International Journal of Advanced Statistics and Probability, 8 (1) (2020) 16-20

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP

Research paper

Gene selection in Cox regression model based on a new adaptive penalized method

Oday Isam Alskal, Zakariya Yahya Algamal *

Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
**Corresponding author E-mail: zakariya.algamal@uomosul.edu.iq*

Abstract

The common issues of high dimensional gene expression data for survival analysis 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 methods. The Cox proportional hazards regression model is the most popular model in regression analysis for censored survival data. In this paper, an adaptive penalized Cox proportional hazards regression model is proposed, with the aim of identification relevant genes and provides high classification accuracy, by combining the Cox proportional hazards regression model with the weighted least absolute shrinkage and selection operator (LASSO) method. Experimental results show that the proposed method significantly outperforms two competitor methods in terms of the area under the curve and the number of the selected genes.

Keywords: *Cox Regression Model; Penalized Method; LASSO; Gene Selection.*

1. Introduction

The problem of analyzing time to event data arises in a number of applied fields, such as medicine, biology, public health, and epidemiology (Cockeran et al., 2019; Emura et al., 2012). Nowadays, high dimensional gene expression data are increasingly used for modeling various clinical outcomes to facilitate disease diagnosis, disease prognosis, and prediction of treatment outcome (Jian Huang et al., 2014).

Regression modeling is a standard practice to study jointly the effects of multiple predictors on a response. The Cox proportional hazards model is ubiquitous in the analysis of time-to-event data. When the number of predictors is large, building a Cox proportional hazards model including all of them is undesirable because it has low prediction accuracy and is hard to interpret (Karabey & Tutkun, 2017; Leng & Helen Zhang, 2006). For these reasons, variable selection has become an important focus in Cox proportional hazards modeling.

Penalized methods are very effective variable selection methods. These methods combine the Cox proportional hazards model with a penalty to perform variable selection and estimation simultaneously. With deferent penalties, several Cox proportional hazards models can be applied, among which are, LASSO, which is called the least absolute shrinkage and selection operator (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), elastic net (Zou & Hastie, 2005), and adaptive LASSO (Zou, 2006). Unquestionably, LASSO is considered to be one of the most popular procedures in the class of penalized methods. However, LASSO has a limitation: It applies the same amount of the penalty to all variables. Thus it is an inconsistent variable selection method (Algamal & Lee, 2015a, 2015b; Zou, 2006).

To increase the power of informative gene selection, in the present study, an adaptive Cox proportional hazards model is proposed. More specifically, a new weight inside LASSO is proposed, which can correctly reduce the estimation error. This weight will reflect the importance amount of each gene. Experimentally, comparisons between our proposed gene selection method and other competitor methods are performed. The experimental results prove that the proposed method is very effective for selecting the relevant genes with high prediction accuracy.

2. Panelized Cox proportional hazards model

Survival analysis is the statistical branch studying time-to-event data, or more precisely the time elapsing from a well-defined initiating event to some particular endpoint. The Cox proportional hazards regression model is one of the most popular and useful models in survival analysis (Cox, 1972).

Consider an analysis with time-to-event outcome, we denote the observed triplet as $\{(t_i, \delta_i, x_i) : i = 1, ..., n\}$ where t_i is the survival time if $\delta_i = 1$ and censored time if $\delta_i = 0$ and $x_i = (x_{i1}, ...x_{ip})$ is a p-dimensional explanatory variables. Under the proportional hazards framework, the Cox proportional hazards model (CPHM) can be defined as

$$h\left(t_i \mid x_i\right) = h_0\left(t_i\right)\exp\left(\beta^T x_i\right), \tag{1}$$

Where $h_0\left(t_i\right)$ is the baseline hazard function and $\beta = (\beta_1, \dots, \beta_p)^T$ is a $p \times 1$ vector of unknown regression coefficients. Assuming that the subjects are statistically independent of each other, the joint probability of all realized events is the following partial likelihood

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left(\beta^T x_i\right)}{\sum_{j \in R_i} \exp\left(\beta^T x_j\right)}, \tag{2}$$

Where R_i is the set of subjects that are at risk just before time t_i .

