Austrian Journal of Statistics
January 2019, Volume 48, 63–72.
http://www.ajs.or.at/
doi:10.17713/ajs.v48i2.806

Bayesian Variable Selection for Multiclass Classification using Bootstrap Prior Technique

Oyebayo Ridwan Olaniran Universiti Tun Hussein Onn Malaysia
Mohd Asrul Affendi Bin Abdullah Universiti Tun Hussein Onn Malaysia

Abstract

In this paper, the one-way ANOVA model and its application in Bayesian multi-class variable selection is considered. A full Bayesian bootstrap prior ANOVA test function is developed within the framework of parametric empirical Bayes. The test function developed was later used for variable screening in multiclass classification scenario. Performance comparison between the proposed method and existing classical ANOVA method was achieved using simulated and real life gene expression datasets. Analysis results revealed lower false positive rate and higher sensitivity for the proposed method.

Keywords: multiclass classification, Bayesian, variable selection, ANOVA.

1. Introduction

In machine or statistical learning, multiclass classification problem involves grouping of observed samples into three or more classes. Variable selection in multiclass classification is the process of identifying relevant subset of input variables that can positively improve the performance of a multiclass classifier. The dual task of classification and subset selection often arise in biological and medical applications, especially in genomic studies (Liu, Bensmail, and Tan 2012). High-dimensional data with large input and small sample size are often observed or reported in genomic studies. Many variable selection methods for multiclass classification task have been developed within the framework of Bayesian and classical approaches.

In a broader sense, variable selection can be divided into three types; filter, wrapper and embedded. Variable selection techniques or methods that are independent of classifier are referred to as filter. Filter methods are fast and easy to apply but posed with the tendency of selecting irrelevant variables. Wrapper method are similar to filter except that the selection procedure is based on the scores generated from a predetermined classifier. The major drawback is that, it is classifier sensitive which implies the subset selected are mostly not optimal. Embedded is an hybrid form of wrapper and filter, with filter or wrapper at the training stage before proceeding to the classification stage (Guyon, Weston, Barnhill, and Vapnik 2002; Guyon and Elisseeff 2003; Guyon, Gunn, Nikravesh, and Zadeh 2008; Peng, Wu, and Jiang 2010). Filter method has been applied in multiclass classification especially for high-dimensional datasets because of its fast computational time. Forman (2003) used Chi-square method, Wright and Simon (2003) used the classical one-way ANOVA method for

preliminary variable selection task. The two approaches are based on ranking the Chi-square or F statistic in descending order with top variables being the best subset. Alternatively, the p-values of the statistics may be reported and variables with p-value lower than a threshold level say 0.05 are selected as best candidate for further classification task (Hwang, Lee, and Park 2017). The classical one-way ANOVA method which authors like Guyon and Elisseeff (2003); Wright and Simon (2003); Qureshi, Oh, Min, Jo, and Lee (2017), among others used suffers from loss of information (Bertolino, Piccinato, and Racugno 1990; Solari, Liseo, and Sun 2008). Solari *et al.* (2008) has worked on Bayesian one-way ANOVA as a safe haven to loss of information issue. They suggested objective prior through Bayes factor as an alternative approach to handle one-way ANOVA model. Objective Bayes (let the data speak for themselves) are no way better than the classical approach as its often used when subjective priors are difficult to compute or elicit (Yahya, Olaniran, and Ige 2014; Olaniran, Olaniran, Yahya, Banjoko, Garba, Amusa, and Gatta 2016; Olaniran and Yahya 2017; Olaniran and Abdullah 2018; Olaniran, Abdullah, Pillay, and Olaniran 2018). Thus, in this paper, we developed a Bayesian one-way ANOVA test function using bootstrap prior Olaniran and Yahya (2017) for variable selection in multiclass classification problem.

