Received: 24 May 2020 | Revised: 1 June 2021 | Accepted: 23 August 2021

DOI: 10.1002/bimj.202000157

Biometrical Journal

RESEARCH ARTICLE

Gene–environment interaction identification via penalized robust divergence

Mingyang Ren[1,2] | Sanguo Zhang[1,2] | Shuangge Ma[3] | Qingzhao Zhang[4]

[1] School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China

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

[3] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[4] Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, Fujian Key Lab of Statistics, Xiamen University, Fujian, P. R. China

Correspondence
Qingzhao Zhang, Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, Fujian Key Lab of Statistics, Xiamen University, Fujian, 361005, P. R. China.
Email: qzzhang@xmu.edu.cn

Funding information
Natural Science Foundation of Beijing Municipality, Grant/Award Number: Z190004; Humanity and Social Science Youth Foundation of Ministry of Education of China, Grant/Award Number: 19YJC910010; National Natural Science Foundation of China, Grant/Award Numbers: 11971404, 12171454, 12026604; NIH, Grant/Award Number: CA204120; Basic Scientific Project 71988101 of National Science Foundation of China, Grant/Award Number: B13028; NISF, Grant/Award Number: 1916251; Key Program of Joint Funds of the National Natural Science Foundation of China, Grant/Award Number: U19B2040; University of Chinese Academy of Sciences Education Foundation, Grant/Award Number: Y95401TXX2

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract

In high-throughput cancer studies, gene–environment interactions associated with outcomes have important implications. Some commonly adopted identification methods do not respect the "main effect, interaction" hierarchical structure. In addition, they can be challenged by data contamination and/or long-tailed distributions, which are not uncommon. In this article, robust methods based on γ-divergence and density power divergence are proposed to accommodate contaminated data/long-tailed distributions. A hierarchical sparse group penalty is adopted for regularized estimation and selection and can identify important gene–environment interactions and respect the "main effect, interaction" hierarchical structure. The proposed methods are implemented using an effective group coordinate descent algorithm. Simulation shows that when contamination occurs, the proposed methods can significantly outperform the existing alternatives with more accurate identification. The proposed approach is applied to the analysis of The Cancer Genome Atlas (TCGA) triple-negative breast cancer data and Gene Environment Association Studies (GENEVA) Type 2 Diabetes data.

KEYWORDS
divergence, gene–environment interaction, hierarchical structure, penalized identification, robustness

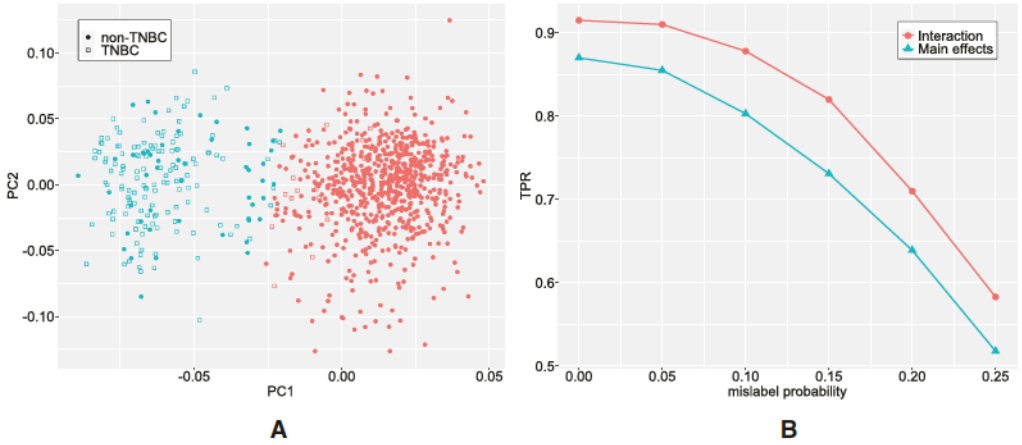

FIGURE 1 (A) The comparison between the labels of TNBC data and cluster results based on the first two principal components of gene variables, in which two colors of points represent different clusters and two shapes represent the labels of samples in this dataset. (B) An example: the TPR for main effects and interactions using conventional logistic regression under different mislabel probabilities

1 | INTRODUCTION

In high-throughput cancer studies, gene–environment interactions (G-E) can have important implications beyond the main effects of genetic (G) and environmental (E) risk factors. The G factors analyzed in the literature include gene expressions, single nucleotide polymorphisms (SNPs), and other omics measurements, which are usually high dimensional. The E factors are usually low-dimensional environmental exposures but may also include clinical measurements. For G-E interaction analysis, the "main effect, interaction" hierarchy, that is, if an interaction is identified as associated with response, the corresponding main effects are automatically identified, has been developed in recent publications and considered as biologically and statistically sensible and necessary (Bien et al., 2013; Wu et al., 2018; Zhu et al., 2014). As simple penalizations are insufficient, several techniques respecting this hierarchy, such as hierarchical sparse group penalizations, have been developed (Liu et al., 2013; Wu & Ma, 2018). For the identification of important interactions, in general, there are two schemes. The first is to conduct marginal analysis, that is, analyzing multiple E factors, one G factor, and their interactions at a time. The other is joint analysis, that is, modeling the joint effects of all E factors, G factors, and interactions in a single model. The two types of analyses have different model assumptions and implications, and both are popular in the current literature (Wu & Ma, 2018).

The extensively adopted likelihood-based and other related approaches can be challenged by data contamination, which is not uncommon. Contamination can have different forms for categorical and continuous variables. Specifically, for categorical responses, there may be samples that belong to different classes present in the data, for example, a tumor instance incorrectly labeled as normal (Shieh & Hung, 2009). In particular, mislabeled responses have been thought to exist in the triple-negative breast cancer (TNBC) data (Lopes et al., 2018) studied in this article. The comparison between the labels of TNBC data and clustering analysis may also suggest this (Figure 1A). As for continuous variables, deviation from normal distribution is common for medical and biological data (Farcomeni & Ventura, 2010), and heavy tails may happen because of biological "anomalies" as well as a mixture of distributions caused by subtypes (Adler et al., 1998; Shen & He, 2015).

Even a single contaminated observation can lead to biased estimation and false marker identification (Figure 1B). To conduct contamination resistant G-E interaction analysis, for example, a robust marginal smoothed penalized rank method has been proposed (Shi et al., 2014). And for joint analysis, a rank-based regression analysis making no stringent distributional assumptions on random error, has been conducted (Wu et al., 2015). A penalized robust approach to dissect G-E interactions based on the least absolute deviation loss for prognosis data has been proposed (Wu et al., 2018). Robust divergence, such as density power divergence (Basu et al., 1998) and γ-divergence (Jones et al., 2001), is an effective technique to deal with contaminated data. In low-dimensional settings, robustness properties of these divergences have been well studied (Fujisawa & Eguchi, 2008; Ghosh & Basu, 2016). Under high-dimensional settings, density power divergence has also been adopted to model the regulatory relationships between gene expression and copy number alterations (Zang et al., 2017) and demonstrates better robustness than alternatives. However, our literature survey suggests that the application of robust divergence to interaction analysis is still very limited.

In this article, we develop a penalized robust divergence approach for identifying G-E interactions in high-dimensional genetic studies, which has the following desirable features. First, robust losses based on density power divergence and γ-divergence are proposed to accommodate contaminated data without assuming contamination distribution and proportion. Our literature search suggests that even for simple, low-dimensional settings, there is no dominating approach (Daszykowski et al., 2007). In addition, the applications of robust analysis techniques to high-dimensional genetic studies remain limited (Farcomeni & Ventura, 2010). As a result, it is of interest to develop and implement new robust analysis approaches to the present setting. Second, a hierarchical sparse group penalization is adopted for regularized estimation and marker selection to achieve the "main effect, interaction" hierarchy, which is not only biologically sensible but may also assist the identification of interactions by main effects and vice versa. Third, we comprehensively conduct both marginal and joint analyses, using similar techniques to achieve the much desired methodological coherence. Fourth, generalized linear models (GLM) are considered to accommodate a variety of responses especially including continuous and binary variables. Overall, this work provides a practical and useful tool for the G-E interaction analysis.

2 | METHODOLOGY

Assume n independent samples $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$, where y_i can be continuous, binary, and so on, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iq})^T$ are q-dimensional clinical/environmental risk factors, and $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{ip})^T$ are p-dimensional genetic markers. Denote $\boldsymbol{w}_{ij}^T = z_{ij}(1, \boldsymbol{x}_i^T)$ and $\boldsymbol{w}_i = (\boldsymbol{w}_{i1}^T, \ldots, \boldsymbol{w}_{ip}^T)^T$. Denote $\boldsymbol{Y}$ as the n-vector composed of y_i's, and $\boldsymbol{X}$, $\boldsymbol{W}_j$, and $\boldsymbol{W}$ as the matrices composed of $\boldsymbol{x}_i'$s, $\boldsymbol{w}_{ij}'$s, and $\boldsymbol{w}_i'$s, respectively.

In marginal analysis, we analyze multiple E factors, one G factor, and their interactions at a time. The covariate effect can be written as $\sum_{k=1}^{q} \theta_k x_{ik} + \zeta_j z_{ij} + \sum_{k=1}^{q} \beta_{kj} x_{ik} z_{ij} = \boldsymbol{\theta}^T \boldsymbol{x_i} + \boldsymbol{b}_j^T \boldsymbol{w}_{ij}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ represents all main E effects, $\boldsymbol{b}_j = (\zeta_j, \boldsymbol{\beta}_j^T)^T$ represents the main effect and interactions corresponding to the jth G variable, and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{qj})^T$. In joint analysis, we model the joint effects of all E factors, G factors, and interactions at a time. The overall covariate effect can be written as $\sum_{k=1}^{q} \theta_k x_{ik} + \sum_{j=1}^{p} (\zeta_j z_{ij} + \sum_{k=1}^{q} \beta_{kj} x_{ik} z_{ij}) = \boldsymbol{\theta}^T \boldsymbol{x_i} + \sum_{j=1}^{p} \boldsymbol{b}_j^T \boldsymbol{w}_{ij}$. Distinctions between the two types of analysis and their individual implications have been well discussed in the literature and will not be reiterated here (Wu & Ma, 2018).

2.1 | Marginal analysis

Consider the jth marginal model for the jth G variable and its G-E interactions. The empirical versions of the density power divergence (Basu et al., 1998) (denoted as DPD for simplicity) and γ-divergence (Jones et al., 2001) loss functions, ignoring terms independent of the unknown parameters, are

$$\ell_{\mathrm{DPD}}(\boldsymbol{\theta}, \boldsymbol{b}_j) = -\frac{1}{n} \sum_{i=1}^{n} \left[\int f(y|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j)^{1+\alpha} \mathrm{d}y - (1 + \frac{1}{\alpha}) f\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j\right)^{\alpha} \right],$$

$$\ell_{\gamma}(\boldsymbol{\theta}, \boldsymbol{b}_j) = -\frac{1}{n} \sum_{i=1}^{n} \frac{f\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j\right)^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j)^{1+\gamma} \mathrm{d}y\right)^{\gamma/(1+\gamma)}},$$

where $f(y_i|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j)$ is the conditional probability density function of y_i given $\boldsymbol{x}_i$ and $\boldsymbol{w}_{ij}$. The parameters $\alpha > 0$ and $\gamma > 0$ balance robustness and efficiency, with a larger value corresponding to more robust but less efficient estimation. To accommodate multiple types of responses, we consider the GLM:

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{w}_{ij}; \boldsymbol{\theta}, \boldsymbol{b}_j) = c(y_i) \exp\left\{ \frac{y_i(\boldsymbol{x}_i^T \boldsymbol{\theta} + \boldsymbol{w}_{ij}^T \boldsymbol{b}_j) - \varphi\left(\boldsymbol{x}_i^T \boldsymbol{\theta} + \boldsymbol{w}_{ij}^T \boldsymbol{b}_j\right)}{\phi} \right\},$$

where $\varphi(\cdot)$ is twice continuously differentiable with $\varphi''(\cdot)$ always positive and ϕ is the dispersion parameter. $\phi = 1$ in a logistic model, and ϕ is usually unknown in a linear model.