The estimation of the regression parameters of Eq. (1) is commonly carried out by minimizing the partial log likelihood function (Eq. (2)) as

$$\hat{\beta}_{CPHM} = \arg\min_{\beta}\left(\log L(\beta)\right) = -\sum_{i=1}^{n}\left[\beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j)\right]. \tag{3}$$

Panelized Cox proportional hazards model (PCPHM) adds a nonnegative penalty term to Eq. (1), such that the size of variable coefficients can be controlled. Several penalty terms have been discussed in the literature considering the Cox proportional hazards model (Du et al., 2010; Fu et al., 2017; Gui & Li, 2005; Hossain & Ahmed, 2014; Hou et al., 2013; H. H. Huang & Liang, 2018; J. Huang et al., 2013; Jiang & Liang, 2018; Kauermann, 2005; Li et al., 2014; Lin & Halabi, 2017; Liu et al., 2014; Park & Ha, 2018; Shi et al., 2019; Suchting et al., 2019; Wang et al., 2019; Wu et al., 2012; Zhang & Lu, 2007). The LASSO method, proposed by Tibshirani (1996), is one of the popular penalty terms. The LASSO performs variable selection and estimation simultaneously by constraining the log-likelihood function of variable coefficients. Generally, the PCPHM is defined as

$$PCPHM = \sum_{i=1}^{n}\left[\beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j)\right] - \lambda P(\beta), \tag{4}$$

Where $\lambda P(\beta)$ is the penalty term that regularized the estimates. The penalty term depends on the positive tuning parameter, $\lambda > 0$, which controls the tradeoff between fitting the data to the model and the effect of the regularization. In other words, it controls the amount of shrinkage. For the $\lambda = 0$, we obtain the CPHM solution in Eq. (3). In contrast, for large values of λ, the influence of the penalty term on the coefficient estimates increases.

Without loss of generality, it is assumed that the explanatory variables are standardized, $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1})\sum_{i=1}^{n} x_{ij}^2 = 1, \quad \forall j \in \{1, 2, \dots, p\}$. The estimation of the vector β using LASSO is obtained by minimizing Eq. (4) as (Bradic et al., 2011; Goeman, 2010; Tibshirani, 1997)

$$\hat{\beta}_{Cox}^{LASSO} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\left[\beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j)\right] + \lambda \sum_{j=1}^{p} \left|\beta_j\right|\right]. \tag{5}$$

Equation (5) can be efficiently solved by the coordinate descent algorithm (Simon et al., 2011).

The LASSO has an advantage in that it is computationally feasible in high-dimensional data. On the other hand, the LASSO has a drawback. The LASSO lacks the oracle properties, as stated in Fan and Li (Fan & Li, 2001) because it is equally penalize all the coefficients. To alleviate this drawback, Zou (2006) and proposed the adaptive LASSO in which adaptive weights are used for penalized different coefficients in the penalty. The basic idea behind the adaptive LASSO is that by assigning a higher weight to the small coefficients and lower weight to the large coefficients, it is possible to reduce the selection bias; therefore, it can consistently select the model. Furthermore, the adaptive LASSO solution is continuous from its definition, which enables it to enjoy oracle properties. In panelized Cox proportional hazards model, Zhang and Lu (2007) proposed the adaptive LASSO (ALASSO) as

$$\hat{\beta}_{Cox}^{ALASSO} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\left[\beta^T x_i - \log \sum_{j \in R_i} \exp(\beta^T x_j)\right] + \lambda \sum_{j=1}^{p} w_j \left|\beta_j\right|\right], \tag{6}$$

Where $w_j = (w_1, \dots, w_p)^T$ is $p \times 1$ data-driven weight vector. It depends on the root n -consistent initial values of $\hat{\beta}$ and $w_j = (|\hat{\beta}_j|)^{-\gamma}$, where γ is a positive constant. For the low dimensional data, initial values of $\hat{\beta}$ can be the unpenalized maximum partial likelihood estimator. While in the case of the high dimensional data, initial values of $\hat{\beta}$ can be the LASSO estimates.