2. One-way ANOVA multi-class variable selection

Suppose we have the training dataset $[\tau_t, y_{t1}, y_{t2}, \ldots, y_{tp}, t = 1, 2, \ldots, n]$, where τ_t is a categorical outcome that assumes $i = 1, 2, \ldots, k$ values and y_t is the vector of continuous input variables. The one-way ANOVA multiclass variable selection takes each input variable y_t as response and τ_t as treatment effect for a one-way ANOVA model given as;

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, i = 1, 2, ..., k, j = 1, 2, ..., n_i \tag{1}$$

where y_{ij} is now the response variable of interest, μ is the overall mean, τ_i is the effect of ith categorical predictor and ϵ_{ij} is the residual error that is distributed $N(0, \sigma^2)$. The traditional Analysis of variance focuses on testing the hypotheses:

$$H_0 : \tau_1 = ... = \tau_k$$

against

$$H_1 : \tau_i \neq \tau_j$$

for at least one pair of $i \neq j$.

The classical approach of testing the hypotheses relies on the use of statistic:

$$F_c = \frac{\sum_{i=1}^{k} n_i(\bar{y_i} - \bar{y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2/(n-k)} \tag{2}$$

(Solari *et al.* 2008). Under H_0, F_c is distributed $F(k-1, n-k)$. The null hypothesis H_0 is rejected if $F_c > F(k-1, n-k)$.

Model (1) can also be reparameterized to a regression format with categorical effect predictors treated as dummy variables. Let $\mathbf{1}_p$ and $\mathbf{0}_p$ be the $p \times 1$ vectors of $1's$ and $0's$. Thus the reparameterized model is given in matrix notation as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3}$$

where

$$\boldsymbol{Y'} = [y_{11}, y_{12}, ..., y_{ij}]$$

,

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \mathbf{1} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \dots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \dots & \mathbf{1}_{n_k} \end{bmatrix}$$

,

$$\boldsymbol{\beta'} = [\mu, \tau_1, \tau_2, ..., \tau_k]$$

.

The Maximum Likelihood Estimate for $\boldsymbol{\beta}$ follows from;

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \prod_{ij=1}^{n} \frac{\gamma}{\sqrt{2\pi}} \exp\left[\frac{-\gamma}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{4}$$

where $\gamma = \sigma^{-2}$ is the model precision.

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \frac{\gamma^{n/2}}{(2\pi)^{n/2}} \exp\left[\frac{-\gamma}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})\right] \tag{5}$$

It is pertinent to note that design matrix $\boldsymbol{X}$ is not of full rank and one of the solution is to reparameterized such that $\sum_{i=1}^{k} \tau_i = 0$. Thus $\boldsymbol{X}$ now becomes $\boldsymbol{X^*}$ with rows of $\boldsymbol{X^*}$ corresponding to τ_k observations replaced with -1 and columns of τ_k omitted completely from $\boldsymbol{X^*}$. Therefore,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^{*'}X^*})^{-1}(\boldsymbol{X^{*'}Y}) \tag{6}$$

The classical ANOVA follows from (6) with:

$$F_c = \frac{n-k}{k-1}\left(\frac{\hat{\boldsymbol{\beta}}\boldsymbol{'X^{*'}Y} - n\bar{Y}^2}{\boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}\boldsymbol{'X^{*'}Y}}\right) \tag{7}$$

2.1. Bayesian one-way ANOVA

The exponent term in (5) can be re-arranged so that we have:

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \frac{\gamma^{n/2}}{(2\pi)^{n/2}} \exp\left[\frac{-\gamma}{2}\left((\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{X\hat{\beta}} + \boldsymbol{X\hat{\beta}})'(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{X\hat{\beta}} + \boldsymbol{X\hat{\beta}})\right)\right]$$

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \frac{\gamma^{n/2}}{(2\pi)^{n/2}} \exp\left[\frac{-\gamma}{2}(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})'(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X'X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]$$

with estimate of variance $s^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})'(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})}{\boldsymbol{n-k}}$, then,

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \frac{\gamma^{n/2}}{(2\pi)^{n/2}} \exp\left[\frac{-\gamma}{2}(n-k)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X'X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]$$

