



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Variable selection methods with FDR control for class imbalanced data via data splitting

Hyunjin Lim

The Graduate School

Yonsei University

Department of Biostatistics and Computing

Variable selection methods with FDR control for class imbalanced data via data splitting

A Master's Thesis

Submitted to the Department of Biostatistics and Computing
and the Graduate School of Yonsei University

in partial fulfillment of the
requirements for the degree of
Master of Science


Hyunjin Lim

June 2024

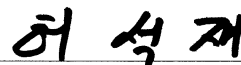
This certifies that the master's thesis of *Hyunjin Lim* is approved.



Inkyung Jung: Thesis Supervisor



ChungMo Nam: Thesis Committee Member #1



SeokJae Heo: Thesis Committee Member #2

The Graduate School

Yonsei University

June 2024

Contents

1. Introduction	1
2. Background	4
2.1 Multiple testing	4
2.2 Benjamini-Hochberg procedure	6
2.3 Knockoff framework	7
2.4 Data splitting	9
2.4.1 Single data splitting	9
2.4.2 Multiple data splitting	13
3. Proposed methods	15
3.1 Penalized logistic regression	15
3.1.1 Lasso	17
3.1.2 Elastic net	18
3.2 Adjustment of imbalance ratio	19
3.2.1 Downsampling	19
3.2.2 LHO-LOO	20

4. Simulation study	23
4.1 Simulation setting	25
4.2 Simulation results	27
5. Application	35
6. Conclusion and discussion	38
7. Supplementary	41
References	48
국문요약	51

List of Tables

Table 1. The result of multiple testing hypotheses	4
Table 2. Summary of standard and proposed data splitting methods	24
Table 3. Results of the average number of cases over imbalance strength.....	26
Table 4. Performance based on simulated data derived from the French pharmacovigilance database	37

List of Figures

Figure 1. DS procedure for FDR control	12
Figure 2. Adjustment methods for class imbalance ratio	21
Figure 3. Adjusted DS procedure for FDR control	22
Figure 4. Empirical FDRs and TPRs over correlation in balanced data	28
Figure 5. Empirical FDRs and TPRs over correlation in imbalanced data	28
Figure 6. Empirical FDRs and TPRs over the imbalance strength with independent assumption	29
Figure 7. Empirical FDRs and TPRs over the imbalance strength with correlation	29
Figure 8. Empirical FDRs and TPRs over correlation in imbalanced data for proposed data splitting methods	32
Figure 9. Empirical FDRs and TPRs over correlation in imbalanced data for proposed MDS and standard methods	32
Figure 10. Empirical FDRs and TPRs over imbalance strength with independent assumption for proposed data splitting methods	33
Figure 11. Empirical FDRs and TPRs over imbalance Strength with correlation for proposed data splitting methods	33
Figure 12. Empirical FDRs and TPRs over imbalance strength with correlation for proposed MDS and standard methods	34

Figure 13. Empirical FDRs and TPRs over the number of true variables in correlation and imbalanced data for proposed MDS and standard methods	34
Figure 14. Number of discovered drug signals using simulated data derived from French pharmacovigilance database	36
Figure 15. FDRs, TPRs, and number of selections for elastic net type data splitting methods under different α values on balanced data with correlation ($\rho = 0.5$)	43
Figure 16. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -3$) with correlation ($\rho = 0.5$)	43
Figure 17. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -3.5$) with correlation ($\rho = 0.5$)	44
Figure 18. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.5$)	44
Figure 19. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4.5$) with correlation ($\rho = 0.5$)	45
Figure 20. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5$) with correlation ($\rho = 0.5$)	45
Figure 21. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5.5, n = 2,000$) with correlation ($\rho = 0.5$)	46

Figure 22. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5.7, n = 2,000$) with correlation ($\rho = 0.5$)	46
Figure 23. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.3$).....	47
Figure 24. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.8$).....	47

Abstract

Variable selection methods with FDR control for class imbalanced data via data splitting

Identifying pertinent variables associated with the response variable in high-dimensional data is crucial across diverse domains. Nonetheless, many of the selected variables might lack actual association with the response variable. Particularly in severely class-imbalanced data, simple Lasso regression often leads to a significant increase in the false discovery rate (FDR). Even with methods implemented to control FDR, the true positive rate (TPR) can be very low. This study proposes two approaches aimed at enhancing TPR when selecting variables while controlling FDR for class-imbalanced data through data splitting strategies: 1) an extension of penalized regression, and 2) adjustment of class imbalance ratio. For comparison, the Benjamini-Hochberg procedure and the Knockoff framework were included. A simulation study showed imbalance ratio adjustment methods improved performance compared to conventional approaches.

Key words: False Discovery Rate (FDR); imbalance ratio; Penalized regression; Knockoff framework

1. Introduction

Identifying pertinent variables associated with the response variable in high-dimensional data is crucial across diverse domains. Methods for performing variable selection include stepwise regression (Efroymson, 1960), Lasso regression (Tibshirani, 1996), and bayesian variable selection methods (O'hara et al., 2009), among others. Additionally, recent research has introduced the application of stable variable selection based on Lasso in low-dimensional data for the purpose of detecting drug-adverse event signals in the field of pharmacovigilance (Ahmed et al., 2018).

However, a significant issue with variable selection is the potential inclusion of numerous variables that are not actually related to the response variable. For example, in the context of pharmacovigilance, clinical analysis is conducted on potential adverse-drug reactions (ADRs) identified through detection techniques such as variable selection (Ahmed et al., 2018). Misidentifying insignificant drugs as signals in the preliminary analysis can lead to subsequent confusion in the analysis process. Therefore, a crucial attribute expected from variable selection is minimizing the potential for false discovery rate (FDR) and controlling it. Efforts to quantify the uncertainty and quality of variable selection results under the specified error level have been ongoing (Dai et al., 2023). In regression-based models, methods for controlling the FDR include the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and Knockoff framework ideas (Candes et al., 2018). Recently, a FDR control method utilizing data splitting (Dai et al., 2023) has been introduced. This method comprises single data splitting (DS) and multiple

data splitting (MDS). DS evaluates variables by applying independent statistical methods to two partitioned datasets and obtaining mirror statistics. It offers an advantage over BHq or Knockoff framework as it does not require independent p-values or the joint distribution of variables. Additionally, the MDS method, which independently and iteratively integrates DS results, ensures the stability of variable selection (Dai et al., 2023).

Class imbalanced data refers to one class of the binary response variable having a much higher proportion than the other class. This characteristic is also observed in drug surveillance data. Instances of adverse events occurring are much rarer than non-occurrences, with the imbalance ratio typically being less than 1:1000. In cases of severe class imbalance, bias towards the majority class can render traditional variable selection methods ineffective (Kamalov et al., 2023). Also, since the minority class becomes the primary focus in most analyses, it is crucial to account for this characteristic.

This study proposes two approaches aimed at enhancing true positive rate (TPR) when selecting variables while controlling FDR for class imbalanced data through data splitting: 1) an extension of penalized regression, and 2) adjustment of class imbalance ratio. In terms of penalties, the Lasso penalty typically used for screening was replaced with the elastic net. While the Lasso selection assumes independence among variables, the elastic net effectively reflects correlations between variables by combining ridge and Lasso penalties. We adopted the stratified splitting method instead of conventional random splitting $\lfloor n/2 \rfloor$ to maintain the imbalance ratio of the raw data in the divided dataset. For adjusting the imbalance ratio, the 1:4 downsampling method, which reduces the sample size of the

majority class, was considered. The ratio was selected based on the claim that improvements in power were minimal for ratios greater than 1:4 in the case-control study (Ahmed et al., 2018). Another method for adjusting class imbalance ratio involves applying the LHO-LOO technique. This method randomly excludes one sample from the minority class and half of the samples from the majority class. This approach can offer broad variations in dataset while preserving the maximum number of samples in the minority class (Fu et al., 2017).

Our paper is structured as follows: Section 2.1 introduces the basic concept of multiple testing in terms of sparse regression. Section 2.2, 2.3, and 2.4 describe methods for controlling the FDR. These include BHq, the Knockoff framework, and data splitting, respectively. Section 3 discusses the concepts applied in the main methodology, which is data splitting. Details on the penalty for sparse selection are provided in section 3.1, and the adjustment methods for the imbalance ratio are placed in section 3.2. In section 4, we evaluate the proposed method's performance in various settings. The results are compared to the existing methods introduced in sections 2. The performance evaluation metrics used are FDR and TPR. In section 5, the proposed methods are applied to simulated data obtained from the French national pharmacovigilance database to detect potential signals. Conclusion and discussion are placed in section 6.

2. Background

2.1 Multiple testing

Let's consider m ($m > 1$) hypotheses are tested simultaneously. As m increases, the probability of making at least one incorrect decision, $1 - (1 - \alpha)^m$, grows remarkably under a significant level of α (Streiner and Norman, 2011). We define m_0 is the number of truly significant hypotheses, and R is the number of rejected hypotheses. Controlling the false discovery rate (FDR) allows for maintaining the overall error rate. Referring to Table 1, the FDR and True Positive Rate (TPR) are defined as follows:

$$FDR = E(FDP), \quad FDP = \frac{V}{R}$$

$$TPR = \frac{S}{m - m_0}$$

where FDP is the proportion of falsely rejected hypotheses among the rejected null hypotheses. Here, if $R = 0$, then $FDP = 0$ is defined.

Table 1. The result of multiple testing hypotheses.

	Accept Null hypothesis	Reject Null hypothesis	Total
Null hypothesis is true	U	V	m_0
Null hypothesis is false	T	S	$m - m_0$
Total	$m - R$	R	m

Translating this into the concept of variable selection. Let denote X be the $n \times p$ design matrix that are not determine as linear combinations of other variables. The response variable y be the $n \times 1$ vector. The logistic regression model is formulated as follows:

$$\log \frac{P(y = 1|X)}{P(y = 0|X)} = \beta_0 + X\boldsymbol{\beta}$$

where β_0 is an intercept term and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the $p \times 1$ vector of regression coefficients. Under y is represented by a GLM, it is known that if y is independent of X_j given the other variables $X_{-j} = \{X_1, X_2, \dots, X_p\} \setminus X_j$, then X_j is considered irrelevant variable (Candas et al., 2018).

$$y \perp X_j | X_{-j} \quad \text{if and only if} \quad \beta_j = 0.$$

Hence, the hypothesis in variable selection becomes equivalent to the multiple testing problem that follows:

$$H_{0j} : \beta_j = 0 \text{ vs } H_{1j} : \beta_j \neq 0.$$

The FDR can be shown as

$$FDR = E(FDP), \quad FDP = \frac{\#\{j : j \in S_0, j \in \hat{S}\}}{\#\{j \in \hat{S}\} \vee 1}$$

where S_0 represents the index set of null features (unrelated), and \hat{S} denotes the index set of variables selected through sparse variable methods.

2.2 Benjamini-Hochberg procedure

The Benjamini-Hochberg procedure (BHq) is the first introduced method for controlling FDR. Initially, it was limited by the assumption of independence among variables. Methods for adjusting significance probabilities in scenarios involving dependence were subsequently proposed by Benjamini and Yekutieli (2001). Let's suppose there are m hypotheses H_1, \dots, H_m to be tested, with p-values p_1, \dots, p_m . BHq utilizes the order statistics of p-value $p_{(i)}$, $i = 1, 2, \dots, m$. We can represent hypotheses corresponding to these order statistics as $H_{(i)}$. BHq's FDR control rule is as follows:

Process 1. Benjamini-Hochberg Procedure

- i. Order the p-value: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(i)} \leq \dots \leq p_{(m)}$.
 - ii. Given the control level $q \in (0,1)$, define $i_{max} = \max \left\{ i : p_{(i)} \leq \frac{i}{m} q \right\}$.
 - iii. Reject all $H_{(i)}$, $i = 1, 2, \dots, i_{max}$.
-

2.3 Knockoff framework

The aim of traditional statistical analysis for y and $X = (X_1, X_2, \dots, X_p)$ is to estimate the conditional distribution of $F_{y|X}$ (Liu and Rigollet, 2019). However, this requires various assumptions. Also, it may be challenging to apply in sparse data. As a solution, “model- X ” has been introduced to model the independent variable X under the knockoff framework. This method generates fake Knockoff variables, $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p)$, enabling variable selection. Random variable \tilde{X} is composed of the following properties (Candes et al., 2018).

