文章编号:2095-6134(2022)03-0289-13

Robust ordinal mislabel logistic regression based on γ-divergence*

GUO Meijun[1,2], REN Mingyang[1,2], LI Shiming[3], ZHANG Sanguo[1,2†]

(1 *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049, *China*;
2 *Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing* 100049, *China*;
3 *Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology & Visual Science Key Laboratory, Beijing Institute of Ophthalmology, Capital Medical University, Beijing* 100730, *China*)

(Received 3 January 2020; Revised 13 April 2020)

Guo M J, Ren M Y, Li S M, et al. Robust ordinal mislabel logistic regression based on γ-divergence[J]. Journal of University of Chinese Academy of Sciences, 2022, 39(3): 289-301. DOI: 10.7523/j.ucas.2020.0056.

Abstract Ordinal multi-classification methods have been studied widely. Traditional ordinal multi-classification methods assume that the sample label is not mislabeled. Due to the complexity of the real data and the limited artificial experience, it is unrealistic to obtain completely accurate labels, in which conventional methods perform poorly. In this article, we propose an ordinal mislabel logistic regression method based on γ-divergence, which possessing strong robustness when dealing with ordinal mislabeled response data. That is to say, when mislabeled, the weight of the sample in parameter estimation equation diminish compared to the case that the sample is properly labeled. Our method not only possesses the robustness but also can ignore the mislabel probabilities in the model. We construct the model by minimizing γ-divergence estimation and solve the model by gradient descent algorithm. Both simulation studies and real data analysis demonstrate that the method, namely robust ordinal mislabel logistic regression, is efficient to analyze ordinal mislabeled response data.

Keywords γ-divergence; logistic regression; mislabeled response; ordinal classification; robustness
CLC number: O212 **Document code**: A **DOI**: 10.7523/j.ucas.2020.0056

基于 γ-散度的稳健有序误标记 logistic 回归

郭美君[1,2],任明旸[1,2],李仕明[3],张三国[1,2]

(1 中国科学院大学数学科学学院,北京 100049;2 中国科学院大数据挖掘与知识管理重点实验室,北京 100049;
3 首都医科大学附属北京同仁医院北京同仁眼科中心,北京眼科及视光学重点实验室,北京 100730)

摘　要　有序多分类方法已经得到了广泛研究。传统的有序多分类方法假设样本的类别标签是不存在误标记的。但是由于实际数据复杂以及人工经验有限,获得标记完全正确的样本是

* Supported by University of Chinese Academy of Sciences (Y95401TXX2), Beijing Natural Science Foundation (Z190004), and Key Program of Joint Funds of the National Natural Science Foundation of China (U19B2040)

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

不现实的，因此，传统的方法就存在局限性。提出一种基于 γ-散度的有序误标记 logistic 回归方法，在处理存在误标记的有序多分类问题时具有很强的稳健性，也就是说，当某一样本被错误标记时它对参数估计的权重小于其被正确标记时的权重。该方法通过最小化 γ-散度构建模型，利用梯度下降算法求解模型，不仅具有很强的稳健性而且在模型中可以忽略误标记概率。模拟研究和真实数据分析都说明该有序误标记 logistic 回归方法在处理存在误标记的有序分类问题时效果很好。

关键词 γ-散度；logistic 回归；误标记；有序分类；稳健性

The classification problem of ordinal response data, examples of this including cancer patients grouped in early, mediocre and terminal stages, customers grouped into low, middle and high credit levels, are widely discussed in recent years. Some methods[1-3] have been proposed to solve the problem. A direct method for ordinal response data is to convert ordinal labels to numerical values, such as the conversion of {excellent, good, moderate, poor} to {1,2,3,4}. As a result, a regression method can be applied to the converted dataset. However, not all class labels can be converted to positive integers. Another method is to convert class labels to common numerical value rather than positive integers by a map. Nevertheless, it is difficult to define the mapping function. Frank and Hall[4] proposed a method that converted ordinal classification problem to several binary classification problems. For each $i \in \{0,1,2,3,\cdots,K\}$, each sample is classified between the meta-class including classes 0 to i and the meta-class including classes $i+1$ to K. The final classification result can be inferred from the K binary predictions. Shashua and Levin[5] extended SVM for ordinal regression by finding K ordinal thresholds, namely $b_1 \leqslant b_2 \leqslant \cdots \leqslant b_K$, spliting the real line into K+1 ordinal parts. Wang et al.[6] proposed nonparallel support vector for ordinal regression by constructing a hyperplane in each class. However, these methods perform poorly when the ordinal label is mislabeled.

More and more literatures studied mislabeled response data by means of binary logistic regression[7]. Tian and Sun[8] proposed a new fuzzy set method to detect suspectable mislabeled points, and then delete their labels and construct a semi-supervised model. Logistic regression based on different mislabel probabilities are also widely used to mislabeled unordered response data. Copas considered equal and constant mislabel probabilities[9] and Komori et al.[10] assumed mislabeling occurs only the 0-group. In addition, Hung et al.[7] proposed a robust mislabel logistic regression based on γ-divergence with the property of strong robustness in dealing with binary mislabeled response data. Neverthelss, above studies is not suitable to mislabeled ordinal response data.

In this article, we apply γ-divergence to ordinal multiple classification logistic regression and construct a robust ordinal mislabel logistic regression model, which not only possess the strong robustness but also the mislabel probabilities need not to be modeled. Our method have better effectiveness on ordinal mislabeled response data. It is noted that although γ-divergence has been previously adopted for regression analysis, to the best of our knowledge, γ-divergence has never been applied to ordinal response data.

1 Methodology

Given a dataset $\{(\boldsymbol{x}_i, y_{0i}), i = 1,2,\cdots,n\}$ in binary logistic regression, where $\boldsymbol{x}_i \in \mathbb{R}^p, y_{0i} \in \{0,1\}$, let $p = P(y_0 = 1 \mid \boldsymbol{X} = \boldsymbol{x}), 1 - p = P(y_0 = 0 \mid \boldsymbol{X} = \boldsymbol{x})$, the log-odds (or logit) of probability p of an event is defined as $\mathrm{logit}(p) = \log[p/(1-p)] = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + b$.

1.1 Ordinal logistic regression

Suppose there are $K + 1$ ordered classes $\{0,1,2,\cdots,K\}$, $\pi_j(\boldsymbol{x}) = P(y_0 = j \mid \boldsymbol{X} = \boldsymbol{x})$. Consider the event that an observation belongs to the meta-class $\{0,1,\cdots,j\}$, that is $y_0 \in \{0,1,\cdots,j\}$, whose probability is $\sum_{k=0}^{j} \pi_k(\boldsymbol{x})$. The ordered logistic regression model (also called cumulative odds logit model[11]) assumes that the log-odds are

$$\begin{aligned} &\text{logit}[\pi_0(\boldsymbol{x})+\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})+\cdots+\pi_j(\boldsymbol{x})] \\ &=\log\left[\frac{\pi_0(\boldsymbol{x})+\pi_1(\boldsymbol{x})+\pi_2(\boldsymbol{x})+\cdots+\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})+\pi_{j+2}(\boldsymbol{x})+\cdots+\pi_K(\boldsymbol{x})}\right] \\ &=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}+b_{j+1}, j=0,\cdots,K-1. \qquad (1) \end{aligned}$$

And the coefficient vector $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ is the same for all j in equation (1). Their intercept terms $b_1, b_2, \cdots, b_K$ split the plane into $K+1$ ordinal parts. Therefore, we only need to estimate parameters $\boldsymbol{\beta}, b_1, b_2, \cdots, b_K$, and $\boldsymbol{\beta}_{j+1}=(\boldsymbol{\beta}^{\mathrm{T}}, b_{j+1})^{\mathrm{T}}$ denotes $(j+1)$ −th hyperplane, $j=0,1,2,\cdots,K-1$.

According to Eq. (1), we obtain the category probability functions as follows:

$$\begin{cases} \pi_0(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)=\dfrac{A\exp(b_1)}{D} \\ \qquad \Pi_{h=2}^{K}[1+A\exp(b_h)], \\ \pi_j(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)=\dfrac{1}{D}[A\exp(b_{j+1})-A\exp(b_j)] \\ \qquad \Pi_{l=1,l\neq j,l\neq j+1}^{K}[1+A\exp(b_l)], \\ \qquad j=1,\cdots,K-1, \\ \pi_K(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)=\dfrac{1}{D}\Pi_{h=1}^{K-1}[1+A\exp(b_h)], \end{cases} \qquad (2)$$

where $D=\Pi_{h=1}^{K}[1+A\exp(b_h)], A=\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})$.

