Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Published online: 13.07.2025
DOI : 10.15672/hujms.1634222

Research Article

# Post high-dimensional shrinkage estimation for sparse generalized linear models

Seyedeh Zahra Aghamohammadi

*Faculty of Quantum and Biosocial Sciences and Technologies, ISI.C., Islamic Azad University, Tehran, Iran.*

## Abstract

In this study, we propose a new selection and estimation procedure for the regression coefficients of high-dimensional generalized linear models in which many coefficients have weak effects (or weak signals). Many existing procedures for selection of regression coefficients in generalized linear models in the high-dimensional situation such as Least Absolute Shrinkage and Selection Operator, Elastic-Net, Smoothly Clipped Absolute Deviation, and Minimax Concave Penalty are mainly focused on selecting variables with strong effects. This may result in biased parameter estimation, particularly when the number of weak signals is extremely high relative to strong signals. Therefore, in this work, we propose an algorithm in which a variable selection is performed first and then an efficient post-selection estimation based on a weighted ridge estimators along with Stein-type shrinkage strategies is employed. We compute the biases and mean square errors for the proposed estimators and we prove the oracle properties of the selection procedure. We investigate the performance of the new procedure relative to the existing penalized regression methods by using Monte Carlo simulations. Finally, we illustrate the methodology by performing genome-wide association analysis on a cancer data set.

**Mathematics Subject Classification (2020).** 62J12, 62F25.

**Keywords.** Generalized linear models; high-dimensional data; penalized likelihood; weak signals; post shrinkage estimation.

## 1. Introduction

Generalized linear models (GLMs) are widely used in medicine, social sciences, engineering, economics, and health sciences, etc. The GLM class provides a common approach to a wide range of response modeling problems. The most popular cases of GLM are logistic regression and Poisson regression. Consider a classical GLM ([17]), let $Y$ be a response variable and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)^{\mathsf{T}} = (X_{ij})_{n\times p_n}$ be an $n \times p_n$ design matrix and $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip_n})$ be the $p_n \times 1$ vector of covariates of the $i$th observation, $i = 1, \cdots, n$. We assume that $Y$ has a density from the exponential family of distributions,

$$f_Y(y, \theta) = \exp[y\theta - \phi(\theta) + c(y)], \tag{1.1}$$

Email addresses: zahraaghamohamadi@yahoo.com (S. Z. Aghamohammadi)
Received: 06.02.2025; Accepted: 29.06.2025

for some known functions $\phi(.)$ and $c(.)$, where $\theta$ is the so-called canonical or natural parameter in parametric GLM. The mean response is $\phi'(\theta)$, the first derivative of $\phi(\theta)$ with respect to $\theta$. In parametric GLM, the mean function is defined via a known link function g(.),

$$\mu = E(Y|\boldsymbol{x}) = \mathrm{g}^{-1}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}),$$

If g is the canonical link, that is, $\mathrm{g}^{-1} = \phi'$, then $\xi(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}$. In this article, we focus on the canonical link function for simplicity of presentation.

Recent advances in technology have facilitated the collection of high-dimensional data in a variety of scientific disciplines. The remarkable feature of these data is that the number of predictors is typically much larger than the total number of observations. For example, in genome-wide association studies, hundreds of thousands of single nucleotide polymorphisms (SNPs) are potentially associated with genetic markers for the study of human diseases [10]. Other common examples of high-dimensional data include DNA sequencing, molecular biology, signal processing, engineering, and astronomy. The analysis of high-dimensional data poses diverse computational and statistical challenges to classical statistical methods and theories [5].

High-dimensional data typically deals with lot of predictors, many of which may be irrelevant. Removing irrelevant variables from the model is essential since the presence of too many variables may cause overfitting, which leads to poor prediction of future outcomes. In the past two decades, numerous penalized regularization techniques have evolved as a powerful tool to solve the problem of variable selection and estimation simultaneously in linear models or even GLMs. The regularization methods are particularly useful to obtain sparse models compared to simply apply traditional criteria such as Akaike's information criterion [1] and Bayesian information criterion [20]. The least absolute shrinkage and selection operation (LASSO) proposed by [22], is one of the most popular approaches because of its consistency and computational efficiency. Some modifications of LASSO have also been developed, including least angle regression [7], Elastic-Net [29], fused LASSO [23], adaptive LASSO [28], Dantzig-selector [6], square-root LASSO [2], and scaled LASSO [21] to improve estimation and prediction in various problems. For GLMs, much of the research has been done to investigate theoretical properties in high dimension, such as [11, 16, 25, 26]. As the dimension $p_n$ increases with the sample size $n$, it is often assumed that only a small number of predictors contribute to the response, leading to the sparsity of the regression coefficient vector $\boldsymbol{\beta}$. Sparsity means that the number of non-zero components of the vector $\boldsymbol{\beta}$ is less than $n$. Sparse representation not only makes regression results interpretable, but can also make the predictive model more accurate. Additional assumptions made on the design matrix include the irrepresentability and the restricted eigenvalue conditions. For more detailed information, we refer the interested reader to [3, 14, 27].

When $p_n > n$, we are interested in recovering the support and nonzero components of the regression coefficient vector $\boldsymbol{\beta}$. As a powerful tool for producing interpretable models, sparse modeling via penalized regularization has gained popularity for analyzing high-dimensional data sets. Following [13], the regression parameter estimation problem is seen when there are many predictors in the model. The predictors can be characterized into the following three groups:

(1) Predictors with strong signals on the response variable and $|\beta_j| > c\sqrt{\log p_n/n}$ for some $c > 0$ and $1 \leq j \leq p_n$.
(2) Predictors with weak signals that may or may not contribute to explaining the response variable and $0 < |\beta_j| < c\sqrt{\log p_n/n}$ for some $c > 0$ and $1 \leq j \leq p_n$.
(3) Predictors with scarce or no signals on the response variable in which their related regression coefficients are exactly zero.

Many existing inference procedures for high-dimensional penalized estimators may ignore contributions from weak signals, and this will result in biased parameter estimation, particularly when weak signals outnumber strong signals. The objective of this paper is to extend the idea of post-selection shrinkage estimation (proposed by [13]) for GLMs that takes into account the joint impact of strong and weak signals. The proposed estimation strategy dominates relative performances over the candidate submodel estimators generated from the LASSO and Elastic-Net methods. To obtain a high-dimension post-selection shrinkage estimator, we offer the weighted ridge (WR) estimator which is able to separate small coefficients from zero coefficients. We also established the asymptotic normality of the post-selection WR estimator when $p_n$ increases polynomially with sample size $n$, i.e. $p_n = \mathcal{O}(n^\alpha)$ for some $\alpha > 0$. These asymptotic properties are employed to develop the asymptotic efficiency of the suggested post-selection shrinkage analytically. We also performed numerical studies to support our theoretical findings.

The remainder of this paper is organized as follows. We outline the proposed method for GLMs in Section 2 and investigate the asymptotic properties in Section 3. Monte-Carlo simulation studies are conducted in Section 4. In Section 5, an analysis of the real data set from genome-wide association studies using our method is presented. We conclude the paper with a brief discussion in Section 6. All technical proofs are relegated to the Appendix.

## 2. Methods

### 2.1. Notation

In this section, we state some standard assumptions and notations used throughout the paper. We use bold upper-case letters for matrices and lower-case letters for vectors. Moreover, $\mathsf{T}$ denotes the matrix transpose and $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix. Design vectors, or columns of $\boldsymbol{X}$, are denoted by $\boldsymbol{X}_j, j = 1, \cdots, p_n$. The index set $\mathcal{M} = \{1, 2, \cdots, p_n\}$ denotes the full model that contains all potential variables. For a subset $\mathcal{A} \subset \mathcal{M}$, use $\boldsymbol{\beta}_\mathcal{A}$ for a subvector of $\boldsymbol{\beta}_\mathcal{M}$ indexed by $\mathcal{A}$, and $\boldsymbol{X}_\mathcal{A}$ for a submatrix of $\boldsymbol{X}$ whose columns are indexed by $\mathcal{A}$. For a vector $\mathbf{v} = (v_1, \cdots, v_{p_n})^\mathsf{T}$, denote $||\mathbf{v}||_2 = \sqrt{\sum_{j=1}^{p_n} v_j^2}$ and $||\mathbf{v}||_1 = \sum_{j=1}^{p_n} |v_j|$. For any square matrix $\boldsymbol{A}$, let $\Lambda_{\min}(\boldsymbol{A})$ and $\Lambda_{\max}(\boldsymbol{A})$ be the smallest and largest eigen values of $\boldsymbol{A}$, respectively. Given $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of $a$ and $b$. For two positive sequences $a_n$ and $b_n$, $a_n \asymp b_n$ if $a_n$ is the same order as $b_n$. We use $I(.)$ to denote the indicator function; $H_\vartheta(.; \Delta)$ denotes the cumulative distribution function (cdf) of a non-central $\chi^2$-distribution with $\vartheta$ degrees of freedom and non-centrality parameter $\Delta$. We also use $\xrightarrow{\mathcal{D}}$ to indicate convergence in distribution.

Let $\mathcal{S} \subset \{1, \cdots, p_n\}$ be the set of the non-zero coefficient indices with $s = |\mathcal{S}|$ denoting the cardinality of $\mathcal{S}$. We assume that the true coefficient vector $\boldsymbol{\beta}^* = (\beta_1^{*\mathsf{T}}, \cdots, \beta_{p_n}^{*\mathsf{T}})^\mathsf{T}$ is sparse, that is $s < n$. To facilitate theoretical results, let the parameter space be $\Omega_n \subseteq \mathbb{R}^{p_n}$. For any $\boldsymbol{\beta} \in \Omega_n$, let $\phi(\boldsymbol{X}\boldsymbol{\beta}) = (\phi(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}), \cdots, \phi(\boldsymbol{X}_n^\mathsf{T}\boldsymbol{\beta}))^\mathsf{T}$, $\phi'(\boldsymbol{X}\boldsymbol{\beta}) = (\phi'(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}), \cdots, \phi'(\boldsymbol{X}_n^\mathsf{T}\boldsymbol{\beta}))^\mathsf{T}$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}\{\phi''(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}), \cdots, \phi''(\boldsymbol{X}_n^\mathsf{T}\boldsymbol{\beta})\}$. Note that $E(Y) = \phi'(\boldsymbol{X}\boldsymbol{\beta}_0)$ and $\text{Cov}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$. Let

$$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^\mathsf{T}\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\boldsymbol{X} \quad \text{and} \quad \boldsymbol{\Sigma}_\mathcal{S} = \frac{1}{n}\boldsymbol{X}_\mathcal{S}^\mathsf{T}\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\boldsymbol{X}_\mathcal{S}. \tag{2.1}$$

Without loss of generality, we partition the $(n \times p_n)$-matrix $\boldsymbol{X}$ as $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{S}_1}|\boldsymbol{X}_{\mathcal{S}_2}|\boldsymbol{X}_{\mathcal{S}_{\text{null}}})^\mathsf{T}$, where $\mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_{\text{null}} = \varnothing$, $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_{\text{null}} = \mathcal{M}$ and $\mathcal{S}_{\text{null}} = \{j : \beta_{0j} = 0\}$. For two matrices $\boldsymbol{X}_{\mathcal{S}_1}$ and $\boldsymbol{X}_{\mathcal{S}_2}$, we define the corresponding sample covariance matrices by

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathcal{S}_1|\mathcal{S}_2} &= \boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_1} - \boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_2}\boldsymbol{\Sigma}_{\mathcal{S}_2\mathcal{S}_2}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}_2\mathcal{S}_1}, \\ \boldsymbol{\Sigma}_{\mathcal{S}_2|\mathcal{S}_1} &= \boldsymbol{\Sigma}_{\mathcal{S}_2\mathcal{S}_2} - \boldsymbol{\Sigma}_{\mathcal{S}_2\mathcal{S}_1}\boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_2}. \end{aligned} \tag{2.2}$$

Let $\boldsymbol{V} = (\boldsymbol{X}_{S_2}, \boldsymbol{X}_{S_{\text{null}}})^{\mathsf{T}}$ be a $p_n - s_1$ submatrix of $\boldsymbol{X}$. Then, another partition can be written as $\boldsymbol{X} = (\boldsymbol{X}_{S_1}, \boldsymbol{V})^{\mathsf{T}}$. Let $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_{S_1}\hat{\boldsymbol{\Sigma}}^{-1}_{S_1S_1}\boldsymbol{X}^{\mathsf{T}}_{S_1}$ . Then, $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{M}_1\boldsymbol{V}$ is a $(p_n-s_1)\times(p_n-s_1)$ dimensional singular matrix with rank $k_1 \geq 0$. We denote $\varrho_1 \leq \cdots \leq \varrho_{k_1}$ as all $k_1$ positive eigenvalues of $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{M}_1\boldsymbol{V}$.

## 2.2. Variable selection and estimation

Let $(Y_i, \boldsymbol{X}_i), i = 1, \cdots, n$, be independent samples of $(Y, \boldsymbol{X})$. For simultaneous parameter estimation and variable selection in GLM, we define the penalized estimator as the argument that minimizes

$$-\ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} P_\lambda(\beta_j), \tag{2.3}$$

where $\ell_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\left[Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta} - \phi(\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})\right]$, $P_\lambda(\beta_j)$ is a penalty function applied to each component of $\boldsymbol{\beta}$ and $\lambda$ is a tuning parameter that controls the amount of penalization. We consider two popular methods, outlined below.