If we let $v = n - k$, implies $n = k + v$,

$$L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma) = \frac{\gamma^{k/2}}{(2\pi)^{n/2}} \exp\left[\frac{-\gamma}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{X'X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] \times \gamma^{v/2} \exp\left(\frac{-\gamma v}{2s^{-2}}\right) \tag{8}$$

The natural conjugate prior to the likelihood in (8) is normal gamma prior given by:

$$p(\boldsymbol{\beta}|\gamma) = \frac{\gamma^{k/2}}{(2\pi)^{k/2}|\Sigma_0|^{\frac{1}{2}}} \exp\left[\frac{-\gamma}{2}(\boldsymbol{\beta}-\boldsymbol{\beta_0})'(\boldsymbol{\Sigma_0})^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta_0})\right] \tag{9}$$

and

$$p(\gamma) = \frac{1}{\Gamma(\frac{v_0}{2})\left(\frac{2s_0^{-2}}{v_0}\right)^{\frac{v_0}{2}}} \gamma^{\frac{v_0}{2}-1} \exp\left[\frac{-\gamma v_0}{2s_0^{-2}}\right] \tag{10}$$

The posterior distribution $p(\boldsymbol{\beta}, \gamma)$ can be obtained from the standard Bayes formula:

$$p(\boldsymbol{\beta}, \gamma|\boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{\beta}, \gamma)L(\boldsymbol{\beta}, \gamma|\boldsymbol{Y}, \boldsymbol{X})}{\int\int p(\boldsymbol{\beta}, \gamma)L(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\beta}, \gamma)d\boldsymbol{\beta}d\gamma} \tag{11}$$

For simplicity, the denominator of (11) is often dropped such that the posterior is of the form;

$$\begin{aligned} p(\boldsymbol{\beta}, \gamma|\boldsymbol{Y}, \boldsymbol{X}) \propto \exp\left[\frac{-\gamma}{2}(\boldsymbol{\beta}-\boldsymbol{\beta_0})'(\boldsymbol{\Sigma_0})^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta_0})\right] \times \gamma^{\frac{v_0}{2}-1} \exp\left(\frac{-\gamma v_0}{2s_0^{-2}}\right) \times \\ \exp\left[\frac{-\gamma}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{X'X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right] \times \gamma^{v/2} \exp\left(\frac{-\gamma v}{2s^{-2}}\right) \end{aligned} \tag{12}$$

From (12), it can be observed that the $p(\boldsymbol{\beta}, \gamma|\boldsymbol{Y}, \boldsymbol{X})$ is also Normal-Gamma distributed. Notationally,

$$p(\boldsymbol{\beta}, \gamma|\boldsymbol{Y}, \boldsymbol{X}) \sim N(\boldsymbol{\beta_n}, \boldsymbol{\Sigma_n}, v_n, s_n^{-2})$$

where

$$\boldsymbol{\Sigma_n} = (\boldsymbol{\Sigma_0^{-1}} + \boldsymbol{X'X})^{-1}$$

$$\boldsymbol{\beta_n} = (\boldsymbol{\Sigma_0^{-1}} + \boldsymbol{X'X})^{-1}(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\beta_0} + \boldsymbol{X'X}\hat{\boldsymbol{\beta}})$$

and $v_n = v_0 + n$,

$$s_n^{-2} = \frac{v_0 + n}{v_0 s_0^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'[\boldsymbol{\Sigma_0} + (\boldsymbol{X'X})^{-1}]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})}$$

.

