Mathematics and Statistics 11(1): 92-100, 2023
DOI: 10.13189/ms.2023.110110 http://www.hrpub.org

ANOVA Assisted Variable Selection in High-dimensional Multicategory Response Data

Demudu Naganaidu[1,*], **Zarina Mohd Khalid**[2]

[1]Centre for Postgraduate Studies, Asia Metropolitan University, Malaysia

[2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Malaysia

Received October 4, 2022; Revised November 25, 2022; Accepted December 23, 2022

Cite This Paper in the following Citation Styles

***(a)**: [1] Demudu Naganaidu, Zarina Mohd Khalid, "ANOVA Assisted Variable Selection in High-dimensional Multicategory Response Data," Mathematics and Statistics, Vol.11, No.1, pp. 92-100, 2023. DOI: 10.13189/ms.2023.110110*

***(b)**: Demudu Naganaidu, Zarina Mohd Khalid, (2023). ANOVA Assisted Variable Selection in High-dimensional Multicategory Response Data. Mathematics and Statistics, 11(1), 92-100. DOI: 10.13189/ms.2023.110110*

Abstract Multinomial logistic regression is preferred in the classification of multicategory response data for its ease of interpretation and the ability to identify the associated input variables for each category. However, identifying important input variables in high-dimensional data poses several challenges as the majority of variables are unnecessary in discriminating the categories. Frequently used techniques in identifying important input variables in high-dimensional data include regularisation techniques such as Least Absolute Selection Shrinkage Operator (LASSO) and sure independent screening (SIS) or combinations of both. In this paper, we propose to use ANOVA, to assist the SIS in variable screening for high-dimensional data when the response variable is multicategorical. The new approach is straightforward and computationally effective. Simulated data without and with correlation are generated for numerical studies to illustrate the methodology, and the results of applying the methods on real data are presented. In conclusion, ANOVA performance is comparable with SIS in variable selection for uncorrelated input variables and performs better when used in combination with both ANOVA and SIS for correlated input variables.

Keywords ANOVA, High-dimensional Data, Sure Independence Screening, LASSO, Multinomial Logistic Regression

1 Introduction

Multinomial logistic regression (MLR) is a statistical tool to model dependent variables with categorical responses. In this model, probability prediction describes the relationship between dependent or response variables (Y) with p input variables or predictors (X). As the probabilities are numerical, MLR is a type of 'Regression'. However, the purpose of the MLR model is for 'Classification' based on these probabilities. Several classification methods are found in the literature to model data with the categorical response variable. Frequently used methods are support vector machines (SVM), neural networks, decision trees, logistic regression, hierarchical classifications and linear discriminant analysis (LDA) [1, 2]. However, logistic regression is preferred because it is easily interpreted and provides input variables or predictors associated with each category [3, 4].

Among the p input variables, not all are helpful for the statistical model. Sparse models make interpretation easy, improve computation time, and maximise model performance. A parsimony model is a desired property in any model building. Besides, the input variables are expected to be independent with no correlation or with low degree of correlation. Several methods such as forward, backward, stepwise, and best subset are methods available for variables selection [5]. Each of these methods imposes some selection criteria. Some of frequently used selection criteria are Akaike's information criterion (AIC_p), Schwarz' Bayesian criterion (SBC_p), and Mallows' C_p criterion [6, 7]. Several modifications and extension to these criteria can be found in [6].

Advancement in computing technologies has led to the development of vast amount of data. The number of features or variables collected for each sample can even exceed the number of samples. Data are characterised as high-dimensional when the number of features or input variables (p), is more than the number of the observations (n), often referred as $p \gg$

n [8, 9]. For example, high-throughput microarray technologies allow researchers to evaluate tens of thousands of genes in a single experiment [10].

High-dimensional data pose several challenges [11, 12, 13]. Computational complexity [12], poor interpretability and high accuracy during training but poor performance with test data [14] are the main challenges. Most importantly, the majority of the input variables are frequently unnecessary in discriminating between samples. As a result, prior to or during analysis, such as regression, classification, or clustering, some techniques for variable selection are required.

Regularisation is one method for variable selection used to solve the $p \gg n$ problem. For example SCAD [15], Danzig selector [16], adaptive LASSO [17], LASSO [18], and their related methods. Although regularisation techniques like LASSO [18] are helpful for automatically selecting variables in high-dimensional data, these methods fail when the data become ultra-high-dimensional space. Ultra-high-dimensional data refers to dataset with $\log(p) = O(n\alpha)$ for some $0 \leq \alpha \leq 1$ [19]. It can be challenging to automatically and correctly choose variables in ultra-high-dimensional space.

