Annals of the Institute of Statistical Mathematics (2021) 73:703–736
https://doi.org/10.1007/s10463-020-00764-1

Robust high-dimensional regression for data with anomalous responses

Mingyang Ren[1,2] · **Sanguo Zhang**[1,2] · **Qingzhao Zhang**[3]

Received: 10 March 2020 / Revised: 28 June 2020 / Published online: 30 September 2020

Abstract

The accuracy of response variables is crucially important to train regression models. In some situations, including the high-dimensional case, response observations tend to be inaccurate, which would lead to biased estimators by directly fitting a conventional model. For analyzing data with anomalous responses in the high-dimensional case, in this work, we adopt γ-divergence to conduct variable selection and estimation methods. The proposed method possesses good robustness to anomalous responses, and the proportion of abnormal data does not need to be modeled. It is implemented by an efficient coordinate descent algorithm. In the setting where the dimensionality p can grow exponentially fast with the sample size n, we rigorously establish variable selection consistency and estimation bounds. Numerical simulations and an application on real data are presented to demonstrate the performance of the proposed method.

Keywords Anomalous responses · Robust · γ-divergence · High-dimensional data

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-020-00764-1) contains supplementary material, which is available to authorized users.

✉ Qingzhao Zhang
qzzhang@xmu.edu.cn

[1] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[2] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100049, China

[3] Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics, MOE Key Lab of Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen 361005, China

1 Introduction

In the regression model, the prediction rule is to be derived from labeled dataset. Traditional regression models assume and expect the correct response variables; however, it is expensive and difficult to obtain accurate responses because of insufficient information, subjective judgment, measurement error and so on, which would lead to biased estimators by directly fitting conventional methods (Piepel 2005).

Anomalous responses would be encountered in the fields of Internet, finance, image processing, biology and so on. For instance, the real data studied in our paper contains mislabeled responses owing to the measurement error of the expression of receptor genes (Lopes et al. 2018). Traditional regression models are not applicable to this kind of data. It is noteworthy that "mislabeled data" in discrete variables like that is an important special case of anomalous responses, which is also called "label noise" (Rebbapragada and Brodley 2007; Frénay and Verleysen 2013) or "misclassification" (Copeland et al. 1977; Grace 2017) in classification problems and "count error" (Cameron and Trivedi 2013) in count data.

Samples with anomalous responses can be considered as outliers. The existing approaches for high-dimensional data can be broadly classified into one of two types. The first type is to filter out outliers. For example, Aggarwal and Yu (2001) developed a distance-based outlier detection method; She and Owen (2011) proposed the individual intercept model to identify outliers for the linear model; Zimek et al. (2012) considered using the clusters to detect mislabeling outliers. However, there might be information loss caused by removing some samples when using this type of method. The second type is to reduce weights on samples with anomalous responses, which might be more applicable to the parameter estimation. Take logistic regression for binary classification as an example. Weight functions have different forms according to certain assumptions on mislabel probabilities, for instance, the mislabel probability regarded as an equal constant (Copas 1988), related to covariates (Hayashi 2012), asymmetric (Komori et al. 2016) and so on. But these methods might not be applicable when the data do not follow corresponding assumptions of mislabel probabilities. Hung et al. (2018) proposed γ-logistic regression. Mislabel probabilities do not need to be modeled using γ-logistic regression because the bias from contamination distribution and contamination proportion can be ignored, which benefits from the robustness of γ-divergence. However, they only focus on binary data without considering "count error" or other types of anomalous responses. In addition, it is not applicable to "large p, small n" data. Besides, Kawashima and Fujisawa (2017) proposed the robust and sparse regression via γ -divergence and presented the robust properties from two viewpoints of latent bias and Pythagorean relation. However, their method cannot deal with high-dimensional data with discrete responses, and they did not theoretically study the consistency of estimation and variable selection. To address these challenges, in this article, we consider a penalized generalized linear model based on γ-divergence.

In summary, our contributions are the following. On one hand, we are the first to adopt γ-divergence on high-dimensional generalized linear model to deal with

multiple types of anomalous responses, which includes mislabeled data as a special case, and contamination proportion need not be modeled. Numerical simulation and real data analysis are presented to demonstrate the good performance of the proposed method. On the other hand, asymptotic properties of γ-divergence in the high-dimensional case are studied. We rigorously establish variable selection consistency and estimation bounds under the setting where the dimensionality p can grow exponentially fast with the sample size n. It is not easy to establish high-dimensional asymptotic properties due to the complexity of γ-divergence.

2 Methodology

2.1 The robust penalized γ-divergence estimation

The γ-divergence approach is firstly introduced in Roberts and Stramer (2001) for the robust estimation of a single distribution parameter. It is extended to the robust regression method with low-dimensional data later (Hung et al. 2018; Fujisawa and Eguchi 2008). It has been shown that the γ-divergence method has multiple statistical and numerical advantages over the nonrobust and robust alternatives.

Let $f_{\boldsymbol{\beta}}$ be the model distribution under the p-dimensional parameter $\boldsymbol{\beta}$ and g be the data generating distribution. For these two density functions, the γ-divergence is defined as

$$D_\gamma(g, f_{\boldsymbol{\beta}}) = \frac{1}{\gamma(\gamma+1)}\left\{\|g\|_{\gamma+1} - \int \left(\frac{f_{\boldsymbol{\beta}}}{\|f_{\boldsymbol{\beta}}\|_{\gamma+1}}\right)^{\gamma} g\right\}, \gamma > 0. \tag{1}$$

The parameter γ balances efficiency and robustness, with a smaller γ corresponding to more efficient but less robust estimation. In the limiting case, $D_\gamma(g, f_{\boldsymbol{\beta}})$ is a version of the Kullback–Leibler divergence as $\gamma \to 0$.

Remark 1 If data is contaminated, $g = cf_{\boldsymbol{\beta}^*} + (1-c)h \doteq g_c$ with contamination distribution h, contamination proportion $1-c$ with $0.5 < c \leqslant 1$ and the true model parameter $\boldsymbol{\beta}^*$. The estimated parameter is obtained by minimizing γ-divergence $D_\gamma(g_c, f_{\boldsymbol{\beta}})$, which is equivalent to minimizing $cD_\gamma(f_{\boldsymbol{\beta}^*}, f_{\boldsymbol{\beta}}) - F_{\boldsymbol{\beta}}(c, h, \boldsymbol{\beta}, \gamma)$ with $F_\beta(c, h, \boldsymbol{\beta}) = \gamma^{-1}(\gamma+1)^{-1}(1-c)\int \frac{f_{\boldsymbol{\beta}}^\gamma}{\|f_{\boldsymbol{\beta}}\|_{\gamma+1}^\gamma} h$. Suppose that $\int f_{\boldsymbol{\beta}^*}^\gamma h$ is sufficiently small for an appropriately large $\gamma > 0$, which implies that the contamination density h mostly lies on the tail of the underlying density $f_{\boldsymbol{\beta}^*}$. Then, for some γ, the bias $F_{\boldsymbol{\beta}}(c, h, \boldsymbol{\beta}, \gamma)$ is negligibly small when $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}^*$. Namely, the estimation of $\boldsymbol{\beta}$ is less affected by $1-c$ and h. More detailed discussions about this property of γ-divergence could be referred to Fujisawa and Eguchi (2008).

Consider n independent samples $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is the random covariate vector with the dimension p and y_i is the response for ith sample. Ignoring terms independent of the unknown parameter, the empirical version of the γ-divergence loss function is

$$\ell(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{f\left(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}\right)^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y\right)^{\gamma/(1+\gamma)}}, \tag{2}$$

where $f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$ is the conditional probability density function of y_i given $\boldsymbol{x}_i$ with the unknown parameter vector $\boldsymbol{\beta}$. To deal with multiple types of responses, we consider generalized linear model (GLM), in which

$$f\left(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}\right) = c\left(y_i\right)\exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi}\right\},$$

where $\theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i$, and $b(\theta)$ is twice continuously differentiable with $b''(\theta)$ always positive. In this paper, we are interested in sparsity estimation of the regression coefficients $\boldsymbol{\beta}$, and assume that the dispersion parameter ϕ is known. If unknown, we can estimate it by solving the estimation equation using the biselection method (Zang et al. 2017) or linear search methods (Ghosh and Basu 2016).

For analyzing high-dimensional data, we propose the following robust penalized γ-divergence loss function

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{3}$$

where $\ell(\boldsymbol{\beta})$ is defined in (2), β_j is the jth component of $\boldsymbol{\beta}$ and p_λ is a concave penalty function such as SCAD with first-order derivative $p'_\lambda(t) = \lambda\{I(t \leqslant \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda}I(t > \lambda)\}$, for $a > 2, t \geqslant 0$ (Fan and Li 2001) or MCP with first-order derivative $p'_\lambda(t) = \lambda(1 - \frac{t}{a\lambda})_+$, for $a > 1, t \geqslant 0$ (Zhang 2010). The objective function $Q(\boldsymbol{\beta})$ consists of the robust loss function dealing with outliers and a concave penalty on $\boldsymbol{\beta}$. The characteristics of the robust loss function are considered in Remark 2 as follows.

Remark 2 For logistic regression, the loss function $\ell(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\omega_{\gamma,i}(\boldsymbol{\beta})$, where the weight function $\omega_{\gamma,i}(\boldsymbol{\beta}) = \left(\frac{\exp\{y_i(\gamma+1)\boldsymbol{\beta}^T\boldsymbol{x}_i\}}{1+\exp\{(\gamma+1)\boldsymbol{\beta}^T\boldsymbol{x}_i\}}\right)^{\gamma/(1+\gamma)}$, which is so-called γ-logistic regression without modeling mislabel probabilities. From the estimation equation $\sum_{i=1}^{n}\omega_{\gamma,i}(\boldsymbol{\beta})\left[y_i - \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}\right]\boldsymbol{x}_i = 0$, the robustness of γ-logistic regression is clear: $\omega_{\gamma,i}(\boldsymbol{\beta})$ could be small with non-matched $(y_i, \boldsymbol{\beta}^T\boldsymbol{x}_i)$. See Hung et al. (2018) for further discussions.

It is remarkable that γ-divergence used in Hung et al. (2018) is focused on analyzing binary data in the low-dimensional case. More generally, GLM combining γ-divergence considered in this article can deal with more types of response variables. γ-logistic regression is a special case of the proposed analysis framework. Although the γ-divergence has been previously adopted in regression analysis, to our best knowledge, this study is the first to adopt γ-divergence to deal with high-dimensional data with multiple types of responses.

2.2 Statistical properties

In this section, we establish the statistical properties of penalized γ-divergence estimation. Write the true coefficient as $\boldsymbol{\beta}^*$ and the important predictor index set is labeled as S. $|S|$ is the cardinality of set S. Let $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ represent the components of $\boldsymbol{\beta}^*$ indexed by S and its complement, respectively. Denote $\boldsymbol{X}_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{n1})^T$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{n2})^T$ are the submatrices of the design matrix $\boldsymbol{X}^T$ formed by columns in S and its complement. Define the oracle estimator as $\hat{\boldsymbol{\beta}}^{\text{oracle}^T} = \{\hat{\boldsymbol{\beta}}_1^{\text{oracle}^T}, \hat{\boldsymbol{\beta}}_2^{\text{oracle}^T}\}$, with $\hat{\boldsymbol{\beta}}_1^{\text{oracle}} = \hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2^{\text{oracle}} = \mathbf{0}$, where

$$\hat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}_1}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \frac{f\left(y_i|\boldsymbol{x}_{i1};\boldsymbol{\beta}_1\right)^{\gamma}}{\left(\int f(y|\boldsymbol{x}_{i1};\boldsymbol{\beta}_1)^{1+\gamma} \mathrm{d}y\right)^{\gamma/(1+\gamma)}} \right\}, \tag{4}$$

$\boldsymbol{\beta}_1 \in \mathbf{R}^{|S|}$. We have $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} I_i(\boldsymbol{\beta}) \boldsymbol{x}_i$ and $\nabla_{\boldsymbol{\beta}} I_i(\boldsymbol{\beta}) = H_i(\boldsymbol{\beta}) \boldsymbol{x}_i$, where $\nabla_{\boldsymbol{\beta}}$ represents the gradient to $\boldsymbol{\beta}$, I_i and H_i are defined by (7) and (9) in "Appendix", respectively. Write $H(\boldsymbol{\beta}) = \operatorname{diag}\{H_1(\boldsymbol{\beta}), \ldots, H_n(\boldsymbol{\beta})\}$. Take

$$\begin{aligned} &\rho_0 = \sigma_{\max}\{n^{-1} E(\boldsymbol{X}_1^T \boldsymbol{X}_1)\}, \quad \rho_1 = \sigma_{\min}\{n^{-1} E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*) \boldsymbol{X}_1)\}, \\ &C = \|E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*) \boldsymbol{X}_1)[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*) \boldsymbol{X}_1)]^{-1}\|_{\infty}, \end{aligned}$$

where $\sigma_{\min}\{\cdot\}$ and $\sigma_{\max}\{\cdot\}$ denote the smallest and largest eigenvalues of the matrix, respectively, and $\|\cdot\|_{\infty}$ denotes the maximum absolute row sum of the matrix.

