Machine Learning (2023) 112:3361–3411
https://doi.org/10.1007/s10994-023-06349-2

Robust estimation in regression and classification methods for large dimensional data

Chunming Zhang[1] Lixing Zhu[2,3] Yanbo Shen[1]

Received: 10 December 2021 / Revised: 26 March 2023 / Accepted: 8 May 2023 /
Published online: 5 July 2023

Abstract
Statistical data analysis and machine learning heavily rely on error measures for regression, classification, and forecasting. Bregman divergence (BD) is a widely used family of error measures, but it is not robust to outlying observations or high leverage points in large- and high-dimensional datasets. In this paper, we propose a new family of robust Bregman divergences called "*robust*-BD" that are less sensitive to data outliers. We explore their suitability for sparse large-dimensional regression models with incompletely specified response variable distributions and propose a new estimate called the "*penalized robust*-BD *estimate*" that achieves the same oracle property as ordinary non-robust penalized least-squares and penalized-likelihood estimates. We conduct extensive numerical experiments to evaluate the performance of the proposed penalized robust-BD estimate and compare it with classical approaches, and show that our proposed method improves on existing approaches. Finally, we analyze a real dataset to illustrate the practicality of our proposed method. Our findings suggest that the proposed method can be a useful tool for robust statistical data analysis and machine learning in the presence of outliers and large-dimensional data.

Keywords Coordinate descent · Hypothesis testing · Loss functions · Penalization · Quasi-likelihood

Editors: Krzysztof Dembczynski and Emilie Devijver.

✉ Chunming Zhang
czhang3@wisc.edu

Lixing Zhu
lzhu@bnu.edu.cn

Yanbo Shen
yshen84@wisc.edu

[1] University of Wisconsin-Madison, Madison, USA

[2] Center for Statistics and Data Science, Beijing Normal University at Zhuhai, Zhuhai City, P.R. China

[3] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

1 Introduction

Advancements in high-throughput technologies have made it possible to collect sophisticated, high-dimensional datasets, such as microarray data, genome-wide human SNP data, high-frequency financial data, functional data, and brain imaging data. In comparison to conventional datasets, where there are fewer variables than observations, large-dimensional datasets involve variables that could be as many as, or even more than, observations. These scenarios correspond to $p_n \ll n$, $p_n \approx n$, and $p_n \gg n$, respectively, with p_n being the number of variables and n being the number of observations. A fundamental problem common to any large-dimensional dataset is that observations are more prone to outliers in either the covariate space or the response space than those in low-dimensional datasets. In such settings, outliers can lead to possible erroneous conclusions concerning statistical estimation, but the mechanism of contamination can be quite complex and intractable in general. It can also be difficult or even impossible to spot outliers in large-dimensional or highly structured data. Hence, exploring and developing robust statistical estimation and inference procedures that are resistant to outliers in large-dimensional data becomes increasingly important.

In recent years, much attention has been focused on developing penalized estimates of parameters in regression models with a large number of parameters. Examples include the Lasso (Tibshirani, 1996), the SCAD (Fan & Peng, 2004), the adaptive Lasso (Zou, 2006), the Dantzig selector (Candes & Tao, 2007), and the group Lasso (Meier et al., 2008), among others. These penalized least-squares estimates or penalized-likelihood estimates are obtained using a quadratic loss function or require full knowledge of the likelihood function. The penalized "classical-BD" estimation was proposed in Zhang et al. (2010) for $p_n \ll n$ and $p_n \approx n$, where the error measure BD includes the quadratic loss, the exponential family of distributions, the (negative) quasi-likelihood, and many others as special cases. However, these non-robust estimates do not handle outlying observations. General tools for investigating the robustness properties of penalized estimates, especially when $p_n \approx n$ and $p_n \gg n$, seem to be much less developed.

It is well-known that the influence functions of classical (non-penalized) regression estimates based on the quadratic loss function and likelihood are unbounded. Large deviations of the response from its mean, as measured by the Pearson residuals, or outlying points in the covariate space, can have a significant influence on the estimates. While robust procedures in Bianco et al. (1996), Künsch et al. (1989), Stefanski et al. (1986) control outliers for the generalized linear model (GLM), these procedures are limited to finite- and low-dimensional problems. It remains unclear to what extent they are useful in large- and high-dimensional settings. The works (Boente et al., 2006; Cantoni & Ronchetti, 2001) developed robust quasi-likelihood estimates of finite-dimensional parameters. However, the robust quasi-likelihood procedure is not available for other types of error measures, such as the hinge loss for the support vector machine (SVM) (Vapnik, 1996) and the exponential loss for AdaBoost (Freund & Schapire, 1997), which are commonly used in classification procedures and machine learning practice. This is because these error measures do ***not*** fall into the (negative) quasi-likelihood category.

This paper aims to investigate the applicability of robust statistical inference for regression estimation and classification procedures in large-dimensional ($p_n \approx n$) and high-dimensional ($p_n \gg n$) settings, where the distribution of the response variable given covariates may be incompletely specified. The proposed work is not a simple endeavor and does not aim to solve all possible issues stemming from the combination of the "*robustness* *property coupled with large-dimensionality*". The paper identifies major challenges and presents new results below.

- In Sect. 2, we contribute to constructing a new class of robust error measures called "*robust*-BD". This is motivated by Bregman divergence (BD), which plays an important role in quantifying error measures for regression estimates and classification procedures. The quadratic loss function and negative quasi-likelihood are two widely used error measures that, along with many others, belong to the family of BD. This newly proposed "*robust*-BD" method broadens the scope of penalized estimation methods, greatly facilitating the investigation of their asymptotic behavior in a systematic way. The new method is applicable to all aforementioned error measures (e.g., the hinge loss and exponential loss, which fail to be (negative) quasi-likelihood but belong to BD). The "*robust*-BD" benefits from the flexibility and extensibility offered by BD. Nonetheless, unlike the "classical-BD", the "*robust*-BD" entails a bias-correction procedure that complicates theoretical derivations as well as practical implementations; see concrete examples in Sect. 2.3. Moreover, when $p_n < n$, justifying the influence function of a "*robust*-BD" estimate for a p_n-dimensional parameter calls for re-examination and re-derivation beyond the framework of Hampel et al. (1986), confined to a fixed-dimensional parameter.
- In Sects. 3 and 4, we study the consistency and oracle property of the proposed "*penalized robust*-BD *estimates*" in $p_n \approx n$ and $p_n \gg n$ settings, respectively. It is shown that the new estimate, combined with an appropriately weighted L_1 penalty, achieves the same oracle property as the ordinary non-robust penalized least-squares and penalized-likelihood estimates, but is less sensitive to outliers, a very desirable property in many applications. It will be seen that the robust counterpart eliminates the finiteness assumption of some higher-order moments of the response variable Y, typically assumed in the non-robust case. Nonetheless, dealing with large-dimensionality in the robust case will face many more theoretical and computational challenges than in the non-robust case. For example, the oracle property in the case $p_n \approx n$ could not be directly extended to the case $p_n \gg n$ without requiring more technical assumptions and invoking more careful mathematical treatments. For practical implementation, two data-driven selection procedures for penalty weights, PMR and MR, will be proposed and justified for both $p_n \approx n$ and $p_n \gg n$. Unlike the selection method in Zhang et al. (2010) for penalty weights which deals with $p_n \approx n$ and requires $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$ for the p_n-dimensional predictor vector $\boldsymbol{X} = (X_1, \ldots, X_{p_n})^T$, both the methods and theory developed in the current work do not impose such a requirement, thus are more widely applicable. In the context of large-dimensional inference, we demonstrate that the Wald-type test statistic based on the "*penalized robust*-BD *estimates*" will be asymptotically distribution-free, whereas the likelihood ratio-type test statistic will **fail** to be.
- In Sect. 5, we devise "*penalized robust*-BD *classifiers*" based on the proposed "*penalized robust*-BD *estimates*" in large- and high-dimensional binary classification. We demonstrate that if a parameter estimate possesses the sparsity property and is consistent at an appropriate rate, then the induced classifier attains classification consistency. Hence, even for data contaminated with outliers, the choice of loss functions for regression estimates invoked in the classifier has an asymptotically relatively negligible impact on classification performance.

There is a diverse and extensive literature on robust procedures for model selection. For instance, Zou and Yuan (2008) proposed the composite quantile regression and the oracle model selection theory for linear models. Theorem 3 of Zhang et al. (2009) shows that the quantile loss function does not belong to the class of BD. Therefore, the "*robust*-BD" and the quantile loss, as two operationally different robust alternatives, work in non-overlapping frameworks with different motivations and demand theoretically distinct manipulations. The work (Dupuis & Victoria-Feser, 2011) developed a fast algorithm for robust forward selection procedure in linear regression, where $p_n < n$.

The rest of the paper is organized as follows. Section 6 presents simulation comparisons of the penalized "*robust*-BD" estimates with the classical ones, including classical-SVM and robust-SVM for Bernoulli responses, to assess the performance in statistical model fitting, variable selection, and classification rules. Section 7 analyzes a real dataset. Limitations and open questions are discussed in Sect. 8. Notations, technical and algorithmic details, figures and tables, and additional analysis are collected in Appendix 1 (in the supplementary materials).

2 Proposed robust penalized regression estimation

Let $\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ be a sample of independent observations from some underlying population, $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} = (X_1, \ldots, X_{p_n})^T \in \mathbb{R}^{p_n}$ is the input vector and Y is the output variable. We assume the parametric model for the conditional mean function,

$$m(\boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = F^{-1}(\beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0), \tag{1}$$

together with the conditional variance function

$$\mathrm{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = V(m(\boldsymbol{x})), \tag{2}$$

where $F(\cdot)$ is a known link function, F^{-1} denotes the inverse function of F, $\beta_{0;0} \in \mathbb{R}^1$ and $\boldsymbol{\beta}_0 = (\beta_{1;0}, \ldots, \beta_{p_n;0})^T \in \mathbb{R}^{p_n}$ are the unknown true intercept and regression parameters, and the functional form of $V(\cdot)$ is known. It is worth noting that (1)–(2) include the GLM as a special case. Moreover, they allow the conditional distribution of Y given $\boldsymbol{X}$ to be incompletely (or partially) specified.

Let $\rho_q(y, \mu)$ be a robust loss function (to be proposed in Sect. 2.1) which aims to guard against outlying observations in the response space. We define the "*penalized robust*-BD *estimate*" $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$ as the minimizer of the criterion function,

$$\ell_n(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}))\, w(\boldsymbol{X}_i) + \lambda_n \sum_{j=1}^{p_n} w_{n,j} |\beta_j|, \tag{3}$$

over $\beta_0 \in \mathbb{R}^1$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^T \in \mathbb{R}^{p_n}$, where $w(\cdot) \geq 0$ is a given bounded weight function that downweights high leverage design points in the p_n-dimensional covariate space, $\{w_{n,j}\}$ represent non-negative weights for the penalty terms, and $\lambda_n \geq 0$ serves as a regularization parameter. In practice, the weight function $w(\cdot)$ can be chosen in various ways, such as through prior knowledge or data-driven methods. An empirical choice of $w(\cdot)$ is provided at the beginning of Sect. 6. Data-driven procedures for properly selected penalty weights $\{w_{n,j}\}$ will be carefully developed in Sects. 3.5 and 4. Optimization solutions of (3) will be discussed in Sect. 6. Hereafter, we write $\widetilde{\boldsymbol{X}} = (1, \boldsymbol{X}^T)^T$ and $\widetilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}^T)^T$ to simplify notations.

2.1 Construction of robust loss functions $\rho_q(\cdot, \cdot)$

We propose a class of robust loss functions ρ_q, which is motivated from Bregman divergence (BD), a notion commonly used in the machine learning applications. The original form of BD, which is a bivariate function introduced by Brègman (Brègman, 1967), is defined by

$$Q_q(v, \mu) = -q(v) + q(\mu) + (v - \mu)q'(\mu), \tag{4}$$

where $q(\cdot)$ is a given concave differentiable function. For an extensive literature on BD, see Altun and Smola (2006), Efron (1986), Gneiting (2011), Grünwald and Dawid (2004), Lafferty et al. (1997), Lafferty (1999), Vemuri et al. (2011) and references therein. The BD is suitable for a broad array of error measures Q_q. For example, $q(\mu) = a\mu - \mu^2$ for any constant a yields the quadratic loss $Q_q(y, \mu) = (y - \mu)^2$. For a binary response variable y, $q(\mu) = \min\{\mu, (1 - \mu)\}$ gives the misclassification loss $Q_q(y, \mu) = \mathrm{I}\{y \neq \mathrm{I}(\mu > 1/2)\}$, where $\mathrm{I}(\cdot)$ denotes the indicator function; $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu)\log(1 - \mu)\}$ gives the Bernoulli deviance-based loss $Q_q(y, \mu) = -2\{y \log(\mu) + (1 - y)\log(1 - \mu)\}$; $q(\mu) = 2\min\{\mu, (1 - \mu)\}$ results in the hinge loss; $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$ yields the exponential loss $Q_q(y, \mu) = \exp[-(y - .5)\log\{\mu/(1 - \mu)\}]$. In the cases of $p_n \ll n$ and $p_n \approx n$, Zhang et al. (2010) developed the penalized "classical-BD" estimation.

Despite a wide range of applications of BD in many different domains, its original form, including the quadratic loss used in the ordinary least squares estimates for regression models, yields estimates not resistant to outliers. The robust loss functions for boosting was studied in Kanamori et al. (2007). To the best of our knowledge, there is very little work in the literature on systematically developing robust forms of BD and related inference, in the presence of outliers. In the present work, we describe the construction of a "*robust*-BD". In accordance with the conditional variance function in (2), let $r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$ denote the Pearson residual, which reduces to the standardized residual for linear models. Following (4), we get partial derivatives $\partial Q_q(y, \mu)/\partial\mu = (y - \mu)q''(\mu)$, which can be rewritten as

$$\frac{\partial}{\partial\mu} Q_q(y, \mu) = r(y, \mu)\{q''(\mu)\sqrt{V(\mu)}\}.$$

To guard against outliers with large Pearson residuals, we replace $r(y, \mu)$ by $\psi(r(y, \mu))$, where $\psi(\cdot)$ is chosen to be a bounded, odd function. There is a wide class of functions $\psi(\cdot)$ satisfying these requirements; feasible choices include the Huber ψ-function (Huber, 1964) defined by

$$\psi(r) = r\,\mathrm{I}(|r| \leq c) + c\ \mathrm{sign}(r)\,\mathrm{I}(|r| > c), \tag{5}$$

and the Tukey biweight function formed by $\psi(r) = r\{1 - (r/c)^2\}^2\,\mathrm{I}(|r| \leq c)$, where c is a positive constant. The proposed robust version of BD, ρ_q, is formed by

$$\rho_q(y, \mu) = \int_y^{\mu} \psi(r(y, s))\{q''(s)\sqrt{V(s)}\}\mathrm{d}s - G(\mu), \tag{6}$$

where the bias-correction term, $G(\mu)$, serves to entail the "*conditional zero-mean property*" (see part **(b)** of Sect. 2.2) of a non-penalized and low-dimensional parameter estimate (i.e. minimizing (3) in the case of $\lambda_n = 0$ and $p_n < n$) and satisfies

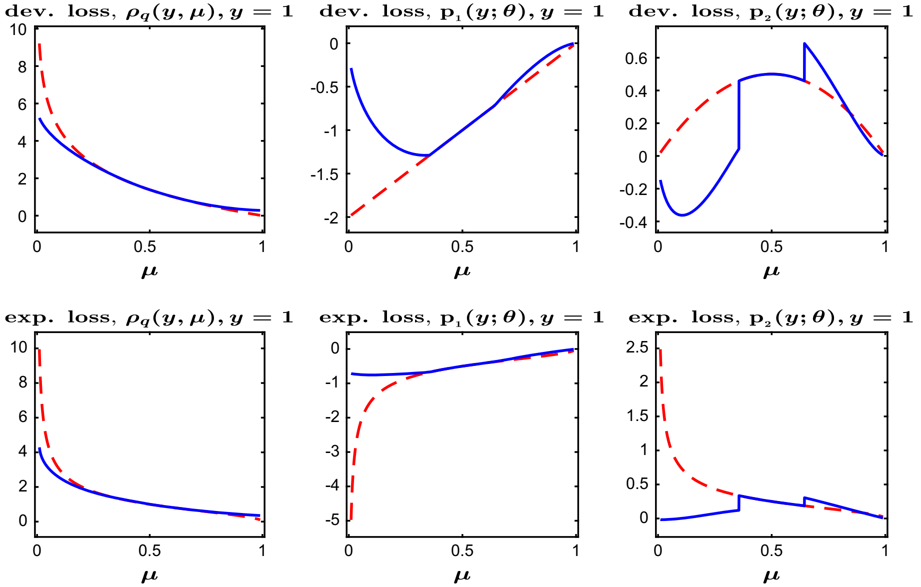


Fig. 1 Plots of $\rho_q(y,\mu)$ (left panels), $p_1(y;\theta)$ (middle panels) and $p_2(y;\theta)$ (right panels) for the Bernoulli response $y=1$. In each panel, solid line: using "*robust*-BD" with Huber ψ-function (5) and $c=1.345$; dashed line: using "*classical*-BD". Top panels: deviance loss used as the BD; bottom panels: exponential loss used as the BD

$$G'(\mu)=g_1(\mu)\{q''(\mu)\sqrt{V(\mu)}\},$$

with

$$g_1(m(\boldsymbol{x}))=\mathrm{E}\{\psi(r(Y,m(\boldsymbol{x})))\mid \boldsymbol{X}=\boldsymbol{x}\}. \tag{7}$$

See Sect. 2.3 for explicit expressions of $G(\mu)$ for Bernoulli responses. We call $\rho_q(\cdot,\cdot)$ defined in (6) the "*robust*-BD", and call the resulting parameter estimate which minimizes (3) the "*penalized robust*-BD *estimate*". An illustrative plot of "*robust*-BD", as compared with the classical-BD, is displayed in Fig. 1.

As a specific example of the class ρ_q of "*robust*-BD" in (6), the robust (negative) quasi-likelihood in Boente et al. (2006) and Cantoni and Ronchetti (2001) can be recovered by setting the generating q-function of BD to be $q(\mu)=\int_a^\mu(s-\mu)/V(s)\mathrm{d}s$, where a is a finite constant such that the integral is well-defined. More generally, the availability of the necessary and sufficient conditions as given in Theorem 3 of Zhang et al. (2009) for an error measure to be a BD enables the construction of the corresponding "*robust*-BD" from expression (6).

2.2 General properties of "*robust*-BD" ρ_q

We make the following comments regarding features of the "*robust*-BD". To facilitate the discussion, we first introduce some necessary notation. Assume that the quantities

$$\mathrm{p}_j(y;\theta) = \frac{\partial^j}{\partial\theta^j}\rho_q(y, F^{-1}(\theta)), j = 0, 1, \ldots, \tag{8}$$

exist finitely up to any order required. Then we have the following expressions,

$$\begin{aligned}
\mathrm{p}_1(y;\theta) &= \{\psi(r(y,\mu)) - g_1(\mu)\}\{q''(\mu)\sqrt{V(\mu)}\}/F'(\mu),\\
\mathrm{p}_2(y;\theta) &= A_0(y,\mu) + \{\psi(r(y,\mu)) - g_1(\mu)\}A_1(\mu),\\
\mathrm{p}_3(y;\theta) &= A_2(y,\mu) + \{\psi(r(y,\mu)) - g_1(\mu)\}A_1'(\mu)/F'(\mu),
\end{aligned} \tag{9}$$

where $\mu = F^{-1}(\theta)$,

$$A_0(y,\mu) = -\left[\psi'(r(y,\mu))\left\{1 + \frac{y-\mu}{\sqrt{V(\mu)}} \times \frac{V'(\mu)}{2\sqrt{V(\mu)}}\right\} + g_1'(\mu)\sqrt{V(\mu)}\right]\frac{q''(\mu)}{\{F'(\mu)\}^2},$$

$A_1(\mu) = [\{q^{(3)}(\mu)\sqrt{V(\mu)} + 2^{-1}q''(\mu)V'(\mu)/\sqrt{V(\mu)}\}F'(\mu) - q''(\mu)\sqrt{V(\mu)}F''(\mu)]/\ \{F'(\mu)\}^3$, and $A_2(y,\mu) = [\partial A_0(y,\mu)/\partial\mu + \partial\{\psi(r(y,\mu)) - g_1(\mu)\}/\partial\mu A_1(\mu)]/F'(\mu)$. Particularly, $\mathrm{p}_1(y;\theta)$ contains $\psi(r)$; $\mathrm{p}_2(y;\theta)$ contains $\psi(r)$, $\psi'(r)$, and $\psi'(r)r$; $\mathrm{p}_3(y;\theta)$ contains $\psi(r)$, $\psi'(r)$, $\psi'(r)r$, $\psi''(r)$, $\psi''(r)r$, and $\psi''(r)r^2$, where $r = r(y,\mu) = (y-\mu)/\sqrt{V(\mu)}$ denotes the Pearson residual. Accordingly, $\{\mathrm{p}_j(y;\theta) : j = 1, 2, 3\}$ depend on y through $\psi(r)$ and its derivatives coupled with r.

(a) ***Relation between the* "*robust*-BD" ρ_q *and* "*classical*-BD" Q_q**. For the particular choice of $\psi(r) = r$, it is clearly noticed from (7) that $g_1(\cdot) = 0$ and thus $G'(\cdot) = 0$. In such a case, the proposed robust $\rho_q(y,\mu)$ in (6) reduces to the conventional form, $Q_q(y,\mu)$, of BD; similarly, $\mathrm{p}_j(y;\theta)$ reduces to $\mathrm{q}_j(y;\theta)$, where

$$\mathrm{q}_j(y;\theta) = \frac{\partial^j}{\partial\theta^j}Q_q(y, F^{-1}(\theta)), j = 0, 1, \ldots. \tag{10}$$

Accordingly, $\mathrm{q}_j(y;\theta)$ is linear in y for fixed θ. As a comparison,

$$\begin{aligned}
\mathrm{q}_1(y;\theta) &= (y-\mu)q^{(2)}(\mu)/F^{(1)}(\mu),\\
\mathrm{q}_2(y;\theta) &= -q^{(2)}(\mu)/\{F^{(1)}(\mu)\}^2 + (y-\mu)A_{1,q}(\mu),\\
\mathrm{q}_3(y;\theta) &\equiv A_2(\mu) + (y-\mu)A_3(\mu),
\end{aligned}$$

where $A_{1,q}(\mu) = \{q^{(3)}(\mu)F^{(1)}(\mu) - q^{(2)}(\mu)F^{(2)}(\mu)\}/\{F^{(1)}(\mu)\}^3$, $A_2(\mu) = \{-2q^{(3)}(\mu)F^{(1)}(\mu) + 3q^{(2)}(\mu)F^{(2)}(\mu)\}/\{F^{(1)}(\mu)\}^4$, and $A_3(\mu) = [q^{(4)}(\mu)\{F^{(1)}(\mu)\}^2 - 3q^{(3)}(\mu)F^{(1)}(\mu)F^{(2)}(\mu) - q^{(2)}(\mu)F^{(1)}(\mu)F^{(3)}(\mu) + 3q^{(2)}(\mu)\{F^{(2)}(\mu)\}^2]/\{F^{(1)}(\mu)\}^5$. In addition, assuming that

$$\mathrm{q}_2(y;\theta) > 0 \text{ for all } \theta \in \mathbb{R} \text{ and all } y \text{ in the range of } Y, \tag{11}$$

we see that $Q_q(Y, F^{-1}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}}))$ is strictly convex in $\widetilde{\boldsymbol{\beta}}$.

(b) ***"Conditional zero-mean property"***. For the proposed class of "*robust*-BD" ρ_q induced by the classical-BD Q_q, it follows from the expression of $\mathrm{p}_1(y;\theta)$ in (9) that $\mathrm{E}\{\mathrm{p}_1(Y;\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} = 0$.

