Predicting ordinal outcomes incorporating nonparametric interactions

Yuting Lu* Yongzhao Shao†

Abstract

Complex diseases such as cancer usually develop through different stages forming ordinal outcomes. Understanding the intrinsic mechanism underlying these disease stages is important for diagnosis, classification and subsequent treatment of these diseases. The etiology and development of complex diseases often involve the interactions between biomolecules, rather than individual molecules such as complicated interactions between tumor cells and immune cells in cancer immunotherapies. Predictive models are of great importance in precision or personalized medicine and other applications. In this paper, we developed a non-parametric approach to predict ordinal outcomes incorporating potentially complicated interactions between biomolecules. Simulation studies demonstrate that our approach performs well in classification, and in identification of truly informative differential pairs of predictors, when there is non-negligible interaction between predictors.

Key Words: Prediction, classification, ordinal outcome, nonparametrics, interaction, precision medicine

1. Introduction

Development and progression of a complex disease typically involve several stages according to disease severity, which naturally form ordinal outcomes. For example, melanoma and other cancers often develop from early to late stages; Alzheimer's disease typically develops progressively from cognitively normal, mild cognitive impairment (MCI) to Alzheimers disease (AD); asthma patients are often diagnosed at various severity from mild to severe (Liu et al., 2011; Patrawalla et al., 2012; Fontanella et al., 2018), etc. Traditional diagnosis, classification, and subsequent treatment often depend heavily on existing medical knowledge and physicians' experience. However, existing medical knowledge for complex diseases has many gaps and uncertainties, consequently, many patients may not be able to receive the most effective treatments or treatment sequences. As a consequence of the rapid development of biological technology and medical science, high-throughput 'omics' data (such as RNA sequencing data, protein expression data), medical imaging, and other clinical data arise as high-dimensional predictors. There are complex interactions between predictors that are important for predicting outcomes such as the interactions between immune cells and tumor cells that determine the success of cancer immunotherapies. In general, it is not straightforward for physicians to make medical decision based on all those high-dimensional data with complex correlations. Therefore, statistical predictive models utilizing such large amount of data with efficient incorporation of various correlations are in high demand to help classify patients systematically, so that patients could be effectively treated with suitable personalized treatment.

The etiology and prognosis of complex diseases involve intricate biological processes. It is well-known that biological processes involve a series of molecular functions, which are fulfilled by the interaction between biomolecules (such as genes), rather than individual biomolecules (Ji et al., 2017). For example, in cancer immunotherapies, we directly

*Department of Population Health, Grossman School of Medicine, New York University, New York, NY, 10016

†Department of Population Health, Grossman School of Medicine, New York University, New York, NY, 10016

treat the immune cells which, in turn, inhibit or kill tumor cells. Thus, the interactions between immune cells and tumor cells determine the success of cancer immunotherapies. Consequently, effective predictive models have to incorporate these important interactions between two kinds of cells (Sun et al., 2019). Thus, incorporating the interactions between predictors in a model is often necessary to improve predictive accuracy. Sometimes, linear interaction terms, such as x_1x_2, might not always be adequate to model complicated interactions. For instance, in the development and prognosis of Alzheimer's disease, age is an important predictor but it has non-linear effect and has complicated interactions with other predictors. In melanoma, the thickness of melanoma is a critical covariate determining the disease stage and survival probability. But "too thin" or "too thick" tumor is often not impacting the outcome linearly. As a result, the interaction between age and APOE, amyloid beta, etc in AD, and the interaction between tumor thickness and other important risk factors in melanoma, may not be simply characterized in a simple linear function of x_1x_2. In addition, differential interactions of gene pairs between various conditions identified by the model could help explore the disease mechanism. The mechanism identified by this model may support to develop new efficient treatments targeting particular biological functions, such as genetic based interventions (He et al., 2020).