To establish statistical properties, the following conditions are needed.

Condition 1. $\lambda^{-1} p_{\lambda}(t)$ is concave in $t \in [0, \infty)$ with a continuous derivative $\lambda^{-1} p'_{\lambda}(t)$ satisfying $\lambda^{-1} p'_{\lambda}(0+) \in (0, \infty)$. $\lambda^{-1} p'_{\lambda}(t)$ is increasing in $\lambda \in (0, \infty)$ and $\lambda^{-1} p'_{\lambda}(0+)$ is independent of λ.
Condition 2. $\varrho = \inf\{t/\lambda : \lambda^{-1} p_{\lambda}(t) = 0, t \geqslant 0\}$ is bounded.
Condition 3. X is a bounded matrix almost surely and the bound is κ.

Condition 1 is considered by Fan and Lv (2011); the SCAD and MCP penalties both satisfy this condition. And these two penalties satisfy Condition 2, which can guarantee unbiasedness, with $\varrho = a$. Condition 3 is assumed to simplify the proof, where κ is a bounded constant related to estimation and probability bounds. The high-dimensional asymptotic properties are as below and the proof is placed in "Appendix".

Theorem 1 *Let Condition 1–3 hold.*

1. For any $\varepsilon < \sqrt{n/|S|}$, *we have*

$$\Pr\{\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leqslant \sqrt{|S|/n}\varepsilon\} \geqslant 1 - \exp\left(-\frac{|S|\rho_1^2}{512\rho_0 M^2}\varepsilon^2\right) - 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2 M^2 \kappa^2}\right) - 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right),$$

where M is a sufficient large bounded constant.

2. *Suppose $2a\lambda < \min_{j\in S}|\beta_{1j}^*|$ and $\varepsilon = o(\sqrt{n\lambda/|S|})$. Then with probability at least*

$$1 - \exp\left(-\frac{|S|\rho_1^2}{512\rho_0 M^2}\varepsilon^2\right) - 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2 M^2\kappa^2}\right) - 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right) - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2 C^2}{2|S|^3}\right) - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2}{8|S|^3}\right) - 2p\cdot \exp\left(-\frac{n\lambda^2}{2M^2\kappa^2(2C+1)^2}\right),$$

$\hat{\beta}^{\text{oracle}}$ is a local minimizer of (3).

The variable selection consistency and estimation bounds are described in Theorem 1. Furthermore, a more straightforward corollary, contributing to understanding of the penalized γ-divergence estimation, could be obtained using the above probability bounds.

Corollary 1 *Suppose that ρ_0 and ρ_1 are bounded away from zero and infinity, $|S| \ll n$, $\log(p) = O(n^\alpha)$ with $\alpha \leqslant 1$, and $C = O(n^{\alpha_1})$ with $\alpha_1 \in [0, 1/2)$. Under Condition 1–3, if $2a\lambda < \min_{j\in S}\left|\beta_{1j}^*\right|$ and $\lambda \gg n^{(\alpha-1)/2+\alpha_1}$, then the true sparsity structure can be identified and $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_p(\sqrt{|S|/n})$.*

Corollary 1 shows that the penalized γ-divergence estimator could achieve the consistency rate of $O_p(\sqrt{|S|/n})$, which is same as that of the penalized maximum likelihood estimator (see Fan and Lv 2011 and others for reference). For properties of the γ-divergence estimator, previous studies only focus on the low-dimensional case, such as asymptotic normality based on M-estimation theory (Hung et al. 2018; Fujisawa and Eguchi 2008). In this paper, we establish variable selection consistency and estimation bounds based on γ-divergence in the high-dimensional case.

2.3 Computation

When we minimize objective function (3), the coordinate descend algorithm can be adopted. Let ∇_j represent the derivative to β_j. A simple calculation shows that

$$\nabla_j \ell(\boldsymbol{\beta}) = -\frac{\gamma}{n}\sum_{i=1}^{n}\omega_{\gamma,i}(\boldsymbol{\beta})\left[K_i(y_i;\boldsymbol{\beta}) - \frac{\int K_i(y;\boldsymbol{\beta})f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y}{\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y}\right]X_{ij} \tag{5}$$

where $\omega_{\gamma,i}(\boldsymbol{\beta}) = \frac{f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y\right)^{\frac{\gamma}{1+\gamma}}}$, $K_i(y_i;\boldsymbol{\beta}) = \frac{y_i-\mu_i}{\mathrm{Var}\,(y_i)q'(\mu_i)}$ and $\mu_i = E(y_i)$ is linked to θ_i through the canonical link function $q(\cdot)$ in GLM. The overall algorithm is described in Algorithm 1. As two special cases, the expressions of the gradient (5) for logistic regression and Poisson regression are given by (28) and (29) in "Appendix B2".

We adopt the MCP penalty with first-order derivative $p'_{\lambda}(t) = \lambda(1-\frac{t}{a\lambda})_+$ for $a > 1$ and $t \geqslant 0$, which contains the tuning parameter λ and the regularization parameter a. Following Zhang (2010), we set $a = 3$. The robust parameter γ can balance robustness and estimation efficiency; however, there is no consistent methods for selecting γ. Bayes Information Criterion (BIC)-type criteria are able to identify the true model consistently (Wang et al. 2007; Wu and Wang 2020). This motivated us to select the optimal (λ,γ) by minimizing the following adjusted Bayes Information Criterion

$$\mathrm{BIC}_{\lambda,\gamma} = \ell(\boldsymbol{\beta}) + \frac{\nu(\lambda,\gamma)\log(n)}{\delta n},$$

where $\ell(\boldsymbol{\beta})$ is defined in (2), $\nu(\lambda,\gamma)$ is the number of nonzero coefficients and $\delta > 1$ is an adjustment coefficient. It is of interest to note that the variation in $\ell(\boldsymbol{\beta})$ is not large under different numbers of nonzero coefficients due to the robustness of γ -divergence. As a result, original BIC would provide too much penalty for variables in the γ-divergence method. Thus, δ in adjusted Bayes Information Criterion should result in a weaker penalty. We set $\delta = 8$ in all numerical studies, which leads to satisfactory performance. As the sensitivity analysis, some simulation results under various choices of δ are shown in Table S3 (Supplementary Materials). Overall, the proposed approach is not much sensitive to the choice of δ when it is in a sensible range.

Algorithm 1 Coordinate Descent Algorithm

Input: Response labels $\boldsymbol{Y}$, predictor variables $\boldsymbol{X}$ and tuning parameters (γ, a, λ).

Output: The estimate of regression coefficients: $\hat{\boldsymbol{\beta}}$.

Initialization: $k = 0$, the convergence threshold $\varepsilon = 10^{-3}$, the cut-off point $\epsilon_c = 10^{-4}$ and given an initial value $\boldsymbol{\beta}^{(0)}$.

repeat

$k = k + 1$;

for $j = 1, \cdots, p$ **do**

Compute $g_j \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} = \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} = \nabla_j \ell(\boldsymbol{\beta}^{(k)}) + p'_{\lambda}(|\beta_j^{(k)}|)\mathrm{sgn}(\beta_j^{(k)})$, where $\nabla_j \ell(\boldsymbol{\beta}^{(k)})$ is defined in (5);

Calculate $\beta_j^{(k+1)} = \beta_j^{(k)} - \alpha \cdot g_j \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}$, where the step length α is obtained by Armijo search.

end for

until $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2 \leqslant \varepsilon$.

cut-off Components of $\boldsymbol{\beta}^{(k+1)}$ at convergence smaller than ϵ_c are set to 0.

return The cut-off estimate $\hat{\boldsymbol{\beta}}$.

The proposed algorithm is computationally affordable. For instance, the analysis of one simulated dataset with $n = 400, p = 1000$ takes less than 10 min on a regular PC. The convergence is achieved in our numerical studies and real data analysis within 30 overall iterations.

3 Simulation studies

In this section, we consider two cases: logistic regression for binary data and Poisson regression for count data. In each case, nonrobust and robust alternatives are compared and two oracle estimators are considered. γ-divergence is used based on known true important variables in the first estimator (Oracle1). In the second estimator (Oracle2), the true label and true important variables are both known and the conventional regression method is used. All simulations are based on 100 repetitions and conducted using MATLAB codes.

We denote $S \equiv \{j : \beta_j \neq 0\}$ as the set of predictor variables that contributes to the model, $\hat{S} \equiv \{j : \hat{\beta}_j \neq 0\}$ as the set of selected predictor variables. In simulation results, the column labeled "TP" shows the mean and standard deviation of numbers of true positives ($|\hat{S} \cap S|$), and the column labeled "FP" shows the mean and standard deviation of numbers of false positives ($|\hat{S} \setminus S|$). We consider $n = 200, 400$ and $p = 1000, 2000$ in all simulations.

3.1 Case 1: Logistic regression for binary data

In this case, we numerically compare the proposed penalized γ-logistic regression with penalized conventional logistic regression and penalized constant-mislabel logistic regression (Copas 1988). We evaluate the performances of variable selection, parameter estimation and prediction of three methods.

Predictor variables x are from p-dimensional normal distribution $N(0, \Sigma)$. Consider two structures of covariance matrix $\Sigma = \left(\sigma_{ij}\right)_{1 \leqslant i,j \leqslant p}$. The first structure is autoregressive correlation (AR) given by $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.25$ and 0.75. The second structure is banded correlation, and two scenarios were considered: in the first scenario, $\sigma_{ij} = 0.33$ if $|i - j| = 1$, and 0 otherwise; in the second, $\sigma_{ij} = 0.6$ if $|i - j| = 1$, 0.33 if $|i - j| = 2$, and 0 otherwise. The response variables are generated from Bernoulli $\{P(y = 1|\boldsymbol{x})\}$, where $P(y = 1|\boldsymbol{x}) = \tau_0\{1 - \pi(\boldsymbol{x};\boldsymbol{\beta}^*)\} + \left\{1 - \tau_1\right\}\pi(\boldsymbol{x};\boldsymbol{\beta}^*)$ with mislabel probabilities $\tau_0 = P\big(y = 1|y_0 = 0, \boldsymbol{x}\big)$, $\tau_1 = P\big(y = 0|y_0 = 1, \boldsymbol{x}\big)$ and the true label y_0's probability of success $\pi(\boldsymbol{x};\boldsymbol{\beta}^*) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}^*)}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta}^*)}$. We consider two mechanisms of mislabel probabilities: (S1) $\tau_0 = m_0$ and $\tau_1 = m_1$; (S2) $\tau_0 = m_0$ and $\tau_1 = m_0 + (m_1 - m_0)\pi(\boldsymbol{x};\boldsymbol{\beta}^*)$. Setting (S1) considers constant-mislabel probabilities, while setting (S2) considers mislabel probabilities related to $\boldsymbol{x}$. There are a total of 16 nonzero effects and coefficients are randomly generated from uniform distribution $U[0.5, 1.5]$. When evaluating the performances of variable selection and parameter estimation, we set $(m_0, m_1) = (0.05, 0.2)$ under (S1) and (S2). And the case

Table 1 Mean (SD) of numbers of true/false positives (TP/FP) for variable selection under logistic regression with $p = 1000$