- (1) Pairwise exchangeability: for any random subset $S \subset \{1, \dots, p\}$, the distribution of $(X, \tilde{X})_{\text{swap}(S)}$ and (X, \tilde{X}) are the same.
- (2) Negative control: if y is present, $\tilde{X} \perp y|X$. \tilde{X} is generated without considering response y .

Given that $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$, a joint distribution, which relies on the property (1), becomes

$$(X, \tilde{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \text{where } \mathbf{G} = \begin{pmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{pmatrix}.$$

Constructing Knockoff variables from observed X is represented as

$$\tilde{X}|X =^d \mathcal{N}(X - X\Sigma^{-1}\text{diag}(s), 2\text{diag}(s) - \text{diag}(s)\Sigma^{-1}\text{diag}(s)).$$

The index set of relevant variables and irrelevant variables are denoted by S_1 and S_0 , respectively. Each X_j is distinguished by W_j to indicate whether it is related to y . We can define $W_j = f_j(Z_j, \tilde{Z}_j)$, where Z_j and \tilde{Z}_j are model-dependent importance scores of X_i and \tilde{X}_i . f is an antisymmetric function. A large positive value of W_j provides the strong reason for the relevance to the y . While irrelevant variable's W_j are symmetric around 0. The selection criterion with the symmetric property is defined as

$$\#\{j: W_j \leq -t\} \geq \#\{j \in S_0 : W_j \leq -t\} =^d \#\{j \in S_0 : W_j \geq t\}, \quad \forall t > 0.$$

The FDP and its estimate $\widehat{FDP}(t)$ are represented by

$$FDP(t) = \frac{\#\{j \in S_0 : W_j \geq t\}}{\#\{j: W_j \geq t\}}, \quad \widehat{FDP}(t) = \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\}},$$

Here, an appropriate cutoff point τ_q to control the FDR at the desired level is defined as

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

where \hat{S} is the subset of selected features by Knockoff filter, which $\hat{S} = \{j: W_j \geq \tau_q\}$.

We can control $\mathbb{E} \left[\frac{|\{j \in \hat{S} \cap S_0\}|}{|\hat{S}| \vee 1} \right] \leq q$.

2.4 Data splitting

Data splitting method comprises single data splitting (DS) and multiple data splitting (MDS). It offers an advantage over BHq or Knockoff filter as it does not require independent p-values or the joint distribution of variables. Let denote X be the $n \times p$ design matrix, the response variable y be the $n \times 1$ vector, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the $p \times 1$ vector of regression coefficients.

2.4.1 Single data splitting

DS divides the data into two parts and applies potentially different statistical models to each. We use a sparse selection method like Lasso regression on the first split data $(y^{(1)}, X^{(1)})$ to proactively choose variables. Subsequently, only the selected variables are included in the second split data $(y^{(2)}, X^{(2)})$, and the basic regression method is performed. At each step, we can obtain independent measurements for a single variable X_j , denoted as $\beta_j^{(1)}$ and $\beta_j^{(2)}$, respectively. The regression coefficients $\beta_j^{(2)}$ for the variables that were not selected in the first step are zero. Ultimately, computing the “Mirror statistic” M_j by $\beta_j^{(1)}$ and $\beta_j^{(2)}$ constitutes the primary idea of data splitting (Dai et al., 2023).

$$M_j = \text{sign}(\hat{\beta}_j^{(1)} \hat{\beta}_j^{(2)}) f(|\hat{\beta}_j^{(1)}|, |\hat{\beta}_j^{(2)}|)$$

The key properties that M_j should have in the DS method are as follows:

- (1) A feature with a larger mirror statistic is more likely to be a relevant feature.
- (2) The sampling distribution of M_j of any null feature is symmetric about 0.

Three practical options of $f(u, v)$ are proposed as

$$f(u, v) = 2 \min(u, v), \quad f(u, v) = uv, \quad f(u, v) = u + v.$$

Property (1) comes from the characteristics of the function $f(u, v)$, which includes non-negativity, symmetry around u and v , and monotonic increase in both u and v . Hence, we can assess the relative importance of each variable using M_j , and variables exceeding the threshold are chosen. Moreover, based on the symmetry described in property (2), we have the capability to establish an upper limit for false positives.

$$\#\{j \in S_0 : M_j > t\} \approx \#\{j \in S_0 : M_j < -t\} \leq \#\{j : M_j < -t\}, \quad \forall t > 0.$$

We define \hat{S} as the set of selected variables, then $\hat{S}_t = \{j : M_j > t\}$. The FDP and its estimate $\widehat{FDP}(t)$ are represented as

$$FDP(t) = \frac{\#\{j : M_j > t, j \in S_0\}}{\#\{j : M_j > t\} \vee 1}, \quad \widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1}.$$

The cutoff τ_q for the control level $q \in (0, 1)$ can be driven as $\tau_q = \min\{t > 0 : \widehat{FDP}(t) \leq q\}$, and the final selection set under control of q changes as $\hat{S}_{\tau_q} = \{j: M_j > \tau_q\}$.

Process 2. FDR control through a single data split

- i. Divide the data into two halves $(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})$ and $(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})$.
- ii. Estimate the coefficients $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$ from each part of the data, applying the sparse variable selection method and basic regression, respectively.
- iii. Computing the M_j for a single variable X_j by

$$M_j = \text{sign}(\hat{\beta}_j^{(1)} \hat{\beta}_j^{(2)}) f(|\hat{\beta}_j^{(1)}|, |\hat{\beta}_j^{(2)}|).$$

- iv. Under the predefined FDR level q , determine the cutoff τ_q as

$$\tau_q = \min\{t > 0 : \widehat{FDP}(t) \leq q\}.$$

- v. Determine the final selected variable set by $\hat{S}_{\tau_q} = \{j : M_j > \tau_q\}$.
-

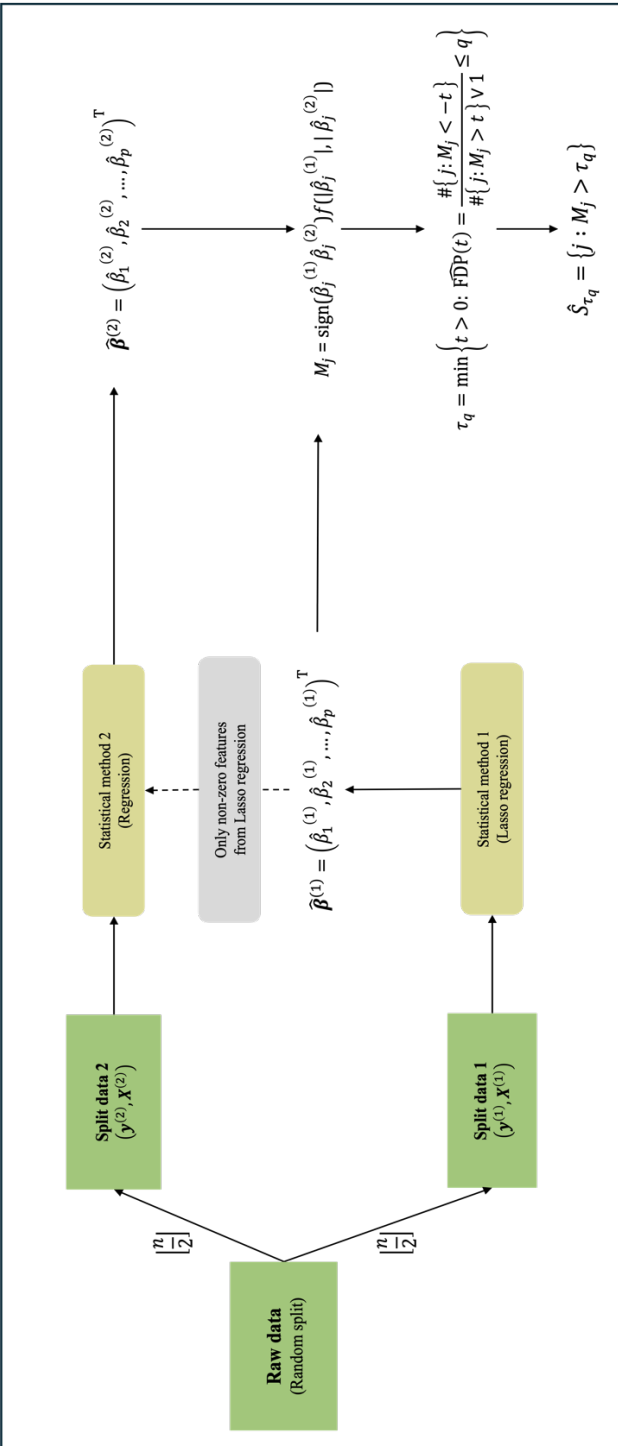


Figure 1. DS procedure for FDR control

2.4.2 Multiple data splitting

Estimating regression coefficient from divided data can lead to increased variance, potentially resulting in lower power compared to using the entire dataset. Additionally, the set of selected variables may be highly heterogeneous over split samples. To overcome these drawbacks, MDS aggregate the independent selection outcomes of DS. Through stability selection at least 50 times iteration, MDS demonstrates significantly enhanced TPR (Dai et al., 2023). MDS, like DS, does not rely on p-values or information about the joint distribution of features.

We maintain the notation of DS. It repeats the single data split method b times on (y, X) . For each trial, we can derive independent selection results $\hat{S}^{(\gamma)}$, $\gamma = 1, 2, \dots, b$. The inclusion rate I_j and its estimate \hat{I}_j associated with $\hat{S}^{(\gamma)}$, can be defined as follows

$$I_j = \mathbb{E} \left[\frac{\mathbb{I}(j \in \hat{S})}{|\hat{S}| \vee 1} \mid X, y \right], \quad \hat{I}_j = \frac{1}{b} \sum_{\gamma=1}^b \frac{\mathbb{I}(j \in \hat{S}^{(\gamma)})}{|\hat{S}^{(\gamma)}| \vee 1}$$

and the order statistics of \hat{I}_j show as $0 \leq \hat{I}_{(1)} \leq \hat{I}_{(2)} \leq \dots \leq \hat{I}_{(p)}$. When \hat{I}_j exceeds a certain cutoff, we define it as a relevant feature. The suitable cutoff for MDS is the average FDP does not exceed q . MDS achieves a lower FDR than the desired level q , while still maintaining strong performance.

Process 3. False discovery rate control for multiple data splits

- i. Order the estimated incorporation rates as $0 \leq \hat{I}_{(1)} \leq \hat{I}_{(2)} \leq \dots \leq \hat{I}_{(p)}$.
- ii. Determine the cutoff $l \in \{1, \dots, p\}$,

$$\hat{I}_{(1)} + \hat{I}_{(2)} + \dots + \hat{I}_{(l)} \leq q.$$

- iii. Define the relevant variable set $\hat{S} = \{j : \hat{I}_{(l)} < \hat{I}_j\}$
-

The concept of stabilizing selection results through repetition is already introduced by Meinshausen and Bühlmann (2010). However, while the stability selection method aims to optimize the regularization parameters in high-dimensional regression, MDS is used to compensate for power loss due to sample splitting. Stability method obtains selection sets using subsamples with replacement for different regularization parameters. Meanwhile, MDS replicates DS with independent and varied sample splits using the entire dataset.