1.2 The minimum γ-divergence estimation and its estimation equation

Let g be the data generating probability density function and $f_{\boldsymbol{\theta}}$ be the model probability density function indexed by the parameter $\boldsymbol{\theta}$, the γ-divergence between g and $f_{\boldsymbol{\theta}}$ is defined as[7]

$$D_\gamma(g,f_{\boldsymbol{\theta}})=\frac{1}{\gamma(\gamma+1)}\left\{\|g\|_{\gamma+1}-\int\left(\frac{f_{\boldsymbol{\theta}}}{\|f_{\boldsymbol{\theta}}\|_{\gamma+1}}\right)^{\gamma}g\right\}, \qquad (3)$$

where

$$\|g\|_{\gamma+1}=\left(\int g^{\gamma+1}\right)^{\frac{1}{\gamma+1}}, \ \|f_{\boldsymbol{\theta}}\|_{\gamma+1}=\left(\int f_{\boldsymbol{\theta}}^{\gamma+1}\right)^{\frac{1}{\gamma+1}}.$$

In the limiting case, $\lim\limits_{\gamma\to 0}D_\gamma(g,f_{\boldsymbol{\theta}})=\int\ln\left(\dfrac{g}{f_{\boldsymbol{\theta}}}\right)g$, which is the KL-divergence[7].

In the presence of contamination, $g=\varepsilon f_{\boldsymbol{\theta}^*}+(1-\varepsilon)\tau$, $\boldsymbol{\theta}^*$ is the true model parameter, $1-\varepsilon$ is the contamination proportion and τ is the contamination density function. The estimation criterion of the minimum γ-divergence estimates parameter by minimizing $D_\gamma(g,f_{\boldsymbol{\theta}})$, which is equivalent to minimizing

$$\varepsilon D_\gamma(f_{\boldsymbol{\theta}^*},f_{\boldsymbol{\theta}})-\frac{H_\gamma(\varepsilon,\tau;\boldsymbol{\theta})}{\gamma(\gamma+1)},$$

by taking the actual g into Eq. (3) and ignoring the terms which not involve $\boldsymbol{\theta}$, where $H_\gamma(\varepsilon,\tau;\boldsymbol{\theta})=(1-\varepsilon)\int\left(\dfrac{f_{\boldsymbol{\theta}}}{\|f_{\boldsymbol{\theta}}\|_{\gamma+1}}\right)^{\gamma}\tau$.

Remark 1.1 Suppose that $\int f_{\boldsymbol{\theta}^*}^{\gamma}\tau$ is sufficiently small for an appropriately large $\gamma>0$, which implies that the contamination density function τ mostly lies on the tail of the underlying density $f_{\boldsymbol{\theta}^*}$. Then for some γ, the bias $H_\gamma(\varepsilon,\tau;\boldsymbol{\theta})$ is negligibly small when $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^*$, this means that the estimation of $\boldsymbol{\theta}$ is less effected by the $1-\varepsilon$ and τ. There are more introductions about γ-divergence and its property[12] and more detailed discussions about the bias $H_\gamma(\varepsilon,\tau;\boldsymbol{\theta})$ [7].

1.3 Robust ordinal mislabel logistic regression method

Consider n independent samples and combine the discussions of $H_\gamma(\varepsilon,\tau;\boldsymbol{\theta})$, the empirical version of the γ-divergence loss function[7] is

$$L(\boldsymbol{\theta})=-\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i\mid\boldsymbol{x}_i;\boldsymbol{\theta})^{\gamma}}{\|f(\cdot\mid\boldsymbol{x}_i;\boldsymbol{\theta})\|_{\gamma+1}}, \qquad (4)$$

where $\|f(\cdot\mid\boldsymbol{x}_i;\boldsymbol{\theta})\|_{\gamma+1}=\left(\int f(\ \mid\boldsymbol{x}_i;\boldsymbol{\theta})^{\gamma+1}\mathrm{d}y\right)^{\frac{\gamma}{\gamma+1}}$.

In conventional ordinal logistic regression model with $K+1$ classes, namely $y_0\in\{0,1,\cdots,K\}$, the probability density function

$$f(y_0\mid\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)=\prod_{j=0}^{K}\pi_j^{Y_{0j}}(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K), \qquad (5)$$

for y_0, where $\pi_j(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K), j=0,1,\cdots,K$ given by the (2), and y_0 can be recorded by K random variables $(Y_{01},Y_{02},\cdots,Y_{0K})$, where $(0,0,0,\cdots,0)$ denotes $y_0=0, y_0=j$ when $Y_{0j}=1, Y_{0k}=0$, $j\neq k$. On the basis of (4), the robust estimator $\hat{\boldsymbol{\beta}}_\gamma, \hat{b}_{j\gamma}, j=1,2,\cdots,K$ can be acquired by minimizing the objective function

$$F(\boldsymbol{\beta},b_1,\cdots,b_K)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{f(y_i \mid \boldsymbol{x}_i;\boldsymbol{\beta},b_1,\cdots,b_K)}{\| f(\cdot \mid \boldsymbol{x}_i;\boldsymbol{\beta},b_1,\cdots,b_K)\|_{1+\gamma}}\right)^{\gamma}, \tag{6}$$

where

$$\| f(\cdot \mid \boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)\|_{1+\gamma} = \left\{\sum_{j=0}^{K}[\pi_j(\boldsymbol{x};\boldsymbol{\beta},b_1,\cdots,b_K)]^{1+\gamma}\right\}^{\frac{1}{1+\gamma}}.$$

Direct differentiation of the objective function $F(\boldsymbol{\beta},b_1,\cdots,b_K)$ leads to the parameters estimation equations $S_\gamma(\hat{\boldsymbol{\beta}}_\gamma)=0, S_\gamma(\hat{b}_{j\gamma})=0, j=1,2,\cdots,K$, where

$$S_\gamma(\boldsymbol{\beta}) = -\frac{\gamma}{1+\gamma}\cdot\frac{1}{n}\sum_{i=1}^{n}\omega_{\gamma i}\left(\frac{\frac{(1+\gamma)\partial C_i}{\partial\boldsymbol{\beta}}}{C_i}-\frac{\frac{\partial B_i}{\partial\boldsymbol{\beta}}}{B_i}\right), \tag{7}$$

$$S_\gamma(b_j) = -\frac{\gamma}{1+\gamma}\cdot\frac{1}{n}\sum_{i=1}^{n}\omega_{\gamma i}\left(\frac{\frac{(1+\gamma)\partial C_i}{\partial b_j}}{C_i}-\frac{\frac{\partial B_i}{\partial b_j}}{B_i}\right), \tag{8}$$

with the weight function

$$\omega_{\gamma i} = \left(\frac{C_i^{1+\gamma}}{B_i}\right)^{\frac{\gamma}{\gamma+1}}, \tag{9}$$

and where $A_i=\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$,

$$B_i = \sum_{j=1}^{K-1}\left\{[A_i(\exp(b_{j+1})-\exp(b_j))]\prod_{\substack{l=1,\\ l\neq j,\\ l\neq j+1}}^{K}[1+A_i\exp(b_l)]\right\}^{1+\gamma} + \left[A_i\exp(b_1)\prod_{h=2}^{K}(1+A_i\exp(b_h))\right]^{1+\gamma}+\prod_{j=1}^{K-1}[1+A_i\exp(b_j)]^{1+\gamma}, \tag{10}$$

$$C_i = \sum_{j=1}^{K-1}\left\{Y_{ij}[A_i(\exp(b_{j+1})-\exp(b_j))]\prod_{\substack{l=1,\\ l\neq j,\\ l\neq j+1}}^{K}[1+A_i\exp(b_l)]\right\} + \left(1-\sum_{j=1}^{K}Y_{ij}\right)A_i\exp(b_1)\prod_{h=2}^{K-1}[1+A_i\exp(b_h)] + Y_{iK}\prod_{h=1}^{K-1}[1+A_i\exp(b_h)] \tag{11}$$

When $\gamma=0$, the above estimation equation degenerates to the estimation equation of conventional ordinal logistic regression with non-robust estimator. From (7)-(9), the robustness of $\hat{\boldsymbol{\beta}}_\gamma, \hat{b}_{j\gamma}, j=1,2,\cdots,K$ is clear, that's to say, when mislabeled, the weight $\omega_{\gamma i}$ of the sample $\boldsymbol{x}_i$ diminish in comparison with the case that the $\boldsymbol{x}_i$ is properly labeled. This point is explained at the Remark 1.2.