Fan and Lv [19] pioneered the sure independence screening (SIS) and iterative sure independence screening (ISIS) methods to solve this problem in a regression context. They claimed that all-important variables in the model are preserved with a probability near 1 by selecting variables using the SIS technique. The approach, which is easy to understand and computationally effective, aims to reduce dimensionality of input variables prior to variable selection by using regularised model learning. Later, Fan et al. [20] expanded the SIS approach for generalised linear models, naming it as vanila SIS (Van-SIS) and introduced two variants of ISIS, naming as first variant of ISIS (Var1-SIS) and second variant of ISIS (Var2-SIS), both variants needed when input variables that are marginally unrelated but jointly related to response variable.

In Var1-SIS and Var2-SIS, the sample data need to be split into two halves randomly for independent learning. Input variables that overlap from the selected variables from each half of sample data are included for the final variable selection via regularisation. However, in highly correlated input variables, this approach may result in poor identification of important input variables due to smaller sample size for learning. In this paper, a solution to the problem is proposed by screening input variables with an analysis of variance (ANOVA) and SIS before variable screening via LASSO. The rest of the article is organized as follows. Section 2 discusses the proposed method, while Section 3 covers numerical results based on simulation data to illustrate the effectiveness of the proposed method. Section 4 presents the proposed method application on real data. Lastly, Section 5 discusses the conclusion and limitations of the proposed method.

2 Materials and Methods

In this section, MLR, ANOVA, SIS and LASSO methods are briefly reviewed. This is followed with ANOVA assisted sure independent screening (ANOVA-SIS) methodology. Two variants of ANOVA-SIS are introduced as Var1-ANOVA-SIS and Var2-ANOVA-SIS.

2.1 Multinomial Logistic Regression

Consider dataset with K category response variable, $Y_i \in [0, 1, ..., K]$ and $X = [X_0, X_1, ..., X_p]$ where $X_0 \equiv 1$ be the multivariate independent variables that influence the response $Y_i \sim multinomial(n = 1, P = (P_{i0}, ..., P_{iK}))$, where

$$\begin{aligned} P(Y_i = 0|X_i) &= P_{i0} \\ P(Y_i = 1|X_i) &= P_{i1} \\ &\cdot \\ &\cdot \\ P(Y_i = K|X_i) &= P_{iK} \end{aligned}$$

For each observation (X_i, Y_i), Y_i can only take one value of $[0, 1, ..., K]$. This can be represented by a vector, $Y_i = (Y_{i0}, Y_{i1}, ..., Y_{iK})$, where

$$Y_{ik} = \begin{cases} 1, & \text{if } Y_{ik} = k, k = 0, 1, ..., K \\ 0, & \text{otherwise.} \end{cases}$$

subject to:

$\sum_{k=0}^{K} Y_{ik} \equiv 1$, $P_{ik} \in [0, 1]$ and $\sum_{k=0}^{K} P_{ik} \equiv 1$

Following Hosmer Jr et al. [21], setting category 0 as the baseline category, each probability P_{ik} can be computed from a softmax function, written as

$$P_{ik} = \frac{exp(X_i\beta_k)}{1 + \sum_{k=1}^{K} exp(X_i\beta_k)}$$

where $\beta_0, \beta_1, ..., \beta_K$ are unknown regression coefficients vectors with $p + 1$ entries, collectively represented by a matrix β of dimension $K \times (p + 1)$,

$$\beta = \begin{bmatrix} \beta_{10} & \beta_{11} & ... & \beta_{1(p+1)} \\ \beta_{20} & \beta_{21} & ... & \beta_{2(p+1)} \\ . & . & ... & . \\ . & . & ... & . \\ \beta_{K0} & \beta_{K1} & ... & \beta_{K(p+1)} \end{bmatrix}$$

and $\beta_0 = 0$, a vector with all $p + 1$ entries is zero. Logit function for category k can be written as:

$$\begin{aligned} g_k(X_i) &= ln\left(\frac{P(Y_i = k|X_i)}{P(Y_i = 0|X_i)}\right) \\ &= ln\left(\frac{exp(X_i\beta_k)}{1 + \sum_{k=1}^{K} exp(X_i\beta_k)} \Big/ \frac{exp(X_i\beta_0)}{1 + \sum_{k=1}^{K} exp(X_i\beta_k)}\right) \\ &= X_i\beta_k, \quad \text{since} X_i\beta_0 = 0, \quad \exp(X_i\beta_0) = 1 \end{aligned}$$

The logit can be interpreted as log of odds ratio of observing category k over category 0 given the input variables, X, expressed as a linear model, $X_i\beta_k$.

For a single observation, X_i is a vector of $1 \times (p + 1)$ dimension that goes through a summation with parameter matrix β^T and gives a vector Z_i with the dimension of $1 \times K$.

$$X_i \xrightarrow{\beta} Z_i$$

Each Z_i goes through the softmax function and gives K probabilities for K categories, P_{ik} with $k = [1, ..., K]$. Probability for baseline category i.e for category 0 then can be derived from $1 - \sum_{k=1}^{K} P_{ik}$. The response variable will be predicted as category k, based on the maximum probability of $P_{ik}, k = 0, 1, ..., K$.