3. Proposed methods

3.1 Penalized logistic regression

Logistic regression is primarily utilized for binary classification problems where the response variable takes values of 0 and 1. To select the optimal subset of explanatory variables associated with the response variable, a penalty term $P_\lambda(\boldsymbol{\beta})$ is incorporated into the log-likelihood function. In logistic regression modeling, various penalty methods have been explored. Shevade and Keerthi (2003) proposed sparse logistic regression using the Lasso penalty, Cawley and Talbot (2007) examined sparse logistic regression with a Bayesian penalty, and Liang et al. (2013) introduced the utilization of the $\ell_{\frac{1}{2}}$ penalty.

The logistic regression model and the probability with Where $\mathbf{x}_i = (x_{i1}, \dots, x_{i2})$ are formulated as follows:

$$\log \frac{P(y = 1|X)}{P(y = 0|X)} = \beta_0 + X\boldsymbol{\beta}, \quad \pi(\mathbf{x}_i) = P(y_i = 1|X = \mathbf{x}_i)$$

Then the log-likelihood and penalized logistic regression (PLR) are defined as

$$\ell(\beta_0, \boldsymbol{\beta}, y_i) = \sum_{i=1}^n \{y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log (1 - \pi(\mathbf{x}_i))\}$$

$$\text{PLR} = \sum_{i=1}^n \{y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log (1 - \pi(\mathbf{x}_i))\} + \lambda P(\boldsymbol{\beta})$$

where λ is a non-negative tuning parameter. It controls the strength of shrinkage in the explanatory variables. If the λ takes larger value, more weight will be given to the penalty term.

3.1.1 Lasso

Tibshirani (1996) suggested the Least Absolute Shrinkage and Selection Operator (Lasso) as a penalty for variable selection. Lasso utilizes the L_1 norm on the logistic regression coefficients. The form of the penalty term is as follows

$$\lambda P(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j|$$

The PLR with Lasso is

$$\text{PLR} = \sum_{i=1}^n \{y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log (1 - \pi(\mathbf{x}_i))\} + \lambda \sum_{j=1}^p |\beta_j|$$

So,

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg \min_{\boldsymbol{\beta}} \left\{ \ell(\beta_0, \boldsymbol{\beta}, y_i) + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

Through the Lasso penalty, some regression coefficients can be precisely set to 0, enabling variable selection. However, this method has drawbacks when there is high correlation among variables (Algamal and Lee, 2015). Additionally, it selects a maximum of n ($n < p$) variables, but in reality, there could be more than n variables with non-zero regression coefficients in the final model (Hou et al., 2023).

3.1.2 Elastic net

Elastic net method is a sparse variable selection method that shrink regression coefficients to zero while considering correlations (Hou et al., 2023). The elastic net penalty defines as

$$\lambda P(\boldsymbol{\beta}) = \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p \beta_j^2 \right)$$

where α ($0 \leq \alpha \leq 1$) is a hyperparameter that determines the balance between Lasso and ridge. Increasing the α value places emphasis on the Lasso penalty. When $\alpha = 0$, elastic net is entirely ridge, and when $\alpha = 1$, it becomes Lasso penalty. Also, we can represent this penalty and PLR by

$$\begin{aligned} \lambda P(\boldsymbol{\beta}) &= \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2, \\ \text{PLR} &= \sum_{i=1}^n \{y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i))\} + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \end{aligned}$$

where λ_1 and λ_2 are punishment parameters of Lasso and ridge method, respectively.

The PLR solution for elastic net can be shown as

$$\hat{\boldsymbol{\beta}}_{\text{Elastic}} = \arg \min_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}, \mathbf{y}_i) + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}.$$

3.2 Adjustment of imbalance ratio

Let U and n denote the entire dataset and the number of observations, respectively. The dataset divided into two halves of equal size using stratified splitting is referred to as U_1 and U_2 . Each set maintains the original class imbalance ratio. The imbalance ratio adjustment method is applied only to the dataset U_1 , where sparse variable selection is employed.

3.2.1 Downsampling

Downsampling is a method of randomly sub-sampling observations from the majority class to reduce its size. This allows adjusting the imbalance ratio to the desired level. After applying the downsampling method, the updated set U_1^* is defined as follows:

$$U_1^* = \{ C_{01}^* \cup C_{11} \}$$

where the majority class set in U_1 as C_{01} with n_{01} observations, and the minority class set as C_{11} with n_{11} observations. And samples were randomly selected from C_{01} , with size $n_{01}^* = k \cdot n_{11}$, which is the set C_{01}^* . In this paper, the value of k is set to 4 that align the ratio between minority and majority classes to 1:4. This decision was made in the argument that there is minimal improvement in power for ratios greater than 1:4 in an epidemiological case-control study.

3.2.2 LHO-LOO

The regularization parameter values of L_1 or L_2 norm penalties directly influence the outcomes of sparse variable selection methods. In other words, the selected variables were heavily influenced by individual observations. However, using inclusion frequency based on sub-sampling can mitigate the importance of regularization parameters (Fu et al., 2017). The use of conventional sub-sampling in imbalanced binary data can exacerbate the imbalance between two classes.

The LHO-LOO (Leaving Half of majority observations Out and Leaving One minority observation Out) strategy involves randomly excluding one observation from the minority class and half of the observations from the majority class, and then combining the remaining samples (Fu et al., 2017). It also applied only to the dataset U_1 , where sparse variable selection is conducted. Thus, the updated U_1^* is defined as:

$$U_1^* = \{C_{01}^* \cup C_{11}^*\}$$

where C_{01}^* is the set of $n_{01}^* = \left\lfloor \frac{n_{01}}{2} \right\rfloor$ samples which were randomly selected from C_{01} and C_{11}^* is the set of $n_{11}^* = n_{11} - 1$ samples from the minority class C_{11} . This approach can introduce broad variations in dataset splitting while preserving the maximum number of samples in the minority class.

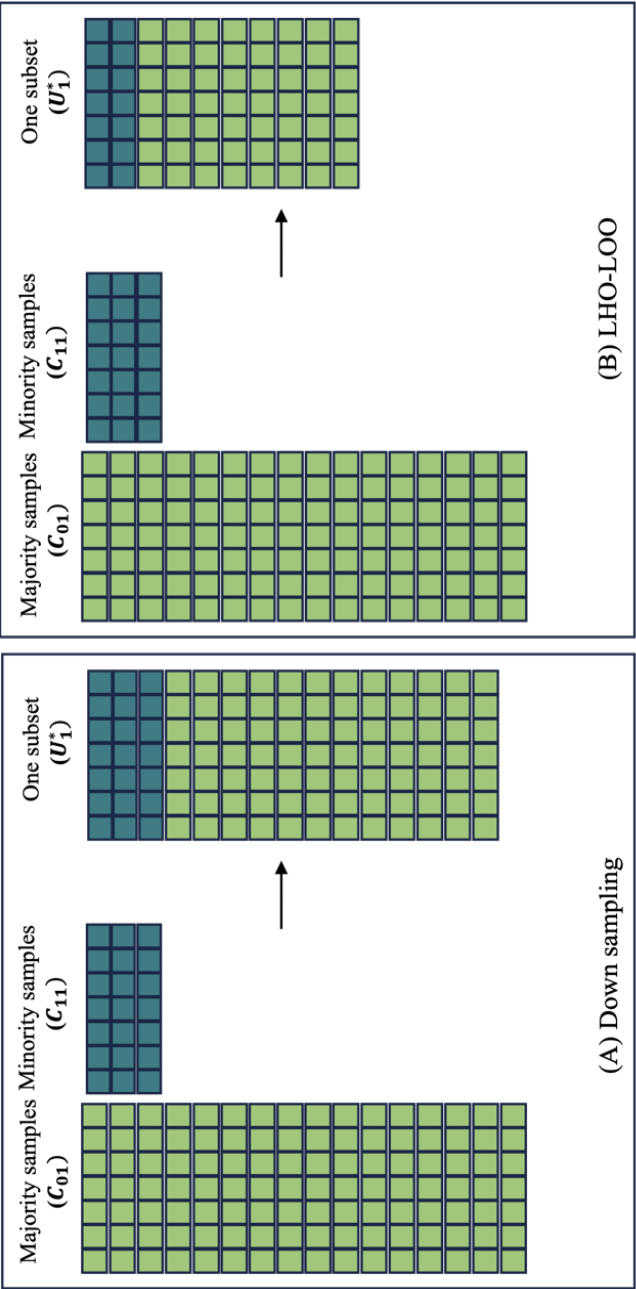


Figure 2. Adjustment methods for class imbalance ratio

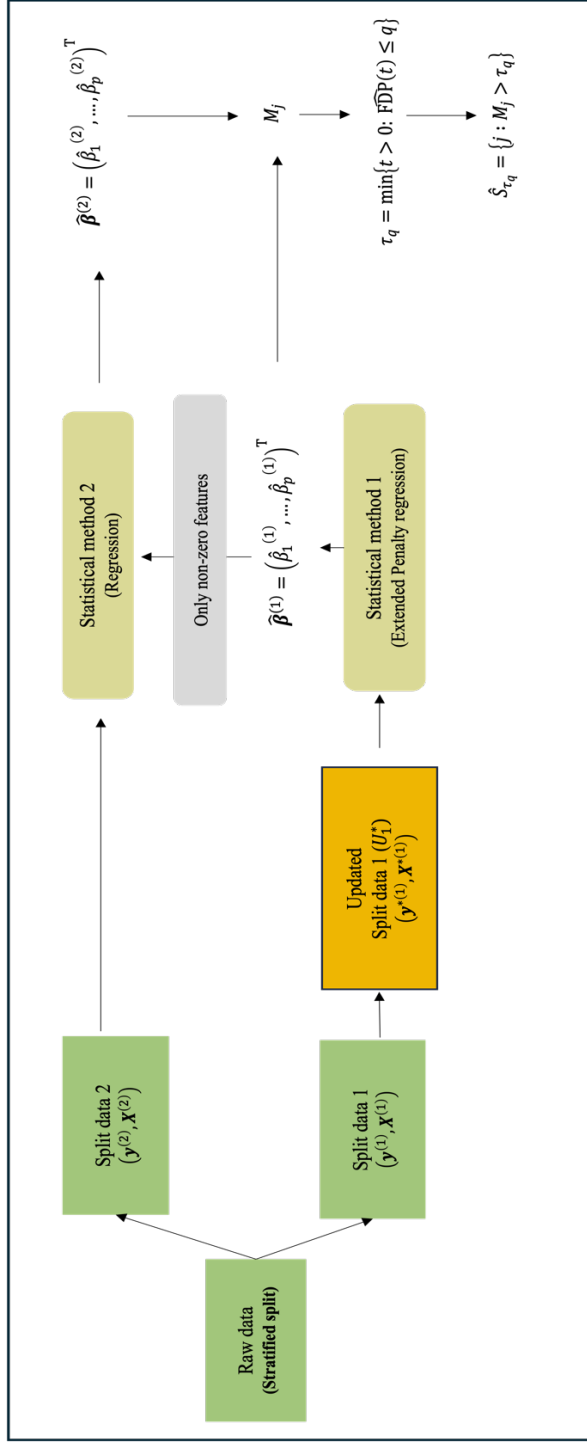


Figure 3. Adjusted DS procedure for FDR control

4. Simulation study

In this section, we explored the extension from Lasso to elastic net penalty mentioned in section 3.1.1 and 3.1.2. Additionally, we discussed the application of methods for adjusting class imbalance ratio mentioned in 3.2.1 and 3.2.2. The information of methods is summarized in the Table 2.

The regularization parameter λ for penalized regression was estimated through cross-validation. The α values, which is the weight provided for ridge and Lasso, are fixed at 1 and 0.5 for Lasso and elastic net, respectively. In elastic net α was determined based on the observation that TPR mostly reached its maximum at $\alpha = 0.5$ under various imbalance ratios. Refer to the supplementary materials for detailed information.

The proposed methods were compared with existing methods such as BHq, Knockoff filter, the original DS, and MDS. Performance measurements were based on FDR and TPR.

Table 2. Summary of standard and proposed data splitting methods.

Single splitting refers to the DS method, and multiple splitting means the MDS method.