2.2. Bootstrap prior one-way ANOVA

The Bayesian solution provided in section (3.1) requires either subjective prior elicitation or objective prior via monte - carlo sampling from the posterior distribution. In variable selection, monte-carlo approach will be computational intensive especially when posed with high-dimensional datasets. The bootstrap prior technique follows from empirical Bayes principle where prior hyperparameters are estimated from the data. Therefore, the empirical Bayes estimates of $\boldsymbol{\beta}, \boldsymbol{\Sigma_n}, v_n, s_n^{-2}$ are;

$$\boldsymbol{\beta}_n^{EB} = (\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{-\mathbf{1}} + \boldsymbol{X'X})^{-1}(\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{-\mathbf{1}}\hat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{X'X}\hat{\boldsymbol{\beta}}) \tag{13}$$

$$\boldsymbol{\Sigma}_n^{EB} = (\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{-\mathbf{1}} + \boldsymbol{X'X})^{-1} \tag{14}$$

$$v_n^{EB} = \hat{v}_0 + n \tag{15}$$

$$(s_n^{-2})^{EB} = \frac{\hat{v}_0 + n}{\hat{v}_0 \hat{s}_0^2 + (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})'[\hat{\boldsymbol{\Sigma}}_{\mathbf{0}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathbf{0}})} \tag{16}$$

The bootstrap Bayesian version of the estimates of $\boldsymbol{\beta}, \boldsymbol{\Sigma_n}$ involves the following steps;

1. Generation of bootstrap samples from the original data B desired number of times,
2. Estimating the hyperparameters (prior parameters) each time the samples are generated using Maximum Likelihood (ML) method,
3. Updating the posterior estimates using the hyperparameters in step (2) above using (13 & 14) and
4. Then obtaining the bootstrap empirical Bayesian estimates $\hat{\boldsymbol{\beta}}^{BT}$ and $\hat{\boldsymbol{\Sigma}}^{BT}$ using;

$$\hat{\boldsymbol{\beta}}^{BT} = B^{-1} \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_b^{EB} \tag{17}$$

$$\hat{\boldsymbol{\Sigma}}^{BT} = B^{-1} \sum_{b=1}^{B} \hat{\boldsymbol{\Sigma}}_b^{EB} \tag{18}$$

The $\hat{\boldsymbol{\beta}}^{BT}$ proposed here has good statistical properties in terms of biasness and Mean Square Error (MSE).

The Bias property can be evaluated as:

$$\begin{aligned}
Bias &= E[\hat{\boldsymbol{\beta}}^{BT}] - \boldsymbol{\beta} \\
&= E\left[B^{-1} \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_b^{EB}\right] - \boldsymbol{\beta} \\
&= E\left\{B^{-1} \sum_{b=1}^{B} \left[(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{b}}^{-\mathbf{1}} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{b}}^{-\mathbf{1}}\hat{\boldsymbol{\beta}}_{\boldsymbol{b}} + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}})\right]\right\} - \boldsymbol{\beta} \\
&= (\boldsymbol{\Sigma}^{-\mathbf{1}} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{\beta} \\
&= (\boldsymbol{\Sigma}^{-\mathbf{1}} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}^{-\mathbf{1}} + \boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} - \boldsymbol{\beta} \\
Bias &= \mathbf{0}
\end{aligned}$$

Also, the MSE is the combination of square of bias and variance of the estimate, then following from the above derivation the MSE is just the variance of the estimate. Thus;

$$MSE[\hat{\boldsymbol{\beta}}^{BT}] = B^{-2} var\left\{\sum_{b=1}^{B} \left[(\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{-\mathbf{1}} + \boldsymbol{X}'\boldsymbol{X})^{-1}(\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{-\mathbf{1}}\hat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}})\right]\right\}$$

The bootstrap prior Bayesian ANOVA follows from (17) with:

$$F_{BT} = \frac{n-k}{k-1}\left[\frac{(\hat{\boldsymbol{\beta}}^{\boldsymbol{BT}})'\boldsymbol{X}^{*'}\boldsymbol{Y} - n\bar{Y}^2}{\boldsymbol{Y}'\boldsymbol{Y} - (\hat{\boldsymbol{\beta}}^{\boldsymbol{BT}})'\boldsymbol{X}^{*'}\boldsymbol{Y}}\right]. \tag{19}$$

Again, the null hypothesis H_0 is rejected if $F_{BT} > F(k-1, n-k)$.