Correlation	n	Methods	S0			S1			S2		
			TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$
AR $\rho = 0.25$	200	Logistic	15.84(0.37)	0.08(0.27)	1.83(0.68)	12.04(2.65)	1.61(1.30)	3.97(0.45)	14.24(2.83)	2.62(1.52)	4.24(0.57)
		Constant	15.81(0.39)	0.04(0.20)	1.76(0.74)	12.96(2.29)	1.34(1.19)	3.65(0.47)	14.80(1.95)	1.45(1.13)	3.63(0.56)
		γ-logistic	15.55(0.92)	0.08(0.32)	1.90(0.51)	15.22(1.10)	0.04(0.20)	2.54(0.53)	15.43(0.67)	0.14(0.35)	2.46(0.51)
		Oracle1	16.00(0.00)	0.00(0.00)	1.64(0.46)	16.00(0.00)	0.00(0.00)	1.96(0.46)	16.00(0.00)	0.00(0.00)	1.94(0.46)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.68(0.44)	16.00(0.00)	0.00(0.00)	1.72(0.41)
	400	Logistic	15.83(0.38)	0.07(0.26)	1.27(0.56)	13.75(1.67)	0.47(0.69)	3.03(0.27)	14.35(1.50)	1.59(1.25)	3.19(0.28)
		Constant	15.82(0.39)	0.05(0.22)	1.32(0.62)	14.77(1.58)	0.41(0.49)	2.72(0.32)	14.81(1.44)	1.38(0.96)	2.82(0.30)
		γ-logistic	15.77(0.45)	0.06(0.24)	1.82(0.48)	15.39(0.68)	0.07(0.26)	1.45(0.50)	15.50(0.92)	0.16(0.37)	1.37(0.38)
		Oracle1	16.00(0.00)	0.00(0.00)	0.99(0.26)	16.00(0.00)	0.00(0.00)	1.17(0.32)	16.00(0.00)	0.00(0.00)	1.24(0.30)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.05(0.30)	16.00(0.00)	0.00(0.00)	1.02(0.29)
AR $\rho = 0.75$	200	Logistic	15.66(0.87)	0.23(0.64)	2.47(0.98)	10.79(1.73)	1.83(1.29)	4.02(0.33)	13.43(2.02)	3.64(1.83)	4.64(0.57)
		Constant	15.58(0.93)	0.18(0.42)	2.44(0.92)	11.48(1.76)	1.49(1.14)	3.81(0.39)	14.76(1.71)	1.64(1.12)	3.65(0.59)
		γ-logistic	15.19(1.25)	0.11(0.36)	2.27(0.61)	14.85(1.88)	0.11(0.31)	3.02(0.61)	15.40(0.73)	0.25(0.44)	2.90(0.64)
		Oracle1	16.00(0.00)	0.00(0.00)	1.89(0.49)	16.00(0.00)	0.00(0.00)	2.24(0.54)	16.00(0.00)	0.00(0.00)	2.35(0.53)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.84(0.45)	16.00(0.00)	0.00(0.00)	1.86(0.40)
	400	Logistic	15.69(0.83)	0.15(0.42)	1.88(0.86)	13.61(1.91)	1.49(1.68)	3.31(0.23)	13.64(1.62)	1.85(1.41)	3.48(0.22)
		Constant	15.59(1.04)	0.13(0.34)	1.75(0.79)	14.60(1.69)	1.40(0.59)	2.99(0.30)	14.79(1.65)	1.50(1.20)	3.12(0.28)
		γ-logistic	15.50(1.16)	0.12(0.33)	1.24(0.39)	15.25(0.84)	0.06(0.24)	1.71(0.61)	15.48(0.76)	0.23(0.42)	1.69(0.51)
		Oracle1	16.00(0.00)	0.00(0.00)	1.15(0.22)	16.00(0.00)	0.00(0.00)	1.30(0.36)	16.00(0.00)	0.00(0.00)	1.41(0.38)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.14(0.24)	16.00(0.00)	0.00(0.00)	1.17(0.29)

Table 1 (continued)

Correlation	n	Methods	S0			S1			S2		
			TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$
Band1	200	Logistic	15.85(0.36)	0.03(0.17)	1.67(0.68)	12.52(3.00)	1.52(1.31)	3.96(0.50)	14.55(3.14)	2.22(1.40)	4.09(0.56)
		Constant	15.86(0.35)	0.03(0.17)	1.59(0.73)	13.54(2.49)	1.29(1.21)	3.58(0.50)	14.84(2.04)	1.14(0.67)	3.63(0.55)
		γ-logistic	15.68(0.82)	0.05(0.22)	1.76(0.47)	15.28(0.79)	0.01(0.10)	2.35(0.49)	15.43(0.65)	0.11(0.31)	2.30(0.46)
		Oracle1	16.00(0.00)	0.00(0.00)	1.54(0.41)	16.00(0.00)	0.00(0.00)	1.85(0.43)	16.00(0.00)	0.00(0.00)	1.79(0.42)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.60(0.39)	16.00(0.00)	0.00(0.00)	1.52(0.38)
	400	Logistic	15.96(0.20)	0.02(0.14)	0.93(0.33)	13.78(1.61)	0.21(0.41)	2.96(0.28)	14.62(1.46)	1.49(1.18)	3.08(0.30)
		Constant	15.95(0.22)	0.03(0.17)	1.02(0.63)	14.82(1.55)	0.16(0.37)	2.65(0.32)	14.96(1.37)	1.33(0.87)	2.71(0.31)
		γ-logistic	15.93(0.26)	0.04(0.20)	0.94(0.24)	15.52(0.64)	0.07(0.26)	1.38(0.47)	15.53(0.99)	0.06(0.24)	1.26(0.33)
		Oracle1	16.00(0.00)	0.00(0.00)	0.92(0.23)	16.00(0.00)	0.00(0.00)	1.14(0.31)	16.00(0.00)	0.00(0.00)	1.17(0.27)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.90(0.28)	16.00(0.00)	0.00(0.00)	0.97(0.23)
Band2	200	Logistic	15.82(0.39)	0.07(0.26)	1.93(0.77)	12.06(2.67)	1.54(1.22)	3.97(0.46)	14.26(2.85)	2.60(1.51)	4.23(0.56)
		Constant	15.80(0.40)	0.05(0.22)	1.86(0.75)	12.99(2.30)	1.27(1.05)	3.65(0.47)	14.82(1.96)	1.27(0.83)	3.63(0.56)
		γ-logistic	15.56(1.18)	0.07(0.26)	1.89(0.50)	15.27(1.08)	0.03(0.17)	2.52(0.53)	15.48(0.61)	0.22(0.42)	2.45(0.51)
		Oracle1	16.00(0.00)	0.00(0.00)	1.63(0.43)	16.00(0.00)	0.00(0.00)	1.95(0.46)	16.00(0.00)	0.00(0.00)	1.94(0.45)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.66(0.42)	16.00(0.00)	0.00(0.00)	1.68(0.41)
	400	Logistic	15.83(0.38)	0.04(0.20)	1.21(0.67)	13.75(1.66)	0.45(0.68)	3.03(0.27)	14.37(1.50)	1.58(1.24)	3.18(0.28)
		Constant	15.83(0.38)	0.06(0.24)	1.13(0.63)	14.78(1.58)	0.39(0.49)	2.71(0.32)	14.85(1.44)	1.38(0.95)	2.82(0.30)
		γ-logistic	15.81(0.39)	0.04(0.20)	1.02(0.28)	15.42(0.67)	0.07(0.26)	1.44(0.50)	15.52(0.93)	0.15(0.34)	1.37(0.38)
		Oracle1	16.00(0.00)	0.00(0.00)	0.96(0.25)	16.00(0.00)	0.00(0.00)	1.17(0.32)	16.00(0.00)	0.00(0.00)	1.23(0.30)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	1.02(0.29)	16.00(0.00)	0.00(0.00)	0.94(0.26)

where all response labels are correct $(m_0, m_1) = (0, 0)$ is considered, denoted by (S0).

In Table 1, "logistic," "constant" and "γ-logistic" in "methods" represent penalized conventional logistic regression, penalized constant-mislabel logistic regression and penalized γ-logistic regression, respectively. Table 1 shows performances of these methods under $p = 1000$ and the results under $p = 2000$ are placed in Supplementary Materials. When mislabeling exists, the proposed approach can identify the majority of the true positives with a small number of false positives. It can be also seen that the results about variable selection of penalized γ-logistic regression are expected to be much closer to the true model as the sample size n increases.

To evaluate the prediction performance, we compute the classification accuracy (CA) of each competitor by Monte Carlo from applying the prediction rule $y = I(\boldsymbol{x}^T\hat{\boldsymbol{\beta}} > 0)$ to a clean test dataset $(\boldsymbol{Y}_0, \boldsymbol{X})$ consisting of 1000 observations, in which the predictor variables $\boldsymbol{X}$ are independent and identically distributed with the training data, and the responses $\boldsymbol{Y}_0$ are generated from Bernoulli $\{\pi(\boldsymbol{x};\boldsymbol{\beta}^*)\}$. For training datasets, we set $m_0 = 0.05$ and $m_1 \in \{0.05, 0.10, \ldots, 0.50\}$. Figures 1 and 2 report the classification accuracy of penalized γ-logistic regression and other methods under S1 and S2 with $(n, p) = (400, 1000)$, respectively. Figure 3 shows an example of the solution paths. More results are placed in Supplementary Materials. Observe that penalized robust logistic regression dominates the alternatives. It can be seen that γ-logistic and constant-mislabel logistic perform similarity when $m_0 = m_1$ as expected. As m_1 increases, the CA loss of the proposed γ-logistic is much less than that of constant-mislabel logistic.

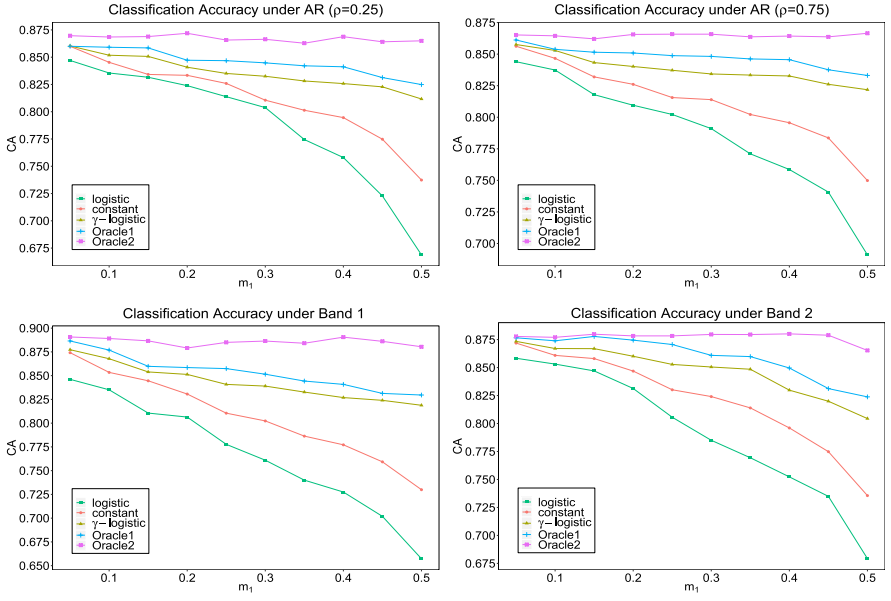

Fig. 1 Simulation results of the classification accuracy under (S1) with $p = 1000$, $n = 400$

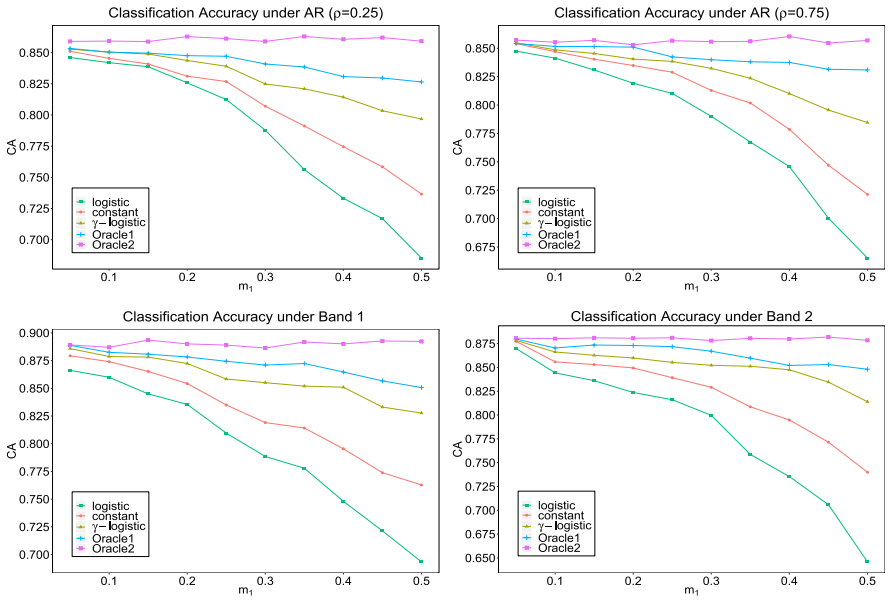

Fig. 2 Simulation results of the classification accuracy under (S2) with $p = 1000$, $n = 400$

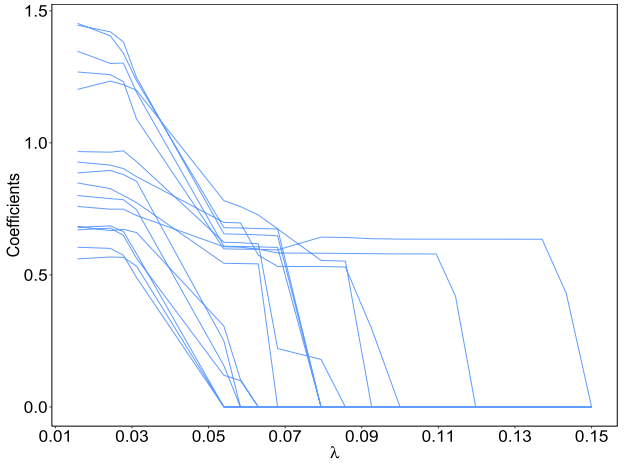

Fig. 3 The path of nonzero effects along λ for logistic regression with $n = 400$, $p = 1000$ under S2 and the first AR correlation

We also consider more realistic predictor variables and use the gene expression data from the TNBC data in the next section. A total of 1000 variables are selected, of which 16 nonzero effects are set. For each replicate, we randomly sample 400 subjects. The scenarios of response variables are same as previous parts. This way, realistic data distributions and correlations can be achieved. Meanwhile,

the abnormality in predictors could be demonstrated (p values of 966 variables are less than 0.05 using Shapiro–Wilk normality test). In this scenario, we evaluate the performances of variable selection and parameter estimation of three methods. These simulation results are summarized in Table 4 ("Appendix"). The robustness for abnormality in predictors of the proposed method could be demonstrated.

