文章编号:2095-6134(2024)02-0151-14

Robust individualized subgroup analysis*

ZHANG Xiaoling, REN Mingyang, ZHANG Sanguo†

(*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049, *China*; *Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing* 100049, *China*)

(Received 8 February 2022; Revised 13 April 2022)

Zhang X L, Ren M Y, Zhang S G. Robust individualized subgroup analysis[J]. Journal of University of Chinese Academy of Sciences, 2024, 41(2):151-164. DOI:10.7523/j.ucas.2022.037.

Abstract Subgroup analysis of heterogeneous groups is a crucial step in the development of individualized treatment and personalized marketing strategies. Regression-based approaches are one of the main schools of subgroup analysis, a paradigm that divides predictor variables into two parts with heterogeneous and homogeneous effects and divides the sample into subgroups based on the heterogeneous effects. However, most of the existing regression-based subgroup analysis methods have two major limitations: First, they still consider the sample homogeneous within subgroups and do not fully consider individual effects; Second, the common contamination phenomenon of homogeneous effect variables is not taken into account, which will lead to large bias in the model results. To address these challenges, we propose a robust individualized subgroup analysis. We use a multidirectional separation penalty function to achieve individualized effects analysis for the heterogeneous part of the model and use γ-divergence to obtain robust estimates for the contaminated homogeneous part. We also propose an efficient alternating iterative two-step algorithm, combining coordinate descent and alternating direction method of multipliers (ADMM) techniques to implement this process. Our proposed method is further illustrated by simulation studies and analysis of a skin cutaneous melanoma dataset.

Keywords subgroup analysis; multidirectional separation penalty; robust regression; variable selection

CLC number: O212.1 **Document code**: A **DOI**: 10.7523/j.ucas.2022.037

稳健的个体化亚组分析

张晓灵,任明旸,张三国

(中国科学院大学数学科学学院,北京 100049; 中国科学院大数据挖掘与知识管理重点实验室,北京 100049)

摘 要 异质群体的亚组分析是实现个体化医疗和个性化营销的关键所在。基于回归的方法是亚组分析的主要流派之一,这种范式将预测变量分为具有异质效应和同质效应的两部分,并根据异质变量是否相同将样本分为不同的亚组。然而,现有的基于回归的亚组分析方法大多有两大局限性:第一,它们仍然认为亚组内的样本是同质的,没有充分考虑个体效应;第二,

* Support by National Natural Science Foundation of China (12171454) and Key R&D Program of Guangxi (2020AB10023)

† Corresponding author, E-mail: sgzhang@ucas.ac.cn

没有考虑到同质变量中常见污染现象,这将导致模型结果出现较大偏差。为应对这些挑战,提出一种稳健的个体化亚组分析方法。使用多向分离惩罚函数估计模型异质部分的个体化效应,并使用γ散度得到同质部分的稳健估计。还提出一种高效的交替迭代的两步算法,这一方法结合了坐标下降法和交替方向乘子法。数值模拟和对皮肤黑色素瘤数据的分析进一步验证了所提方法的有效性。

关键词 亚组分析;多向分离惩罚;稳健回归;变量选择

In recent years, there has been a growing demand to explore individualized models, which have a wide range of applications for personalized medicine, personalized education and personalized marketing. In the era of big data, heterogeneous data is one of the key challenges in data analytics, which is to correctly identify subgroups from a heterogeneous population in order to target treatment or marketing for each subgroup. For example, in the fight against diseases such as cancer, the effectiveness of a new medicine to treat a disease is evaluated for the whole population. However, if there is significant heterogeneity in treatment effects due to genetic variation or environmental influences, then new treatments are likely to be particularly effective for some patient subgroups and ineffective or less effective for others. Therefore, subgroups of patients with desired outcomes need to be identified based on appropriate statistical methods.

In order to solve this problem, we need to identify potential subgroup structures. In general, subgroup identification can be achieved by clustering samples. To group different individuals, Hocking et al. [1] and Lindsten et al. [2] used L_p fused penalties, treating clustering as a problem of penalized regression. Pan et al. [3] and Ma and Huang[4] used a non-convex fused penalty to reduce bias. However, the fused penalty approach focuses on subgroups rather than model selection of individual coefficients. In other words, although existing subgroup analysis methods have explored possible heterogeneous effects across samples, they have not adequately considered individual effects. For example, in genetic studies to identify biomarkers associated with a particular disease, a gene may be a relevant biomarker for one individual in the population but not for other individuals. Furthermore, for different genes, the effects on heterogeneous covariates may vary between individuals. Therefore, individualized variable selection is important, as different individuals may have different sets of biomarker genes. In addition, the rise of precision medicine and personalized marketing strategies has driven us to develop more effective personalized treatments and recommendations by selecting unique characteristics for each individual. The rich collection of data information makes the use of individualized models feasible and convincing, as traditional aggregate models cannot incorporate heterogeneous effects across individuals. Tang et al. [5] proposed an effective method for individualized model selection using a multidirectional separation penalty function to select significant covariates for different individuals and to simultaneously identify subgroups based on the effects of heterogeneous covariates.

Although the individualized model proposed by Tang et al. [5] can realize the feature selection of individuals and the subgroup division of heterogeneous covariates, it is only for ideal data. However, in the real data, especially in the data of genes and diseases, there are often more complex situations. For example, in The Cancer Genome Atlas (TCGA) collection data on skin cucumber melanoma (SKCM) data, studies have shown that[6] the environmental variables are heterogeneous, important to some samples and not important to some samples. And the remaining high-dimensional gene expression variables are homogeneous, and some homogeneous variables are completely unimportant, so variable selection is needed. In addition, due to some technical reasons, there are often some measurement errors in gene expression data, resulting in long-tailed distributions or contamination.

Therefore, robust methods need to be used to process this part of data. The method of individualized model[5] can not select homogeneous variables, and it is not suitable for dealing with contaminated data.

Outliers in data are often encountered with biomedicine, image processing and other areas. However, traditional linear models require assumptions and expectations of the correct variables. The data may be contaminated due to inadequate access to information, errors in subjective judgement and measurement errors, which will lead to bias in the estimates derived from traditional linear models. Robust estimation methods have thus been developed to reduce the impact of outliers on data analysis. The initially used is the least absolute deviation (LAD), but the calculation is more complex. In recent years, the divergence-based methods have been developed. Two of the more common methods are density power divergence and γ-divergence. The MDPD (minimum density power divergence) method was first proposed by Basu et al.[7], it is mainly used to solve the problem of parameter estimation for density distributions. Many statistical models constructed on the basis of density power divergence have been shown in the literature to exhibit excellent robustness. Fujisawa and Eguchi[8] proposed a robust parameter estimation method for Gaussian mixture models based on density power divergence; Ghosh and Basu[9] and Durio and Isaia[10] have extended the method to regression problems and have shown good robustness; Zang et al.[11] proposed a high-dimensional robust parameter estimation method based on density power divergence and applied it to a high-dimensional linear regression model with multiple response variables. The MDPD method is more robust than the LAD method and can deal well with data contamination and heavy-tailed distribution of residuals. Jones et al.[12] first proposed the γ-divergence method for robust estimation of parameters from a single distribution. It was later extended to robust regression methods for low-dimensional data by Fujisawa and Eguchi[13], so that the estimates obtained have strong robustness, and the potential bias can be sufficiently small even in the case of heavy contamination. None of the other robust methods can achieve these properties, and the estimates are affected by the proportion of contaminated data. It has been shown that the estimation has good statistical and numerical advantages. Kawashima and Fujisawa[14] proposed a robust sparse regression based on γ-divergence to establish robust properties from Pythagorean relations. Hung et al.[15] proposed γ-logistic regression. As γ-logistic regression can ignore the bias caused by the contamination distribution and the proportion of contamination, the probability of the wrong category in the model does not need to be modelled. The MDPD logistic regression was also compared with the γ-logistic regression, showing that the γ-logistic regression has stronger robustness. Ren et al.[16] used γ-divergence on a high-dimensional generalized linear model to deal with multiple types of anomalous responses and rigorously established consistency in variable selection and estimation bounds. However, γ-divergence has not been used for individualized subgroup analysis.