3. The proposed weight

In the context of gene expression data problems, the goal of gene selection is to improve prediction performance, to provide faster and more cost-effective genes, and to achieve a better knowledge of the underlying problem. High dimensionality can negatively influence the performance of the Cox proportional hazards regression model by increasing the risk of overfitting and lengthening the computational time. Therefore, removing irrelevant and noisy genes from the original microarray gene expression data is essential for applying Cox proportional hazards regression model 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 PCPHM with LASSO can be applied directly to the high dimensional gene expression data, this method may select irrelevant genes because LASSO has the inconsistent property in gene selection. In other words, the estimates of the PCPHM with LASSO can be biased for large coefficients because larger coefficients will take larger penalties. Second, in PCPHM, 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 LASSO, where

$$w_j = \frac{1}{\hat{sd}_j}, \quad j = 1, 2, ..., p, \tag{7}$$

Where $\hat{sd}_j$ is the standard deviation for each gene. According to Eq. (7), 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 LASSO can reduce the inconsistent property in gene selection.

The detailed of the our proposed weight computation is described in as

Step 1: Find $w_j, \quad j = 1, 2, ..., \mathrm{p}$.

Step 2: Define $\tilde{\mathbf{x}}_i = w_j \mathbf{x}_i$

Step 3: Solve the PCPHM,

$$\hat{\beta}_{Cox}^{Proposed} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \left[\beta^T \tilde{x}_i - \log \sum_{j \in R_i} \exp(\beta^T \tilde{x}_j) \right] + \lambda \sum_{j=1}^{p} w_j \left| \beta_j \right| \right]. \tag{8}$$

4. Real application

To evaluate the performance of the proposed method, three real gene datasets were used. A brief introduction and summary of the used datasets are given in Table 1. The first dataset is the Diffuse large B-cell lymphoma dataset (DLBCL) (Rosenwald et al., 2002). There are 240 lymphoma patients' samples. Each patient's data consists 7399 gene expression measurements, and its survival time, including censored or not.

The second dataset is the Dutch breast cancer dataset (DBC) (van Houwelingen et al., 2006). In this dataset, there was 295 breast cancer patients' information collected in this dataset. Each patient's data consist 4919 gene expression measurements.

The third dataset is the Lung cancer dataset (LC) (Beer et al., 2002). This dataset contains 86 lung cancer patients' information including 7129 gene expression measurements, survival time and whether the survival time is censored.

Table 1: The Detail of the Used Three Real Microarray Datasets

Dataset	Sample	Gene	Censored
DLBCL	240	7399	102
DBC	295	4919	207
LC	86	7129	62

To demonstrate the usefulness of the proposed method, comparative experiments with the LASSO and ALASSO are conducted. To do so, each gene expression dataset is randomly partitioned into the training dataset and the test dataset, where 70% of the sample 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 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 method was fixed as $0 \le \lambda \le 50$. To assess how well the model predicts the outcome, the idea of time-dependent receiver-operator characteristics (ROC) curves for censored data and area under the curve (AUC) as our criteria. The real application results are summarized in Tables 2 – 4.

Table 2 shows the average results of different used methods applied to the three real datasets. It is obviously that the numbers of genes selected by LASSO are much more than those of the ALASSO and the proposed method. Among the other two methods, the proposed method selected the least subset of genes. For example, in LC dataset, the proposed method selected 20 gens out of 7129 genes comparing to 61 and 75 selected genes by ALASSO and LASSO, respectively.

Table 2: The Selected Genes Results

	LASSO	ALASSO	Proposed
DLBCL	122	94	55
DBC	87	44	34
LC	75	61	20

In order to test the prediction accuracy of the different used methods, their average values of AUC for both the training and testing dataset were given in Tables 3 and 4, respectively. In the observation of Table 3, in terms of AUC, the proposed method achieved a maximum accuracy of 95.5%, 96.1% and 97.4% for DLBCL, DBC, and LC datasets, respectively. Furthermore, it is clear from the results that the proposed method outperformed the ALASSO for all datasets. This improvement in AUC is mainly due to the proposed method ability in taking into account the new weight. Moreover, the proposed method improved the classification accuracy compared to LASSO. The improvements were 8.5%, 7.7%, and 7.3% for the DLBCL, DBC, and LC datasets, respectively.