(c) ***Bounded influence function***. When $\lambda_n = 0$ in (3) and $p_n < n$, the "*robust*-BD *estimate*" defined by minimizing (3) is characterized by the score function and influence function below,

$$\boldsymbol{\psi}_{\rho_q}(Y, \boldsymbol{X}) = \mathrm{p}_1(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}) \widetilde{\boldsymbol{X}}, \tag{12}$$

$$\mathrm{IF}(Y, \boldsymbol{X}; \boldsymbol{\psi}_{\rho_q}) = \{M(\boldsymbol{\psi}_{\rho_q})\}^{-1} \boldsymbol{\psi}_{\rho_q}(Y, \boldsymbol{X}), \tag{13}$$

where $M(\boldsymbol{\psi}_{\rho_q}) = -\mathrm{E}\{\partial \boldsymbol{\psi}_{\rho_q}(Y, \boldsymbol{X}) / \partial \widetilde{\boldsymbol{\beta}}_0\} = -\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}) \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T\}$;note that justifying the influence function of a "*robust*-BD" estimate for a p_n-dimensional parameter calls for re-examination and re-derivation beyond the framework of Hampel et al. (1986) for a fixed-dimensional parameter. Thus, the boundedness of $\psi(\cdot)$ and the weight function $w(\cdot)$ ensure a bounded influence function. In contrast, for non-robust counterparts, with $\psi(r) = r$ and $w(\cdot) \equiv 1$, the influence function is unbounded. As to the score function, the "*conditional zero-mean property*" in part **(b)** ensures that under the parametric model (1), $\mathrm{E}\{\boldsymbol{\psi}_{\rho_q}(Y, \boldsymbol{X})\} = \mathbf{0}$ holds for the proposed class of "*robust*-BD" ρ_q induced by the classical-BD Q_q.

(d) ***Conditions under which*** $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} \ge 0$. This is a very minimal condition relevant to discussing Theorems 1 and 2, assumption (19) and numerical minimization of (3). First, as observed from (9), the sign of $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\}$ is invariant with the choice of generating q-functions of BD. Second, one sufficient condition for $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} \ge 0$ is that the conditional distribution of Y given $\boldsymbol{X}$ is symmetric about $m(\boldsymbol{X})$. Third, another sufficient condition for $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} \ge 0$ is that $\mathrm{E}[\psi(r(Y, m(\boldsymbol{X}))) \frac{\partial}{\partial m(\boldsymbol{X})} \log\{f(Y \mid \boldsymbol{X}, m(\boldsymbol{X}))\} \mid \boldsymbol{X}] \ge 0$, which holds for $\psi(r) r \ge 0$ (applicable to Huber and Tukey ψ-functions), and the conditional distribution of Y given $\boldsymbol{X}$ belongs to the exponential family, where f denotes the conditional density or probability of Y given $\boldsymbol{X}$. Fourth, in the particular choice of $\psi(r) = r$, which is unbounded, a direct computation gives that $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} = -q''(m(\boldsymbol{X})) / \{F'(m(\boldsymbol{X}))\}^2 \ge 0$, for any conditional distribution of Y given $\boldsymbol{X}$.

2.3 Difference between "*robust*-BD" and "classical-BD"

To better distinguish between the "*robust*-BD" and "classical-BD", we derive below their closed-form expressions for the Bernoulli responses, using the canonical link $\theta = \log\{\mu/(1-\mu)\}$, Huber ψ-function, and the deviance loss and the exponential loss as the BD. In that case, assume $c \ge 1$ in the Huber ψ-function (5), and define two constants $C_1 = 1/(1 + c^2)$ and $C_2 = 1 - C_1$. Results in the case of $0 < c < 1$ can be similarly obtained.

For the deviance loss employed as the BD, the "*robust*-BD" in (6) takes the form,

$$\rho_q(y,\mu) = p^*(y,\mu) - G(\mu),$$

where

$$p^*(y,\mu) = \begin{cases} -2\log(1-\mu)(1-y) - [4c\{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_1})\} & \\ +2\log(C_1)]y, & \text{if } 0 \le \mu \le C_1, \\ -2\log(1-\mu)(1-y) - 2\log(\mu)y, & \text{if } C_1 < \mu < C_2, \\ -2\log(\mu)y + [4c\{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_2})\} & \\ -2\log(C_1)](1-y), & \text{if } C_2 \le \mu \le 1, \end{cases}$$

$$G(\mu) = \begin{cases} -2(1-\mu) - 2c\{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_1}) - \sqrt{V(\mu)}\}, & \text{if } 0 \le \mu \le C_1, \\ 0, & \text{if } C_1 < \mu < C_2, \\ -2\mu + 2c\{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_2}) + \sqrt{V(\mu)}\}, & \text{if } C_2 \le \mu \le 1. \end{cases}$$

The two related derivative quantities are

$$\mathrm{p}_{1}(y;\theta) = \begin{cases} -2(y-\mu)\{\mu + c\sqrt{V(\mu)}\}, & \text{if } 0 \le \mu \le C_1, \\ -2(y-\mu), & \text{if } C_1 < \mu < C_2, \\ -2(y-\mu)\{(1-\mu) + c\sqrt{V(\mu)}\}, & \text{if } C_2 \le \mu \le 1, \end{cases}$$

$$\mathrm{p}_{2}(y;\theta) = \begin{cases} 2\sqrt{V(\mu)}\{(2\mu - y)\sqrt{V(\mu)} + c/2(\mu - y)(1-2\mu) + cV(\mu)\}, \\ 2V(\mu), \text{if } C_1 < \mu < C_2, \\ 2\sqrt{V(\mu)}\{(1-2\mu+y)\sqrt{V(\mu)} + c/2(\mu-y)(1-2\mu) + cV(\mu)\}, \end{cases} .$$

The "classical-BD" is $Q_q(y,\mu) = -2\log(1-\mu)(1-y) - 2\log(\mu)y$, and the two related quantities are $\mathrm{q}_{1}(y;\theta) = -2(y-\mu)$ and $\mathrm{q}_{2}(y;\theta) = 2V(\mu)$.

Analogously, for the exponential loss used as the BD, the "*robust*-BD" counterpart is represented by $\rho_q(y,\mu) = p^*(y,\mu) - G(\mu)$, where

$$p^*(y,\mu) = \begin{cases} \sqrt{\mu/(1-\mu)}(1-y) - c/2\{\log(\mu/(1-\mu)) + 2\log(c) - 2\}y, & \text{if } 0 \le \mu \le C_1, \\ \{\mu(1-y) + (1-\mu)y\}/\sqrt{V(\mu)}, & \text{if } C_1 < \mu < C_2, \\ \sqrt{(1-\mu)/\mu}y + c/2\{\log(\mu/(1-\mu)) - 2\log(c) + 2\}(1-y), & \text{if } C_2 \le \mu \le 1, \end{cases}$$

$$G(\mu) = \begin{cases} \{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_1})\} + c/2\{\log(1-\mu) - \log(C_2)\}, & \text{if } 0 \le \mu \le C_1, \\ 0, & \text{if } C_1 < \mu < C_2, \\ -\{\sin^{-1}(\sqrt{\mu}) - \sin^{-1}(\sqrt{C_2})\} + c/2\{\log(\mu) - \log(C_2)\}, & \text{if } C_2 \le \mu \le 1. \end{cases}$$

The two related derivative quantities are

$$\mathrm{p}_{1}(y;\theta) = \begin{cases} -(y-\mu)\{\sqrt{\mu/(1-\mu)} + c\}/2, & \text{if } 0 \le \mu \le C_1, \\ -(y-\mu)/\{2\sqrt{V(\mu)}\}, & \text{if } C_1 < \mu < C_2, \\ -(y-\mu)\{\sqrt{(1-\mu)/\mu} + c\}/2, & \text{if } C_2 \le \mu \le 1, \end{cases}$$

$$\mathrm{p}_{2}(y;\theta) = \begin{cases} \sqrt{\mu/(1-\mu)}(1-y)/4 - (1-2\mu)\sqrt{V(\mu)}/4 + c/2V(\mu), & \text{if } 0 \le \mu < C_1, \\ \{\mu(1-y) + (1-\mu)y\}/\{4\sqrt{V(\mu)}\}, & \text{if } C_1 < \mu < C_2, \\ \sqrt{(1-\mu)/\mu}y/4 + (1-2\mu)\sqrt{V(\mu)}/4 + c/2V(\mu), & \text{if } C_2 < \mu \le 1. \end{cases}$$

The "classical-BD" is $Q_q(y, \mu) = \{\mu(1-y) + (1-\mu)y\}/\sqrt{V(\mu)}$, and the two related quantities are $\mathrm{q}_1(y;\theta) = -(y-\mu)/\{2\sqrt{V(\mu)}\}$ and $\mathrm{q}_2(y;\theta) = \{\mu(1-y) + (1-\mu)y\}/\{4\sqrt{V(\mu)}\}$.

In summary, for both types of "classical-BD", which are unbounded, the corresponding versions of "*robust*-BD" are bounded. See the left panels of Fig. 1 with the response $y = 1$, where the "classical-BD" goes to infinity as μ approaches zero. (The case of $y = 0$ will be similar.) From the middle and right panels, $\mathrm{p}_1(y;\theta)$ and $\mathrm{p}_2(y;\theta)$ associated with the "*robust*-BD" are always bounded. In contrast, for the exponential loss, the non-robust counterparts for the "classical-BD" are unbounded. Moreover, we observe from each panel that the "*robust*-BD" and "classical-BD" differ at lower and upper tails of μ, but coincide at the intermediate values of μ.

3 Robust estimation with large-dimensions: $p_n \approx n$

This section investigates the statistical properties of the "*penalized robust*-BD *estimate*" defined by minimizing (3) in sparse large-dimensional parametric models with $p_n \approx n$. Throughout the paper, it is assumed that some entries in $\boldsymbol{\beta}_0$ are exactly zero. Without loss of generality, write $\boldsymbol{X} = (\boldsymbol{X}^{(\mathrm{I})T}, \boldsymbol{X}^{(\mathrm{II})T})^T$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{(\mathrm{I})T}, \boldsymbol{\beta}_0^{(\mathrm{II})T})^T$, where the $\boldsymbol{\beta}_0^{(\mathrm{I})}$ part collects all non-zero coefficients, and $\boldsymbol{\beta}_0^{(\mathrm{II})} = \mathbf{0}$. Let s_n denote the number of non-zero coordinates of $\boldsymbol{\beta}_0$, and set $\widetilde{\boldsymbol{\beta}}_0 = (\beta_{0;0}, \boldsymbol{\beta}_0^T)^T$. Correspondingly, write $\widetilde{\boldsymbol{X}}^{(\mathrm{I})} = (1, \boldsymbol{X}^{(\mathrm{I})T})^T$ and $\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})} = (\beta_0, \boldsymbol{\beta}^{(\mathrm{I})T})^T$.

It will be demonstrated that the impact of penalty weights $\{w_{n,j}\}$ in (3) on the penalized "*robust*-BD" estimate is primarily captured by two quantities, defined by

$$w_{\max}^{(\mathrm{I})} = \max_{1 \le j \le s_n} w_{n,j}, \quad w_{\min}^{(\mathrm{II})} = \min_{s_n+1 \le j \le p_n} w_{n,j}.$$

3.1 Consistency

Theorem 1 guarantees the existence of a $\sqrt{n/s_n}$-consistent local minimizer of (3). In particular, Theorem 1 allows the dimension to diverge with n at the rate $p_n = o\{n^{(3+\delta)/(4+\delta)}\}$ for any $\delta > 0$, as long as the number of truly non-zero parameters fulfills that $s_n = O\{n^{1/(4+\delta)}\}$.

Theorem 1 (*existence and consistency*: $p_n \approx n$) *Assume Conditions* A0, A1, A2, A4, A5, A6, A7 *in Appendix* 1.1, $w_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and there exists a constant* $M \in (0, \infty)$ *such that* $\lim_{n\to\infty} \mathrm{P}(w_{\min}^{(\mathrm{II})}\lambda_n > M) = 1$. *If* $s_n^4/n \to 0$ *and* $s_n(p_n - s_n) = o(n)$, *then there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* (3) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n})$, *where* $\|\cdot\|_2$ *denotes the Euclidean norm.*

To clarify the distinction between Theorem 5 of Zhang et al. (2010) and Theorem 1 of this paper, we make the following comparison. (i) Theorem 5 of Zhang et al. (2010) uses the classical-BD, corresponding to $\psi(r) = r$ in (6), and assumes the finiteness of some moments of Y. (ii) Condition A5 of our Theorem 1 uses the robust-BD and assumes the boundedness of $\psi(r)$, thus excluding $\psi(r) = r$, but avoids the moment assumption. Thus, our Theorem 1 is not applicable to the case of $\psi(r) = r$ in (6) associated with a classical-BD of Zhang et al. (2010).

3.2 Oracle property

To investigate the asymptotic distribution of the penalized *robust*-BD estimate $\widehat{\widetilde{\boldsymbol{\beta}}}$, we define three square matrices of size (s_n+1) by $\mathbf{W}_n^{(\mathrm{I})} = \mathrm{diag}(0, w_{n,1}, \ldots, w_{n,s_n})$ and

$$\begin{aligned}\boldsymbol{\Omega}_n^{(\mathrm{I})} &= \mathrm{E}\{\mathrm{p}_1^2(Y; \widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w^2(\boldsymbol{X})\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\},\\ \mathbf{H}_n^{(\mathrm{I})} &= \mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X})\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\}.\end{aligned}$$

Both $\boldsymbol{\Omega}_n^{(\mathrm{I})}$ and $\mathbf{H}_n^{(\mathrm{I})}$ depend on the choice of BD, weight function $w(\cdot)$, and the ψ-function $\psi(\cdot)$.

Following Theorems 1, 2 obtains the oracle property of the $\sqrt{n/s_n}$-consistent local minimizer. Namely, if the "*robust*-BD" is used as the loss function for parameter estimation, then the penalized *robust*-BD estimates of the zero parameters take exactly zero values with probability tending to one, and the penalized *robust*-BD estimates of the non-zero parameters are asymptotically Gaussian with the same means and variances as if the zero coefficients were known in advance.

Theorem 2 (*oracle property*: $p_n \approx n$) *Assume Conditions* A0, A1, A2, A4, A5, B5, A6, A7 *in Appendix* 1.1.

(i) If $s_n^2/n = O(1)$ and $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/\sqrt{s_n p_n} \xrightarrow{\mathrm{P}} \infty$ as $n \to \infty$, then any $\sqrt{n/s_n}$-consistent local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})T}, \widehat{\boldsymbol{\beta}}^{(\mathrm{II})T})^T$ of (3) satisfies $\mathrm{P}(\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \mathbf{0}) \to 1$.

(ii) Moreover, if $w_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$, $s_n^5/n \to 0$ and $\min_{1\le j\le s_n}|\beta_{j;0}|/\sqrt{s_n/n} \to \infty$, then for any fixed integer k and any $\mathsf{k}\times(s_n+1)$ matrix A_n such that $A_n A_n^T \to \mathbb{G}$ for a $\mathsf{k}\times\mathsf{k}$ nonnegative-definite symmetric matrix $\mathbb{G}$, then $\sqrt{n}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}[\mathbf{H}_n^{(\mathrm{I})}\{\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\} + \lambda_n \mathbf{W}_n^{(\mathrm{I})}\,\mathrm{sign}\{\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\}] \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbb{G})$, where $\mathrm{sign}\{\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\} = (\mathrm{sign}(\beta_0), \mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_{s_n}))^T$.

Remark 1 In Theorem 2, Condition B5 extends the positive-definiteness assumption of $\mathrm{E}\{\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\}$ in the non-robust and fixed-dimensional case to the robust and large-dimensional case. The assumption $\min_{1\le j\le s_n}|\beta_{j;0}|/\sqrt{s_n/n} \to \infty$ is relevant to the magnitude of coefficients for significant variables which can be selected, and is fulfilled when $\min_{1\le j\le s_n}|\beta_{j;0}| \ge Cn^{-4/5}$ for a constant $C > 0$.

3.3 Comparison with the penalized "classical-BD" estimate

Comparisons are made between the penalized "*robust*-BD" and "classical-BD" estimates. (I) The penalized "classical-BD" estimate in Zhang et al. (2010) requires $\mathrm{E}(Y^2) < \infty$ for the consistency and requires finiteness of some higher-order moments of Y for the oracle property. These requirements are avoided in the "*robust*-BD" counterpart. (II) The two types of penalized estimates appear to share similar forms of the asymptotic distribution, except that matrices $\boldsymbol{\Omega}_n^{(\mathrm{I})}$ and $\mathbf{H}_n^{(\mathrm{I})}$ for the "classical-BD" estimate are given by

$$\mathrm{E}\{\mathrm{q}_{_1}^2(Y;\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\}, \qquad \mathrm{E}\{\mathrm{q}_{_2}(Y;\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\},$$

respectively. Hence, the differences are captured by the distinction between the robust versions, $\{\mathrm{p}_{_j}(y;\theta)\}_{j=1}^2$ (defined in (8)) and weight function $w(\cdot)$, used in the penalized "*robust*-BD" estimate and the non-robust counterparts, $\{\mathrm{q}_{_j}(y;\theta)\}_{j=1}^2$ (defined in (10)) and $w(\cdot)\equiv 1$, used in the penalized "classical-BD" estimate.

As observed from (12)–(13), a bounded function $\mathrm{p}_{_1}(y;\theta)$ is introduced from a bounded function $\psi(r)$ to control deviations in the Y-space, and leverage points are down-weighted by the weight function $w(\boldsymbol{X})$. In contrary, $\mathrm{q}_{_1}(y;\theta)$ is not guaranteed to be bounded. It is then clear that for penalized "*robust*-BD" estimates of non-zero parameters, the choice of a bounded score function ensures robustness by putting a bound on the influence function. Such property is not possessed by the penalized "classical-BD" counterparts.

3.4 Hypothesis testing

We consider the hypothesis testing about $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}$ formulated as

$$H_0: A_n\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})} = \mathbf{0}\text{versus} \tag{14}$$

where A_n is a given $\mathsf{k}\times(s_n+1)$ matrix such that $A_nA_n^T$ equals a $\mathsf{k}\times\mathsf{k}$ positive-definite matrix $\mathbb{G}$. This form of linear hypotheses allows one to simultaneously test whether a subset of variables used are statistically significant by taking some specific form of the matrix A_n; for instance $A_n=[\mathbf{I}_{\mathsf{k}},\mathbf{0}_{\mathsf{k},s_n+1-\mathsf{k}}]$ yields $A_nA_n^T=\mathbf{I}_{\mathsf{k}}$ for a $\mathsf{k}\times\mathsf{k}$ identity matrix.

In the context of non-robust penalized-likelihood estimation, Fan and Peng (2004) showed that the likelihood ratio-type test statistic asymptotically follows a chi-squared distribution under the null. It is thus natural to explore the extent to which the likelihood ratio-type test can feasibly be extended to the "*robust*-BD". Our derivations (with details omitted) indicate that the resulting asymptotic null distribution is generally not chi-squared, but a sum of weighted chi-squared variables, with weights involving unknown quantities, thus not distribution free, and holds under restrictive conditions.

To ameliorate this undesirable property, we propose a robust generalized Wald-type test statistic of the form,

$$W_n = n\{A_n\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}\}^T\{A_n(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}\widehat{\boldsymbol{\Omega}}_n^{(\mathrm{I})}(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}A_n^T\}^{-1}\{A_n\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}\},$$

where $\widehat{\boldsymbol{\Omega}}_n^{(\mathrm{I})} = n^{-1}\sum_{i=1}^n \mathrm{p}_{_1}^2(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})})\,w^2(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}$ and

$\widehat{\mathbf{H}}_n^{(\mathrm{I})} = n^{-1}\sum_{i=1}^n \mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})})\,w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}$. This test is asymptotically distribution-free, as Theorem 3 justifies that under the null, W_n is asymptotically chi-squared with k degrees of freedom.

Theorem 3 (*Wald-type test under H_0 based on robust*-BD: $p_{_n}\approx n$) *Assume Conditions* A0, A1, A2, C4, A5, B5, A6, A7 *in Appendix* 1.1, *and* $w_{\max}^{(\mathrm{I})} = o_{\mathrm{P}}\{1/(\lambda_n\sqrt{ns_n})\}$. *If* $s_n^5/n\to 0$ *and* $\min_{1\le j\le s_n}|\beta_{j;0}|/\sqrt{s_n/n}\to\infty$ *as* $n\to\infty$, *then* $W_n\xrightarrow{\mathcal{L}}\chi^2_{\mathsf{k}}$ *under the null* H_0 *in* (14).

Since the influence function of $\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}$ is bounded, Proposition 2 of Cantoni and Ronchetti (2001) can be modified to show that the asymptotic level of the robust test statistic W_n under a sequence of ϵ-contaminations is bounded. Similarly, the asymptotic power is stable under contamination. Details are omitted for lack of space.

3.5 Proposed selection for penalty weights

In practice, the weights $\{w_{n,j}\}$ in the penalty part of (3) need to be selected and their validity also needs to be justified. To accommodate the general error measures, we propose two procedures. The first one, called the "*penalized marginal regression*" (PMR) method, selects the data-dependent penalty weights $\widehat{w}_{n,j}$, for each individual $j = 1, \ldots, p_n$, according to

$$\widehat{w}_{n,j} = 1/|\widehat{\beta}_j^{\mathrm{PMR}}|, \tag{15}$$

where $\widehat{\beta}_j^{\mathrm{PMR}}$ satisfies that $(\widehat{\alpha}_j^{\mathrm{PMR}}, \widehat{\beta}_j^{\mathrm{PMR}}) \in \mathbb{R}^2$ minimize the criterion function

$$\ell_{n,j}^{\mathrm{PMR}}(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n} Q_q(Y_i, F^{-1}(\alpha + X_{i,j}\beta)) + \kappa_n|\beta| \tag{16}$$

over (α, β), with some sequence $\kappa_n > 0$, and $X_{i,j}$ denoting the jth variable in the ith sample.

An alternative procedure, called the "*marginal regression*" (MR) method, selects the penalty weights $\widehat{w}_{n,j}$, for each individual $j = 1, \ldots, p_n$, by means of

$$\widehat{w}_{n,j} = 1/|\widehat{\beta}_j^{\mathrm{MR}}|, \tag{17}$$

where $\widehat{\beta}_j^{\mathrm{MR}}$ satisfies that $(\widehat{\alpha}_j^{\mathrm{MR}}, \widehat{\beta}_j^{\mathrm{MR}}) \in \mathbb{R}^2$ minimize the criterion function

$$\ell_{n,j}^{\mathrm{MR}}(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n} Q_q(Y_i, F^{-1}(\alpha + X_{i,j}\beta)) \tag{18}$$

over (α, β).

Note that (16) and (18) each involves a univariate predictor with the intercept term. Thus fast bivariate optimization solutions of (16) and (18) would be feasible even when $p_n > n$. Compared with the PMR method, the MR method gains computational superiority with less computational cost.

Theorems 4 and 5 indicate that under the assumptions on the correlation between the predictor variables and the response variable, the penalty weights selected by either the PMR or MR method satisfy the conditions on $\{w_{n,j}\}$ in Theorem 1.

Theorem 4 (PMR *for penalty weights*: $p_n \approx n$) *Assume* (11) *and Conditions* A0, A1, A2, B3, A4, A6, A7 *in Appendix* 1.1. *Assume* E1 *in Appendix* 1.1, *where* $\mathcal{A}_n = \lambda_n\sqrt{n}$, $\mathcal{A}_n/\kappa_n \to \infty$ *and* $\mathcal{B}_n/\kappa_n = O(1)$ *for* κ_n *in* (16). *Suppose* $\lambda_n\sqrt{n} = O(1)$, $\lambda_n = o(\kappa_n)$ *and* $\log(p_n) = o(n\kappa_n^2)$. *Then there exist local minimizers* $\{\widehat{\beta}_j^{\mathrm{PMR}}\}_{j=1}^{p_n}$ *of* (16) *such that the penalty weights* $\{\widehat{w}_{n,j}\}_{j=1}^{p_n}$ *defined in* (15) *satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and* $\widehat{w}_{\min}^{(\mathrm{II})}\lambda_n \xrightarrow{\mathrm{P}} \infty$ *as needed in Theorem* 1, *where* $\widehat{w}_{\max}^{(\mathrm{I})} = \max_{1 \le j \le s_n} \widehat{w}_{n,j}$ *and* $\widehat{w}_{\min}^{(\mathrm{II})} = \min_{s_n+1 \le j \le p_n} \widehat{w}_{n,j}$.

Theorem 5 (MR *for penalty weights*: $p_n \approx n$) *Assume* (11) *and Conditions* A0, A1, A2, B3, A4, A6, A7 *in Appendix* 1.1. *Assume* E2 *and* E1 *in Appendix* 1.1, *where* $\mathcal{A}_n = \lambda_n\sqrt{n}$ *and* $\mathcal{B}_n = \lambda_n$. *Suppose* $\lambda_n\sqrt{n} = O(1)$, $\lambda_n \to 0$, $\lambda_n n/s_n \to \infty$ *and* $s_n^2 \log(p_n) = o(\lambda_n^2 n^2)$. *Then there exist local minimizers* $\widehat{\beta}_j^{\mathrm{MR}}$ *of* (18) *such that the penalty weights* $\widehat{w}_{n,j}$ *defined in* (17) *satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and* $\widehat{w}_{\min}^{(\mathrm{II})}\lambda_n \xrightarrow{\mathrm{P}} \infty$ *as needed in Theorem* 1.

