5th International Conference on Mathematical Applications in Engineering IOP Publishing
IOP Conf. Series: Journal of Physics: Conf. Series **1489** (2020) 012030 doi:10.1088/1742-6596/1489/1/012030

Subset Selection in High-Dimensional Genomic Data using Hybrid Variational Bayes and Bootstrap priors

O R Olaniran[1], **M A A Abdullah**[2]

[1]Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, PMB 1515, Nigeria.

[2]Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Malaysia.

Email: Olaniran.or@unilroin.edu.ng

Abstract. In this study, the Variational Bayes (VB) approach was hybridized with the bootstrap prior procedure to improve the accuracy of subset selection as well as optimizing the algorithm time in modelling high-dimensional genomic data with inherent sparse structure. The new hybrid VB approach is shown to yields a minimal sufficient statistic which under mild regularity conditions converges to the true sparse structure. Simulation and real-life high-dimensional genomic data experiments revealed comparable empirical performance with other competing frequentist and Bayesian methods. In addition, a new fast algorithm that illustrates the procedure was developed and implemented in the environment of R statistical software as package "VBbootprior".

Key Words: Bayesian Inference, Big data, Variational Bayes, Bootstrap prior, R package

1. Introduction

Variable selection is one of the fundamental aspects of statistical modelling as evidenced by numerous authors. The classes of model selection procedures include criteria based approaches [1 - 3], penalized regression approaches [4 - 6] and Bayesian modelling approaches [7 - 10]. There have been several improvements in the area yet there is still no standard view on how to effectively perform subset selection, especially in high-dimensional settings. The primary focus of this paper is to develop a robust strategy for selecting variables in high-dimensional genomic data with inherent classification problem.

The Bayesian model selection strategy is more advantageous due to its ability to incorporate many types of variation as well as the incorporation of prior information. The weak aspect of the procedure is its sensitivity to the knowledge and thus proper specification of prior is highly necessary. One of the common prior used in Bayesian model selection involving linearity of variables and parameters is Zellner's g-prior [11].

In low dimensional setting, Bayesian model selection is performed by exploring all models involved vis-à-vis Bayesian inference. In contrast, for high-dimensional setting the Markov Chain Monte Carlo (MCMC) methods are employed. MCMC for moderate to large scale problems can be computationally inefficient [12-18]. For this reason, an enormous amount of effort has been put into developing MCMC and similar stochastic search based methods which can be used to explore the model space in a computationally efficient manner [19-23].

The Mean-field variational Bayes (VB) is a computationally efficient but approximate alternative to MCMC for Bayesian inference [12]. Although, it is difficult to fairly compare MCMC and VB as they differ theoretically. The VB approach is deterministic while MCMC is a stochastic approach, but in most cases, VB tends to be faster than MCMC, especially in high-dimensional settings [24-25].

2. Variational Bayes Bootstrap Prior (VBprior) for Subset Selection in Binary Classification

Suppose we have a binary class problem defined as $\{y \in [0,1] | \boldsymbol{z}\}$ governed by the logistic discriminant model:

$$p(y=1|\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}\beta)}{1+\exp(\boldsymbol{z}\beta)}$$

where y is the response variable, **z** is the matrix of predictors and $\boldsymbol{\beta}$ is the weight vector.

The classifier often return estimated probability $\hat{p}(y=1|\boldsymbol{z})$ which is then converted to $[0,1]$ using the scheme below:

$$\psi = \begin{cases} 1, \ p(y=1|\boldsymbol{z}) \geq cutpoint \\ 0, \ p(y=1|\boldsymbol{z}) < cutpoint \end{cases}$$

The cut-point probability is often fixed at 0.5.