Table 3: The AUC Results for the Training Dataset

	LASSO	ALASSO	Proposed
DLBCL	0.871	0.912	0.955
DBC	0.884	0.924	0.961
LC	0.901	0.937	0.974

It can also be seen from Table 4 that the proposed method has the best results in terms of the AUC for the testing dataset. The proposed method has the largest AUC of 93.4%, 94.7%, and 95.8% for the DLBCL, DBC, and LC datasets, respectively. This indicated that the proposed method significantly succeeded in identifying the patients who are in fact having the cancer with a probability of greater than 0.93.

Table 4: The AUC Results for the Testing Dataset

	LASSO	ALASSO	Proposed
DLBCL	0.854	0.905	0.934
DBC	0.814	0.911	0.947
LC	0.882	0.923	0.958

5. Conclusion

This paper presents an adaptive penalized Cox proportional hazards regression model by combining the Cox proportional hazards regression model with the weighted LASSO to identify the relevant genes in gene expression data. Our proposed method was experimentally tested and compared with other existing methods. The superior prediction performance of the proposed method was shown through the AUC. Meeting this criterion nominates the proposed method as a promising gene selection method.

References

[1] Algamal, Z. Y., & Lee, M. H. (2015a). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications, 42*(23), 9326–9332. https://doi.org/10.1016/j.eswa.2015.08.016.

[2] Algamal, Z. Y., & Lee, M. H. (2015b). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine, 67*, 136-145. https://doi.org/10.1016/j.compbiomed.2015.10.008.

[3] Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., . . . Thomas, D. G. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine, 8*(8), 816. https://doi.org/10.1038/nm733.

[4] Bradic, J., Fan, J., & Jiang, J. (2011). Regularization for Cox's Proportional Hazards Model with Np-Dimensionality. *Ann Stat, 39*(6), 3092-3120. https://doi.org/10.1214/11-AOS911.

[5] Cockeran, M., Meintanis, S. G., & Allison, J. S. (2019). Goodness-of-fit tests in the Cox proportional hazards model. *Communications in Statistics - Simulation and Computation*, 1-12. https://doi.org/10.1080/03610918.2019.1639738.

[6] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[7] Du, P., Ma, S., & Liang, H. (2010). Penalized Variable Selection Procedure for Cox Models with Semiparametric Relative Risk. *Ann Stat, 38*(4), 2092-2117. https://doi.org/10.1214/09-AOS780.

[8] Emura, T., Chen, Y. H., & Chen, H. Y. (2012). Survival prediction based on compound covariate under Cox proportional hazard models. *PLoS One, 7*(10), e47627. https://doi.org/10.1371/journal.pone.0047627.

[9] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348-1360. https://doi.org/10.1198/016214501753382273.

[10] Fu, Z., Parikh, C. R., & Zhou, B. (2017). Penalized variable selection in competing risks regression. *Lifetime Data Anal, 23*(3), 353-376. https://doi.org/10.1007/s10985-016-9362-3.

[11] Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biom J, 52*(1), 70-84. https://doi.org/10.1002/bimj.200900028.

[12] Gui, J., & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics, 21*(13), 3001-3008. https://doi.org/10.1093/bioinformatics/bti422.

[13] Hossain, S., & Ahmed, S. E. (2014). Penalized and Shrinkage Estimation in the Cox Proportional Hazards Model. *Communications in Statistics - Theory and Methods, 43*(5), 1026-1040. https://doi.org/10.1080/03610926.2013.826368.

[14] Hou, W., Song, L., Hou, X., & Wang, X. (2013). Penalized Empirical Likelihood via Bridge Estimator in Cox's Proportional Hazard Model. *Communications in Statistics - Theory and Methods, 43*(2), 426-440. https://doi.org/10.1080/03610926.2012.657325.

[15] Huang, H. H., & Liang, Y. (2018). Hybrid L1/2 +2 method for gene selection in the Cox proportional hazards model. *Comput Methods Programs Biomed, 164*, 65-73. https://doi.org/10.1016/j.cmpb.2018.06.004.

