Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n2p561

Classification of gene expression autism data based on adaptive penalized logistic regression
By Algamal

Published: 14 October 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 02, October 2017, 561-571
DOI: 10.1285/i20705948v10n2p561

Classification of gene expression autism data based on adaptive penalized logistic regression

Zakariya Yahya Algamal[*]

Department of Statistics and Informatics,University of Mosul

Published: 14 October 2017

The common issues of high-dimensional gene expression data are that many of genes may not be relevant to their diseases. Gene selection has been proved to be an effective way to improve the result of many classification methods. In this paper, an adaptive penalized logistic regression is proposed, with the aim of identification relevant genes and provides high classification accuracy of autism data, by combining the logistic regression with the weighted L1-norm. Experimental results show that the proposed method significantly outperforms two competitor methods in terms of classification accuracy, G-mean, and area under the curve. Thus, the proposed method can be useful for other cancer classification using DNA gene expression data in the real clinical practice.

keywords: penalized logistic regression; lasso; SCAD; autism data; gene selection.

1 Introduction

The autism spectrum disorder (ASD) term is often referred to neurodevelopmental syndrome which is characterized by impairments in socialization and communication, and behaviors and interests. It usually appears before 3 years of age. In addition, the prevalence of autism in boys is nearly four times greater than in girls. ASD is strongly heritable and it is among the most highly heritable common neuropsychiatric disorders

[*] Corresponding author: zakariya.algamal@uomosul.edu.iq

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

(Latkowski and Osowski, 2015a,b). Studying the gene expression microarray is an important research direction in ASD, which can help in identifying the most related genes to ASD, and, therefore, can help in diagnosis.

One of the major advancement made in the field of biology and genetics research is the emergence of DNA microarray technology. This technology facilitates the determination of the expression values of thousands of genes simultaneously (Zheng and Liu, 2011). The gene expression data is used for various analyses to understand the biological significance of the tissue from which the genes were extracted for the experiment (Apolloni et al., 2016; Algamal and Lee, 2015a). These gene expression datasets are applied to numerous areas of application, such as cancer classification and tumor detection. In cancer classification, the taxonomy of normal and abnormal patterns of the cells is one of the most important and significant processes during the cancer diagnosis and drug discovery (Algamal and Lee, 2015c,b; Algamal, 2012). It can help to improve the health care of patients, and, therefore, the high prediction of cancer has great value in the treatment or the therapy (Algamal and Lee, 2015b). Gene expression dataset often contains a large number of genes, d , with only a few samples, n, making the gene expression dataset matrix has rows less than columns, $d > n$ (Algamal and Lee, 2015c). Over the last two decades, gene selection has received increasing attention, motivated by the desire to understand structure in the high-dimensional gene expression datasets. With these types of datasets, typically many genes are irrelevant and redundant which could potentially vitiate the classification performance. Accordingly, it is preferred to reduce the dimensionality of these datasets. Reduction of the dimensions is often achieved by gene selection, which is maintaining a direct relationship between a gene and a classification performance (Algamal, 2012).

According to the mechanism of selection, gene selection methods, in general, can be classified into three categories: filter methods, wrapper methods, and embedded methods. Filter methods are one of the most popular gene selection methods, which are based on a specific criterion by gaining information of the each gene. These methods work separately and they are not dependent on the classification method. For the wrapper methods, on the other hand, the gene selection process is based on the performance of a classification algorithm to optimize the classification performance. In embedded methods, gene selection process is incorporated into the classification methods, which can perform gene selection and classification simultaneously. These methods provide higher computational efficiency comparing with the wrapper methods (Algamal and Lee, 2017; Algamal et al., 2017b; Algamal and Ali, 2017b,a; Kahya et al., 2017a; Algamal, 2008; Kahya et al., 2017b; Al-Fakih et al., 2015; Algamal et al., 2017a, 2016b, 2015, 2016a; Algamal, 2011; Algamal and Allyas, 2017; Al-Fakih et al., 2016; Algamal, 2017).