Adopting the hierarchical sparse group penalty, we propose the penalized robust objective function for marginal analysis:

$$Q(\boldsymbol{\theta},\boldsymbol{b}_j)=\ell(\boldsymbol{\theta},\boldsymbol{b}_j)+\rho(\|\boldsymbol{b}_j\|;\sqrt{q+1}\lambda,a)+\sum_{k=2}^{q+1}\rho(|b_{kj}|;\lambda,a), \tag{1}$$

where $\ell(\boldsymbol{\theta},\boldsymbol{b}_j)$ is $\ell_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b}_j)$ or $\ell_\gamma(\boldsymbol{\theta},\boldsymbol{b}_j)$, and $\rho(t;\lambda,a)=\lambda\int_0^{|t|}(1-\frac{x}{\lambda a})_+dx$ is the minimax concave penalty (MCP) (Zhang, 2010) with first-order derivative $\rho'(t)=\lambda(1-\frac{t}{a\lambda})_+$ for $a>1$ and $t\geqslant 0$. λ is a data-based tuning parameter, and a is the regularization parameter. The same tuning parameters are used in all p marginal models to ensure that all G variables are analyzed on a fair ground.

Rationale. The divergence-based robust goodness-of-fit measures have been developed under low-dimensional settings. Intuitively, they down-weigh the "influence" of abnormal observations and hence achieve robustness (also see Remark 1). Theoretical and numerical studies under low-dimensional settings show that they can be advantageous over other robust methods under certain scenarios and outperform nonrobust approaches with the presence of data contamination/long-tailed distributions (Basu et al., 1998; Fujisawa & Eguchi, 2008; Ghosh & Basu, 2016; Hung et al., 2018). We note that the data settings considered here are significantly more challenging than the low-dimensional settings. The first term of the penalty determines whether $\boldsymbol{b}_j\equiv 0$, namely the jth G variable has no effect at all. If $\boldsymbol{b}_j\neq 0$, the second term penalizes the interaction terms to determine either the interaction effects are nonzero. The sum of these two terms can identify important G variables as well as important interactions and can respect the "main effect, interaction" hierarchy. Interactions and main G effects corresponding to the nonzero components of b_j's are considered as important. The estimation-based identification strategy, which has been extensively adopted in joint analysis (Kim et al., 2017; Liu et al., 2013) but only a few marginal analysis (Bien et al., 2013), is notably different from the significance-based one. It is also noted that selection of E variables is not conducted, as in many studies, E variables are "preselected" and are all of interest. Penalty can also be revised to conduct E variable selection if needed.

2.2 | Joint analysis

The empirical versions of the DPD and γ-divergence loss functions, respectively, are

$$\ell_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b})=-\frac{1}{n}\sum_{i=1}^{n}\left[\int f(y|\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\theta},\boldsymbol{b})^{1+\alpha}\mathrm{d}y-(1+\frac{1}{\alpha})f(y_i|\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\theta},\boldsymbol{b})^{\alpha}\right],$$

$$\ell_\gamma(\boldsymbol{\theta},\boldsymbol{b})=-\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i|\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\theta},\boldsymbol{b})^{\gamma}}{\left(\int f(y|\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\theta},\boldsymbol{b})^{1+\gamma}\mathrm{d}y\right)^{\gamma/(1+\gamma)}},$$

where $\boldsymbol{b}=(\boldsymbol{b}_1^T,\ldots,\boldsymbol{b}_p^T)^T$, and other notations are similar to those in marginal analysis.

The penalized robust objective function is

$$Q(\boldsymbol{\theta},\boldsymbol{b})=\ell(\boldsymbol{\theta},\boldsymbol{b})+\sum_{j=1}^{p}\rho(\|\boldsymbol{b}_j\|;\sqrt{q+1}\lambda,a)+\sum_{j=1}^{p}\sum_{k=2}^{q+1}\rho(|b_{kj}|;\lambda,a), \tag{2}$$

where $\ell(\boldsymbol{\theta},\boldsymbol{b})$ is $\ell_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b})$ or $\ell_\gamma(\boldsymbol{\theta},\boldsymbol{b})$.

It may seem that the forms of the objective functions in the marginal and joint analyses are very similar, but the two paradigms are significantly different in multiple aspects. The prominent feature of marginal analysis is that, in each individual analysis, the number of variables including interactions and main effects is considerably smaller than the sample size, and it is still highly popular in biomedical studies because of its computational simplicity. On the contrary, joint analysis can better reflect the fact that molecular mechanisms under complex diseases involve multiple G variables and their interactions, while the dimensionality, complexity, and computational cost are much higher than with marginal analysis. Joint analysis has been popular in statistical literature, and its popularity is increasing fast in recent biomedical studies.

To our best knowledge, this work is the first to adopt DPD and γ-divergence in G-E interaction analysis to accommodate contaminated data with multiple types of responses. In addition, in previous studies, the frameworks of marginal and joint analyses have usually been significantly different (incoherent). In contrast, the proposed marginal and joint analyses naturally fall in the same penalized estimation and selection paradigms, which also simplify computational development.

ALGORITHM 1 Group Coordinate Descent Algorithm for Marginal Analysis

Input: Response $\boldsymbol{Y}$, predictors $\boldsymbol{X}$, $\boldsymbol{W}_j$, robust parameters (γ, α), and tuning parameters (a, λ).

Output: Regression coefficients $\boldsymbol{\theta}$ and $\boldsymbol{b}_j$.

Initialization: $m = 0$, $\varepsilon = 10^{-3}$, $\boldsymbol{\theta}^{(0)} = 0$ and $\boldsymbol{b}_j^{(0)} = 0$.

Repeat:

(i) Update $\boldsymbol{\theta}^{(m+1)}$ using gradient descent and Armijo search with $\boldsymbol{b}$ fixed at $\boldsymbol{b}_j^{(m)}$;

(ii) Update $\boldsymbol{b}_j^{(m+1)}$ using gradient descent and Armijo search with $\boldsymbol{\theta}$ fixed at $\boldsymbol{\theta}^{(m+1)}$;

Until: $\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\|_2^2 + \|\boldsymbol{b}_j^{(m+1)} - \boldsymbol{b}_j^{(m)}\|_2^2 \leqslant \varepsilon^2$.

Return: $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{b}_j^{(m+1)}$ at convergence.

ALGORITHM 2 Group Coordinate Descent Algorithm for Joint Analysis

Input: Response $\boldsymbol{Y}$, predictors $\boldsymbol{X}$, $\boldsymbol{W}$, robust parameters (γ, α), and tuning parameters (a, λ).

Output: Regression coefficients $\boldsymbol{\theta}$ and $\boldsymbol{b}$.

Initialization: $m = 0$, $\varepsilon = 10^{-3}$, and $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{b}^{(0)}$ using group LASSO.

Repeat:

(i) Update $\boldsymbol{\theta}^{(m+1)}$ using gradient descent and Armijo search with $\boldsymbol{b}$ fixed at $\boldsymbol{b}^{(m)}$;

(ii) For $j = 1, \ldots, p$;

Update $\boldsymbol{b}_j^{(m+1)}$ using gradient descent and Armijo search with $\boldsymbol{\theta}$, $\boldsymbol{b}_k (k = 1, \ldots, j-1)$ and $\boldsymbol{b}_l (l = j+1, \ldots, p)$ fixed at $\boldsymbol{\theta}^{(m+1)}$, $\boldsymbol{b}_k^{(m+1)}$ and $\boldsymbol{b}_l^{(m)}$, respectively;

Until: $\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\|_2^2 + \|\boldsymbol{b}^{(m+1)} - \boldsymbol{b}^{(m)}\|_2^2 \leqslant \varepsilon^2$.

Return: $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{b}^{(m+1)}$ at convergence.

Remark 1. Robustness of the DPD and γ-divergence methods benefits from down-weighed outliers in estimating equations. It is also noted that the two divergence methods have different behaviors in estimating parameters. Take linear regression under joint analysis as an example. When the hierarchical sparse group penalty is not considered, the two robust methods have the same estimating equation for the regression coefficients $(\boldsymbol{\theta}, \boldsymbol{b})$:

$$(\boldsymbol{X}^T : \boldsymbol{W}^T)^T \tilde{S}(\boldsymbol{\theta}, \boldsymbol{b}) = 0,$$

where

$$\tilde{S}(\boldsymbol{\theta}, \boldsymbol{b}) = \exp\left\{-\frac{\iota\left(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\theta} - \boldsymbol{W}^T\boldsymbol{b}\right)^2}{2\sigma^2}\right\} \odot (\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\theta} - \boldsymbol{W}^T\boldsymbol{b}),$$

$\iota = \gamma$ or α, σ^2 is the variance of the response, $\odot$ is the componentwise product, and the operations are elementwise except for the matrix–vector multiplication of $\boldsymbol{X}^T\boldsymbol{\theta}$ and $\boldsymbol{W}^T\boldsymbol{b}$. It can be seen that weights of instances with large standardized residuals are reduced. And the two divergence methods are connected via the equal weight function when $\gamma = \alpha$. However, the resulted estimating equations for the dispersion parameter are different in the bias correction scheme. Specifically, the DPD method corrects bias by subtracting a bias correction term, while the γ-divergence method uses the expanded parameter $\frac{\sigma^2}{\gamma+1}$ (see (A.6) in the Appendix). As for logistic regression, more discussion can be found in the previous studies (Hung et al., 2018). A consequence is that the parameter estimation and variable selection of the two methods differ. In addition, it is difficult (or impossible) to conclude that one divergence dominates the other under all settings. So when analyzing data in practice, it is sensible to comprehensively consider both divergence.

2.3 | Computation

For marginal analysis, we minimize objective function (1) using the group coordinate descent algorithm summarized in Algorithm 1. For joint analysis, the algorithm minimizing objective function (2) is summarized in Algorithm 2.

MCP contains tuning parameters λ and regularization parameter a. We set $a = 3$ following previous studies (Zhang, 2010). The robust parameter γ and α balance robustness and estimation efficiency. In the literature, there is a lack of consensus on the selection of γ and α. In practice, we search for the optimal values of (λ, γ, or α) jointly using the Bayesian information criterion. In our numerical study, γ and α are selected from the grid $\{0.01, 0.1, 0.3, 0.5, 0.8, 1, 1.5\}$. The proposed algorithm is computationally affordable. For instance, the joint and marginal analysis of one simulated dataset with $n = 500, q = 5, p = 1,000$ under the logistic model takes about 15 and 10 min, respectively, on a regular PC.