Remark 1.2 The weight of $\boldsymbol{x}$ in the parameter estimation equation is $\omega_\gamma = \left(\frac{C^{1+\gamma}}{B}\right)^{\frac{\gamma}{1+\gamma}}$. Let the $\omega_{\gamma,k}$ represents the weight of $\boldsymbol{x}$ when it is judged as the k-th class. When $\boldsymbol{x}$ is mislabeled, the weight diminish in comparison with the situation that the sample $\boldsymbol{x}$ is properly labeled, which can be described as the following proposition: if the true label of the sample $\boldsymbol{x}$ is j, i.e. $\pi_j > \pi_h$, then $\omega_{\gamma,j} > \omega_{\gamma,h}, j=0,1,2,\cdots,K, h\neq j$. For example, when $K=2$, we assume that the true label of the $\boldsymbol{x}$ is 1, we demonstrate $\pi_1>\pi_0, \pi_1>\pi_2 \Rightarrow \omega_{\gamma,1}>\omega_{\gamma,0}, \omega_{\gamma,1}>\omega_{\gamma,2}$ correspondingly. $\pi_0 = \frac{A\exp(b_1)}{1+A\exp(b_1)}$, $\pi_1 = \frac{A\exp(b_2)}{1+A\exp(b_2)}-\frac{A\exp(b_1)}{1+A\exp(b_1)}$, we get $\exp(b_2) > 2\exp(b_1)+A\exp(b_1+b_2)$ when $\pi_1>\pi_0$. $\omega_{\gamma,0} = \left\{\frac{[A\exp(b_1)(1+A\exp(b_2))]^{1+\gamma}}{B}\right\}^{\frac{\gamma}{1+\gamma}}$, $\omega_{\gamma,1} = \left\{\frac{[A\exp(b_2)-A\exp(b_1)]^{1+\gamma}}{B}\right\}^{\frac{\gamma}{1+\gamma}}$, if $\omega_{\gamma,1}>\omega_{\gamma,0}$ we need $\exp(b_2)>2\exp(b_1)+A\exp(b_1+b_2)$, so $\pi_1>\pi_0 \Rightarrow \omega_{\gamma,1}>\omega_{\gamma,0}$. Similarly, when $\pi_1>\pi_2$, we can deduce $\omega_{\gamma,1}>\omega_{\gamma,2}$. Similar conclusions can be drawn under other K.

1.4 Computation

Due to the complexity of the second derivative of the objective function based on γ-divergence, in this paper, we adopt the gradient descent algorithm to solve model, which is summarized in Algorithm 1.

Algorithm 1 Gradient descent algorithm

```
Require: The dataset: $(\boldsymbol{x}_i, y_i), i=1,2,\cdots,n$, the tuning parameter: $\gamma$.
Ensure: The regression coefficients: $\boldsymbol{\beta}, b_1, b_2, \cdots, b_K$. Initializing the $s=0, \varepsilon=10^{-3}$ and given an initial value $\boldsymbol{\beta}^{(0)}, b_j^{(0)}, j=1,2,\cdots,K$.
repeat
    $s=s+1$;
```

```
  for $k = 1,2,\cdots,p$ do
    $\frac{\partial F}{\partial \beta_k} = -\frac{\gamma}{1+\gamma}\cdot\frac{1}{n}\sum_{i=1}^{n}\omega_{\gamma i}\left(\frac{\frac{(1+\gamma)\partial C_i}{\partial\beta_k}}{C_i}-\frac{\frac{\partial B_i}{\partial\beta_k}}{B_i}\right)$;
end for
for $j=1,2,\cdots,K$ do
    $\frac{\partial F}{\partial b_j} = -\frac{\gamma}{1+\gamma}\cdot\frac{1}{n}\sum_{i=1}^{n}\omega_{\gamma i}\left(\frac{\frac{(1+\gamma)\partial C_i}{\partial b_j}}{C_i}-\frac{\frac{\partial B_i}{\partial b_j}}{B_i}\right)$;
end for
    $(\boldsymbol{\beta}^{(s+1)},\boldsymbol{b}^{(s+1)}) = (\boldsymbol{\beta}^{(s)},\boldsymbol{b}^{(s)}) - \alpha\frac{\partial F}{\partial(\boldsymbol{\beta},\boldsymbol{b})}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(s)},\boldsymbol{b}=\boldsymbol{b}^{(s)}}$;
Where $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$, $\boldsymbol{b} = (b_1,b_2,\cdots,b_K)^{\mathrm{T}}$, $\omega_{\gamma i}, B_i, C_i$ are defined as (9)-(11), $\alpha$ is the step length obtained by Armijo search. We consider three classification problem in our simulation study, the calculation results of $\frac{\partial C_i}{\partial\beta_k},\frac{\partial B_i}{\partial\beta_k},\frac{\partial C_i}{\partial b_j},\frac{\partial B_i}{\partial b_j}$ are placed in Appendix (A1)-(A6).
until
    $\frac{\|(\boldsymbol{\beta}^{\mathrm{T}(s+1)}),\boldsymbol{b}^{\mathrm{T}(s+1)}) - (\boldsymbol{\beta}^{\mathrm{T}(s)},\boldsymbol{b}^{\mathrm{T}(s)})\|_2}{\|(\boldsymbol{\beta}^{\mathrm{T}(s)},\boldsymbol{b}^{\mathrm{T}(s)}\|_2}\leqslant\varepsilon.$
return $\boldsymbol{\beta}^{\mathrm{T}(s+1)},\boldsymbol{b}^{\mathrm{T}(s+1)}$.
```

The proposed objective function has a complex form, so it is difficult to establish the convergence property. Experimental results demonstrate that, for all of our simulated and real datasets, convergence is successfully achieved within 50 overall iterations (mostly within 20 iterations).

The definition of γ-divergence and studies suggest that γ balances between robustness and efficiency. However, Fujisawa and Eguchi[12] said there could be not consistent best way to select an appropriate tuning parameter γ. In practice, there are some methods to select the tuning parameter γ, such as the adaptive selection procedure by Mollah et al.[13], the cross validation[14] and a sequence of the parameter[15]. In this article, we consider a sequence of γ as 1,2,3,4 and 5 following Zang et al.[15], and proceed the simulation study under each determined γ.

2 Simulation studies

In the numerical simulation study, we consider ordinal three classification problem with possibly mislabeling, that's to say, the labels of ordinal response datas $y \in \{0,1,2\}$, then $\pi_0(\boldsymbol{x}), \pi_1(\boldsymbol{x}), \pi_2(\boldsymbol{x})$, the probability density function $f(y \mid \boldsymbol{x};\boldsymbol{\beta}, b_1, b_2)$, and C_i, B_i, A_i can be obtained by the (2), (5), (10)-(11), and the calculation results of them are placed in Appendix (A7)-(A13). Simulation results are reported with 500 replicates in the simulation studies.

2.1 Simulation settings

In each simulation run, we generate 300 random samples $\boldsymbol{x}_s$ in $\mathbb{R}^{14}$, and $\boldsymbol{x}_s \sim N(\boldsymbol{0},\boldsymbol{\Sigma})$, we consider three different structures of covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{1\leqslant i,j\leqslant 14}$,

1) the explanatory variables are independent, i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$,

2) the auto-regressive correlation (AR) given by $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$ and $\rho = 0.75$,

3) the banded correlation (Band) given by $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.6$ if $|i-j| \leqslant 2, \rho = 0$ if $|i-j| > 2$.

We generate ordered mislabeled labels $\{0,1,2\}$ in the light of the following settings of mislabel probabilities:

$$\begin{cases}\eta_{01} = \eta_{10} = m/2,\\ \eta_{12} = \eta_{21} = m/2,\\ \eta_{02} = \eta_{20} = m/4,\end{cases} \tag{12}$$

where

$$\eta_{ij} = P(y = j \mid y_0 = i, \boldsymbol{X} = \boldsymbol{x}), i,j = 0,1,2.$$

We assume $m = \{0.1, 0.15, 0.2, 0.25, 0.3\}$ in the simulation. Considering the mislabel probabilities, the labels can be generated by following formulas:

$$\begin{cases}P(y = 0 \mid \boldsymbol{X} = \boldsymbol{x}) = (1-\eta_{01}-\eta_{02})\pi_0 + \eta_{10}\pi_1 + \eta_{20}\pi_2,\\ P(y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = (1-\eta_{10}-\eta_{12})\pi_1 + \eta_{01}\pi_0 + \eta_{21}\pi_2,\\ P(y = 2 \mid \boldsymbol{X} = \boldsymbol{x}) = (1-\eta_{20}-\eta_{21})\pi_2 + \eta_{02}\pi_0 + \eta_{12}\pi_1.\end{cases} \tag{13}$$

