathbf{D}_2$, with sample sizes $|\mathbf{D}_1| = n_1$, $|\mathbf{D}_2| = n_2$, respectively, such that $n_1 + n_2 = n$.

As an example, consider an equal splitting with $n_1 = n_2 = n/2$. We apply a variable selection scheme, $\mathcal{S}_\lambda$, where λ denotes the tuning parameters, to $\mathbf{D}_2$ to select a subset of important predictors $S \subset \{1, \ldots, p\}$, with $|S| < n$ for dimension reduction. Then using $\mathbf{D}_1 = (\mathbf{Y}^1, \mathbf{X}^1)$, for each $j = 1, 2, \ldots, p$, we fit a low dimensional GLM by regressing $\mathbf{Y}^1$ on $\mathbf{X}^1_{S_{+j}}$, where $S_{+j} = \{j\} \cup S$. Denote the maximum likelihood estimate (MLE) of each fitted model as $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$, and define $\widetilde{\beta}_j = \left(\widetilde{\boldsymbol{\beta}}_{S_{+j}}\right)_j$, the element of $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$ corresponding to predictor $\mathbf{X}_j$. We denote by $\widetilde{\beta}_0$ the estimator of the intercept from the model $\mathbf{Y}^1 \sim \mathbf{X}^1_S$. Thus, the one-time estimator based on a single data split is defined as

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_{S_{+j}} = \underset{\boldsymbol{\beta}_{S_{+j}}}{\operatorname{argmin}} \ell_{S_{+j}}(\boldsymbol{\beta}_{S_{+j}}) = \underset{\boldsymbol{\beta}_{S_{+j}}}{\operatorname{argmin}} \ell(\boldsymbol{\beta}_{S_{+j}}; \mathbf{Y}^1, \mathbf{X}^1_{S_{+j}}); \\ \widetilde{\beta}_j = \left(\widetilde{\boldsymbol{\beta}}_{S_{+j}}\right)_j; \quad \widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p). \end{aligned} \tag{3}$$

In the linear regression setting (Fei et al., 2019), $\widetilde{\beta}_j$ in (3) has an explicit form, $\widetilde{\beta}_j = \left\{(\mathbf{X}^1_{S_{+j}}{}^{\mathrm{T}}\mathbf{X}^1_{S_{+j}})^{-1}\mathbf{X}^1_{S_{+j}}{}^{\mathrm{T}}\mathbf{Y}^1\right\}_j$.

The rationale for this one-time estimator is that if the subset of important predictors, S, is equal to or contains the active set, S^*, then $\widetilde{\beta}_j$ would be a consistent estimator regardless of whether variable j is selected or not (Fei et al., 2019). We show in Theorem 1 that the one-time estimator is indeed consistent and asymptotically normal in the GLM setting.

However, the estimator based on a single split is highly variable, making it difficult to separate true signals from noises. This phenomenon is analogous to using a single tree in the bagging algorithm (Bühlmann and Yu, 2002). To reduce this variability, we resort to a multi-sample splitting scheme. We randomly split the data multiple times, repeat the estimation procedure, and average the resulting estimates to obtain the smoothed coefficient estimates. Specifically, for each $b = 1, 2, \ldots, B$, where B is large, we randomly split the data, $\mathbf{D}^{(n)}$, into $\mathbf{D}^b_1$ and $\mathbf{D}^b_2$, with $|\mathbf{D}^b_1| = n_1$ and $|\mathbf{D}^b_2| = n_2$ such that the splitting proportion is $q = n_1/n$, $0 < q < 1$. Denote the candidate set of variables selected by applying $\mathcal{S}_\lambda$ to $\mathbf{D}^b_2$ as S^b, and the estimates via (3), as $\widetilde{\boldsymbol{\beta}}^b = (\widetilde{\beta}^b_0, \widetilde{\beta}^b_1, \ldots, \widetilde{\beta}^b_p)$. Then the smoothed estimator, termed the SSGLM estimator, is defined to be

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p), \text{ where } \widehat{\beta}_j = \frac{1}{B}\sum_{b=1}^{B} \widetilde{\beta}^b_j. \tag{4}$$

The procedure is described in Algorithm 1.

3. Theoretical Results

We specify the following regularity conditions.

(A1) (*Bounded observations*) $\|\mathbf{x}\|_\infty \le C_0$ and $\mathbf{E}\,|Y| < \infty$. Without loss of generality, we assume $C_0 = 1$.

(A2) (*Bounded eigenvalues and effects*) The eigenvalues of $\Sigma = \mathbf{E}\,(\overline{\mathbf{x}}^{\mathrm{T}}\overline{\mathbf{x}})$, where $\overline{\mathbf{x}} = (1, \mathbf{x})$, are bounded below and above by constants $c_{\min}, c_{\max}$, such that

$$0 < c_{\min} \le \lambda_{\min}(\Sigma) < \lambda_{\max}(\Sigma) \le c_{\max} < \infty.$$

Algorithm 1 SSGLM Estimator

Require: A variable selection procedure denoted by $\mathcal{S}_\lambda$
Input: Data $(\mathbf{Y}, \mathbf{X})$, a splitting proportion $q \in (0,1)$, and the number of random splits B
Output: Coefficient vector estimator $\widehat{\boldsymbol{\beta}}$

1: **for** $b = 1, 2, \ldots, B$ **do**
2: Split the samples into $\mathbf{D}_1$ and $\mathbf{D}_2$, with $|\mathbf{D}_1| = qn$, $|\mathbf{D}_2| = (1-q)n$
3: Apply $\mathcal{S}_\lambda$ to $\mathbf{D}_2$ to select predictors indexed by $S \subset \{1, \ldots, p\}$
4: **for** $j = 0, 1, \ldots, p$ **do**
5: With $S_{+j} = \{j\} \cup S$, fit model (1) by regressing $\mathbf{Y}^1$ on $\mathbf{X}^1_{S_{+j}}$, where $\mathbf{D}_1 = (\mathbf{Y}^1, \mathbf{X}^1)$, and compute the MLE $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$ as in (3)
6: Compute $\widetilde{\beta}^b_j = \left(\widetilde{\boldsymbol{\beta}}_{S^b_{+j}}\right)_j$, which is the coefficient for predictor $\mathbf{X}_j$ ($\widetilde{\beta}^b_0$ represents the intercept)
7: **end for**
8: Output $\widetilde{\boldsymbol{\beta}}^b = (\widetilde{\beta}^b_0, \widetilde{\beta}^b_1, \ldots, \widetilde{\beta}^b_p)$
9: **end for**
10: Compute $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$, where $\widehat{\beta}_j = \frac{1}{B}\sum_{b=1}^B \widetilde{\beta}^b_j$

Algorithm 2 Model-free Variance Estimator

Input: n, n_1, B, $\widetilde{\boldsymbol{\beta}}^b, b = 1, 2, \ldots, B$ and $\widehat{\boldsymbol{\beta}}$
Output: Variance estimator $\widehat{V}^B_j$ for $\widehat{\beta}_j$, $j = 0, 1, \ldots, p$

1: For $i = 1, 2, \ldots, n$ and $b = 1, 2, \ldots, B$, define $J_{bi} = \mathbf{I}\left((Y_i, \mathbf{x}_i) \in \mathbf{D}^b_1\right) \in \{0, 1\}$, and $J_{\cdot i} = \left(\sum_{b=1}^B J_{bi}\right)/B$
2: **for** $j = 0, 1, \ldots, p$ **do**
3: Compute

$$\widehat{V}_j = \frac{n(n-1)}{(n-n_1)^2}\sum_{i=1}^n \widehat{\mathrm{cov}}^2_{ij},$$

where

$$\widehat{\mathrm{cov}}_{ij} = \frac{1}{B}\sum_{b=1}^B (J_{bi} - J_{\cdot i})\left(\widetilde{\beta}^b_j - \widehat{\beta}_j\right)$$

4: Compute

$$\widehat{V}^B_j = \widehat{V}_j - \frac{n}{B^2}\frac{n_1}{n-n_1}\sum_{b=1}^B (\widetilde{\beta}^b_j - \widehat{\beta}_j)^2$$

5: **end for**
6: Set $\widehat{V}^B = \left(\widehat{V}^B_1, \widehat{V}^B_2, \ldots, \widehat{V}^B_p\right)$

In addition, there exists a constant $c_\beta > 0$ such that $|\boldsymbol{\beta}^*|_\infty \le c_\beta$.

(A3) (*Sparsity and sure screening property*) Recall that $S^* = \left\{j \neq 0 : \beta_j^* \neq 0\right\}$ and $s_0 = |S^*|$. Let $\widehat{S}_{\lambda_n}$ be the index set of predictors selected by $\mathcal{S}$ with a tuning parameter λ_n. Assume $\log p = o(n^{1/2})$, there exists a sequence $\{\lambda_n\}_{n\ge 1}$ and constants $0 \le c_1 < 1/2$, $c_2, K_1, K_2 > 0$ such that $s_0 \le K_1 n^{c_1}$, $|\widehat{S}_{\lambda_n}| \le K_1 n^{c_1}$, and

$$\mathrm{P}\left(S^* \subseteq \widehat{S}_{\lambda_n}\right) \ge 1 - K_2 (p \vee n)^{-1-c_2}.$$

Assumption (A1) states that the predictors are uniformly bounded, which is reasonable as predictors are often normalized during data pre-processing. As defined in (A2), $\Sigma = \mathrm{diag}(1, \Sigma_x)$, where Σ_x is the variance-covariance matrix of $\mathbf{x}$. The boundedness of the eigenvalues of the variance-covariance matrix of $\mathbf{x}$ has been commonly assumed in the high dimensional literature (Zhao and Yu, 2006; Belloni and Chernozhukov, 2011; Fan et al., 2014; Van de Geer et al., 2014). (A3) restricts the orders of p and n as well as the sparsity of $\boldsymbol{\beta}^*$. Both (A1) and (A2) ensure the convergence of the MLEs for the low dimensional GLMs (3) with a diverging number of predictors (Portnoy, 1985; He and Shao, 2000). (A3) requires $\mathcal{S}$ to possess the sure screening property, which substantially relaxes the restrictive selection consistency assumption in Fei et al. (2019).

Variable selection methods that satisfy the sure screening property are available. For example, Assumptions (A1) and (A2), along with a "beta-min" condition, which stipulates that $\min_{j\in S^*} |\beta_j^*| > c_0 n^{-\kappa}$ with $c_0 > 0, 0 < \kappa < 1/2$, ensure that the commonly used sure independence screening (SIS) procedure (Fan and Song, 2010) satisfies the sure screening property; see Theorem 4 in Fan and Song (2010). While a "beta-min" condition is common in deriving the sure screening property, it is not required for the de-biased type of estimators. We take $\mathcal{S}$ to be the SIS procedure when conducting variable selection in simulations and the data analysis. Theorems 1 and 2 correspond to the one-time estimator and the SSGLM estimator, respectively.

Theorem 1 *Given model (1) and assumptions (A1) – (A3), consider the one-time estimator* $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^{\mathrm{T}}$ *as defined in (3). Denote* $p_s = |S|$ *and* $\widetilde{\sigma}_j^2 = \left(\{I^*_{S_{+j}}\}^{-1}\right)_{jj}$, $j \in \{0, 1, \ldots, p\}$. *Then as* $n \to \infty$,

i. $\|\widetilde{\boldsymbol{\beta}}_{S_{+j}} - \boldsymbol{\beta}^*_{S_{+j}}\|_2^2 = o_p(p_s/n)$, *if* $p_s \log p_s / n \to 0$;

ii. $\sqrt{n_1}\left(\widetilde{\beta}_j - \beta_j^*\right)/\widetilde{\sigma}_j \xrightarrow{d} N(0,1)$, *if* $p_s^2 \log p_s / n \to 0$.