Let $x \in \mathbb{N}$ be the subset of z that is correctly related to y. Thus we define the probability distribution of x as Poisson with parameter λ, thus;

$$x \sim Poisson(\lambda)$$

with hierarchical prior; $\lambda \sim Gamma(a, \theta)$ and $\theta \sim Gamma(b, c)$.We want to estimate the posterior of $p(\lambda, \theta|x)$ given the distribution $q(\lambda, \theta)$. The first step is to obtain the factorization method;

So we choose;

$$p(\lambda, \theta|x) \approx q(\lambda, \theta) = q(\lambda)q(\theta)$$

From the general approach we can obtain;

$$q(\theta) \propto exp\{E_{q(\lambda)}[lnp(x|\lambda,\theta)] + lnp(\lambda) + lnp(\theta)\}$$

$$q(\theta) \propto exp\{E_{q(\lambda)}[lnp(x|\lambda,\theta)] + lnp(\theta)\}$$

Since the expectation is only over λ, hence it doesn't affect $lnp(\theta)$. Thus;

$$q(\theta) \propto exp\{E_{q(\lambda)}[lnp(x|\lambda,\theta)]\}p(\theta)$$

$$p(x|\lambda,\theta) = p(x|\lambda)p(\lambda|\theta) = \frac{\exp(-\lambda)\lambda^x}{x!} \times \frac{\theta^{\ a}}{\Gamma a}\ \lambda^{a-1}\exp(-\theta\lambda).$$

For iid samples;

$$p(x_1, \dots, x_N|\lambda, \theta) = \prod_{i=1}^{N} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \frac{\theta^{\ a}}{\Gamma a}\ \lambda^{a-1}\exp(-\theta\lambda)$$

$$lnp(x_1, \dots, x_N|\lambda, \theta) = \sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \frac{\theta^{\ a}}{\Gamma a}\ \lambda^{a-1}\exp(-\theta\lambda)\right)$$

$$q(\theta) \propto exp\left\{E_{q(\lambda)}\left[\sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \lambda^{a-1}\exp(-\theta\lambda)\right)\right]\right\}p(\theta)$$

$$p(\theta) = \frac{c^{\ b}}{\Gamma b}\ \theta^{b-1}\exp(-c\theta)$$

$$q(\theta) \propto exp\left\{E_{q(\lambda)}\left[\sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \lambda^{a-1}\exp(-\theta\lambda)\right)\right]\right\}\theta^{b-1}\exp(-c\theta)$$

$$q(\theta) \propto exp\left\{E_{q(\lambda)}\left[\sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i+a-1}\exp(-\theta\lambda)}{x_i!}\right)\right]\right\}\theta^{b-1}\exp(-c\theta)$$

$$q(\theta) \propto \left[\prod_{i=1}^{N} \frac{E_{q(\lambda)}[\lambda^{x_i+a-1}]}{x_i!}\ (1+\theta)\exp\{E_{q(\lambda)}[-\lambda]\}\right]\theta^{b-1}\exp(-c\theta)$$

$$q(\theta) = Gamma(\theta|a', b')$$

$$a' = b + a + \sum_{i=1}^{N} x_i$$

$$b' = c + E_{q(\lambda)}[-\lambda]$$

Next to obtain optimal λ;

$$q(\lambda) \propto exp\{E_{q(\theta)}[lnp(x|\lambda, \theta)] + lnp(\lambda)\}$$

$$q(\lambda) \propto exp\{E_{q(\theta)}[lnp(x|\lambda, \theta)]\}p(\lambda)$$

$$p(x|\lambda, \theta) = p(x|\lambda)p(\lambda|\theta) = \frac{\exp(-\lambda)\lambda^{x}}{x!} \times \frac{\theta^{\ a}}{\Gamma a}\ \lambda^{a-1}\exp(-\theta\lambda)$$

$$q(\lambda) \propto exp\left\{E_{q(\theta)}\left[\sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \lambda^{a-1}\exp(-\theta\lambda)\right)\right]\right\}\lambda^{a-1}\exp(-\theta\lambda)$$

$$q(\lambda) \propto exp\left\{E_{q(\theta)}\left[\sum_{i=1}^{N} ln\left(\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \times \lambda^{a-1}\exp(-\theta\lambda)\right)\right]\right\}\lambda^{a-1}\exp(-\theta\lambda)$$