Besides, we further consider the detection of the suspect mislabeled samples by searching for instances with small values of the weight function, in which the instances whose weights of less a cutoff value are considered as candidates of mislabeled subjects. Under the first AR correlation, $p = 1000$ and $n = 400$, we examine the relationships between the cutoff value and the performance of detecting mislabeled samples (Table 5, "Appendix"). It can be shown that the true positive rate (TPR) and false positive rate (FPR) are increased with increasing cutoff values of weights, which reflects that there is no consistently good cutoff values. In practice, the cutoff value could be selected according to the preference of TPR or FPR.

3.2 Case 2: Poisson regression for count data

In this case, we numerically compare the proposed penalized γ-Poisson regression with penalized conventional Poisson regression. We evaluate the performances of variable selection and parameter estimation. And detailed implementation algorithms of γ-Poisson regression are relegated to "Appendix B2".

Predictor variables x are from p-dimensional normal distribution $N(0, \Sigma)$, $\Sigma = DRD$, where R is correlation matrix and D is standard deviation diagonal matrix. The structures of correlation matrix $R = (r_{ij})_{1 \leqslant i,j \leqslant p}$ are same as the covariance matrix in Case 1. There are a total of 16 nonzero effects. In order to have a reasonable range for the response variable, we set $D = \text{diag}\{0.5, \cdots, 0.5\}_{p \times p}$ and nonzero coefficients are randomly generated from uniform distribution $U[0.5, 1]$. The response variable $y_i \sim$ Poisson $\left\{ \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}^*\right) + \tau_i \right\}$, where $\tau_i = 0$ or 5. We consider two mechanisms of contamination: (S1) m_1 of τ_is in each dataset are randomly set as 5; (S2) $\mathrm{P}\left(\tau_i = 5\right) = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}^*\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}^*\right)} m_1$. We set $m_1 = 0, 0.2$, in which $m_1 = 0$ represents that responses are not contaminated (denoted by S0).

In Table 2, "Poisson" and "γ-Poisson" in "methods" represent penalized conventional Poisson regression and penalized γ-Poisson regression, respectively. Table 2 shows performances of these methods under $p = 1000$ and more results are placed in Supplementary Materials. The penalized γ-Poisson regression outperforms penalized conventional Poisson regression for identifying more true positives and less false positives under both contaminations. When responses are not contaminated, the results of these two methods are similar.

4 Real data analysis

In this section, the Cancer Genome Atlas (TCGA) data (https://cancergenome.nih.gov/) on Breast Invasive Carcinoma (BRCA) is used. A total of 57,251 variables for a total of 1,222 samples (1,102 with a primary solid tumor, 7 with metastases

Table 2 Mean (SD) of numbers of true/false positives (TP/FP) for variable selection under Poisson regression with $p = 1000$

Correlation	n	Methods	S0			S1			S2		
			TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta} - \beta_0\|\|_2$
AR $\rho = 0.25$	200	Poisson	15.91(0.29)	0.33(0.87)	0.84(0.37)	15.20(1.18)	4.92(3.91)	1.32(0.60)	15.57(0.71)	3.86(1.36)	0.97(0.42)
		γ-Poisson	15.95(0.22)	0.56(1.08)	0.97(0.37)	15.90(0.30)	0.76(1.06)	1.31(0.54)	15.93(0.27)	0.65(0.76)	0.95(0.41)
		Oracle1	16.00(0.00)	0.00(0.00)	0.55(0.24)	16.00(0.00)	0.00(0.00)	0.88(0.16)	16.00(0.00)	0.00(0.00)	0.73(0.16)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.52(0.18)	16.00(0.00)	0.00(0.00)	0.57(0.21)
	400	Poisson	15.99(0.10)	0.05(0.22)	0.57(0.26)	15.91(0.29)	3.22(2.21)	0.75(0.33)	15.93(0.27)	3.23(1.97)	0.67(0.32)
		γ-Poisson	15.98(0.14)	0.02(0.14)	0.62(0.31)	15.98(0.14)	0.07(0.26)	0.70(0.35)	16.00(0.00)	0.05(0.22)	0.65(0.29)
		Oracle1	16.00(0.00)	0.00(0.00)	0.45(0.13)	16.00(0.00)	0.00(0.00)	0.58(0.12)	16.00(0.00)	0.00(0.00)	0.51(0.19)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.43(0.14)	16.00(0.00)	0.00(0.00)	0.49(0.15)
AR $\rho = 0.75$	200	Poisson	15.89(0.36)	0.38(0.89)	1.15(0.50)	14.97(1.23)	6.84(5.84)	1.45(0.56)	15.35(0.83)	5.77(4.72)	1.31(0.46)
		γ-Poisson	15.91(0.29)	0.61(1.13)	1.27(0.54)	15.89(0.31)	0.91(1.18)	1.38(0.25)	15.94(0.24)	0.73(0.94)	1.29(0.21)
		Oracle1	16.00(0.00)	0.00(0.00)	0.76(0.37)	16.00(0.00)	0.00(0.00)	0.84(0.16)	16.00(0.00)	0.00(0.00)	0.82(0.14)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.74(0.35)	16.00(0.00)	0.00(0.00)	0.79(0.38)
	400	Poisson	15.95(0.22)	0.07(0.26)	0.92(0.39)	15.79(0.41)	4.95(3.80)	1.24(0.31)	15.81(0.39)	3.54(2.50)	1.16(0.41)
		γ-Poisson	15.92(0.27)	0.03(0.17)	0.93(0.37)	15.99(0.10)	0.18(0.42)	1.15(0.11)	16.00(0.00)	0.11(0.31)	1.14(0.29)
		Oracle1	16.00(0.00)	0.00(0.00)	0.71(0.11)	16.00(0.00)	0.00(0.00)	0.91(0.20)	16.00(0.00)	0.00(0.00)	0.84(0.18)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.70(0.14)	16.00(0.00)	0.00(0.00)	0.73(0.16)
Band 1	200	Poisson	15.92(0.27)	0.30(0.85)	0.65(0.30)	15.34(1.16)	4.06(5.11)	1.23(0.62)	15.71(0.64)	2.84(1.50)	0.83(0.40)
		γ-Poisson	15.97(0.17)	0.53(1.05)	0.67(0.27)	15.90(0.30)	0.62(0.98)	0.91(0.24)	15.93(0.26)	0.57(0.82)	0.76(0.21)
		Oracle1	16.00(0.00)	0.00(0.00)	0.42(0.17)	16.00(0.00)	0.00(0.00)	0.63(0.16)	16.00(0.00)	0.00(0.00)	0.55(0.17)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.43(0.20)	16.00(0.00)	0.00(0.00)	0.40(0.16)
	400	Poisson	16.00(0.00)	0.04(0.20)	0.43(0.21)	15.95(0.22)	2.66(1.97)	0.73(0.33)	15.98(0.14)	1.70(1.56)	0.51(0.28)
		γ-Poisson	16.00(0.00)	0.02(0.14)	0.44(0.28)	16.00(0.00)	0.04(0.20)	0.62(0.18)	16.00(0.00)	0.05(0.22)	0.49(0.19)
		Oracle1	16.00(0.00)	0.00(0.00)	0.34(0.15)	16.00(0.00)	0.00(0.00)	0.49(0.18)	16.00(0.00)	0.00(0.00)	0.41(0.21)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.36(0.19)	16.00(0.00)	0.00(0.00)	0.32(0.12)

Table 2 (continued)

Correlation	n	Methods	S0			S1			S2		
			TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$	TP	FP	$\|\|\hat{\beta}-\beta_0\|\|_2$
Band 2	200	Poisson	15.94(0.30)	0.32(0.86)	0.78(0.35)	15.25(1.17)	4.25(4.29)	1.29(0.60)	15.62(0.69)	3.51(1.08)	1.12(0.42)
		γ-Poisson	15.96(0.20)	0.55(1.07)	0.88(0.34)	15.90(0.30)	0.78(1.03)	0.96(0.24)	15.93(0.26)	0.63(0.84)	0.90(0.21)
		Oracle1	16.00(0.00)	0.00(0.00)	0.51(0.22)	16.00(0.00)	0.00(0.00)	0.68(0.16)	16.00(0.00)	0.00(0.00)	0.58(0.16)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.50(0.19)	16.00(0.00)	0.00(0.00)	0.47(0.15)
	400	Poisson	15.99(0.10)	0.05(0.22)	0.52(0.24)	15.93(0.26)	3.53(2.98)	0.74(0.33)	15.95(0.22)	2.93(1.77)	0.69(0.30)
		γ-Poisson	15.99(0.10)	0.02(0.14)	0.57(0.30)	15.99(0.10)	0.08(0.27)	0.71(0.29)	16.00(0.00)	0.04(0.20)	0.65(0.19)
		Oracle1	16.00(0.00)	0.00(0.00)	0.41(0.18)	16.00(0.00)	0.00(0.00)	0.58(0.18)	16.00(0.00)	0.00(0.00)	0.41(0.15)
		Oracle2	–	–	–	16.00(0.00)	0.00(0.00)	0.38(0.17)	16.00(0.00)	0.00(0.00)	0.42(0.20)

and 113 with normal breast tissue) are included in the BRCA gene expression data, which can be downloaded using the R package *brca.data*.

4.1 Triple-negative breast cancer data

We focus on the Triple-Negative Breast Cancer (TNBC) built from the BRCA data. TNBC, the most heterogeneous group of breast cancers, presents a significantly shorter survival comparing those with non-triple-negative after the first metastatic event. The TNBC is characterized by lack of expression of three receptors (estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor type 2 (HER2)) (Foulkes et al. 2010). TNBC individuals are with ER, PR and HER2 negative and non-TNBC individuals are with at least one of the three genes positive.

However, Hammond et al. (2010) reported that up to 20% of immunohistochemical (IHC) ER and PR determinations worldwide might be inaccurate for some reasons, such as the variation in interpretation criteria, preanalytic variables. In addition, distinct HER2 labels can be provided by three available variable sources, namely, the HER2 (IHC) level, HER2 (IHC) status and HER2 (fluorescence in-situ hybridization, FISH) (Lopes et al. 2018; Wolff et al. 2007).

Following Lopes et al. (2018), only 1,102 samples from primary solid tumor and 19,688 variables (including the three TNBC-associated key variables ER, PR and HER2) are considered for analysis, corresponding to the protein coding genes reported from Ensembl genome browser (2000) and Consensus CDS (2017). Log-transformed gene expression data are normalized to have zero mean and unit variance. We conduct a marginal screening and keep the top 3500 genes for downstream analysis to remove noise.

4.2 Gene identification and outlier detection

From variables selected by penalized γ-logistic regression, 26 genes are identified, in which 14 genes are down-regulated and 12 genes are up-regulated in TNBC listed in Table 3. As a contrast, the penalized conventional logistic regression identifies 18 genes (11 down-regulated and 7 up-regulated). Quite a lot genes are reported as strong TNBC-regulated many times previously, in which it is of interest to note that some genes are identified by γ-logistic regression but conventional logistic regression fail, such as down-regulated genes *AGR3*, *TGFB3*, *AR* and *SPDEF* and up-regulated genes *CT83*, *FAM171A1*, *FZD9*, *VGLL1* and *PPP1R14C*. The *AGR3* is considered to be a suitable serum-based biomarker for early cancer detection because of the low expression in TNBC cell lines (Guo et al. 2017). The *TGFB3*, encoding a secreted ligand of the transforming growth factor$-\beta$ protein superfamily, is found to be linked with the carcinogenesis of TNBC and involved in the cell cycle pathway (Naorem et al. 2019). Li (2017) concludes that the *AR* expression, as a subclassification marker, contributes to good prognosis in TNBC and that AR-positive TNBC patients might respond to anti-androgen endocrine therapy. Jung et al. (2016) report that *SPDEF* is expressed at high levels in non-TNBC cell lines. For up-regulated

Table 3 Identified genes in TNBC based on all samples (γ-logistic: highlighted in bold; logistic: highlighted with star) and stability results

Down-regulated			Up-regulated					
Gene	γ-logistic	Logistic	Gene	γ-logistic	Logistic	Gene	γ-logistic	Logistic
AGR2*	1.000	1.000	C16orf95	0.017		PDX1		0.009
AGR3	1.000		**CDCA2***	0.974	0.922	POM121L2*		0.991
AR	1.000		**CHODL**	1.000		SLC15A1		0.353
CA12*	1.000	1.000	**CT83**	1.000		SLC6A15		0.190
CAPN13	0.034		**FAM171A1**	1.000		SUV39H2		0.078
ERBB2	0.026		FAM64A	0.009		TLX1		0.026
FOXA1*	1.000	1.000	**FOXC1***	1.000	1.000	TMEFF1*		0.603
GATA3*	1.000	1.000	**FZD9**	1.000		VAX1		0.293
GUCY1A2	0.034		HORMAD1	0.060		ZIC1		0.578
JAM3	0.060		MOGS	0.009		ZIC4		0.009
MLPH*	1.000	1.000	**PAPSS1**	0.862				
MYCT1	0.017		**PPP1R14C**	1.000				
PGAP3*	0.871	1.000	ROPN1	0.043				
PRR15*	1.000	1.000	**SFT2D2**	1.000				
SPARCL1	0.931		**SRSF12***	1.000	1.000			
SPDEF	1.000		**TTLL4***	1.000	1.000			
TBC1D9*	1.000	0.526	**VGLL1**	1.000	0.328			
TFF3*	1.000	1.000	CLDN10		0.026			
TGFB3	1.000		COL9A3*		0.612			
CPE*		1.000	DMRTA2		0.009			
HTRA1*		0.991	FTCD		0.103			
TTC6		0.017	ILF2		0.009			
VAV3		0.017	LYPD1		0.198			

genes, the *CT83* has been identified as a potential target for triple-negative breast cancer (Jessica et al. 2019). In addition, it has been reported that *FAM171A1* (Sandra et al. 2017), *FZD9* (Tudoran et al. 2015), *VGLL1* (Chen et al. 2019) and *PPP1R14C* (Al-Zahrani et al. 2018) would overexpress in TNBC cell lines.