Theorem 2 *Given model (1) and under assumptions (A1) – (A3) and a partial orthogonality condition that* $\{x_j, j \in S^*\}$ *are independent of* $\{x_k, k \notin S^*\}$, *consider the smoothed estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$ *as defined in (4). For each* j, *define* $\check{\sigma}_j^2 = \left(\{I^*_{S^*_{+j}}\}^{-1}\right)_{jj}$. *Then, as* $n, B \to \infty$,

$$\sqrt{n}(\widehat{\beta}_j - \beta_j^*)/\check{\sigma}_j \xrightarrow{d} N(0,1).$$

The added partial orthogonality condition for Theorem 2 is a technical assumption to ensure the validity of the theorem, which has been assumed in the high dimensional literature (Fan and Lv, 2008; Fan and Song, 2010; Wang and Wang, 2014). However, our numerical experiments suggest the robustness of our results to the violation of this condition. In addition, while both of the one-time estimator $\widetilde{\beta}_j$ and the SSGLM estimator $\widehat{\beta}_j$ possess asymptotic consistency and normality, the key advantage of $\widehat{\beta}_j$ over $\widehat{\beta}_j$ lies in the efficiency. An immediate observation is that $\widehat{\beta}_j$ is estimated using all n samples but $\widetilde{\beta}_j$ is estimated with only n_1 samples, which explains the different normalization constants in their respective variances, $\widehat{\sigma}_j^2/n$ and $\widetilde{\sigma}_j^2/n_1$. In addition, with $\widetilde{\sigma}_j^2$ depending solely on a one-time variable selection S, its variability is high given the wide variability of S. On the other hand, $\widehat{\sigma}_j^2$ implicitly averages over the multiple selections, S^b's, and gains efficiency via "the effect of bagging" (Bühlmann and Yu, 2002); also see Web Table 1 of Fei et al. (2019) for empirical evidence under the linear regression setting. Moreover, the high variability of $\widetilde{\beta}_j$ may lead to a large false positive rate; see Figure 1 of Fan and Lv (2008).

We defer the proofs to the Appendix, but provide some intuition here. The randomness of the selection $\widehat{S}_\lambda$ presents difficulties when developing the theoretical properties, but why sure screening works is that, given any subset $S \supseteq S^*$, the estimator $\widetilde{\boldsymbol{\beta}}_S$ is consistent, though less efficient (with additional noise variables) than the "oracle estimator" $\widetilde{\boldsymbol{\beta}}_{S^*}$ acting upon the true active set. The proof also shows that $\widehat{\sigma}_j^2$ depends on the unknown S^*, taking into account the variation in B random splits. Therefore, direct computation of $\widehat{\sigma}_j^2$ in an analytical form is not feasible. Alternatively, we estimate the variance component via the infinitesimal jackknife method (Efron, 2014; Fei et al., 2019).

4. Variance Estimator and Inference by SSGLM

The infinitesimal jackknife method has been applied to estimate the variance of the bagged estimator with bootstrap-type resampling (sampling with replacement) (Efron, 2014; Fei et al., 2019). The idea is to treat each $\widetilde{\beta}_j^b$ as a function of the sub-sample $\mathbf{D}_1^b$, or its empirical distribution represented by the sampling indicator vector $\mathbf{J}_b = (J_{b1}, J_{b2}, \ldots, J_{bn})$, where $J_{bi} \in \{0, 1\}$ is an indicator of whether the i^{th} observation is sampled in $\mathbf{D}_1^b$. We further denote $J_{\cdot i} = \left(\sum_{b=1}^B J_{bi}\right)/B$. With slightly overused notation, let

$$\widetilde{\beta}_j^b = t(\mathbf{D}_1^b) = t(\mathbf{J}_b; \mathbf{D}^{(n)});$$
$$\widehat{\beta}_j = \frac{1}{B}\sum_{b=1}^B \widetilde{\beta}_j^b \xrightarrow{p} \mathbf{E}^* t(\mathbf{J}_b; \mathbf{D}^{(n)}), \text{ as } B \to \infty,$$

where $t(\cdot)$ is a general function that maps the data to the estimator, the expectation $\mathbf{E}^*$ and the convergence are with respect to the probability measure induced by the randomness of $\mathbf{J}_b$'s. We can generalize the infinitesimal jackknife to estimate the variance, $\text{Var}\left(\widehat{\beta}_j\right)$, analogous to equation (8) of Wager and Athey (2018), as follows

$$\widehat{V}_j = \frac{n-1}{n}\left(\frac{n}{n-n_1}\right)^2 \sum_{i=1}^n \widehat{\text{cov}}_{ij}^2, \tag{5}$$

where

$$\widehat{\text{cov}}_{ij} = \frac{1}{B}\sum_{b=1}^{B}(J_{bi} - J_{\cdot i})\left(\widetilde{\beta}_j^b - \widehat{\beta}_j\right)$$

is the covariance between the estimates $\widetilde{\beta}_j^b$'s and the sampling indicators J_{bi}'s with respect to the B splits. Here, $n(n-1)/(n-n_1)^2$ is a finite-sample correction term with respect to the sub-sampling scheme. Theorem 1 of Wager and Athey (2018) implies that this variance estimator is consistent, in the sense that $\widehat{V}_j/\text{Var}\left(\widehat{\beta}_j\right) \xrightarrow{p} 1$ as $B \to \infty$.

We further propose a bias-corrected version of (5):

$$\widehat{V}_j^B = \widehat{V}_j - \frac{n}{B^2}\frac{n_1}{n-n_1}\sum_{b=1}^{B}(\widetilde{\beta}_j^b - \widehat{\beta}_j)^2. \tag{6}$$

The derivation is similar to that in Section 4.1 of Wager et al. (2014), but it is adapted to the sub-sampling scheme. The difference between $\widehat{V}_j$ and $\widehat{V}_j^B$ converges to zero as $n, B \to \infty$, as it can be re-written as $\frac{n}{B}\frac{n_1}{n-n_1}\widehat{v}_j$, where $\widehat{v}_j = B^{-1}\sum_{b=1}^{B}(\widetilde{\beta}_j^b - \widehat{\beta}_j)^2$ is the sample variance of $\widetilde{\beta}_j^b$'s from B splits. Thus both variance estimators are asymptotically equal. See Algorithm 2 for the complete procedure of estimating the variance component of SSGLM.

For finite samples, we give the order of B to control the Monte Carlo errors of these two variance estimators. First, with $n_1 = qn$ for a fixed $0 < q < 1$, the bias of $\widehat{V}_j$ is of order $n\widehat{v}_j/B$ (Wager et al., 2014). Thus, setting $B = O(n^{1.5})$ will reduce the bias to the desired level of $O(n^{-0.5})$. On the other hand, $\widehat{V}_j^B$ effectively removes this bias, as it only requires $B = O(n)$ to control the Monte Carlo Mean Squared Error (MSE) to $O(n^{-1})$ (Wager et al., 2014). A comparison between $\widehat{V}_j$ and $\widehat{V}_j^B$, given in Simulation **Example 1**, also shows the preference of $\widehat{V}_j^B$ to $\widehat{V}_j$.

For $0 < \alpha < 1$, the asymptotic $100(1-\alpha)\%$ confidence interval for $\beta_j^*, j = 1, \ldots, p$, is given by

$$\left(\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B}\ ,\ \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B}\right),$$

and the p-value for testing $H_0 : \beta_j^* = 0$ is

$$2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j|/\sqrt{\widehat{V}_j^B}\right)\right\},$$

where Φ is the CDF of the standard normal distribution.

5. Extension to Subvectors With Fixed Dimensions

We extend the SSGLM procedure to derive confidence regions for a subset of predictors and to test for contrasts of interest. Consider $\boldsymbol{\beta}_{S^{(1)}}^*$ with $|S^{(1)}| = p_1 \geq 2$, which is finite and does not increase with n or p. Accordingly, the SSGLM estimator for it is presented in **Algorithm 3**, and the extension of Theorem 2 is stated below.

Theorem 3 *Given model (1) under assumptions (A1)–(A3) and a fixed finite subset $S^{(1)} \subset \{1, 2, \ldots, p\}$ with $|S^{(1)}| = p_1$. Let $\widehat{\boldsymbol{\beta}}^{(1)}$ be the smoothed estimator for $\boldsymbol{\beta}^*_{S^{(1)}}$ as defined in Algorithm 3. Then as $n, B \to \infty$,*

$$\sqrt{n} I^{(1)} \left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*_{S^{(1)}} \right) \xrightarrow{d} N(0, \mathbf{I}_{p_1}),$$

where $\mathbf{I}_{p_1}$ is a $p_1 \times p_1$ identity matrix, and $I^{(1)}$ is a $p_1 \times p_1$ positive definite matrix depending on $S^{(1)}$ and S^* and is defined in the proof.

There is a direct extension of the one-dimensional infinitesimal jackknife for estimating the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}^{(1)}$, $\widehat{\Sigma}^{(1)} = \widehat{\mathrm{COV}}_{(1)}^{\mathrm{T}} \widehat{\mathrm{COV}}_{(1)}$, where

$$\widehat{\mathrm{COV}}_{(1)} = \left(\widehat{\mathrm{cov}}_1^{(1)}, \widehat{\mathrm{cov}}_2^{(1)}, \ldots, \widehat{\mathrm{cov}}_n^{(1)} \right)^{\mathrm{T}}, \text{ with}$$

$$\widehat{\mathrm{cov}}_i^{(1)} = \sum_{b=1}^{B} (J_{bi} - J_{\cdot i})(\widehat{\boldsymbol{\beta}}^b_{S^{(1)}} - \widehat{\boldsymbol{\beta}}^{(1)})/B.$$

To test $H_0 : Q\boldsymbol{\beta}^{(1)} = R$, where Q is a $r \times p_1$ contrast matrix and R is a $r \times 1$ vector, a Wald-type test statistic can be formulated as

$$T = \left(Q\widehat{\boldsymbol{\beta}}^{(1)} - R \right)^{\mathrm{T}} \left[Q\widehat{\Sigma}^{(1)} Q^{\mathrm{T}} \right]^{-1} \left(Q\widehat{\boldsymbol{\beta}}^{(1)} - R \right), \tag{7}$$

which follows χ^2_r under H_0. Therefore, with a significance level $\alpha \in (0, 1)$, we reject H_0 when T is larger than the $(1 - \alpha) \times 100$ percentile of χ^2_r.