Logistic regression is a widely-used classification method in different classification areas, especially in gene expression data classification. As the number of the genes increases, the training time of applying logistic regression increases and also its computational complexity increases (Algamal and Lee, 2015c,d; Inan and Erdogan, 2013; Asar and Gen, 2016; Asar, 2017). Unfortunately, logistic regression cannot automatically handle gene selection although it has been proven advantageous in handling gene expression data classification (Liang et al., 2013). Penalized methods are very effective embedded gene selection methods, which connected with many popular classification methods. In recent years, logistic regression as among all the classification methods, those based on sparseness, received much attention. It combines the logistic regression with a penalty to perform gene selection and classification simultaneously. With deferent penalties, several logistic regression models can be applied, among which are, L1-norm, which is called the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and adaptive L1-norm (Zou, 2006). Unquestionably, L1-norm is considered to be one of the most popular procedures in the class of sparse methods. Nonetheless, L1-norm applies the same amount of the sparseness to all genes, resulting in inconsistent gene selection (Fan and Li, 2001; Zou, 2006). To increase the power of informative gene selection, in the present study, an adaptive logistic regression is proposed. More specifically, a new weight inside L1-norm is proposed, which can correctly discriminate the healthy children from children with autism. This weight will reflect the importance amount of each gene. Experimentally, comparisons between our proposed gene selection method and other two competitor methods are performed. The experimental results prove that the proposed method is very effective for selecting the relevant genes with high classification accuracy.

2 Penalized Logistic Regression

Logistic regression is a statistical method, which can be used to model a binary classification problem. The regression function has a nonlinear relation with the linear combination of the genes. In cancer classification, the response variable of the logistic regression has two values either 1 for the tumor class or 0 for the normal class. Let $y_i \in \{0, 1\}$, then the logistic regression model is defined as

$$\ln[\frac{p_i}{1-p_i}] = \mathbf{x}_i^T \beta, \;\; i = 1, 2, ..., n, \tag{1}$$

where $\mathbf{x}_i^T$ is a $1 \times p$ vector of genes and $\beta = (\beta_1, ..., \beta_k)^T \in R^k$ is a vector of unknown gene coefficients, and

$$E(y_i = 1|\mathbf{x}_i) = \frac{Exp(\mathbf{x}_i^T \beta)}{1 + Exp(\mathbf{x}_i^T \beta)}. \tag{2}$$

The log-likelihood function can be written as:

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \ln{(p_i)} + (1 - y_i) \ln(1 - p_i)\}. \tag{3}$$

Logistic regression offers the advantage of simultaneously estimating the probabilities p_i and $1-p_i$ for each class and classifying subjects. The predicted class is then obtained by $I\{\hat{p}_i > 0.5\}$, where I is an indicator function. PLR adds a nonnegative penalty term to Eq. (3), such that the size of gene coefficients in high-dimension can be controlled.

Without loss of generality, it is assumed that the genes are standardized, then the estimation of the vector β is obtained by maximizing the penalized logistic regression as

$$\hat{\beta}_{PLR} = \arg\max_{\beta} \left[\sum_{i=1}^{n} \{ y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \} - \lambda g(\beta) \right] \tag{4}$$

where $\lambda\, g(\beta)$ is the penalty term that sparse the estimates. The penalty term depends on the positive tuning parameter, λ, which controls the tradeoff between fitting the data to the model and the effect of the penalty. In other words, it controls the amount of shrinkage. For the $\lambda = 0$, we obtain the maximum likelihood method (MLE) solution. Conversely, for large values of λ the influence of the penalty term on the coefficient estimates increases. Eq. (4) can be efficiently solved by the coordinate descent algorithm (Friedman et al., 2010).