It should be pointed out that the PMR method here differs from the weights selection method in Zhang et al. (2010), which is restricted to $p_n \approx n$, excludes the intercept term in (16) and requires $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$ to satisfy the conditions on $\{w_{n,j}\}$ in Theorem 1. In contrast, the PMR method here removes that requirement and will also be applicable to $p_n \gg n$; see Theorem 7.

4 Robust estimation with high-dimensions: $p_n \gg n$

This section explores the behavior of penalized *robust*-BD estimates in sparse high-dimensional parametric models when p_n is allowed to grow faster than n. Evidently, the technical conditions (e.g., in Theorem 1) on p_n in Sect. 3 are violated for $p_n \gg n$. Thus, directly carrying through the proofs in Sect. 3 to the counterpart of $p_n \gg n$ is infeasible. To facilitate the study in the case of $p_n \gg n$, we impose a convexity assumption on the "*robust*-BD":

$$\mathrm{p}_2(y;\theta) > 0 \text{ for all } \theta \in \mathbb{R} \text{ and all } y \text{ in the range of } Y. \tag{19}$$

Under this assumption, $\rho_q(Y, F^{-1}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}}))$ is strictly convex in $\widetilde{\boldsymbol{\beta}}$.

Apparently, assumption (19) is stronger than $\mathrm{E}\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} \geq 0$ discussed in Sect. 2.2; relaxing (19) will be desirable but is not pursued in this paper. We consider here particular cases where assumption (19) is practically/approximately achievable and theoretically relevant for high-dimensional settings.

Case 1: For any observation $(\boldsymbol{X}, Y)$ such that the conditional distribution of Y given $\boldsymbol{X}$ is symmetric about $m(\boldsymbol{X})$, the use of a quadratic loss function combined with an identity link, and a constant conditional variance ensures that $\mathrm{p}_2(y;\theta) = 2\psi'(r(y,\theta)) \geq 0$, for a monotone non-decreasing ψ-function. Thus, the Gaussianity assumption on the conditional distribution of Y given $\boldsymbol{X}$ is relaxed.

Case 2: Recall that if $\psi(r) = r$, then $\mathrm{p}_j(y;\theta) = \mathrm{q}_j(y;\theta)$, and thus condition (19) is equivalent to condition (11). Indeed, condition (11) holds broadly for nearly all commonly used BD. Examples include the quadratic loss function, the deviance-based loss and exponential loss functions for the Bernoulli responses, and the (negative) quasi-likelihood for over-dispersed Poisson responses, among many others. The implication is that for high-dimensional data, if we are most concerned with dealing with outliers arising from the explanatory variables, we may employ $\psi(r) = r$ for Yalone while retaining the weight function $w(\cdot)$ on the covariates $\boldsymbol{X}$. This is particularly relevant to samples with Bernoulli or Binomial responses, where both the parameter space and the response space are bounded, and thus is applicable to a wide class of classification procedures.

Theorem 6 states that the oracle property remains true, under suitable conditions, for the penalized *robust*-BD estimates in the $p_n \gg n$ settings. Due to the technical challenge from $p_n \gg n$, Theorem 6 contains stronger assumptions than those in Theorem 2 (with $p_n \approx n$). Lemma 4 in Appendix 1.1 provides key proofs for Theorem 6.

Theorem 6 (*oracle property*: $p_n \gg n$) *Assume* (19) *and Conditions* A0, A1, A2, A4, A5$'$, B5, A6, A7 *in Appendix* 1.1. *Suppose* $s_n^4/n \to 0, \log(p_n - s_n)/n = O(1), \log(p_n - s_n)/\{n\lambda_n^2 (w_{\min}^{(\mathrm{II})})^2\} = o_{\mathrm{P}}(1)$ *and* $\min_{1\le j\le s_n} |\beta_{j;0}|/\sqrt{s_n/n} \to \infty$. *Assume* $w_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and* $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/s_n \xrightarrow{\mathrm{P}} \infty$.

(i) Then there exists a global minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}$ of (3) such that $\mathrm{P}(\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \mathbf{0}) \to 1$.

(ii) Moreover, if $s_n^5/n \to 0$, then for any fixed integer k and any $\mathsf{k} \times (s_n + 1)$ matrix A_n such that $A_n A_n^T \to \mathbb{G}$ for a $\mathsf{k} \times \mathsf{k}$ nonnegative-definite symmetric matrix $\mathbb{G}$, then $\sqrt{n}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}[\mathbf{H}_n^{(\mathrm{I})}(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + \lambda_n \mathbf{W}_n^{(\mathrm{I})}\,\mathrm{sign}\{\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\}] \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbb{G})$.

The conditions $s_n^5/n = o(1)$, $\log(p_n - s_n) = O(n)$ and $\log(p_n - s_n)/\{(w_{\min}^{(\mathrm{II})})^2\lambda_n^2 n\} = o_{\mathrm{P}}(1)$ impose the constraints on s_n and p_n simultaneously. To illustrate, take $s_n = n^a$, where $0<a<1/5$. Then we can take $w_{\min}^{(\mathrm{II})} = n^b/(\lambda_n\sqrt{n})$ for $b > a$, namely, $n\lambda_n^2\{w_{\min}^{(\mathrm{II})}\}^2 = n^{2b}$. A sufficient condition for the model dimension p_n is that $\log(p_n - s_n) = O(n)$ and $\log(p_n - s_n) = o(n^{2b})$. So $\log(p_n - s_n) = \min\{o(n^{2b}), O(n)\}$. This indicates that Theorem 6 allows $p_n = \exp\{o(n^{2b}) \wedge O(n)\}$, which grows nearly exponentially fast with n.

Regarding the selection of penalty weights, the PMR and MR methods proposed in Sect. 3.5 with $p_n \approx n$ continue to work well for selecting weights with $p_n \gg n$. The validity is presented in Theorems 7 and 8, which again do not require $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$.

Theorem 7 (PMR *for penalty weights*: $p_n \gg n$) *Assume* (11) *and Conditions* A0, A1, A2, B3, A4, A6, A7 *in Appendix* 1.1. *Assume* E1 *in Appendix* 1.1, *where* $\mathcal{A}_n = \lambda_n\sqrt{n}$, $\mathcal{A}_n/\kappa_n \to \infty$, *and* $\mathcal{B}_n/\kappa_n = O(1)$ *for* κ_n *in* (16). *Suppose* $\lambda_n\sqrt{n} = O(1)$, $\lambda_n\sqrt{n}/s_n = o(\kappa_n)$ *and* $\log(p_n) = o(n\kappa_n^2)$. *Then there exist local minimizers* $\widehat{\beta}_j^{\mathrm{PMR}}$ *of* (16) *such that the penalty weights* $\widehat{w}_{n,j}$ *defined in* (15) *satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and* $\widehat{w}_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/s_n \xrightarrow{\mathrm{P}} \infty$ *as needed in Theorem* 6.

Theorem 8 (MR *for penalty weights*: $p_n \gg n$) *Assume* (11) *and Conditions* A0, A1, A2, B3, A4, A6, A7 *in Appendix* 1.1. *Assume* E2 *and* E1 *in Appendix* 1.1, *where* $\mathcal{A}_n = \lambda_n\sqrt{n}$ *and* $\mathcal{B}_n = \lambda_n\sqrt{n}/s_n$. *Suppose* $\lambda_n\sqrt{n} = O(1)$, $\lambda_n\sqrt{n}/s_n \to 0$, $\lambda_n n/s_n \to \infty$ *and* $s_n^2\log(p_n) = o(\lambda_n^2 n^2)$. *Then there exist local minimizers* $\widehat{\beta}_j^{\mathrm{MR}}$ *of* (18) *such that the penalty weights* $\widehat{w}_{n,j}$ *defined in* (17) *satisfy that* $\widehat{w}_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n\sqrt{n})\}$ *and* $\widehat{w}_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/s_n \xrightarrow{\mathrm{P}} \infty$ *as needed in Theorem* 6.

5 Robust estimation in classification: $p_n \approx n$, $p_n \gg n$

This section deals with the binary response variable Y which takes values 0 and 1. In this case, the mean regression function $m(\boldsymbol{x})$ in (1) becomes the class label probability, $\mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. From the penalized *robust*-BD estimate $\widehat{\widetilde{\boldsymbol{\beta}}}$ proposed in either Sect. 3 or Sect. 4, we can construct the "*penalized robust*-BD *classifier*", $\widehat{\phi}(\boldsymbol{x}) = \mathrm{I}\{\widehat{m}(\boldsymbol{x}) > 1/2\}$, for a future input variable $\boldsymbol{x}$, where $\widehat{m}(\boldsymbol{x}) = F^{-1}(\widetilde{\boldsymbol{x}}^T\widehat{\widetilde{\boldsymbol{\beta}}})$.

In the classification literature, the misclassification loss of a classification rule ϕ at a sample point $(\boldsymbol{x}, y)$ is $l(y, \phi(\boldsymbol{x})) = \mathrm{I}\{y \neq \phi(\boldsymbol{x})\}$. The risk of ϕ is the expected misclassification loss, $R(\phi) = \mathrm{E}\{l(Y, \phi(\boldsymbol{X}))\} = \mathrm{P}(\phi(\boldsymbol{X}) \neq Y)$. The optimal Bayes rule, which minimizes the risk over ϕ, is $\phi_{\mathrm{B}}(\boldsymbol{x}) = \mathrm{I}\{m(\boldsymbol{x}) > 1/2\}$. For a test sample $(\boldsymbol{X}^o, Y^o)$, which is an i.i.d. copy of samples in the training set $\mathcal{T}_n = \{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$, the optimal Bayes risk is then $R(\phi_{\mathrm{B}}) = \mathrm{P}(\phi_{\mathrm{B}}(\boldsymbol{X}^o) \neq Y^o)$. Meanwhile, the conditional risk of the classification rule $\widehat{\phi}$ is $R(\widehat{\phi} \mid \mathcal{T}_n) = \mathrm{P}(\widehat{\phi}(\boldsymbol{X}^o) \neq Y^o \mid \mathcal{T}_n)$.

Theorem 9 demonstrates that $\widehat{\phi}$ attains the classification consistency, provided that the estimate $\widehat{\widehat{\boldsymbol{\beta}}}$ possesses the sparsity property and is consistent at an appropriate rate. As observed, the conclusion of Theorem 9 is applicable to penalized *robust*-BD estimates in both Theorem 2(i) with $p_{_n} \approx n$ and Theorem 6(i) with $p_{_n} \gg n$.

Theorem 9 (*consistency of the penalized robust*-BD *classifier*) *Assume Conditions* A0, A1 *and* A4 *in Appendix* 1.1. *Suppose that the estimate* $\widehat{\widehat{\boldsymbol{\beta}}} = (\widehat{\widehat{\boldsymbol{\beta}}}^{(\mathrm{I})T}, \widehat{\boldsymbol{\beta}}^{(\mathrm{II})T})^T$ *satisfies* $\mathrm{P}(\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \boldsymbol{0}) \to 1$ *and* $\|\widehat{\widehat{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(r_n)$. *If* $r_n\sqrt{s_n} = o(1)$, *then the classification rule* $\widehat{\phi}$ *constructed from* $\widehat{\widehat{\boldsymbol{\beta}}}$ *is consistent in the sense that* (i) $\mathrm{E}\{R(\widehat{\phi} \mid \mathcal{T}_n)\} - R(\phi_{\mathrm{B}}) \to 0$ *as* $n \to \infty$, *which in turn yields* (ii) $R(\widehat{\phi} \mid \mathcal{T}_n) \xrightarrow{\mathrm{P}} R(\phi_{\mathrm{B}})$.

6 Simulation study

We conduct simulation studies to evaluate the performance of the penalized *robust*-BD estimates in the absence and presence of outliers. The Huber ψ-function is used with $c = 1.345$. For practical applications, we suggest an empirical choice of the weight function, $w(\boldsymbol{x}) = 1/\{1 + \sum_{j=1}^{p_n} (\frac{x_j - m_{\cdot,j}}{s_{\cdot,j}})^2\}^{1/2}$, where $\boldsymbol{x} = (x_1, \ldots, x_{p_n})^T$, $m_{\cdot,j}$ and $s_{\cdot,j}$ denote the sample median and sample median absolute deviation of $\{X_{i,j} : i = 1, \ldots, n\}$ respectively, $j = 1, \ldots, p_{_n}$. This form of $w(\boldsymbol{x})$ is a generalization of a weight function used on page 2864 of Boente et al. (2006) for a one-dimensional covariate. The classical non-robust counterparts correspond to $\psi(r) = r$ and $w(\boldsymbol{x}) \equiv 1$. In the simulation, $p_{_n} = 50$ and 500 are treated respectively.

For illustrative purpose, 4 types of penalization techniques combined with the loss term in (3) are compared: (I) the SCAD penalty, with an accompanying parameter $a = 3.7$, combined with the local linear approximation; (II) the L_1 penalty; (III) the weighted-L_1 penalty with weights selected by the proposed MR method; (IV) the weighted-L_1 penalty with weights selected by the proposed PMR method. For comparison, the oracle (non-penalized) estimator (abbreviated as "Oracle") of parameters using the true model containing truly significant variables is included. The tuning constants λ_n in each simulation for methods (I)–(III) are selected separately by minimizing the classical-BD (for non-robust methods) and "*robust*-BD" (for robust methods) on a test set of size n; λ_n and κ_n for method (IV) are searched on a surface of grid points. Numerical algorithms for the proposed penalized estimator in (3) are given in Appendix 1.3.

The number of Monte Carlo runs is 500. To measure the performance of a parameter estimate $\widehat{\widehat{\boldsymbol{\beta}}}$ through simulation, we use the average value and standard deviation (sd) of estimation errors, $\mathrm{EE}(\widehat{\widehat{\boldsymbol{\beta}}}) = \|\widehat{\widehat{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$. Variable selection is assessed by C-Z, the average number of regression coefficients which are correctly estimated to be zero when the true coefficients are zero, and C-NZ, the average number of coefficients which are correctly estimated to be non-zero when the true coefficients are non-zero. For binary responses, MCR denotes the average of the misclassification rates on a test set of size 5000.

6.1 Overdispersed Poisson responses

We generate overdispersed Poisson counts $\{Y_i\}_{i=1}^{n}$ with $n = 100$, satisfying $\mathrm{var}(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 2\,m(\boldsymbol{x}_i)$. In the predictor $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p_n})^T$, $X_{i,1} = i/n - 0.5$; for $j = 2, \ldots, p_n$, $X_{i,j} = \Phi(Z_{i,j}) - 0.5$, where Φ is the standard Gaussian distribution function, and $(Z_{i,2}, \ldots, Z_{i,p_n})^T \sim N(\mathbf{0}, \Sigma_{p_n - 1})$, with $\Sigma_{p_n-1}(j,k) = 0.2^{|j-k|}$, $j,k = 1, \ldots, p_n - 1$. The link function is $\log\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0$, with $\beta_{0;0} = 2.5$ and $\boldsymbol{\beta}_0 = (2, 2, 0, \ldots, 0)^T$. The (negative) quasi-likelihood with $V(\mu) = \mu$ is utilized as the BD.

Study 1 (*raw data without outliers*). For simulated data without contamination, the results are summarized in Table 1. The lose of efficiency (at the true model) from classical to robust estimates using the (non-penalized) oracle estimation method can be seen from the increase of $\mathrm{EE}(\widehat{\widehat{\boldsymbol{\beta}}})$. For penalized methods, the robust estimates exhibit similar performance to those of the non-robust counterparts, with little loss in estimation efficiency and selecting relevant and irrelevant variables. Thus, there is no serious adverse effect of applying penalized robust estimation to clean datasets.

Study 2 (*contaminated data with outliers*). For each data set generated from the model, we create a contaminated data set, where 8 data points $(X_{i,j}, Y_i)$ are subject to contamination: They are replaced by $(X_{i,j}^*, Y_i^*)$, where $Y_i^* = Y_i\,\mathrm{I}(Y_i > 100) + A\,\mathrm{I}(Y_i \le 100)$ with $A = 50$, $i = 1, \ldots, 8$,

$$\begin{aligned}
&X_{1,1}^* = .5\,\mathrm{sign}(U_1 - 0.5), \quad X_{2,2}^* = .5\,\mathrm{sign}(U_2 - 0.5), \quad X_{3,3}^* = .5\,\mathrm{sign}(U_3 - 0.5),\\
&X_{4,5}^* = .5\,\mathrm{sign}(U_4 - 0.5), \quad X_{5,7}^* = .5\,\mathrm{sign}(U_5 - 0.5), \quad X_{6,8}^* = .5\,\mathrm{sign}(U_6 - 0.5),\\
&X_{7,9}^* = .5\,\mathrm{sign}(U_7 - 0.5),
\end{aligned}$$

with $\{U_i\} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1)$. Table 2 summarizes the results over 500 sets of contaminated data. A comparison of each penalized quasi-likelihood estimate across Tables 1 and 2 indicates that the presence of contamination substantially increases $\mathrm{EE}(\widehat{\widehat{\boldsymbol{\beta}}})$. Among the 4 penalized estimates, the L_1 penalty tends to have higher false positive rates. As observed from Table 2 with contaminated cases, the non-robust estimates are more sensitive to outliers than the robust counterparts. This lends support to Theorems 2 and 6. To provide a closer view of the estimates, Fig. 2 draws the boxplots of biases $(\widehat{\beta}_j - \beta_{j;0}), j = 0, 1, \ldots, 5$, corresponding to results in Tables 1 and 2, using the PMR selection method for penalty weights in the weighted-L_1 penalty.

6.2 Bernoulli responses

We generate samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$ with $n = 200$ from the model, $Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \sim \mathrm{Bernoulli}\{m(\boldsymbol{x}_i)\}$, where $\mathrm{logit}\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0$ with $\beta_{0;0} = 2$ and $\boldsymbol{\beta}_0 = (2, 2, 0, \ldots, 0)^T$. The predictor $\boldsymbol{X}_i \sim N(\mathbf{0}, \Sigma_{p_n})$, with $\Sigma_{p_n}(j,k) = 0.1^{|j-k|}$, $j,k = 1, \ldots, p_n$. Both the deviance and exponential loss functions are employed as the BD.

Table 1 (Simulation study: overdispersed Poisson responses, $n = 100$) Summary results for Study 1 (raw data without outliers)

Procedure	p_n	Method	Regression	Variable selection	
			$\mathrm{EE}(\widehat{\boldsymbol{\beta}})$ (sd)	C-Z (sd)	C-NZ (sd)
non-robust	50	SCAD	0.205 (0.1)	45.8 (3.6)	3.0 (0.0)
		L_1	0.415 (0.1)	39.8 (5.4)	3.0 (0.0)
		wL_1, MR	0.250 (0.1)	46.0 (3.0)	3.0 (0.0)
		wL_1, PMR	0.202 (0.1)	46.9 (2.6)	3.0 (0.0)
		Oracle	0.178 (0.1)	48.0 (0.0)	3.0 (0.0)
	500	SCAD	0.205 (0.1)	494.0 (7.0)	3.0 (0.0)
		L_1	0.581 (0.1)	479.7 (12.2)	3.0 (0.0)
		wL_1, MR	0.314 (0.1)	494.9 (3.5)	3.0 (0.0)
		wL_1, PMR	0.205 (0.1)	496.7 (3.1)	3.0 (0.0)
		Oracle	0.172 (0.1)	498.0 (0.0)	3.0 (0.0)
robust	50	SCAD	0.244 (0.2)	47.5 (1.2)	3.0 (0.0)
		L_1	0.411 (0.1)	39.4 (5.8)	3.0 (0.0)
		wL_1, MR	0.242 (0.1)	45.7 (3.6)	3.0 (0.0)
		wL_1, PMR	0.208 (0.1)	46.6 (3.1)	3.0 (0.0)
		Oracle	0.202 (0.1)	48.0 (0.0)	3.0 (0.0)
	500	SCAD	0.443 (0.4)	495.8 (4.0)	3.0 (0.2)
		L_1	0.587 (0.2)	477.6 (12.5)	3.0 (0.0)
		wL_1, MR	0.297 (0.1)	494.9 (3.7)	3.0 (0.0)
		wL_1, PMR	0.215 (0.1)	496.5 (3.1)	3.0 (0.0)
		Oracle	0.193 (0.1)	498.0 (0.0)	3.0 (0.0)

Study 1 (*raw data without outliers*). For simulated data without contamination, the results are summarized in Table 3. The robust estimates perform as well as the non-robust counterparts, with respect to parameter estimation, variable selection and classification accuracy. Indeed, the optimal Bayes rule gives misclassification rates 0.137 for $p_n = 50$, and 0.138 for $p_n = 500$. Thus, the choice of loss functions has an asymptotically relatively negligible impact on classification performance. This agrees with results of Theorem 9 on the asymptotic classification consistency.

Study 2 (*contaminated data with outliers*). For each data set generated from the model, we create a contaminated data set. The contamination scheme is to replace the original 8 data points $(X_{i,j}, Y_i)$ by $(X^*_{i,j}, Y^*_i)$, where $Y^*_i = 1 - Y_i$, $i = 1, \ldots, 8$,

$$
\begin{aligned}
&X^*_{1,1} = 3\ \mathrm{sign}(U_1 - 0.5), \quad X^*_{2,1} = 3\ \mathrm{sign}(U_2 - 0.5), \quad X^*_{3,1} = 3\ \mathrm{sign}(U_3 - 0.5),\\
&X^*_{4,3} = 3\ \mathrm{sign}(U_4 - 0.5), \quad X^*_{5,5} = 3\ \mathrm{sign}(U_5 - 0.5), \quad X^*_{6,9} = 3\ \mathrm{sign}(U_6 - 0.5),\\
&X^*_{7,9} = 3\ \mathrm{sign}(U_7 - 0.5),
\end{aligned}
$$

with $\{U_i\} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0, 1)$. Table 4 summarizes the results over 500 sets of contaminated data. Regarding the robustness-efficiency tradeoff, analogous conclusions to Sect. 6.1 can be reached. Moreover, comparing Tables 3 and 4 reveals that (i) contamination increases the misclassification rates of all 5 methods; (ii) for contaminated cases in Table 4, robust procedures tend to reduce the misclassification rates; (iii) for robust estimation, the deviance loss is computationally

Table 2 (Simulation study: overdispersed Poisson responses, $n = 100$) Summary results for Study 2 (contaminated data with outliers)

Procedure	p_n	Method	Regression	Variable selection	
			EE($\widehat{\widehat{\boldsymbol{\beta}}}$) (sd)	C-Z (sd)	C-NZ (sd)
non-robust	50	SCAD	1.955 (0.3)	40.9 (4.6)	2.9 (0.3)
		L_1	2.012 (0.2)	40.2 (5.2)	2.9 (0.3)
		wL_1, MR	1.912 (0.3)	44.3 (3.7)	2.8 (0.4)
		wL_1, PMR	1.846 (0.3)	45.8 (3.4)	2.7 (0.4)
		Oracle	1.455 (0.2)	48.0 (0.0)	3.0 (0.0)
	500	SCAD	2.246 (0.2)	482.4 (12.4)	2.7 (0.5)
		L_1	2.292 (0.2)	484.8 (12.3)	2.6 (0.5)
		wL_1, MR	2.194 (0.2)	493.1 (6.1)	2.4 (0.5)
		wL_1, PMR	2.105 (0.2)	495.6 (5.2)	2.3 (0.5)
		Oracle	1.475 (0.2)	498.0 (0.0)	3.0 (0.0)
robust	50	SCAD	0.309 (0.2)	47.6 (1.1)	3.0 (0.0)
		L_1	0.689 (0.2)	39.0 (6.1)	3.0 (0.0)
		wL_1, MR	0.603 (0.3)	43.6 (4.1)	3.0 (0.1)
		wL_1, PMR	0.558 (0.3)	44.3 (4.3)	3.0 (0.1)
		Oracle	0.242 (0.1)	48.0 (0.0)	3.0 (0.0)
	500	SCAD	0.799 (0.5)	494.5 (3.2)	2.9 (0.3)
		L_1	1.093 (0.3)	481.8 (11.7)	3.0 (0.0)
		wL_1, MR	1.037 (0.5)	491.6 (5.4)	2.9 (0.3)
		wL_1, PMR	0.996 (0.5)	491.6 (5.5)	2.9 (0.3)
		Oracle	0.255 (0.1)	498.0 (0.0)	3.0 (0.0)

more stable yielding relatively lower misclassification rates than the exponential loss, see also Fig. 1, and thus the deviance loss is recommended for practical applications.