$$q(\lambda) \propto \left[\prod_{i=1}^{N} \frac{[\lambda^{x_i+a-1}]}{x_i!}\exp(-\lambda)\ exp\left[E_{q(\theta)}[\theta]\right]\right]\lambda^{a-1}\exp(-\theta\lambda)$$

$$q(\lambda) = Gamma(\lambda|c', d')$$

$$c' = b + a + \sum_{i=1}^{N} x_i$$

$$d' = E_{q(\theta)}(\theta)$$

The next step is to calculate variation inference objective which is the likelihood $L = L(a', b', c', d')$.

Algorithm 1: Variational Bayes Bootstrap Prior for Subset Selection in Binary Classification

```
Define the data as input for $\mathbf{q(\theta) = Gamma(\theta|a', b')}$ and $\mathbf{q(\lambda) = Gamma(\lambda|c', d')}$
Obtain $\mathbf{a', b', c', d'}$ as output.
    1. Initialize $\mathbf{a_0{}', b_0{}', c'_0, d_0{}'}$
    2. For iteration $\mathbf{t = 1, \dots, T}$
        - Update $\mathbf{q(\theta)}$ by setting;
              $\mathbf{a_t{}' = b + a + \sum_{i=1}^{N} x_i}$
              $\mathbf{b_t{}' = c'_{t-1} + E_{q(\lambda)}[-\lambda]}$
              $\mathbf{b_t{}' = c'_{t-1} - \frac{c'_{t-1}}{d'_{t-1}}}$
        - Update $\mathbf{q(\lambda)}$ by setting;
              $\mathbf{c_t{}' = b_{t-1}{}' + a_{t-1}{}' + \sum_{i=1}^{N} x_i}$
              $\mathbf{d_t{}' = E_{q(\theta)}[\theta]}$
              $\mathbf{d_t{}' = \frac{a_{t-1}{}'}{b_t - 1'}}$
    3. Evaluate $\mathbf{L(a', b', c', d')}$ to assess convergence.
```

3. Data Calibration

This section presents the application of VB on published real data. Each dataset represents a microarray study of thousands of gene expression profiles (p) measured on (n) individuals. The dataset was extracted from the Gene Expression Omnibus (GEO) database a subsection of National Center Biotechnology Information (NCBI). The data were partitioned using 5-folds cross-validation. One-fifth of each dataset was used to test while the VB approached was developed on the remaining part of the dataset. The datasets description is summarized in **Table 1**.

Table 1. Dataset description with the number of samples (*n*) and number of predictors (*p*).

Author of Data	Cancer type	n	p
[26]	Colon Cancer	62	2000
[27]	Lymphoma Cancer	58	6817
[28]	Breast Cancer	168	2905
[29]	Lung Cancer	181	12533

4. Results

This section presents the results of the train and test data for each of the datasets. **Table 2.** Presents the results of best subsets after applying the VB algorithm. The best subsets were then later used to perform classification of the disease as shown in the subsequent Tables.

Table 2. Number of Best Subsets of Predictors

Cancer type	p	x
Colon Cancer	2000	3
Lymphoma Cancer	6817	3
Breast Cancer	2905	3
Lung Cancer	12533	2

Table 3. Performance metrics results (%) for Colon and Lymphoma cancer

	Cancer type			
	Colon		Lymphoma	
Performance Metrics	Train	Test	Train	Test
Sensitivity	85.18	84.67	98.72	96.67
Specificity	94.95	94.29	98.67	88.33
Positive Predictive Value	90.40	91.67	99.57	96.90
Negative Predictive Value	92.20	89.64	96.08	90.00
Accuracy	91.55	90.26	98.70	94.92
Balance Accuracy	90.06	89.48	98.69	92.50
Prevalence	35.49	35.64	75.32	75.25
Detection Rate	30.27	29.10	74.35	72.67
Area Under the ROC Curve (AUC)	90.06	89.48	98.69	92.50

Table 2 showed that out of many predictors 2 or 3 of them can guarantee accuracy in the prediction of test data of at least 85%.