Despite the advantage of the Lasso, it has three shortcomings (Wang et al., 2011). First, it cannot select more genes than the number of samples. Second, in microarray gene data, there is grouping among genes, where genes that share a common biological pathway have a high pairwise correlation with each other. The Lasso tries to select only one gene or a few of them among a group of correlated genes. To overcome the first two limitations, Zou and Hastie (2005) proposed the elastic net penalty, for which the penalty is a linear combination of L1-norm and L2-norm. Last, the Lasso has a bias gene selection, because it penalizes all gene coefficients equally. In other words, the Lasso does not have the oracle properties, which refer to the probability of selecting the right set of genes (with nonzero coefficients) converged to one, and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients are known in advance (Fan and Li, 2001). In relation to the last limitation of the Lasso, oracle properties, Zou (2006) proposed the adaptive Lasso in which adaptive weights are used for penalizing different coefficients in the L1-norm penalty.

3 The proposed method

In the context of gene expression classification problems, the goal of gene selection is to improve classification performance, to provide faster and more cost-effective genes, and to achieve a better knowledge of the underlying classification problem. High dimensionality can negatively influence the classification performance of a classifier by increasing the risk of overfitting and lengthening the computational time. In addition, it makes various classification methods not applicable for analyzing microarray gene expression data directly. Therefore, removing irrelevant and noisy genes from the original microarray gene expression data is essential for applying classification methods to analyze the microarray gene expression data.

It is worthwhile to highlight that our contribution of this paper comes from the following issues. First, although PLR with L1-norm can be applied directly to the high dimensional gene expression data, this method may select irrelevant genes because L1-norm has the inconsistent property in gene selection. In other words, the estimates of the PLR with L1-norm can be biased for large coefficients because larger coefficients will take larger penalties. Second, in PLR, the genes are usually standardized. However, the standardization process may be unreasonable when the variances of genes showing important effect.

Motivated by these issues, a consistent identification of the true underlying genes is essential to improve the classification accuracy. As a result, the standard deviation for each gene is proposed as a weight inside L1-norm, where

$$w_j = \frac{1}{\hat{sd_j}}, \quad j = 1, 2, ..., d, \tag{5}$$

where $\hat{sd_j}$ is the standard deviation for each gene. According to Eq. (5), the gene with low value of standard deviation will receive relatively large amount of weight, while the gene with high value of standard deviation will receive small amount of weight. By this weighting procedure, the L1-norm can reduce the inconsistent property in gene selection.

After assigning each gene with its related weight, the PLR with weighted L1-norm is utilized to select the informative genes with high classification accuracy. The detailed of the adaptive PLR (APLR) computation is described in Algorithm 1. The APLR equation has a convex form, which ensures the existence of global maximum point and can be efficiently solved. As a result, the coordinate descent method can be used to solve APLR.

```
Algorithm 1: The computation of APLR
Step 1: Find $[w_j, \quad j = 1, 2, ..., d.$
Step 2: Define $\tilde{\mathbf{x}}_i = w_j \mathbf{x}_i$
Step 3: Solve the APLR,
```

$$\hat{\beta}_{\text{APLR}} = \arg\min_{\beta} \left\{ -\sum_{i=1}^{n} \{y_i \ln(p_i) + (1 - y_i) \ln(1 - pi_i)\} + \lambda \sum_{j=1}^{p} w_{\text{j}} \, |\beta_{\text{j}}| \right\}$$

4 Experimental Study

4.1 Dataset description

The autism gene expression dataset was first presented by Latkowski and Osowski (2015a). This dataset is publicly available and it was retrieved from NCBI repository database on September 5, 2015 ("NCBI database," 2015). It consists of the gene expression values of 146 male children (observations) from peripheral blood lymphocytes (PBL). The total RNA was extracted for microarray experiments with Affymetrix Human U133 Plus 2.0 39 expression arrays. This dataset comprises 54,613 of genes, 82 of children with autism, and 64 of healthy children. Furthermore, this dataset has been recently analyzed by Latkowski and Osowski (2015a) and Latkowski and Osowski (2015b).