		Splitting repetition	Splitting method	penalty	Adjustment method
Proposed data splitting		Single	Random	Elastic net	-
		Multiple			
		Single	Stratified	Lasso	Downsampling
					LHO-LOO
				Elastic net	Downsampling
					LHO-LOO
		Multiple		Lasso	Down sampling
					LHO-LOO
				Elastic net	Downsampling
					LHO-LOO
Standard method	Original data splitting	Single	Random	Lasso	-
		Multiple			
	other	Knockoff framework			
		BHq			

4.1 Simulation setting

Suppose the response variable y be the $n \times 1$ vector and X be the $n \times p$ matrix which follows a multivariate normal distribution $N(\mathbf{0}, \Sigma)$. The covariance matrix is $\Sigma = \rho^{|h-g|}$, $h, g = 1, 2, \dots, p$. We consider the number of sample size $n = 1,000$, and the number of features $p = 500$. Response variable y follows a binomial distribution.

$$y|X \sim \text{Binom}\left(\frac{\exp(X\boldsymbol{\beta} + \eta)}{1 + \exp(X\boldsymbol{\beta} + \eta)}\right)$$

Here, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, each β_j randomly has 0 or ω ($\omega > 0$) that represent the amplitude of signal. η is a parameter of imbalance strength.

The information about the settings used in the simulation follow:

- The correlation coefficient $\rho \in \{0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8\}$.
- The number of true relevant features $z \in \{10, 20, 30, 40, 50\}$.
- The parameter of imbalance strength $\eta \in \{0, -3, -3.5, -4, -4.5, -5\}$.
- The parameter of amplitude of true relevant features $\omega = 0.5$.
- Total sample size $n = 1,000$; The number of variables $p = 500$.
- Control level $q = 0.2$; 500 iterations for simulation.
- 50 repetitions for MDS.
- Hyperparameters for Lasso and elastic net $\alpha = 1$; $\alpha = 0.5$, respectively.
- Function f for calculating M_j in data splitting $f(u, v) = u + v$.

Table 3. Results of the average number of cases over imbalance strength.

The number of true features $z \in \{10, 20, 30, 40, 50\}$; Imbalance parameter $\eta \in \{-5, -4.5, -4, -3.5, -3, 0\}$; Correlation coefficient $\rho = 0.5$; $n = 1,000$; $p = 500$; $\omega = 0.5$; The number of iterations = 500.

z	$\eta = -5$	$\eta = -4.5$	$\eta = -4$	$\eta = -3.5$	$\eta = -3$	$\eta = 0$
10	20.78	32.39	48.74	72.29	103.91	500.52
20	45.27	63.02	86.48	115.96	151.95	499.9
30	71.39	93.34	119.81	151.88	188.24	500.64
40	96.62	120.27	149.1	181.53	218.08	499.45
50	121.05	145.88	174.37	205.87	240.67	500.37

4.2 Simulation results

Figures 4 and 5 show the results for class imbalance ratios of 1:1 and 1:11, respectively. As correlation increases, the TPR typically decreases, with performance degradation exacerbated under imbalanced data. Excluding cases with strong correlation, (MDS, Elastic net) method exhibit high TPR under imbalance. At $\rho = 0.8$, (MDS, Lasso) demonstrate a higher TPR than (MDS, Elastic net), but it suffers from inflated false discoveries, leading to a failure in FDR control.

Figures 6 and 7 demonstrate performance changes according to imbalance situation with $\rho = 0$ and $\rho = 0.5$, respectively. We consider $\eta \in \{-3, -3.5, -4, -4.5, -5\}$, each representing imbalance ratio of $\{1:6, 1:8, 1:11, 1:15, 1:22\}$. Even as the imbalance ratio increases, elastic net methods consistently maintain stable FDR control and exhibit a more gradual decrease in TPR compared to Lasso methods. Thus, it can be concluded that the elastic net type methods consider correlation and have better performance in imbalanced setting compared to Lasso based methods.

Due to the violation of the independence assumption of p-values, BHq has low TPR and unstable FDR as correlation increases. This phenomenon occurs regardless of the presence of imbalance. Also, as known in the knockoff framework, performance degradation occurs because of the reconstruction issues when strong linear dependencies among variables exists (Liu and Rigollet, 2019; Gimenez and Zou, 2019).

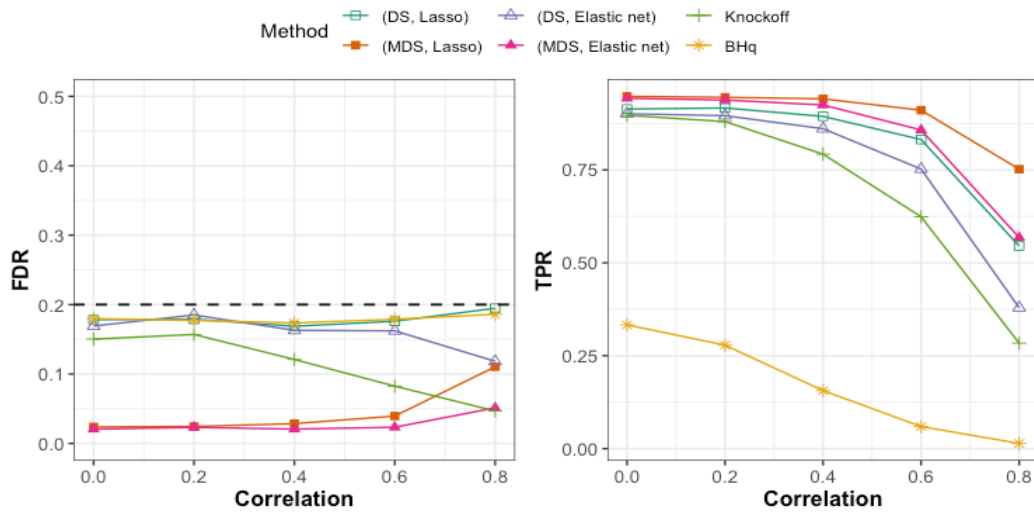


Figure 4. Empirical FDRs and TPRs over correlation in balanced data ($\eta = 0$). The considered correlation ρ are $\{0, 0.2, 0.4, 0.6, 0.8\}$ and designated FDR control level is $q = 0.2$.

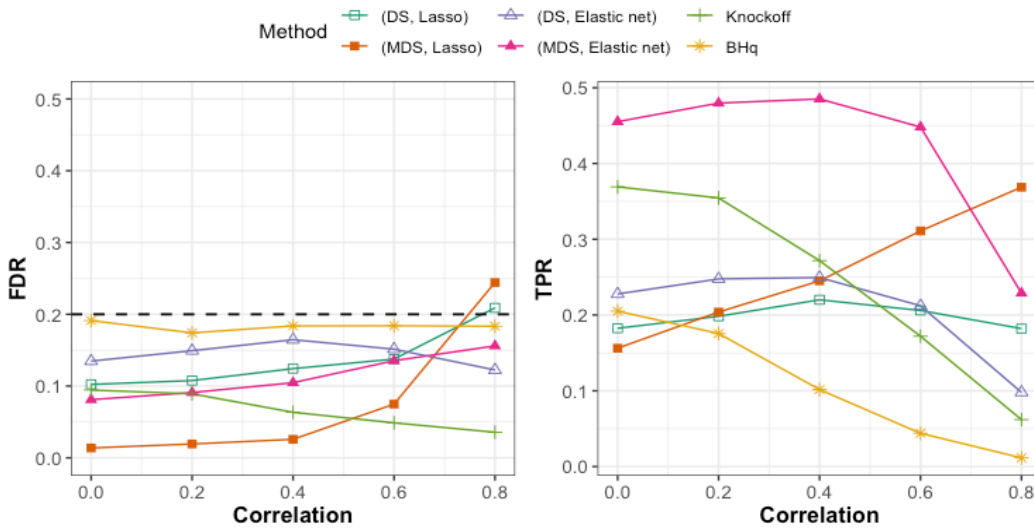


Figure 5. Empirical FDRs and TPRs over correlation in imbalanced data ($\eta = -4$). The considered correlation ρ are $\{0, 0.2, 0.4, 0.6, 0.8\}$ and designated FDR control level is $q = 0.2$.

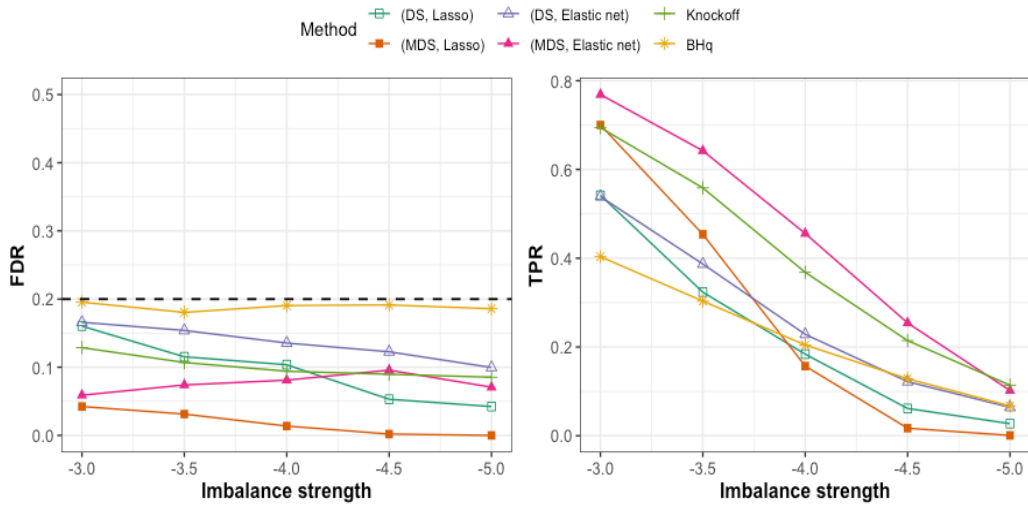


Figure 6. Empirical FDRs and TPRs over the imbalance strength with independent assumption. The considered η are $\{-3, -3.5, -4, -4.5, -5\}$, indicating ratio of $\{1: 6, 1: 8, 1: 11, 1: 15, 1: 22\}$. Designated FDR control level is $q = 0.2$.

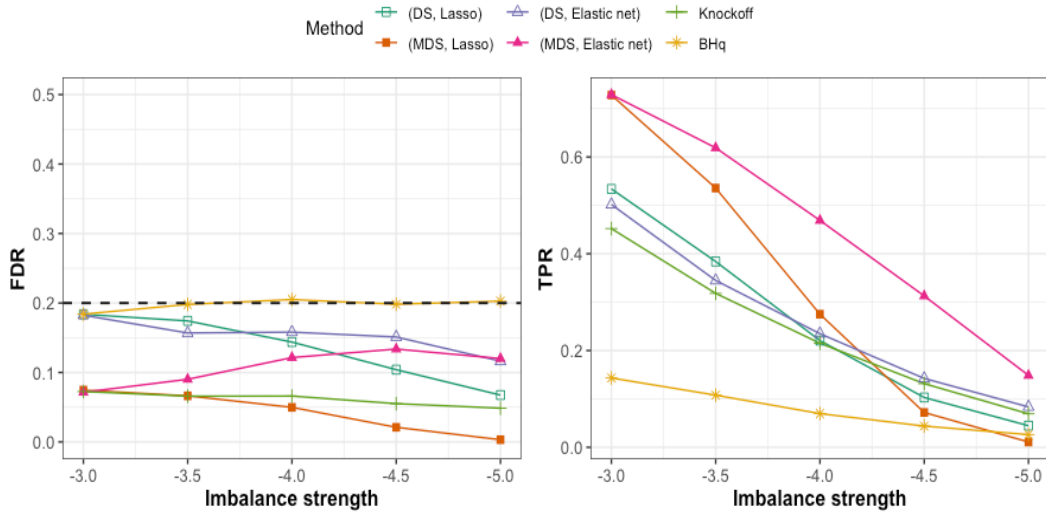


Figure 7. Empirical FDRs and TPRs over the imbalance strength with correlation ($\rho = 0.5$). The considered η are $\{-3, -3.5, -4, -4.5, -5\}$, indicating ratio of $\{1: 6, 1: 8, 1: 11, 1: 15, 1: 22\}$. Designated FDR control level is $q = 0.2$.