We apply the "leave-one-out" approach to assess the stability of our method and findings. The proposed method is applied in the case where one sample is removed from the dataset firstly and then this step is repeated over all samples. Genes' frequency of being identified is computed (Table 3). It can be seen that all genes, identified by the proposed method, have stability measures close to 1. We have also examined those genes not identified and found that their stability measures are equal or close to 0, which suggests satisfactory stability of our method. For comparison, the penalized logistic regression is applied in the same process, but it does not work well in stability.

We detect the suspect individuals by searching for instances with small values of the weight function. 57 instances whose weights of less 0.5 are considered as candidates of mislabeled subjects and a careful inspection might contribute to disclosing the outlierness of suspect individuals detected (Table B3 in "Appendix"). In these suspect individuals, abnormal ER or PR gene expression values regarding their TNBC labels can be observed for some individuals, such as "TCGA-GM-A2DI" (ER-, 23.49; PR-, 12.05) and "TCGA-BH-A1EW" (ER-, 29.98; PR-, 18.9). The opposite situation can also be observed for non-TNBC patients "TCGA-AR-A1AH" (ER+, 0.03), "TCGA-AR-A0TP" (ER+, 0.04) and "TCGA-A2-A4S1" (ER+, 0.29), which might be identified as TNBC if labeled correctly. Besides, the inconsistency of HER2 labels is observed. For instance, individual "TCGA-C8-A3M7" (HER2-, 25.47) is identified as HER2- by IHC testing, while its HER2 value most probably indicates positive for the gene expression.

5 Discussion

We have proposed a robust high-dimensional regression method for generalized linear model with anomalous responses. Mislabeled responses, as an important special case, have been the focus in this paper, but continuous responses could also be dealt with in the proposed framework. We have rigorously established variable selection consistency and estimation bounds in the setting where the dimensionality p can grow exponentially fast with the sample size n. The reasonable performance of the proposed method has been shown in simulations and real data analysis. Although this work focuses on data with anomalous responses, abnormality in the predictors could also be analyzed using the proposed framework. And relevant simulations in "Appendix" show that even analyzing data with normal responses, the proposed method is close to the conventional methods.

This study can be potentially extended in multiple directions. Imbalanced data is common in classification problems. How to deal with imbalanced data with mislabeled responses could also be an interesting problem. In technologies of imbalanced data, data-level and algorithm-level methods are two main approaches. For data-level methods, Stefanowski (2016) proposes algorithms consists of cleaning overlapping instances and removing noisy objects that might affect models negatively. These ideas combining with γ-divergence may offer a solution. Specifically, when removing overlapping or noisy samples, weight function from γ-divergence could be used to improve detection. As for algorithm-level methods, the cost-sensitive approach is the most popular branch (Thai-Nghe et al. 2010), in which the loss function incorporates varying penalties for considered groups of examples and the importance of less represented objects is improved with the assignment of higher cost. And the weight function derived from γ-divergence could be added to these varying penalties according to a proper way, which might reduce the impact of mislabeled responses. Another interesting research is robust high-dimensional statistical inference (e.g., constructing confidence intervals or statistical testing) based on the penalized γ-divergence. For regularized estimators, inference procedures based

on asymptotic properties perform poorly, especially when the signal-to-noise ratio is high and the between covariate correlations are not low (Minnier et al. 2011). Recently, many powerful techniques have been proposed, see Dezeure et al. (2015) for an overview. However, little work exists on the robust high-dimensional statistical inference. This raises many interesting theoretical and methodological questions for the future.

Acknowledgements We thank the Associate Editor and two referees for very constructive comments and suggestions which have improved the presentation of the paper. This work was supported by the National Natural Science Foundation of China (11971404), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (19YJC910010), the University of Chinese Academy of Sciences (Y95401TXX2), Beijing Natural Science Foundation (Z190004) and Key Program of Joint Funds of the National Natural Science Foundation of China (U19B2040).

Appendix

Appendix A. *Proofs of Theorem 1*

In this section, we give rigorous proofs of Theorem 1. To prove Theorem 1, some notations are needed. Define $\mu_i = E(y_i) = b'(\theta_i)$, which is linked to θ_i through the canonical link function $q(\mu_i) = \theta_i$. Let $\nabla_{\boldsymbol{\beta}}$ represent the derivative to $\boldsymbol{\beta}$. Then,

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} I_i(\boldsymbol{\beta}) \boldsymbol{x}_i, \tag{6}$$

where

$$\begin{aligned} &I_i(\boldsymbol{\beta}) = -\gamma \omega_{\gamma,i}(\boldsymbol{\beta}) \left[K_i(y_i;\boldsymbol{\beta}) - \frac{\Psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})} \right], \quad \omega_{\gamma,i}(\boldsymbol{\beta}) = \frac{f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma} \mathrm{d}y \right)^{\frac{\gamma}{1+\gamma}}}, \\ &K_i(y_i;\boldsymbol{\beta}) = \frac{y_i - \mu_i}{v(\mu_i)}, \quad v(\mu_i) = \mathrm{Var}\left(y_i\right) q'\left(\mu_i\right), \\ &\Psi_{1i}(\boldsymbol{\beta}) = \int K_i(y;\boldsymbol{\beta}) f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma} \mathrm{d}y, \quad \Psi_{2i}(\boldsymbol{\beta}) = \int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma} \mathrm{d}y. \end{aligned} \tag{7}$$

Denote $\nabla_{\boldsymbol{\beta}} I_i(\boldsymbol{\beta}) = H_i(\boldsymbol{\beta}) \boldsymbol{x}_i$ and $\nabla_{\boldsymbol{\beta}} H_i(\boldsymbol{\beta}) = J_i(\boldsymbol{\beta}) \boldsymbol{x}_i$. $H_i(\boldsymbol{\beta})$ and $J_i(\boldsymbol{\beta})$ need to be calculated in Theorem 1. Some calculations show that

$$\begin{aligned} &\nabla_{\boldsymbol{\beta}} \omega_{\gamma,i}(\boldsymbol{\beta}) = -I_i(\boldsymbol{\beta}) \boldsymbol{x}_i, \quad \nabla_{\boldsymbol{\beta}} K_i(y_i;\boldsymbol{\beta}) = k_{1i}(y_i;\boldsymbol{\beta}) \boldsymbol{x}_i, \\ &\nabla_{\boldsymbol{\beta}} \Psi_{1i}(\boldsymbol{\beta}) = \psi_{1i}(\boldsymbol{\beta}) \boldsymbol{x}_i, \quad \nabla_{\boldsymbol{\beta}} \Psi_{2i}(\boldsymbol{\beta}) = (1+\gamma) \Psi_{1i}(\boldsymbol{\beta}) \boldsymbol{x}_i, \end{aligned}$$

where

$$
\begin{aligned}
k_{1i}(y_i;\boldsymbol{\beta}) &= \zeta_{1i}(y_i;\boldsymbol{\beta})[q^{-1}(\theta_i)]', \\
\zeta_{1i}(y_i;\boldsymbol{\beta}) &= \frac{-v(\mu_i) - (y_i - \mu_i)v'(\mu_i)}{v^2(\mu_i)}, \quad \theta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i, \\
\psi_{1i}(\boldsymbol{\beta}) &= \int (k_{1i}(y;\boldsymbol{\beta}) + (1+\gamma)K_i^2(y;\boldsymbol{\beta}))f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y.
\end{aligned} \tag{8}
$$

Then we can get $\nabla_{\boldsymbol{\beta}} I_i(\boldsymbol{\beta}) = H_i(\boldsymbol{\beta})\boldsymbol{x}_i$, where

$$
\begin{aligned}
H_i(\boldsymbol{\beta}) &= -\gamma\omega_{\gamma,i}(\boldsymbol{\beta})\left[\gamma\Omega_{1i}^2(\boldsymbol{\beta}) + \Omega_{2i}(\boldsymbol{\beta})\right], \\
\Omega_{1i}(\boldsymbol{\beta}) &= K_i(y_i;\boldsymbol{\beta}) - \frac{\Psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})}, \\
\Omega_{2i}(\boldsymbol{\beta}) &= k_{1i}(y_i;\boldsymbol{\beta}) - \frac{\psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})} + (1+\gamma)\left[\frac{\Psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})}\right]^2.
\end{aligned} \tag{9}
$$

More calculations show that

$$
\nabla_{\boldsymbol{\beta}} k_{1i}(y_i;\boldsymbol{\beta}) = k_{2i}(y_i;\boldsymbol{\beta})\boldsymbol{x}_i, \quad \nabla_{\boldsymbol{\beta}}\psi_{1i}(\boldsymbol{\beta}) = \psi_{2i}(\boldsymbol{\beta})\boldsymbol{x}_i,
$$

where

$$
\begin{aligned}
k_{2i}(y_i;\boldsymbol{\beta}) &= \zeta_{1i}(y_i;\boldsymbol{\beta})[q^{-1}(\theta_i)]'' + \zeta_{2i}(y_i;\boldsymbol{\beta})\{[q^{-1}(\theta_i)]'\}^2, \\
\zeta_{2i}(y_i;\boldsymbol{\beta}) &= \frac{(y_i - \mu_i)\{2v(\mu_i)[v'(\mu_i)]^2 - v^2(\mu_i)v''(\mu_i)\} + 2v^2(\mu_i)v'(\mu_i)}{v^4(\mu_i)}, \\
\psi_{2i}(\boldsymbol{\beta}) &= \int [k_{2i}(y;\boldsymbol{\beta}) + 3(1+\gamma)K_i(y;\boldsymbol{\beta})k_{1i}(y;\boldsymbol{\beta}) \\
&\quad + (1+\gamma)^2K_i^3(y;\boldsymbol{\beta})]f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y.
\end{aligned} \tag{10}
$$

Then we can get $\nabla_{\boldsymbol{\beta}} H_i(\boldsymbol{\beta}) = J_i(\boldsymbol{\beta})\boldsymbol{x}_i$, where

$$
\begin{aligned}
J_i(\boldsymbol{\beta}) &= \gamma I_i(\boldsymbol{\beta})[\gamma\Omega_{1i}^2(\boldsymbol{\beta}) + \Omega_{2i}(\boldsymbol{\beta})] - 2\gamma^2\omega_{\gamma,i}(\boldsymbol{\beta})\Omega_{1i}(\boldsymbol{\beta})\Omega_{2i}(\boldsymbol{\beta}) \\
&\quad - \gamma\omega_{\gamma,i}(\boldsymbol{\beta})\{k_{2i}(y_i;\boldsymbol{\beta}) - \frac{\psi_{2i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})} - (1+\gamma)\frac{\psi_{1i}(\boldsymbol{\beta})\Psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}^2(\boldsymbol{\beta})} \\
&\quad + 2(1+\gamma)^2[\frac{\Psi_{1i}(\boldsymbol{\beta})}{\Psi_{2i}(\boldsymbol{\beta})}]^3\},
\end{aligned} \tag{11}
$$

$\Omega_{1i}(\boldsymbol{\beta})$ and $\Omega_{2i}(\boldsymbol{\beta})$ are defined in (9), $k_{2i}(y_i;\boldsymbol{\beta})$ and $\psi_{2i}(\boldsymbol{\beta})$ are defined in (10), $\psi_{1i}(\boldsymbol{\beta})$ is defined in (8), $\Psi_{1i}(\boldsymbol{\beta})$ and $\Psi_{2i}(\boldsymbol{\beta})$ are defined in (7). And then we give the proof of Theorem 1.