For the penalized methods, the tuning parameter λ_n is searched in the interval [0.0034, 0.1975], and κ_n is searched in the interval $[1/2^8, 1/2^5]$. To compare the classification performance with methods such as the classical-SVM and robust-SVM (using either the linear or Gaussian kernel, combined with auxiliary parameters c^* and/or s) in Wu and Liu (2007), Table 5 summarizes the results under the same set-ups as in Tables 3 and 4. Compared with classical- and robust-SVMs (in Table 5), the robust-BD method (in Tables 3, 4) clearly lowers MCRs.

6.3 Gaussian responses

To further illustrate the benefits of the robust method relative to the non-robust method, Appendix 1.2 gives additional simulation studies for Gaussian responses.

7 Real data application

To illustrate the application of the penalized *robust*-BD methods for classifying high-dimension low sample size data, we consider the Lymphoma data studied in Alizadeh et al. (2000), which identified two molecularly distinct forms of diffuse large B-cell Lymphoma

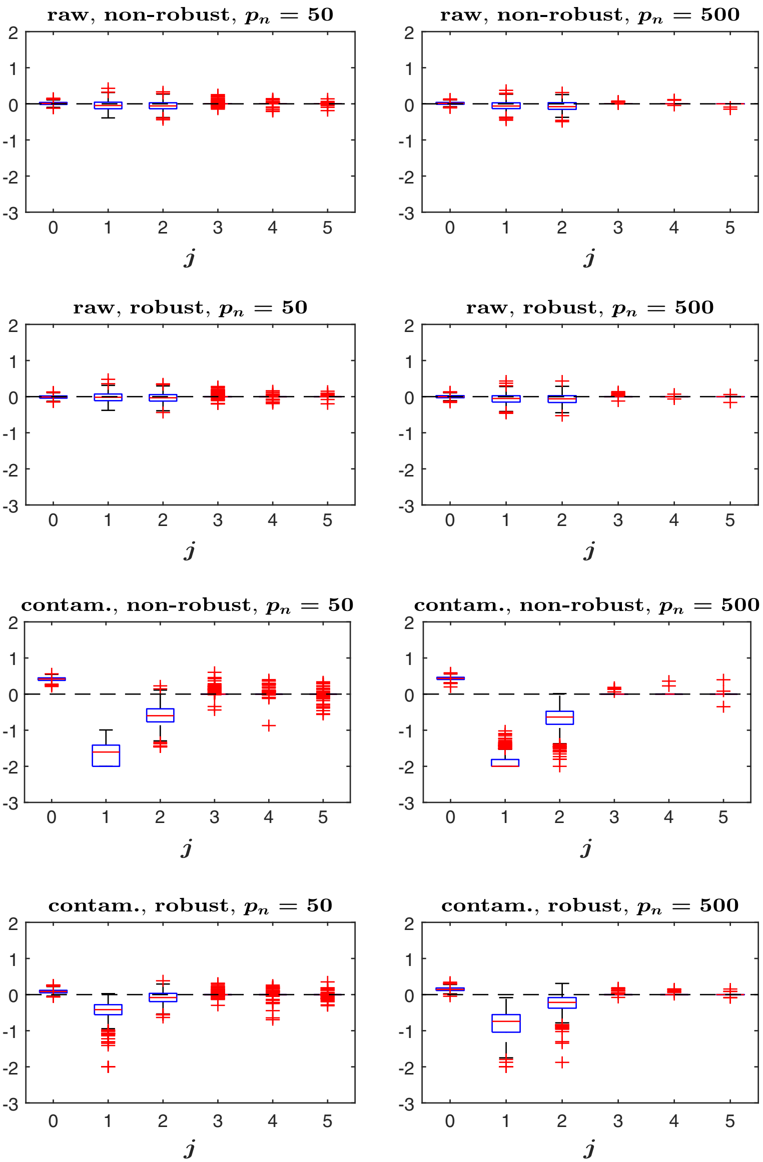


Fig. 2 (Simulation study: overdispersed Poisson responses, $n = 100, p_n = 50$ (left panel) and $p_n = 500$ (right panel)) Boxplots of $(\widehat{\beta}_j - \beta_{j;0})$, $j = 0, 1, \ldots, 5$, corresponding to results in Tables 1 and 2, using the PMR selection method for penalty weights in the weighted-L_1 penalty. The first row: raw data and using non-robust method; the second row: raw data and using robust method; the third row: contaminated data and using non-robust method; the fourth row: contaminated data and using robust method

(DLBCL). These two forms, called "germinal centre B-like (gc-B)" DLBCL and "activated B-like (a-B)" DLBCL, had gene expression patterns indicative of different stages of B-cell differentiation.

Table 3 (Simulation study: Bernoulli responses, $n = 200$) Summary results for Study 1 (raw data without outliers)

				Regression	Variable selection		Classification
Procedure	loss	p_n	Method	EE($\widehat{\widehat{\boldsymbol{\beta}}}$) (sd)	C-Z (sd)	C-NZ (sd)	MCR
non-robust	Dev	50	SCAD	0.574 (0.3)	47.5 (1.0)	3.0 (0.0)	0.141
			L_1	1.071 (0.3)	36.4 (5.7)	3.0 (0.0)	0.151
			wL_1, MR	0.654 (0.3)	45.8 (2.7)	3.0 (0.0)	0.144
			wL_1, PMR	0.628 (0.2)	45.9 (2.2)	3.0 (0.0)	0.143
			Oracle	0.553 (0.3)	48.0 (0.0)	3.0 (0.0)	0.141
		500	SCAD	0.580 (0.4)	497.5 (1.3)	3.0 (0.0)	0.142
			L_1	1.422 (0.2)	473.0 (13.0)	3.0 (0.0)	0.161
			wL_1, MR	0.855 (0.3)	493.7 (4.7)	3.0 (0.0)	0.146
			wL_1, PMR	0.829 (0.3)	494.1 (4.8)	3.0 (0.0)	0.146
			Oracle	0.569 (0.4)	498.0 (0.0)	3.0 (0.0)	0.141
	Exp	50	SCAD	0.740 (0.5)	47.7 (0.8)	3.0 (0.0)	0.142
			L_1	0.843 (0.2)	38.3 (5.5)	3.0 (0.0)	0.151
			wL_1, MR	0.638 (0.3)	45.7 (2.7)	3.0 (0.0)	0.144
			wL_1, PMR	0.635 (0.3)	46.0 (2.5)	3.0 (0.0)	0.144
			Oracle	0.704 (0.4)	48.0 (0.0)	3.0 (0.0)	0.141
		500	SCAD	0.740 (0.5)	497.5 (1.1)	3.0 (0.0)	0.142
			L_1	1.054 (0.3)	478.3 (11.2)	3.0 (0.0)	0.159
			wL_1, MR	0.723 (0.3)	493.9 (4.9)	3.0 (0.0)	0.147
			wL_1, PMR	0.708 (0.3)	494.5 (4.5)	3.0 (0.0)	0.146
			Oracle	0.719 (0.5)	498.0 (0.0)	3.0 (0.0)	0.142
robust	Dev	50	SCAD	0.669 (0.5)	47.9 (0.3)	3.0 (0.0)	0.142
			L_1	1.367 (0.2)	36.7 (4.5)	3.0 (0.0)	0.151
			wL_1, MR	0.951 (0.3)	47.6 (0.6)	3.0 (0.0)	0.143
			wL_1, PMR	0.959 (0.3)	47.7 (0.5)	3.0 (0.0)	0.143
			Oracle	0.685 (0.5)	48.0 (0.0)	3.0 (0.0)	0.142
		500	SCAD	0.661 (0.5)	498.0 (0.3)	3.0 (0.0)	0.142
			L_1	2.016 (0.1)	493.3 (2.5)	3.0 (0.0)	0.159
			wL_1, MR	1.721 (0.2)	498.0 (0.1)	3.0 (0.0)	0.151
			wL_1, PMR	1.734 (0.2)	498.0 (0.1)	3.0 (0.0)	0.151
			Oracle	0.679 (0.5)	498.0 (0.0)	3.0 (0.0)	0.142
	Exp	50	SCAD	0.578 (0.3)	47.9 (0.3)	3.0 (0.0)	0.141
			L_1	1.394 (0.2)	43.5 (2.1)	3.0 (0.0)	0.149
			wL_1, MR	1.084 (0.3)	48.0 (0.2)	3.0 (0.0)	0.145
			wL_1, PMR	1.099 (0.3)	48.0 (0.2)	3.0 (0.0)	0.145
			Oracle	0.593 (0.4)	48.0 (0.0)	3.0 (0.0)	0.141
		500	SCAD	0.570 (0.4)	498.0 (0.0)	3.0 (0.0)	0.141
			L_1	2.405 (0.1)	498.0 (0.1)	3.0 (0.1)	0.213
			wL_1, MR	2.045 (0.2)	498.0 (0.0)	3.0 (0.1)	0.174
			wL_1, PMR	2.065 (0.2)	498.0 (0.0)	3.0 (0.1)	0.176
			Oracle	0.602 (0.4)	498.0 (0.0)	3.0 (0.0)	0.142

Table 4 (Simulation study: Bernoulli responses, $n = 200$) Summary results for Study 2 (contaminated data with outliers)

				Regression	Variable selection		Classification
Procedure	loss	p_n	Method	$\mathrm{EE}(\widehat{\widehat{\boldsymbol{\beta}}})$ (sd)	C-Z (sd)	C-NZ (sd)	MCR
non-robust	Dev	50	SCAD	1.106 (0.3)	47.3 (1.4)	3.0 (0.0)	0.145
			L_1	1.768 (0.2)	38.9 (5.5)	3.0 (0.0)	0.162
			wL_1, MR	1.479 (0.3)	45.5 (2.8)	3.0 (0.0)	0.151
			wL_1, PMR	1.464 (0.3)	45.7 (2.7)	3.0 (0.0)	0.151
			Oracle	1.101 (0.3)	48.0 (0.0)	3.0 (0.0)	0.144
		500	SCAD	1.112 (0.3)	497.0 (1.8)	3.0 (0.0)	0.145
			L_1	2.027 (0.2)	479.4 (12.1)	3.0 (0.0)	0.177
			wL_1, MR	1.658 (0.3)	493.9 (4.3)	3.0 (0.0)	0.157
			wL_1, PMR	1.634 (0.3)	494.2 (3.8)	3.0 (0.0)	0.156
			Oracle	1.107 (0.3)	498.0 (0.0)	3.0 (0.0)	0.144
	Exp	50	SCAD	1.507 (0.4)	47.3 (1.2)	3.0 (0.1)	0.156
			L_1	1.821 (0.3)	39.8 (5.8)	3.0 (0.0)	0.173
			wL_1, MR	1.685 (0.4)	45.2 (3.0)	3.0 (0.0)	0.164
			wL_1, PMR	1.669 (0.4)	45.6 (2.8)	3.0 (0.0)	0.163
			Oracle	1.504 (0.4)	48.0 (0.0)	3.0 (0.0)	0.152
		500	SCAD	1.521 (0.5)	497.2 (1.9)	3.0 (0.2)	0.161
			L_1	1.927 (0.3)	481.6 (11.7)	3.0 (0.0)	0.187
			wL_1, MR	1.757 (0.4)	492.8 (5.9)	3.0 (0.0)	0.172
			wL_1, PMR	1.732 (0.4)	493.6 (5.2)	3.0 (0.0)	0.170
			Oracle	1.498 (0.4)	498.0 (0.0)	3.0 (0.0)	0.153
robust	Dev	50	SCAD	0.682 (0.4)	47.9 (0.5)	3.0 (0.0)	0.143
			L_1	1.686 (0.3)	37.6 (4.7)	3.0 (0.0)	0.157
			wL_1, MR	1.395 (0.3)	47.4 (0.7)	3.0 (0.0)	0.146
			wL_1, PMR	1.401 (0.3)	47.5 (0.7)	3.0 (0.0)	0.146
			Oracle	0.681 (0.4)	48.0 (0.0)	3.0 (0.0)	0.143
		500	SCAD	0.681 (0.3)	498.0 (0.2)	3.0 (0.0)	0.143
			L_1	2.208 (0.1)	491.3 (3.0)	3.0 (0.0)	0.168
			wL_1, MR	2.095 (0.2)	498.0 (0.2)	3.0 (0.1)	0.164
			wL_1, PMR	2.105 (0.2)	498.0 (0.2)	3.0 (0.1)	0.165
			Oracle	0.685 (0.4)	498.0 (0.0)	3.0 (0.0)	0.143
	yExp	50	SCAD	0.976 (0.3)	47.9 (0.5)	3.0 (0.0)	0.144
			L_1	1.818 (0.2)	42.4 (2.4)	3.0 (0.0)	0.157
			wL_1, MR	1.766 (0.3)	47.9 (0.3)	3.0 (0.0)	0.155
			wL_1, PMR	1.778 (0.3)	47.9 (0.3)	3.0 (0.0)	0.155
			Oracle	0.804 (0.3)	48.0 (0.0)	3.0 (0.0)	0.143
		500	SCAD	1.049 (0.4)	498.0 (0.0)	3.0 (0.2)	0.146
			L_1	2.578 (0.1)	498.0 (0.1)	3.0 (0.1)	0.232
			wL_1, MR	2.545 (0.2)	498.0 (0.0)	2.8 (0.4)	0.226
			wL_1, PMR	2.559 (0.2)	498.0 (0.0)	2.8 (0.4)	0.228
			Oracle	0.788 (0.3)	498.0 (0.0)	3.0 (0.0)	0.143

Table 5 (Simulation study: Bernoulli responses, $n = 200$) Compare MCR using classical-SVM and robust-SVM for Study 1 (raw data without outliers) in Table 3 and Study 2 (contaminated data with outliers) in Table 4

Data	p_n	Method	MCR
Raw	50	classical-SVM, linear kernel, $c^* = 0.1$	0.203
		classical-SVM, linear kernel, $c^* = 10$	0.227
		classical-SVM, Gaussian kernel, $c^* = 0.1$	0.281
		classical-SVM, Gaussian kernel, $c^* = 10$	0.208
		robust-SVM, linear kernel, $c^* = 0.1$, $s = -0.5$	0.225
		robust-SVM, linear kernel, $c^* = 0.1$, $s = 0$	0.230
		robust-SVM, linear kernel, $c^* = 10$, $s = -0.5$	0.232
		robust-SVM, linear kernel, $c^* = 10$, $s = 0$	0.233
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = -0.5$	0.281
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = 0$	0.281
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = -0.5$	0.217
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = 0$	0.245
	500	classical-SVM, linear kernel, $c^* = 0.1$	0.305
		classical-SVM, linear kernel, $c^* = 10$	0.305
		classical-SVM, Gaussian kernel, $c^* = 0.1$	0.281
		classical-SVM, Gaussian kernel, $c^* = 10$	0.278
		robust-SVM, linear kernel, $c^* = 0.1$, $s = -0.5$	0.295
		robust-SVM, linear kernel, $c^* = 0.1$, $s = 0$	0.292
		robust-SVM, linear kernel, $c^* = 10$, $s = -0.5$	0.295
		robust-SVM, linear kernel, $c^* = 10$, $s = 0$	0.293
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = -0.5$	0.281
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = 0$	0.281
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = -0.5$	0.279
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = 0$	0.281

The publicly available dataset contains 4026 genes across 47 samples, of which 24 are "gc-B" and 23 are "a-B". We use the 10-nearest neighbor method to impute the missing expression data. After imputing, the dataset is standardized to zero mean and unit variance across genes. We intend to predict whether a sample can be categorized as "gc-B" or "a-B".

To evaluate the performance of the penalized estimates of parameters in the model, $\text{logit}\{\text{P}(Y = 1 \mid X_1, \ldots, X_{4026})\} = \beta_0 + \sum_{j=1}^{4026} \beta_j X_j$, we randomly split the data into a training set with 31 samples (containing 16 cases of "gc-B" and 15 cases of "a-B") and a test set with 16 samples (containing 8 cases of "gc-B" and 8 cases of "a-B"). For each training set, λ_n is selected by minimizing a 3-fold cross validated estimate of the misclassification rate; λ_n and κ_n for the proposed PMR method are searched on a surface of grid points. The test error (TE) gives the misclassification rate of the penalized classifier to the test set. Both the Huber ψ-function (with $c = 1.345$) and Tukey ψ-function (with $c = 4.685$) are utilized in the robust estimates, and the weight function $w(\boldsymbol{x})$ is the same as used in Sect. 6. Table 6 tabulates the average of TE and the average number of selected genes over 100 random splits. The penalized estimates/classifiers induced by the deviance loss and the exponential loss yield similar performance. The L_1 penalty selects approximately twice as many genes as the other penalty choices. On the basis of TE and sparse modeling

Table 5 continued

Data	p_n	Method	MCR
Contaminated	50	classical-SVM, linear kernel, $c^* = 0.1$	0.219
		classical-SVM, linear kernel, $c^* = 10$	0.245
		classical-SVM, Gaussian kernel, $c^* = 0.1$	0.281
		classical-SVM, Gaussian kernel, $c^* = 10$	0.217
		robust-SVM, linear kernel, $c^* = 0.1$, $s = -0.5$	0.237
		robust-SVM, linear kernel, $c^* = 0.1$, $s = 0$	0.241
		robust-SVM, linear kernel, $c^* = 10$, $s = -0.5$	0.243
		robust-SVM, linear kernel, $c^* = 10$, $s = 0$	0.246
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = -0.5$	0.281
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = 0$	0.281
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = -0.5$	0.221
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = 0$	0.244
	500	classical-SVM, linear kernel, $c^* = 0.1$	0.322
		classical-SVM, linear kernel, $c^* = 10$	0.322
		classical-SVM, Gaussian kernel, $c^* = 0.1$	0.282
		classical-SVM, Gaussian kernel, $c^* = 10$	0.279
		robust-SVM, linear kernel, $c^* = 0.1$, $s = -0.5$	0.307
		robust-SVM, linear kernel, $c^* = 0.1$, $s = 0$	0.302
		robust-SVM, linear kernel, $c^* = 10$, $s = -0.5$	0.306
		robust-SVM, linear kernel, $c^* = 10$, $s = 0$	0.302
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = -0.5$	0.282
		robust-SVM, Gaussian kernel, $c^* = 0.1$, $s = 0$	0.282
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = -0.5$	0.280
		robust-SVM, Gaussian kernel, $c^* = 10$, $s = 0$	0.282

simultaneously, the robust estimate combined with the weighted-L_1 penalty appears to perform the best. The results also reveal that the choice of the ψ-functions in (6) has a negligible impact on the performance of the robust penalized estimates.

8 Discussion

The conventional penalized least-squares and penalized-likelihood estimates of parameters exhibit the oracle property for sparsity recovery but lack resistance to outlying observations. This paper proposes a class of robust error measures called "*robust*-BD" and introduces the "*penalized robust*-BD *estimate*". The "*robust*-BD" induces a bounded influence function that makes the resulting penalized estimation less sensitive to outliers. Since BD is widely used in machine learning practice, the proposed "*penalized robust*-BD *estimate*" combined with suitably chosen weights for penalties can be broadly applicable in regression and classification problems of large dimensions ($p_n \approx n$) and high dimensions ($p_n \gg n$), achieving both the oracle property and the robustness to outliers in either the covariate space or the response space.

There are several limitations to our study and ongoing challenges that should be considered. Firstly, in the current scope of the paper, technical conditions, particularly A1, A2,

Table 6 (Real data) Classification for the Lymphoma data

Procedure	$\psi(r)$	Penalty	Deviance loss		Exponential loss	
			TE	# genes	TE	# genes
non-robust	r	SCAD	0.213	6.32	0.205	10.47
		L_1	0.121	14.42	0.121	12.56
		wL_1, MR	0.136	6.68	0.141	5.51
		wL_1, PMR	0.123	7.30	0.128	6.19
robust	Huber	SCAD	0.211	3.35	0.201	3.61
		L_1	0.123	10.63	0.122	9.19
		wL_1, MR	0.160	4.48	0.156	4.21
		wL_1, PMR	0.141	5.26	0.142	4.72
robust	Tukey	SCAD	0.170	3.73	0.175	4.09
		L_1	0.126	10.95	0.125	9.14
		wL_1, MR	0.151	4.78	0.155	4.46
		wL_1, PMR	0.141	5.39	0.143	5.20

A5 and B3 in Appendix 1.1 exclude heavy-tailed covariates X_j and responses Y. Similarly, the proposed method does not handle very high levels of contamination in the data. Moreover, relaxing (19) could be pursued in a separate study. Secondly, Algorithm 1 in Appendix 1.3.2 shows that the computational complexity depends on various key factors, including the sample size (n), dimensionality (p_n), the response variable type (Gaussian, Bernoulli, or Poisson), selected penalty weights ($\{\widehat{w}_{n,j}\}_{j=1}^{p_n}$), proportion of contamination in the dataset, approximation accuracy of the quadratic function to the loss function in (3), and convergence property of the coordinate-descent (CD) algorithm (Friedman et al., 2010). In particular, the quadratic approximation may reduce the accuracy of the "*penalized robust*-BD *estimator*" and slow down the convergence of the resulting CD algorithm. This is demonstrated in numerical experiments that indicate an increase in CPU runtime as the response type changes from Gaussian to Bernoulli to Poisson, and as non-robust procedures using classical-BD are replaced with robust counterparts using "*robust*-BD". Despite these limitations, our work contributes to the "*robust*-BD" estimation of p_n-dimensional parameters with some provable results when $p_n \approx n$ and $p_n \gg n$, and thus fills a gap in the literature. The goal of our work is to better understand the flexibility, challenges, and limitations of *robust*-BD estimation as a data-analytic tool for big data analysis.

A number of open questions need to be discussed. (a) In the linear regression model, the quantification of the robustness property of estimates in terms of gross error sensitivity, rejection point, or local-shift sensitivity has been studied. However, beyond the linear model, such as the GLM, relatively little theoretical work has been done about this property, even in the case of fixed dimensions $p_n = p$. Rigorously exploring the robustness property, including the breakdown point, for the proposed class of "*penalized robust*-BD *estimates*" in the current model, assuming (1)–(2), remains a nontrivial task. (b) The existence of consistent local solutions of penalized estimates has also appeared in several other existing works, such as the nonconcave penalized likelihood method in Fan and Peng (2004). Devising efficient numerical procedures for obtaining a local solution that is consistent will be desirable but challenging. See Gong et al. (2013) for recent progress made in addressing this issue. (c) For $p_n \gg n$, it is of interest to explore some variable screening procedure for dimension reduction before applying the proposed "*penalized robust*-BD" estimation method. (d) The weight function $w(\boldsymbol{x})$ used in the numerical evaluations in Sects. 6 and 7 is a feasible choice for *robust*-BD estimation in large ($p_{_n} \approx n$) and high ($p_{_n} \gg n$) dimensions. However, an optimal method for selecting the weight function would be desirable, depending on specific criteria, such as the robustness property, which has yet to be explored. In low dimensions ($p_{_n} < n$), the weight function (e.g., $w(\boldsymbol{x}) = 1/\{1 + (\boldsymbol{x} - \widehat{\boldsymbol{m}})^T \widehat{\Sigma}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{m}})\}^{1/2}$) can rely on robust estimates $\widehat{\boldsymbol{m}}$ and $\widehat{\Sigma}$ of the location vector and scatter matrix of $\boldsymbol{X}$, and pages 137–138 of Heritier et al. (2009) suggest alternative weight functions for robust estimators of linear model parameters in fixed ($p_{_n} = p$) dimensions. (e) Finite sample other than asymptotic results may be obtained for certain types of covariates X_j and responses Y under more stringent assumptions. A complete and thorough study of these theoretical/methodological development and computational advancement is beyond the scope of the current paper and needs to be examined in future research.

9 Supplementary information

Appendices 1.1, 1.2 and 1.3 collect proofs of Theorems 1 to 9, Figs. 1 and 2 and Tables 1, 2, 3, 4, 5 and 6, additional numerical studies (Figs. 3 and 4 and Tables 7 and 8 in Sect. 6.3; real data analysis), and algorithmic details, respectively.

Declarations

Conflict of interest 'Not applicable'.

Ethics approval 'Not applicable'.

Consent to participate 'Not applicable'.

Consent for publication All authors of the paper consent for publication.

Supplementary Appendix

1. Proofs, figures and tables, algorithm

Notations and symbols.