The main contribution of this paper is as follows. First, we propose a robust regression-based individualized subgroup analysis method. Specifically, a multidirectional separation penalty is introduced to analyze heterogeneous individualized effects, and γ-divergence and regularization techniques are introduced to simultaneously achieve variable selection and robust analysis of homogeneous effects with possible contamination data. Second, an effective twostep algorithm with stepwise alternating iterations, combining coordinate descent and ADMM techniques, is proposed to address the difficulties of objective function optimization. Numerical simulations demonstrate the effectiveness of this algorithm. Third, the real data for skin cutaneous melanoma (SKCM) effectively explores the individualized heterogeneous effects of this disease and provides a practical analytical framework for the analysis of such complex diseases.

1 Methodology

1.1 Model settings

We formulate the problem under the heterogeneous regression model. For the ith individual, y_i is a response variable, $\boldsymbol{X}_i = (x_{i1},\cdots,x_{ip})^{\mathrm{T}}$ is a p-dimensional vector of predictors with heterogeneous effects, and $\boldsymbol{Z}_i = (z_{i1},\cdots,z_{iq})^{\mathrm{T}}$ is a $q-$dimensional vector of homogengous effects. The model is denoted as

$$y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i,\ i = 1,\cdots,n, \quad (1)$$

where each individual has a unique heterogeneous effect $\boldsymbol{\beta}_i = (\beta_{i1},\cdots,\beta_{ip})^{\mathrm{T}}$ related to some certain variables $\boldsymbol{X}_i$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_n)$. The homogeneous effect $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_q)^{\mathrm{T}}$ is related to $\boldsymbol{Z}_i$. The random errors ε_i are independent and $E(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) < \infty$.

The heterogeneous linear model (1) can be decomposed into

$$\begin{cases} y_i^{(1)} = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i^{(1)}, \\ y_i^{(2)} = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i^{(2)}, \\ y_i = y_i^{(1)} + y_i^{(2)}, \\ E(\varepsilon_i^{(1)}) = 0, \mathrm{Var}(\varepsilon_i^{(1)}) < \infty, \\ E(\varepsilon_i^{(2)}) = 0, \mathrm{Var}(\varepsilon_i^{(2)}) < \infty. \end{cases} \quad (2)$$

We can see that the decomposition of this model (2) combines the existing subgroup analysis heterogeneous model and the classical homogeneous linear model.

1.2 Robust individual estimation based on γ-divergence

We propose a robust individualized subgroup analysis method based on γ-divergence, which can deal with the linear model with contamination in the homogeneous part and has a individual penalty in the heterogeneous part. According to (2), our objective function is

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau}) = L_1(\boldsymbol{\beta}) + S(\boldsymbol{\beta},\boldsymbol{\tau}) + L_2(\boldsymbol{\alpha}) + P(\boldsymbol{\alpha}). \quad (3)$$

The first term $L_1(\boldsymbol{\beta})$ is the square loss function,

$$L_1(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i^{(1)} - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i)^{\mathrm{T}}(y_i^{(1)} - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i). \quad (4)$$

The second term $S(\boldsymbol{\beta},\boldsymbol{\tau})$ is the multidirectional separation penalty (MDSP). The MDSP was first proposed by Tang et al.[5]. We use the MDSP to get the heterogeneity estimate $\hat{\boldsymbol{\beta}}$. We consider different subgrouping with respect to different heterogeneous predictors. We assume that every heterogeneous variable $\boldsymbol{X}_i$ has B_k subgroups,

$$\beta_{ik} = \begin{cases} \tau_k^{(l)}, & \text{if } i \in g_k^{(l)}, \quad l = 1,\cdots,B_k - 1, \\ 0, & \text{if } i \in g_k^{(0)}, \end{cases} \quad \text{for } i = 1,\cdots,n, \quad (5)$$

where $\tau_k^{(l)}(l = 1,\cdots,B_k - 1)$ is unknow nonzero sub-homogeneous effect shared by individuals with lth subgroup, and each potential subgroup is represented by the index partition set $\{g_k^{(l)}\}_{l=0,1,\cdots,B_k-1}$ heterogeneous covariate as an example: one is a subgroup of zero effect ($\beta_{ik} = 0$, $i \in g_k$) and the other is a subgroup of non-zero effect ($\beta_{ik} = \tau_k$, $i \in g_k^c$). And the multidirectional separation (MDSP) function $S(\boldsymbol{\tau},\boldsymbol{\beta})$ is defined as

$$S(\boldsymbol{\beta},\boldsymbol{\tau}) = \sum_{i=1}^{n}\sum_{k=1}^{p} s_{\lambda_2}(\beta_{ik},\tau_k), \quad (6)$$

$$s_{\lambda_2}(\beta_{ik},\tau_k) = \lambda_2 \min(|\beta_{ik}|, |\beta_{ik} - \tau_k|), \quad (7)$$

where λ_2 is a tuning parameter for individual penalty. Given τ_k, the MDSP function provides β_{ik} another shrinking direction τ_k except 0, which can reduce the bias and reduce the sparsity of weak signals. The MDSP term $\sum_{i=1}^{n} s_{\lambda_2}(\beta_{ik},\tau_k)$ is a center-based clustering. The center τ_k of the subgroup and the subgroup members obtained from each contraction direction are iteratively updated. This allows each individual to shrink in the best direction, thus improving the individual model fit, while further adaptively estimating $\hat{\tau}$ to obtain subgroups of individuals. More specific details about MDSP can be found in Tang et al.[5].

The third term $L_2(\boldsymbol{\alpha})$ is the γ-divergence loss function. According to the existing literature[13] that uses γ-divergence to deal with linear regression, they argue that the response variable y will deviate partially from the normal distribution due to the y presence of outliers. And by adjusting the appropriate γ values, robust estimates can be obtained with deviating from the normal distribution.

Therefore, we adopted one of the same settings and practices as theirs. According to the second equation of model (2), we can get conditional probability of $y_i^{(2)}$

$$f(y_i^{(2)} \mid \boldsymbol{Z}_i;\boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y_i^{(2)} - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2}{2\sigma^2}\right\}. \tag{8}$$

We use γ-divergence to deal with contamination in homogenerous variables. For two density functions $f_\theta(x)$ and $g(x)$, the γ-divergence is defined as

$$D_\gamma(g(x), f_\theta(x)) = \frac{1}{\gamma(\gamma+1)} \left\{\| g(x) \|_\gamma - \int\left(\frac{f_\theta(x)}{\| f_\theta(x) \|_{\gamma+1}}\right)^\gamma g(x)\,\mathrm{d}x\right\}, \gamma > 0, \tag{9}$$

where f_θ is the model distribution under p dimensional parameter θ, g is the distribution of data generation, and $\| g(x) \|_\gamma = \left(\int g(x)^\gamma \mathrm{d}x\right)^{\frac{1}{\gamma}}$, $\| f_\theta(x) \|_{\gamma+1} = \left(\int f_\theta^{\gamma+1}(x)\,\mathrm{d}x\right)^{\frac{1}{\gamma+1}}$. And the parameter γ balances robustness and efficiency, which means a bigger γ corresponding to more robust but less efficient estimation. Noted that as $\gamma \to 0$, $D_\gamma(g, f_\theta)$ is a version of the Kullback-Leibler divergence in the limiting case.

Neglecting the terms independent of the unknown parameters, the empirical version[13] of the γ-divergence loss function is obtained by

$$L_2(\boldsymbol{\alpha}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{f(y_i^{(2)} \mid \boldsymbol{Z}_i;\boldsymbol{\alpha})^\gamma}{\left(\int f(y^{(2)} \mid \boldsymbol{Z}_i;\boldsymbol{\alpha})^{(1+\gamma)}\mathrm{d}y^{(2)}\right)^{\gamma/(1+\gamma)}}, \tag{10}$$

where $f(y_i^{(2)} \mid \boldsymbol{Z}_i;\boldsymbol{\alpha}_i)$ is the conditional probability of y_i giving $\boldsymbol{Z}_i$. Following Fujisawa and Eguchi[13], then we can get the γ-loss function

$$L_2(\boldsymbol{\alpha}) = -\frac{1}{n}\left(\frac{1+\gamma}{2\pi\sigma^2}\right)^{\frac{\gamma}{2(1+\gamma)}}\sum_{i=1}^{n}\exp\left(-\frac{\gamma}{2\sigma^2}(y_i^{(2)} - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right). \tag{11}$$

The last term is the Lasso penalty to select variables,

$$P(\boldsymbol{\alpha}) = \lambda_1\sum_{j=1}^{q} \mid \alpha_j \mid, \tag{12}$$

where λ_1 is a tuning parameter for Lasso penalty.

