High-dimensional variable selection for ordinal outcomes with error control

Han Fu and Kellie J. Archer

Corresponding author: Kellie J. Archer, The Ohio State University, Columbus, OH 43210, USA. Tel.: +1-614-247-6167; Fax: +1-614-247-1846; E-mail: archer.43@osu.edu

Abstract

Many high-throughput genomic applications involve a large set of potential covariates and a response which is frequently measured on an ordinal scale, and it is crucial to identify which variables are truly associated with the response. Effectively controlling the false discovery rate (FDR) without sacrificing power has been a major challenge in variable selection research. This study reviews two existing variable selection frameworks, model-X knockoffs and a modified version of reference distribution variable selection (RDVS), both of which utilize artificial variables as benchmarks for decision making. Model-X knockoffs constructs a ‘knockoff’ variable for each covariate to mimic the covariance structure, while RDVS generates only one null variable and forms a reference distribution by performing multiple runs of model fitting. Herein, we describe how different importance measures for ordinal responses can be constructed that fit into these two selection frameworks, using either penalized regression or machine learning techniques. We compared these measures in terms of the FDR and power using simulated data. Moreover, we applied these two frameworks to high-throughput methylation data for identifying features associated with the progression from normal liver tissue to hepatocellular carcinoma to further compare and contrast their performances.

Key words: false discovery rate; ordinal regression; knockoff filter; L_1 penalization; boosting; ordinal forests

Introduction

Ordinal measurement scales are commonly employed in the medical literature [1]. Health status and patient outcomes are often evaluated on an ordinal scale [2]. For example, in a hepatocellular carcinoma (HCC) study, liver tissue samples can be classified into one of three ordinal class categories including normal, cirrhotic without concomitant cancer and cirrhotic with HCC [3]. In order to better understand disease mechanisms, some genomic applications have been conducted to identify important genomic features whose characteristics, such as methylation or expression levels, are associated with a certain clinical response. Such applications often possess two properties, high dimensionality and sparsity, which means that a massive number of features (far more than the sample size) are collected but we suspect only a small portion of them are truly associated with

the outcome. Thus, effective variable selection procedures for high-dimensional data are needed to select a set of candidate variables for further confirmatory investigation.

Various high-dimensional variable selection or screening methods have been proposed and widely used during the past decades. Examples include the least-absolute shrinkage and selection operator (LASSO) [4], boosting [5] and sure independence screening [6]. These methods focus mainly on providing a small set of potentially relevant variables, but not necessarily the ‘smallest’. In other words, the set provided by these methods is very likely to contain redundant variables or noise variables, which are also known as false positives. Controlling the false discovery rate (FDR), the expected proportion of false positives among all discoveries, was initially suggested by Benjamini and Hochberg [7] and has quickly established itself in the statistics literature. As the canonical approach to FDR control,

Han Fu is a PhD student in biostatistics at the Ohio State University. Her main research interest lies in cancer genomics, high-dimensional data analysis, and precision medicine.

Kellie J. Archer is a professor and the chair of the Division of Biostatistics at The Ohio State University. Her primary research area has been developing statistical methods for analyzing high-throughput genomic data.

Submitted: 23 September 2019; Received (in revised form): 6 January 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

the Benjamini–Hochberg (BH) procedure [7] requires a set of valid P -values, and it controls the FDR exactly when the P -values are independent or have positive regression dependency on a subset [8]. A modified version, the Benjamini–Yekutieli (BY) procedure, controls the FDR when the P -values have arbitrary forms of dependence [8], but it may suffer from power loss compared with BH [9]. Both procedures assume that valid P -values can be obtained, which may be difficult for high-dimensional conditional modeling problems. Using marginal p -values from feature-by-feature hypothesis tests is a possible solution for high-dimensional data, but power and interpretability are sacrificed to some extent in most cases [10].

The model-X knockoff approach [10] was recently developed as a general variable selection framework with exact finite-sample FDR control that can be applied to high-dimensional conditional testing problems. The framework generalizes the original knockoff procedure [9] designed for linear models to arbitrary nonlinear relationships between the response and the covariates, without relying on P -value calculation. Briefly, a set of ‘knockoff’ copies are constructed within the framework to mirror the structure of the original covariates while remaining independent of the response. The knockoff variables can thus be used as negative controls for the real covariates, so that true signals will be teased apart from noise variables to ensure FDR control. Both Gaussian and binomial responses were considered in the original article [10], and although no special attention was paid to ordinal outcomes, the framework still applies. Another framework that shares similar underlying philosophies with knockoffs, the reference distribution variable selection (RDVS) method [11], generates one artificial null variable to form an empirical null distribution for making inferences as in the hypothesis testing setting. The RDVS was originally described as a Bayesian screening approach using a Gaussian spatial process model, introduced in the field of computer experiments where a limited number of covariates are typically considered, but it naturally extends to a general selection framework that can accommodate arbitrary outcome types and data dimensionality. As with the model-X knockoff approach, the original RDVS method did not accommodate ordinal response data but nonetheless can be used in ordinal response modeling.

In this article, we review the model-X knockoff and reference distribution method frameworks and examine their performances with respect to FDR error control for ordinal response models in comparison to their corresponding traditional base learners that do not apply a framework. Specifically, we describe multiple variable importance measures for ordinal outcomes using various techniques, including L_1 penalized regression, boosting and ordinal forests, to make both frameworks applicable to ordinal response data. The remainder of the article is organized as follows. Section 2 reviews the two variable selection frameworks as well as the importance measures that can be used for ordinal outcomes. Section 2 also describes the simulation study designed to compare various base learners with and without using these frameworks. Results from both the simulation study and the case application are reported in Section 3. Section 4 concludes the article with a discussion.

Methods

Problem statement

Consider the problem where we have n independent observations and p predictors ($p > n$) for an ordinal response Y , which takes on one of K ordinal levels. Let $X = (X_1, \dots, X_p)$ denote the

random vector of the predictors and X denote the $n \times p$ design matrix. We would like to identify the subset of predictors that are truly associated with Y while making as few false discoveries as possible. Strictly speaking, our main objective is to identify the ‘smallest’ subset S such that given covariates in the subset, Y is conditionally independent of all other variables. The subset S is called the Markov blanket in the graphical model literature [12]. The variables in the subset S , $\{X_j\}_{j \in S}$, are truly associated with the response and those not in S , $\{X_j\}_{j \notin S}$, are not associated with the response (null).