The LASSO estimator takes the form given in (2.3) with an $L_1$-norm penalty: $\text{Pen}_\lambda(\beta_j) = \lambda|\beta_j|$. This continuously shrinks the coefficients towards 0 when $\lambda$ increases, and some coefficients are shrunk to exact 0 if $\lambda$ is sufficiently large. The theoretical properties of LASSO have been well studied, and an extensive treatment can be found in [4].

The Elastic Net (ENet) estimator is (2.3) with a penalty.

$$P_\lambda(\beta_j) = \lambda(\alpha|\beta_j| + (1-\alpha)\beta_j^2). \tag{2.4}$$

That is, $L_1$ and $L_2$-norm penalties combined with an additional parameter $\alpha \in [0, 1]$ ($\alpha = 1$ and $\alpha = 0$ correspond to LASSO and Ridge, respectively). This combines some of the benefits of Ridge while giving sparse solutions. In the $p_n > n$ setting, LASSO can select at most $n$ variables, but ENet has no such limitation.

Given $a > 2$ and $\lambda > 0$, the penalty function SCAD is according to the following formula:

$$P_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & |\beta_j| < \lambda \\ -\frac{(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \lambda \leq |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda \end{cases} \tag{2.5}$$

Then $P_\lambda(|\beta_j|) = \lambda I(|\beta_j| < \lambda) + \frac{a\lambda - |\beta_j|}{a-1} I(\lambda \leq |\beta_j| < a\lambda)$.

In minimax concave penalty (MCP) selection the penalty takes the following form:

$$P_\lambda(\beta_j) = \begin{cases} \lambda\beta_j - \frac{1}{2a}\beta_j^2 & \beta_j \leq a\lambda \\ \frac{a}{2}\lambda^2 & \beta_j > a\lambda \end{cases} \tag{2.6}$$

for $\lambda > 0$ and $a > 1$. For both SCAD and MCP, the regularization parameter $a$ controls the degree of concavity, with a smaller $a$ corresponding to a penalty that is more concave. Both penalties begin by applying the same rate of penalization as LASSO, and then gradually reduce the penalization rate to zero as $|\beta_j|$ increases.

**2.2.1. Variable selection procedure for $\mathcal{S}_1$ and $\mathcal{S}_2$.** We summarize the variable selection procedure for $\mathcal{S}_1$ and $\mathcal{S}_2$.

**Step 1** (*Detection of* $\mathcal{S}_1$). Obtain a candidate subset $\mathcal{S}_1$ of strong signals using a penalized regression method. We consider the following penalized likelihood estimator (PLE):

$$\hat{\boldsymbol{\beta}}^{\text{PLE}} = \arg\min_{\boldsymbol{\beta}}\{-\ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p_n} P_\lambda(\beta_j)\}, \tag{2.7}$$

where $P_\lambda(\beta_j)$ is a penalty for each individual $\beta_j$ to shrink the weak effects toward zeros and select the strong signals, the tuning parameter $\lambda > 0$ controlling the size of the candidate subset $\hat{\mathcal{S}}_1$.

**Step 2** (*Detection of* $\mathcal{S}_2$). To identify $\hat{\mathcal{S}}_2$, we first solve a regression problem with a ridge penalty on only the variables in $\hat{\mathcal{S}}_1^c$. That is,

$$\hat{\boldsymbol{\beta}}^r = \arg\min_{\boldsymbol{\beta}} \Big\{ -\ell(\boldsymbol{\beta}) + r_n ||\boldsymbol{\beta}_{\hat{\mathcal{S}}_1^c}||_2^2 \Big\}, \tag{2.8}$$

where $r_n > 0$ is a tuning parameter that controls the overall strength of the variables selected in $\hat{\mathcal{S}}_1^c$. Then, a post-selection weighted ridge (WR) estimator $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}$ has the form

$$\hat{\beta}_j^{\mathrm{WR}} = \begin{cases} \hat{\beta}_j^r, & j \in \hat{\mathcal{S}}_1, \\ \hat{\beta}_j^r 1(|\hat{\beta}_j^r| > a_n), & j \in \hat{\mathcal{S}}_1^c, \end{cases} \tag{2.9}$$

where $a_n$ is a thresholding parameter. Then, the candidate subset $\hat{\mathcal{S}}_2$ is obtained by

$$\hat{\mathcal{S}}_2 = \{ j \in \hat{\mathcal{S}}_1^c : \hat{\beta}_j^{\mathrm{WR}} \neq 0, 1 \leq j \leq p \}. \tag{2.10}$$

The post-selection estimation strategy is only used when the threshold parameter $a_n$ satisfies $|\hat{\mathcal{S}}_2| > 2$. In particular, we set

$$a_n = cn^{-\kappa}, \quad 0 < \kappa \leq 1/2. \tag{2.11}$$

Note that $\hat{\mathcal{S}}_{null} = \mathcal{S} - (\hat{\mathcal{S}}_1 \bigcup \hat{\mathcal{S}}_2)$ and this is the set of variables that will be discarded as irrelevant signals.

**2.2.2. Post-selection estimation strategies.** In this section, we propose a shrinkage estimate based on two post-selection estimators $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}$. Recall that $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}} = (\hat{\beta}_j^r, j \in \hat{\mathcal{S}}_1)^{\mathsf{T}}$ and $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_2}^{\mathrm{WR}} = (\hat{\beta}_j^r 1(|\hat{\beta}_j^r| > a_n), j \in \hat{\mathcal{S}}_2)^{\mathsf{T}}$.

We obtain the post-selection shrinkage estimator $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{SE}}$ by

$$\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{SE}} = \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}} - \Big(\frac{\hat{s}_2 - 2}{\hat{T}_n}\Big)(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{R}E}), \tag{2.12}$$

where $\hat{s}_2 = |\hat{\mathcal{S}}_2|$ , the post selection restirected estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}$ restricted to $\hat{\mathcal{S}}_1$ is constructed by

$$\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{RE}} = \hat{\boldsymbol{\Sigma}}_{\hat{\mathcal{S}}_1}^{-1} \boldsymbol{X}_{\hat{\mathcal{S}}_1}^{\mathsf{T}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}) \hat{\boldsymbol{Z}} \tag{2.13}$$

where $\hat{\boldsymbol{Z}}$ is a $n$-dimensional vector of working response with elements

$$\hat{Z}_i = \boldsymbol{X}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}} + (Y_i - \hat{\mu}_i) \mathrm{g}'(\hat{\mu}_i), \quad i = 1, 2, \cdots, n.$$

and $\hat{T}_n$ is as defined by

$$\hat{T}_n = (\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_2}^{\mathrm{WR}})^{\mathsf{T}} \Big( \boldsymbol{X}_{\hat{\mathcal{S}}_2}^{\mathsf{T}} \boldsymbol{M}_{\hat{\mathcal{S}}_1} \boldsymbol{X}_{\hat{\mathcal{S}}_2} \Big)^{-1} \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_2}^{\mathrm{WR}}, \tag{2.14}$$

with $\boldsymbol{M}_{\hat{\mathcal{S}}_1} = \boldsymbol{I}_n - \boldsymbol{X}_{\hat{\mathcal{S}}_1} \hat{\boldsymbol{\Sigma}}_{\hat{\mathcal{S}}_1}^{-1} \boldsymbol{X}_{\hat{\mathcal{S}}_1}^{\mathsf{T}}$. A generalized inverse is used if $\hat{\boldsymbol{\Sigma}}_{\hat{\mathcal{S}}_1}$ is not singular. To avoid over-shrinking where $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}}$ has a different sign from $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{SE}}$, we consider a positive shrinkage estimator given by a convex combination of $\boldsymbol{\beta}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}}$ and $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{RE}}$,

$$\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{PSE}} = \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}} - \Big(\frac{\hat{s}_2 - 2}{\hat{T}_n} \wedge 1\Big)(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{RE}}), \tag{2.15}$$

Again, we emphasize here that $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{PSE}}$ is particularly important for controlling the over-shrinking problem inherited in the shrinkage estimator.

### 2.3. Selection of threshold parameters

To implement the procedure of simultaneous estimation and selection of variables based on penalized loss function, it is necessary to choose the appropriate value of the threshold parameter $a_n$. One of the ways to select $a_n$ is by minimizing the generalized cross-validation (GCV) criterion. Tibshirani [22] and Fan and Li [8] used the same criterion for the selection of the threshold parameter. However, Wang et al. [24] pointed out that even if the sample size goes to infinity, the GCV criterion has a non-ignorable overfitting. Further, for selection of threshold parameter they have proposed generalized information criterion and showed that it is able to identify the true model consistently. [19] defined a BIC-type selector for selecting the threshold parameter and, through simulation study, demonstrated that it gives good results. This motivated us to select the optimal sn by minimizing the BIC-type criterion.

$$BCI(a_n) = \ell_n(\hat{\boldsymbol{\beta}}_{a_n}) + \frac{1}{n} DF_{a_n} \log(n) \tag{2.16}$$

where $\hat{\boldsymbol{\beta}}_{a_n}$ is penalized estimator of $\boldsymbol{\beta}$ at threshold parameter $a_n$, $\ell_n(\hat{\boldsymbol{\beta}}_{a_n})$ is loss function of estimator evaluated at $\hat{\boldsymbol{\beta}}_{a_n}$ and $DF_{a_n}$ denotes number of non-zero components in $\hat{\boldsymbol{\beta}}_{a_n}$.

## 3. Asymptotic properties

In this section, we study the asymptotic properties of post-selection shrinkage estimators. To investigate the asymptotic theory, we need the following regularity conditions.
**(B1)** $p_n = \exp(\mathcal{O}(n^{\nu}))$ for some $0 < \nu < 1$.
**(B2)** $\varrho_{1n} = \mathcal{O}(n^{-\eta})$, where $0 < \tau < \eta \leq 1$.
**(B3)** There exists a positive definite matrix $\boldsymbol{\Sigma}_n$ such that $\lim_{n\to\infty} \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, where the eigenvalues of $\boldsymbol{\Sigma}$ satisfy $0 < \kappa_1 < \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) < \kappa_2 < \infty$.
**(B4)** Sparse Riesz condition. For the random design matrix $\boldsymbol{X}$, any $\mathcal{S} \subset \{1, \cdots, p_n\}$ with $|\mathcal{S}| = q, q \leq p_n$, and any vector $\mathbf{v} \in \mathbb{R}^q$, there exist $0 < c_* < c^* < \infty$ such that $c_* \leq ||\boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\mathbf{v}||_2/||\mathbf{v}||_2 \leq c^*$ holds with probability tending to 1.
**(B5)** Assume that $||\boldsymbol{\beta}_{\mathcal{S}_2}||_2 \sim O(n^{\tau})$ for some $0 < \tau < 1$, where $||.||_2$ is the Euclidean norm.

Here, condition **(B1)** allows high dimensionality such that the number of predictors can increase with sample size at an almost exponential rate. The condition **(B2)** guarantees that the positive eigenvalues of $\boldsymbol{Z}^T\boldsymbol{M}_1\boldsymbol{Z}$ cannot be too small with a rate associated with the strength of the weak signals in $\mathcal{S}_2$. The condition **(B3)** assumes that the eigenvalues of $\boldsymbol{\Sigma}$ are bounded away from zero and infinity. This is reasonable since the number of nonzero covariates is small in a sparse model. **(B4)** guarantees that $\mathcal{S}_1$ can be recovered with probability tending to 1 as $n \to \infty$. **(B5)**, which bounds the total size of weak signals on $\mathcal{S}_2$, is required for selection consistency on $\mathcal{S}_2$.

The following theorems will make it easier to compute the ADB and ADR of the proposed estimators (see [13]).

**Theorem 3.1.** *Suppose that assumptions* **(B1)**-**(B5)** *hold. If we choose*

$$r_n = c_2 a_n^{-2} (\log\log n)^3 \log(n \vee p_n) \tag{3.1}$$

*for some constant* $c_2 > 0$ *and* $a_n$ *defined in (2.11) with* $\nu < (\eta - \alpha - \tau)/3$, *then* $\hat{\mathcal{S}}_2$ *in (2.10) satisfies*

$$\lim_{n\to\infty} P(\hat{\mathcal{S}}_2 = \mathcal{S}_2 | \hat{\mathcal{S}}_1 = \mathcal{S}_1) = 1. \tag{3.2}$$

*where* $\tau, \eta$ *and* $\alpha$ *are defined in (***B1***), (***B2***) and (***B5***), respectively.*

**Theorem 3.2.** *Let* $s_n^2 = \boldsymbol{d}_n^{\mathsf{T}}\Sigma_n^{-1}\boldsymbol{d}_n$ *for any* $(p_{1n} + p_{2n}) \times 1$ *vector* $\boldsymbol{d}_n$ *satisfying* $||\boldsymbol{d}_n||_2 \leq 1$. *Suppose assumptions (***B1***)-(***B5***) hold and a pre-selected model such as* $\mathcal{S}_1 \subset \hat{\mathcal{S}}_1 \subset \mathcal{S}_1 \cup \mathcal{S}_2$

*is obtained with probability 1. If we choose $r_n$ in Theorem 3.1 with $\nu < \{(\eta-\alpha-\tau)/, 1/4-\tau/2\}$, then we have the asymptotic normality,*

$$n^{1/2}s_n^{-1}\boldsymbol{d}_n^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{\mathcal{S}^c_{null}}^{WR} - \boldsymbol{\beta}_{\mathcal{S}^c_{null}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{3.3}$$

### 3.1. Asymptotic distributional bias and risk analysis

In order to compare the estimators, we use the asymptotic distributional bias (ADB) and the asymptotic risk (ADR) expressions of the proposed estimators.