Proof of Theorem 1 First, let $\Theta = \{\boldsymbol{\beta}_1 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 = r\}$ and $r = \varepsilon\sqrt{|S|/n}$ with $\varepsilon < \sqrt{n/|S|}$. It suffices to show that

$$
\Pr\{\inf_{\boldsymbol{\beta}_1\in\Theta} \ell_1(\boldsymbol{\beta}_1) > \ell_1(\boldsymbol{\beta}_1^*)\} \geqslant 1 - \tau(r),
$$

where $\tau(r) = \exp\left(-\frac{n\rho_1^2 r^2}{512\rho_0 M^2}\right) + 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2 M^2 \kappa^2}\right) + 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right)$, $\boldsymbol{\beta}_1 \in \mathbf{R}^{|S|}$ and

$$\ell_1(\boldsymbol{\beta}_1) = -\frac{1}{n}\sum_{i=1}^{n} \frac{f\left(y_i|\boldsymbol{x}_{i1};\boldsymbol{\beta}_1\right)^{\gamma}}{\left(\int f(y|\boldsymbol{x}_{i1};\boldsymbol{\beta}_1)^{1+\gamma}\mathrm{d}y\right)^{\gamma/(1+\gamma)}}. \tag{12}$$

This implies that with probability at least $1-\tau$, $\ell_1(\boldsymbol{\beta}_1)$ has a local minimum $\hat{\boldsymbol{\beta}}_1 \in \Theta$.

Let $\boldsymbol{u} \in \mathbf{R}^{|S|}$ with $\|\boldsymbol{u}\|_2 = 1$. Define $\check{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1^* + r\boldsymbol{u}$ and $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$. Consider $Q(\boldsymbol{u}) = n\{\ell_1(\check{\boldsymbol{\beta}}_1) - \ell_1(\boldsymbol{\beta}_1^*)\}$. It is equivalent to show that

$$\Pr\left(\inf_{\|\boldsymbol{u}\|_2=1} Q(\boldsymbol{u}) > 0\right) \geqslant 1 - \tau(r). \tag{13}$$

In fact

$$\begin{aligned} Q(\boldsymbol{u}) &= r\boldsymbol{u}^T\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*) + \frac{1}{2}r^2\boldsymbol{u}^T\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1\boldsymbol{u} + \frac{1}{2}r^2\boldsymbol{u}^T\boldsymbol{X}_1^T[H(\tilde{\boldsymbol{\beta}}) - H(\boldsymbol{\beta}^*)]\boldsymbol{X}_1\boldsymbol{u} \\ &=: Q_1 + Q_2 + Q_3, \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^*$ and $\check{\boldsymbol{\beta}}$, and $I(\boldsymbol{\beta}) = (I_1(\boldsymbol{\beta}), \ldots, I_n(\boldsymbol{\beta}))^T$ with I_i is defined in (7). Each element of $I(\boldsymbol{\beta}^*)$ and $H(\boldsymbol{\beta}^*)$ are bounded by a sufficient large bounded constant M.

For Q_1, the following inequality can be obtained for any $t > 0$ using $\|\boldsymbol{X}_1\boldsymbol{u}\|_2^2 \leqslant n\rho_0$ and Hoeffding's inequality,

$$\Pr(|Q_1| \geqslant rt) \leqslant 2\exp\left(-\frac{2t^2}{\|\boldsymbol{X}_1\boldsymbol{u}\|_2^2 M^2}\right) \leqslant 2\exp\left(-\frac{2t^2}{n\rho_0 M^2}\right).$$

Let $t = \frac{1}{32}n\rho_1 r$, then

$$\Pr(Q_1 \geqslant -\frac{1}{32}n\rho_1 r^2) \geqslant 1 - \exp\left(-\frac{n\rho_1^2 r^2}{512\rho_0 M^2}\right). \tag{14}$$

For Q_2, note that $Q_2 \geqslant \frac{1}{2}r^2\lambda_{\min}\{\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1\}$. From Bonferroni's inequality and Hoeffding's inequality, we have

$$\begin{aligned} &\Pr\left(\|\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E[\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1]\|_F^2 \geqslant \frac{n^2\rho_1^2}{4}\right) \\ &\leqslant \sum_{j\in S}\sum_{k\in S} \Pr\left(\left|\sum_{i=1}^{n} H_i(\boldsymbol{\beta}^*)x_{ij}x_{ik} - \sum_{i=1}^{n} E[H_i(\boldsymbol{\beta}^*)x_{ij}x_{ik}]\right| \geqslant \frac{n\rho_1}{2|S|}\right) \\ &\leqslant 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2 M^2\kappa^2}\right), \end{aligned}$$

where x_{ij} is the (i,j) component of $\boldsymbol{X}^T$. By the inequality

$$\lambda_{\min}\{\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1\} \geqslant n\rho_1 - \|\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E[\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1]\|_F,$$

we have

$$\Pr(Q_2 \geqslant \frac{1}{4}n\rho_1 r^2) \geqslant 1 - 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2M^2\kappa^2}\right). \tag{15}$$

As for Q_3, we have $|H_i(\tilde{\boldsymbol{\beta}}) - H_i(\boldsymbol{\beta}^*)| = |J_i(\boldsymbol{\beta}^\circ)\boldsymbol{x}_{i1}^T(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)| \leqslant M\kappa r$, where J_i is defined in (11). $\boldsymbol{\beta}^\circ$ is between $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}_1$ denotes the components of $\tilde{\boldsymbol{\beta}}$ indexed by S. $J_i(\boldsymbol{\beta})$ is bounded by M. Note that $|Q_3| \leqslant \frac{1}{2}r^3 M\kappa\sigma_{\max}\{\boldsymbol{X}_1^T\boldsymbol{X}_1\}$. By the inequality

$$\sigma_{\max}\{\boldsymbol{X}_1^T\boldsymbol{X}_1\} \leqslant n\rho_0 + \|\boldsymbol{X}_1^T\boldsymbol{X}_1 - E[\boldsymbol{X}_1^T\boldsymbol{X}_1]\|_F,$$

similarly as the derivation on Q_2, it can be derived that

$$\Pr(Q_3 \geqslant -\frac{3}{4}nr^3\rho_0 M\kappa) \geqslant 1 - 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right).$$

With $r \ll \frac{\rho_1}{\rho_0}$, it can be obtained that

$$\Pr(Q_3 \geqslant -\frac{3}{32}n\rho_1 r^2) \geqslant 1 - 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right). \tag{16}$$

From (14)–(16), we have

$$Q(u) \geqslant \frac{1}{8}n\rho_1 r^2, \tag{17}$$

with probability at least

$$1 - \exp\left(-\frac{n\rho_1^2 r^2}{512\rho_0 M^2}\right) - 2|S|^2 \exp\left(-\frac{n\rho_1^2}{2|S|^2M^2\kappa^2}\right) - 2|S|^2 \exp\left(-\frac{n\rho_0^2}{2|S|^2\kappa^2}\right).$$

Define $\xi(r) = \frac{1}{8}n\rho_1 r^2$. Recall that $r = \varepsilon\sqrt{|S|/n}$, $\xi(r) \geqslant 0$. Therefore, (13) is proved and Part 1 of Theorem 1 is established.

Now Part 2 is considered. Let $\hat{\boldsymbol{\beta}}$ denote $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ for simplicity. By the Karush–Kuhn–Tucker (KKT) conditions, we need to prove that $\hat{\boldsymbol{\beta}}$ satisfies

$$\boldsymbol{X}_1^T I(\hat{\boldsymbol{\beta}}) = n\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) \tag{18}$$

and

$$\|\boldsymbol{X}_2^T I(\hat{\boldsymbol{\beta}})\|_\infty \leqslant n\lambda, \tag{19}$$

where $I(\boldsymbol{\beta}) = (I_1(\boldsymbol{\beta}), \ldots, I_n(\boldsymbol{\beta}))^T$ with I_i defined in (7) and $\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) = (p'_\lambda(|\hat{\beta}_{11}|)\,\text{sgn}\,(\hat{\beta}_{11}), \ldots, p'_\lambda(|\hat{\beta}_{1s}|)\,\text{sgn}\,(\hat{\beta}_{1s}))^T$ with $s = |S|$. If

$a\lambda < \min_{j\in S}|\hat{\beta}_{1j}|$, $\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) = \mathbf{0}$, and (18) holds certainly. Note $2a\lambda < \min_{j\in S}\left|\boldsymbol{\beta}^*_{1j}\right|$ and $r^2 \ll \lambda$, it can be concluded that the event $\{\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1\|_2 \leqslant r\}$ belongs to the event $\{a\lambda < \min_{j\in S}|\hat{\beta}_{1j}|\}$. That is,

$$\Pr(a\lambda < \min_{j\in S}|\hat{\beta}_{1j}|) \geqslant \Pr(\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1\|_2 \leqslant r) \geqslant 1 - \tau(r). \tag{20}$$

Now consider the probability of

$$\|\boldsymbol{X}_2^T I(\hat{\boldsymbol{\beta}})\|_\infty \leqslant n\lambda.$$

With Taylor expansion, we have

$$n\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) = \boldsymbol{X}_1^T I(\boldsymbol{\beta}^*) - \boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1) - \boldsymbol{z}_1, \tag{21}$$

where $\boldsymbol{z}_1 = \boldsymbol{X}_1^T(H(\bar{\boldsymbol{\beta}}_{t_1}) - H(\boldsymbol{\beta}^*))\boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1)$, $\bar{\boldsymbol{\beta}}_{t_1}$ is between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$. Then, we can have

$$\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1 = (\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*) - n\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) - \boldsymbol{z}_1). \tag{22}$$

Therefore, $\boldsymbol{X}_2^T I(\hat{\boldsymbol{\beta}})$ can be rewritten as

$$\begin{aligned}\boldsymbol{X}_2^T I(\hat{\boldsymbol{\beta}}) &= \boldsymbol{X}_2^T I(\boldsymbol{\beta}^*) - \boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1) - \boldsymbol{z}_2 \\ &= \boldsymbol{X}_2^T I(\boldsymbol{\beta}^*) - \boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*) - n\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|) \\ &\quad - \boldsymbol{z}_1) - \boldsymbol{z}_2,\end{aligned}$$

where $\boldsymbol{z}_2 = \boldsymbol{X}_2^T(H(\bar{\boldsymbol{\beta}}_{t_2}) - H(\boldsymbol{\beta}^*))\boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1)$, $\bar{\boldsymbol{\beta}}_{t_2}$ lies between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$. From (20) we know that $\|\boldsymbol{p}'_\lambda(|\hat{\boldsymbol{\beta}}_1|)\|_\infty = 0$ with probability at least $1 - \tau(r)$. Then, the KKT Condition (19) is guaranteed if

$$\lambda \geqslant [\|\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1}\|_\infty + 1](\|\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*)/n\|_\infty + \|\boldsymbol{z}\|_\infty), \tag{23}$$

where $\boldsymbol{z} = \frac{1}{n}X^T[H(\boldsymbol{\beta}^* + \eta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)) - H(\boldsymbol{\beta}^*)]\boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*_1)$ with $\eta \in [0, 1]$.

Next, we define

$$\begin{aligned}q_0 &= \|\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1} - E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty, \\ q_1 &= \|(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1} - [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty, \\ q_2 &= \|\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)\|_\infty, \\ q_3 &= \|\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)\|_\infty, \\ \varphi &= \|[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty.\end{aligned}$$

We can find that

$$
\begin{aligned}
q_0 &= \|[\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)] \cdot \left((\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1} - [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\right) \\
&\quad + E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1} \cdot [-\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 + E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)] \\
&\quad \cdot [\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1]^{-1} \\
&\quad + [\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1 - E(\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)] \cdot [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty \\
&\leqslant q_2 q_1 + C q_3 \|[\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1]^{-1}\|_\infty + q_2 \varphi \\
&\leqslant q_1 q_2 + C q_3 (\varphi + q_1) + q_2 \varphi.
\end{aligned}
$$

Note that

$$
\begin{aligned}
q_1 &= \|(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1} \cdot (E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1) - \boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1) \cdot [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty \\
&\leqslant \|(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1}\|_\infty \cdot \|E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1) - \boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1\|_\infty \\
&\quad \cdot \|[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty \\
&\leqslant \|(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1} - [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1} + [E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_\infty \cdot q_3 \varphi \\
&\leqslant (q_1 + \varphi) q_3 \varphi.
\end{aligned}
$$

Hence, as long as $q_3\varphi < 1$, then we have $q_1 \leqslant q_3\varphi^2/(1 - q_3\varphi)$ and $q_0 \leqslant \varphi(q_2 + Cq_3)/(1 - q_3\varphi)$. Now the events $\{q_2 \leqslant C/(2\varphi)\}$ and $\{q_3 \leqslant 1/(4\varphi)\}$ are considered. Similar to the proof for Q_2, and note that $\varphi \leqslant \sqrt{|S|} \cdot \|[E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)]^{-1}\|_2 \leqslant \sqrt{|S|} \cdot \sigma_{\min}^{-1}\{E(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)\} \leqslant \sqrt{|S|}/(n\rho_1)$, it can be obtained

$$
\begin{aligned}
\Pr(q_2 \leqslant \frac{C}{2\varphi}) &\geqslant 1 - 2|S|^2 \exp\left(-\frac{nC^2}{2\varphi^2|S|^2}\right) \geqslant 1 - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2 C^2}{2|S|^3}\right), \\
\Pr(q_3 \leqslant \frac{1}{4\varphi}) &\geqslant 1 - 2|S|^2 \exp\left(-\frac{n}{8\varphi^2|S|^2}\right) \geqslant 1 - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2}{8|S|^3}\right).
\end{aligned}
$$