However, it was found that $y_i^{(1)}$ and $y_i^{(2)}$ are not recognizable, we can not solve (3) directly. Therefore, we design a two-step algorithm that can give reasonable solutions, which is described in Section 1.3. More discussion on the identifiability of the objective function and the effectiveness of the algorithm is shown in the final remark of Section 1.3.1.

Regarding the theoretical properties of parameter estimation, there is really no discussion of related issues in this article, and this theoretical nature of parameter estimation presented in this article is indeed a great challenge. Based on the theoretical properties of the estimates obtained by γ-divergence, Fujisawa and Eguchi[13] proved in detail the theoretical properties of the estimates obtained by linear regression based on γ-divergence; Hung et al.[15] applied γ-divergence to logistic regression and proved the consistency and robustness of the obtained estimates; Ren et al.[16] further applied γ-divergence to high-dimensional generalized linear regression and gives the theoretical properties of the estimates. Regarding the MDSP penalty term, it was first proposed by Tang et al.[5] to apply MDSP to heterogeneous linear models to achieve individualized variable selection, they use a weighted least squares loss function and prove the large sample theoretical properties of the estimation. We proposed a robust individualized subgroup analysis method based on the above, and the theoretical properties of this estimation will be the subject of further research in the future.

1.3 Computation

1.3.1 Overall two-step algorithm

To solve the problem that $y_i^{(1)}$ and $y_i^{(2)}$ are not recognizable, we proposed a two-step method. Our idea is to divide the solution of the model (2) into two parts, iterating alternately. First, the initial value of the coefficient $\boldsymbol{\alpha}_{(0)}, \boldsymbol{\beta}_{(0)}$ is obtained by the method in Tang et al.[5]. Next, the two parts of the model are solved alternately. In the first part, we think of $\boldsymbol{\beta}$ as a known, then we obtain a general linear model, which only contains homogeneous coefficients $\boldsymbol{\alpha}$. So the objective function in this step is

$$Q(\boldsymbol{\alpha}) = L_2(\boldsymbol{\alpha}) + P(\boldsymbol{\alpha}), \quad (13)$$

where $L_2(\boldsymbol{\alpha})$ is defined by (11), $P(\boldsymbol{\alpha})$ is defined by (12). Then we get the estimation $\hat{\boldsymbol{\alpha}}$ by solving the γ-divergence loss function with Lasso penalty. The optimization algorithm is described in Section 1.3.2. In the second part, we fix $\boldsymbol{\alpha}$, then we get heterogeneity estimation $\hat{\boldsymbol{\beta}}$ by the following objective function

$$Q(\boldsymbol{\beta},\boldsymbol{\tau}) = L_1(\boldsymbol{\beta}) + S(\boldsymbol{\beta},\boldsymbol{\tau}), \quad (14)$$

where $L_1(\boldsymbol{\beta})$ is defined by (4), and $S(\boldsymbol{\beta},\boldsymbol{\tau})$ is MDSP. The two parts iterate alternately until convergence. The optimization algorithm is described in Section 1.3.3. The detailed procedure is described as follows.

Initialize: Based on ADMM algorithm in Tang et al.[5], the initial values of homogeneous coefficient $\boldsymbol{\alpha}_{(0)}$ and heterogeneous coefficient $\boldsymbol{\beta}_{(0)}$ are solved.

Step 1: We get the value of $\boldsymbol{\alpha}_{(l)}$ from the previous step, the coefficients of homogeneous covariates $\boldsymbol{\alpha}_{(l)}$ are reestimated based on γ-divergence. We reestimate the coefficients of homogeneous covariates $\boldsymbol{\alpha}_{(l+1)}$ by solving (13).

Step 2: We get the value of $\boldsymbol{\beta}_{(l)}$ from the previous step, solve the objective function with multidirectional separation penalty, and reestimate the coefficient of heterogeneous covariate $\boldsymbol{\beta}_{(l+1)}$ by solving (14). Thus, the step 1 and step 2 are repeated until convergence.

Remark 1: We acknowledge that there is a recognizability problem in the design of the underlying true model, and there is indeed a gap between the objective function and the current algorithm, and the solution optimized by the algorithm may not correspond exactly to the objective function, but the result obtained by the algorithm is also a reasonable solution with good numerical results. It is the direction of our future exploration to explore the problem of recognizability of the underlying true model in depth.

Remark 2: To eliminate the problem of recognizability of the current model, another possible model is

$$Q(\boldsymbol{\alpha},\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i \mid \boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\alpha},\boldsymbol{\beta})^{\gamma}}{\left(\int f(y \mid \boldsymbol{Z}_i,\boldsymbol{X}_i;\boldsymbol{\alpha},\boldsymbol{\beta})^{(1+\gamma)}\mathrm{d}y\right)^{\frac{\gamma}{(1+\gamma)}}} + \lambda\left(\sum_{j=1}^{q}\mid\alpha_j\mid + \sum_{i=1}^{p}\mid\beta_i\mid\right). \quad (15)$$

This objective function (15) can be solved by the classical coordinate descent method, which is omitted here. This scheme we also tried, and the numerical results are not good. The numerical results show that both the estimation of the homogeneous coefficient $\boldsymbol{\alpha}$ and the heterogeneous coefficient $\boldsymbol{\beta}$ have poorer results compared to our proposed method. And the MSE of the estimated homogeneity coefficients $\boldsymbol{\alpha}$ for our proposed method is significantly lower than the model (15). The lower value of RI obtained by the model (15) suggests that the direct use of γ-divergence to the whole y may be less suitable in the presence of individual effects. Our intuitive interpretation is that if the γ-divergence is added directly to the whole y, because the whole y contains a portion of individual effects, that portion is not normally distributed at all, and such the y will cause the robustness of the γ-divergence to fail completely. On the other hand, we may regard the heterogeneous data as outliers, we can not judge whether the data is contaminated or caused by heterogeneous variables. So we use γ-divergence loss function in the homogeneous part to greatly reduce the not robust results caused by homogeneous data contamination, and use MDSP in the heterogeneous part to realize the individualized selection of heterogeneous covariates.

1.3.2 CD algorithm

To minimize the (13), we use coordinate descent algorithm to solve this problem. According to (11), the derivative of $Q(\boldsymbol{\alpha})$ to α_j is

$$\begin{aligned}\nabla_{\alpha_j} Q &= \frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_j} \\ &= -\frac{1}{n}\left(\frac{1+\gamma}{2\pi\sigma^2}\right)^{\frac{\gamma}{2(1+\gamma)}}\frac{\gamma}{\sigma^2}\sum_{i=1}^{n}(y_i^{(2)} - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})z_{ij} \\ &\quad \exp\left(-\frac{\gamma}{2\sigma^2}(y_i^{(2)} - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right) + \lambda_1\mathrm{sgn}(\alpha_j),\end{aligned} \quad (16)$$

then we let $Q(\boldsymbol{\alpha})$ derivative of σ^2, and we let it

equals 0

$$\frac{\partial Q(\boldsymbol{\alpha})}{\partial \sigma^2} = -\frac{1}{n}\left(\frac{1+\gamma}{2\pi\sigma^2}\right)^{\frac{\gamma}{2(1+\gamma)}}\frac{\gamma}{2}\left(\frac{1}{\sigma^2}\right)^2 \sum_{i=1}^{n}\left[(y_i^* - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2 - \frac{\sigma^2}{1+\gamma}\right] \exp\left(-\frac{\gamma}{2\sigma^2}(y_i^* - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right) = 0. \quad (17)$$

With fixed γ and λ_1, we calculate the gradient according to (16), select the step length using the armijo criterion, and update $\boldsymbol{\alpha}$ according to the coordinate descent method, then use the bisection method to estimate σ^2.

