Check for updates

Received: 18 June 2019 | Revised: 21 July 2020 | Accepted: 22 July 2020

DOI: 10.1111/biom.13346

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A weak-signal-assisted procedure for variable selection and statistical inference with an informative subsample

Fang Fang[1] | **Jiwei Zhao**[2] | **S. Ejaz Ahmed**[3] | **Annie Qu**[4]

[1] Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, China

[2] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin

[3] Faculty of Mathematics and Science, Brock University, St. Catharines, Ontario, Canada

[4] Department of Statistics, University of California, Irvine, California

Correspondence
Annie Qu, Department of Statistics, University of California, Irvine, CA 92697.
Email: aqu2@uci.edu

Funding information
Natural Sciences and Engineering Research Council of Canada; National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR001412; National Natural Science Foundation of China, Grant/Award Numbers: 11831008, 11601156, 11771146; National Science Foundation, Grant/Award Numbers: DMS-1613190, 1821198, 1953526

Abstract

This paper is motivated from an HIV-1 drug resistance study where we encounter three analytical challenges: to analyze data with an informative subsample, to take into account the weak signals, and to detect important signals and also conduct statistical inference. We start with an initial estimation method, which adopts a penalized pairwise conditional likelihood approach for variable selection. This initial estimator incorporates the informative subsample issue. To accounting for the effect of weak signals, we use a key idea of partial ridge regression. We also propose a one-step estimation method for each of the signal coefficients and then construct confidence intervals accordingly. We apply the proposed method to the Stanford HIV-1 drug resistance study and compare the results with existing approaches. We also conduct comprehensive simulation studies to demonstrate the superior performance of our proposed method.

KEYWORDS

informative subsample, pairwise conditional likelihood, penalization, post-selection inference, variable selection, weak signal

1 | INTRODUCTION

1.1 | Significant challenges in scientific problems

One of major goals in data-driven scientific discovery is to identify useful associations among different features and to quantify the uncertainty of the corresponding associations. However, due to the nature of complex study designs in observational studies, there are several challenges to overcome in a variety of disciplines. For example, in high throughput genotyping study, it is usually believed that there only exist a small number of mutations that truly affect the outcome, while there may also exist some so-called *weak* signals that could mystify the identification and estimation of the *strong* signals.

The other challenges are probably more prominent, and these occur when data collection is determined by clinical-care decisions, in that only those selected for a medical intervention or treatment contribute to the data collection. Therefore, the selected sample could differ clinically from those not selected, and it would lead to a nonrepresentatively informative subsample.

Our motivating example comes from the study of identifying genotype mutations for human immunodeficiency virus type 1 (HIV-1) drug resistance (Rhee *et al.*, 2003;

 © 2020 The International Biometric Society

Shafer, 2006). It is essential to understanding the genetic basis of HIV-1 drug resistance to develop new antiretroviral drugs so the use of existing drugs can be optimized. However, as pointed out by Abraham *et al.* (2011), the study of HIV-1 drug resistance is heavily influenced by the sample selection. Standard laboratory tests measure genotypic resistance only among subjects with a high viral load. Consequently, individuals who undergo genotype testing are likely very different clinically from those who do not have genotype testing. Therefore, the study only based on selected samples may prompt biased outcome or even lead to incorrect conclusion.

Besides the HIV-1 drug resistance study, the informative subsample issue also exists in many other biomedical studies:

- *Booster dose*: In medicine, a booster dose is an extra administration of a vaccine after the prime dose. Its intention is to increase the immunity against the antigen back to protective level (Crum-Cianflone and Wallace, 2014). For example, a booster dose of the measles vaccine could produce protective antibody responses and immunologic memory in children receiving HIV therapy (Abzug *et al.*, 2012). In a study conducted by the investigators from the International Maternal, Pediatric, Adolescent AIDS Clinical Trials (IMPAACT) Network, 193 children were sampled and monitored, and then a subset of 65 children was given another dose of the measles, mumps, and rubella (MMR) vaccine after approximately 4 years' follow-up. It is clear that the 128 children who were not given a booster dose are not selected from a random sample of the original population under investigation. Therefore, they are a biased sample. The statistical analysis based on a biased sample could lead to biased estimation and incorrect conclusion.
- *Rescue therapy*: In a multicenter clinical study of investigating biomarkers of periodontal disease progression (Teles *et al.*, 2016), around 375 subjects with periodontal disease and 125 periodontally healthy subjects were enrolled. All subjects were monitored clinically and had samples taken every 2 months for 12 months for analysis of inflammatory and immunological biomarkers and microbial species. Subjects displaying periodontal disease progression greater than or equal to a 4-mm increase of clinical attachment loss (CAL) received periodontal rescue therapy at progressing periodontal sites and then continue with monitoring. Therefore, the periodontal sites which did not go through the rescue therapy had the CAL increase less than 4 mm. Apparently, these sites also formed a biased sample. Similarly, the analysis without an additional adjustment would also produce biased results.
- *Second-line treatment*: In complex cancer studies such as ovarian cancer (Markman and Bookman, 2000) and gastric cancer (Wesolowski *et al.*, 2009), although some chemotherapy regimens have been well established as first-line therapy and have been shown to increase survival, almost all patients with epithelial ovarian cancer or metastatic gastric cancer will ultimately relapse and develop progressive disease after first-line therapy. Thus, there is a common need in the cancer studies to consider the use of second-line chemotherapeutic options. The patients who show benefits after the initial treatment remain good candidates for additional therapy. The clinical outcome from the patients who went through the second-line therapy should be treated differently from the ones who did not, as otherwise it would produce a biased sampling issue.

1.2 | Bridging the hypothetical population and the source population

One potential remedy to circumvent this sample selection difficulty is to introduce the concept of the *hypothetical population* which encompasses hypothetical representative random sample, and the concept of the *source population* where the actual observations form a random sample. We introduce a dichotomous random variable R denoting whether a subject is selected into the random sample of the source population. One might connect this setting to the missing data framework and treat R as the missing data indicator. However, there are differences between these two, since once a sample is not selected we do not have any information on subjects in our context, whereas missing data could still provide partial information from the outcome or some of the covariate variables.

To better understand the concepts of hypothetical and source populations, we introduce the notation $p(Y \mid \mathbf{X})$ to be the probability density function (pdf) of Y given $\mathbf{X}$ for the hypothetical population. The standard analysis based on the observed samples can only estimate or make an inference for the source population $p(Y \mid \mathbf{X}, R = 1)$, denoted as $p_1(Y \mid \mathbf{X})$ throughout the paper. Note that $p_1(Y \mid \mathbf{X})$ equals $p(Y \mid \mathbf{X})$ multiples the term $\Pr(R = 1 \mid Y, \mathbf{X}) / \Pr(R = 1 \mid \mathbf{X})$. It is obvious that, if $\Pr(R = 1 \mid Y, \mathbf{X})$ only depends on $\mathbf{X}$, that is, $\Pr(R = 1 \mid Y, \mathbf{X}) = \Pr(R = 1 \mid \mathbf{X})$, say, $a(\mathbf{X})$, then the source population is the same as the hypothetical population for the purpose of estimating the pdf of Y given $\mathbf{X}$. This implies that the naive analysis only based on the observed samples would produce valid estimation for the unknown parameter in $p(Y \mid \mathbf{X})$ only when the assumption $\Pr(R = 1 \mid Y, \mathbf{X}) = a(\mathbf{X})$ is correct.

In practice, it is difficult to specify the correct form to estimate $\Pr(R = 1 \mid Y, \mathbf{X})$ in general. Different study

designs would provide some prior information for the functional form of $\Pr(R = 1 \mid Y, \mathbf{X})$. For example, the HIV-1 drug resistance study ascertains that it is sensible to assume $\Pr(R = 1 \mid Y, \mathbf{X}) = b(Y)$; in other words, it only depends on the outcome Y—the measurement to quantify the degree of the drug resistance.

1.3 | Novel contributions of proposed approach

In the above discussion, we either assumed $\Pr(R = 1 \mid Y, \mathbf{X}) = a(\mathbf{X})$ or $\Pr(R = 1 \mid Y, \mathbf{X}) = b(Y)$. It is difficult or even infeasible to verify the correct form of $\Pr(R = 1 \mid Y, \mathbf{X})$ in practice, not to mention to estimate $a(\mathbf{X})$ or $b(Y)$ accurately. Therefore, it is ideal to propose a statistical procedure which avoids estimating $\Pr(R = 1 \mid Y, \mathbf{X})$. In this paper, we develop a new method which imposes the assumption on $\Pr(R = 1 \mid Y, \mathbf{X})$ at a minimum level so that the procedure would be robust to the model misspecification of $\Pr(R = 1 \mid Y, \mathbf{X})$ at the maximum level.

The conventional penalized conditional likelihood theory relies on a strong assumption on the minimum signal strength, which aims to estimate the coefficients for *strong* signals without considering *weak* signals as weak signals tend to be shrunk towards zero. However, weak signals exist in the HIV-1 drug resistance study (Shi and Qu, 2017) and present quite ubiquitously in general. Donoho and Jin (2008) illustrated that a single weak signal might not contribute significantly to the response, but all of them combined together could have significant influence towards scientific discovery. In general, identification of weak signals could facilitate us to demystify the entire scope of studies. In this paper, we develop a procedure to identify the weak signals in the penalized conditional likelihood framework. We show that, compared to the methods without identifying the weak signal, our method is able to provide a more precise and efficient estimator.

In addition, we also develop a postdetection statistical inference procedure for detected strong and weak signals. Postselection inference has drawn lots of attention in the statistical community recently, for example, Efron (2014), Lockhart *et al.* (2014), Zhang and Zhang (2014), Van de Geer *et al.* (2014), Javanmard and Montanari (2014) and among others. Since our procedure is built upon a conditional likelihood, which is not a conventional likelihood, it brings additional challenges including derivations of an appropriate score function and generating a score-type test statistic. We develop a one-step estimator (Bickel, 1975), which is asymptotically normally distributed, and construct a confidence interval of the parameter estimator corresponding to either strong or weak signals.