Remark 2. The proposed algorithm borrows strength from the existing framework of group coordinate descent, with the difference that the observation weights derived from the robust divergence need to be updated at each iteration (see (A.4) in the Appendix). For both joint and marginal analyses, the key are Step (i) and Step (ii). The details of these two steps under linear regression for continuous data and logistic regression for binary data are provided in the Appendix. Under some distributions, the dispersion parameter ϕ needs to be estimated. It can be achieved by solving the estimating equation with the bisection method, which is described in the Appendix. Convergence is achieved in all of our numerical studies and real data analysis within 50 overall iterations.

3 | SIMULATION STUDIES

We adopt two strategies. First, we simulate data from parametric distributions, which has been done in a large number of published studies. Second, we simulate data based on the TNBC data described in Section 4.1 to closely mimic practical data.

Under the first strategy, we first simulate five normally distributed E risk factors. The correlation between the ith and jth E factors is $\rho^{|i-j|}$ with $\rho = 0.2$. Then two E variables are dichotomized at 0, so there are three continuous and two binary E variables. For G variables, we simulate from 1000-dimensional multivariate normal distribution $N(0, \Sigma)$. Consider two structures of the covariance matrix $\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant p}$. The first structure is autoregressive correlation (AR) given by $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.25$ and 0.75 (denoted as AR1 and AR2, respectively). The second structure is banded correlation. Here two scenarios are considered. The first scenario is given by $\sigma_{ij} = 0.3$ if $|i - j| = 1$, and 0 otherwise. Under the second scenario, $\sigma_{ij} = 0.6$ if $|i - j| = 1$, 0.3 if $|i - j| = 2$, and 0 otherwise. There are a total of 1003 main effects and 5000 interactions. A total of 35 nonzero effects are set, including five main E effects, 10 main G effects, and 20 interactions. The sample size is set as $n = 200$ or 400 under linear regression and $n = 300$ or 500 under logistic regression. We note that it is commonly recognized that logistic regression is "more difficult" than linear regression, thus demanding larger sample size, which can be observed similarly in Fujisawa and Eguchi (2008) and Hung et al. (2018). More importantly, the goal of varying sample sizes is to examine, for a specific model, the dependence on sample size. It is not our goal to compare linear against logistic regression. Under the second strategy, we sample 1000-dimensional gene expression variables and two-dimensional environmental variables from the TNBC data. There are a total of 22 nonzero effects: two main E effects, 10 main G effects, and 10 interactions, and the sample size is set as 500.

The nonzero regression coefficients are all generated from Unif[0.2, 0.8]. All simulations are based on 100 replicates. Linear regression for continuous responses and logistic regression for binary responses are considered.

Under linear regression for continuous responses, three scenarios for the random error distributions are considered: (S0) standard normal distribution, (S1) $0.6N(0, 1) + 0.4\log N(0, 1)$ distribution, and (S2) t-distribution with degree of freedom 5. For comparison, we consider two alternatives: least squares regression (LS) and least absolute deviation regression (LAD) with the sparse group penalty.

Under logistic regression for binary responses, the response $y_i \sim \text{Bernoulli}\{P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}_i)\}$, where $P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}_i) = \eta_{i0}\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}, \boldsymbol{b})\} + \{1 - \eta_{i1}\}\pi(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}, \boldsymbol{b})$, the mislabel probability of the ith sample is $\eta_{i0} = P(y_i = 1|y_{i0} = 0, \boldsymbol{x}_i, \boldsymbol{w}_i)$ or $\eta_{i1} = P(y_i = 0|y_{i0} = 1, \boldsymbol{x}_i, \boldsymbol{w}_i)$, and $\pi(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}, \boldsymbol{b}) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\theta} + \boldsymbol{w}_i^T\boldsymbol{b})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\theta} + \boldsymbol{w}_i^T\boldsymbol{b})}$. We consider three scenarios of mislabel probability: (S0) $\eta_{i0} = \eta_{i1} = 0$; (S1) $\eta_{i0} = m_0$ and $\eta_{i1} = m_1$; (S2) $\eta_{i0} = m_0$ and $\eta_{i1} = m_0 + (m_1 - m_0)\pi(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\theta}, \boldsymbol{b})$. Under (S0), all response labels are correct. (S1) has constant mislabel probabilities. (S2) has mislabel probabilities dependent on $\boldsymbol{x}$, where the mislabel probabilities of samples with higher probabilities of success are higher. We set $m_0 = 0.05$ and $m_1 = 0.2$. For comparison, we also consider two alternatives: logistic regression and robust constant logistic regression (Copas, 1988) with the sparse group penalty, which are referred to "logistic" and "constant," respectively. We acknowledge that for both types of response, there are other alternatives. The aforementioned comparisons can be the most relevant and explicitly demonstrate benefits of the proposed robustness.

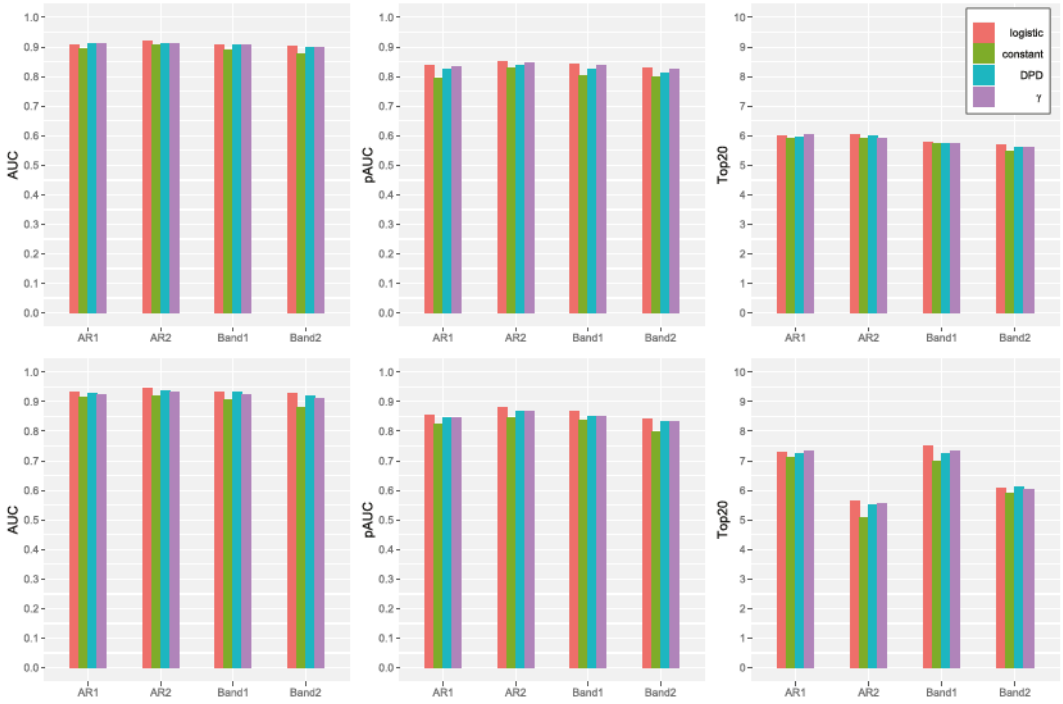

FIGURE 2 Simulation results for the logistic model under S0 and $n = 300$. Three subgraphs at the top correspond to the main effects and the rest correspond to the interactions

3.1 | Marginal analysis

First, a sequence of λ is considered, and identification performance is evaluated at each value. Then the AUC (area under the receiver operating characteristic curve) is computed. Besides, the partial AUC (Walter, 2005) is also considered, where we consider the AUC with 0 and 0.3/0.5 marking the range of the false positive rate (FPR) value (denoted by pAUC1/pAUC2). We also consider Top20 (Top40), defined as the number of true positives when 20 (40) variables (main effects or interactions) are identified.

Representative results are presented in Figures 2–4 (for simplicity, only AUC, pAUC2, Top20 are shown), and additional results are in presented in the Supporting Information. (Figures S1–S3 and Tables S1–S7). When data are not contaminated, LS and logistic regression have the best performance (Tables S3 and S6), as expected. But with contamination, the proposed methods significantly outperform the alternatives. For example, when $n = 500$ for binary response under S2 and Band1 correlation structure (Table S5), the mean AUC values of interaction identification are 0.825 (logistic), 0.833 (constant), 0.937 (DPD), and 0.937 (γ), and the mean (pAUC1, pAUC2) values are (0.708, 0.733) (logistic), (0.726, 0.754) (constant), (0.844, 0.879) (DPD), and (0.840, 0.876) (γ). In addition, the mean (Top20, Top40) values are (6.1, 8.1) (logistic), (6.4, 8.3) (constant), (9.9, 12.2) (DPD), and (9.8, 12.3) (γ).

The performance of the proposed and alternative methods is further examined with selected tunings. Representative results are provided in Table S8 (Supporting Information). The true positive rate (TPR) and FPR values show the competitive performance of the proposed methods with Bayesian information criterion (BIC)-selected tunings. For instance, with interaction effects under linear regression and S2, the (TPR, FPR) values are (0.663, 0.073) for LS, (0.692, 0.017) for LAD, (0.806, 0.014) for DPD, and (0.828, 0.016) for γ-divergence.

3.2 | Joint analysis

The TPR and FPR values for the main and interactions effects at the optimal tuning parameter values are used to evaluate identification accuracy. Representative results are provided in Tables 1–3, and additional results are summarized in the Supporting Information (Tables S9–S12).

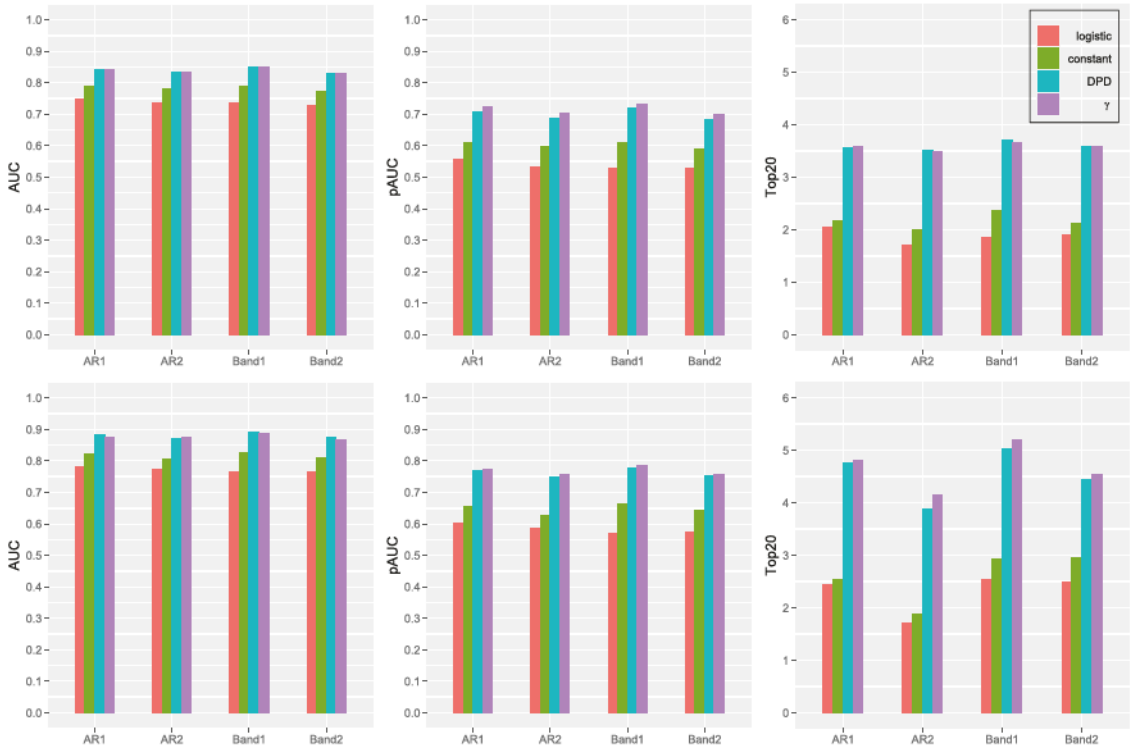

FIGURE 3 Simulation results for the logistic model under S1 and $n = 300$. Three subgraphs at the top correspond to the main effects and the rest correspond to the interactions