Table 4. Performance metrics results (%) for Breast and Lung cancer

	Cancer type			
	Breast		Lung	
Performance Metrics	Train	Test	Train	Test
Sensitivity	100.00	89.33	100.00	99.31
Specificity	100.00	86.67	100.00	91.43
Positive Predictive Value	100.00	95.00	100.00	98.04
Negative Predictive Value	100.00	94.17	100.00	97.50
Accuracy	100.00	91.56	100.00	97.78
Balance Accuracy	100.00	88.00	100.00	95.37
Prevalence	51.04	51.33	82.87	82.87
Detection Rate	51.04	47.33	82.87	82.31
Area Under the ROC Curve (AUC)	100.00	88.00	100.00	95.37

Table 5. Test Data Accuracy Comparison of VB with Random Forest (RF)

Cancer type	VB	RF
Colon Cancer	90.26	85.48
Lymphoma Cancer	94.92	89.46
Breast Cancer	91.56	61.83
Lung Cancer	97.98	99.44
Average	93.68	84.05

The results in **Table 3** and **Table 4** showed that overfitting is not a problem with VB procedure since the test data results are not better than the train data results except for prevalence and detection rate which are a function of the sample rather than the algorithm. One of the outstanding results can be observed with Lung cancer data where only 2 subsets of the 12533 gene expression profile taken resulted in almost 98% accuracy. **Table 5**. showed the side by side comparison with RF, the accuracies of VB were found to be larger than RF except for Lung cancer data.

5. Conclusion

In this paper, a new stage-wise feature selection named VB (Accuracy Based Feature Selection) was developed and its performance in the area of biomarker gene identification was achieved. The performance results revealed high accuracy in both test and train data. A comparative analysis of VB and Random Forest (RF) by [30] established the supremacy of the proposed methods for 75% of the data used.

Acknowledgements
The authors would like to thank the Tertiary Education Trust Fund (TetFund) and University of Ilorin, (UNILORIN) for financially supporting this research.

References

[1] Akaike H 1973 Information theory and an extension of the maximum likelihood principle *Proc. of the 2nd Int. Symp. on Information Theory* (Akademiai Kiad6, Budapest) 267 – 281.
[2] Mallows C L 1973 Some comments on Cp *Technometrics* **15** 661– 75.
[3] Schwarz G 1978 Estimating the dimension of a model *The Annals of Statist.* **6** 461–64.
[4] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. of the Royal Statist. Soc., Series B* **58** 267–88.
[5] Fan J and Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties *J. of the Amer. Statist. Ass.* **96** 1348–60.
[6] Fan J and Peng H 2004 Nonconcave penalized likelihood with a diverging number of parameters *The Annals of Statist.* **32** 928–61.
[7] Bottolo L and Richardson S 2010 Evolutionary stochastic search for Bayesian model exploration *Bayesian Analysis* **5** 583–618.
[8] Hans C Dobra A and West M 2007 Shotgun stochastic search for “large p” regression *J. of the Amer. Statist. Ass.* **102** 507–16.
[9] Li F and Zhang N R 2010 Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics *J. of the Amer. Statist. Ass.* **105** 1202–14.
[10] Stingo, F C and Vannucci M 2011 Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data *Bioinformatics* **27** 495–501.
[11] Zellner A 1986 On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* ed. P K Goel and A Zellner (North-Holland/Elsevier) 233–43.