The reminder of the paper is organized as follows. The proposed method consisting of three steps, including the initial penalized conditional likelihood method, the partial ridge regression to take into account the weak signals, and the postselection inference procedure, is illustrated in Section 2. The algorithm to execute our proposed method is contained in Section 3. The application to the HIV-1 drug resistance study is detailed in Section 4. Some results from numerical studies with simulated data are shown in Section 5. The paper is concluded with a discussion in Section 6.

2 | METHODOLOGY

For ease of presentation, throughout the paper we assume a linear model between outcome Y and p-dimensional covariate $\mathbf{X}$ as

$$Y = \beta_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma^2)$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ as a vector of unknown parameters. This assumption is not a dealbreaker for our proposed method. The proposed method could be applied to other models as well such as the generalized linear model. We use notation * in the superscript of the parameter to denote its underlying true value.

We assume there is a random sample from the hypothetical population with size N. Following the discussion in Section 1, the binary variable R equals 1 means that the corresponding subject is selected into the random sample of the source population, of size n with $\{y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$, which are independent and identically distributed. In practice, one might only observe n without any knowledge of N. To link the source population to the hypothetical population in our proposed modeling framework, we assume that the binary variable R satisfies

$$\Pr(R = 1 \mid Y, \mathbf{X}) = a(\mathbf{X})b(Y), \tag{2}$$

that is, we assume the quantity $\Pr(R = 1 \mid Y, \mathbf{X})$ can be decomposed into a product of a function of $\mathbf{X}$-only and a function of Y-only, where the forms of $a(\mathbf{X})$ or $b(Y)$ do not need to be specified explicitly or estimated. Note that, in terms of the relation between hypothetical population and source population, the assumption (2) can also be written as

$$p_1(Y \mid \mathbf{X}) \propto p(Y \mid \mathbf{X})b(Y).$$

The main idea is to utilize a conditional likelihood so that the estimation of $a(\mathbf{X})$ or $b(Y)$ is not needed. The assumption (2) also includes $\Pr(R = 1 \mid Y, \mathbf{X}) = a(\mathbf{X})$ or $= b(Y)$

as special cases, hence is more generally applicable than either $\Pr(R=1 \mid Y, \mathbf{X}) = a(\mathbf{X})$ or $= b(Y)$, as discussed in Section 1.

We first introduce a preliminary estimation method based on the penalized conditional likelihood idea. This method could tackle the informative subsampling issue, but could only detect strong signals.

2.1 | A preliminary estimation method

Based on definition of $p_1(Y \mid \mathbf{X})$ in Section 1 and under the assumption (2), we formulate the maximum likelihood estimation based on the observed data $\{y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$ as

$$\prod_{i=1}^{n} p_1(y_i \mid \mathbf{x}_i) = \prod_{i=1}^{n} \left\{ \frac{\Pr(r_i = 1 \mid y_i, \mathbf{x}_i)}{\Pr(r_i = 1 \mid \mathbf{x}_i)} p(y_i \mid \mathbf{x}_i; \theta) \right\} = \prod_{i=1}^{n} \left\{ \frac{a(\mathbf{x}_i) b(y_i)}{\Pr(r_i = 1 \mid \mathbf{x}_i)} p(y_i \mid \mathbf{x}_i; \theta) \right\}.$$

However, this is difficult to optimize due to the unknown quantities $a(\mathbf{x}_i)$, $b(y_i)$, and $\Pr(r_i = 1 \mid \mathbf{x}_i)$.

We decompose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as its rank statistic $\mathbf{D}$ and its corresponding order statistic $\mathbf{O}$, then

$$\prod_{i=1}^{n} p_1(y_i \mid \mathbf{x}_i) = p_1(\mathbf{D} = \mathbf{d} \mid \mathbf{O} = \mathbf{o}, \tilde{\mathbf{x}}) p_1(\mathbf{O} = \mathbf{o}, \tilde{\mathbf{x}}), \quad (3)$$

where $\tilde{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{d}$ and $\mathbf{o}$ denote the observed values of $\mathbf{D}$ and $\mathbf{O}$, respectively. Instead of considering the whole component $\prod_{i=1}^{n} p_1(y_i \mid \mathbf{x}_i)$ on the left-hand side of (3), we only consider the first term on the right-hand side of (3), which equals

$$p_1(\mathbf{D} = \mathbf{d} \mid \mathbf{O} = \mathbf{o}, \tilde{\mathbf{x}}) = \frac{\prod_{i=1}^{n} p_1(y_i \mid \mathbf{x}_i)}{\sum_c \prod_{i=1}^{n} p_1(y_{(i)} \mid \mathbf{x}_i)} = \frac{\prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i; \theta)}{\sum_c \prod_{i=1}^{n} p(y_{(i)} \mid \mathbf{x}_i; \theta)}, \quad (4)$$

where the summation in the denominator corresponds to all possible permutations c of $\{1, \ldots, n\}$. The equality in (4) is due to the fact that all $a(\mathbf{x}_i)$, $b(y_i)$, and $\Pr(r_i = 1 \mid \mathbf{x}_i)$ are canceled out from the numerator and the denominator. The intuition is that the first term on the right-hand side of (3), $p_1(\mathbf{D} = \mathbf{d} \mid \mathbf{O} = \mathbf{o}, \tilde{\mathbf{x}})$, has no information on $a(\mathbf{x}_i)$, $b(y_i)$, or $\Pr(r_i = 1 \mid \mathbf{x}_i)$.

We further refine (4) because of the computational burden as its denominator involves an infeasible computational cost of order $n!$. We consider a surrogate of $p_1(\mathbf{D} = \mathbf{d} \mid \mathbf{O} = \mathbf{o}, \tilde{\mathbf{x}})$ using its pairwise information studied first by Liang and Qin (2000). For any pair (i, j), let $\mathbf{D}_{ij}$ denote the pairwise rank statistic of $\mathbf{Y}_{ij} = (Y_i, Y_j)$, that is, $\mathbf{D}_{ij} = (1, 2)$ or $(2, 1)$. Instead of considering the full conditional likelihood in (4), we study the product of all possible combinations of the pairwise conditional likelihood

$$\prod_{1 \le i < j \le n} p_1(\mathbf{D}_{ij} = \mathbf{d}_{ij} \mid \mathbf{Y}_{ij} = \mathbf{y}_{ij}, \mathbf{x}_i, \mathbf{x}_j) = \prod_{1 \le i < j \le n} \frac{p(y_i \mid \mathbf{x}_i; \theta) p(y_j \mid \mathbf{x}_j; \theta)}{p(y_i \mid \mathbf{x}_i; \theta) p(y_j \mid \mathbf{x}_j; \theta) + p(y_i \mid \mathbf{x}_j; \theta) p(y_j \mid \mathbf{x}_i; \theta)}, \quad (5)$$

where $\mathbf{y}_{ij}$ and $\mathbf{d}_{ij}$ are the observed values of $\mathbf{Y}_{ij}$ and $\mathbf{D}_{ij}$, respectively.

Under a linear model (1), applying the logarithmic transformation to (5), we obtain the negative log surrogate likelihood function, to be minimized, is

$$\mathcal{L}(\boldsymbol{\gamma}) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \log \left[1 + \exp \left\{ -(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \boldsymbol{\gamma} \right\} \right], \quad (6)$$

where $\boldsymbol{\gamma} = \sigma^{-2} \boldsymbol{\beta}$. It can be seen that the pairwise surrogate conditional likelihood approach only requires to estimate $\boldsymbol{\gamma}$, a rescaled version of $\boldsymbol{\beta}$.

From estimating $\boldsymbol{\gamma}$, we can achieve the signal detection through regularization. To do so, we aim to minimize the following penalized negative log surrogate conditional likelihood discussed in Zhao *et al.* (2018):

$$\mathcal{L}(\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_\lambda(|\gamma_j|) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \log \left[1 + \exp \left\{ -(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \boldsymbol{\gamma} \right\} \right] + \sum_{j=1}^{p} p_\lambda(|\gamma_j|), \quad (7)$$

where $p_\lambda(t)$ could be a penalty function such as LASSO, SCAD or MCP, and $\lambda > 0$ is a tuning parameter. We denote this preliminary estimator by minimizing (7) as $\hat{\boldsymbol{\gamma}}^{\mathrm{ini}}$. If the LASSO penalty (Tibshirani, 1996) is used, $p_\lambda(t) = \lambda |t|$, and its derivative $p'_\lambda(t) = \lambda$ for $t > 0$. If the SCAD penalty (Fan and Li, 2001) is used, $p'_\lambda(t) = \lambda \mathbb{1}(t \le \lambda) + \frac{(a\lambda - t)_+}{a-1} \mathbb{1}(t > \lambda)$ for some $a > 2$ and $t > 0$, where $\mathbb{1}(\cdot)$ is the indicator function, and the standard choice of $a = 3.7$ in the SCAD penalty. If the MCP penalty (Zhang, 2010) is used, $p'_\lambda(t) = \frac{(a\lambda - t)_+}{a}$ for some $a > 0$ and $t > 0$, and a convention MCP penalty chooses $a = 3$. Throughout this paper, we also use

these conventional penalty values in our numerical comparisons.

Denote the support set $S = \{i : \gamma_i^* \neq 0\}$ and $s^* = |S|$, where $|A|$ is the cardinality of a generic set A. In the context of nonignorable nonresponse, Zhao *et al.* (2018) showed that, to achieve the selection consistency, one needs the condition $\log(n)(s^*)^2\sqrt{\frac{\log p}{n}} = o(1)$ and the minimum signal strength condition $\min_j |\gamma_j^*| > \lambda$ with $\lambda \asymp \sqrt{\frac{\log p}{n}}$, where $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. The minimum signal strength condition is critical for theoretical derivation, but it is difficult to be verified ina finite sample size, and there is no guarantee that it will be satisfied in practice. To the contrary, if this condition is not satisfied, or in other words, if weak signals exist, the method in Zhao *et al.* (2018) would fail in theory and the weak signal coefficient would shrink towards zero numerically. Recently, Shi and Qu (2017) and Li *et al.* (2019) provide methods for weak signal detection and estimation, but their methods are only applicable to noninformative subsampling settings.