Maximum likelihood estimation (MLE) is the common method used in estimating the matrix β in MLR model. MLE can be written as

$$\begin{aligned}
L(\beta|X_i; Y_i) &= \prod_{i=1}^{n}\prod_{k=0}^{K} P_{ik}^{Y_{ik}} \\
\mathcal{L}(\beta|X_i; Y_i) &= ln\prod_{i=1}^{n}\prod_{k=0}^{K} P_{ik}^{Y_{ik}} \text{ , take natural log} \\
&= \sum_{i=1}^{n}\sum_{k=0}^{K} Y_{ik} ln P_{ik} \\
&= \sum_{i=1}^{n}(1 - \sum_{k=1}^{K} Y_{ik}) ln P_{i0} + Y_{i1} ln P_{i1} + ...+ \\
&\quad Y_{iK} ln P_{iK} \\
&= \sum_{i=1}^{n} Y_{i1} ln\frac{P_{i1}}{P_{i0}} + ... + Y_{iK} ln\frac{P_{iK}}{P_{i0}} + \\
&\quad ln\frac{1}{1 + \sum_{k=1}^{K} g_k(X_i)} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik} g_k(X_i) - \sum_{i=1}^{n} ln(1 + \sum_{k=1}^{K} g_k(X_i))
\end{aligned}$$

Minimising $\mathcal{L}(\beta||X_i; Y_i)$ the MLR estimator is the

$$\hat{\beta}^{MLE} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{k=1}^{K} d(Y_{ik}, X_i^T \beta_k) \qquad (1)$$

where $d(Y_{ik}, X_i^T \beta_k) = -\mathcal{L}(\beta|X_i; Y_i)$ is the deviance for the i^{th} observation.

No closed form solutions can be found to solve (1). For optimised solution, one has to resort to iterative algorithms such as Newton method or gradient descent method.

2.2 Analysis of Variance

Analysis of variance (ANOVA) is a statistical technique used to compare the means of more than two groups by analysing variances. ANOVA answers the statistical question on the null hypothesis: the assumption that all groups are equal and drawn from the same population. Any difference among groups comes from random sampling differences. Essentially ANOVA, answers the question of whether group means differ from each other.

Following [22, Chapter 15], consider N observations were sampled randomly from G groups with $n_1, n_2, ..., n_G$ representing the sample size from each group. Let y_{gj} be the j^{th} observation from group g. The data from the G groups can be presented as per the Table 1. If the G group means are represented by $\mu_1, \mu_2, ..., \mu_G$, the null hypothesis can be equally tested as follows:

$$\begin{aligned}
&H_0 : \mu_1 = \mu_2 = ... = \mu_G \\
&H_1 : \mu_i \neq \mu_j \qquad \text{At least one pair of } \mu_i \mu_j
\end{aligned}$$

Table 1. Observations from G Groups

Group			
1	2	...	G
y_{11}	y_{21}	...	y_{G1}
y_{12}	y_{22}	...	y_{G2}
.	.	...	.
.	.	...	.
.	.	...	.
y_{1n_1}	y_{2n_2}	...	y_{Gn_G}

Each group g sample mean is denoted as $\bar{y}_1, \bar{y}_2, ..., \bar{y}_G$, which are derived as follows:

$$\bar{y}_g = \frac{\sum_{j=1}^{n_g} y_{gj}}{n_g} \left(g = 1, 2, ..., G \text{ and } N = \sum_{k=1}^{G} n_g \right)$$

The overall mean is expressed as

$$\bar{\bar{y}} = \frac{\sum_{j=1}^{n_g}\sum_{j=1}^{n_g} y_{gj}}{N} = \frac{\sum_{j=1}^{n_g} n_g \bar{y}_g}{N}$$

Variability in each group g, computed with sum of squared of each observations about their sample mean $\bar{y_g}$. that is,

$$SS_g = \sum_{j=1}^{n_g} (y_{gj} - \bar{y_g})^2$$

The total within-groups variability, denoted as SSW is each SS_g, that is,

$$SSW = SS_1 + SS_2 + ... + SS_G = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{gj} - \bar{y}_g)^2$$

and variations between groups SSG are computed as follows:

$$SSG = \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{\bar{y}})^2$$

Overall measure of variability SST can be computed by taking sum of squared all sample observations from overall sample mean, as expressed as

$$SST = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{gj} - \bar{\bar{y}})^2$$

The mean squares within group MSW and between groups MSG are given by

$$MSW = \frac{SSW}{N-G}$$
$$MSG = \frac{SSG}{G-1}$$

The ratio of the mean squares is basis for testing null hypothesis. The ratio denoted as F_{STAT}, is a test statistic.

$$F_{STAT} = \frac{MSG}{MSW}$$

The F_{STAT} test statistic follows an **F distribution**, with $G-1$ degrees of freedom in the numerator and $N-G$ degrees of freedom in the denominator. A test of significance level α is provided by the decision rule, reject H_0 if $F_{STAT} \geq F_{G-1,N-G,\alpha}$.