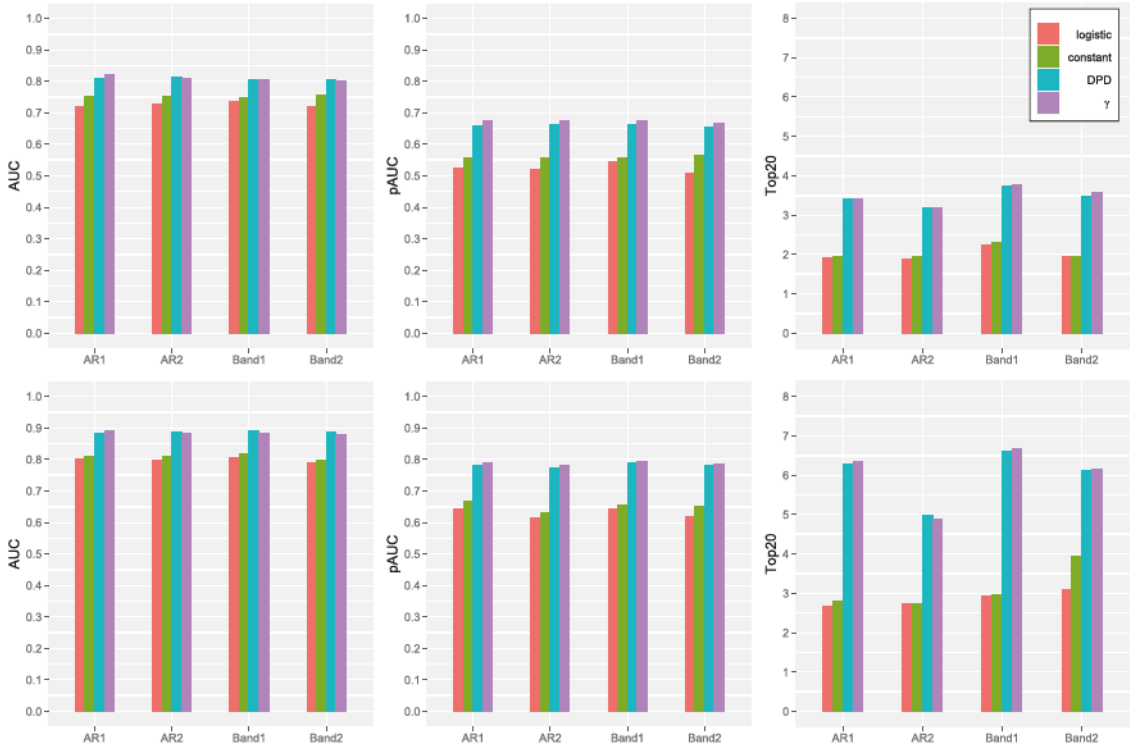

FIGURE 4 Simulation results for the logistic model under S2 and $n = 300$. Three subgraphs at the top correspond to the main effects and the rest correspond to the interactions

TABLE 1 Simulation: Mean (SD) of TPR and FPR for main effects and interactions under logistic regression and S0, joint analysis

			Main effect		Interaction	
Correlation	n	Method	TPR	FPR	TPR	FPR
AR1	300	logistic	0.736 (0.127)	0.068 (0.034)	0.855 (0.132)	0.067 (0.030)
		constant	0.777 (0.112)	0.097 (0.040)	0.843 (0.104)	0.095 (0.036)
		DPD	0.732 (0.120)	0.084 (0.037)	0.798 (0.121)	0.082 (0.033)
		γ	0.733 (0.121)	0.084 (0.034)	0.795 (0.119)	0.083 (0.031)
	500	logistic	0.894 (0.090)	0.078 (0.047)	0.958 (0.078)	0.079 (0.039)
		constant	0.907 (0.074)	0.096 (0.046)	0.961 (0.034)	0.092 (0.043)
		DPD	0.895 (0.083)	0.081 (0.046)	0.956 (0.045)	0.091 (0.041)
		γ	0.887 (0.090)	0.060 (0.061)	0.953 (0.055)	0.072 (0.042)
AR2	300	logistic	0.705 (0.127)	0.060 (0.033)	0.786 (0.154)	0.061 (0.027)
		constant	0.719 (0.120)	0.091 (0.033)	0.797 (0.124)	0.090 (0.030)
		DPD	0.670 (0.128)	0.077 (0.041)	0.744 (0.141)	0.077 (0.039)
		γ	0.677 (0.124)	0.079 (0.032)	0.750 (0.143)	0.077 (0.029)
	500	logistic	0.870 (0.094)	0.067 (0.047)	0.946 (0.091)	0.079 (0.040)
		constant	0.875 (0.086)	0.086 (0.049)	0.941 (0.062)	0.073 (0.045)
		DPD	0.860 (0.088)	0.071 (0.041)	0.930 (0.060)	0.089 (0.042)
		γ	0.867 (0.089)	0.068 (0.037)	0.935 (0.072)	0.083 (0.031)
Band1	300	logistic	0.789 (0.123)	0.068 (0.034)	0.827 (0.148)	0.068 (0.028)
		constant	0.777 (0.103)	0.098 (0.036)	0.854 (0.106)	0.096 (0.035)
		DPD	0.736 (0.108)	0.084 (0.037)	0.810 (0.120)	0.083 (0.034)
		γ	0.744 (0.105)	0.085 (0.030)	0.814 (0.116)	0.084 (0.028)
	500	logistic	0.904 (0.087)	0.088 (0.057)	0.931 (0.076)	0.079 (0.045)
		constant	0.918 (0.067)	0.094 (0.050)	0.964 (0.039)	0.093 (0.044)
		DPD	0.915 (0.07)	0.091 (0.053)	0.960 (0.043)	0.095 (0.050)
		γ	0.913 (0.077)	0.086 (0.050)	0.942 (0.070)	0.072 (0.036)
Band2	300	logistic	0.728 (0.120)	0.066 (0.031)	0.776 (0.129)	0.086 (0.028)
		constant	0.745 (0.108)	0.097 (0.035)	0.813 (0.109)	0.095 (0.033)
		DPD	0.705 (0.110)	0.083 (0.035)	0.776 (0.117)	0.082 (0.034)
		γ	0.707 (0.108)	0.084 (0.031)	0.773 (0.108)	0.083 (0.028)
	500	logistic	0.897 (0.104)	0.074 (0.053)	0.946 (0.093)	0.085 (0.044)
		constant	0.906 (0.075)	0.093 (0.045)	0.959 (0.045)	0.092 (0.042)
		DPD	0.899 (0.080)	0.091 (0.052)	0.956 (0.049)	0.095 (0.052)
		γ	0.889 (0.088)	0.079 (0.057)	0.949 (0.075)	0.089 (0.038)

Simulation suggests that, when errors have a normal distribution for continuous responses or there is no mislabeled data for binary responses (Tables 1 and S9), regular likelihood–based estimation has overall good performance. For examples, when $n = 300$ for binary response under the AR1 correlation structure, the mean TPR values of interaction identification are 0.855 (logistic), 0.843 (constant), 0.798 (DPD), and 0.795 (γ). When $n = 200$ for continuous response under the AR(0.25) correlation structure, the mean TPR values of main effect identification are 0.947 (LS), 0.949 (LAD), 0.913 (DPD), and 0.923 (γ). However, if errors have a long tail for continuous responses or binary responses are mislabeled, the proposed robust methods outperform the alternatives (results are summarized in Tables 2, 3, S10, S11, S12). For instance, when $n = 400$ for the continuous response under S1 and the Band1 correlation structure, the mean TPR values of interaction identification are 0.822(LS), 0.845(LAD), 0.997(DPD), and 0.999(γ). In higher dimensional case, these all methods can also be used, but they become progressively less effective. Therefore, in practice, the prescreening is commonly conducted to reduce dimensionality to improve performances.

TABLE 2 Simulation: Mean (SD) of TPR and FPR for main effects and interactions under logistic regression and S1, joint analysis

Correlation	n	Method	Main effect		Interaction	
			TPR	FPR	TPR	FPR
AR1	300	logistic	0.289 (0.137)	0.007 (0.015)	0.300 (0.146)	0.008 (0.016)
		constant	0.358 (0.134)	0.015 (0.033)	0.370 (0.155)	0.017 (0.033)
		DPD	0.535 (0.140)	0.063 (0.050)	0.526 (0.160)	0.063 (0.046)
		γ	0.532 (0.143)	0.062 (0.049)	0.528 (0.156)	0.063 (0.046)
	500	logistic	0.535 (0.132)	0.009 (0.027)	0.560 (0.142)	0.011 (0.027)
		constant	0.656 (0.165)	0.034 (0.097)	0.668 (0.172)	0.032 (0.102)
		DPD	0.775 (0.127)	0.080 (0.106)	0.784 (0.148)	0.081 (0.099)
		γ	0.777 (0.139)	0.071 (0.068)	0.794 (0.156)	0.083 (0.065)
AR2	300	logistic	0.226 (0.116)	0.007 (0.016)	0.242 (0.138)	0.008 (0.015)
		constant	0.292 (0.128)	0.014 (0.030)	0.307 (0.151)	0.014 (0.031)
		DPD	0.464 (0.151)	0.088 (0.040)	0.458 (0.160)	0.086 (0.039)
		γ	0.451 (0.128)	0.090 (0.035)	0.511 (0.164)	0.087 (0.031)
	500	logistic	0.434 (0.142)	0.009 (0.022)	0.450 (0.155)	0.011 (0.022)
		constant	0.610 (0.136)	0.035 (0.057)	0.580 (0.139)	0.031 (0.053)
		DPD	0.736 (0.126)	0.091 (0.045)	0.748 (0.127)	0.089 (0.043)
		γ	0.728 (0.141)	0.084 (0.031)	0.756 (0.145)	0.092 (0.034)
Band1	300	logistic	0.300 (0.122)	0.007 (0.016)	0.324 (0.146)	0.008 (0.017)
		constant	0.366 (0.147)	0.015 (0.035)	0.390 (0.169)	0.016 (0.035)
		DPD	0.534 (0.159)	0.062 (0.046)	0.539 (0.169)	0.062 (0.042)
		γ	0.531 (0.160)	0.061 (0.044)	0.542 (0.171)	0.062 (0.042)
	500	logistic	0.536 (0.121)	0.008 (0.024)	0.556 (0.137)	0.011 (0.022)
		constant	0.661 (0.158)	0.033 (0.096)	0.668 (0.156)	0.032 (0.099)
		DPD	0.769 (0.132)	0.077 (0.105)	0.770 (0.135)	0.078 (0.102)
		γ	0.762 (0.141)	0.059 (0.063)	0.774 (0.150)	0.071 (0.064)
Band2	300	logistic	0.262 (0.132)	0.007 (0.015)	0.275 (0.153)	0.008 (0.016)
		constant	0.344 (0.133)	0.014 (0.032)	0.360 (0.151)	0.016 (0.033)
		DPD	0.506 (0.153)	0.090 (0.039)	0.498 (0.170)	0.088 (0.036)
		γ	0.492 (0.148)	0.094 (0.033)	0.544 (0.177)	0.090 (0.028)
	500	logistic	0.490 (0.145)	0.008 (0.023)	0.508 (0.151)	0.011 (0.024)
		constant	0.530 (0.145)	0.018 (0.032)	0.521 (0.165)	0.023 (0.042)
		DPD	0.774 (0.121)	0.099 (0.053)	0.786 (0.120)	0.093 (0.051)
		γ	0.769 (0.124)	0.087 (0.035)	0.792 (0.131)	0.097 (0.038)