Based on the preliminary estimator $\hat{\gamma}^{\text{ini}}$ and the idea of partial ridge regression, in the following we propose a new method to detect weak signals under an informative subsample setting.

2.2 | Proposed signal detection method

We now decompose the support set $S = S_1 \bigcup S_2$, where S_1 denotes covariates set with strong signals and S_2 denotes covariates set with weak signals. The covariates set with no signals is denoted as S^c. Based on the estimator $\hat{\gamma}^{\text{ini}}$, we have $\widehat{S}_1 = \{j : 1 \leq j \leq p, \hat{\gamma}_j^{\text{ini}} \neq 0\}$ and $\widehat{S}_1^c = \{j : 1 \leq j \leq p, \hat{\gamma}_j^{\text{ini}} = 0\}$. We now propose a two-step estimation method to achieve the proposed estimator $\hat{\gamma}^{\text{pro}}$. First, we obtain a partial ridge estimator $\widetilde{\gamma}$ by minimizing

$$\mathcal{L}(\gamma) + s\sum_{j\in\widehat{S}_1^c}\gamma_j^2 = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \log\left[1+\exp\left\{-(y_i-y_i)(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}\gamma\right\}\right] + s\sum_{j\in\widehat{S}_1^c}\gamma_j^2, \tag{8}$$

where $s > 0$ is a tuning parameter and the shrinkage only applies to the coefficients in $\widehat{S}_1^c$. That is, we do not impose any penalization for strong signals. The above idea is connected to the regularization after the retention approach proposed in Weng *et al.* (2019), while the difference is that their first retention step targets for large marginal-correlation coefficients with the response. Gao *et al.* (2017) also used this idea to propose a shrinkage estimation for high-dimensional linear models. In the second step, our proposed estimator $\hat{\gamma}^{\text{pro}}$ is obtained by

$$\hat{\gamma}_j^{\text{pro}} = \begin{cases} \widetilde{\gamma}_j, & \text{if } j \in \widehat{S}_1, \\ \widetilde{\gamma}_j, & \text{if } j \in \widehat{S}_1^c \text{ and } |\widetilde{\gamma}_j| > a, \\ 0, & \text{if } j \in \widehat{S}_1^c \text{ and } |\widetilde{\gamma}_j| \leq a, \end{cases} \tag{9}$$

where the truncation threshold $a > 0$ is also a tuning parameter. Then we form $\widehat{S}_2 = \{j : j \in \widehat{S}_1^c \text{ and } |\widetilde{\gamma}_j| > a\}$, hence $\widehat{S}^c = (\widehat{S}_1 \bigcup \widehat{S}_2)^c$. The thresholding step in (9) can also be understood as a weighted ridge regression with tuning $s/\mathbb{1}(|\widetilde{\gamma}_j| > a)$. Hence, the theoretical property of $\hat{\gamma}^{\text{pro}}$ heavily relies on output of variable selection and estimation for ridge regression. In the online supporting information, we outline the theoretical properties of $\hat{\gamma}^{\text{pro}}$. As for the tuning parameter selection, typically a data-driven method such as cross-validation is effective. In Section 3, we provide more details on how to apply the cross-validation to select the tuning parameters s and a simultaneously.

Through (8) and (9), the proposed estimator $\hat{\gamma}^{\text{pro}}$ is able to detect weak signals and provide an estimation to both strong and weak signals. In the following, we also investigate the statistical inference of the proposed estimation. Following the one-step estimation (Bickel, 1975) of $\hat{\gamma}^{\text{pro}}$, $\hat{\gamma}^{\text{pro1}}$, and the corresponding asymptotic theory, we propose a confidence interval for the strong and weak signal coefficient estimators.

2.3 | Statistical inference based on the one-step estimator

The essence of conducting statistical inference based on $\hat{\gamma}^{\text{pro}}$ is to find a valid variance estimator for $\hat{\gamma}^{\text{pro}}$. However, this is not a trivial task since the regularized estimator may not have a tractable limiting distribution even in the simplest linear regression setting (Knight and Fu, 2000).

Developing valid statistical inference with a regularized estimator has drawn a considerable amount of attention in recent years. For example, Lockhart *et al.* (2014), Tibshirani *et al.* (2016), and Lee *et al.* (2016) considered the conditional inference given that some covariates have been selected; Van de Geer *et al.* (2014) and Javanmard and Montanari (2014) proposed a debiased method for constructing confidence intervals. But their work is based on a linear model or generalized linear model with no informative subsample. Zhang and Zhang (2014) developed a projection approach to project a high-dimensional model to a low-dimensional problem and construct confidence intervals, but their method is not for the penalized estimator.

Here our primary aim is to develop a valid variance estimation for the estimator of $\boldsymbol{\gamma}$. Our proposal could be applied to $\boldsymbol{\gamma}$ with any fixed dimension, but for simplicity we only confine our presentation to the univariate case, that is, we decompose $\boldsymbol{\gamma} = (\gamma_1, \boldsymbol{\gamma}_{-1}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\gamma}_{-1}$ is a $(p-1)$-dimensional vector containing all elements in $\boldsymbol{\gamma}$ except γ_1. We first use the decorrelated score function studied in Ning and Liu (2017), $S_D(\gamma_1, \boldsymbol{\gamma}_{-1}) = \nabla_1 \mathcal{L}(\gamma_1, \boldsymbol{\gamma}_{-1}) - \mathbf{w}^{\mathrm{T}} \nabla_{-1} \mathcal{L}(\gamma_1, \boldsymbol{\gamma}_{-1}) = \mathbf{v}^{\mathrm{T}} \nabla \mathcal{L}(\gamma_1, \boldsymbol{\gamma}_{-1})$, where $\mathbf{I} = E_{\gamma}\{\nabla^2 \mathcal{L}(\gamma_1, \boldsymbol{\gamma}_{-1})\}$, $\mathbf{w}^{\mathrm{T}} = \mathbf{I}_{1,-1}\mathbf{I}_{-1,-1}^{-1}$ is a $(p-1)$-dimensional vector, $\mathbf{v} = (1, -\mathbf{w}^{\mathrm{T}})^{\mathrm{T}}$, where the sub-scripts 1 and -1 represent the first coordinate and all coordinates except the first one of the corresponding vector, and the subscripts $(1,-1)$ and $(-1,-1)$ are defined similarly as for the corresponding matrix. This is due to the fact that S_D is uncorrelated with the nuisance score function $\nabla_{-1}\mathcal{L}(\gamma_1, \boldsymbol{\gamma}_{-1})$ under the conventional negative log-likelihood function.

To proceed with $S_D(\gamma_1, \boldsymbol{\gamma}_{-1})$, we fist need to estimate $\boldsymbol{\gamma}$ and $\mathbf{w}$. We use $\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}$ obtained from Section 2.2 as the estimator of $\boldsymbol{\gamma}$. For estimating $\mathbf{w}$, following Ning and Liu (2017), we adopt a computationally efficient way of using the Dantzig-type estimator as the following:

$$\widehat{\mathbf{w}} = \arg\min \|\mathbf{w}\|_1 \text{ s.t. } \|\nabla^2_{1,-1}\mathcal{L}(\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}) - \mathbf{w}^{\mathrm{T}}\nabla^2_{-1,-1}\mathcal{L}(\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}})\|_{\infty} \le \lambda', \tag{10}$$

where the sparsity level of $\mathbf{w}^*$, the true value of $\mathbf{w}$, is defined as $s^*_{\mathbf{w}} = \|\mathbf{w}^*\|_0$. Here for a generic d-dimensional vector $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}}$, we define $\|\mathbf{v}\|_1 = \sum_{i=1}^{d} |v_i|$, $\|\mathbf{v}\|_{\infty} = \max_{1\le i\le d} |v_i|$. We also define $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$ with $\mathrm{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$. Once $\widehat{\mathbf{w}}$ is obtained, we can form the score function for γ_1 such as $S_D(\gamma_1) = S_D(\gamma_1, \widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}_{-1}) = \nabla_1 \mathcal{L}(\gamma_1, \widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}_{-1}) - \widehat{\mathbf{w}}^{\mathrm{T}}\nabla_{-1}\mathcal{L}(\gamma_1, \widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}_{-1}) = \widehat{\mathbf{v}}^{\mathrm{T}}\nabla\mathcal{L}(\gamma_1, \widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}_{-1})$.