Combinations of Lasso and elastic net type data splitting applied with downsampling and LHO-LOO are shown in Figure 8. The imbalance ratio is assumed to be approximately 1:11 ($\eta = -4$). The MDS approaches demonstrate superior performance compared to DS. Methods, which are (DS, Lasso, Downsampling), (MDS, Lasso, Downsampling), (DS, Lasso, LHO-LOO), and (MDS, Lasso, LHO-LOO), exhibit a slightly higher TPR compared to the unadjusted methods (DS, Lasso) and (MDS, Lasso). These methods show a higher TPR at $\rho = 0.8$, but they do not completely control the FDR. However, the elastic net type adjusted methods conduct detection while controlling FDR all the correlation settings.

When the ratio exceeds 1:10, the TPR of Lasso type methods sharply decline. In all imbalance ratio scenarios $\{1:6, 1:8, 1:11, 1:15, 1:22\}$, the TPRs of the imbalance adjusted elastic net type are higher than that of Lasso type in figure 10 and 11. Especially in the 1:11 scenario, TPRs of (MDS, Elastic net, Downsampling) and (MDS, Elastic net, LHO-LOO) are approximately 0.2 greater than (MDS, Lasso, Downsampling) and (MDS, Lasso, LHO-LOO) in figure 10. Between (MDS, Elastic net, Downsampling) and (MDS, Elastic net, LHO-LOO), the method employing the LHO-LOO technique represents slightly higher TPR.

Under the assumption of variable independence in class imbalanced data, the TPR of the Knockoff framework is similar with (MDS, Elastic net, LHO-LOO). However, at a correlation of 0.5, the TPR of the Knockoff framework is lower than that of (MDS, Elastic net, LHO-LOO) by about 0.2. In the case of BHq, this difference increases to over 0.4, while FDR control very unstable. We can see the TPR of basic methods placed ranging

from 0.1 to 0.3 even at a low imbalance ratio with $\rho = 0$. However, the adjusted methods of the elastic net type maintain the TPR above 0.5 in same situation. Also, as the number of true variables raises, the data splitting based methods consistently exhibit enhanced TPR and stable FDR in figure 13.

The conclusions from simulation study can be summarized as follows:

In class imbalanced data environments,

- 1) about the data splitting, MDS exhibited higher TPR and lower FDR than DS.
- 2) add the correlation settings $\rho \leq 0.6$, (MDS, Elastic net) represents superior TPR than traditional selection methods such as (DS, Lasso), (MDS, Lasso), Knockoff framework, and BHq.
- 3) adjusting for class imbalance yields better TPR under FDR control, particularly in combination with elastic net based data splitting than Lasso.
- 4) among the considered class imbalance ratios, the LHO-LOO adjustment slightly outperforms both (MDS, Elastic net, Downsampling) and (MDS, Elastic net, LHO-LOO) at $\rho = 0$ and $\rho = 0.5$.

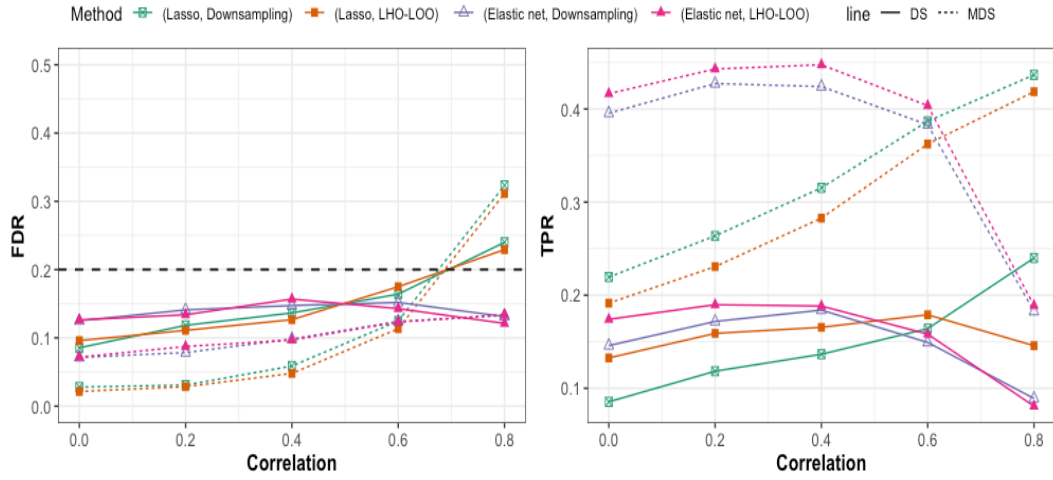


Figure 8. Empirical FDRs and TPRs over correlation in imbalanced data ($\eta = -4$) for proposed data splitting methods. The considered correlation ρ are $\{0, 0.2, 0.4, 0.6, 0.8\}$ and designated FDR control level is $q = 0.2$.

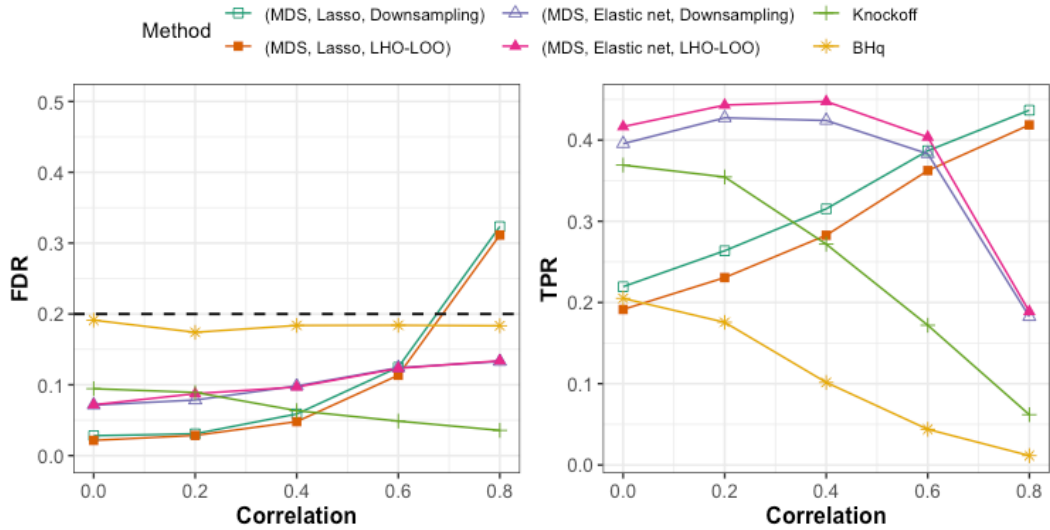


Figure 9. Empirical FDRs and TPRs over correlation in imbalanced data ($\eta = -4$) for proposed MDS and standard methods. The considered correlation ρ are $\{0, 0.2, 0.4, 0.6, 0.8\}$ and designated FDR control level is $q = 0.2$.

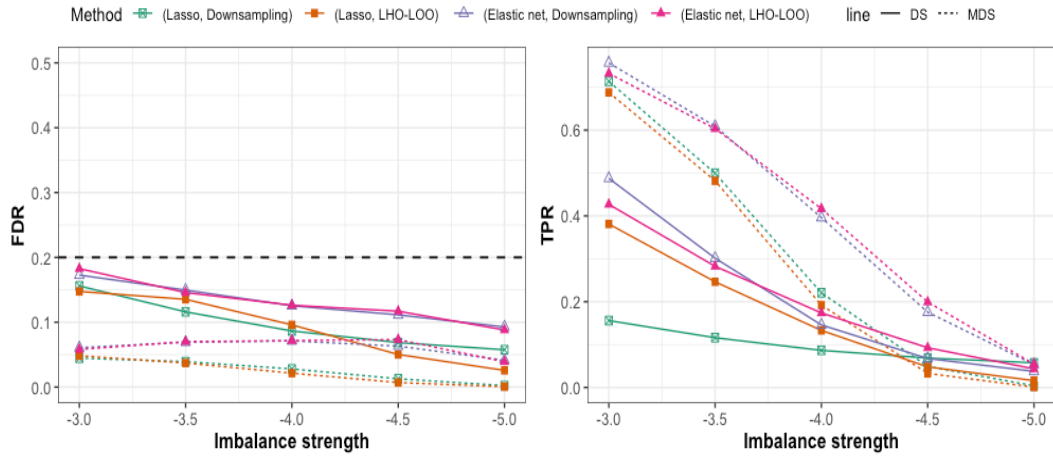


Figure 10. Empirical FDRs and TPRs over imbalance strength with independent assumption for proposed data splitting methods. The considered η are $\{-3, -3.5, -4, -4.5, -5\}$ and designated FDR control level is $q = 0.2$.

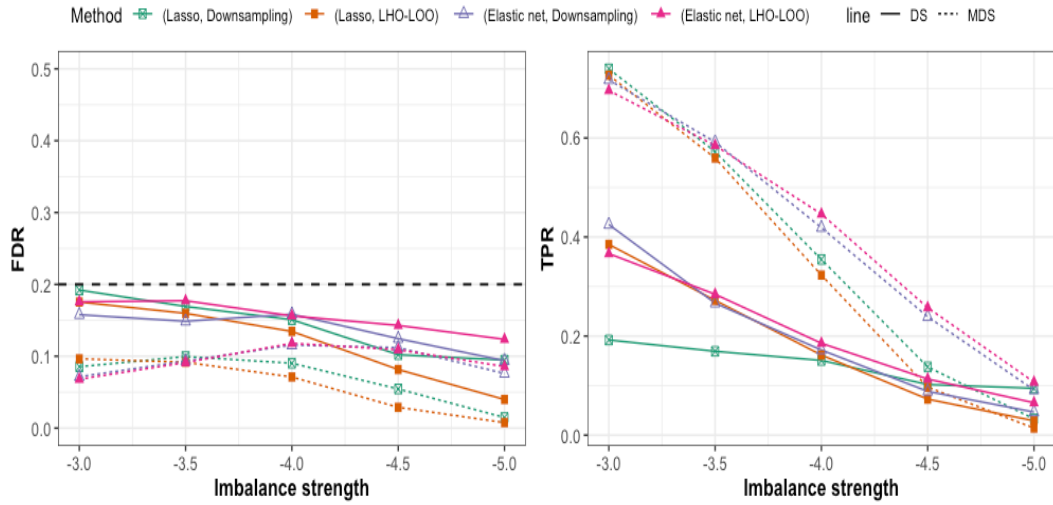


Figure 11. Empirical FDRs and TPRs over the imbalance strength with correlation ($\rho = 0.5$) for proposed data splitting methods. The considered η are $\{-3, -3.5, -4, -4.5, -5\}$, indicating ratio of $\{1: 6, 1: 8, 1: 11, 1: 15, 1: 22\}$. Designated FDR control level is $q = 0.2$.