4 | DATA ANALYSIS

4.1 | Triple-negative breast cancer data

TNBC is the most heterogeneous group of breast cancer, and patients have a significantly shorter survival after the first metastatic event than those with nontriple-negative cancers. The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) data on TNBC contains a total of 1222 samples (1102 with primary solid tumors, seven with metastases, and 113 with normal breast tissues) as well as 57,251 gene expression measurements and seven environmental/clinical variables. The response variable is the TNBC status, which is a binary variable. The data can be downloaded using the R package *brca.data* (https://github.com/averissimo/brca.data/releases/download/1.0/brca.data_1.0.tar.gz).

TNBC is characterized by a lack of expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor type 2 (HER2) (Foulkes et al., 2010). Non-TNBC patients have at least one of them positive. It has been reported that up to 20% of immunohistochemical (IHC) ER and PR determinations may be inaccurate (Hammond et al., 2010). Distinct HER2 labels can be provided by three available sources, which are the HER2 (IHC)

TABLE 3 Simulation: Mean (SD) of TPR and FPR for main effects and interactions under logistic regression and S2, joint analysis

Correlation	n	Method	Main effect		Interaction	
			TPR	FPR	TPR	FPR
AR1	300	logistic	0.250 (0.123)	0.007 (0.016)	0.310 (0.130)	0.008 (0.015)
		constant	0.291 (0.135)	0.018 (0.034)	0.349 (0.142)	0.018 (0.035)
		DPD	0.473 (0.145)	0.096 (0.039)	0.505 (0.140)	0.093 (0.037)
		γ	0.480 (0.148)	0.096 (0.033)	0.510 (0.142)	0.093 (0.030)
	500	logistic	0.489 (0.127)	0.009 (0.022)	0.560 (0.124)	0.011 (0.022)
		constant	0.575 (0.138)	0.033 (0.061)	0.636 (0.132)	0.029 (0.061)
		DPD	0.710 (0.125)	0.079 (0.098)	0.751 (0.121)	0.079 (0.093)
		γ	0.715 (0.119)	0.078 (0.079)	0.749 (0.112)	0.080 (0.078)
AR2	300	logistic	0.200 (0.127)	0.006 (0.014)	0.248 (0.154)	0.007 (0.014)
		constant	0.254 (0.115)	0.015 (0.034)	0.311 (0.147)	0.015 (0.033)
		DPD	0.407 (0.152)	0.089 (0.038)	0.450 (0.151)	0.086 (0.035)
		γ	0.423 (0.154)	0.091 (0.032)	0.458 (0.146)	0.088 (0.027)
	500	logistic	0.372 (0.137)	0.008 (0.023)	0.436 (0.157)	0.010 (0.022)
		constant	0.478 (0.143)	0.091 (0.046)	0.440 (0.107)	0.055 (0.013)
		DPD	0.635 (0.162)	0.085 (0.042)	0.670 (0.159)	0.091 (0.042)
		γ	0.627 (0.163)	0.079 (0.047)	0.664 (0.158)	0.085 (0.037)
Band1	300	logistic	0.239 (0.110)	0.007 (0.017)	0.312 (0.126)	0.008 (0.017)
		constant	0.309 (0.137)	0.018 (0.038)	0.374 (0.155)	0.019 (0.038)
		DPD	0.470 (0.152)	0.096 (0.041)	0.499 (0.153)	0.093 (0.039)
		γ	0.483 (0.149)	0.096 (0.036)	0.506 (0.144)	0.093 (0.031)
	500	logistic	0.484 (0.140)	0.009 (0.021)	0.547 (0.122)	0.011 (0.021)
		constant	0.598 (0.145)	0.034 (0.063)	0.648 (0.133)	0.031 (0.065)
		DPD	0.760 (0.115)	0.090 (0.043)	0.780 (0.111)	0.093 (0.043)
		γ	0.752 (0.124)	0.074 (0.097)	0.784 (0.106)	0.086 (0.097)
Band2	300	logistic	0.235 (0.116)	0.007 (0.016)	0.295 (0.144)	0.008 (0.016)
		constant	0.304 (0.140)	0.016 (0.033)	0.371 (0.154)	0.017 (0.033)
		DPD	0.460 (0.140)	0.093 (0.040)	0.504 (0.146)	0.090 (0.037)
		γ	0.466 (0.143)	0.095 (0.036)	0.510 (0.140)	0.092 (0.032)
	500	logistic	0.424 (0.129)	0.009 (0.026)	0.392 (0.125)	0.011 (0.026)
		constant	0.526 (0.135)	0.098 (0.038)	0.430 (0.095)	0.053 (0.014)
		DPD	0.690 (0.144)	0.092 (0.040)	0.708 (0.133)	0.097 (0.036)
		γ	0.706 (0.163)	0.084 (0.056)	0.696 (0.152)	0.091 (0.042)

level, HER2 (IHC) status, and HER2 (fluorescence in situ hybridization, FISH) (Lopes et al., 2018; Wolff et al., 2007), sometimes leading to conflict and mislabeling.

Following published studies, 1102 samples with primary solid tumors and 19,688 gene expression measurements are analyzed. For environmental/clinical variables, age (normalized) and race (BLACK OR AFRICAN AMERICAN coded as 1, and other races coded as 0) are analyzed (Lopes et al., 2018). When matching the clinical/environmental data with genetic data, complete records are available for 924 samples. Log-transformed gene expression data are normalized to have zero means and unit variances. The prescreening is further conducted, and the top 2000 genes are kept.

4.1.1 | Marginal analysis

The proposed methods based on DPD and γ-divergence identify 35 main gene effects, 59 G-E interactions, and 32 main gene effects, 51 G-E interactions, respectively. Detailed estimation results are provided in Table 4.

Among the genes identified by the proposed methods, some findings have also been reported in previous publications. For instance, it has been reported that the transcription factor BCL11A is overexpressed in TNBC including basal-like breast

TABLE 4 Marginal analysis of TNBC: Identified main effects and interactions

	DPD			γ		
		Interaction			Interaction	
Gene	Main	Age	Race	Main	Age	Race
AGR2	−0.2511	−0.0016	−0.0675	−0.0969	−0.0035	−0.0245
AGR3	−0.2557	−0.0050	−0.0643	−0.0987	−0.0047	−0.0233
AR	−0.2311		−0.0616	−0.0882		−0.0198
B3GNT5	0.2437	−0.0176	0.0568	0.0939	−0.0045	0.0217
BCL11A	0.2348	−0.0136	0.0632	0.0903	−0.0017	0.0230
C5AR2	−0.2230	−0.0114	−0.0563			
CA12	−0.2645	−0.0042	−0.0760	−0.1006	−0.0011	−0.0252
CHODL	0.2678	−0.0898		0.1058	−0.0346	
CLCN4	0.2265	−0.0108	0.0621	0.0871	−0.0020	0.0234
CXXC5	−0.2451		−0.0849	−0.0950		−0.0316
DLI1	−0.2381		−0.0879			
EN1	0.2382	−0.0205	0.0765	0.0917	−0.0063	0.0275
ESR1	−0.2430	−0.0515	−0.0577	−0.0934	−0.0248	−0.0204
FAM171A1	0.2309	−0.0045	0.0620			
FBP1	−0.2292		−0.0712	−0.0881	−0.0022	−0.0257
FOXA1	−0.3175	−0.0220	−0.1258	−0.1161	−0.0018	−0.0382
FOXC1	0.2413	−0.0192	0.0725	0.0925	−0.0046	0.0266
GATA3	−0.2701	−0.0037	−0.0850	−0.1028		−0.0290
HAPLN3	0.2280	−0.0124	0.0612	0.0875	−0.0018	0.0217
HORMAD1	0.2323	−0.0142	0.0715	0.0894	−0.0041	0.0259
MLPH	−0.2781	−0.0042	−0.0904	−0.1046		−0.0300
PPP1R14C	0.2403	−0.0011	0.0656	0.0996	−0.0088	0.0146
PRR15	−0.2628		−0.0860	−0.1018	−0.0020	−0.0319
PSAT1	0.2317	−0.0113	0.0540	0.0892	−0.0019	0.0194
RGMA	0.2254	−0.0028	0.0702			
RHOB	−0.2401	−0.0028	−0.0842	−0.0927		−0.0286
ROPN1	0.2301	−0.0183	0.0650	0.0885	−0.0044	0.0234
SLC44A4	−0.2500		−0.0774	−0.0961		−0.0273
SLC7A8	−0.2265		−0.0562	−0.0863		−0.0174
SPDEF	−0.2657		−0.1047	−0.1030	−0.0013	−0.0393
SRSF12	0.2913	−0.0927		0.1156	−0.0389	
TBC1D9	−0.2515		−0.0764	−0.0964		−0.0267
TFF3	−0.2607	−0.0011	−0.1011	−0.1013		−0.0370
UGT8	0.2381	−0.0185	0.0654	0.0918	−0.0041	0.0239
VGLL1	0.2461	−0.0302	0.0650	0.1183	−0.0381	
SFT2D2				0.0867		0.0258

cancer, whose amplified genomic locus occurs in many basal-like breast cancer tumors, and the overexpression of exogenous BCL11A promotes the formation of tumor, whose knockdown in TNBC cell lines can suppress tumorigenic potential (Khaled et al., 2015). It has been found that the downregulation of EN1 can reduce colony formation, tumorigenicity, and cellular viability of TNBC cell lines significantly. Besides, it has been shown that fructose-1,6-bisphosphatase (FBP1), as the rate-limiting enzyme in gluconeogenesis and a tumor suppressor, regardless of histological type, is upregulated in tumor tissues of TNBC (Li et al., 2016). GATA3 is an effective marker for TNBC diagnostically, and immunohistochemical detection of GATA3 expression contributes to identifying the primary site of metastatic tumors (Krings et al., 2014). It has been reported that HORMAD1 is overexpressed in TNBC, and its expression makes breast cancer cells sensitive to homologous repair–defect targeting agents by resulting in the deficiency of homologous recombination (Wang et al.,

2018). In addition, FOXA1, MLPH, and SLC44A4 have been reported as downregulated in TNBC (He et al., 2015; Lin and Hsu, 2015). FOXC1, FAM171A1, RGMA, PSAT1, and UGT8 have been identified as upregulated in TNBC (Bao et al., 2019; Santuario et al., 2017).