The above score function provides us an estimator of γ_1 by solving $S_D(\gamma_1) = 0$, but it may suffer from the multiple roots issue. Instead we consider a one-step estimator (Bickel, 1975) of γ_1, $\widehat{\gamma}_1^{\mathrm{pro1}}$, which solves $S_D(\widehat{\gamma}_1^{\mathrm{pro}}) + \widehat{I}_{1|-1}(\gamma_1 - \widehat{\gamma}_1^{\mathrm{pro}}) = 0$, i.e., $\widehat{\gamma}_1^{\mathrm{pro1}} = \widehat{\gamma}_1^{\mathrm{pro}} - S_D(\widehat{\gamma}_1^{\mathrm{pro}})/\widehat{I}_{1|-1}$, where $\widehat{I}_{1|-1} = \nabla^2_{1,1}\mathcal{L}(\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}) - \widehat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{-1,1}\mathcal{L}(\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}})$. Under the conditions verified in the online supporting information and the condition that $n^{-1/2}\max(s^*, s^*_{\mathbf{w}})\log p = o(1)$, following Theorem 3.2 in Ning and Liu (2017), we have

$$\sqrt{n}(\widehat{\gamma}_1^{\mathrm{pro1}} - \gamma_1^*)\widehat{I}_{1|-1}/\sqrt{\sigma_s^*} \xrightarrow{d} N(0,1), \tag{11}$$

where $\sigma_s^* = (\mathbf{v}^*)^{\mathrm{T}}\boldsymbol{\Sigma}^*\mathbf{v}^*$, $\boldsymbol{\Sigma}^* = \lim_{n\to\infty} \mathrm{Var}_{\boldsymbol{\gamma}^*}(n^{1/2}\nabla\mathcal{L}(\boldsymbol{\gamma}^*)) = 4E\{\frac{\partial \xi_{12}}{\partial \boldsymbol{\gamma}}(\boldsymbol{\gamma}^*)\frac{\partial \xi_{13}}{\partial \boldsymbol{\gamma}}(\boldsymbol{\gamma}^*)^{\mathrm{T}}\}$ and $\xi_{ij}(\boldsymbol{\gamma}) = \log[1 + \exp\{-(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\boldsymbol{\gamma}\}]$. Immediately, we can construct a $(1-\alpha)\times 100\%$ confidence interval for γ_1^* as

$$[\widehat{\gamma}_1^{\mathrm{pro1}} - n^{-1/2}\Phi^{-1}(1-\alpha/2)\widehat{I}^{-1}_{1|-1}\widehat{\sigma}_s^{1/2}, \widehat{\gamma}_1^{\mathrm{pro1}} + n^{-1/2}\Phi^{-1}(1-\alpha/2)\widehat{I}^{-1}_{1|-1}\widehat{\sigma}_s^{1/2}], \tag{12}$$

where $\widehat{\sigma}_s$ is a consistent estimator for σ_s^* with more details provided in Section 3.

Based on the asymptotically normal distribution of $\widehat{\gamma}_1^{\mathrm{pro1}}$ in (11), we can conduct conventional statistical inference for γ_1^*. In addition, we construct a Wald-type test statistic $U_n = n^{1/2}\widehat{\gamma}_1^{\mathrm{pro1}}\widehat{I}_{1|-1}/\sqrt{\widehat{\sigma}_s}$ for the hypothesis testing

$$H_0 : \gamma_1^* = 0 \text{ versus } H_1 : \gamma_1^* \neq 0. \tag{13}$$

That is, with a significance level α, based on the test statistic U_n, if $|U_n| > \Phi^{-1}(1-\alpha/2)$, where Φ is the cumulative distribution function of the standard normal distribution, then the null hypothesis H_0 in (13) is rejected.

Note that since γ_1^* is a rescaled β_1^*, the hypothesis testing (13) is equivalent to testing

$$H_0 : \beta_1^* = 0 \text{ versus } H_1 : \beta_1^* \neq 0, \tag{14}$$

where we decompose $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{-1}$ is defined similar to $\boldsymbol{\gamma}_{-1}$. Therefore, we can perform testing the original parameter $\boldsymbol{\beta}$ through the knowledge of a derived parameter $\boldsymbol{\gamma}$, in the presence of informative subsample.

3 | ALGORITHM

In this section, we summarize the computing algorithms for the proposed estimation and inference procedure. We also discuss the computational issues including the selection of tuning parameters.

Algorithm 1 Calculate $\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}$

Require: Conditional likelihood $\mathcal{L}(\boldsymbol{\gamma})$, penalty $p_\lambda(\cdot)$, tuning parameters λ, s and a.

(i) Calculate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ini}}$ by minimizing the objective function $\mathcal{L}(\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_\lambda(|\gamma_j|)$ in (7).

(ii) Determine $\widehat{S}_1 = \{j : 1 \le j \le p, \widehat{\gamma}_j^{\mathrm{ini}} \neq 0\}$ and $\widehat{S}_1^c = \{j : 1 \le j \le p, \widehat{\gamma}_j^{\mathrm{ini}} = 0\}$.

(iii) Estimate $\widetilde{\boldsymbol{\gamma}}$ by minimizing $\mathcal{L}(\boldsymbol{\gamma}) + s\sum_{j\in\widehat{S}_1^c}\gamma_j^2$ in (8).

(iv) Calculate $\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}$ by (9), and determine $\widehat{S}_2 = \{j : j \in \widehat{S}_1^c \text{ and } |\widetilde{\gamma}_j| > a\}$.

Return: $\widehat{S}_1$, $\widehat{S}_2$, $\widehat{\boldsymbol{\gamma}}^{\mathrm{ini}}$ and $\widehat{\boldsymbol{\gamma}}^{\mathrm{pro}}$.

Remark 1. The optimization problem for (7) in step (i) is not trivial. If we denote $m = n(n-1)/2$ as the number of terms in the summation for pairwise y_i and y_j, $1 \le i < j \le n$ such that $y_i - y_j \neq 0$, $\mathbf{v}_k = (\mathbf{x}_i - \mathbf{x}_j)|y_i - y_j|$ and $u_k = \mathbb{1}(y_i - y_j > 0)$, then following (6) we have $\mathcal{L}(\boldsymbol{\gamma}) = -\frac{1}{m}\sum_{k=1}^{m}[u_k \log \frac{\exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\gamma})}{1+\exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\gamma})} + (1-u_k)\log\{1 - \frac{\exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\gamma})}{1+\exp(\mathbf{v}_k^{\mathrm{T}}\boldsymbol{\gamma})}\}]$, which has exactly the same form as the negative log-likelihood of a logistic regression without the intercept term assuming the data $\{u_k, \mathbf{v}_k\}_{k=1}^{m}$ artificially independent and identically distributed. Hence, the optimization problem in step (i) can be solved by the logistic regression with a penalty function. Therefore, we can implement R package TVsMiss", where the Bayesian information criterion is used to select the tuning parameter λ.

Remark 2. Similarly, the optimization of (8) in step (iii) can be directly executed by R package glmnet". We use $\mathcal{K}$-fold cross-validation to select the tuning parameters (s, a). Specifically, we randomly split the data $\{y_i, \mathbf{x}_i\}_{i=1}^{n}$ into $\mathcal{K}$ equal-sized samples $\mathcal{S}_1, \ldots, \mathcal{S}_{\mathcal{K}}$. For any given (s, a), we obtain $\hat{\boldsymbol{\gamma}}_{(-k)}^{\mathrm{pro}}$ from (8) and (9) based on data not in $\mathcal{S}_l$, $l = 1, \ldots, \mathcal{K}$. The cross-validation criterion is defined as

$$\mathrm{CV}(s,a) = \frac{1}{\mathcal{K}}\sum_{k=1}^{\mathcal{K}} \frac{2}{|\mathcal{S}_k|(|\mathcal{S}_k|-1)} \sum_{i<j, i\in\mathcal{S}_k, j\in\mathcal{S}_k} \times \log\left[1 + \exp\{-(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_{(-k)}^{\mathrm{pro}}\}\right].$$

We conduct a two-dimensional grid search to find the optimal (s, a) which minimizes $\mathrm{CV}(s, a)$.

Algorithm 2 Calculate confidence interval of γ_1^* and test statistic U_n for (14)

Require: Conditional likelihood $\mathcal{L}(\boldsymbol{\gamma})$, estimator $\hat{\boldsymbol{\gamma}}^{\mathrm{pro}}$, parameter λ', confidence level $1-\alpha$.

(i) Estimate $\mathbf{w}$ by the Dantzig type estimator $\hat{\mathbf{w}}$ via (10).

(ii) Obtain $S_D(\gamma_1) = \nabla_1\mathcal{L}(\gamma_1, \hat{\boldsymbol{\gamma}}_{-1}^{\mathrm{pro}}) - \hat{\mathbf{w}}^{\mathrm{T}}\nabla_{-1}\mathcal{L}(\gamma_1, \hat{\boldsymbol{\gamma}}_{-1}^{\mathrm{pro}})$ and calculate $S_D(\hat{\gamma}_1^{\mathrm{pro}})$.

(iii) Calculate $\hat{I}_{1|-1}$, $\hat{\gamma}_1^{\mathrm{pro1}}$, and estimate σ_s^* by $\hat{\sigma}_s = (1, -\hat{\mathbf{w}}^{\mathrm{T}})\hat{\boldsymbol{\Sigma}}\begin{pmatrix}1\\ -\hat{\mathbf{w}}\end{pmatrix}$.

(iv) Calculate the confidence interval as in (12).

(v) Calculate $U_n = n^{1/2}\hat{\gamma}_1^{\mathrm{pro1}}\hat{I}_{1|-1}/\sqrt{\hat{\sigma}_s}$, $T_n = 1$ if $|U_n| > \Phi^{-1}(1-\alpha/2)$ and $T_n = 0$ otherwise.

Return: The confidence interval in (iv) and T_n in (v).

Remark 3. The l_1 minimization problem in step (i) is convex and can be solved by function "fastlp" in R package fastclime". More details of Dantzig selector optimization can be found in Candes and Tao (2007) and Cai and Liu (2011). The tuning parameter λ' in (10) is selected by cross-validation as in Ning and Liu (2017).

Remark 4. Details of the variance estimator $\hat{\boldsymbol{\Sigma}}$ in step (iii) can be found in Zhao and Shao (2017).