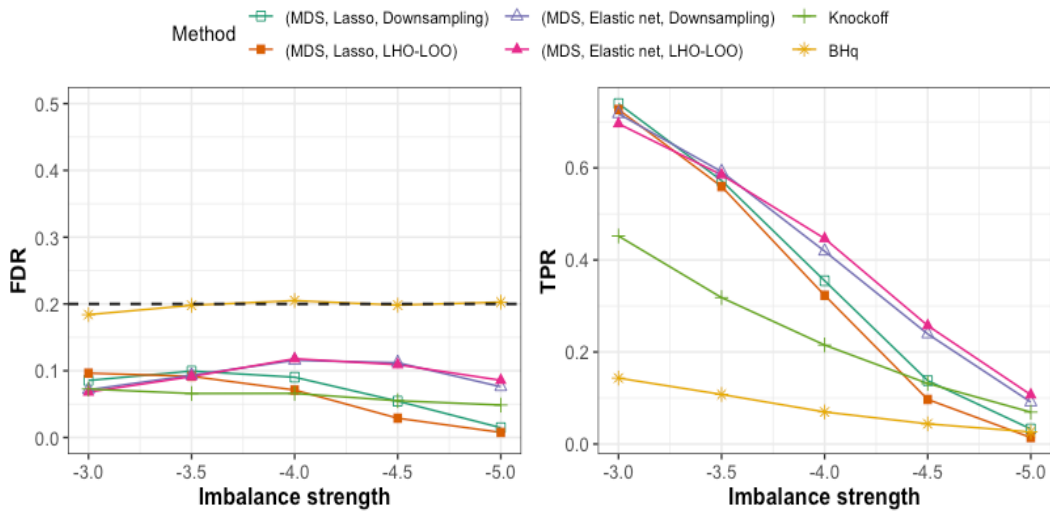


Figure 12. Empirical FDRs and TPRs over imbalance strength with correlation ($\rho = 0.5$) for proposed MDS and standard methods. The considered η are $\{-3, -3.5, -4, -4.5, -5\}$.

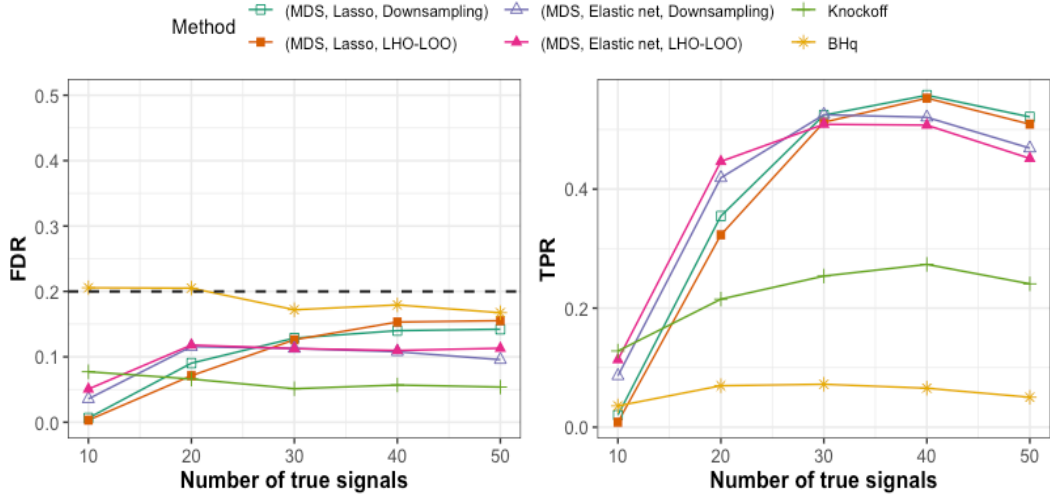


Figure 13. Empirical FDRs and TPRs over number of true variables in correlation and imbalanced data ($\rho = 0$, $\eta = -4$) for proposed MDS and standard methods. The considered number of true variables z are $\{10, 20, 30, 40, 50\}$.

5. Application

The performance of the proposed methods was validated through empirical analysis using simulated data derived from the French pharmacovigilance database. In France, healthcare professionals are required to spontaneously report adverse drug reactions (ADRs) (Thiessard et al., 2005).

The data we used is included in the “adapt4pv” package in R. The X matrix is a large sparse and binary matrix consisting of 117,160 rows and 300 columns. The rows and columns represent individual reports and drugs, respectively. y is a vector of length 117,160. It composes of binary values, where $y_i = 1$ if an adverse event occurred in the i th report, and $y_i = 0$ otherwise. The dataset includes only 300 drugs out of the larger number of drugs in the actual database. It contains 30 true signals based on the positive control group identified by the Comité Technique de Pharmacovigilance. Primary event is an adverse event occurrence. Approximately 3% of the total reports (3,557 out of 117,160) represents an adverse event. This indicates that the imbalance ratio in the simulated data is highly skewed at approximately 1:32.

The FDR control level is set at 0.2. In the table 4, FD and TD represent the number of false discoveries and true discoveries, respectively. The BHq successfully detected all 30 true signals; however, the FDR was notably high at approximately 0.5. Among the proposed DS methods, the elastic net based LHO-LOO demonstrated the best performance, effectively controlling the FDR below 0.2. This is consistent with the simulation results, where the proposed (MDS, Elastic net, LHO-LOO) demonstrated the best performance.

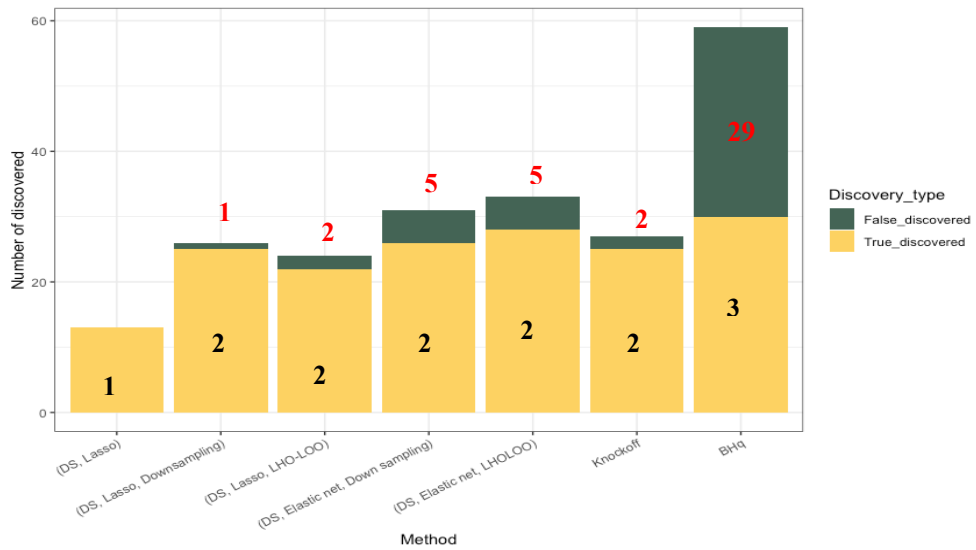


Figure 14. Number of discovered drug signals using simulated data derived from the French pharmacovigilance database. Methods are Lasso type and proposed elastic net type DS and the standard methods as knockoff and BHq. The designated FDR control level is $q = 0.2$.

Table 4. Performance based on simulated data derived from the French pharmacovigilance database. FD and TD represent the number of false discoveries and true discoveries, respectively.

	Method	FDR	TPR	# Selection	FD	TD
Proposed	(DS, Lasso, Downsampling)	0.038	0.833	26	1	25
	(DS, Lasso, LHO-LOO)	0.083	0.733	24	2	22
	(DS, Elastic net, Downsampling)	0.161	0.867	31	5	26
	(DS, Elastic net, LHO-LOO)	0.152	0.933	33	5	28
Standard	Knockoff	0.074	0.833	27	2	25
	BHq	0.492	1	59	29	30

6. Conclusion and discussion

In imbalanced data, variable selection method such as Lasso logistic regression tends to inflate the FDR. Even with the implementation of methods such as the BHq, Knockoff framework, or Lasso type data splitting to control the FDR, the TPR remains notably low. So, we proposed methods to increase the TPR in variable selection under FDR control via data splitting in class imbalanced data and conducted simulation study. Our approach includes: 1) extending the penalty of penalized regression from Lasso to elastic net, and 2) applying imbalance adjustment techniques such as 1:4 downsampling and LHO-LOO. Based on the results of simulation study, we conclude the following:

Regardless of the degree of data class imbalance, MDS consistently outperforms DS significantly. In case where correlation exists within imbalanced data, the performance degradation is more pronounced. We assume imbalance ratio of 1:11. With increasing correlation, existing methods such as BHq and Knockoff framework control the FDR but exhibit a significant decrease in TPR. Lasso type original data splitting show an increase in TPR but also surpass the controlled level of FDR. This holds true with imbalance adjusted lasso type methods. This suggests a failure to consider correlation after addressing the imbalance ratio. However, in the same 1:11, proposed elastic net type methods consistently uphold FDR levels below the threshold across all correlation settings. Furthermore, they demonstrate TPRs close to 0.5 in all correlations of 0.6 or less. This signifies enhanced performance compared to adjusted lasso type methods, which yield

TPRs under 0.3. Between the two imbalance adjustment methods of elastic net type MDS, which are (MDS, Elastic net, Downsampling) and (MDS, Elastic net, LHO-LOO), there is a marginal difference in TPR. The LHO-LOO method tends to slightly outperform across most correlation and imbalance ratio combinations. Consequently, the proposed imbalance adjustment methods based on elastic net show effective FDR control in imbalance situations, especially in the presence of correlation, and demonstrate superior performance compared to existing methods.

This study is significant as it explores FDR control and TPR performance in regression based variable selection for class imbalanced data. Additionally, it discusses a novel combination approach to data splitting and provides a comprehensive comparison existing methods.

In binary data prediction and classification, regularization parameter λ is selected via cross-validation which minimize classification error. Ahmed⁴ applied stability selection to alleviate the burden of parameters in sparse selection methods. Among our methods, MDS involves repeating at least 50 times to integrate the results, ensuring the stability of the selection process. Therefore, we conducted parameter tuning via cross validation for λ of Lasso and elastic net. In addition, we found that setting the α of the elastic net to 0.5 resulted in its most robust performance for moderately correlated imbalance dataset. However, when correlation is high, the TPR increase continuously and faster than the FDR. Therefore, when applying the elastic net type methods, it is crucial to select an appropriate α based on the correlation of the data.

Additionally, we only considered 1:4 downsampling and LHO-LOO as imbalance adjustment methods. Other widely recognized methods that include Synthetic Minority Oversampling Technique (SMOTE) and random oversampling, which increase the sample size of the minority class to address class imbalance. In the future, performance comparison studies, which applied different imbalance adjustment methods, could be conducted.

7. Supplementary

Tuning the parameters of Lasso and elastic net penalties is a crucial aspect. It is common practice to select parameters that minimize the cross-validation error. For binary response data, the measurement is represented as a misclassification rate. However, parameters considered from a prediction and classification perspective may not necessarily yield the same performance in variable selection. Ahmed⁴ also mentioned this concern occurring in their Lasso logistic model-based signal detection. They alleviated the burden of parameter estimation by repeating their selection procedure and integrating the results obtained through sub-sampling, while using the λ from the k-fold cross-validation. The MDS in section 2.4.2, also repeats DS more than 50 times and averages them to calculate the inclusion rate for each variable. Therefore, in this paper, we performed 10-fold cross validation to estimate the penalty parameters.

The elastic net penalty is defined as

$$\lambda P(\boldsymbol{\beta}) = \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p \beta_j^2 \right)$$

where α ($0 \leq \alpha \leq 1$) is a hyperparameter that determines the balance between Lasso and ridge. When $\alpha = 1$, it represents the Lasso penalty. Tuning for elastic net was conducted by the “cva.glmnet” function from the “glmnetUtils” package in R. This function enables the estimation of both α and λ . For each specified α , it computes the performance across

different λ values and selects the (α, λ) combination with the lowest cross-validation error as the parameter estimation values.

We conducted simulations related to the α value of elastic net. Using the data generated in section 4, we compared the performance across α values under each imbalance ratios. The considered imbalance parameter $\eta \in \{0, -3, -3.5, -4, -4.5, -5, -5.5, -5.7\}$, indicating the ratios as $\{1:1, 1:6, 1:8, 1:11, 1:15, 1:22, 1:31, 1:36\}$, α values are $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and $\rho \in \{0.3, 0.5, 0.8\}$.