There are important knowledge gaps in existing statistical models for ordinal outcomes with complicated interactions. The typical multinomial logistic regression model generalizes binary or ordinary logistic regression to model outcome with more than two levels. But it ignores the intrinsic ordering between the ordinal levels, which is unique and important for ordinal outcomes. Ordered logistic regression, also called proportional odds regression, takes into account the ordering of levels. But very often it can only include interaction in the form of linear terms in x_1x_2. Though tree-based methods, such as decision tree and random forest (Friedman et al. 2001), are nonparametric classification and prediction approaches, they have certain drawbacks. A single fitted tree might be unstable (e.g. due to overfitting) especially when there are too many covariates, while random forest is often too complicated to yield a simple biological explanation.

Recently, three nonparametric classification methods have been developed: FANS (Fan et al., 2016), ordinal FANS (Ferber, 2016), and JDINAC (Ji et al., 2017). FANS is a binary high-dimensional classification model that generalized naive Bayes and logistic regression. Ordinary logistic regression is

$$\log \frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

which uses simple linear combination of the original covariates to formulate a classifier. Substituting the original covariate x with $\log \frac{f(x)}{g(x)}$ in ordinary logistic regression, FANS is formed as

$$\log \frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j \log \frac{f_j(x_j)}{g_j(x_j)}, \tag{1}$$

where f_j, g_j are the densities of covariate x_j in class 1 and class 0, respectively. Both f_j, g_j were estimated nonparametrically, thus the classifier formed by FANS is in general not a linear classifier of the original covariates. As is well-known, the likelihood ratio test is the most powerful test to distinguish two models according to the Neyman-Pearson lemma. FANS, using the likelihood ratio $\log \frac{f_j(x_j)}{g_j(x_j)}$ as classifiers, could yield powerful classification procedure. Specifically, if x is normally distributed, FANS will generally become ordinary logistic regression. Ordinal FANS (Ferber, 2016) and JDINAC (Ji et al., 2017) are both extended from FANS. Ordinal FANS extended FANS to ordinal outcomes using powerful

predictors made from log density ratio as well. However, it only considers individual genes without their interactions, which might not perform well in prediction if interaction plays an important role as is often the case in complex diseases. JDINAC built classifiers that take into account the pairwise interaction between covariates. Motivated from FANS, the interaction is modeled by the log of bivariate joint density ratio, and its model is

$$\log \frac{P(Y=1|\mathbf{G},\mathbf{z})}{P(Y=0|\mathbf{G},\mathbf{z})} = \alpha_0 + \sum_{k=1}^{K} \alpha_k z_k + \sum_{i=1}^{p}\sum_{j>i}^{p} \beta_{ij} \log \frac{f_{ij}(G_i,G_j)}{g_{ij}(G_i,G_j)},$$

where $\mathbf{z}$ is covariate, G_i, G_j denote two different genes. f_{ij}, g_{ij} were also estimated non-parametrically. Such interaction terms could not only potentially capture the non-linear relationship between genes, but also act as a powerful classifier to improve classification accuracy. Unfortunately, it only works when the outcome is binary. In addition, there are a lot of tools to perform differential gene network analysis (Shojaie, 2020), but most of them assumed linear interaction, which could be restrictive in real applications.

To fill the knowledge gap, in this paper, we present a non-parametric predicting approach incorporating potentially complicated pairwise interactions between predictors for the ordinal outcomes. Numerical results demonstrate that our method performs better in terms of lower misclassification rate and higher Somers' index than existing approaches that don't include interaction. Additionally, our model can also identify important differential pairs efficiently, which could be useful for differential network construction and mechanism investigation.

2. Method

2.1 Model set-up

For each individual $i(i = 1, \ldots, n)$, suppose we observed the outcome y_i which is a K-level ordinal outcome taking values from $\{1, \ldots, K\}$, p-dimensional gene-level activities (such as mRNA expression level, protein expression) $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, and q covariates $z_{i1}, \ldots, z_{iq}$ referring to age, gender, or the original gene-level activities etc. Let $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{iq})^T$. Our approach is structured as the following $K-1$ binary logit models:

$$\begin{aligned}
\log \frac{P(Y=1|\mathbf{x},\mathbf{z})}{P(Y>1|\mathbf{x},\mathbf{z})} &= \mathbf{z}^T\boldsymbol{\alpha}^{(1)} + \sum_{s=1}^{p-1}\sum_{t>s}^{p} \beta_{st}^{(1)} \log \frac{f_{st}^{(1)}(x_s,x_t)}{g_{st}^{(1)}(x_s,x_t)}, \\
\log \frac{P(Y\le 2|\mathbf{x},\mathbf{z}))}{P(Y>2|\mathbf{x},\mathbf{z}))} &= \mathbf{z}^T\boldsymbol{\alpha}^{(2)} + \sum_{s=1}^{p-1}\sum_{t>s}^{p} \beta_{st}^{(2)} \log \frac{f_{st}^{(2)}(x_s,x_t)}{g_{st}^{(2)}(x_s,x_t)}, \qquad (2) \\
&\vdots \\
\log \frac{P(Y\le K-1|\mathbf{x},\mathbf{z}))}{P(Y=K|\mathbf{x},\mathbf{z}))} &= \mathbf{z}^T\boldsymbol{\alpha}^{(K-1)} + \sum_{s=1}^{p-1}\sum_{t>s}^{p} \beta_{st}^{(K-1)} \log \frac{f_{st}^{(K-1)}(x_s,x_t)}{g_{st}^{(K-1)}(x_s,x_t)},
\end{aligned}$$

where $\boldsymbol{\alpha}^{(k)} = (\alpha_0^{(k)}, \alpha_1^{(k)}, \ldots, \alpha_q^{(k)})^T$ is the coefficient of $\mathbf{z}_i$ in the k-th equation, $\alpha_0^{(k)}$ is the intercept, $\beta_{st}^{(k)}$ is the coefficient of the (s,t)-th log density ratio in the k-th equation. $\mathbf{z}^T\boldsymbol{\alpha}^{(k)}$ represents the main effect of covariates including original gene-activities, while $\log \frac{f_{st}^{(k)}(x_s,x_t)}{g_{st}^{(k)}(x_s,x_t)}$ models the interaction in a more general way, which could handle linear or non-linear interactions. The k-th logit model regresses a new binary outcome, obtained by

partitioning the original samples into $Y \leq k$ and $Y > k$, on corresponding covariates and log density ratios. The log density ratio in (2) is defined as

$$\log \frac{f_{st}^{(k)}(x_s, x_t)}{g_{st}^{(k)}(x_s, x_t)} = \log \frac{P(X_s = x_s, X_t = x_t | Y \leq k)}{P(X_s = x_s, X_t = x_t | Y > k)}, \tag{3}$$

for $k = 1, \ldots, K-1$. $f_{st}^{(k)}, g_{st}^{(k)}$ are the class-specific joint density of X_s and X_t, showing the strength of interaction between the s-th and t-th genes(or mRNA, protein, etc.), for $Y \leq k$ and $Y > k$, respectively. The density ratio implies the differential association of the two genes between $Y \leq k$ and $Y > k$. Combining the main effect $\mathbf{z}^T \boldsymbol{\alpha}^{(k)}$ and interactions $\sum_{s=1}^{p-1} \sum_{t>s}^{p} \beta_{st}^{(k)} \log \frac{f_{st}^{(k)}(x_s,x_t)}{g_{st}^{(k)}(x_s,x_t)}$ should formulate a powerful multivariate classifier. One benefit of designing the approach in this way is that the whole training set contributes to each logit model, so that the estimation of each logit model are in the same precision.

Model (2) contains a lot of existing models. When X_s, X_t are independent, the joint density breaks down into the product of marginal densities, i.e.,

$$f_{st}^{(k)}(x_s, x_t) = f_s^{(k)}(x_s) f_t^{(k)}(x_t), \quad g_{st}^{(k)}(x_s, x_t) = g_s^{(k)}(x_s) g_t^{(k)}(x_t).$$

Thus model (2) will become ordinal FANS, or FANS if $K = 2$. When (X_s, X_t) follows bivariate normal distribution that shares the same variance Σ in both $Y \leq k$ and $Y > k$, the interaction term in model (2) will be the ordinary linear interaction $x_s x_t$. Model (2) include JDINAC as a special case when $K = 2$.