The model-X knockoff method tests conditional hypotheses that are mathematically equivalent to the objective stated above under mild conditions [10] (see Section 2.3.1 for details), and thus addresses the problem of interest precisely. In contrast, the RDVS framework tests hypotheses that are of a marginal flavor because it compares each single statistic to an empirical reference distribution separately. In this regard, the RDVS is not adequately addressing the problem proposed. However, it pools information across predictors by utilizing statistics from multivariable models, instead of looking at one variable at a time. We therefore want to examine its performance in solving the conditional problem, i.e. identifying variables in the Markov blanket given the outcome is independent of others.

Model assessment

FDR is a commonly used method to quantify false positive findings nowadays. This section presents the concept of the FDR and power in the context of variable selection. Suppose we have a selection method that identifies a subset of the p predictors, denoted by $\hat{S}$, and claims that $\{X_j\}_{j \in \hat{S}}$ are significant while $\{X_j\}_{j \notin \hat{S}}$ are null. Denote the number of truly null variables that are claimed to be significant (i.e. false positives or Type I errors) by n_{01} , the number of truly null variables claimed to be null by n_{00} , the number of truly associated variables claimed to be significant by n_{11} , and the number of truly associated variables claimed to be null by n_{10} . In the language of sets, $n_{00} = \#\{j : j \in (S \cup \hat{S})^c\}$, $n_{01} = \#\{j : j \in \hat{S} \setminus S\}$, $n_{10} = \#\{j : j \in S \setminus \hat{S}\}$ and $n_{11} = \#\{j : j \in (\hat{S} \cap S)\}$, where S is the Markov blanket. Then, the false discovery proportion (FDP) is defined as the proportion of false positives among all discoveries, i.e. $n_{01}/(n_{01} + n_{11})$. The FDR is the expected value of the FDP, and we can use the FDP to estimate the FDR. The power (or sensitivity), on the other hand, can be estimated by the proportion of true signals being identified, defined as $n_{11}/(n_{11} + n_{10})$. In the ideal case, we would discover all true signals without false positives, such that the FDR is zero and power is one.

Selection frameworks

Model-X knockoffs

The knockoff framework [10] estimates the covariate distribution F_X from the observations, which are assumed to be drawn from some joint distribution of (X, Y) . It tests the hypotheses of conditional independence, i.e. whether X_j is independent of Y conditional on all other variables. A false positive under this framework is thus defined by selecting a variable X_j that is independent of Y given the others. In the special case of generalized linear models, the hypotheses above are equivalent with the traditional parametric hypotheses—whether the regression coefficient is zero or not. Under weak regularity conditions, the set of relevant variables defined by the hypotheses is guaranteed to be unique, and coincides with the Markov blanket S described in Section 2.1 according to the graphical model literature [13].

Specifically, the knockoff method [10] requires the construction of knockoff variables $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_p)$, which satisfy two properties:

- (1) Pairwise exchangeability: for any subset $B \subset \{1, \dots, p\}$, $(X, \tilde{X})_{\text{swap}(B)} \stackrel{d}{\longleftrightarrow} (X, \tilde{X})$, i.e. swapping the entries X_j and $\tilde{X}_j$ for each $j \in B$ leaves the joint distribution invariant;
- (2) Conditional independence: $\tilde{X} \perp Y \mid X$.

Given these two properties, knockoff variables are constructed to mimic the covariance structure of the original covariates while remaining as different as possible from the corresponding covariate to boost the power. Since knockoff variables are conditionally independent of the response, they are able to serve as benchmarks to make controlled variable selection possible. In this article, knockoffs are constructed using the second-order approximate semidefinite program approach by Candès *et al.* [10] with the *knockoff* R package [14], assuming that the covariates follow a multivariate normal distribution. A James-Stein-type shrinkage estimator for the covariance matrix is used to produce a positive-definite and well-conditioned matrix. All covariates are centered and scaled prior to the analysis.

After generating the knockoff variables, the original design matrix is then augmented with the knockoffs to form an extended $n \times 2p$ matrix $[X, \tilde{X}]$. Some variable importance measures designed for ordinal responses, which will be described in Section 2.4, can be used to assess the importance of the $2p$ features in terms of explaining the ordinal outcome Y . The importance measures, Z_j , may be based on a fitted-regression model, such as the absolute value of the regression coefficients, or using side-effects that result from machine learning techniques, such as risk reduction in boosting. Since only the general selection frameworks are described in this section, we suppose, for now, that we have decided on a certain variable importance measure and obtain the $2p$ measurements accordingly, denoted by $Z_1, \dots, Z_{2p}$. Then, the statistic for X_j used for variable selection, for $j \in \{1, \dots, p\}$, is calculated as the difference in importance measurements between the original variable and its knockoff variable, denoted by $W_j = Z_j - Z_{j+p}$. Other statistics are acceptable if they attain the flip-sign property required by model-X knockoffs [10], which says that swapping a variable with its knockoff changes the sign of the statistic. A large positive value typically provides evidence of an important variable, while the statistics for null variables tend to have symmetric distributions around zero with small magnitudes. In this article, only the differences in importance measurements are considered.

After computing importance difference statistics W_j for $j \in \{1, \dots, p\}$, the following procedure guarantees strict FDR control according to [10]. A data-dependent threshold is first given by

$$\tau = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q \right\} \quad (1)$$

where q denotes the target FDR level. Then, the variables whose statistic exceeds the threshold will be selected. Since the model-X knockoff framework places no restriction on the data dimensionality or the conditional distribution of $Y \mid X$, the framework is known in theory to apply seamlessly to high-dimensional ordinal response data with FDR control, as long as appropriate importance measures are available.

Reference distribution variable selection

RDVS was originally described as a Bayesian screening approach designed exclusively for Gaussian processes [11] but naturally extends to a general selection framework, which can accommodate arbitrary outcome types and feature importance measures. The framework requires us to generate only one new 'null' variable $X^\dagger$ for n observations. Ideally, the null variable should be orthogonal to all variables in the design space, so that it will not interfere with the inference of the original variables. In practice, however, it would be difficult to identify such a variable and one may not exist. To address this, the RDVS method randomly samples from the design space of X multiple times (M) after centering and scaling all covariates, so that there are M realizations of the null variable, denoted by $X^{\dagger(1)}, \dots, X^{\dagger(M)}$. Thus, the potential correlation with original variables can be reduced through an averaging effect. In other words, although only one null variable needs to be generated each time, multiple replications of null variable generation and model fitting are required.