2.3 Sure Independence Screening

Following the SIS method in [19, 20], the marginal utility L_j is the negative log-likelihood computed with one input variable. The marginal utility for the j^{th} independent variable x_j for $j = 1, .., p$, with response variable $y_i, i = 1, .., n$, is defined by:

$$L_0 = \underset{\beta_0}{\operatorname{argmin}}\, l(\beta) = \mathcal{L}(\beta_0, Y_i), \quad \text{and}$$
$$L_j = \underset{\beta_0,\beta_j}{\operatorname{argmin}}\, l(\beta) = \mathcal{L}(\beta_0 + X_j\beta_j, Y_i)$$

The marginal utilities $L_1, .., L_p$ are then ranked in ascending order giving, $L_{v_1}, L_{v_2}, .., L_{v_q}, \ldots, L_{v_p}$ from where q vector of input variables $(x_{v_1}, x_{v_2}, .., x_{v_q})$ is selected. Here, $q = \lfloor n/4\log n \rfloor$, for multicategory classification, as suggested by Fan et al. [20]. With, $q < n$, computational complexity is reduced, and low dimensional statistical methods can be applied.

2.4 LASSO

Least absolute shrinkage and selection operator or LASSO [18] impose penalty term with L1-norm on model coefficients. The L1-norm constraint yields a sparse solution by assigning zero coefficients to a subset of the variables; thus LASSO provides an automatic variable selection. L1-norm regularization or LASSO penalty term is frequently used when dealing with high-dimensional data [8] to produce sparse model and highly interpretable model.

Penalised multinomial logistic regression (PMLR) is achieved by adding a non negative term to (1) given regularised parameter β^{PMLR}. $\lambda \geq 0$ is a regularisation parameter.

$$\hat{\beta}^{PMLR} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n}\sum_{k=1}^{K} d(Y_{ik}, X_i^T\beta_k) + \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\beta_{kj}| \right] \tag{2}$$

where $j \in \{1, 2, ..., p\}$ excluding the intercept parameter β_0. If $\lambda = 0$, $\hat{\beta}^{PMLR} = \hat{\beta}^{MLE}$.

However Fan and Lv [19] and Fan et al. [20] demonstrated that directly applying LASSO without independent variable screening can result in recruiting many unwanted input variables.

2.5 ANOVA Assisted Sure Independence Screening

In variable selection for the MLR model, the input variable that is useful in discriminating a category from the G categories is potentially an important variable. Categorical responses can be used to partition each input variable observation into its respective category bucket as per Table 1. The categories can be assumed as a factor with G levels, which is equivalent to placing each observation of the input variable X_j into a specific group $g, (g = 1, 2, ..., G)$. This way of partitioning is similar to the setup of ANOVA. As such, if the input variable is random with no contribution to any particular category, the means computed from each group are expected not to differ significantly from each other. Treating all input variables as the dependent variable and partitioned into G groups, the ANOVA test can be conducted component-wise. Since the computation of each ANOVA only involves one input variable at a time, the computational complexity is reduced $O(np)$ in a similar way in SIS [19] method.

Only variables that produce significant results from the ANOVA test will be selected as the candidate variable, i.e input variables that contribute for discriminating at least one group from the rest of the groups. Here the selection of variables is refereed as ANOVA sure independence screening (ANOVA-SIS). The first $d \leq n$ vector of input variables selected after reordering is based on the F_{STAT} values in descending order. Variables selected via ANOVA-SIS may still include some unimportant variables. Regularization via LASSO as in (2) will further remove unimportant variables [19, 20]. In this way, the estimation procedure is formally defined as Van-ANOVA-SIS. The final model is built from the set of variables selected from Van-ANOVA-SIS for purpose of prediction of categories from given input variables.

The above two steps variable selection is outlined in details in two steps as follows:

- Step 1 : Variable Screening by ANOVA
  - Step 1.1: Split data set to traning data and test data in ratio of $80:20$.
  - Step 1.2: Conduct ANOVA test component-wise for each input variable on the training dataset. For each significant test, record the input variable and the associated F_{STAT} values.
  - Step 1.3: Sort the F_{STAT} values in descending order and its corresponding variables recorded in step 1.2. The larger F_{STAT} value, is the better discriminating power by the input variable.
  - Step 1.4: Select the first d variables $\mathcal{A} = \{X_{i_1}, X_{i_2}, ..., X_{i_d}\}$ from step 1.3. Here, $d = \lfloor n/4\log n \rfloor$.
- Step 2 : Variable Selection via LASSO
  - Step 2.1: The d variables selected from step 1.4 are low-dimensional data with $d \leq p$. LASSO is applied to further select variables by choosing

the appropriate tuning parameter, λ, from cross validation method.