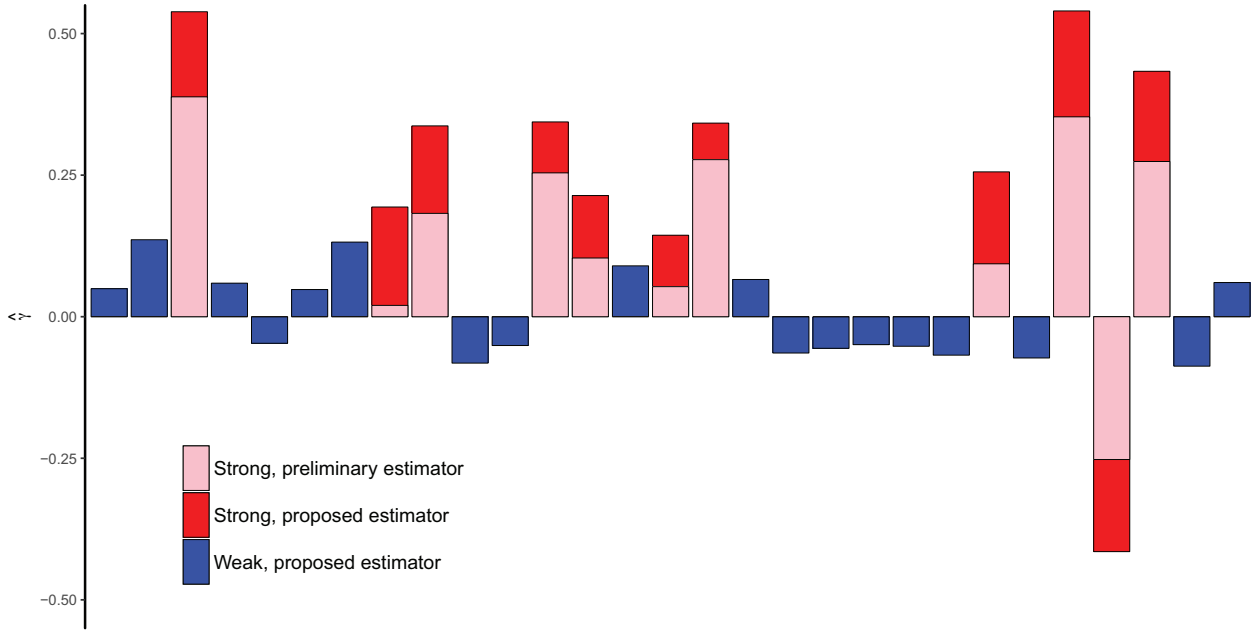
4 | APPLICATION TO THE HIV-1 DRUG RESISTANCE STUDY

In this section, we provide an application of the HIV-1 drug resistance study using the proposed method. In controlling the HIV-1 progression and prolonging survival, the use of antiretroviral drugs could be less effective by the development of drug resistance (Clavel and Hance, 2004). Knowledge regarding the drug resistance is critical for the development of antiretroviral drugs and management of infections with drug-resistant viruses. The Stanford HIV-1 drug resistance study (Rhee *et al.*, 2003; Shafer, 2006) investigated associating genetic variation on the antiretroviral treatments from patients with observed variant, drug susceptibility data, and virological and clinical responses to a new treatment regimen. The details of HIV-1 sequences, isolates, and genotypes are provided in Rhee *et al.* (2006).

The drugs currently used to treat HIV-1 infection consist of a variety of distinct classes: nucleoside and nucleotide analogues, nonnucleoside reverse-transcriptase inhibitors, protease inhibitors, and entry inhibitors. Some combinations of these drugs are also used for treating HIV-1 infection, the so-called highly active antiretroviral therapy, but there is no evidence that any combinations of currently available drugs are strongly synergistic in vitro. Rhee *et al.* (2006) analyzed 16 drugs, and for each drug they investigated the association between HIV-1 protease and reverse transcriptase mutations with its susceptibility.

We analyze the association of protease mutations with susceptibility to amprenavir, a protease inhibitor drug, studied in Rhee *et al.* (2006). This dataset has $N = 848$ HIV-1 infected patients whose drug resistance was assayed and mutation information at 99 protease codons in the viral genome determined by the genotyping technology. Out of 99 protease codons, 79 contain mutations. In particular, the outcomes of drug susceptibility were expressed in units of IC_{50}, the amount of drug inhibits viral replication by 50% increase in units of fold compared with standard drug-sensitive wild-type virus. Higher IC_{50} corresponds to higher levels of drug resistance. Identifying codon mutations which are truly associated with amprenavir resistance is crucial for understanding the biological mechanisms of drug resistance.

Abraham *et al.* (2011) pointed out that the study of HIV-1 drug resistance is heavily influenced by the selected samples in that participants with a relatively higher viral load



Codon	P2	P6	P10	P15	P17	P24	P30	P32	P33	P37	P45	P46	P47	P48	P50	P54	P63	P64	P65	P66	P70	P71	P76	P83	P84	P88	P90	P93	P98
MTC11			*				*	*	*	*		*	*	*	*	*		*				*	*		*	*	*	*	
Wetal17			*					*				*	*		*	*							*		*		*		
SQ17			s	w	w	w	s	s	s	w	w	s	s	s	s	s	w	w		w		w	s	w	s	s	s	w	
proposed	w	w	s	w	w	w	w	s	s	w	w	s	s	w	s	s	w	w	w	w	w	w	s	w	s	s	s	w	w

FIGURE 1 The comparison of signal detection results with the literature. MTC11 is the finding from Minnier *et al.* (2011), Wetal17 is from Wensing *et al.* (2017), and SQ17 is from Shi and Qu (2017). The notation "s" represents strong signal, "w" weak signal, and "*" simply means signal without knowing strong or weak. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

(response variable Y) are selected and their measurements are collected in the study, but how the selection process depends on Y is largely unknown. Hence, the assumption $\Pr(R = 1 \mid Y, \mathbf{X}) = b(Y)$ is very plausible in practice. This flexible assumption implies that the sample selection depends on Y, the viral load, via an arbitrary form. This could be a case that participants with a relatively higher viral load (than a certain threshold) are selected. In practice, $b(Y)$ could be a complicated function. On the other hand, the assumption $\Pr(R = 1 \mid Y, \mathbf{X}) = a(\mathbf{X})$ is a trivial case of (2) in the sense that the source population is the same as the hypothetical population, that is, $p_1(Y \mid \mathbf{X}) = p(Y \mid \mathbf{X})$, and there is no informative subsampling issue. Some previous analyses on this HIV-1 drug resistance study (Wu, 2009; Minnier *et al.*, 2011; Shi and Qu, 2017) overlooked the informative subsample issue, where this trivial assumption is implicitly used. Therefore, we adopt the more generally applicable assumption (2), where the random samples with $N = 848$ represent the hypothetical population of our interest, while the random samples with $n = 702$ represent the source population which is different from the hypothetical population. We illustrate the proposed methodology for this application as follows.

Similar to the existing literature, we fit a linear model where the response variable Y is the log-transformed IC_{50}, and each dichotomous predictor contains the mutation information (0 = no mutation/1 = mutation) at each of the relevant codons. For simplicity, we only report the results using the MCP penalty.

The preliminary estimator $\hat{\gamma}^{\mathrm{ini}}$ detected 11 signals, and all of which were identified as strong by our proposed estimator $\hat{\gamma}^{\mathrm{pro}}$. In addition, $\hat{\gamma}^{\mathrm{pro}}$ also identified 18 weak signals. In the literature, Shi and Qu (2017), hereafter referred to as SQ17", identified 28 signals including 13 strong ones and 15 weak ones. The 13 strong signals detected by SQ17" are also selected by our method, where 11 of them (P10, P32, P33, P46, P47, P50, P54, P76, P84, P88, P90) were identified as strong signals and two of them (P30, P48) were identified as weak ones. Our method provides a smaller set of strong signals compared to existing ones, and the strong signals replicated by our method are very plausible to be strong ones in practice. This is confirmed by the most