The specific procedure is described as follows. In the m th run where a new $X^{\dagger(m)}$ is sampled, for example, from a standard normal distribution, $m = 1, \dots, M$, we have an augmented $n \times (p+1)$ design matrix $[X, X^{\dagger(m)}]$. Again, we suppose that we have decided on a certain variable importance measure in terms of explaining the response Y and obtain the $(p+1)$ measurements accordingly, denoted by $Z_1^{(m)}, \dots, Z_{p+1}^{(m)}$, where $Z_{p+1}^{(m)}$ corresponds to $X^{\dagger(m)}$. For example, we can fit Y against $[X, X^{\dagger(m)}]$ using some generalized linear model and use the absolute values of regression coefficients as the importance measurements $Z_1^{(m)}, \dots, Z_{p+1}^{(m)}$. Repeating this step M times, we have M observations of the feature importance measurements for every single variable including the null variable $X^\dagger$. For the p original variables, the sample mean of M measurements,

$$W_j = \frac{1}{M} \sum_{m=1}^M Z_j^{(m)}, j \in \{1, \dots, p\} \quad (2)$$

can be used as a summary statistic reflecting relative variable importance. The M measurements for $X^\dagger$, i.e. $Z_{p+1}^{(1)}, \dots, Z_{p+1}^{(M)}$, are combined to form a reference distribution for the null variable, against which the relative variable importance can be assessed. The RDVS method then uses this empirical distribution as a null distribution and borrows ideas from one-sided hypothesis testing. Assuming a larger statistic indicates a more important variable, we can calculate empirical p -values using the proportion of times the samples in the reference distribution exceed the observed statistic W_j , given by

$$pval_j = \# \{m : Z_{p+1}^{(m)} > W_j\} / M, j \in \{1, \dots, p\} \quad (3)$$

Considering the potentially strong dependency among the resulting empirical P -values, the BY procedure [8] can be applied to adjust the P -values for approximate FDR control. The variables with values smaller than the target FDR level q will be selected.

In the spirit of one-sided hypothesis testing, the RDVS framework is roughly testing whether the statistic W_j for the variable j comes from the same distribution as that of the artificial variable $X^\dagger$. However, the hypotheses proposed may not be valid under certain circumstances where the covariates are not independent of one another. The variable $X^\dagger$ is known to be independent of both $X = (X_1, \dots, X_p)$ and Y due to the way it is created; while any truly null variable(s) in X are conditionally independent of Y , they

Table 1. The similarity and difference between the procedure of the knockoff framework and that of the RDVS framework

	Knockoffs	RDVS
Augmentation	$\mathbf{X}^* = [\mathbf{X}, \tilde{\mathbf{X}}]_{(n \times 2p)}$	$\mathbf{X}^* = [\mathbf{X}, \mathbf{X}^\dagger]_{(n \times (p+1))}$
Importance measurements	$Z_1, \dots, Z_{2p}$	$Z_1^{(1)}, \dots, Z_{p+1}^{(1)}, \dots, Z_1^{(M)}, \dots, Z_{p+1}^{(M)}$
Statistics	$W_j = Z_j - Z_{j+p}$	$W_j = \frac{1}{M} \sum_{m=1}^M Z_j^{(m)}$
Selection	$W_j > \min \left\{ t > 0 : \frac{1 + \#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}} \leq q \right\}$	Benjamini–Yekutieli adjusted $p\text{val}_j < q$, where $p\text{val}_j$ = empirical P-values from comparing W_j against $\{Z_{p+1}^{(1)}, \dots, Z_{p+1}^{(M)}\}$

may be dependent on some significant variables in the Markov blanket S . In this case, the truly null variables will tend to have higher probabilities of being selected than the artificial $\mathbf{X}^\dagger$, and hence the empirical distribution of the importance measure for $\mathbf{X}^\dagger$ cannot well represent the null distribution anymore. In contrast, the knockoff framework preserves the covariance structure of the original variables, which makes it more suitable for the case of dependent covariates. A recent article [15] explained and empirically demonstrated the advantage of the knockoffs in providing the correct reference distribution over other augmentation methods including adding independent dummy variables as in the RDVS.

As a summary, Table 1 presents the similarity and difference between the procedures of the two frameworks. They have similar procedures in general which makes them comparable, yet there are disparities between the two frameworks in each specific step.

Variable importance measures

Several variable importance measures designed for ordinal outcomes are described in this section. The design matrix used in this section, denoted by $\mathbf{X}^*$, represents the $n \times 2p$ augmented design matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$ for the knockoff framework, the $n \times (p+1)$ augmented design matrix $[\mathbf{X}, \mathbf{X}^\dagger]$ for the RDVS and the original (centered and scaled) design matrix $\mathbf{X}$ if no selection frameworks were applied. The following measures described can be used to assess the importance of each variable in $\mathbf{X}^*$ in fitting or predicting the response Y .

Measures based on continuation ratio models with L_1 penalization

The continuation ratio (CR) model is a type of generalized linear model for ordinal outcomes. The link function is the logit of a conditional probability, formulated as below

$$\begin{aligned} \text{logit}(P(Y = k | Y \leq k, \mathbf{X}^* = \mathbf{x})) &= \log \left(\frac{P(Y = k | Y \leq k, \mathbf{X}^* = \mathbf{x})}{P(Y < k | Y \leq k, \mathbf{X}^* = \mathbf{x})} \right) \\ &= \alpha_k + \beta^T \mathbf{x} \end{aligned} \quad (4)$$

where $k \in \{2, \dots, K\}$. The CR model can be fitted by augmenting the dataset to represent $K - 1$ conditionally independent binary response datasets, and thus standard techniques for modeling binary data can be applied. In addition, the CR model is frequently used when progression through the ordinal levels cannot be reversed, such as stage of cancer [16].

For the purpose of variable selection, Archer and Williams [2] proposed the L_1 penalized CR model, which added the L_1 penalization to the likelihood for the CR model, as in the LASSO.

The model is given by

$$\arg\max_{\beta} \left(L(\beta | y, \mathbf{x}) - \lambda \sum_{l=1}^p |\beta_l| \right) \quad (5)$$

where $L(\beta | y, \mathbf{x})$ is the likelihood function for a regular CR model and λ is the regularization parameter. In practice, the penalized CR model was fitted by augmenting the dataset then applying techniques for solving the logistic LASSO with the *glmnet* R package.