**Definition 3.3.** For any estimator $\boldsymbol{\beta}_{1n}^*$ and $p_{1n}$-dimensional vector $\boldsymbol{d}_{1n}$, satisfying $||\boldsymbol{d}_{1n}||_2 \leq 1$, the ADB and ADR of $\boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\beta}_{1n}^*$, respectively, are defined as

$$ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\beta}_{1n}^*) = \lim_{n\to\infty} E[\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_{01})\}], \tag{3.4}$$

$$ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\beta}_{1n}^*) = \lim_{n\to\infty} E[\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_{01})\}^2], \tag{3.5}$$

where $s_{1n}^2 = \boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathcal{S}_1|\mathcal{S}_2}^{-1}\boldsymbol{d}_{1n}$. Let $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_{p_{2n}})^{\mathsf{T}} \in \mathbb{R}^{p_{2n}}$ and

$$\boldsymbol{\Delta}_{\boldsymbol{d}_{1n}} = \frac{\boldsymbol{d}_{1n}^{\mathsf{T}}(\boldsymbol{\Sigma}_{\mathcal{S}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_2}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\Sigma_{\mathcal{S}_2\mathcal{S}_1}\boldsymbol{\Sigma}_{\mathcal{S}_1}^{-1})\boldsymbol{d}_{1n}}{\boldsymbol{d}_{1n}^{\mathsf{T}}(\boldsymbol{\Sigma}_{\mathcal{S}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}_1\mathcal{S}_2}\boldsymbol{\Sigma}_{\mathcal{S}_2|\mathcal{S}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}_2\mathcal{S}_1}\boldsymbol{\Sigma}_{\mathcal{S}_1}^{-1})\boldsymbol{d}_{1n}}. \tag{3.6}$$

We have the following Theorems on the expression of ADBs and ADRs of the post-selection estimators.

**Theorem 3.4.** *Let $\boldsymbol{d}_{1n}$ be any $p_{1n}$-dimensional vector satisfying $0 < ||\boldsymbol{d}_{1n}||_2 \leq 1$ and $s_{1n}^2 = \boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathcal{S}_1|\mathcal{S}_2}^{-1}\boldsymbol{d}_{1n}$. Under the assumption (**B5**), we have*

$$ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{WR}) = 0, \tag{3.7}$$

$$ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{RE}) = s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2, \tag{3.8}$$

$$ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) = (p_2-2)s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 E[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d_2}})], \tag{3.9}$$

$$\begin{aligned} ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) = s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 \Bigg[ (p_2-2)\Bigg\{ & E[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ & + E\big[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi_{p_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_2-2))\big]\Bigg\} - H_{p_2}\Big(p_2-2; \boldsymbol{\Delta}_{\boldsymbol{d}_2}\Big)\Bigg], \end{aligned} \tag{3.10}$$

*where $\boldsymbol{d}_{2n} = \Sigma_{\mathcal{S}_2\mathcal{S}_1}\Sigma_{\mathcal{S}_1}^{-1}\boldsymbol{d}_{1n}$ and $E[\chi_{p_2}^{-2i}(\boldsymbol{\Delta}_{\boldsymbol{d_2}})] = \int_0^\infty x^{-2i}dH_{p_2}(x; \boldsymbol{\Delta}_{\boldsymbol{d_2}})$.*

***Proof:*** See the Appendix for a detailed proof.

**Theorem 3.5.** *Let (***B5***) is replaced by* $\beta_{0j} = \delta_j/\sqrt{n}$, *for* $j \in \mathcal{S}_2$, *with* $|\delta_j| < \delta_{\max}$, *for some* $\delta_{\max} > 0$, *we have*

$$ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{WR}) = 1, \tag{3.11}$$

$$ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{RE}) = 1 + (1-c)^{1/2}[2 + (1-c)^{1/2}(1 + 2\boldsymbol{\Delta}_{\boldsymbol{d}_2})], \tag{3.12}$$

$$\begin{aligned} ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) &= 1 + (1-c)^{1/2}(p_2-2)\Big[(1-c)^{1/2}(p_2-2)\Big\{E[\chi^{-4}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ &\quad + (s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2 E[\chi^{-4}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]\Big\} + 2E[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]\Big], \end{aligned} \tag{3.13}$$

$$\begin{aligned} ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) &= 1 + (1-c)(p_2-2)^2\Bigg\{E[\chi^{-4}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ &\quad + (s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2 E[\chi^{-4}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ &\quad + E[\chi^{-4}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))]\Bigg\} \\ &\quad + 2(1-c)^{1/2}(p_2-2)\Bigg\{E[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ &\quad + E[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))] \\ &\quad - (p_2-2)E[\chi^{-4}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))] - (1-c)^{1/2} \\ &\quad \times \Bigg[E[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))] \\ &\quad + (s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2 E[\chi^{-2}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))]\Bigg]\Bigg\}, \\ &\quad + (1-c)^{1/2}\Bigg[(1-c)^{1/2}\Bigg(E[\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})] \\ &\quad + (s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2 H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Bigg) \\ &\quad + 2\Bigg\{H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2}) - (p_2-2) \\ &\quad \times E[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2}) < (p_{2n}-2))]\Bigg\}\Bigg], \end{aligned} \tag{3.14}$$

*where* $c = \lim_{n\to\infty} \boldsymbol{d}_{1n}^{\mathsf{T}}\Sigma_{\mathcal{S}_1}^{-1}\boldsymbol{d}_{1n}/(\boldsymbol{d}_{1n}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathcal{S}_1|\mathcal{S}_2}^{-1}\boldsymbol{d}_{1n}) \le 1$ *and* $s_{2n}^2 = \boldsymbol{d}_{2n}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathcal{S}_2|\mathcal{S}_1}^{-1}\boldsymbol{d}_{2n}$.

***Proof:*** See the Appendix for a detailed proof.

From Theorem 3.5, we can compare the ADRs of the estimators.

**Corollary 3.6.** *Under assumptions in Theorem 3.5, we have*

(1) *If* $||\boldsymbol{\delta}||_2 \le 1$, *then* $ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) \le ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) \le ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{WR})$;
(2) *If* $||\boldsymbol{\delta}||_2 = o(1)$ *and* $p_{2n} \to \infty$, *then for* $\boldsymbol{\delta} = 0$,

$$ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{RE}) < ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) \le ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{WR}).$$

Corollary 3.6 shows that the performance of the post-selection PSE is closely related to the RE. On one hand, if $\hat{\mathcal{S}}_1 \subset \mathcal{S}_1 \cup \mathcal{S}_2$ and $(\mathcal{S}_1 \cup \mathcal{S}_2) \cap \hat{\mathcal{S}}_1^c$ are large, then the post-selection PSE

tends to dominate the RE. Furthermore, if a variable selection method generates the right submodel and $||\boldsymbol{\delta}||_2 = o(1)$ , that is, $\lim_{n\to\infty} \hat{\mathcal{S}}_1 = \mathcal{S}_1 \cup \mathcal{S}_2$, then a post-selection likelihood estimator $\hat{\boldsymbol{\beta}}_{1n}^{RE}$ is the most efficient compared to all other post-selection estimators.

**Remark 3.7.** The simultaneous variable selection and parameter estimation may not lead to a good estimation strategy when weak signals are co-exist with zero signals. Even though selected candidate subset models can be provided by some existing variable selection techniques when $p_n > n$, the prediction performance can be improved by the post-selection shrinkage strategy, especially when an under-fitted subset model is selected by an aggressive variable selection procedure.

## 4. Simulation study

We consider the estimation problem to the high-dimensional GLMs in which $p_n > n$. The response variable was generated from the two logistic and Poisson regression models with the following regression coefficient vector under the three effect sizes such as strong, weak, and no effect,

$$\boldsymbol{\beta}_0 = (\overbrace{5,5,5}^{\mathcal{S}_1}, \underbrace{\overbrace{0.5,\ldots,0.5}^{\mathcal{S}_2}}_{30}, \underbrace{\overbrace{0,0,0,0,\ldots,0}^{\mathcal{S}_{\text{null}}}}_{p_n-p_1-p_2})^{\mathsf{T}}. \tag{4.1}$$

The covariate $\boldsymbol{x}_i$ is generated from a $p_n$-dimensional multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{I}_{p_n\times p_n}$. We separated the method under a high-dimensional setting into two steps, namely:

(1) A variable selection step to detect significant predictors and to reduce the dimensions to a low-dimensional model.
(2) A post-selection parameter estimation step, using the resulting model attained from step 1 above.

In variable selection, we adopt the LASSO, ENet, SCAD and MCP method to eliminate predictors with no signals and to keep predictors with both strong and weak signals. All parameter tuning in variable selection approaches are chosen using cross-validation (CV). We also apply the function cv.glmnet from the R statistical package glmnet with 10-fold CV for tuning parameter selection. In particular, cv.glmnet is applied to obtain both LASSO and ENet estimators. We use the ncvreg package in the R software to obtain SCAD and MCP estimators, respectively.

We use different $(n, p_n)$ combinations. For each combination, we run Monte-Carlo studies with 1000 replicates. Let $\boldsymbol{\beta}_{1n}^*$ be either $\hat{\boldsymbol{\beta}}_{1n}^{\text{PSE}}$ or $\hat{\boldsymbol{\beta}}_{1n}^{\text{RE}}$ after the variable selection. The performance of an estimator $\boldsymbol{\beta}_{1n}^*$ is evaluated by the relative mean squared error (RMSE) criterion with respect to $\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}}$ as follows:

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}}, \boldsymbol{\beta}_{1n}^*) = \frac{E[||\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}} - \boldsymbol{\beta}_0||_2]}{E[||\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_0||_2]}$$

The $\text{RMSE}(\boldsymbol{\beta}_{1n}^*) > 1$ means the superiority of $\boldsymbol{\beta}_{1n}^*$ over $\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}}$. Larger RMSE indicates the stronger degree of superiority of the estimator $\boldsymbol{\beta}_{1n}^*$ over $\hat{\boldsymbol{\beta}}_{1n}^{\text{WR}}$. The results on RMSE from 1000 iterations are reported in Table 1. To check the behavior of LASSO, ENet, SCAD and MCP for variable selection, we further report the average number of selected important covariates as $|\hat{\mathcal{S}}_1|$. From tabulated values it is observed that $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\text{PSE}}$ outperforms LASSO, ENet, SCAD and MCP in identifying signals in $\mathcal{S}$. Figures 1-10 show results when the LASSO, ENet, SCAD and MCP is utilized to generate the submodel. We summarize the simulation results as follows:

(1) For all combination of $n$ and $p_n$, it is clear that $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{\mathcal{S}}_1}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{\hat{\mathcal{S}}_1}$ perform better than LASSO, ENet, SCAD and MCP. This suggests that these estimators provide better predictive accuracy and stability.
(2) When data includes both strong signals and weak signals, all of LASSO, ENet, SCAD and MCP tend to ignore those weak covariates. In this case, the post-selection shrinkage estimator dominates these estimators in terms of the risk performance. This is because shrinkage estimators can recover some information ignored by LASSO, ENet, SCAD and MCP when they underfit the model.
(3) Post-selection shrinkage estimators are more stable than variable selection estimators in terms of risk performance. They are not seriously affected by a heavily underfitted model.
(4) The relative performances of the proposed estimators become better when $p_n$ grows for fixed $n$. This trend suggests that LASSO, ENet, SCAD and MCP struggle with larger feature spaces, likely due to their tendency to aggressively shrink weaker covariates. In contrast, the post-selection estimators show relatively stable RMSE behavior, indicating their ability to retain relevant information even in high-dimensional settings.
(5) SCAD and MCP generally has lower RMSE values than LASSO and ENet, particularly for small sample sizes, indicating its advantage in balancing feature selection and regularization.

Figures 1-5 visualize the RMSE trends of the logistic regression model for different values of $p_n$ when comparing LASSO (Figure 1), ENet (Figure 3), SCAD (Figure 4) and MCP (Figure 5) against the proposed estimators. The plots indicate how RMSE varies as $p_n$ increases for different sample sizes $n$. The same resultshold for the Poisson regression model.

In order to compare the sparsity of the coefficient estimators, we also evaluate the False Positive Rate (FPR) defined as

$$\mathrm{FPR}(\hat{\boldsymbol{\beta}}_0) = \frac{|\{j = 0, \ldots, p; \hat{\beta}_{0j} \neq 0 \wedge \beta_{0j} = 0\}|}{|\{j = 0, \ldots, p; \beta_{0j} = 0\}|}. \tag{4.2}$$

The FPR is the proportion of non-informative variables that are incorrectly included in the model. This value is desired to be as small as possible. However, a large FPR indicates that unnecessary associations are included, which 'only' complicates the model (see [15]). Note that if $\boldsymbol{\beta}_0$ does not contain any zero components, FPRe is not defined. This evaluation measure is denoted for the generated data in each of 1000 simulation replications separately, and then summarized in Table 1 and in Figures 4 and 10. The lower the value of this criterion, the better the performance of the method.