[16] Huang, J., Liu, L., Liu, Y., & Zhao, X. (2014). Group selection in the Cox model with a diverging number of covariates. *statistica Sinica*. https://doi.org/10.5705/ss.2013.061.

[17] Huang, J., Sun, T., Ying, Z., Yu, Y., & Zhang, C. H. (2013). Oracle Inequalities for the Lasso in the Cox Model. *Ann Stat, 41*(3), 1142-1165. https://doi.org/10.1214/13-AOS1098.

[18] Jiang, H. K., & Liang, Y. (2018). The L1/2 regularization network Cox model for analysis of genomic data. *Comput Biol Med, 100*, 203-208. https://doi.org/10.1016/j.compbiomed.2018.07.009.

[19] Karabey, U., & Tutkun, N. A. (2017). Model selection criterion in survival analysis. *1863*, 120003. https://doi.org/10.1063/1.4992296.

[20] Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis, 49*(1), 169-186. https://doi.org/10.1016/j.csda.2004.05.006.

[21] Leng, C., & Helen Zhang, H. (2006). Model selection in nonparametric hazard regression. *Journal of Nonparametric Statistics, 18*(7-8), 417-429. https://doi.org/10.1080/10485250601027042.

[22] Li, Y., Dicker, L., & Zhao, S. D. (2014). The Dantzig Selector for Censored Linear Regression Models. *Stat Sin, 24*(1), 251-2568. https://doi.org/10.5705/ss.2011.220.

[23] Lin, C. Y., & Halabi, S. (2017). A Simple Method for Deriving the Confidence Regions for the Penalized Cox's Model via the Minimand Perturbation. *Commun Stat Theory Methods, 46*(10), 4791-4808. https://doi.org/10.1080/03610926.2015.1085568.

[24] Liu, C., Liang, Y., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., & Zhang, H. (2014). The L1/2 regularization method for variable selection in the Cox model. *Applied Soft Computing, 14*, 498-503. https://doi.org/10.1016/j.asoc.2013.09.006.

[25] Park, E., & Ha, I. D. (2018). Penalized variable selection for accelerated failure time models. *Communications for Statistical Applications and Methods, 25*(6), 591-604. https://doi.org/10.29220/CSAM.2018.25.6.591.

[26] Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., . . . Giltnane, J. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine, 346*(25), 1937-1947. https://doi.org/10.1056/NEJMoa012914.

[27] Shi, Y., Xu, D., Cao, Y., & Jiao, Y. (2019). Variable Selection via Generalized SELO-Penalized Cox Regression Models. *Journal of Systems Science and Complexity, 32*(2), 709-736. https://doi.org/10.1007/s11424-018-7276-8.

[28] Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software, 39*(5), 1. https://doi.org/10.18637/jss.v039.i05.

[29] Suchting, R., Hebert, E. T., Ma, P., Kendzor, D. E., & Businelle, M. S. (2019). Using Elastic Net Penalized Cox Proportional Hazards Regression to Identify Predictors of Imminent Smoking Lapse. *Nicotine Tob Res, 21*(2), 173-179. https://doi.org/10.1093/ntr/ntx201.

[30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[31] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine, 16*(4), 385-395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

[32] van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J., & Wessels, L. F. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in medicine, 25*(18), 3201-3216. https://doi.org/10.1002/sim.2353.

[33] Wang, D., Wu, T. T., & Zhao, Y. (2019). Penalized empirical likelihood for the sparse Cox regression model. *Journal of Statistical Planning and Inference, 201*, 71-85. https://doi.org/10.1016/j.jspi.2018.12.001.

[34] Wu, T. T., Gong, H., & Clarke, E. M. (2012). A Transcriptome Analysis by Lasso Penalized Cox Regression for Pancreatic Cancer Survival. *Journal of Bioinformatics and Computational Biology, 09*(supp01), 63-73. https://doi.org/10.1142/S0219720011005744.

[35] Zhang, H. H., & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691-703. https://doi.org/10.1093/biomet/asm037.

[36] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association, 101*(476), 1418-1429. https://doi.org/10.1198/016214506000000735.

[37] Zou, H., & 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. https://doi.org/10.1111/j.1467-9868.2005.00503.x.