The stability of findings are assessed by applying the "leave-out" approach. Specifically, the proposed method is applied with 1% sample removed from the dataset, and this step is repeated many times. Genes and interactions' frequencies of being identified are computed (Table S13, Supporting Information). It can be seen that almost all genes and interactions identified by the proposed analysis have stability measures close to 1. For comparison, we have also examined those genes not identified and found that their stability measures are equal or close to 0, which suggests satisfactory stability.

Data are also analyzed using the alternatives. The summary comparison results are provided in Table 6, and detailed estimation results using the alternatives are available in the Supporting Information (Table S14). It is observed that the proposed methods and alternatives make different discoveries. More specifically, the two divergence methods generate highly overlapping with each other but moderate overlapping with the alternatives in both main G effects and interaction identification.

4.1.2 | Joint analysis

The proposed methods based on DPD and γ-divergence identify 31 main gene effects, 58 G-E interactions, and 39 main gene effects, 70 G-E interactions, respectively. Detailed estimation results are provided in Table 5.

For the identified genes, relevant findings have also been made in the literature. For example, it has been reported that asporin (ASPN) is highly upregulated in invasive ductal carcinoma, possibly associated with invasion, and related to the epithelial mesenchymal transition (Castellana et al., 2012). CENPW, playing crucial roles in the formation of a functional kinetochore involved in cell division during mitosis, is suppressed in the [ER-,PR-,HER2+] subgroup but elevated in the [ER-,PR-,HER2-] subgroup (Li et al., 2016). The COL9A3 signature has been constructed for efficient and sensitive prognosis prediction of TNBC patients, and COL9A3 has been reported as potentially contributing to the pathogenesis of mammary tumors (Lv et al., 2019). CPA4 is differentially expressed in Flag-TBC1D3-cells, and oncogene TBC1D3 promotes the migration of breast cancer cells (Wang et al., 2017). Besides, gene expression signatures of CXXC5 have been used for breast cancer diagnosis and prognostic testing (Bydoun et al., 2014). DCLRE1C encodes proteins that are part of the TP53 and β-estradiol centered network and operate in the DNA double-strand break repair pathway, the defect of which has been strongly associated with breast cancer predispositions (Tervasmäki et al., 2014). ESR1 mutations are a rare event in treatment-naive patients but common in ER+ metastatic breast cancer patients. The incidence of ESR1 mutations in pretreated ER+ metastatic breast cancer patients is approximately 12%. Thus, advanced breast cancer harboring ESR1 mutations can affect a large number of patients (Niu et al., 2015). ITGB5 encodes a secreted ligand of the transforming growth factor-β and shows decreased expression in TNBC cells (Niu et al., 2015). LYPD1 is a transmembrane protein involved in ligand-dependent signal transduction and plays critical roles in cancer progression (Burnett et al., 2015). In addition, CCL13, DPF3, HAPLN3, and PCDHB9 have been previously reported as upregulated in TNBC (Coyle et al., 2018; Santuario et al., 2017). AADAT and PGAP3 have significant associations with breast cancer risk (Waddell et al., 2010).

The stability of findings are assessed using the same approach as for marginal analysis, which shows satisfactory performance of the proposed method (Table S15, Supporting Information). In addition, we consider detecting suspicious individuals by searching for instances with small values of the weight function, and instances whose weights are less 0.5 are considered as candidates of mislabeled subjects. After removing these suspicious individuals, we reanalyze data using the γ-logistic or DPD method, and the identified main gene effects and interactions are summarized in Table S16 (Supporting Information). The new results are notably different. For instance, 33 main gene effects are identified by the DPD method, among which 18 are included in the previous results (Table 5). And among the 37 genes identified by γ-logistic, 10 are shown in Table 5. We do note that there may not be a universally good cutoff value of weights. In fact, this analysis is simply to show that possible contamination does occur. A more rigorous "outlier" detection will be needed if desirable.

Data are also analyzed using the alternatives. The summary comparison results are provided in Table 6, and detailed estimation results using the alternatives are available in the Supporting Information (Table S17). It is observed that the proposed methods and alternatives make different discoveries. Similar to the marginal analysis, the two divergence methods generate highly overlapping findings but have moderate overlapping with the alternatives.

Remark 3. It is noted that the results of marginal and joint analyses are quite different, which is expected and shown in the literature, as their analysis schemes are fundamentally different. With the same reason, it is not sensible to compare marginal and joint analysis results.

TABLE 5 Joint analysis of TNBC: Identified main effects and interactions

	DPD			γ		
		Interaction			Interaction	
Gene	Main	Age	Race	Main	Age	Race
AADAT	0.1596	0.0764	−0.0235	0.1859	0.0949	−0.0084
ASPN	−0.0263	−0.0026	−0.0137	−0.0278	−0.0021	−0.0146
CCL13	0.0020	−0.0201	−0.0165			
CENPW	0.0798	−0.0257	−0.0784	0.0798	−0.0257	−0.0784
COL9A3	0.0615	−0.0059	−0.0433	0.0923	0.0159	−0.0099
CPA4	0.0365	−0.0538	−0.0339	0.0300	−0.0160	−0.0520
CXXC5	−0.6245	−0.1322	−0.3555	−0.1780	−0.0404	−0.2420
DCLRE1C	0.0739	−0.0035	−0.0492	0.0739	−0.0035	−0.0492
DNAJB11	0.0774	−0.0300	−0.0724	0.0774	−0.0300	−0.0724
DPF3	0.0419	−0.0127	−0.0086			
ERBB2	−0.0709	−0.0015	−0.0430	−0.0824		−0.0443
ESR1	−0.2735	−0.1500	−0.0914	−0.3243	−0.0750	−0.0839
HAPLN3	0.0996	−0.0097	−0.0029			
HDAC2	0.0108	−0.0095	−0.0097	0.0098	−0.0036	−0.0052
ITGB5	−0.0616	−0.0183	−0.0687	−0.0743	−0.0045	−0.0824
LYPD1	0.1918	−0.0665	−0.0037	0.2082	−0.0621	−0.0012
MISP3	0.0014	−0.0031	−0.0041	−0.0343	−0.0113	−0.0183
MMP12	0.0086	−0.0058		0.0090	−0.0054	
MMS22L	0.0794	−0.0512	−0.0462	0.0646	−0.0280	−0.0702
PCDHB9	0.0053	−0.0035	−0.0018	0.0053	−0.0035	−0.0018
PGAP3	−0.1338		−0.0346	−0.1376	−0.0018	−0.0360
PTCHD1	0.0569	−0.0031	−0.0639	0.0569	−0.0031	−0.0639
RPL39L	0.0071	−0.0014	−0.0093	0.0071	−0.0014	−0.0093
SIX3	0.0184	−0.0020	−0.0037	0.0199	−0.0062	
SLC15A1	0.0835		−0.0408	0.1021		−0.0406
SLC38A3	0.0855	−0.0406	−0.0028	0.1008	−0.0417	
SLC6A11	0.1375	0.0881	−0.0074	0.1375	0.0881	−0.0074
TLX1	0.0089		−0.0027	0.0089		−0.0025
TMEM217	0.0551	−0.0195	−0.0063	0.0551	−0.0195	−0.0063
TRPV6	−0.0096	−0.0068	−0.0062			
ZIC1	0.2012	−0.0693	−0.1491	0.2084	−0.0750	−0.1522
ASB12				−0.0101	−0.0750	0.1653
GZMB				0.1433	−0.0527	−0.0089
IL22RA2				0.0443	−0.0348	−0.0109
KCNS1				0.0665	−0.0027	0.0507
LYAR				0.0044	−0.0025	
PDIA5				0.0777	−0.0079	−0.0599
PLCG2				0.0792	−0.0398	−0.0061
PML				0.1148	−0.0593	−0.0055
PPP1R14C				0.0498		−0.0305
RAD51AP2				0.0950	−0.0012	0.1095
UBASH3B				0.0512	−0.0107	−0.0170
ZNF883				0.0424	−0.0323	−0.0149

TABLE 6 Analysis of TNBC: Numbers of main G effects and interactions identified by different methods and their overlaps

	Main effects				Interactions			
Marginal	**Logistic**	**Constant**	**DPD**	**γ**	**Logistic**	**Constant**	**DPD**	**γ**
Logistic	30	15	22	19	48	15	35	28
Constant		29	20	19		33	24	22
DPD			35	31			59	47
γ				32				51
Joint	**Logistic**	**Constant**	**DPD**	**γ**	**Logistic**	**Constant**	**DPD**	**γ**
Logistic	28	18	25	24	52	9	45	42
Constant		36	21	29		17	11	15
DPD			31	27			58	47
γ				39				70

4.2 | GENEVA type 2 diabetes data

As part of the Gene Environment Association Studies (GENEVA), the Health Professionals Follow-up Study (HPFS) was organized by the National Institutes of Health (NIH). GENEVA type 2 diabetes data, available at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1with the permission of National Human Genome Research Institute, is analyzed.

The response variable of interest is the continuously distributed body mass index (BMI), which is the principal measure of adiposity and plays an important role in diabetes. E factors considered include age, family history of diabetes among first degree relatives (famdb), total physical activity (act), trans fat intake (trans), cereal fiber intake (ceraf), and heme iron intake (heme). All of these E factors have been suggested as potentially associated with BMI. For G factors, we analyze SNPs on chromosome 4, which is suggested as having an important role in many disorders. The data contain 2558 subjects and 40,568 SNPs. Prescreening is conducted, and the top 2000 SNPs are kept.

4.2.1 | Marginal analysis

The proposed methods based on DPD and γ-divergence identify 35 main gene effects, 196 G-E interactions, and 43 main gene effects, 209 G-E interactions, respectively. Table S18 (Supporting Information) shows the detailed estimation results.

It is again observed that the findings are biologically sensible. UGT2B7 are catalytic enzymes in Mitiglinide carboxyl-glucuronidation in human liver, exhibiting high Mitiglinide glucuronosyltransferase activity in Mitiglinide glucuronide formation, and Mitiglinide is a new potassium channel antagonist for the treatment of type 2 diabetes mellitus (Yu et al., 2007). Published data analyses have examined the allelic association, confirming a significant association with the disease and revealing a significant association of BANK1 with diabetes, which suggests the possibility that BANK1 is a susceptibility gene. Some studies have further provided evidence of new genetic associations of BANK1 gene with diabetes (Zouidi et al., 2014). PPA2 has a function in feeding behavior by controlling the phosphate level of the cell, and PPA2 is a negative regulator of the insulin metabolic signaling pathway, which may contribute to abnormal BMI (Noratto et al., 2016). Elovl6 is a microsomal enzyme-converting palmitoleates saturated and monounsaturated fatty acids into oleate species, which plays a critical role in the development of obesity-induced insulin resistance by modifying fatty acid composition. Elovl6 is a fundamental factor linking dysregulated lipid metabolism to β-cell dysfunction, islet inflammation, and β-cell apoptosis in type 2 diabetes (Zhao et al., 2017).