For a vector $\boldsymbol{a} = (a_1, \ldots, a_d)^T$, $\|\boldsymbol{a}\|_1 = \sum_{j=1}^d |a_j|$, $\|\boldsymbol{a}\|_2 = (\sum_{j=1}^d a_j^2)^{1/2}$ and $\|\boldsymbol{a}\|_\infty = \max_{1 \le j \le d} |a_j|$. Let $\mathbf{I}_{\mathsf{k}}$ denote a $\mathsf{k} \times \mathsf{k}$ identity matrix, and $\mathbf{0}_{p,q}$ denote a $p \times q$ matrix of zero entries. For a matrix M, its eigenvalues, minimum eigenvalue, maximum eigenvalue are labeled by $\lambda_j(M)$, $\lambda_{\min}(M)$, $\lambda_{\max}(M)$ respectively; $\mathrm{tr}(M)$ denotes the trace of a square matrix M; let $\|M\| = \|M\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|M\boldsymbol{x}\|_2 = \{\lambda_{\max}(M^T M)\}^{1/2}$ be the matrix L_2 norm, and $\|M\|_F = \{\mathrm{tr}(M^T M)\}^{1/2}$ be the Frobenius norm. Throughout the proof, C is used as a generic finite constant. The sign function $\mathrm{sign}(x)$ equals $+1$ if $x > 0$, 0 if $x = 0$, and -1 if $x < 0$. For a function $g(x)$, the first-order derivative is $g'(x)$ or $g^{(1)}(x)$, the second-order derivative is $g''(x)$ or $g^{(2)}(x)$, and the jth other derivative is $g^{(j)}(x)$. The chi-squared distribution with k degrees of freedom is denoted by χ_k^2.

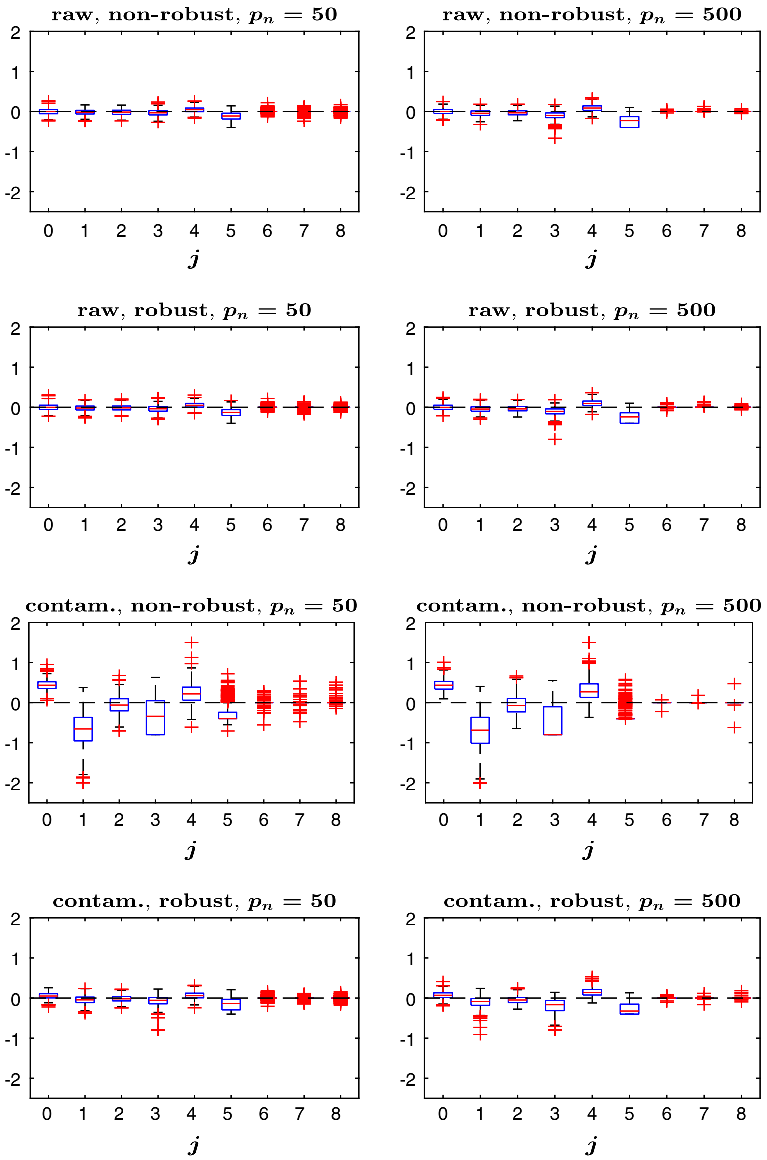
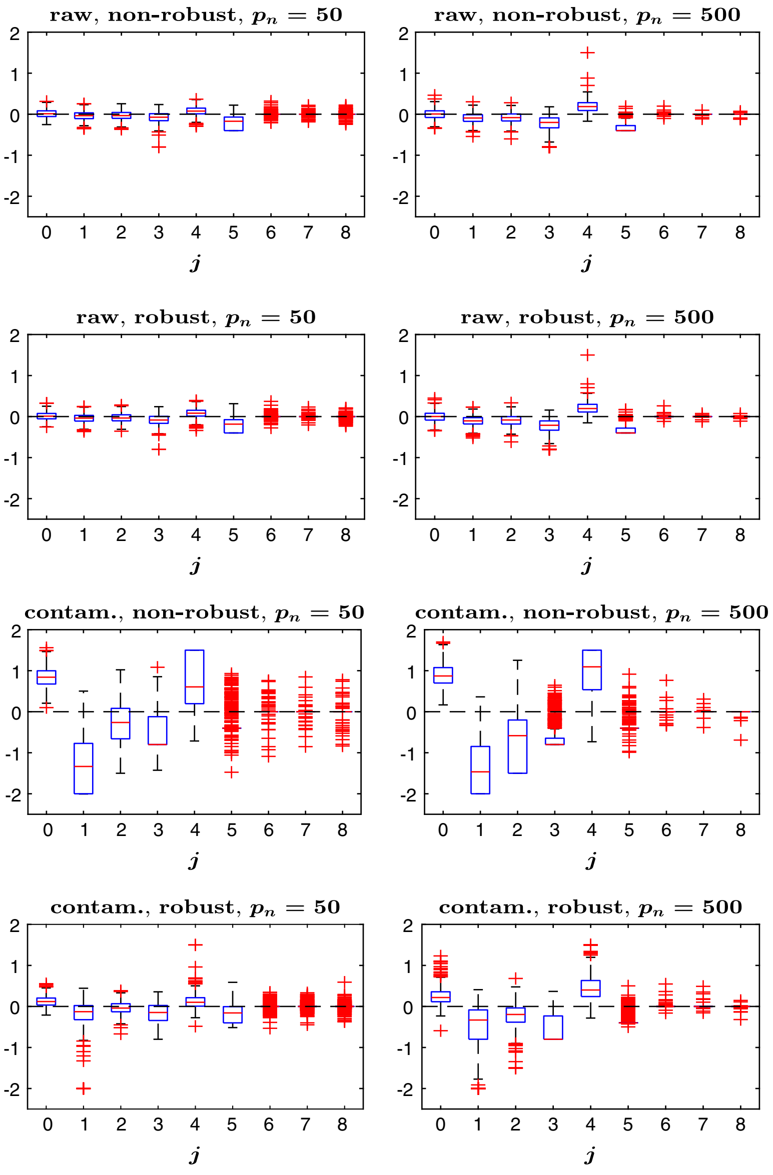


Fig. 3 (Simulation study: Gaussian responses, $n = 200$, $p_n = 50$ (left panel) and $p_n = 500$ (right panel)) Boxplots of $(\widehat{\beta}_j - \beta_{j;0}), j = 0, 1, \ldots, 8,$ corresponding to results in Tables 7 and 8, using the PMR selection method for penalty weights in the weighted-L_1 penalty. The first row: raw data and using non-robust method; the second row: raw data and using robust method; the third row: contaminated data and using non-robust method; the fourth row: contaminated data and using robust method

The conditional expectation and condition variance of Y given $\boldsymbol{X}$ are denoted by $\mathrm{E}(Y \mid \boldsymbol{X})$ and $\mathrm{var}(Y \mid \boldsymbol{X})$ respectively. Notations in the asymptotic derivations follow (van der Vaart, 1998), where $\xrightarrow{\mathrm{P}}$ denotes converges in probability, $\xrightarrow{\mathcal{L}}$ means converges in distribution, $o_{\mathrm{P}}(1)$ is a term which converges to zero in probability, and $O_{\mathrm{P}}(1)$ is a term which is bounded in probability.


raw, non-robust, $p_n = 50$

raw, non-robust, $p_n = 500$

raw, robust, $p_n = 50$

raw, robust, $p_n = 500$

contam., non-robust, $p_n = 50$

contam., non-robust, $p_n = 500$

contam., robust, $p_n = 50$

contam., robust, $p_n = 500$

Fig. 4 (Simulation study: Gaussian responses, $n = 100$, $p_n = 50$ (left panel) and $p_n = 500$ (right panel)) The caption is similar to that in Fig. 3, except $n = 100$

1.1. Proofs of Theorems 1 up to 9

We first impose some regularity conditions, which are not the weakest possible but facilitate the technical derivations.

Condition A.

A0. $s_n \geq 1$ and $p_n - s_n \geq 1$. $\sup_{n \geq 1} \|\boldsymbol{\beta}_0^{(\mathrm{I})}\|_1 < \infty$.

Table 7 (Simulation study: Gaussian responses, $n = 200$) Summary results for Study 1 (raw data without outliers)

Procedure	p_n	Method	Regression	Variable selection	
			$\mathrm{EE}(\widehat{\widetilde{\boldsymbol{\beta}}})$ (sd)	C-Z (sd)	C-NZ (sd)
non-robust	50	SCAD	0.236 (0.1)	38.0 (5.9)	6.0 (0.0)
		L_1	0.316 (0.1)	32.5 (6.5)	6.0 (0.0)
		wL_1, MR	0.276 (0.1)	37.7 (6.9)	5.9 (0.2)
		wL_1, PMR	0.279 (0.1)	38.8 (6.3)	5.9 (0.3)
		Oracle	0.166 (0.1)	45.0 (0.0)	6.0 (0.0)
	500	SCAD	0.305 (0.1)	473.1 (9.0)	6.0 (0.1)
		L_1	0.439 (0.1)	469.5 (11.5)	6.0 (0.0)
		wL_1, MR	0.387 (0.1)	485.0 (8.8)	5.7 (0.4)
		wL_1, PMR	0.378 (0.1)	487.0 (8.4)	5.7 (0.5)
		Oracle	0.172 (0.1)	495.0 (0.0)	6.0 (0.0)
robust	50	SCAD	0.187 (0.1)	44.6 (1.1)	6.0 (0.1)
		L_1	0.329 (0.1)	32.0 (7.3)	6.0 (0.0)
		wL_1, MR	0.287 (0.1)	38.2 (6.2)	5.9 (0.3)
		wL_1, PMR	0.289 (0.1)	39.4 (5.5)	5.9 (0.4)
		Oracle	0.176 (0.1)	45.0 (0.0)	6.0 (0.0)
	500	SCAD	0.239 (0.1)	494.7 (1.5)	5.8 (0.4)
		L_1	0.467 (0.1)	466.9 (19.9)	6.0 (0.1)
		wL_1, MR	0.397 (0.1)	484.4 (10.4)	5.7 (0.4)
		wL_1, PMR	0.390 (0.1)	486.0 (9.8)	5.7 (0.5)
		Oracle	0.185 (0.1)	495.0 (0.0)	6.0 (0.0)

A1. $\|\boldsymbol{X}\|_\infty = \max_{1\le j\le p_n} |X_j|$ is bounded almost surely.

A2. $\mathrm{E}(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)$ exists and is nonsingular in the case of $p_n + 1 \le n$; $\mathrm{E}\{\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\}$ exists and is nonsingular in the case of $p_n + 1 > n$.

A4. There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_0$, such that $F^{-1}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}})$ is bounded almost surely for all $\widetilde{\boldsymbol{\beta}}$ in the subset.

A5. $w(\cdot) \ge 0$ is a bounded function. Assume that $\psi(r)$ is a bounded, odd function, and twice differentiable, such that $\psi'(r)$, $\psi'(r)r$, $\psi''(r)$, $\psi''(r)r$ and $\psi''(r)r^2$ are bounded; $V(\cdot) > 0$, $V^{(2)}(\cdot)$ is continuous. The matrix $\mathbf{H}_n^{(\mathrm{I})}$ is positive definite, with eigenvalues uniformly bounded away from zero.

A5′. $w(\cdot) \ge 0$ is a bounded function.

A6. $q^{(4)}(\cdot)$ is continuous, and $q^{(2)}(\cdot) < 0$. $g_1^{(2)}(\cdot)$ is continuous.

A7. $F(\cdot)$ is monotone and a bijection, $F^{(3)}(\cdot)$ is continuous, and $F^{(1)}(\cdot) \ne 0$.

Condition B.

B3.

There exists a constant $C \in (0, \infty)$ such that $\sup_{n\ge 1} \mathrm{E}\{|Y - m(\boldsymbol{X})|^j\} \le j!C^j$ for all $j \ge 3$. Also, $\inf_{n\ge 1, 1\le j\le p_n} \mathrm{E}\{\mathrm{var}(Y \mid \boldsymbol{X})X_j^2\} > 0$.

Table 8 (Simulation study: Gaussian responses, $n = 200$) Summary results for Study 2 (contaminated data with outliers)

Procedure	p_n	Method	Regression	Variable selection	
			EE($\widehat{\widetilde{\boldsymbol{\beta}}}$) (sd)	C-Z (sd)	C-NZ (sd)
non-robust	50	SCAD	1.291 (0.4)	36.1 (6.5)	5.4 (0.6)
		L_1	1.341 (0.4)	35.4 (6.4)	5.5 (0.6)
		wL_1, MR	1.255 (0.4)	41.0 (4.8)	5.2 (0.7)
		wL_1, PMR	1.207 (0.3)	42.6 (4.1)	5.0 (0.7)
		Oracle	1.003 (0.3)	45.0 (0.0)	6.0 (0.0)
	500	SCAD	1.547 (0.4)	471.2 (21.6)	5.2 (0.7)
		L_1	1.593 (0.4)	476.9 (17.7)	5.1 (0.7)
		wL_1, MR	1.428 (0.4)	487.3 (7.8)	4.8 (0.7)
		wL_1, PMR	1.311 (0.4)	491.2 (7.7)	4.6 (0.7)
		Oracle	1.026 (0.3)	495.0 (0.0)	6.0 (0.0)
robust	50	SCAD	0.249 (0.1)	43.9 (3.5)	5.9 (0.2)
		L_1	0.373 (0.1)	31.3 (8.8)	6.0 (0.0)
		wL_1, MR	0.375 (0.1)	36.2 (7.4)	5.8 (0.4)
		wL_1, PMR	0.379 (0.1)	37.7 (6.5)	5.8 (0.4)
		Oracle	0.196 (0.1)	45.0 (0.0)	6.0 (0.0)
	500	SCAD	0.299 (0.1)	494.8 (1.2)	5.7 (0.4)
		L_1	0.527 (0.1)	466.0 (19.8)	6.0 (0.2)
		wL_1, MR	0.547 (0.2)	480.4 (12.5)	5.6 (0.5)
		wL_1, PMR	0.544 (0.2)	482.9 (12.1)	5.5 (0.6)
		Oracle	0.192 (0.1)	495.0 (0.0)	6.0 (0.0)

B5.

The matrices $\boldsymbol{\Omega}_n^{(\mathrm{I})}$ and $\mathbf{H}_n^{(\mathrm{I})}$ are positive definite, with eigenvalues uniformly bounded away from zero. Also, $\|(\mathbf{H}_n^{(\mathrm{I})})^{-1}\boldsymbol{\Omega}_n^{(\mathrm{I})}\|_2$ is bounded away from infinity.

Condition C.

C4.

There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_0$, such that $F^{-1}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\beta}})$ is bounded almost surely for all $\widetilde{\boldsymbol{\beta}}$ in the subset. Moreover, the subset contains the origin.

Condition D.

D5.

The eigenvalues of $\mathbf{H}_n^{(\mathrm{I})}$ are uniformly bounded away from zero. Also, $\|(\mathbf{H}_n^{(\mathrm{I})})^{-1/2}(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{1/2}\|_2$ is bounded away from infinity.

Condition E.

E1.
$\min_{1 \le j \le s_n} |\mathrm{cov}(X_j, Y)| \succeq \mathcal{A}_n$ and $\max_{s_n+1 \le j \le p_n} |\mathrm{cov}(X_j, Y)| = o(\mathcal{B}_n)$ for some positive sequences $\mathcal{A}_n$ and $\mathcal{B}_n$, where the symbol $s_n \succeq t_n$, for two nonnegative sequences s_n and t_n, means that there exists a constant $c > 0$ such that $s_n \ge c\, t_n$ for all $n \ge 1$.

E2.
$\sup_{n \ge 1, 1 \le j \le s_n} \mathrm{E}\{\mathrm{q}_2(Y; \alpha_0) X_j^2\} < \infty$; $\inf_{n \ge 1, s_n+1 \le j \le p_n} \mathrm{E}\{\mathrm{q}_2(Y; \alpha_0) X_j^2\} = \eta > 0$, where $\alpha_0 = F(\mathrm{E}(Y))$.

Proof of Theorem 1 We first need to show Lemma 1. □

Lemma 1 (*existence and consistency*: $p_n \ll n$) *Assume Conditions* A0, A1, A2, A4, A5, A6 *and* A7 *in Appendix* 1.1, *the matrix* $\mathbf{H}_n = \mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^{(\mathrm{I})T} \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}) \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T\}$ *is positive definite with eigenvalues uniformly bounded away from zero, and* $w_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n \sqrt{n} \sqrt{s_n/p_n})\}$. *If* $p_n^4/n \to 0$ *as* $n \to \infty$, *then there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* (3) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(\sqrt{p_n/n})$.

Proof We follow the idea of the proof of Theorem 1 in Fan and Peng (2004). Let $r_n = \sqrt{p_n/n}$ and $\widetilde{\boldsymbol{u}}_n = (u_0, u_1, \ldots, u_{p_n})^T \in \mathbb{R}^{p_n+1}$. It suffices to show that for any given $\epsilon > 0$, there exists a sufficiently large constant C_ϵ such that, for large n we have

$$\mathrm{P}\Big\{ \inf_{\|\widetilde{\boldsymbol{u}}_n\|_2 = C_\epsilon} \ell_n(\widetilde{\boldsymbol{\beta}}_0 + r_n \widetilde{\boldsymbol{u}}_n) > \ell_n(\widetilde{\boldsymbol{\beta}}_0) \Big\} \ge 1 - \epsilon. \tag{20}$$

This implies that with probability at least $1 - \epsilon$, there exists a local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}$ of $\ell_n(\widetilde{\boldsymbol{\beta}})$ in the ball $\{\widetilde{\boldsymbol{\beta}}_0 + r_n \widetilde{\boldsymbol{u}}_n : \|\widetilde{\boldsymbol{u}}_n\|_2 \le C_\epsilon\}$ such that $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(r_n)$. To show (20), consider

$$\begin{aligned} & D_n(\widetilde{\boldsymbol{u}}_n) \\ &= \frac{1}{n} \sum_{i=1}^n \{\rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_i^T (\widetilde{\boldsymbol{\beta}}_0 + r_n \widetilde{\boldsymbol{u}}_n)))\, w(\boldsymbol{X}_i) \\ &\quad - \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0))\, w(\boldsymbol{X}_i)\} \\ &\quad + \lambda_n \sum_{j=1}^{p_n} w_{n,j} (|\beta_{j;0} + r_n u_j| - |\beta_{j;0}|) \\ &\equiv I_1 + I_2, \end{aligned} \tag{21}$$

where $\|\widetilde{\boldsymbol{u}}_n\|_2 = C_\epsilon$.

First, we consider I_1. By Taylor's expansion, I_1 has the decomposition,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \tag{22}$$

where $I_{1,1} = r_n/n \sum_{i=1}^n \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i) \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n$, $I_{1,2} = r_n^2/(2n) \sum_{i=1}^n \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)$ $w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n)^2$, and

$I_{1,3} = r_n^3/(6n)\sum_{i=1}^n \mathrm{p}_3(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}^*)\, w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n)^3$ for $\widetilde{\boldsymbol{\beta}}^*$ located between $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\boldsymbol{\beta}}_0 + r_n\widetilde{\boldsymbol{u}}_n$. Hence

$$|I_{1,1}| \le r_n \left\| \frac{1}{n}\sum_{i=1}^n \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i \right\|_2 \|\widetilde{\boldsymbol{u}}_n\|_2 = O_{\mathrm{P}}(r_n\sqrt{p_n/n})\|\widetilde{\boldsymbol{u}}_n\|_2. \tag{23}$$

For the term $I_{1,2}$ in (22),

$$\begin{aligned} I_{1,2} &= \frac{r_n^2}{2n}\sum_{i=1}^n \mathrm{E}\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n)^2\} \\ &\quad + \frac{r_n^2}{2n}\sum_{i=1}^n \Big[\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n)^2 \\ &\quad - \mathrm{E}\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n)^2\}\Big] \\ &\equiv I_{1,2,1} + I_{1,2,2}, \end{aligned}$$

where $I_{1,2,1} = 2^{-1} r_n^2 \widetilde{\boldsymbol{u}}_n^T \mathbf{H}_n \widetilde{\boldsymbol{u}}_n$. Meanwhile, we have

$$\begin{aligned} |I_{1,2,2}| &\le r_n^2 \Big\| \frac{1}{n}\sum_{i=1}^n \Big[\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T \\ &\quad - \mathrm{E}\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\}\Big] \Big\|_F \|\widetilde{\boldsymbol{u}}_n\|_2^2 \\ &= r_n^2 O_{\mathrm{P}}(p_n/\sqrt{n})\|\widetilde{\boldsymbol{u}}_n\|_2^2. \end{aligned}$$

Thus,

$$I_{1,2} = \frac{r_n^2}{2}\widetilde{\boldsymbol{u}}_n^T \mathbf{H}_n \widetilde{\boldsymbol{u}}_n + O_{\mathrm{P}}(r_n^2 p_n/\sqrt{n})\|\widetilde{\boldsymbol{u}}_n\|_2^2. \tag{24}$$

For the term $I_{1,3}$ in (22), we observe that

$$|I_{1,3}| \le r_n^3 \frac{1}{n}\sum_{i=1}^n |\mathrm{p}_3(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}^*)|\, w(\boldsymbol{X}_i)|\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{u}}_n|^3 = O_{\mathrm{P}}(r_n^3 p_n^{3/2})\|\widetilde{\boldsymbol{u}}_n\|_2^3,$$

which follows from Conditions A0, A1, A4 and A5.

Next, we consider I_2 in (21). Note $I_2 = \lambda_n \sum_{j=1}^{s_n} w_{n,j}(|\beta_{j;0} + r_n u_j| - |\beta_{j;0}|) + \lambda_n r_n \sum_{j=s_n+1}^{p_n} w_{n,j}|u_j|$. Clearly, by the triangle inequality,

$$I_2 \ge -\lambda_n r_n \sum_{j=1}^{s_n} w_{n,j}|u_j| \equiv I_{2,1},$$

in which

$$|I_{2,1}| \le \lambda_n r_n w_{\max}^{(\mathrm{I})} \|\boldsymbol{u}_n^{(\mathrm{I})}\|_1, \tag{25}$$

where $\boldsymbol{u}_n^{(\mathrm{I})} = (u_1, \ldots, u_{s_n})^T$. By (23)–(25) and $p_n^4/n \to 0$, we can choose some large C_ϵ such that $I_{1,1}$, $I_{1,3}$ and $I_{2,1}$ are all dominated by the first term of $I_{1,2}$ in (24), which is positive by the eigenvalue assumption. This implies (20). □

We now show Theorem 1. Write $\widetilde{\boldsymbol{u}}_n=(\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})T},\boldsymbol{u}_n^{(\mathrm{II})T})^T$, where $\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}=(u_0,u_1,\ldots,u_{s_n})^T$ and $\boldsymbol{u}_n^{(\mathrm{II})}=(u_{s_n+1},\ldots,u_{p_n})^T$. Following the proof of Lemma 1, it suffices to show (20) for $r_n=\sqrt{s_n/n}$.