2.2 Estimation

Parameters will be estimated through maximum likelihood. Since $K-1$ equations in (2) involve different parameters, the estimation would be conducted individually. In the k-th equation, denote all log of bivariate density ratios by a $\frac{p(p-1)}{2}$ vector $\mathbf{w}^{(k)}$,

$$\mathbf{w}^{(k)} = \left(\log \frac{f_{12}^{(k)}(x_1, x_2)}{g_{12}^{(k)}(x_1, x_2)}, \ldots, \log \frac{f_{p-1,p}^{(k)}(x_{p-1}, x_p)}{g_{p-1,p}^{(k)}(x_{p-1}, x_p)} \right),$$

and denote the corresponding coefficient by $\boldsymbol{\beta}^{(k)} = (\beta_{12}^{(k)}, \ldots, \beta_{p-1,p}^{(k)})^T, k = 1, \ldots, K-1$. Then the log-likelihood of the k-th logit model is

$$l(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{n} [I(Y_i \leq k) \log(p_i^{(k)}) + I(Y_i > k) \log(1 - p_i^{(k)})], \tag{4}$$

where

$$p_i^{(k)} = P(Y_i \leq k | \mathbf{w}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{z}_i^T \boldsymbol{\alpha}^{(k)} + \mathbf{w}_i^{(k)} \boldsymbol{\beta}^{(k)})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\alpha}^{(k)} + \mathbf{w}_i^{(k)} \boldsymbol{\beta}^{(k)})}, \tag{5}$$

$$1 - p_i^{(k)} = P(Y_i > k | \mathbf{w}_i, \mathbf{z}_i) = \frac{1}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\alpha}^{(k)} + \mathbf{w}_i^{(k)} \boldsymbol{\beta}^{(k)})}. \tag{6}$$

Note that both $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$ and densities $f_{st}^{(k)}, g_{st}^{(k)}$ are unknown, so a two-stage estimation strategy is used: splitting the training data into two parts, first we use one part to fit $f_{st}^{(k)}$ and $g_{st}^{(k)}$ by kernel density estimation, next we evaluate the densities on the other part to fit the model and get the estimation of $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$. Because the number of gene pairs can

be much larger than the sample size in high-dimensional settings, L_1-penalty (Tibshirani, 1996) is adopted in the estimation. Therefore, the solution of unknown parameters is

$$(\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)}) = \underset{\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}}{\operatorname{argmax}} (l(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}) - \lambda \sum_{s,t} |\beta_{st}^{(k)}|). \tag{7}$$

Tuning parameter λ could be selected by minimizing AIC, BIC, or cross-validation error, etc. Parameter $\beta_{st}^{(k)} \neq 0$ indicates X_s, X_t have differential dependency patterns between $Y \leq k$ and $Y > k$.

2.3 Prediction

Given the estimated joint densities $\hat{f}_{st}^{(k)}$, $\hat{g}_{st}^{(k)}$ and parameters $\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)}$ for the k-th logit model, the predicted probability of the i-th subject that belongs to $Y \leq k$, i.e., $\hat{P}(Y_i \leq k|\mathbf{x}_i, \mathbf{z}_i)$, could be calculated by equation (5). And correspondingly, $\hat{P}(Y_i > k|\mathbf{x}_i, \mathbf{z}_i) = 1 - \hat{P}(Y_i \leq k|\mathbf{x}_i, \mathbf{z}_i)$. Define

$$\begin{aligned} p_{i1}^{(k)} &= \cdots = p_{ik}^{(k)} = \hat{P}(Y_i \leq k|\mathbf{x}_i, \mathbf{z}_i), \\ p_{i(k+1)}^{(k)} &= \cdots = p_{iK}^{(k)} = 1 - \hat{P}(Y_i \leq k|\mathbf{x}_i, \mathbf{z}_i). \end{aligned}$$

$p_{im}^{(k)}$ is the score representing how likely Y_i is predicted to be m in the k-th logit model. For subject i, we can calculate $p_{im}^{(k)}$ from all $K-1$ logit models and aggregate the scores in each class, as illustrated below:

Class	1	2	...	K
Model1	$p_{i1}^{(1)}$	$p_{i2}^{(1)}$	...	$p_{iK}^{(1)}$
Model2	$p_{i1}^{(2)}$	$p_{i2}^{(2)}$	...	$p_{iK}^{(2)}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Model$K-1$	$p_{i1}^{(K-1)}$	$p_{i2}^{(K-1)}$	...	$p_{iK}^{(K-1)}$
Add up	$\sum_{k=1}^{K-1} p_{i1}^{(k)}$	$\sum_{k=1}^{K-1} p_{i2}^{(k)}$	...	$\sum_{k=1}^{K-1} p_{iK}^{(k)}$.

(8)

Thus the predicted outcome $\hat{y}_i$ is determined by the one maximizing aggregated scores, i.e.,

$$\hat{y}_i = \underset{m}{\operatorname{argmax}} \sum_{k=1}^{K-1} p_{im}^{(k)}.$$

2.4 Selection of differential pairs

A pair of covariates (such as genes) is selected if any of its coefficients in the $K-1$ models is non-zero. Specifically, define the differential dependency weight by

$$w_{st} = \sum_{l=1}^{L} I\left(\sum_{k=1}^{K-1} |\hat{\beta}_{st,l}^{(k)}| \neq 0\right), \tag{9}$$

where $\hat{\beta}_{st,l}^{(k)}$ is the estimated regression coefficient from the k-th logit model on the l-th repetition, L is the number of splitting, $l = 1, \ldots, L$. $I(\sum_{k=1}^{K-1} |\hat{\beta}_{st,l}^{(k)}| \neq 0)$ indicates whether the log density ratio of (X_s, X_t) has any non-zero coefficient in one of $K-1$ logit models. If there exists a k such that $\hat{\beta}_{st,l}^{(k)} \neq 0$, then $I(\sum_{k=1}^{K-1} |\hat{\beta}_{st,l}^{(k)}| \neq 0) = 1$, otherwise

$I(\sum_{k=1}^{K-1} |\hat{\beta}_{st,l}^{(k)}| \neq 0) = 0$. A set of highly discriminative gene pairs would be selected by specifying a cutoff for differential dependency weights.

The following algorithm summarizes the whole procedure for estimation, prediction and selection of differential pairs:

Step 1 Randomly split the training set into two parts: (D_l, D_l^c);

Step 2 In part D_l, estimate $f_{st}^{(k)}, g_{st}^{(k)}$ by bivariate kernel density estimation, $s = 1, \ldots, p-1, t = s+1, \ldots, p, k = 1, \ldots, K-1$;

Step 3 In part D_l^c, evaluate bivariate densities at observed values in D_l^c using $\hat{f}_{st}^{(k)}, \hat{g}_{st}^{(k)}$ fitted in **Step 2**, then fit each logit model in (2) with the fitted densities and observed covariates to get coefficients' estimate $\hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)}$. Each logit model could be fit in `R` with packages `glmnet` (Freidman et al., 2010) or `glmpath` (Park and Hastie, 2018);

Step 4 In testing set, evaluate bivariate densities using $\hat{f}_{st}^{(k)}, \hat{g}_{st}^{(k)}$ fitted in **Step 2**. Plug the estimated densities and other observed covariates into the corresponding logit models fitted in **Step 3**, and calculate $\hat{P}(Y_i \leq k | \mathbf{x}_i, \mathbf{z}_i)$. Subsequently, calculate $p_{im,l}^{(k)}$ for each subject i in the testing set on the l-th repetition, as shown in (8);

Step 5 Repeat **Step 1-Step 4** for L times. For subject i in the testing set, the predicted class label(outcome) is $\hat{y}_i = \text{argmax}_m \sum_{l=1}^{L} \sum_{k=1}^{K-1} p_{im,l}^{(k)}$. Also, a set of highly discriminative gene pairs can then be selected with differential dependency weights w_{st} as defined in equation (9).