Algorithm 3 SSGLM for Subvector $\boldsymbol{\beta}^{(1)}$

Require: A selection procedure $\mathcal{S}_\lambda$

Input: Data $(\mathbf{Y}, \mathbf{X})$, a data splitting proportion $q \in (0, 1)$, the number of splits B, and an index set $S^{(1)}$ for the predictors of interest

Output: Estimates of the coefficients of predictors indexed by $S^{(1)}$, $\widehat{\boldsymbol{\beta}}^{(1)}$

1: **for** $b = 1, 2, \ldots, B$ **do** Split the samples into two parts $\mathbf{D}_1$ and $\mathbf{D}_2$, with $|\mathbf{D}_1| = qn$ and $|\mathbf{D}_2| = (1 - q)n$

2: Apply $\mathcal{S}_\lambda$ to $\mathbf{D}_2$ to select a subset of important predictors $S \subset \{1, \ldots, p\}$

3: Fit a GLM by regressing $\mathbf{Y}^1$ on $\mathbf{X}^1_{S^{(1)} \cup S}$, where $\mathbf{D}_1 = (\mathbf{Y}^1, \mathbf{X}^1)$ and compute the MLEs, denoted by $\widetilde{\boldsymbol{\beta}}^{(1)}$

4: Define $\widetilde{\boldsymbol{\beta}}^b_{S^{(1)}} = \left(\widetilde{\boldsymbol{\beta}}^{(1)} \right)_{S^{(1)}}$

5: **end for**

6: Compute $\widehat{\boldsymbol{\beta}}^{(1)} = \left(\sum_{b=1}^{B} \widetilde{\boldsymbol{\beta}}^b_{S^{(1)}} \right) / B$

6. Simulations

We compared the finite sample performance of the proposed SSGLM procedure, under various settings, with two existing methods, the de-biased LASSO for GLMs (Van de Geer

et al., 2014; Dezeure et al., 2015) and the de-correlated score test (Ning and Liu, 2017), in estimation accuracy and computation efficiency. We also investigated how the choice of $q = n_1/n$, the splitting proportion, may impact the performance of SSGLM, explored various selection methods as part of the SSGLM procedure and their impacts on estimation and inference, illustrated our method with both logistic and Poisson regression settings, and assessed the power and type I error of the procedure. We adopted some challenging simulation settings in Bühlmann et al. (2014). For example, the indices of the active set, as well as the non-zero effect sizes, were randomly generated, and various correlation structures were explored.

Example 1 investigated the performance of SSGLM with various splitting proportions and the convergence of the proposed variance estimators. We set $n_1 = qn,\ q = 0.1, 0.2, \ldots, 0.9$. Under a linear regression model, $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \ldots, n$ with i.i.d. $\varepsilon_i \sim N(0, 1)$, we set $n = 500,\ p = 1,000,\ s_0 = 10$ with an AR(1) correlation structure, i.e. $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5,\ i, j = 1, 2, \ldots, p$. The indices in the active set S^* randomly varied from $\{1, \ldots, p\}$, and the non-zero effects of $\beta_j^*, j \in S^*$ were generated from $\text{Unif}[(-1.5, -0.5) \cup (0.5, 1.5)]$. For each q, we computed the MSE for $\widehat{\beta}_j^{(k)}$, the smoothed estimate of β_j from the k-th simulation, $k = 1, 2, \ldots, K$,

$$\text{MSE}_j = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\beta}_j^{(k)} - \beta_j^*)^2, \quad \text{MSE}_{\text{avg}} = \frac{1}{p}\sum_{j=1}^{p}\text{MSE}_j.$$

The left panel of Figure 1 showed that the minimum MSE was achieved when $q = 0.5$, suggesting the rationality of equal-size splitting in practice.

However, the MSE was, in general, less sensitive to q when q was getting larger, hinting that a large n_1 may lead to adequate accuracy. Intuitively, there is a minimum sample size $n_2 = (1 - q)n$ required for the selections to achieve the "sure screening" property. For example, LASSO with smaller sample size would select less variables given the same tuning parameter. On the other hand, larger $n_1 = qn$ improves the power of the low dimensional GLM estimators directly. Thus the optimal split proportion is achieved when n_1 is as large as possible, while n_2 is large enough for the sure screening selection to hold. This intuition is also validated in Figure 1, as efficiency is gained faster at the beginning due to better GLM estimators with larger n_1. This gain is then outweighed by the bias due to poor selections with small n_2. Our conclusion is that an optimal split proportion exists, but depends on the specific selection method, the true model size, and other factors, rather than being fixed.

We further examined the convergence of the two variance estimators $\widehat{V}_j$ and $\widehat{V}_j^B$ proposed in (5) and (6) with respect to the number of splits, B. Under the same setting, and with $q = 0.5$, we calculated both $\widehat{V}_j$ and $\widehat{V}_j^B$ for $B = 100,\ 200,\ \ldots,\ 2,000$, and compared these estimates with the empirical variance of $\widehat{\beta}_j$'s (considered to be the *truth*) based on 200 simulation replicates. The right panel of Figure 1 plots the averages over all signals $j = 1, 2, \ldots, p$ and shows $\widehat{V}_j$ converges to the truth much slower than $\widehat{V}_j^B$, and $\widehat{V}_j^B$ has small biases even with a relatively small B.

Example 2 implemented various selection methods, LASSO, SCAD, MCP, Elastic net, and Bayesian LASSO, when conducting variable selection for SSGLM, and compared their impacts on estimation and inference. Ten-fold cross-validation was used for the tuning parameters in each selection procedure. We assumed a Poisson model with $n = 300$, $p = 400$,

and $s_0 = 5$. For $i = 1, \ldots, n$,

$$\log\Big(\mathbf{E}\left(Y_i|\mathbf{x}_i\right)\Big) = \beta_0 + \mathbf{x}_i\boldsymbol{\beta}. \tag{8}$$

Table 1 reports the selection frequency for each j out of B splits. Larger $|\beta_j^*|$ yielded a higher selection frequency. For example, predictors with an absolute effect larger than 0.6 were selected frequently. The average size of the selected models by each method varied from 23 (for LASSO) to 8 (for Bayesian LASSO). However, in terms of the bias, coverage probabilities, and mean squared errors, the impact of the different variable selection methods seemed to be negligible. Thus, SSGLM was fairly robust to the choice of variable selection method.

Example 3 also assumed model (8). We set $n = 400$, $p = 500$, and $s_0 = 6$, with non-zero coefficients between 0.5 and 1, and three correlation structures: Identity; AR(1) with $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5$; Compound Symmetry (CS) with $\Sigma_{ij} = \rho^{I(i\neq j)}, \rho = 0.5$.

Table 2 shows that SSGLM consistently provided nearly unbiased estimates. The obtained standard errors (SEs) were close to the empirical standard deviations (SDs), leading to confidence intervals with coverage probabilities that were close to the 95% nominal level.

Example 4 assumed a logistic regression model for binary outcomes, with $n = 400$, $p = 500$, and $s_0 = 4$,

$$\text{logit}\Big(\mathbf{P}(Y_i = 1|\mathbf{x}_i)\Big) = \beta_0 + \mathbf{x}_i\boldsymbol{\beta}. \tag{9}$$

The index set for predictors with nonzero coefficients, $S^* = \{218,\ 242,\ 269,\ 417\}$, were randomly generated, and $\boldsymbol{\beta}^*_{S^*} = (-2,\ -1,\ 1,\ 2)$. We report the performance of SSGLM when inferring the subvector $\boldsymbol{\beta}^*_{S^*}$, in Tables 3 and 4. Our method gave nearly unbiased estimates under different correlation structures and sufficient power for the various contrasts.

Example 5 compared our method with the de-biased LASSO estimator (Van de Geer et al., 2014) and the de-correlated score test (Ning and Liu, 2017) in terms of power and type I error. We assumed model (9) with $n = 200$, $p = 300$, $s_0 = 3$, and $\boldsymbol{\beta}^*_{S^*} = (2,\ -2,\ 2)$ with AR(1) correlation structures. Table 5 summarises the power of detecting each true signal and the average type I error for the noise variables under the AR(1) correlation structure with four correlation values, $\rho = 0.25, 0.4, 0.6, 0.75$.

Our method was shown to be the most powerful, while maintaining the type I error around the nominal 0.05 level. The power was over 0.9 for the first three scenarios and was above 0.8 with $\rho = 0.75$. The de-biased LASSO estimators controlled the type I error well, but the power dropped from 0.9 to approximately 0.67 as the correlation among the predictors increased. The de-correlated score tests had the least power and the highest type I error. While these two competing methods have the same efficiency asymptotically, they do differ by specific implementations, for example, the choice of tuning parameters. Indeed, de-biased methods may be sensitive to tuning parameters, which could explain the gap in the finite sample performance.

Table 5 summarizes the average computing time (in seconds) of the three methods per dataset (R-3.6.2 on an 8-core MacBook Pro). On average, our method took 17.7 seconds, which was the fastest among the three methods. The other two methods were slower for the smaller ρ's (75 and 37 seconds, respectively) and faster for the larger ρ's (41 and 18 seconds, respectively), likely because the node-wise LASSO procedure that was used for estimating the precision matrix tended to be faster when handling more highly correlated predictors.

7. Data Example

We analyzed a subset of the BLCSC data (Christiani, 2017), consisting of $n = 1,459$ individuals, among whom 708 were lung cancer patients and 751 were controls. After cleaning, the data contained $6,829$ SNPs, along with important demographic variables including age, gender, race, education level, and smoking status (Table 6). As smoking is known to play a significant role in the development of lung cancer, we were particularly interested in estimating the interactions between the SNPs and smoking status, in addition to their main effects.

We assumed a high-dimensional logistic model with the binary outcome being an indicator of lung cancer status. Predictors included demographic variables, the SNPs (with prefix "AX"), and the interactions between the SNPs and smoking status (with prefix "SAX"; $p = 13,663$). We applied the SSGLM with $B = 1,000$ random splits and drew inference on all 13,663 predictors. Table 7 lists the top predictors ranked by their p-values. We identified 9 significant coefficients after using Bonferroni correction for multiple comparisons. All were interaction terms, providing strong evidence of SNP-smoking interactions, which have rarely been reported. These nine SNPs came from three genes, TUBB, ERBB2, and TYMS. TUBB mutations are associated with both poor treatment response to paclitaxel-containing chemotherapy and poor survival in patients with advanced non-small-cell lung cancer (NSCLC) (Monzó et al., 1999; Kelley et al., 2001). Rosell et al. (2001) has proposed using the presence of TUBB mutations as a basis for selecting initial chemotherapy for patients with advanced NSCLC. In contrast, intragenic ERBB2 kinase mutations occur more often in the adenocarcinoma lung cancer subtype (Stephens et al., 2004; Beer et al., 2002). Lastly, advanced NSCLC patients with low/negative thymidylate synthase (TYMS) are shown to have better responses to Pemetrexed–based chemotherapy and longer progression free survival (Wang et al., 2013).

For comparisons, we applied the de-sparsified estimator for GLM (Bühlmann et al., 2014). A direct application of the "lasso.proj" function in the "hdi" R package (Dezeure et al., 2015) was not feasible given the data size. Instead, we used a shorter sequence of candidate λ values and 5-fold instead of 10-fold cross validation for the node-wise LASSO procedure. This procedure costs approximately one day of CPU time. After correcting for multiple testing, there were two significant coefficients, both of which were interaction terms corresponding to SNPs AX.35719413_C and AX.83477746_A. Both SNPs were from the TUBB gene, and the first SNP was also identified by our method.

To validate our findings, we applied the prediction accuracy measures for nonlinear models proposed in Li and Wang (2019). We calculated the R^2, the proportion of variation explained in $\mathbf{Y}$, for the models we chose to compare. We report five models and their corresponding R^2 values: **Model 1.** the baseline model including only the demographic variables ($R^2 = 0.0938$); **Model 2.** the baseline model plus the significant interactions after the Bonferroni correction in Table 7 ($R^2 = 0.1168$); **Model 3.** Model 2 plus the main effects of its interaction terms ($R^2 = 0.1181$); **Model 4.** the baseline model plus the significant interactions from the de-sparsified LASSO method ($R^2 = 0.1018$); **Model 5.** Model 4 plus the corresponding main effects ($R^2 = 0.1076$). Model 2 based on our method explained 25% more variation in $\mathbf{Y}$ than the baseline model (from 0.0938 to 0.1168), while Model 4 based on the de-sparsified LASSO method only explains 8.5% more variation

(from 0.0938 to 0.1018). We also plotted Receiver-Operating Characteristic (ROC) curves for models 1, 2, and 4 (Figure 2). Their corresponding areas under the curves (AUCs) were 0.645, 0.69, and 0.668, respectively.