– Step 2.2: Final variables selected from step 2.1 by discarding variables with zero coefficients of β, leaving the final input variables in the model with size $d' \leq d$.

– Step 2.3: Predict the category k for test dataset from step 1.1 based on the final variables in step 2.2.

2.6 First Variant of ANOVA-SIS

Unlike splitting data into two halves as outlined by Fan et al. [20], the first variant of ANOVA-SIS (Var1-ANOVA-SIS) simply selects the set of variables size d from SIS and from ANOVA-SIS, each with $d = \lfloor n/4 \log n \rfloor$. Let $\mathcal{A}_1 = \{X_{i_1}, X_{i_2}, ..., X_{i_d}\}$ be the variables selected from SIS and $\mathcal{A}_2 = \{X_{i_1}, X_{i_2}, ..., X_{i_d}\}$ be the variables selected from Van-ANOVA-SIS. Input variables appear in both $\mathcal{A}_1$ and $\mathcal{A}_2$ and are then selected as the candidate input variables $\mathcal{A}$, where $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2$.

With similar argument by Fan et al [20], input variables that appear in $\mathcal{A}$ are much fewer as these input variables have to appear twice at random in the sets of $\mathcal{A}_1$ and $\mathcal{A}_2$. Final selection of important input variables is performed by applying LASSO.

2.7 Second Variant of ANOVA-SIS

The option of $d = \lfloor n/4 \log n \rfloor$ in Var1-ANOVA-SIS reduces the probability of including important input variables. The second variant of ANOVA-SIS (Var2-ANOVA-SIS) overcomes this by having a larger value of $d = \lfloor n/ \log n \rfloor$, to increase the probability in selecting the important input variables. However, this may also increase some unimportant variables but can be eliminated by applying LASSO in final selection of important input variables.

3 Numerical Study

To evaluate the performance of Van-ANOVA-SIS, Var1-ANOVA-SIS and Var2-ANOVA-SIS, simulated data generated in the context of multinomial logistics regression with $p = 1000$ input variables $X_1, ..., X_p$. 100 simulation data generated with each consist of sample size of 200. The size of true model set to 5, i.e the numbers of non-zero coefficients with important input variables are fixed as X_1, X_2, X_3, X_4 and X_5 by choosing non zero coefficients in matrix β. The coefficients and important input variables remain same for each simulation and the 5 coefficients are generated randomly as $(-1)^U (4 \log n/\sqrt{n} + |Z|)$ with $Z \sim \mathcal{N}(0, 1)$ and U is a random integer between $[-10, 10]$

Four different scenarios were considered for the input variables in :

- Scenario 1: $X_1, ..., X_P$ are independent and identically distributed $N(0, 1)$ random variables.
- Scenario 2: $X_1, ..., X_P$ are jointly multivariate normal distribution and marginally $N(0, 1)$ with no correlation among variables.
- Scenario 3: $X_1, ..., X_P$ are jointly multivariate normal distribution and marginally $N(0, 1)$ with correlation, $\text{corr}(X_i, X_4) = \frac{1}{\sqrt{2}}$ for all $i \neq 4$ and $\text{corr}(X_i, X_j) = \frac{1}{2}$ if i and j are distinct elements of $\{1, ..., p\} \setminus \{4\}$
- Scenario 4: $X_1, ..., X_P$ are jointly multivariate normal distribution and marginally $N(0, 1)$ with correlation, $corr\,(X_i, X_5) = 0$ for all $i \neq 5$, $corr\,(X_i, X_4) = \frac{1}{\sqrt{2)}}$ for all $i \notin \{4, 5\}$, and $corr\,(X_i, X_j) = \frac{1}{2}$ if i and j are distinct elements of $\{1, ..., p\} \setminus \{4, 5\}$.

Scenarios 1, 3 and 4 are same scenario setting found in [20].

All analyses were implemented in Jupyter Notebook using Python programming language (version 3.9.7). Throughout of the analysis, the tuning parameter λ is fixed at 0.0055 based on the cross validation. Gradient descent method used to optimise the penalised MLR model is stated in (2). After performing 100 simulations for each of the scenarios above, the performance was compared with Van-SIS, Var1-SIS, Var2-SIS and LASSO on the following metrics:

1. Proportions of important input variables before applying LASSO,
2. Proportions of important input variables after applying LASSO,
3. Median of final model size, and
4. Average test error

The comparisons were summarised in Table 2 to Table 5.

Table 2 shows the results of the seven methods, Van-ANOVA-SIS, Var1-ANOVA-SIS, Var2-ANOVA-SIS, Van-SIS, Var1-SIS, Var2-SIS and LASSO for scenario 1. In this scenario Van-ANOVA-SIS and Van-SIS recorded almost equivalent proportions in including important input variables before and after applying LASSO, i.e 80% and 81% respectively from 100 simulations. Similarly, the average test errors from the 100 simulations are respectively 0.0883 and 0.0878 while the median final model size of each method is 7. It can be said that the performance of Van-ANOVA-SIS method is at par with the Van-SIS method in scenario 1.