**Table 1.** The RMSE values and FPR averaged over $N = 1000$ runs.

| Model | $n$ | $p_n$ | Method | $\lvert\hat{\mathcal{S}}_1\rvert$ | FPR | $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}$ | $\hat{\boldsymbol{\beta}}^{RE}_{\hat{\mathcal{S}}_1}$ | $\hat{\boldsymbol{\beta}}^{PSE}_{\hat{\mathcal{S}}_1}$ |
|---|---|---|---|---|---|---|---|---|
| Logistic | 200 | 200 | LASSO | 35.988 | 0.1736 | 1.1607 | 3.5262 | 3.9596 |
| | | | ENet | 36.002 | 0.1796 | 1.6677 | 4.1716 | 4.6844 |
| | | | SCAD | 32.084 | 0.1692 | 1.0913 | 3.2415 | 3.7217 |
| | | | MCP | 30.111 | 0.1702 | 1.2048 | 3.6405 | 3.888 |
| | | 250 | LASSO | 38.996 | 0.1797 | 0.3301 | 0.9326 | 1.0300 |
| | | | ENet | 36.993 | 0.2027 | 0.4242 | 1.0657 | 1.1816 |
| | | | SCAD | 34.865 | 0.2333 | 0.2194 | 0.2212 | 1.0217 |
| | | | MCP | 35.536 | 0.2123 | 0.3375 | 0.2663 | 1.0301 |
| | | 300 | LASSO | 37.994 | 0.1535 | 0.2415 | 0.9033 | 0.9961 |
| | | | ENet | 37.991 | 0.1685 | 0.2848 | 0.9316 | 0.9960 |
| | | | SCAD | 36.087 | 0.1551 | 0.2205 | 0.8875 | 0.8706 |
| | | | MCP | 35.988 | 0.1634 | 0.2591 | 0.8963 | 0.9231 |
| | 300 | 300 | LASSO | 33.008 | 0.1610 | 1.5494 | 1.6230 | 1.7295 |
| | | | ENet | 33.100 | 0.1647 | 0.9420 | 1.0515 | 1.1155 |
| | | | SCAD | 32.998 | 0.1584 | 0.9233 | 1.1988 | 1.1306 |
| | | | MCP | 33.022 | 0.1611 | 0.9537 | 1.1402 | 1.5238 |
| | | 450 | LASSO | 36.003 | 0.1247 | 1.4410 | 1.3448 | 1.4495 |
| | | | ENet | 36.003 | 0.1270 | 1.7995 | 1.5887 | 1.7131 |
| | | | SCAD | 35.978 | 0.1254 | 1.3371 | 1.2243 | 1.3942 |
| | | | MCP | 33.222 | 0.1198 | 1.5277 | 1.2305 | 1.4236 |
| | | 500 | LASSO | 35.004 | 0.1177 | 0.9994 | 1.0890 | 1.1470 |
| | | | ENet | 37.000 | 0.1177 | 0.9814 | 1.0656 | 1.1201 |
| | | | SCAD | 37.023 | 0.1163 | 0.9405 | 1.1005 | 1.1187 |
| | | | MCP | 37.155 | 0.1094 | 0.9528 | 1.1257 | 1.1195 |
| | 400 | 400 | LASSO | 33.996 | 0.0900 | 1.9431 | 2.2596 | 2.4025 |
| | | | ENet | 34.000 | 0.0916 | 2.3976 | 2.5437 | 2.7099 |
| | | | SCAD | 34.013 | 0.0902 | 1.8862 | 2.1902 | 2.3348 |
| | | | MCP | 34.211 | 0.0898 | 1.8233 | 2.2023 | 2.4324 |
| | | 500 | LASSO | 35.003 | 0.0899 | 1.3406 | 1.3907 | 1.4169 |
| | | | ENet | 33.001 | 0.0214 | 0.3831 | 2.4878 | 2.6432 |
| | | | SCAD | 34.200 | 0.0452 | 1.2204 | 1.4333 | 1.5204 |
| | | | MCP | 33.355 | 0.0400 | 1.3202 | 1.5475 | 2.2001 |
| | | 600 | LASSO | 34.998 | 0.0846 | 1.1135 | 1.2607 | 1.3337 |
| | | | ENet | 33.050 | 0.0194 | 0.1621 | 1.6631 | 1.7571 |
| | | | SCAD | 32.293 | 0.0525 | 1.2363 | 1.4302 | 1.3585 |
| | | | MCP | 33.114 | 0.0873 | 1.4238 | 1.5204 | 1.6228 |
| | 500 | 500 | LASSO | 36.000 | 0.1199 | 2.9295 | 3.4447 | 3.5971 |
| | | | ENet | 35.997 | 0.1284 | 1.5312 | 2.1733 | 2.2699 |
| | | | SCAD | 35.828 | 0.1196 | 1.3341 | 2.5232 | 2.9555 |
| | | | MCP | 35.532 | 0.1225 | 1.5582 | 2.1184 | 2.2122 |
| | | 700 | LASSO | 34.000 | 0.0869 | 1.6610 | 2.2539 | 2.3670 |
| | | | ENet | 34.000 | 0.0884 | 1.6324 | 2.3195 | 2.4313 |
| | | | SCAD | 33.945 | 0.0874 | 1.5423 | 2.1444 | 2.2236 |
| | | | MCP | 33.852 | 0.0923 | 1.5572 | 2.2208 | 2.3151 |
| | | 1000 | LASSO | 35.002 | 0.0775 | 2.1598 | 2.6015 | 2.7025 |
| | | | ENet | 33.000 | 0.0258 | 0.7346 | 2.4387 | 2.5410 |
| | | | SCAD | 34.200 | 0.0386 | 0.9231 | 2.2341 | 2.5970 |
| | | | MCP | 35.010 | 0.0444 | 0.8862 | 2.2243 | 2.6133 |

**Table 2.** The RMSE values and FPR averaged over $N = 1000$ runs.(countinued)

| Model | $n$ | $p_n$ | Method | $\lvert\hat{\mathcal{S}}_1\rvert$ | FPR | $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}$ | $\hat{\boldsymbol{\beta}}^{RE}_{\hat{\mathcal{S}}_1}$ | $\hat{\boldsymbol{\beta}}^{PSE}_{\hat{\mathcal{S}}_1}$ |
|---|---|---|---|---|---|---|---|---|
| Poisson | 200 | 200 | LASSO | 39.587 | 0.0350 | 2.9705 | 3.6283 | 4.8609 |
| | | | ENet | 37.011 | 0.0538 | 3.7216 | 4.2209 | 5.1548 |
| | | | SCAD | 38.052 | 0.0723 | 2.8763 | 3.8890 | 4.9543 |
| | | | MCP | 38.000 | 0.0334 | 2.5411 | 3.4420 | 4.1401 |
| | | 250 | LASSO | 41.889 | 0.0184 | 0.5227 | 1.3390 | 1.5883 |
| | | | ENet | 38.002 | 0.0297 | 0.8017 | 2.4403 | 2.6913 |
| | | | SCAD | 39.022 | 0.0336 | 0.7095 | 2.2331 | 2.5542 |
| | | | MCP | 38.110 | 0.0281 | 0.6642 | 2.1244 | 2.1181 |
| | | 300 | LASSO | 39.915 | 0.0037 | 0.3341 | 0.9546 | 1.0702 |
| | | | ENet | 38.866 | 0.0055 | 0.5934 | 1.4728 | 1.8991 |
| | | | SCAD | 40.000 | 0.0077 | 0.3422 | 0.9624 | 1.6582 |
| | | | MCP | 39.333 | 0.0035 | 0.3852 | 0.9288 | 1.5531 |
| | 300 | 300 | LASSO | 37.005 | 0.0281 | 3.1058 | 3.6625 | 4.0410 |
| | | | ENet | 35.020 | 0.0311 | 3.7355 | 4.0668 | 4.3787 |
| | | | SCAD | 36.056 | 0.0433 | 3.9875 | 4.0225 | 4.2231 |
| | | | MCP | 35.118 | 0.0651 | 3.8657 | 4.7975 | 4.5532 |
| | | 450 | LASSO | 39.001 | 0.0143 | 3.0217 | 3.4878 | 3.9898 |
| | | | ENet | 39.008 | 0.0198 | 3.5671 | 4.2902 | 4.4413 |
| | | | SCAD | 40.023 | 0.0226 | 3.6654 | 4.0100 | 4.2341 |
| | | | MCP | 39.066 | 0.1736 | 3.2234 | 3.8975 | 4.1284 |
| | | 500 | LASSO | 40.030 | 0.0021 | 2.9901 | 3.2003 | 3.6333 |
| | | | ENet | 38.994 | 0.0033 | 3.0889 | 3.9177 | 4.2441 |
| | | | SCAD | 39.00 | 0.0044 | 3.1011 | 4.0024 | 4.1335 |
| | | | MCP | 39.00 | 0.0073 | 3.2325 | 3.9997 | 4.0228 |
| | 400 | 400 | LASSO | 41.000 | 0.0238 | 3.8195 | 4.1112 | 4.4496 |
| | | | ENet | 41.300 | 0.0414 | 4.1911 | 4.6937 | 5.1204 |
| | | | SCAD | 40.00 | 0.0523 | 4.2291 | 4.6617 | 5.0042 |
| | | | MCP | 41.240 | 0.03452 | 3.6614 | 4.0978 | 5.1041 |
| | | 500 | LASSO | 40.988 | 0.0042 | 3.5093 | 4.0017 | 4.2588 |
| | | | ENet | 40.000 | 0.0063 | 4.3858 | 4.7390 | 5.0136 |
| | | | SCAD | 41.035 | 0.0022 | 4.2135 | 4.6541 | 5.2336 |
| | | | MCP | 40.001 | 0.0064 | 3.9889 | 4.5311 | 5.1203 |
| | | 600 | LASSO | 39.000 | 0.0017 | 3.2685 | 3.9765 | 4.1302 |
| | | | ENet | 39.002 | 0.0039 | 3.8716 | 4.5792 | 5.1119 |
| | | | SCAD | 40.000 | 0.0040 | 3.7887 | 4.6531 | 5.1090 |
| | | | MCP | 40.001 | 0.0056 | 3.6228 | 4.3444 | 4.9892 |
| | 500 | 500 | LASSO | 40.013 | 0.0205 | 4.1924 | 4.6721 | 4.9023 |
| | | | ENet | 40.001 | 0.0317 | 4.8817 | 5.2996 | 5.8943 |
| | | | SCAD | 40.022 | 0.0421 | 4.3668 | 5.3000 | 5.9768 |
| | | | MCP | 41.000 | 0.0388 | 4.2265 | 5.4502 | 5.7673 |
| | | 700 | LASSO | 39.008 | 0.0194 | 3.5377 | 3.4966 | 3.5138 |
| | | | ENet | 38.000 | 0.0217 | 4.1444 | 4.7397 | 4.9906 |
| | | | SCAD | 39.00 | 0.0343 | 3.4431 | 4.2954 | 4.5456 |
| | | | MCP | 40.00 | 0.0356 | 4.1205 | 4.2322 | 4.3326 |
| | | 1000 | LASSO | 40.996 | 0.0031 | 4.0316 | 4.4415 | 4.6712 |
| | | | ENet | 39.005 | 0.0038 | 5.3108 | 5.5024 | 5.8112 |
| | | | SCAD | 40.00 | 0.0043 | 5.5542 | 5.4023 | 5.6641 |
| | | | MCP | 39.044 | 0.0038 | 5.4856 | 5.4115 | 5.6562 |