Previous literature has identified several SNPs as potential risk factors for lung cancer. We studied a controversial SNP, rs3117582, from the TUBB gene on chromosome 6. This SNP was identified in association with lung cancer risk in a case/control study by Wang et al. (2008), while on the other hand, Wang et al. (2009) found no evidence of association between the SNP and risk of lung cancer among *never-smokers*. Our goal was to test this SNP and its interaction with smoking in the presence of all the other predictors under the high dimensional logistic model. Slightly overusing notation, we denoted the coefficients corresponding to rs3117582 and its interaction with smoking as $\boldsymbol{\beta}^{(1)} = (\beta_1, \beta_2)$, and tested $H_0 : \beta_1 = \beta_2 = 0$. Applying the proposed method, we obtained

$$(\widehat{\beta}_1, \widehat{\beta}_2) = (-0.067, 0.005),\ \widehat{\mathrm{COV}}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \begin{pmatrix} 0.44, & -0.43 \\ -0.43, & 0.50 \end{pmatrix}.$$

The test statistic of the overall effect was $T = 0.062$ by (7) with a p-value of 0.97, which concluded that, among the patients in BLCSC, rs3117582 was not significantly related to lung cancer, regardless of the smoking status.

8. Conclusions

Our approach for drawing inference, by adopting a "split and smoothing" idea, improves upon Fei et al. (2019) which used bootstrap resampling, and recasts a high dimensional inference problem into a sequence of low dimensional estimations. Unlike many of the existing methods (Zhang and Zhang, 2014; Bühlmann et al., 2014; Javanmard and Montanari, 2018), our method is more computationally feasible as it does not require estimating high dimensional precision matrices. Our algorithm enables us to make full use of parallel computing for improved computational efficiency, because fitting the p low dimensional GLMs and randomly splitting the data B times are both separable tasks, which can be implemented in parallel.

We have derived the variance estimator using the infinitesimal jackknife method adapted to the splitting and smoothing procedure (Efron, 2014; Wager and Athey, 2018). This estimator is free of parametric assumptions, resembles bagging (Bühlmann and Yu, 2002), and leads to confidence intervals with correct coverage probabilities. Moreover, we have relaxed the stringent "selection consistency" assumption on variable selection as required in Fei et al. (2019). We have shown that our procedure works with a mild "sure screening" assumption for the selection method.

There are open problems to be addressed. First, our method relies on a sparsity condition for the model parameters. We envision that relaxation of the condition may take a major effort, though our preliminary simulations (Example B.2 in Appendix B) suggest that our procedure might work when the sparsity condition fails. Second, as our model is fully parametric, in-depth research is needed to develop a more robust approach when the model is mis-specified. Finally, while our procedure is feasible when p is large (tens of thousands), the computational cost increases substantially when p is extraordinarily large

(millions). Much effort is warranted to enhance its computational efficiency. Nevertheless, our work does provide a starting point for future investigations.

Acknowledgements

We are grateful towards Dr. Boaz Nadler and three referees for the insightful comments that have helped improve the manuscript. We thank Mr. Stephen Salerno, Department of Biostatistics, University of Michigan, for carefully proofreading the manuscript and for the excellent edits that have bettered the presentation of the manuscript. We thank our long time collaborator, Dr. David Christiani, Harvard Medical School, for providing the BLCSC data. The work is partially supported by grants from NIH (R01CA249096,R01AG056764 and U01CA209414).

References

David G Beer, Sharon LR Kardia, Chiang-Ching Huang, Thomas J Giordano, Albert M Levin, David E Misek, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8):816–824, 2002.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34 (4):606–619, 2016.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1: 255–278, 2014.

Xing Chen and Gui-Ying Yan. Semi-supervised learning for potential human microRNA-disease associations inference. *Scientific Reports*, 4:5501, 2014.

David C. Christiani. The Boston lung cancer survival cohort. `http://grantome.com/grant/NIH/U01-CA209414-01A1`, 2017. URL `http://grantome.com/grant/NIH/U01-CA209414-01A1`. [Online; accessed November 27, 2018].

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

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

Jianqing Fan, Yingying Fan, and Emre Barut. Adaptive robust variable selection. *The Annals of Statistics*, 42(1):324–351, 2014.

Zhe Fei, Ji Zhu, Moulinath Banerjee, and Yi Li. Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. *Biometrics*, 75(2):551–561, 2019.

Jerome H Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

Carmen Garrigos, Ana Salinas, Ricardo Melendez, Marta Espinosa, Iván Sánchez, et al. Clinical validation of single nucleotide polymorphisms (snps) as predictive biomarkers in localized and metastatic renal cell cancer (RCC)., 2018.

Nicolas Goossens, Shigeki Nakagawa, Xiaochen Sun, and Yujin Hoshida. Cancer biomarker discovery and validation. *Translational cancer research*, 4(3):256, 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

Daniel Sik Wai Ho, William Schierding, Melissa Wake, Richard Saffery, and Justin O'Sullivan. Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics*, 10, 2019.

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

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Michael J Kelley, Sufeng Li, and David H Harpole. Genetic analysis of the β-tubulin gene, tubb, in non-small-cell lung cancer. *Journal of the National Cancer Institute*, 93(24): 1886–1888, 2001.

Eva YHP Lee and William J Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology*, 2(10):a003236, 2010.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Gang Li and Xiaoyan Wang. Prediction accuracy measures for a nonlinear model and for right-censored time-to-event data. *Journal of the American Statistical Association*, 114 (528):1815–1825, 2019.

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

Jessica Minnier, Lu Tian, and Tianxi Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496): 1371–1382, 2011.

Mariano Monzó, Rafael Rosell, José Javier Sánchez, Jin S Lee, Aurora O'Brate, José Luis González-Larriba, et al. Paclitaxel resistance in non–small-cell lung cancer associated with beta-tubulin gene mutations. *Journal of Clinical Oncology*, 17(6):1786–1786, 1999.

Chulso Moon, Yun Oh, and Jack A Roth. Current status of gene therapy for lung cancer and head and neck cancer. *Clinical cancer research*, 9(14):5055–5067, 2003.

Wojciech Niemiro. Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, 20(3):1514–1533, 1992.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Stephen Portnoy. Asymptotic behavior of m estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985.

Rafael Rosell, Miquel Tarón, and Aurora O'brate. Predictive molecular markers in non–small cell lung cancer. *Current Opinion in Oncology*, 13(2):101–109, 2001.

Philip Stephens, Chris Hunter, Graham Bignell, Sarah Edkins, Helen Davies, Jon Teague, et al. Lung cancer: intragenic erbb2 kinase mutations in tumours. *Nature*, 431(7008): 525–526, 2004.

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(3):1166–1202, 2014.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15 (1):1625–1651, 2014.

Jingshen Wang, Xuming He, and Gongjun Xu. Debiased inference on treatment effect in a high-dimensional model. *Journal of the American Statistical Association*, 115(529): 442–454, 2020.

Mingqiu Wang and Xiuli Wang. Adaptive lasso estimators for ultrahigh dimensional generalized linear models. *Statistics & Probability Letters*, 89:41–50, 2014.

Ting Wang, Chang Chuan Pan, Jing Rui Yu, Yu Long, Xiao Hong Cai, Xu De Yin, et al. Association between tyms expression and efficacy of pemetrexed–based chemotherapy in advanced non-small cell lung cancer: A meta-analysis. *PLoS One*, 8(9):e74284, 2013.

Yufei Wang, Peter Broderick, Emily Webb, Xifeng Wu, Jayaram Vijayakrishnan, Athena Matakidou, et al. Common 5p15. 33 and 6p21. 33 variants influence lung cancer risk. *Nature Genetics*, 40(12):1407–1409, 2008.

Yufei Wang, Peter Broderick, Athena Matakidou, Timothy Eisen, and Richard S Houlston. Role of 5p15. 33 (TERT-CLPTM1L), 6p21. 33 and 15q25. 1 (CHRNA5-CHRNA3) variation and lung cancer risk in never-smokers. *Carcinogenesis*, 31(2):234–238, 2009.

Daniela M Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

Lu Xia, Bin Nan, and Yi Li. A revisit to de-biased lasso for generalized linear models. *arXiv preprint arXiv:2006.12778*, 2020.

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

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

Figure 1: Left: Average MSEs of all predictors at split proportions q's from 0.1 to 0.9. Right: Convergence of two variance estimators as B increases.

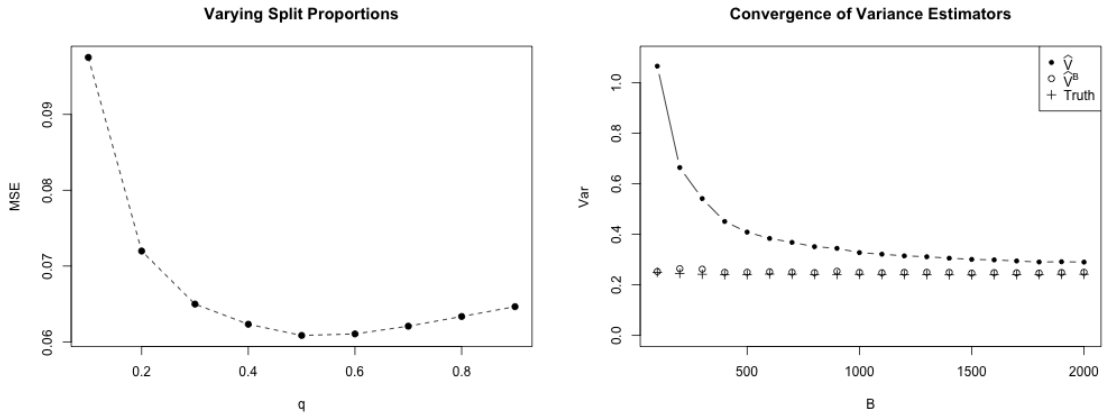

Figure 2: ROC curves of the three selected models.

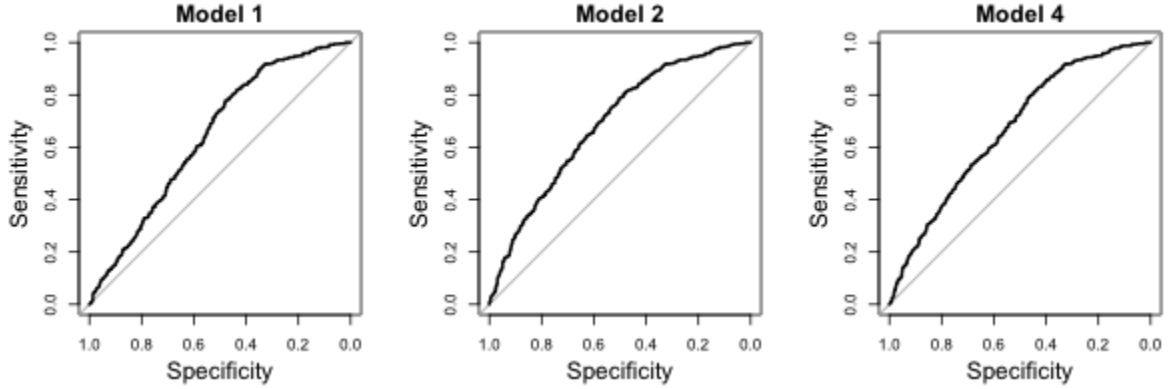

Table 1: Comparisons of different selection procedures to implement our proposed method. The first column is the indices of the non-zero signals. The last row for the selection frequency is the average number of predictors being selected by each procedure. The last row for the coverage probability is the average coverage probability of all predictors.