Var1-SIS performed poorly among all methods in scenario 1. Only 2% of the simulations included all important input variables both before and after applying LASSO, and an average test error of 0.2440. Median final model size is only 3 missing many important input variables. This is because the splitting of the data into halves has reduced the number of samples, hence the power of learning by the method in selecting the important input variables is reduced. Comparatively the Var1-ANOVA-SIS method which does the variable selection by choosing input variables that overlap from both SIS and ANOVA, recorded a proportion of 80% in including the important input variables both before and after ap- plying LASSO and average test error

of 0.0885. Here, Var1-ANOVA-SIS provides two advantages. First, the number of samples does not require to be halved hence better learning from samples available in hand. This is particularly pertinent when dealing with limited sample sizes, as is often the case with high-dimensional data [23]. Second, selecting overlap variables from two different variable selection methods i.e SIS and ANOVA, increases the probability of the selected variable as the important variable.

Var2-ANOVA-SIS recorded a higher proportion of 98% in including the important input variables before and after applying LASSO as a result of the increase in size d, double the size d of Var1-ANOVA-SIS method. Thus, the median final model size for Var2-ANOVA-SIS was higher at 18 relative to Var1-ANOVA-SIS. However, the average test error remained the same for both methods, indicating that a model with more variables does not necessarily make a good model. The LASSO method recorded 100% in including important variables in the final model, but included many other unimportant input variables. Median final model size for LASSO is 41 versus only 5 correct input variables. Here selecting an optimal size for d is important in ensuring a higher chance of capturing true important variables in the model and at the same time, unimportant variables are reduced.

Table 3 summarises the performance of the seven methods based on scenario 2. Except for LASSO which produces similar results as in scenario 1, all other 6 methods have shown improvement in including the correct input variables in the final model. The Var2-ANOVA-SIS method showed an improvement with 99% for including important variables in the selected model, however, the median final model size increased from 18 in scenario 1 to 19 in scenario 2. Similarly, the average test error increased from 0.0885 to 0.1033. It is observed that models with input variables that are jointly multivariate normal and with fewer unimportant variables, the probability of selecting relevant input variables increases significantly. This is followed by a lower average test error.

From Table 4 and Table 5, except for the LASSO method, all six methods performance has deteriorated. This is due to correlated input variables in scenario 3 and scenario 4. In scenario 3, Van-ANOVA-SIS and Van-SIS again produced almost similar results. Var1-ANOVA-SIS recorded 1% in including all the input variables from the 100 simulations, but Var1-SIS recorded 0%. As the size d in Var2-ANOVA-SIS increased relative to Var1-ANOVA-SIS, 6% of the simulations included the relevant input variables. On the other hand, the Var2-SIS method recorded 0%, meaning this method fails to include all the important variables even at least for once. Scenario 4, which has input variables much more correlated compared to scenario 3, has failed in including all important input variables for the six methods except for LASSO. However, average test error improved significantly for all six methods except for the LASSO method, with an increased average test of 0.1275 in scenario 4 compared with 0.1140 in scenario 3.

The LASSO method was able to include all pertinent input variables in all 4 scenarios. Interestingly the median final model size was reduced in the presence of correlated input variables, while the average test error was significantly affected. This indicates that the LASSO method is a very useful method of variable selection in the presence of correlated input variables.

Based on the numerical studies the proposed methods produced better results when comparing pair-wise, i.e. Van-ANOVA-SIS with Van-SIS, Var1-ANOVA-SIS with Var1-SIS and Var2-ANOVA-SIS with Var2-SIS. In the presence of correlated input variables, ANOVA-assisted SIS variable selection has a higher probability of including the true model as evidenced in Table 3 compared with variable selection with the SIS method alone.

4 Real data examples

As an illustration of application on real data, the proposed methods are applied on SRBCT microarray dataset reported in Khan et al. [24]. The dataset reports the classification of children's cancer to the small round blue cell tumours (SRBCT) into four categories of cancer known as neuroblastoma(NB), rhabdomysarcoma(RMS),non-Hodgkin lymphoma (NHL), and the Ewing family of tumours (EWS) using gene expression profiles. There are 2308 genes in 83 samples in this data collection. There were 29 cases of EWS, 11 cases of NHL, 18 cases of NB, and 25 cases of RMS, all of which were classed as 0 through 3 respectively. The data which initially made available on http://research.nhgri.nih.gov/microarray/Supplement/ (now no longer available) was copied from https://rdrr.io/cran/plsgenomics/man/SRBCT.html.