Then we have $q_0 \leqslant \varphi(q_2 + Cq_3)/(1 - q_3\varphi) \leqslant C$ with probability at least

$$
1 - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2 C^2}{2|S|^3}\right) - 2|S|^2 \exp\left(-\frac{n^3\rho_1^2}{8|S|^3}\right). \tag{24}
$$

With $\|\boldsymbol{X}_2^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1(\boldsymbol{X}_1^T H(\boldsymbol{\beta}^*)\boldsymbol{X}_1)^{-1}\|_\infty \leqslant q_0 + C$ and above discussions, go back to (23), it is sufficient to show

$$
\lambda \geqslant (2C + 1)(\|\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*)/n\|_\infty + \|\boldsymbol{z}\|_\infty).
$$

Therefore, focus on the events that

$$
\begin{aligned}
\|\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*)/n\|_\infty &\leqslant \frac{\lambda}{2(2C + 1)}, \\
\|\boldsymbol{z}\|_\infty &\leqslant \frac{\lambda}{2(2C + 1)}.
\end{aligned}
$$

According to the Bonferroni's inequality and Hoeffding's inequality, it can be obtained

$$\begin{aligned} &\Pr\left(\|\boldsymbol{X}_1^T I(\boldsymbol{\beta}^*)\|_\infty \leqslant \frac{\lambda}{2(2C+1)} \right) \\ &\quad \geqslant 1 - 2p \cdot \exp\left(-\frac{n\lambda^2}{2M^2\kappa^2(2C+1)^2} \right). \end{aligned} \tag{25}$$

As for $\boldsymbol{z} = (z_1, \ldots, z_p)^T$, we have

$$z_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \eta J_i(\boldsymbol{\beta}^* + \tilde{\eta}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)) \boldsymbol{x}_{i1}^T (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \boldsymbol{x}_{i1}^T (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)/n,$$

where x_{ij} is the (i, j) component of $\boldsymbol{X}^T$, $\tilde{\eta} \in [0, \eta]$, and $J_i(\boldsymbol{\beta})$ is defined in (11). By Cauchy–Swarchz inequality, it can be obtained

$$|z_j| \leqslant M\kappa\sigma_{\max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{i1} \boldsymbol{x}_{i1}^T \right\} \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2^2.$$

Recall that $\Pr(\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \leqslant r) \geqslant 1 - \tau(r)$. Note that $r^2 \ll \lambda$, it can be known that $M\kappa\rho_0 r^2 \leqslant \frac{\lambda}{2(2C+1)}$. Following the proof of part 1, we have

$$\Pr\left(\|\boldsymbol{z}\|_\infty \leqslant \frac{\lambda}{2(2C+1)} \right) \geqslant 1 - \tau(r). \tag{26}$$

Part 2 is proved by combining (20) and (24)–(26). □

Appendix B. Additional numeric results and implementation algorithm of γ -Poisson regression

Appendix B1. Additional numeric results

See Tables 4, 5 and 6.

Table 4 Mean (SD) of numbers of true/false positives (TP/FP) for variable selection under logistic regression based on TNBC real data

	Methods	TP	FP	$\|\hat{\beta} - \beta_0\|_2$
S0	Logistic	13.66(1.17)	2.91(1.46)	3.87(0.98)
	Constant	13.71(1.23)	2.83(1.40)	3.64(0.92)
	γ-logistic	14.89(1.55)	1.12(1.90)	2.57(0.61)
	Oracle1	16.00(0.00)	0.00(0.00)	1.89(0.49)
	Oracle2	–	–	–
S1	Logistic	10.79(2.93)	5.83(2.29)	6.42(1.33)
	Constant	11.18(2.76)	5.49(2.14)	5.71(1.39)
	γ-logistic	12.75(1.88)	3.11(1.53)	4.92(1.21)
	Oracle1	16.00(0.00)	0.00(0.00)	2.64(0.85)
	Oracle2	16.00(0.00)	0.00(0.00)	1.94(0.51)
S2	Logistic	11.36(2.67)	5.20(2.30)	6.37(1.96)
	Constant	11.99(2.30)	4.94(2.19)	5.65(1.87)
	γ-logistic	13.57(1.08)	3.04(1.23)	3.92(1.33)
	Oracle1	16.00(0.00)	0.00(0.00)	2.25(0.76)
	Oracle2	16.00(0.00)	0.00(0.00)	1.86(0.42)

Table 5 Mean (SD) of TPR and FPR for the detection of mislabeled samples under logistic regression

Cutoff values	S1		S2	
	TPR	FPR	TPR	FPR
0.05	0.6129(0.0691)	0.0259(0.0093)	0.6966(0.0726)	0.0272(0.0099)
0.10	0.6616(0.0695)	0.0345(0.0109)	0.7350(0.0698)	0.0355(0.0101)
0.15	0.6876(0.0723)	0.0408(0.0119)	0.7574(0.0655)	0.0414(0.0110)
0.20	0.7018(0.0714)	0.0449(0.0119)	0.7715(0.0663)	0.0462(0.0113)
0.25	0.7112(0.0712)	0.0494(0.0130)	0.7820(0.0653)	0.0498(0.0119)
0.30	0.7213(0.0721)	0.0528(0.0136)	0.7910(0.0628)	0.0535(0.0125)
0.35	0.7301(0.0723)	0.0564(0.0137)	0.7991(0.0615)	0.0569(0.0127)
0.40	0.7360(0.0730)	0.0599(0.0138)	0.8069(0.0588)	0.0608(0.0131)
0.45	0.7456(0.0717)	0.0641(0.0137)	0.8166(0.0562)	0.0654(0.0138)
0.50	0.7577(0.0682)	0.0681(0.0141)	0.8259(0.0544)	0.0689(0.0148)
0.55	0.7679(0.0688)	0.0731(0.0146)	0.8359(0.0510)	0.0743(0.0150)
0.60	0.7796(0.0663)	0.0786(0.0152)	0.8452(0.0500)	0.0797(0.0152)
0.65	0.7909(0.0660)	0.0848(0.0164)	0.8542(0.0504)	0.0856(0.0160)
0.70	0.8028(0.0636)	0.0937(0.0171)	0.8642(0.0490)	0.0933(0.0168)
0.75	0.8168(0.0591)	0.1046(0.0170)	0.8748(0.0461)	0.1026(0.0161)

Table 6 Summary of some outliers detected as suspect individuals in TNBC data

Individual	ER	PR	HER2	HER2 level (IHC)	HER2 status (IHC)	HER2 (FISH)	Type
TCGA-E9-A22G	0.44(-)	0.02(-)	15.32		+		NO TNBC
TCGA-A7-A13E	0.82(+)	0.06(-)	46.08	2+	Equivocal	–	NO TNBC
TCGA-A2-A04U	0.02(-)	0.02(-)	9.64	1+	–	+	NO TNBC
TCGA-BH-A5IZ	5.12(+)	0.03(-)	28.08		–	–	NO TNBC
TCGA-AR-A251	1.57(+)	0.10(-)	14.02	2+	Equivocal	–	NO TNBC
TCGA-AR-A1AH	0.03(+)	0.03(-)	34.12		–		NO TNBC
TCGA-AN-A0FJ	0.08(+)	0.04(-)	14.28	1+	+		NO TNBC
TCGA-AR-A0TP	0.04(+)	0.03(-)	13.39		–		NO TNBC
TCGA-A2-A0YJ	0.09(+)	0.03(-)	240.24	0	–		NO TNBC
TCGA-OL-A5S0	0.09(+)	0.06(-)	31.92			+	NO TNBC
TCGA-E2-A1II	0.14(-)	0.19(+)	10.73	1+	–		NO TNBC
TCGA-LL-A5YP	0.16(+)	0.05(-)	15.10	1+	–	+	NO TNBC
TCGA-A7-A13D	0.52(-)	0.81(+)	42.28	2+	Equivocal	–	NO TNBC
TCGA-AR-A1AJ	1.47(+)	0.07(-)	9.74		–		NO TNBC
TCGA-D8-A1JM	5.00(+)	0.01(-)	21.85	1+	–		NO TNBC
TCGA-B6-A0IJ	1.18(+)	0.46(+)	11.12				NO TNBC
TCGA-A2-A1G1	0.53(-)	0.17(-)	819.76	2+	Equivocal	+	NO TNBC
TCGA-AO-A0JL	0.63(-)	0.08(-)	63.60	1+	–	+	NO TNBC
TCGA-AC-A62X	0.19(+)	0.02(-)	28.53				NO TNBC
TCGA-D8-A1XW	0.32(-)	0.11(+)	21.03	1+	–		NO TNBC
TCGA-LL-A6FR	0.33(-)	0.04(+)	32.13	2+	Equivocal	+	NO TNBC
TCGA-S3-AA0Z	16.67(+)	0.07(+)	33.07	1+	Equivocal	–	NO TNBC
TCGA-AN-A0FX	1.13(-)	0.64(-)	24.02	1+	+		NO TNBC

Table 6 (continued)

Individual	ER	PR	HER2	HER2 level (IHC)	HER2 status (IHC)	HER2 (FISH)	Type
TCGA-E9-A1NC	0.11(-)	0.07(+)	15.91		+		NO TNBC
TCGA-LL-A8F5	1.08(+)	0.04(-)	11.86	1+	–		NO TNBC
TCGA-AR-A24Q	1.00(+)	0.36(-)	20.67		–		NO TNBC
TCGA-A2-A3Y0	2.18(+)	0.03(-)	11.34	1+	–		NO TNBC
TCGA-E2-A14Y	0.67(+)	0.03(+)	487.90	2+	Equivocal	+	NO TNBC
TCGA-A1-A0SB	3.16(+)	0.03(-)	32.35		–		NO TNBC
TCGA-E9-A1ND	1.44(-)	0.05(-)	13.05		+		NO TNBC
TCGA-AN-A0FL	0.09(-)	1.07(-)	15.07	1+	+		NO TNBC
TCGA-A2-A25F	0.62(-)	0.23(+)	5.19		–		NO TNBC
TCGA-BH-A0DL	6.99(+)	0.04(-)	9.92		–		NO TNBC
TCGA-A2-A4RX	0.68(+)	0.93(+)	26.64	1+	–		NO TNBC
TCGA-A2-A4S1	0.29(+)	0.01(-)	0.61		–		NO TNBC
TCGA-BH-A6R9	0.59(-)	0.25(+)	8.18		–		NO TNBC
TCGA-E2-A1B0	0.14(-)	0.26(-)	563.81	3+	+		NO TNBC
TCGA-AO-A1KO	10.78(+)	9.12(+)	14.91	1+	–		NO TNBC
TCGA-D8-A13Y	15.48(+)	4.17(+)	4.83	1+	–		NO TNBC
TCGA-D8-A1JK	0.40(-)	0.72(+)	22.19	1+	–		NO TNBC
TCGA-AR-A24U	1.22(-)	0.26(-)	410.17	3+	+		NO TNBC
TCGA-AR-A1AO	1.47(+)	1.13(-)	14.89	1+	–		NO TNBC
TCGA-C8-A12P	0.15(-)	0.20(-)	259.71	3+	+		NO TNBC
TCGA-BH-A1EW	29.98(-)	18.9(-)	42.47		–		TNBC
TCGA-GM-A2DI	23.49(-)	12.05(-)	20.30			–	TNBC
TCGA-C8-A3M7	4.27(-)	0.76(-)	25.47		–		TNBC

Table 6 (continued)

Individual	ER	PR	HER2	HER2 level (IHC)	HER2 status (IHC)	HER2 (FISH)	Type
TCGA-JL-A3YW	0.35(+)	0.09(+)	31.47	1+	+		NO TNBC
TCGA-LL-A73Z	7.19(+)	2.10(+)	28.34	2+	Equivocal	–	NO TNBC
TCGA-OL-A5RY	0.99(+)	0.38(-)	658.80			+	NO TNBC
TCGA-D8-A1XT	0.30(-)	0.13(-)	692.72	3+	+		NO TNBC
TCGA-BH-A1FN	14.34(+)	3.30(+)	10.64				NO TNBC
TCGA-AC-A2FK	4.44(+)	18.2(+)	58.01			–	NO TNBC
TCGA-AN-A0AM	73.18(+)	0.09(-)	12.23		–		NO TNBC
TCGA-BH-A209	27.39(+)	7.08(+)	10.00				NO TNBC
TCGA-UU-A93S	0.30(-)	0.12(-)	1668.35	3+	+		NO TNBC
TCGA-E2-A108	12.43(+)	2.96(+)	14.32	1+	–		NO TNBC
TCGA-AO-A0JC	13.67(+)	1.31(+)	5.20	0	–		NO TNBC

Appendix B2. Detailed implementation algorithm of γ-logistic and γ-Poisson regression

In this section, the concrete expressions of the gradient (5) for logistic regression and Poisson regression are given. Recall that the gradient (5) is

$$\nabla_j \ell(\boldsymbol{\beta}) = -\frac{\gamma}{n}\sum_{i=1}^{n}\omega_{\gamma,i}(\boldsymbol{\beta})\left[K_i(y_i;\boldsymbol{\beta}) - \frac{\int K_i(y;\boldsymbol{\beta}) f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y}{\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y}\right]X_{ij}$$

where $\omega_{\gamma,i}(\boldsymbol{\beta}) = \frac{f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y\right)^{\frac{\gamma}{1+\gamma}}}$, $K_i(y_i;\boldsymbol{\beta}) = \frac{y_i-\mu_i}{\mathrm{Var}\,(y_i)q'(\mu_i)}$ and $\mu_i = E(y_i)$ is linked to θ_i through the canonical link function $q(\cdot)$ in GLM.