Index j	β_j^*	LASSO	SCAD	MCP	EN	Bayesian
Selection frequency						
12	0.4	0.59	0.55	0.49	0.60	0.60
71	0.6	0.93	0.92	0.90	0.95	0.94
351	0.8	0.99	0.99	0.99	1.00	1.00
377	1.0	1.00	1.00	1.00	1.00	1.00
386	1.2	1.00	1.00	1.00	1.00	1.00
Average model size		23.12	13.15	10.89	10.31	7.98
Bias						
12	0.4	0.003	0.003	0.003	0.003	0.001
71	0.6	0.007	0.008	0.008	0.008	-0.010
351	0.8	-0.001	0.001	0	0	0.001
377	1.0	-0.005	-0.005	-0.006	-0.005	0.001
386	1.2	0.002	0.001	0.001	0.001	0.004
Coverage probability						
12		0.90	0.90	0.91	0.91	0.95
71		0.94	0.94	0.95	0.94	0.94
351		0.95	0.95	0.95	0.94	0.95
377		0.94	0.93	0.93	0.94	0.92
386		0.94	0.95	0.95	0.95	0.94
Average		0.93	0.94	0.94	0.94	0.94
MSE						
12		0.111	0.110	0.110	0.109	0.106
71		0.104	0.103	0.102	0.102	0.101
351		0.103	0.103	0.103	0.103	0.100
377		0.101	0.100	0.100	0.100	0.109
386		0.097	0.096	0.096	0.096	0.102
Average		0.105	0.104	0.103	0.103	0.102

Table 2: SSGLM under the Poisson regression and three correlation structures. Bias, average standard error (SE), empirical standard deviation (SD), coverage probability (Cov prob), and selection frequency (Sel freq) are reported. The last column summarizes the average of all noise variables.

	Index j	0 (Int)	74	109	347	358	379	438	-
	β_j^*	1.000	0.810	0.595	0.545	0.560	0.665	0.985	0
Identity	Bias	-0.010	0	0	0.001	0.005	0.005	0.006	0
	SE	0.050	0.035	0.034	0.035	0.035	0.034	0.035	0.034
	SD	0.064	0.036	0.038	0.031	0.033	0.038	0.036	0.036
	Cov prob	0.870	0.920	0.900	0.960	0.990	0.910	0.950	0.936
	Sel freq	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.015
AR(1)	Bias	0.006	0.003	-0.002	-0.001	-0.001	-0.005	0.003	0
	SE	0.052	0.035	0.035	0.035	0.035	0.035	0.035	0.035
	SD	0.056	0.031	0.041	0.035	0.037	0.037	0.037	0.036
	Cov prob	0.930	0.970	0.890	0.960	0.950	0.930	0.960	0.937
	Sel freq	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.015
CS	Bias	-0.003	-0.005	0.004	-0.002	0.005	-0.004	-0.001	0.001
	SE	0.033	0.043	0.043	0.042	0.043	0.043	0.044	0.042
	SD	0.038	0.046	0.044	0.052	0.040	0.047	0.043	0.044
	Cov prob	0.960	0.900	0.930	0.900	0.970	0.910	0.950	0.934
	Sel freq	1.000	1.000	0.999	0.997	0.998	0.999	1.000	0.016

Table 3: SSGLM under the logistic regression, with estimation and inference for the subvector $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_{S^*}$. We compare the SSGLM (left) to the oracle model (right), where the oracle estimator is from the low dimensional GLM given the true set S^*, and the empirical covariance matrix is with respect to the simulation replications.

Index j	218	242	269	417	Index j	218	242	269	417
β_j^*	-2	-1	1	2	β_j^*	-2	-1	1	2
					Identity				
$\widehat{\boldsymbol{\beta}}^{(1)}$	-2.048	-1.043	0.999	2.096	Oracle	-1.995	-1.026	0.973	2.043
$\widehat{\Sigma}^{(1)}$	0.146	0.010	-0.009	-0.020	Empirical	0.155	0.006	-0.009	-0.027
	0.010	0.134	-0.004	-0.011		0.006	0.129	-0.011	-0.015
	-0.009	-0.004	0.134	0.009		-0.009	-0.011	0.152	0.010
	-0.020	-0.011	0.009	0.143		-0.027	-0.015	0.010	0.134
					AR(1)				
$\widehat{\boldsymbol{\beta}}^{(1)}$	-2.073	-1.014	1.002	2.110	Oracle	-2.024	-0.991	0.977	2.062
$\widehat{\Sigma}^{(1)}$	0.145	0.012	-0.011	-0.023	Empirical	0.141	0.012	-0.016	-0.028
	0.012	0.137	-0.006	-0.011		0.012	0.112	-0.006	0
	-0.011	-0.006	0.135	0.010		-0.016	-0.006	0.129	0.009
	-0.023	-0.011	0.010	0.147		-0.028	0	0.009	0.136
					CS				
$\widehat{\boldsymbol{\beta}}^{(1)}$	-2.095	-1.033	1.070	2.102	Oracle	-2.037	-1.024	1.027	2.028
$\widehat{\Sigma}^{(1)}$	0.223	-0.026	-0.048	-0.063	Empirical	0.192	-0.030	-0.044	-0.045
	-0.026	0.208	-0.043	-0.047		-0.030	0.187	-0.037	-0.044
	-0.048	-0.043	0.207	-0.028		-0.044	-0.037	0.165	-0.011
	-0.063	-0.047	-0.028	0.224		-0.045	-0.044	-0.011	0.179

Table 4: SSGLM under Logistic regression, with rejection rates of testing the contrasts. When the truth is 0, the rejection rates estimate the type I error probability; when the truth is nonzero, they estimating the testing power.

H_0	Truth	Identity	AR(1)	CS
$\beta^*_{218} + \beta^*_{417} = 0$	0	0.05	0.04	0.03
$\beta^*_{242} + \beta^*_{269} = 0$	0	0.06	0.04	0.025
$\beta^*_{218} + \beta^*_{269} = 0$	-1	0.56	0.57	0.42
$\beta^*_{242} + \beta^*_{417} = 0$	1	0.55	0.58	0.48
$\beta^*_{242} = 0$	-1	0.83	0.80	0.61
$\beta^*_{269} = 0$	1	0.74	0.81	0.70
$\beta^*_{218} = 0$	-2	1	1	1
$\beta^*_{417} = 0$	2	1	1	1

Table 5: Comparisons of SSGLM, Lasso-pro, and De-correlated score test (Dscore) in power, type I error and computing time. AR(1) correlation structure with different ρ's for $\mathbf{X}$ are assumed.

		Power		Type I error	Time
Truth	$\beta^*_{10} = 2$	$\beta^*_{20} = -2$	$\beta^*_{30} = 2$	$\beta^*_j = 0$	(secs)
$\rho = 0.25$ Proposed	0.920	0.930	0.950	0.049	17.7
Lasso-pro	0.900	0.930	0.900	0.042	74.7
Dscore	0.790	0.880	0.890	0.177	37.0
$\rho = 0.4$ Proposed	0.940	0.960	0.965	0.049	17.6
Lasso-pro	0.920	0.910	0.920	0.043	66.0
Dscore	0.770	0.905	0.840	0.175	30.7
$\rho = 0.6$ Proposed	0.940	0.950	0.880	0.054	17.7
Lasso-pro	0.850	0.750	0.850	0.045	53.3
Dscore	0.711	0.881	0.647	0.268	20.1
$\rho = 0.75$ Proposed	0.863	0.847	0.923	0.060	17.7
Lasso-pro	0.690	0.670	0.650	0.053	41.0
Dscore	0.438	0.843	0.530	0.400	17.9

Table 6: Demographic characteristics of the BLCSC SNP data.

	Controls (751)	Cases (708)
Race		
White	726	668
Black	5	22
Other	20	18
Education		
<High school	64	97
High school	211	181
>High school	476	430
Age		
Mean(sd)	59.7(10.6)	60(10.8)
Gender		
Female	460	437
Male	291	271
Pack years		
Mean(sd)	18.8(25.1)	46.1(38.4)
Smoking		
Ever	498	643
Never	253	65

Table 7: SSGLM fitted to the BLCSC data. SNP variables start with "AX"; interaction terms start with "SAX"; "Smoke" is binary (1=ever smoked, 0=never smoked). Rows are sorted by p-values.

Variable	$\widehat{\beta}$	SE	T	p-value	Adjusted P	Sel freq
SAX.88887606_T	0.33	0.02	17.47	$< 10^{-3}$	< 0.01	0.08
SAX.11279606_T	0.53	0.06	8.23	$< 10^{-3}$	< 0.01	0.00
SAX.88887607_T	0.29	0.04	6.97	$< 10^{-3}$	< 0.01	0.01
SAX.15352688_C	0.56	0.08	6.90	$< 10^{-3}$	< 0.01	0.01
SAX.88900908_T	0.54	0.09	5.95	$< 10^{-3}$	< 0.01	0.02
SAX.88900909_T	0.51	0.09	5.69	$< 10^{-3}$	< 0.01	0.02
SAX.32543135_C	0.78	0.14	5.49	$< 10^{-3}$	< 0.01	0.25
SAX.11422900_A	0.32	0.06	5.24	$< 10^{-3}$	< 0.01	0.09
SAX.35719413_C	0.47	0.10	4.63	$< 10^{-3}$	0.049	0.00
SAX.88894133_C	0.43	0.10	4.53	$< 10^{-3}$	0.08	0.00
SAX.11321564_T	0.47	0.11	4.44	$< 10^{-3}$	0.12	0.00
...						
AX.88900908_T	0.40	0.11	3.84	$< 10^{-3}$	1.00	0.00
Smoke	0.89	0.23	3.82	$< 10^{-3}$	1.00	-
...						

Appendix A: Proofs of Theorems

Proof of Theorem 1:

From the data split, $\mathbf{D}_1$ and $\mathbf{D}_2$ are mutually exclusive, thus S, from $\mathbf{D}_2$, is independent of $\mathbf{D}_1 = (\mathbf{Y}^1, \mathbf{X}^1)$. We show the asymptotics of $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$ in (3) with a diverging number of parameters p_s, by using the techniques and results from He and Shao (2000); Niemiro (1992). Without loss of generality, and to simplify notation, we let $j = 1 \in S$, then $S_{+j} = S$. The argument is the same if $1 \notin S$ and for any other j.