[12] Ormerod J T, You C and Müller S 2017 A variational Bayes approach to variable selection. *Electron. J. Statist.* **11** 3549–94.
[13] Banjoko A W, Yahya W B, Garba M K, Olaniran O R, Olorede K O and Dauda K A 2015: Efficient Support Vector Machine Classification of Diffuse Large B-Cell Lymphoma and Follicular Lymphoma mRNA Tissue Samples. *Annals. Computer Science Series*. **13** 69 – 79.
[14] Jamil S A M, Abdullah M A A, Kek, S L, Olaniran O R and Amran S E 2017 Simulation of parametric model towards the fixed covariate of right-censored lung cancer data. *Journal of Physics: Conference Series* **890** (IOP Publishing) p. 012172.
[15] Olaniran O R and Abdullah M A A 2019a Bayesian Variable Selection for Multiclass Classification using Bootstrap Prior Technique. *Austrian J. of Statist.*, **48** 63-72.
[16] Olaniran O R and Abdullah M A A 2019b *BayesRandomForest*: An R implementation of Bayesian Random Forest for Regression Analysis of High-dimensional Data *Proceedings of the International Conference on Computing, Mathematics and Statistics (iCMS 2017)* ed. Ahmad A R, Kor L, Ahmad I and Idrus Z (Springer, Singapore).
[17] Olaniran O R and Abdullah M A A 2019c *BayesRandomForest*: Bayesian Random Forest for the Classification of High-dimensional mRNA Cancer Samples *Proceedings of the International Conference on Computing, Mathematics and Statistics (iCMS 2017)* ed. Ahmad A R, Kor L, Ahmad I and Idrus Z (Springer, Singapore).
[18] Olaniran O R and Abdullah M A A 2018a BayesRandomForest: An R implementation of Bayesian Random Forest for Regression Analysis of High-dimensional Data, *Romanian Statist. Rev.*, **66** 95-102.
[19] Olaniran O R and Abdullah M A A 2018b Bayesian Analysis of Extended Cox Model with Time-Varying Covariates using Bootstrap Prior. *J. of Modern App. Statist. Methods*. *In press.*
[20] Olaniran O R, Abdullah M A A, Pillay K G and Olaniran S F 2018 Empirical Bayesian Binary Classification Forests Using Bootstrap Prior. *Int. J. of Eng. & Tech.* **7** 170-5.
[21] Olaniran O R and Yahya W B 2017 Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. *J. of Modern App. Statist. Methods* **16** 618-38.
[22] Olaniran O R and Abdullah M A A 2017 Gene Selection for Colon Cancer Classification using Bayesian Model Averaging of Linear and Quadratic Discriminants, *Journal of Science and Technology* (Penerbit UTHM) **9** 140 – 4.
[23] Olaniran O R, Olaniran S F, Yahya W B, Banjoko A W, Garba M K, Amusa L B and Gatta N F 2016 Improved Bayesian Feature Selection and Classification Methods Using Bootstrap Prior Techniques. *Annals. Computer Science Series*, **14** 46 – 51.
[24] Yahya W B, Olaniran O R, Garba M K, Oloyede I, Banjoko A W, Dauda K A and Olorede K O 2016 A Test Procedure for Ordered Hypothesis of Population Proportions Against a Control. *Turkiye Klinikleri J. of Biostatistics* **8** 1 – 12.
[25] Yahya W B, Olaniran O R and Ige S O 2014 On Bayesian Conjugate Normal Linear Regression and Ordinary Least Square Regression Methods: A Monte Carlo Study. *Ilorin J. Sci* **1** 216-27.
[26] Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D and Levine A J 1999 Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*. **96** 6745–50.
[27] Shipp M A, Ross K N, Tamayo P, Weng A P, Kutok J L, Aguiar R C, Gaasenbeek M, Angelo M, Reich M, Pinkus G S et al. 2002 Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine* **8** 68 – 74.
[28] Gravier E, Pierron G, Vincent-Salomon A, Gruel N, Raynal V, Savignoni A, De Rycke Y et al. 2010. A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer*. **49** 1125–34.
[29] Gordon G J, Jensen R V, Hsiao L-L, Gullans S R, Blumenstock J E, Ramaswamy S, Richards W G, Sugarbaker D J and Bueno R 2002. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer research. **62**, 4963–67.
[30] Breiman L 2001 Random forests. *Machine Learning*, **45** 5–32.