4.2 Performance evaluation

In order to evaluate the predictive performance of the proposed method, three performance metrics are implemented, specifically: (1) classification accuracy (CA), (2) geometric mean of sensitivity and specificity (G-mean), and (3) area under the curve (AUC). The CA stands for the proportion of correctly classified children with autism and healthy children, which measures the classification power of the classifier. The CA can define as:

$$\mathrm{CA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \times 100\%, \tag{6}$$

where TP is the number of true positive, FP is the number of false positive, TN is the number of true negative, and FN is the number of false negative. A typical classification method should maximize the accuracy on the both of children with autism and healthy children. As a consequence, the G-mean has been proposed as a metric to highlight the joint performance of sensitivity and specificity. It is defined as:

$$\mathrm{G-mean}{=}\sqrt{\mathrm{Sensitivity} \times \mathrm{Specificity}}, \tag{7}$$

where sensitivity is the fraction of children with autism that were successfully classified, and specificity is the fraction of healthy children that were properly classified. The AUC was used to quantitatively evaluate the overall classification performance of the proposed method. Its value can vary from 0 to 1, the closer value to 1, the better overall classification performance.

4.3 Experimental setting

To demonstrate the usefulness of the proposed method, comparative experiments with the PLR-lasso and the PLR-SCAD are conducted. To do so, the autism gene expression dataset is randomly partitioned into the training dataset and the test dataset, where 70% of the children are selected for training dataset and the rest 30% are selected for testing dataset. For a fair comparison and for alleviating the effect of the data partition, all the used classification methods are evaluated, for their classification performance metrics using 10 folds cross validation, averaged over 100 partitioned times.Depending on the training dataset, the tuning parameter value, λ, for each used classification method was fixed as $0 \leq \lambda \leq 100$. For the SCAD penalty, the constant a was set to equal 3.7 as it suggested by Fan and Li (2001).

5 Experimental results

5.1 Classification performance

Table 1 summarizes, on average, the number of selected genes (# genes), the classification accuracy, and the G-mean for the training dataset of applying the APLR, PLR-SCAD, and the PLR-lasso. In addition, it summarizes the classification accuracy for the testing dataset. The number in parenthesis is the corresponding standard deviation.

Regarding classification accuracy and based on the training dataset, the proposed method, APLR, achieved 93.27%, defeating PLR-SCAD and the PLR-lasso by 4.14% and 9.41%, respectively. The G-mean of the APLR yields 0.927, which indicates that the APLR has a separation capability between healthy children and children with autism. In addition, PLR-SCAD secondly comes with 89.13% and better than PLR-lasso. This is not surprising because the PLR-SCAD has the effectiveness of consistent selection. Depending on the testing dataset, the APLR is better than the others in terms of classification accuracy because it achieved 91.28%, which is 4.17% and 10.57% better than PLR-SCAD and the PLR-lasso, respectively.

In terms of the number of selected genes, the results in Table 1 show that the APLR selected significantly less genes than the other two methods, where it selects 9 genes while PLR-SCAD and the PLR-lasso, respectively, selects 12 and 17 genes. Overall, the classification performance of our proposed adaptive penalized support vector machine method provides best overall classification performance compared to PLR-SCAD and PLR-lasso. This is an implication that our proposed method can take consideration of the information of each gene by its weight.

Table 1: Classification performance of the APLR, PLR-SCAD, and PLR-lasso over 100 times.

Methods	training dataset		testing dataset	
	# genes	CA	G-mean	CA
APLR	9	93.27 (0.070)	0.927 (0.050)	91.28 (0.006)
PLR-SCAD	12	89.13 (0.010)	0.887 (0.007)	87.11 (0.007)
PLR-lasso	17	83.86 (0.010)	0.825 (0.006)	80.71 (0.009)

5.2 Statistical significance test

For further ability confirmation of the proposed method in selecting the most relevant genes with high classification performance, a pairwise comparison between the proposed method and each competitor method was utilized using paired two-tailed t-test. This test was performed depending on the area under the curve of the training dataset. Figure 1 shows the boxplot of the AUC for each used method. It is clearly seen that the AUC of the proposed method is comparable to the results obtained from PLR-SCAD and the PLR-lasso.