To proceed, we first restrict on the event of $\Omega = \{S \supseteq S^*\}$. With Assumption (A3), $\mathrm{P}(\Omega) \geq 1 - K_2(p \vee n_2)^{-1-c_2}$. Recall that

$$\widetilde{\boldsymbol{\beta}}_{S_{+j}} = \underset{\boldsymbol{\beta} \in \mathbf{R}^{|S|+1}}{\operatorname{argmin}} \ell_S(\boldsymbol{\beta}_S) = \underset{\boldsymbol{\beta} \in \mathbf{R}^{|S|+1}}{\operatorname{argmin}} \ell(\boldsymbol{\beta}_S; \mathbf{Y}^1, \mathbf{X}^1_S);$$
$$\widetilde{\beta}_1 = \left(\widetilde{\boldsymbol{\beta}}_{S_{+j}}\right)_1.$$

To apply Theorems 2.1 and 2.2 of He and Shao (2000) in the GLM case, we can verify that our Assumptions (A1) and (A2) will lead to their conditions (C1), (C2), (C4) and (C5) with $C = 1, r = 2$ and $A(n, p_s) = p_s$. To verify their (C3), we first note that their D_n is our I^*_S. Then for any $\boldsymbol{\beta}_S$, $\alpha \in \mathbf{R}^{p_s}$ such that $\|\alpha\|_2 = 1$, a second order Taylor expansion of $U_S(\boldsymbol{\beta}_S)$ around $\boldsymbol{\beta}^*_S$ leads to

$$\left|\alpha^{\mathrm{T}} \mathbf{E}_{\boldsymbol{\beta}^*} \left(U_S(\boldsymbol{\beta}_S) - U_S(\boldsymbol{\beta}^*_S)\right) - \alpha^{\mathrm{T}} I^*_S \left(\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\right)\right| \leq O(\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\|_2^2).$$

Hence,

$$\sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq (p_s/n)^{1/2}} \left|\alpha^{\mathrm{T}} \mathbf{E}_{\boldsymbol{\beta}^*} \left(U_S(\boldsymbol{\beta}_S) - U_S(\boldsymbol{\beta}^*_S)\right) - \alpha^{\mathrm{T}} I^*_S \left(\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\right)\right| \leq O(p_s/n) = o(n^{1/2}),$$

which means their (C3) follows. Thus, by Theorem 2.1 of He and Shao (2000),

$$\|\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_2^2 = o_p(p_s/n_1),$$

given $p_s \log p_s/n_1 \to 0$. Furthermore, by Theorem 2.2 of He and Shao (2000), if $p_s^2 \log p_s/n_1 \to 0$, then

$$\|n_1^{1/2}(\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S) + n_1^{-1/2}\{I^*_S\}^{-1} U_S(\boldsymbol{\beta}^*_S)\|_2 = o_p(1).$$

Releasing the restriction on Ω and with $\mathrm{P}(\Omega^c) = \mathrm{P}(S \not\supseteq S^*) \leq K_2(p \vee n_2)^{-1-c_2}$, we would still have $\|\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_2^2 = o_p(p_s/n_1)$, given $p_s \log p_s/n_1 \to 0$. To see this, for any $\epsilon > 0$, we can consider

$$\begin{aligned} & \mathrm{P}(\|(n_1/p_s)^{1/2}(\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_2 > \epsilon) \\ < \; & \mathrm{P}(\|(n_1/p_s)^{1/2}(\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_2 > \epsilon | \Omega)\mathrm{P}(\Omega) + \mathrm{P}(\Omega^c) \\ < \; & \mathrm{P}(\|(n_1/p_s)^{1/2}(\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S)\|_2 > \epsilon | \Omega) + K_2(p \vee n_2)^{-1-c_2}, \end{aligned}$$

where both terms in the last inequality converge to 0 as $n_1 \to \infty$ and $n_2 = (1-q)n_1/q$, with $0 < q < 1$ a constant. Similarly, we can show $\|n_1^{1/2}(\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S) + n_1^{-1/2}\{I^*_S\}^{-1} U_S(\boldsymbol{\beta}^*_S)\|_2 = o_p(1)$, if $p_s^2 \log p_s/n_1 \to 0$, which can also be written as

$$\widetilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S = -n_1^{-1}\{I^*_S\}^{-1} U_S(\boldsymbol{\beta}^*_S) + r_{n_1}, \tag{10}$$

with $\|r_{n_1}\|_2^2 = o_p(1/n_1)$. Consequently, by taking $\alpha = (0, 1, 0, \ldots, 0)^{\mathrm{T}}$ and left-multiplying both sides of (10) by $n^{1/2}\alpha^{\mathrm{T}}$, we have

$$\sqrt{n_1}\left(\widetilde{\beta}_1 - \beta_1^*\right)/\widetilde{\sigma}_1 \xrightarrow{d} N(0,1),$$

where $\widetilde{\sigma}_1^2 = \left(\{I_S^*\}^{-1}\right)_{11}$.

■

The following lemma, which is needed for the proof of Theorem 2, bounds the estimates of coefficients, when the selected subset S^b misses important predictors in S^* for some $1 \le b \le B$. Although $S^* \not\subseteq S^b$ with probability going to zero by Assumption (A3), we need to establish an upper bound in order to control the bias of $\widehat{\beta}_j$ for any j.

Lemma 1 *With model (1) and Assumptions (A1) and (A2), consider the GLM estimator $\widetilde{\boldsymbol{\beta}}_S$ with respect to subset S as defined in (3). Denote by $p_s = |S|$. If $p_s \log p_s/n \to 0$, then with probability going to 1, $|\widetilde{\boldsymbol{\beta}}_S|_\infty \le C_\beta$, where $C_\beta > 0$ is a constant depending on $c_{\min}, c_{\max}, c_\beta$, and $A(0)$.*

Proof By definition,

$$\widetilde{\boldsymbol{\beta}}_{S_{+j}} = \underset{\boldsymbol{\beta}_S \in \mathbf{R}^{p_s+1}}{\operatorname{argmin}}\ \ell_S(\boldsymbol{\beta}_S) = \underset{\boldsymbol{\beta}_S \in \mathbf{R}^{p_s+1}}{\operatorname{argmin}}\ \ell(\boldsymbol{\beta}_S; \mathbf{Y}^1, \mathbf{X}_S^1).$$

If $S^* \subseteq S$, the result immediately follows from Theorem 1 by taking $C_\beta = 2c_\beta$. When $S^* \not\subseteq S$, the minimizer $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$ is not an unbiased estimator of $\boldsymbol{\beta}_S^*$ anymore. However, we show that the boundedness of $\widetilde{\boldsymbol{\beta}}_{S_{+j}}$ is guaranteed from the strong convexity of the objective function $\ell_S(\boldsymbol{\beta}_S)$.

To see this, we note that the observed information is $\nabla^2\ell_S(\boldsymbol{\beta}_S) = \widehat{I}_S(\boldsymbol{\beta}_S) = \frac{1}{n}\overline{\mathbf{X}}_S{}^{\mathrm{T}}\mathbf{V}_S\overline{\mathbf{X}}_S$, where $\mathbf{V}_S = \operatorname{diag}\{A''(\overline{\mathbf{x}}_{1S}\boldsymbol{\beta}_S), \ldots, A''(\overline{\mathbf{x}}_{nS}\boldsymbol{\beta}_S)\}$ consisting of all positive diagonal entries, because of the positivity assumption on $A''(\cdot)$. Then for any column vector $w \in \mathbf{R}^{p_s+1}$, $\mathbf{V}_S^{1/2}\overline{\mathbf{X}}_S w = 0$ if and only if $\overline{\mathbf{X}}_S w = 0$, implying that the positive definiteness of $\nabla^2\ell_S(\boldsymbol{\beta}_S)$ is equivalent to that of $\widehat{\Sigma}_S = \frac{1}{n}\overline{\mathbf{X}}_S{}^{\mathrm{T}}\overline{\mathbf{X}}_S$. On the other hand, with $p_s \log p_s/n \to 0$, Lemma 1 of Fei et al. (2019) implies that, with probability going to 1, $\|\widehat{\Sigma}_S - \Sigma_S\| \le \varepsilon$ for $\varepsilon = \min(1/2, c_{\min}/2)$, and, hence,

$$\lambda_{\min}(\widehat{\Sigma}_S) \ge \lambda_{\min}(\Sigma_S) - \varepsilon \ge \lambda_{\min}(\Sigma) - \varepsilon \ge c_{\min}/2 > 0.$$

Thus, with probability going to 1, $\widehat{\Sigma}_S$ is positive definite, yielding that

$$\ell_S(\boldsymbol{\beta}_S) = n^{-1}\sum_{i=1}^{n}\{A(\overline{\mathbf{x}}_{iS}\boldsymbol{\beta}_S) - Y_i\overline{\mathbf{x}}_{iS}\boldsymbol{\beta}_S\}$$

is strongly convex with respect to $\boldsymbol{\beta}_S$. Hence, $\widetilde{\boldsymbol{\beta}}_{S_{+j}} \in \{\boldsymbol{\beta}_S : \ell_S(\boldsymbol{\beta}_S) \le A(0)\}$, which is a strongly convex set with probability going to 1. As $A(0)$ does not depend on S or the data, there exists a constant $C_\beta > 0$ (which only depends on $A(0)$, but does not depend on S or the data), such that $|\widetilde{\boldsymbol{\beta}}_S|_\infty \le C_\beta$ holds with probability going to 1. ■

Proof of Theorem 2:

We define the *oracle* estimators of β_j^* on the full data $(\mathbf{Y}, \mathbf{X})$ and the b-th subsample $\mathbf{D}_1^b$ respectively, where the candidate set is the true set S^*:

$$\check{\boldsymbol{\beta}}_{S^*_{+j}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{s_0+1}}{\operatorname{argmin}}\, \ell_{S^*_{+j}}(\boldsymbol{\beta}_{S^*_{+j}}) = \underset{\boldsymbol{\beta}\in\mathbf{R}^{s_0+1}}{\operatorname{argmin}}\, \ell_{S^*_{+j}}(\boldsymbol{\beta}_{S^*_{+j}}; \mathbf{Y}, \mathbf{X}_{S^*_{+j}}),\ \check{\beta}_j = \left(\check{\boldsymbol{\beta}}_{S^*_{+j}}\right)_j;$$
$$\check{\boldsymbol{\beta}}^b_{S^*_{+j}} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{s_0+1}}{\operatorname{argmin}}\, \ell^b_{S^*_{+j}}(\boldsymbol{\beta}_{S^*_{+j}}) = \underset{\boldsymbol{\beta}\in\mathbf{R}^{s_0+1}}{\operatorname{argmin}}\, \ell_{S^*_{+j}}(\boldsymbol{\beta}_{S^*_{+j}}; \mathbf{Y}^{1(b)}, \mathbf{X}^{1(b)}_{S^*_{+j}}),\ \check{\beta}^b_j = \left(\check{\boldsymbol{\beta}}^b_{S^*_{+j}}\right)_j.$$

By Theorem 1 and given $s_0^2 \log s_0/n \to 0$, for each $j \in \{1, \ldots, p\}$,

$$\sqrt{n}(\check{\beta}_j - \beta_j^*)/\check{\sigma}_j \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty, \tag{11}$$

where $\check{\sigma}_j^2 = \left(\{I^*_{S^*_{+j}}\}^{-1}\right)_{jj}$.

With the oracle estimators $\check{\beta}_j$'s and $\check{\beta}_j^b$'s, we have the following decomposition:

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\beta}_j - \beta_j^*\right)\\
=&\sqrt{n}\left(\check{\beta}_j - \beta_j^*\right) + \sqrt{n}\left(\widehat{\beta}_j - \check{\beta}_j\right)\\
=&\sqrt{n}\left(\check{\beta}_j - \beta_j^*\right) + \sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\widetilde{\beta}_j^b - \check{\beta}_j\right)\\
=&\underbrace{\sqrt{n}\left(\check{\beta}_j - \beta_j^*\right)}_{\text{I}} + \underbrace{\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\check{\beta}_j^b - \check{\beta}_j\right)}_{\text{II}} + \underbrace{\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\left\{\widetilde{\beta}_j^b - \check{\beta}_j^b\right\}\right)}_{\text{III}}.
\end{aligned} \tag{12}$$

The first two terms in (12), which do not involve various selections S^b's, deal with the oracle estimators and the true active set S^*. We need to show the following, which will lead to the results stated in the theorem by using Slutsky's theorem.