3. Numerical studies

The proposed method has two aims: predicting class labels for new observations and identifying important differentially correlated gene pairs. Therefore, in this section, we examine our method's performance through the accuracy of classification and differential pair selection. Besides, our approach will be compared with ordinal FANS, which doesn't take into account possible interactions.

In each replication, generated data is randomly divided into training set (with sample size n_{train}) and testing set (with sample size n_{test}). Model fitting and differential pair selection are conducted in the training set, while the prediction of class labels is evaluated in the testing set. Two simulation scenarios are considered:

Scenario 1: Set $K = 3, p = 20, q = 3$, generate

- y from $\{1, 2, 3\}$;
- $\mathbf{x} \sim N_p(\boldsymbol{\mu}_x^{(k)}, \Sigma^{(k)})$, where $\boldsymbol{\mu}_x^{(1)} = 0 \times 1_{20}, \boldsymbol{\mu}_x^{(2)} = (0 \times 1_{14}, 0.05 \times 1_6), \boldsymbol{\mu}_x^{(3)} = (0 \times 1_{14}, 0.1 \times 1_6)$. Let $\Sigma^{(k)} = diag(\Sigma_1, \Sigma_2^{(k)}, \Sigma_3^{(k)})$, where $\Sigma_1 = (\sigma_{ij})_{14 \times 14}, \sigma_{ii} = 1, \sigma_{ij} = 0.5$ for $i \neq j, i, j = 1, \ldots, 14$. $\Sigma_2^{(k)} = \Sigma_3^{(k)}$, and

$$\Sigma_2^{(1)} = \begin{pmatrix} 1 & -0.9 & 0.9 \\ -0.9 & 1 & -0.9 \\ 0.9 & -0.9 & 1 \end{pmatrix}, \Sigma_2^{(2)} = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{pmatrix},$$

$$\Sigma_2^{(3)} = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix};$$

- $\mathbf{z} \sim N_q(\boldsymbol{\mu}_z^{(k)}, I_q)$, where $\boldsymbol{\mu}_z^{(1)} = 0.5 \times 1_3, \boldsymbol{\mu}_z^{(2)} = 1_3, \boldsymbol{\mu}_z^{(3)} = 1.5 \times 1_3$;
- $n_{train} = 450, n_{test} = 150$, random splits $L = 20$, replicate 100 times.

Scenario 2: Set $K = 4, p = 10, q = 3$, generate

- y from $\{1, 2, 3, 4\}$;
- $\mathbf{x} \sim N_p(\boldsymbol{\mu}_x^{(k)}, \Sigma^{(k)})$, where $\boldsymbol{\mu}_x^{(1)} = 0 \times 1_{10}, \boldsymbol{\mu}_x^{(2)} = (0 \times 1_4, 0.15 \times 1_6), \boldsymbol{\mu}_x^{(3)} = (0 \times 1_4, 0.3 \times 1_6), \boldsymbol{\mu}_x^{(4)} = (0 \times 1_4, 0.45 \times 1_6)$. Let $\Sigma^{(k)} = diag(\Sigma_1, \Sigma_2^{(k)}, \Sigma_3^{(k)})$, where $\Sigma_1 = (\sigma_{ij})_{4\times4}, \sigma_{ii} = 1, \sigma_{ij} = 0.5$ for $i \neq j, i, j = 1, \ldots, 4$. $\Sigma_2^{(k)} = \Sigma_3^{(k)}$, and

$$\Sigma_2^{(1)} = \begin{pmatrix} 1 & -0.9 & 0.9 \\ -0.9 & 1 & -0.9 \\ 0.9 & -0.9 & 1 \end{pmatrix}, \Sigma_2^{(2)} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix},$$

$$\Sigma_2^{(3)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Sigma_2^{(4)} = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix};$$

- $\mathbf{z} \sim N_q(\boldsymbol{\mu}_z^{(k)}, I_q)$, where $\boldsymbol{\mu}_z^{(1)} = 0.5 \times 1_3, \boldsymbol{\mu}_z^{(2)} = 1_3, \boldsymbol{\mu}_z^{(3)} = 1.5 \times 1_3, \boldsymbol{\mu}_z^{(4)} = 2 \times 1_3$;
- $n_{train} = 600, n_{test} = 200$, random splits $L = 20$, replicate 100 times.