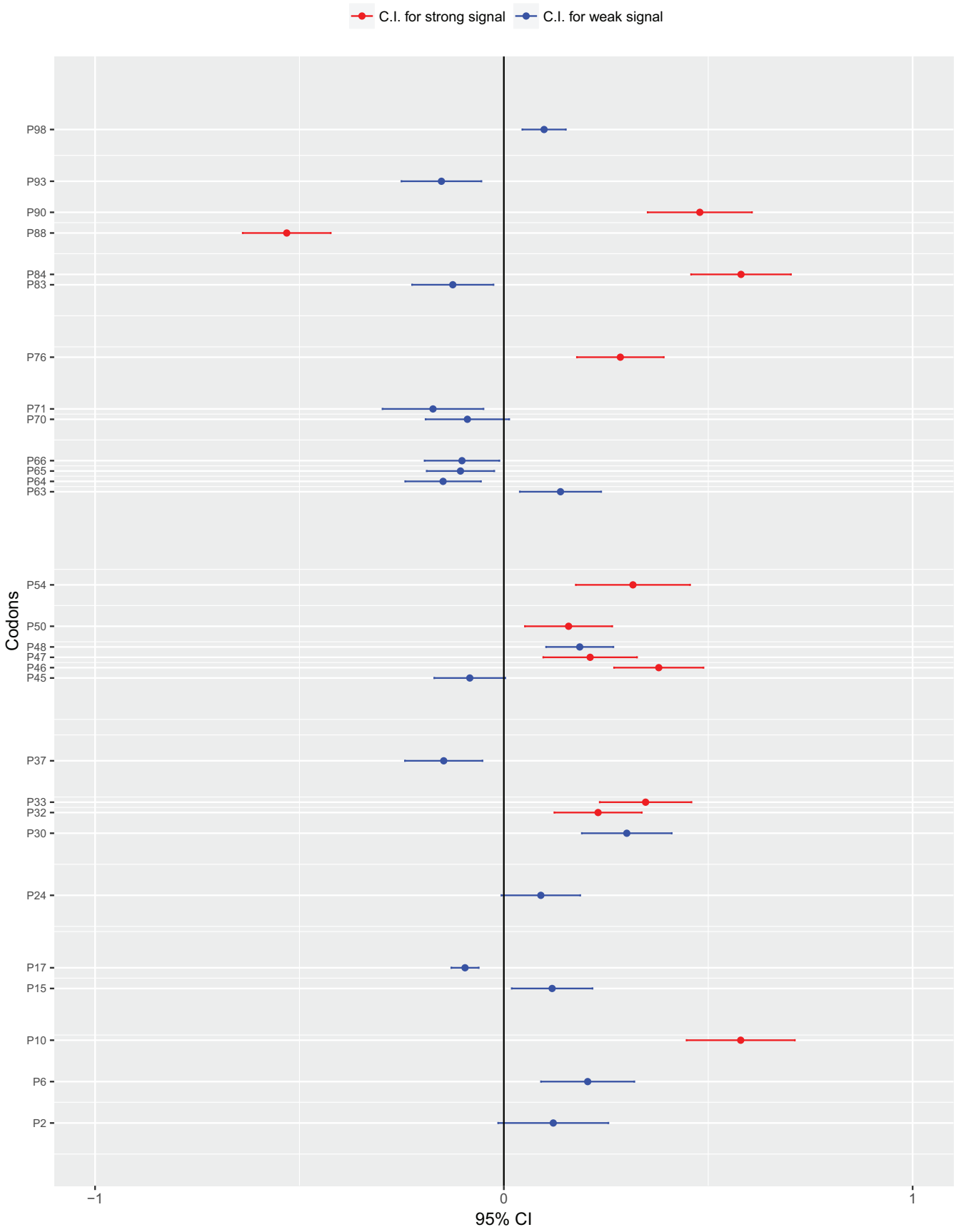


FIGURE 2 The one-step estimator $\hat{\gamma}^{\text{prol}}$ and its 95% confidence interval for each of the detected strong and weak signals. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 1 The comparison of the mean, first quartile (Q1), median (Q2), 3rd quartile (Q3) of the FP, FN, and F-measure for signal detection with simulated datasets in Case (C1)

		Setting 1: $p = 10$											
		FP				**FN**				**F**			
N	**Method**	**Mean**	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**
200	$\hat{\gamma}^{\text{nv}}$	1.18	0	1	2	1.15	0	1	2	0.80	0.71	0.80	0.91
	$\hat{\gamma}^{\text{ini}}$	0.01	0	0	0	2.84	3	3	3	0.69	0.67	0.67	0.67
	$\hat{\gamma}^{\text{pro}}$	1.36	0	1	3	1.05	0	1	2	0.80	0.75	0.80	0.86
500	$\hat{\gamma}^{\text{nv}}$	1.55	0	1	3	0.14	0	0	0	0.88	0.80	0.91	0.92
	$\hat{\gamma}^{\text{ini}}$	0	0	0	0	2.85	3	3	3	0.69	0.67	0.67	0.67
	$\hat{\gamma}^{\text{pro}}$	1.15	0	1	2	0.37	0	0	1	0.89	0.80	0.91	1.00
1000	$\hat{\gamma}^{\text{nv}}$	1.22	0	1	2	0.01	0	0	0	0.92	0.86	0.92	1.00
	$\hat{\gamma}^{\text{ini}}$	0.00	0	0	0	2.22	1	3	3	0.76	0.67	0.67	0.91
	$\hat{\gamma}^{\text{pro}}$	0.76	0	0	1	0.06	0	0	0	0.94	0.92	1.00	1.00
		Setting 2: $p = 50$											
		FP				**FN**				**F**			
N	**Method**	**Mean**	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**	**Mean**	**Q1**	**Q2**	**Q3**
200	$\hat{\gamma}^{\text{nv}}$	1.36	0	0	2	2.06	1	2	3	0.70	0.67	0.67	0.77
	$\hat{\gamma}^{\text{ini}}$	0.04	0	0	0	2.85	3	3	3	0.68	0.67	0.67	0.67
	$\hat{\gamma}^{\text{pro}}$	8.62	0	3	12	1.60	1	2	3	0.57	0.42	0.63	0.67
500	$\hat{\gamma}^{\text{nv}}$	2.61	0	2	4	0.58	0	0	1	0.79	0.71	0.80	0.91
	$\hat{\gamma}^{\text{ini}}$	0	0	0	0	2.91	3	3	3	0.68	0.67	0.67	0.67
	$\hat{\gamma}^{\text{pro}}$	4.91	0	1	5	0.91	0	1	2	0.73	0.63	0.77	0.86
1000	$\hat{\gamma}^{\text{nv}}$	2.84	1	2	4	0.04	0	0	0	0.83	0.75	0.86	0.92
	$\hat{\gamma}^{\text{ini}}$	0.00	0	0	0	2.46	2	3	3	0.73	0.67	0.67	0.80
	$\hat{\gamma}^{\text{pro}}$	2.79	0	0	2	0.25	0	0	0	0.86	0.80	0.91	1.00

recent update of HIV-1 drug resistance report in Wensing *et al.* (2017), hereafter referred to as Wetal17". Our method also contains a larger pool of possible weak signals which might be mildly associated with drug resistance. In many studies, it is important not to miss any statistically weak signals, since they could be clinically valuable predictors. Our results share many common findings with Minnier *et al.* (2011), hereafter referred to as MTC11", which identified 18 signals but did not distinguish strong and weak signals. Also, compared to $\hat{\gamma}^{\text{ini}}$, the proposed method $\hat{\gamma}^{\text{pro}}$ provides each of the strong signal estimations with larger magnitudes. Apparently eliminating the weak mutations could lead to an attenuation of the drug resistance effect to the strong mutations. Figure 1 summarizes the signal detection results and the comparison results among $\hat{\gamma}^{\text{ini}}$, $\hat{\gamma}^{\text{pro}}$, and other existing estimators.

Additionally the confidence interval for each of the detected signals based on (12) using the one-step estimation $\hat{\gamma}^{\text{pro1}}$ is plotted in Figure 2. We observe that the one-step estimates for the strong signals in general still have larger magnitudes than that of weak signals, and none of their confidence intervals contain zero. On the other hand, the one-step estimates for the weak signals are much closer to zero in that some of their confidence intervals include zero. The interval length for weak signal is not necessarily wider than that of strong signal, indicating that the proposed method produces reasonably weak signal inference compared to strong signal inference, both of which are credible for clinical guidance purpose.

5 | SIMULATION STUDIES

In this section, we conduct simulation studies to illustrate the finite sample performance of our proposed methodology as well as its comparison with some existing methods. Specifically we consider three cases as detailed below.

First, we generate the data following the assumption (2), and we mainly compare $\hat{\gamma}^{\text{pro}}$ with the estimator $\hat{\gamma}^{\text{ini}}$, which does not detect weak signals, and the estimator $\hat{\gamma}^{\text{nv}}$, which does not distinguish the hypothetical population and the source population. We also compare $\hat{\gamma}^{\text{pro1}}$ with the estimator $\hat{\gamma}^{\text{ini1}}$ for the variance estimation and hypothesis testing. In the second case, we generate the data following the assumption that $p_1(Y \mid \mathbf{X}) = p(Y \mid \mathbf{X})$ such that the estimator $\hat{\gamma}^{\text{nv}}$ should work well. The main purpose is

TABLE 2 The comparison of the parameter estimate, standard deviation, and root mean squared error (RMSE) with simulated datasets in Case (C1)

		Estimate			Standard deviation			RMSE		
Strong signals		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
Setting 1: $p = 10$										
$N = 200$	$\hat{\gamma}^{\text{nv}}$	0.931	0.928	0.931	0.051	0.049	0.052	0.086	0.087	0.086
	$\hat{\gamma}^{\text{ini}}$	0.940	0.941	0.944	0.162	0.157	0.161	0.173	0.168	0.170
	$\hat{\gamma}^{\text{pro}}$	1.021	1.020	1.025	0.154	0.150	0.153	0.155	0.151	0.155
$N = 500$	$\hat{\gamma}^{\text{nv}}$	0.933	0.935	0.932	0.030	0.030	0.030	0.073	0.072	0.074
	$\hat{\gamma}^{\text{ini}}$	0.927	0.930	0.929	0.090	0.090	0.091	0.116	0.114	0.115
	$\hat{\gamma}^{\text{pro}}$	0.997	1.000	0.999	0.095	0.097	0.097	0.096	0.097	0.097
$N = 1000$	$\hat{\gamma}^{\text{nv}}$	0.935	0.935	0.934	0.021	0.021	0.021	0.068	0.068	0.069
	$\hat{\gamma}^{\text{ini}}$	0.944	0.944	0.944	0.067	0.067	0.069	0.087	0.087	0.089
	$\hat{\gamma}^{\text{pro}}$	1.005	1.005	1.004	0.064	0.064	0.066	0.064	0.064	0.066
Setting 2: $p = 50$										
$N = 200$	$\hat{\gamma}^{\text{nv}}$	0.932	0.930	0.930	0.050	0.049	0.051	0.085	0.085	0.086
	$\hat{\gamma}^{\text{ini}}$	0.917	0.921	0.918	0.171	0.171	0.170	0.190	0.189	0.189
	$\hat{\gamma}^{\text{pro}}$	1.030	1.032	1.030	0.160	0.163	0.161	0.163	0.166	0.164
$N = 500$	$\hat{\gamma}^{\text{nv}}$	0.932	0.931	0.932	0.030	0.031	0.031	0.074	0.075	0.075
	$\hat{\gamma}^{\text{ini}}$	0.931	0.931	0.932	0.086	0.085	0.089	0.111	0.110	0.112
	$\hat{\gamma}^{\text{pro}}$	1.015	1.015	1.015	0.093	0.093	0.095	0.094	0.094	0.096
$N = 1000$	$\hat{\gamma}^{\text{nv}}$	0.934	0.932	0.932	0.021	0.021	0.021	0.069	0.071	0.071
	$\hat{\gamma}^{\text{ini}}$	0.937	0.937	0.937	0.065	0.063	0.063	0.090	0.090	0.089
	$\hat{\gamma}^{\text{pro}}$	1.020	1.019	1.020	0.070	0.070	0.069	0.072	0.073	0.072
		Estimate			**Standard deviation**			**RMSE**		
Weak signals		γ_4	γ_5	γ_6	γ_4	γ_5	γ_6	γ_4	γ_5	γ_6
Setting 1: $p = 10$										
$N = 200$	$\hat{\gamma}^{\text{nv}}$	0.060	0.060	0.057	0.059	0.059	0.059	0.071	0.072	0.073
	$\hat{\gamma}^{\text{ini}}$	0.008	0.008	0.007	0.034	0.036	0.033	0.098	0.099	0.099
	$\hat{\gamma}^{\text{pro}}$	0.057	0.058	0.053	0.060	0.062	0.059	0.074	0.075	0.075
$N = 500$	$\hat{\gamma}^{\text{nv}}$	0.086	0.084	0.084	0.035	0.036	0.036	0.038	0.039	0.039
	$\hat{\gamma}^{\text{ini}}$	0.007	0.006	0.006	0.029	0.028	0.030	0.098	0.098	0.098
	$\hat{\gamma}^{\text{pro}}$	0.079	0.077	0.077	0.042	0.044	0.045	0.047	0.049	0.050
$N = 1000$	$\hat{\gamma}^{\text{nv}}$	0.089	0.090	0.090	0.021	0.021	0.022	0.024	0.024	0.024
	$\hat{\gamma}^{\text{ini}}$	0.027	0.027	0.029	0.048	0.047	0.050	0.088	0.087	0.087
	$\hat{\gamma}^{\text{pro}}$	0.093	0.093	0.094	0.029	0.029	0.030	0.029	0.030	0.031
Setting 2: $p = 50$										
$N = 200$	$\hat{\gamma}^{\text{nv}}$	0.022	0.021	0.021	0.043	0.040	0.039	0.089	0.089	0.088
	$\hat{\gamma}^{\text{ini}}$	0.006	0.005	0.005	0.030	0.025	0.025	0.098	0.098	0.099
	$\hat{\gamma}^{\text{pro}}$	0.031	0.029	0.029	0.055	0.049	0.050	0.088	0.087	0.087
$N = 500$	$\hat{\gamma}^{\text{nv}}$	0.054	0.053	0.054	0.042	0.042	0.041	0.062	0.063	0.062
	$\hat{\gamma}^{\text{ini}}$	0.004	0.003	0.004	0.023	0.018	0.020	0.098	0.099	0.099
	$\hat{\gamma}^{\text{pro}}$	0.049	0.048	0.050	0.045	0.042	0.042	0.068	0.067	0.065
$N = 1000$	$\hat{\gamma}^{\text{nv}}$	0.079	0.077	0.077	0.029	0.029	0.028	0.036	0.037	0.036
	$\hat{\gamma}^{\text{ini}}$	0.018	0.016	0.016	0.038	0.037	0.037	0.091	0.092	0.092
	$\hat{\gamma}^{\text{pro}}$	0.078	0.076	0.076	0.039	0.038	0.037	0.045	0.045	0.044