(a) $\text{I}/\check{\sigma}_j = \sqrt{n}\left(\check{\beta}_j - \beta_j^*\right)/\check{\sigma}_j \xrightarrow{d} N(0,1)$;

(b) $\text{II} = \frac{\sqrt{n}}{B}\sum_{b=1}^{B}\left\{\check{\beta}_j^b - \check{\beta}_j\right\} = o_p(1)$;

(c) $\text{III} = \frac{\sqrt{n}}{B}\sum_{b=1}^{B}\left\{\widetilde{\beta}_j^b - \check{\beta}_j^b\right\} = o_p(1)$.

First, (a) holds because of (11). To show (b), i.e. $\text{II} = o_p(1)$, we first denote $\xi_{b,n} = \sqrt{n}\left(\check{\beta}_j^b - \check{\beta}_j\right)$, then $\text{II} = \left(\sum_{b=1}^{B}\xi_{b,n}\right)/B$. Since the sampling indicator vectors, $\mathbf{J}_b$'s (defined in Section 4) are i.i.d, $\xi_{b,n}$'s are i.i.d conditional on data $\mathbf{D}^{(n)} = (\mathbf{Y}, \mathbf{X})$. The conditional distribution of $\sqrt{n}\left(\check{\beta}_j^b - \check{\beta}_j\right)$ given $\mathbf{D}^{(n)}$ is asymptotically the same as the unconditional distribution of $\sqrt{n}\left(\check{\beta}_j - \beta_j^*\right)$, which converges to zero Gaussian by (11). With the uniform boundedness of $\check{\beta}_j^b$ and $\check{\beta}_j$ as shown in Lemma 1, we can show that $\mathbf{E}\left(\xi_{b,n}|\mathbf{D}^{(n)}\right) \to 0$

and $\mathrm{Var}\,(\xi_{b,n}|\mathbf{D}^{(n)}) \to \check{\sigma}_j^2$ uniformly over $\mathbf{D}^{(n)}$ as $n \to \infty$. Furthermore, $\mathbf{E}\,(\mathrm{II}|\mathbf{D}^{(n)}) = \mathbf{E}\,(\xi_{b,n}|\mathbf{D}^{(n)})$, and $\mathrm{Var}\,(\mathrm{II}|\mathbf{D}^{(n)}) = \mathrm{Var}\,(\xi_{b,n}|\mathbf{D}^{(n)})/B$. Denote by Ω_n the sample space of $\mathbf{D}^{(n)}$. For any $\delta, \zeta > 0$, there exist $N_0, B_0 > 0$ such that when $n > N_0, B > B_0$,

$$\begin{aligned}
&\mathrm{P}(|\mathrm{II}| \geq \delta) \leq \int_{\Omega_n} \mathrm{P}\left(|\mathrm{II}| \geq \delta \,\middle|\, \mathbf{D}^{(n)}\right) \mathrm{dP}(\mathbf{D}^{(n)}) \\
\leq &\int_{\Omega_n} \mathrm{P}\left(|\mathrm{II} - \mathbf{E}\,(\mathrm{II}|\mathbf{D}^{(n)})| \geq \delta/2 \,\middle|\, \mathbf{D}^{(n)}\right) \mathrm{dP}(\mathbf{D}^{(n)}) \\
\leq &\int_{\Omega_n} \frac{\mathrm{Var}\,\left(\mathrm{II} \,\middle|\, \mathbf{D}^{(n)}\right)}{\delta^2/4} \mathrm{dP}(\mathbf{D}^{(n)}) \leq \frac{\check{\sigma}_j^2}{B_0\delta^2/4} \int_{\Omega_n} \mathrm{dP}(\mathbf{D}^{(n)}) \leq \zeta.
\end{aligned}$$

Thus, $\mathrm{II} = o_p(1)$.

To prove (c), i.e. $\mathrm{III} = o_p(1)$, we first note that each subsample $\mathbf{D}_1^b$ can be regarded as a random sample of $n_1 = qn$ $(0 < q < 1)$ i.i.d. observations from the population distribution for which Assumption (A3) holds, that is $|S^b| \leq K_1 n^{c_1}$ and $\mathrm{P}\left(S^* \subseteq S^b\right) \geq 1 - K_2(p \vee n)^{-1-c_2}$. We show that for any b, conditional on $S^b \supseteq S^*$, $\sqrt{n}\left(\widetilde{\beta}_j^b - \check{\beta}_j^b\right) = o_p(1)$.

To see this, we first arrange the order of the components of $\mathbf{x} = (x_1, \ldots, x_p)$ such that the first s_0 components are signal variables. In other words, $S^* = \{1, \ldots, s_0\}$. From (10) in in the proof of Theorem 1 and omitting superscript b, we have that

$$\begin{aligned}
\widetilde{\beta}_j - \beta_j^* &= -n_1^{-1}\widetilde{e}_j^{\mathrm{T}}\{I^*_{S_{+j}}\}^{-1}U_{S_{+j}}(\boldsymbol{\beta}^*_{S_{+j}}) + \widetilde{r}_{n_1}, \\
\check{\beta}_j - \beta_j^* &= -n_1^{-1}\check{e}_j^{\mathrm{T}}\{I^*_{S^*_{+j}}\}^{-1}U_{S^*_{+j}}(\boldsymbol{\beta}^*_{S^*_{+j}}) + \check{r}_{n_1},
\end{aligned} \tag{13}$$

where $\widetilde{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ is a unit vector of length $|S_{+j}|$ to index the position of variable j in S_{+j}, $\check{e}_j$ is a unit vector of length $|S^*_{+j}|$ to index the position of variable j in S^*_{+j}, and the residuals $\|\widetilde{r}_{n_1}\|_2^2 = o_p(1/n_1), \|\check{r}_{n1}\|_2^2 = o_p(1/n_1)$. Here, $I^*_{S_{+j}}$ and $I^*_{S^*_{+j}}$ are two submatrices of the expected information at $\boldsymbol{\beta}^*$, i.e. $I^* = \mathbf{E}\,\{\frac{1}{n}\overline{\mathbf{X}}^{\mathrm{T}}\mathbf{V}\overline{\mathbf{X}}\}$, where $\mathbf{V} = \mathrm{diag}\{\nu(\mu_1), \ldots, \nu(\mu_n)\}$ is an $n \times n$ diagonal matrix with $\mu_i = g^{-1}(\overline{\mathbf{x}}_i\boldsymbol{\beta}^*)$; see Section 2.1 for the notation.

Therefore, the jk-th $(j, k = 0, 1, \ldots, p)$ entry of I^*, a $(p+1) \times (p+1)$ matrix, is $\mathbf{E}\,(x_j\nu(\mu)x_k)$ with $\mu = g^{-1}(\overline{\mathbf{x}}\boldsymbol{\beta}^*)$. Now for any $j \in S^*$, $k \in S^c$, the complement of S^*, the partial orthogonality condition (Fan and Lv, 2008; Fan and Song, 2010) that $\{x_j, j \in S^*\}$ are independent of $\{x_k, k \in S^c\}$ implies that $\mathbf{E}\,(x_j\nu(\mu)x_k) = 0$, as μ only depends on $x'_j, j' \in S^*$ and $\mathbf{E}\,(x_k) = 0$ with centered predictors. Therefore, I^* is block-diagonal with two blocks indexed by S^* and S^c. That is,

$$I^* = \begin{pmatrix} \mathbf{E}\,(\frac{1}{n}\overline{\mathbf{X}}^{\mathrm{T}}_{S^*}\mathbf{V}\overline{\mathbf{X}}_{S^*}) & 0 \\ 0 & \mathbf{E}\,(\frac{1}{n}\mathbf{X}^{\mathrm{T}}_{S^c}\mathbf{V}\mathbf{X}_{S^c}) \end{pmatrix}.$$

where the submatrices $\overline{\mathbf{X}}_{S^*}$ and $\mathbf{X}_{S^c}$ are as defined in Section 2.1. Hence, $I^*_{S_{+j}}$ is block-diagonal with two blocks indexed by S^* and $S_{+j} \setminus S^*$, and $I^*_{S^*_{+j}}$ is block-diagonal with two blocks indexed by S^* and $S^*_{+j} \setminus S^* = \emptyset$ if $j \in S^*$ or $= \{j\}$ otherwise.

Therefore, $\{I^*\}^{-1}$, $\{I^*_{S_{+j}}\}^{-1}$ and $\{I^*_{S^*_{+j}}\}^{-1}$ are all block-diagonal. Furthermore, the blocks corresponding to S^* in $\{I^*_{S_{+j}}\}^{-1}$ and $\{I^*_{S^*_{+j}}\}^{-1}$ are identical and are equal to $\{\mathbf{E}\,(\frac{1}{n}\overline{\mathbf{X}}^{\mathrm{T}}_{S^*}\mathbf{V}\overline{\mathbf{X}}_{S^*})\}^{-1}$.

Write $U(\boldsymbol{\beta}^*) = (u_0, u_1, u_2, \ldots, u_p)^{\mathrm{T}}$, $\widetilde{e}_j^{\mathrm{T}}\{I^*_{S_{+j}}\}^{-1} = (\widetilde{i}_{jk})_{k\in S_{+j}}$ and $\check{e}_j^{\mathrm{T}}\{I^*_{S^*_{+j}}\}^{-1} = (\check{i}_{jk})_{k\in S^*_{+j}}$. Then, it follows that $\widetilde{i}_{jk} = \check{i}_{jk}$ for $k \in S^*$, which leads to

$$\begin{aligned}\sqrt{n_1}\left(\widetilde{\beta}_j - \check{\beta}_j\right) &= -\frac{1}{\sqrt{n_1}}\sum_{k\in S_{+j}}\widetilde{i}_{jk}u_k + \frac{1}{\sqrt{n_1}}\sum_{k\in S^*_{+j}}\check{i}_{jk}u_k + r'_{n_1}\\ &= -\frac{1}{\sqrt{n_1}}\sum_{k\in S\setminus S^*}\widetilde{i}_{jk}u_k + \frac{1(j\notin S^*)}{\sqrt{n_1}}\left(\check{i}_{jj} - \widetilde{i}_{jj}\right)u_j + r'_{n_1}\end{aligned}$$

where $r'_{n_1} = \sqrt{n_1}(\widetilde{r}_{n_1} - \check{r}_{n_1}) = o_p(1)$, and $\widetilde{r}_{n_1}$ and $\check{r}_{n_1}$ are as in (13).

With Assumption (A3), $|S \setminus S^*| \le K_1 n^{c_1} = o(\sqrt{n_1})$ with $0 \le c_1 < 1/2$ and, thus, $\mathrm{Var}\left(\sum_{k\in S\setminus S^*}\widetilde{i}_{jk}u_k\right) = o(n_1)$. By the Chebyshev inequality, the first term on the right hand side converges to 0 in probability. Thus, each of these three terms is $o_p(1)$ and we have $\sqrt{n_1}\left(\widetilde{\beta}_j - \check{\beta}_j\right) = o_p(1)$. As $n_1/n = q$ where $0 < q < 1$, the original statement holds.