The multiclass variable selection procedure will be repeated for all p input variables in the training set. This implies that the test will be carried out p times which will lead to multiple

testing issue. To avert this, the False Discovery Rate (FDR) approach of Benjamini and Yekutieli (2001) is adopted as it has been adjudged to be more powerful than other method. The FDR procedure correct the multiple comparison issue by adjusting the p-values returned from the test functions. The function "p.adjust" in R with the option "fdr" was used to adjust the resulting p-values yielded by the method.

3. Simulation study

The R software was used to investigate the performance of classical ANOVA and bootstrap prior ANOVA in multiclass variable selection. The simulation procedure used was adapted from Olaniran *et al.* (2016) with little modifications. We simulated $n = 100$ observations representing the number of patients samples with $k = 3, 9$ distinct biological groups corresponding to different types of disease outcomes. For each observation, $p = 100, 1000, 5000$ covariates, $Y = (y_1, ...y_p)$, representing the observed gene expression profiles were simulated. For $k = 3$ and $Y = (y_1, ..., y_5)$, the dataset $Y|\tau = 1, 2, 3$ were simulated from multivariate normal distributions with means μ_1, μ_2, μ_3 and variance-covariance matrix Σ with values; $\mu_1 = 1, \mu_2 = 2, \mu_3 = 0, \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma = I$. The remaining dataset $Y|\tau = 1, 2, 3$ for $Y = (y_6, ..., y_p)$ were simulated from multivariate normal distributions with means $\mu_1 = \mu_2, = \mu_3 = 0$ and variance-covariance matrix $\Sigma = I$. Similarly, for $k = 9$ and $Y = (y_1, ..., y_5)$, the dataset $Y|\tau = 1, 2, 3, \ldots, 9$ were simulated from multivariate normal distributions with means $[\mu_1, \mu_2, \mu_3] = [\mu_4, \mu_5, \mu_6] = [\mu_4, \mu_5, \mu_6] = [1, 2, 0]$ and variance-covariance matrix $\Sigma = I$. The remaining dataset $Y|\tau = 1, 2, 3, \ldots, 9$ for $Y = (y_6, ..., y_{1000})$ were simulated from multivariate normal distributions with means $[\mu_1, \mu_2, \mu_3] = [\mu_4, \mu_5, \mu_6] = [\mu_7, \mu_8, \mu_9] = [0, 0, 0]$ and variance-covariance matrix $\Sigma = I$. In the all cases, the first 5 variables $y_1, ..., y_5$ are regarded as relevant variables while the remaining $p - 5$ are the irrelevant variables as they constitute same mean structure $\mathbf{0}$ irrespective of the class. The bootstrap size B and number of simulation iterations were fixed at 1000. The simulation results are presented in Tables 1 - 4. Table 1 presents the comparison between classical F_c and the proposed F_{BT} for a single gene y with $k = 3, 9$.

Table 1: Comparison results between F_c and F_{BT} for $p = 1$ and $k = 3, 9$

		H_0 **is true**		H_1 **is true**	
		F_c	F_{BT}	F_c	F_{BT}
$k = 3$	F	1.0239	1.0158	45.804	45.788
	$P(H\|Y, X)$	0.4994	0.5034	0.0000	0.0000
	Error	0.0527	0.0506	0.0000	0.0000
$k = 9$	F	0.8017	0.7918	12.5300	12.5098
	$P(H\|Y, X)$	0.6732	0.6772	0.0002	0.0002
	Error	0.0790	0.0776	0.0007	0.0008

Table 1 presents the simulation results for a single gene that corresponds to testing for $p = 1$. The underlying null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$ for $k = 3$ and $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_9$ for $k = 9$ against alternative hypothesis $H_1 : \mu_i \neq \mu_j$ for at least one pair of $i \neq j$. The table is partitioned into two conditions that correspond to the situation where the null hypothesis is true or false. The level of significance used for the testing is 0.05. The F values of F_{BT} is lower than F_c at various levels of k and conditions. The $P(H|Y, X)$ is often interpreted as the p-value for frequentist procedure F_c and the probability of the null hypothesis for Bayesian procedure F_{BT}. At 5% significance level, F_{BT} has a larger probability of H_0 being true when it is indeed true than F_c. These results subsequently corresponds to lower Type I errors for F_{BT} at $k = 3, 9$ when H_0 is true. For the second condition when H_1 is true and $k = 3$, the two procedures return similar probability of H_0 being true and thus similar Type II error