1.3.3　ADMM algorithm

In order to optimize the objective function (14), the constraint set is introduced and transformed into solving the following constrained optimization problem

$$\min_{\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\tau}} L_\gamma(\boldsymbol{\beta}) + S_{\lambda_2}(\boldsymbol{\beta},\boldsymbol{\tau}) \quad \text{s.t.} \quad \boldsymbol{\beta} = \boldsymbol{\nu}. \quad (18)$$

where $\boldsymbol{\beta}_{np\times 1} = (\beta_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$, $\boldsymbol{\nu}_{np\times 1} = (\nu_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$. The augmented Lagrange multiplier method is used to solve (18) by introducing Lagrange multiplier $\boldsymbol{\Lambda}, \kappa$.

$$\min_{\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\tau}} L_\gamma(\boldsymbol{\beta}) + S_{\lambda_n}(\boldsymbol{\beta},\boldsymbol{\tau}) \quad \text{s.t.} \quad \boldsymbol{\beta} = \boldsymbol{\nu}. \quad (19)$$

$$\boldsymbol{L}_\gamma(\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\tau}) = L_\gamma(\boldsymbol{\beta}) + S_{\lambda_n}(\boldsymbol{\beta},\boldsymbol{\tau}) + \boldsymbol{\Lambda}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\nu}) + \frac{\kappa}{2}\|\boldsymbol{\beta} - \boldsymbol{\nu}\|_2^2, \quad (20)$$

where $\Lambda_{np\times 1} = (\Lambda_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$, κ is a fixed number. Then we use ADMM algorithm to solve the following optimization problem:

$$\boldsymbol{\beta}_{(l+1)} = \operatorname{argmin}_{\boldsymbol{\beta}} L_\gamma(\boldsymbol{\beta}) + \frac{\kappa}{2}\|\boldsymbol{\beta} - \boldsymbol{\nu}_{(l)} + \kappa^{-1}\Lambda^{(l)}\|_2^2, \quad (21)$$

$$\{\boldsymbol{\nu}_{(l+1)}, \boldsymbol{\tau}_{(l+1)}\} = \operatorname{argmin}_{\nu,\tau} S_{\lambda_n}(\boldsymbol{\nu},\boldsymbol{\tau}) + \frac{\kappa}{2}\|\boldsymbol{\beta}_{(l+1)} - \boldsymbol{\nu} + \kappa^{-1}\Lambda^{(l)}\|_2^2, \quad (22)$$

$$\Lambda_{(l+1)} = \Lambda_{(l)} + \kappa(\boldsymbol{\beta}_{(l+1)} - \boldsymbol{\nu}_{(l+1)}). \quad (23)$$

For more details about ADMM algorithm for (14), please refer to Tang et al.[5]. The overall algorithm in our approach is described in Table 1.

Table 1　Algorithm 1

Algorithm 1

Input: Response variable y_i, covariates $\boldsymbol{X}_i, \boldsymbol{Z}_i$.

Output: Estimation of coefficient, $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$.

Initialization: $l = 1$, the threshold $e = 10^{-3}$.

The initial value $\boldsymbol{\alpha}_{(0)}, \boldsymbol{\beta}_{(0)}$ are abtained by ADMM algorithm in[5];

repeat

$l = l + 1$;

Update $\boldsymbol{\alpha}_{(l+1)}$ with CD algorithm via (13);

Update $\boldsymbol{\beta}_{(l+1)}$ with ADMM algorithm via (14);

until $\|\boldsymbol{\beta}_{(l+1)} - \boldsymbol{\beta}_{(l)}\|_2 \leqslant e$, $\|\boldsymbol{\alpha}_{(l+1)} - \boldsymbol{\alpha}_{(l)}\|_2 \leqslant e$.

Regarding the algorithm for the solution, we use a two-step iterative method. Given the initial values, the first step uses the alternating direction method of multipliers (ADMM) to solve an optimization problem containing the MDSP of the optimization problem. In this step, we use the ADMM algorithm borrowed from the method used by Tang et al.[5] in their solution. Boyd et al.[17] proved that the ADMM algorithm can guarantee the convergence of residuals, objective function, and pairwise variables under the assumption of general. Tang et al.[5] show that ADMM algorithm can converge to a stable point when solving the penalty term with MDSP, and in practice, it can converge to a local minimum by iteration. And most of the individuals are insensitive to the initial values except those near the subgroup boundary. Therefore, using the ADMM algorithm to solve this problem can ensure convergence.

In the second step, we use the coordinate descent method to solve a loss function based on γ-divergence. We use this algorithm by drawing from existing papers on γ-divergence-based regression[14,16] which use the CD algorithm in solving such problems. The convergence of the algorithm of coordinate descent is a very general framework, for the main term is the likelihood function are applicable, so its convergence can be guaranteed. Both optimization subproblems are convergent, so it is a natural result that the whole algorithm is convergent. For all of our simulated and real data sets, convergence was successfully achieved within 50 overall iterations (mostly within 20 iterations).

Regarding the efficiency of the overall two-step method, we performed numerical simulations on a computer configured with an Apple M1 (ARM64) chip (total number of cores 8, memory 16 GB), and

the time for 100 repetitions at a given setting was 1 to 2 hours, with an average time of less than 1 minute for one calculation. The complexity of the algorithm is $O(n^3)$, and the relatively high complexity of the algorithm is due to the high complexity of the ADMM algorithm used in solving the objective function containing the MDSP. However, in general, the time required to solve is shorter and the algorithm is more efficient.

1.3.4 Tuning parameter

In this article, we should tune three parameters $\gamma, \lambda_1, \lambda_2$. It is worth noting that the first tuning parameter γ balance the estimation efficiency and robustness. However, there is no consistent way to select γ. The second parameter λ_1 is Lasso penalty parameter, and it can continuously shrink coeffients toward zero, which really improves prediction ability via the bias variance trade-off. The last parameter λ_2 is MDSP parameter, which control the individual variables selection. Bayes Information Criterion (BIC) criteria are able to identify the true model consistently. Here, we use BIC criteria to select optimal parameters $\lambda_1, \lambda_2, \gamma$.

Regarding the parameter γ, according to the definition of γ-divergence, and studies related to γ-divergence, it is shown that the value of γ balances the robustness of the model. However, regarding the selection of parameter γ, Basu et al.[7] pointed out that there is no consistent best way to choose a suitable parameter γ. Therefore, there are many ways to choose γ. The first method is to use cross-validation together with other parameters to be optimized to select[18]. The second method does not select the parameters based on the data, but artificially gives the γ value directly[14]. The third approach is to use specified rules (e.g. BIC criterion) for parameter selection together with other parameters that need to be optimized[11,15-16,19]. Since cross-validation is slow and relying on artificially given parameter values without data-based selection is unreliable, we draw on the existing literature[11,16] to use the BIC criterion for parameter γ selection.

2 Simulation

2.1 Model for simulation

We used the simulation model given by

$$y_i = x_{i1}\beta_{i1} + x_{i2}\beta_{i2} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + e,$$
$$e \sim N(0,1), i = 1,\cdots,n.$$

We set the sample size $n = 180,120$, and the number of homogeneous explanatory variables $q = 20,40$, respectively. The true coefficients were give by

$$\boldsymbol{\beta}_1 = (\theta_1,\cdots,\theta_1,\cdots,\theta_1,0,\cdots,0),$$
$$\boldsymbol{\beta}_2 = (0,\cdots,0,\theta_2,\cdots,\theta_2,\cdots,\theta_2),$$
$$\boldsymbol{\alpha} = (\theta_3,\cdots,\theta_3,0,\cdots,0)^{\mathrm{T}},$$

where $\theta_1 = 2, \theta_2 = -1, \theta_3 = 1$. The number of nonzero elements θ_1 in $\boldsymbol{\beta}_1$ is the same as the number of θ_2 in $\boldsymbol{\beta}_2$, which is $2n/3$. And the number of θ_3 is 10.