γ-Logistic Regression. In logistic regression, $f(y|\boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{\exp\{y\boldsymbol{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}$, and $K_i(y;\boldsymbol{\beta}) = y - \frac{\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}$. Then, it can be obtained that:

$$\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y = \frac{1+\exp\{(1+\gamma)\boldsymbol{x}_i^T\boldsymbol{\beta}\}}{(1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\})^{(1+\gamma)}},$$
$$\int K_i(y;\boldsymbol{\beta})f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y = \frac{\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}(\exp\{\gamma\boldsymbol{x}_i^T\boldsymbol{\beta}\}-1)}{(1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\})^{(2+\gamma)}}, \tag{27}$$
$$\omega_{\gamma,i}(\boldsymbol{\beta}) = \left(\frac{\exp\{y_i(1+\gamma)\boldsymbol{x_i}^T\boldsymbol{\beta}\}}{1+\exp\{(1+\gamma)\boldsymbol{x_i}^T\boldsymbol{\beta}\}}\right)^{\frac{\gamma}{\gamma+1}}.$$

Some calculations show that the gradient (5) for γ-logistic regression can be given as follows.

$$\nabla_j\ell(\boldsymbol{\beta}) = -\frac{\gamma}{n}\sum_{i=1}^{n}\omega_{\gamma,i}(\boldsymbol{\beta})\big[y_i - \pi(\boldsymbol{x_i};(1+\gamma)\boldsymbol{\beta})\big]X_{ij}, \tag{28}$$

where $\omega_{\gamma,i}(\boldsymbol{\beta})$ is defined in (27), $\pi(\boldsymbol{x_i};(1+\gamma)\boldsymbol{\beta}) = \frac{\exp\{(1+\gamma)\boldsymbol{x_i}^T\boldsymbol{\beta}\}}{1+\exp\{(1+\gamma)\boldsymbol{x_i}^T\boldsymbol{\beta}\}}$.

γ-Poisson Regression. In this subsection, we consider detailed implementation algorithms of γ-Poisson regression, in which the calculation of two terms $\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y$ and $\int K_i(y;\boldsymbol{\beta})f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y$ in the loss function (2) and the derivative to β_j (5) are different from the γ-logistic regression.

In Poisson regression, $f(y|\boldsymbol{x}_i;\boldsymbol{\beta}) = \frac{\mu_i^y}{y!}\exp\left(-\mu_i\right)$, and $K_i(y;\boldsymbol{\beta}) = y-\mu_i$, where $\mu_i = \exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$. Then, it can be obtained that:

$$\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y = \sum_{y=0}^{+\infty}\left(\frac{\mu_i^y}{y!}\exp\left(-\mu_i\right)\right)^{1+\gamma},$$
$$\int K_i(y;\boldsymbol{\beta})f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y = \sum_{y=0}^{+\infty}\left(y-\mu_i\right)\left(\frac{\mu_i^y}{y!}\exp\left(-\mu_i\right)\right)^{1+\gamma}.$$

The algebraic expression of infinite sum could not be obtained, so we calculate the infinite sum using numerical approximation with finite sum:

$$\int f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y \approx \sum_{y=0}^{B}\left(\frac{\mu_i^y}{y!}\exp\left(-\mu_i\right)\right)^{1+\gamma},$$
$$\int K_i(y;\boldsymbol{\beta})f(y|\boldsymbol{x}_i;\boldsymbol{\beta})^{1+\gamma}\mathrm{d}y \approx \sum_{y=0}^{B}\left(y-\mu_i\right)\left(\frac{\mu_i^y}{y!}\exp\left(-\mu_i\right)\right)^{1+\gamma}.$$

In the simulation, we set $B = 1000$. For evaluating the performance of numerical approximation, we consider the ratio of the remaining items to main items defined by

$$\tau_{\text{approx1}} = \frac{\sum_{y=B+1}^{1000B} \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}}{\sum_{y=0}^{B} \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}},$$

$$\tau_{\text{approx2}} = \frac{\sum_{y=B+1}^{1000B} \left(y-\mu_i\right) \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}}{\sum_{y=0}^{B} \left(y-\mu_i\right) \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}}.$$

The values of $(\tau_{\text{approx1}}, \tau_{\text{approx2}})$ are shown in Table S4 (Supplementary Materials), in which it can be shown that the values are less than 10^{-8} under all settings. It reflects that this numerical approximation works in γ-Poisson regression.

Therefore, the gradient (5) for γ-Poisson regression can be given as follows.

$$\nabla_j \ell(\boldsymbol{\beta}) = -\frac{\gamma}{n} \sum_{i=1}^{n} \omega_{\gamma,i}(\boldsymbol{\beta}) \left[y_i - \mu_i - \frac{\varsigma_i}{\iota_i} \right] X_{ij}, \tag{29}$$

where

$$\begin{aligned} \omega_{\gamma,i}(\boldsymbol{\beta}) &= \frac{\mu_i^{\gamma y_i} \cdot \exp\left(-\gamma \mu_i\right)}{(y_i!)^{\gamma} \cdot \iota_i^{\gamma/(1+\gamma)}} \\ &= \exp\left\{ \gamma [y_i \log(\mu_i) - \mu_i - \log(y_i!) - \log(\iota_i)/(1+\gamma)] \right\}, \\ \iota_i &= \sum_{y=0}^{B} \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}, \quad \text{and} \quad \varsigma_i = \sum_{y=0}^{B} \left(y-\mu_i\right) \left(\frac{\mu_i^y}{y!} \exp\left(-\mu_i\right) \right)^{1+\gamma}. \end{aligned}$$

References

Aggarwal, C., Yu, P. (2001). Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on management of data* (pp. 37–46).

Al-Zahrani, K., Cook, D., Vanderhyden, B., Sabourin, L. (2018). Assessing the efficacy of androgen receptor and Sox10 as independent markers of the triple-negative breast cancer subtype by transcriptome profiling. *Oncotarget*, *9*(70), 3348–3359.

Cameron, A., Trivedi, P. (2013). *Regression analysis of count data*. Cambridge: Cambridge University Press.

Chen, B., Tang, H., Chen, X., Zhang, G., Wang, Y., Xie, X., Liao, N. (2019). Transcriptomic analyses identify key differentially expressed genes and clinical outcomes between triple-negative and non-triple-negative breast cancer. *Cancer Management and Research*, *11*, 179–190.

Copas, J. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B*, *50*(2), 225–265.

Copeland, K., Checkoway, H., McMichael, A., Holbrook, R. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*, *105*(5), 488–495.

Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and r-software hdi. *Statistical Science*, *30*(4), 533–558.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 54–67.

Foulkes, W., Smith, I., Reis-Filho, J. (2010). Triple-negative breast cancer. *New England Journal of Medicine*, *363*(20), 1938–1948.

Frénay, B., Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869.
Fujisawa, H., Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, *99*(9), 2053–2081.
Ghosh, A., Basu, A. (2016). Robust estimation in generalized linear models: The density power divergence approach. *Test*, *25*(2), 269–290.
Grace, Y. (2017). *Statistical analysis with measurement error or misclassification strategy, method and application*. New York: Springer.
Guo, J., Gong, G., Zhang, B. (2017). Screening and identification of potential biomarkers in triple-negative breast cancer by integrated analysis. *Oncology Reports*, *38*(4), 2219–2228.
Hammond, M., Elizabeth, H., Hayes, D., Dowsett, M., Allred, D., Hagerty, K., Badve, S., Fitzgibbons, P., Francis, G., Goldstein, N., Hayes, M. (2010). American society of clinical oncology/college of American pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Archives of Pathology and Laboratory Medicine*, *131*(1), 18.
Hayashi, K. (2012). A boosting method with asymmetric mislabeling probabilities which depend on covariates. *Computational Statistics*, *27*(2), 203–218.
Hung, H., Jou, Z., Huang, S. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, *74*(1), 145–154.
Jessica, K., Nicolas, W., Titus, J., Niels, G. (2019). Large-scale in-silico identification of a tumor-specific antigen pool for targeted immunotherapy in triple-negative breast cancer. *Oncotarget*, *10*(26), 2515–2529.
Jung, H., Lee, S., Kim, J., Ahn, J., Park, Y., Im, Y. (2016). Statins affect ETS1-overexpressing triple-negative breast cancer cells by restoring DUSP4 deficiency. *Scientific Reports*, *6*, 33–35.
Kawashima, T., Fujisawa, H. (2017). Robust and sparse regression via γ-divergence. *Entropy*, *19*(11), 608.
Komori, O., Eguchi, S., Ikeda, S., Okamura, H., Ichinokawa, M., Nakayama, S. (2016). An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution*, *7*(2), 249–260.
Li, Z. (2017). Expression and clinical significance of androgen receptor in triple negative breast cancer. *Cancers*, *9*(1), 585–590.
Lopes, M., Veríssimo, A., Carrasquinha, E., Casimiro, S., Beerenwinkel, N., Vinga, S. (2018). Ensemble outlier detection and gene selection in triple-negative breast cancer data. *BMC Bioinformatics*, *19*(1), 168.
Minnier, J., Tian, L., Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, *106*(496), 1371–1382.
Naorem, L., Muthaiyan, M., Venkatesan, A. (2019). Integrated network analysis and machine learning approach for the identification of key genes of triple negative breast cancer. *Journal of Cellular Biochemistry*, *120*(4), 6154–6167.
Piepel, G. (2005). Robust regression and outlier detection. *Technometrics*, *31*(2), 260–261.
Rebbapragada, U., Brodley, C. (2007). Class noise mitigation through instance weighting. In *European conference on machine learning* (pp. 260–261). New York: Springer.
Roberts, G., Stramer, O. (2001). A comparison of related density based minimum divergence estimators. *Biometrika*, *88*(3), 865–873.
Sandra, K., Cardona-Huerta, S., Yadira, X., Trevino, V., Hernandez-Cabrera, F., Rojas-Martinez, A., Uscanga-Perales, G., Jorge, L., Martinez-Jacobo, L., Padilla-Rivas, G. (2017). A new gene expression signature for triple-negative breast cancer using frozen fresh tissue before neoadjuvant chemotherapy. *Molecular Medicine*, *23*(1), 101–111.
She, Y., Owen, A. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, *106*(494), 626–639.
Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data. In *Challenges in computational statistics and data mining* (pp. 333–363). New York: Springer.
Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In *The 2010 international joint conference on neural networks* (pp. 1–8).
The Consensus CDS (CCDS) Project. (2000). https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi. Accessed May 2017.
The Ensembl Genome Browser. (2000). http://www.ensembl.org/index.html. Accessed May 2017.
Tudoran, O., Soritau, O., Balacescu, L., Visan, S., Barbos, O., Cojocneanu-Petric, R., Balacescu, O., Berindan-Neagoe, I. (2015). Regulation of stem cells-related signaling pathways in response to doxorubicin treatment in Hs578T triple-negative breast cancer cells. *Molecular and Cellular Biochemistry*, *409*(1), 163–176.

Wang, H., Li, R., Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568.

Wolff, A., Hammond, M., Schwartz, J., Hagerty, K., Allred, D., Cote, R., Dowsett, M., Fitzgibbons, P., Hanna, W., Langer, A. (2007). Guideline summary: American society of clinical oncology/college of American pathologists guideline recommendations for human epidermal growth factor receptor HER2 Testing in Breast Cancer. *Journal of Oncology Practice*, *3*(1), 48–50.

Wu, Y., Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. *Annual Review of Statistics and its Application*, *7*, 209–226.

Zang, Y., Zhao, Q., Zhang, Q., Li, Y., Zhang, S., Ma, S. (2017). Inferring gene regulatory relationships with a high-dimensional robust approach. *Genetic Epidemiology*, *41*(5), 437–454.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*(2), 894–942.

Zimek, A., Schubert, E., Kriegel, H. (2012). A survey on unsupervised outlier detection in high dimensional numerical data. *Statistical Analysis and Data Mining*, *5*(5), 363–387.