was achieved. However, when $k = 9$ and H_1 is true, approximately similar probability of null hypothesis and Type II error were obtained. Thus regarding validity, F_{BT} is more valid than F_c regarding relatively closer Type I error rate to the imposed 0.05 level especially when $k = 3$. Also, regarding the power of detecting a true difference when it exists, F_{BT} performance is relatively similar to F_c.

Table 2: Performance results of the simulated data at $p = 100$

		FDR Unadjusted		**FDR Adjusted**	
	Metrics	F_c	F_{BT}	F_c	F_{BT}
	TPR	1.0000	1.0000	1.0000	1.0000
$k = 3$	FPR	0.0505	0.0253	0.0053	0.0032
	TNR	0.9495	0.9747	0.9947	0.9968
	FNR	0.0000	0.0000	0.0000	0.0000
	TPR	1.0000	1.0000	1.0000	1.0000
	FPR	0.0516	0.0158	0.0042	0.0000
	TNR	0.9484	0.9842	0.9958	1.0000
$k = 9$	FNR	0.0000	0.0000	0.0000	0.0000

The performance metrics used to assess the methods are True Positive Rate (Sensitivity or Power) TPR which measures the expected proportion of active variables that are declared active, False Positive Rate (False Discovery) FPR which measures the expected proportion of inactive variables that are declared active, True Negative Rate (Specificity) TNR which measures the expected proportion of inactive variables that are declared inactive and False Negative Rate FNR which measures the expected proportion of active variables that are declared inactive.

The performance result for $p = 100$ corresponding to moderate high-dimensional modeling scenario is presented in Table 2. The two approaches maintained same sensitivity level at various k levels. This implies the two approach will always detect relevant variables. However,the most important criterion is the false positive or false discovery rate, a good selection procedure should have a reasonably lower false positive rate as well as high power. The false positive rate of F_{BT} is lower than F_c at various levels of k. Similar results were equally observed when $p = 5000$ in Table 4. However, when $p = 1000$ in Table 3, the FPR is approximately similar.

Table 3: Performance results of the simulated data at $p = 1000$

		FDR Unadjusted		**FDR Adjusted**	
	Metrics	F_c	F_{BT}	F_c	F_{BT}
	TPR	1.0000	1.0000	1.0000	1.0000
$k = 3$	FPR	0.0481	0.0275	0.0003	0.0003
	TNR	0.9519	0.9725	0.9997	0.9997
	FNR	0.0000	0.0000	0.0000	0.0000
	TPR	1.0000	1.0000	1.0000	1.0000
	FPR	0.0449	0.0143	0.0004	0.0001
	TNR	0.9551	0.9857	0.9996	0.9999
$k = 9$	FNR	0.0000	0.0000	0.0000	0.0000

4. Application to gene expression cancer RNA-Seq dataset

The dataset used here is a subset of RNA-Seq (HiSeq) PANCAN data set (Weinstein, Col-

lisson, Mills, Shaw, Ozenberger, Ellrott, Shmulevich, Sander, Stuart, Network *et al.* 2013). It represent a random extraction of 16384 gene expressions profiles of 801 patients with five different form of tumors labels BRCA, KIRC, COAD, LUAD and PRAD. The two methods were used to identify the most relevant biomarker genes for possible classification of tumors.