The variables are generated by two ways. Let $\boldsymbol{X} = (\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)^{\mathrm{T}}, \boldsymbol{Z} = (\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n)^{\mathrm{T}}$. The first way was heterogeneity explanatory variables $\boldsymbol{X}$ and homogeneous explanatory variables $\boldsymbol{Z}$ are independent. And $\boldsymbol{X}$ are generated from a normal distribution $N(0, \Sigma_1)$ with $\Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$. The homogeneous explanatory variable $\boldsymbol{Z}$ are also generated from a normal distribution $N(0, \Sigma_2)$. And we consider two structures of covariance matrix Σ_2. The first structure is autoregressive correlation (AR), which is given by $\Sigma_2 = (\rho^{|i-j|})_{1\leqslant i,j\leqslant p}$ with $\rho = 0.2$. The banded correlation is the second structure. We consider $\sigma_{ij} = 0.5$, if $|i-j| = 1$, and 0 otherwise. The second way to generate variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are dependent. They are generated from a multivariate normal distribution with mean 0 and covariance $R(\rho)$, where $R(\rho)$ is the correlation matrix with AR structure like Σ_2, and $\rho = 0.2$.

Outliers in homogeneous variables were incorporated into simulations. We investigated two outlier ratios ($r = 0.1$ and 0.3). The outliers are generated from $N(\boldsymbol{\mu}, c\Sigma_2)$, where $\boldsymbol{\mu} = (1,\cdots,1)^{\mathrm{T}}, c = 2$.

2.2 Performance measure

The mean squared error (MSE) were examined

to verify the predictive performance and fitness of regression coefficient:

$$\mathrm{MSE}(\boldsymbol{\alpha}) = \frac{1}{q}\sum_{j=1}^{q}(\alpha_j^* - \hat{\alpha}_j)^2,$$

$$\mathrm{MSE}(\boldsymbol{\beta}) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\beta_{ij}^* - \hat{\beta}_{ij})^2,$$

where $(x_i^*, z_i^*, y_i^*)(i = 1, 2, \cdots, n)$ is the test sample generated from the simulation model without outliers, $\boldsymbol{\beta}_{ij}^*$ and α_j^* are the true coefficients. The true positive rate (TPR) and true negative rate (TNR) of coefficients were:

$$\mathrm{TPR}(\hat{\boldsymbol{\alpha}}) = \frac{|\{j \in \{1, \cdots, q\}: \hat{\alpha}_j \neq 0 \wedge \alpha_j^* \neq 0\}|}{|\{j \in \{1, \cdots, q\}: \alpha_j^* \neq 0\}|},$$

$$\mathrm{TNR}(\hat{\boldsymbol{\alpha}}) = \frac{|\{j \in \{1, \cdots, q\}: \hat{\alpha}_j = 0 \wedge \alpha_j^* = 0\}|}{|\{j \in \{1, \cdots, q\}: \alpha_j^* = 0\}|},$$

$$\mathrm{TPR}(\hat{\boldsymbol{\beta}}) = \frac{|\{i \in \{1, \cdots, n\}, j \in \{1, \cdots, p\}: \hat{\beta}_{ij} \neq 0 \wedge \beta_{ij}^* \neq 0\}|}{|\{i \in \{1, \cdots, n\}, j \in \{1, \cdots, p\}: \beta_{ij}^* \neq 0\}|},$$

$$\mathrm{TNR}(\hat{\boldsymbol{\beta}}) = \frac{|\{i \in \{1, \cdots, n\}, j \in \{1, \cdots, p\}: \hat{\beta}_{ij} = 0 \wedge \beta_{ij}^* = 0\}|}{|\{i \in \{1, \cdots, n\}, j \in \{1, \cdots, p\}: \beta_{ij}^* = 0\}|}.$$

The rand index (RI) is a measure of the similarity between two data clusterings. It is definated by

$$\mathrm{RI} = 1 - \frac{I(M_1) - I(M_2)}{C_n^2},$$

where function $I(M)$ is the number of upper triangular nonzero elements in matrix M. $M_s(s = 1, 2)$ represents class matrix, its element $a_{ij} = 1$ means subject i and j are in the same group, otherwise $a_{ij} = 0$. If subject i and subject j have completely identical heterogeneity coefficients, we consider they belong to the same class. M_1 is real class matrix, M_2 is estimate class matrix.

2.3 Comparative method

We compare the performance of our proposed method γ-MDSP with five subgrouping based variable selection approaches: 1) the original MDSP method; 2) the pairwise fused Lasso with an Lasso penalty (FLPa): $\lambda_1 \sum_{i=1}^{n} \|\beta_i\|_1 + \lambda_2 \sum_{k=1}^{p} \sum_{i<j} |\beta_{ik} - \beta_{jk}|$, was sovled by R package penalized (version 0.9-50); 3) fusion and feature selection with a truncated Lasso penalty (FTLP): $\lambda_1 \sum_{k=1}^{p} \sum_{i=1}^{n} J_\tau(|\beta_{ik}|) + \lambda_2 \sum_{k=1}^{p} \sum_{i<j} J_\tau(|\beta_{ik} - \beta_{jk}|)$, where $J_\tau(a) = \min(\frac{a}{\tau}, 1)$, was implemented by R package penalized (version 0.9-50).

In addition, we also focus on the following methods about subgrouping and clustering aspect: 1) clustering based on response variables with a Lasso penalty (CVL). First we cluster based on the response variables, and then we use Lasso penalty on each cluster to realize variables selection; 2) clustering based on residual with an Lasso penalty (CRL). First, under the homogeneity assumption $y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i$, we use Lasso penalty to select variables, then we cluster them based on residuals $y_i - \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\alpha}}$, and we also apply Lasso penalty on each cluster.

We also compare the γ-divergence with some other robust loss functions: 1) least absolute deviation (LAD) solved by R package L1pack (version 0.38.196); 2) Huber function solved by R package MASS (version 7.3-55).

Moreover, we used MDSP based on known true data and true important homogeneous variables in the first estimator (Oracle1). In the second estimator (Oracle2), we know true important homogeneous variables and original MDSP method is used.

In our method, we need an initial point to obtain the estimate, and in this experiment, we used the estimate of MDPS as an initial point. The tuning parameter γ, MDSP penalty λ_1 and Lasso penalty λ_2 are selected via line search to minimize Bayesian information criterion.

2.4 Results

Table 2 is the results that heterogeneous variable $\boldsymbol{X}$ is independent of homogeneous variable $\boldsymbol{Z}$. Table 2 shows the results of that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are dependent. In the main text we only show Table 2 and Table 3. And these tables provide the MSE, TPR, and TNR of coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the last column displays the RI for all methods.