For the term $I_{1,1}$ in (22),

$$\begin{aligned}I_{1,1}=&\frac{r_n}{n}\sum_{i=1}^n \mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}_0)\,w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\\&+\frac{r_n}{n}\sum_{i=1}^n \mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}_0)\,w(\boldsymbol{X}_i)\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})}\equiv I_{1,1}^{(\mathrm{I})}+I_{1,1}^{(\mathrm{II})}.\end{aligned}$$

It follows that

$$|I_{1,1}^{(\mathrm{I})}|\le r_nO_{\mathrm{P}}(\sqrt{s_n/n})\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2,\qquad |I_{1,1}^{(\mathrm{II})}|\le r_nO_{\mathrm{P}}(1/\sqrt{n})\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1.$$

For the term $I_{1,2}$ in (22), similar to the proof of Lemma 1, $I_{1,2}=I_{1,2,1}+I_{1,2,2}$. We observe that

$$\begin{aligned}I_{1,2,1}\ge&\frac{r_n^2}{2}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})T}\mathbf{H}_n^{(\mathrm{I})}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\\&+\frac{r_n^2}{n}\sum_{i=1}^n \mathrm{E}[\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\}\{\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})}\}]\\\equiv& I_{1,2,1}^{(\mathrm{I})}+I_{1,2,1}^{(\mathrm{cross})}.\end{aligned}$$

Then there exists a constant $C>0$ such that

$$I_{1,2,1}^{(\mathrm{I})}\ge Cr_n^2\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2^2,\quad |I_{1,2,1}^{(\mathrm{cross})}|\le O_{\mathrm{P}}(r_n^2\sqrt{s_n})\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2\cdot\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1.$$

For the term $I_{1,2,2}$,

$$\begin{aligned}&I_{1,2,2}\\&=\frac{r_n^2}{2n}\sum_{i=1}^n[\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})})^2-\mathrm{E}\{\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)(\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})})^2\}]\\&\quad+\frac{r_n^2}{2n}\sum_{i=1}^n\Big[\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)2(\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})})(\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})})\\&\quad-\mathrm{E}\{\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)2(\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})})(\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})})\}\Big]\\&\quad+\frac{r_n^2}{2n}\sum_{i=1}^n[\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)(\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})})^2-\mathrm{E}\{\mathrm{p}_{2i}\,w(\boldsymbol{X}_i)(\boldsymbol{X}_i^{(\mathrm{II})T}\boldsymbol{u}_n^{(\mathrm{II})})^2\}]\\&\equiv I_{1,2,2}^{(\mathrm{I})}+I_{1,2,2}^{(\mathrm{cross})}+I_{1,2,2}^{(\mathrm{II})},\end{aligned}$$

where

$$\begin{aligned}|I_{1,2,2}^{(\mathrm{I})}|&\le r_n^2O_{\mathrm{P}}(s_n/\sqrt{n})\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2^2,\\|I_{1,2,2}^{(\mathrm{cross})}|&\le r_n^2O_{\mathrm{P}}(\sqrt{s_n/n})\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1,\\|I_{1,2,2}^{(\mathrm{II})}|&\le r_n^2O_{\mathrm{P}}(1/\sqrt{n})\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^2.\end{aligned}$$

For the term $I_{1,3}$ in (22), since $s_n p_n = o(n)$, $\|\widetilde{\boldsymbol{\beta}}^*\|_1$ is bounded and thus

$$|I_{1,3}| \le O_{\mathrm{P}}(r_n^3)\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_1^3 + O_{\mathrm{P}}(r_n^3)\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^3 \equiv I_{1,3}^{(\mathrm{I})} + I_{1,3}^{(\mathrm{II})},$$

where

$$|I_{1,3}^{(\mathrm{I})}| \le O_{\mathrm{P}}(r_n^3 s_n{}^{3/2})\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2^3, \qquad |I_{1,3}^{(\mathrm{II})}| \le O_{\mathrm{P}}(r_n^3)\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1^3.$$

For the term I_2 in (21), $I_2 \ge I_{2,1}^{(\mathrm{I})} + I_{2,1}^{(\mathrm{II})}$, where $I_{2,1}^{(\mathrm{I})} = -\lambda_n r_n \sum_{j=1}^{s_n} w_{n,j}|u_j|$ and $I_{2,1}^{(\mathrm{II})} = \lambda_n r_n \sum_{j=s_n+1}^{p_n} w_{n,j}|u_j|$. Thus, we have

$$|I_{2,1}^{(\mathrm{I})}| \le \lambda_n r_n w_{\max}^{(\mathrm{I})}\sqrt{s_n}\|\boldsymbol{u}_n^{(\mathrm{I})}\|_2, \qquad I_{2,1}^{(\mathrm{II})} \ge \lambda_n r_n w_{\min}^{(\mathrm{II})}\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1.$$

It can be shown that either $I_{1,2,1}^{(\mathrm{I})}$ or $I_{2,1}^{(\mathrm{II})}$ dominates all other terms in groups $\mathcal{G}_1 = \{I_{1,2,2}^{(\mathrm{I})}, I_{1,3}^{(\mathrm{I})}\}$, $\mathcal{G}_2 = \{I_{1,1}^{(\mathrm{II})}, I_{1,2,2}^{(\mathrm{II})}, I_{1,3}^{(\mathrm{II})}, I_{1,2,1}^{(\mathrm{cross})}, I_{1,2,2}^{(\mathrm{cross})}\}$ and $\mathcal{G}_3 = \{I_{1,1}^{(\mathrm{I})}, I_{2,1}^{(\mathrm{I})}\}$. Namely, $I_{1,2,1}^{(\mathrm{I})}$ dominates $\mathcal{G}_1$, and $I_{2,1}^{(\mathrm{II})}$ dominates $\mathcal{G}_2$. For $\mathcal{G}_3$, since $\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2 \le C_\epsilon$, we have that

$$|I_{1,1}^{(\mathrm{I})}| \le O_{\mathrm{P}}(r_n\sqrt{s_n/n})C_\epsilon, \qquad |I_{2,1}^{(\mathrm{I})}| \le \lambda_n r_n \sqrt{s_n} w_{\max}^{(\mathrm{I})} C_\epsilon.$$

Hence, if $\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1 \le C_\epsilon/2$, then $\|\widetilde{\boldsymbol{u}}_n^{(\mathrm{I})}\|_2 > C_\epsilon/2$, and thus $\mathcal{G}_3$ is dominated by $I_{1,2,1}^{(\mathrm{I})}$, which is positive; if $\|\boldsymbol{u}_n^{(\mathrm{II})}\|_1 > C_\epsilon/2$, then $\mathcal{G}_3$ is dominated by $I_{2,1}^{(\mathrm{II})}$, which is positive. This completes the proof. □

Proof of Theorem 2 We first need to show Lemma 2. □

Lemma 2 *Assume Condition A in Appendix 1.1. If* $s_n^2/n = O(1)$ *and* $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/\sqrt{s_n p_n} \xrightarrow{\mathrm{P}} \infty$, *then with probability tending to one, for any given* $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})T}, \boldsymbol{\beta}^{(\mathrm{II})T})^T$ *satisfying* $\|\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n})$ *and any constant* $C > 0$, *it follows that* $\ell_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}, \boldsymbol{0}) = \min_{\|\boldsymbol{\beta}^{(\mathrm{II})}\|_2 \le C\sqrt{s_n/n}} \ell_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}, \boldsymbol{\beta}^{(\mathrm{II})})$.

Proof It suffices to prove that with probability tending to one, for any $\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}$ satisfying $\|\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n})$, the following inequalities hold for $s_n + 1 \le j \le p_n$,

$$\begin{aligned} \partial\ell_n(\widetilde{\boldsymbol{\beta}})/\partial\beta_j &< 0, \quad \text{for } \beta_j < 0, \\ \partial\ell_n(\widetilde{\boldsymbol{\beta}})/\partial\beta_j &> 0, \quad \text{for } \beta_j > 0, \end{aligned}$$

namely, with probability tending to one,

$$\begin{aligned} \max_{s_n+1 \le j \le p_n} \sup_{\|\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n});\, \beta_j < 0} \frac{\partial}{\partial\beta_j}\ell_n(\widetilde{\boldsymbol{\beta}}) &< 0, \\ \min_{s_n+1 \le j \le p_n} \inf_{\|\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n});\, \beta_j > 0} \frac{\partial}{\partial\beta_j}\ell_n(\widetilde{\boldsymbol{\beta}}) &> 0. \end{aligned} \tag{26}$$

Proofs for showing both inequalities are similar; we only need to show (26). Note that for

$\beta_j \neq 0$,

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\ell_n(\widetilde{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}})\,w(\boldsymbol{X}_i)X_{i,j} + \lambda_n w_{n,j}\ \mathrm{sign}(\beta_j)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}_0)\,w(\boldsymbol{X}_i)X_{i,j}\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}^*)\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^T(\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0)\}X_{i,j} + \lambda_n w_{n,j}\ \mathrm{sign}(\beta_j),\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}^*$ lies between $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\boldsymbol{\beta}}$. It follows that

$$\begin{aligned}&\max_{s_n+1\le j\le p_n}\ \sup_{\|\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0\|_2=O_{\mathrm{P}}(\sqrt{s_n/n});\,\beta_j<0}\ \frac{\partial}{\partial\beta_j}\ell_n(\widetilde{\boldsymbol{\beta}})\\ &\le \max_{s_n+1\le j\le p_n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}_0)\,w(\boldsymbol{X}_i)X_{i,j}\\ &\quad+\max_{s_n+1\le j\le p_n}\ \sup_{\|\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0\|_2=O_{\mathrm{P}}(\sqrt{s_n/n})}\ \frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^T\widetilde{\boldsymbol{\beta}}^*)\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^T(\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0)\}X_{i,j}\\ &\quad-\min_{s_n+1\le j\le p_n}\{\lambda_n w_{n,j}\}\\ &\equiv I_1+I_2-\lambda_n\min_{s_n+1\le j\le p_n}w_{n,j}=I_1+I_2-\lambda_n w_{\min}^{(\mathrm{II})}.\end{aligned}$$

The first term I_1 satisfies that

$$|I_1|\le O_{\mathrm{P}}(\{\log(p_{_n}-s_n+1)/n\}^{1/2}). \tag{27}$$

For the term I_2,

$$|I_2|\le O_{\mathrm{P}}(\sqrt{s_n p_{_n}/n}). \tag{28}$$

Therefore, by (27) and (28), the left side of (26) is

$$\le O_{\mathrm{P}}(\sqrt{s_n p_{_n}/n})-\lambda_n w_{\min}^{(\mathrm{II})}=\sqrt{s_n p_{_n}/n}\{O_{\mathrm{P}}(1)-\lambda_n\sqrt{n}w_{\min}^{(\mathrm{II})}/\sqrt{s_n p_{_n}}\}.$$

By $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/\sqrt{s_n p_{_n}}\xrightarrow{\mathrm{P}}\infty$, (26) is proved. □

We now show Theorem 2. By Lemma 2, the first part of Theorem 2 holds that $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})},\mathbf{0}^T)^T$. To verify the second part of Theorem 2, notice the estimating equations $\frac{\partial\ell_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})},\mathbf{0})}{\partial\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}}\Big|_{\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}=\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}}=\mathbf{0}$, since $\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}$ is a local minimizer of $\ell_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})},\mathbf{0})$. Denote $\boldsymbol{d}_n(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}) = \lambda_n\mathbf{W}_n^{(\mathrm{I})}\,\mathrm{sign}\{\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}\}$ which is equal to $\boldsymbol{d}_n$ when $\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}=\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}$. Since $\min_{1\le j\le s_n}|\beta_{j;0}|/\sqrt{s_n/n}\to\infty$ and $\|\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2=O_{\mathrm{P}}(\sqrt{s_n/n})$, it follows that

$$
\begin{aligned}
&\mathrm{P}(\,\mathrm{sign}\{\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}\} \neq \mathrm{sign}\{\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\})\\
&\quad = \mathrm{P}(\,\mathrm{sign}(\widehat{\beta}_j) \neq \mathrm{sign}(\beta_{j;0}) \text{ for some } j \in \{1, \ldots, s_n\})\\
&\quad \le \mathrm{P}\Big(\max_{1 \le j \le s_n} |\widehat{\beta}_j - \beta_{j;0}| \ge \min_{1 \le j \le s_n} |\beta_{j;0}|\Big) \to 0.
\end{aligned}
$$

Thus with probability tending to one, $\boldsymbol{d}_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}) = \boldsymbol{d}_n(\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) = \boldsymbol{d}_n$. Taylor's expansion applied to the loss part on the left side of the estimating equations yields

$$
\begin{aligned}
\mathbf{0} = &\left\{\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + \boldsymbol{d}_n\right\}\\
&+ \left\{\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\right\}(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\\
&+ \frac{1}{2n}\sum_{i=1}^{n} \mathrm{p}_3(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{*(\mathrm{I})})\, w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\}^2\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}\\
\equiv &\left\{\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + \boldsymbol{d}_n\right\} + K_2(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + K_3,
\end{aligned}
\tag{29}
$$

where both $\widetilde{\boldsymbol{\beta}}^{*(\mathrm{I})}$ and $\widetilde{\boldsymbol{\beta}}^{**(\mathrm{I})}$ lie between $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}$. Below, we will show

$$\|K_2 - \mathbf{H}_n^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(s_n/\sqrt{n}), \tag{30}$$

$$\|K_3\|_2 = O_{\mathrm{P}}(s_n^{5/2}/n). \tag{31}$$

First, to show (30), note that $K_2 - \mathbf{H}_n^{(\mathrm{I})} = K_2 - \mathrm{E}(K_2) \equiv L_1$. Similar arguments for the proof of Lemma 1 give $\|L_1\|_2 = O_{\mathrm{P}}(s_n/\sqrt{n})$.

Second, a similar proof used for $I_{1,3}$ in (22) completes (31).

Third, by (29)–(31) and $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n})$, we see that

$$\mathbf{H}_n^{(\mathrm{I})}(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + \boldsymbol{d}_n = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + \boldsymbol{u}_n, \tag{32}$$

where $\|\boldsymbol{u}_n\|_2 = O_{\mathrm{P}}(s_n^{5/2}/n)$. Note that by Condition B5,

$$
\begin{aligned}
&\|\sqrt{n}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}\boldsymbol{u}_n\|_2 \le \sqrt{n}\|A_n\|_F \lambda_{\max}((\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2})\|\boldsymbol{u}_n\|_2\\
&\quad = \sqrt{n}\{\mathrm{tr}(A_n A_n^T)\}^{1/2}/\lambda_{\min}^{1/2}(\boldsymbol{\Omega}_n^{(\mathrm{I})})\|\boldsymbol{u}_n\|_2 = O_{\mathrm{P}}(s_n^{5/2}/\sqrt{n}) = o_{\mathrm{P}}(1).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\sqrt{n}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}\{\mathbf{H}_n^{(\mathrm{I})}(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + \boldsymbol{d}_n\}\\
&\quad = -\frac{1}{\sqrt{n}}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + o_{\mathrm{P}}(1).
\end{aligned}
$$

To complete proving the second part of Theorem 2, we apply the Lindeberg-Feller central limit theorem (van der Vaart, 1998) to $\sum_{i=1}^{n}\boldsymbol{Z}_i$, where $\boldsymbol{Z}_i = -n^{-1/2}A_n(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}\mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}$. It suffices to check two conditions: (I) $\sum_{i=1}^{n}\mathrm{cov}(\boldsymbol{Z}_i) \to \mathbb{G}$; (II)

$\sum_{i=1}^n \mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta}) = o(1)$ for some $\delta > 0$. Condition (I) follows from the fact that $\mathrm{var}\{\mathrm{p}_1(Y; \widetilde{\boldsymbol{X}}^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X})\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\} = \boldsymbol{\Omega}_n^{(\mathrm{I})}$. To verify condition (II), notice that using Conditions B5 and A5,

$$\begin{aligned}
&\mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta})\\
&\le n^{-(2+\delta)/2}\mathrm{E}\Bigg\{\|A_n\|_F^{2+\delta}\Bigg[\|(\boldsymbol{\Omega}_n^{(\mathrm{I})})^{-1/2}\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\|_2\\
&\left|\{\psi(r(Y, m(\boldsymbol{X}))) - g_1(m(\boldsymbol{X}))\}\frac{\{q''(m(\boldsymbol{X}))\sqrt{V(m(\boldsymbol{X}))}\}}{F'(m(\boldsymbol{X}))}\, w(\boldsymbol{X})\right|\Bigg]^{2+\delta}\Bigg\}\\
&\le Cn^{-(2+\delta)/2}\mathrm{E}[\{\lambda_{\min}^{-1/2}(\boldsymbol{\Omega}_n^{(\mathrm{I})})\|\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\|_2\}^{2+\delta}|\{\psi(r(Y, m(\boldsymbol{X}))) - g_1(m(\boldsymbol{X}))\}\times\\
&\{q''(m(\boldsymbol{X}))\sqrt{V(m(\boldsymbol{X}))}\}/F'(m(\boldsymbol{X}))|^{2+\delta}]\\
&\le Cs_n^{(2+\delta)/2}n^{-(2+\delta)/2}\mathrm{E}[|\{\psi(r(Y, m(\boldsymbol{X}))) - g_1(m(\boldsymbol{X}))\}\times\\
&\{q''(m(\boldsymbol{X}))\sqrt{V(m(\boldsymbol{X}))}\}/F'(m(\boldsymbol{X}))|^{2+\delta}]\\
&\le O\{(s_n/n)^{(2+\delta)/2}\}.
\end{aligned}$$

Thus, we get $\sum_{i=1}^n \mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta}) \le O\{n(s_n/n)^{(2+\delta)/2}\} = O\{s_n^{(2+\delta)/2}/n^{\delta/2}\}$, which is $o(1)$. This verifies Condition (II). □

Proof of Theorem 3 Before showing Theorem 3, Lemma 3 is needed. □

Lemma 3 *Assume conditions of Theorem 3. Then*

$$\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})} = -\frac{1}{n}(\mathbf{H}_n^{(\mathrm{I})})^{-1}\sum_{i=1}^n \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + o_{\mathrm{P}}(n^{-1/2}),$$

$$\sqrt{n}\{A_n(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}\widehat{\boldsymbol{\Omega}}_n^{(\mathrm{I})}(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}A_n^T\}^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathsf{k}}).$$

Proof Following (32) in the proof of Theorem 2, we observe that $\|\boldsymbol{u}_n\|_2 = O_{\mathrm{P}}(s_n^{5/2}/n) = o_{\mathrm{P}}(n^{-1/2})$. Furthermore, $\|\boldsymbol{d}_n\|_2 \le \sqrt{s_n}\lambda_n w_{\max}^{(\mathrm{I})} = o_{\mathrm{P}}(n^{-1/2})$. Condition B5 completes the proof for the first part.

To show the second part, denote $U_n = A_n(\mathbf{H}_n^{(\mathrm{I})})^{-1}\boldsymbol{\Omega}_n^{(\mathrm{I})}(\mathbf{H}_n^{(\mathrm{I})})^{-1}A_n^T$ and $\widehat{U}_n = A_n(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}\widehat{\boldsymbol{\Omega}}_n^{(\mathrm{I})}(\widehat{\mathbf{H}}_n^{(\mathrm{I})})^{-1}A_n^T$. Notice that the eigenvalues of $(\mathbf{H}_n^{(\mathrm{I})})^{-1}\boldsymbol{\Omega}_n^{(\mathrm{I})}(\mathbf{H}_n^{(\mathrm{I})})^{-1}$ are uniformly bounded away from zero. So are the eigenvalues of U_n. From the first part, we see that

$$A_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) = -\frac{1}{n}A_n(\mathbf{H}_n^{(\mathrm{I})})^{-1}\sum_{i=1}^n \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})} + o_{\mathrm{P}}(n^{-1/2}).$$

It follows that

$$\sqrt{n}U_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) = \sum_{i=1}^n \boldsymbol{Z}_i + o_{\mathrm{P}}(1),$$

where $\boldsymbol{Z}_i = -n^{-1/2}U_n^{-1/2}A_n(\mathbf{H}_n^{(\mathrm{I})})^{-1}\mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\, w(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})}$. To show $\sum_{i=1}^n \boldsymbol{Z}_i\xrightarrow{\mathcal{L}} N$

$(\mathbf{0}, \mathbf{I}_{\mathsf{k}})$, similar to the proof for Theorem 2, we check two conditions: (III) $\sum_{i=1}^{n} \operatorname{cov}(\boldsymbol{Z}_i) \to \mathbf{I}_{\mathsf{k}}$; (IV) $\sum_{i=1}^{n} \mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta}) = o(1)$ for some $\delta > 0$. Condition (III) is straightforward since $\sum_{i=1}^{n} \operatorname{cov}(\boldsymbol{Z}_i) = U_n^{-1/2} U_n U_n^{-1/2} = \mathbf{I}_{\mathsf{k}}$. To check condition (IV), similar arguments used in the proof of Theorem 2 give that $\mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta}) = O\{(s_n/n)^{(2+\delta)/2}\}$. This and the boundedness of the ψ-function yield $\sum_{i=1}^{n} \mathrm{E}(\|\boldsymbol{Z}_i\|_2^{2+\delta}) \le O\ \{s_n^{(2+\delta)/2} / n^{\delta/2}\} = o(1)$. Hence

$$\sqrt{n} U_n^{-1/2} A_n (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathsf{k}}). \tag{33}$$

Also, it can be concluded that $\|\widehat{U}_n - U_n\|_2 = o_{\mathrm{P}}(1)$ and that the eigenvalues of $\widehat{U}_n$ are uniformly bounded away from zero and infinity with probability tending to one. Consequently,

$$\|\widehat{U}_n^{-1/2} U_n^{1/2} - \mathbf{I}_{\mathsf{k}}\|_2 = o_{\mathrm{P}}(1). \tag{34}$$

Combining (33), (34) and Slutsky's theorem completes the proof that $\sqrt{n} \widehat{U}_n^{-1/2} A_n (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathsf{k}})$. □

We now show Theorem 3, which follows directly from the null hypothesis H_0 in (14) and the second part of Lemma 3. This completes the proof. □

Proof of Theorem 4 The proof of Theorem 4 is similar to that used in Theorem 7, except that in the Part 2, $\mathcal{C}_n$ is changed from $\lambda_n \sqrt{n}/s_n$ to λ_n. □

Proof of Theorem 5 The proof of Theorem 5 is similar to that used in Theorem 8, except that in the Part 2, $\mathcal{B}_n$ is changed from $\lambda_n \sqrt{n}/s_n$ to λ_n. □

Proof of Theorem 6 Assumption (19) implies that $\ell_n(\widetilde{\boldsymbol{\beta}})$ in (3) is convex in $\widetilde{\boldsymbol{\beta}}$. By Karush-Kuhn-Tucker conditions (Wright 1997, Theorem A.2), a set of sufficient conditions for an estimate $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{p_n})^T$ being a global minimizer of (3) is that

$$\begin{gathered}
\frac{1}{n} \sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widehat{\widetilde{\boldsymbol{\beta}}})\, w(\boldsymbol{X}_i) = 0, \\
\frac{1}{n} \sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widehat{\widetilde{\boldsymbol{\beta}}})\, w(\boldsymbol{X}_i) X_{i,j} = -\lambda_n w_{n,j} \operatorname{sign}(\widehat{\beta}_j), \text{ for } 1 \le j \le p_n \text{ with } \widehat{\beta}_j \ne 0, \\
\left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widehat{\widetilde{\boldsymbol{\beta}}})\, w(\boldsymbol{X}_i) X_{i,j} \right| \le \lambda_n w_{n,j}, \text{ for } 1 \le j \le p_n \text{ with } \widehat{\beta}_j = 0.
\end{gathered} \tag{35}$$

Before proving Theorem 6, we first show Lemma 4. □

Lemma 4 (*existence and consistency*: $p_n \gg n$) *Assume* (19) *and Conditions* A0, A1, A2, A4, A5′, B5, A6, A7 *in Appendix* 1.1. *Suppose* $s_n^4/n \to 0$, $\log(p_n - s_n)/n = O(1)$, $\log(p_n - s_n)/\{n\lambda_n^2 (w_{\min}^{(\mathrm{II})})^2\} = o_{\mathrm{P}}(1)$ *and* $\min_{1 \le j \le s_n} |\beta_{j;0}| / \sqrt{s_n/n} \to \infty$. *Assume* $w_{\max}^{(\mathrm{I})} = O_{\mathrm{P}}\{1/(\lambda_n$

$\sqrt{n})\}$ and $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/s_n \xrightarrow{\mathrm{P}} \infty$. Then with probability tending to one, there exists a global minimizer $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})T}, \widehat{\boldsymbol{\beta}}^{(\mathrm{II})T})^T$ of $\ell_n(\widetilde{\boldsymbol{\beta}})$ in (3) which satisfies that

(i) $\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \mathbf{0}$,

(ii) $\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}$ is the minimizer of the oracle subproblem,

$$\ell_n^O(\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}) = \frac{1}{n}\sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})}))\, w(\boldsymbol{X}_i) + \lambda_n \sum_{j=1}^{s_n} w_{n,j}|\beta_j|. \quad (36)$$

Proof Let $\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})} = (\widehat{b}_0, \widehat{b}_1, \ldots, \widehat{b}_{s_n})^T$ be the minimizer of the subproblem (36). By Karush-Kuhn-Tucker necessary conditions (Wright 1997, Theorem A.1), $\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}$ satisfies that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\, w(\boldsymbol{X}_i) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\, w(\boldsymbol{X}_i)X_{i,j} = -\lambda_n w_{n,j}\ \mathrm{sign}(\widehat{b}_j),\ \text{for}\ 1 \le j \le s_n\ \text{with}\ \widehat{b}_j \ne 0,$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\, w(\boldsymbol{X}_i)X_{i,j}\right| \le \lambda_n w_{n,j},\ \text{for}\ 1 \le j \le s_n\ \text{with}\ \widehat{b}_j = 0.$$

In the following, we will verify conditions

$$\widehat{b}_1 \ne 0, \ldots, \widehat{b}_{s_n} \ne 0, \quad (37)$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\, w(\boldsymbol{X}_i)X_{i,j}\right| \le \lambda_n w_{n,j},\ \text{for}\ s_n + 1 \le j \le p_n. \quad (38)$$

It then follows, from (37), (38) and (35), that $(\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})T}, \mathbf{0}^T)^T$ is the global minimizer of (3). This will in turn imply Lemma 4.