Therefore, there are two natural approaches for obtaining variable importance measures based on the L_1 penalized CR model. One uses the absolute value of coefficients resulting from fitting the response Y against the design matrix $\mathbf{X}^*$. The regularization parameter λ can be selected from a grid of values by minimizing either the AIC or BIC. Then, the importance measure for X_j can be expressed by

$$Z_j^{\text{AIC}} = |\hat{\beta}_j^{\text{AIC}}| \quad (6)$$

or

$$Z_j^{\text{BIC}} = |\hat{\beta}_j^{\text{BIC}}| \quad (7)$$

where $\hat{\beta}_j^{\text{AIC}}$ and $\hat{\beta}_j^{\text{BIC}}$ are the estimated coefficients for X_j from the final model tuned by AIC or BIC, respectively, for $j \in \{1, \dots, p^*\}$ (p^* is the number of columns in $\mathbf{X}^*$ and takes the value of p , $p+1$, or $2p$ depending on whether a traditional, RDVS, or model-X knockoff framework is used). A larger Z_j^{AIC} or Z_j^{BIC} indicates a more important feature. The regularization parameter λ can be also tuned by cross-validation (CV). Specifically, we choose λ that leads to the smallest 5-fold cross-validated classification error from a reasonable grid of 100 values. The coefficients corresponding to the optimal λ tuned by CV can be thus used to measure importance, expressed by

$$Z_j^{\text{CV}} = |\hat{\beta}_j^{\text{CV}}| \quad (8)$$

for $j \in \{1, \dots, p^*\}$.

Another variable importance measure can be based on the solution path, i.e. the penalty at which variable j first enters the model. The formal definition is

$$Z_j^{\text{path}} = \sup \left\{ \lambda : \hat{\beta}_j(\lambda) \neq 0 \right\}, \quad (9)$$

for $j \in \{1, \dots, p^*\}$. A larger Z_j^{path} means that the variable X_j enters the model earlier and thus is more important.

Risk reduction in boosting

Boosting is a machine learning ensemble method, which converts weak learners to strong learners to increase predictive ability and reduce over-fitting. Model-based boosting can use component-wise least-squares estimates as base-learners to fit generalized linear models, such as the cumulative logit (CL) model, which is commonly used for ordinal data. The CL model links the logit of cumulative probabilities to covariates, formulated as

$$\text{logit}(P(Y \leq k | \mathbf{X}^* = \mathbf{x})) = \log\left(\frac{P(Y \leq k | \mathbf{X}^* = \mathbf{x})}{P(Y > k | \mathbf{X}^* = \mathbf{x})}\right) = \alpha_k + \beta^T \mathbf{x} \quad (10)$$

where $k \in \{1, \dots, K-1\}$.

We use the *mboost* R package [17] to implement the boosting method for ordinal responses. The gradient boosting algorithm was applied [5] to the CL models using the default component-wise smooth P-spline base-learners [18]. The high flexibility allows us to fit more complex relationships between Y and $\mathbf{X}^*$ than that specified by a regular CL model. There is a unique base-learner for each variable in $\mathbf{X}^*$ since no interaction term is specified. The 5-fold cross-validated estimation of the empirical risk was used to select an optimal number of boosting iterations. The variable importance can be measured by the accumulated in-bag risk reductions per boosting step for each base-learner or variable, provided by the *varimp* function in the *mboost* package. We denote the measurements by Z_j^{mboost} for $j \in \{1, \dots, p^*\}$. The larger the risk reduction Z_j^{mboost} , the more important the variable X_j appears to be in predicting Y . The boosting measure has been shown to perform well within the knockoff framework when the outcome is continuous, binary or time to event [19].

Ordinal random forests

The random forest (RF) methodology [20] is another ensemble learning method for classification and regression. As an extension of bagging classification trees, it has the characteristics of random feature selection at each node, so that resulting trees are uncorrelated, which improves prediction accuracy. The RF method has been recently generalized to incorporate the ordering information for ordinal responses, using either conditional inference trees [21] or classical regression trees [22]. The latter method 'ordinal forests' [22] provides a variable importance measure for high-dimensional data, which can be incorporated into the frameworks above. Similar to the ordered probit regression model, the ordinal forest method assumes a latent continuous variable underlying the ordered response. The method [22] uses the out-of-bag misclassification error estimates to construct the forest and to compute the variable importance measurements, which can be obtained using the *ordfor* function in the *ordinalForest* R package [23]. We denote the measurements by $Z_j^{\text{ordinalForest}}$ for $j \in \{1, \dots, p^*\}$. An important variable tends to have a large positive value of $Z_j^{\text{ordinalForest}}$, while a null variable tends to have a measurement around zero. The method is not sparse by itself in the sense that most of the measurements are close to but not exactly zero.

Other measures

Elastic net regularization, which is a linear combination of the L_1 and L_2 penalties, overcomes some limitations of the LASSO, such as selecting only one variable from a group of high-correlated

variables [24]. The elastic-net penalty can be expressed below [25]

$$\lambda \left[\frac{1}{2} (1 - \alpha) \sum_{l=1}^p \beta_l^2 + \alpha \sum_{l=1}^p |\beta_l| \right], \quad (11)$$

where $0 \leq \alpha \leq 1$ is a mixing parameter, and $\alpha = 0$ corresponds to the ridge penalty and $\alpha = 1$ the LASSO penalty. We explored the performance of the RDVS with the value of α varying from 0 to 1. It turns out that the RDVS framework with the AIC- or BIC-based importance measure performs equally well for all different values of α except for $\alpha = 0$, i.e. the pure ridge case. The ridge penalty ($\alpha = 0$) leads to a higher FDR than other α values since it does not produce sparse solutions. We therefore maintain use of the L_1 -norm penalty ($\alpha = 1$) in this article for ease of interpretation.

We also applied a Bayesian procedure to fit CL models and assess variable importance by regression coefficients. The outcome Y is set to have a multinomial distribution, and the link function is given in Equation (10). For the purpose of variable selection, Laplace (or double-exponential) priors are used for regression parameters β_j , $j \in \{1, \dots, p^*\}$, as commonly seen in Bayesian LASSO models [4]. Specifically, the following priors are specified for model parameters: $\alpha_{0k} \sim N(0, \sigma^2)$, where $\alpha_{01} < \alpha_{02} < \dots < \alpha_{0,K-1}$, $\beta_j \sim \text{Double exponential}(0, \mu)$, $j \in \{1, \dots, p^*\}$, $\mu \sim \text{Gamma}(a, \text{rate} = b)$, where $(\alpha_{01}, \alpha_{02}, \dots, \alpha_{0,K-1})$ and $(\beta_1, \beta_2, \dots, \beta_{p^*})$ comprise the vectors α_k and β in Equation (10), respectively. The absolute values of posterior mean estimates for β_j can be used to measure variable importance, i.e.

$$Z_j^{\text{Bayesian}} = |\bar{\beta}_j| \quad (12)$$

where $\bar{\beta}_j$ is the posterior mean, for $j \in \{1, \dots, p^*\}$. Again, a larger Z_j^{Bayesian} means that X_j is more important. Due to the computational inefficiency and convergence issues of MCMC especially when p^* is large, the results for the Bayesian measure are not presented in this article. However, a small-scale experiment with $p = 500$ showed that the Bayesian measure together with knockoffs could achieve decent power with FDR properly controlled.