Table 2 Independent X and Z, $q = 20, r = 0.1$

Correlation	n	Method	MSE(α)	MSE(β)	TPR(α)	TPR(β)	TNR(α)	TNR(β)	RI
AR	180	γ-MDSP	0.029(0.030)	1.531(0.692)	1.000(0.000)	0.419(0.059)	0.610(0.189)	0.637(0.089)	0.661(0.016)
		MDSP	0.463(0.080)	6.719(2.046)	0.472(0.129)	0.470(0.050)	0.518(0.142)	0.545(0.055)	0.643(0.009)
		FLPa	0.455(0.349)	2.073(0.464)	0.550(0.164)	0.531(0.028)	0.510(0.139)	0.499(0.046)	0.648(0.005)
		FTLP	0.428(0.060)	2.011(0.174)	0.508(0.132)	0.536(0.032)	0.508(0.138)	0.504(0.050)	0.649(0.006)
		CVL	0.494(0.012)	1.612(0.090)	0.050(0.066)	0.097(0.115)	0.963(0.052)	0.909(0.110)	0.33(0.007)
		CRL	0.521(0.079)	1.629(0.649)	0.089(0.089)	0.258(0.124)	0.891(0.094)	0.750(0.131)	0.322(0.066)
		LAD	0.222(0.187)	12.070(3.258)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.647(0.000)
		Huber	0.481(0.162)	11.787(3.001)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.647(0.000)
		Oracle1	0.019(0.008)	1.239(0.127)	1.000(0.000)	0.402(0.075)	0.516(0.125)	1.000(0.000)	0.665(0.025)
		Oracle2	0.442(0.092)	14.604(5.662)	0.480(0.131)	0.477(0.047)	0.504(0.107)	1.000(0.000)	0.619(0.016)
	120	γ-MDSP	0.135(0.182)	1.494(0.443)	0.983(0.065)	0.382(0.084)	0.587(0.172)	0.662(0.095)	0.657(0.018)
		MDSP	1.278(1.170)	8.035(2.267)	0.514(0.090)	0.474(0.053)	0.504(0.086)	0.559(0.067)	0.642(0.011)
		FLPa	0.869(0.587)	3.241(1.157)	0.500(0.093)	0.522(0.031)	0.533(0.092)	0.483(0.045)	0.647(0.007)
		FTLP	0.839(0.452)	3.193(0.995)	0.488(0.098)	0.530(0.033)	0.492(0.090)	0.491(0.045)	0.648(0.006)
		CVL	0.496(0.011)	1.601(0.091)	0.038(0.047)	0.107(0.150)	0.953(0.059)	0.893(0.151)	0.330(0.007)
		CRL	0.505(0.050)	1.938(2.418)	0.077(0.075)	0.175(0.131)	0.909(0.078)	0.831(0.125)	0.334(0.038)
		LAD	0.877(0.192)	10.120(2.909)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Huber	0.949(0.208)	10.143(3.197)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Oracle1	0.140(0.137)	1.281(0.154)	0.986(0.045)	0.391(0.074)	0.504(0.070)	1.000(0.000)	0.661(0.026)
		Oracle2	1.262(0.996)	19.692(8.350)	0.498(0.074)	0.473(0.052)	0.508(0.080)	1.000(0.000)	0.615(0.016)
Band	180	γ-MDSP	0.027(0.041)	1.419(1.213)	0.995(0.022)	0.402(0.083)	0.635(0.179)	0.665(0.064)	0.664(0.02)
		MDSP	0.133(0.056)	3.360(1.084)	0.962(0.070)	0.465(0.054)	0.718(0.134)	0.558(0.076)	0.643(0.015)
		FLPa	0.319(0.353)	1.916(0.583)	0.870(0.167)	0.524(0.040)	0.710(0.149)	0.500(0.026)	0.649(0.009)
		FTLP	0.141(0.062)	1.626(0.143)	0.939(0.096)	0.546(0.031)	0.753(0.124)	0.501(0.051)	0.654(0.007)
		CVL	0.434(0.053)	1.489(0.155)	0.294(0.183)	0.218(0.149)	0.928(0.081)	0.788(0.147)	0.330(0.007)
		CRL	0.313(0.067)	8.130(38.451)	0.542(0.139)	0.303(0.135)	0.900(0.066)	0.707(0.130)	0.355(0.042)
		LAD	0.284(0.259)	15.177(3.348)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.647(0.000)
		Huber	0.123(0.415)	14.924(3.741)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.647(0.000)
		Oracle1	0.019(0.008)	1.307(0.135)	1.000(0.000)	0.380(0.072)	0.498(0.138)	1.000(0.000)	0.667(0.025)
		Oracle2	0.170(0.097)	7.397(3.195)	0.914(0.109)	0.445(0.064)	0.712(0.148)	1.000(0.000)	0.615(0.022)
	120	γ-MDSP	0.069(0.067)	1.467(0.241)	1.000(0.000)	0.402(0.084)	0.550(0.181)	0.654(0.084)	0.669(0.019)
		MDSP	0.668(0.679)	4.392(1.773)	0.760(0.160)	0.469(0.073)	0.552(0.089)	0.563(0.070)	0.643(0.015)
		FLPa	0.428(0.426)	1.949(0.472)	0.785(0.164)	0.538(0.031)	0.595(0.109)	0.504(0.039)	0.652(0.005)
		FTLP	0.446(0.387)	2.033(0.587)	0.811(0.143)	0.542(0.036)	0.595(0.107)	0.504(0.050)	0.652(0.007)
		CVL	0.418(0.055)	1.489(0.130)	0.366(0.142)	0.225(0.144)	0.938(0.046)	0.771(0.153)	0.330(0.006)
		CRL	0.374(0.156)	2.046(3.146)	0.537(0.126)	0.362(0.156)	0.896(0.083)	0.637(0.156)	0.330(0.035)
		LAD	0.947(0.853)	13.395(4.185)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.649(0.000)
		Huber	0.538(0.911)	13.357(3.510)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Oracle1	0.159(0.169)	1.312(0.149)	0.980(0.076)	0.367(0.058)	0.532(0.082)	1.000(0.000)	0.665(0.021)
		Oracle2	1.061(1.169)	9.961(5.181)	0.680(0.148)	0.459(0.056)	0.578(0.097)	1.000(0.000)	0.618(0.019)

Table 3　Dependent X and Z, $q=20, r=0.1$

Correlation	n	Method	MSE($\boldsymbol{\alpha}$)	MSE($\boldsymbol{\beta}$)	TPR($\boldsymbol{\alpha}$)	TPR($\boldsymbol{\beta}$)	TNR($\boldsymbol{\alpha}$)	TNR($\boldsymbol{\beta}$)	RI
AR	180	γ-MDSP	0.045(0.043)	1.165(1.675)	1.000(0.000)	0.414(0.086)	0.617(0.191)	0.646(0.092)	0.655(0.021)
		MDSP	0.463(0.111)	5.541(1.765)	0.476(0.132)	0.483(0.065)	0.492(0.114)	0.547(0.067)	0.647(0.012)
		FLPa	0.501(0.245)	2.457(0.456)	0.567(0.197)	0.534(0.050)	0.553(0.141)	0.491(0.048)	0.651(0.009)
		FTLP	0.455(0.087)	2.463(0.394)	0.494(0.138)	0.538(0.043)	0.474(0.134)	0.498(0.056)	0.652(0.008)
		CVL	0.492(0.012)	1.581(0.105)	0.050(0.058)	0.115(0.131)	0.956(0.046)	0.883(0.139)	0.326(0.011)
		CRL	0.495(0.029)	1.444(0.210)	0.098(0.067)	0.208(0.090)	0.898(0.072)	0.788(0.106)	0.333(0.056)
		LAD	0.213(0.079)	5.927(1.794)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.639(0.000)
		Huber	0.203(0.102)	5.728(1.780)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.637(0.000)
		Oracle1	0.034(0.016)	1.190(0.118)	1.000(0.000)	0.367(0.077)	0.502(0.145)	1.000(0.000)	0.658(0.027)
		Oracle2	0.484(0.120)	11.109(4.246)	0.496(0.135)	0.485(0.054)	0.502(0.125)	1.000(0.000)	0.622(0.020)
	120	γ-MDSP	0.184(0.101)	1.536(4.174)	0.930(0.166)	0.375(0.110)	0.540(0.179)	0.657(0.101)	0.652(0.053)
		MDSP	1.661(1.236)	7.469(2.745)	0.518(0.085)	0.468(0.056)	0.510(0.105)	0.562(0.075)	0.643(0.013)
		FLPa	0.830(0.507)	3.972(1.917)	0.554(0.137)	0.515(0.042)	0.524(0.106)	0.485(0.060)	0.649(0.008)
		FTLP	0.861(0.517)	4.174(1.656)	0.500(0.090)	0.530(0.044)	0.504(0.089)	0.492(0.057)	0.650(0.008)
		CVL	0.528(0.208)	1.598(0.106)	0.044(0.065)	0.093(0.123)	0.948(0.066)	0.906(0.123)	0.327(0.009)
		CRL	0.497(0.029)	3.568(10.567)	0.066(0.058)	0.208(0.142)	0.941(0.055)	0.794(0.152)	0.324(0.033)
		LAD	0.266(0.115)	4.365(1.333)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Huber	0.291(0.155)	5.078(1.555)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Oracle1	0.279(0.331)	1.308(0.167)	0.958(0.086)	0.363(0.078)	0.508(0.078)	1.000(0.000)	0.657(0.030)
		Oracle2	2.222(2.160)	16.165(6.557)	0.476(0.102)	0.473(0.064)	0.488(0.075)	1.000(0.000)	0.614(0.021)
Band	180	γ-MDSP	0.029(0.009)	1.521(0.445)	1.000(0.000)	0.402(0.096)	0.605(0.190)	0.632(0.103)	0.658(0.023)
		MDSP	0.175(0.089)	3.197(1.066)	0.914(0.107)	0.456(0.071)	0.656(0.123)	0.567(0.092)	0.641(0.021)
		FLPa	0.216(0.166)	2.005(0.637)	0.886(0.120)	0.544(0.039)	0.708(0.114)	0.494(0.068)	0.655(0.006)
		FTLP	0.158(0.072)	1.782(0.222)	0.914(0.103)	0.547(0.036)	0.719(0.141)	0.501(0.053)	0.656(0.007)
		CVL	0.449(0.039)	1.568(0.135)	0.226(0.150)	0.177(0.152)	0.927(0.065)	0.824(0.157)	0.326(0.011)
		CRL	0.332(0.062)	2.639(7.963)	0.498(0.124)	0.288(0.146)	0.893(0.072)	0.707(0.148)	0.351(0.059)
		LAD	0.185(0.281)	6.718(1.853)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.644(0.000)
		Huber	0.383(0.754)	7.367(2.304)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.647(0.000)
		Oracle1	0.035(0.018)	1.301(0.187)	1.000(0.000)	0.386(0.089)	0.492(0.154)	1.000(0.000)	0.654(0.034)
		Oracle2	0.184(0.079)	6.740(4.270)	0.906(0.113)	0.446(0.070)	0.664(0.121)	1.000(0.000)	0.615(0.024)
	120	γ-MDSP	0.233(0.181)	1.445(1.741)	0.905(0.173)	0.392(0.094)	0.580(0.182)	0.667(0.124)	0.656(0.028)
		MDSP	0.781(0.601)	3.878(2.232)	0.742(0.150)	0.472(0.070)	0.584(0.106)	0.570(0.094)	0.642(0.020)
		FLPa	0.554(0.571)	2.797(1.355)	0.746(0.149)	0.530(0.041)	0.602(0.119)	0.510(0.062)	0.650(0.008)
		FTLP	0.379(0.260)	2.195(0.641)	0.809(0.148)	0.548(0.037)	0.616(0.139)	0.499(0.048)	0.655(0.007)
		CVL	0.435(0.038)	1.564(0.130)	0.245(0.121)	0.160(0.156)	0.942(0.058)	0.841(0.152)	0.328(0.010)
		CRL	0.395(0.207)	1.498(0.917)	0.485(0.125)	0.286(0.148)	0.898(0.075)	0.719(0.138)	0.332(0.040)
		LAD	0.740(0.391)	5.657(1.849)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.649(0.000)
		Huber	0.392(0.291)	5.790(2.057)	0.000(0.000)	0.000(0.000)	1.000(0.000)	1.000(0.000)	0.642(0.000)
		Oracle1	0.161(0.159)	1.301(0.168)	0.978(0.058)	0.368(0.083)	0.512(0.096)	1.000(0.000)	0.658(0.031)
		Oracle2	0.864(0.779)	7.133(4.663)	0.726(0.154)	0.450(0.070)	0.562(0.116)	1.000(0.000)	0.614(0.024)