Classification is assessed via misclassification rate, class-specific misclassification rate, Somers' index. Misclassification rate is define as

$$\frac{\#\{\hat{y}_i \neq y_i\}}{n_{test}},$$

and class-specific misclassification rate is

$$1 - \frac{\#\{\hat{y}_i = k \text{ and } y_i = k\}}{\#\{y_i = k\}}.$$

The higher the misclassification rate, the worse the classification accuracy. Somers' index D (range from -1 to 1) measures the concordance between two ordinal variables, which in our case are Y and $\hat{Y}$ (Somers, 1962). '-1' means disagreement, '1' means perfect agreement. The closer $D_{Y\hat{Y}}$ is to 1, the better the prediction is. Differential pair selection is examined by true discovery rate(TDR), true positive rate(TPR) and true negative rate(TNR) that are defined as

$$TDR = \frac{\sum_{i\neq j} I(\delta_{ij}\hat{\delta}_{ij} \neq 0)}{\sum_{i\neq j} I(\hat{\delta}_{ij} \neq 0)},$$

$$TPR = \frac{\sum_{i\neq j} I(\delta_{ij}\hat{\delta}_{ij} \neq 0)}{\sum_{i\neq j} I(\delta_{ij} \neq 0)},$$

$$TNR = \frac{\sum_{i\neq j} I(\delta_{ij} = 0)I(\hat{\delta}_{ij} = 0)}{\sum_{i\neq j} I(\delta_{ij} = 0)}.$$

$(\delta_{ij})_{p\times p}$ is the differential adjacency matrix, $\delta_{ij} \neq 0$ indicate the pair (X_i, X_j) are deferentially dependent between at least two groups; $(\hat{\delta}_{ij})_{p\times p}$ is the estimated differential

adjacency matrix. The closer TDR, TPR, TNR are to 1, the better the performance of identifying differential pairs is.

Table 1 shows the performance of our approach compared with ordinal FANS under two scenarios. Generally, our method has lower misclassification rate than ordinal FANS. Specifically, in scenario 1, ordinal FANS predicted all observations to $Y = 2$; in scenario 2, ordinal FANS almost classified all observations to $Y = 2$ and $Y = 3$. Our approach has Somers' index $D_{Y\hat{Y}}$ closer to 1 than ordinal FANS, which suggests that our approach has considerably higher classification accuracy than ordinal FANS. These results indicate that ordinal FANS failed to classify observations in these scenarios since it didn't capture the interaction between covariates. Our method is demonstrated to perform better in prediction when interaction plays an important role. In addition, our approach has TDR, TPR and TNR close to 1, which shows that truly differential gene pairs could be efficiently selected by our approach.

Table 1: Performance of our approach and ordinal FANS under two scenarios on average

		Our approach	Ordinal FANS
Scenario 1	Misclassification rate	0.23	0.64
	Class-specific misclassification rate	(0.24, 0.20, 0.25)	(1.00, 0.00, 1.00)
	Somer's $D_{Y\hat{Y}}$	0.74	0.00
	TDR	1.00	-
	TPR	1.00	-
	TNR	1.00	-
Scenario 2	Misclassification rate	0.37	0.73
	Class-specific misclassification rate	(0.13, 0.43, 0.48, 0.43)	(1.00, 0.49, 0.46, 0.99)
	Somer's $D_{Y\hat{Y}}$	0.68	0.19
	TDR	1.00	-
	TPR	0.97	-
	TNR	1.00	-

4. Discussion

Many complex diseases are actually categorized into several stages, which can be considered as ordinal outcomes. Existing prediction models for ordinal outcomes generally only consider the effect of each individual predictor and simple multiplicative interactions, i.e., linear terms of form x_1x_2. But the intrinsic mechanism of these diseases always involves intricate interactions between biomolecules, instead of individual biomolecules. In this paper, we presented a predictive modeling approach incorporating nonparametric interactions as predictors for ordinal outcomes. Interactions are evaluated by logarithms of bivariate density ratios, which are fitted without imposing any assumption of the densities of original covariates. Lasso penalty is used so that highly differential pairs could be selected. The splitting procedure makes the result more stable, and maximizes the utility of limited data. Numerical studies indicate that our approach has higher prediction accuracy than ordinal FANS which didn't consider interaction. The ignorance of interactions could reduce prediction accuracy substantially. Also, our approach shows great performance in selecting important differential pairs.