Marginal testing

Practitioners frequently resort to marginal screening when dealing with high-dimensional data. As a baseline for comparing the performance of multivariable methods and augmentation frameworks for variable selection, we applied several univariate testing approaches which produce marginal p -values for ordinal responses. In other words, the P variables are tested one-by-one for unconditional associations with Y . The resulting marginal P -values are then used to control the FDR using the BH procedure [7], where variables having smaller BH-adjusted P -values than the pre-specified target FDR level q are selected. Specifically, we examined three different marginal testing procedures. The Jonckheere–Terpstra test, which tests the equality of population medians among K groups against ordered alternatives [26], was applied using a permutation version for estimating P -values for each individual covariate using the *jonckheere.test* function in the *clinfun* R package [27]. Other marginal procedures include fitting univariate ordinal regression models including the CL model and the CR model. Both models were fit using the *vglm* function in the VGAM R package [28].

Simulation study

A simulation study was performed to compare different selection frameworks as well as different importance measures. The data-generating process is described below. The design matrix $\mathbf{X}$ contained n vectors of observed covariates, each of which had a p -dimensional Gaussian distribution $MVN(0, \Sigma)$ and Σ had an autoregressive structure with the $(i, j)^{\text{th}}$ element being $\rho^{|i-j|}$. We set the correlation parameter ρ to be either 0.2 or 0.5 to reflect small to moderate correlation among predictors as we might see in real genomic datasets. The sample size n was set to 300 and the number of predictors $p = 1000$ to simulate the situation where p exceeds n . Among the $p = 1000$ variables, $k = 50$ variables were randomly selected to be important and truly associated with the outcome Y . For these 50 variables, the corresponding coefficient β_j 's had the signal amplitude A taking on random signs, i.e. $\beta_j \in (\pm A)$. The amplitude represents the signal strength and was varied from 0.2 to 1. The other 950 variables were simply 'null' noise variables with the coefficients being zero. After obtaining the coefficient vector $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$, we sampled a latent continuous response U through the formula below

$$U = \mathbf{X}\beta + \epsilon \quad (13)$$

where ϵ followed a logistic (0,1) distribution, so that $U \mid \mathbf{X}$ followed a logistic $(\mathbf{X}\beta, 1)$ distribution. Then, the ordinal response Y with three ordinal levels was generated by categorizing U using its 33.3th and 66.7th percentiles. The simulation settings correspond to a CL model and the percentiles here correspond to the α_k 's in Equation (10). We set the target FDR q at 20%, i.e. we wanted to control the FDR below or at least around 20%. For the RDVS framework, we repeated the model fitting $M = 100$ times.

Our goal was to compare base learners combined with one of the two selection frameworks, model-X knockoffs and RDVS described in Section 2.3, to base learners without using any framework with respect to FDR and power. That is, the L_1 penalized CR model and boosting introduced in Section 2.4 can be used to perform variable selection without applying a framework. For example, non-zero coefficient estimates from the L_1 penalized CR model tuned by AIC (Z_j^{AIC}), BIC (Z_j^{BIC}) or CV (Z_j^{CV}), and risk reductions from boosting (Z_j^{mbboost}). These methods can be used to select variables since they are automatically sparse, i.e. only a small number of estimates are non-zero. We considered that the variables having non-zero importance measurements were 'selected' by these methods and other variables were not.

We then compared the FDR and power resulting from different frameworks and different importance measures. The FDR and power were plotted against the amplitude A to exhibit the growing trends over signal strength. For each fixed A , the values reported were averaged over 30 repeated runs. All simulations were performed on the Ohio Supercomputer Center cluster owens [29].

Real application

We applied the methods described to identify CpG sites whose methylation levels were associated with the progression from normal liver tissue to HCC. We used a subset of a dataset, available from Gene Expression Omnibus under accession number GSE18081, from a study of HCC where liver tissue samples were assayed using the Illumina GoldenGate Methylation BeadArray Cancer Panel I [3]. The dataset includes the methylation percentages of 1505 CpG sites from 56 liver tissues. Twenty liver

tissues were from hepatitis C virus (HCV) related HCC tumors and their adjacent non-tumorous HCV-cirrhotic tissues, 16 were independent HCV-cirrhotic tissues from patients without concomitant HCC, and the other 20 tissues were normal liver tissues. The outcome can be considered as an ordinal variable with three levels (normal < cirrhosis non-HCC < HCC). We would like to fit the ordinal outcome using the CpG site methylation percentages as predictors to identify a set of genes associated with the ordinal phenotype.

Prior to applying any methods, we removed the CpG sites with missing values or extreme values (mean percent methylation larger than 95% or smaller than 5%), leaving 1078 CpG sites for the analyses. Then, we applied different selection methods to the dataset, as we have done in the simulation study. To be consistent with the simulations, we set the target FDR q to be 20%. The models were fit $M = 100$ times when using the RDVS framework. Genes associated with CpG sites selected by different framework-measure combinations were recorded for comparison.

The original study [3] used marginal one-sided Jonckheere-Terpstra tests and the q -value method [30] to control the FDR at 10%, and identified 235 CpG sites that had a significant increasing trend and 266 sites with a significant decreasing trend in methylation proportion among tissues ranging from normal, to HCV-cirrhosis, to HCV-HCC. To assess the consistency with those previous findings for each method considered in this article, we calculated the proportion of times that the CpGs selected by each method also appeared in their results. The significant list given by Archer et al. [3] does not necessarily represent the ground truth, but if it did, the calculated proportion could be used to assess our methods in terms of the true positive proportion $(1 - \text{FDR})$.

Results

Simulation study

The trends of the FDR and power with varying signal amplitude (A) are presented for base learners (where neither the RDVS nor the knockoff framework applied), the BH procedure with marginal P -values, the RDVS framework and the knockoff framework, in Figures 1–4, respectively. In these plots, different lines correspond to different variable importance measures as defined in Section 2.4. The top panels present the case of slightly correlated covariates (autocorrelation $\rho = 0.2$) and the bottom present the moderately correlated case ($\rho = 0.5$). Each error bar indicates one standard deviation of the corresponding quantity over 30 replications. The x-axis—signal amplitude A —was slightly jittered to ensure the error bars for each method was distinguishable.