In weakly imbalanced data, the TPR increases with larger α values and then stabilizes. However, as the imbalance ratio increases, the TPR peaks around $\alpha = 0.5$. The results of changing ρ under the same imbalance conditions are shown in Figure 23 and Figure 24. α appears to be more sensitive to the degree of imbalance than to the correlation coefficient. In strong correlations at $\alpha = 0.8$, TPR continues to rise, but FDR also inflated significantly. FDR should be recognized as being in control when α is below 0.6. Based on these findings, we set α to 0.5 for elastic net type, and then estimated the appropriate λ for each case through cross-validation.

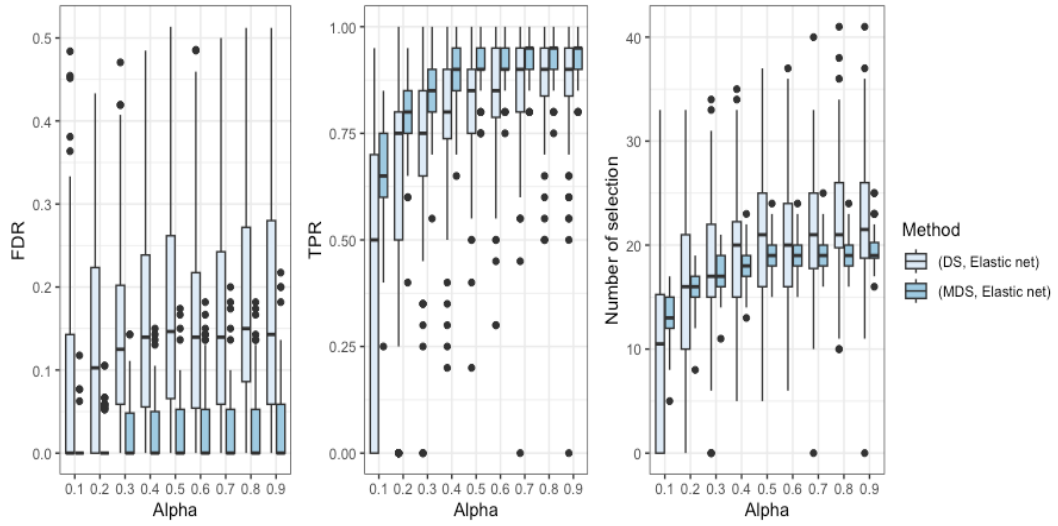


Figure 15. FDRs, TPRs, and number of selections for elastic net type data splitting methods under different $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ values on balanced data with correlation $\rho = 0.5$.

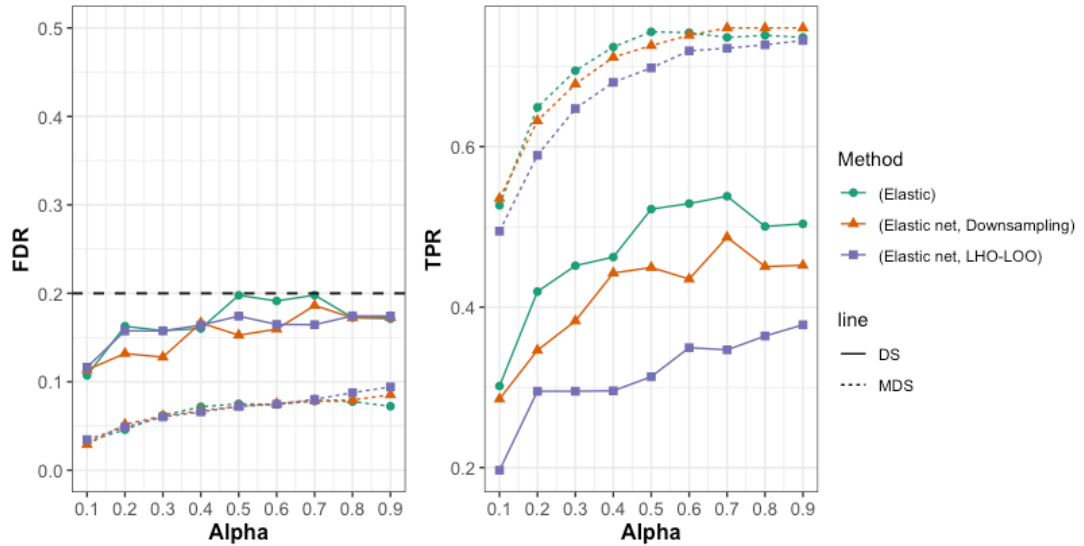


Figure 16. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -3$) with correlation ($\rho = 0.5$).

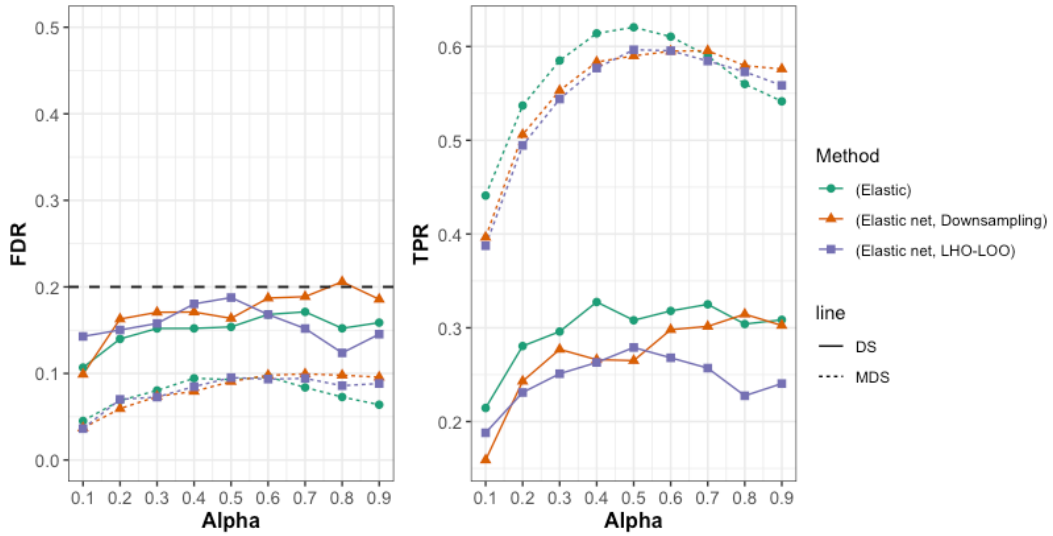


Figure 17. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -3.5$) with correlation ($\rho = 0.5$).

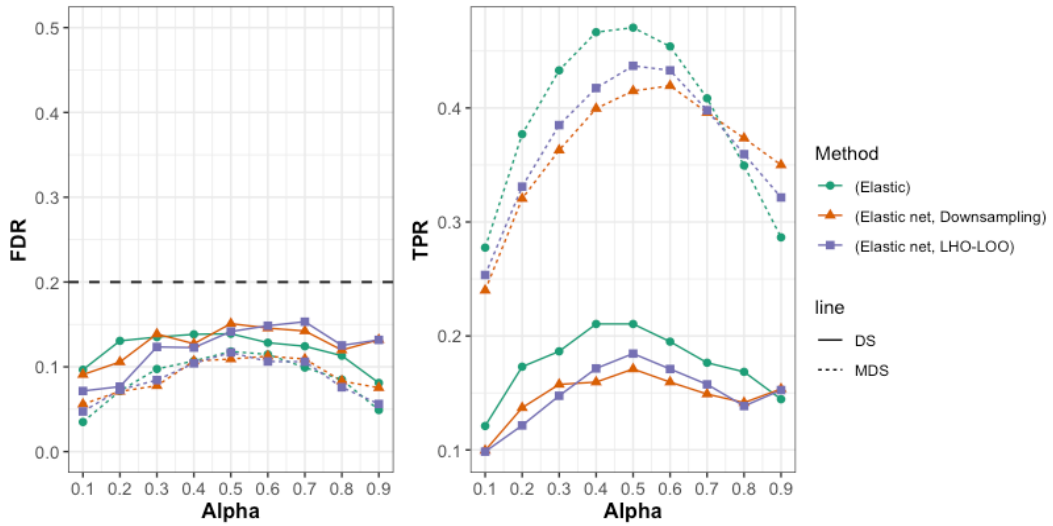


Figure 18. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.5$).

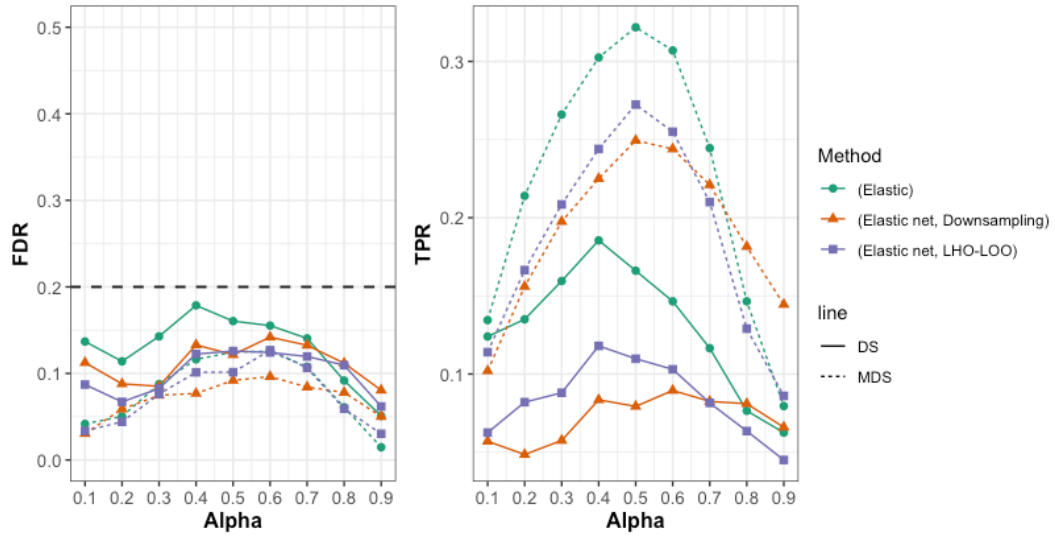


Figure 19. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4.5$) with correlation ($\rho = 0.5$).

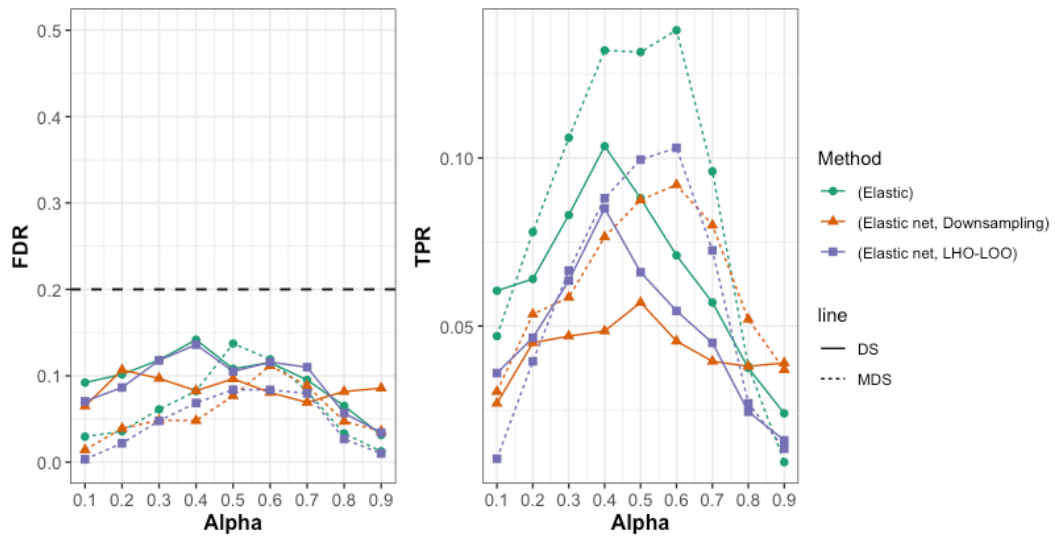


Figure 20. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5$) with correlation ($\rho = 0.5$).