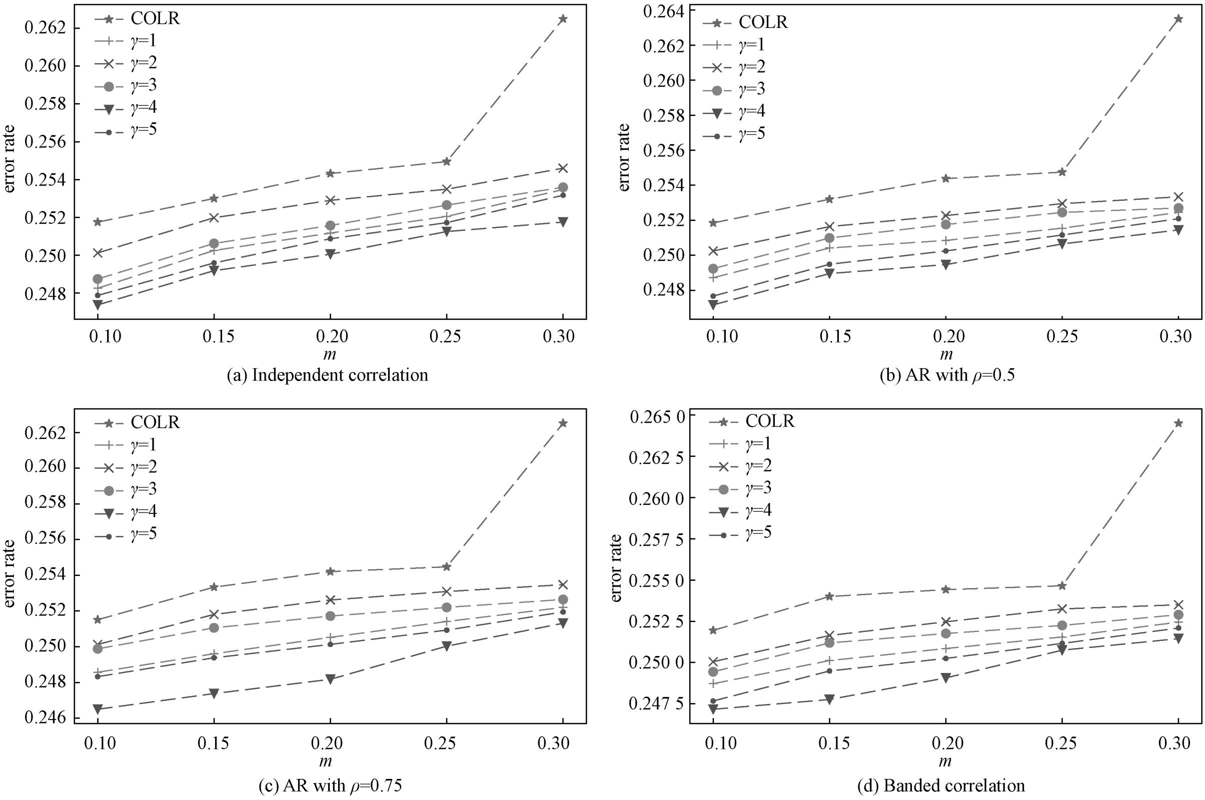
2.2 Simulation results

We apply the robust ordinal mislabel logistic regression method on the simulated dataset. And we compare the conventional ordinal logistic regression (COLR) method which takes no account of mislabeling with the robust ordinal mislabel logistic regression (robust ordinal mislabel) method from different aspects. Firstly, we evaluate the two methods via two indexes, i. e. the mean absolute error (MAE) and standard deviation (SD) of the estimates of $\boldsymbol{\beta}, b_1, b_2$. Secondly, we compare the classification accuracy of the two methods on 800 test samples. The true values, namely $\boldsymbol{\beta}_0, b_{01}, b_{02}$ follow the first part of uniform distribution $[-1, 0.5]$ and the last part of uniform distribution $[0.5, 1]$, and $\boldsymbol{\beta}_{01} = (\boldsymbol{\beta}_0^{\mathrm{T}}, b_{01})^{\mathrm{T}}$, $\boldsymbol{\beta}_{02} = (\boldsymbol{\beta}_0^{\mathrm{T}}, b_{02})^{\mathrm{T}}$ are two parallel classification hyperplanes. Partial results are presented in the article.

Figure 1 displays the variation of classification error rate of 800 test samples with m under different structures of $\boldsymbol{\Sigma}$. It can be seen clearly that the classification error rate obtained by COLR method increases more greatly than the classification error rate acquired by robust ordinal mislabel method with the increase of the mislabel probability (MP). Furthermore, we can select γ by the performance of classification from Fig. 1. We can also note that there exists the best effectiveness when $\gamma = 4$.

When $\boldsymbol{\Sigma} = \boldsymbol{I}$, the comparisons between the true values and the estimates of $\boldsymbol{\beta}, b_1, b_2$ under different γ and m are presented in Fig. 2 (a)–Fig. 2 (d) and the comparisons between the true values and the estimates of $\boldsymbol{\beta}, b_1, b_2$ under different γ and m are displayed in Fig. 2 (e) – Fig. 2 (h) when the structure of $\boldsymbol{\Sigma}$ is AR with $\rho = 0.5$. Figure 2 show that, for all given m and γ and the structure of $\boldsymbol{\Sigma}$ is AR ($\rho = 0.5$) or $\boldsymbol{I}$, the estimates of $\boldsymbol{\beta}, b_1, b_2$ from the robust ordinal mislabel method are closer to the true values of $\boldsymbol{\beta}, b_1, b_2$ than from COLR method.

When the structure of $\boldsymbol{\Sigma}$ is Band, MAE of parameter estimation under different γ and m values are displayed in Fig. 3 (a)–Fig. 3 (d). When the structure of $\boldsymbol{\Sigma}$ is AR with $\rho = 0.75$, the MAE of parameter estimation are presented in Fig. 3 (e) – Fig. 3 (h). Figure 3 illustrate that the MAE of parameter estimation from the robust ordinal mislabel method is smaller than from COLR method. Both Fig. 2 and Fig. 3 demonstrate the robust ordinal mislabel method possesses better performance in estimating coefficients of the classification hyperplanes.