The proposed method has the smallest MSE of $\boldsymbol{\alpha}$ in all settings besides Oracle1. The MSE ($\boldsymbol{\alpha}$) of our method are significantly improved compared to other methods. This is because our method introduce the $\gamma-$divergence to increase the model robustness. Also, our method has the smallest MSE ($\boldsymbol{\beta}$) besides Oracle1, which has a significant improvement on original MDSP approach. This is because we get more precise $\hat{\boldsymbol{\alpha}}$ and we reduce errors of $\hat{\boldsymbol{\beta}}$ by iteration.

In addition, the results show that our approach has the biggest TPR ($\boldsymbol{\alpha}$) in all methods besides Oracle1. As for RI, we can find that our method has

almost the biggest value in all settings. Although our method has lower TNR than CVL and CRL, CVL and CRL have the worest performance on TPR, which makes the worst result of RI in all approaches. We should notice that the results of TNR ($\boldsymbol{\alpha}$) of all methods are not too high, and this is a challenge for above approaches.

In general, the proposed method has the closest values to Oracle1. Our method is robust against the contamination data and misspecification of subgroup numbers in terms of the consistently smallest MSE and highest TPR among all approaches.

3 Real data application

In this section, we apply the proposed robust individualized subgroup analysis method to the data of SKCM. Skin cutaneous melanoma is one of the highly lethal and aggressive skin diseases. The identification of biomarkers is of great importance for clinical treatment. But handling high-dimensional data analysis and identifying potential genes on the dataset is challenging Zhang et al.[20]. The data includes 342 individuals, the response variable is Breslow's depth, and the covariants includes 9 environmental effects and 20189 mRNA expression. The response variable Breslow's depth in the dataset is considered to be a prognostic factor for melanoma in medicine, that is, the greater the Breslow depth, the lower the survival rate. The heterogeneous effect of environmental factors on individual skin melanoma has been mentioned many times in the field of biomedicine[21]. Therefore, it is reasonable to take environmental variables as heterogeneous variables in genetic and environmental data.

Consider a linear model with heterogeneous variables, $y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i$, denote y_i as the Breslow's depth (log-transformed) of ith-individual, $\boldsymbol{X}_i$ as the environmental variables of ith-individual, and $\boldsymbol{Z}_i$ as the mRNA expression. We removed covariates with missing value rates greater than 0.15 and removed other individuals with missing values. The final remaining 294 individuals with 7 environmental covariates. Meanwhile, we used prescreening for the selection of genetic variables, resulting in 50 remaining mRNA variables. To identify environmental factors significantly associated with melanoma in each individual, we used 7 environmental covariates as heterogeneous potential predictors and added genetic covariates as homogeneous variables.

The gene coefficients are shown in the Table 4 and the results indicate that 11 of the 50 genes are significantly associated with melanoma, and the selected genes have been shown in the available literature to be highly correlated with skin cutaneous melanoma. Lambert[22] mentioned the gene CCNK as one of the 20 most significantly expressed genes in squamous cell carcinoma of the skin, and also CCNK is a gene involved in DNA repair and has a better role in patient prognosis[23]. CNOT11, one of the eight subunits of the mammalian CCR4 – NOT complex, plays a dual role in the control of tumor progression[24]. In characterizing the transcriptome of Spitzoid neoplasms cutaneous melanocytic proliferations with digital mRNA expression profiles, EIF2B4 is one of the most informative genes[25]. In addition, LSM12 was detected more frequently in male carriers containing preferred partner BRCA1 fusion mutations, and cutaneous melanoma were the frequent tumors demonstrating BRCAm in males[26]. Also, PSMD9 showed significant upregulated expression in primary tumors[27]. Also, the TRIAP1 gene is involved in cell cycle regulation and cellular stress response, regulating P53-mediated apoptosis or cell death in response to stresses such as UV irradiation or DNA damage[28].

The three heterogeneous variables AJCC NODES PATHOLOGIC PN, AJCC TUMOR PATHOLOGIC PT, and GENDER were selected for analysis, and subgroups were classified according to whether the coefficients of the heterogeneous variables were zero or not. The 294 individuals were classified into a total of four subgroups, and the values of the non-zero coefficients are shown in Table 5. The four subgroups classified were tested for differences with the rest of the variables and the results showed significant differences in AJCC METASTASIS PATHOLOGIC PM, CLARK LEVEL

Table 4 Coefficients of gene

Gene	Coefficient	Gene	Coefficient	Gene	Coefficient	Gene	Coefficient
AAMP		EIF1		OXA1L		SEC11A	
ANAPC5		**EIF2B4**	0.002 312	**PIGH**	0.002 881	SPPL3	
ANKLE2		EIF3I		PITHD1		**SPRYD7**	0.002 098
CAMTA2	0.001 525	FOXK2		PITPNA		SRRT	
CCNK	0.001 122	GORASP2		**PSMD9**	0.002 817	STK16	
CCT7		**IAH1**	0.001 721	RABGEF1		TMED2	
CERS5		LRRC47		RBM4		**TMEM120B**	0.005 418
CLK3		**LSM12**	0.001 624	RER1		TMEM248	
CNOT11	0.001 021	MAPK1IP1L		RNF10		**TRIAP1**	0.003 136
COX7A2L		MBD1		RNF34		TTC4	
CSNK1D		MLX		SDF2		ZDHHC5	
DDX23		NAT10		SDHAF2		ZNF207	
DIABLO		NIPA2					

Table 5 Information about subgroups

Subgroup	1	2	3	4
Number of subgroup	156	90	11	37
Coefficients				
AJCC NODES PATHOLOGIC PN	0.000 0	0.083 9	0.083 9	0.083 9
AJCC TUMOR PATHOLOGIC PT	0.293 9	0.000 0	0.000 0	0.293 9
GENDER	0.000 0	0.001 0	0.000 0	0.000 0
Significance test	*p*-value			
AJCC METASTASIS PATHOLOGIC PM	0.005 3			
CLARK LEVEL AT DIAGNOSIS	3.94E−09			
SAMPLE TYPE	7.66E−06			

AT DIAGNOSIS, and SAMPLE TYPE. This suggests that environmental variables have a strong heterogeneous effect on melanoma patients, so that the risk of melanoma is diverse at the individual level and other environmental factors. Therefore, in the treatment and prevention of melanoma, individualized treatment plans should be carried out at the individual level.