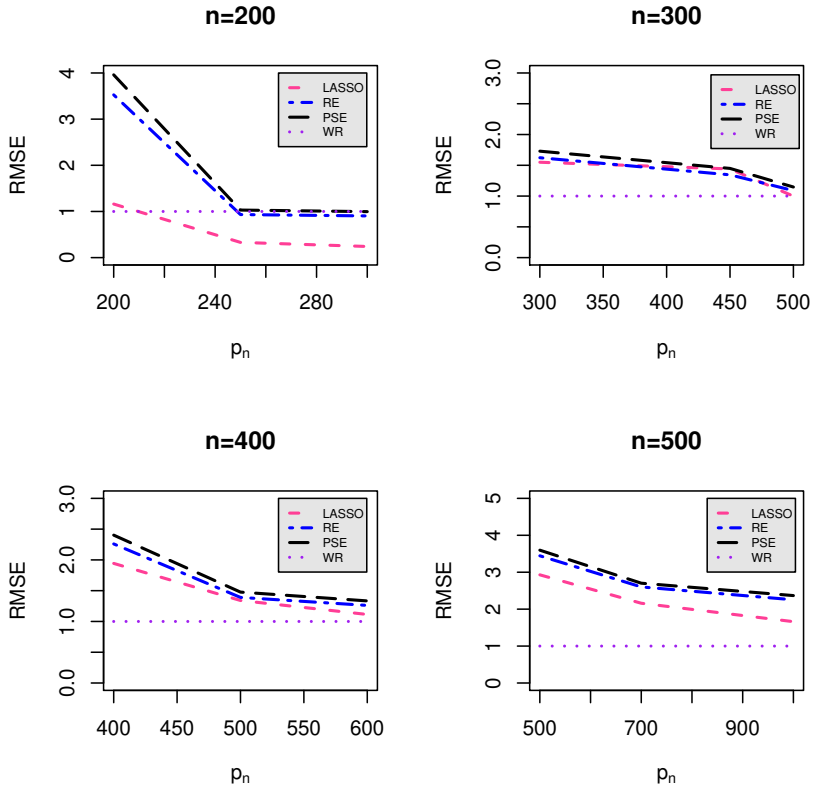


**Figure 1.** The RMSE of the proposed estimators compared with LASSO for different $n$ and $p_n$ in logistic model.

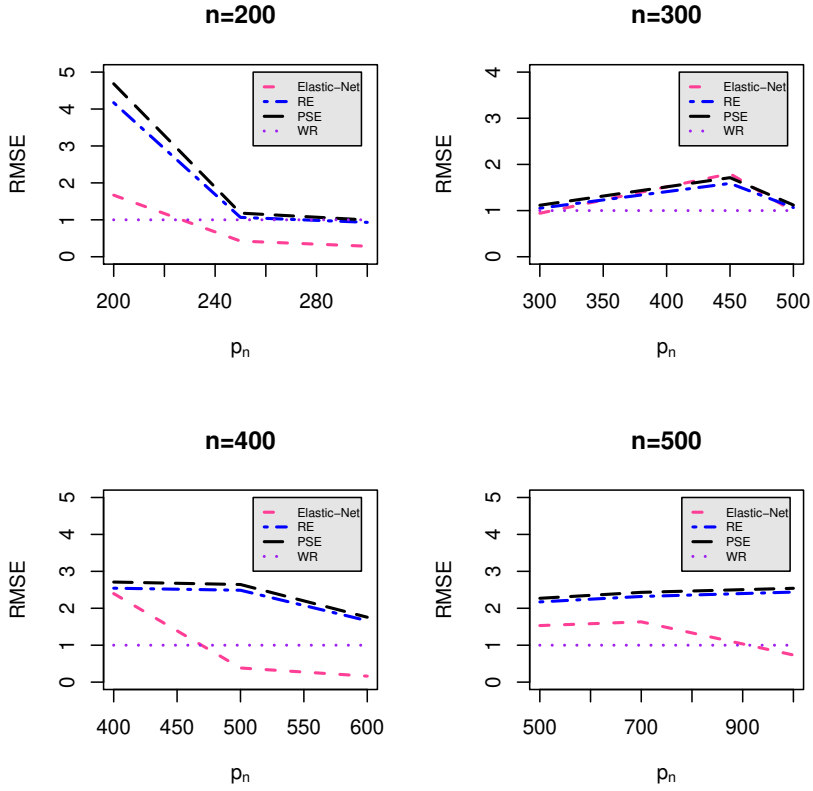


**Figure 2.** The RMSE of the proposed compared with ENet for different $n$ and $p_n$ in logistic model.

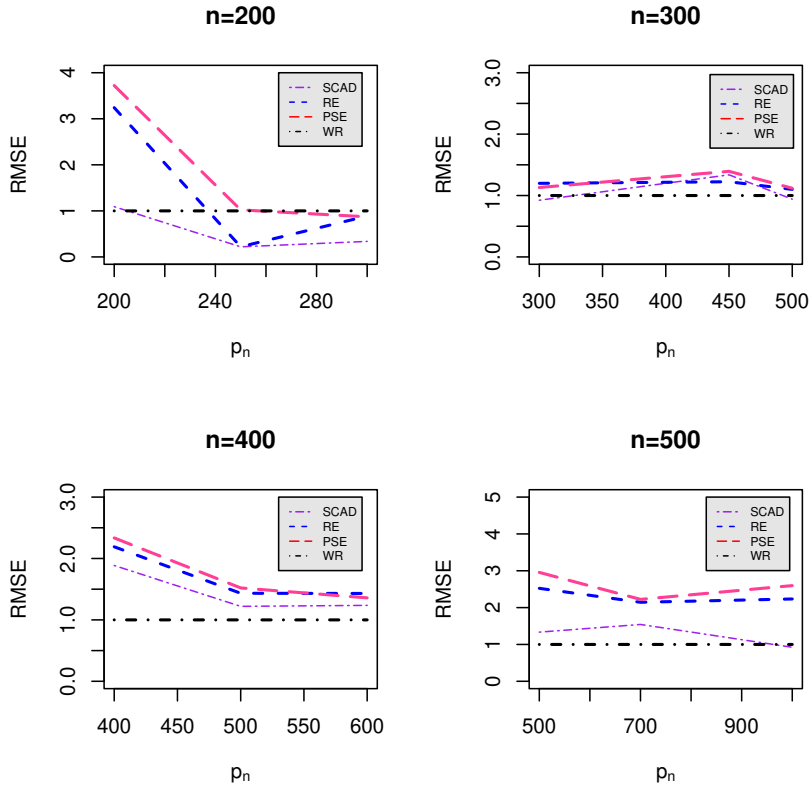


**Figure 3.** The RMSE of the proposed compared with SCAD for different $n$ and $p_n$ in logistic model.

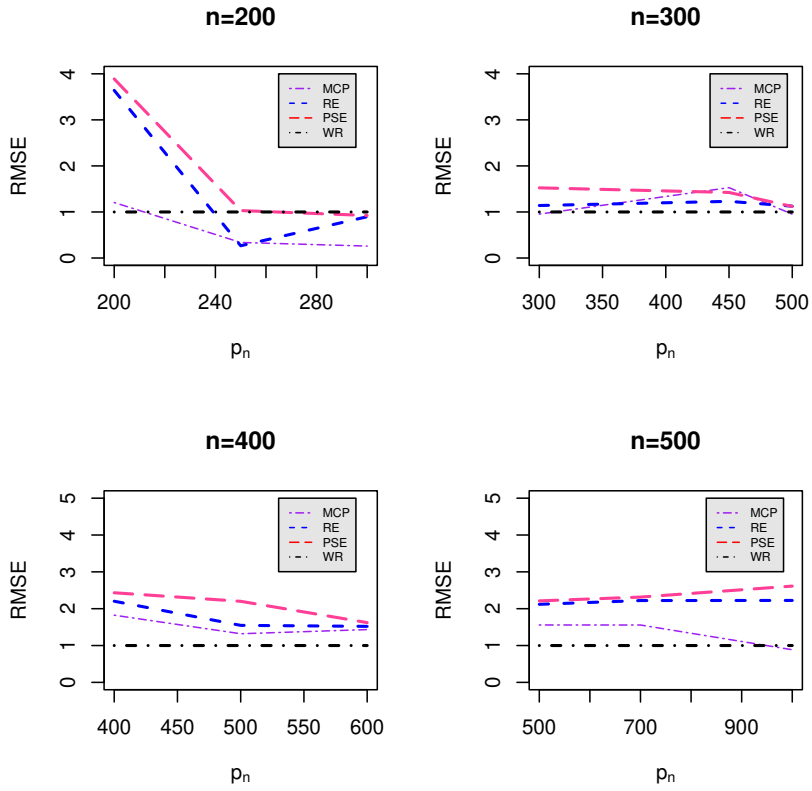


**Figure 4.** The RMSE of the proposed compared with MCP for different $n$ and $p_n$ in logistic model.

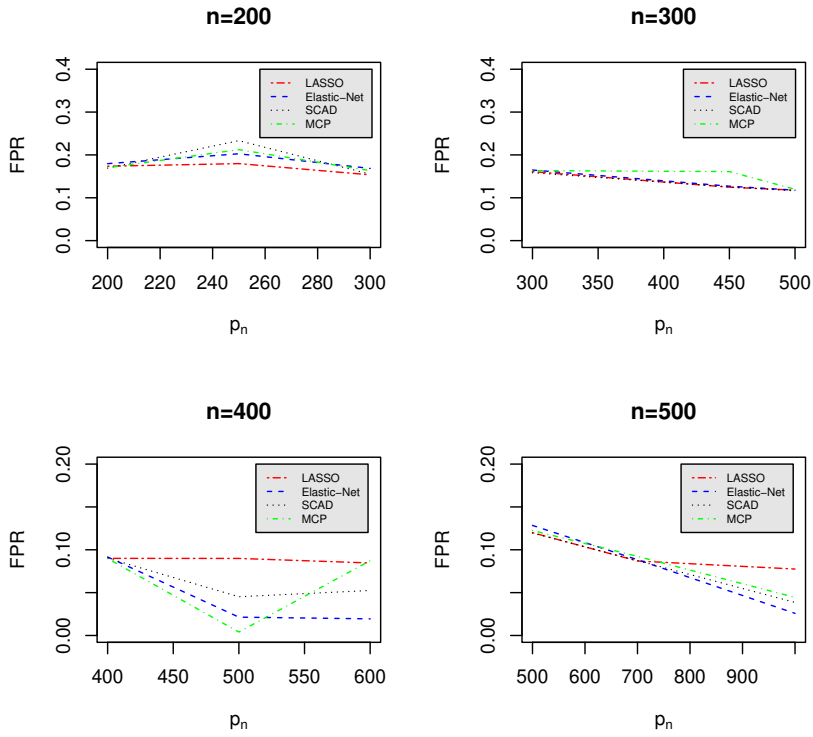


**Figure 5.** The FPR for LASSO, ENet, SCAD and MCP methods for different $n$ and $p_n$ in logistic model.

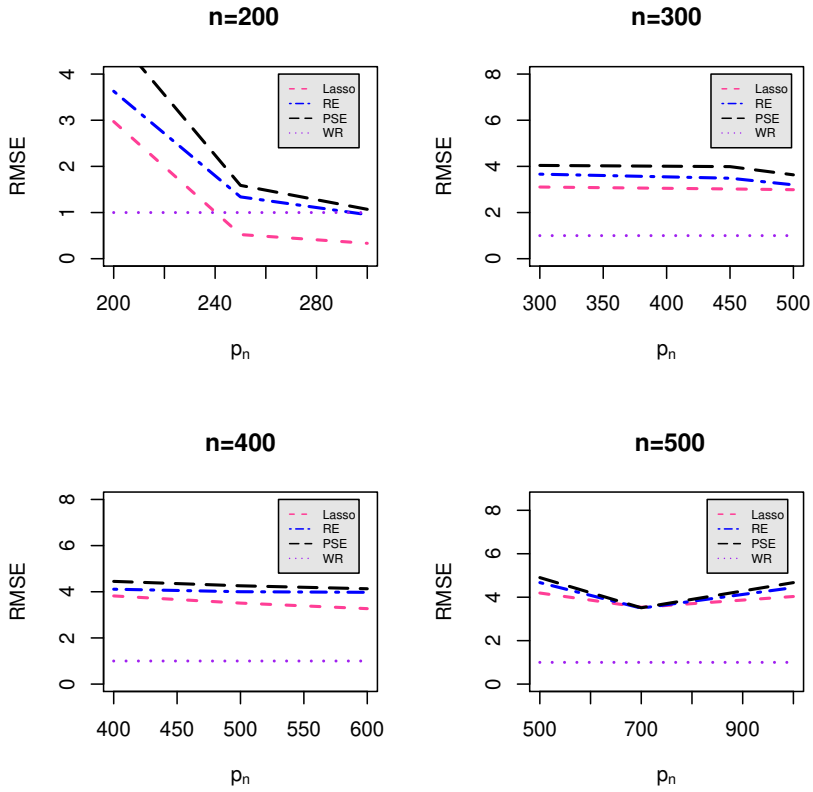


**Figure 6.** The RMSE of the proposed estimators compared with LASSO for different $n$ and $p_n$ in Poisson model.

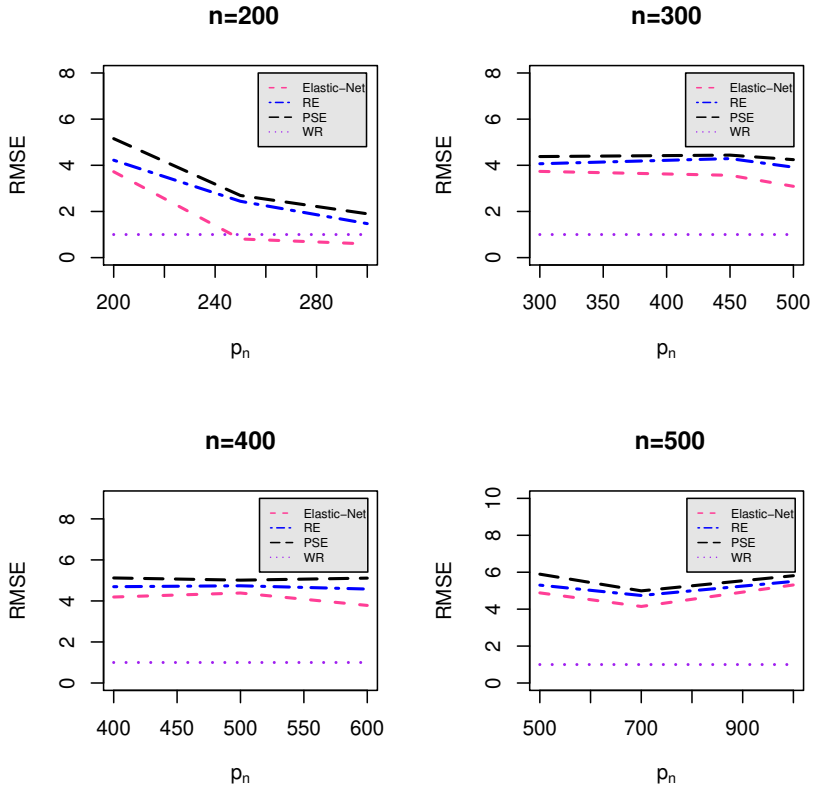


**Figure 7.** The RMSE of the proposed compared with ENet for different $n$ and $p_n$ in Poisson model.

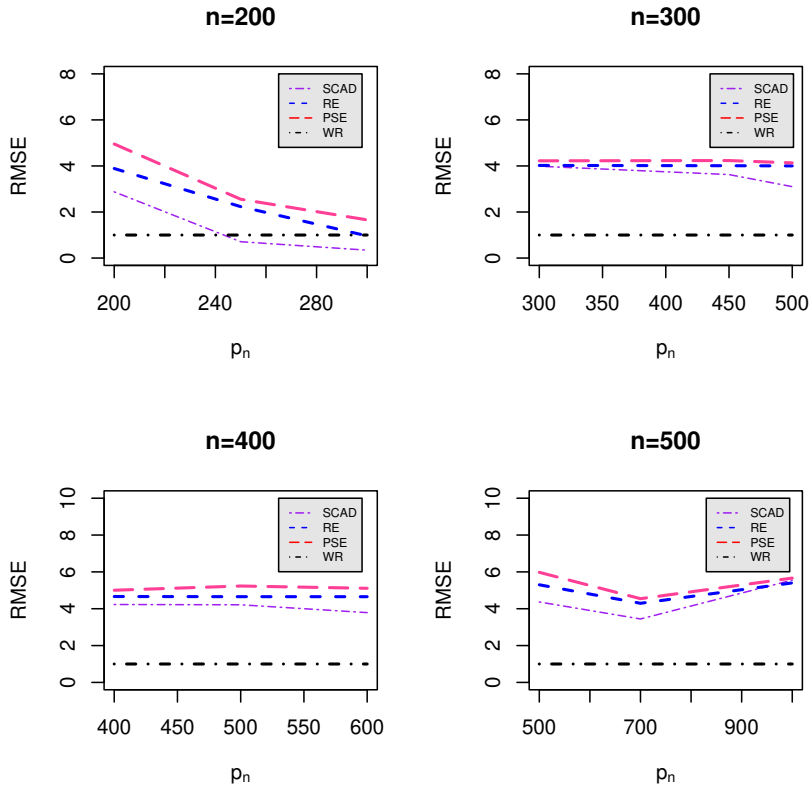


**Figure 8.** The RMSE of the proposed compared with SCAD for different $n$ and $p_n$ in Poisson model.

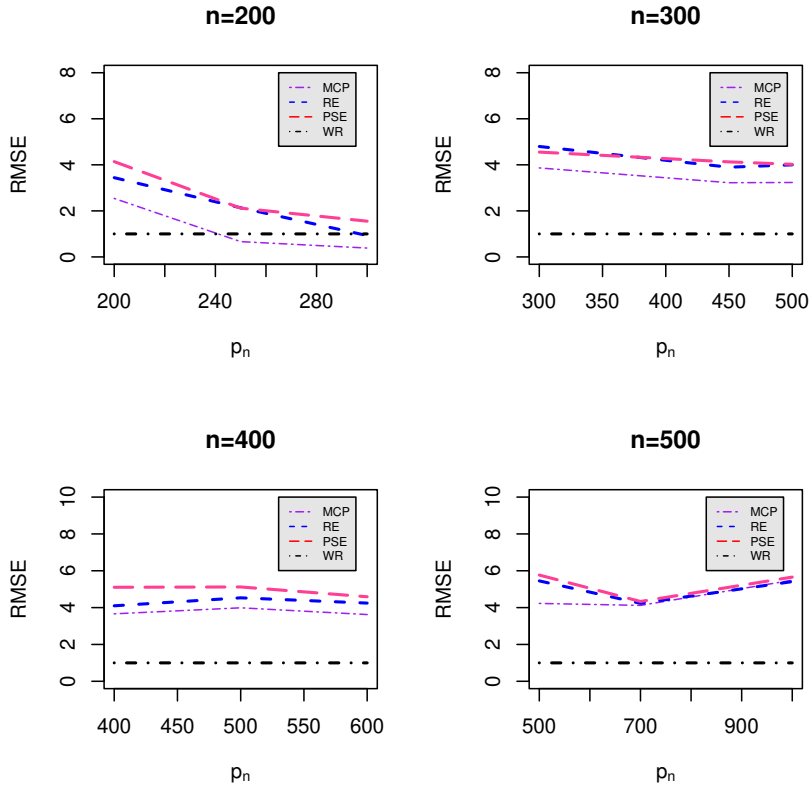


**Figure 9.** The RMSE of the proposed compared with MCP for different $n$ and $p_n$ in Poisson model.

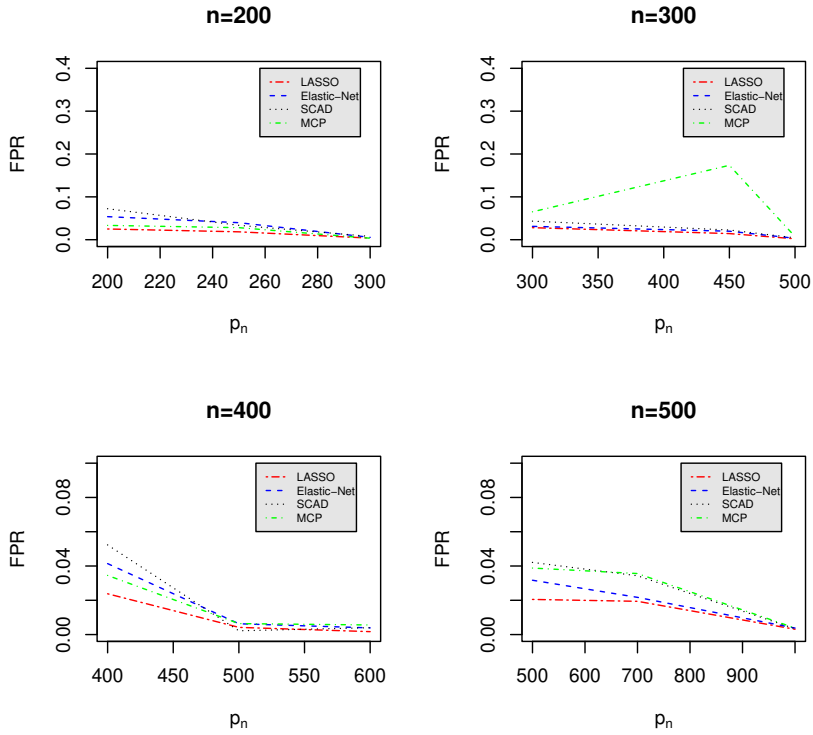


**Figure 10.** The FPR for LASSO, ENet, SCAD and MCP methods for different $n$ and $p_n$ in Poisson model.

## 5. Application

In this section, we investigate the practical usefulness of our methodology for gene expression studies which involve high-dimensional data.

### 5.1. Genome-wide studies in cancer

We apply our methodology to a real data set on genome-wide association (GWA) studies in cancer. The response variable, presence or absence of the illness is denoted by $Y \in \{0, 1\}$ (where $Y = 1$ denotes "diseased", and $Y = 0$ "healthy") and is binary, therefore we model the relationship using logistic regression. The linear logistic regression model is defined as follows:
$Y_1, \cdots, Y_n$ independent,

$$P(Y_i = 1|\boldsymbol{X}_i) = \pi(\boldsymbol{X}_i), \quad \log\Big(\frac{\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\Big) = \sum_{j=1}^{p_n} \beta_{0j} X_{ij}. \tag{5.1}$$

The dataset is ultra-high-dimensional given that it contains 102 observations (52 positive, 50 control) on 6033 genes and is available from the R package spls. To obtain the post-selection shrinkage estimators, we first select the candidate subsets from four variable selection approaches, namely, LASSO, ENet, SCAD and MCP, respectively. All tuning parameters are computed using 10-fold cross-validation. To evaluate the prediction accuracy of the listed estimators, we randomly divided the data into two parts: 70% of the dataset was designated as the training set and 30% was designated as the test set. Table 3 reports the selected number of genes for different values $p_n$. As seen, the ENet strategy selects too many predictors, which may yield an overfitted model, whereas SCAD and MCP select fewer substantial predictors, which may produce an under-fit model. We observe that the suggested post-shrinkage estimator outperforms both submodels and full models estimators in all cases. For these data, ENet performs relatively better than three selected penalized methods used to construct the post-shrinkage estimators, perhaps due to an inherited large amount of bias being more aggressive in variable selection. Interestingly, all three penalized methods are superior to LASSO. Nevertheless, the post-shrinkage estimator is utperforming the listed penalty estimators either we use ENet or LASSO to construct it.