TABLE 3 The comparison between one-step estimators $\hat{\gamma}^{\text{ini1}}$ and $\hat{\gamma}^{\text{pro1}}$, the standard deviation, the standard error, and the 95% coverage probability for the strong signals with simulated datasets in Case (C1)

Setting 1: $p = 10$							
		One-step estimate			**Standard deviation**		
		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.959	0.958	0.963	0.147	0.143	0.146
	$\hat{\gamma}^{\text{pro1}}$	1.056	1.055	1.059	0.165	0.160	0.163
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.931	0.934	0.932	0.089	0.090	0.091
	$\hat{\gamma}^{\text{pro1}}$	1.020	1.023	1.021	0.097	0.100	0.100
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.952	0.952	0.951	0.070	0.070	0.071
	$\hat{\gamma}^{\text{pro1}}$	1.013	1.013	1.013	0.064	0.065	0.067
		Standard error			**Coverage probability**		
		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.130	0.130	0.130	87.0	86.5	87.3
	$\hat{\gamma}^{\text{pro1}}$	0.137	0.136	0.136	89.1	90.4	89.1
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.084	0.084	0.084	82.3	82.6	82.7
	$\hat{\gamma}^{\text{pro1}}$	0.086	0.086	0.086	91.4	91.1	91.3
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.060	0.060	0.060	80.7	83.4	80.2
	$\hat{\gamma}^{\text{pro1}}$	0.064	0.064	0.064	94.8	94.5	93.7
Setting 2: $p = 50$							
		One-step estimate			**Standard deviation**		
		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.919	0.919	0.918	0.150	0.150	0.150
	$\hat{\gamma}^{\text{pro1}}$	1.058	1.057	1.057	0.193	0.197	0.195
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.920	0.920	0.921	0.086	0.085	0.088
	$\hat{\gamma}^{\text{pro1}}$	1.020	1.020	1.020	0.104	0.106	0.106
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.938	0.937	0.938	0.070	0.069	0.069
	$\hat{\gamma}^{\text{pro1}}$	1.013	1.012	1.013	0.069	0.070	0.069
		Standard error			**Coverage probability**		
		γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.131	0.131	0.131	81.1	81.8	81.4
	$\hat{\gamma}^{\text{pro1}}$	0.139	0.138	0.139	82.6	82.0	82.5
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.085	0.085	0.085	83.1	82.6	80.6
	$\hat{\gamma}^{\text{pro1}}$	0.089	0.089	0.089	90.3	89.4	89.7
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.060	0.060	0.060	75.8	76.7	76.8
	$\hat{\gamma}^{\text{pro1}}$	0.065	0.065	0.065	92.8	92.5	92.8

to assess the efficiency loss of the proposed estimator $\hat{\gamma}^{\text{pro}}$ if the hypothetical population happens to be the same as the source population. In the third case, the data generation violates the assumption (2). This can be regarded as a sensitivity analysis to evaluate the proposed methodology under misspecification of the assumption (2).

For our data generation mechanism for all three cases, each of the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is distributed as $x_{ik} = \xi_{ik,1}^2 + \xi_{ik,2}$, $i = 1, \ldots, N$, $k = 1, \ldots, p$, where $\xi_{ik,1}$ and $\xi_{ik,2}$ are independent samples of the standard normal distribution. We generate response $y_i = \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and the regression coefficient $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} = (1, 1, 1, 0.1, 0.1, 0.1, 0, \ldots, 0)^{\mathrm{T}}$. We let $N =$ 200, 500, 1000 and $p = 10, 50$. In each of the three cases, we generate the sample selection mechanism $\Pr(R = 1 \mid Y, \mathbf{X})$ such that around 75% to 80% samples are selected. Note that, although the knowledge of N is not required in our algorithm implementation as well as in application, we adopt to gradually increase N, the same mechanism in gradually increasing the sample size, to assess the performance of our proposed methodology.

For signal detection, we evaluate the performance by the F-measure defined as the harmonic mean of precision and sensitivity, $F = \frac{2\#TP}{2\#TP+\#FP+\#FN}$, where $\#TP$ is the number of true positives (the ones with true nonzero value and also correctly identified as nonzero), $\#FP$ is the number of false

TABLE 4 The comparison of the type I error and the power of the methods based on $\hat{\gamma}^{\text{ini1}}$ and $\hat{\gamma}^{\text{pro1}}$, respectively, for testing $H_0 : \boldsymbol{\beta}_1^* = 0$ vs. $H_1 : \boldsymbol{\beta}_1^* \neq 0$ with simulated datasets in Case (C1)

		Setting 1: $p = 10$				
		Type I error	**Power**			
Sample size	**Method**	**0**	**0.05**	**0.10**	**0.15**	**0.20**
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.021	0.047	0.170	0.438	0.699
	$\hat{\gamma}^{\text{pro1}}$	0.077	0.139	0.378	0.669	0.870
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.022	0.114	0.526	0.899	0.995
	$\hat{\gamma}^{\text{pro1}}$	0.061	0.235	0.723	0.966	0.999
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.022	0.302	0.892	1.000	1.000
	$\hat{\gamma}^{\text{pro1}}$	0.045	0.401	0.935	1.000	1.000
		Setting 2: $p = 50$				
		Type I error	**Power**			
Sample size	**Method**	**0**	**0.05**	**0.10**	**0.15**	**0.20**
$N = 200$	$\hat{\gamma}^{\text{ini1}}$	0.005	0.019	0.088	0.232	0.484
	$\hat{\gamma}^{\text{pro1}}$	0.034	0.079	0.222	0.428	0.690
$N = 500$	$\hat{\gamma}^{\text{ini1}}$	0.018	0.100	0.383	0.833	0.983
	$\hat{\gamma}^{\text{pro1}}$	0.044	0.182	0.571	0.929	0.994
$N = 1000$	$\hat{\gamma}^{\text{ini1}}$	0.030	0.236	0.857	1.000	1.000
	$\hat{\gamma}^{\text{pro1}}$	0.062	0.373	0.916	1.000	1.000

positives (the ones with a atrue zero value but falsely identified as nonzero), and $\#FN$ is the number of false negatives (the ones with a true nonzero value but falsely identified as zero). Below we report the results and the conclusions for the three cases separately. All of the summarized results are based on 1000 simulation replications.

Due to the page limit, we only report the results for case (C1) here. The results for (C2) and (C3) are contained in the online supporting information. For the first case, our sample is selected according to (C1): $r_i = 1$ if and only if $y_i > 1$ and $x_{i1} > -1.5$.

In Table 1, we summarize the sample mean, the first quartile, median, and the third quartile for FP, FN, and the F-measure in each of the simulation settings. In terms of a larger F-measure, the method $\hat{\gamma}^{\text{pro}}$, incorporating the weak signal detection, outperforms $\hat{\gamma}^{\text{ini}}$ obviously. With $p = 50$, the estimator $\hat{\gamma}^{\text{pro}}$ performs slightly worse than $\hat{\gamma}^{\text{ini}}$ and $\hat{\gamma}^{\text{nv}}$ when the sample size is small, but clearly much better once the sample size N increases. Additionally, from the summary of FP and FN values, the estimator $\hat{\gamma}^{\text{ini}}$ which does not attempt to detect the weak signals, hence induces a much more sparse estimation than $\hat{\gamma}^{\text{pro}}$, has FP = 0 almost in every setting but much larger FN values.