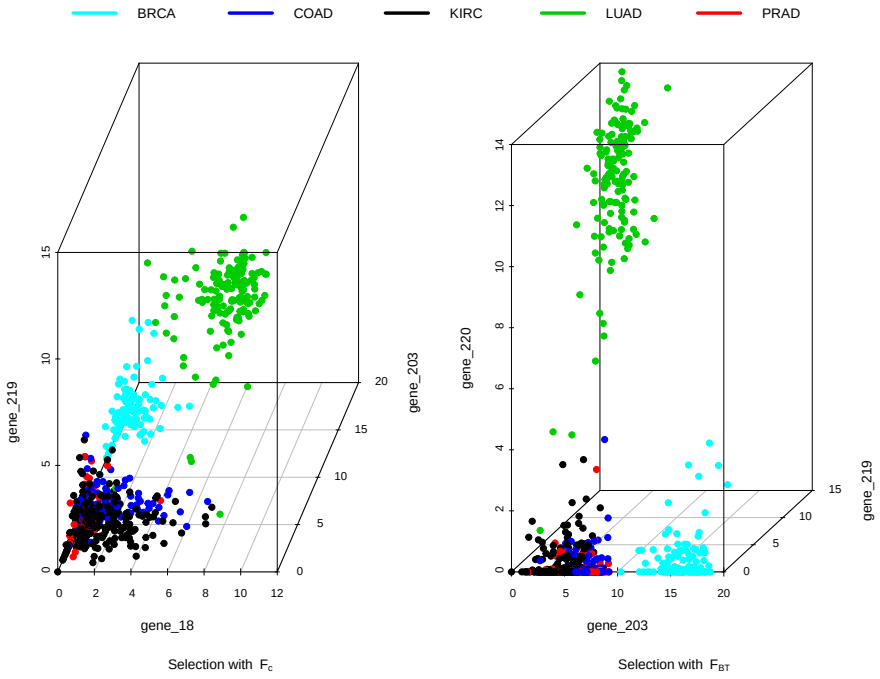

Figure 1: 3D classification plot for selection with F_c and F_{BT}

F_c identified 15798 genes while F_{BT} identified 4511 genes as relevant both at 5% threshold. The large number of gene subset identified by F_c can be attributed to high false positive rate as observed in the simulation studies. In addition, the p-values yielded by the two methods were used to rank the genes in increasing order of relevance. The three best subset genes were later used to plot the graph in Figure 1.

The plot showed that the two methods could only classify the tumors into three clear groups using the best three genes. But the classification from F_{BT} is more distinct compared to that of F_c. Also, two of the best three genes overlapped for the two methods. Tumor LUAD and BRCA were more evidently a product of high level of expressions for $gene_{220}$ and $gene_{219}$ using F_{BT} and F_c selections. Therefore, further clinical examination can be followed up on the identified genes for tumors with labels LUAD and BRCA.

5. Conclusion

In this paper, we considered Bayesian variable selection for multiclass classification task using the bootstrap prior technique. The bias derivation showed that the approach used is unbiased as well as maintaining lower mean square error property of Bayesian techniques. Simulation studies revealed that the proposed method F_{BT} has lower false positive rate in addition to the high power property in ANOVA based methods. Additionally, the proposed method has higher accuracy when selecting biomarker subsets useful for disease classification as observed in the PANCAN dataset. Hybridizing the proposed F_{BT} method with a classification technique is a good area of research that can be considered in future. Also, we have considered one way ANOVA because of its applicability to multiclass variable selection. The bootstrap prior technique can also be extended to multi-factor ANOVA.

Acknowledgement

We will like to appreciate Universiti Tun Hussein Onn (UTHM), Malaysia for supporting this research with grant $[Vot, U607]$.

Table 4: Performance results of the simulated data at $p = 5000$

		FDR Unadjusted		**FDR Adjusted**	
	Metrics	F_c	F_{BT}	F_c	F_{BT}
$k = 3$	TPR	1.0000	1.0000	1.0000	1.0000
	FPR	0.0503	0.0268	0.0001	0.0000
	TNR	0.9497	0.9732	0.9999	1.0000
	FNR	0.0000	0.0000	0.0000	0.0000
$k = 9$	TPR	1.0000	1.0000	1.0000	1.0000
	FPR	0.0483	0.0154	0.0001	0.0000
	TNR	0.9517	0.9846	0.9999	1.0000
	FNR	0.0000	0.0000	0.0000	0.0000

References

Benjamini Y, Yekutieli D (2001). "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of statistics*, pp. 1165–1188.