Identification stability is evaluated, and the results are provided in Table S19 (Supporting Information). The proposed methods have satisfactory stability. Data are also analyzed using the alternatives. The summary comparison results are provided in Table S20 (Supporting Information), and detailed estimation results using the alternatives are available in Supporting Information (Table S21). The two divergence methods generate highly overlapping results but have small overlapping with the alternatives.

4.2.2 | Joint analysis

The proposed methods based on DPD and γ-divergence identify 25 main gene effects, 115 G-E interactions, and 29 main gene effects, 142 G-E interactions, respectively. Table S22 (Supporting Information) shows the detailed estimation results.

Literature search suggests that the identified genes may have important implications. For instance, the activation effect of the CSN1S2-derived bioactive peptides for glucokinase-binding affinity of glucose has been indicated, and the protein regulated by gene CSN1S2 has been suggested to be a common nutrient used for the treatment of diabetes mellitus (Fatchiyah et al., 2017). CCNI is the most expressed cyclin genes and the most highly expressed in pancreatic islets (Taneera et al., 2013). In genome-wide association studies, C4orf22 has been found to reduce the risk of insulin resistance (Daily et al., 2019). The abnormally high levels of SCD5 in diabetes and SCD5 may be a common molecular link among diabetes (Bellenghi et al., 2015). Besides, the ratio of ADH1B protein expression in adipose tissue from low BMI individuals is approximately fivefold higher than that observed in high BMI individuals, and ADH1B expression, measured both by Illumina BeadArrays and qRT-PCR, is negatively correlated with BMI (Winnier et al., 2015). It has been shown that MTHFD2L is associated with diabetes in other genome-wide association studies (Chidambaram et al., 2016).

Table S23 (Supporting Information) shows satisfactory stability of the proposed methods. Data are also analyzed using the alternatives. The summary comparison results are provided in Table S24 (Supporting Information), and detailed estimation results using the alternatives are available in the Supporting Information (Table S25). Similarly, the proposed methods and alternatives make different discoveries.

5 | DISCUSSION

Identifying G-E interactions associated with outcomes has important implications. In this article, we have proposed a framework of G-E interaction analysis based on robust divergence to accommodate contaminated data. A sparse group penalty has been adopted to respect the "main effect, interaction" hierarchical structure. Both joint and marginal analysis have been conducted. In addition, categorical and continuous responses, as two important special cases, have been examined in detail. And some other responses can also be accommodated under the proposed framework. In simulation, the proposed methods have notable advantages over the alternatives when data are contaminated. In real data analysis, sensible biological implications and identification stability have provided support to the validity of the proposed methods. It is noted that results from the DPD and γ-divergence methods are often different, and we recognize that it is not easy to conclude which method is better when analyzing a practical data. As such, the two results have been comprehensively considered. There are also some limitations. For example, it is difficult to provide clear interpretations of the identified interactions due to limited studies in the literature, which needs to be refined by further biological studies.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their careful review and insightful comments. This work was supported by the Beijing Natural Science Foundation (Z190004), National Natural Science Foundation of China (11971404, 12171454, 12026604), Humanity and Social Science Youth Foundation of Ministry of Education of China (19YJC910010), Basic Scientific Project 71988101 of National Science Foundation of China, the 111 Project (B13028), the University of Chinese Academy of Sciences (Y95401TXX2), Key Program of Joint Funds of the National Natural Science Foundation of China (U19B2040), NIH (CA204120), and NISF (1916251).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study have been derived from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/). The data can be downloaded using the R package brca.data (https://github.com/averissimo/brca.data/releases/download/1.0/brca.data_1.0.tar.gz).

OPEN RESEARCH BADGES

 This article has earned an open data badge for making publicly available the code necessary to reproduce the reported results. Some of the data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID

Sanguo Zhang https://orcid.org/0000-0002-0688-0016

REFERENCES

Adler, R., Feldman, R., & Taqqu, M. (Eds.) (1998). *A practical guide to heavy tails: Statistical techniques and applications*. Springer Science & Business Media.

Bao, C., Lu, Y., Chen, J., Chen, D., Lou, W., Ding, B., Xu, L.& Fan, W. (2019). Exploring specific prognostic biomarkers in triple-negative breast cancer. *Cell Death & Disease*, *10*(11), 1–14.

Basu, A., Harris, I., Hjort, N., & Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Bellenghi, M., Puglisi, R., Pedini, F., De F., Felicetti, F., Bottero, L., Sangaletti, S., Errico, M., Petrini, M.& Gesumundo, C. (2015). SCD5-induced oleic acid production reduces melanoma malignancy by intracellular retention of SPARC and cathepsin B. *The Journal of Pathology*, *236*(3), 315–325.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, *41*(3), 1111–1141.

Bien, J., Taylor, J., & Tibshirani, R. (2013). Convex hierarchical testing of interactions. *The Annals of Applied Statistics*, 27–42.

Burnett, R., Craven, K., Krishnamurthy, P., Goswami, C., Badve, S., Crooks, P., Mathews, W., Bhat-Nakshatri, P.& Nakshatri, H. (2015). Organ-specific adaptive signaling pathway activation in metastatic breast cancer cells. *Oncotarget*, *6*(14), 12682.

Bydoun, M., Marcato, P. & Dellaire, G. (2014). *Breast cancer genomics*. Elsevier.

Castellana, B., Escuin, D., Peiró, G., Garcia-Valdecasas, B., Vázquez, T., Pons, C., Pérez-Olabarria, M., Barnadas, A.& Lerma, E. (2012). ASPN and GJB2 are implicated in the mechanisms of invasion of ductal breast carcinomas. *Journal of Cancer*, *3*, 175.

Chidambaram, M., Liju, S., Saboo, B., Sathyavani, K., Viswanathan, V., Pankratz, N., Gross, M., Mohan, V., & Radha, V. (2016). Replication of genome-wide association signals in Asian Indians with early-onset type 2 diabetes. *Acta Diabetologica*, *53*(6), 915–923.

Copas, J. (1988). Binary regression models for contaminated data. *Annals of Statistics*, *41*(3), 1111–1141.

Coyle, K., Dean, C., Thomas, M., Vidovic, D., Giacomantonio, C., Helyer, L., & Marcato, P. (2018). DNA methylation predicts the response of triple-negative breast cancers to all-transretinoic acid. *Cancers*, *10*(11), 397.

Daily, Ja., Liu, M.& Park, S. (2019). High genetic risk scores of SLIT3, PLEKHA5 and PPP2R2C variants increased insulin resistance and interacted with coffee and caffeine consumption in middle-aged adults. *Molecular and Cellular Endocrinology*, *29*(1), 79–89.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis–A review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 203–219.

Farcomeni, A., & Ventura, L. (2010). An overview of robust methods in medical research. *Statistical Methods in Medical Research*, *21*(2), 111–133.

Fatchiyah, F., Rahasta, A., & Cairns, J. (2017). Virtual screening and prediction of binding of caprine CSN1S2 protein tryptic peptides to glucokinase. *Acta Informatica Medica*, *25*(4), 225.

Foulkes, W., Smith, I., & Reisfilho, J. (2010). Triple-negative breast cancer. *Wiener Medizinische Wochenschrift*, *160*(7-8), 174–181.

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.

Hammond, M., 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.

He, J., Yang, J., Chen, W., Wu, H., Yuan, Z., Wang, K., Li, G., Sun, J.& Yu, L. (2015). Molecular features of triple negative breast cancer: Microarray evidence and further integrated analysis. *PloS ONE*, *10*(6), e0129842.

Hung, H., Jou, Z., & Huang, S. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, *74*(1), 145–154.

Jones, M., Hjort, N., Harris, I.& Basu, A. (2001). A comparison of related density based minimum divergence estimators. *Biometrika*, *88*(3), 865–873.

Khaled, W., Lee, S., Stingl, J., Chen, X., Ali, H., Rueda, O., Hadi, F., Wang, J., Yu, Y. & Chin, S., (2015). BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nature Communications*, *6*, 5987.

Kim, G., Lai, C., Arnett, D., Parnell, L., Ordovas, J., Kim, Y., & Kim, J. (2017). Detection of gene-environment interactions in a family-based population using SCAD. *Statistics in Medicine*, *36*(22), 3547–3559.

Krings, G., Nystrom, M., Mehdi, I., Vohra, P.& Chen, Y. (2014). Diagnostic utility and sensitivities of GATA3 antibodies in triple-negative breast cancer. *Human Pathology*, *45*(11), 2225–2232.

Li, Y., Tang, X., Bai, Z. & Dai, X. (2016). Exploring the intrinsic differences among breast tumor subtypes defined using immunohistochemistry markers based on the decision tree. *Scientific Reports*, *6*, 35773.

Li, K., Ying, M., Feng, D., Du, J., Chen, S., Dan, B., Wang, C. & Wang, Y. (2016). Fructose-1,6-bisphosphatase is a novel regulator of Wnt/β-catenin pathway in breast cancer. *Biomedicine & Pharmacotherapy*, *84*, 1144–1149.

Lin, I., & Hsu, M. (2015). Genome-wide gene expression analysis to identify epistatic gene-pairs associated with prognosis of breast cancer. In M. Gunduz (Ed.), *A concise review of molecular pathology of breast cancer* (p. 57). IntechOpen.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., & Ma, S. (2013). Identification of gene-environment interactions in cancer studies using penalization. *Genomics*, *102*(4), 189–194.

Lopes, M., Verssimo, 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.

Lv, X., He, M., Zhao, Y., Zhang, L., Zhu, W., Jiang, L., Yan, Y., Fan, Y., Zhao, H.& Zhou, S. (2019). Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer. *Cancer Cell International*, *19*(1), 172.

Niu, J., Andres, G., Kramer, K., Kundranda, M., Alvarez, R., Klimant, E., Parikh, A., Tan, B., Staren, E.& Markman, M. (2015). Incidence and clinical significance of ESR1 mutations in heavily pretreated metastatic breast cancer patients. *OncoTargets and Therapy*, *8*, 3323.

Noratto, G., Chew, B. P., & Ivanov, I. (2016). Red raspberry decreases heart biomarkers of cardiac remodeling associated with oxidative and inflammatory stress in obese diabetic db/db mice. *Food & Function*, *7*(12), 4944–4955.

Santuario F., Cardona H., Perez P., Trevino, V., Hernandez C., Rojas-Martinez, A., Uscanga-Perales, G., Martinez-Rodriguez, J., 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.

Shen, J., & He, X. (2015). Inference for subgroup analysis with a structured logistic normal mixture model. *Journal of the American Statistical Association*, *110*(509), 303–312.

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., & Ma, S. (2014). A penalized robust method for identifying gene-environment interactions. *Genetic Epidemiology*, *38*(3), 220–230.

Shieh, A., & Hung, Y. (2009). Detecting outlier samples in microarray data. *Statistical Applications in Genetics and Molecular Biology*, *8*(1), 1–24.

Taneera, J., Fadista, J., Ahlqvist, E., Zhang, M., Wierup, N., Renström, E., & Groop, L. (2013). Expression profiling of cell cycle genes in human pancreatic islets with and without type 2 diabetes. *Molecular and Cellular Endocrinology*, *375*(2), 35–42.

Tervasmäki, A., Winqvist, R., Jukkola-Vuorinen, A., & Pylkäs, K. (2014). Recurrent CYP2C19 deletion allele is associated with triple-negative breast cancer. *BMC Cancer*, *14*(1), 902.