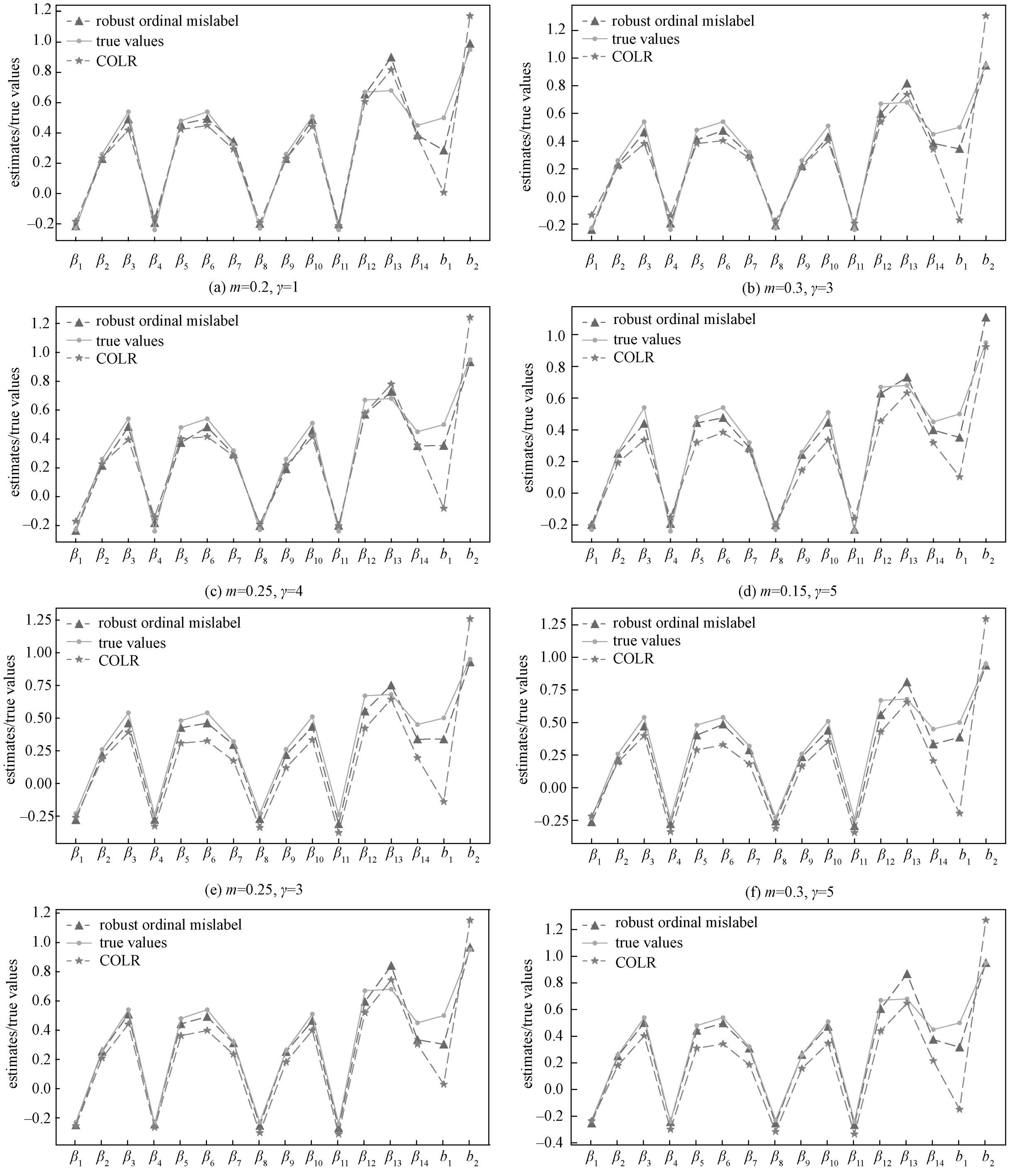
(a) Independent correlation

(b) AR with ρ=0.5

(c) AR with ρ=0.75

(d) Banded correlation

Fig. 1 Variation of classification error rate with m under different $\boldsymbol{\Sigma}$



(a) m=0.2, γ=1

(b) m=0.3, γ=3

(c) m=0.25, γ=4

(d) m=0.15, γ=5

(e) m=0.25, γ=3

(f) m=0.3, γ=5

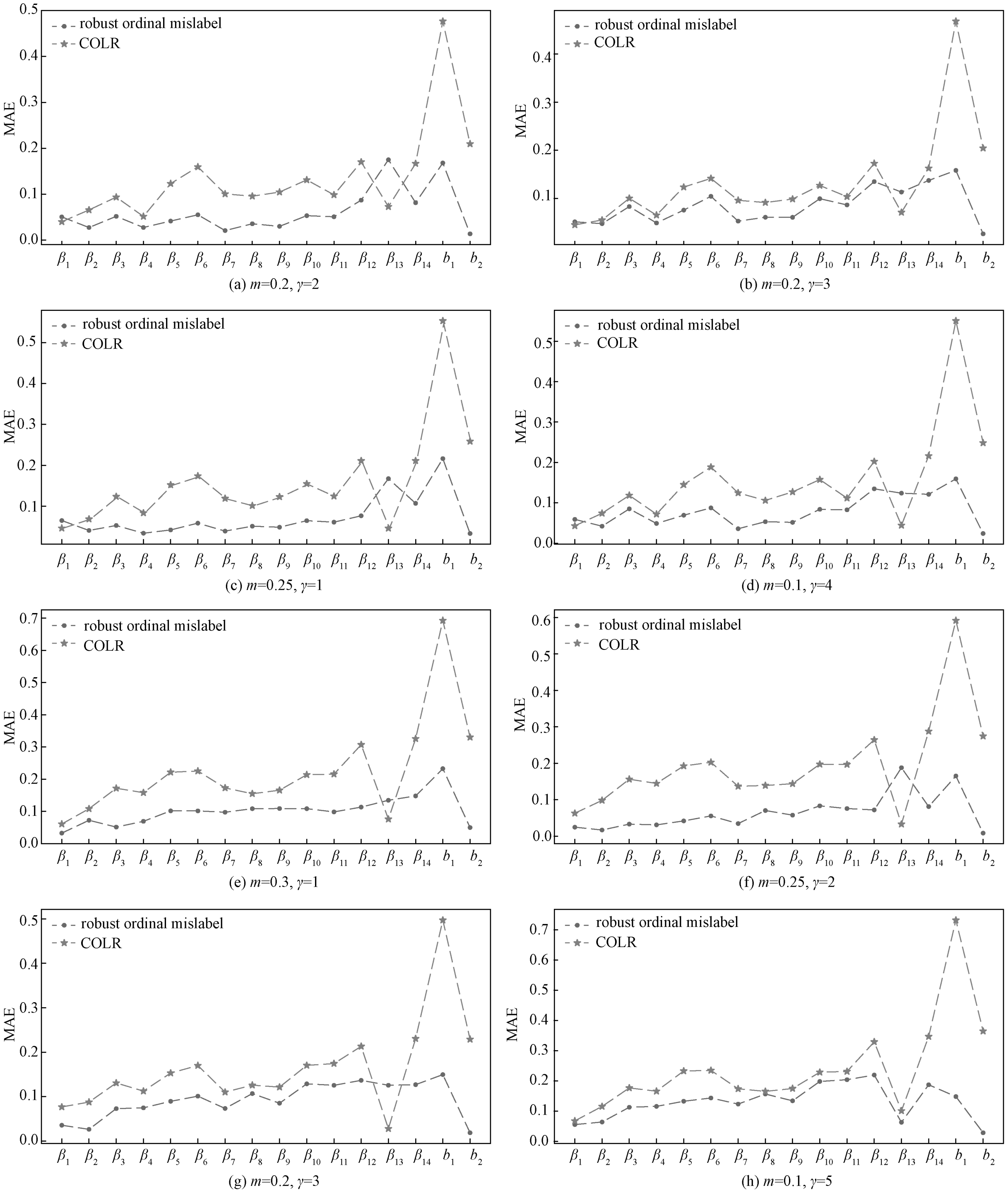
(g) m=0.2, γ=1

(h) m=0.3, γ=2

Fig. 2 The comparisons between the true values and the estimates of $\boldsymbol{\beta}$, b_1, b_2 under different settings

Table 1 displays MAE of parameter estimation and SD of the estimates from the COLR method and the robust ordinal mislabel method under different $\boldsymbol{\Sigma}$, MP and γ values. We can obviously find that the MAE and the SD from robust ordinal mislabel method are smaller than from COLR method respectively. It also demonstrates that the robust ordinal mislabel method performs well and more robustly than COLR method in estimating coefficients.

In addition, we verify the effectiveness of the proposed method on samples conforming to the heavy tailed distribution, such as t-distribution, mixture of normal distribution. In each simulation, we generate



(a) m=0.2, γ=2

(b) m=0.2, γ=3

(c) m=0.25, γ=1

(d) m=0.1, γ=4

(e) m=0.3, γ=1

(f) m=0.25, γ=2

(g) m=0.2, γ=3

(h) m=0.1, γ=5

Fig. 3 The comparisons of MAE from two methods under different settings

two groups of random samples $\boldsymbol{x}_s$ in $\mathbb{R}^{14}$, 300 in each group. $\boldsymbol{x}_s$ are from the distribution $t(\boldsymbol{0},\boldsymbol{\Sigma},df=3)$ in the first group and $\boldsymbol{x}_s$ are from mixture of normal distribution (0.2,$\boldsymbol{0}$,$\boldsymbol{I}$; 0.3,$\boldsymbol{0}$, AR (ρ = 0.5); 0.3,$\boldsymbol{0}$, AR (ρ = 0.75); 0.2,$\boldsymbol{0}$, Band) in the second group.

When $\gamma = 4$ and $\boldsymbol{x}_s$ are from the mixture of normal distribution, the MAE and SD of parameter estimation are displayed in Table 2. When $\gamma = 2$ and $\boldsymbol{x}_s$ are from the distribution $t(\boldsymbol{0},\boldsymbol{\Sigma},3)$, the MAE and SD of parameter estimation are displayed in Table 3. From the resluts, we can obviously find that both MAE and SD from the COLR mtehod are larger than both MAE and SD from the robust ordinal mislabel method when $\boldsymbol{x}_s$ are from heavy tailed distributions. This also means that the robust ordinal mislabel