First, we prove that (37) holds with probability tending to one. Applying Lemma 1 to the subproblem (36), we conclude that $\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(\sqrt{s_n/n})$. Since $\min_{1 \le j \le s_n} |\beta_{j;0}| / \sqrt{s_n/n} \to \infty$ as $n \to \infty$, it is seen that

$$\begin{aligned} &\mathrm{P}(\,\mathrm{sign}(\widehat{\beta}_j) \ne \mathrm{sign}(\beta_{j;0})\ \text{for some}\ j \in \{1, \ldots, s_n\}) \\ &\le \mathrm{P}\Big(\max_{1 \le j \le s_n} |\widehat{\beta}_j - \beta_{j;0}| \ge \min_{1 \le j \le s_n} |\beta_{j;0}|\Big) \to 0. \end{aligned}$$

Hence (37) holds with probability tending to one.

Second, we prove that (38) holds with probability tending to one. It suffices to prove that

$$\mathrm{P}\left(\max_{s_n+1 \le j \le p_n} \left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\, w(\boldsymbol{X}_i)X_{i,j}\right| < \lambda_n w_{\min}^{(\mathrm{II})}\right) \to 1. \quad (39)$$

By Taylor's expansion, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})})\,w(\boldsymbol{X}_i)X_{i,j}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\,w(\boldsymbol{X}_i)X_{i,j}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*})\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}(\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\}X_{i,j},$$

with $\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*}$ located between $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}$ and $\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}$. Then (39) holds if we can prove

$$\mathrm{P}\left(\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\,w(\boldsymbol{X}_i)X_{i,j}\right|<\frac{\lambda_n}{2}w_{\min}^{(\mathrm{II})}\right)\to 1, \tag{40}$$

and

$$\mathrm{P}\left(\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*})\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}(\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\}X_{i,j}\right|<\frac{\lambda_n}{2}w_{\min}^{(\mathrm{II})}\right)\to 1. \tag{41}$$

We first prove (40). Set $\mathrm{p}_{1i}=\mathrm{p}_{_1}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})$. Since $\log(p_{_n}-s_n)=O(n)$ and $\log(p_{_n}-s_n)=o_{_\mathrm{P}}\{n\lambda_n^2(w_{\min}^{(\mathrm{II})})^2\}$, we see that

$$\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{1i}\,w(\boldsymbol{X}_i)X_{i,j}\right|=O_{\mathrm{P}}\{\sqrt{\log(p_{_n}-s_n+1)/n}\}=o_{_\mathrm{P}}(\lambda_n w_{\min}^{(\mathrm{II})}).$$

This implies (40).

Second, we prove (41). Since $\|\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_1<\infty$ and $\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2=O_{\mathrm{P}}(\sqrt{s_n/n})$, it follows that

$$\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}\|_1\le\|\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_1+\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_1\le\|\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_1+\sqrt{s_n}\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|=O_{\mathrm{P}}(1),$$

and then $\|\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*}\|_1=O_{\mathrm{P}}(1)$, thus

$$\max_{s_n+1\le j\le p_n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*})\,w(\boldsymbol{X}_i)\{\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}(\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\}X_{i,j}\right|$$
$$\le C\sqrt{s_n}\|\widehat{\widetilde{\boldsymbol{b}}}_n^{(\mathrm{I})}-\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|\left\{\frac{1}{n}\sum_{i=1}^{n}|\mathrm{p}_{_2}(Y_i;\widetilde{\boldsymbol{X}}_i^{(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^{(\mathrm{I})*})|\,w(\boldsymbol{X}_i)\right\}$$
$$=\sqrt{s_n}O_{\mathrm{P}}(\sqrt{s_n/n})O_{\mathrm{P}}(1)=O_{\mathrm{P}}(s_n/\sqrt{n})=o_{_\mathrm{P}}\{\lambda_n w_{\min}^{(\mathrm{II})}\}.$$

Here $w_{\min}^{(\mathrm{II})}\lambda_n\sqrt{n}/s_n\xrightarrow{\mathrm{P}}\infty$ is used. Hence (41) is proved. □

The first part of Theorem 6 follows from the first part of Lemma 4. The second part of Theorem 6 follows directly from applying Theorem 2 to the oracle subproblem (36). □

Proof of Theorem 7 It is easy to see that $\widehat{\beta}_j^{\mathrm{PMR}}=\arg\min_\beta \ell_j^{\mathrm{PMR}*}(\beta)$, where $\ell_{n,j}^{\mathrm{PMR}*}(\beta)=\ell_{n,j}^{\mathrm{PMR}}(\widehat{\alpha}_j(\beta),\beta)$, and $\widehat{\alpha}_j(\beta)$ satisfies $n^{-1}\sum_{i=1}^{n}\mathrm{q}_{_1}(Y_i;\widehat{\alpha}_j(\beta)+X_{i,j}\beta)=0$ for $j=1,\ldots,p_{_n}$. From (11), $\widehat{\alpha}_1(0)=\cdots=\widehat{\alpha}_{p_n}(0)$. Let $\widehat{\alpha}_0=\widehat{\alpha}_1(0)$. Then $\widehat{\alpha}_0\xrightarrow{\mathrm{P}}\alpha_0$, where $\alpha_0=F(\mu_{_0})$ with

$\mu_0 = \mathrm{E}(Y)$. The rest of the proof contains two parts.

Part 1. For $\mathcal{A}_n = \lambda_n \sqrt{n}$, we will show that $\widehat{w}^{(\mathrm{I})}_{\max} \mathcal{A}_n = O_{\mathrm{P}}(1)$. It suffices to show that there exist local minimizers $\widehat{\beta}_j^{\mathrm{PMR}}$ of $\ell_{n,j}^{\mathrm{PMR}*}(\beta)$ such that $\lim_{\delta \to 0+} \inf_{n \geq 1} \mathrm{P}(\min_{1 \leq j \leq s_n} |\widehat{\beta}_j^{\mathrm{PMR}}| > \mathcal{A}_n \delta) = 1$. It suffices to prove that, for $1 \leq j \leq s_n$, there exist some b_j with $|b_j| = 2\delta$ such that

$$\lim_{\delta \to 0+} \inf_{n \geq 1} \mathrm{P}\Big(\min_{1 \leq j \leq s_n} \Big\{ \inf_{|\beta| \leq \delta} \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n \beta) - \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n b_j) \Big\} > 0 \Big) = 1, \tag{42}$$

and there exists some large enough $C_n > 0$ such that

$$\lim_{\delta \to 0+} \inf_{n \geq 1} \mathrm{P}\Big(\min_{1 \leq j \leq s_n} \Big\{ \inf_{|\beta| \geq C_n} \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n \beta) - \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n b_j) \Big\} > 0 \Big) = 1. \tag{43}$$

Equations (42) and (43) imply that with probability tending to one, there must exist local minimizers $\widehat{\beta}_j^{\mathrm{PMR}}$ of $\ell_{n,j}^{\mathrm{PMR}*}(\beta)$ such that $\mathcal{A}_n \delta < |\widehat{\beta}_j^{\mathrm{PMR}}| < \mathcal{A}_n C_n$ for $1 \leq j \leq s_n$.

First, we prove (43). For every $n \geq 1$, when $|\beta| \to \infty$,

$$\min_{1 \leq j \leq s_n} \{\ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n \beta) - \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n b_j)\} \geq \kappa_n \mathcal{A}_n |\beta| - \max_{1 \leq j \leq s_n} \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n b_j) \xrightarrow{\mathrm{P}} \infty.$$

Thus (43) holds.

Second, we prove (42). Since $\mathcal{A}_n = O(1)$, we see that $|\mathcal{A}_n \beta| \leq O(1)\delta \to 0$ as $\delta \to 0+$. For $1 \leq j \leq s_n$, by Taylor's expansion,

$$\begin{aligned} \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n \beta) = {} & \frac{1}{n} \sum_{i=1}^n Q_q(Y_i, \widehat{\mu}_0) + \mathcal{A}_n \beta \frac{1}{n} \sum_{i=1}^n \mathrm{q}_1(Y_i; \widehat{\alpha}_0)\{X_{i,j} - \mathrm{E}(X_j)\} \\ & + \frac{\mathcal{A}_n^2 \beta^2}{2} \frac{1}{n} \sum_{i=1}^n \mathrm{q}_2(Y_i; \theta_{ij}^*)\{\widehat{\alpha}_j'(\mathcal{A}_n \beta_j^*) + X_{i,j}\}^2 + \mathcal{A}_n \kappa_n |\beta|, \end{aligned}$$

where $\widehat{\mu}_0 = F^{-1}(\widehat{\alpha}_0)$, $\theta_{ij}^* = \theta_{ij}(\mathcal{A}_n \beta_j^*)$, $\theta_{ij}(\beta) = \widehat{\alpha}_j(\beta) + X_{i,j}\beta$ and β_j^* is between 0 and β. Thus we have that

$$\begin{aligned} & \min_{1 \leq j \leq s_n} \Big\{ \inf_{|\beta| \leq \delta} \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n \beta) - \ell_{n,j}^{\mathrm{PMR}*}(\mathcal{A}_n b_j) \Big\} \\ & \geq \mathcal{A}_n \min_{1 \leq j \leq s_n} \inf_{|\beta| \leq \delta} \Big[(\beta - b_j) \frac{1}{n} \sum_{i=1}^n \mathrm{q}_1(Y_i; \widehat{\alpha}_0)\{X_{i,j} - \mathrm{E}(X_j)\} \Big] \\ & \quad + \frac{\mathcal{A}_n^2}{2} \min_{1 \leq j \leq s_n} \inf_{|\beta| \leq \delta} \Big[\beta^2 \frac{1}{n} \sum_{i=1}^n \mathrm{q}_2(Y_i; \theta_{ij}^*)\{\widehat{\alpha}_j'(\mathcal{A}_n \beta_j^*) + X_{i,j}\}^2 \\ & \quad - b_j^2 \frac{1}{n} \sum_{i=1}^n \mathrm{q}_2(Y_i; c_{ij}^*)\{\widehat{\alpha}_j'(\mathcal{A}_n b_j^*) + X_{i,j}\}^2 \Big] \\ & \quad + \mathcal{A}_n \min_{1 \leq j \leq s_n} \inf_{|\beta| \leq \delta} \{\kappa_n(|\beta| - |b_j|)\} \equiv I_1 + I_2 + I_3, \end{aligned}$$

where $c_{ij}^* = \theta_{ij}(\mathcal{A}_n b_j^*)$, with b_j^* between 0 and b_j. Let $\widehat{C}_0 = q''(\widehat{\mu}_0)/F'(\widehat{\mu}_0) \neq 0$. Then $\widehat{C}_0 \xrightarrow{\mathrm{P}} C_0$, where $C_0 = q''(\mu_0)/F'(\mu_0)$. We obtain

$$\begin{aligned} I_1 = {} & \mathcal{A}_n \min_{1\le j\le s_n} \inf_{|\beta|\le\delta}\{\widehat{C}_0(\beta-b_j)\mathrm{cov}(X_j,Y)\} \\ & + \mathcal{A}_n \min_{1\le j\le s_n} \inf_{|\beta|\le\delta}\left(\widehat{C}_0(\beta-b_j)\frac{1}{n}\sum_{i=1}^{n}[(Y_i-\mu_0)\{X_{i,j}-\mathrm{E}(X_j)\}-\mathrm{cov}(X_j,Y)]\right) \\ & - \mathcal{A}_n \max_{1\le j\le s_n} \sup_{|\beta|\le\delta}\left[\widehat{C}_0(\widehat{\mu}_0-\mu_0)(\beta-b_j)\frac{1}{n}\sum_{i=1}^{n}\{X_{i,j}-\mathrm{E}(X_j)\}\right] \equiv I_{1,1}+I_{1,2}+I_{1,3}. \end{aligned}$$

Choosing $b_j = -2\delta\,\mathrm{sign}\{\widehat{C}_0\,\mathrm{cov}(X_j,Y)\}$, which satisfies $|b_j| = 2\delta$, gives

$$\begin{aligned} I_{1,1} &= \mathcal{A}_n \min_{1\le j\le s_n} \inf_{|\beta|\le\delta}\{\beta\widehat{C}_0\,\mathrm{cov}(X_j,Y) + 2\delta|\widehat{C}_0\,\mathrm{cov}(X_j,Y)|\} \\ &\ge \mathcal{A}_n\,\delta|\widehat{C}_0| \min_{1\le j\le s_n}|\mathrm{cov}(X_j,Y)| \ge |\widehat{C}_0|c\mathcal{A}_n^2\,\delta. \end{aligned}$$

We can see that $|I_{1,2}| = O_{\mathrm{P}}(\mathcal{A}_n\{\log(s_n)/n\}^{1/2})\delta$, by the Bernstein's inequality (van der Vaart and Wellner 1996, Lemma 2.2.11). Similarly, $|I_{1,3}| \le o_{\mathrm{P}}(\mathcal{A}_n\{\log(s_n)/n\}^{1/2})\delta$. For terms I_2 and I_3, we observe that $|I_2| \le O_{\mathrm{P}}(\mathcal{A}_n^2)\,\delta^2$ and $|I_3| = O(\mathcal{A}_n\,\kappa_n)\delta$. The conditions $\log(p_n) = o(n\kappa_n^2)$ and $\mathcal{A}_n/\kappa_n \to \infty$ imply that $\{\log(s_n)/n\}^{1/2}/\mathcal{A}_n = o(1)$. Together with the condition $\mathcal{A}_n/\kappa_n \to \infty$, we can choose a small enough $\delta > 0$ such that with probability tending to one, $I_{1,2}, I_{1,3}, I_2$ and I_3 are dominated by $I_{1,1}$, which is positive. Thus (42) is proved.

Part 2. For $\mathcal{C}_n = \lambda_n\sqrt{n}/s_n$, we will show that $\widehat{w}^{(\mathrm{II})}_{\min}\mathcal{C}_n \xrightarrow{\mathrm{P}} \infty$. It suffices to prove that for any $\epsilon > 0$, there exist local minimizers $\widehat{\beta}_j^{\mathrm{PMR}}$ of $\ell_{n,j}^{\mathrm{PMR}*}(\beta)$ such that $\lim_{n\to\infty}\mathrm{P}(\max_{s_n+1\le j\le p_n}|\widehat{\beta}_j^{\mathrm{PMR}}| \le \mathcal{C}_n\epsilon) = 1$. Similar to the proof of Lemma 1, we will prove that for any $\epsilon > 0$,

$$\lim_{n\to\infty}\mathrm{P}\left(\min_{s_n+1\le j\le p_n}\left\{\inf_{|\beta|=\epsilon}\ell_{n,j}^{\mathrm{PMR}*}(\mathcal{C}_n\beta) - \ell_{n,j}^{\mathrm{PMR}*}(0)\right\} > 0\right) = 1. \tag{44}$$

Since $\mathcal{C}_n \to 0$ as $n \to \infty$, we have that by Taylor's expansion,

$$\begin{aligned} & \min_{s_n+1\le j\le p_n}\left\{\inf_{|\beta|=\epsilon}\ell_{n,j}^{\mathrm{PMR}*}(\mathcal{C}_n\beta) - \ell_{n,j}^{\mathrm{PMR}*}(0)\right\} \\ & \ge \mathcal{C}_n \min_{s_n+1\le j\le p_n}\inf_{|\beta|=\epsilon}\left[\beta\frac{1}{n}\sum_{i=1}^{n}\mathrm{q}_1(Y_i;\widehat{\alpha}_0)\{X_{i,j}-\mathrm{E}(X_j)\}\right] \\ & \quad + \frac{\mathcal{C}_n^2}{2}\min_{s_n+1\le j\le p_n}\inf_{|\beta|=\epsilon}\left[\beta^2\frac{1}{n}\sum_{i=1}^{n}\mathrm{q}_2(Y_i;\theta_{ij}(\mathcal{C}_n\beta_j^*))\{\widehat{\alpha}_j'(\mathcal{C}_n\beta_j^*)+X_{i,j}\}^2\right] \\ & \quad + \mathcal{C}_n\inf_{|\beta|=\epsilon}(\kappa_n|\beta|) \equiv I_1+I_2+I_3, \end{aligned}$$

where β_j^* is between 0 and β. Similar to the proof in **Part 1**,

$$\begin{aligned} I_1 = {} & \mathcal{C}_n \min_{s_n+1\le j\le p_n}\inf_{|\beta|=\epsilon}\{\widehat{C}_0\beta\mathrm{cov}(X_j,Y)\} \\ & + \mathcal{C}_n \min_{s_n+1\le j\le p_n}\inf_{|\beta|=\epsilon}\left(\widehat{C}_0\beta\frac{1}{n}\sum_{i=1}^{n}[(Y_i-\mu_0)\{X_{i,j}-\mathrm{E}(X_j)\}-\mathrm{cov}(X_j,Y)]\right) \\ & - \mathcal{C}_n \max_{s_n+1\le j\le p_n}\sup_{|\beta|=\epsilon}\left[\widehat{C}_0(\widehat{\mu}_0-\mu_0)\beta\frac{1}{n}\sum_{i=1}^{n}\{X_{i,j}-\mathrm{E}(X_j)\}\right] \equiv I_{1,1}+I_{1,2}+I_{1,3}. \end{aligned}$$

Then $|I_{1,1}| \le o(\mathcal{C}_n\mathcal{B}_n\epsilon)$, $|I_{1,2}| \le O_{\mathrm{P}}[\mathcal{C}_n\{\log(p_n - s_n + 1)/n\}^{1/2}]\epsilon$ and $|I_{1,3}| \le o_{\mathrm{P}}[\mathcal{C}_n\{\log(p_n - s_n + 1)/n\}^{1/2}]\epsilon$. Hence $|I_1| \le O_{\mathrm{P}}[\mathcal{C}_n\{\log(p_n - s_n + 1)/n\}^{1/2}]\epsilon + o(\mathcal{C}_n\mathcal{B}_n)\epsilon$. For the term I_2, we have that $|I_2| \le O_{\mathrm{P}}(\mathcal{C}_n^2)\epsilon^2$. Note $I_3 = \mathcal{C}_n\kappa_n\epsilon$. Since $\log(p_n) = o(n\kappa_n^2)$, $\mathcal{B}_n = O(\kappa_n)$ and $\mathcal{C}_n = o(\kappa_n)$, it follows that with probability tending to one, terms I_1 and I_2 are dominated by I_3, which is positive. So (44) is proved. □

Proof of Theorem 8 It is easy to see that $\widehat{\beta}_j^{\mathrm{MR}} = \arg\min_\beta \ell_j^{\mathrm{MR}*}(\beta)$, where $\ell_{n,j}^{\mathrm{MR}*}(\beta) = \ell_{n,j}^{\mathrm{MR}}(\widehat{\alpha}_j(\beta), \beta)$, and $\widehat{\alpha}_j(\beta)$ satisfies $n^{-1}\sum_{i=1}^n \mathrm{q}_1(Y_i; \widehat{\alpha}_j(\beta) + X_{i,j}\beta) = 0$ for $j = 1, \ldots, p_n$. From (11), $\widehat{\alpha}_1(0) = \cdots = \widehat{\alpha}_{p_n}(0)$. Let $\widehat{\alpha}_0 = \widehat{\alpha}_1(0)$. Then $\widehat{\alpha}_0 \xrightarrow{\mathrm{P}} \alpha_0$, where $\alpha_0 = F(\mu_0)$ with $\mu_0 = \mathrm{E}(Y)$. Let $h_{n,j}(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\ell_{n,j}^{\mathrm{MR}*}(\beta) = n^{-1}\sum_{i=1}^n \mathrm{q}_1(Y_i; \widehat{\alpha}_j(\beta) + X_{i,j}\beta)\ \{\widehat{\alpha}_j'(\beta) + X_{i,j}\}$. Then $h_{n,j}'(\beta) = n^{-1}\sum_{i=1}^n \mathrm{q}_2(Y_i; \widehat{\alpha}_j(\beta) + X_{i,j}\beta)\{\widehat{\alpha}_j'(\beta) + X_{i,j}\}^2$ and $h_{n,j}''(\beta) = n^{-1}\sum_{i=1}^n \mathrm{q}_{3i}(\beta)$. The minimizer $\widehat{\beta}_j^{\mathrm{MR}}$ of (17) satisfies the estimating equations, $h_{n,j}(\widehat{\beta}_j^{\mathrm{MR}}) = 0$. The rest of the proof consists of two parts.

Part 1. For $\mathcal{A}_n = \lambda_n\sqrt{n}$, we will show that $\widehat{w}_{\max}^{(\mathrm{I})}\mathcal{A}_n = O_{\mathrm{P}}(1)$, which is $\mathcal{A}_n/\min_{1\le j\le s_n}|\widehat{\beta}_j^{\mathrm{MR}}| = O_{\mathrm{P}}(1)$. That is, $\lim_{\delta\to 0+}\sup_{n\ge 1}\mathrm{P}(\min_{1\le j\le s_n}|\widehat{\beta}_j^{\mathrm{MR}}| < \mathcal{A}_n\delta) = 0$. Using the Bonferroni inequality, it suffices to show that

$$\lim_{\delta\to 0+}\sup_{n\ge 1}\sum_{j=1}^{s_n}\mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| < \mathcal{A}_n\,\delta) = 0.$$

With assumption (11) for the convex BD, $h_{n,j}(\cdot)$ is an increasing function. Thus

$$\mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| < \mathcal{A}_n\,\delta) \le \mathrm{P}\{h_{n,j}(-\mathcal{A}_n\,\delta) \le 0 \le h_{n,j}(\mathcal{A}_n\,\delta)\}.$$

Note that $\mathcal{A}_n = O(1)$ gives $\mathcal{A}_n\,\delta \to 0$ as $\delta \to 0+$. By Taylor's expansion, for $1 \le j \le s_n$, we have that

$$\begin{aligned} h_{n,j}(\pm\mathcal{A}_n\,\delta) &= \frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_i; \widehat{\alpha}_0)\{X_{i,j} - \mathrm{E}(X_j)\} + (\pm\mathcal{A}_n\,\delta)\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_i; \widehat{\alpha}_0)\{\widehat{\alpha}_j'(0) + X_{i,j}\}^2 \\ &\quad + \frac{1}{2}(\mathcal{A}_n\,\delta)^2\frac{1}{n}\sum_{i=1}^n \mathrm{q}_{3i}(\mathcal{A}_n\,\delta_j^*) \equiv I_{1j} + I_{2j} + I_{3j}, \end{aligned}$$

with $\delta_j^* \in (0, \delta)$. Let $\widehat{C}_0 = q''(\widehat{\mu}_0)/F'(\widehat{\mu}_0) \neq 0$, where $\widehat{\mu}_0 = F^{-1}(\widehat{\alpha}_0)$. Then $\widehat{C}_0 \xrightarrow{\mathrm{P}} C_0$, where $C_0 = q''(\mu_0)/F'(\mu_0)$. We obtain

$$\begin{aligned} I_{1j} &= \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\mu}_0)\widehat{C}_0\{X_{i,j} - \mathrm{E}(X_j)\} \\ &= \widehat{C}_0\,\mathrm{cov}(X_j, Y) + \widehat{C}_0\frac{1}{n}\sum_{i=1}^n [(Y_i - \mu_0)\{X_{i,j} - \mathrm{E}(X_j)\} - \mathrm{cov}(X_j, Y)] \\ &\quad - \widehat{C}_0(\widehat{\mu}_0 - \mu_0)\frac{1}{n}\sum_{i=1}^n \{X_{i,j} - \mathrm{E}(X_j)\} \equiv I_{1j,1} + I_{1j,2} + I_{1j,3}. \end{aligned}$$

Because $\mathcal{A}_n = O(1)$, $|\mathrm{cov}(X_j, Y)| \ge c\,\mathcal{A}_n$, $1 \le j \le s_n$, and both $\max_{1\le j\le s_n}\mathrm{E}[n^{-1}\sum_{i=1}^n \mathrm{q}_2(Y_i; \widehat{\alpha}_0)\{\widehat{\alpha}_j'(0) + X_{i,j}\}^2]$ and $\max_{1\le j\le s_n}\mathrm{E}\{n^{-1}\ \sum_{i=1}^n |\mathrm{q}_{3i}(\mathcal{A}_n$

$\delta_j^*)|\}$ are bounded, we can choose δ small enough such that, uniformly for all $1 \le j \le s_n$, the term $I_{1j,1} = \widehat{C}_0 \operatorname{cov}(X_j, Y)$ dominates I_{2j} and I_{3j}. By assuming $\widehat{C}_0 \operatorname{cov}(X_j, Y) < 0$ without loss of generality,

$$\begin{aligned} \mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| \le \mathcal{A}_n \delta) &\le \mathrm{P}(0 \le h_{n,j}(\mathcal{A}_n \delta)) \\ &\le \mathrm{P}(I_{1j,2} + I_{1j,3} \ge C\mathcal{A}_n) \le 4 \exp\left(\frac{-n^2 \mathcal{A}_n^2}{C_1 n + C_2 n \mathcal{A}_n}\right), \end{aligned} \tag{45}$$

for some positive constants C, C_1 and C_2, where the last inequality applies the Bernstein inequality. By (45), for a small enough $\delta > 0$,

$$\sum_{j=1}^{s_n} \mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| < \mathcal{A}_n \delta) \le 4 s_n \exp\left(\frac{-n^2 \mathcal{A}_n^2}{C_1 n + C_2 n \mathcal{A}_n}\right) = o(1). \tag{46}$$

The equality in (46) follows from $\mathcal{A}_n = O(1)$, $\lambda_n n \to \infty$ and $\log(s_n) = o(\lambda_n^2 n^2)$, where the latter two are implied by the conditions $\lambda_n n / s_n \to \infty$ and $\log(p_n) = o(\lambda_n^2 n^2 / s_n^2)$.