Waddell, N., Jeremy, A., Sibylle, C., Leonard, S., Anna, M., Joan, R., Cameron, N., Mohammed, O., Guillaume, A., x00026; Charis, E. (2010). Subtypes of familial breast tumours revealed by expression and copy number profiling. *Breast Cancer Research & Treatment*, *123*(3), 661–677.

Walter, S. (2005). The partial area under the summary ROC curve. *Statistics in Medicine*, *24*(13), 2025.

Wang, X., Tan, Y., Cao, X., Kim, J., Chen, T., Hu, Y., Wexler, M., & Wang, X. (2018). Epigenetic activation of HORMAD1 in basal-like breast cancer: Role in Rucaparib sensitivity. *Oncotarget*, *9*(53), 30115.

Wang, B., Zhao, H., Zhao, L., Zhang, Y., Wan, Q., Shen, Y., Bu, X., Wan, M., & Shen, C. (2017). Up-regulation of OLR1 expression by TBC1D3 through activation of TNFα/NF-κB pathway promotes the migration of human breast cancer cells. *Cancer Letters*, *408*, 60–70.

Winnier, D., Fourcaudot, M., Norton, L., Abdul-Ghani, M., Hu, S., Farook, V., Coletta, D., Kumar, S., Puppala, S., & Chittoor, G. (2015). Transcriptomic identification of ADH1B as a novel candidate gene for obesity and insulin resistance in human adipose tissue in Mexican Americans from the Veterans Administration Genetic Epidemiology Study (VAGES). *PloS ONE*, *10*(4), e0119941.

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, C., Jiang, Y., Ren, J., Cui, Y.& Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in Medicine*, *37*(3), 437–456.

Wu, M., & Ma, S. (2018). Robust genetic interaction analysis. *Briefings in Bioinformatics*, *20*(2), 624–637.

Wu, C., Shi, X., Cui, Y., & Ma, S. (2015). A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, *34*(30), 4016–4030.

Yu, L., Lu, S., Lin, Y., & Zeng, S. (2007). Carboxyl-glucuronidation of mitiglinide by human UDP-glucuronosyltransferases. *Biochemical Pharmacology*, *73*(11), 1842–1851.

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. *The Annals of Statistics*, *38*(2), 894–942.

Zhao, H., Matsuzaka, T., Nakano, Y., et al. (2017). Elovl6 deficiency improves glycemic control in diabetic db/db mice by expanding β-cell mass and increasing insulin secretory capacity. *Diabetes*, *66*(7), 1833–1846.

Zhu, R., Zhao, H., & Ma, S. (2014). Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genetic Epidemiology*, *38*(4), 353–368.

Zouidi, F., Stayoussef, M., Bouzid, D., Fourati, H., Abida, O., João, C., Ayed, M., Fakhfakh, R., Thouraya, K.& Monjia, H. (2014). Association of BANK1 and cytokine gene polymorphisms with type 1 diabetes in Tunisia. *Gene*, *536*(2), 296–301.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ren M, Zhang S, Ma S, Zhang Q. Gene–environment interaction identification via penalized robust divergence. *Biometrical Journal*. 2022;*64*:461–480. https://doi.org/10.1002/bimj.202000157

APPENDIX A: DETAILS FOR ALGORITHMS

The details of Steps (i) and (ii) in Algorithm 2 for joint analysis are as follows. The implementation of Algorithm 1 for marginal analysis is very similar. Steps (i) and (ii) are realized by gradient descent and Armijo search.

A.1 | Details for algorithms under the logistic model

The loss functions with DPD and γ-divergence, respectively, are

$$\ell_\gamma(\boldsymbol{\theta},\boldsymbol{b}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp\{y_i(\gamma+1)(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b})\}}{1+\exp\{(\gamma+1)(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b})\}}\right)^{\frac{\gamma}{\gamma+1}},$$

$$\ell_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1+\exp\{(1+\alpha)\left(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b}\right)\}}{\left(1+\exp\{(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b})\}\right)^{1+\alpha}} - \left(1+\frac{1}{\alpha}\right)\frac{\exp\{\alpha y_i\left(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b}\right)\}}{(1+\exp\{(\boldsymbol{x}_i^T\boldsymbol{\theta}+\boldsymbol{w}_i^T\boldsymbol{b})\})^{\alpha}}\right).$$

The two robust methods have the same form of estimating equation:

$$(\boldsymbol{X}^T \vdots \boldsymbol{W}^T)^T S(\boldsymbol{\theta},\boldsymbol{b}) = 0, \tag{A.1}$$

where $S(\boldsymbol{\theta},\boldsymbol{b}) = S_\gamma(\boldsymbol{\theta},\boldsymbol{b})$ or $S_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b})$, and

$$\begin{aligned} S_\gamma(\boldsymbol{\theta},\boldsymbol{b}) &= \gamma\omega_\gamma(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b})\odot\{\boldsymbol{Y}-\pi(\boldsymbol{X},\boldsymbol{W};(1+\gamma)\boldsymbol{\theta},(1+\gamma)\boldsymbol{b})\},\\ S_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b}) &= (1+\alpha)\{\omega_\alpha(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b})\odot[\boldsymbol{Y}-\pi(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b})]-\delta_\alpha(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b})\}. \end{aligned} \tag{A.2}$$

Denote $\odot$ as the componentwise product and

$$\begin{aligned} \omega_\alpha(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b}) &= \frac{\exp\{\alpha\boldsymbol{Y}\odot(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})\}}{(1+\exp\{(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})\})^{\alpha}},\\ \omega_\gamma(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b}) &= \left(\frac{\exp\{(\gamma+1)\boldsymbol{Y}\odot(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})\}}{1+\exp\{(\gamma+1)(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})\}}\right)^{\frac{\gamma}{\gamma+1}},\\ \pi(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b}) &= \frac{\exp(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})}{1+\exp(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})},\\ \delta_\alpha(\boldsymbol{X},\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{b}) &= \frac{\exp\{\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b}\}(\exp\{\alpha(\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b})\}-1)}{1+\exp\{\boldsymbol{X}^T\boldsymbol{\theta}+\boldsymbol{W}^T\boldsymbol{b}\}}, \end{aligned} \tag{A.3}$$

with the operations being elementwise except for the matrix–vector multiplication of $\boldsymbol{X}^T\boldsymbol{\theta}$ and $\boldsymbol{W}^T\boldsymbol{b}$.

We propose the following iterations.

Step (i):

$$\theta^{(m+1)} = \theta^{(m)} + \frac{s_0}{n}X^T\boldsymbol{S}^{(m)}, \tag{A.4}$$

where $\boldsymbol{S}^{(m)} = S(\boldsymbol{\theta}^{(m)},\boldsymbol{b}^{(m)})$ is defined in (A.2), and s_0 is the step length obtained by Armijo search.

Step (ii):

For $j = 1,\dots,p$

$$\begin{aligned} \zeta_j^{(m+1)} &= \zeta_j^{(m)} + s\left\{\frac{1}{n}\boldsymbol{W}_j^T\boldsymbol{S}^{(m)} - \frac{\zeta_j^{(m)}}{\|\boldsymbol{b}_j^{(m)}\|}\sqrt{q+1}\lambda_1\left(1-\frac{\|\boldsymbol{b}_j^{(m)}\|}{\sqrt{q+1}\lambda_1 a}\right)_+\right\},\\ \boldsymbol{\beta}_j^{(m+1)} &= \boldsymbol{\beta}_j^{(m)} + \frac{s}{n}\boldsymbol{W}_j^T\boldsymbol{S}^{(m)}\\ &\quad - s\left\{\frac{\boldsymbol{\beta}_j^{(m)}}{\|\boldsymbol{b}_j^{(m)}\|}\sqrt{q+1}\lambda_1(1-\frac{\|\boldsymbol{b}_j^{(m)}\|}{\sqrt{q+1}\lambda_1 a})_+ + \mathrm{sign}(\boldsymbol{\beta}_j^{(m)})\odot\lambda_1(1-\frac{\boldsymbol{\beta}_j^{(m)}}{\lambda_1 a})_+\right\}, \end{aligned} \tag{A.5}$$

where s is the step length obtained by Armijo search. Computations iterated until convergence.

A.2 | Details for algorithms under the linear model

The loss functions with DPD and γ-divergence, respectively, are

$$\ell_\gamma(\boldsymbol{\theta},\boldsymbol{b}) = -\frac{1}{n}\left(\frac{1+\gamma}{2\pi\sigma^2}\right)^{\frac{\gamma}{2(1+\gamma)}} \sum_{i=1}^{n}\left\{\exp\left[-\frac{\gamma\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b}\right)^2}{2\sigma^2}\right]\right\},$$

$$\ell_{\mathrm{DPD}}(\boldsymbol{\theta},\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha}\exp\left[-\frac{\alpha\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b}\right)^2}{2\sigma^2}\right]\right\},$$

where σ is the standard deviation of the response and needs to be estimated.

The estimating equation for σ based on γ-divergence and density power divergence, respectively, are

$$\begin{aligned}
\gamma &: \sum_{i=1}^{n}\left[\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b}\right)^2 - \frac{\sigma^2}{1+\gamma}\right]\exp\left\{-\frac{\gamma\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b}\right)^2}{2\sigma^2}\right\} = 0,\\
\mathrm{DPD} &: \sum_{i=1}^{n}\left[1-\frac{(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b})^2}{\sigma^2}\right]\exp\left\{-\frac{\alpha\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\theta}-\boldsymbol{w}_i^T\boldsymbol{b}\right)^2}{2\sigma^2}\right\} = \frac{n\alpha}{(1+\alpha)^{\frac{3}{2}}}.
\end{aligned} \tag{A.6}$$

Add the following step before Steps (i) and (ii) in each iteration:

Step ($*$):

Calculate $\sigma^{(m)}$ by solving (A.6) with the biselection method, with $\boldsymbol{\theta},\boldsymbol{b}$ fixed at $\boldsymbol{\theta}^{(m)},\boldsymbol{b}^{(m)}$.

The two robust methods have the same estimating equation for $(\boldsymbol{\theta},\boldsymbol{b})$:

$$(\boldsymbol{X}^T : \boldsymbol{W}^T)^T\tilde{S}(\boldsymbol{\theta},\boldsymbol{b}) = 0, \tag{A.7}$$

where

$$\tilde{S}(\boldsymbol{\theta},\boldsymbol{b}) = \exp\left\{-\frac{\iota\left(\boldsymbol{Y}-\boldsymbol{X}^T\boldsymbol{\theta}-\boldsymbol{W}^T\boldsymbol{b}\right)^2}{2\sigma^2}\right\}\odot\left(\boldsymbol{Y}-\boldsymbol{X}^T\boldsymbol{\theta}-\boldsymbol{W}^T\boldsymbol{b}\right). \tag{A.8}$$

$\iota=\gamma$ or α, $\odot$ is the componentwise product, and the operations are elementwise except for the matrix–vector multiplication of $\boldsymbol{X}^T\boldsymbol{\theta}$ and $\boldsymbol{W}^T\boldsymbol{b}$.

Steps (i) and (ii) under the linear model are same as those under the logistic model, when replacing $S(\boldsymbol{\theta},\boldsymbol{b})$ in (A.4) and (A.5) with $\tilde{S}(\boldsymbol{\theta},\boldsymbol{b})$ defined by (A.8). Then repeat Step ($*$), Step (i), and Step (ii) until convergence.