Before we applied any selection frameworks, the FDR was relatively high as shown in Figure 1, especially for the AIC- and CV-based selection methods. The corresponding power for these methods was high as well, which indicates that they are quite liberal. In contrast, the BIC-based method is fairly conservative with both FDR and power close to zero. The power increased as the signal strength increased since the selection methods can better distinguish between signals and noise variables with a larger signal strength. As the covariate correlation increased from $\rho = 0.2$ to $\rho = 0.5$, the power was slightly lower but the difference was inapparent.

For the marginal models adjusted by the BH procedure shown in Figure 2, the Jonckheere-Terpstra test appeared to be less powerful than the univariate CL and CR models. The FDR of the

Figure 1. FDR (left) and power (right) for L_1 penalized CR models tuned by AIC, BIC or CV, and for boosting with ordinal responses (base learners). The design matrix has an AR(1) covariance structure with autocorrelation 0.2 (top) and 0.5 (bottom). The error bars indicate $\pm$ one standard deviation over 30 replications. The signal amplitude was jittered to make error bars distinguishable.

univariate CL and CR models was well controlled at the target level of 20% for the slightly correlated datasets ($\rho = 0.2$) but not the moderately correlated cases ($\rho = 0.5$), which suggests that the FDR control of the BH procedure becomes unstable as the dependence gets stronger. The power still increased as the signal amplitude became larger, but only a small level (around 20%) was achieved in the end.

Compared to the base learners, the FDR for the AIC and CV measures was reduced after using the RDVS framework as shown in Figure 3, but only the CV measure in the slightly correlated scenario ($\rho = 0.2$) yielded strict FDR control at 20%. The AIC, path-based and ordinal forest measures produced relatively high FDR, though the FDR had a decreasing trend as the amplitude increased. The BIC and boosting measures that selected almost all variables, completely failed to control the FDR. Compared with the marginal methods in Figure 2, one reason for this failure could be that the empirical P -values produced by these measures were not uniformly distributed under the null; the distribution was concentrated near zero instead. Although the CV measure in the case of $\rho = 0.2$ controlled the FDR and obtained relatively high power achieving around 40%, the resulting FDPs were highly variable as suggested by the long error bars. In this case, the FDR control is not very helpful since the realized FDP has an unacceptably wide range.

When we applied the knockoff framework, we can see from Figure 4 that the average FDR for all six measures has been well-controlled at or below 20% in both correlation scenarios. The FDR showed an increasing trend as the amplitude grew rather than remaining stable at the nominal level as in the BH procedure, which makes knockoffs preferable in the regime where the signal strength is so low that we have little chance of identifying the true signals. The resulting error bars are shorter than those from the other methods, indicating less variable FDP and more reliable FDR control. The power increased as the amplitude increased, ultimately achieving around 20% for some of the measures. Some importance measures from the L_1 penalized CR model (AIC and CV) yielded greater power than the others.

The selection methods for base learners in Figure 1 were more computationally efficient compared to using either selection framework, since no augmentation of the design matrix or repeated model fitting was needed. Compared with the RDVS, the knockoff appeared to run faster for most of the importance measures as the models were only fit once within the knockoff framework. This may not be true if the importance measure does not scale well to high-dimensional data, because the knockoff uses $n \times 2p$ augmented matrices, while the RDVS uses $n \times (p + 1)$ matrices that can make a large difference if the original dimension p is high. Within a fixed framework, the measures

Figure 2. FDR (left) and power (right) for marginal tests with BH adjustment. The design matrix has an AR(1) covariance structure with autocorrelation 0.2 (top) and 0.5 (bottom). The error bars indicate $\pm$ one standard deviation over 30 replications. The signal amplitude was jittered to make error bars distinguishable.

from the L_1 penalized CR models (AIC, BIC, CV and path) were computationally faster than the ordinal forest measure, which was faster than the boosting measure. Regarding the marginal methods, univariate CL and CR models are slightly faster than the Jonckheere–Terpstra tests.

Real application

Figure 5 demonstrates the selection results of the real dataset by different methods, where a cross mark indicates that the gene (mapping to the selected CpG site) in the corresponding column was selected by the method in the row. The genes were sorted by the number of times they were selected. Some methods including the three marginal methods and some RDVS methods were far too liberal and selected many more genes than the other methods. Due to the space limitation, only genes selected by more than eight methods are presented. Genes that were reported in the original study as significant [3] appear in pink if their methylation proportion had an increasing trend for progression to HCC and in blue if their methylation proportion had a decreasing trend. The proportions of times that the genes selected by each method were also identified as significant in the original article [3] was calculated and listed to the right of Figure 5. The methods were then sorted by this proportion in descending order. The boosting-based method as well as the ‘knockoff + boosting’ combination turned out

to have the highest proportion of 100%, indicating all of the findings were identified in the previous study, followed by some RDVS methods and L_1 penalized methods. The results from the marginal Jonckheere–Terpstra tests differed from the previous findings because a more liberal FDR threshold of 20% (rather than 10%) and two-sided tests were applied in this current article. Overall, our results coincide well with the previous findings [3], and these discoveries may be reproducible in independent validation studies due to the error-controlling procedures used.

Other evidence has been found to support the roles of the selected genes in liver cancer development. The most frequently identified gene, endoplasmic reticulum to nucleus signaling 1 (ERN1), also known as IRE1, has been shown to play an important role in the ER stress-activated autophagy, which protects the HCC cells from death [31]. The myeloperoxidase (MPO) gene is known to be associated with myeloperoxidase deficiency and Alzheimer Disease, and a certain MPO genotype is found to increase the risk of HCC occurrence and liver-related death [32]. DNA damage inducible transcript 3 (DDIT3), encodes a transcription factor that regulates cell migration, proliferation, cell apoptosis and survival in many cancers [33–35]. Previous research has suggested that differentially methylated DDIT3 may lead to liver carcinogenesis [36]. There is weak evidence supporting that glycosylphosphatidylinositol anchored molecule like (GML) is differentially expressed comparing HCC with non-HCC

Figure 3. FDR (left) and power (right) using the RDVS framework with various importance measures. The design matrix has an AR(1) covariance structure with autocorrelation 0.2 (top) and 0.5 (bottom). The error bars indicate $\pm$ one standard deviation over 30 replications. The signal amplitude was jittered to make error bars distinguishable.

hepatitis/cirrhosis [37]. There is also evidence showing that some other commonly identified genes are relevant to HCC, including CDKN2B (or P15) [38], HLA-DPA1 and IL8 [39], IL16 [40], TJP2 [41], SOX17 [42], PADI4 [43] and DLC-1 [44]. The DLC-1 gene functions as a tumor suppressor gene in a number of common cancers including HCC.