All input variables are normalised to have zero mean and a variance of one. Response variable recoded with number values between 0 to 3. Data were split into train data and test data with $80 : 20$ ratio, resulting in 66 samples for training and 17 samples for testing. Table 6 provides the summary of the results. Initial training and testing were conducted with $d = \lfloor n/4 \log n \rfloor$ as recommended for logistic regression models by Fan et al. [20]. However, this has produced very high test error due to possible missing of important genes in training the models. Upon increasing the number of input variables to $d = \lfloor n/2 \log n \rfloor$, Van-ANOVA-SIS test error reduced dramatically with only 7 genes included in the model. Further increase of $d = \lfloor n/ \log n \rfloor$, improved Van-SIS test error as equivalent to Van-ANOVA-SIS with 12 genes selected in the final model. Therefore, the proposed Van-ANOVA-SIS method has performed better than the Van-SIS method. By increasing value of d to 35, both Van-ANOVA-SIS and Van-SIS recorded zero test error with each selecting 16 genes for the final model.

5 Conclusions

ANOVA is comparable to SIS as a variable screening tool for uncorrelated input variables in the case of high-dimensional data with multi categorical response variables. A set of input variables from intersection of variables selected separately by ANOVA and SIS methods, increases the probability of such variables as important variables in the model. In correlated high-dimensional data, variable screening with both ANOVA and SIS, has higher probability in picking the true model. Mod-

els selected with both ANOVA and SIS demonstrated better accuracy in classification versus models selected with SIS.

The proposed method in this paper is limited to be used for data with categorical responses and with no outliers. Future research can be extended to include outliers in both input variables and response variable.

Acknowledgements

The authors would like to express their appreciation for the support from Universiti Teknologi Malaysia, UTMER2021 Project Number PY/2021/01377.

REFERENCES

[1] Asenso, T. Q., Zhang, H. & Liang, Y., "Pliable lasso for the multinomial logistic regression," Communications In Statistics - Theory And Methods, Vol. 51, No. 11, pp. 3596-3611, 2022. DOI:10.1080/03610926.2020.1800041

[2] Kim, H., Choi, B. S., & Huh, M. Y., "Booster in High Dimensional Data Classification," IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 1, pp. 29-40, 2016. DOI:10.1109/TKDE.2015.2458867

[3] Kim, Y., Kwon, S. & Heun Song, S., "Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data," Computational Statistics And Data Analysis, Vol. 51, No.3, pp. 1643-1655, 2006. https://doi.org/10.1016/j.csda.2006.06.007

[4] Lei, D., Zhang, H., Liu, H., Li, Z. & Wu, Y., "Maximal Uncorrelated Multinomial Logistic Regression," IEEE Access, Vol. 7, pp. 89924-89935, 2019. DOI:10.1109/ACCESS.2019.2921820

[5] Tutz, G., Pößnecker, W. & Uhlmann, L., "Variable selection in general multinomial logit models,", Computational Statistics And Data Analysis, Vol. 82, pp. 207-222, 2015. https://doi.org/10.1016/j.csda.2014.09.009

[6] Oda, R., "Consistent variable selection criteria in multivariate linear regression even when dimension exceeds sample size," Hiroshima Mathematical Journal, Vol. 50, No. 3, pp.339-374, 2020. DOI: 10.32917/hmj/1607396493

[7] Fan, J. & Song, R., "Sure independence screening in generalized linear models with NP-dimensionality," Annals Of Statistics, Vol. 38, No. 6, pp. 3567-3604, 2010. DOI: 10.1214/10-AOS798

[8] Alfons, A., Croux, C. & Gelper, S., "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," Annals Of Applied Statistics, Vol. 7, No. 1, pp. 226-248, 2013. DOI: 10.1214/12-AOAS575

[9] Hastie, T., Tibshirani, R., Friedman, J. & Friedman, J., "The elements of statistical learning," 2nd Edition, Springer New York, NY, 2009.

[10] Bootkrajang, J. & Kabán, A., "Classification of mislabelled microarrays using robust sparse logistic regression,", Bioinformatics, Vol. 29, No 7, pp. 870-877, 2013. https://doi.org/10.1093/bioinformatics/btt078

[11] Abramovich, F., Grinshtein, V. & Levy, T., "Multiclass classification by sparse multinomial logistic regression,", IEEE Transactions On Information Theory, Vol. 67, No. 7, pp. 4637-4646, 2021. DOI: 10.1109/TIT.2021.3075137

[12] Zhang, C., Zhou, Y., Guo, J., Wang, G. & Wang, X., "Research on classification method of high-dimensional class-imbalanced datasets based on SVM," International Journal Of Machine Learning And Cybernetics, Vol. 10, No.7, pp. 1765-1778, 2019. https://doi.org/10.1007/s13042-018-0853-2

[13] Donoho, D. L., "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality,"American Math. Society Lecture-Math Challenges Of The 21st Century, 2000.