Now define $\eta_b = 1\left\{S^* \not\subseteq S^b\right\}\sqrt{n}\left(\widetilde{\beta}_j^b - \check{\beta}_j^b\right)$, while omitting subscripts j in η for simplicity, then $\mathrm{III} = \left(\sum_{b=1}^B \eta_b\right)/B$. When $S^* \not\subseteq S^b$, $\widetilde{\beta}_j^b$ is not an unbiased estimator of β_j^* any more, instead we try to bound it by some constant. By Lemma 1, there exists $C_\beta \ge c_\beta$ such that $\sup_b\left|\widetilde{\beta}_j^b - \check{\beta}_j^b\right| \le \sup_b\left|\widetilde{\beta}_j^b - \beta_j^*\right| + \sup_b\left|\check{\beta}_j^b - \beta_j^*\right| \le 2C_\beta + 1$. Therefore, by (A3),

$$\begin{aligned}\mathbf{E}(\eta_b) &\le \mathrm{P}\left(S^* \not\subseteq S^b\right)\sqrt{n}\sup_{1\le b\le B}\left|\widetilde{\beta}_j^b - \check{\beta}_j^b\right| \le 2C_\beta\sqrt{n}K_2(p\vee n)^{-1-c_2},\\ \mathrm{Var}(\eta_b) &\le \mathrm{P}\left(S^* \not\subseteq S^b\right)n\sup_{1\le b\le B}\left(\widetilde{\beta}_j^b - \check{\beta}_j^b\right)^2 \le 4C_\beta^2 nK_2(p\vee n)^{-1-c_2}. \qquad (14)\end{aligned}$$

With dependent η_b's, we further have

$$\begin{aligned}\mathbf{E}(\mathrm{III}) &= \mathbf{E}\left\{\left(\sum_{b=1}^B \eta_b\right)/B\right\} \le 2C_\beta\sqrt{n}K_2(p\vee n)^{-1-c_2},\\ \mathrm{Var}(\mathrm{III}) &\le \frac{1}{B^2}\sum_{b=1}^B\sum_{b'=1}^B|\mathrm{Cov}(\eta_b,\eta_{b'})| \le 4C_\beta^2 nK_2(p\vee n)^{-1-c_2},\end{aligned}$$

where the last inequality holds because of $|\mathrm{Cov}(\eta_b,\eta_{b'})| \le \{\mathrm{Var}(\eta_b)\mathrm{Var}(\eta_{b'})\}^{1/2}$ and (14). Then we show $\mathrm{III} = o_p(1)$. More specifically, for any $\delta > 0, \zeta > 0$, take $N_0 = \lfloor(C_\beta\delta)^{1/2+c_2}\rfloor$, where, for a real number a, $\lfloor a\rfloor$ denotes the integer part of a. When $n > N_0$, $\mathbf{E}(\mathrm{III}) \le \delta/2$. Also let $N_1 = \lfloor\{\zeta\delta^2/(16C_\beta^2K_2)\}^{c_2}\rfloor$. Then when $n > \max(N_0, N_1)$, we have

$$\begin{aligned}\mathrm{P}(|\mathrm{III}| \ge \delta) &\le \mathrm{P}(|\mathrm{III} - \mathbf{E}(\mathrm{III})| \ge \delta/2)\\ &\le \frac{\mathrm{Var}\ (\mathrm{III})}{\delta^2/4} \le \frac{16C_\beta^2K_2}{\delta^2}n(p\vee n)^{-1-c_2}\\ &< \frac{16C_\beta^2K_2}{\delta^2}n^{-c_2} < \zeta,\end{aligned}$$

where the first inequality is due to $|\mathbf{E}(\mathrm{III})| \le \delta/2$ when $n > N_0$, the second one is due to the Chebyshev inequality and the last one is due to $n > N_1$. ■

Proof of Theorem 3:

Following the previous proof, we replace the arguments in j with those in $S^{(1)}$. The *oracle* estimators are

$$\check{\boldsymbol{\beta}}_{S^{(1)}\cup S^*} = \operatorname{argmin} \ell_{S^{(1)}\cup S^*}(\boldsymbol{\beta}_{S^{(1)}\cup S^*}; \mathbf{Y}, \mathbf{X}_{S^{(1)}\cup S^*}),\ \check{\boldsymbol{\beta}}_{S^{(1)}} = \left(\check{\boldsymbol{\beta}}_{S^{(1)}\cup S^*}\right)_{S^{(1)}};$$
$$\check{\boldsymbol{\beta}}^b_{S^{(1)}\cup S^*} = \operatorname{argmin} \ell_{S^{(1)}\cup S^*}(\boldsymbol{\beta}_{S^{(1)}\cup S^*}; \mathbf{Y}^{1(b)}, \mathbf{X}^{1(b)}_{S^{(1)}\cup S^*}),\ \check{\boldsymbol{\beta}}^b_{S^{(1)}} = \left(\check{\boldsymbol{\beta}}^b_{S^{(1)}\cup S^*}\right)_{S^{(1)}}.$$

Notice that $|S^{(1)}| = p_1 = O(1)$, as $n \to \infty$, $|S^* \cup S^{(1)}| = O\big(|S^*|\big) = o(n)$, so that the above quantities are well-defined. The oracle estimator follows

$$\sqrt{n}\left\{(I^*_{S^{(1)}\cup S^*})^{-1/2}\right\}_{S^{(1)}}\left(\check{\boldsymbol{\beta}}_{S^{(1)}} - \boldsymbol{\beta}^*_{S^{(1)}}\right) \xrightarrow{d} N(0, \mathbf{I}_{p_1}) \quad \text{as} \quad n \to \infty.$$

Here, for a square matrix, say, Q, $(Q)_S$ is a submatrix of Q with rows and columns indexed by S. Denote by $I^{(1)} = \left\{(I^*_{S^{(1)}\cup S^*})^{-1/2}\right\}_{S^{(1)}}$.

Similar to (12), we have a decomposition

$$\sqrt{n}I^{(1)}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*_{S^{(1)}}) = \sqrt{n}I^{(1)}\left(\check{\boldsymbol{\beta}}_{S^{(1)}} - \boldsymbol{\beta}^*_{S^{(1)}}\right) + \sqrt{n}I^{(1)}\left(\frac{1}{B}\sum_{b=1}^{B}\widetilde{\boldsymbol{\beta}}^b_{S^{(1)}} - \check{\boldsymbol{\beta}}_{S^{(1)}}\right).$$

Analogous to the derivations in the previous proof, it follows that the second term is $o_p(1)$. Hence, the theorem holds. ■

Appendix B: Additional Simulations

To assess the robustness of our method, we perform some simulations when the parametric model is mis-specified and when the sparsity condition is violated.

Example B.1 assumes that $Y|\mathbf{x}$ follows a negative binomial distribution:

$$\mathrm{P}(Y = y) = \frac{\Gamma(y + r)}{\Gamma(r)y!}p^r(1 - p)^y,$$
$$\mathbf{E}\,Y = \mu = r(1 - p)/p = \mathbf{x}\boldsymbol{\beta}^*,$$

with $r = 10$, sample size $n = 300$, $p = 500$, and $s_0 = 5$. However, we model the data using SSGLM under the Poisson regression (8) with $B = 300$. Table 8 summarizes the results based on 200 simulated datasets. The $\widehat{\beta}_j$'s have small biases. The estimated standard errors are slightly less than the empirical standard deviations. Nevertheless, the coverage probabilities are still close to the 0.95 nominal level.

Example B.2 assumes a non-sparse truth $\boldsymbol{\beta}^*$ under the Poisson truth (8). With $n = 300$ and $p = 500$, we let $s_0 = 100$. Among the 100 predictors with non-zero effects, 96 β^*_j's are small, which are randomly drawn from Unif$[-0.5, 0.5]$, and the other 4 have values $-1.5, -1, 1, 1.5$ (as shown in Figure 3). With many small but non-zero signals, SSGLM still gives nearly unbiased estimates to all of them. See Table 9, where the columns represent 4 large size β^*_j's, and the averages over all small signals and all noise variables, respectively.

Figure 3: SSGLM under non-sparse truth, with $p = 500$ and $s_0 = 100$.

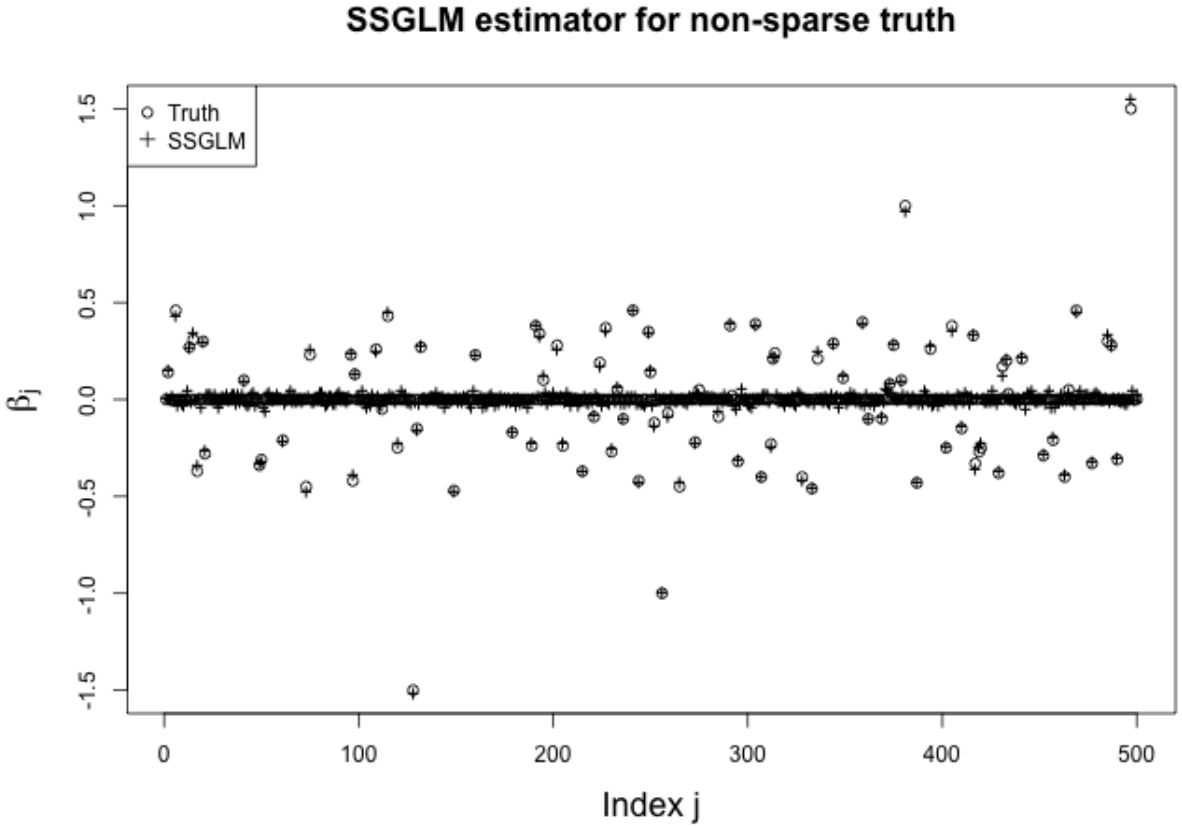

Table 8: SSGLM under mis-specified model.

Index j	90	179	206	237	316	Noise
β_j^*	-1.000	-0.500	0.500	1.000	1.500	0.000
Bias	-0.020	0.020	0.018	0.001	0.010	-0.001
SE	0.240	0.235	0.232	0.236	0.243	0.233
SD	0.258	0.243	0.249	0.249	0.250	0.231
Cov prob	0.955	0.945	0.900	0.930	0.925	0.946
Sel freq	0.724	0.177	0.216	0.723	0.977	0.021

Table 9: SSGLM under non-sparse truth.

Index j	128	256	381	497	Small	Noise
β_j^*	-1.50	-1.00	1.00	1.50	-	0
Bias	-0.01	0.003	-0.03	0.05	0.003	9×10^{-4}
SE	0.31	0.30	0.30	0.31	0.29	0.30
SD	0.31	0.30	0.32	0.30	0.29	0.29
Cov prob	0.93	0.93	0.93	0.94	0.94	0.94
Sel freq	0.87	0.50	0.49	0.87	0.06	0.03