Discussion and conclusion

In this article, we reviewed two variable selection frameworks, RDVS and knockoffs, together with several variable importance measures designed for high-dimensional ordinal outcome data. The two frameworks share a general idea of augmenting the data with artificial null variables to serve as benchmarks for the purpose of variable selection. Both frameworks are model-free and applicable to many well-designed importance measures for various outcome types including ordinal data. Knockoffs are more mathematically rigorous in terms of basing the definitions of the FDR and hypotheses on conditional independence, and thus guarantee exact finite-sample FDR control for the problem of interest, which was empirically verified by the present article and previous literature where continuous or binary outcomes were considered [9, 10, 19]. The RDVS framework, whose hypotheses are of a marginal flavor, resulted in some inflation in FDR, except when the CV-based measure was used in a slight correlated scenario. Among different importance measures, some

measures based on the penalized CR model (AIC, path and CV) that were relatively powerful and also computationally efficient are recommended for future research in ordinal response modeling. In the real application, some genes that mapped to selected CpG sites have been identified to be potentially associated with HCC, which have important implications for their roles in liver cancer and may help to discover novel biomarkers and mechanisms of HCC development.

We followed the second-order approximation approach [10] to construct knockoff variables in this article. Nevertheless, the approximation relies on a simplified assumption that the covariate vector (each row of X) can be well-described by a p -dimensional multivariate normal distribution, which may not hold in cases like methylation percentages for CpG sites. In addition, the approach is occasionally unstable for complex datasets and causes power loss for the subsequent selection procedure. Novel approaches using deep generative models have been developed to relax the distributional assumption and generate valid knockoffs in general settings [45–47]. The generative adversarial networks have been adapted for knockoff generation with four different and interacting neural networks involved [45]. Liu and Zheng [46] proposed an alternative solution with less computational burden based on the variational autoencoder [48]. In parallel, Romano et al. [47] introduced a machine that moves beyond the second-order approximation and matches higher-order moments of $(X, \tilde{X})$ through generative moment

Figure 4. FDR (left) and power (right) using the knockoff framework with various importance measures. The design matrix has an AR(1) covariance structure with autocorrelation 0.2 (top) and 0.5 (bottom). The error bars indicate $\pm$ one standard deviation over 30 replications. The signal amplitude was jittered to make error bars distinguishable.

Figure 5. Genes mapping to CpG sites identified to be important for HCC more than eight times using different methods, where a cross mark indicates that the gene in the corresponding column was selected by the method in the row. The knockoff and ordinal forest combination did not select any variable and was thus omitted. Pink columns indicate genes that were identified in the original study [3] as having increasing trends for progression to HCC, and the blue columns indicate genes having decreasing trends. The numbers to the right represent the proportion of times that the discoveries were identified in the original results [3].

matching networks. These novel construction approaches could be applied in future research to yield more accurate and powerful knockoffs for complex datasets such as high-throughput methylation data.

Other relevant variable selection techniques include permutation methods and the conditional randomization test [10]. A permutation method resamples the observed data points to construct a sampling distribution of test statistics under the null hypothesis. Given the empirical null distribution, FDR can be estimated or controlled mostly in the marginal modeling regime [49, 50]. The ‘fixed-X’ knockoff article [9] empirically compared the permutation method with the knockoff method and showed that the permutation method lost FDR control when covariates displayed non-vanishing correlations. The conditional randomization test [10] is somewhat related to the RDVS framework, but instead of obtaining the sampling distribution of statistics for a variable marginally independent of both X and Y , it calculates the sampling distribution for each variable under the null hypothesis of conditional independence. Assuming the covariate distribution F_X is known, the empirical null distribution can be obtained by independently sampling each variable conditional on all the other variables and recomputing the statistic with the sampled value. Then, conditional randomization p -values can be calculated from the empirical distribution. This method is superior to the RDVS framework in terms of allowing rigorous testing of conditional hypothesis, but it requires one to know the covariate distribution and to repetitively sample from the conditional distributions for all covariates which is computationally expensive.

One of the shortcomings of the article is that the power is greatly sacrificed for error control in the present simulation study. Though the controlled selection problem for high-dimensional, sparse, correlated ordinal response data is difficult by itself, designing powerful variable importance measures suitable for the problem would be helpful to boost the power. In addition, we did not apply one consistent ordinal regression model but instead used multiple models including the CR and CL models, making different measures less comparable. The simulation design, which is essentially based on the CL model, might favor the CL-based importance measures ($Z_j^{\text{uni_cumulative}}$, Z_j^{mboost} and Z_j^{Bayesian}) over the others. However, no obvious advantage has been observed for these measures according to the simulation results, which may ease the concern. Further efforts could be made to unify the choice of models, although the change in results, if any, is expected to be negligible.

The R programs for producing the results in this article are available at <https://github.com/hanfu-bios/ordinalVS>.

Key Points

- Compare the performance of two existing variable selection frameworks with traditional selection methods.
- Describe multiple variable importance measures for ordinal outcomes to make the frameworks applicable to ordinal response data.
- Identify features associated with the progression from normal liver tissue to hepatocellular carcinoma using high-throughput methylation data.
- The knockoff framework strictly controls FDR at a target value while the RDVS framework fails to do that in most cases.