[14] Piao, Y., Piao, M., Park, K. & Ryu, K. "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data,", Bioinformatics, Vol. 28, No. 24, pp. 3306-3315, 2012. DOI: 10.1093/bioinformatics/bts602

[15] Fan, J. & Li, R., "Variable selection via nonconcave penalized likelihood and its oracle properties,", Journal Of The American Statistical Association, Vol. 96, No. 456, pp. 1348-1360, 2001. https://doi.org/10.1198/016214501753382273

[16] Candes, E. & Tao, T., "The Dantzig selector: Statistical estimation when p is much larger than n," Annals Of Statistics, Vol. 35, No. 6 , pp. 2313-2351, 2007. DOI: 10.1214/009053606000001523

[17] Zou, H., "The adaptive lasso and its oracle properties,", Journal Of The American Statistical Association, Vol. 101, No. 476, pp. 1418-1429, 2006. https://doi.org/10.1198/016214506000000735

[18] Tibshirani, R., "Regression Shrinkage and Selection Via the Lasso," Journal Of The Royal Statistical Society: Series B (Methodological), Vol. 58, No. 1, pp. 267-288, 1996. https://www.jstor.org/stable/2346178

[19] Fan, J. & Lv, J., "Sure independence screening for ultrahigh dimensional feature space," Journal Of The Royal Statistical Society. Series B: Statistical Methodology, Vol. 70, No. 5, pp. 849-911, 2008. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[20] Fan, J., Samworth, R. & Wu, Y., "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," Journal Of Machine Learning Research, Vol. 10, pp. 2013-2038, 2009. https://pubmed.ncbi.nlm.nih.gov/21603590/

[21] Hosmer Jr, D., Lemeshow, S. Sturdivant, R., "Applied logistic regression," 3rd Edition, John Wiley Sons, 2013.

[22] Newbold, P., "Statistics for business and economics," 8th Edition, Pearson, 2013.

[23] Hall, P., Marron, J. S., & Neeman, A., "Geometric representation of high dimension, low sample size data," Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 67, No.3, pp. 427-444, 2005. DOI: https://doi.org/10.1111/j.1467-9868.2005.00510.x

[24] Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C. & Others, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, Vol. 7, No. 6 , pp. 673-679, 2001. DOI: 10.1038/89044

Table 2. Scenario 1

	Prop. incl models	Prop. incl final models	Median final model size	Average test error
Van-ANOVA-SIS	0.8	0.8	7	0.0883
Var1-ANOVA-SIS	0.8	0.8	7	0.0885
Var2-ANOVA-SIS	0.98	0.98	18	0.0885
Van-SIS	0.81	0.81	7	0.0878
Var1-SIS	0.02	0.02	3	0.244
Var2-SIS	0.27	0.27	4	0.1658
LASSO	N/A	1	41	0.1038

Table 3. Scenario 2

	Prop. incl models	Prop. incl final models	Median final model size	Average test error
Van-ANOVA-SIS	0.97	0.97	7	0.0785
Var1-ANOVA-SIS	0.97	0.97	7	0.0795
Var2-ANOVA-SIS	0.99	0.99	19	0.1033
Van-SIS	0.97	0.97	7	0.0797
Var1-SIS	0.07	0.07	3.5	0.2355
Var2-SIS	0.49	0.49	5	0.1362
LASSO	N/A	1	41	0.116

Table 4. Scenario 3

	Prop. incl models	Prop. incl final models	Median final model size	Average test error
Van-ANOVA-SIS	0.01	0.01	7	0.2923
Var1-ANOVA-SIS	0.01	0.01	6	0.3005
Var2-ANOVA-SIS	0.06	0.06	17	0.246
Van-SIS	0.01	0.01	7	0.29
Var1-SIS	0	0	2	0.3505
Var2-SIS	0	0	3	0.3255
LASSO	N/A	1	33.5	0.114

Table 5. Scenario 4

	Prop. incl models	Prop. incl final models	Median final model size	Average test error
Van-ANOVA-SIS	0	0	7	0.2123
Var1-ANOVA-SIS	0	0	6	0.2132
Var2-ANOVA-SIS	0	0	12.5	0.184
Van-SIS	0	0	7	0.212
Var1-SIS	0	0	2	0.3302
Var2-SIS	0	0	3	0.2588
LASSO	N/A	1	29	0.1275

Table 6. SRBCT Dataset

	d=n/4logn = 3		d=n/2logn = 7		d=n/logn = 15		d=35	
	Test error	# Vars	Test error	# Vars	Test error	# Vars	Test error	# Vars
Van-SIS	0.3529	3	0.2941	6	0.0588	12	0	16
Var1-SIS	0.7058	0	0.4706	1	0.2353	3	0.2353	9
Var2-SIS	Test error = 0.2941, # Vars = 6, (d=n/logn)							
Van-ANOVA-SIS	0.2353	3	0.0588	7	0.0588	12	0	16
Var1-ANOVA-SIS	0.5882	1	0.2353	2	0.2941	6	0.2353	11
Var2-ANOVA-SIS	Test error = 0.2941, # Vars = 6 , (d=n/logn)							
LASSO	Test error = 0, # Vars = 29							

Vars = Number of variables selected