Part 2. For $\mathcal{B}_n = \lambda_n \sqrt{n}/s_n$, we will prove that $\widehat{w}_{\min}^{(\mathrm{II})} \mathcal{B}_n \xrightarrow{\mathrm{P}} \infty$, which is $\max_{s_n+1 \le j \le p_n} |\widehat{\beta}_j^{\mathrm{MR}}| / \mathcal{B}_n = o_{\mathrm{P}}(1)$. Namely, for any $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(\max_{s_n+1 \le j \le p_n} |\widehat{\beta}_j^{\mathrm{MR}}| \ge \mathcal{B}_n \epsilon) = 0$. By the Bonferroni inequality, it suffices to show that

$$\lim_{n\to\infty} \sum_{j=s_n+1}^{p_n} \mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| \ge \mathcal{B}_n \epsilon) = 0.$$

Since $h_{n,j}(\cdot)$ is increasing, we have that for $j = s_n + 1, \ldots, p_n$,

$$\mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| \ge \mathcal{B}_n \epsilon) \le \mathrm{P}\{h_{n,j}(-\mathcal{B}_n \epsilon) \ge 0\} + \mathrm{P}\{h_{n,j}(\mathcal{B}_n \epsilon) \le 0\}. \tag{47}$$

Similar to **Part 1**, $\mathcal{B}_n = o(1)$ gives that for $j = s_n + 1, \ldots, p_n$,

$$\begin{aligned} h_{n,j}(\pm \mathcal{B}_n \epsilon) = &\frac{1}{n} \sum_{i=1}^{n} \mathrm{q}_1(Y_i; \widehat{\alpha}_0) \{X_{i,j} - \mathrm{E}(X_j)\} \\ &+ (\pm \mathcal{B}_n \epsilon) \frac{1}{n} \sum_{i=1}^{n} \mathrm{q}_2(Y_i; \widehat{\alpha}_0) \{\widehat{\alpha}_j'(0) + X_{i,j}\}^2 \\ &+ \frac{1}{2} (\mathcal{B}_n \epsilon)^2 \frac{1}{n} \sum_{i=1}^{n} \mathrm{q}_{3i}(\mathcal{B}_n \epsilon_j^*) \equiv I_{1j} + J_{2j} + J_{3j}, \end{aligned}$$

with $\epsilon_j^* \in (0, \delta)$. Since $\mathcal{B}_n = o(1)$, $|\operatorname{cov}(X_j, Y)| = o(\mathcal{B}_n)$, $s_n + 1 \le j \le p_n$, and from Condition E2, $|J_{2j}| \ge \mathcal{B}_n \epsilon \eta$, as $n \to \infty$, J_{2j} dominates $I_{1j,1}$ and J_{3j}. Applying the Bernstein's inequality, for large n,

$$\begin{aligned} \mathrm{P}\{h_{n,j}(\mathcal{B}_n \epsilon) \le 0\} &\le \mathrm{P}(I_{1j,2} + I_{1j,3} \le -C\mathcal{B}_n \epsilon) \\ &\le 4 \exp\left(\frac{-\epsilon^2 n^2 \mathcal{B}_n^2}{C_1 n + C_2 \epsilon n \mathcal{B}_n}\right), \end{aligned} \tag{48}$$

for some positive constants C, C_1 and C_2, where I_{1j}, $I_{1j,2}$ and $I_{1j,3}$ are as defined in **Part 1**. Similarly,

$$\mathrm{P}\{h_{n,j}(-\mathcal{B}_n\epsilon) \geq 0\} \leq \mathrm{P}(I_{1j,2} + I_{1j,3} \geq C\mathcal{B}_n\epsilon)4 \\ \exp\left(\frac{-\epsilon^2 n^2 \mathcal{B}_n^2}{C_1 n + C_2 \epsilon n \mathcal{B}_n}\right). \tag{49}$$

Thus by (47), (48) and (49),

$$\sum_{j=s_n}^{p_n} \mathrm{P}(|\widehat{\beta}_j^{\mathrm{MR}}| \geq \mathcal{B}_n\epsilon) \leq 8(p_n - s_n) \\ \exp\left(\frac{-\epsilon^2 n^2 \mathcal{B}_n^2}{C_1 n + C_2 \epsilon n \mathcal{B}_n}\right) = o(1). \tag{50}$$

The equality in (50) follows from the conditions $\mathcal{B}_n = o(1)$, $\lambda_n n/s_n \to \infty$ and $\log(p_n) = o(\lambda_n^2 n^2/s_n^2)$. □

Proof of Theorem 9 For part (i), note that for $\boldsymbol{X}^o = (\boldsymbol{X}^{o(\mathrm{I})T}, \boldsymbol{X}^{o(\mathrm{II})T})^T$, $\widetilde{\boldsymbol{X}}^o = (1, \boldsymbol{X}^{oT})^T$ and $\widetilde{\boldsymbol{X}}^{o(\mathrm{I})} = (1, \boldsymbol{X}^{o(\mathrm{I})T})^T$,

$$\begin{aligned} |\widehat{m}(\boldsymbol{X}^o) - m(\boldsymbol{X}^o)| &= |F^{-1}(\widetilde{\boldsymbol{X}}^{o(\mathrm{I})T}\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}) - F^{-1}(\widetilde{\boldsymbol{X}}^{o(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})| \\ &\leq |(F^{-1})'(\widetilde{\boldsymbol{X}}^{o(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^*)| \|\widetilde{\boldsymbol{X}}^{o(\mathrm{I})T}\|_2 \|\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2, \end{aligned}$$

for some $\widetilde{\boldsymbol{\beta}}^*$ located between $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}$. By Condition A4, we conclude that $(F^{-1})'(\widetilde{\boldsymbol{X}}^{o(\mathrm{I})T}\widetilde{\boldsymbol{\beta}}^*) = O_{\mathrm{P}}(1)$. This along with $\|\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}\|_2 = O_{\mathrm{P}}(r_n)$ and $\|\widetilde{\boldsymbol{X}}^{o(\mathrm{I})}\|_2 = O_{\mathrm{P}}(\sqrt{s_n})$ implies that $|\widehat{m}(\boldsymbol{X}^o) - m(\boldsymbol{X}^o)| = O_{\mathrm{P}}(r_n\sqrt{s_n}) = o_{\mathrm{P}}(1)$. The rest of the proof is similar to that of Theorem 9 in Zhang et al. (2010) and is omitted.

For part (ii), using the proof similar to Lemma A1 of Zhang et al. (2010), we obtain that for any BD $\mathtt{Q}$ satisfying (4),

$$\mathrm{E}\{\mathtt{Q}(Y^o, \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n, \boldsymbol{X}^o\} = \mathrm{E}\{\mathtt{Q}(Y^o, m(\boldsymbol{X}^o)) \mid \boldsymbol{X}^o\} + \mathtt{Q}(m(\boldsymbol{X}^o), \widehat{m}(\boldsymbol{X}^o)).$$

It follows that

$$\begin{aligned} &\mathrm{E}\{\mathtt{Q}(Y^o, \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n\} \\ &\quad = \mathrm{E}[\mathrm{E}\{\mathtt{Q}(Y^o, \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n, \boldsymbol{X}^o\} \mid \mathcal{T}_n] \\ &\quad = \mathrm{E}[\mathrm{E}\{\mathtt{Q}(Y^o, m(\boldsymbol{X}^o)) \mid \boldsymbol{X}^o\} + \mathtt{Q}(m(\boldsymbol{X}^o), \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n] \\ &\quad = \mathrm{E}[\mathrm{E}\{\mathtt{Q}(Y^o, m(\boldsymbol{X}^o)) \mid \boldsymbol{X}^o\} \mid \mathcal{T}_n] + \mathrm{E}\{\mathtt{Q}(m(\boldsymbol{X}^o), \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n\} \\ &\quad = \mathrm{E}[\mathrm{E}\{\mathtt{Q}(Y^o, m(\boldsymbol{X}^o)) \mid \boldsymbol{X}^o\}] + \mathrm{E}\{\mathtt{Q}(m(\boldsymbol{X}^o), \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n\} \\ &\quad = \mathrm{E}\{\mathtt{Q}(Y^o, m(\boldsymbol{X}^o))\} + \mathrm{E}\{\mathtt{Q}(m(\boldsymbol{X}^o), \widehat{m}(\boldsymbol{X}^o)) \mid \mathcal{T}_n\}. \end{aligned}$$

Setting $\mathtt{Q}$ to be the misclassification loss implies

$$R(\widehat{\phi} \mid \mathcal{T}_n) \geq R(\phi_{\mathrm{B}}),$$

which combined with part (i) completes the proof. □

1.2 Additional numerical studies

1.2.1 Gaussian responses in Sect. 6.3

Random samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ of size $n = 200$ are generated from the model,

$$\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p_n})^T \sim N(\mathbf{0}, \Sigma_{p_n}), \quad Y_i \mid \boldsymbol{X}_i \sim N(\beta_{0;0} + \boldsymbol{X}_i^T \boldsymbol{\beta}_0, \sigma^2),$$

where $\beta_{0;0} = 1$, $\boldsymbol{\beta}_0 = (2, 1.5, 0.8, -1.5, 0.4, 0, \ldots, 0)^T$ with $\sigma^2 = 1$. Here $\Sigma_{p_n}(j,k) = \rho^{|j-k|}$, $j,k = 1, \ldots, p_n$, with $\rho = 0.1$. The qudartic loss is used as the BD.

Study 1 (*raw data without outliers*). For simulated data in the non-contaminated case, the results are summarized in Table 7. The robust estimators perform very similar to the non-robust counterparts.

Study 2 (*contaminated data with outliers*). For each data set generated from the model, we create a contaminated data set, where 7 data points $(X_{i,j}, Y_i)$ are contaminated as follows: They are replaced by $(X_{i,j}^*, Y_i^*)$, where $Y_i^* = Y_i \mathrm{I}\{|Y_i - m(\boldsymbol{X}_i)|/\sigma > 2\} + 15\mathrm{I}\{|Y_i - m(\boldsymbol{X}_i)|/\sigma \le 2\}$, $i = 1, \ldots, 7$,

$$\begin{aligned}
&X_{1,1}^* = 5 \ \mathrm{sign}(U_1 - .5), \quad X_{2,1}^* = 5 \ \mathrm{sign}(U_2 - .5), \quad X_{3,1}^* = 5 \ \mathrm{sign}(U_3 - .5),\\
&X_{4,3}^* = 5 \ \mathrm{sign}(U_4 - .5), \quad X_{5,5}^* = 5 \ \mathrm{sign}(U_5 - .5), \quad X_{6,9}^* = 5 \ \mathrm{sign}(U_6 - .5),\\
&X_{7,9}^* = 5 \ \mathrm{sign}(U_7 - .5),
\end{aligned}$$

with $\{U_i\} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1)$. Table 8 summarizes the results over 500 sets of contaminated data. A comparison of each estimator in Tables 7 and 8 indicates that the presence of contamination substantially increases the estimation errors $\mathrm{EE}(\widehat{\widehat{\boldsymbol{\beta}}})$ and reduces either C-Z or C-NZ. On the other hand, it is clearly observed that the non-robust estimates are more sensitive to outliers than the robust counterparts.

To further assess the impact of the sample size n on the parameter estimates, we display boxplots of $(\widehat{\beta}_j - \beta_{j;0})$, $j = 0, 1, \ldots, 8$, using the PMR selection method for the weighted-L_1 penalty, in Fig. 3 using $n = 200$ and Fig. 4 using $n = 100$, respectively. The comparison supports the consistency of both the classical and robust estimates of large dimensional model parameters for clean data as n increases, in addition to the stability of the robust estimates under a small amount of contaminated outliers.

1.2.2 Real data analysis

We consider the classification of Colon cancer data discussed in Alon et al. (1999) and available at http://genomics-pubs.princeton.edu/oncology/. It consists of 2000 genes and 62 samples, where 22 samples are from normal colon tissues and 40 samples are from tumor tissues. Similar to the analysis in Sect. 7, the data set is randomly split into two parts, with 45 samples as training samples and the rest 17 as test samples. Table 9 summarizes the average of the test errors (TE) and the average number of selected genes over 100 random splits. We observe that robust procedures tend to select fewer genes than the non-robust procedures, without getting much increase in the test errors. This lends further support to the practicality of the proposed *penalized robust*-BD *estimation*.

1.3 Numerical procedure for *penalized robust*-BD *estimator* in (3)

1.3.1 Optimization algorithm

Numerically, the *penalized robust*-BD *estimators* in (3) combined with penalties used in Sects. 6 and 7 are implemented by extending the coordinate descent (CD) iterative algorithm (Friedman et al., 2010), with the initial value $(b, 0, \ldots, 0)^T$, where $b = \log\{(\overline{Y}_n + 0.1)/(1 - \overline{Y}_n + 0.1)\}$ and $b = \log(\overline{Y}_n + 0.1)$ for Bernoulli and count responses respectively, using the sample mean $\overline{Y}_n$ of $\{Y_i\}_{i=1}^n$. Namely, the loss term

$$L(\widetilde{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}))\, w(\boldsymbol{X}_i)$$

in (3) is locally approximated by a weighted form of quadratic loss functions, and the optimization solution of (3) is obtained by the CD method. Particularly, the gradient vector and Hessian matrix of $L(\widetilde{\boldsymbol{\beta}})$ are

$$L'(\widetilde{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}})\, w(\boldsymbol{X}_i) \widetilde{\boldsymbol{X}}_i,$$

$$L''(\widetilde{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^{n} \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}})\, w(\boldsymbol{X}_i) \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^T.$$

The quadratic approximation is supported by the fact that the Hessian matrix of $L(\widetilde{\boldsymbol{\beta}})$ evaluated at the true parameter vector $\widetilde{\boldsymbol{\beta}}_0$ is

$$\begin{aligned} L''(\widetilde{\boldsymbol{\beta}}_0) &= n^{-1} \sum_{i=1}^{n} \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{\beta}}_0)\, w(\boldsymbol{X}_i) \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^T \\ &= \mathrm{E}[\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\}\, w(\boldsymbol{X}) \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T] + o_{\mathrm{P}}(1), \end{aligned}$$

which, combined with the property $\mathrm{E}\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\beta}}_0) \mid \boldsymbol{X}\} \ge 0$ discussed in part **(d)** of Sect. 2.2, indicates that with probability tending to one, the matrix $L''(\widetilde{\boldsymbol{\beta}}_0)$ is positive semidefinite.

Both $\mathrm{p}_1(y; \theta)$ and $\mathrm{p}_2(y; \theta)$ in $L'(\widetilde{\boldsymbol{\beta}})$ and $L''(\widetilde{\boldsymbol{\beta}})$ are calculated using (9), which incorporates the Huber and Tukey ψ-functions whose derivatives $\psi'(r)$ can be substituted by its subgradient or approximation.

1.3.2 Pseudo codes, source codes and computational complexity analysis

Algorithm 1 summarizes the complete procedure for numerically solving the "*penalized robust*-BD *estimator*" in (3).

Table 9 (Real data) Classification for the Colon cancer data

			Deviance loss		Exponential loss	
Procedure	$\psi(r)$	Penalty	TE	# genes	TE	# genes
non-robust	r	SCAD	0.224	9.95	0.217	10.68
		L_1	0.190	18.64	0.204	16.26
		wL_1, MR	0.155	10.65	0.156	9.51
		wL_1, PMR	0.161	10.89	0.165	9.43
robust	Huber	SCAD	0.268	3.40	0.243	5.02
		L_1	0.282	5.84	0.226	10.42
		wL_1, MR	0.231	4.93	0.172	6.83
		wL_1, PMR	0.214	5.97	0.180	6.79
robust	Tukey	SCAD	0.247	3.91	0.263	3.84
		L_1	0.230	11.30	0.228	10.91
		wL_1, MR	0.185	7.45	0.166	7.54
		wL_1, PMR	0.184	7.90	0.169	7.31

Algorithm 1 Numerical procedure for the "*penalized robust*-BD *estimator*"

```
Require: Data $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$, dimension $p_n$, type of the response variable $Y$, link function $F$ in (1), variance function $V$ in (2), the generating $q$-function of BD, weight function $w(\boldsymbol{x})$, Huber or Tukey $\psi$-function with a constant $c > 0$.
1: Compute the Pearson residual $r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$, and construct the robust-BD $\rho_q(y, \mu)$ from (6).
2: Select the data-driven penalty weights $\{\widehat{w}_{n,j}\}_{j=1}^{p_n}$ by the PMR method in (15)–(16) or the MR method in (17)–(18); select the data-driven regularization parameter $\widehat{\lambda}_n$ on a grid via cross-validation or other prediction-based criterion. For the PMR method, each $\kappa_n$ on a grid $\mathcal{K}$ yields a selected $\lambda_n$ and a test error; take the optimal $\widehat{\kappa}_n$ which minimizes the test error over the grid $\mathcal{K}$ to select the final $\widehat{\lambda}_n$.
3: Use the data-driven $\{\widehat{w}_{n,j}\}_{j=1}^{p_n}$ and $\widehat{\lambda}_n$ to compute the criterion function (3).
4: Minimize the criterion function (3) over $(\beta_0, \boldsymbol{\beta})$ via the optimization algorithm in Section A.4.1.
Ensure: The "penalized robust-BD estimator" $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})$.
```

To illustrate the computational complexity analysis, Tables 10 and 11 compare runtime of the non-robust and robust procedures. All computations are performed using MATLAB (Version: 9.12.0.1956245 (R2022a) Update 2) on Windows 11, 12th Gen Intel(R) Core(TM) i9-12900, 2400 Mhz, 16 Core(s), 24 Logical Processors. MATLAB source codes are available at GitHub https://github.com/ChunmingZhangUW/Robust_penalized_BD_high_dim_GLM. For either clean or contaminated data, the algorithmic complexity depends on

Table 10 The total CPU time (in seconds) for overdispersed Poisson responses in 500 replications

	Data	Procedure	n	p_n	runtime
Table 1	raw	non-robust	100	50	82
	Raw	non-robust	100	500	865
	raw	robust	100	50	14404
	raw	robust	100	500	16319
Table 2	contaminated	non-robust	100	50	94
	contaminated	non-robust	100	500	1070
	contaminated	robust	100	50	18185
	contaminated	robust	100	500	21437

Table 11 The total CPU time (in seconds) for Gaussian responses in 500 replications

	Data	Procedure	n	p_n	Runtime
Table 7	raw	non-robust	200	50	53
	raw	non-robust	200	500	672
	raw	robust	200	50	91
	raw	robust	200	500	800
Table 8	contaminated	non-robust	200	50	63
	contaminated	non-robust	200	500	911
	contaminated	robust	200	50	96
	contaminated	robust	200	500	899
Figure 4	raw	non-robust	100	50	49
	raw	non-robust	100	500	429
	raw	robust	100	50	106
	raw	robust	100	500	1154
Figure 4	contaminated	non-robust	100	50	67
	contaminated	non-robust	100	500	610
	contaminated	robust	100	50	120
	contaminated	robust	100	500	1246

the type of response variables, the dimensionality and the procedure. Poisson-type responses are more computationally intensive than Gaussian responses; robust procedures are slower than the non-robust counterparts; higher dimensions demand more computational costs than lower dimensional settings.

Author Contributions CM Zhang contributed to the proof, computation, and writing; LXZ contributed to the discussion, proof, and writing; YBS contributed to the implementation of SVM and robust-SVM.

Funding The work was partially supported by the U.S. National Science Foundation grants DMS-2013486 and DMS-1712418, and provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Data availability The Lymphoma data studied in Sect. 7 is publicly available from Alizadeh et al. (2000); the Colon cancer dataset in Appendix 1.2.2 is at http://genomics-pubs.princeton.edu/oncology/.

Code availability MATLAB codes are available at https://github.com/ChunmingZhangUW/Robust_penalized_BD_high_dim_GLM.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., … Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA, 96*, 6745–6750.

Altun, Y., & Smola, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In G. Lugosi & H. U. Simon (Eds.), *Learning theory: 19th annual conference on learning theory* (pp. 139–153). Springer.

Bianco, A. M., & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994)* (Vol. 109, pp. 17–34), Lecture Notes in Statist., Springer.

Boente, G., He, X., & Zhou, J. (2006). Robust estimates in generalized partially linear models. *Annals of Statistics, 34*, 2856–2878.

Brègman, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics, 7*, 620–631.

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

Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association, 96*, 1022–1030.

Dupuis, D. J., & Victoria-Feser, M.-P. (2011). Fast robust model selection in large datasets. *Journal of the American Statistical Association, 106*, 203–212.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81*, 461–470.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics, 32*, 928–961.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*, 746–762.

Gong, P., Zhang, C., Lu, Z., Huang, J., & Ye, J. (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *The 30th international conference on machine learning (ICML 2013)*.

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics, 32*, 1367–1433.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. Wiley.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73–101.

Kanamori, T., Takenouchi, T., Eguchi, S., & Murata, N. (2007). Robust loss functions for boosting. *Neural Computation, 19*, 2183–2244.

Künsch, H., Stefanski, L., & Carroll, R. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association, 84*, 460–466.

Lafferty, J. D., Della Piestra, S., & Della Piestra, V. (1997). Statistical learning algorithms based on Bregman distances. In *Proceedings of the 5th Canadian workshop on information theory.*

Lafferty, J. (1999). Additive models, boosting and inference for generalized divergences. In: *Proceedings of the twelfth annual conference on computational learning theory* (pp. 125–133). ACM Press.

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B, 70*, 53–71.

Stefanski, L., Carroll, R., & Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika, 73*, 413–424.

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

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer.

Vapnik, V. (1996). *The nature of statistical learning theory*. Springer.

Vemuri, B. C., Liu, M., Amari, S.-I., & Nielsen, F. (2011). Total Bregman divergence and its applications to DTI analysis. *IEEE Transactions on Medical Imaging, 30*, 475–483.

Wright, S. J. (1997). *Primal-dual interior-point methods*. SIAM.

Wu, Y. C., & Liu, Y. F. (2007). Robust truncated hinge Loss support vector machines. *Journal of the American Statistical Association, 102*, 974–983.

Zhang, C. M., Jiang, Y., & Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics, 37*, 119–139.

Zhang, C. M., Jiang, Y., & Chai, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika, 97*, 551–566.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418–1429.

Zou, H., & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics, 36*, 1108–1126.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.