Table 1　MAE and SD from two methods under different settings

Σ	MP	COLR		robust ordinal mislabel									
				γ = 1		γ = 2		γ = 3		γ = 4		γ = 5	
		MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD
Independent	m=0.10	0.080 2	0.371 2	0.055 0	0.336 0	0.069 1	0.370 0	0.055 1	0.351 5	0.054 0	0.296 3	0.055 0	0.296 3
	m=0.15	0.088 4	0.376 3	0.059 2	0.337 1	0.072 4	0.377 0	0.060 1	0.354 2	0.057 3	0.301 0	0.057 8	0.317 0
	m=0.20	0.104 1	0.379 6	0.061 4	0.385 0	0.075 2	0.377 4	0.068 2	0.354 6	0.058 2	0.307 3	0.060 5	0.313 4
	m=0.25	0.120 3	0.382 7	0.069 8	0.344 2	0.076 6	0.377 3	0.075 6	0.360 0	0.060 0	0.314 3	0.067 5	0.332 9
	m=0.30	0.136 6	0.383 9	0.078 4	0.346 5	0.078 4	0.379 3	0.078 8	0.367 1	0.072 4	0.320 4	0.073 2	0.336 5
AR ρ=0.50	m=0.10	0.079 2	0.385 0	0.054 6	0.359 8	0.071 1	0.369 9	0.055 1	0.367 1	0.054 3	0.344 6	0.054 4	0.353 1
	m=0.15	0.100 2	0.389 2	0.059 4	0.362 0	0.075 9	0.371 7	0.059 6	0.368 9	0.056 7	0.348 4	0.057 4	0.356 2
	m=0.20	0.122 5	0.391 1	0.060 6	0.364 2	0.076 0	0.371 8	0.065 8	0.370 6	0.060 5	0.352 7	0.060 5	0.359 7
	m=0.25	0.148 3	0.398 2	0.062 0	0.365 5	0.079 6	0.373 0	0.070 1	0.371 9	0.060 9	0.355 2	0.061 0	0.361 3
	m=0.30	0.173 4	0.407 5	0.061 6	0.367 0	0.081 3	0.373 2	0.073 6	0.372 8	0.061 2	0.359 0	0.061 4	0.363 8
AR ρ=0.75	m=0.10	0.100 5	0.393 8	0.063 2	0.380 0	0.086 1	0.381 1	0.064 9	0.380 0	0.055 4	0.379 1	0.062 1	0.379 7
	m=0.15	0.132 8	0.401 3	0.065 4	0.380 5	0.091 5	0.382 8	0.068 9	0.381 8	0.064 2	0.379 4	0.065 3	0.379 8
	m=0.20	0.161 7	0.408 2	0.067 9	0.380 8	0.096 9	0.384 2	0.078 4	0.383 7	0.066 5	0.380 1	0.067 1	0.380 3
	m=0.25	0.188 7	0.413 3	0.069 9	0.381 2	0.101 5	0.385 7	0.081 3	0.383 8	0.069 3	0.380 2	0.069 8	0.380 5
	m=0.30	0.217 9	0.423 5	0.075 8	0.382 3	0.112 2	0.391 1	0.085 5	0.390 3	0.072 4	0.381 2	0.074 6	0.382 2
Band	m=0.10	0.086 8	0.389 6	0.057 7	0.366 0	0.074 2	0.373 2	0.058 5	0.371 1	0.056 4	0.354 3	0.057 5	0.361 3
	m=0.15	0.108 7	0.394 0	0.058 4	0.366 1	0.079 0	0.371 3	0.060 0	0.313 0	0.058 3	0.355 9	0.058 4	0.362 6
	m=0.20	0.135 0	0.398 4	0.060 5	0.368 2	0.081 9	0.374 9	0.067 7	0.373 0	0.060 3	0.360 1	0.060 3	0.365 4
	m=0.25	0.158 2	0.403 0	0.061 3	0.369 5	0.086 4	0.376 6	0.071 8	0.374 4	0.060 9	0.361 8	0.061 3	0.366 3
	m=0.30	0.185 5	0.412 1	0.065 2	0.371 2	0.092 4	0.378 1	0.075 4	0.375 9	0.064 4	0.365 8	0.064 9	0.369 3

Table 2　MAE and SD from two methods under different distributions of X

MP	$\boldsymbol{X}\sim N(\boldsymbol{0},\mathrm{AR}(\rho=0.5))$				$\boldsymbol{X}\sim$ mixture of normal distribution			
	COLR		γ = 4		COLR		γ = 4	
	MAE	SD	MAE	SD	MAE	SD	MAE	SD
m=0.10	0.079 2	0.385 0	0.054 3	0.344 6	0.079 1	0.364 2	0.053 4	0.362 1
m=0.15	0.100 2	0.389 2	0.056 7	0.348 4	0.105 9	0.374 7	0.056 9	0.367 9
m=0.20	0.122 5	0.391 1	0.060 5	0.352 7	0.132 4	0.379 8	0.065 5	0.372 6
m=0.25	0.148 3	0.398 2	0.060 9	0.355 2	0.148 6	0.384 3	0.064 1	0.378 0
m=0.30	0.173 4	0.407 5	0.061 2	0.359 0	0.179 5	0.388 2	0.063 6	0.380 8

method has better effectiveness than COLR method in estimating coefficients. Simultaneously, we can find from the Table 2 and Table 3 that MAE of parameter estimation is smaller when samples are from normal distribution in contrast with the situation that samples are from t-distribution or mixture of normal distribution.

3　Real data analysis

In this section, we demonstrate the effectiveness of the proposed method on real data. Our dataset, i. e. the childhood eye data, is from Beijng Tongren Hospital[16], which contains 6 years of data from the first grade to sixth grade, and the variables include ocular biometry, near work, food habits, living habits, habits of wearing spectacles in this school year, accommodative response, times outdoors, and parental myopia and so on.

In the childhood vision research, we regard the change of spherical equivalent (CSE) after mydriasis from first grade to sixth grade as dependent variable. We divide the range of myopia into high myopia, moderate myopia and low myopia. The CSE <-2.5 denotes high myopia, corresponding to class 2 and the label is 2 in this class. The $-2.5\leqslant$ CSE $\leqslant-1.5$ denotes moderate myopia and corresponds to class 1. The CSE>-1.5 denotes low myopia and corresponds to class 0. The level of myopia is an ordinal response data and spherical equivalent might be inaccurate because of limitation of mydriasis examination, so there might be some mislabeled responses in the data.

We use the factors of vision from the first grade students as the independent variables to construct

Table 3 MAE and SD from two methods under different distributions of X

$\boldsymbol{\Sigma}$	MP	$\boldsymbol{X} \sim N(\boldsymbol{0},\boldsymbol{\Sigma})$				$\boldsymbol{X} \sim t(\boldsymbol{0},\boldsymbol{\Sigma},df = 3)$			
		COLR		$\gamma = 2$		COLR		$\gamma = 2$	
		MAE	SD	MAE	SD	MAE	SD	MAE	SD
Independent	m=0.10	0.080 2	0.371 2	0.069 1	0.370 0	0.084 8	0.360 2	0.069 3	0.360 3
	m=0.15	0.088 4	0.376 3	0.072 4	0.377 0	0.092 6	0.367 6	0.072 8	0.363 2
	m=0.20	0.104 1	0.379 6	0.075 2	0.377 4	0.105 2	0.370 4	0.075 6	0.364 2
	m=0.25	0.120 3	0.382 7	0.076 6	0.377 3	0.129 6	0.371 5	0.076 6	0.364 8
	m=0.30	0.136 6	0.383 9	0.078 4	0.379 3	0.137 2	0.379 2	0.078 9	0.368 1
AR ρ=0.50	m=0.10	0.079 2	0.385 0	0.071 1	0.369 9	0.079 6	0.364 2	0.071 3	0.362 1
	m=0.15	0.100 2	0.389 2	0.075 9	0.371 7	0.115 9	0.364 7	0.079 6	0.362 4
	m=0.20	0.122 5	0.391 1	0.076 0	0.371 8	0.128 9	0.369 8	0.081 5	0.364 3
	m=0.25	0.148 3	0.398 2	0.079 6	0.373 0	0.158 6	0.374 3	0.081 6	0.371 0
	m=0.30	0.173 4	0.407 5	0.081 3	0.373 2	0.175 5	0.375 2	0.081 6	0.371 8
AR ρ=0.75	m=0.10	0.100 5	0.393 8	0.086 1	0.381 1	0.102 1	0.370 1	0.085 9	0.361 2
	m=0.15	0.132 8	0.401 3	0.091 5	0.382 8	0.139 5	0.372 8	0.094 3	0.362 8
	m=0.20	0.161 7	0.408 2	0.096 9	0.384 2	0.171 0	0.373 1	0.098 9	0.363 2
	m=0.25	0.188 7	0.413 3	0.101 5	0.385 7	0.198 2	0.381 6	0.101 5	0.372 8
	m=0.30	0.217 9	0.423 5	0.112 2	0.391 1	0.219 6	0.384 1	0.113 1	0.380 4
Band	m=0.10	0.086 8	0.389 6	0.074 2	0.373 2	0.087 8	0.363 1	0.074 6	0.351 0
	m=0.15	0.108 7	0.394 0	0.079 0	0.371 3	0.109 6	0.367 3	0.079 6	0.363 2
	m=0.20	0.135 0	0.398 4	0.081 9	0.374 9	0.135 8	0.371 2	0.081 8	0.370 3
	m=0.25	0.158 2	0.403 0	0.086 4	0.376 6	0.164 5	0.376 1	0.086 8	0.371 4
	m=0.30	0.185 5	0.412 1	0.092 4	0.378 1	0.194 1	0.382 1	0.094 3	0.375 9

regression model, which can analyze the effectivenss of variables on the myopia progression in primary school. These variables are divided into three categories: continuous variables, nominal variables and multi-class variables. The data used in the model contains 1 370 samples and 28 independent variables and the meanings of the variables are presented in Appendix B.