The relative residual sum of squares (RRSS) of estimator $\boldsymbol{\beta}_{\mathcal{J}}^*$ over the WR estimator $\boldsymbol{\beta}_{0\mathcal{J}}^{\mathrm{WR}}$ is computed as follows:

$$\mathrm{RRSS}(\boldsymbol{\beta}_{\mathcal{J}}^*) = \frac{\sum_{i=1}^n ||y_i - \sum_{j\in\mathcal{J}} \boldsymbol{X}_{\mathcal{J}} \boldsymbol{\beta}_{0\mathcal{J}}^{\mathrm{WR}}||_2}{\sum_{i=1}^n ||y_i - \sum_{j\in\mathcal{J}} \boldsymbol{X}_{\mathcal{J}} \boldsymbol{\beta}_{\mathcal{J}}^*||_2}, \tag{5.2}$$

where $\mathcal{J}$ is the index of the subset model, and $\boldsymbol{\beta}_{\mathcal{J}}^*$ can be LASSO, ENet, SCAD, MCP or post-selection RE and selection PSEs. RRSS $> 1$ indicates the supremacy of $\boldsymbol{\beta}_{\mathcal{J}}^*$ over $\boldsymbol{\beta}_{0\mathcal{J}}^{\mathrm{WR}}$. We compute the RRSS values for the underlying estimators under subset chosen by LASSO, ENet, SCAD and MCP. The RRSS values are computed using cross-validation following 50 random partitions of the data set. In each partition, the training set consists of 60% observations and the test set consists of the remaining 40% observations. The results of RRSS for different $p_n$'s are reported in Figures 11 and 12. RRSS values of $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{PSE}}$ are observed to give the highest value in both cases. This is not surprising because the selected submodel generated by LASSO, ENet, SCAD or MCP is the right one and does not account for any bias. It is clear that in both cases $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_1}^{\mathrm{RE}}$ dominates the corresponding penalized estimators based on RRSS with $\hat{\mathcal{S}}_1$ detected by LASSO, ENet, SCAD, or MCP. The prediction error accuracy (PEA) of the data based on the LASSO, ENet, SCAD, and MCP methods is reported in Table 3. It can be seen that the PEA of LASSO and SCAD is

higher than that of the second method. Thus, the data analysis agrees with the simulation and theoretical findings.

**Table 3.** Results of classification & prediction error accuracy from real data set.

| $p_n$ | Method | Selected number of genes | PEA |
|---|---|---|---|
| 2050 | LASSO | 26 | 0.1878 |
| | ENet | 31 | 0.1633 |
| | SCAD | 27 | 0.0836 |
| | MCP | 22 | 0.2197 |
| 2500 | LASSO | 18 | 0.3523 |
| | ENet | 45 | 0.1241 |
| | SCAD | 28 | 0.0825 |
| | MCP | 17 | 0.1913 |
| 3550 | LASSO | 19 | 0.1183 |
| | ENet | 38 | 0.0195 |
| | SCAD | 29 | 0.1388 |
| | MCP | 21 | 0.1700 |
| 4300 | LASSO | 22 | 0.1247 |
| | ENet | 54 | 0.0571 |
| | SCAD | 32 | 0.1345 |
| | MCP | 25 | 0.2024 |
| 6033 | LASSO | 21 | 0.1078 |
| | ENet | 52 | 0.1300 |
| | SCAD | 24 | 0.1088 |
| | MCP | 33 | 0.1339 |

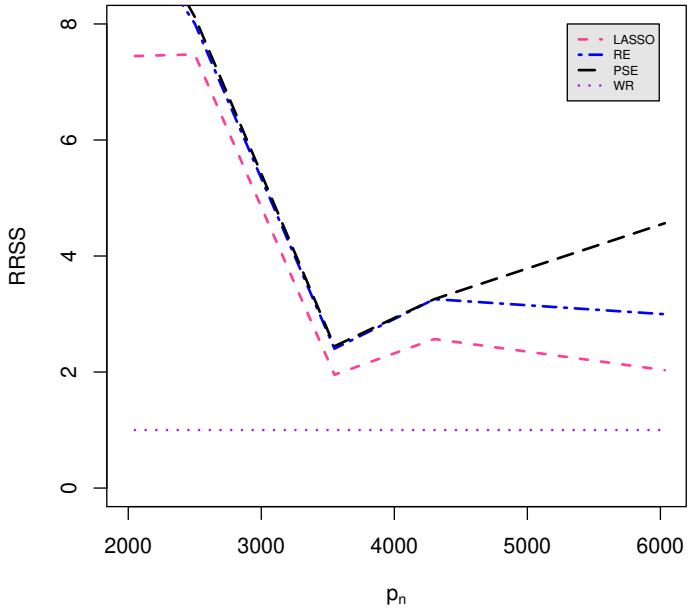


**Figure 11.** Relative residual sum of squares (RRSS) of the proposed estimators compared with LASSO for different $p_n$.