Our proposed approach has two primary applications. First, after fitting model to the observed data, our approach can predict group labels or determine disease stages on new data. For example, new patients can be classified into a disease category using our model with genetic, genomics and clinical information. Our approach might be used to segment patients into subgroups. Efficient treatment may be developed based on the mechanism suggested by the fitted model and assigned to each subgroup of patients. Also, our approach identifies important differentially expressed gene pairs, which may assist and facilitate the investigation of the etiology of a complex disease, from the perspective of how genes function with each other in the development of a disease. However, one potential concern is the high computational cost in real applications in the presence of a large number of candidate genes. Pre-screening of genes might be needed, such as focusing on one specific disease-relevant pathway first. By splitting the optimization problem into subproblems, parallel computation might be employed in real applications.

Acknowledgments

This research is partially supported by the NIH grants P50CA225450, P30CA016087, P30AG066512, P30ES000260, P01AG060882. The authors would like to thank Dr. Yong He for useful communication and conversations.

Conflict of Interest: None declared.

REFERENCES

Fan, J., Feng, Y., Jiang, J., & Tong, X. (2016), "Feature Augmentation via Nonparametrics and Selection (FANS) in high-dimensional classification," *Journal of the American Statistical Association*, 111(513), 275-287.

Ferber, K. L. (2016), "Methods for Predicting an Ordinal Response with High-Throughput Genomic Data." *PhD Thesis*.

Fontanella, S., Frainay, C., Murray, C. S., Simpson, A., & Custovic, A. (2018), "Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort," *PLoS medicine*, 15(11), e1002691.

Friedman, J., Hastie, T., & Tibshirani, R. (2001), *The elements of statistical learning (Vol. 1, No. 10)*. New York: Springer series in statistics.

Friedman, J., Hastie, T., & Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33(1), 1.

He, X., Sun, X., & Shao, Y. (2020), "Network-based survival analysis to discover target genes for developing cancer immunotherapies and predicting patient survival," *Journal of Applied Statistics*, 1-22.

Ji, J., He, D., Feng, Y., He, Y., Xue, F., & Xie, L. (2017), "JDINAC: joint density-based non-parametric differential interaction network analysis and classification using high-dimensional sparse omics data," *Bioinformatics*, 33(19), 3080-3087.

Liu, M., Rogers, L., Cheng, Q., Shao, Y., Fernandez-Beros, M. E., Hirschhorn, J. N., ... & Seldin, M. F. (2011), "Genetic variants of TSLP and asthma in an admixed urban population," *PloS one*, 6(9), e25099.

Park, M. and Hastie, T. (2018), "glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model," R package version 0.98.

Patrawalla, P., Kazeros, A., Rogers, L., Shao, Y., Liu, M., Fernandez-Beros, M. E., ... & Reibman, J. (2012), "Application of the asthma phenotype algorithm from the Severe Asthma Research Program to an urban population," *PloS one*, 7(9), e44540.

Shojaie, A. (2020), "Differential network analysis: A statistical perspective," *Wiley Interdisciplinary Reviews: Computational Statistics*, e1508.

Somers, R. H. (1962), "A new asymmetric measure of association for ordinal variables," *American sociological review*, 799-811.

Sun, X., Liu, X., Xia, M., Shao, Y., & Zhang, X. D. (2019), "Multicellular gene network analysis identifies a macrophage-related gene signature predictive of therapeutic response and prognosis of gliomas," *Journal of translational medicine*, 17(1), 159.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.