The performance of $\hat{\gamma}^{\text{nv}}$, $\hat{\gamma}^{\text{ini}}$, and $\hat{\gamma}^{\text{pro}}$ for both strong and weak signals is presented in Table 2 with estimation, standard deviation, and root mean squared error (RMSE). It is apparent that the estimator $\hat{\gamma}^{\text{pro}}$ has consistently smaller estimation bias and RMSE than $\hat{\gamma}^{\text{ini}}$. For the strong signals, the estimator $\hat{\gamma}^{\text{nv}}$ has nonnegligible estimation bias even when the sample size goes large, and it has smaller standard deviation than $\hat{\gamma}^{\text{pro}}$ as expected. When the sample size is small, $\hat{\gamma}^{\text{nv}}$ has smaller RMSE than $\hat{\gamma}^{\text{pro}}$. But as N goes large, the performance of $\hat{\gamma}^{\text{nv}}$ deteriorates rapidly and has larger RMSE than $\hat{\gamma}^{\text{pro}}$ when $N = 1000$ and $p = 10$. For the weak signals, $\hat{\gamma}^{\text{nv}}$ and $\hat{\gamma}^{\text{pro}}$ have almost comparable performances.

In Table 3, we report the performance of the one-step estimators for strong signals including estimation, standard deviation, standard error, which is the estimated standard deviation $n^{-1/2}\hat{I}_{1|-1}^{-1}\hat{\sigma}_s^{1/2}$ in (12), and coverage probability of the confidence interval based on (12). We compare the proposed one-step estimator $\hat{\gamma}^{\text{pro1}}$ based on $\hat{\gamma}^{\text{pro}}$ with the one-step estimator $\hat{\gamma}^{\text{ini1}}$ based on the preliminary estimator $\hat{\gamma}^{\text{ini}}$. The estimator $\hat{\gamma}^{\text{pro1}}$ is almost unbiased for the strong signals, while $\hat{\gamma}^{\text{ini1}}$ has much larger bias even with a large sample size. The standard error for the estimator $\hat{\gamma}^{\text{pro1}}$ is close to its standard deviation especially when the sample size is large. The coverage probability for the estimator $\hat{\gamma}^{\text{pro1}}$ is much closer to the nominal level 95% than that of $\hat{\gamma}^{\text{ini1}}$. As the sample size increases, it is almost identical to 95%.

We also evaluate the performance of the Wald test to perform hypothesis testing $H_0 : \boldsymbol{\beta}_1^* = 0$ versus $H_1 : \boldsymbol{\beta}_1^* \neq 0$. The type I error and power of the two methods based on $\hat{\gamma}^{\text{ini1}}$ and $\hat{\gamma}^{\text{pro1}}$, respectively, are presented in Table 4. The type I error of the proposed Wald test based on $\hat{\gamma}^{\text{pro1}}$ is close to 5% as the sample size increases. The power of rejecting the null hypothesis increases rapidly when $\boldsymbol{\beta}_1^*$ increases. In addition, the power based on $\hat{\gamma}^{\text{pro1}}$ is much

larger than the power based on $\hat{\gamma}^{\text{inil}}$ especially when the sample size is small.

In summary, through incorporating the weak signals, the proposed method $\hat{\gamma}^{\text{pro}}$ has asymptotically consistent estimation and produces smaller RMSE than the method $\hat{\gamma}^{\text{ini}}$. The one-step estimation $\hat{\gamma}^{\text{prol}}$ based on $\hat{\gamma}^{\text{pro}}$ also has better performance than its counterpart $\hat{\gamma}^{\text{inil}}$, in terms of estimation efficiency, coverage probability of confidence interval, and testing power. The naive estimator $\hat{\gamma}^{\text{nv}}$, which treats the hypothetical population and the source population the same, has nonnegligible estimation bias even when the sample size goes sufficiently large, which demonstrates the need of the proposed methodologies.

6 | DISCUSSION

In this paper, we propose a weak-signal-assisted procedure for both signal detection and statistical inference with an informative subsample. One distinct feature of our proposed methodology compared to existing literature is that we rigorously consider the sample selection issue, which is a common problem in HIV studies.

Compared with a vanilla method, which considers the sample selection issue but does not take weak signals into account, our proposed methodology identifies both strong and weak signals and provides consistent estimates for both types of signals. To address the high-dimensional problem for postselection statistical inference in general, we adopt the one-step estimation based on a score-type function in the high-dimensional space. Our proposed one-step estimator provides standard error estimation, which can be used for confidence interval construction for either strong or weak signals.

We apply the proposed methodology to the HIV-1 drug resistance study, and we are able to replicate most of the findings in the literature. Additionally, we estimate strong signals with a larger magnitude, indicating that the estimation of the strong signal could be attenuated if the weak signals are ignored.

ACKNOWLEDGMENTS

The authors would like to thank the co-editor, the associate editor, and two reviewers for their insightful comments which have helped improve the manuscript substantially. Fang's research was supported by the National Natural Science Foundation of China (11831008, 11601156, 11771146) and the 111 Project (B14019). Zhao's research was supported by the National Science Foundation (DMS-1953526). Ahmed's research was supported by the Natural Sciences and Engineering Research Council of Canada. Qu's research was supported by the National Science Foundation (DMS-1821198).

ORCID

Jiwei Zhao https://orcid.org/0000-0002-9298-9412
Annie Qu https://orcid.org/0000-0002-8396-7828

REFERENCES

Abraham, A.G., Lau, B., Deeks, S., Moore, R.D., Zhang, J., Eron, J., et al. (2011) Missing data on the estimation of the prevalence of accumulated human immunodeficiency virus drug resistance in patients treated with antiretroviral drugs in North America. *American Journal of Epidemiology*, 174, 727–735.

Abzug, M.J., Qin, M., Levin, M.J., Fenton, T., Beeler, J.A., Bellini, W.J., et al. (2012) Immunogenicity, immunologic memory, and safety following measles revaccination in HIV-infected children receiving highly active antiretroviral therapy. *The Journal of Infectious Diseases*, 206, 512–522.

Bickel, P.J. (1975) One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70, 428–434.

Cai, T. and Liu, W. (2011) A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106, 1566–1577.

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2313–2351.

Clavel, F. and Hance, A.J. (2004) HIV drug resistance. *New England Journal of Medicine*, 350, 1023–1035.

Crum-Cianflone, N.F. and Wallace, M.R. (2014) Vaccination in HIV-infected adults. *AIDS Patient Care and STDs*, 28, 397–410.

Donoho, D. and Jin, J. (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105, 14790–14795.

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

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Gao, X., Ahmed, S. and Feng, Y. (2017) Post selection shrinkage estimation for high-dimensional data analysis (with discussion). *Applied Stochastic Models in Business and Industry*, 33, 97–135.

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

Knight, K. and Fu, W. (2000) Asymptotics for Lasso-type estimators. *Annals of Statistics* pages 1356–1378.

Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. (2016) Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 44, 907–927.

Li, Y., Hong, H.G., Ahmed, S.E. and Li, Y. (2019) Weak signals in high-dimensional regression: detection, estimation and prediction. *Applied Stochastic Models in Business and Industry*, 35, 283–298.

Liang, K.-Y. and Qin, J. (2000) Regression analysis under nonstandard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 773–786.

Lockhart, R., Taylor, J., Tibshirani, R.J. and Tibshirani, R. (2014) A significance test for the Lasso. *Annals of Statistics*, 42, 413.

Markman, M. and Bookman, M.A. (2000) Second-line treatment of ovarian cancer. *The Oncologist*, 5, 26–35.

Minnier, J., Tian, L. and Cai, T. (2011) A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106, 1371–1382.

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

Rhee, S.-Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J. and Shafer, R.W. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31, 298–303.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D.L. and Shafer, R.W. (2006) Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103, 17355–17360.

Shafer, R.W. (2006) Rationale and uses of a public HIV drug-resistance database. *The Journal of Infectious Diseases*, 194, S51–S58.

Shi, P. and Qu, A. (2017) Weak signal identification and inference in penalized model selection. *The Annals of Statistics*, 45, 1214–1253.

Teles, R., Benecha, H.K., Preisser, J.S., Moss, K., Starr, J.R., Corby, P., et al. (2016) Modelling changes in clinical attachment loss to classify periodontal disease progression. *Journal of Clinical Periodontology*, 43, 426–434.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Tibshirani, R.J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016) Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111, 600–620.

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

Weng, H., Feng, Y. and Qiao, X. (2019) Regularization after retention in ultrahigh dimensional linear regression models. *Statistica Sinica*, 29, 387–407.

Wensing, A.M., Calvez, V., Günthard, H.F., Johnson, V.A., Paredes, R., Pillay, D., Shafer, R.W. and Richman, D.D. (2017) 2017 update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine*, 24, 132–133.

Wesolowski, R., Lee, C. and Kim, R. (2009) Is there a role for second-line chemotherapy in advanced gastric cancer? *The Lancet Oncology*, 10, 903–912.

Wu, M. (2009). *A parametric permutation test for regression coefficients in LASSO regularized regression*. Ph.D. thesis, Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

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

Zhao, J. and Shao, J. (2017) Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of Systems Science and Complexity*, 30, 139–153.

Zhao, J., Yang, Y. and Ning, Y. (2018) Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statistica Sinica*, 28, 2125–2148.

SUPPORTING INFORMATION

A web appendix referenced in Sections 2 and 5 is available with this paper at the Biometrics website on Wiley Online Library. The R code, the dataset and their descriptions are also available in the datacode.zip file at the Biometrics website on Wiley Online Library.

How to cite this article: Fang F, Zhao J, Ahmed SE, Qu A. A weak-signal-assisted procedure for variable selection and statistical inference with an informative subsample. *Biometrics*. 2020;1–15. https://doi.org/10.1111/biom.13346