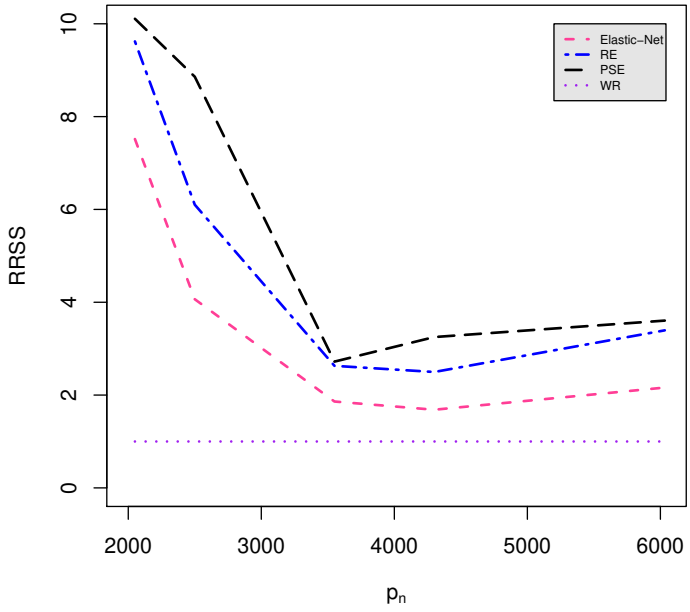


**Figure 12.** Relative residual sum of squares (RRSS) of the proposed estimators compared with Elastic-Net for different $p_n$.

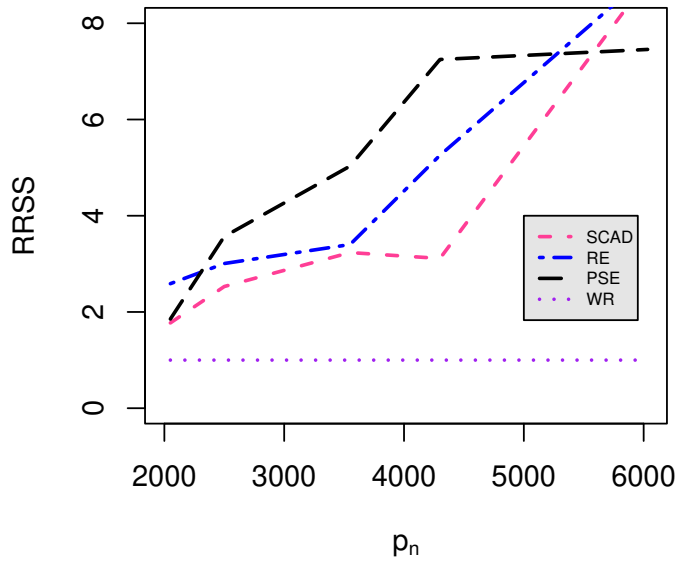


**Figure 13.** Relative residual sum of squares (RRSS) of the proposed estimators compared with SCAD for different $p_n$.

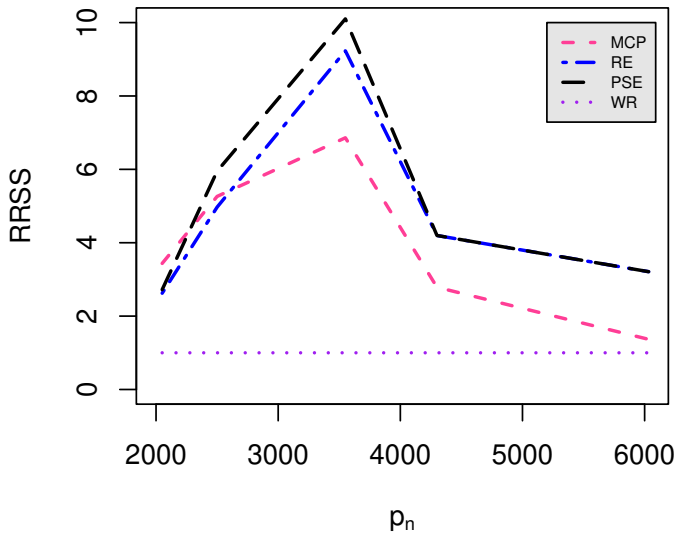


**Figure 14.** Relative residual sum of squares (RRSS) of the proposed estimators compared with MCP for different $p_n$.

## 6. Conclusion

For a GLM where the number of predictors is much larger than the number of observations, we offer a post-selection high-dimensional shrinkage estimation strategy. The asymptotic risk properties of the underlying estimators are developed and are evaluated with the risk of the subset candidate model, LASSO, ENet, SCAD and MCP estimators, respectively. We conclude that the suggested estimation strategy is very competitive with SCAD and MCP estimators and in many cases performs better. In particular, it performs very well when there are weak signals based on regression coefficients. The proposed strategy is intuitively appealing and can be easily realized. Theoretical and simulation results demonstrated that the post-selection shrinkage estimator has favorable performance and is a good alternative to LASSO, ENet, SCAD and MCP estimators. It can save the loss of efficiency of SCAD due to the effect of variable selection. When $p_n > n$, we are interested in recovering the support and nonzero components of the regression coefficient vector $\boldsymbol{\beta}$. As a powerful tool for producing interpretable models, sparse modeling via penalized regularization has gained popularity for analyzing high-dimensional data sets.

The proposed post-selection shrinkage estimator has a superior prediction performance over other penalized regression estimators. The penalized estimators do not allow other predictors to contribute once a sparse model or subset model is generated. The proposed post-selection shrinkage estimators inherit the advantage of Stein-type estimators and take into account possible contributions of other irrelevant variables.

This paper contributes to the investigation of post-penalized estimation analysis in a high-dimensional setting. In summary, we investigate the asymptotic normality of the WR estimator when $p_n$ increases with $n$. In addition, the relative efficiency of the proposed high-dimensional shrinkage estimator to its competitors is assessed analytically and numerically. We show that the performance of the shrinkage strategy is favorable relative to that of other estimators. Our Monte Carlo simulation studies suggest that post-selection shrinkage strategies perform better than penalty estimators for both variable selection

and prediction purposes in many instances. The reported results of the data set are consistent with our simulation study, demonstrating the superior performance of suggested post-selection shrinkage strategies.

In this work, we focus on the high-dimensional post-selection shrinkage estimation within the context of GLMs. For future work, one may consider combining all the estimators produced by multiple (more than two) variable selection techniques into a single estimator to improve the overall prediction error. In another study, it would be interesting to include penalty estimators that correspond to Bayes procedures based on priors with polynomial tails.

## Acknowledgements

The author would like to thank the Editor, associate Editor, and anonymous reviewer for helpful comments which led to improvements on an earlier version of the manuscript.

**Author contributions.** The co-author has contributed equally in all aspects of the preparation of this submission.

**Conflict of interest statement.** The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Funding.** No funding resource is available for this paper.

**Data availability.** Real dataset on "genome-wide association (GWA)" is available from the R package spls.

## References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control **19**, 716–723, 1974.

[2] A. Belloni, V. Chernozhukov and L. Wang, *Square-root lasso: Pivotal recovery of sparse signals via conic programming*, Biometrika **98**(4), 791–806, 2011.

[3] P. Bickel, Y. Ritov and A. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Ann. Stat. **37**, 1705–1732, 2009.

[4] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, Berlin, 2011.

[5] D.L. Donoho, *High-dimensional data analysis: the curses and blessings of dimensionality*, AMS Math Challenges Lecture, 1–32, 2000.

[6] E. Candès and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, Ann. Stat. **35**(6), 2313–2351, 2007.

[7] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Least angle regression*, Ann. Stat. **32**(2), 407–499, 2004.

[8] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. **96**(456), 1348–1360, 2001.

[9] J. Fan and R. Li, *Statistical challenges with high dimensionality: Feature selection in knowledge discovery*, Proc. Int. Cong. Math. **III**, 595–622, 2006.

[10] J. Fan and J. Lv, *A selective overview of variable selection in high dimensional feature space*, Stat. Sin. **20**(1), 101–148, 2010.

[11] J. Fan and R. Song, *Sure independence screening in generalized linear models with np-dimensionality*, Ann. Stat. **38**, 3567–3604, 2010.

[12] J. Fan and J. Lv, *Non-concave penalized likelihood with np-dimensionality*, IEEE Trans. Reliab. **57**, 5467–5484, 2011.

[13] X. Gao, S.E. Ahmed and Y. Feng, *Post selection shrinkage estimation for high-dimensional data analysis*, Appl. Stoch. Models Bus. Ind. **33**, 97–120, 2017.
[14] J. Huang, S.G. Ma and C.H. Zhang, *Adaptive Lasso for sparse high-dimensional regression models*, Stat. Sin. **18**, 1603–1618, 2008.
[15] S.F. Kurnaz, I. Hoffmann and P. Filzmoser, *Robust and sparse estimation methods for high-dimensional linear and logistic regression*, J. Chemom. **31**(1112), e2936, 2017.
[16] S. Kwon and Y. Kim, *Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions*, Stat. Sin. **22**, 629–653, 2012.
[17] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, London, 1989.
[18] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the Lasso*, Ann. Stat. **34**, 1436–1462, 2006.
[19] D.M. Sakate and D.N. Kashid, *Variable selection via penalized minimum u-divergence estimation in logistic regression*, J. Appl. Stat. **41**(6), 1233–1246, 2014.
[20] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. **6**, 461–464, 1978.
[21] T. Sun and C.H. Zhang, *Scaled sparse linear regression*, Biometrika **99**(4), 879–898, 2012.
[22] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. R. Stat. Soc. Ser. B **58**(1), 267–288, 1996.
[23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, *Sparsity and smoothness via the fused Lasso*, J. R. Stat. Soc. Ser. B **67**(1), 91–108, 2005.
[24] H. Wang, R. Li and C.L. Tsai, *Tuning parameter selectors for the smoothly clipped absolute deviation method*, Biometrika **94**(3), 553–568, 2007.
[25] M. Wang, L. Song and X. Wang, *Bridge estimation for generalized linear models with a diverging number of parameters*, Stat. Probab. Lett. **80**, 1584–1596, 2010.
[26] X. Wang and M. Wang, *Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure*, J. Appl. Stat. **43**(5), 796–809, 2016.
[27] P. Zhao and B. Yu, *On model selection consistency of Lasso*, J. Mach. Learn. Res. **7**(11), 2541–2563, 2006.
[28] H. Zou, *The adaptive Lasso and its oracle properties*, J. Am. Stat. Assoc. **101**(476), 1418–1429, 2006.
[29] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B **67**(2), 301–320, 2005.

## APPENDIX

The technical proofs of the Theorems 3.4 and 3.5 are included in this section.

**Proof of Theorem 3.4**
Here, we provide the proof of the ADB expressions of the proposed estimators. Based on Theorem 3.2, it is clear that

$$\lim_{n\to\infty} E[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1)] = E[\lim_{n\to\infty}\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1)\}] = E[\mathcal{Z}] = 0,$$

where $\mathcal{Z} \sim \mathcal{N}(0,1)$. Then,

$$\begin{aligned}
ADB(\boldsymbol{d}_{1n}^{\intercal}\hat{\boldsymbol{\beta}}_{1n}^{RE}) &= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{RE}-\boldsymbol{\beta}_1\Big)\Big] \\
&= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big\{(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1)-(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE})\Big\}\Big] \\
&= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\Big)\Big] - \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\Big] \\
&= ADB(\boldsymbol{d}_{1n}^{\intercal}\hat{\boldsymbol{\beta}}_{1n}^{WR}) - \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\Big] \\
&= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{2n}^{\intercal}\hat{\boldsymbol{\beta}}_{2n}^{WR}\Big] \\
&= \lim_{n\to\infty} (s_{2n}/s_{1n}) E\Big[n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\intercal}\hat{\boldsymbol{\beta}}_{2n}^{WR}\Big] \\
&= (s_2/s_1)s_2^{-1}\boldsymbol{d}_2^{\intercal}\boldsymbol{\beta}_2 = s_1^{-1}\boldsymbol{d}_2^{\intercal}\boldsymbol{\beta}_2,
\end{aligned}$$

where $\boldsymbol{d}_{2n} = \hat{\boldsymbol{\Sigma}}_{\mathbb{S}_2\mathbb{S}_1}\hat{\boldsymbol{\Sigma}}_{\mathbb{S}_1}^{-1}\boldsymbol{d}_{1n}$, $\boldsymbol{d}_{1n}^{\intercal}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}) = -\boldsymbol{d}_1^{\intercal}\hat{\boldsymbol{\Sigma}}_{\mathbb{S}_1}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbb{S}_2\mathbb{S}_1}\hat{\boldsymbol{\beta}}_{2n}^{WR} = -\boldsymbol{d}_{2n}^{\intercal}\hat{\boldsymbol{\beta}}_{2n}^{WR}$ and $s_{2n}^2 = \boldsymbol{d}_2^{\intercal}\hat{\boldsymbol{\Sigma}}_{\mathbb{S}_2|\mathbb{S}_1}^{-1}\boldsymbol{d}_{2n}$.