We analyse the real data from two cases.

Example 3.1 (assessment of robustness) We first delete 10% samples before running and then calculate the mean (Mean) and SD of ten runs of each variable. The Mean and SD are presented in Table 4. Our method performs more robustly than COLR by SD obviously. Most important variables[17], such as "AL", "SE", "DUCVA", "D_COMR2", "BREAK3", have smaller SD by means of robust ordinal mislabel method than COLR method. It demonstrates that our method have strong stability.

Example 3.2 We take the Mean as the estimates of $\boldsymbol{\beta}, b_1, b_2$. The absolute value of the Mean can reflect the importance of variables. Our method can better reflect the importance of important variables, such as "AL", "SE", which are regarded as important variables[18], having greater absolute estimate with the help of robust ordinal mislabel method than COLR method. "LT" is positively correlated with the degree of myopia[19]. However, the estimate of it from robust ordinal mislabel method is positive than from COLR method is negative. It shows that the robust ordinal mislabel method proposed in the paper can more accurately reflect the relationship between independent variables and dependent variables. In addition, "K1" is negatively correlated with the degree of myopia. It can reflect this point from the case of $\gamma = 1,4,5$. However, the case of $\gamma = 2,3$ can not reflect the situation. The absolute value of the estimate of "K1" is the largest for the case of $\gamma = 4$ and the result is consistent with classification error rate under different γ.

4 Conclusion and future work

In this paper, we proposed a robust ordinal mislabel logistic regression method based on γ-divergence. The model is obtained by minimizing γ-divergence estimation. Both theoretical analysis and simulation studies make clear that the proposed method is quite efficient for solving classification

Table 4　The Mean and SD of the variables from two methods

variables	COLR		robust ordinal mislabel									
			γ = 1		γ = 2		γ = 3		γ = 4		γ = 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AR	−0.000 8	0.116 0	−0.006 9	0.058 3	−0.005 8	0.037 3	−0.005 9	0.050 5	−0.015 9	0.042 3	−0.010 1	0.027 2
LT	−0.021 8	0.081 6	0.006 2	0.058 6	0.004 1	0.061 9	0.005 7	0.069 1	0.013 2	0.065 1	0.010 8	0.021 9
AL	0.119 0	0.273 2	0.269 4	0.193 1	0.247 8	0.187 5	0.268 1	0.180 2	0.285 3	0.181 3	0.277 0	0.174 6
K1	−0.023 4	0.062 4	−0.031 6	0.044 1	0.032 6	0.025 0	0.033 5	0.021 6	−0.042 9	0.024 1	−0.041 6	0.056 4
K2	0.003 9	0.086 6	0.005 4	0.040 4	0.007 7	0.075 3	0.011 4	0.090 3	0.013 5	0.086 9	0.011 9	0.077 4
BL	−0.005 7	0.078 0	−0.009 8	0.141 5	−0.007 6	0.090 7	−0.007 9	0.068 7	−0.016 4	0.028 3	−0.013 6	0.061 0
SE	−0.137 0	0.265 2	−0.165 0	0.222 3	−0.151 0	0.166 0	−0.157 3	0.153 2	−0.174 0	0.148 3	−0.169 0	0.144 9
OMT	−0.007 3	0.009 8	−0.004 6	0.008 8	−0.004 0	0.008 1	−0.00 8	0.007 2	−0.008 9	0.009 8	−0.008 2	0.008 7
DUCVA	−0.135 8	0.191 2	0.134 0	0.183 1	0.140 4	0.152 9	−0.134 1	0.168 1	0.135 7	0.164 0	0.134 4	0.153 6
NUCVA	−0.015 3	0.016 2	−0.018 9	0.060 9	−0.009 7	0.085 6	−0.016 9	0.105 3	−0.024 7	0.146 5	−0.019 7	0.087 6
GLASS	−0.051 6	0.073 9	−0.0674	0.007 1	−0.072 4	0.003 0	−0.055 7	0.001 4	−0.058 4	0.005 8	−0.056 2	0.029 1
DESK_L	−0.045 4	0.079 4	−0.058 8	0.079 0	−0.046 3	0.061 2	−0.046 6	0.049 2	−0.055 6	0.039 3	−0.052 6	0.050 6
TUTOR2	−0.008 4	0.010 4	−0.004 9	0.010 4	−0.004 0	0.011 5	−0.009 5	0.023 5	−0.003 8	0.009 8	−0.006 5	0.103 6
GENDER	−0.013 1	0.022 6	−0.047 5	0.014 5	−0.061 8	0.016 5	−0.009 8	0.010 4	−0.014 3	0.014 1	−0.003 1	0.020 5
COLA3	−0.004 8	0.000 7	−0.009 4	0.001 3	−0.012 7	0.001 5	−0.014 5	0.001 8	−0.015 9	0.002 3	−0.017 9	0.002 1
EGGEAT1	−0.007 2	0.009 2	−0.007 3	0.004 1	−0.006 3	0.006 7	−0.006 5	0.008 6	−0.005 8	0.008 3	−0.008 3	0.010 8
EGGEAT2	−0.011 7	0.088 7	−0.018 9	0.017 1	−0.021 4	0.024 5	−0.011 4	0.087 9	−0.024 1	0.063 8	−0.027 8	0.050 3
EGGEAT4	−0.017 0	0.089 8	−0.040 2	0.080 4	−0.087 1	0.071 1	−0.056 1	0.064 5	−0.068 2	0.063 4	−0.039 0	0.060 5
MEATS1	−0.009 0	0.078 2	−0.008 7	0.053 1	−0.008 3	0.064 8	−0.009 8	0.069 5	−0.007 3	0.075 2	−0.008 4	0.075 8
READLY1	−0.082 2	0.049 6	−0.091 6	0.021 5	−0.120 8	0.013 4	−0.131 8	0.011 2	−0.135 1	0.028 8	−0.142 6	0.120 0
READLY3	−0.088 3	0.071 4	−0.076	0.046 2	−0.105 7	0.044 3	−0.116 9	0.085 4	−0.123 1	0.115 9	−0.134 1	0.124 9
READLY4	−0.122 9	0.119 7	−0.087 2	0.1046	−0.124 6	0.105 2	−0.137 9	0.104 6	−0.143 9	0.105 3	−0.155 0	0.113 7
BREAK3	−0.119 5	0.112 4	−0.085 7	0.106 8	−0.122 1	0.116 8	−0.135 7	0.106 8	−0.140 0	0.104 5	−0.146 7	0.111 7
MYOPICS	−0.045 3	0.092 4	−0.015 5	0.092 8	−0.027 1	0.091 6	−0.031 5	0.091 6	−0.031 8	0.091 0	−0.031 4	0.099 6
D_COMR1	−0.302 4	0.161 1	−0.334 3	0.104 0	−0.428 1	0.075 4	−0.459 2	0.070 4	−0.453 8	0.053 9	−0.455 6	0.057 4
D_COMR2	−0.154 6	0.136 9	−0.170 6	0.080 4	−0.241 6	0.047 3	−0.274 0	0.029 5	−0.302 9	0.019 2	−0.345 4	0.034 1
D_COMR3	−0.030 4	0.002 6	−0.027 1	0.002 7	−0.040 7	0.004 6	−0.047 4	0.006 2	−0.074 4	0.005 4	−0.065 2	0.003 0
PUPIL_SIZE	−0.026 8	0.071 1	−0.028 7	0.070 6	−0.021 4	0.063 2	−0.029 4	0.070 7	−0.021 3	0.071 7	−0.020 74	0.062 0

problems with ordinal and mislabeled response data. Firstly, the mislabel probabilities need not to be modeled in our method. Secondly, our method performs more robustly than conventional ordinal logistic regression by the ways of simulation results and real data analysis. However, the robust ordinal mislabel logistic regression based on γ-divergence is applied to low dimensional data. There exists mislabeled response data on some high dimensional ordinal data, we can consider further research for ordinal mislabel method to high dimensional data.

References

[1] Dembczyński K, Kotłowski W, Słowiński R. Ordinal classification with decision rules[C]//International Workshop on Mining Complex Data. Springer, Berlin, Heidelberg, 2007: 169-181. DOI:10.1007/978-3-540-68416-9_14.