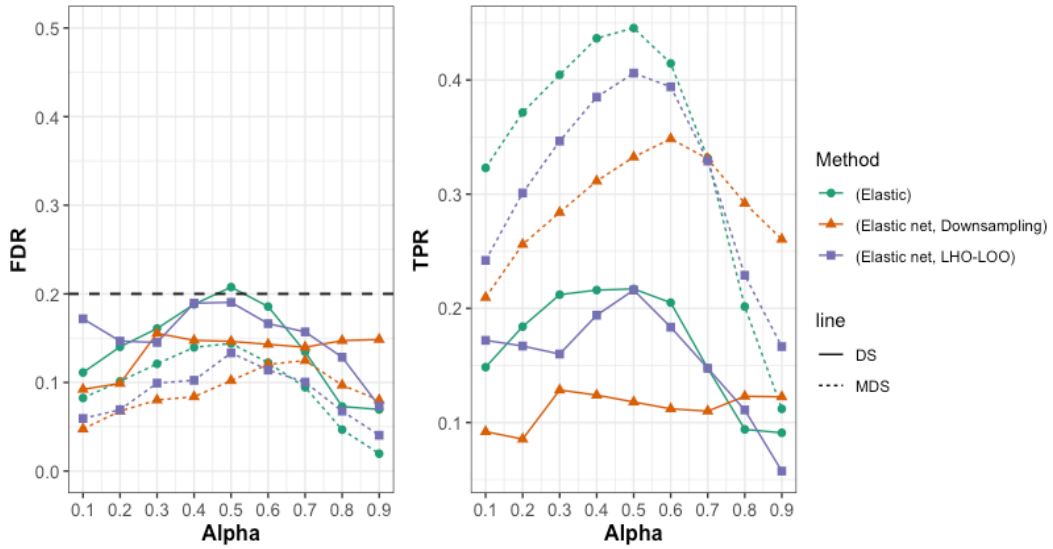


Figure 21. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5.5$, $n = 2,000$) with correlation ($\rho = 0.5$).

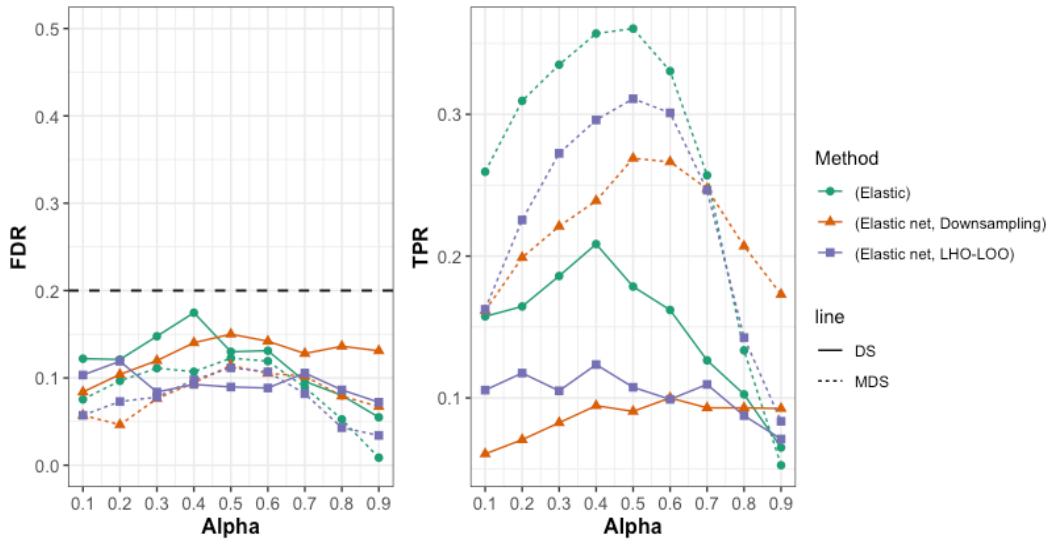


Figure 22. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -5.7$, $n = 2,000$) with correlation ($\rho = 0.5$).

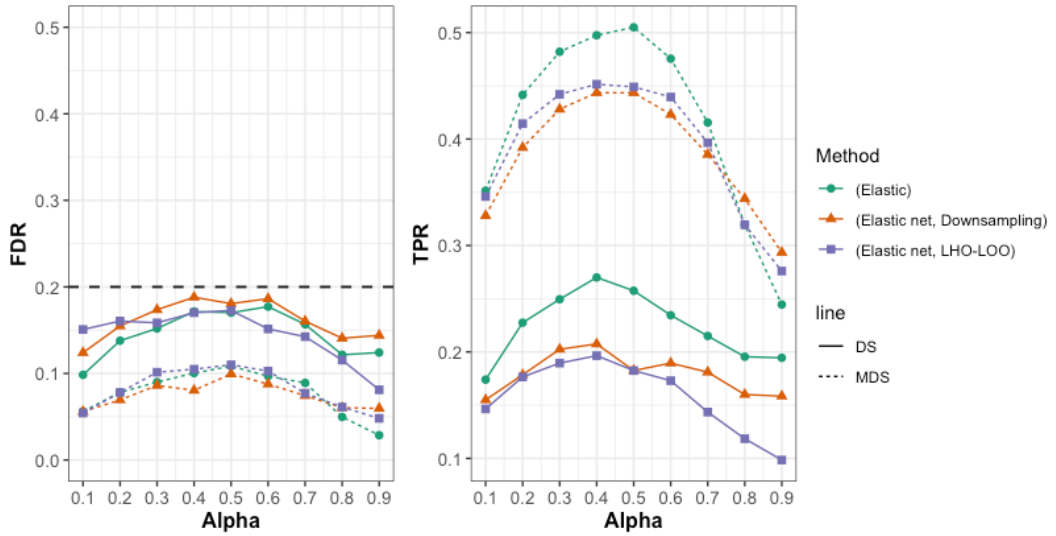


Figure 23. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.3$).

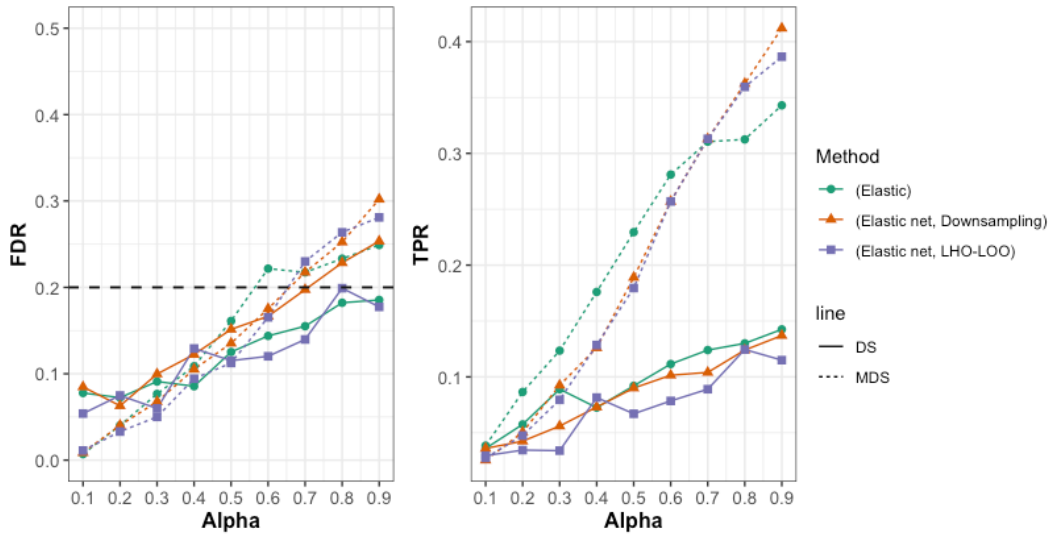


Figure 24. FDRs, and TPRs for elastic net type proposed data splitting methods under different α values on imbalanced data ($\eta = -4$) with correlation ($\rho = 0.8$).

References

- Algama, Zakariya Yahya, and Muhammad Hisyam Lee. "Applying penalized binary logistic regression with correlation based elastic net for variables selection." *Journal of Modern Applied Statistical Methods* 14.1 (2015): 15.
- Ahmed, Ismail, Antoine Pariente, and Pascale Tubert-Bitter. "Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions." *Statistical Methods in Medical Research* 27.3 (2018): 785-797.
- Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (1995): 289-300.
- Benjamini, Yoav, and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *Annals of Statistics* (2001): 1165-1188.
- Cawley, Gavin C., and Nicola LC Talbot. "Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters." *Journal of Machine Learning Research* 8.4 (2007).
- Candes, Emmanuel, et al. "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018): 551-577.
- Dai, Chenguang, et al. "False discovery rate control via data splitting." *Journal of the American Statistical Association* 118.544 (2023): 2503-2520.
- Efroymson, Michael Alin. "Multiple regression analysis." *Mathematical Methods for Digital Computers* (1960): 191-203.

Fu, Guang-Hui, et al. "Stable variable selection of class-imbalanced data with precision-recall criterion." *Chemometrics and Intelligent Laboratory Systems* 171 (2017): 241-250.

Gimenez, Jaime Roquero, and James Zou. "Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization." *PMLR* 89 (2019): 2184-2192.

Hou, Dake, et al. "A comparative study of different variable selection methods based on numerical simulation and empirical analysis." *PeerJ Computer Science* 9 (2023): e1522.

Kamalov, Firuz, Fadi Thabtah, and Ho Hon Leung. "Feature selection in imbalanced data." *Annals of Data Science* 10.6 (2023): 1527-1541.

Liang, Yong, et al. "Sparse logistic regression with a $\ell_{\frac{1}{2}}$ penalty for gene selection in cancer classification." *BMC bioinformatics* 14 (2013): 1-12.

Liu, Jingbo, and Philippe Rigollet. "Power analysis of knockoff filters for correlated designs." *Advances in Neural Information Processing Systems* 32 (2019).

Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4 (2010): 417-473.

O'hara, Robert B., and Mikko J. Sillanpää. "A review of Bayesian variable selection methods: what, how and which." *Bayesian Analysis* 4 (2009): 85-117.

Shevade, Shirish Krishnaj, and S. Sathya Keerthi. "A simple and efficient algorithm for gene selection using sparse logistic regression." *Bioinformatics* 19.17 (2003): 2246-2253.

Streiner, David L., and Geoffrey R. Norman. "Correction for multiple testing: is there a resolution?." *Chest* 140.1 (2011): 16-18.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996): 267-288.

Thiessard, Frantz, et al. "Trends in spontaneous adverse drug reaction reports to the French Pharmacovigilance system (1986-2001)." *Drug Safety* 28 (2005): 731-740.

국 문 요 약

클래스 불균형 자료에서 데이터 분할을 통한 FDR 통제하의 변수 선택 방법

고차원 자료에서 반응변수와 유의미한 연관성을 갖는 독립변수를 선택하는 것은 다양한 분야에 적용되고 있다. 그러나 선택된 변수 중에서 실제로 응답 변수와 관련이 없는 변수가 다수 포함될 수 있는 문제점이 있다. 특히 심각한 클래스 불균형 자료에서 단순 Lasso regression의 변수 선택은 잘못된 발견 비율 (False Discovery Rate, FDR)을 증가시킨다. FDR 통제 방법을 구현하더라도 실제 양성 비율 (True Positive Rate, TPR)이 매우 낮을 수 있다. 본 논문에서는 클래스 불균형 자료에서 FDR 통제 하에 데이터 분할을 통한 변수 선택 시 TPR을 향상시키기 위한 두 가지 접근 방법을 제안한다. 1) 패널티화 회귀의 확장, 2) 불균형 비율 조정 방법의 적용. 비교에는 Benjamini-Hochberg procedure과 Knockoff framework를 사용한다. 시뮬레이션 연구를 통해 기존 방법보다 제안된 조정 방법에서 높은 성능임을 확인하였다.

핵심되는 말: 잘못된 발견 비율 (False Discovery Rate, FDR); 불균형 비율; 벌점화 회귀; 녹오프 프레임워크