4 Discussion

In this paper, we consider a robust individualized linear model based on γ-divergence. For heterogeneous variables, we use MDSP penalty term to realize individualized penalty. For contaminated homogeneous variables, we use gamma divergence to obtain robust estimation of anti pollution.

We further propose a two-step iterative method to calculate the above process. In addition, we have carried out numerical simulation, and the results show that our method has better effect. In addition, the analysis of SKCM data shows that this method has strong applicability. And in the future work, it can be extended to the generalized linear model to make it more generalized.

References

[1] Hocking T D, Joulin A, Bach F, et al. Clusterpath an algorithm for clustering using convex fusion penalties[C/OL]//Proceedings of the 28th International Conference on Machine Learning. June 28-July 2, 2011, Bellevue, Washington, USA. ICML, 2011:1. https://hal.archives-ouvertes.fr/hal-00591630.

[2] Lindsten F, Ohlsson H, Ljung L. Clustering using sum-of-norms regularization: with application to particle filter output computation [C]//2011 IEEE Statistical Signal Processing Workshop. June 28-30, 2011, Nice, France. IEEE, 2011: 201-204. DOI:10.1109/SSP.2011.5967659.

[3] Pan W, Shen X T, Liu B H. Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty [J]. Journal of Machine Learning Research, 2013, 14(1): 1865. https://pubmed.ncbi.nlm.nih.gov/24358018.

[4] Ma S J, Huang J. A concave pairwise fusion approach to subgroup analysis [J]. Journal of the American Statistical Association, 2017, 112(517): 410-423. DOI: 10.1080/01621459.2016.1148039.

[5] Tang X W, Xue F, Qu A. Individualized multidirectional

variable selection[J]. Journal of the American Statistical Association, 2021, 116(535): 1280-1296. DOI: 10.1080/01621459.2019.1705308.

[6] Bhalla S, Kaur H, Dhall A, et al. Prediction and analysis of skin cancer progression using genomics profiles of patients [J]. Scientific Reports, 2019, 9(1): 1-16. DOI:10.1038/s41598-019-52134-4.

[7] Basu A, Harris I R, Hjort N L, et al. Robust and efficient estimation by minimising a density power divergence[J]. Biometrika, 1998, 85(3): 549-559. DOI:10.1093/biomet/85.3.549.

[8] Fujisawa H, Eguchi S. Robust estimation in the normal mixture model [J]. Journal of Statistical Planning and Inference, 2006, 136(11): 3989-4011. DOI: 10.1016/j.jspi.2005.03.008.

[9] Ghosh A, Basu A. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression[J]. Electronic Journal of Statistics, 2013, 7: 2420-2456. DOI:10.1214/13-EJS847.

[10] Durio A, Isaia E D. The minimum density power divergence approach in building robust regression models [J]. Informatica, 2011, 22(1): 43-56. DOI: 10.15388/Informatica.2011.313.

[11] Zang Y G, Zhao Q, Zhang Q Z, et al. Inferring gene regulatory relationships with a high-dimensional robust approach[J]. Genetic Epidemiology, 2017, 41(5): 437-454. DOI:10.1002/gepi.22047.

[12] Jones M C, Hjort N L, Harris I R, et al. A comparison of related density-based minimum divergence estimators[J]. Biometrika, 2001, 88(3): 865-873. DOI:10.1093/biomet/88.3.865.

[13] Fujisawa H, Eguchi S. Robust parameter estimation with a small bias against heavy contamination [J]. Journal of Multivariate Analysis, 2008, 99(9): 2053-2081. DOI: 10.1016/j.jmva.2008.02.004.

[14] Kawashima T, Fujisawa H. Robust and sparse regression via γ-divergence[J]. Entropy, 2017, 19(11): 608. DOI: 10.3390/e19110608.

[15] Hung H, Jou Z Y, Huang S Y. Robust mislabel logistic regression without modeling mislabel probabilities [J]. Biometrics, 2018, 74(1): 145-154. DOI:10.1111/biom.12726.

[16] Ren M Y, Zhang S G, Zhang Q Z. Robust high-dimensional regression for data with anomalous responses[J]. Annals of the Institute of Statistical Mathematics, 2021, 73(4): 703-736. DOI:10.1007/s10463-020-00764-1.

[17] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1):1-122. DOI:10.1561/2200000016.

[18] Smith S A, O'Meara B C. treePL: divergence time estimation using penalized likelihood for large phylogenies [J]. Bioinformatics, 2012, 28(20): 2689-2690. DOI: 10.1093/bioinformatics/bts492.

[19] Mollah M N H, Eguchi S, Minami M. Robust prewhitening for ICA by minimizing β-divergence and its application to FastICA[J]. Neural Processing Letters, 2007, 25(2): 91-110. DOI:10.1007/s11063-006-9023-8.

[20] Zhang L, Wang Q, Wang L J, et al. OSskcm: an online survival analysis webserver for skin cutaneous melanoma based on 1085 transcriptomic profiles [J]. Cancer Cell International, 2020, 20: 176. DOI: 10.1186/s12935-020-01262-3.

[21] Volkovova K, Bilanicova D, Bartonova A, et al. Associations between environmental factors and incidence of cutaneous melanoma. Review[J]. Environmental Health: A Global Access Science Source, 2012, 11(Suppl 1): S12. DOI: 10.1186/1476-069X-11-S1-S12.

[22] Lambert S R. Molecular profiling of cutaneous squamous cell carcinoma[D]. London, Queen Mary University of London, 2010. https://qmro.qmul.ac.uk/xmlui/handle/123456789/564.

[23] Ramazzotti D, Lal A, Wang B, et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival[J]. Nature Communications, 2018, 9: 4453. DOI: 10.1038/s41467-018-06921-8.

[24] Bartlam M, Yamamoto T. The structural basis for deadenylation by the CCR4-NOT complex[J]. Protein & Cell, 2010, 1(5): 443-452. DOI: 10.1007/s13238-010-0060-8.

[25] Hillen L M, Geybels M S, Spassova I, et al. A digital mRNA expression signature to classify challenging spitzoid melanocytic neoplasms[J]. FEBS Open Bio, 2020, 10(7): 1326-1341. DOI:10.1002/2211-5463.12897.

[26] Sun P, Li Y, Chao X, et al. Clinical characteristics and prognostic implications of BRCA-associated tumors in males: a pan-tumor survey[J]. BMC Cancer, 2020, 20(1): 994. DOI:10.1186/s12885-020-07481-1.

[27] Miñoza J M A, Rico J A, Zamora P R F, et al. Biomarker discovery for meta-classification of melanoma metastatic progression using transfer learning[EB/OL]. (2021-05-27)[2022-04-08]. https://www.preprints.org/manuscript/202105.0670/v1.

[28] López S, Smith-Zubiaga I, García de Galdeano A, et al. Comparison of the transcriptional profiles of melanocytes from dark and light skinned individuals under basal conditions and following ultraviolet-B irradiation[J]. PLoS One, 2015, 10(8): e0134911. DOI: 10.1371/journal.pone.0134911.