Table 2 reports the paired two-tailed t-test results at significance level $\alpha = 0.05$. As shown in Table 2, the AUC of the proposed method is statistically significant better than those of PLR-SCAD and PLR-lasso.

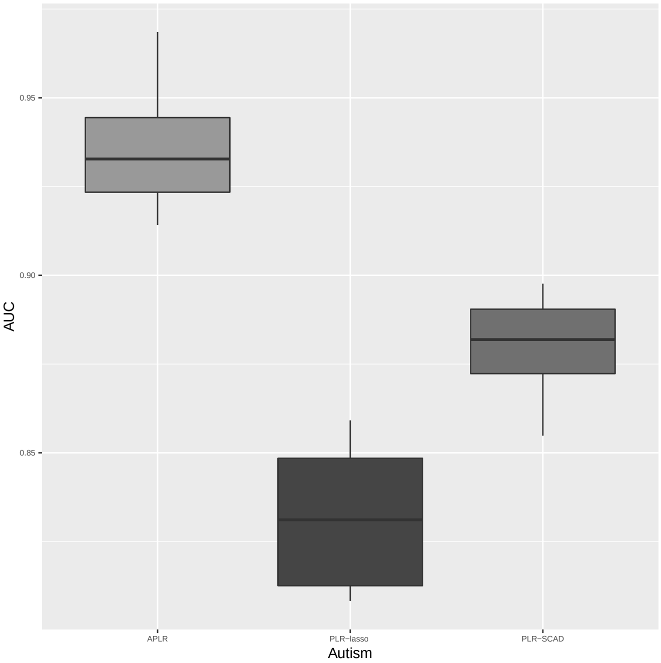

Figure 1: Boxplot of the AUC for the autism dataset achieved by the three used methods.

Table 2: P-values for the paired t-test of our proposed method results with two competitor methods. (*) means that the two methods have significant differences

Dataset	APLR vs PLR-SCAD	APLR vs PLR-lasso
Autism	0.0012(*)	0.0002(*)

6 Conclusion

This paper presents an adaptive penalized support vector machine by combining the logistic regression with the weighted L1-norm to identify the relevant genes in gene expression autism data. Our proposed method was experimentally tested and compared with other existing methods. The superior classification performance of the proposed method was shown through four aspects: high classification accuracy, G-mean and AUC. Meeting these four aspects simultaneously nominates the proposed method as a promising gene selection method. Overall, the proposed method clearly illustrates its applicability and usefulness in other types of high-dimensional classification data related to the biological field.

References

Al-Fakih, A. M., Algamal, Z. Y., Lee, M. H., Abdallah, H. H., Maarof, H., and Aziz, M. (2016). Quantitative structure-activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression. *Journal of Chemometrics*, 30(7):361–368.

Al-Fakih, A. M., Aziz, M., Abdallah, H. H., Algamal, Z. Y., Lee, M. H., and Maarof, H. (2015). High dimensional qsar study of mild steel corrosion inhibition in acidic medium by furan derivatives. *International Journal of Electrochemical Science*, 10:3568–3583.

Algamal, Z. and Allyas, I. I. (2017). Prediction of blood lead level in maternal and fetal us-ing generalized linear model. *International Journal of Advanced Statistics and Probability*, 5(2):65.

Algamal, Z. Y. (2008). Exponentiated exponential distribution as a failure time distribution. *IRAQI Journal of Statistical science*, 14:63–75.

Algamal, Z. Y. (2011). Paired bootstrapping procedure in gamma regression model using r. *Journal of Basrah Researches*, 37(4):201–211.

Algamal, Z. Y. (2012). Diagnostic in poisson regression models. *Electronic Journal of Applied Statistical Analysis*, 5(2):178–186.