Bertolino F, Piccinato L, Racugno W (1990). "A Marginal Likelihood Approach to Analysis of Variance." *The Statistician*, pp. 415–424.

Forman G (2003). "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of machine learning research*, **3**(Mar), 1289–1305.

Guyon I, Elisseeff A (2003). "An Introduction to Variable and Feature Selection." *Journal of machine learning research*, **3**(Mar), 1157–1182.

Guyon I, Gunn S, Nikravesh M, Zadeh LA (2008). *Feature extraction: foundations and applications*, volume 207. Springer.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). "Gene Selection for Cancer Classification Using Support Vector Machines." *Machine learning*, **46**(1-3), 389–422.

Hwang K, Lee K, Park S (2017). "Variable Selection Methods for Multi-class Classification Using Signomial Function." *Journal of the Operational Research Society*, **68**(9), 1117–1130.

Liu Z, Bensmail H, Tan M (2012). "Efficient Feature Selection and Multiclass Classification with Integrated Instance and Model Based Learning." *Evolutionary Bioinformatics*, **8**, EBO–S9407.

Olaniran OR, Abdullah MAA (2018). "Bayesian Analysis of Extended Cox Model with Time-varying Covariates Using Bootstrap Prior." *Journal of Modern Applied Statistical Methods*, **in-press**.

Olaniran OR, Abdullah MAA, Pillay KG, Olaniran SF (2018). "Empirical Bayesian Binary Classification Forests Using Bootstrap Prior." *International Journal of Engineering & Technology*, **7**(4.30), 170 – 175.

Olaniran OR, Olaniran SF, Yahya WB, Banjoko AW, Garba MK, Amusa LB, Gatta NF (2016). "Improved Bayesian Feature Selection and Classification Methods using Bootstrap Prior Techniques." *Annals. Computer Science Series*, **14**(2).

Olaniran OR, Yahya WB (2017). "Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique." *Journal of Modern Applied Statistical Methods*, **16**(2), 34.

Peng Y, Wu Z, Jiang J (2010). "A Novel Feature Selection Approach for Biomedical Data Classification." *Journal of Biomedical Informatics*, **43**(1), 15–23.

Qureshi MNI, Oh J, Min B, Jo HJ, Lee B (2017). "Multi-modal, Multi-measure, and Multi-class Discrimination of ADHD with Hierarchical Feature Extraction and Extreme Learning Machine Using Structural and Functional Brain MRI." *Frontiers in human neuroscience*, **11**, 157.

Solari F, Liseo B, Sun D (2008). "Some Remarks on Bayesian Inference for One-way ANOVA Models." *Annals of the Institute of Statistical Mathematics*, **60**(3), 483–498.

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR, *et al.* (2013). "The Cancer Genome Atlas Pan-cancer Analysis Project." *Nature genetics*, **45**(10), 1113.

Wright GW, Simon RM (2003). "A Random Variance Model for Detection of Differential Gene Expression in Small Microarray Experiments." *Bioinformatics*, **19**(18), 2448–2455.

Yahya W, Olaniran O, Ige S (2014). "On Bayesian Conjugate Normal Linear Regression and Ordinary Least Square Regression Methods: A Monte Carlo Study." *Ilorin Journal of Science*, **1**(1), 216–227.

Affiliation:

Oyebayo Ridwan Olaniran
Department of Mathematics and Statistics
Faculty of Applied Science and Technology
Universiti Tun Hussein Onn Malaysia
Pagoh, Educational Hub, 84600 Pagoh, Johor, Malaysia
E-mail: rid4stat@yahoo.com

Mohd Asrul Affendi Bin Abdullah
Department of Mathematics and Statistics
Faculty of Applied Science and Technology
Universiti Tun Hussein Onn Malaysia
Pagoh, Educational Hub, 84600 Pagoh, Johor, Malaysia
E-mail: afendi@uthm.edu.my