[2] Cardoso J S, Costa J F. Learning to classify ordinal data: the data replication method[J]. Journal of Machine Learning Research, 2007, 8(Jul): 1393-1429.

[3] Chang K Y, Chen C S, Hung Y P. A ranking approach for human ages estimation based on face images[C]//2010 20th International Conference on Pattern Recognition. August 23-26, 2010, Istanbul, Turkey. IEEE, 2010: 3396-3399. DOI: 10.1109/ICPR.2010.829.

[4] Frank E, Hall M. A simple approach to ordinal classification [C]//European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2001: 145-156. DOI: 10.1007/3-540-44795-4_13.

[5] Shashua A, Levin A. Ranking with large margin principle: two approaches [C] // Advances in Neural Information Processing Systems. 2003: 961-968.

[6] Wang H D, Shi Y, Niu L F, et al. Nonparallel support vector ordinal regression[J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3306-3317. DOI: 10.1109/TCYB.2017.2682852.

[7] 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.

[8] Tian Y, Sun M, Deng Z B, et al. A new fuzzy set and nonkernel SVM approach for mislabeled binary classification

with applications[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1536-1545. DOI: 10.1109/TFUZZ.2017.2752138.

[9] Qian W M, Li Y M. Parameter estimation in linear regression models for longitudinal contaminated data [J]. Applied Mathematics: A Journal of Chinese Universities, 2005, 20(1): 64-74. DOI:10.1007/s11766-005-0038-0.

[10] Komori O, Eguchi S, Ikeda S, et al. An asymmetric logistic regression model for ecological data[J]. Methods in Ecology and Evolution, 2016, 7(2): 249-260. DOI:10.1111/2041-210X.12473.

[11] Cox C. Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach[J]. Statistics in Medicine, 1995, 14(11): 1191-1203. DOI: 10.1002/sim.4780141105.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Kakita T, Hiraoka T, Oshika T. Influence of overnight orthokeratology on axial elongation in childhood myopia[J]. Investigative Ophthalmology & Visual Science, 2011, 52(5): 2170-2174. DOI:10.1167/iovs.10-5485.

[17] Hoyt C S, Stone R D, Fromer C, et al. Monocular axial myopia associated with neonatal eyelid closure in human infants[J]. American Journal of Ophthalmology, 1981, 91(2): 197-200. DOI:10.1016/0002-9394(81)90173-2.

[18] Li S M, Liu L R, Li S Y, et al. Design, methodology and baseline data of a school-based cohort study in Central China: the Anyang childhood eye study [J]. Ophthalmic Epidemiology, 2013, 20(6): 348-359. DOI: 10.3109/09286586.2013.842596.

[19] Chang K Y, Chen C S, Hung Y P. Ordinal hyperplanes ranker with cost sensitivities for age estimation[C]//CVPR 2011. June 20-25, 2011, Colorado Springs, CO, USA. IEEE, 2011: 585-592. DOI:10.1109/CVPR.2011.5995437.

Appendices

Appendix A

We consider ordinal three classification problem in simulation studies, the $\frac{\partial C_i}{\partial \beta_k}$, $\frac{\partial B_i}{\partial \beta_k}$, $\frac{\partial C_i}{\partial b_j}$, $\frac{\partial B_i}{\partial b_j}$ in Algorithm 1 are as follows,

$$\frac{\partial C_i}{\partial b_1} = (1 - Y_{i1} - Y_{i2})A_i\exp(b_1)(1 + A_i\exp(b_2)) - Y_{i1}A_i\exp(b_1) + Y_{i2}A_i\exp(b_2). \tag{A1}$$

$$\frac{\partial B_i}{\partial b_1} = \{-[A_i(\exp(b_2) - \exp(b_1))]^{\gamma}A_i\exp(b_1) + (1 + A_i\exp(b_1))^{\gamma}A_i\exp(b_1) + [A_i\exp(b_1)(1 + A_i\exp(b_2))]^{1+\gamma}\}(1 + \gamma). \tag{A2}$$

$$\frac{\partial C_i}{\partial b_2} = (1 - Y_{i1} - Y_{i2})A_i^2\exp(b_1 + b_2) + Y_{i1}A_i\exp(b_2). \tag{A3}$$

$$\frac{\partial B_i}{\partial b_2} = \{[A_i(\exp(b_2) - \exp(b_1))]^{\gamma}A_i\exp(b_2) + [A_i\exp(b_1)(1 + A_i\exp(b_2))]^{\gamma}A_i^2\exp(b_1 + b_2)\}(\gamma + 1). \tag{A4}$$

$$\frac{\partial C_i}{\partial \beta_k} = \{Y_{i1}[A_i(\exp(b_2) - \exp(b_1))] + Y_{i2}A_i\exp(b_1) + (1 - Y_{i1} - Y_{i2})[A_i^2\exp(b_1 + b_2) + A_i\exp(b_1)(1 + A_i\exp(b_2))]\}x_{ij}. \tag{A5}$$

$$\frac{\partial B_i}{\partial \beta_k} = (1 + \gamma)\{[A_i(\exp(b_2) - \exp(b_1))]^{\gamma+1} + A_i\exp(b_1)(1 + A_i\exp(b_1))^{\gamma} + [A_i\exp(b_1)(1 + A_i\exp(b_2))]^{1+\gamma} + [A_i\exp(b_1)(1 + A_i\exp(b_2))]^{\gamma}A_i\exp(b_1 + b_2)\}x_{ij}. \tag{A6}$$

$$\pi_0 = P(y_0 = 0 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{A\exp(b_1)}{1 + A\exp(b_1)}. \tag{A7}$$

$$\pi_1 = P(y_0 = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{A\exp(b_2)}{1 + A\exp(b_2)} - \frac{A\exp(b_1)}{1 + A\exp(b_1)}. \tag{A8}$$

$$\pi_2 = P(y_2 = 2 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + A\exp(b_2)}. \tag{A9}$$

The objective function:

$$F(\boldsymbol{\beta}, b_1, b_2) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, b_1, b_2)}{\| f(\cdot \mid \boldsymbol{x}_i; \boldsymbol{\beta}, b_1, b_2) \|_{1+\gamma}}\right)^{\gamma} = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{C_i^{1+\gamma}}{B_i}\right)^{\frac{\gamma}{1+\gamma}}, \tag{A10}$$

where $\| f(\cdot \mid \boldsymbol{x}; \boldsymbol{\beta}, b_1, b_2) \|_{1+\gamma} = \left\{ \sum_{j=0}^{2} \pi_j^{1+\gamma} \right\}^{\frac{1}{1+\gamma}}$,

$$A_i = \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i), \tag{A11}$$

$$B_i = \{A_i\exp(b_1)[1 + A_i\exp(b_2)]\}^{1+\gamma} + [A_i(\exp(b_2) - \exp(b_1))]^{1+\gamma} + [1 + A_i\exp(b_1)]^{1+\gamma}, \tag{A12}$$

$$C_i = (1 - Y_{i1} - Y_{i2})A_i\exp(b_1)[1 + A_i\exp(b_2)] + Y_{i1}[A_i(\exp(b_2) - \exp(b_1))] + Y_{i2}[1 + A_i\exp(b_1)]. \tag{A13}$$

Appendix B

The meanings of the variables in Table 4 are as follows. The level number after some variable name is the result of the dummy variable processing.

Binary variables (1: yes, 0: no):

MOTHER_A: Whether the natural mother has amblyopia?

GLASS: Do you wear glasses?

DESK_L: When you are reading or doing close work, do you use a desk lamp?

OMT: Do your children have received the other myopia treatment?

TUTOR2: Do your children take part in the tutoring class in their spare time (indoor learning class)?

GENDER: Is gender a woman?

Ordinal variables (1:lower, 4:upper):

COLA: Quantity of drinking a carbonated drink frequency in the last 4 weeks.

EGGEAT: Quantity of eating eggs frequency in the last 4 weeks.

MEATS: Quantity of eating meat frequency in the last 4 weeks.

READLY: Quantity of weekly reading.

BREAK: Quantity of keeping reading or doing close work frequency.

D_COMR: When your child uses a computer, distance from computer from the computer (1:nearer, 4: farer).

Continuous variables:

MYOPICS: The number of myopia parents.

SE: Spherical equivalent after mydriasis.

DUCVA: Distant uncorrected visual acuity.

AL: Axial length.

LT: Lens thickness.

K1: The flat keratometry readings.

K2: The steep keratometry readings.

AR: Accommodative response.

PUPIL_SIZE: Pupil diameter.

BL: Birth length.

NUCVA: Near uncorrected visual acuity.