Acknowledgements

We would like to thank Jared Huling for helpful discussions and useful comments about the methods considered in this article. We would also like to thank the reviewers for valuable feedback and suggestions regarding an earlier version of this article.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R03CA245771. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Forrest M, Andersen B. Ordinal scale and statistics in medical research. *Br Med J (Clin Res Ed)* 1986;**292**(6519):537–8.
2. Archer KJ, Williams AA. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat Med* 2012;**31**(14):1464–74.
3. Archer KJ, Mas VR, Maluf DG, et al. High-throughput assessment of CpG site methylation for distinguishing between HCV-cirrhosis and HCV-associated hepatocellular carcinoma. *Mol Genet Genomics* 2010;**283**(4):341–9.
4. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol* 1996;**58**(1):267–88.
5. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci* 2007;**22**(4):477–505.
6. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodology* 2008;**70**(5):849–911.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;**57**(1):289–300.
8. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001;**29**(4):1165–88.
9. Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 2015;**43**(5):2055–85.
10. Candès EJ, Fan Y, Janson L, et al. Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. *J R Stat Soc Series B Stat Methodology* 2018;**80**(3):551–77.
11. Linkletter C, Bingham D, Hengartner N, et al. Variable selection for Gaussian process models in computer experiments. *Dent Tech* 2006;**48**(4):478–90.
12. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, CA. 1988.
13. Edwards D. *Introduction to Graphical Modelling*. New York, NY: Springer, 2000.
14. Patterson E, Sesia M. *Knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.2. <https://CRAN.R-project.org/package=knockoff>, 2018.
15. Sesia M, Sabatti C, Candès EJ. Rejoinder: ‘gene hunting with hidden Markov model knockoffs’. *Biometrika* 2019;**106**(1):35–45.
16. Greenland S. Alternative models for ordinal logistic regression. *Stat Med* 1994;**13**(16):1665–77.

17. Hothorn T, Buehlmann P, Kneib T, et al. *mboost: Model-Based Boosting*, R Package Version 2.9-1, <https://CRAN.R-project.org/package=mboost>, 2018.
18. Schmid M, Hothorn T. Boosting additive models using component-wise P-splines. *Comput Stat Data Anal* 2008; **53**(2):298–311.
19. Shen A, Fu H, He K, et al. False discovery rate control in cancer biomarker selection using knockoffs. *Cancer* 2019;**11**(6): 744.
20. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.
21. Janitzka S, Tutz G, Boulesteix AL. Random forest for ordinal responses: prediction and variable selection. *Comput Stat Data Anal* 2016;**96**:57–73.
22. Hornung R. Ordinal forests. *J Classif* 2019;**1**–14. doi: <https://doi.org/10.1007/s00357-018-9302-x>
23. Hornung R. *ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables*, R Package Version 2.3-1, <https://CRAN.R-project.org/package=ordinalForest>, 2019.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology* 2005; **67**(2):301–20.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;**33**(1):1.
26. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. *Biometrika* 1954;**41**(1/2):133–45.
27. Seshan VE. *clinfun: Clinical Trial Design and Data Analysis Functions*, R Package Version 1.0.15, <https://CRAN.R-project.org/package=clinfun>, 2018.
28. Yee TW. *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, NY: Springer, 2015.
29. Ohio Supercomputer Center. Columbus, OH: 1987, Available from: <http://osc.edu/ark:/19495/f5s1ph73>.
30. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci* 2003;**100**(16):9440–5.
31. Hu F, Han J, Zhai B, et al. Blocking autophagy enhances the apoptosis effect of bufalin on human hepatocellular carcinoma cells through endoplasmic reticulum stress and JNK activation. *Apoptosis* 2014;**19**(1):210–23.
32. Nahon P, Sutton A, Rufat P, et al. Myeloperoxidase and superoxide dismutase 2 polymorphisms comodule the risk of hepatocellular carcinoma and death in alcoholic cirrhosis. *Hepatology* 2009;**50**(5):1484–93.
33. Jauhainen A, Thomsen C, Strömbom L, et al. Distinct cytoplasmic and nuclear functions of the stress induced protein DDIT3/CHOP/GADD153. *PLoS ONE* 2012;**7**(4):e33208.
34. Marciniak SJ, Yun CY, Oyadomari S, et al. CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. *Genes Dev* 2004;**18**(24): 3066–77.
35. He K, Zheng X, Li M, et al. mTOR inhibitors induce apoptosis in colon cancer cells via CHOP-dependent DR5 induction on 4E-BP1 dephosphorylation. *Oncogene* 2016;**35**(2):148–57.
36. Li CW, Chang PY, Chen BS. Investigating the mechanism of hepatocellular carcinoma progression by constructing genetic and epigenetic networks using NGS data identification and big database mining method. *Oncotarget* 2016;**7**(48):79453–73.
37. Yang B, Guo M, Herman JG, et al. Aberrant promoter methylation profiles of tumor suppressor genes in hepatocellular carcinoma. *Am J Pathol* 2003;**163**(3):1101–7.
38. Wang L, Sun L, Huang J, et al. Cyclin-dependent kinase inhibitor 3 (CDKN3) novel cell cycle computational network between human non-malignancy associated hepatitis/cirrhosis and hepatocellular carcinoma (HCC) transformation. *Cell Prolif* 2011;**44**(3):291–9.
39. Budhu A, Forgues M, Ye QH, et al. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell* 2006;**10**(2):99–111.
40. Li S, Deng Y, Chen ZP, et al. Genetic polymorphism of interleukin-16 influences susceptibility to HBV-related hepatocellular carcinoma in a Chinese population. *Infect Genet Evol* 2011;**11**(8):2083–8.
41. Laquaglia MJ, Grijalva JL, Mueller KA, et al. YAP subcellular localization and hippo pathway transcriptome analysis in pediatric hepatocellular carcinoma. *Sci Rep* 2016;**6**: 30238.
42. Jia Y, Yang Y, Liu S, et al. SOX17 antagonizes WNT/ β -catenin signaling pathway in hepatocellular carcinoma. *Epigenetics* 2010;**5**(8):743–9.
43. Chang X, Han J, Pang L, et al. Increased PADI4 expression in blood and tissues of patients with malignant tumors. *BMC Cancer* 2009;**9**:40.
44. Wong CM, Lee JM, Ching YP, et al. Genetic and epigenetic alterations of DLC-1 gene in hepatocellular carcinoma. *Cancer Res* 2003;**63**(22):7646–51.
45. Jordon J, Yoon J, Schaar M. KnockoffGAN: generating knockoffs for feature selection using generative adversarial networks. *International Conference on Learning Representations*, New Orleans, LA: OpenReview, 2019.
46. Liu Y, Zheng C. Auto-encoding knockoff generator for FDR controlled variable selection. *arXiv preprint arXiv*. 2018;**1809**:10765.
47. Romano Y, Sesia M, Candès EJ. Deep knockoffs. *arXiv preprint arXiv*. 2018;**1811**:06687. doi: [10.1080/01621459.2019.1660174](https://doi.org/10.1080/01621459.2019.1660174)
48. Kingma DP, Welling M. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014. 1312.6114
49. Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 2005;**21**(23):4280–8.
50. Yang YH, Lin WY, Lee WCA. Fuzzy permutation method for false discovery rate control. *Sci Rep* 2016;**6**:28507.