Now, we compute the ADB of $\hat{\boldsymbol{\beta}}_{1n}^{SE}$ as follows

$$\begin{aligned}
ADB(\boldsymbol{d}_{1n}^{\intercal}\hat{\boldsymbol{\beta}}_{1n}^{SE}) &= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)\Big] \\
&= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-[(p_{2n}-2)\hat{T}_n^{-1}](\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE})-\boldsymbol{\beta}_1\Big)\Big] \\
&= \lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\Big)\Big] \\
&\quad - \lim_{n\to\infty} (p_{2n}-2)E\Big[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_1^{RE}\Big)\hat{T}_n^{-1}\Big] \\
&= E[\mathcal{Z}] - (p_2-2)E\Big[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\intercal}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE})T_n^{-1}\Big\}\Big] \\
&= (p_2-2)E\Big[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{2n}^{\intercal}\hat{\boldsymbol{\beta}}_{2n}^{WR}T_n^{-1}\Big\}\Big] \\
&= (p_2-2)(s_2/s_1)E\Big[\lim_{n\to\infty}\Big\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\intercal}\hat{\boldsymbol{\beta}}_{2n}^{WR}T_n^{-1}\Big\}\Big] \\
&= (p_2-2)s_1^{-1}\boldsymbol{d}_2^{\intercal}\boldsymbol{\beta}_2 E\Big[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Big].
\end{aligned}$$

Finally, we obtain ADB of $\hat{\boldsymbol{\beta}}_{1n}^{PSE}$,

$$
\begin{aligned}
ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) &= \lim_{n\to\infty} E\Bigg[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{PSE}-\boldsymbol{\beta}_1\Big)\Bigg] \\
&= \lim_{n\to\infty} E\Bigg[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big\{\hat{\boldsymbol{\beta}}_{1n}^{SE}+[1-(p_{2n}-2)\hat{T}_n^{-1}](\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}) \\
&\quad \times I(\hat{T}_n<(p_{2n}-2))-\boldsymbol{\beta}_1\Big\}\Bigg] \\
&= \lim_{n\to\infty} E\Bigg[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)\Bigg] \\
&\quad + \lim_{n\to\infty} E\Bigg[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}[1-(p_{2n}-2)\hat{T}_n^{-1}]\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)I(\hat{T}_n<(p_{2n}-2))\Bigg] \\
&\quad + E\Bigg[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1)\Big\}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+ E\Bigg[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE})I(\hat{T}_n<(p_{2n}-2))\Big\}\Bigg] \\
&- (p_2-2)E\Bigg[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE})T_n^{-1}I(\hat{T}_n<(p_{2n}-2))\Big\}\Bigg] \\
&= ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) - E\Bigg[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}I(\hat{T}_n<(p_{2n}-2))\Big\}\Bigg] \\
&+ (p_2-2)E\Bigg[\lim_{n\to\infty}\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\Big\}\Bigg] \\
&= ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) - (s_2/s_1)E\Bigg[\mathcal{Z}I\big(\chi_{P_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2)\big)\Bigg] \\
&- s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2}) \\
&+ (p_2-2)(s_2/s_1)E\Bigg[\mathcal{Z}\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I\big(\chi_{P_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2)\big)\Bigg] \\
&+ (p_2-2)s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 E\Bigg[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I\big(\chi_{P_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2)\big)\Bigg] \\
&= ADB(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) - s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2}) \\
&+ (p_2-2)s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2 E\Bigg[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I\big(\chi_{p_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2)\big)\Bigg] \\
&= s_1^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2\Bigg[(p_2-2)\Big\{E[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]+E[\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I\big(\chi_{P_2}^2(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2)\big)]\Big\} \\
&- H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Bigg]
\end{aligned}
$$

and the proof is completed. □

**Proof of Theorem 3.5**

Here, we provide the proof of the ADR expressions of the proposed estimators. It is clear that

$$\lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)\right\}^2\right] = E\left[\lim_{n\to\infty}\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)\right\}^2\right] = E[\mathcal{Z}^2] = 1,$$

$$\begin{aligned}
ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{RE}) &= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{RE}-\boldsymbol{\beta}_1\right)\right\}^2\right] \\
&= \lim_{n\to\infty} s_{1n}^{-2}E\left[\left\{n^{1/2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left[\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)-\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\right]\right\}^2\right] \\
&= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)\right\}^2\right] \\
&+ \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\right\}^2\right] \\
&- 2\lim_{n\to\infty} E\left[ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)^{\mathsf{T}}\boldsymbol{d}_{1n}\right] \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

From (3.11), we have $I_1 = \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1)\right\}^2\right] = 1$. Also,

$$\begin{aligned}
I_2 &= \lim_{n\to\infty} s_{1n}^{-2}E\left[n^{1/2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\right]^2 \\
&= \lim_{n\to\infty}(s_{2n}^2/s_{1n}^2)E\left[n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right]^2.
\end{aligned}$$

Since $s_{2n}^2/s_{1n}^2 \to 1-c$, then

$$I_2 = (1-c)\lim_{n\to\infty} E\left[\chi_1^2(\boldsymbol{\Delta}_{d_{2n}})\right] = (1-c)(1+2\boldsymbol{\Delta}_{d_2}).$$

Furthermore,

$$\begin{aligned}
I_3 &= -2\lim_{n\to\infty} E\left[ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)^{\mathsf{T}}\boldsymbol{d}_{1n}\right] \\
&= 2\lim_{n\to\infty}(s_{2n}/s_{1n})E\left[n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}n^{1/2}s_{1n}^{-1}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)^{\mathsf{T}}\boldsymbol{d}_{1n}\right] \\
&= 2(1-c)^{1/2}.
\end{aligned}$$

Now, we investigate (3.13). By using Eq. (3.5), we have

$$\begin{aligned}
ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) &= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1)\right\}^2\right] \\
&= \lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big\{\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\big)-[(p_{2n}-2)/\hat{T}_n]\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)\Big\}\right]^2 \\
&= \lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\big)\right]^2 \\
&+ \lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}[(p_{2n}-2)T_n^{-1}]\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)\right]^2 \\
&- 2\lim_{n\to\infty} E\left[ns_{1n}^{-2}[(p_{2n}-2)\hat{T}_n^{-1}]\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\big)^{\mathsf{T}}\boldsymbol{d}_{1n}\right] \\
&= J_1+J_2+J_3.
\end{aligned}$$

Again, $J_1 = \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\big)\right\}^2\right] = 1$. Then, we have

$$\begin{aligned}
J_2 &= \lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}[(p_{2n}-2)\hat{T}_n^{-1}]\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)\right]^2 \\
&= \lim_{n\to\infty}(p_{2n}-1)^2E\left[n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}T_n^{-1}\right]^2 \\
&= (s_2^2/s_1^2)(p_2-1)^2E\left[\lim_{n\to\infty}\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\hat{T}_n^{-1}\right\}\right]^2 \\
&= (s_2^2/s_1^2)(p_2-1)^2\Big\{E[\mathcal{Z}^2\chi_{p_2}^{-4}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2E[\chi_{p_2}^{-4}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]\Big\} \\
&= (1-c)(p_2-2)^2\Big\{E[\chi_{p_2+2}^{-4}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2E[\chi_{p_2}^{-4}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})]\Big\},
\end{aligned}$$

and

$$\begin{aligned}
J_3 &= -2\lim_{n\to\infty} E\left[ns_{1n}^{-2}[(p_{2n}-2)\hat{T}_n^{-1}]\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\big)^{\mathsf{T}}\boldsymbol{d}_{1n}\right] \\
&= 2\lim_{n\to\infty}(s_{2n}/s_{1n})(p_{2n}-2)E\left[n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}s_{1n}^{-1}(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1)^{\mathsf{T}}\hat{T}_n^{-1}\right] \\
&= 2(1-c)^{1/2}(p_2-2)\left\{E\Big[\mathcal{Z}^2\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Big]+s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2E\Big[\mathcal{Z}\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Big]\right\} \\
&= 2(1-c)^{1/2}(p_2-2)E\Big[\chi_{p_2+2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})\Big].
\end{aligned}$$

$$
\begin{aligned}
ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{PSE}) &= \lim_{n\to\infty} E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{PSE}-\boldsymbol{\beta}_1\Big)\Big\}^2\Big]\\
&= \lim_{n\to\infty} E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big[\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)\\
&\quad+[1-(p_{2n}-2)\hat{T}_n^{-1}]\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)I(\hat{T}_n<(p_{2n}-2))\Big]\Big\}^2\Big]\\
&= \lim_{n\to\infty} E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)\Big\}^2\Big]\\
&+ \lim_{n\to\infty} E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}[1-(p_{2n}-2)\hat{T}_n^{-1}]\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)I(\hat{T}_n<(p_{2n}-2))\Big\}^2\Big]\\
&+ 2\lim_{n\to\infty} E\Big[n^{1/2}s_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}[1-(p_{2n}-2)\hat{T}_n^{-1}]\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)^{\mathsf{T}}\\
&\quad\times I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\Big]\\
&= ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) + \lim_{n\to\infty} E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)I(\hat{T}_n<(p_{2n}-2))\Big\}^2\Big]\\
&+ \lim_{n\to\infty}(p_{2n}-2)^2E\Big[\Big\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\Big\}^2\Big]\\
&- 2\lim_{n\to\infty}(p_{2n}-2)E\Big[\Big\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)^2\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\Big\}\Big]\\
&+ 2\lim_{n\to\infty} E\Big[\Big\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)^{\mathsf{T}}I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\Big\}\Big]\\
&- 2\lim_{n\to\infty}(p_{2n}-2)E\Big[\Big\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\Big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\Big)\Big(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\Big)^{\mathsf{T}}\\
&\quad\times\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\Big\}\Big]\\
&= ADR(\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{1n}^{SE}) + D_1 + D_2 + D_3 + D_4 + D_5,
\end{aligned}
$$

$$
\begin{aligned}
D_1 &= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= \lim_{n\to\infty}(s_{2n}/s_{1n})^2E\left[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= (s_2/s_1)^2\left\{E\left[\mathcal{Z}^2I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2})\right\}\\
&= (1-c)\left\{E\left[\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})\right]+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2H_{p_2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2})\right\},\\
D_2 &= \lim_{n\to\infty} E\left[\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)(p_{2n}-2)\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= \lim_{n\to\infty}(p_{2n}-2)^2(s_{1n}/s_{2n})^2E\left[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= (p_2-2)^2(s_2/s_1)^2E\left[\lim_{n\to\infty}\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^2\right]\\
&= (p_2-2)^2(1-c)E\left[\mathcal{Z}^2\chi^{-4}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\\
&= (p_2-2)^2(1-c)E\left[\chi^{-4}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right],\\
D_3 &= -2\lim_{n\to\infty}(p_{2n}-2)E\left[\left\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\big(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\big)^2\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\right\}\right]\\
&= -2(p_2-2)(s_2/s_1)^2E\left[\lim_{n\to\infty}\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right\}^2\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right]\\
&= -2(p_2-2)(1-c)\left\{E\left[\mathcal{Z}^2\chi^{-2}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right.\\
&\quad\left.+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2E\left[\chi^{-2}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right\}\\
&= -2(p_2-2)(1-c)\left\{E\left[\chi^{-2}_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right.\\
&\quad\left.+(s_2^{-1}\boldsymbol{d}_2^{\mathsf{T}}\boldsymbol{\beta}_2)^2E\left[\chi^{-2}_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right\},
\end{aligned}
$$

$$
\begin{aligned}
D_4 &= 2\lim_{n\to\infty} E\left[\left\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\right)^{\mathsf{T}} I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\right\}\right]\\
&= 2\lim_{n\to\infty}(s_{2n}/s_{1n})E\left[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right\}\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\right)\right\}^{\mathsf{T}} I(\hat{T}_n<(p_{2n}-2))\right]\\
&= 2(s_2/s_1)E\Bigg[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right\}\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)\right.\\
&\quad\left.-(p_{2n}-2)\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^{\mathsf{T}} I(\hat{T}_n<(p_{2n}-2))\Bigg]\\
&= 2(1-c)^{1/2}\Bigg\{E\left[\lim_{n\to\infty}\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right\}\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)I(\hat{T}_n<(p_{2n}-2))\right\}^{\mathsf{T}}\right]\\
&\quad-(p_2-2)E\left[\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\right\}\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^{\mathsf{T}}\right]\Bigg\}\\
&= 2(1-c)^{1/2}\left\{E\left[\mathcal{Z}^2 I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]-(p_2-2)E\left[\mathcal{Z}^2\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right\}\\
&= 2(1-c)^{1/2}\left\{H_{p_2+2}(p_2-2;\boldsymbol{\Delta}_{\boldsymbol{d}_2})-(p_2-2)E\left[\chi_{p_2+2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
D_5 &= -2\lim_{n\to\infty}(p_{2n}-2)E\left[\left\{ns_{1n}^{-2}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{SE}-\boldsymbol{\beta}_1\right)^{\mathsf{T}}\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\boldsymbol{d}_{1n}\right\}\right]\\
&= 2(p_2-2)(s_2/s_1)E\Bigg[\lim_{n\to\infty}\left\{n^{1/2}s_{2n}^{-1}\boldsymbol{d}_{2n}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{2n}^{WR}\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}\\
&\quad\times\left\{n^{1/2}s_{1n}^{-1}\boldsymbol{d}_{1n}^{\mathsf{T}}\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\boldsymbol{\beta}_1\right)-(p_{2n}-2)\left(\hat{\boldsymbol{\beta}}_{1n}^{WR}-\hat{\boldsymbol{\beta}}_{1n}^{RE}\right)\hat{T}_n^{-1}I(\hat{T}_n<(p_{2n}-2))\right\}^{\mathsf{T}}\Bigg]\\
&= 2(p_2-2)(s_2/s_1)\Bigg\{E\left[\mathcal{Z}^2\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\\
&\quad-(p_2-2)E\left[\mathcal{Z}^2\chi_{p_2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\Bigg\}\\
&= 2(p_2-2)(1-c)^{1/2}\Bigg\{E\left[\chi_{p_2+2}^{-2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\\
&\quad-(p_2-2)E\left[\chi_{p_2+2}^{-4}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})I(\chi^2_{p_2+2}(\boldsymbol{\Delta}_{\boldsymbol{d}_2})<(p_2-2))\right]\Bigg\}.
\end{aligned}
$$

□