Algamal, Z. Y. (2017). Using maximum likelihood ratio test to discriminate between the inverse gaussian and gamma distributions. *International Journal of Statistical Distributions*, 1(1):27–32.

Algamal, Z. Y. and Ali, H. T. M. (2017a). Bootstrapping pseudo - r2 measures for binary response variable model. *Biomedical Statistics and Informatics*, 2(3):107–110.

Algamal, Z. Y. and Ali, H. T. M. (2017b). An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. *Electronic Journal of Applied Statistical Analysis*, 10(1):242–256.

Algamal, Z. Y. and Lee, M. H. (2015a). Applying penalized binary logistic regression with correlation based elastic net for variables selection. *Journal of Modern Applied Statistical Methods*, 14(1):168–179.

Algamal, Z. Y. and Lee, M. H. (2015b). High dimensional logistic regression model using adjusted elastic net penalty. *Pakistan Journal of Statistics and Operation Research*, 11(4):1–10.

Algamal, Z. Y. and Lee, M. H. (2015c). Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42(23):93269332.

Algamal, Z. Y. and Lee, M. H. (2015d). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Comput Biol Med*, 67:136–45.

Algamal, Z. Y. and Lee, M. H. (2017). A novel molecular descriptor selection method in qsar classification model based on weighted penalized logistic regression. *Journal of Chemometrics*, page e2915.

Algamal, Z. Y., Lee, M. H., and Al-Fakih, A. M. (2016a). High-dimensional quantitative structure-activity relationship modeling of influenza neuraminidase a/pr/8/34 (h1n1) inhibitors based on a two-stage adaptive penalized rank regression. *Journal of Chemometrics*, 30(2):50–57.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., and Aziz, M. (2015). High-dimensional qsar prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive lasso. *Journal of Chemometrics*, 29(10):547–556.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., and Aziz, M. (2016b). High-dimensional qsar modelling using penalized linear regression model with l1/2-norm. *SAR and QSAR in Environmental Research*, 27(9):703–19.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., and Aziz, M. (2017a). High-dimensional qsar classification model for anti-hepatitis c virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty. *Journal of Chemometrics*, 31:1–8.

Algamal, Z. Y., Qasim, M. K., and Ali, H. T. M. (2017b). A qsar classification model for neuraminidase inhibitors of influenza a viruses (h1n1) based on weighted penalized support vector machine. *SAR and QSAR in Environmental Research*, pages 1–12.

Apolloni, J., Leguizamn, G., and Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing*, 38:922–932.

Asar, Y. (2017). Some new methods to solve multicollinearity in logistic regression. *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/03610918.2015.1053925.

Asar, Y. and Gen, A. (2016). New shrinkage parameters for the liu-type logistic estimators. *Communications in Statistics - Simulation and Computation*, 45(3):1094–1103.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Inan, D. and Erdogan, B. E. (2013). Liu-type logistic estimator. *Communications in Statistics - Simulation and Computation*, 42(7):1578–1586.

Kahya, M. A., Al-Hayani, W., and Algamal, Z. Y. (2017a). Classification of breast cancer histopathology images based on adaptive sparse support vector machine. *Journal of Applied Mathematics & Bioinformatics*, 7(1):49–69.

Kahya, M. A., Al-Hayani, W., and Algamal, Z. Y. (2017b). Gene selection inside pathways using weighted l1-norm support vector machine. *American Journal of Computational and Applied Mathematics*, 7(4):87–94.

Latkowski, T. and Osowski, S. (2015a). Computerized system for recognition of autism on the basis of gene expression microarray data. *Computers in Biology and Medicine*, 56:82–91.

Latkowski, T. and Osowski, S. (2015b). Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*, 42(2):864–872.

Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., and Zhang, H. (2013). Sparse logistic regression with a l1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics*, 14(1):198–211.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, 5(1):468–485.

Zheng, S. and Liu, W. (2011). An experimental comparison of gene selection by lasso and dantzig selector for cancer classification. *Computers in Biology and Medicine*, 41(11):1033–1040.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.